HOMEWORK 5. ISYE 6501

Guillermo de la Hera Casado

September 24th, 2019

Question 8.1. Describe a situation for which linear regression would be appropiate. List up to 5 predictors

I work for a large Pay-TV company in Africa. Regression analysis is a very relevant technique to predict the Lifetime Value of our customers. That's how much revenue can be extracted from that subscriber during overall product consumption up to cancellation. It's a continuous response that can be estimated for future observations from historical beahavior.

Regarding the predictors, I would suggest the following:

- Previous dormancy patterns
- Tenure
- The package customer is subscribed to
- demographics of subscriber
- Add-ons

Question 8.2. Use regression to predict the observed crime rate in a city

Building the basic model

From the Grubbs test performed in HW3 with this data, we learnt that the value: *1993* could be potentially an outlier with p value = 0.079.

However, since we will use for this homework confidence level = 95%, we cannot reject the null hypothesis [value is not an outlier]. Now we can move on to create the regression model to predict the crime rate, without performing any modification on the dataset:

```
set.seed(101) # Set Seed so that same sample can be reproduced in future also
usCrime <- read.table("uscrime.txt", header = TRUE, sep = "\t")
lm_uscrime <- lm(Crime~., data=usCrime)
summary(lm_uscrime)
```

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

```
print(BIC(lm_uscrime))
```

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

```
print(AIC(lm_uscrime))
```

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

From the summary we can infer the following learnings:

- The median residual (difference between predicted value and actual value of crime rate) is: -6.69
- For few coefficients we can reject the null hypothesis at 95% confidence level. Therefore we say that the following coefficients are meaningful and not equal to zero: **M, Ed, Ineq, Prob**. Then we have **Po1, U2** that are significant at 90% confidence level, so would to keep them in mind.
- As we can expect Multiple R squared is higher than the adjusted R-squared, since there is more than one predictor. As discussed during office hours we will pay attention to the adjusted R-squared, as it doesn't only speak about reduction of error but also reduction of complexity in the model. That helps to assess whether each extra coefficient added to the model is making a relevant impact in explaining the variance or not vs removing it. Adjusted R squared = 70.78%, that is a great value. It means that the model is able to explain 70.78% of the variance in the response variable, with the current set up.
- We can also see that the p-value in the F-statistic is quite low. That confirms that at least one of the coefficients in the model is significant.

Building a simplified model with only statistically relevant predictors

What if we fit the model only with those coefficients that proved to be significant from 90% confidence level onwards? Let's see the impact on adjusted R-squared, BIC and AIC

```
lm_uscrime_simple <- lm(Crime~M+Ed+Ineq+Prob+Po1+U2, data=usCrime)
summary(lm_uscrime_simple)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Ineq + Prob + Po1 + U2, 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 ***
## 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 ***
## U2             89.37      40.91   2.185  0.03483 *
## ---
## 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
```

```
print(BIC(lm_uscrime_simple))
```

```
## [1] 654.9673
```

```
print(AIC(lm_uscrime_simple))
```

```
## [1] 640.1661
```

As we can see:

- The median residual (difference between predicted value and actual value of crime rate) is: -19.68
- Adjusted R-squared increases from 71% to 73%!
- BIC decreases from 681.48 to 654.97. The difference is: 26.5. As per explained by Dr. Sokol, that means that the simplified model, that has the smaller BIC is **"very likely"** better than the overall model.
- AIC decreases from 650 to 640. Even though, what's the likelihood of the overall model to be better than the simplified model? As per the formula explained by Dr. Sokol, just only: *e^ (640-650/2) = 0.7%*

Therefore we can infer that **the simplified model works better on reducing both error and complexity** and we will choose it for the next sections.

Cross validation

The limitation on the above procedure is that we are fitting the model with all the training data and then comparing the actuals vs predicted values to calculate the residuals. That may lead to overfitting.

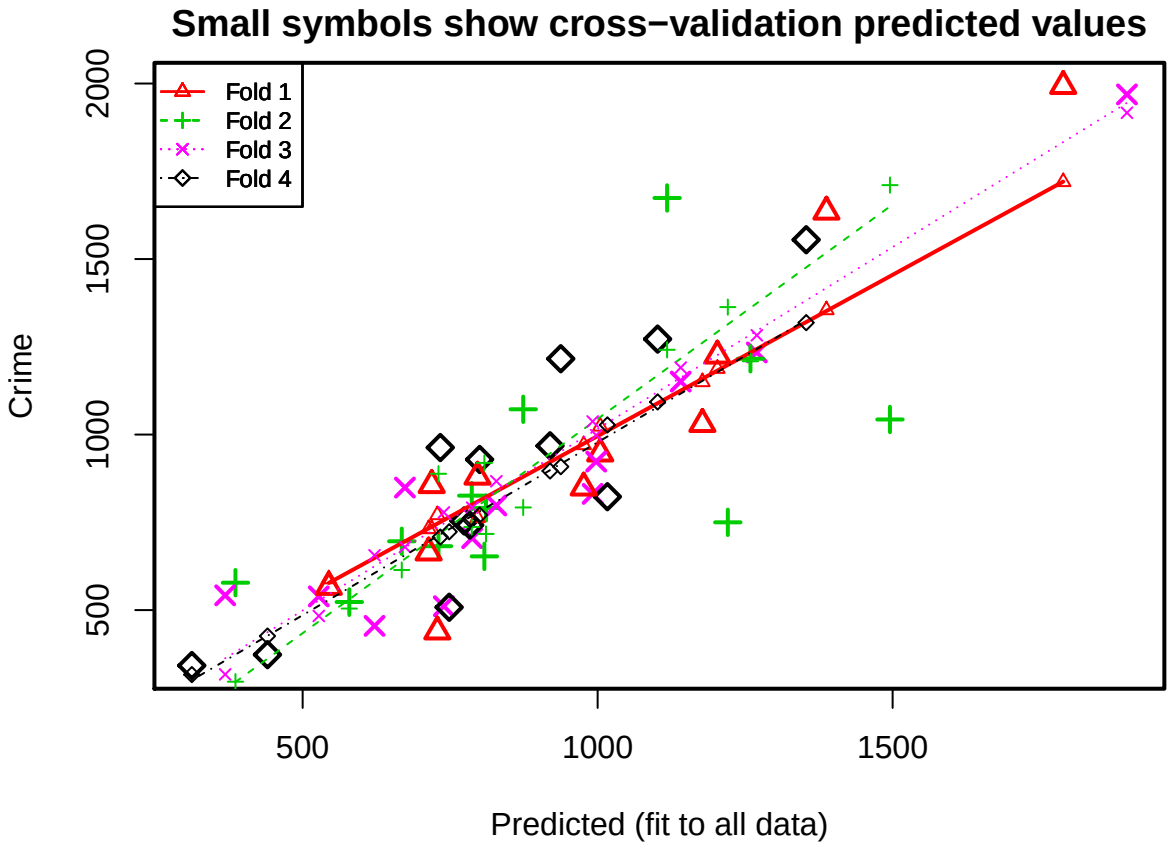
It would be interesting to calculate the residuals on unseen data (test set) and work out the adjusted R squared coming out of it. For that, we will use the *cv.lm* function in DAAG package, selecting 4 folds.

```
lm_uscrime_cv <-cv.lm(usCrime, lm_uscrime_simple, m=4)
```

```
## Analysis of Variance Table
##
## Response: Crime
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## M          1   55084   55084    1.37 0.24914
## Ed         1  725967  725967   18.02 0.00013 ***
## Ineq       1   37674   37674    0.94 0.33928
## Prob       1  990334  990334   24.59 1.4e-05 ***
## Po1        1 3268577 3268577   81.15 3.6e-11 ***
## U2         1  192233  192233    4.77 0.03483 *
## Residuals 40 1611057   40276
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Warning in cv.lm(usCrime, lm_uscrime_simple, 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    1388 719 713.6 1004.4 1203.0  728 1789 544.4 796 1178  976
## cvpred       1355 731 731.1 1023.2 1187.6  771 1720 588.4 763 1150  970
## Crime        1635 856 664.0  946.0 1225.0  439 1993 566.0 880 1030  849
## CV residual   280 125 -67.1  -77.2   37.4 -332  273 -22.4 117 -120 -121
##
## Sum of squares = 334042    Mean square = 30367    n = 11
##
## fold 2
## Observations in test set: 12
##                  1   3    6   11   19    25     28   29    30   33   35
## Predicted    810.8 386  730 1118 1221 579.1 1259.0 1495 668.0  874  808
## cvpred       716.9 296  888 1241 1363 504.3 1208.7 1711 614.2  792  919
## Crime        791.0 578  682 1674  750 523.0 1216.0 1043 696.0 1072  653
```

```
## CV residual  74.1 282 -206  433 -613  18.7    7.3 -668  81.8  280 -266
##                39
## Predicted   786.7
## cvpred      736.6
## Crime       826.0
## CV residual  89.4
##
## Sum of squares = 1300449    Mean square = 108371    n = 12
##
## fold 3
## Observations in test set: 12
##                  4      5    10  12   13  15    17    34   37     40  42
## Predicted   1897.2 1269.8 787.3 673  739 828 527.4 997.5  992 1140.8 369
## cvpred      1916.6 1282.8 791.8 680  778 867 483.3 998.2 1037 1190.7 317
## Crime       1969.0 1234.0 705.0 849  511 798 539.0 923.0  831 1151.0 542
## CV residual   52.4  -48.8 -86.8 169 -267 -69  55.7 -75.2 -206  -39.7 225
##                45
## Predicted    622
## cvpred       656
## Crime        455
## CV residual -201
##
## Sum of squares = 261503    Mean square = 21792    n = 12
##
## fold 4
## Observations in test set: 12
##                7    8  18  21   23    24    27  31  32   36   43   46
## Predicted   733 1354 800 783  938 919.4 312.2 440 774 1102 1017  748
## cvpred      708 1319 771 759  909 896.3 316.2 426 740 1093 1027  723
## Crime       963 1555 929 742 1216 968.0 342.0 373 754 1272  823  508
## CV residual 255  236 158 -17  307  71.7  25.8 -53  14  179 -204 -215
##
## Sum of squares = 369549    Mean square = 30796    n = 12
##
## Overall (Sum over all 12 folds)
##    ms
## 48203
```

Let's use the results from the cross validation to calculate SS residuals, SS total and adjusted R-squared:

```
n <- length(usCrime$Crime)
k <- length(lm_uscrime_simple$coefficients)-1
cvpred = lm_uscrime_cv$cvpred
CV_residual <- cvpred - usCrime$Crime
SSyy <- sum((usCrime$Crime - mean(usCrime$Crime))^2)
SSE <- sum(CV_residual^2)
Adj_R_Squared <- 1 - (SSE/SSyy) * (n-1) / (n-(k+1))
print(Adj_R_Squared)
```

```
## [1] 0.621
```

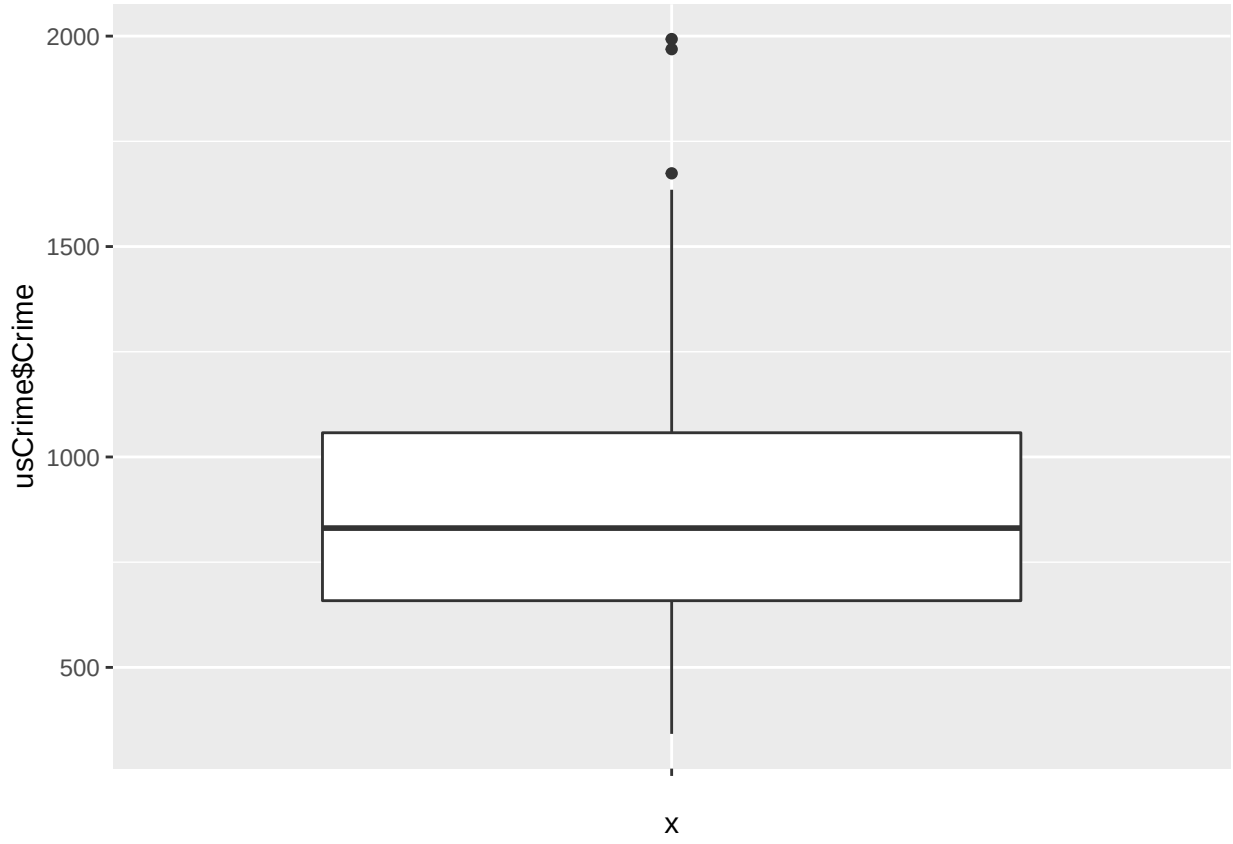
The adjusted R-squared on unseen obervations is **62.1%**. This means that the model is able to explain that percentage of variance on unseen data. It's not as good as the 73% that we got on the training data, but probably much more realistic when the model is tested in a real life scenario.

Fitting the observation on the best model

The next step will be to predict the value for the required observation, and see where it sits vs overall response distribution.

```
test_point <- 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)

ggplot(data=usCrime, aes(x="", y = usCrime$Crime)) + geom_boxplot()
```




```
pred_model <- predict(lm_uscrime_simple, test_point)
pred_model
```

```
##    1
## 1304
```

The predicted crime rate for this observation would be: **1304**. Looking at the boxplot for the response value, we can see that the predicted value would be outside of the Interquantile Range (IQR) but **still contained within the whisker** (so between third quantile and third quantile + 1.5*IQR) Therefore the model wouldn't be producing an outlier in this specific case.

Bonus point - Checking normality of residuals on the simplified model

Another important bit of testing the regression quality is to verify that the residuals are independent and normally distributed. From the below, we can see that the Standard Residuals of the sample match closely the normal distribution, except for one data point that may likely be the outlier mentioned above.

```r
usCrime.stdres = rstandard(lm_uscrime_simple)
qqnorm(usCrime.stdres, ylab = "Standarized Residuals", xlab = "Normal Scores", main = "Checking Normali
qqline(usCrime.stdres)
```