Универзитет “Св. Кирил и Методиј” - Скопје

Факултет за информатички науки и компјутерско инженерство

Семинарска работа по предметот Бизнис статистика

Михајло Крагујевски

183002

Предметен професор:

Д-р Весна Димитриевска Ристовска

Вовед

Извор на конкретно податочно множество:
Предвидувања за прием на дипломирани студенти на магистарски студии.
https://www.kaggle.com/mohansacharya/graduate-admissions

Ова податочно множество е изградено со цел да им помогне на студентите од универзитетите со потесен избор на нивните профили. Се состои од 9 обележја и 400 единки.

Предвидените резултати им даваат фер идеа за нивните шанси за прием во одредена магистарска програма на некој универзитет.

Базата на податоци се состои од 9 обележја кои се сметаат за важни за време на апликацијата за магистарски програми. Вклучените параметри се:

1. Сериски број

2. Дипломска евиденција – GRE (од 340)

3. Тест за англиски како странски јазик – TOEFL (од 120)

4. Универзитетско рејтинг – University rating (од 5)

5. Изјава за намена – SOP (од 5)

6. Писмо за препорака – LOR (од 5)

7. Додипломски успех – UGPA (од 10)

8. Истражувачко искуство (0 или 1)

9. Шанси за прием (во ранг од 0 до 1).

До крајот на овој труд ќе ви биде јасно за тоа какви се резултати се потребни, на различни тестови, за студентите да имаат подобри можности за прием и да стапат во добро организиран универзитет.

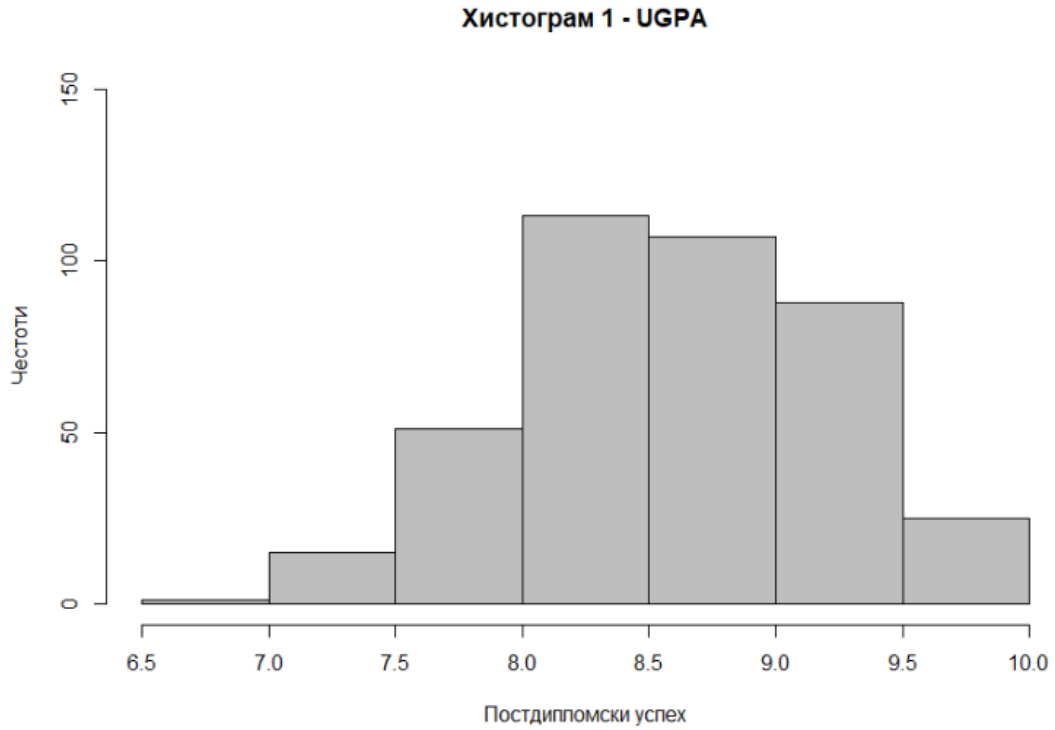
Барање 1, Табела на честоти, средни точки, релативни и кумулативни фреквенции

Најпрво започнувам со бирање две обележја за да вршам пресметки на нив, ги бирам UGPA и Chance of Admit. Правам две променливи за да ги сместам параметрите, chance – Шанси за прием и GPA – Додипломски успех. Следно, ја правам табелата со распределби на честоти за додипломскиот успех, (види Слика 1.) со тоа што правам интервалите да одат од 6,5 – 10,0 притоа со интервал од 0,5 помеѓу.

```
> # Tabela so raspredelba na chestoti na UGPA
> len = length(GPA)
> brIntervali = seq(6.5, 10, by=0.5)
> d.int = cut(GPA, brIntervali, right=FALSE)
> freq = table(d.int)
> Rfreq = freq/len
> CumFreq = cumsum(freq)
> R_CumFreq = cumsum(freq)/len
> R_CumFreq2 = cumsum(Rfreq)
> Pfreq = Rfreq*100
> P_Cumfreq = R_CumFreq*100
> d.table = cbind(freq, Rfreq, CumFreq, Pfreq, P_Cumfreq)
> d.table
         freq  Rfreq CumFreq Pfreq P_Cumfreq
[6.5,7)     1 0.0025       1  0.25       0.25
[7,7.5)    13 0.0325      14  3.25       3.50
[7.5,8)    44 0.1100      58 11.00      14.50
[8,8.5)   117 0.2925     175 29.25      43.75
[8.5,9)   107 0.2675     282 26.75      70.50
[9,9.5)    91 0.2275     373 22.75      93.25
[9.5,10)   27 0.0675     400  6.75     100.00
>
```

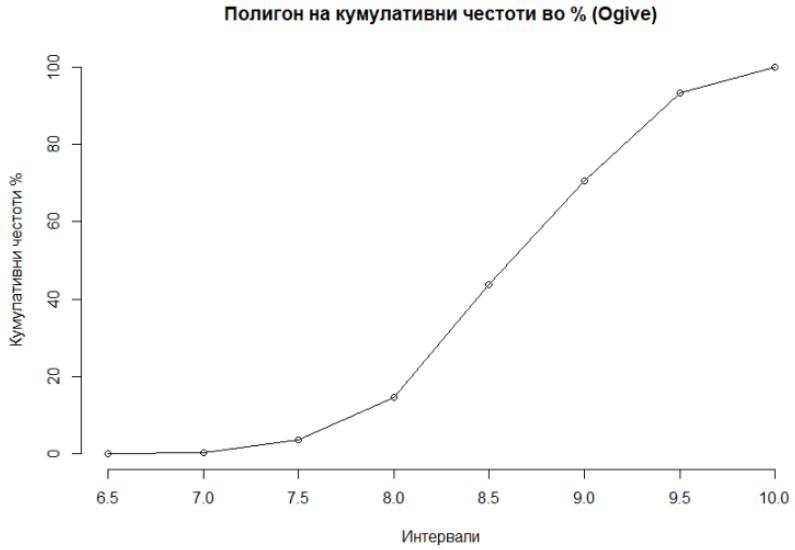
Слика 1. Табела на распределба на честоти на Додипломскиот успех

Потоа ги добивам средните точки на интервалите од распределбата на честоти со помош на еден for - циклус и тие се 6.75 7.25 7.75 8.25 8.75 9.25 9.75. Хистограмот го правам со hist() така што на y – оската се честотите а додека па на x – оската се интервалите на додипломските успеси од 6,5 - 10 со breaks = 10 за подобра прегледност на податоците.



Од сликата може јасно да се види дека најмалку често појавување на просекот е во интервал од 6.5 – 7.0 а додека па најмногу појавувања има во интервалот 8.0 – 8.5.

Следно ова обележје го претставувам со полигон на честоти – ogive (види Слика 2). Го користам процентот на кумулативната честота на y – оската, додека па на x – оската се интервалите од обележјето, каде што јасно се забележува растот на кривата од почетниот интервал па се до крајниот.



Слика 2. Ogive на додипломскиот успех

Наредно, истово го повторувам за второто обележје.

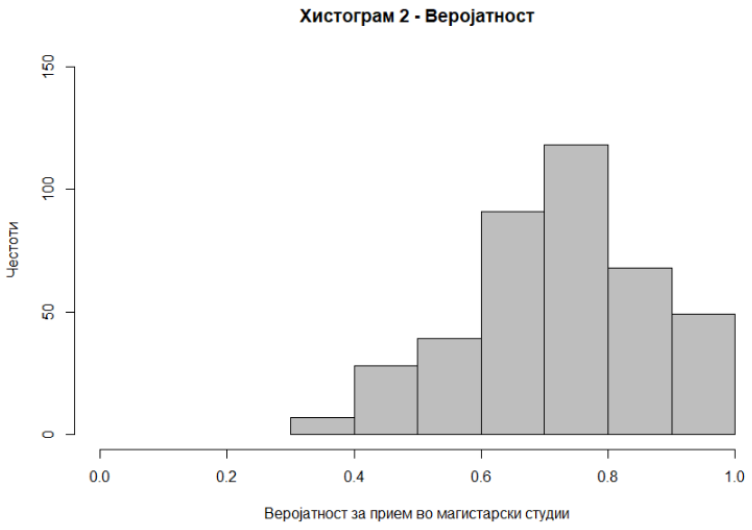
```
> # Tabela so raspredelba na chestoti za shansi za priem
> len = length(chance)
> brIntervali = seq(0, 1, by=0.2)
> d.int = cut(chance, brIntervali, right=FALSE)
> freq = table(d.int)
> Rfreq = freq/len
> CumFreq = cumsum(freq)
> R_CumFreq = cumsum(freq)/len
> R_CumFreq2 = cumsum(Rfreq)
> Pfreq = Rfreq*100
> P_Cumfreq = R_CumFreq*100
> d.table = cbind(freq, Rfreq, CumFreq, Pfreq, P_Cumfreq)
> d.table
          freq  Rfreq CumFreq Pfreq P_Cumfreq
[0,0.2)      0 0.0000       0  0.00      0.00
[0.2,0.4)    7 0.0175       7  1.75      1.75
[0.4,0.6)   67 0.1675      74 16.75     18.50
[0.6,0.8)  198 0.4950     272 49.50     68.00
[0.8,1)    128 0.3200     400 32.00    100.00
>
```

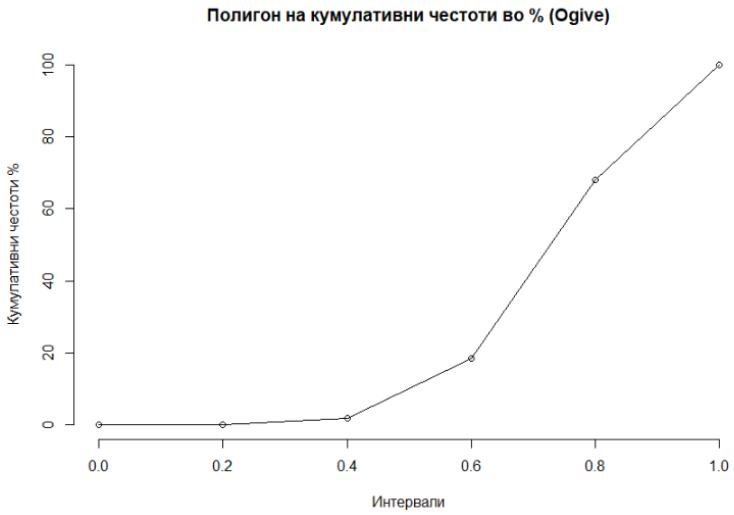
Распределбата на честоти ја правам така што бројот на интервали го бирам да оди од 0-1 со интервал од 0.2 (види Слика 3). Може уште тука да се забележи дека најголема фреквенција се јавува во интервал од 0.6 – 0.8, што е многу добро бидејќи постојат 198 студенти кои имаат високи шанси за прием.

Слика 3. Табела на распределба на честоти за шансите за прием

Ги пресметувам средните точки на интервалите и добивам 0.1 0.3 0.5 0.7 и 0.9.

Хистограмот на ист начин го конструирам меѓутоа тука ги менувам интервалите на x – оската да оди од 0 – 1 со breaks = 5, со тоа што x – оската се шансите (или веројатноста, бидејќи се вредности од 0 до 1) за прием, а y-оската останува непроменета бидејќи тие параметри сеуште ми вршат работа. Од сликата може попрецизно да се воочи дека најголем број студенти има што имаат веројатност за прием во интервал од 0.7 – 0.8.





Полигонот на честоти по ист принцип, ги земам кумулативните честоти изразени во проценти и интервалите. Го добивам овој графички приказ каде збирот на секоја веројатност во проценти дава 100.

Барање 2, Стебло – лист дијаграм на двете обележја.

За цел креирање на стебло – лист дијаграм, ја користам функцијата myStem и пакетот data.table.
leftDigits претставува позицијата на | во однос на децималната точка а додека па rounding = 1 е бројот на децимали.
За првото обележје, односно за постдипломскиот успех резултатот од стебло – лист дијаграмот е прикажан на сликата. Резултат од ова е дека најмногу податоци има со почетна децимала 8, што значи има најголем број студенти во рангот од 8 – 9 просек, при што и заклучивме со сите претходни испитувања на ова обележје.

```
> # Steblo List na GPA
> cost = sort(GPA)
> library(data.table)
> myStem = function(x, leftDigits, rounding = 1)
+ {
+   data = data.table("x" = x)
+   data[, left := floor(x/10^leftDigits)]
+   data[, right := (round(x - left*10^leftDigits, rounding))*10^rounding]
+   data = data[, paste(sort(right), collapse = " "), by = left]
+   data[, out := paste(left, " | ", V1), by = left]
+   cat(data$out, sep = "\n")
+ }
> myStem(cost, 0, 2)
6  |  80
7  |  20 25 28 30 34 34 36 40 43 43 46 46 46 50 50 54 56 60 64 64 64 65 65 65 65 66 66 6
6 67 68 70 70 80 80 80 84 86 86 86 86 87 88 88 88 89 89 89 89 90 90 90 90 92 95 96 97 98
8  |  0 0 0 0 0 0 0 0 0 0 1 1 2 2 2 3 3 3 4 4 4 6 7 7 9 10 10 10 10 12 12 12 12 12 13 13 1
4 14 15 16 16 16 17 17 18 18 20 20 20 20 20 21 21 22 22 22 23 24 24 24 25 26 26 26 26 27
 27 27 27 27 28 30 30 30 30 32 32 33 33 33 33 34 34 34 34 36 36 36 37 37 40 40 40 40 42
 42 42 43 43 43 44 44 44 45 45 45 45 45 45 46 46 46 47 47 48 48 48 49 50 50 50 50 50 52
 53 53 54 54 54 54 54 55 56 56 56 56 56 56 56 56 57 57 60 60 62 62 64 64 64 64 64 64 65 65
 65 65 65 66 66 66 66 67 67 67 67 67 68 68 68 69 69 69 70 70 71 72 73 73 74 74 75 75 76
 76 76 76 76 76 76 76 77 77 77 78 78 78 79 79 79 79 80 80 80 80 80 83 84 84 85 87 87 87
 88 89 90 90 90 92 94 96 96 96 97 97 97 98
9  |  0 0 0 0 0 1 1 2 2 2 4 4 4 4 4 5 6 6 7 8 8 8 10 10 10 10 10 11 11 11 11 11 11 12 12
 12 13 13 13 14 14 14 15 15 16 16 16 17 17 18 18 19 19 20 22 22 22 23 23 23 23 24 24 26
 26 28 28 30 31 32 34 34 35 36 36 36 36 38 40 40 42 43 44 44 45 45 45 45 46 47 48 50 50
 53 54 56 60 60 62 64 65 66 66 66 68 70 70 74 76 76 78 80 80 80 82 87 91 92
>
```

Слика 4. Стебло - лист дијаграм на додипломскиот успех

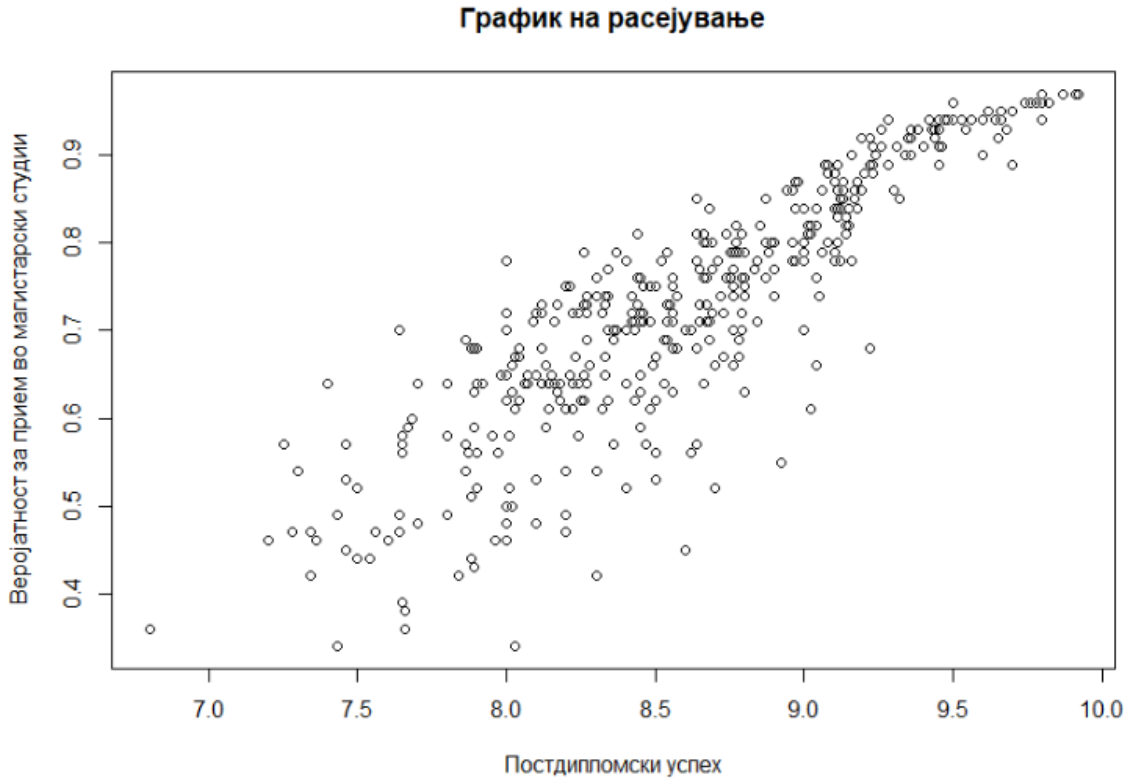
За второто обележје го користам истиот принцип на пресметување, но резултатот тука е различен. Добиваме само една цифра која стои лево од | и таа е нула, тоа е бидејќи вредностите за овој параметар се движат од 0 до 1 па логично е да сите вредности и "припаѓаат" на нулата.

```
> # Steblo List na Chance of admit
> cost = sort(chance)
> library(data.table)
> myStem = function(x, leftDigits, rounding = 1)
+ {
+   data = data.table("x" = x)
+   data[, left := floor(x/10^leftDigits)]
+   data[, right := (round(x - left*10^leftDigits, rounding))*10^rounding]
+   data = data[, paste(sort(right), collapse = " "), by = left]
+   data[, out := paste(left, " | ", V1), by = left]
+   cat(data$out, sep = "\n")
+ }
> myStem(cost, 0, 2)
0  |  34 34 36 36 38 38 39 42 42 42 43 44 44 44 45 45 46 46 46 46 46 47 47 47 47 47 48 4
8 48 49 49 49 49 50 50 51 52 52 52 52 52 53 53 53 54 54 54 54 54 55 56 56 56 56 56 56 57
 57 57 57 57 57 57 58 58 58 58 58 59 59 59 59 60 61 61 61 61 61 61 62 62 62 62 62
 62 62 62 63 63 63 63 63 63 64 64 64 64 64 64 64 64 64 64 64 64 64 64 64 64 65 65
 65 65 65 65 65 65 66 66 66 66 66 66 66 67 67 67 67 67 67 67 68 68 68 68 68 68 68
 68 68 69 69 69 69 69 69 69 70 70 70 70 70 70 70 70 70 70 70 70 71 71 71 71 71 71 71
 71 71 71 71 71 71 71 71 72 72 72 72 72 72 72 72 72 72 72 72 72 72 73 73 73 73 73 73
 73 73 73 73 73 73 73 74 74 74 74 74 74 74 74 74 74 74 75 75 75 75 75 75 75 75 76 76 76
 76 76 76 76 76 76 76 76 76 77 77 77 77 77 77 77 77 78 78 78 78 78 78 78 78 78 78 78
 79 79 79 79 79 79 79 79 79 79 79 79 80 80 80 80 80 80 80 80 80 80 80 81 81 81 81 81 81
 81 81 82 82 82 82 82 82 82 82 82 83 83 83 84 84 84 84 84 84 84 84 84 84 85 85 85 85 85 85 86
 86 86 86 86 86 86 86 87 87 87 87 87 88 88 88 88 89 89 89 89 89 89 89 89 89 90 90 90 90
 90 90 90 90 91 91 91 91 91 91 91 92 92 92 92 92 92 92 93 93 93 93 93 93 93 93 93 94 94 94
 94 94 94 94 94 94 94 94 94 95 95 95 95 96 96 96 96 96 96 96 97 97 97 97
>
```

Слика 5. Стебло - лист дијаграм на шансите за прием

Барање 3, График на расејување

Од графикот на расејување на податоци за двете обележја може да се заклучи дека веројатноста за прием е право пропорционална со успехот, односно како расте успехот така расте и веројатноста што е и логично бидејќи студенти со помал просек имаат помали шанси за прием за разлика од студенти со висок успех кои што имаат високи шанси да бидат примени во некоја магистарска програма.



Барање 4, Мода, медијана и просек

```
> # Moda
> Mode <- function(x)
+ {
+   ux <- unique(x)
+   ux[which.max(tabulate(match(x, ux)))]
+ }
> Mode(GPA)
[1] 8
> Mode(chance)
[1] 0.64
>
> # Medijana
> median(GPA)
[1] 8.61
> median(chance)
[1] 0.73
>
> # Prosek
> prosek1 = mean(GPA)
> prosek1
[1] 8.598925
> prosek2 = mean(chance)
> prosek2
[1] 0.72435
>
```

Слика 6. Мода, медијана и просек на двете обележја

За мода ја наоѓам вредноста што најчесто се јавува во обележјето, со помош на формула, во овој случај за првото обележје тоа е 8 а за второто 0.64.

Медијана барам со функцијата median() и за првото обележје тоа е 8.61 односно тоа значи дека 50% од вредностите се под овој податок и 50% се над, додека за второто обележје тоа е 0.73.

Просек барам со функцијата mean() и за првото обележје добивам 8.59 додека па за второто 0.72. (види Слика 6)

Барање 5, Квартали, опсег и интерквартален распон

Кварталите ги добивам со функцијата quantile(), посебно за двете обележја, каде што табеларно ми се прикажани кој податок одговара на одредениот квартал. Тука може да се забележи дека вториот квартал (50%) има исти вредности како и медијаната од претходно, што потврдува дека пресметките се точни.

Опсегот по дефиниција е разликата помеѓу најголемиот и најмалиот елемент, па тоа и го правам користејќи ги функциите max() и min(). За првото обележје добивам 3,12 додека па за второто добивам 0,63.

Интерквартален распон пресметувам со формулата IQR() и го претставува распонот на средните 50% од податоците, односно $Q_3 - Q_1$. За првото обележје добивам 9,0625 - 8,1700 = 0,8925, додека па за второто 0,83 – 0,64 = 0,19.

```
>
> # Kvartali
> quantile(GPA)
    0%    25%    50%    75%   100%
6.8000 8.1700 8.6100 9.0625 9.9200
> quantile(chance)
  0%  25%  50%  75% 100%
0.34 0.64 0.73 0.83 0.97
>
> # Opseg
> opseg1 = max(GPA) - min(GPA)
> opseg1
[1] 3.12
> opseg2 = max(chance) - min(chance)
> opseg2
[1] 0.63
>
> # Interkvartalen Raspon
> interkvartalenRaspon1 = IQR(GPA)
> interkvartalenRaspon1
[1] 0.8925
> interkvartalenRaspon2 = IQR(chance)
> interkvartalenRaspon2
[1] 0.19
>
```

Слика 7. Квартали, опсег (распон) и интерквартален распон

Барање 6, Дисперзија и стандардна девијација

```
>
> # Disperzija
> dis1 = var(GPA)
> dis1
[1] 0.3555941
> dis2 = var(chance)
> dis2
[1] 0.02033742
>
> # Standardna devijacija
> std1 = sd(GPA)
> std1
[1] 0.5963171
> std2 = sd(chance)
> std2
[1] 0.1426093
>
```

Слика 8. Дисперзија и стандардна девијација

Дисперзија наоѓам со функцијата var() што го претставува варирањето околу просекот, а стандардна девијација е корен од дисперзијата.
Во R има посебна функција за стандардна девијација и тоа е sd().

Од добиените резултати, бидејќи добив ниски вредности на стандардната девијација за двете обележја, може да се заклучи дека останатите вредности вариираат во близина на просекот

Барање 7, Коефициент на корелација

Коефициент на корелација барам по готовата функција во R, cor(), на двете обележја. Пресметаниот коефициент изнесува 0.8732891, што може да се забележи дека јачината на линеарната врска меѓу овие две обележја е силна односно велиме дека има силна позитивна линеарна поврзаност.

| ВТОР ДЕЛ

Барање 1, Интервал на доверба за математичко очекување од податоците на GPA

```
>
> # Interval na doverba
> n = 400                                  # Golemina na primerok
> error = qnorm(0.975)*std1/sqrt(n)        # std1 - Prethodno presmetanata standardna devijacija na GPA
>                                          # prosek1 - Prethodno presmetaniot prosek na GPA
> prosek1 - error                          # Dolna granica na intervalot vo koj se naogja prosekot
[1] 8.540487
> prosek1 + error                          # Gorna granica na intervalot vo koj se naogja prosekot
[1] 8.657363
>
```

Слика 9, Интервал на доверба

Интервал на доверба $\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ претставува множество вредности за кои можеме да утврдиме, со одредена веројатност, дека го содржи непознатиот параметар. Истово може да се запише како $\bar{x} \pm ME$ каде ME е маргината на грешка, $ME = z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$.

Во мојов случај, примерокот е со големина од 400 (вкупниот број на податоци), за да добијам ниво на доверба од 95%, мора $\alpha = 1 - 0{,}95 = 0{,}05$. Читам од таблица за нормална распределба дека 1 – 0,025 = 0,975 па $z_{0{,}025} = 1{,}96$. Според формулата ги имам сите потребни податоци за да ја пресметам маргината на грешка, па $1{,}96\frac{0{,}5963171}{8{,}598925} = 0{,}13592$. Следно ги наоѓам долната и горната граница и го добивам следниот интервал на доверба, (8.540487, 8.657363). Со други зборови, со 95% сигурност, просекот на популацијата се наоѓа во овој интервал.

Барање 2, Поставување и тестирање на хипотези

Најпрво ја изразувам претпоставката која треба да се тестира.

$H_0 : \mu = 9$

$H_a : \mu \neq 9$

```
> # Postavuvanje i testiranje na hipotezi
> mi = 9.0                                 # Pretpostavuvam deka prosekot na GPA e 9
>
> z.test = function(GPA, mi, dis1){        # dis1 - Prethodno presmetanata disperzija na GPA
+   zeta = (prosek1 - mi) / (sqrt(dis1 / length(GPA)))
+   return(zeta)
+ }
> Z = z.test(GPA, mi, dis1)
> Z
[1] -13.45174
> if(Z < -1.96 || Z > 1.96)
+ {
+   print('H0 se otfrla')
+ }else
+ {
+   print('H0 se prifakja')
+ }
[1] "H0 se otfrla"
>
```

Слика 10, Тестирање на хипотези

Претпоставувам дека просекот на GPA од сите апликанти е 9 и на почеток ја земам оваа вредност за точна.

Следно, избирам ниво на значајност $\alpha = 0{,}05$ која ќе ми ги даде критичните вредности на тестот. Правам двостран тест за статистика и го добивам критичниот домен $(-\infty, -1{,}96)\ U\ (1{,}96, +\infty)$.

Според формула, за z_0 добивам -13,45174 каде оваа вредност припаѓа во критичниот домен и поради тоа се отфрла нултата хипотеза и се прифаќа алтернативната, а тоа е дека просекот на GPA од сите апликанти не е еднаков на 9.

Барање 3, Тест за распределба

```
>
> # Test za raspredelba
> shapiro.test(GPA)       # Shapiro-Wilk normality test

        Shapiro-Wilk normality test

data:  GPA
W = 0.9933, p-value = 0.07235

>
> ks.test(GPA, "pnorm", prosek1, std1)

        One-sample Kolmogorov-Smirnov test

data:  GPA
D = 0.044395, p-value = 0.4097
alternative hypothesis: two-sided

Warning message:
In ks.test(GPA, "pnorm", prosek1, std1) :
  ties should not be present for the Kolmogorov-Smirnov test
>
>
> library(e1071)          #install.packages("e1071", dep = TRUE, type = "source")
>
> kurtosis(GPA)           # Excess kurtosis
[1] -0.4803728
> skewness(GPA)           #Naklonosta na histogramot
[1] -0.06549644
>
```

Слика 11, Тестови за распределба

Повторно, за почеток ја изразувам претпоставката, овој пат тестирам дали GPA податоците имаат (приближно) нормална распределба:

H_0 : GPA податоците имаат нормална распределба

H_a : GPA податоците немаат нормална распределба

Постојат повеќе методи за тестира1ње на нормална распределба, јас ги одбрав двата најкористени и тоа се Shapiro – Wilk's тест и Kolmogorov – Smirnov тест.

Shapiro – Wilk's методот е општо препорачан за тест за нормалност и притоа обезбедува подобра моќност од K – C. Тој се заснова на корелацијата помеѓу податоците и соодветните нормални оцени.
Користев готова функција во R, shapiro.test(). Од излезот (види Слика 11)
p - вредноста > 0,05, што подразбира дека дистрибуцијата на податоците не се значително различни од нормалната распределба. Со други зборови, H_0 се прифаќа и можеме да претпоставиме дека има нормална распределба.

За Kolmogorov – Smirnov тест, уште наречен K-C тест, R исто така има вградена функција, ks.test(). Оваа функција прима два аргументи, првиот е обележјето кое се тестира, вториот е видот на распределба за која се проверува. Бидејќи тестирам за нормална распределба, мора исто така да ги назначам просекот и стандардната девијација како дополнителни аргументи во функцијата, ако тоа не го направам, тестот ќе се направи на стандардна gaussian распределба. Добиената p – вредност е поголема од 0,05, при што заклучуваме дека дистрибуцијата на GPA податоците не е значително различна од нормална, односно нултата хипотеза се прифаќа.

Покрај овие два тестови за распределба, направив и тестови за обликот на графикот на обележјето, skewness и kurtosis. За да се употребат овие формули во R потребен е пакетот “e1071”.

Вишокот Kurtosis ја опишува формата на опашката на дистрибуцијата на податоците. Нормалната распределба има нула вишок kurtosis и поради тоа има стандардна форма на опашката. Добиената негативна вредност укажува на дистрибуција на тенка опашка, со други зборови се нарекува platykurtic.

Интуитивно, the skewness или искривеноста на графикот е мерка на симетрија. Од добиениот резултат може да се увиди дека негативната вредност укажува на тоа дека просекот на податоците е помал од медијаната, што е и точно.

Барање 4, Тестирање хипотези за независност

H_0: Веројатноста за прием на магистарски студии не зависи од додипломскиот успех на студентите.
H_a: Веројатноста за прием на магистарски студии зависи од додипломскиот успех на студентите.

За да ја тестирам независноста на овие две обележја користев два методи.

Според првиот метод, употребувајќи го пакетот ‘MASS’ и функцијата table(), направив табела на контингенција на двете обележја (види Слика 12).

```
> # Testiranje na hipotezi za nezavisnost
> library(MASS)
> tbl = table(Osnovna$CGPA, Osnovna$Chance.of.Admit)
> tbl                  # Tabela na kontingencija

      0.34 0.36 0.38 0.39 0.42 0.43 0.44 0.45 0.46 0.47 0.48 0.49 0.5 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59 0.6 0.61 0.62 0.63 0.64 0.65 0.66 0.67
  6.8    0    1    0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.2    0    0    0    0    0    0    0    0    1    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.25   0    0    0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    1    0    0   0    0    0    0    0    0    0    0
  7.28   0    0    0    0    0    0    0    0    0    1    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.3    0    0    0    0    0    0    0    0    0    0    0    0   0    0    0    0    1    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.34   0    0    0    0    1    0    0    0    0    1    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.36   0    0    0    0    0    0    0    0    1    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.4    0    0    0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    1    0    0    0
  7.43   1    0    0    0    0    0    0    0    0    0    0    1   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.46   0    0    0    0    0    0    0    1    0    0    0    0   0    0    0    0    1    0    0    0    1    0   0    0    0    0    0    0    0    0
  7.5    0    0    0    0    0    0    1    0    0    0    0    0   0    0    0    1    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.54   0    0    0    0    0    0    1    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.56   0    0    0    0    0    0    0    0    0    1    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.6    0    0    0    0    0    0    0    0    1    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.64   0    0    0    0    0    0    0    0    0    1    0    1   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.65   0    0    0    1    0    0    0    0    0    0    0    0   0    0    0    0    0    0    1    1    1    0   0    0    0    0    0    0    0    0

      0.68 0.69 0.7 0.71 0.72 0.73 0.74 0.75 0.76 0.77 0.78 0.79 0.8 0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89 0.9 0.91 0.92 0.93 0.94 0.95 0.96 0.97
  6.8    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.2    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.25   0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.28   0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.3    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.34   0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.36   0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.4    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.43   0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.46   0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.5    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.54   0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.56   0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.6    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.64   0    0   1    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
  7.65   0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0    0    0   0    0    0    0    0    0    0    0
[ reached getOption("max.print") -- omitted 152 rows ]
```

Слика 12, Табела на контингенција

```
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 11054, df = 9853, p-value < 2.2e-16

Warning message:
In chisq.test(tbl) : Chi-squared approximation may be incorrect
>
```

Следно, употребив Пирсонов χ^2 тест, преку готовата функција chisq.test(), каде што добив многу мала вредност за p, помала од 0,05, што ми укажува дека треба да се отфрли нултата хипотеза и да се прифати алтернативната, а тоа е дека двете обележја се зависни.

Вториот метод се состои од корелациски тест, користејќи ја функцијата cor.test(). За ова ми е потребен пакетот "ggpubr".

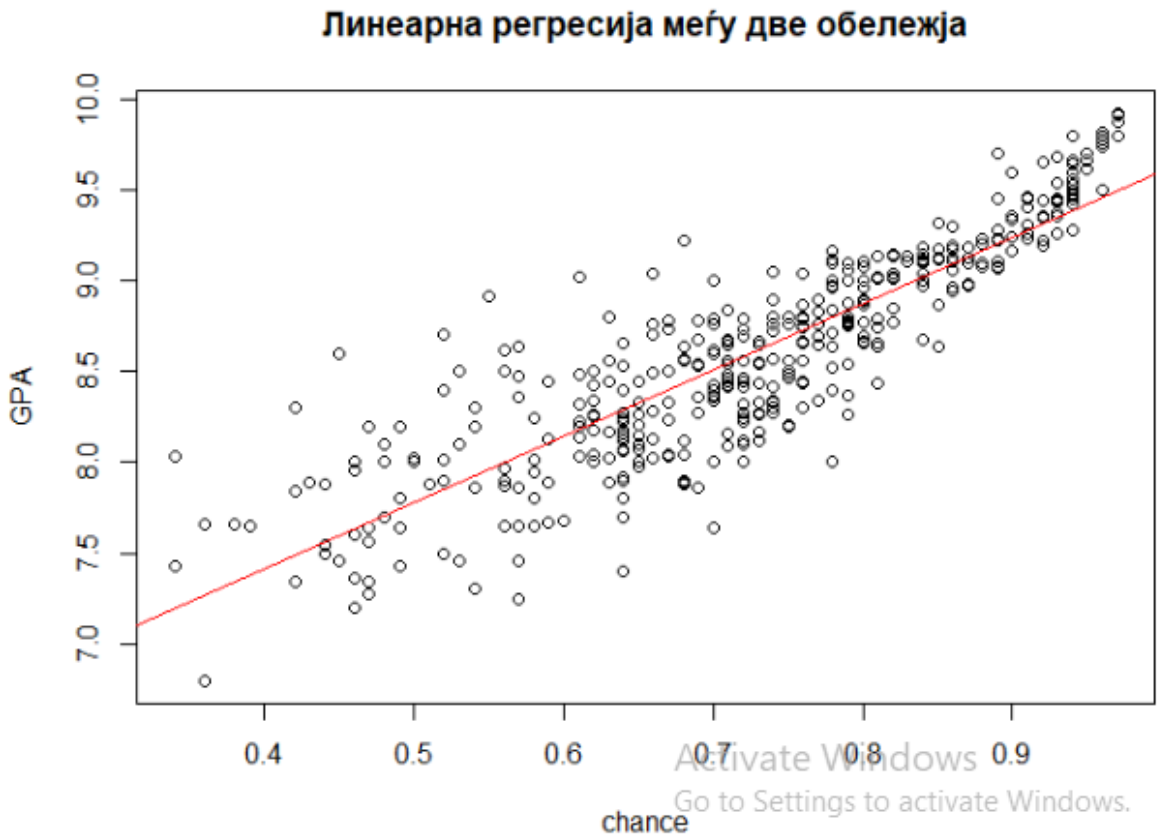
```
> 
> library("ggpubr")      #install.packages("ggpubr")
> cor.test(GPA, chance, method=c("pearson", "kendall", "spearman"))

        Pearson's product-moment correlation

data:  GPA and chance
t = 35.759, df = 398, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8478354 0.8947275
sample estimates:
      cor
0.8732891
```

Според добиените вредности (вредност на т-статистика = 35,759, степени на слобода = 398 и P - вредност < 2,2e-16) може да се заклучи дека p вредноста е значително помала од 0,05, што значи се отфрла нултата хипотеза и се прифаќа дека двете обележја се зависни.
Покрај ова, пресметано е и 95% интервал на доверба (0,8478354, 0,8947275) и коефициентот на корелација 0,8732891, која е истата вредност со претходно пресметаниот коефициент, користејќи ја функцијата cor(), што ги потврдува пресметките.

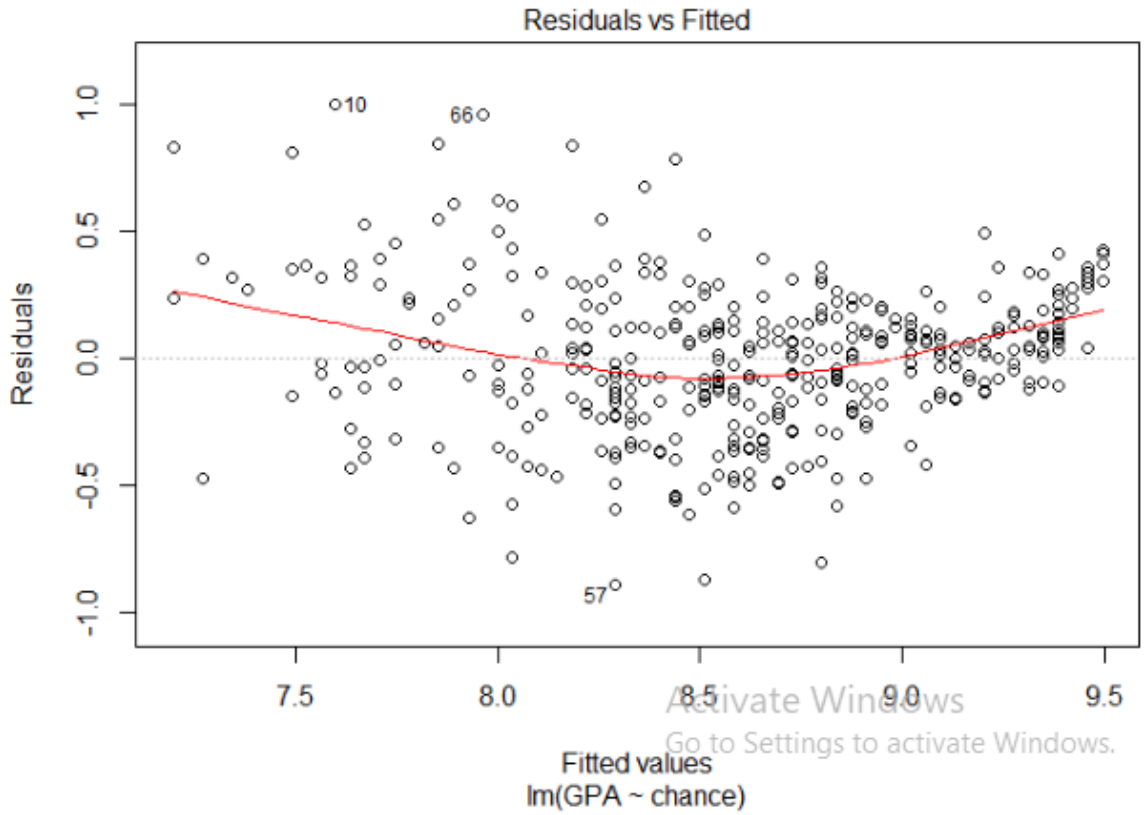
Барање 5, Регресиона анализа

```
> # Regresiona analiza
> model = lm(GPA~chance)
> round(coefficients(model), 3)
(Intercept)      chance
      5.954       3.652
> plot(GPA~chance, main = "Линеарна регресија меѓу две обележја")
> abline(reg = model, col="red")
> Z = coefficients(model)
> summary(model)

Call:
lm(formula = GPA ~ chance)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89091 -0.16933  0.01528  0.16943  1.00290

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.95386    0.07538   78.98   <2e-16 ***
chance       3.65163    0.10212   35.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2909 on 398 degrees of freedom
Multiple R-squared:  0.7626,    Adjusted R-squared:  0.762
F-statistic:  1279 on 1 and 398 DF,  p-value: < 2.2e-16

> model

Call:
lm(formula = GPA ~ chance)

Coefficients:
(Intercept)       chance
      5.954        3.652

> Z
(Intercept)      chance
   5.953863    3.651635
> plot(model)
Hit <Return> to see next plot: termplot(model)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
>
```

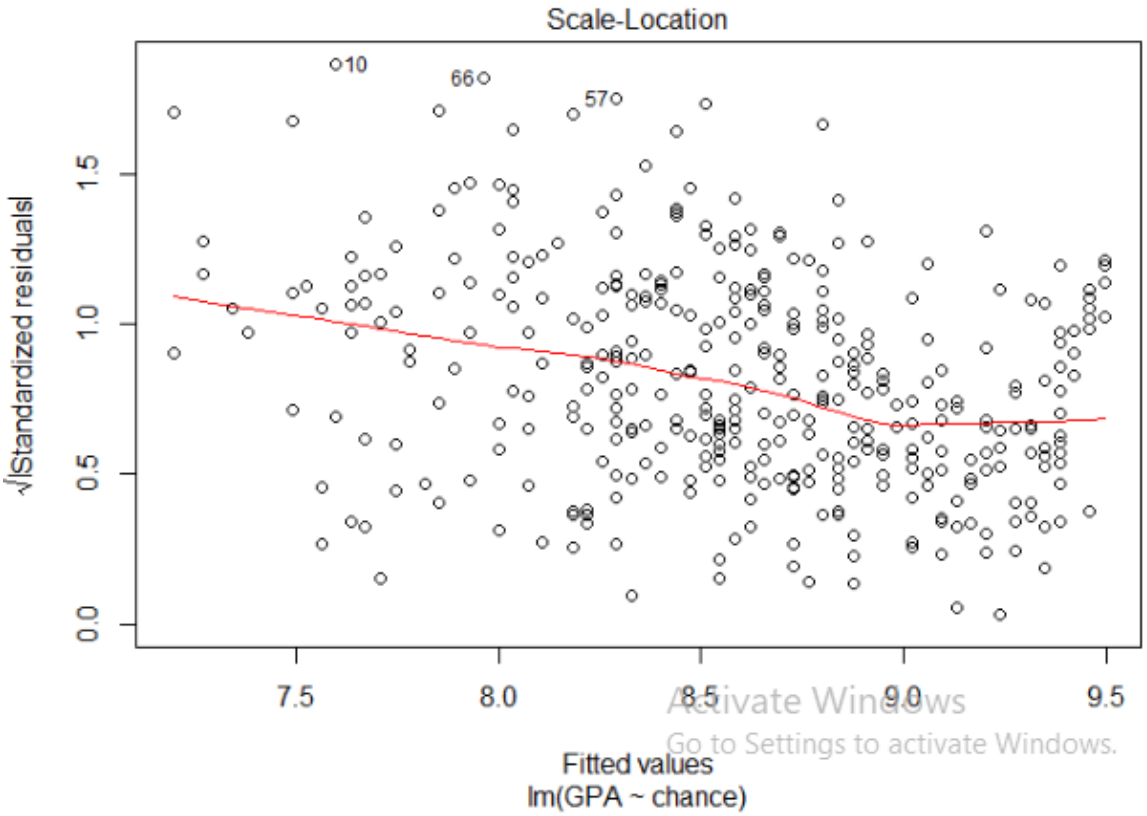
За да направам регресионен модел, најпрво ја користам функцијата lm() која што наоѓа линеарна зависност. Следно преку функцијата round() ја заокружувам вредноста на коефициентите на три децимали. Plot() ми го исцртува графикот, додека па со abline() се исцртува правата на линеарната регресија. Z ми е променлива која ги содржи коефициентите на правата. Преку summary() се испишува дополнителен опис за регресиониот модел како стандардната грешка, коефициент на детерминираност и сл. Следно ја повикувам функцијата за да се добијат коефициентите на правата и потоа ги печатам.
Равенката на проценетата права е $y = 5.954 + 3.652x$, каде $b_0 = 5.954$ и претставува пресек со y – оската, додека па $b_1 = 3.652$ и го претставува правецот на правата или наклонот, во овој случај бидејќи е позитивна вредност, можеме да заклучиме дека правата ќе расте од лево кон десно.
Преку функциите plot() и termplot() се исцртуваат дополнителни графици за недостатоците на моделот и стандардизирани остатоци наспроти теоретски квартали.



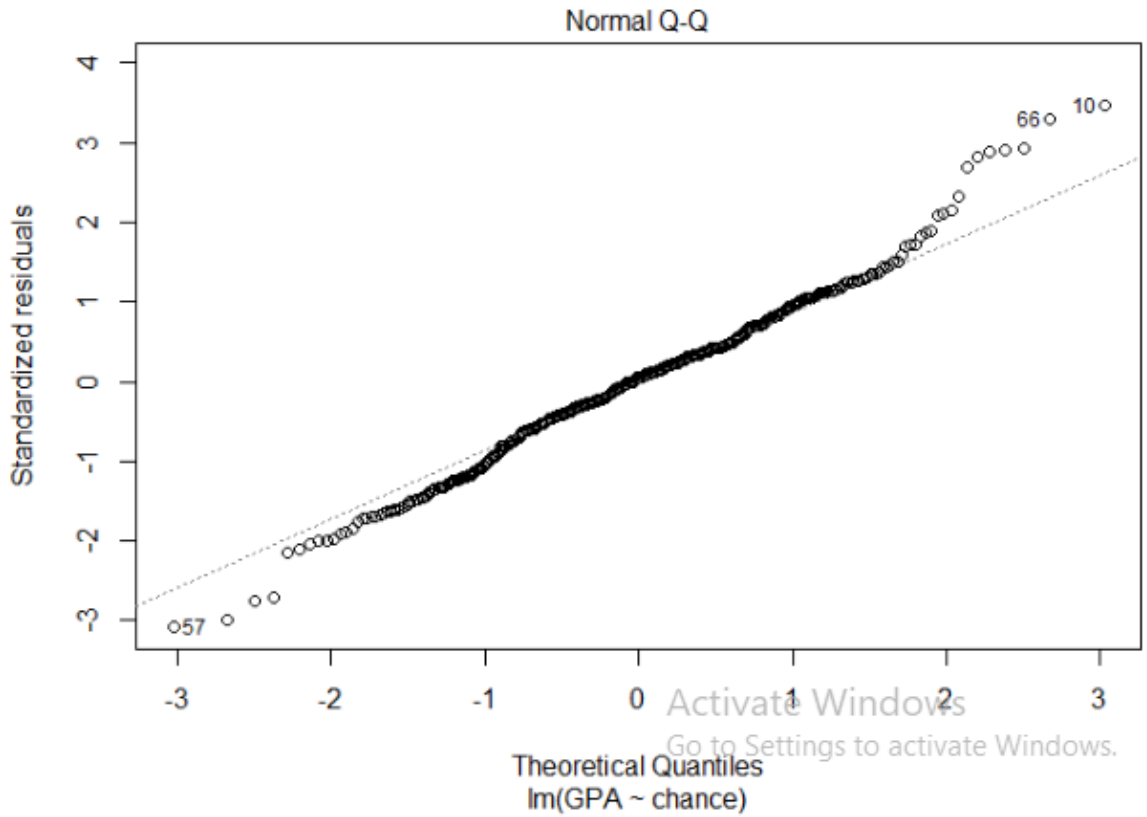
Слика 14, Линеарна регресија

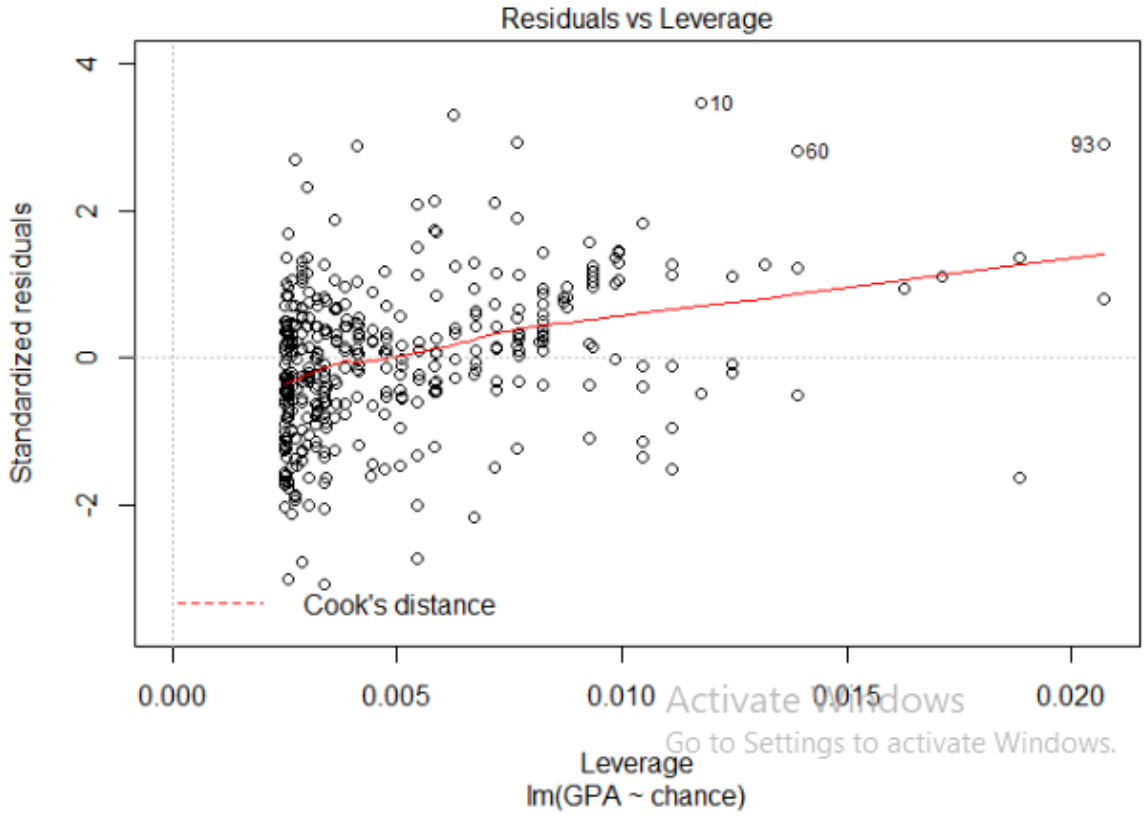


Слика 13, Недостатоци на моделот



Слика 15, Стандардизирани остатоци





Заклучок

Како за крај би сакал да ги сумаризирам изведениве пресметки во една целина. Со помош на табелата на честоти, хистограмот и полигоните добиваме јасна слика за тоа дека успехот од додипломските студии игра најголем удел во приемот. Најголем број на студенти, кои се запишуваат на магистарски студии, имаат завршено со додипломски успех во интервал од 8.0 – 9.0 и веројатноста дека овие студенти ќе бидат примени во програмата е прилично голема, просечно 0.73, што е добро, тоа значи дека колку повисок успех, толку поголеми се шансите за влез, а графикот на расејување како и тестовите за независност ја потврдија оваа претпоставка.