Model Selection

Divya Krishnan

December 14, 2015

Model Selection

```
# Standard libraries
library(RCurl)
library(leaps)
library(car)
library(randomForest)
library(pROC)
library(boot)
library(tree)
library(AER)
```

```
## Warning: package 'sandwich' was built under R version 3.2.3
```

```
library(bestglm)
# Setting seed
set.seed(1)
```

In this problem we will revisit the state dataset. This data, available as part of the base R package, contains various data related to the 50 states of the United States of America.

Suppose you want to explore the relationship between a state's Murder rate and other characteristics of the state, for example population, illiteracy rate, and more. Follow the questions below to perform this analysis.

(a) Examine the bivariate relationships present in the data. Briefly discuss notable results. You might find the scatterplotMatrix() function available in the car package helpful.

Since the state dataset has a lot of categorical variables, we choose to consider only state.x77 which has all numeric variables including the response variable, Murder. Also the variables in the state.x77 dataset seem to be more suited for predictor variables.

There seems to be positive bivariate relationship between -

(High-school graduates and income), (High-school graduates and life expectancy) and (Murder and illiteracy).

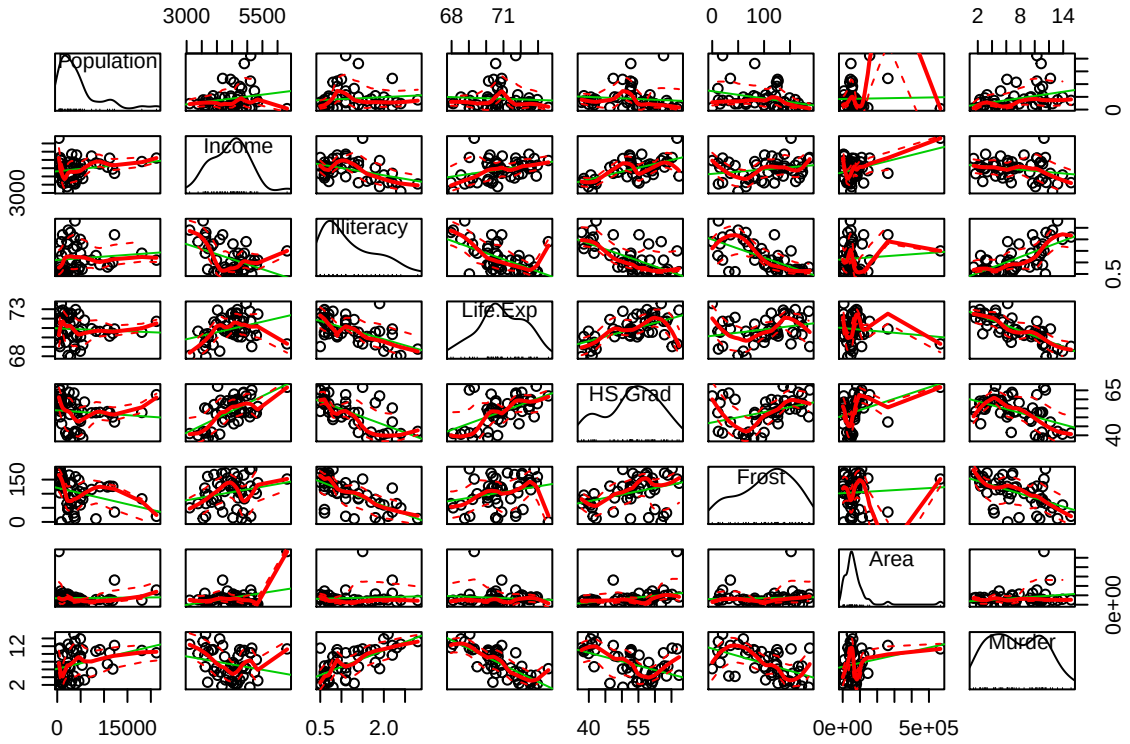
There seems to be a negative bivariate relationship between -

(Life expectancy and illiteracy), (High school graduates and illiteracy), (Frost and illiteracy), (Murder and life expectancy) and (Murder and Frost).

```
# Getting state dataset
data(state)
# Creating a data frame of the various variables
  # inside state data set including the matrix variable state.x77
state<-data.frame(cbind(abb=state.abb,area=state.area,
                  longitude=state.center$x,latitude=state.center$y,
                  division=state.division,name=state.name,
                  region=state.region),state.x77)
str(state)
```

```
## 'data.frame':    50 obs. of  15 variables:
##  $ abb       : Factor w/ 50 levels "AK","AL","AR",..: 2 1 4 3 5 6 7 8 9 10 ...
##  $ area      : Factor w/ 50 levels "104247","10577",..: 24 33 4 26 8 1 23 9 31 32 ...
##  $ longitude : Factor w/ 50 levels "-100.099","-105.513",..: 37 14 7 42 11 2 19 22 30 32 ...
##  $ latitude  : Factor w/ 44 levels "27.8744","30.6181",..: 6 44 9 11 15 20 28 20 1 5 ...
##  $ division  : Factor w/ 9 levels "1","2","3","4",..: 4 9 8 5 9 8 1 3 3 3 ...
##  $ name      : Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ region    : Factor w/ 4 levels "1","2","3","4": 2 4 4 2 4 4 1 2 2 2 ...
##  $ Population: num  3615 365 2212 2110 21198 ...
##  $ Income    : num  3624 6315 4530 3378 5114 ...
##  $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
##  $ Life.Exp  : num  69 69.3 70.5 70.7 71.7 ...
##  $ Murder    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
##  $ HS.Grad   : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
##  $ Frost     : num  20 152 15 65 20 166 139 103 11 60 ...
##  $ Area      : num  50708 566432 113417 51945 156361 ...
```

```
# Using only the state.x77 dataset and rearranging the columns
st<-cbind(state[,8:11],state[,13:15],state[12])
# Scatterplot of bivariate realtionships in state.x77
scatterplotMatrix(st)
```




```
# Correlation Matrix
cor(st)
```

```
##             Population     Income  Illiteracy    Life.Exp     HS.Grad
```

```
## Population  1.00000000  0.2082276  0.10762237 -0.06805195 -0.09848975
## Income      0.20822756  1.0000000 -0.43707519  0.34025534  0.61993232
## Illiteracy  0.10762237 -0.4370752  1.00000000 -0.58847793 -0.65718861
## Life.Exp   -0.06805195  0.3402553 -0.58847793  1.00000000  0.58221620
## HS.Grad    -0.09848975  0.6199323 -0.65718861  0.58221620  1.00000000
## Frost      -0.33215245  0.2262822 -0.67194697  0.26206801  0.36677970
## Area        0.02254384  0.3633154  0.07726113 -0.10733194  0.33354187
## Murder      0.34364275 -0.2300776  0.70297520 -0.78084575 -0.48797102
##                 Frost        Area     Murder
## Population -0.3321525  0.02254384  0.3436428
## Income      0.2262822  0.36331544 -0.2300776
## Illiteracy -0.6719470  0.07726113  0.7029752
## Life.Exp    0.2620680 -0.10733194 -0.7808458
## HS.Grad     0.3667797  0.33354187 -0.4879710
## Frost       1.0000000  0.05922910 -0.5388834
## Area        0.0592291  1.00000000  0.2283902
## Murder     -0.5388834  0.22839021  1.0000000
```

```
# Checking significant correlations
ifelse(cor(st)>=0.5,1,ifelse(cor(st)<=-0.5,-1,0))
```

```
##            Population Income Illiteracy Life.Exp HS.Grad Frost Area Murder
## Population          1      0          0        0       0     0    0      0
## Income              0      1          0        0       1     0    0      0
## Illiteracy          0      0          1       -1      -1    -1    0      1
## Life.Exp            0      0         -1        1       1     0    0     -1
## HS.Grad             0      1         -1        1       1     0    0      0
## Frost               0      0         -1        0       0     1    0     -1
## Area                0      0          0        0       0     0    1      0
## Murder              0      0          1       -1       0    -1    0      1
```

Multivariate Linear Regression Fit a multiple linear regression model. How much variance in the murder rate across states do the predictor variables explain?

The r-squared value is 0.8083 and adjusted r-squared value is 0.7763. Since it is multiple linear regression, we should consider adjusted r-squared value to account for model complexity. As per adjusted r-squared value, about 77.63% of variance in the dataset is explained by the predictor variables.

```
# Fitting multiple linear regression
fit.mlm<-lm(Murder ~ .,data=st)
# R-squared value
summary(fit.mlm)$r.squared
```

```
## [1] 0.8082607
```

```
# Adjusted r-squared value
summary(fit.mlm)$adj.r.squared
```

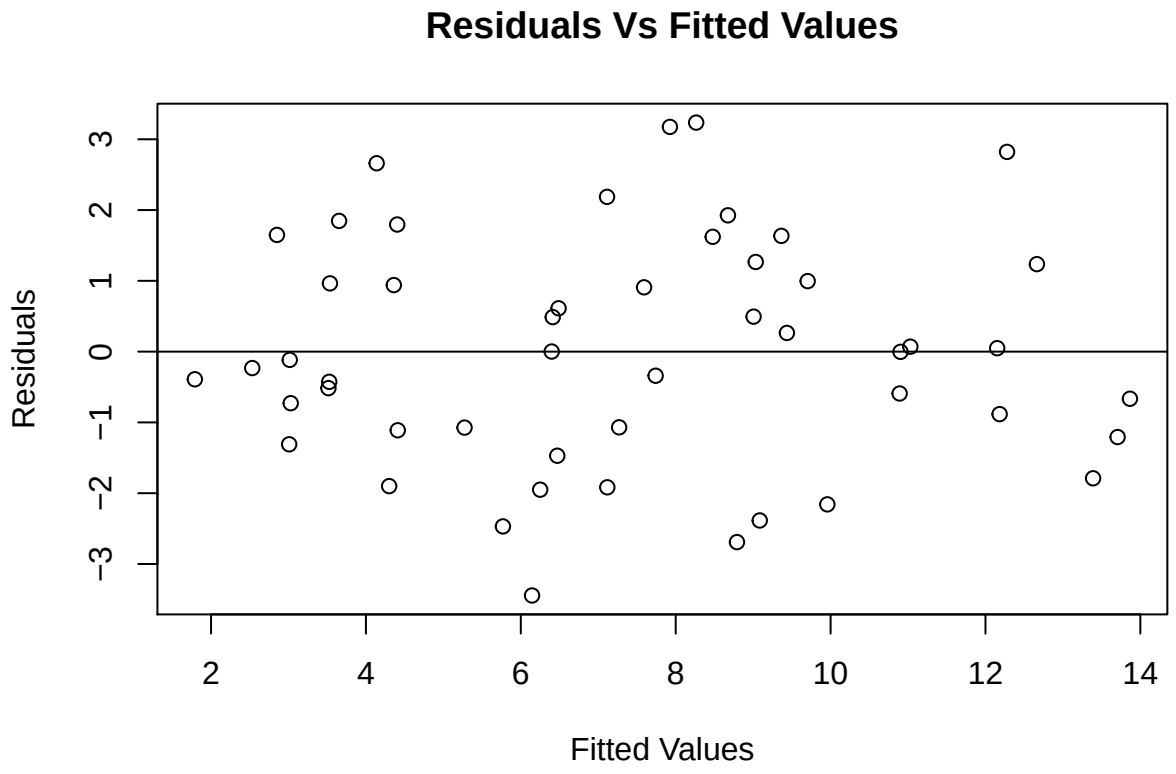
```
## [1] 0.7763042
```

Residual Analysis Evaluate the statistical assumptions in your regression analysis from part (b) by performing a basic analysis of model residuals and any unusual observations. Discuss any concerns you have about your model.

The plot of residuals vs fitted values seems normally distributed and doesn't show any patterns. Hence linear model seems appropriate. But the residual plot shows that there is lot of variance in the residuals.

The p-value suggests that only population and life expectancy have statistically significant association. Population seems to have a positive relationship with Murder but the coefficient estimate is very close to zero(0.000188) whereas life expectancy seems to have a negative relationship with the coefficient estimate of -1.655.

```
# Residual plot
plot(fit.mlm$residuals ~ fit.mlm$fitted.values,
     xlab="Fitted Values",ylab="Residuals",main="Residuals Vs Fitted Values")
abline(lm(fit.mlm$residuals ~ fit.mlm$fitted.values))
```




```
# Model summary
summary(fit.mlm)
```

```
##
## Call:
## lm(formula = Murder ~ ., data = st)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.4452 -1.1016 -0.0598  1.1758  3.2355
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.222e+02  1.789e+01   6.831 2.54e-08 ***
## Population   1.880e-04  6.474e-05   2.905  0.00584 **
## Income      -1.592e-04  5.725e-04  -0.278  0.78232
## Illiteracy   1.373e+00  8.322e-01   1.650  0.10641
## Life.Exp    -1.655e+00  2.562e-01  -6.459 8.68e-08 ***
## HS.Grad      3.234e-02  5.725e-02   0.565  0.57519
## Frost       -1.288e-02  7.392e-03  -1.743  0.08867 .
## Area         5.967e-06  3.801e-06   1.570  0.12391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.746 on 42 degrees of freedom
## Multiple R-squared:  0.8083, Adjusted R-squared:  0.7763
## F-statistic: 25.29 on 7 and 42 DF,  p-value: 3.872e-13
```

Stepwise Model Selection Use a stepwise model selection procedure of your choice to obtain a "best" fit model. Is the model different from the full model you fit in part (b)? If yes, how so?

The best fit model seems to be lm(Murder ~ Life.Exp + Frost + Population + Area + Illiteracy). The model is different from the full model as it has excluded percentage of high school graduates and income variable. In the best model, life expectancy, population and area have statistically significant association with the response variable. The adjusted r-squared value suggest that the best model explains 78.48% (slightly more than the full model) of the variance in the dataset.

```
# Null model
nullModel<-lm(Murder ~ 1,data=st)
# Full model
fullModel<-lm(Murder ~ .,data=st)
# Stepwise model selection
stStep<-step(nullModel,scope=list(lower=nullModel,upper=fullModel),direction="both")
```

```
## Start:  AIC=131.59
## Murder ~ 1
##
##              Df Sum of Sq    RSS     AIC
## + Life.Exp    1    407.14 260.61  86.550
## + Illiteracy  1    329.98 337.76  99.516
## + Frost       1    193.91 473.84 116.442
## + HS.Grad     1    159.00 508.75 119.996
## + Population  1     78.85 588.89 127.311
## + Income      1     35.35 632.40 130.875
## + Area        1     34.83 632.91 130.916
## <none>                    667.75 131.594
##
## Step:  AIC=86.55
## Murder ~ Life.Exp
##
##              Df Sum of Sq    RSS     AIC
## + Frost       1     80.10 180.50  70.187
```

```
## + Illiteracy  1     60.55 200.06  75.329
## + Population  1     56.62 203.99  76.303
## + Area        1     14.12 246.49  85.764
## <none>                    260.61  86.550
## + HS.Grad     1      1.12 259.48  88.334
## + Income      1      0.96 259.65  88.366
## - Life.Exp    1    407.14 667.75 131.594
##
## Step:  AIC=70.19
## Murder ~ Life.Exp + Frost
##
##              Df Sum of Sq    RSS     AIC
## + Population  1    23.710 156.79  65.146
## + Area        1    21.084 159.42  65.976
## <none>                    180.50  70.187
## + Illiteracy  1     6.066 174.44  70.477
## + Income      1     5.560 174.94  70.622
## + HS.Grad     1     2.068 178.44  71.610
## - Frost       1    80.104 260.61  86.550
## - Life.Exp    1   293.331 473.84 116.442
##
## Step:  AIC=65.15
## Murder ~ Life.Exp + Frost + Population
##
##              Df Sum of Sq    RSS     AIC
## + Area        1    19.040 137.75  60.672
## + Illiteracy  1    11.826 144.97  63.225
## <none>                    156.79  65.146
## + HS.Grad     1     1.821 154.97  66.561
## + Income      1     0.739 156.06  66.909
## - Population  1    23.710 180.50  70.187
## - Frost       1    47.198 203.99  76.303
## - Life.Exp    1   296.694 453.49 116.247
##
## Step:  AIC=60.67
## Murder ~ Life.Exp + Frost + Population + Area
##
##              Df Sum of Sq    RSS     AIC
## + Illiteracy  1     8.723 129.03  59.402
## <none>                    137.75  60.672
## + Income      1     1.241 136.51  62.220
## + HS.Grad     1     0.771 136.98  62.392
## - Area        1    19.040 156.79  65.146
## - Population  1    21.666 159.42  65.976
## - Frost       1    52.970 190.72  74.940
## - Life.Exp    1   272.927 410.68 113.290
##
## Step:  AIC=59.4
## Murder ~ Life.Exp + Frost + Population + Area + Illiteracy
##
##              Df Sum of Sq    RSS    AIC
## <none>                    129.03 59.402
## - Illiteracy  1     8.723 137.75 60.672
## + HS.Grad     1     0.763 128.27 61.105
```

```
## + Income      1     0.026 129.01 61.392
## - Frost       1    11.030 140.06 61.503
## - Area        1    15.937 144.97 63.225
## - Population  1    26.415 155.45 66.714
## - Life.Exp    1   140.391 269.42 94.213
```

```
# Summary of the best model
summary(stStep)
```

```
##
## Call:
## lm(formula = Murder ~ Life.Exp + Frost + Population + Area +
##     Illiteracy, data = st)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2976 -1.0711 -0.1123  1.1092  3.4671
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.202e+02  1.718e+01   6.994 1.17e-08 ***
## Life.Exp    -1.608e+00  2.324e-01  -6.919 1.50e-08 ***
## Frost       -1.373e-02  7.080e-03  -1.939  0.05888 .
## Population   1.780e-04  5.930e-05   3.001  0.00442 **
## Area         6.804e-06  2.919e-06   2.331  0.02439 *
## Illiteracy   1.173e+00  6.801e-01   1.725  0.09161 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.712 on 44 degrees of freedom
## Multiple R-squared:  0.8068, Adjusted R-squared:  0.7848
## F-statistic: 36.74 on 5 and 44 DF,  p-value: 1.221e-14
```

10-fold Cross Validation Assess the model (from part (d)) generalizability. Perform a 10-fold cross validation to estimate model performance. Report the results.

The 10-fold cross validation shows that the standard k-fold CV estimate of the model is 3.546 and the bias-corrected version is 3.484.

```
# Fitting the best model
glm.fit<-glm(Murder ~ Life.Exp + Frost + Population + Area + Illiteracy,
             data=st)
# Model summary
summary(glm.fit)
```

```
##
## Call:
## glm(formula = Murder ~ Life.Exp + Frost + Population + Area +
##     Illiteracy, data = st)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2976  -1.0711  -0.1123   1.1092   3.4671
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.202e+02  1.718e+01   6.994 1.17e-08 ***
## Life.Exp    -1.608e+00  2.324e-01  -6.919 1.50e-08 ***
## Frost       -1.373e-02  7.080e-03  -1.939  0.05888 .
## Population   1.780e-04  5.930e-05   3.001  0.00442 **
## Area         6.804e-06  2.919e-06   2.331  0.02439 *
## Illiteracy   1.173e+00  6.801e-01   1.725  0.09161 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.932536)
##
##     Null deviance: 667.75  on 49  degrees of freedom
## Residual deviance: 129.03  on 44  degrees of freedom
## AIC: 203.3
##
## Number of Fisher Scoring iterations: 2
```

```
# Compute k-fold CV estimate of the test MSE
cv.err.k10<-cv.glm(st, glm.fit, K=10)
# Delta vector containing cv results
cv.err.k10$delta
```

```
## [1] 3.842053 3.755144
```

Regression Tree Fit a regression tree using the same covariates in your "best" fit model from part (d). Use cross validation to select the "best" tree.

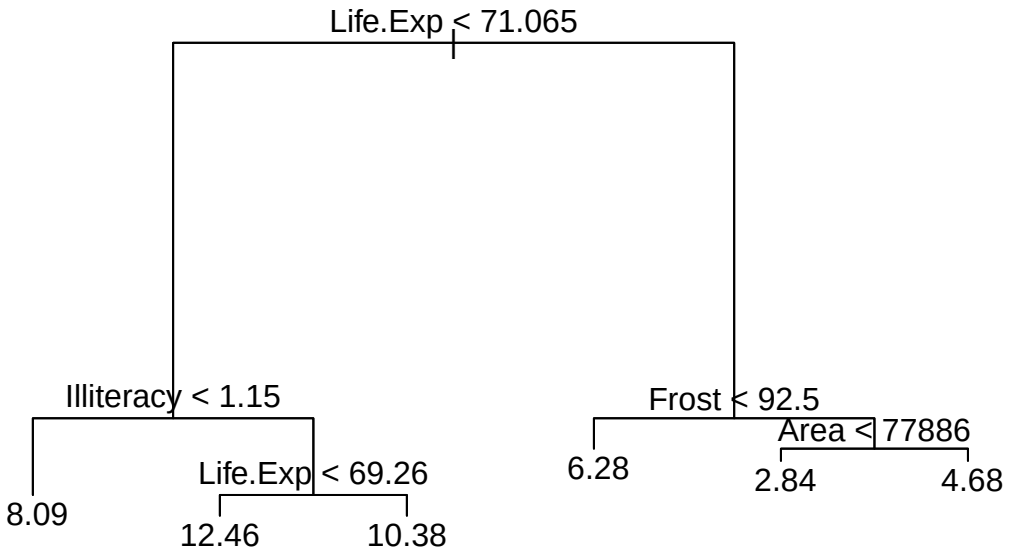
The regression tree used only 2 predictors to contsruct the tree - life exectancy and frost. The regression tree had a size of 6. But the cross validation revealed that the tree with size 5 has minimum deviance of 412.2660 (although the difference is only about 0.2, from tree of size 6). Hence, the tree was pruned to size 5 to obtain the best model. The best model also uses life expectancy and frost as predictors but has only 5 terminal nodes.

```
# Creating training set using 80% of the dataset
train<-sample(1:nrow(st),round(0.8*nrow(st),0))
# Regression tree creation
tree.st<-tree(Murder ~ Life.Exp + Frost + Population + Area +
                Illiteracy,data=st,subset=train)
# Tree summary
summary(tree.st)
```

```
##
## Regression tree:
## tree(formula = Murder ~ Life.Exp + Frost + Population + Area +
##     Illiteracy, data = st, subset = train)
## Variables actually used in tree construction:
## [1] "Life.Exp"   "Illiteracy" "Frost"      "Area"
## Number of terminal nodes:  6
## Residual mean deviance:  2.998 = 101.9 / 34
## Distribution of residuals:
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -3.090  -1.028   0.050   0.000   0.650   4.020
```

```
# Plotting the tree
plot(tree.st)
text(tree.st,pretty=0)
```




```
# Cross validation
cv.err.regression<-cv.tree(tree.st, FUN=prune.tree,K=10)
# Cross validation results
cv.err.regression
```

```
## $size
## [1] 6 5 4 3 2 1
##
## $dev
## [1] 322.8984 311.9054 309.0279 330.2934 387.2959 549.3553
##
## $k
## [1]      -Inf   8.46400  14.42133  21.16800  53.40167 261.36000
##
## $method
## [1] "deviance"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

```
# Plotting to understand the best tree
plot(cv.err.regression$size,cv.err.regression$dev,type='b',
     xlab="Size",ylab="Deviance",main="Plot of Deviance Vs Size")
```


```
# Pruning tree based on minimum deviance
prune.st<-prune.tree(tree.st,best=5)
# Plotting best tree based on minimum deviance
plot(prune.st)
text(prune.st,pretty=0)
```

Model Performance Comparison Compare the models from part (d) and (f) based on their performance. Which do you prefer? Be sure to justify your preference.

The test MSE for the best model according to part (d) is 3.5 (based on 10-fold cross validation). The test MSE for the regression tree according to part (f) is 5.56 and the test MSE for the pruned tree is 8.46. So neither the regression tree noe the pruned tree b ased on cross validation performs better than the best model as suggested by the step function. Hence, the model from part (d) is preferred as it has low test error.

```
# Performance of model from part (d)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = Murder ~ Life.Exp + Frost + Population + Area +
##     Illiteracy, data = st)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2976  -1.0711  -0.1123   1.1092   3.4671
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.202e+02  1.718e+01   6.994 1.17e-08 ***
## Life.Exp    -1.608e+00  2.324e-01  -6.919 1.50e-08 ***
## Frost       -1.373e-02  7.080e-03  -1.939  0.05888 .
```

```
## Population   1.780e-04  5.930e-05   3.001  0.00442 **
## Area         6.804e-06  2.919e-06   2.331  0.02439 *
## Illiteracy   1.173e+00  6.801e-01   1.725  0.09161 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 2.932536)
##
##     Null deviance: 667.75  on 49  degrees of freedom
## Residual deviance: 129.03  on 44  degrees of freedom
## AIC: 203.3
##
## Number of Fisher Scoring iterations: 2
```

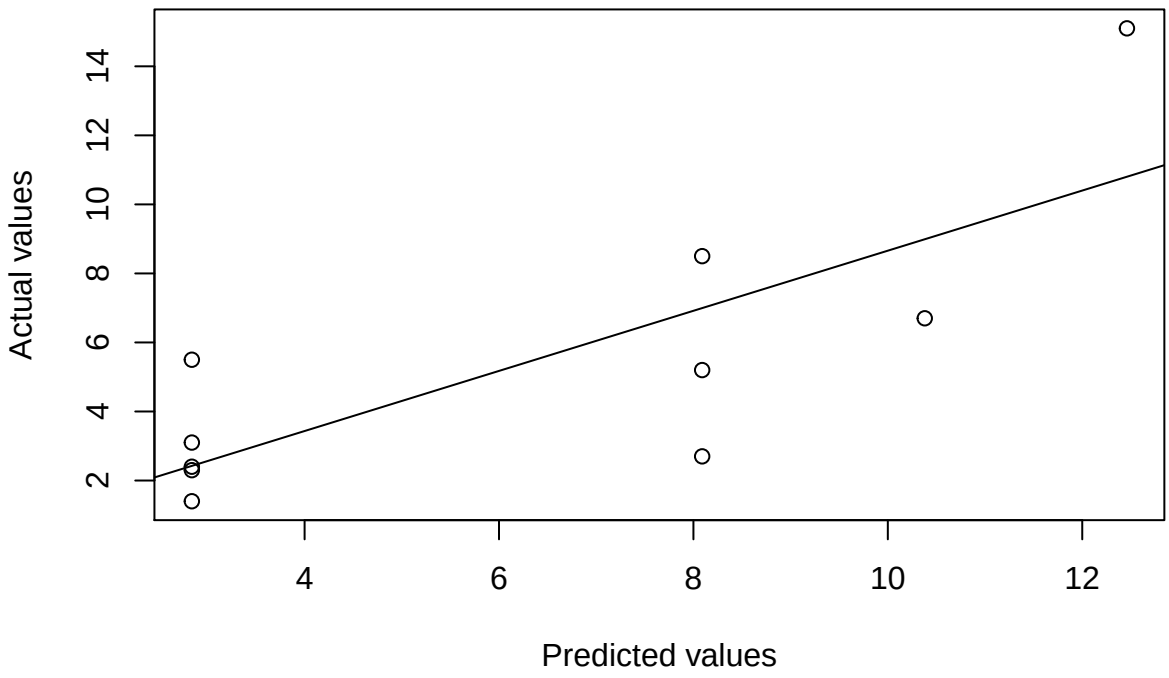
```
# Test MSE for the best model according to step function
cv.err.k10$delta
```

```
## [1] 3.842053 3.755144
```

```
# Test set
st.test<-st[-train,"Murder"]
# Performance of tree model from part (f)
summary(tree.st)
```

```
##
## Regression tree:
## tree(formula = Murder ~ Life.Exp + Frost + Population + Area +
##     Illiteracy, data = st, subset = train)
## Variables actually used in tree construction:
## [1] "Life.Exp"   "Illiteracy" "Frost"      "Area"
## Number of terminal nodes:  6
## Residual mean deviance:  2.998 = 101.9 / 34
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -3.090  -1.028   0.050   0.000   0.650   4.020
```

```
# Predicting the Murder rate
yhat.tree.st<-predict(tree.st,newdata=st[-train,])
# Plot of Actual values Vs Predicted values
plot(yhat.tree.st,st.test,xlab="Predicted values",ylab="Actual values")
abline(lm(st.test ~ yhat.tree.st))
```



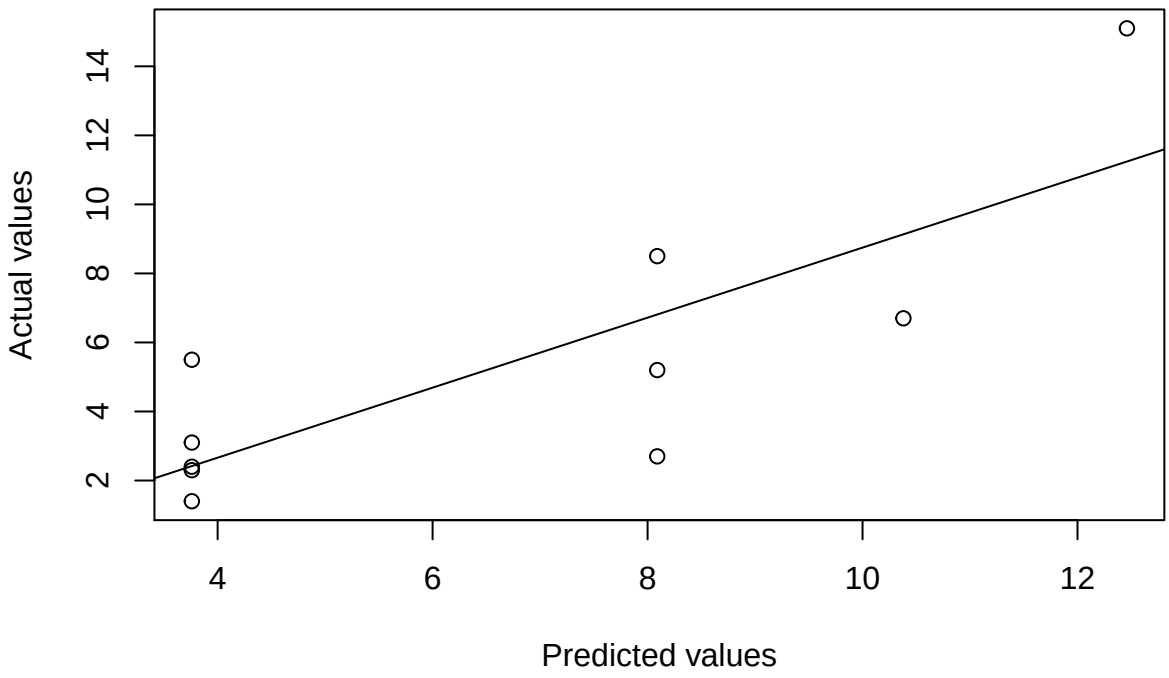

```
# Test MSE associated with the regression tree
mean((yhat.tree.st-st.test)^2)
```

```
## [1] 6.77863
```

```
# Performance of pruned tree from part (f)
summary(prune.st)
```

```
##
## Regression tree:
## snip.tree(tree = tree.st, nodes = 7L)
## Variables actually used in tree construction:
## [1] "Life.Exp"   "Illiteracy" "Frost"
## Number of terminal nodes:  5
## Residual mean deviance:  3.154 = 110.4 / 35
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -3.090  -1.040  -0.180   0.000   0.785   4.020
```

```
# Predicting the Murder rate
yhat.prune.st<-predict(prune.st,newdata=st[-train,])
# Plot of Actual values Vs Predicted values
plot(yhat.prune.st,st.test,xlab="Predicted values",ylab="Actual values")
abline(lm(st.test ~ yhat.prune.st))
```




```
# Test MSE associated with the pruned tree
mean((yhat.prune.st-st.test)^2)
```

```
## [1] 7.10983
```