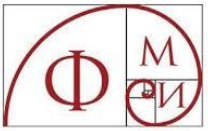
Софийски университет “Св. Климент Охридски”
Факултет по математика и информатика



Курсов проект по Статистика и емпирични методи летен семестър 2020/2021

Тема - Нагласите на студентите за предсрочните избори 2021

Изработил: Веселин Славов Тодоров
Специалност: Информационни системи
Курс: 2
Факултетен номер: 71923

Гр. София
Май 2021

Част 1. Избор на данни. Задаване на цели на проекта.

Поради неуспешното съставяне на правителство, през лятото на 2021 ще се проведат предсрочни избори. И тъй като това управлението на държавата е важна част от живота ни, ме накара да избера тази тема за своя проект. За да мога да събера мнения за своя проект създадох анкета, която разпространих в група студенти от 1ви до 4ти курс. Важно е да се каже, че анкетата няма за цел политическа агитация, а мнението на младите хора относно ситуацията в която се намира нашето управление.

Главната цел на този проект е да се определи дали младите вярват на сегашните управляващи и дали смятат че е нужна промяна. Ще намеря отговор и на въпросите:
- Доволни ли са от избраното правителството избрано на 4.4.2021?
- Какво доверие имат на сегашното правителство?
- Какво доверие имат към президентската институция?
- Мнението за броя на изборните секции в чужбина?

Анкетата се състои от 5 въпроса. Общият брой на събраните отговори е 43.
Най – напред, последователно ще направя анализ на всяка променлива (всеки
въпрос). След това ще направя проверка за зависимост между отделните
въпроси. И накрая ще направя заключение, като отговоря на въпросите по – горе.

Самата анкета може да намерите, като отворите следния линк:
https://forms.gle/NKK9CS6TFJ34eZo96

Част 2. Описателен анализ на променливите. Анализ на взаимодействието им.

За да можем да използваме данните и да ги изследваме с помощта на R, първо трябва да ги вкараме в програмата. Тъй като ще въведа таблицата с отговори, която е с формат .xlsx, то първо трябва да инсталирам пакетите за обработка на такива файлове. Следният код показва как става въвеждането на данните от таблицата, като функцията library(readxl) зарежда пакетите от библиотеката readxl.

```
#Добавяне на данните от таблицата
answers <- read_excel("C:\\Users\\vesel\\Desktop\\Statistics-Practicum\\DebatesProject\\Answers.xlsx")
colnames(answers) <- paste("Question", 0:ncol(answers), sep="")
str(answers)
View(answers)
```

Получава се структура от данни tibble, която е специален случай на data.frame и е предназначена за набора от пакети tidyverse. За разлика от data frame, при tibble редовете и колоните може да са с различна дължина. За по – добра четливост задавам имена на колоните: “Question”, като е последвана от номера на колоната (въпрос 0 е клеймото за време).

2.1 Анализ на едномерна променлива – графично представяне и статистики
за локацията и дисперсията.
Анкетата съдържа 5 въпроса, които включват три дискретни числови величини и останалите две са категорийни.
● Категорийните променливи носят по – малко информация, отколкото числовите. Но за сметка на това представят по – голяма стабилност за прогнозните модели. За анализ на този вид променливи трябва да представим данните в таблици и след това да ги изобразим графично, като най – подходящ инструмент за тази цел е barplot. Може да се използва и pie.

● Числовите променливи ще изследвам по следния начин:
 Оценка на центъра (локацията) на разпределение.
- средна стойност(очакване)
- медиана
- мода - Оценка на вариацията на разпределение.
- Обхват (Range)
- Вариация (дисперсия) и стандартно отклонение
- The five number summary

Ще изобразя графично разпределението на отговорите на съответните въпроси, представляващи числови променливи. А след изследването на въпросите поотделно ще направя и тест, който ще покаже дали са нормално разпределени, или не

Въпрос 1: Ще гласувате ли на предсрочните избори през лятото на 2021?

Този въпрос представлява категорийна величина, тъй като анкетираните трябва да изберат от трите отговора спрямо нагласите им да гласуват на предсрочните избори, т.е.
тяхната нагласа попада в 3 различни категории.

Код на R:

```
# Въпрос 1: Ще гласувате ли на предсрочните избори през лятото на 2021? - Категорийна променлива

utf8_print(answers$Question1)
#[1] "Да"           "Да"           "Да"           "Да"           "Да"           "Да"           "Да"           "Да"           "Да"
#[10] "Да"           "Да"           "Не съм решил" "Да"           "Да"           "Да"           "Да"           "Да"           "Не"
#[19] "Да"           "Не"           "Да"           "Да"           "Да"           "Да"           "Да"           "Да"           "Да"
#[28] "Да"           "Да"           "Да"           "Да"           "Да"           "Не"           "Да"           "Да"           "Да"
#[37] "Да"           "Да"           "Не съм решил" "Да"           "Да"           "Да"           "Да"

table_q1 <- table(answers$Question1)
table_q1

#Да    Не    Не съм решил
#38    3     2

# Графично представяне
barplot(round(prop.table(table_q1)*100, 2), col = c("green", "red", "red"), main = "Ще гласувате ли на предсрочните избори?", xlab = "Отговор",
        ylab = "Проценти", ylim = c(0, 100))
```

Графика:

Отговорите на този въпрос са 43.
От графиката се вижда, че от анкетираните:

- около 90% са твърдо решени да гласуват
- малко над 5% са твърдо решени да не гласуват
- малко под 5% не са решили дали да гласуват

Въпрос 2: Колко процента смятате че избрания на 04.04.2021 парламент е достатъчно добър, за да съществува? (0-100)

Вторият въпрос представлява числова дискретна променлива, поради причината че искаме целочислени стойности между 0 и 100, която ще изследвам, следвайки стъпките, споменати по – горе.

Започвам с въвеждането на всички отговори във вектор. След това ще намеря средната стойност, медианата и модата, както и центъра на разпределение:

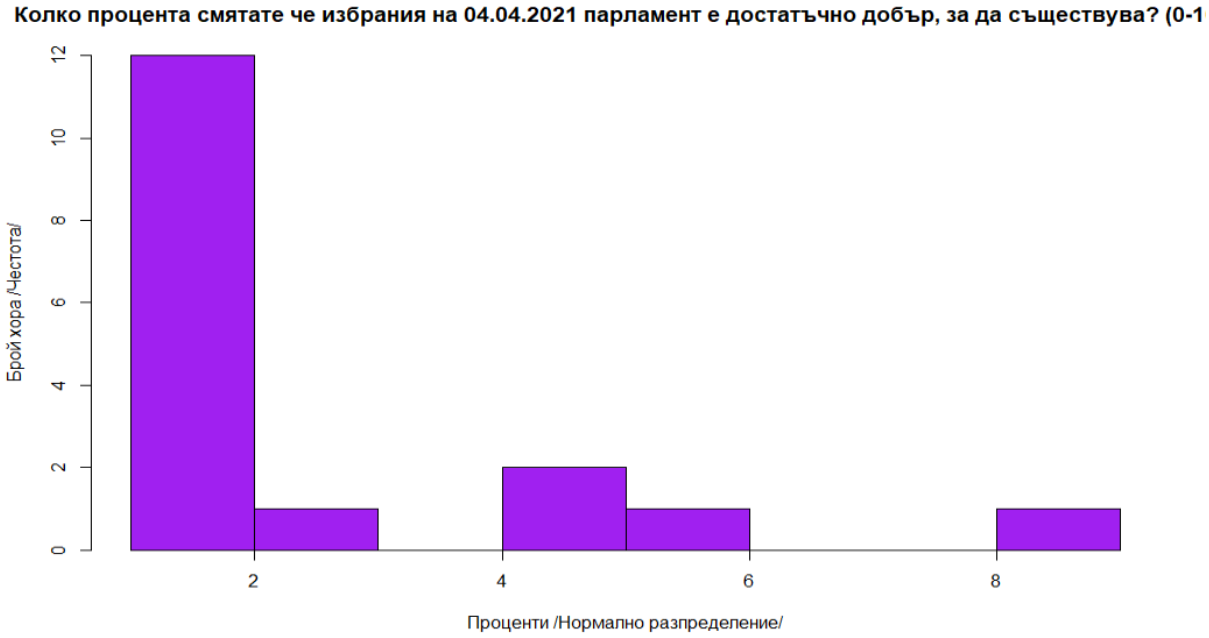
```
#Въпрос 2: Колко процента смятате че избрания на 04.04.2021 парламент е достатъчно добър, за да съществува?
percent_acceptance_q2 <- c(0, 0, 0, 0, 0, 0, 0, 0, 0,
                           50, 50, 50, 50, 50, 30, 30, 30, 30, 30, 40, 40, 40, 40, 20, 20, 20,
                           10, 10, 1, 1, 69, 69, 60, 49, 40, 50, 80, 70, 100, 12, 33, 63, 51)

percent_acceptance_q2 <- sort(percent_acceptance_q2)
length(percent_acceptance_q2)
#[1] 43

mean(percent_acceptance_q2) # средна стойност
# [1] 32.27907
median(percent_acceptance_q2) # медиана
# [1] 30
table_q2 <- table(percent_acceptance_q2)        # мода - най - често срещана стойност
names(table_q2)[table_q2 == max(table_q2)]
# [1] "0"

# Следващите две функции описват центъра на разпределение
summary(percent_acceptance_q2)
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#    0.00    5.50   30.00   32.28   50.00  100.00

quantile(percent_acceptance_q2, prob = seq(0.1, 0.9, by = 0.1))
# 10%  20%  30%  40%  50%  60%  70%  80%  90%
# 0.0  0.4 11.2 28.0 30.0 40.0 50.0 50.0 67.8
```

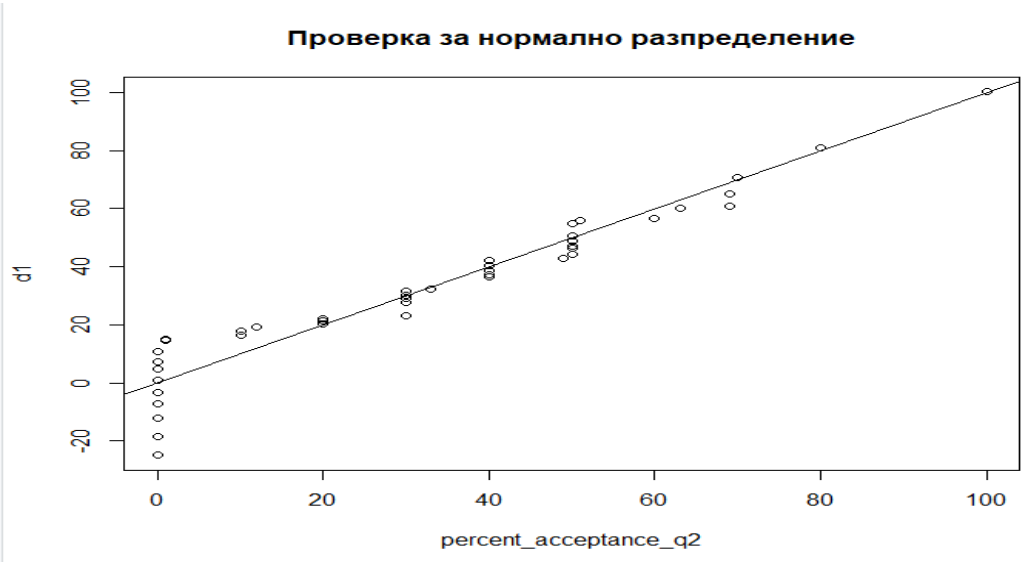
Сега ще опиша вариацията (дисперсията) и стандартното отклонение, което ще покаже колко далече са наблюденията от очакването. Освен това чрез хистограма ще изобразя разпределението на данните графично.

Код на R:

```
# Вариация (дисперсия) на разпределението

range(percent_acceptance_q2)   # range - показва най - голямата и най - малката стойност
# [1]  0 100
var(percent_acceptance_q2)     # дисперсия
# [1] 683.9203
sd(percent_acceptance_q2)      # стандартно отклонение
# [1] 26.15187
fivenum(percent_acceptance_q2)
# [1]   0.0   5.5  30.0  50.0 100.0

# Графично представяне
hist(table_q2, main = "Колко процента смятате че избрания на 04.04.2021 парламент е достатъчно добър, за да съществува? (0-100)",
     xlab = "Проценти /Нормално разпределение/",ylab = "Брой хора /честота/", col = "purple")

# Проверка за нормално разпределение
d1 <- rnorm(n = 10^2, mean = mean(percent_acceptance_q2), sd = sd(percent_acceptance_q2))
qqplot(percent_acceptance_q2, d1, main = "Проверка за нормално разпределение")
abline(a = 0, b = 1) # чертае линия
```

Графики:





Повечето от хората имат доверие към новоизбрания парламент по-малко от 20%, като най – честият отговор е 0%. Изчислявайки средната стойност от всички отговори, става ясно, че средно на човек се падат по около 32% доверие, което показва че рейтингът на новоизбрания парламент е доста нисък, което е тревожен факт. Освен това отговорите не са нормално разпределени, както се вижда от втората графика, но това ще докажем по – късно с помощта на shapiro.test.

Въпрос 3: Колко процента имате доверие на сегашния парламент?

Третият въпрос също представлява числова дискретна променлива, поради причината че искаме целочислени стойности между 0 и 100, която ще изследвам, следвайки стъпките, споменати по – горе. Започвам с въвеждането на всички отговори във вектор. След това ще намеря средната стойност, медианата и модата, както и центъра на разпределение:

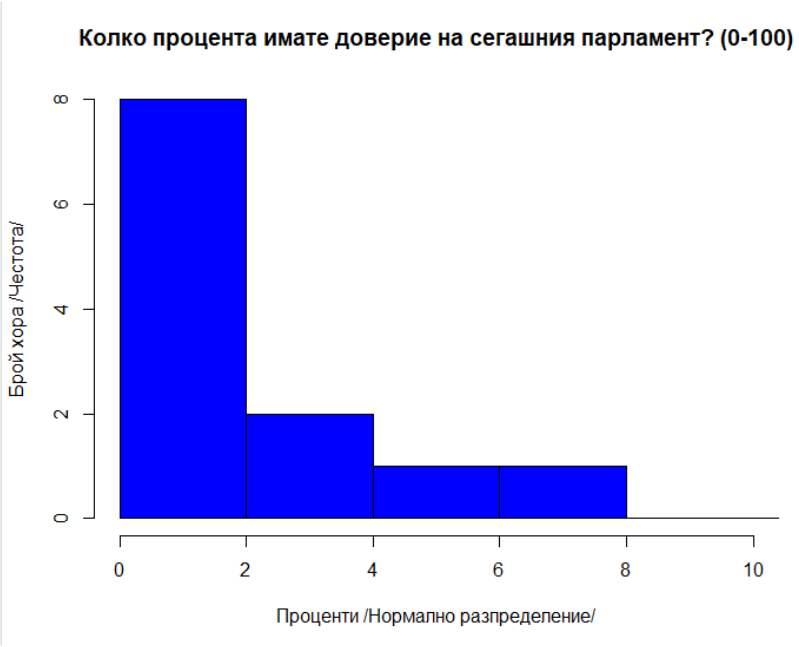
```
#Въпрос 3: Колко процента имате доверие на сегашния парламент? (0-100) - дискретна променлива
percent_acceptance_q3 <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                           1, 10, 10, 10, 10, 20, 20, 20, 20, 20, 20, 20,
                           25, 30, 30, 33, 40, 40, 40, 47, 50, 50, 50, 50, 50,
                           50, 53, 60, 69, 69)

percent_acceptance_q3 <- sort(percent_acceptance_q3)
length(percent_acceptance_q3)
#[1] 43

mean(percent_acceptance_q3) # средна стойност
# [1] 23.65116
median(percent_acceptance_q3) # медиана
# [1] 20
table_q3 <- table(percent_acceptance_q3)        # мода - най - често срещана стойност
names(table_q3)[table_q3 == max(table_q3)]
# [1] "0"

# Следващите две функции описват центъра на разпределение
summary(percent_acceptance_q3)
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#    0.00    5.50   30.00   32.28   50.00  100.00

quantile(percent_acceptance_q3, prob = seq(0.1, 0.9, by = 0.1))
#   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#   0.00    0.00   20.00   23.65   43.50   69.00
```

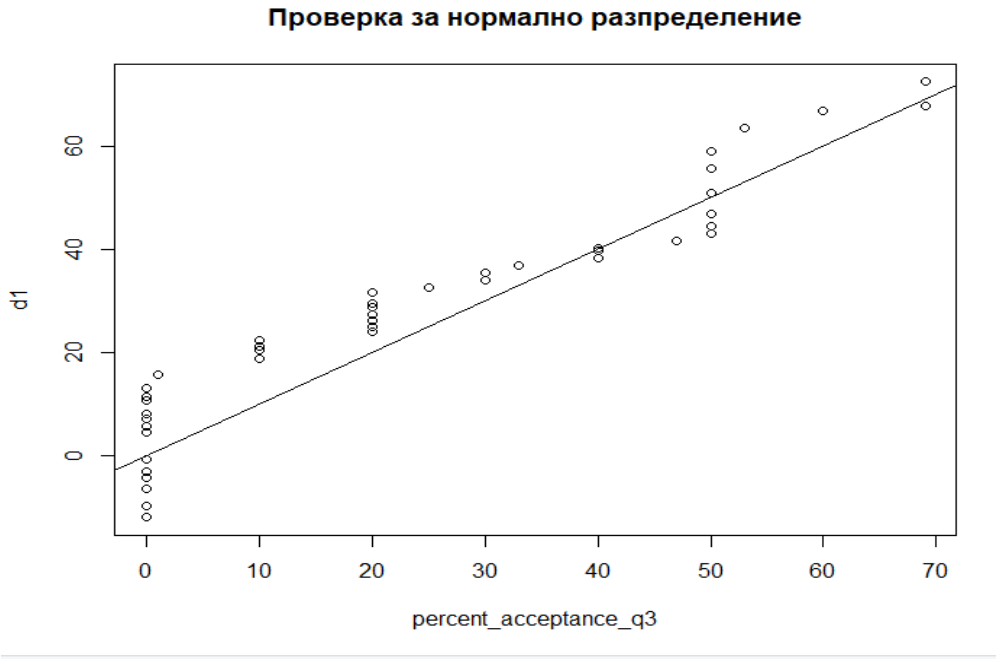
Сега ще опиша вариацията (дисперсията) и стандартното отклонение, което ще покаже колко далече са наблюденията от очакването. Освен това чрез хистограма ще изобразя разпределението на данните графично.

Код на R:

```r
# Вариация (дисперсия) на разпределението
range(percent_acceptance_q3)   # range - показва най - голямата и най - малката стойност
# [1]  0 69
var(percent_acceptance_q3)     # дисперсия
# [1] 490.5183
sd(percent_acceptance_q3)      # стандартно отклонение
# [1] 22.14765
fivenum(percent_acceptance_q3)
# [1]   0.0   0.0  20.0  43.5 69.0

# Графично представяне
hist(table_q3, main = "Колко процента имате доверие на сегашния парламент? (0-100)", xlab = "Проценти /Нормално разпределение/",
     ylab = "Брой хора /честота/", col = "blue", xlim = c(0,15))

# Проверка за нормално разпределение
d1 <- rnorm(n = 10^2, mean = mean(percent_acceptance_q3), sd = sd(percent_acceptance_q3))
qqplot(percent_acceptance_q3, d1, main = "Проверка за нормално разпределение")
abline(a = 0, b = 1) # чертае линия
```

Графики:





Основно хората имат много ниско доверие към управляващите, като най – честият отговор е 0%. Изчислявайки средната стойност от всички отговори, става ясно, че средно на човек се падат по около 23% доверие, което показва че доверието към управляващите е доста ниско, което означава че избирателите желаят промяна. Освен това отговорите не са нормално разпределени, както се вижда от втората графика, като това ще докажем по – късно с помощта на shapiro.test отново.

Въпрос 4: Колко процента имате доверие на президентът Румен Радев?

Четвъртия въпрос представлява числова дискретна променлива, поради причината че искаме целочислени стойности между 0 и 100, която ще изследвам, следвайки стъпките, споменати по – горе. Има за цел да се разбере доверието на избирателите към президента Румен Радев, което ще допринесе за резултатите на президентските избори през есента на 2021. Започвам с въвеждането на всички отговори във вектор. След това ще намеря средната стойност, медианата и модата, както и центъра на разпределение:

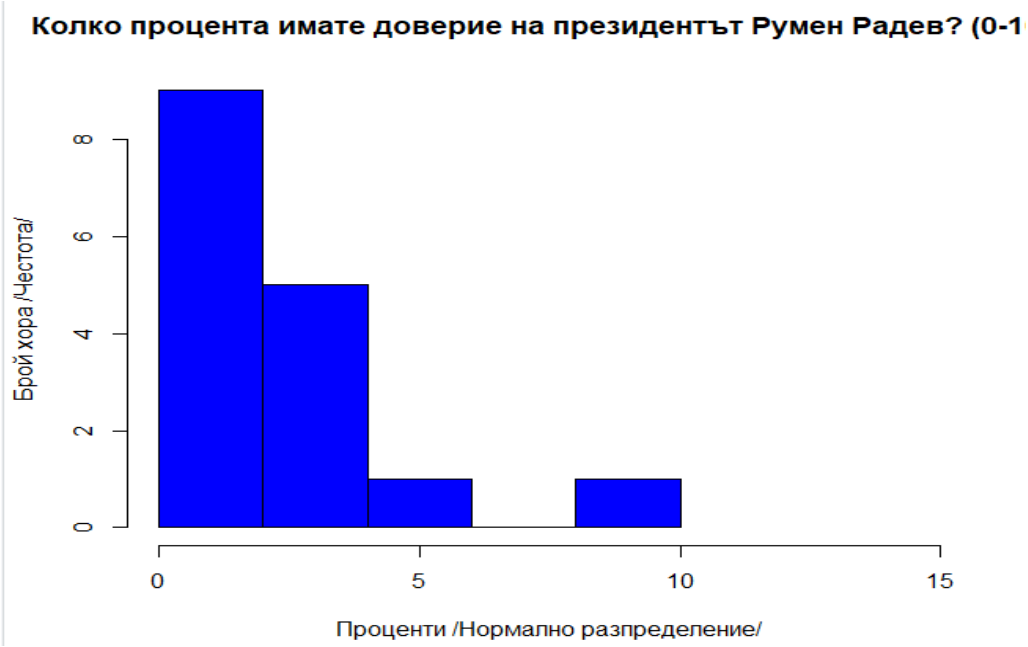
```r
#Въпрос 4: Колко процента имате доверие на президентът Румен Радев? (0-100) - Дискр
percent_acceptance_q4 <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 10, 10, 10, 10,
                           20, 20, 25, 30, 40, 40, 40, 40, 50, 50, 50, 51,
                           60, 60, 65, 65, 70, 70, 70, 70, 70, 75, 80, 80, 80,
                           90, 95, 100, 100, 100)

percent_acceptance_q4 <- sort(percent_acceptance_q4)
length(percent_acceptance_q4)
#[1] 43

mean(percent_acceptance_q4) # средна стойност
# [1] 44.09302
median(percent_acceptance_q4) # медиана
# [1] 50
table_q4 <- table(percent_acceptance_q4)          # мода - най - често срещана стойнс
names(table_q4)[table_q4 == max(table_q4)]
# [1] "0"

# Следващите две функции описват центъра на разпределение
summary(percent_acceptance_q4)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.00   10.00   50.00   44.09   70.00  100.00

quantile(percent_acceptance_q4, prob = seq(0.1, 0.9, by = 0.1))
# 10% 20% 30% 40% 50% 60% 70% 80% 90%
# 0   4  16  38  50  60  70  73  88

# Вариация (дисперсия) на разпределението
range(percent_acceptance_q4)    # range - показва най - голямата и най - малката стс
# [1]  0 100
var(percent_acceptance_q4)      # дисперсия
# [1] 1136.324
sd(percent_acceptance_q4)       # стандартно отклонение
# [1] 33.70941
fivenum(percent_acceptance_q4)
# [1]   0  10  50  70 100
```

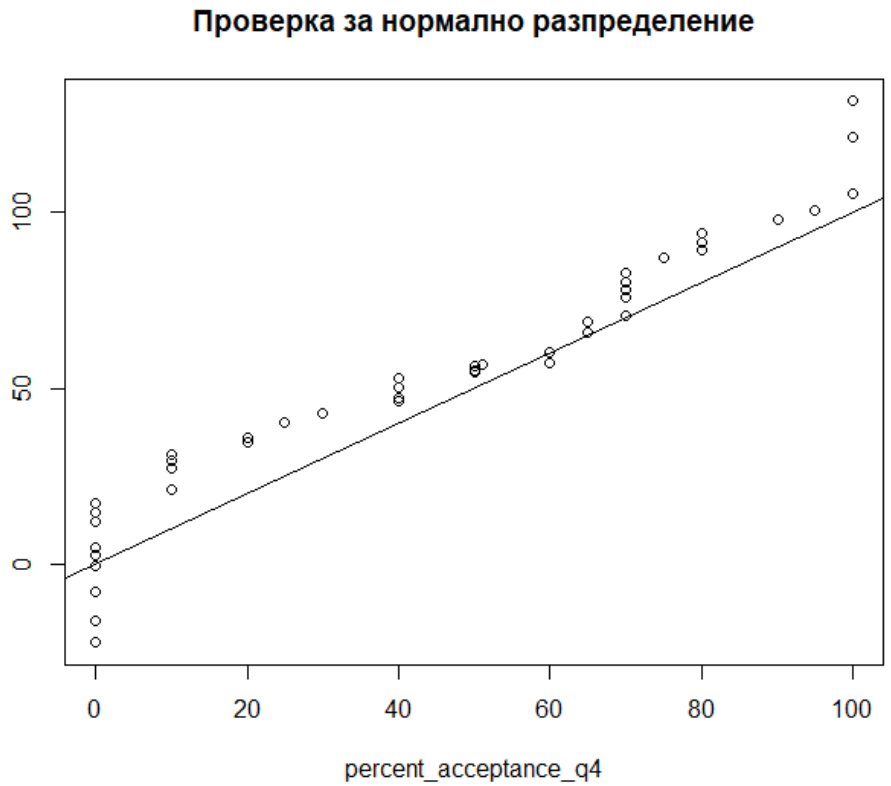
Сега ще опиша вариацията (дисперсията) и стандартното отклонение, което ще покаже колко далече са наблюденията от очакването. Освен това чрез хистограма ще изобразя разпределението на данните графично.

Код на R:

```r
# Вариация (дисперсия) на разпределението
range(percent_acceptance_q4)    # range - показва най - голямата и най - малката стойност
# [1]  0 100
var(percent_acceptance_q4)      # дисперсия
# [1] 1136.324
sd(percent_acceptance_q4)       # стандартно отклонение
# [1] 33.70941
fivenum(percent_acceptance_q4)
# [1]   0  10  50  70 100

# Графично представяне
hist(table_q4, main = "Колко процента имате доверие на президентът Румен Радев? (0-100)", xlab = "Проценти /Нор
     ylab = "Брой хора /честота/", col = "blue", xlim = c(0,15))

# Проверка за нормално разпределение
d1 <- rnorm(n = 10^2, mean = mean(percent_acceptance_q4), sd = sd(percent_acceptance_q4))
qqplot(percent_acceptance_q4, d1, main = "Проверка за нормално разпределение")
abline(a = 0, b = 1) # чертае линия
```

Графики:


Колко процента имате доверие на президентът Румен Радев? (0-1
Брой хора /Честота/
0
2
4
6
8
0
5
10
15
Проценти /Нормално разпределение/


Проверка за нормално разпределение
0
50
100
0
20
40
60
80
100
percent_acceptance_q4

Основно хората имат неутрално доверие към президента, като има стойности в целия интервал от стойности (0-100. Изчислявайки средната стойност от всички отговори, става ясно, че средно на човек се падат по около 44% доверие, което показва че доверието към управляващите е доста ниско, което означава че избирателите желаят промяна. Освен това отговорите не са нормално разпределени, както се вижда от втората графика, като това ще докажем по – късно с помощта на shapiro.test отново.

Въпрос 5: Смятате ли, че трябва да се увеличи броят на секциите за гласуване в чужбина?

Този въпрос представлява категорийна величина, тъй като анкетираните трябва да изберат от трите отговора спрямо нагласите им да се създадат повече изборни секции в чужбина, т.е. тяхната нагласа попада в 3 различни категории.

Код на R:

```
# Въпрос 5: Смятате ли, че трябва да се увеличи броят на секциите за гласуване в чужбина? - Категорийна променлива

utf8_print(answers$Question5)
#[1] "Да"           "Да"           "Да"           "Да"           "Да"           "Да"           "Да"           "Да"           "Да"
#[10] "Да"           "Да"           "Не съм решил" "Да"           "Да"           "Да"           "Да"           "Да"           "Не"
#[19] "Да"           "Не"           "Да"           "Да"           "Да"           "Да"           "Да"           "Да"           "Да"
#[28] "Да"           "Да"           "Да"           "Да"           "Да"           "Не"           "Да"           "Да"           "Да"
#[37] "Да"           "Да"           "Не съм решил" "Да"           "Да"           "Да"           "Да"

table_q5 <- table(answers$Question5)
table_q5

#Да    Не    Не мога да преценя
#28    12    3

# Графично представяне
barplot(round(prop.table(table_q5)*100, 2), col = c("green", "red", "red"), main = "Смятате ли, че трябва да се увеличи броят на секциите за гласуване в чужбина?",
        xlab = "Отговор", ylab = "Проценти", ylim = c(0, 100))
```

Графика:

Отговорите на този въпрос са 43.
От графиката се вижда, че от анкетираните:

- малко над 60% смятат че трябва да се увеличи броя секции
- около 30% смятат че не трябва да се увеличи броя секции
- около 10% нямат впечатления чрез които да могат да преценят

2.2. Проверка за нормално разпределение.
След като вече въведох и изобразих данните, ще използвам shapiro.test за всеки въпрос, за да проверя дали са равномерно разпределени. Нека хипотезата H0 гласи, че отговорите са равномерно разпределени.
Нивото на съгласие е alpha = 0,05.

```
> #Тест за нормално разпределение
> shapiro.test(table_q1)

        Shapiro-Wilk normality test

data:  table_q1
W = 0.77082, p-value = 0.04658

> #Тест за нормално разпределение
> shapiro.test(table_q2)

        Shapiro-Wilk normality test

data:  table_q2
W = 0.71548, p-value = 0.0001741

> #Тест за нормално разпределение
> shapiro.test(table_q3)

        Shapiro-Wilk normality test

data:  table_q3
W = 0.7197, p-value = 0.0008798

> #Тест за нормално разпределение
> shapiro.test(table_q4)

        Shapiro-Wilk normality test

data:  table_q4
W = 0.77919, p-value = 0.001461

> #Тест за нормално разпределение
> shapiro.test(table_q5)

        Shapiro-Wilk normality test

data:  table_q5
W = 0.97453, p-value = 0.6939
```

Забелязваме, че стойностите за p са различни. Тези, чиито p – value са по –големи от нивото на съгласие, считаме разпределенията им за нормални. За нормално разпределените променливи ще използваме параметрични тестове,
а за останалите – непараметрични. Забелязваме, че стойността на p при въпрос 5 е над нивото на съгласие (>0,05). Следователно, можем да считаме, че отговорите на този въпрос са нормално разпределени.

2.1. Анализ на взаимодействието между две променливи.
✔ Категорийна vs Числова
✔ Числова vs Числова. За изпълнението на тази част ще направя проверка за това дали има
зависимост между две променливи, като използвам различни хипотези. За графичното им изобразяване ще използвам различни техники.

Категорийна vs Числова

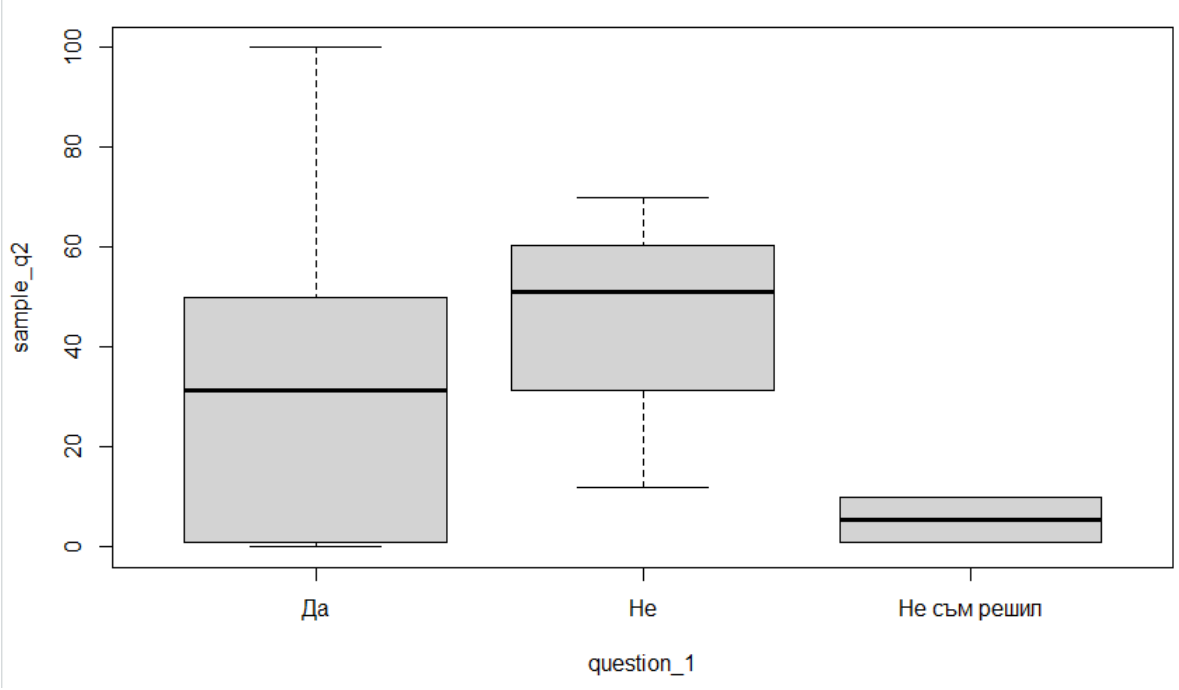
Ще изследвам:
Въпрос 1: Ще гласувате ли на предсрочните избори през лятото на 2021?
и
Въпрос 2: Колко процента смятате че избрания на 04.04.2021 парламент е достатъчно добър, за да съществува? (0-100)
с помощта на boxplot.

Тази техника се използва за обобщаване на данните и бързо определяне на това дали те са симетрични или имат outliers.

Код:

```
#Категорийна vs числова
#1. Въпрос 1 - Въпрос 2
question_1 <- c(rep("Да", 38), rep("Не", 3), rep("Не съм решил", 2))
sample_q2 <- sample(x = percent_acceptance_q2)
boxplot(sample_q2 ~ question_1)
```

Графика:



Удебелената черна линия във всяка група е медианата. От двете страни на медианата са съответно първи и трети квартил. Дължините на опашките пък представляват минималната и максималната стойност. В първата група медианата се доближава до минималната стойност. Във втората група медианата се доближава до максималната стойност .Във третата група медианата съвпада със средната стойност от отговорите. А в първата група стойностите са най – големи, защото отговорът „Да“ преобладава и както може да се види от графиката, максималната стойност е по – далеч от медианата.
Изводът е, че хората не са доволни от избрания на 04.04 парламент и смятат да гласуват отново на предсрочните избори 2021.

Числова vs Числова:

Ще направя един пример за корелационен анализ и един пример за линейна регресия (регресионен анализ)

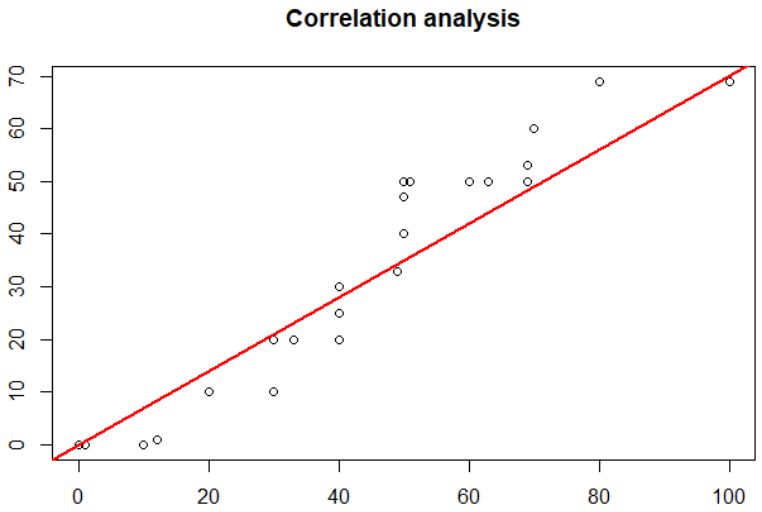
Корелационен анализ:
Въпрос 2: Колко процента смятате че избрания на 04.04.2021 парламент е достатъчно добър, за да съществува? (0-100)
Въпрос 3: Колко процента имате доверие на сегашния парламент? (0-100)

Изобразявам графично връзката между двете променливи. Вижда се, че въпреки високия коефициент на корелация, тя не е линейна и не мога да използвам линеен модел, за да моделирам връзката.

```
#Числова vs Числова
#2. Въпрос 2 - Въпрос 3

# Корелационен анализ
rho <- round(cor(percent_acceptance_q2, percent_acceptance_q3), 3) #коефициент на корелация
par(mfrow = c(1, 1))
plot(percent_acceptance_q2, percent_acceptance_q3, main = "Correlation analysis")
abline(a = 3, b = 4, col = "red", lwd = 2)
cor(percent_acceptance_q2, percent_acceptance_q3)
#[1] 0.9675945

cor.test(percent_acceptance_q2, percent_acceptance_q3, method = "spearman")
#Spearman's rank correlation rho

#data:  percent_acceptance_q2 and percent_acceptance_q3
#S = 244.52, p-value < 2.2e-16
#alternative hypothesis: true rho is not equal to 0
#sample estimates:
#      rho
#0.9815369
```

Ще направя корелационен анализ, за да измеря силата на връзката им.

rho представлява коефициент на корелация и в случая той показва,че има много силна корелация (зависимост) между тези два въпроса. Това може да се види и на графиката(стойностите са ориентирани в средата):



Изводът, който може да се направи е, че процентът на неодобрение на избрания парламент, е силно зависимо и повлияно от силното недоверие.

Регресионен анализ:
Въпрос 3: Колко процента имате доверие на сегашния парламент? (0-100)
Въпрос 4: Колко процента имате доверие на президентът Румен Радев? (0-100)

Ще изследвам двата въпроса с помощта на линейна регресия, за да видя дали има връзка между тях.

Нулевата хипотеза е: Хората, които имат по-ниско доверие към парламента, имат по-високо доверие към президента.

Алтернативната хипотеза - H1 - е, че това не е така. H0 се отхвърля при стойност на p-value < 0.05.

Най – напред въвеждам данните от двата въпроса в data frame (DF), като първо изравнявам дължините на двата вектора. Построяваме линейния модел по следния начин:

```
#3. Въпрос 3 - Въпрос 4
# Линейна регресия
DF <- data.frame(percent_acceptance_q3, percent_acceptance_q4)
model <- lm(percent_acceptance_q3~percent_acceptance_q4)
model
#Call:
#       lm(formula = percent_acceptance_q3 ~ percent_acceptance_q4)
#
#Coefficients:
#       (Intercept)  percent_acceptance_q4
#-4.4510                 0.6373
```

Коефициентът пред percent_acceptance_q4 е 0.6373:
percent_acceptance_q3 = 0.6373*percent_acceptance_q4 - 4.4510

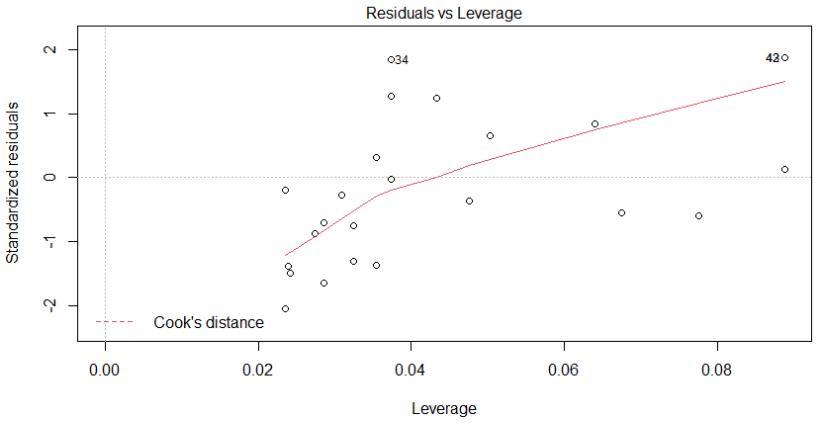
След това ще проверим до колко този модел описва добре данните и какви са оценките на коефициентите му:

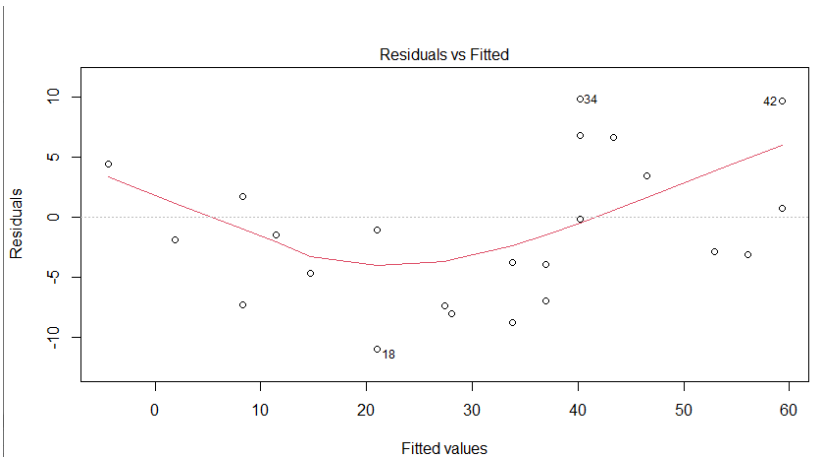
```
summary(model)

#Call:
#       lm(formula = percent_acceptance_q3 ~ percent_acceptance_q4)
#
#Residuals:
#       Min       1Q   Median       3Q      Max
#-11.0425  -3.4427  -0.1627   4.4510   9.8373
#
#Coefficients:
#       Estimate Std. Error t value Pr(>|t|)
#(Intercept)          -4.45103    1.37742  -3.231  0.00243 **
#       percent_acceptance_q4  0.63734    0.02492  25.571  < 2e-16 ***
#       ---
#       Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
#Residual standard error: 5.445 on 41 degrees of freedom
#Multiple R-squared:  0.941,    Adjusted R-squared:  0.9396
#F-statistic: 653.9 on 1 and 41 DF,  p-value: < 2.2e-16
```

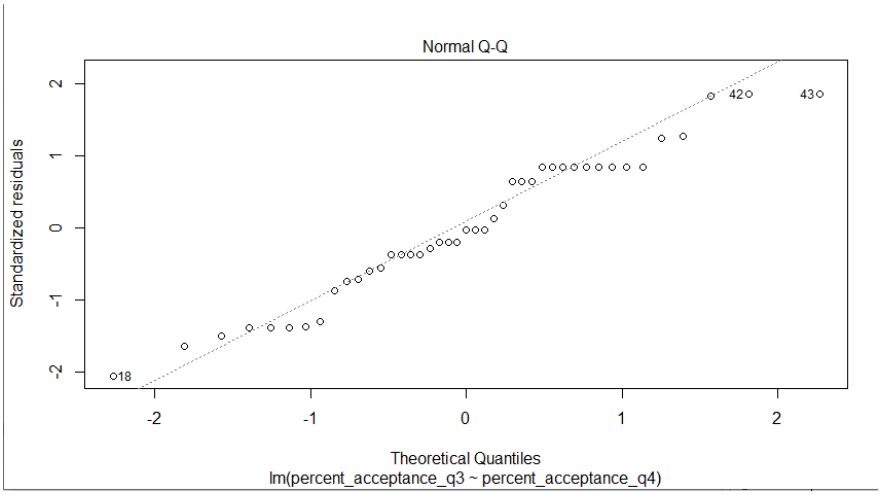
Стойностите на всички коефициенти са различни от 0. Тъй като стойността на p (p – value) в колоната Pr(>|t|) за единия параметър е 0.00243 << 0.05, то той е статистически значими. Коефициентът пред percent_acceptance_q4 е -2.2e-16 < 0.05. Тоест, той също е значим за анализа.

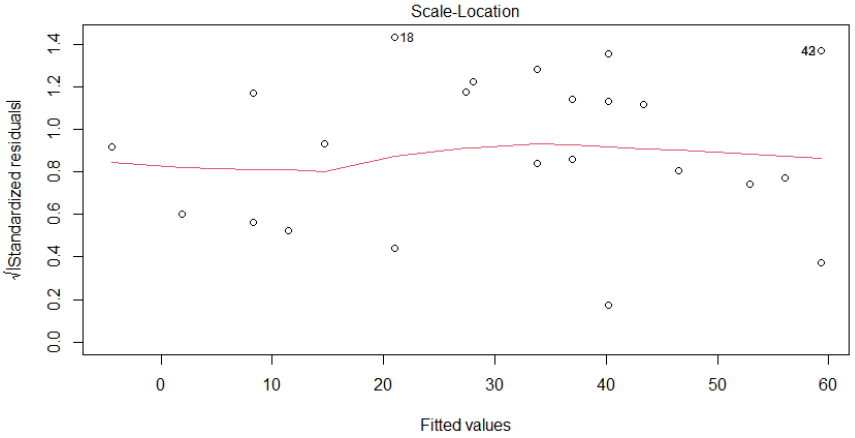
Статистиката “Multiple R-squared“ или “Adjusted R-squared“ показват колко добре моделът описва данните. И двете приемат стойности в интервала [0, 1]. Стойността на първата статистика е далеч от 1 и това означава, че моделът не е толкова добър. Стойността на “Adjusted R-squared“ е отрицателна, тъй като тя приема и статистически незначими стойности. Използвам функцията resid() за създаване на списък с остатъците и след това правя графика за тях

```
> resid(lm(percent_acceptance_q3~percent_acceptance_q4))
         1          2          3          4          5          6          7          8          9         10         11         12
 4.4510326  4.4510326  4.4510326  4.4510326  4.4510326  4.4510326  4.4510326  4.4510326  4.4510326 -1.9223556 -1.9223556 -1.9223556
        13         14         15         16         17         18         19         20         21         22         23         24
-1.9223556 -7.2957438  1.7042562 -1.4824379 -4.6691320 -11.0425202 -1.0425202 -1.0425202 -1.0425202 -7.4159084 -7.4159084 -7.4159084
        25         26         27         28         29         30         31         32         33         34         35         36
-8.0532472 -8.7892966 -3.7892966 -6.9759907 -3.9759907 -0.1626848 -0.1626848 -0.1626848  6.8373152  9.8373152  6.6506211  3.4639270
        37         38         39         40         41         42         43
 3.4639270  3.4639270 -2.9094612 -3.0961553  0.7171507  9.7171507  9.7171507
```









Residuals vs Fitted – Графика на разсейване на остатъците по оста y и прогнозните отговори по оста x. Използва се за откриване на нелинейност, неравномерни отклонения и външни разходи. Вижда се, че остатъците са разпръснати на случаен принцип около остатъчния ред = 0. Това предполага, че връзката е линейна. Освен това остатъците образуват „хоризонтална лента“ около остатъчната линия и това означава, че вариантите на условията за грешки са равни. Има няколко остатъци, които се открояват от основния модел, но това не винаги е показател за наличие на проблем.

Normal Q-Q – Метод на най – малките квадрати, който показва, че остатъците не са нормално разпределени спрямо прекъснатата линия, тъй като се наблюдават тежки опашки. Това може да се докаже и чрез shapiro.test(res), където res е променлива за остатъците.

Scale – Location – Тази графика (на мащабното местоположение) е подобна на първата, но опростява анализа на предположението за хомоскедастичност. Той отнема квадратния корен на абсолютната стойност на стандартизираните остатъци вместо да начертава самите остатъци. Червената линия е приблизително хоризонтална. Това означава, че средната величина на стандартизираните остатъци не се променя много като функция на fitted стойностите. Освен това тази линия не се променя особено в зависимост от тези стойности. Значи променливостта на величините не се променя много като функция от fitted стойностите.

Residuals vs Leverage – Тази графика се използва за откриване на хетероскедастичност и нелинейност. Разпространението на стандартизираните остатъци не трябва да се променя като функция на leverage. В случая тази функция намалява, което показва наличието на хетероскедастичност. Точките с висок leverage (лост) могат да повлияят на модела, тъй като тяхното изтриване би могло да го промени изцяло. Затова разглеждаме разстоянието на Кук, което измерва ефекта от изтриването на точка върху комбинирания вектор на параметъра. В този случай няма точки извън пунктираната линия, затова не можем и да определим тяхното влияние.

Част 3: Заключение.

С помощта на различните подходи – тестове и графики за анализ на променливите и взаимодействието между тях, стигнах до някои важни изводи, които отговарят на въпросите, поставени още в началото. Целта на проекта беше да се определи дали младите вярват на сегашните управляващи и дали смятат че е нужна промяна.

- Около 90% от анкетираните не са доволни от избрания парламент и смятат да гласуват на следващите избори.
- Основната част от анкетираните нямат почти никакво доверие към парламентарната институция, като няма анкетиран който да има пълно доверие на управляващите.
- Отношението към президентската институция е почти неутрално, клонящо към липса на доверие.
- Мнението за броя на изборни секции в чужбина е разминаващо се и няма утвърдено мнозинство.

В заключение може да се каже, че няма логическа свързаност между отговорите на въпросите. Човек се страхува от промяната, но в случаи като този я желае.

Използвана литература:

Упражнения по СЕМ - практикум