Московский государственный университет имени М. В. Ломоносова
Факультет Вычислительной математики и кибернетики

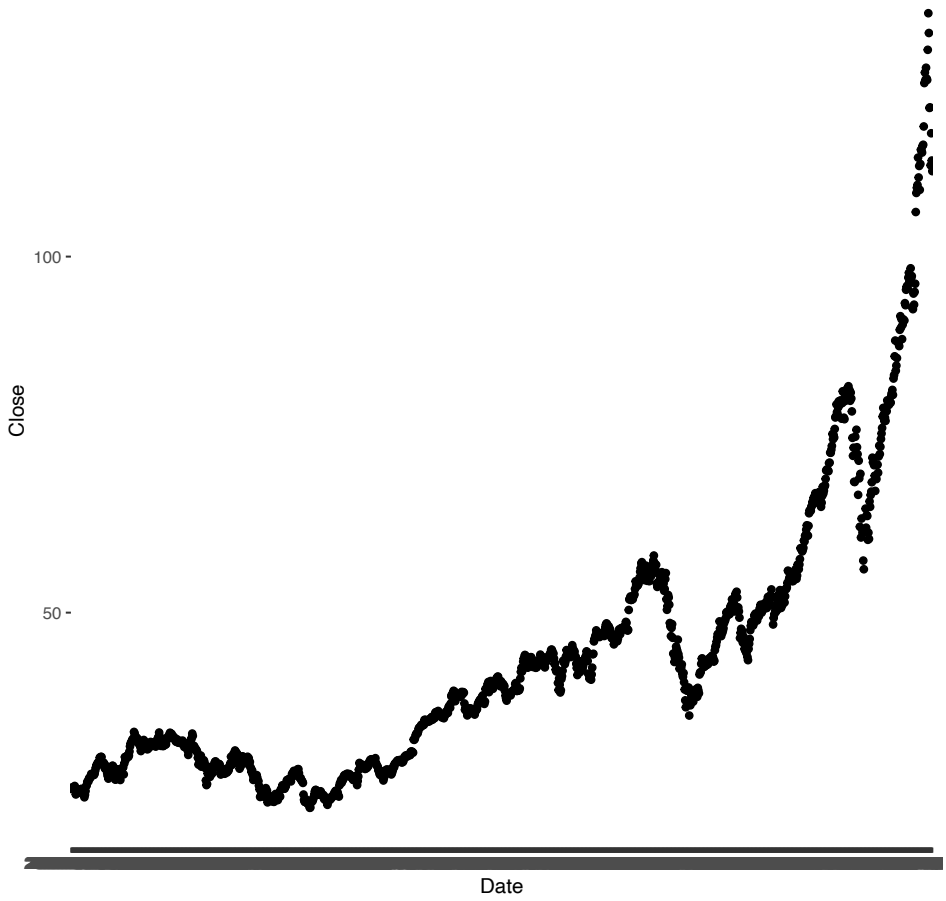
Никонов Максим Викторович
316 группа

2020

Задание 2: Проверить наличие мультиколлинеарности в собственных данных. Изучить фактор инфляции дисперсии VIF

Построим ggplot Close от Date

```
ggplot(data = data, aes(x = Date, y = Close)) + geom_point()
```



Зададим в переменную poly нашу полиномиальную регрессию цены от количества

```
poly <- lm(Volume ~ Close + I(Close^2) + I(Close^3),
data)
summary(poly)
```

Данные из протокола

```
Residuals:
       Min         1Q     Median         3Q        Max
-129102670  -44700887  -14815297   28134861  463542440

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.904e+08  2.324e+07  16.800  < 2e-16 ***
Close       -1.129e+07  1.295e+06  -8.720  < 2e-16 ***
I(Close^2)   1.427e+05  2.142e+04   6.664 3.72e-11 ***
I(Close^3)  -5.154e+02  1.054e+02  -4.890 1.12e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67410000 on 1506 degrees of freedom
Multiple R-squared:  0.1336, Adjusted R-squared:  0.1319
              F-statistic: 77.41 on 3 and 1506 DF,  p-value: < 2.2e-16
```

Проверим фактор дисперсии VIF

```
vif(poly)
```

Данные из протокола

```
> vif(poly)
     Close I(Close^2) I(Close^3)
  207.9222   875.6382   265.2496
```

Получаем, что ни один коэфициент не значим, но регрессия значима, что является одним из признаков мультиколлинеарности

Итоговый вид программы

```
library("dplyr")
library("ggplot2")
library("car")

head <- read.csv("/Users/Nikon/Desktop/CMC MSU/MC/5
sem/Data/AAPL.csv",
                 header = TRUE)
num <- random <- sample(1:2, nrow(head), replace=TRUE)
num2 <- random <- sample(1:2, nrow(head), replace=TRUE)
head <- cbind(head, num)
head <- cbind(head, num2)
data <- head[-c(2,3,4,6)]
data <- mutate_each(data, "factor", Date)

ggplot(data = data, aes(x = Date, y = Close)) +
geom_point()
poly <- lm(Volume ~ Close + I(Close^2) + I(Close^3),
data)
summary(poly)
vif(poly)
```

Используемые packages. ggplot2, dplyr, car
Тестирование. Не требует
Неразрешенные вопросы. Нет
Новые функции. vif
Статус компиляции. ОК. Данные из протокола:

```
> ^M
^[[1m^[[7m%^[[27m^[[1m^[[0m
^M ^M^[]7;file://MBP-
Nikon.Dlink/Users/Nikon/Desktop/CMC%20MSU/MC/5%20sem/R/tz10/1^G^M^[[0m^[[27m^[[24m^[[JNik
on@MBP-Nikon 1 % ^[[K^[[?2004he^Hexit^[[?2004l^M^M

                        Script done on Wed Nov 18 15:26:25 2020
```

Задание 3-4: Изучить форму связи между переменными для выбранных данных.
Изучить взаимодействие между предикторами для выбранных данных.

Мы рассмотрим каким образом на объем продаж влияет сезонность (время года и год) и цена акции

Для иллюстрации возьмем столбец Date по годам

```
data$Date <- as.POSIXct(data$Date, format = "%Y")
```

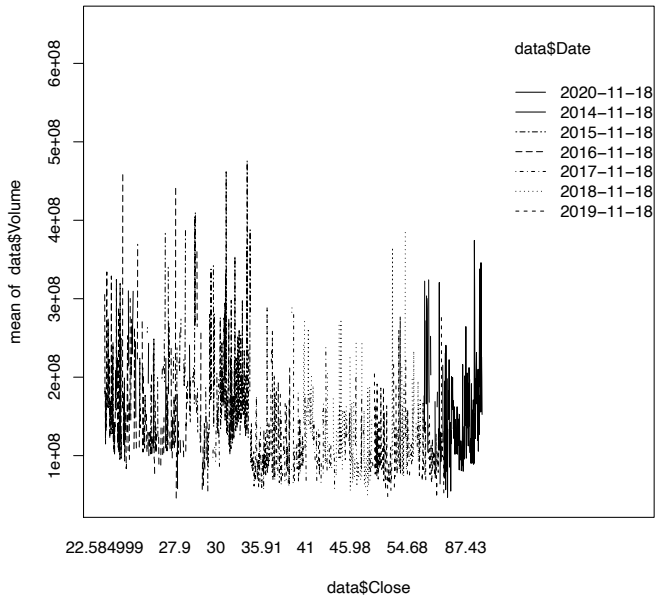
Запишем временной ряд

```
mk <- lm(data$Volume ~ data$Date + data$Close)
```

Для иллюстрации связи между переменными используем **interaction.plot**

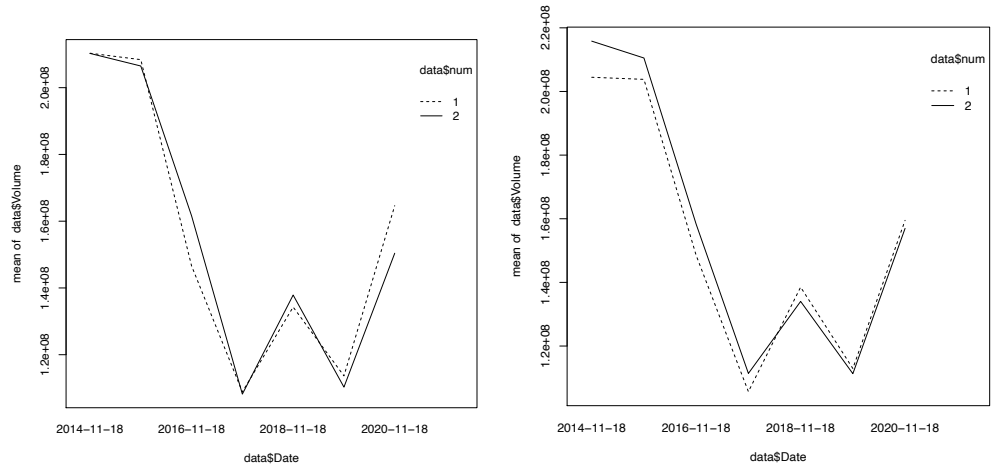
```
interaction.plot(data$Close, data$Date, data$Volume)
```

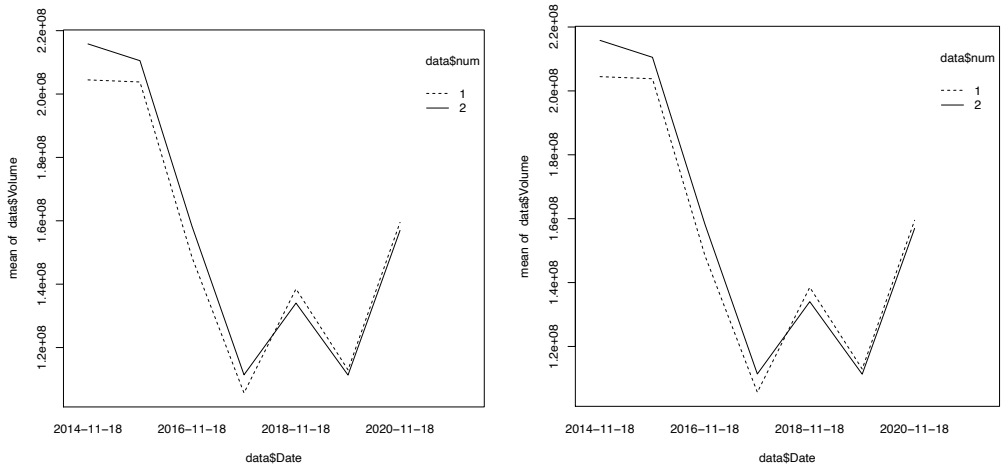
Получим картину очень похожую на Гауссовский шум, без учета значимости переменных



Поскольку больше предикторов в данных нет, то для иллюстрации связи между предикторами возьмем в качестве предиктора столбец num, который постоянно заново генерируется

```
interaction.plot(data$Date, data$num, data$Volume)
```





На тех участках, где прямые были параллельны, очевидно, есть линейная зависимость, на остальных участках зависимости не наблюдается

Итоговый вид программы

```
library(faraway)
library("dplyr")


head <- read.csv("/Users/Nikon/Desktop/CMC MSU/MC/5 sem/Data/AAPL.csv",
                 header = TRUE)
num <- random <- sample(1:2, nrow(head), replace=TRUE)
num2 <- random <- sample(1:2, nrow(head), replace=TRUE)
head <- cbind(head, num)
head <- cbind(head, num2)
data <- head[-c(2,3,4,6)]
data$Date <- as.POSIXct(data$Date, format = "%Y")
data <- mutate_each(data, "factor", Date, Close)
mk <- lm(data$Volume ~ data$Date + data$Close)
interaction.plot(data$Close, data$Date, data$Volume)
interaction.plot(data$Date, data$num, data$Volume)
```

Используемые packages. dplyr, faraway
Тестирование. Не требует
Неразрешенные вопросы. Нет
Новые функции. interaction.plot
Статус компиляции. ОК. Данные из протокола:

```
^[[1m^[[7m%^[[27m^[[1m^[[0m
^M ^M^[]7;file://MBP-
Nikon.Dlink/Users/Nikon/Desktop/CMC%20MSU/MC/5%20sem/R/tz10/2^G^M^[[0m^[[27m^[[24m^[[JNik
on@MBP-Nikon 2 % ^[[K^[[?2004he^Hexit^[[?2004l^M^M

                        Script done on Wed Nov 18 15:44:58 2020
```

Задание 5: Построить график автокорреляцоионной функции

В переменную state запишем временной ряд

```
mk <- lm(data$Volume ~ data$Date + data$Close, state)
```

В переменную M cor от временного ряда

```
M <- cor(state)
```

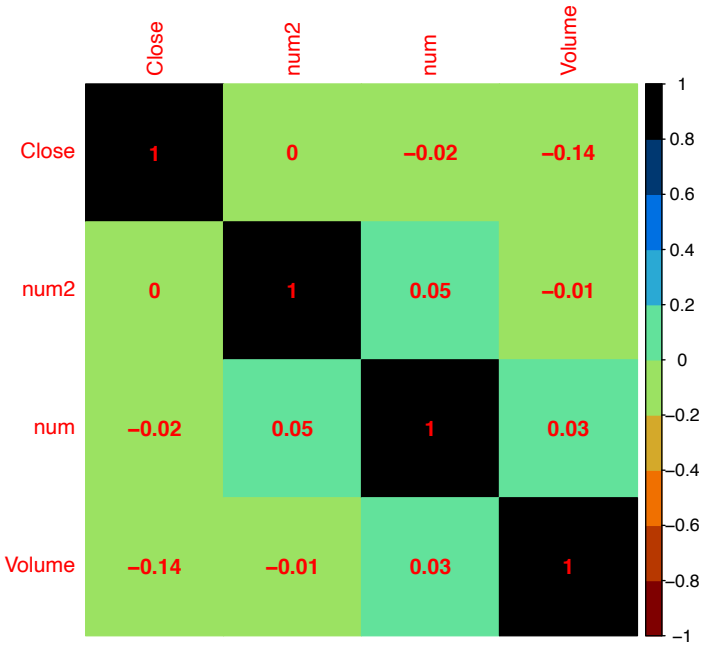
Зададим палитру цветов в переменную colfor

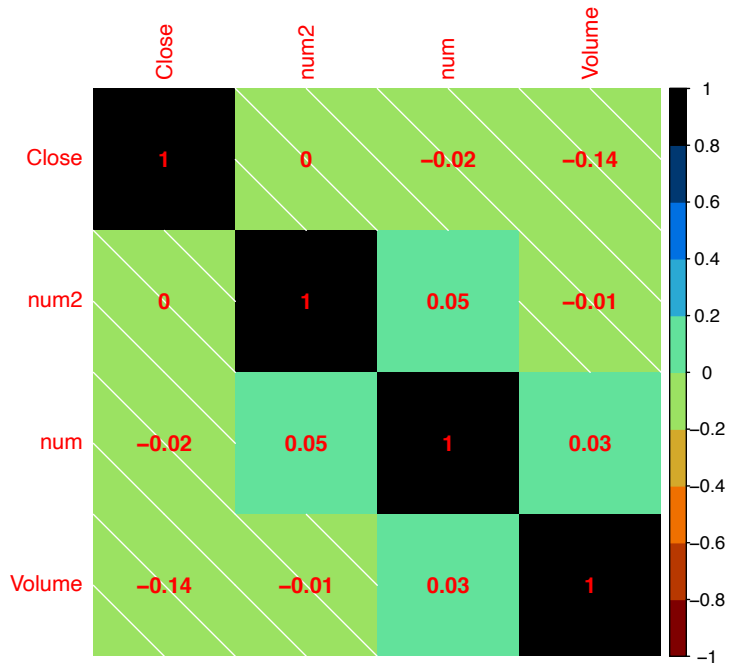
```
col4 <- colorRampPalette(c("#7F0000","#FF7F00", "#7FFF7F",
"#007FFF","#000000"))
```

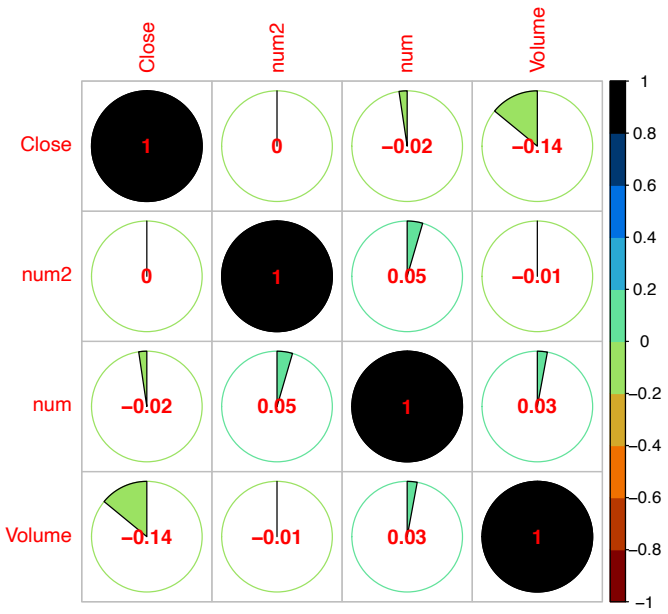
Далее наглядно покажет корреляцию данных несколькими методами

```
corrplot(M, method = "color", col=col4(10), cl.length =
11,
         order = "AOE", addCoef.col = "red")
corrplot(M, method="shade", col=col4(10), cl.length = 11,
         order = "AOE", addCoef.col = "red")
corrplot(M, method="pi", col=col4(10), cl.length = 11,
         order = "AOE", addCoef.col = "red")
```

Соответсвующие иллюстарции см. ниже


Close
num2
num
Volume
Close
num2
num
Volume
1
0
−0.02
−0.14
0
1
0.05
−0.01
−0.02
0.05
1
0.03
−0.14
−0.01
0.03
1
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

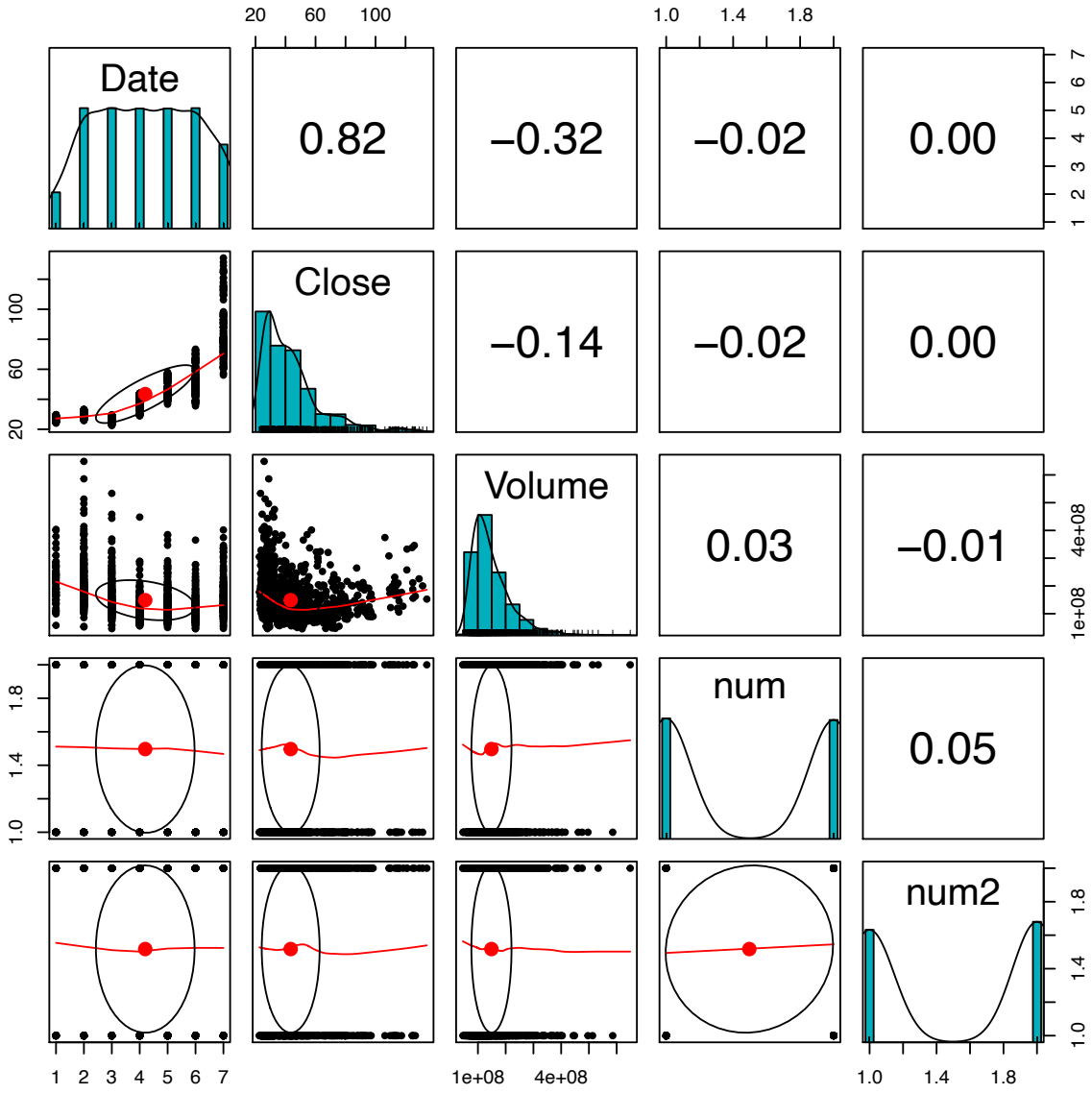

Close
num2
num
Volume
Close
num2
num
Volume
1
0
−0.02
−0.14
0
1
0.05
−0.01
−0.02
0.05
1
0.03
−0.14
−0.01
0.03
1
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1



Из полученных данных можно сделать вывод, что данные практически не коррелируют, однако коррелируют полностью данные расположенные по диагонали, что логично, ведь это одни и те же данные

В конце проиллюстрируем все данные через **pairs.panels**, все данные в коридоре

```
pairs.panels(data,
             method = "pearson",
             hist.col = "#00AFBB",
             density = TRUE,
             ellipses = TRUE
)
```



Итоговый вид программы

```
library("dplyr")
library(psych)
library(corrplot)

head <- read.csv("/Users/Nikon/Desktop/CMC MSU/MC/5 sem/Data/AAPL.csv",
                 header = TRUE)
num <- random <- sample(1:2, nrow(head), replace=TRUE)
num2 <- random <- sample(1:2, nrow(head), replace=TRUE)
head <- cbind(head, num)
head <- cbind(head, num2)
data <- head[-c(2,3,4,6)]
data$Date <- as.POSIXct(data$Date, format = "%Y")
data <- mutate_each(data, "factor", Date)
state <- as.data.frame(data[,c("Close", "Volume", "num", "num2")])
mk <- lm(data$Volume ~ data$Date + data$Close, state)
M <- cor(state)
col4 <- colorRampPalette(c("#7F0000","#FF7F00", "#7FFF7F", "#007FFF","#000000"))
corrplot(M, method = "color", col=col4(10), cl.length = 11,
         order = "AOE", addCoef.col = "red")
corrplot(M, method="shade", col=col4(10), cl.length = 11,
         order = "AOE", addCoef.col = "red")
corrplot(M, method="pi", col=col4(10), cl.length = 11,
         order = "AOE", addCoef.col = "red")

pairs.panels(data,
             method = "pearson",
             hist.col = "#00AFBB",
             density = TRUE,
             ellipses = TRUE
```

Используемые packages. psych, dplyr, corplot
Тестирование. Не требует
Неразрешенные вопросы. Нет
Новые функции. corrplot, pairs.panels
Статус компиляции. ОК. Данные из протокола:

```
Desktop/CMC%20MSU/MC/5%20sem/R/tz10/5^G^M^[[0m^[[27m^[[24m^[[JNikon@MBP-Nikon 5 %
^[[K^[[?2004h R --no-save <task5.r ^[[22Dexit                ^[[17D^[[?2004l^M^M

                        Script done on Wed Nov 18 16:07:22 2020
```