Homework5

08/02/2019

Contents

Question 8.1

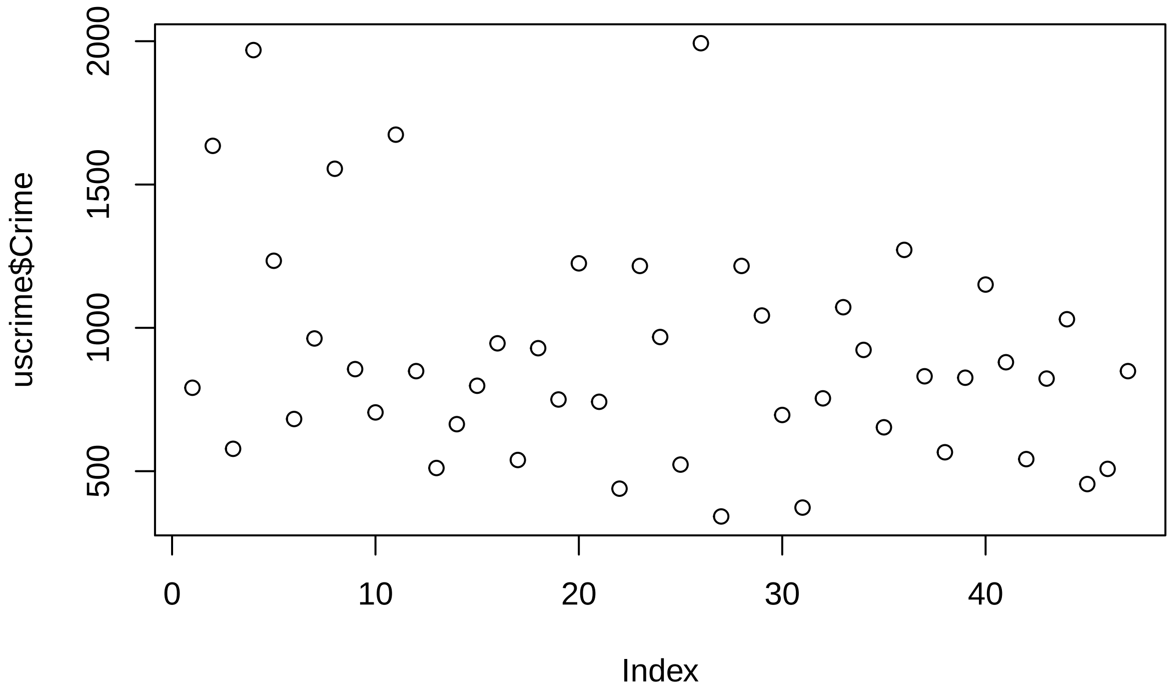
I work in analytics in a telephone company and the linear regression is a easy model to use in a production environment and easy to understand, so we use it regularly. For this question, I have a nice example. In our telephone company, the prepaid users are the majority of users and they make the most of the revenue of the company. The prepaid users are the ones without a contract and they recharge their phone voice plans or data plans according to their needs. So in this case we want to predict monthly sales of prepaid recharges using data from previous years. As possible predictors we can use the following ones:

1. The monthly recharges of voice plans (minutes).
2. The monthly recharges of data plans (GB).
3. The monthly new users we have.
4. The montly users we loose (churn).
5. The monthly active users.

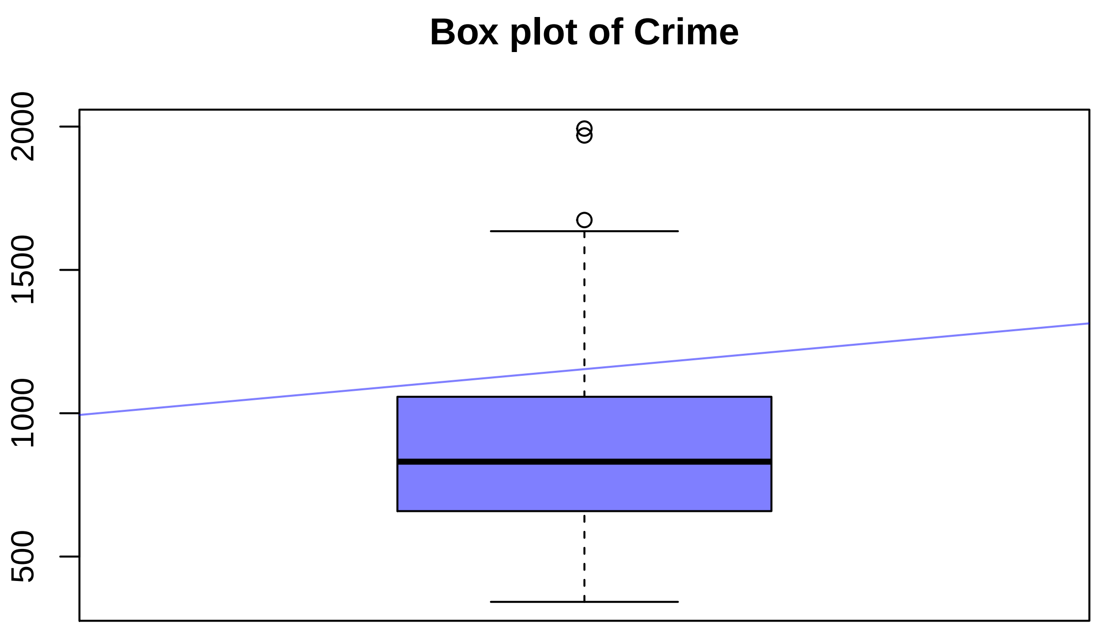
Question 8.2

use regression (a useful R function is lm or glm) to predict the observed crime rate in a city.

```
set.seed(42)
# import data
uscrime <- read.delim("~/Documents/R/GeorgiaTech/DataPreparation/uscrime.txt")
# see graphically
plot(uscrime$Crime)
```




```r
# we already know there are 2 outliers from homework 3 but it is a good practice to check for it.
boxplot(uscrime$Crime,col=rgb(0,0,1,0.5), main="Box plot of Crime")
# This graph shows that most of the data behaves normal and shows 2 possible outliers.
qqline(uscrime$Crime,col=rgb(0,0,1,0.5))
```




```r
# build the model with all avaialble predictors.
model <- lm( Crime ~ ., uscrime)

# lets see
print (summary(model))
```

```
##
## 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
```

Analysis 8.2

To apply a linear regression model to a data set is pretty straight forward.

First use all the variable we have to build the model and we analyze the model there are some predictors with high `p-values` that we can remove because they won't be significant coefficients for the model. The adjusted R^2 value `0.7` seems high enough but by removing coefficients the model can be simplified. The `p-values` represent the probability of a coefficient being zero, so only keep the coefficients with relative low `p-values`.

```r
# use only the predictors that show a low probability of being zero
better_model <- lm(Crime ~ M + Ed + U2 + Ineq + Prob, uscrime)
summary(better_model)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + U2 + Ineq + Prob, data = uscrime)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -478.8 -233.6  -46.5  143.2  797.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -3336.52    1435.26  -2.325  0.02512 *
## M              85.33      54.39   1.569  0.12437
## Ed            214.69      73.20   2.933  0.00547 **
## U2            160.01      65.54   2.441  0.01903 *
## Ineq           29.50      21.56   1.368  0.17880
## Prob        -6897.24    2427.81  -2.841  0.00697 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 328.6 on 41 degrees of freedom
## Multiple R-squared:  0.3565, Adjusted R-squared:  0.278
## F-statistic: 4.542 on 5 and 41 DF,  p-value: 0.002186
```

With this result the R^2 metric has lowered to `0.2`, maybe if instead of using a `0.05` threshold for the p-values it is used a `0.1` threshold, the R^2 gets better.

```r
better_model_2 <- lm(Crime ~ M + Ed + U2 +Po1+ Ineq + Prob, uscrime)
summary(better_model_2)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + U2 + Po1 + Ineq + Prob, 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 ***
## U2             89.37      40.91   2.185  0.03483 *
## Po1           115.02      13.75   8.363 2.56e-10 ***
## 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
```

With this change the adjusted R^2 value has risen to `0.73`, so the threshold of the p-value can have significant relevance over the model quality. It is better to use the adjusted R^2 because the other one will always increase when the predictors increase, the adjusted one takes into account the number of predictors used, so it is more reliable.

The adjusted R^2 value is not the unique measure for a linear regression model, let's try Dr. Sokol lectures and apply `Akaike's information criterion - AIC and the Bayesian information criterion - BIC`

```r
# model 1
AIC(model)
```

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

```r
BIC(model)
```

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

```r
# model 2
AIC(better_model)
```

```
## [1] 685.6872
```

```r
BIC(better_model)
```

```
## [1] 698.6382
```

```r
# model 3
AIC(better_model_2)
```

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

```r
BIC(better_model_2)
```

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

```r
# The lower AIC/BIC is the best one, so better_model_2 is the best
```

Now let's try to predict, with the given data and the models.

```r
data_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

pred_model1 <- predict(model, data_point)
pred_model1
```

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

```r
pred_model2 <- predict(better_model,data_point)
pred_model2
```

```
##        1
## 898.1004
```

```r
pred_model3 <- predict(better_model_2, data_point)
pred_model3
```

```
##        1
## 1304.245
```

```r
# basic stats to compare the predictions
avg <- mean(uscrime$Crime)
mx <- max(uscrime$Crime)
mn <- min(uscrime$Crime)
avg
```

```
## [1] 905.0851
```

```r
mx
```

```
## [1] 1993
```

```r
mn
```

```
## [1] 342
```

As it can be seen, the predicted value with the first model `155` seems not real because it is less than the minimum value of the whole data set. The predicted value with model 2 and 3, are near the average and between the range, so those values make sense. The best model still is model 3 because has lower AIC, BIC and better adjusted R^2.

The equation of the best model is the following one:

$ y = -5040.50 + 105.02M + 196.47Ed+ 89.39U2 + 115.02Po1 + 67.65Ineq - 3801.84Prob $

In linear regression it is important to understand what the coefficients mean, so here it is the description on each coefficient used.

M = percentage of males aged 14–24 in total state population Ed = mean years of schooling of the population aged 25 years or over Po1 = per capita expenditure on police protection in 1960 U2 = unemployment rate of urban males 35–39 Ineq = income inequality: percentage of families earning below half the median income. Prob = probability of imprisonment: ratio of number of commitments to number of offenses

Seeing this information it could be very tempting to say, for example, unemployment leads to crime, but remember that correlation doesn't mean causation.

Regression output

```
# confidence interval
confint(better_model_2)
```

```
##                    2.5 %      97.5 %
## (Intercept) -6859.156298 -3221.85366
## M              37.719271   172.31986
## Ed            106.019238   286.92316
## U2              6.692602   172.03948
## Po1            87.227152   142.82123
## Ineq           39.488023    95.81841
## Prob        -6890.236192  -713.43637
```

```
# Adjusted R squared
summary(better_model_2)$adj.r.squared
```

```
## [1] 0.7307463
```

```
# Residual Standard Error (RSE)
RSE <- sigma(better_model_2)
RSE
```

```
## [1] 200.6899
```

```
error_rate <- RSE/avg
```

The quality of a linear regression fit is typically assessed using two quantities: the residual standard error (RSE) and the R^2.

The R^2 can be interpreted as 73% of the variance in the measure of crime can be predicted by M, Ed, Po1, U2, Ineq and Prob.

The RSE/mean can be interpreted as 23% errror rate of the model (The less is better)

The 1% threshold chosen for the p-values means that each coefficient will have 1% of chance of not being significant.

Cross Validation

Since the data set only has 47 points, it is highly probably that the reported quality of the model is overfitted so let's do cross validation to have a more realistic quality.

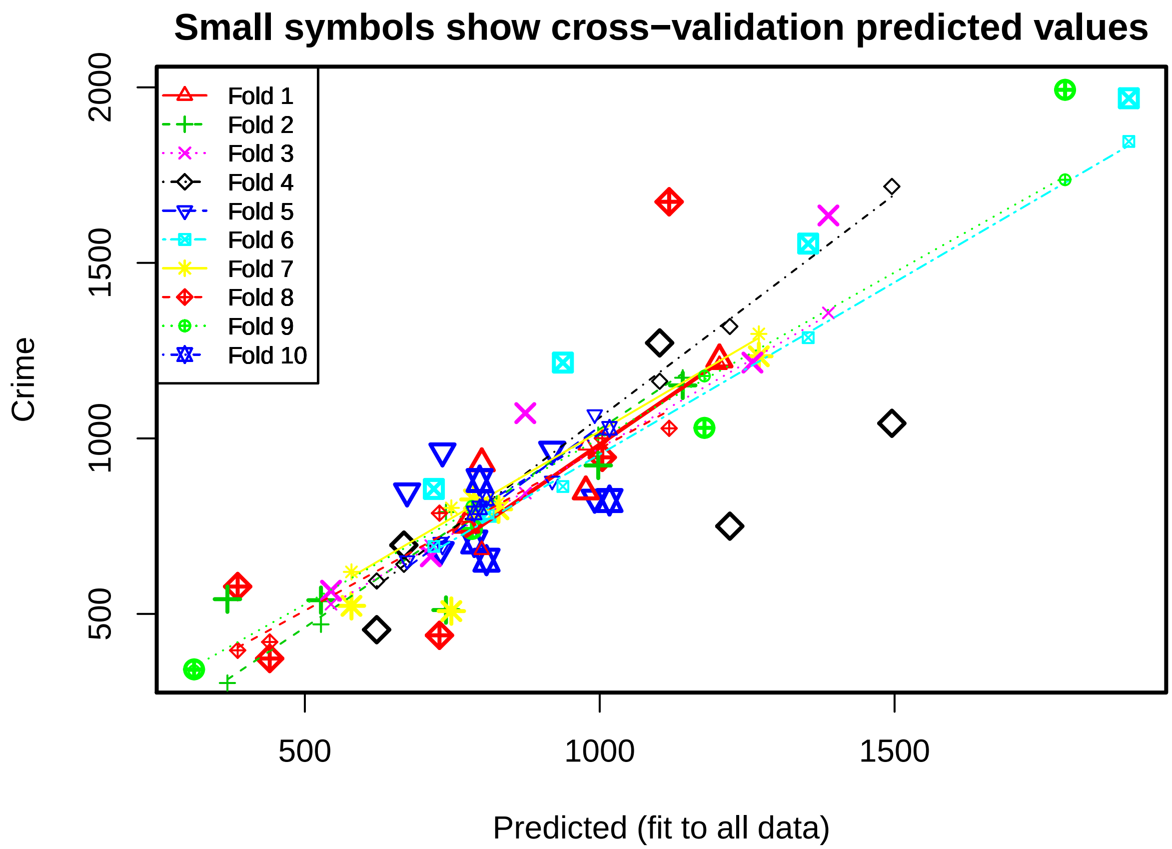
```
# TA suggestion
library(DAAG)
```

```
## Loading required package: lattice
```

```
# m is the number of folds. The m value was chosen arbitrarily
cross_lm <- cv.lm(uscrime,better_model_2,m=10)
```

```
## 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 ***
## U2         1  736262  736262   18.28 0.00011 ***
## Po1        1 2654976 2654976   65.92 5.5e-10 ***
## Ineq       1  848273  848273   21.06 4.3e-05 ***
## Prob       1  249308  249308    6.19 0.01711 *
## Residuals 40 1611057   40276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Warning in cv.lm(uscrime, better_model_2, m = 10):
##
##  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: 4
##              18     20    32   47
## Predicted   800 1203.0 773.7  976
## cvpred      682 1209.6 774.5  980
## Crime       929 1225.0 754.0  849
## CV residual 247   15.4 -20.5 -131
##
## Sum of squares = 78635    Mean square = 19659    n = 4
##
## fold 2
## Observations in test set: 5
##               13    17     34     40  42
## Predicted    739 527.4  997.5 1140.8 369
## cvpred       790 470.2 1009.6 1172.8 303
## Crime        511 539.0  923.0 1151.0 542
## CV residual -279  68.8  -86.6  -21.8 239
##
## Sum of squares = 147616    Mean square = 29523    n = 5
##
## fold 3
## Observations in test set: 5
##                2    14      28   33    38
## Predicted   1388 713.6 1259.00  874 544.4
## cvpred      1358 694.5 1220.27  844 527.7
## Crime       1635 664.0 1216.00 1072 566.0
```

```
## CV residual  277 -30.5   -4.27  228  38.3
##
## Sum of squares = 131281    Mean square = 26256    n = 5
##
## fold 4
## Observations in test set: 5
##               19   29    30   36   45
## Predicted   1221 1495 668.0 1102  622
## cvpred      1319 1718 641.3 1163  594
## Crime        750 1043 696.0 1272  455
## CV residual -569 -675  54.7  109 -139
##
## Sum of squares = 813002    Mean square = 162600    n = 5
##
## fold 5
## Observations in test set: 5
##                 6   7  12  24   37
## Predicted   730.3 733 673 919  992
## cvpred      704.7 694 652 879 1068
## Crime       682.0 963 849 968  831
## CV residual -22.7 269 197  89 -237
##
## Sum of squares = 175791    Mean square = 35158    n = 5
##
## fold 6
## Observations in test set: 5
##                  1    4    8   9   23
## Predicted   810.83 1897 1354 719  938
## cvpred      786.87 1846 1287 692  863
## Crime       791.00 1969 1555 856 1216
## CV residual   4.13  123  268 164  353
##
## Sum of squares = 238444    Mean square = 47689    n = 5
##
## fold 7
## Observations in test set: 5
##                  5    15    25    39   46
## Predicted   1269.8 828.3 579.1 786.7  748
## cvpred      1297.7 815.7 619.7 790.6  802
## Crime       1234.0 798.0 523.0 826.0  508
## CV residual  -63.7 -17.7 -96.7  35.4 -294
##
## Sum of squares = 101422    Mean square = 20284    n = 5
##
## fold 8
## Observations in test set: 5
##               3   11     16   22    31
## Predicted   386 1118 1004.4  728 440.4
## cvpred      396 1029 1000.5  787 420.1
## Crime       578 1674  946.0  439 373.0
## CV residual 182  645  -54.5 -348 -47.1
##
## Sum of squares = 575876    Mean square = 115175    n = 5
##
```

```
## fold 9
## Observations in test set: 4
##                 21   26     27   44
## Predicted   783.3 1789 312.20 1178
## cvpred      806.5 1737 337.99 1178
## Crime       742.0 1993 342.00 1030
## CV residual -64.5  256   4.01 -148
##
## Sum of squares = 91639    Mean square = 22910    n = 4
##
## fold 10
## Observations in test set: 4
##                 10   35    41   43
## Predicted   787.3  808 796.4 1017
## cvpred      787.6  828 802.2 1029
## Crime       705.0  653 880.0  823
## CV residual -82.6 -175  77.8 -206
##
## Sum of squares = 85976    Mean square = 21494    n = 4
##
## Overall (Sum over all 4 folds)
##    ms
## 51908
```

```
# measure quality by calculating R^2 and RSE
#R^2 = 1 - RSE/TSS
```

$R^2 = (TSS - RSS)/TSS = 1 - RSS/TSS$ $RSE = \sqrt{\frac{1}{n-2}RSS}$

Where TSS is the total sum of squares. $\sum_{i=1}^{n}(y_i - \bar{y})^2$ and RSS is the residual sum of squares. $\sum_{i=1}^{n}(y_i - \hat{y})^2$

```
RSS <- attr(cross_lm,"ms")*nrow(uscrime)
TSS <- sum((uscrime$Crime - mean(uscrime$Crime))^2)
R_square <- 1 - RSS/TSS
R_square
```

```
## [1] 0.645
```

```
RSE <- sqrt((1/(nrow(uscrime)-2))*RSS)
RSE
```

```
## [1] 233
```

```
error_rate <- RSE/avg
error_rate
```

```
## [1] 0.257
```

Regression output CV

Now with cross-validation we have a more realistic fit of the model selected

The R^2 can be interpreted as 65% of the variance in the measure of crime can be predicted by M, Ed, Po1, U2, Ineq and Prob. Here we have the not overfitted quality, and according to Dr. Sokol, it is still good enough to use it.

The RSE/mean can be interpreted as 26% errror rate of the model (The less is better). This measure only changed 3%.

Also tested cv with multiple `m` folds and the results varied 1%, so it was not sensible to m.