Homework 5-Regression

Mark Pearl

9/25/2019

8.1 Application for Regression in Real-Life

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.

For a recent project at work for an NHL team, regression is heavily used for points production for each line. There are several categorical and continuous variables to account for such as historical point production, line rank (i.e 1st to 4th line), aggressivness, country of origin, etc. The response variable is the aggregated points across all players on a line or defensemen pairing.

The data would be trained on the categorical attributes which remain static along with historical season results for each player. This is a great tool for understanding line combinations against other teams, and what-if-analysis to determine which players on a team should be paired up together to optimize the total results for the entire team.

```
uscrime_data <- read.table('C:/Users/mjpearl/Desktop/omsa/ISYE-6501-OAN/hw5/data/uscrime.txt',header = T
head(uscrime_data)
```

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

8.2 LM Model for US Crime Data

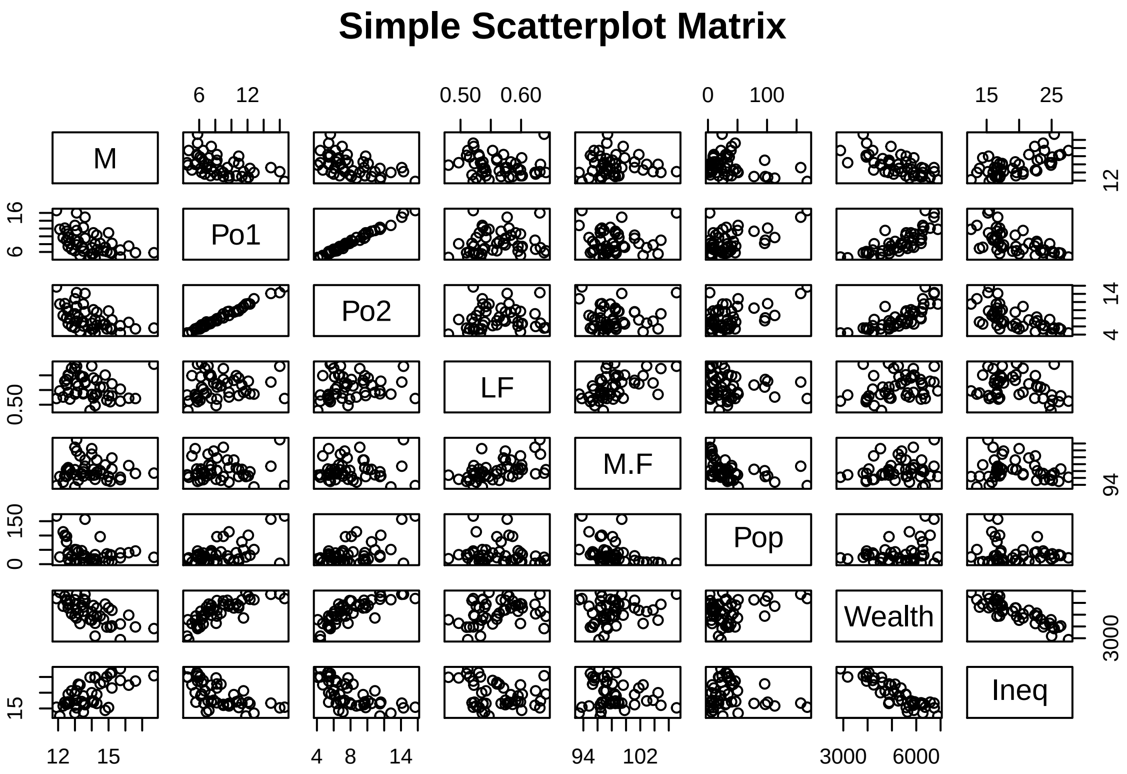
The following plots will conduct exploratory analysis on the data to get a sense of the data's distribution for each variable. We will then use a simple linear regression approach using lm() to predict against the target variable for crime with the one row of data we've been provided for test data.

```
summary(uscrime_data)
```

```
##        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  
```

There's potentially a few variables in this dataset such as Population which could require scaling or normalization. In addition, based on our last findings from a previous homework, it might be beneficial to remove outlier values towards the upper quartile.

```r
## 75% of the sample size for the uscrime dataset
pairs(~M+Po1+Po2+LF+M.F+Pop+Wealth+Ineq,data=uscrime_data,
   main="Simple Scatterplot Matrix")
```



As we can see, there seems to be a positive correlation between Po1 and Po2 and a negative correlation between Wealth and Inequality. There are several approaches we can use to deal with these features. One approach through feature engineering would be to conduct PCA (Principal Component Analysis) on the correlated features to produce a net new feature which alleviates the co-linearity. Another approach is to use L1 or L2 regularization to penalize the weights of these features so that they don't impact the results of our response variable.

```
#Since we're using a specific row for the test dataset, we'll use the full output to train the model
lm.fit = lm(Crime ~ .,data = uscrime_data)
summary(lm.fit)
```

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

Based on the output of our fitted model on the training data, we can see modest performance with an R-Sqaure of 0.78. This means that the model is able to fairly accomodate the dataset's variance. We also see that there's several features with large p-values indicating that they don't provide any predictive value. The p-value results could be distorted for several of the features which have high co-linearity or due to overfitting. In a future test we will try and remove some features to assess the impact.

```
test <- data.frame(M = 14.0, So = 0, Ed = 10.0, Po1 = 12.0, Po2 = 15.5, LF = 0.640, M.F = 94.0, Pop = 15
lm_predict <- predict(lm.fit, test)
lm_predict
```

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

We can see with the predict function that are value doesn't fall within the range of values for the Crime variable in the training dataset, however since we haven't split our test dataset from the original data, we don't have a way of measure it's accuracy. We would be able to calculate the MSE or RMSE if we were using test data from the original dataset.

Now let's try and re-run the fit with the Po2 and Wealth variables removed from the training dataset.

```
drops <- c("Po2","Wealth")
train <- uscrime_data[ , !(names(uscrime_data) %in% drops)]
lm.fit2 = lm(Crime ~ .,data = train)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -469.4  -93.1   12.6  117.3  506.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6041.0176  1515.7345  -3.986 0.000351 ***
## M              84.0350    40.8957   2.055 0.047879 *
## So             35.2894   143.7092   0.246 0.807543
## Ed            185.9198    59.8202   3.108 0.003861 **
## Po1           105.0940    21.7659   4.828 3.06e-05 ***
## LF           -127.9865  1392.3561  -0.092 0.927317
## M.F            20.1254    20.1066   1.001 0.324141
## Pop            -0.6822     1.2761  -0.535 0.596494
## NW              1.3912     6.0482   0.230 0.819502
## U1          -5748.4126  4146.8729  -1.386 0.174980
## U2            180.7362    80.8400   2.236 0.032251 *
## Ineq           60.7323    17.9172   3.390 0.001829 **
## Prob        -4517.0792  2160.3360  -2.091 0.044315 *
## Time           -0.5337     6.6346  -0.080 0.936366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 207.9 on 33 degrees of freedom
## Multiple R-squared:  0.7927, Adjusted R-squared:  0.711
## F-statistic: 9.707 on 13 and 33 DF,  p-value: 7.32e-08
```

Based on the updated plot, we can see this did improve the standard error for the residuals. However, we don’t notice a significant improvement for the R2 score and a worse F-statistics measure. This is the result of overfitting, which makes sense since we have such few observations in our training dataset.