R Notebook

## Question 11.1

Using the crime data set uscrime.txtfrom Questions 8.2, 9.1, and 10.1, build a regression model using

#### Stepwise regression

setwd("c:\\stuff")  
uscrime=read.table("uscrime.txt",header=T)  
  
current=names(uscrime)[-16]  
for (i in 1:12) {  
 form=as.formula(paste("Crime ~ ",paste(current, collapse="+")))  
 #print(form)  
 reg=lm(form,data=uscrime)  
 prt=summary(reg)$coefficients[,'Pr(>|t|)']  
 print(prt)  
 max=0.05  
 new\_i=-1  
 for (i in 2:length(prt)) {  
 print(paste(prt[i],max))  
 if (prt[i] > max) {  
 max=prt[i]  
 new\_i = i  
 }  
 }  
 if (new\_i != -1) {  
 #remove the factor  
 print("removing factor")  
 print(new\_i)  
 print(paste(paste(prt[new\_i],""),current[(new\_i-1)]))  
 current=current[-(new\_i-1)]  
 }  
}

## (Intercept) M So Ed Po1   
## 0.0008929887 0.0434433942 0.9797653725 0.0048614327 0.0788919769   
## Po2 LF M.F Pop NW   
## 0.3588295738 0.6546540941 0.3989953316 0.5738452309 0.5212791189   
## U1 U2 Wealth Ineq Prob   
## 0.1762380311 0.0501612829 0.3607537824 0.0039831365 0.0406269260   
## Time   
## 0.6307084351   
## [1] "0.0434433942179666 0.05"  
## [1] "0.979765372505204 0.05"  
## [1] "0.00486143272383178 0.979765372505204"  
## [1] "0.0788919768740791 0.979765372505204"  
## [1] "0.358829573837579 0.979765372505204"  
## [1] "0.654654094141773 0.979765372505204"  
## [1] "0.398995331608846 0.979765372505204"  
## [1] "0.573845230939972 0.979765372505204"  
## [1] "0.521279118878553 0.979765372505204"  
## [1] "0.176238031075354 0.979765372505204"  
## [1] "0.0501612829157054 0.979765372505204"  
## [1] "0.360753782383893 0.979765372505204"  
## [1] "0.00398313645843849 0.979765372505204"  
## [1] "0.0406269260020164 0.979765372505204"  
## [1] "0.63070843513499 0.979765372505204"  
## [1] "removing factor"  
## [1] 3  
## [1] "0.979765372505204 So"  
## (Intercept) M Ed Po1 Po2   
## 0.0007113266 0.0395203338 0.0041698787 0.0741517742 0.3508553433   
## LF M.F Pop NW U1   
## 0.6157358444 0.3878763591 0.5675732764 0.4800023797 0.1411060133   
## U2 Wealth Ineq Prob Time   
## 0.0420807933 0.3430412342 0.0018339190 0.0353441414 0.6245556161   
## [1] "0.0395203337591127 0.05"  
## [1] "0.00416987866960955 0.05"  
## [1] "0.0741517742323576 0.05"  
## [1] "0.35085534334732 0.0741517742323576"  
## [1] "0.615735844433309 0.35085534334732"  
## [1] "0.387876359086413 0.615735844433309"  
## [1] "0.5675732764493 0.615735844433309"  
## [1] "0.480002379704809 0.615735844433309"  
## [1] "0.14110601334741 0.615735844433309"  
## [1] "0.0420807933073864 0.615735844433309"  
## [1] "0.34304123416835 0.615735844433309"  
## [1] "0.00183391902970522 0.615735844433309"  
## [1] "0.03534414143995 0.615735844433309"  
## [1] "0.624555616050579 0.615735844433309"  
## [1] "removing factor"  
## [1] 15  
## [1] "0.624555616050579 Time"  
## (Intercept) M Ed Po1 Po2   
## 0.0001354104 0.0415288501 0.0034971802 0.0809142388 0.4090321567   
## LF M.F Pop NW U1   
## 0.6172149452 0.3222343448 0.4725633219 0.5449145118 0.1466431193   
## U2 Wealth Ineq Prob   
## 0.0425533253 0.3572368593 0.0014063961 0.0224087415   
## [1] "0.0415288501090542 0.05"  
## [1] "0.00349718022869499 0.05"  
## [1] "0.0809142388071344 0.05"  
## [1] "0.409032156702767 0.0809142388071344"  
## [1] "0.617214945239772 0.409032156702767"  
## [1] "0.322234344793523 0.617214945239772"  
## [1] "0.472563321859992 0.617214945239772"  
## [1] "0.544914511787899 0.617214945239772"  
## [1] "0.146643119348015 0.617214945239772"  
## [1] "0.0425533253170034 0.617214945239772"  
## [1] "0.357236859278827 0.617214945239772"  
## [1] "0.00140639608936938 0.617214945239772"  
## [1] "0.0224087415169915 0.617214945239772"  
## [1] "removing factor"  
## [1] 6  
## [1] "0.617214945239772 LF"  
## (Intercept) M Ed Po1 Po2 M.F   
## 0.000122424 0.026506163 0.003008320 0.089178459 0.471885878 0.385673681   
## Pop NW U1 U2 Wealth Ineq   
## 0.412946946 0.594678766 0.166335748 0.036995800 0.358685071 0.001255586   
## Prob   
## 0.023695708   
## [1] "0.0265061630975728 0.05"  
## [1] "0.00300832024547944 0.05"  
## [1] "0.0891784586988011 0.05"  
## [1] "0.471885877612328 0.0891784586988011"  
## [1] "0.385673681172337 0.471885877612328"  
## [1] "0.412946945739773 0.471885877612328"  
## [1] "0.594678765895897 0.471885877612328"  
## [1] "0.166335747544869 0.594678765895897"  
## [1] "0.0369958004170268 0.594678765895897"  
## [1] "0.358685071155953 0.594678765895897"  
## [1] "0.00125558595927249 0.594678765895897"  
## [1] "0.0236957082099952 0.594678765895897"  
## [1] "removing factor"  
## [1] 8  
## [1] "0.594678765895897 NW"  
## (Intercept) M Ed Po1 Po2   
## 9.682308e-05 8.413108e-03 3.018999e-03 9.283846e-02 5.204697e-01   
## M.F Pop U1 U2 Wealth   
## 4.166545e-01 4.253410e-01 1.546886e-01 3.054410e-02 3.784269e-01   
## Ineq Prob   
## 4.712554e-04 2.521143e-02   
## [1] "0.00841310825219112 0.05"  
## [1] "0.00301899862959163 0.05"  
## [1] "0.0928384603430016 0.05"  
## [1] "0.520469660971848 0.0928384603430016"  
## [1] "0.416654470248506 0.520469660971848"  
## [1] "0.425341020993339 0.520469660971848"  
## [1] "0.154688614131234 0.520469660971848"  
## [1] "0.0305441028905388 0.520469660971848"  
## [1] "0.378426892213736 0.520469660971848"  
## [1] "0.000471255395145949 0.520469660971848"  
## [1] "0.025211429118713 0.520469660971848"  
## [1] "removing factor"  
## [1] 5  
## [1] "0.520469660971848 Po2"  
## (Intercept) M Ed Po1 M.F   
## 4.278582e-05 8.053544e-03 3.192430e-03 8.427503e-06 3.378232e-01   
## Pop U1 U2 Wealth Ineq   
## 4.540482e-01 1.454417e-01 2.753934e-02 3.700439e-01 3.486622e-04   
## Prob   
## 2.341087e-02   
## [1] "0.0080535437415174 0.05"  
## [1] "0.00319242970232973 0.05"  
## [1] "8.42750348924233e-06 0.05"  
## [1] "0.337823205884842 0.05"  
## [1] "0.454048161002607 0.337823205884842"  
## [1] "0.145441713538961 0.454048161002607"  
## [1] "0.0275393382095845 0.454048161002607"  
## [1] "0.370043930440032 0.454048161002607"  
## [1] "0.000348662226474622 0.454048161002607"  
## [1] "0.0234108737278983 0.454048161002607"  
## [1] "removing factor"  
## [1] 6  
## [1] "0.454048161002607 Pop"  
## (Intercept) M Ed Po1 M.F   
## 5.181152e-06 6.491114e-03 3.079916e-03 3.987213e-06 1.472515e-01   
## U1 U2 Wealth Ineq Prob   
## 1.167448e-01 2.551218e-02 4.124562e-01 3.605724e-04 2.916452e-02   
## [1] "0.00649111419381439 0.05"  
## [1] "0.00307991608671434 0.05"  
## [1] "3.98721262208169e-06 0.05"  
## [1] "0.147251535007199 0.05"  
## [1] "0.116744758435378 0.147251535007199"  
## [1] "0.0255121821701065 0.147251535007199"  
## [1] "0.412456171096788 0.147251535007199"  
## [1] "0.000360572417973314 0.412456171096788"  
## [1] "0.0291645175624472 0.412456171096788"  
## [1] "removing factor"  
## [1] 8  
## [1] "0.412456171096788 Wealth"  
## (Intercept) M Ed Po1 M.F   
## 4.039544e-06 8.280837e-03 1.533513e-03 8.260573e-08 1.087397e-01   
## U1 U2 Ineq Prob   
## 7.621739e-02 1.371415e-02 8.633441e-05 1.505273e-02   
## [1] "0.00828083693120303 0.05"  
## [1] "0.00153351333307674 0.05"  
## [1] "8.26057277341362e-08 0.05"  
## [1] "0.108739697628897 0.05"  
## [1] "0.0762173921579561 0.108739697628897"  
## [1] "0.0137141489636276 0.108739697628897"  
## [1] "8.63344065925845e-05 0.108739697628897"  
## [1] "0.0150527348181969 0.108739697628897"  
## [1] "removing factor"  
## [1] 5  
## [1] "0.108739697628897 M.F"  
## (Intercept) M Ed Po1 U1   
## 1.434953e-06 2.596534e-03 5.622597e-05 4.913267e-08 2.482240e-01   
## U2 Ineq Prob   
## 3.310219e-02 2.608177e-05 1.883455e-02   
## [1] "0.00259653364091096 0.05"  
## [1] "5.62259671804597e-05 0.05"  
## [1] "4.91326689266037e-08 0.05"  
## [1] "0.248224023870202 0.05"  
## [1] "0.0331021939679553 0.248224023870202"  
## [1] "2.60817654070251e-05 0.248224023870202"  
## [1] "0.0188345465733403 0.248224023870202"  
## [1] "removing factor"  
## [1] 5  
## [1] "0.248224023870202 U1"  
## (Intercept) M Ed Po1 U2   
## 1.715273e-06 3.054440e-03 8.072016e-05 2.561505e-10 3.483130e-02   
## Ineq Prob   
## 1.879377e-05 1.711387e-02   
## [1] "0.00305444025385066 0.05"  
## [1] "8.07201602423655e-05 0.05"  
## [1] "2.56150546826762e-10 0.05"  
## [1] "0.0348312990067525 0.05"  
## [1] "1.87937651482061e-05 0.05"  
## [1] "0.0171138690389445 0.05"  
## (Intercept) M Ed Po1 U2   
## 1.715273e-06 3.054440e-03 8.072016e-05 2.561505e-10 3.483130e-02   
## Ineq Prob   
## 1.879377e-05 1.711387e-02   
## [1] "0.00305444025385066 0.05"  
## [1] "8.07201602423655e-05 0.05"  
## [1] "2.56150546826762e-10 0.05"  
## [1] "0.0348312990067525 0.05"  
## [1] "1.87937651482061e-05 0.05"  
## [1] "0.0171138690389445 0.05"  
## (Intercept) M Ed Po1 U2   
## 1.715273e-06 3.054440e-03 8.072016e-05 2.561505e-10 3.483130e-02   
## Ineq Prob   
## 1.879377e-05 1.711387e-02   
## [1] "0.00305444025385066 0.05"  
## [1] "8.07201602423655e-05 0.05"  
## [1] "2.56150546826762e-10 0.05"  
## [1] "0.0348312990067525 0.05"  
## [1] "1.87937651482061e-05 0.05"  
## [1] "0.0171138690389445 0.05"

print("FINAL REGRESSION")

## [1] "FINAL REGRESSION"

summary(reg)

##   
## Call:  
## lm(formula = form, data = uscrime)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -470.68 -78.41 -19.68 133.12 556.23   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5040.50 899.84 -5.602 1.72e-06 \*\*\*  
## M 105.02 33.30 3.154 0.00305 \*\*   
## Ed 196.47 44.75 4.390 8.07e-05 \*\*\*  
## Po1 115.02 13.75 8.363 2.56e-10 \*\*\*  
## U2 89.37 40.91 2.185 0.03483 \*   
## Ineq 67.65 13.94 4.855 1.88e-05 \*\*\*  
## Prob -3801.84 1528.10 -2.488 0.01711 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 200.7 on 40 degrees of freedom  
## Multiple R-squared: 0.7659, Adjusted R-squared: 0.7307   
## F-statistic: 21.81 on 6 and 40 DF, p-value: 3.418e-11

## Question 11.1 (cont.)

Using the crime data set uscrime.txtfrom Questions 8.2, 9.1, and 10.1, build a regression model using:1.Stepwise regression2.Lasso3.Elastic netFor Parts 2 and 3, remember to scale the data first –otherwise, the regression coefficients will be on different scales andthe constraint won’t have the desired effect.

Reference::: <https://web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf>

We do LASSO first this is alpha=1

If you use glmnet, the scaling is performed by the package. You don’t need to worry about scaling the test set because the “coefficients are always returned on the original scale”.

For this not doing a training and test set

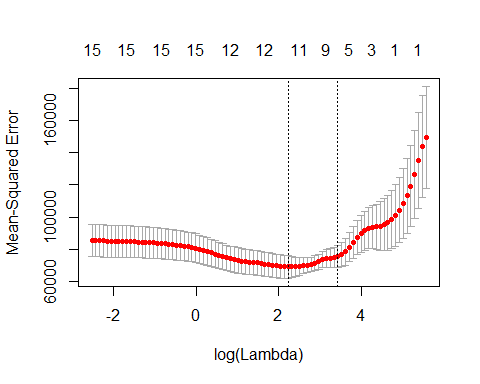
rm(list=ls())  
set.seed(1)  
setwd("c:\\stuff")  
uscrime=read.table("uscrime.txt",header=T)  
  
#uscrime=scale(uscrime,center=FALSE)  
indep=as.matrix(uscrime[,1:15])  
depend=as.matrix(uscrime[,16])  
library(glmnet)

## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-18

fit=glmnet(indep,depend,family="gaussian",alpha=1)  
  
#run cross validation to choose a model  
cvfit=cv.glmnet(indep,depend,family="gaussian",alpha=1)  
#plot the values and see the selected lambdas  
plot(cvfit)

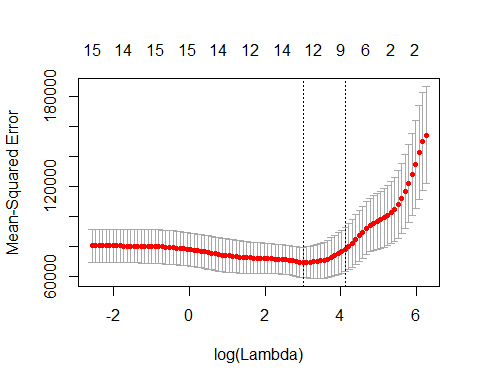


#get the coefficients for the cross validated model  
coef(cvfit, s = "lambda.min")

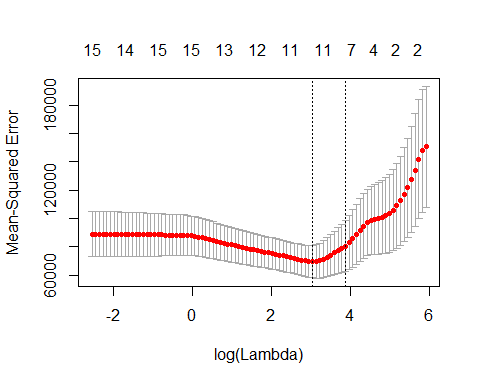
## 16 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) -5.000723e+03  
## M 7.108443e+01  
## So 4.447322e+01  
## Ed 1.234084e+02  
## Po1 1.027122e+02  
## Po2 .   
## LF .   
## M.F 1.871453e+01  
## Pop .   
## NW 6.010913e-01  
## U1 -2.014423e+03  
## U2 8.512460e+01  
## Wealth 4.980586e-03  
## Ineq 4.809861e+01  
## Prob -3.675090e+03  
## Time .

Next we do Elastic net with an alpha of 0.2, 0.5, 0.7

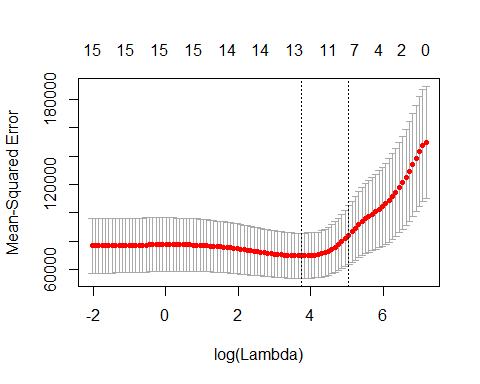
cvfit=cv.glmnet(indep,depend,family="gaussian",alpha=.5)  
plot(cvfit)



cvfit=cv.glmnet(indep,depend,family="gaussian",alpha=.7)  
plot(cvfit)



cvfit=cv.glmnet(indep,depend,family="gaussian",alpha=.2)  
plot(cvfit)



coef(cvfit, s = "lambda.min")

## 16 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) -4.323628e+03  
## M 5.609156e+01  
## So 7.038283e+01  
## Ed 7.603884e+01  
## Po1 5.425561e+01  
## Po2 3.883364e+01  
## LF 4.404154e+02  
## M.F 2.027080e+01  
## Pop .   
## NW 2.434964e+00  
## U1 -1.388974e+03  
## U2 6.769596e+01  
## Wealth 1.197364e-02  
## Ineq 3.082774e+01  
## Prob -3.639193e+03  
## Time .

0.2 looks the best with the mean squared error around the lambda

coef(cvfit, s = "lambda.min")

## 16 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) -4.323628e+03  
## M 5.609156e+01  
## So 7.038283e+01  
## Ed 7.603884e+01  
## Po1 5.425561e+01  
## Po2 3.883364e+01  
## LF 4.404154e+02  
## M.F 2.027080e+01  
## Pop .   
## NW 2.434964e+00  
## U1 -1.388974e+03  
## U2 6.769596e+01  
## Wealth 1.197364e-02  
## Ineq 3.082774e+01  
## Prob -3.639193e+03  
## Time .