HW5

Q8.1

I work at a packaging automation company that designs, builds, and tests machines. We can use various setup factors on the machines to predict expected reject rate during production. Various settings can be adjusted to see how it affects the reject rate and therefore the throughput.

1. Sensor window opening size in mm
2. Delay distance set in the motion control platform
3. Speed
4. Amplitude setting on magnetic stripe encoder

Q8.2

Using all predictors we get an adjusted r-squared value of 0.7078, which predicts 155 as Crime which does not seem reasonable.

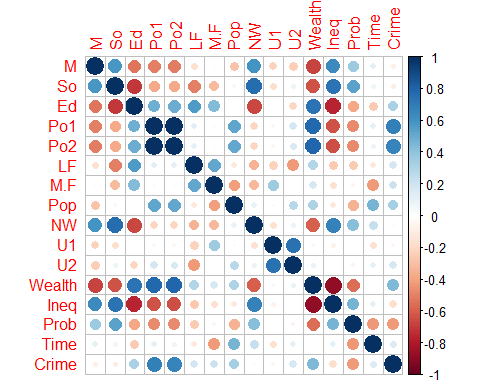
Running the model again with predictors with p-value 0.05 or less, and using some trial and error by including and excluding values close to 0.05, we get adjusted r-squared of 0.7307 using the following formula: Crime~M+Ed+U2+Ineq+Prob+Po1.

Looking at the cross-validated results, we get a much bigger drop off using all 15 attributes, but a much lower drop-off with the 6 attributes mentioned above. These 6 attributes predict a crime rate of 1304.

data <- read.table('uscrime.txt', header = TRUE)  
  
library(corrplot)

## corrplot 0.92 loaded

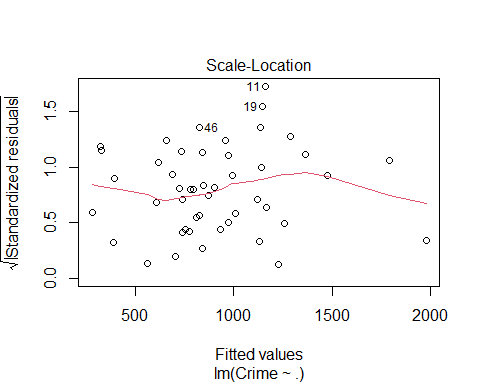
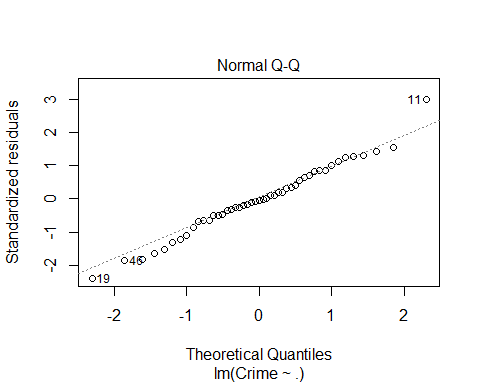
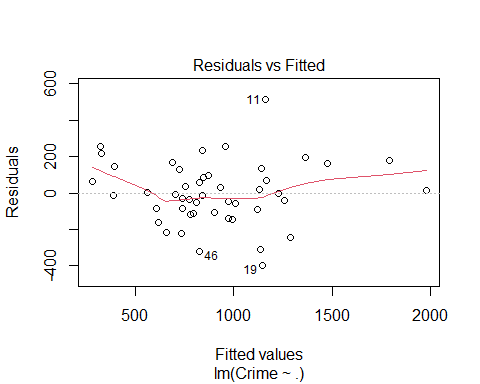
corrplot(cor(data))



lm1 <- lm(Crime~., data = data)  
summary(lm1)

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

plot(lm1)

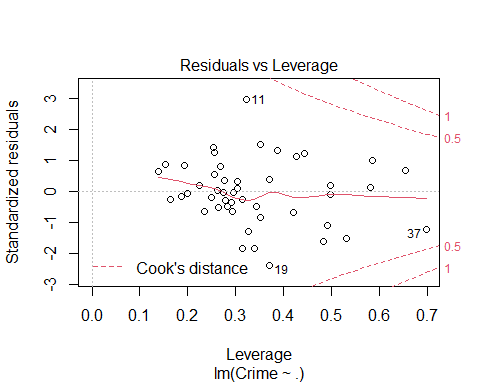


test\_pt <-data.frame(M = 14.0,So = 0, Ed = 10.0, Po1 = 12.0, Po2 = 15.5,LF = 0.640, M.F = 94.0, Pop = 150, NW = 1.1, U1 = 0.120, U2 = 3.6, Wealth = 3200, Ineq = 20.1, Prob = 0.04,Time = 39.0)  
  
result1 <- predict(lm1,test\_pt)  
  
#keep only the attributes with p-value 0.05 or less  
lm2 <- lm(Crime~M+Ed+U2+Ineq+Prob+Po1, data=data)  
summary(lm2)

##   
## Call:  
## lm(formula = Crime ~ M + Ed + U2 + Ineq + Prob + Po1, data = data)  
##   
## 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 \*\*\*  
## 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 \*   
## Po1 115.02 13.75 8.363 2.56e-10 \*\*\*  
## ---  
## 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

result2 <- predict(lm2,test\_pt)  
  
#check model quality with cross validation  
library("DAAG")

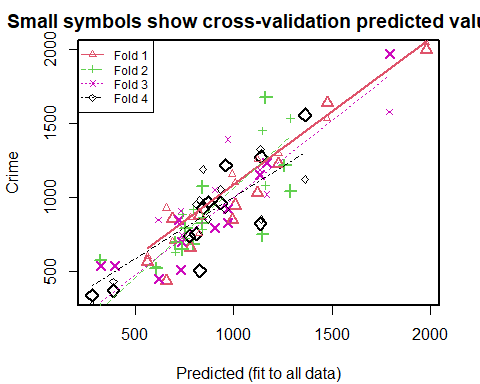
## Loading required package: lattice



set.seed(12)  
cv1 <- cv.lm(data,lm1,m=4)

## Analysis of Variance Table  
##   
## Response: Crime  
## Df Sum Sq Mean Sq F value Pr(>F)   
## M 1 55084 55084 1.26 0.2702   
## So 1 15370 15370 0.35 0.5575   
## Ed 1 905668 905668 20.72 7.7e-05 \*\*\*  
## Po1 1 3076033 3076033 70.38 1.8e-09 \*\*\*  
## Po2 1 153024 153024 3.50 0.0708 .   
## LF 1 61134 61134 1.40 0.2459   
## M.F 1 111000 111000 2.54 0.1212   
## Pop 1 42649 42649 0.98 0.3309   
## NW 1 14197 14197 0.32 0.5728   
## U1 1 7065 7065 0.16 0.6904   
## U2 1 269663 269663 6.17 0.0186 \*   
## Wealth 1 34748 34748 0.79 0.3795   
## Ineq 1 547423 547423 12.52 0.0013 \*\*   
## Prob 1 222620 222620 5.09 0.0312 \*   
## Time 1 10304 10304 0.24 0.6307   
## Residuals 31 1354946 43708   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Warning in cv.lm(data, lm1, m = 4):   
##   
## As there is >1 explanatory variable, cross-validation  
## predicted values for a fold are not a linear function  
## of corresponding overall predicted values. Lines that  
## are shown for the different folds are approximate



##   
## fold 1   
## Observations in test set: 11   
## 2 9 14 16 20 22 26 38 41 44 47  
## Predicted 1474 689 780 1006 1227.8 657 1977.4 562.7 824 1121 992  
## cvpred 1535 706 867 1100 1298.9 931 2043.3 602.8 757 1257 1159  
## Crime 1635 856 664 946 1225.0 439 1993.0 566.0 880 1030 849  
## CV residual 100 150 -203 -154 -73.9 -492 -50.3 -36.8 123 -227 -310  
##   
## Sum of squares = 512057 Mean square = 46551 n = 11   
##   
## fold 2   
## Observations in test set: 12   
## 1 3 6 11 19 25 28 29 30 33 35 39  
## Predicted 755.0 322 793 1161 1146 605.9 1258.48 1287 703 841 738 839.3  
## cvpred 727.7 265 920 1082 1449 535.1 1219.78 1534 634 784 886 868.7  
## Crime 791.0 578 682 1674 750 523.0 1216.00 1043 696 1072 653 826.0  
## CV residual 63.3 313 -238 592 -699 -12.1 -3.78 -491 62 288 -233 -42.7  
##   
## Sum of squares = 1382466 Mean square = 115205 n = 12   
##   
## fold 3   
## Observations in test set: 12   
## 4 5 10 12 13 15 17 34 37 40 42 45  
## Predicted 1791 1167 736.5 722 733 903 393 971.5 971 1131.5 326.3 617  
## cvpred 1576 1021 745.1 824 912 1050 103 823.4 1392 1186.8 -85.5 848  
## Crime 1969 1234 705.0 849 511 798 539 923.0 831 1151.0 542.0 455  
## CV residual 393 213 -40.1 25 -401 -252 436 99.6 -561 -35.8 627.5 -393  
##   
## Sum of squares = 1491541 Mean square = 124295 n = 12   
##   
## fold 4   
## Observations in test set: 12   
## 7 8 18 21 23 24 27 31 32 36 43 46  
## Predicted 934.2 1362 844 774.9 958 869 279.5 388.0 808 1138 1134 827  
## cvpred 1055.1 1123 1189 725.3 922 851 272.7 433.1 953 852 1324 984  
## Crime 963.0 1555 929 742.0 1216 968 342.0 373.0 754 1272 823 508  
## CV residual -92.1 432 -260 16.7 294 117 69.3 -60.1 -199 420 -501 -476  
##   
## Sum of squares = 1065774 Mean square = 88814 n = 12   
##   
## Overall (Sum over all 12 folds)   
## ms   
## 94720

#calculate r-squared using r^2 = = 1 - SSEresidual/SSEtotal  
SSEtotal <- sum((data$Crime - mean(data$Crime))^2)  
  
SSEres1 <- attr(cv1,"ms")\*nrow(data)  
  
r\_lm2\_cv <- 1- SSEres1/SSEtotal  
r\_lm2\_cv

## [1] 0.353