Homework5

Pablo

08/02/2019

Contents

Question 8.1

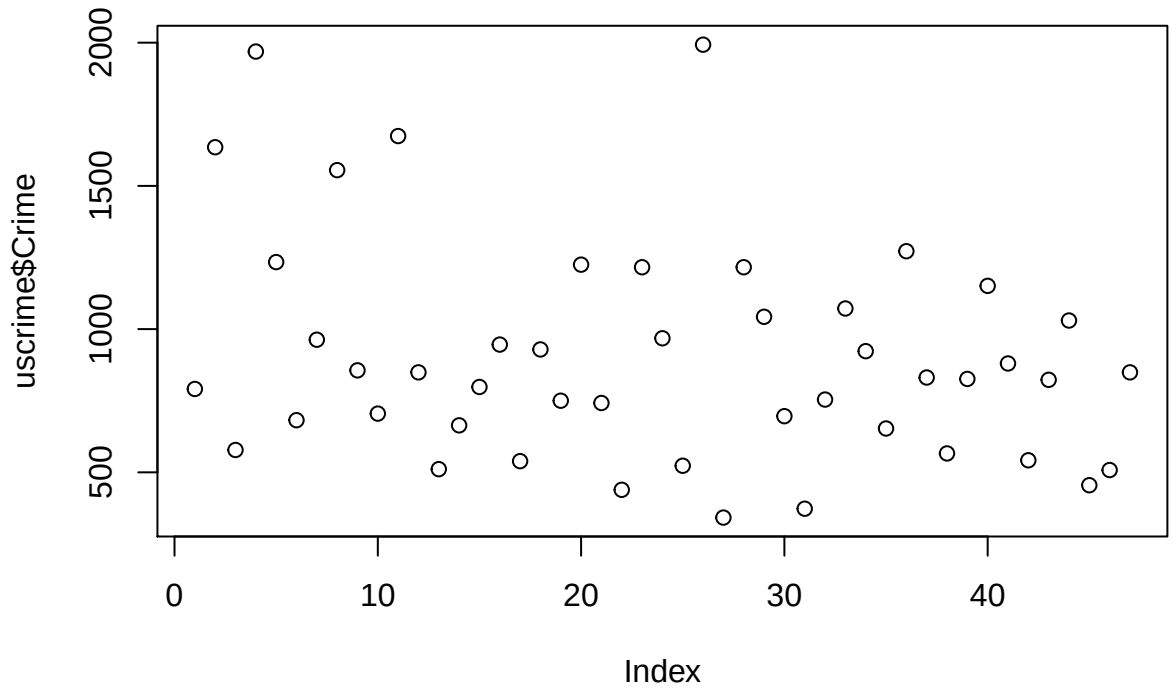
I work in analytics in a telephone company and the linear regression is a easy model to use in a production environment and easy to understand, so we use it regularly. For this example, I have a nice example. In our telephone company, the prepaid users are the majority of users and they make the most 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.

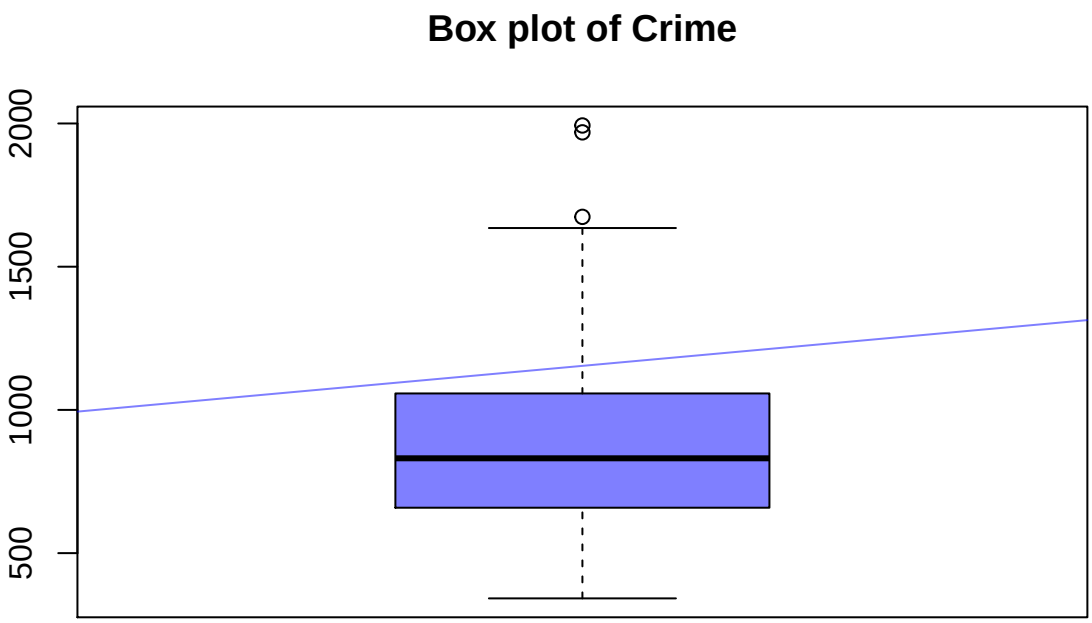
Question 8.2

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)
```




```
# 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))
```




```
# 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`.

```
# 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.

```
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`

```
# 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)
sigma(better_model_2)/mean(uscrime$Crime)
```

```
## [1] 0.2217359
```

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 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.