VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

DUOMENŲ MOKSLAS. BAKALAURAS

4 Laboratorinis darbas

Išgyvenamumo analizė

Ataskaita

Užduotį atliko: Ugnius Vilimas

Rytis Baltaduonis

Justinus Pipiras

Vilnius 2024

Turinys

Įvadas

Tikslas

Atlikti išgyvenamumo analizę mūsų pasirinktam duomenų rinkiniui apie moterų išgyvenamumą nuo krūties vėžio.

Uždaviniai

- Pasirinkti tinkamus duomenis daryti išgyvenamumo analizę.
- Atlikti pradinę duomenų analizę.
- Sukurti modelį.
- Interpretuoti rezultatus.
- Parašyti išvadas.

Duomenys

Pasirinkome duomenis apie moterų krūties vėžio išgyvenamumą nuo 1978 iki 1993. Duomenys imti iš R paketo pavadinimu „survival“, duomenų pavadinimas – „Rotterdam“. Mūsų duomenų rinkinį sudaro 2982 eilutės ir 15 stulpelių su, per 15 metų surinktais duomenimis apie krūties vėžiu sirgusių moterų būklę ir charakteristikomis. Analizei atlikti mes pasirinkome naudoti tokius kintamuosius:

- age – paciento amžius.
- meno – menopauzės statusas (0 - prieš, 1 - po).
- size – naviko dydis (milimetrais).
- grade – naviko stadija.
- nodes – teigiamų limfmazgių skaičius.
- pgr – progesterono receptoriai.
- er – estrogenų receptoriai.
- hormon – hormoninis gydimas (0 - nebuvo, 1- buvo).
- chemo – chemoterapija (0 - nebuvo, 1 - buvo).
- dtime – laikas iki tiriamo įvykio.
- death – ar žmogus mirė ar ne (0 – pasveiko, 1 – mirė).

Kaip priklausomąjį kintamąjį imame „death“, kiti kintamieji bus nepriklausomi.

Pradinė analizė

Pirmiausia atliekame pradinę analizę ir pradedame viską nuo priklausomojo kintamojo susipažinimo. Mūsų atveju jis yra kategorinis kintamasis, kuris nusako ar moteris pasveiko ar mirė nuo krūties vėžio, todėl galime sudaryti lentelę. 0 reiškia, kad pacientė pasveiko, o 1 reiškia, kad pacientė mirė.

1 lentelė pasveikusių ir nepasveikusių pacienčių pasiskirstymas

Pasveiko - 0	Mirė - 1
1710	1272

Matome, kad pasveikusių ir mirusių pacientų skaičius yra gana vienodas. Tai mums gali padėti sudarant modelį.

Toliau pažiūrėsime kitus kategorinius parametrus ir jų pasiskirstymus tokius kaip: menopauzės, hormonų, chemoterapijos statusas, paplitimo dydis bei krūties vėžio stadija. Visi jie pavaizduoti lentelėmis

2 lentelė pacientų pasiskirstymas pagal menopauzės statusą

Prieš menopauzę - 0	Po menopauzės - 1
1312	1670

3 lentelė pacientų pasiskirstymas pagal hormonų vartojimą

Nevartojo hormonų - 0	Vartojo hormonus - 1
2643	339

4 lentelė pacientų pasiskirstymas pagal atliktą chemoterapiją

Nedaroma chemoterapija - 0	Daroma chemoterapija - 1
2402	580

5 lentelė pacientų auglio dydžio pasiskirstymas

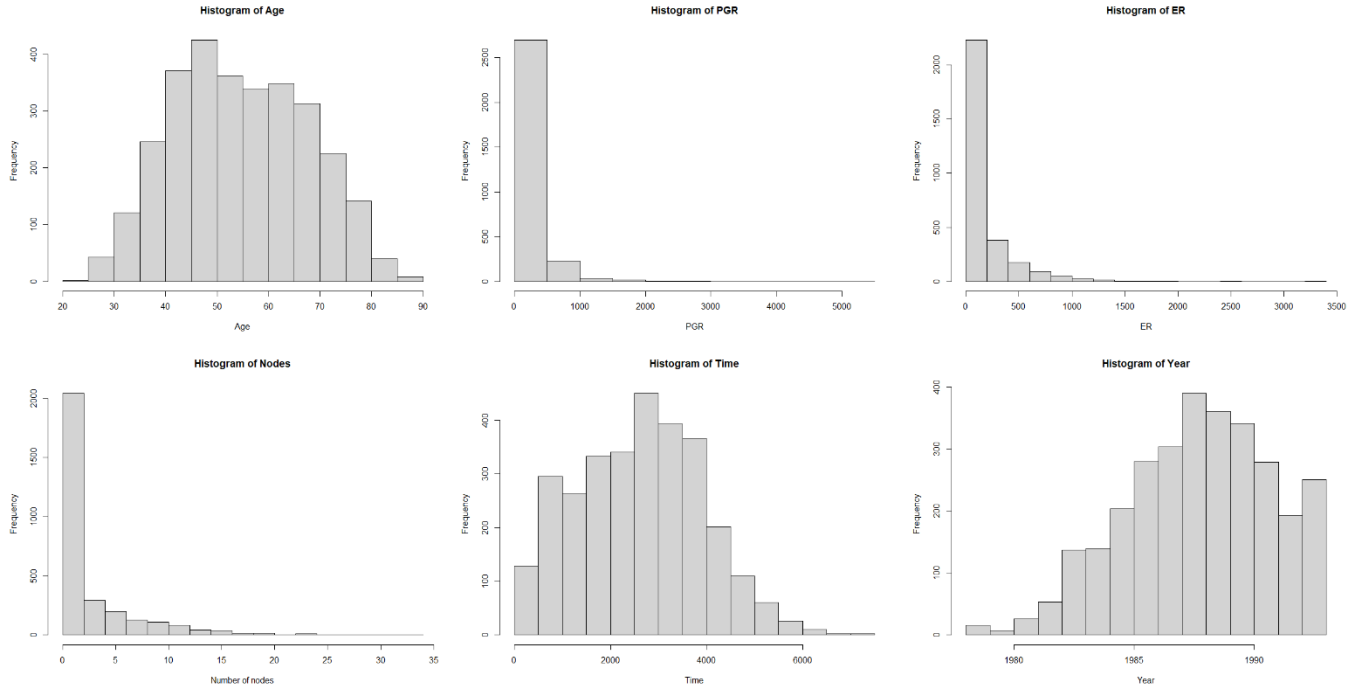
<=20mm	20-50mm	>50mm
1387	1291	304

6 lentelė pacientų pasiskirstymas pagal vėžio stadiją

Antra vėžio stadija - 2	Trečia vėžio stadija - 3
794	2188

Aiškiai matome, kad daugumoje pasiskirstymai yra nelabai tolygūs, bet tai neturėtų trukdyti mūsų modelio kūrimui ir rezultatų interpretavimui.

Toliau galime pasižiūrėti į mūsų kiekybinių duomenų pasiskirstymo grafikus. Turėjome tokius kiekybinius duomenis: paciento amžius, sirgimo metai, limfmazgių kiekis, progesterono ir estrogenų receptorių kiekis bei laikas, po kurio atsitiko įvykis. Visi jie pavaizduoti dažnių histogramomis.



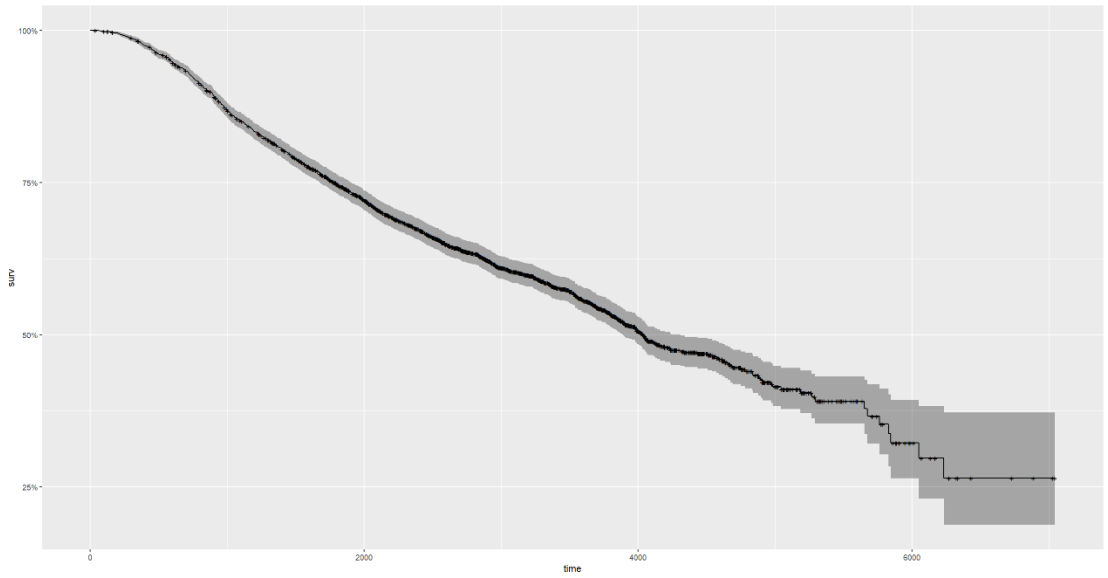
1 pav. visų kiekybinių kintamųjų dažnių histogramos

Aiškiai matome, kad turime reikšmių, kurios yra neįprastos ir ekstremalios, tačiau pačias išskirtis tikrinsime jau tik sudarę modelį.

Toliau atliekame cenzūravimą iš dešinės, kurio metu sujungiame priklausomąjį kintamąjį „death“, kuris nurodo ar žmogus pasveiko ar mirė, su nepriklausomu kintamuoju „Time“, kuris nurodo kiek dienų reikėjo įvykiui atsitikti.

Analizė su „R“

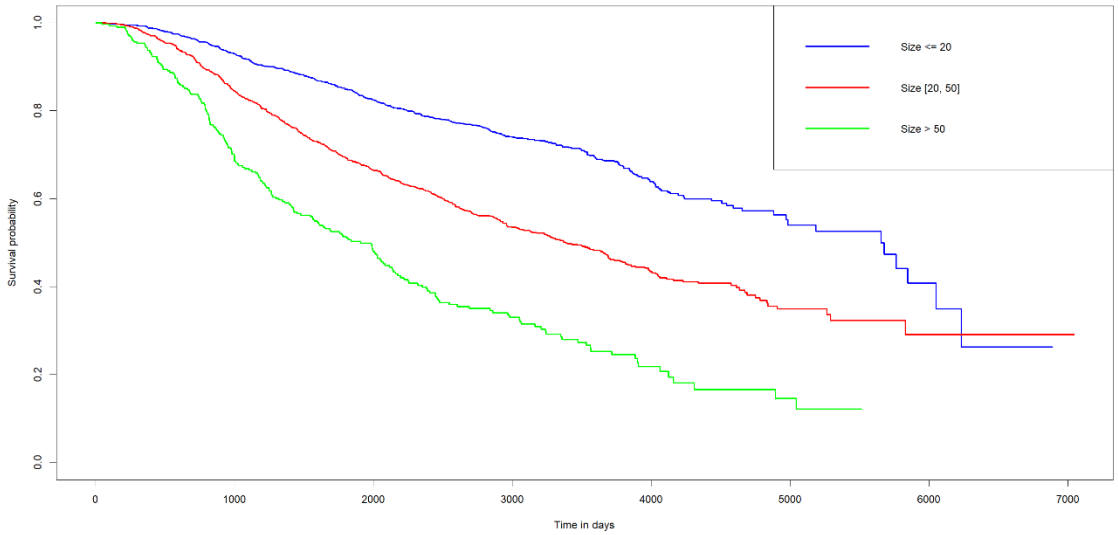
Pirmiausia pavaizduojame „Kaplan-Meier“ įvertį grafike, kuris rodo, kaip keičiasi išgyvenimo tikimybė per laiką.



2 pav. „Kaplan-Meier“ grafikas

Mūsų grafike galime pamatyti, kad jeigu jau yra sergama bent 6000 dienų (~16 metų), tikimybė išgyventi yra tik apie 25%.

Čia matome tokį patį „Kaplan-Meier“ grafiką, tik jau išskirstytą pagal tris kategorijas: Auglio dydis iki 20mm, tarp 20 ir 50mm bei daugiau nei 50mm. Grafike kategorijos išskirstytos spalvomis.



3 pav. „Kaplan-Meier“ grafikas pagal auglio dydžio kategorijas

Toliau atliekame homogeniškumo hipotezės tikrinimą. Pasitelkiame logranginį kriterijų, taip pat naudojome ir Gehan-Wilcoxon kriterijaus Peto ir Peto modifikaciją.

```
> survdiff(Surv(Galutinis$dtime, Galutinis$death)~Galutinis$size, rho = 0)
Call:
survdiff(formula = Surv(Galutinis$dtime, Galutinis$death) ~ Galutinis$size,
    rho = 0)

                        N Observed Expected (O-E)^2/E (O-E)^2/V
Galutinis$size=<=20  1387      414    655.7      89.1       185
Galutinis$size=20-50 1291      646    525.7      27.5        47
Galutinis$size=>50    304      212     90.6     162.6       176

 Chisq= 281  on 2 degrees of freedom, p= <2e-16
> survdiff(Surv(Galutinis$dtime, Galutinis$death)~Galutinis$size, rho = 1)
Call:
survdiff(formula = Surv(Galutinis$dtime, Galutinis$death) ~ Galutinis$size,
    rho = 1)

                        N Observed Expected (O-E)^2/E (O-E)^2/V
Galutinis$size=<=20  1387      308      502      75.1       192
Galutinis$size=20-50 1291      503      408      22.0        47
Galutinis$size=>50    304      172       73     135.3       178

 Chisq= 287  on 2 degrees of freedom, p= <2e-16
>
```

4 pav. homogeniškumo tikrinimas, logranginis ir Gehan-Wilcoxon kriterijai

Iš 4 pav. matome, kad abiejų kriterijų p reikšmės yra mažesnės negu mūsų nustatytas reikšmingumo lygis alfa (0.05), tai reiškia, kad kuo didesnis auglio dydis, tuo mažesnis išgyvenamumas.

Taigi dabar galime pasižiūrėti modelį ir reikšmingus kintamuosius, kuriuos ir paliksime modelyje.

```
Call:
coxph(formula = Surv(dtime, death) ~ year + age + meno + size +
    grade + nodes + pgr + er + hormon + chemo, data = Galutinis)

  n= 2982, number of events= 1272

               coef  exp(coef)   se(coef)      z Pr(>|z|)
year      -3.042e-02  9.700e-01  1.010e-02 -3.011 0.002602 **
age        1.334e-02  1.013e+00  3.840e-03  3.473 0.000516 ***
meno       5.486e-02  1.056e+00  1.009e-01  0.544 0.586520
size20-50  4.204e-01  1.523e+00  6.578e-02  6.390 1.65e-10 ***
size>50    8.066e-01  2.240e+00  9.129e-02  8.835  < 2e-16 ***
grade      3.304e-01  1.391e+00  7.096e-02  4.656 3.23e-06 ***
nodes      7.270e-02  1.075e+00  4.887e-03 14.878  < 2e-16 ***
pgr       -3.647e-04  9.996e-01  1.220e-04 -2.989 0.002797 **
er        -4.985e-05  1.000e+00  1.099e-04 -0.453 0.650219
hormon     8.837e-03  1.009e+00  9.162e-02  0.096 0.923163
chemo      4.429e-02  1.045e+00  8.202e-02  0.540 0.589162
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5 pav. modelis su visais kintamaisiais

Matome, kad turime keletą kintamųjų, kurių p reikšmės viršija mūsų nustatytą reikšmingumo lygį alfa, todėl jas pašalinsime ir atliksime pažingsninę regresiją.

Toliau atliekame pažingsninę regresiją ir paliekame tik reikšmingus kintamuosius.

```
Call:
coxph(formula = Surv(dtime, death) ~ year + size + grade + nodes +
    pgr + age, data = Galutinis)

  n= 2982, number of events= 1272

                coef  exp(coef)   se(coef)      z Pr(>|z|)
year       -0.0304964  0.9699639  0.0097281 -3.135  0.00172 **
size20-50   0.4196132  1.5213730  0.0655379  6.403 1.53e-10 ***
size>50     0.8065130  2.2400831  0.0904456  8.917  < 2e-16 ***
grade       0.3318539  1.3935493  0.0708031  4.687 2.77e-06 ***
nodes       0.0734350  1.0761986  0.0047409 15.490  < 2e-16 ***
pgr        -0.0003825  0.9996176  0.0001169 -3.272  0.00107 **
age         0.0140644  1.0141638  0.0022500  6.251 4.08e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Concordance= 0.694  (se = 0.008 )
Likelihood ratio test= 532.7  on 7 df,   p=<2e-16
Wald test            = 618.3  on 7 df,   p=<2e-16
Score (logrank) test = 689.3  on 7 df,   p=<2e-16
```

6 pav. modelis tik su reikšmingais kintamaisiais

Čia matome, kad yra likę tik reikšmingi kintamieji: size, grade, nodes, pgr ir age. Jų p reikšmės yra mažesnės už alfa. Kitus kintamuosius teko pašalinti iš modelio.

Tada tikriname multikolinearumą su VIF.

```
> VIF(cox2)
    year    nodes      pgr      age
1.013191 1.008160 1.003665 1.013974
```

7 pav.

Toliau testuojame PH prielaidą. Funkcija „cox.zph“ realizuotas PH prielaidos tikrinimo kriterijus. Jis yra skaičiuojamas kiekvienai kovariantei bei grindžiamas Schoenfeld liekanų koreliacija su transformuotu laiko kintamuoju (pagal nutylėjimą naudojamas K-M įvertis).

```
> cox.zph(cox1)
        chisq df       p
year    0.347  1 0.55559
size    5.128  2 0.07699
grade   3.226  1 0.07248
nodes   3.760  1 0.05250
pgr    42.645  1 6.6e-11
age    14.750  1 0.00012
GLOBAL 61.394  7 7.9e-11
```

8 pav. PH prielaida

Prielaidai patenkinti ir modelio tinkamumui nusakyti, p reikšmės turi gautis didesnės negu alfa. Matome, kad 2 regresoriai netenkina šitos sąlygos, todėl turėsime atlikti sluoksniavimą ir tikrinti prielaidą dar kartą.

Taigi toliau atliekame sluoksniavimą ir tuos 2 regresorius, kurie nepatenkino PH prielaidos sąlygos, išskirstome į keturias kategorijas.

```
> xtabs(~ age.cat, data=Galutinis)
age.cat
   1    2    3    4
 412 1157  999  414
> xtabs(~ pgr.cat, data=Galutinis)
pgr.cat
   1    2    3    4
2242  456  171  113
>
```

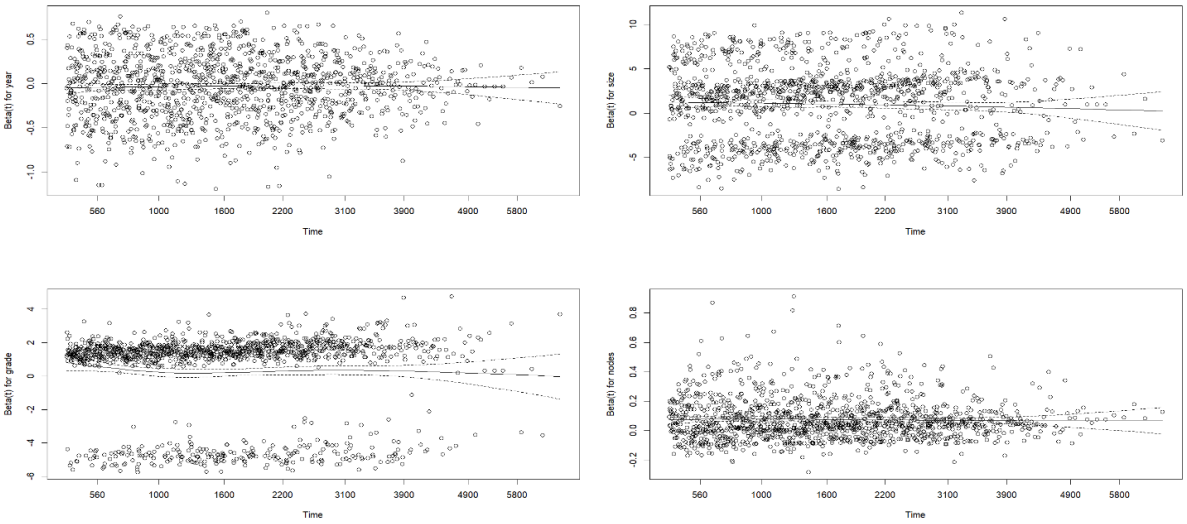
9 pav. Sluoksniavimas

Čia matome, kaip programa šiuos regresorius išskirsto į kategorijas ir dabar vėl galėsime tikrinti PH prielaidą.

```
> cox.zph(cox2)
        chisq df    p
year    0.455  1 0.50
size    3.944  2 0.14
grade   1.416  1 0.23
nodes   1.730  1 0.19
GLOBAL  6.055  5 0.30
```

10 pav. antroji PH prielaida

Jau dabar matome, kad visi regresoriai yra didesni už alfa ir PH prielaida yra patenkinama, todėl galime daryti išvadą, kad modelį galima interpretuoti ir gauti tinkamus rezultatus.



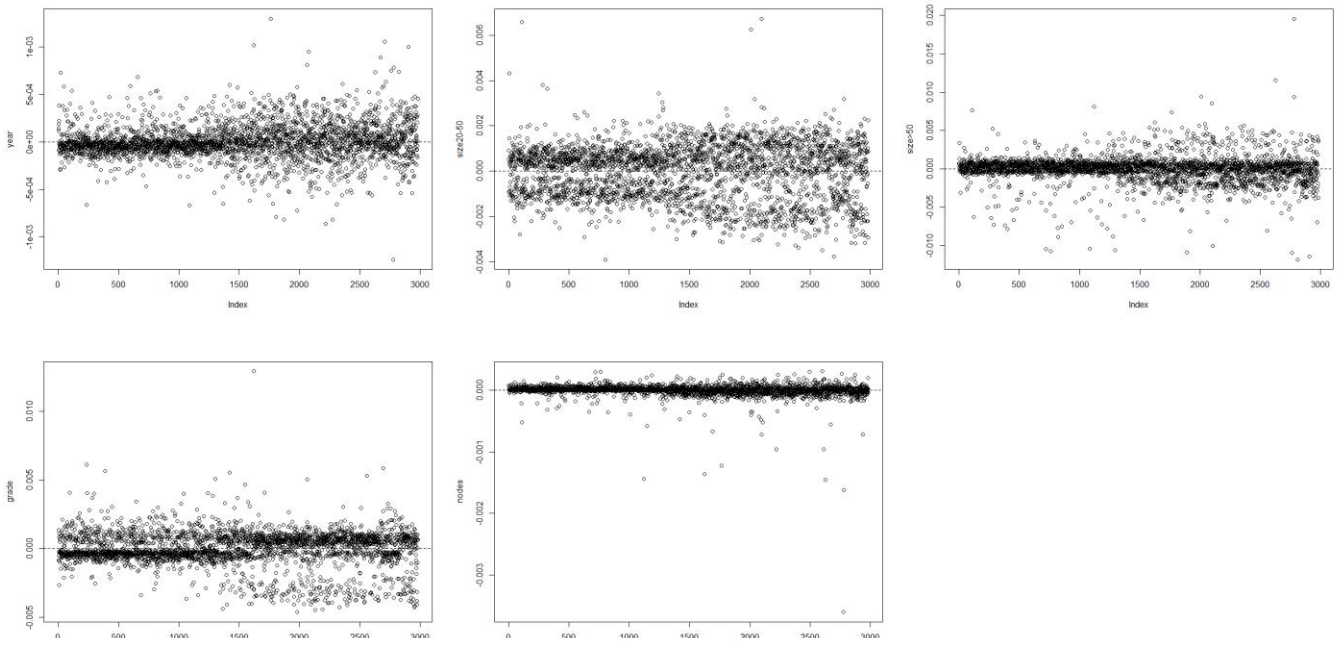
11 pav. prielaidos tikrinimas grafiškai

Galiausiai tikriname išskirtis. Pirmiausia pagal formule susiskaičiuojame rėžį, kuris indikuos, kad modelyje yra išskirčių. Viršutinio rėžio formulė yra: 2/n^0.5 , kur n yra visų stebėjimų skaičius.

```
> 2/sqrt(nrow(Galutinis))
[1] 0.03662488
```

12 pav. rėžio formulė ir atsakymas

Pagal šį rėžį dabar žiūrėsime ar neturime išskirčių duomenyse.

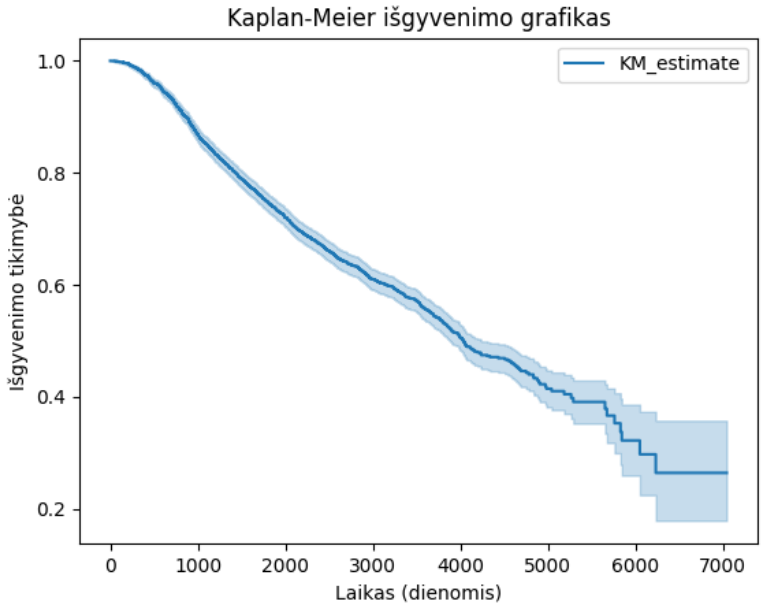


13 pav. liekanų grafikas

Iš šios standartizuotų koeficientų matricos matome, kad mūsų apskaičiuotas rėžis visada yra aukščiausiai ir nėra taško, kuris viršytų rėžį, todėl darome išvadą, kad išskirčių neturėsime.

Analizė su „Python“

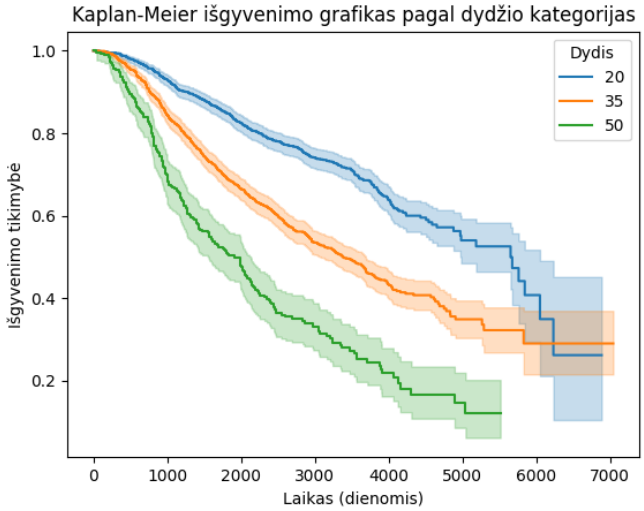
Pirmiausia pavaizduojame „Kaplan-Meier“ įvertį grafike, kuris rodo, kaip keičiasi išgyvenimo tikimybė per laiką.



14 pav. „Kaplan-Meier“ grafikas

Ir vėl mūsų grafike galime pamatyti, kad jeigu jau yra sergama bent 6000 dienų (~16 metų), tikimybė išgyventi yra tik apie 25%.

Čia irgi matome tokį patį „Kaplan-Meier“ grafiką, tik jau išskirstytą pagal tris kategorijas: Auglio dydis iki 20mm, tarp 20 ir 50mm bei daugiau nei 50mm. Grafike kategorijos išskirstytos spalvomis.



15 pav. „Kaplan-Meier“ grafikas pagal auglio dydį

Toliau atliekame homogeniškumo hipotezės tikrinimą. Pasitelkiame logranginį kriterijų, taip pat naudojome ir Gehan-Wilcoxon kriterijaus Peto ir Peto modifikaciją. Kaip ir r programoje, tikrinsime ir su rho = 0 ir su rho = 1.

```
<lifelines.StatisticalResult: logrank_test>
               t_0 = -1
 null_distribution = chi squared
degrees_of_freedom = 1
         test_name = logrank_test

---
 test_statistic      p  -log2(p)
        6556.25 <0.005       inf
```

16 pav. homogeniškumo tikrinimas, logranginis ir Gehan-Wilcoxon kriterijai kai rho = 0

```
                              p    -log2(p)
covariate
size       1.046598e-57  189.284194
```

17 pav. homogeniškumo tikrinimas, logranginis ir Gehan-Wilcoxon kriterijai kai rho = 1

Iš pav. matome, kad abiejų kriterijų p reikšmės yra mažesnės negu mūsų nustatytas reikšmingumo lygis alfa (0.05), tai reiškia, kad kuo didesnis auglio dydis, tuo mažesnis išgyvenamumas.

Taigi dabar galime pasižiūrėti modelį ir reikšmingus kintamuosius, kuriuos ir paliksime modelyje.

Toliau atliekame pažingsninę regresiją ir paliekame tik reikšmingus kintamuosius.

```
          cmp to     z      p  -log2(p)
covariate
year         0.00 -3.17 <0.005      9.33
size         0.00  9.22 <0.005     64.82
grade        0.00  4.69 <0.005     18.48
nodes        0.00 15.49 <0.005    177.37
pgr          0.00 -3.28 <0.005      9.91
age          0.00  6.25 <0.005     31.14
---
Concordance = 0.69
Partial AIC = 18533.98
log-likelihood ratio test = 532.64 on 6 df
-log2(p) of ll-ratio test = 369.09
```

18 pav. modelis

Čia matome, kad yra likę tik reikšmingi kintamieji: size, grade, nodes, pgr ir age. Jų p reikšmės yra mažesnės už alfa. Kitus kintamuosius teko pašalinti iš modelio.

`

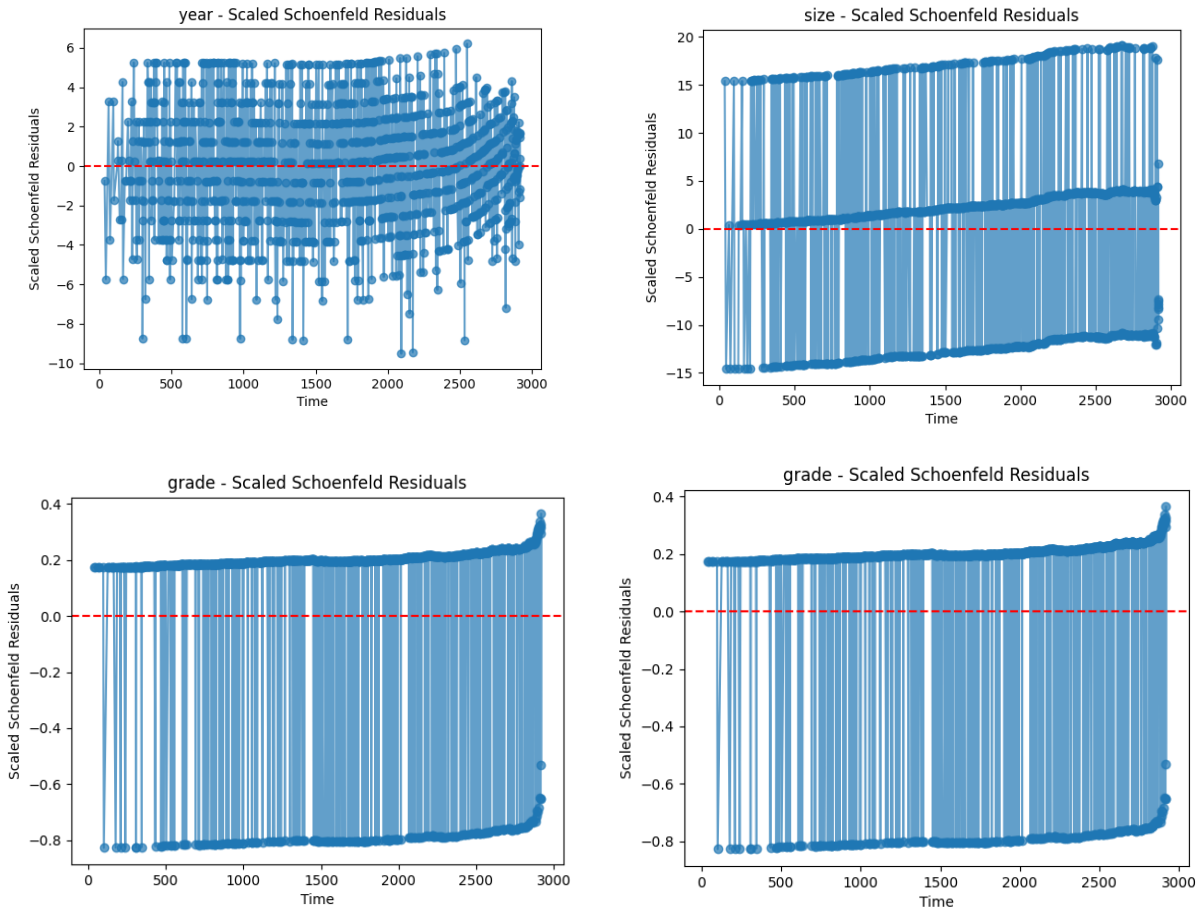
Toliau testuojame PH prielaidą. Jis yra skaičiuojamas kiekvienai kovariantei bei grindžiamas Schoenfeld liekanų koreliacija su transformuotu laiko kintamuoju (pagal nutylėjimą naudojamas K-M įvertis). Per „Python“ programą yra labai paprasta kodo eilutė, kuri mums duoda tikslų ir aiškų atsakymą.

```
result1 = cox.check_assumptions(df1)
print("----------------------------",result1)
```

```
Proportional hazard assumption looks okay.
```

19 pav. PH prielaidos tikrinimas. Kodo eilutė ir rezultatas

Kai matome, kad programa galime daryti išvadą, kad modelį galima interpretuoti ir gauti tinkamus rezultatus.



20 pav. prielaidos tikrinimas grafiškai

Išvados

Gauti rezultatai buvo beveik identiški tiek su „R“ programavimo kalba, tiek su „Python programavimo kalba“. Gavome rezultatus, kad kuo pacientas ilgiau sirgs krūties vėžiu, tuo tikimybė išgyventi mažės. Pagal mūsų gautus grafikus, pažiūrėjus tam tikras grafiko dalis, galime spręsti, jog išgyvenimo priklausomybė labai aiškiai mažėja laikui bėgant. Interpretacija, tokia, kad tikimybė išgyventi yra lygiai 50% yra tada, kai žmogaus sirgimo laikas pasiekia 4000 dienų, o tikimybė išgyventi yra lygiai 25% yra tada, kai žmogaus sirgimo laikas pasiekia 1600 dienų. Tikra smagu gauti tokius rezultatus, ir daryti tokias išvadas, kad žmogus prasirgęs net 5 metus, vis tiek turi labai gerus šansus išgyventi ir pasveikti. Konkordacijos koeficientas gaunamas 0,7, tai reiškia, kad gana gerai galime spėti išgyvenamumą.

Šaltiniai

- https://cran.r-project.org/web/packages/survival/survival.pdf