Lab work 1

Artem Storozhuk

DATA

First of all, we need to read our data (apparently)

CHECKING FILES IN DOWNLOADED ZIP ARCHIVE

```
filenames = list.files(path='./quantquote-daily-sp500-83986/daily/', full.names=TRUE)[21:30]
filenames
```

```
##  [1] "./quantquote-daily-sp500-83986/daily/table_agn.csv"
##  [2] "./quantquote-daily-sp500-83986/daily/table_aig.csv"
##  [3] "./quantquote-daily-sp500-83986/daily/table_aiv.csv"
##  [4] "./quantquote-daily-sp500-83986/daily/table_aiz.csv"
##  [5] "./quantquote-daily-sp500-83986/daily/table_akam.csv"
##  [6] "./quantquote-daily-sp500-83986/daily/table_all.csv"
##  [7] "./quantquote-daily-sp500-83986/daily/table_altr.csv"
##  [8] "./quantquote-daily-sp500-83986/daily/table_alxn.csv"
##  [9] "./quantquote-daily-sp500-83986/daily/table_amat.csv"
## [10] "./quantquote-daily-sp500-83986/daily/table_amd.csv"
```

READING FILES, TAKING OUT FIRST AND SIXTH COLUMNS AND PUTTING THEM INTO THE OTHER FRAME

```
datalist = lapply(filenames, function(x) {
  x0 <- read.csv(file=x, header=F)[, c(1, 6)];
  colnames(x0) <- c("data", unlist(strsplit(x,"[_.]"))[3]);
  x0
})
```

There wont be any output, because the output is way too big (im sure you don‘t want to scroll all of it)

MERGING FRAMES INTO ONE

```
y <- Reduce(function(x, y) { merge(x, y, by='data') }, datalist)
head(y,5)
```

```
##       data     agn     aig     aiv     aiz  akam     all    altr   alxn    amat
## 1 20040205 41.1405 1127.78 15.6552 21.5306 14.48 34.6745 20.4520 5.0400 18.3241
## 2 20040206 41.9096 1142.97 15.9251 21.5653 15.04 35.5850 21.5611 5.0625 19.1730
## 3 20040209 42.0982 1137.59 15.7857 21.2523 15.23 35.4167 21.4389 5.2450 18.9672
## 4 20040210 42.5383 1162.91 15.7767 21.1392 15.04 35.6998 21.5046 5.5675 18.8729
## 5 20040211 42.5383 1176.83 15.6552 21.0610 15.12 35.9064 21.8900 5.5375 19.0787
##     amd
## 1 14.18
## 2 14.91
## 3 15.06
## 4 15.53
## 5 15.68
```

CREATING A FILE WITH DATA FOR REGRESSION MODEL

```
Data <- y[-nrow(y), -1]
Data$alxn <- y$alxn[-1]
head(Data, 5)
```

```
##       agn     aig     aiv     aiz  akam     all    altr   alxn    amat   amd
## 1 41.1405 1127.78 15.6552 21.5306 14.48 34.6745 20.4520 5.0625 18.3241 14.18
## 2 41.9096 1142.97 15.9251 21.5653 15.04 35.5850 21.5611 5.2450 19.1730 14.91
## 3 42.0982 1137.59 15.7857 21.2523 15.23 35.4167 21.4389 5.5675 18.9672 15.06
## 4 42.5383 1162.91 15.7767 21.1392 15.04 35.6998 21.5046 5.5375 18.8729 15.53
## 5 42.5383 1176.83 15.6552 21.0610 15.12 35.9064 21.8900 5.4725 19.0787 15.68
```

TRAINING MODELS

The time has come to train models!

MODEL TRAINED ON LAST FIFTY OBSERVATIONS

```
nn <- nrow(Data)
model1 <- lm(alxn~.-alxn,data=Data[(nn-50):nn,])
summary(model1)
```

```
##
## Call:
## lm(formula = alxn ~ . - alxn, data = Data[(nn - 50):nn, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9679 -2.0519  0.2954  1.2041  7.2325
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -145.5183    27.1559  -5.359 3.52e-06 ***
## agn           -0.2162     0.1632  -1.325 0.192672
## aig           -0.3329     0.9445  -0.352 0.726319
## aiv            0.1639     0.8160   0.201 0.841773
## aiz            1.6194     1.0997   1.473 0.148509
## akam           1.7538     0.4524   3.877 0.000375 ***
## all            0.8870     0.7469   1.188 0.241835
## altr           1.0790     1.0955   0.985 0.330457
## amat           2.8457     1.8831   1.511 0.138402
## amd           -2.2991     2.5357  -0.907 0.369872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.778 on 41 degrees of freedom
## Multiple R-squared:  0.9369, Adjusted R-squared:  0.923
## F-statistic: 67.59 on 9 and 41 DF,  p-value: < 2.2e-16
```

R^2 = 0.9369 - it is a great result for prediction. We can see that “akam” has a significant influence, p-value is close to 0, which means that regressors are affecting response.

MODEL TRAINED ON SIGNIFICANT REGRESSORS

```
model2 <- lm(alxn~akam, data=Data[(nn-50):nn,])
summary(model2)
```

```
##
## Call:
## lm(formula = alxn ~ akam, data = Data[(nn - 50):nn, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.092  -3.227   1.279   5.011  11.727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -80.2317    24.0466  -3.337  0.00162 **
## akam          4.0942     0.5414   7.563 8.96e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.87 on 49 degrees of freedom
## Multiple R-squared:  0.5386, Adjusted R-squared:  0.5292
## F-statistic:  57.2 on 1 and 49 DF,  p-value: 8.958e-10
```

R^2 = 0.53 - it is a really bad result, which means that the other regressors have an influence, but not as big as "akam" regressor.

MODEL TRAINED ON ALL DATA

```
model3 <- lm(alxn~.-alxn, data=Data[20:nn,])
summary(model3)
```

```
##
## Call:
## lm(formula = alxn ~ . - alxn, data = Data[20:nn, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.110  -5.629  -0.101   4.167  36.055
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -32.196194   1.661365 -19.379  < 2e-16 ***
## agn           1.159186   0.024769  46.800  < 2e-16 ***
## aig          -0.009952   0.001190  -8.364  < 2e-16 ***
## aiv          -0.276629   0.109590  -2.524   0.0117 *
## aiz          -0.292682   0.037552  -7.794 9.65e-15 ***
## akam         -0.294419   0.025949 -11.346  < 2e-16 ***
## all           1.206399   0.059420  20.303  < 2e-16 ***
## altr          0.684441   0.060619  11.291  < 2e-16 ***
## amat         -2.298739   0.163843 -14.030  < 2e-16 ***
## amd          -0.533250   0.036979 -14.420  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9.248 on 2365 degrees of freedom
## Multiple R-squared:  0.9172, Adjusted R-squared:  0.9169
## F-statistic:  2912 on 9 and 2365 DF,  p-value: < 2.2e-16
```

Rˆ2 = 0.91 - it is still a good result, but still we can see a remarkable deterioration

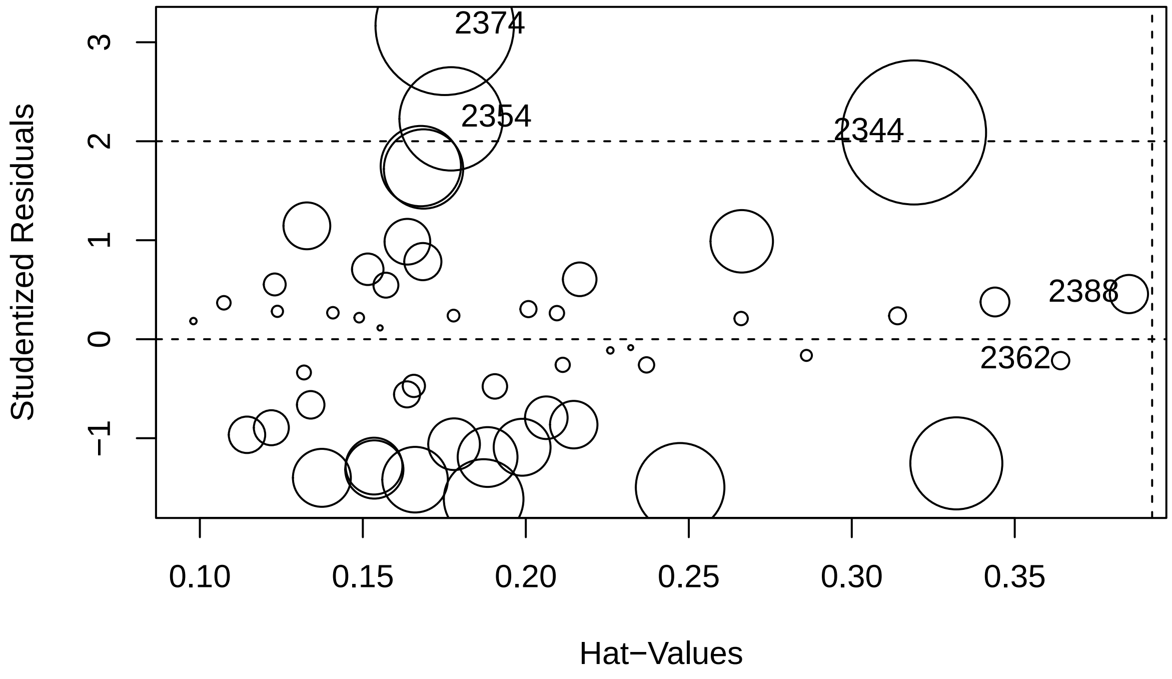
PREDICTION

BUBBLE CHART

```
library(car)
```

```
## Loading required package: carData
```

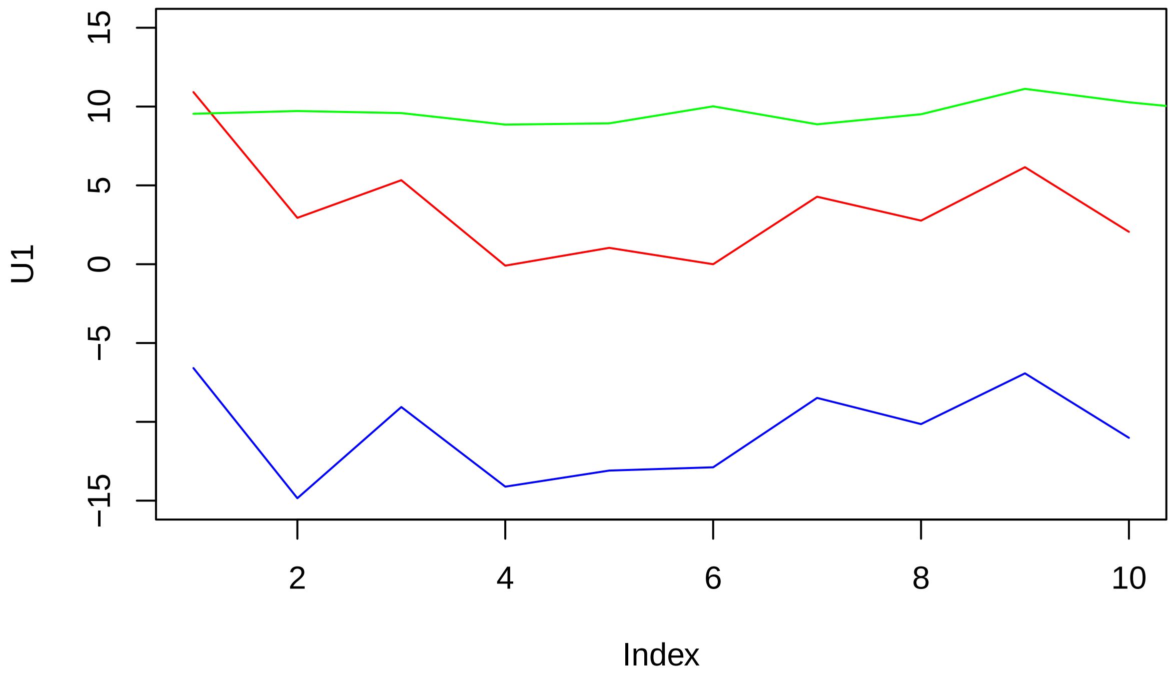
```
influencePlot(model1)
```



```
##          StudRes       Hat      CookD
## 2344  2.0887155 0.3191076 0.18896562
## 2354  2.2258856 0.1770752 0.09723294
## 2362 -0.2163368 0.3640802 0.00274329
## 2374  3.1660466 0.1751124 0.17440677
## 2388  0.4561850 0.3850348 0.01328623
```

PLOT

```
U1 <- Data$alxn[(nn-60):(nn-51)]-predict(model1,Data[(nn-60):(nn-51),])
U2 <- Data$alxn[(nn-60):(nn-51)]-predict(model2,Data[(nn-60):(nn-51),])
U3 <- Data$alxn[1:19]-predict(model3, Data[1:19,])
plot(U1,type="l",col="red",ylim=c(-15,15))
lines(U2, col='blue')
lines(U3, col='green')
```



We can see that model trained on last fifty observations gives us a better result, also we can see that there exists a systematical error which means that we have some problems with regression bias