Week 5 Assignment - Simple Regression

Omer Farooq (EDx ID: mfarooq4)

02/11/2020

Table of Contents

QUESTION 8.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use.

At my job at the T-Mobile HQ in the Seattle area, my team helps get analytics products built for our network supply chain team. This team manages the planning, procurement and logistics of getting the right equipment to the right locations so that T-Mobile's network could get built or improved. The equipment used to build or enhance the cellular network includes items like radios, antenna, cables, etc. Some of these items are very expensive with very long lead times (time from order to delivery) of several months. Another interesting fact is that the technology changes pretty rapidly and we are constantly enhancing the existing network with new radio, antenna or a receiver. Not knowing the price of the newly released item poses a risk that an unexpected high or low price could affect the budget. Similarly, an unexepctedly longer lead time of a new item could affect the supply chain planning.

A regression model to predict the price and lead time of a new equipment item would be very helpful. We could train the model on past data using predictors like specifications of the equipment (frequency, range, etc.), quantity consumed, material class of the equipment (a classification hierarchy in our data for each type of equipment), dimensions, weight, vendor, etc.

QUESTION 8.2

Using crime data from http://www.statsci.org/data/general/uscrime.txt (file uscrime.txt,description at http://www.statsci.org/data/general/uscrime.html), use

regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data:

- 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

Show your model (factors used and their coefficients), the software output, and the quality of fit.

Note that because there are only 47 data points and 15 predictors, you'll probably notice some overfitting. We'll see ways of dealing with this sort of problem later in the course.

Loaded all needed library.

```
library(corrplot) #for correlation plot
library (caret) #for cross-validation
library(MASS) #for stepwise regression
```

Next, I loaded the Crimes data and printed a sample and summary of the data. The summary of Crime column is to be noted. Min is 342 and max is 1993 with median 831 and mean 905. This tells us the acceptable range of our predicted value for the given parameters given all provided parameters are within the ranges of available data.

```
#setting the seed so that results are the same at every run
set.seed(101)

#loading data
crimedata <- read.delim("data_5.1/uscrime.txt")

#quick glance at the data
head(crimedata)

##      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq     Prob
## 1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1 0.084602
## 2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4 0.029599
## 3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0 0.083401
## 4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7 0.015801
## 5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4 0.041399
## 6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6 0.034201
##      Time Crime
## 1 26.2011   791
## 2 25.2999  1635
## 3 24.3006   578
## 4 29.9012  1969
## 5 21.2998  1234
## 6 20.9995   682
```

```
#basic stats of the temps data
summary(crimedata)
```

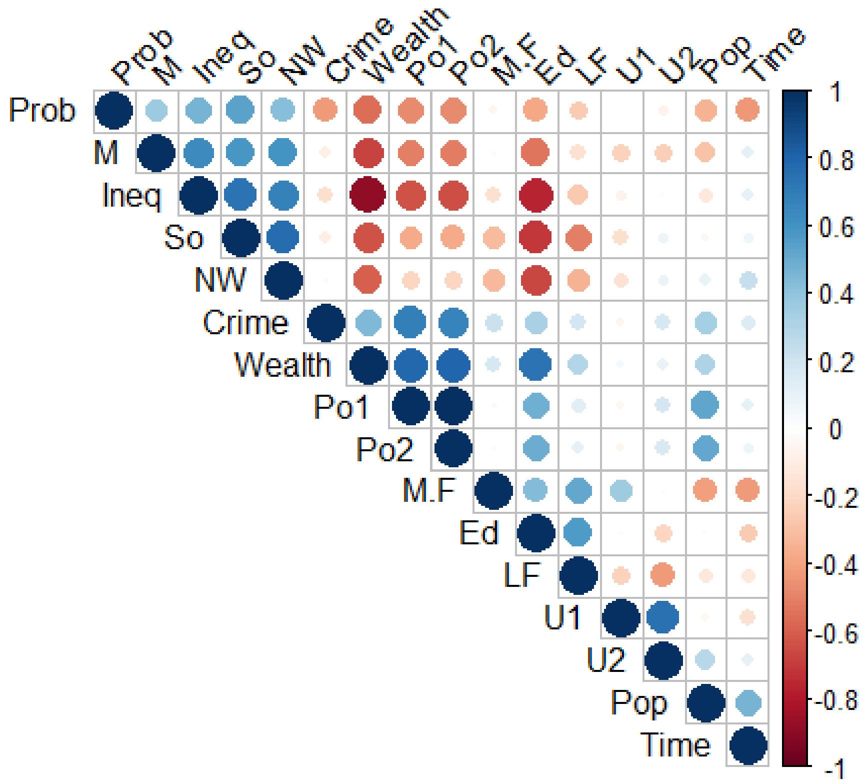
```
##        M               So               Ed              Po1
##  Min.   :11.90   Min.   :0.0000   Min.   : 8.70   Min.   : 4.50
##  1st Qu.:13.00   1st Qu.:0.0000   1st Qu.: 9.75   1st Qu.: 6.25
##  Median :13.60   Median :0.0000   Median :10.80   Median : 7.80
##  Mean   :13.86   Mean   :0.3404   Mean   :10.56   Mean   : 8.50
##  3rd Qu.:14.60   3rd Qu.:1.0000   3rd Qu.:11.45   3rd Qu.:10.45
##  Max.   :17.70   Max.   :1.0000   Max.   :12.20   Max.   :16.60
##       Po2              LF              M.F              Pop
##  Min.   : 4.100   Min.   :0.4800   Min.   : 93.40   Min.   :  3.00
##  1st Qu.: 5.850   1st Qu.:0.5305   1st Qu.: 96.45   1st Qu.: 10.00
##  Median : 7.300   Median :0.5600   Median : 97.70   Median : 25.00
##  Mean   : 8.023   Mean   :0.5612   Mean   : 98.30   Mean   : 36.62
##  3rd Qu.: 9.700   3rd Qu.:0.5930   3rd Qu.: 99.20   3rd Qu.: 41.50
##  Max.   :15.700   Max.   :0.6410   Max.   :107.10   Max.   :168.00
##        NW              U1                U2            Wealth
##  Min.   : 0.20   Min.   :0.07000   Min.   :2.000   Min.   :2880
##  1st Qu.: 2.40   1st Qu.:0.08050   1st Qu.:2.750   1st Qu.:4595
##  Median : 7.60   Median :0.09200   Median :3.400   Median :5370
##  Mean   :10.11   Mean   :0.09547   Mean   :3.398   Mean   :5254
##  3rd Qu.:13.25   3rd Qu.:0.10400   3rd Qu.:3.850   3rd Qu.:5915
##  Max.   :42.30   Max.   :0.14200   Max.   :5.800   Max.   :6890
##       Ineq            Prob              Time           Crime
##  Min.   :12.60   Min.   :0.00690   Min.   :12.20   Min.   : 342.0
##  1st Qu.:16.55   1st Qu.:0.03270   1st Qu.:21.60   1st Qu.: 658.5
##  Median :17.60   Median :0.04210   Median :25.80   Median : 831.0
##  Mean   :19.40   Mean   :0.04709   Mean   :26.60   Mean   : 905.1
##  3rd Qu.:22.75   3rd Qu.:0.05445   3rd Qu.:30.45   3rd Qu.:1057.5
##  Max.   :27.60   Max.   :0.11980   Max.   :44.00   Max.   :1993.0
```

I built a dataframe of the provided predictors to use in the models later on.

```
#data frame with data we need to predict crime for
predictdata <-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 = 0.040,Time = 39.0)
```

Before I jumped into models, I checked the pearson correlation matrix of the Crimes data. Value is 1 and -1 indicate positive and negative correlation where 0 indicates no correlation. The last column was of interest where correlation of Crime column with each predictor was given. Po1, Po2, Wealth and Prob showed some correlation to Crime column. The models built below tested these correlations further.

```
#pearson correlation matrix
corrmat <- cor(crimedata)
round(corrmat, 2)

##            M    So    Ed   Po1   Po2    LF   M.F   Pop    NW    U1    U2 Wealth
## M       1.00  0.58 -0.53 -0.51 -0.51 -0.16 -0.03 -0.28  0.59 -0.22 -0.24  -0.67
## So      0.58  1.00 -0.70 -0.37 -0.38 -0.51 -0.31 -0.05  0.77 -0.17  0.07  -0.64
## Ed     -0.53 -0.70  1.00  0.48  0.50  0.56  0.44 -0.02 -0.66  0.02 -0.22   0.74
## Po1    -0.51 -0.37  0.48  1.00  0.99  0.12  0.03  0.53 -0.21 -0.04  0.19   0.79
## Po2    -0.51 -0.38  0.50  0.99  1.00  0.11  0.02  0.51 -0.22 -0.05  0.17   0.79
## LF     -0.16 -0.51  0.56  0.12  0.11  1.00  0.51 -0.12 -0.34 -0.23 -0.42   0.29
## M.F    -0.03 -0.31  0.44  0.03  0.02  0.51  1.00 -0.41 -0.33  0.35 -0.02   0.18
## Pop    -0.28 -0.05 -0.02  0.53  0.51 -0.12 -0.41  1.00  0.10 -0.04  0.27   0.31
## NW      0.59  0.77 -0.66 -0.21 -0.22 -0.34 -0.33  0.10  1.00 -0.16  0.08  -0.59
## U1     -0.22 -0.17  0.02 -0.04 -0.05 -0.23  0.35 -0.04 -0.16  1.00  0.75   0.04
## U2     -0.24  0.07 -0.22  0.19  0.17 -0.42 -0.02  0.27  0.08  0.75  1.00   0.09
## Wealth -0.67 -0.64  0.74  0.79  0.79  0.29  0.18  0.31 -0.59  0.04  0.09   1.00
## Ineq    0.64  0.74 -0.77 -0.63 -0.65 -0.27 -0.17 -0.13  0.68 -0.06  0.02  -0.88
## Prob    0.36  0.53 -0.39 -0.47 -0.47 -0.25 -0.05 -0.35  0.43 -0.01 -0.06  -0.56
## Time    0.11  0.07 -0.25  0.10  0.08 -0.12 -0.43  0.46  0.23 -0.17  0.10   0.00
## Crime  -0.09 -0.09  0.32  0.69  0.67  0.19  0.21  0.34  0.03 -0.05  0.18   0.44


##         Ineq  Prob  Time Crime
## M       0.64  0.36  0.11 -0.09
## So      0.74  0.53  0.07 -0.09
## Ed     -0.77 -0.39 -0.25  0.32
## Po1    -0.63 -0.47  0.10  0.69
## Po2    -0.65 -0.47  0.08  0.67
## LF     -0.27 -0.25 -0.12  0.19
## M.F    -0.17 -0.05 -0.43  0.21
## Pop    -0.13 -0.35  0.46  0.34
## NW      0.68  0.43  0.23  0.03
## U1     -0.06 -0.01 -0.17 -0.05
## U2      0.02 -0.06  0.10  0.18
## Wealth -0.88 -0.56  0.00  0.44
## Ineq    1.00  0.47  0.10 -0.18
## Prob    0.47  1.00 -0.44 -0.43
## Time    0.10 -0.44  1.00  0.15
## Crime  -0.18 -0.43  0.15  1.00

#plotting the correlation matrix
corrplot(corrmat, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



I decided to try several regression model and compare results to see how they performed. Following is the list of different model tried.

1. Simple regression using lm() using all variable using all data
2. Simple regression using lm() with selected variable using all data
3. Simple regression using lm() using cross-validation with selected variable (using 80/20 training/testing split)
4. Simple regression using cross validation using caret package
5. Simple regression using stepwise method backward & forward

```
#empty matrix to log models resuls
model_results <- matrix(NA, nrow=5, ncol=7)
colnames(model_results) <- c("MODEL","R-SQUARED","ADJ R-SQUARED","F-
STATISTIC", "AIC", "BIC","PREDICTION")
```

Model 1 - Simple regression using lm() using all variable using all data

First model was a simple regression using lm() function using all data as training dataset and all variables. R-Squared of 80% showed a great fit but given the size the data, this indicated overfitting. Even the adjusted R-Squared was high. Most importantly, the predicted Crime value of 155 was very low even compared to the lowest Crime value in the available data. This model did not seem to perform well. But the output shows the predictors that are significant than others.

```
model1 <- lm(Crime~. , data=crimedata)
sum_model1 <- summary(model1)
sum_model1
```

```
##
## 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
```

```
AIC(model1)
```

```
## [1] 650.0291
```

```
BIC(model1)
```

```
## [1] 681.4816
```

```
#predicting crime value
predict(model1, predictdata)
```

```
##        1
## 155.4349
```

```
#logging results
model_results[1,] <- c("Simple Reg w/ lm() w/ all var",
round(sum_model1$r.squared,2),
round(sum_model1$adj.r.squared,2),round(sum_model1$fstatistic[1],2),round(AIC
(model1),2),round(BIC(model1),2), round(predict(model1,predictdata),2))
```

Model 2 - Simple regression using lm() using selected variables using all data

Next model was similar to model 1 except that I used the suggest 6 variables from model 1 only. I still used all data to train the model. R-Sqaured and Adj R-Squared of 76.5% and 73% were lower than model 1 but still showed overfitting. Predicted value was 1304 which was closer to the 3rd quartile of the crimes data and thus in the acceptable range.

```
model2 <- lm(Crime~ M + Ed + Po1 + U2 + Ineq + Prob , data=crimedata)
sum_model2 <- summary(model2)
sum_model2

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

AIC(model2)

## [1] 640.1661

BIC(model2)

## [1] 654.9673

#predicting crime value
predict(model2, predictdata)

##        1
## 1304.245

#logging results
model_results[2,] <- c("Simple Reg w/ lm() w/ selected var",
round(sum_model2$r.squared,2),
round(sum_model2$adj.r.squared,2),round(sum_model2$fstatistic[1],2),round(AIC
(model2),2),round(BIC(model2),2), round(predict(model2,predictdata),2))
```

Model 3 - Simple regression using lm() with selected variable using 80/20 training/testing split

Next, I built model similar to model 2 except that I splitted the data into 80% training and 20% testing. Results were pretty similar to model 2 with R-squared of 75.8% and adjusted R-Squared of 71%. Predicted values dropped slightly compared to model 2 to 1269, and was still in the acceptable range.

```
#splitting data to training and validation
set.seed(101)
sample <- sample.int(n = nrow(crimedata), size = floor(.80*nrow(crimedata)),
replace = F)
train_data <- crimedata[sample,]
test_data  <- crimedata[-sample,]
nrow(train_data)

## [1] 37

nrow(test_data)

## [1] 10

#building model 3 on training data
model3 <- lm(Crime~ M + Ed + Po1 + U2 + Ineq + Prob , data=train_data)
sum_model3 <- summary(model3)
sum_model3

##
## Call:
## lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -432.18 -124.12  -21.34   96.59  573.68
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4631.33     979.78  -4.727 5.03e-05 ***
## M             110.37      35.36   3.121 0.003965 **
## Ed            176.34      47.41   3.719 0.000821 ***
## Po1           110.07      15.02   7.328 3.67e-08 ***
## U2             94.55      47.19   2.003 0.054232 .
## Ineq           53.21      14.72   3.616 0.001084 **
## Prob        -3451.98    1557.01  -2.217 0.034341 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 192.5 on 30 degrees of freedom
## Multiple R-squared:  0.7586, Adjusted R-squared:  0.7103 
## F-statistic: 15.71 on 6 and 30 DF,  p-value: 4.484e-08
```

```
AIC(model3)
```

```
## [1] 502.5064
```

```
BIC(model3)
```

```
## [1] 515.3937
```

```
#checking model performance on testing data
eval3 <- predict(model3, test_data)
pred3 <- data.frame(cbind(actuals=test_data$Crime, predicteds=eval3))
cor(pred3)
```

```
##              actuals predicteds
## actuals    1.0000000  0.8092192
## predicteds 0.8092192  1.0000000
```

```
head(pred3)
```

```
##    actuals predicteds
## 2     1635  1343.8833
## 5     1234  1230.4937
## 15     798   780.7135
## 19     750  1231.8635
## 20    1225  1208.1651
## 23    1216   880.5614
```

```
#predicting crime value
predict(model3, predictdata)
```

```
##        1 
## 1269.989
```

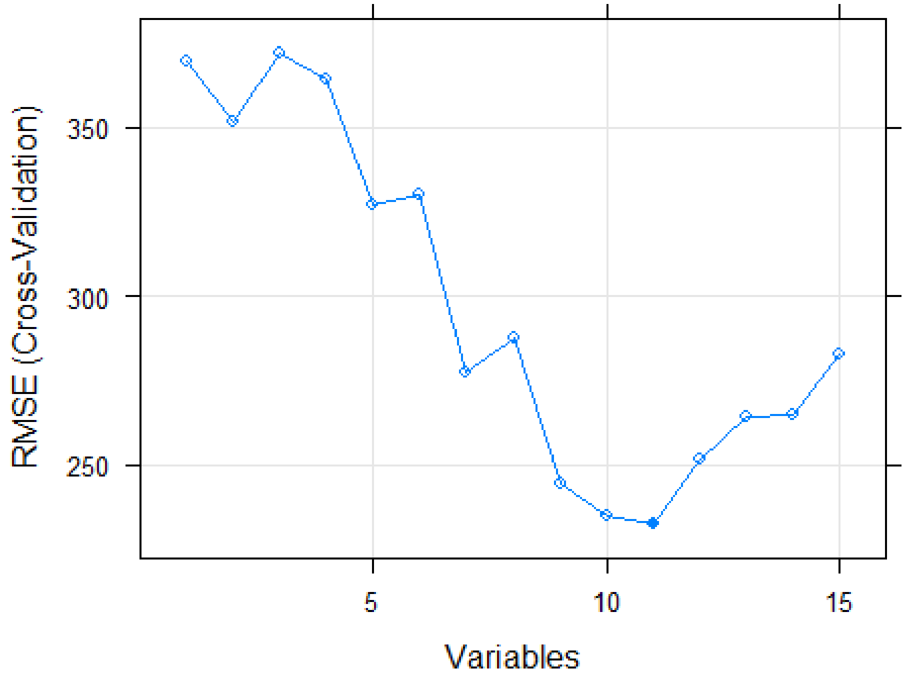
```
#logging results
model_results[3,] <- c("Simple Reg w/ lm() w/ selected var w/ train/test",
round(sum_model3$r.squared,2),
round(sum_model3$adj.r.squared,2),round(sum_model3$fstatistic[1],2),round(AIC
(model3),2),round(BIC(model3),2), round(predict(model3,predictdata),2))
```

Model 4 - Simple regression using cross validation with caret package

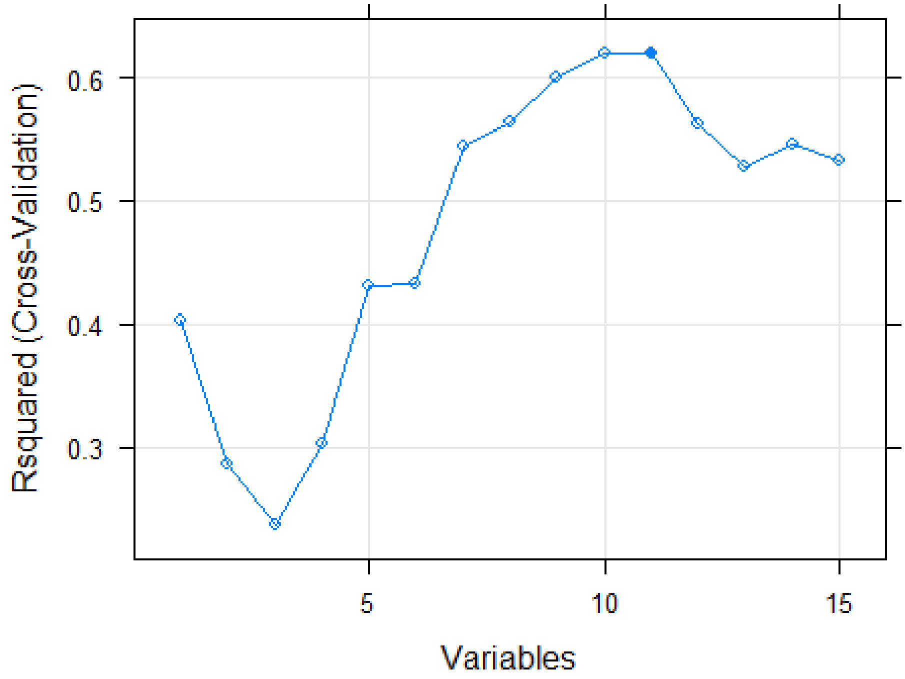
Next model I tried was a regression with k-fold cross validation using caret package (ref: https://r-forge.r-project.org/scm/viewvc.php/*checkout*/pkg/caret/inst/doc/caretSelection.pdf?revision=77&root=caret&pathrev=90 pages 5 and 6). This was a 10 fold cross-validation and the caret rfecontrol function checked model performance for diffferent combination of variables and suggested the best. 11 variables were suggested (RMSE was lowest for 11 varaibles).

The predicted value from this model was 641 which was very close to the 1st quartile of the data. R-squared of 62% for the selected model showed overfitting like other models. I manually calculated Adj R-Squared which was 42%.

```
set.seed(101) #to keep output consistent

#building model 4
subsets <- c(1:15)
ctrl <- rfeControl(functions = lmFuncs, method = "cv", number = 10, verbose =
FALSE)
model4 <- rfe(crimedata[,-16], crimedata[,16], sizes = subsets, rfeControl =
ctrl)
model4

## 
## Recursive feature selection
## 
## Outer resampling method: Cross-Validated (10 fold) 
## 
## Resampling performance over subset size:
## 
##  Variables  RMSE Rsquared   MAE RMSESD RsquaredSD  MAESD Selected
##          1 369.6   0.4028 300.5 124.54     0.2634  89.93         
##          2 351.5   0.2865 280.7 126.85     0.3334  91.60         
##          3 372.2   0.2365 296.2 127.46     0.2693  94.52         
##          4 363.9   0.3031 294.3  94.42     0.3308  72.00         
##          5 327.4   0.4315 273.8 111.03     0.3125  84.96         
##          6 329.9   0.4326 274.7 109.47     0.2906  87.00         
##          7 277.5   0.5441 231.7 111.93     0.3432  97.34         
##          8 287.8   0.5645 238.5 130.36     0.3831 105.68         
##          9 244.5   0.5998 204.8 119.52     0.3163  94.68         
##         10 235.1   0.6206 192.3 120.85     0.3114  97.35         
##         11 232.3   0.6202 191.3 111.36     0.3133  90.45        *
##         12 251.9   0.5630 204.0 113.46     0.3261  93.40         
##         13 264.4   0.5281 217.1 129.19     0.3308 112.88         
##         14 265.0   0.5461 213.8 135.84     0.3401 115.25         
##         15 282.9   0.5332 227.3 125.51     0.3105 106.87         
## 
## The top 5 variables (out of 11):
##    U1, Prob, LF, Po1, Ed

#model suggest best predictors (it's suggest 11 predictors)
predictors(model4)

##  [1] "U1"   "Prob" "LF"   "Po1"  "Ed"   "U2"   "Po2"  "So"   "M"    "Ineq"
## [11] "M.F"

#plots of model4 output
plot(model4, type=c("g","o"))
```



```r
plot(model4, metric = "Rsquared",type=c("g","o"))
```



```
#predicting crime value
predict(model4,predictdata)

##        1
## 641.0715

#calculating adj r-sq using https://mathcracker.com/r-squared-adjusted-r-
squared-calculator
model4_adjrsq <- ((1-model4$results$Rsquared[11])*(47-1)) / (47-4-1)

#logging results
model_results[4,] <- c("Simple Reg w/ cross valid using Caret",
round(model4$results$Rsquared[11],2),
round(model4_adjrsq,2),"","","",round(predict(model4,predictdata),2))
```

5. Simple regression using stepwise method backward & forward

Lastly, I built a regression model with both ways stepwise variables selection (ref: https://www.statmethods.net/stats/regression.html). I used the stepAIC function from MASS package to perform the both ways stepwise regression. The final model with 8 variables was selected.

The predicted value was 1038 which was similar to other models and R-Squared and Adj R-Squared showed overfitting (high values).

```
#building model 5
model5 <- stepAIC(model1, direction="both")

## Start:  AIC=514.65
## Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 +
##     U2 + Wealth + Ineq + Prob + Time
##
##          Df Sum of Sq     RSS    AIC
## - So      1        29 1354974 512.65
## - LF      1      8917 1363862 512.96
## - Time    1     10304 1365250 513.00
## - Pop     1     14122 1369068 513.14
## - NW      1     18395 1373341 513.28
## - M.F     1     31967 1386913 513.74
## - Wealth  1     37613 1392558 513.94
## - Po2     1     37919 1392865 513.95
## <none>                1354946 514.65
## - U1      1     83722 1438668 515.47
## - Po1     1    144306 1499252 517.41
## - U2      1    181536 1536482 518.56
## - M       1    193770 1548716 518.93
## - Prob    1    199538 1554484 519.11
## - Ed      1    402117 1757063 524.86
## - Ineq    1    423031 1777977 525.42
##
## Step:  AIC=512.65
## Crime ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 +
##     Wealth + Ineq + Prob + Time
```

```
##
##          Df Sum of Sq     RSS    AIC
## - Time    1     10341 1365315 511.01
## - LF      1     10878 1365852 511.03
## - Pop     1     14127 1369101 511.14
## - NW      1     21626 1376600 511.39
## - M.F     1     32449 1387423 511.76
## - Po2     1     37954 1392929 511.95
## - Wealth  1     39223 1394197 511.99
## <none>                1354974 512.65
## - U1      1     96420 1451395 513.88
## + So      1        29 1354946 514.65
## - Po1     1    144302 1499277 515.41
## - U2      1    189859 1544834 516.81
## - M       1    195084 1550059 516.97
## - Prob    1    204463 1559437 517.26
## - Ed      1    403140 1758114 522.89
## - Ineq    1    488834 1843808 525.13
##
## Step:  AIC=511.01
## Crime ~ M + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 +
##     Wealth + Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## - LF      1     10533 1375848 509.37
## - NW      1     15482 1380797 509.54
## - Pop     1     21846 1387161 509.75
## - Po2     1     28932 1394247 509.99
## - Wealth  1     36070 1401385 510.23
## - M.F     1     41784 1407099 510.42
## <none>                1365315 511.01
## - U1      1     91420 1456735 512.05
## + Time    1     10341 1354974 512.65
## + So      1        65 1365250 513.00
## - Po1     1    134137 1499452 513.41
## - U2      1    184143 1549458 514.95
## - M       1    186110 1551425 515.01
## - Prob    1    237493 1602808 516.54
## - Ed      1    409448 1774763 521.33
## - Ineq    1    502909 1868224 523.75
##
## Step:  AIC=509.37
## Crime ~ M + Ed + Po1 + Po2 + M.F + Pop + NW + U1 + U2 + Wealth +
##     Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## - NW      1     11675 1387523 507.77
## - Po2     1     21418 1397266 508.09
## - Pop     1     27803 1403651 508.31
## - M.F     1     31252 1407100 508.42
## - Wealth  1     35035 1410883 508.55
## <none>                1375848 509.37
## - U1      1     80954 1456802 510.06
## + LF      1     10533 1365315 511.01
## + Time    1      9996 1365852 511.03
## + So      1      3046 1372802 511.26
## - Po1     1    123896 1499744 511.42
## - U2      1    190746 1566594 513.47
## - M       1    217716 1593564 514.27
## - Prob    1    226971 1602819 514.54
## - Ed      1    413254 1789103 519.71
## - Ineq    1    500944 1876792 521.96
##
## Step:  AIC=507.77
## Crime ~ M + Ed + Po1 + Po2 + M.F + Pop + U1 + U2 + Wealth + Ineq +
##     Prob
##
##          Df Sum of Sq     RSS    AIC
## - Po2     1     16706 1404229 506.33
## - Pop     1     25793 1413315 506.63
## - M.F     1     26785 1414308 506.66
## - Wealth  1     31551 1419073 506.82
## <none>                1387523 507.77
## - U1      1     83881 1471404 508.52
## + NW      1     11675 1375848 509.37
## + So      1      7207 1380316 509.52
## + LF      1      6726 1380797 509.54
## + Time    1      4534 1382989 509.61
## - Po1     1    118348 1505871 509.61
## - U2      1    201453 1588976 512.14
## - Prob    1    216760 1604282 512.59
## - M       1    309214 1696737 515.22
## - Ed      1    402754 1790276 517.74
## - Ineq    1    589736 1977259 522.41
##
## Step:  AIC=506.33
## Crime ~ M + Ed + Po1 + M.F + Pop + U1 + U2 + Wealth + Ineq +
##     Prob
##
##          Df Sum of Sq     RSS    AIC
## - Pop     1     22345 1426575 505.07
## - Wealth  1     32142 1436371 505.39
## - M.F     1     36808 1441037 505.54
## <none>                1404229 506.33
## - U1      1     86373 1490602 507.13
## + Po2     1     16706 1387523 507.77
## + NW      1      6963 1397266 508.09
## + So      1      3807 1400422 508.20
## + LF      1      1986 1402243 508.26
## + Time    1       575 1403654 508.31
## - U2      1    205814 1610043 510.76
## - Prob    1    218607 1622836 511.13
## - M       1    307001 1711230 513.62
## - Ed      1    389502 1793731 515.83
## - Ineq    1    608627 2012856 521.25
## - Po1     1   1050202 2454432 530.57
##
## Step:  AIC=505.07
## Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Wealth + Ineq + Prob
##
```

```
##          Df Sum of Sq     RSS    AIC
## - Wealth  1     26493 1453068 503.93
## <none>                1426575 505.07
## - M.F     1     84491 1511065 505.77
## - U1      1     99463 1526037 506.24
## + Pop     1     22345 1404229 506.33
## + Po2     1     13259 1413315 506.63
## + NW      1      5927 1420648 506.87
## + So      1      5724 1420851 506.88
## + LF      1      5176 1421398 506.90
## + Time    1      3913 1422661 506.94
## - Prob    1    198571 1625145 509.20
## - U2      1    208880 1635455 509.49
## - M       1    320926 1747501 512.61
## - Ed      1    386773 1813348 514.35
## - Ineq    1    594779 2021354 519.45
## - Po1     1   1127277 2553852 530.44
##
## Step:  AIC=503.93
## Crime ~ M + Ed + Po1 + M.F + U1 + U2 + Ineq + Prob
##
##          Df Sum of Sq     RSS    AIC
## <none>                1453068 503.93
## + Wealth  1     26493 1426575 505.07
## - M.F     1    103159 1556227 505.16
## + Pop     1     16697 1436371 505.39
## + Po2     1     14148 1438919 505.47
## + So      1      9329 1443739 505.63
## + LF      1      4374 1448694 505.79
## + NW      1      3799 1449269 505.81
## + Time    1      2293 1450775 505.86
## - U1      1    127044 1580112 505.87
## - Prob    1    247978 1701046 509.34
## - U2      1    255443 1708511 509.55
## - M       1    296790 1749858 510.67
## - Ed      1    445788 1898855 514.51
## - Ineq    1    738244 2191312 521.24
## - Po1     1   1672038 3125105 537.93
```

```
model5$anova # display results
```

```
## 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
```

```
sum_model5 <- summary(model5)
sum_model5
```

```
## 
## 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
```

```
AIC(model5)
```

```
## [1] 639.3151
```

```
BIC(model5)
```

```
## [1] 657.8166
```

```
#predicting Crime Value
predict(model5,predictdata)
```

```
##        1 
## 1038.413
```

```
#logging results
model_results[5,] <- c("Simple Reg using both ways stepwise",
round(sum_model5$r.squared,2),
round(sum_model5$adj.r.squared,2),round(sum_model5$fstatistic[1],2),round(AIC
(model5),2),round(BIC(model5),2), round(predict(model5,predictdata),2))
```

Finally, I printed the key outputs from all 5 models including R-Sq, Adj R-Sq, F-Statistic, AIC, BIC and the predicted value. It was clear that all models except may be k-vold cross validation (model 4) showed overfitting. The predicted values, though were in acceptable range except for model 1, they varied significantly model to model with output ranging from 1st to 3rd quartile. As noted throughout, the models show overfitting due to small n (amount of data points) and we will need other techniques (regularization, drop-out layers etc.) or more data to get better results.

```
model_results

##      MODEL                                              R-SQUARED ADJ R-SQUARED
## [1,] "Simple Reg w/ lm() w/ all var"                    "0.8"     "0.71"
## [2,] "Simple Reg w/ lm() w/ selected var"               "0.77"    "0.73"
## [3,] "Simple Reg w/ lm() w/ selected var w/ train/test" "0.76"    "0.71"
## [4,] "Simple Reg w/ cross valid using Caret"            "0.62"    "0.42"
## [5,] "Simple Reg using both ways stepwise"              "0.79"    "0.74"


##      F-STATISTIC AIC      BIC      PREDICTION
## [1,] "8.43"      "650.03" "681.48" "155.43"
## [2,] "21.81"     "640.17" "654.97" "1304.25"
## [3,] "15.71"     "502.51" "515.39" "1269.99"
## [4,] ""          ""       ""       "641.07"
## [5,] "17.74"     "639.32" "657.82" "1038.41"
```