УВОД У СТАТИСТИКУ час 14

31. maj ’17.

Пример 1

У извесном производном погону испитивана је веза између дневне температуре измерене у подне (у °C) и броја дефектних делова произведених тог дана (укупно 22 дана). Добијени су следећи подаци:

Рбр. дана	Температура	Бр. дефектних делова
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32.0	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26.0	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24

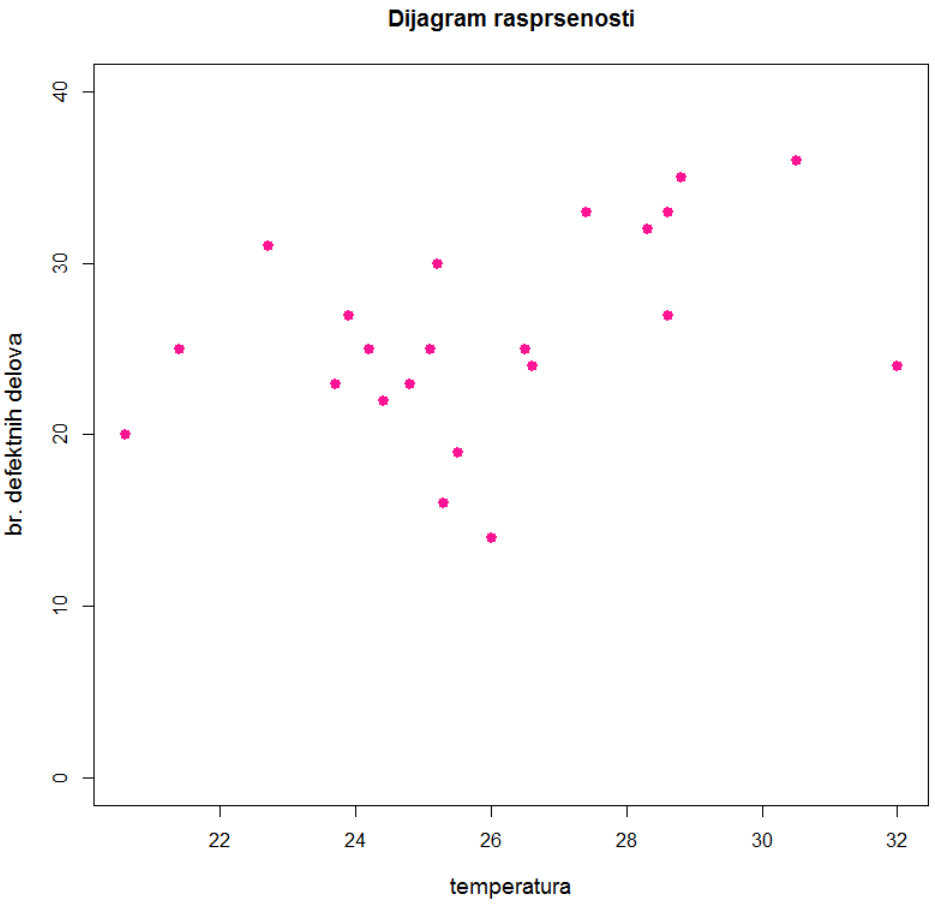
Израчунати узорачки коефицијент (линеарне) корелације два посматрана обележја.

Решење:

Реализована вредност узорачког коефицијента корелације је:

$$r_{X,Y} \approx 0.41894,$$

која указује на релативно слабу позитивну (линеарну) корелацију између обележја:
дневна температура (X) и број дефектних делова произведених тог дана (Y).
На дијаграму десно (scatter plot) приказани су парови вредности из реализованог узорка као тачке у Декартовом координатном систему. △



Пример 2

Следећи подаци издвојени су из опсежније студије приказане у једном броју часописа *Motor Trend* (САД) из 1974. г. Испитивана је потрошња горива и још 10 других карактеристика дизајна и перформанси аутомобила на узорку од 32 аутомобила (модели из сезоне ’73/74. г).

Издвојена су следећа обележја:

- mpg – број пређених миља по (US) галону горива
- disp – кубикажа мотора (у in^3)
- hp – бруто снага мотора
- drat – задњи осовински однос
- wt – тежина (у $1000\ \text{lbs}$)
- qsec – дужина временског периода потребног да аутомобил кренувши из стања мировања пређе $1/4$ миље.

Израчунати узорачки коефицијент корелације за сваки пар описаних обележја.

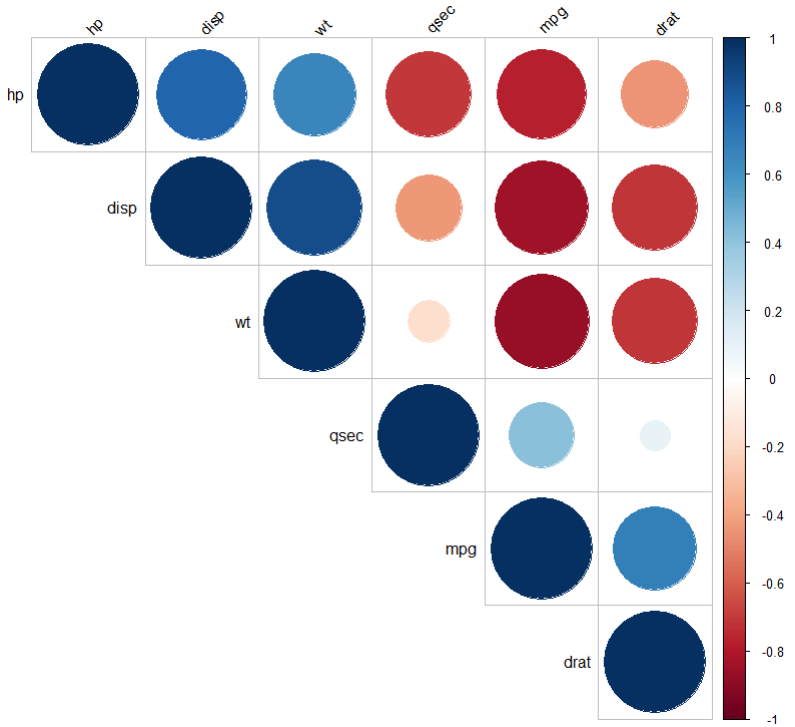
Решење:

Матрица лево је (корелациона) матрица која садржи узорачке коефицијенте корелације за све могуће парове обележја. Ова матрица је, наравно, симетрична. Са десне стране налази се један од начина визуелизације корелационе матрице, који се састоји у томе да се нумеричке вредности коефицијената корелације замењују симболима који одговарају нивоу, тј. јачини корелације.

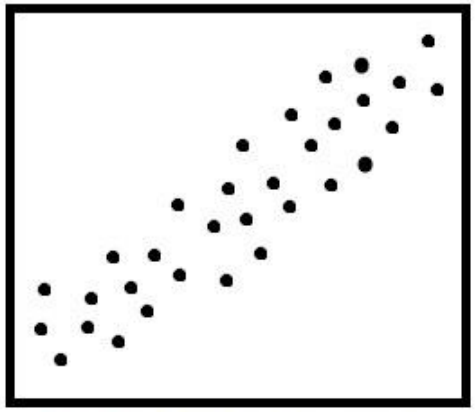
```
             mpg       disp         hp       drat         wt       qsec
mpg    1.0000000 -0.8475514 -0.7761684  0.6811719 -0.8676594  0.4186840
disp  -0.8475514  1.0000000  0.7909486 -0.7102139  0.8879799 -0.4336979
hp    -0.7761684  0.7909486  1.0000000 -0.4487591  0.6587479 -0.7082234
drat   0.6811719 -0.7102139 -0.4487591  1.0000000 -0.7124406  0.0912048
wt    -0.8676594  0.8879799  0.6587479 -0.7124406  1.0000000 -0.1747159
qsec   0.4186840 -0.4336979 -0.7082234  0.0912048 -0.1747159  1.0000000
```

```
     mpg disp hp drat wt qsec
mpg  1
disp +   1
hp   ,   ,    1
drat ,   ,    .  1
wt   +   +    ,  ,    1
qsec .   .    ,          1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```

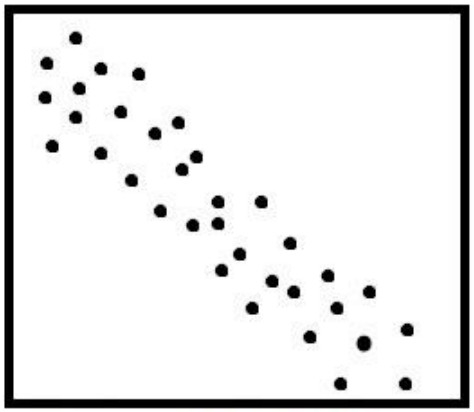
Графички приказ корелационе матрице, којим се истичу најјаче корелирана обележја, је корелограм:



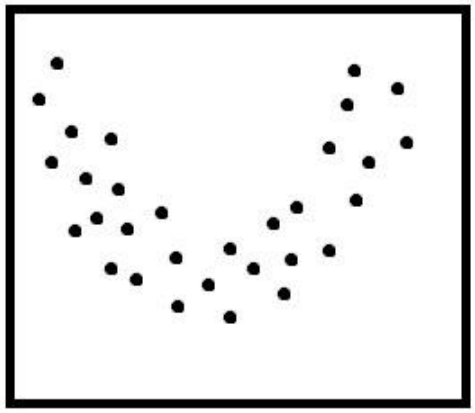
Различити типови зависности два обележја:



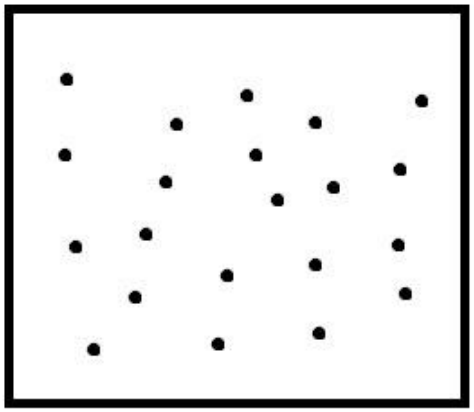
позитивна линеарна веза



негативна линеарна веза

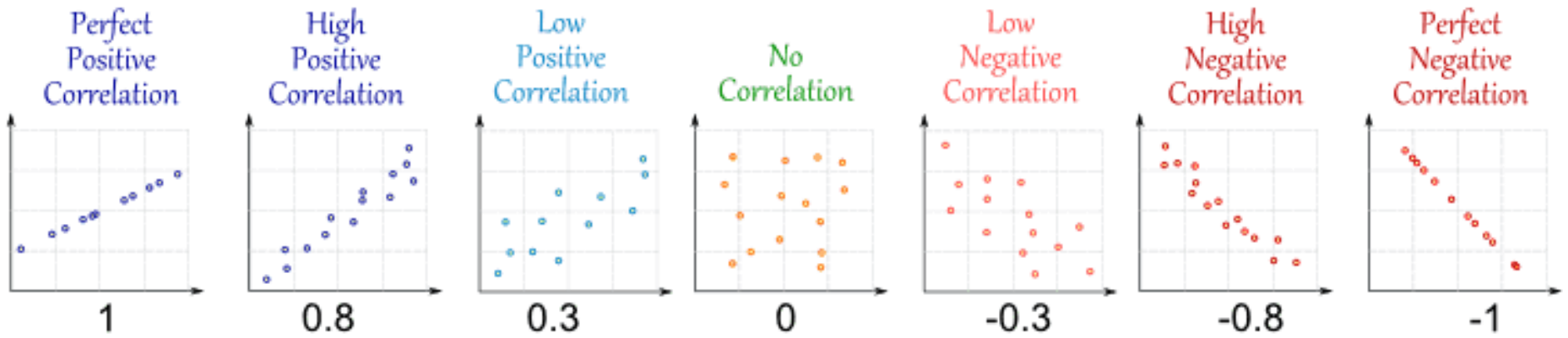


нелинеарна веза



одсуство зависности

Проста линеарна регресија

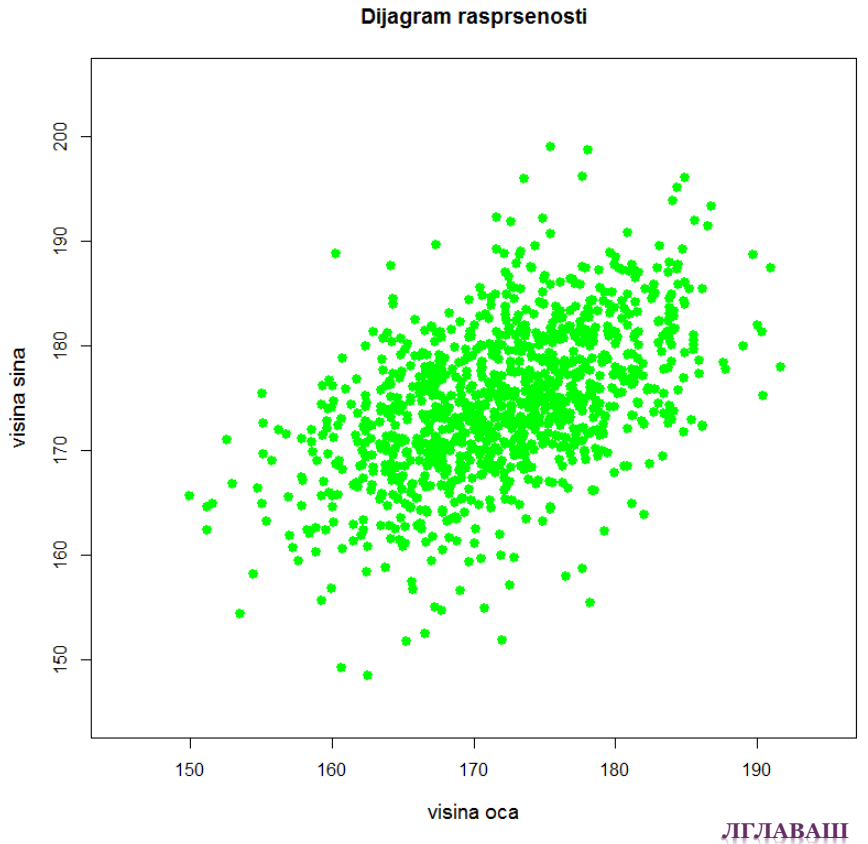


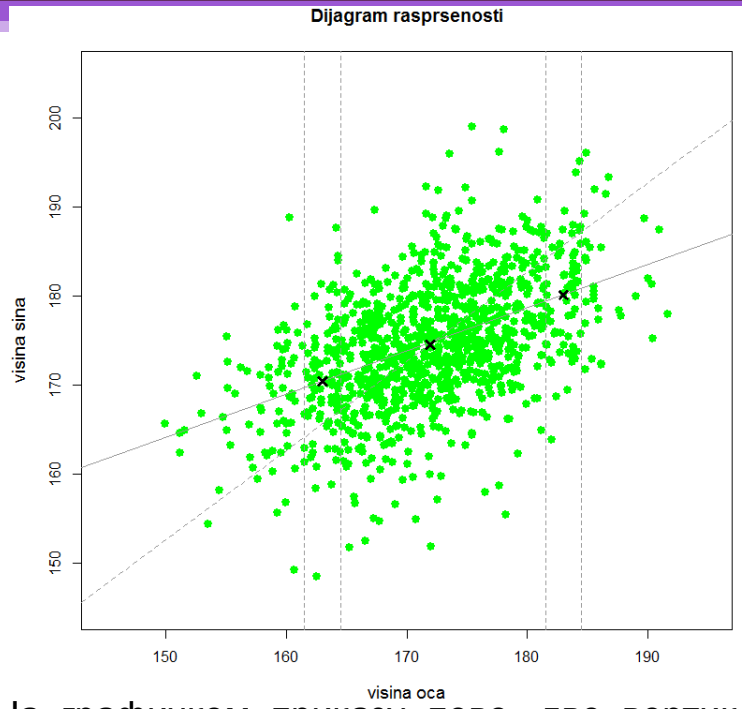
Пример 3

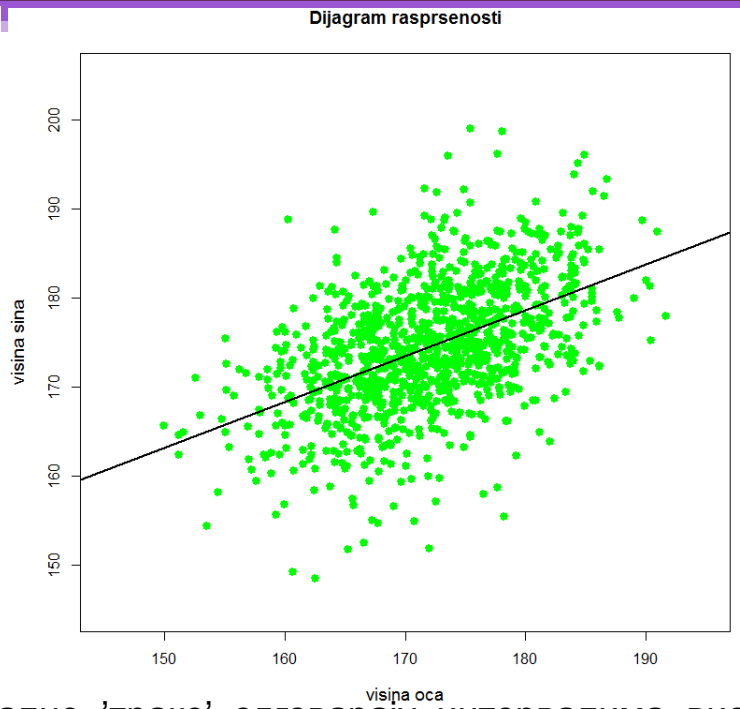
Статистичари викторијанске Енглеске (друга половина XIX в) били су фасцинирани идејом о могућем квантификовању наследних утицаја. У покушајима реализације ове идеје прикупљали су велике количине података.

Sir Francis Galton, научник из XIX в. (Енглеска, 1822-1911. г), постигао је известан напредак на овом пољу разматрајући у којој мери деца личе на своје родитеље. Он је први користио термин „регресија“. Његов ученик Karl Pearson спровео је истраживање чији ће резултати укратко бити наведени. Наиме, као део студије, он је прикупио мерења висине за 1078 очева и њихових синова у одраслом добу (погледати базу података и Пример 2 у Презентацији 4).

Веза између обележја: висина оца (X) и висина сина (Y) може се прво визуелизовати путем дијаграма распршености (десно).
На први поглед, евидентна је позитивна (линеарна) повезаност X и Y.

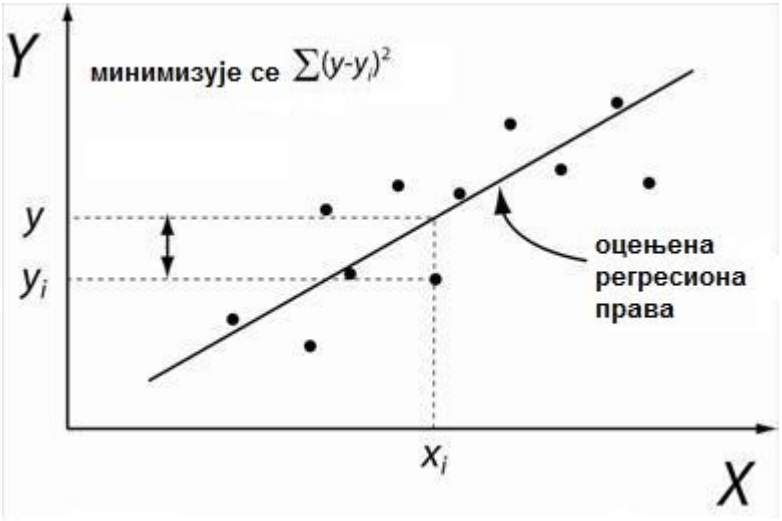






На графичком приказу лево, две вертикалне 'траке' одговарају интервалима висина очева $[161.5, 164.5)$, односно $[181.5, 184.5)$ (то су симетрични интервали око вредности 163, односно 183). Тачке означене крстићем у свакој од трака имају, редом, x-координате 163, односно 183, а y-координате су просечне вредности висина синова, чији очеви имају вредности висина у наведеним интервалима (то су, заправо, аритметичке средине y-координата тачака које припадају вертикалним тракама). Те просечне вредности уствари су извесна условна очекивања висина синова при услову да су задате висине њихових очева. Уколико би се на читавом распону висина очева исцртале траке и одредиле одговарајуће тачке аналогне овима означеним крстићем оне би требало да буду груписане око регресионе праве (тј. регресиона права била би „поравната" верзија линије која спаја ове тачке). Испрекидана права по углом од $45°$ пролази кроз тачку $(\bar{x}_n, \bar{y}_n) = (171.9, 174.5)$.

На графичком приказу десно уцртана је права линеарне регресије, чији су праметри (регресиони коефицијенти) одређени методом најмањих квадрата.



Излаз функције којом се креира адекватан модел линеарне регресије (статистички пакет R):

```
Call:
lm(formula = visina.sina ~ visina.oca)

Residuals:
     Min       1Q   Median       3Q      Max
-22.5957  -3.8614   0.0091   4.1230  22.7570

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 86.10257    4.65558   18.49   <2e-16 ***
visina.oca   0.51391    0.02706   18.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.191 on 1076 degrees of freedom
Multiple R-squared:  0.2511,    Adjusted R-squared:  0.2504
F-statistic: 360.8 on 1 and 1076 DF,  p-value: < 2.2e-16
```

Реализоване вредности статистика које се користе:

$\bar{x}_n$	$\bar{y}_n$	$\frac{1}{n}\sum_{j=1}^{n} x_j y_j$	$\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x}_n)^2$	$\frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y}_n)^2$	$r_{X,Y}$	$\hat{a}$	$\hat{b}$
171.925	174.457	30018.56	48.56851	51.08527	0.50272	0.51391	86.10246

Сада се може извршити тестирање хипотезе $H_0(\rho_{X,Y} = 0)$ против $H_1(\rho_{X,Y} > 0)$. Под претпоставком нормалне расподељености резидуала користи се тест статистика:

$$T = R_{X,Y} \cdot \sqrt{\frac{n-2}{1-R_{X,Y}^2}},$$

која, при тачној H_0, има Студентову расподелу са $n-2$ степена слободе. Реализована вредност тест статистике је: $t \approx 19.07623$. Излаз функције која врши тестирање:

```
        Pearson's product-moment correlation

data:  visina.oca and visina.sina
t = 18.994, df = 1076, p-value < 2.2e-16
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.4625878 1.0000000
sample estimates:
      cor
0.5010942
```

па би нулту хипотезу за праг значајности нпр. $\alpha = 0.01$ требало одбацити.

Ради потврђивања адекватности / коректности модела линеарне регресије приступа се анализи резидуала. У ситуацији када је регресиони модел коректан тзв. стандардизовани резидуали:

$$\frac{Y_j - \hat{Y}_j}{\sqrt{SS_R/(n-2)}}, j = 1, 2, \dots, n,$$

где је SS_R сума квадрата резидуала, требало би да су приближно независне случајне величине са стандардном нормалном расподелом, те би, стога, требало да су на случајан начин расподељене око 0 са око 95% вредности концентрисаних у интервалу $[-2, 2]$ (на основу 2-сигма правила).

Додатно, на дијаграму стандардизованих резидуала не би требало да се појави никакав посебан (систематски) образац у распореду вредности нити варијације резидуала са променом величине оцењене вредности ($\hat{Y}_j$). Свака аномалија на дијаграму требало би да побуди сумњу нпр. о неједнакости дисперзија резидуала (својство хетероскедастичности), односно, у крајњем случају, о ваљаности самог претпостављеног модела просте линеарне регресије. △