VILNIAUS UNIVERSITETAS MATEMATIKOS IR INFORMATIKOS FAKULTETAS

DUOMENŲ MOKSLAS

AFT REGRESIJOS MODELIS

2 laboratorinis darbas

Atliko: Simona Gelžinytė,
Ugnė Kniukštaitė,
Laineda Morkytė,
Austėja Valeikaitė DM 4k. 2gr.

Vilnius

2023

TURINYS

1. ĮVADAS

1.1 Tyrimo tikslas

Pritaikyti AFT regresijos modelį pasirinktiems duomenims.

1.2 Tyrimo uždaviniai

- Atlikti pirminę duomenų analizę;
- Patikrinti modelio prielaidas;
- Sukonstruoti modelį;
- Įvertinti modelio gerumą;
- Pateikti gauto modelio interpretacijas;
- Apibendrinti gautus rezultatus, pateikti išvadas.

1.3 Duomenys ir programinė įranga

Pasirinktas duomenų rinkinys apie storosios žarnos vėžio adjuvantinę chemoterapiją. Pateikti įvairūs požymiai apie pacientą. Priklausomas kintamasis – dienų skaičius iki paciento mirties ir 12 kovariančių:

- Statusas – cenzūravimo statusas: 0 – cenzūruota, 1 - mirė;
- Gydymo tipas (konservavimas, levamizolas – mažo toksiškumo, levamizolas +5 – vidutinio toksiškumo);
- Lytis (1 – vyras, 0 – moteris);
- Amžius;
- Obstrukcija – storosios žarnos obstrukcija dėl naviko (0 – nėra, 1 - yra);
- Perforacija[1] – storosios žarnos perforacija (0 – nėra, 1 - yra);
- Prilipimas – prilipimas prie gretimų organų (0 – nėra, 1 - yra);
- Limfmazgiai – limfmazgių, kuriuose aptiktas vėžys, skaičius;
- Diferenciacija – naviko diferenciacija (1 – gera, 2 – vidutinė, 3 – bloga);

[1] prakiurimas, tuščiavidurio organo sienos pažeidimas, po kurio organo ertmė susisiekia su gretimais audiniais ar ertmėmis.

- Apimtis – vietinio išplitimo mastas (1 – submukozė (pogleivinė dalis), 2 – raumenys, 3 – serozė, 4 – gretimos struktūros);
- Operacija – laikas nuo operacijos iki registracijos (0 – trumpas, 1 – ilgas);
- Limfmazgiai4 – daugiau nei 4 teigiami limfmazgiai.

Visos kovariantės kategorinės išskyrus amžių ir vėžinius limfmazgius. Iš viso yra 929 stebėjimai, pašalinus praleistas reikšmes stebėjimų liko 888. Tyrimo metu naudota ,,R“ programinė įranga.

2. AFT REGRESIJOS MODELIS

Parinkus AFT regresijos modelį pereiname visus modelio parinkimo etapus:

- Pradinė duomenų analizė – susipažįstame su kintamaisiais;
- Prielaidų tikrinimas –multikolinearumas ir skirstinio parinkimas;
- Reikšmingų kovariančių atranka;
- Parametrų ir kovariančių koeficientų įvertinimas, interpretacija;
- Modelio gerumo įvertinimas.

Duomenys buvo padalinti į mokymo ir testavimo aibes 80 : 20 santykiu.

2.1 Pradinė duomenų analizė

Prieš pradedant taikyti modelį, susipažinome su duomenimis – kiekybiniams kintamiesiems nusibraižėme stačiakampes diagramas, o kategoriniams – pasižiūrėjome dažnių lenteles priklausomai nuo paciento statuso, t. y. cenzūruotas ar miręs.

1 lentelė. Cenzūruotų stebėjimų ir įvykių dažnių lentelė

Cenzūruotas stebėjimas	Įvykis
458	430

Iš pateiktos lentelės (1 lentelė) matome, jog iš viso mirė 430 pacientų, o likusiems 458 pacientams – įvykis dar neįvyko.

2 lentelė. Kategorinių kintamųjų dažniai (I)

	Gydymo tipas			Lytis		Obstrukcija	
Statusas	**Konservavimas**	**Levamizolas**	**Levamizolas +5**	**Moteris**	**Vyras**	**Nėra**	**Yra**
0	141	145	172	220	238	379	79
1	164	149	117	208	222	338	92

Pagal gydymo tipą dar nemirusiems pacientams dažniau buvo paskirtas levamizolas +5, konservavimo tipo gydymas ir levamizolas buvo paskirti beveik vienodai. Tiems pacientams, kurie mirė dažniausiai priskirtas gydymo tipas – konservavimas, o rečiausias – levamizolas. Tyrime moterų ir vyrų skaičius panašus. Taip pat daugelis pacientų neturėjo storosios žarnos obstrukcijos.

3 lentelė. Kategorinių kintamųjų dažniai (II)

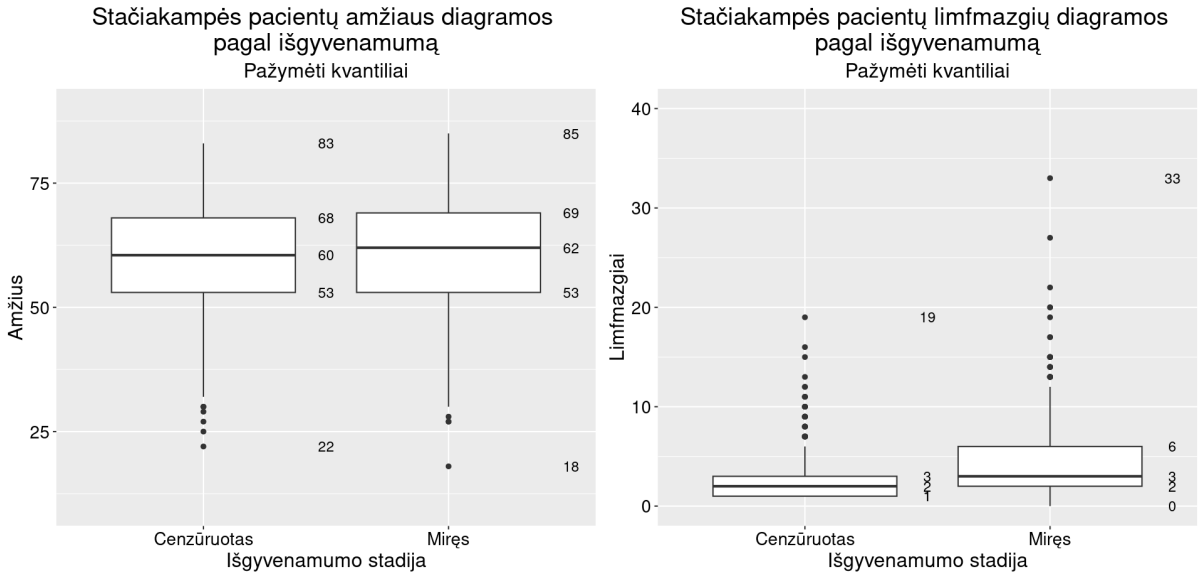
	Perforacija		Prilipimas		Diferenciacija			Operacija	
Statusas	**Nėra**	**Yra**	**Nėra**	**Yra**	**Gera**	**Vidutinė**	**Bloga**	**Trumpas**	**Ilgas**
0	446	12	404	54	49	347	62	349	109
1	415	15	356	74	41	306	83	301	129

Beveik visi pacientai neturėjo perforacijos. Taip pat daugeliui pacientų nepasireiškė prilipimas prie gretimų organų. Naviko diferenciacija tiek dar gyviems, tiek jau mirusiems pacientams panaši – daugiausiai pasitaikė vidutinės diferenciacijos navikų. Daugeliui pacientų laikas nuo operacijos iki registracijos buvo trumpas.

4 lentelė. Kategorinių kintamųjų dažniai (III)

	Apimtis				4 limfmazgiai	
Statusas	**Submukozė**	**Raumenys**	**Serozė**	**Gretimos struktūros**	**Nėra**	**Yra**
0	16	67	362	13	390	68
1	3	35	268	24	263	167

Vietinis išplitimas tiek dar gyviems, tiek mirusiems pacientams panašus – dažniausiai buvo išplitę serozės dalyje, tačiau matome, jog išplitimas pogleivinėje dalyje buvo dažnesnis dar nemirusiems pacientams, o jau mirusiems pacientams dažnesnis išplitimas buvo gretimose struktūrose nei cenzūruotiems.

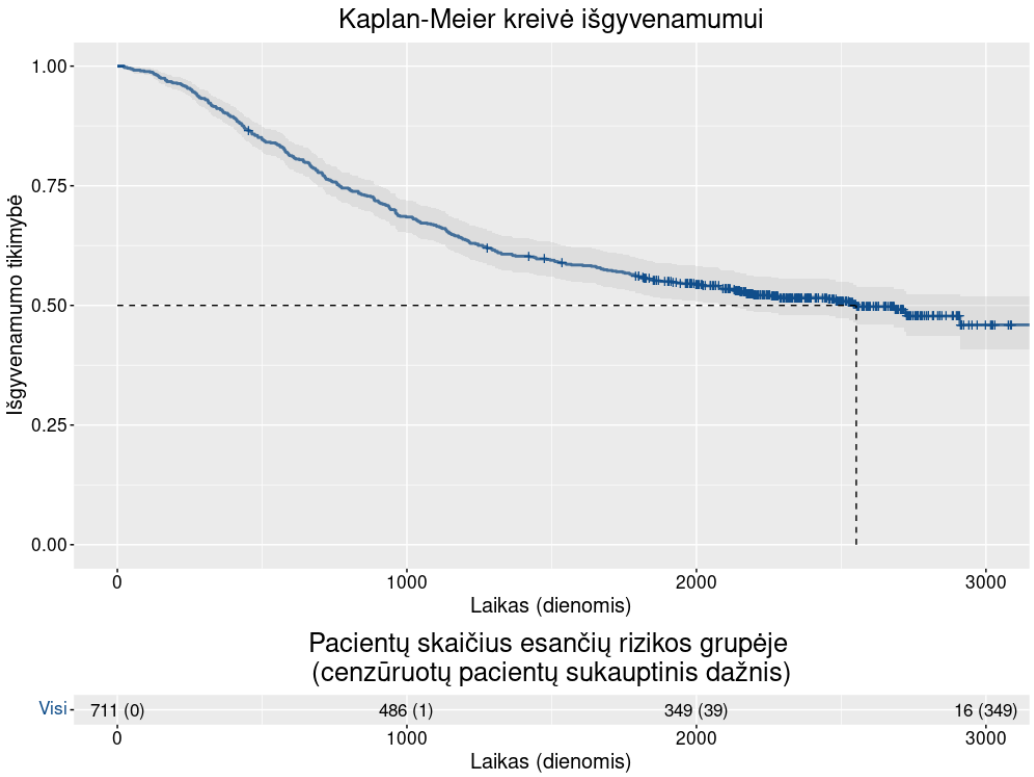


1 pav. Kiekybinių kintamųjų stačiakampės diagramos

Iš stačiakampės diagramos (1 pav.) pagal amžių matome, jog tiek cenzūruotų, tiek mirusių pacientų amžius labai panašus. Taip pat pastebime, jog mirusiems pacientams limfmazgių skaičius buvo didesnis nei cenzūruotiems.

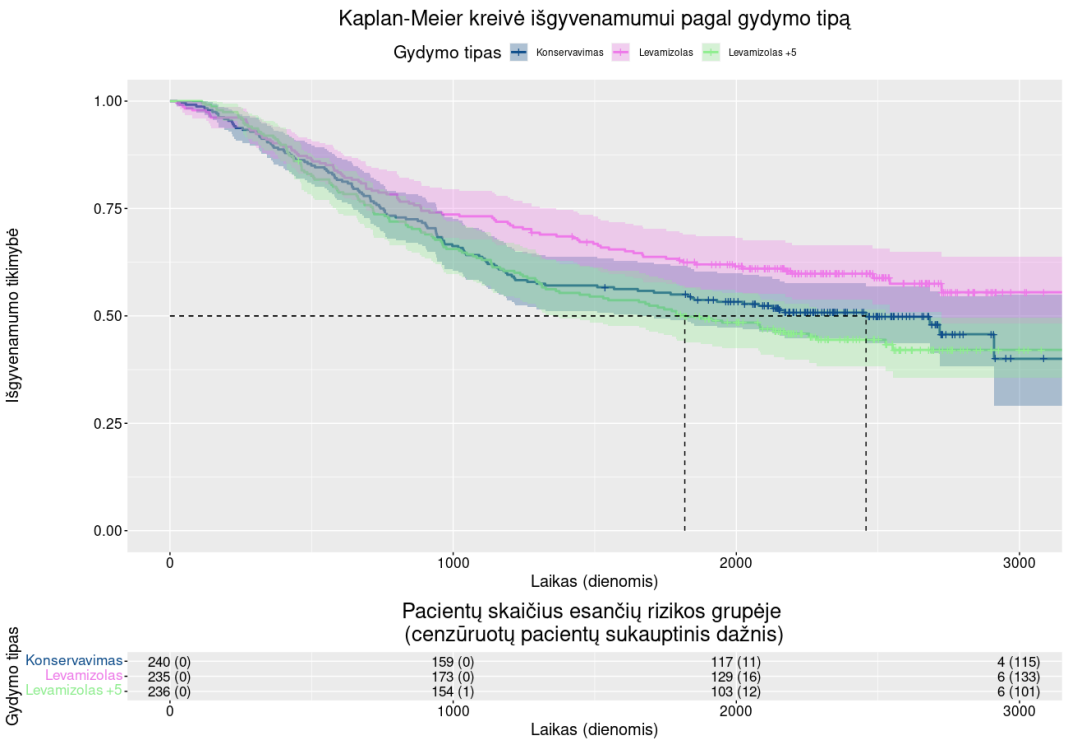
2.2 Kaplan – Meier kreivės

Taip pat buvo nubraižytos Kaplan – Meier kreivės, kad būtų galima pažiūrėti, kaip nuo kategorinių kintamųjų priklauso išgyvenamumo tikimybė bei apskritai kaip ji kinta.



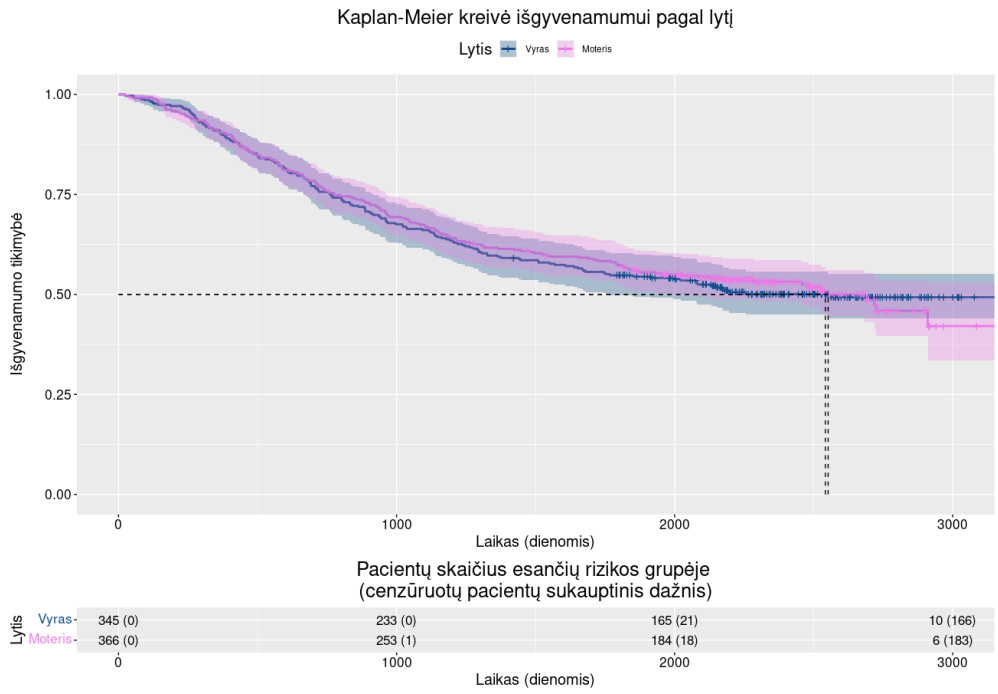
2 pav. Kaplan – Meier kreivė išgyvenamumui

Iš (2 pav.) matome, jog tikimybė išgyventi daugiau nei 2552 dienas (apie 7 m.) nukrenta iki 50 %. Paskutinė mirtis buvo 3329 dieną. Apribotas vidurkio laikas yra 2107 – tai plotas po išgyvenamumo kreive nuo nulinio laiko momento iki paskutinio įvykio momento.



3 pav. Kaplan – Meier kreivė išgyvenamumui pagal gydymo tipą

Iš (3 pav.) matome, jog 120 iš 240 pacientų buvo įvykis, kai jam buvo pritaikytas konservavimo gydymo tipas, 96 iš 235 pacientų buvo įvykis, kai buvo paskirtas levamizolas, o 130 iš 236 pacientų buvo įvykis, kai buvo paskirtas levamizolas +5.

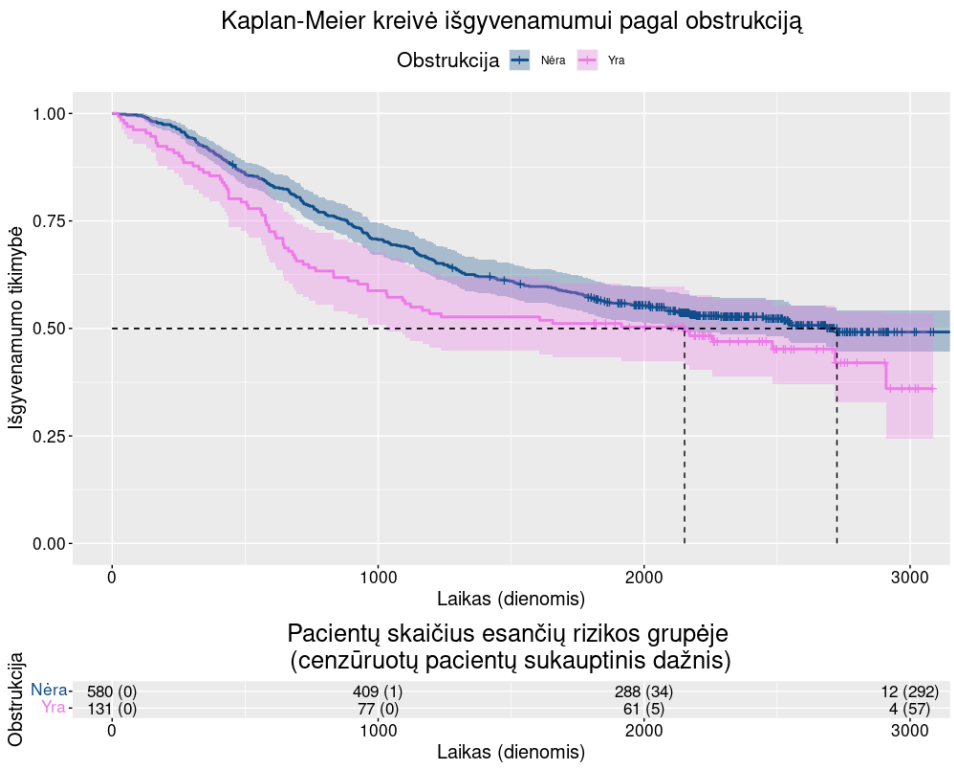


4 pav. Kaplan – Meier kreivė išgyvenamumui pagal lytį

Iš (4 pav.) matome, jog grupes tarpusavyje sunku atskirti, todėl pritaikome *twostage* testą ir tikriname hipotezę:

$$\begin{cases} H_0: nėra\ skirtumų\ tarp\ grupių \\ H_1: yra\ skirtumai\ tarp\ grupių \end{cases}.$$

Gauta p reikšmė – 0,02 (atmetame nulinę hipotezę), todėl grupės statistiškai reikšmingai skiriasi. Iš 345 vyrų mirė 170, o iš 366 moterų mirė 177. Tikimybė, jog moteris išgyvens ilgiau nei 2552 (apie 7 m.) dienas nukrenta iki 50 %, o vyrams tikimybė išgyventi ilgiau nei 2542 (apie 7 m.) dienas nukrenta iki 50 %.

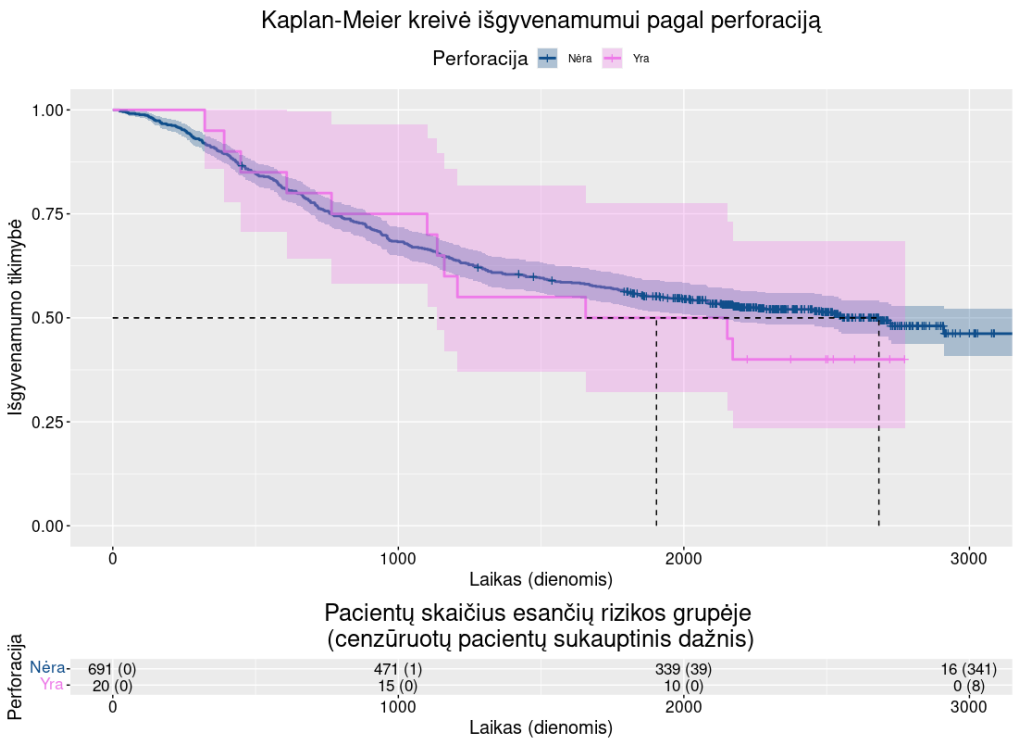


5 pav. Kaplan – Meier kreivė išgyvenamumui pagal obstrukciją

Iš (5 pav.) matome, jog grupių kreivės nesikerta – dėl to grupių palyginimui bus naudojamas log ranginis kriterijus. Tikriname hipotezę:

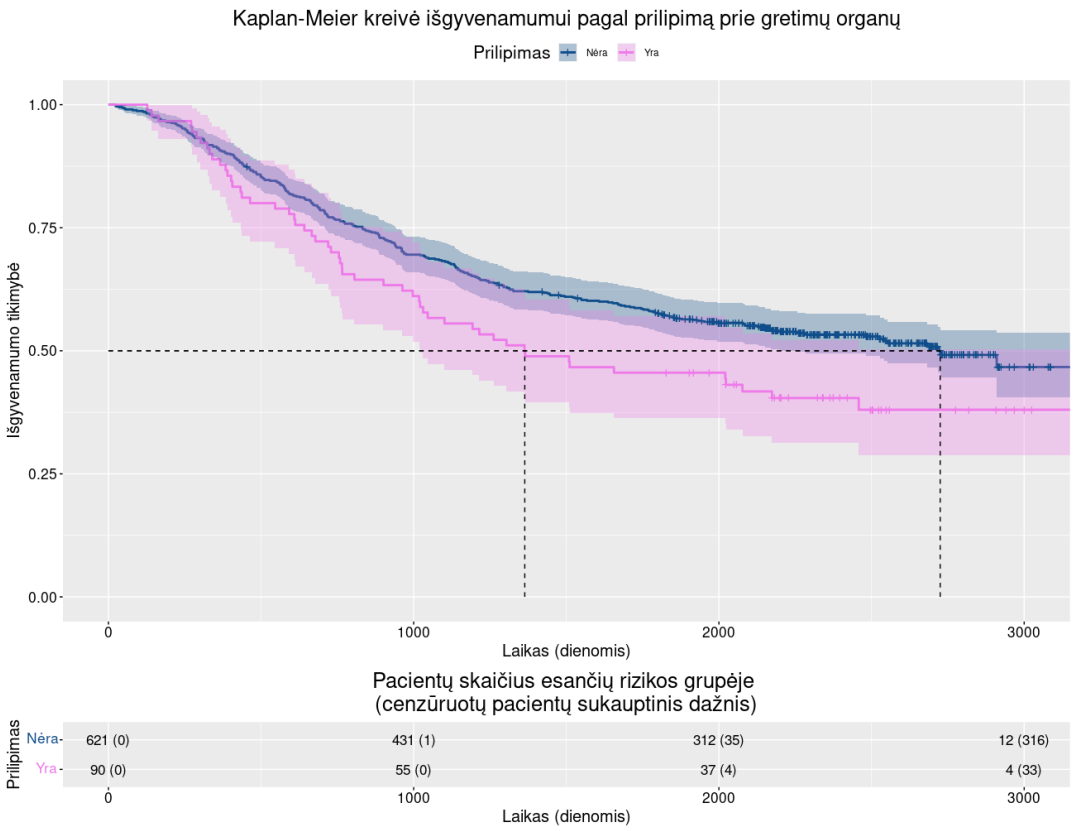
$$\begin{cases} H_0: nėra\ skirtumų\ tarp\ grupių \\ H_1: yra\ skirtumai\ tarp\ grupių \end{cases}.$$

Gavome p reikšmę – 0,055 (nulinė hipotezė neatmetama), vadinasi grupės statistiškai reikšmingai nesiskiria. Iš 580 atvejų, kai pacientui nebuvo obstrukcijos, mirė 276, iš 131 atvejo kai obstrukcija buvo – mirė 71 pacientas. Tikimybė, jog pacientas išgyvens ilgiau nei 2725 dienas (apie 7,5 m.), kai obstrukcijos nėra, nukrenta iki 50 %, o kai ji yra – tikimybė išgyventi ilgiau nei 2152 dienas (apie 6 m.) nukrenta iki 50 %.



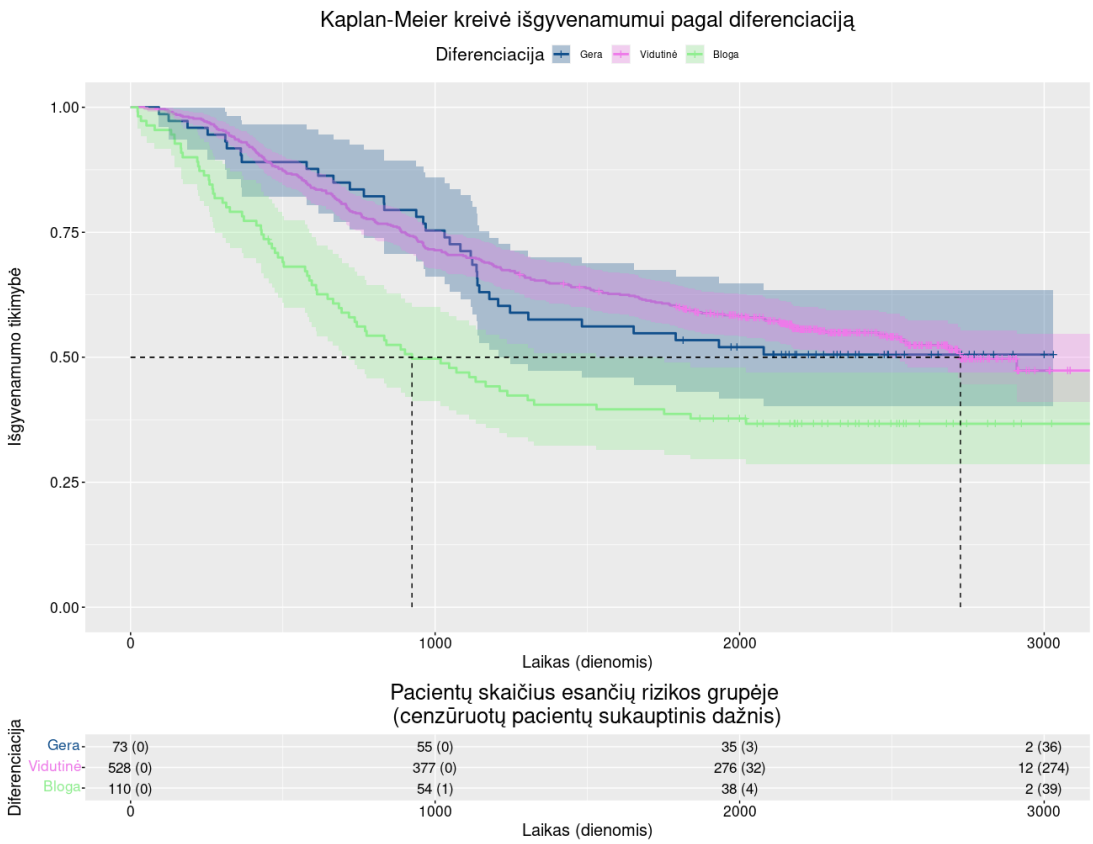
6 pav. Kaplan – Meier kreivė išgyvenamumui pagal perforaciją

Iš (6 pav.) matome, jog grupes tarpusavyje sunku atskirti, todėl pritaikome *twostage* testą. Gauta p reikšmė – 0,166, taigi grupės statistiškai reikšmingai nesiskiria. Iš 691 atvejo, kai pacientui nebuvo perforacijos, mirė 335, iš 20 atvejo kai perforacija buvo – mirė 12 pacientų. Tikimybė, jog pacientas išgyvens ilgiau nei 2683 dienas (apie 7,5 m.), kai perforacijos nėra, nukrenta iki 50 %, o kai ji yra – tikimybė išgyventi ilgiau nei 1904 dienas (apie 5 m.) nukrenta iki 50 %.



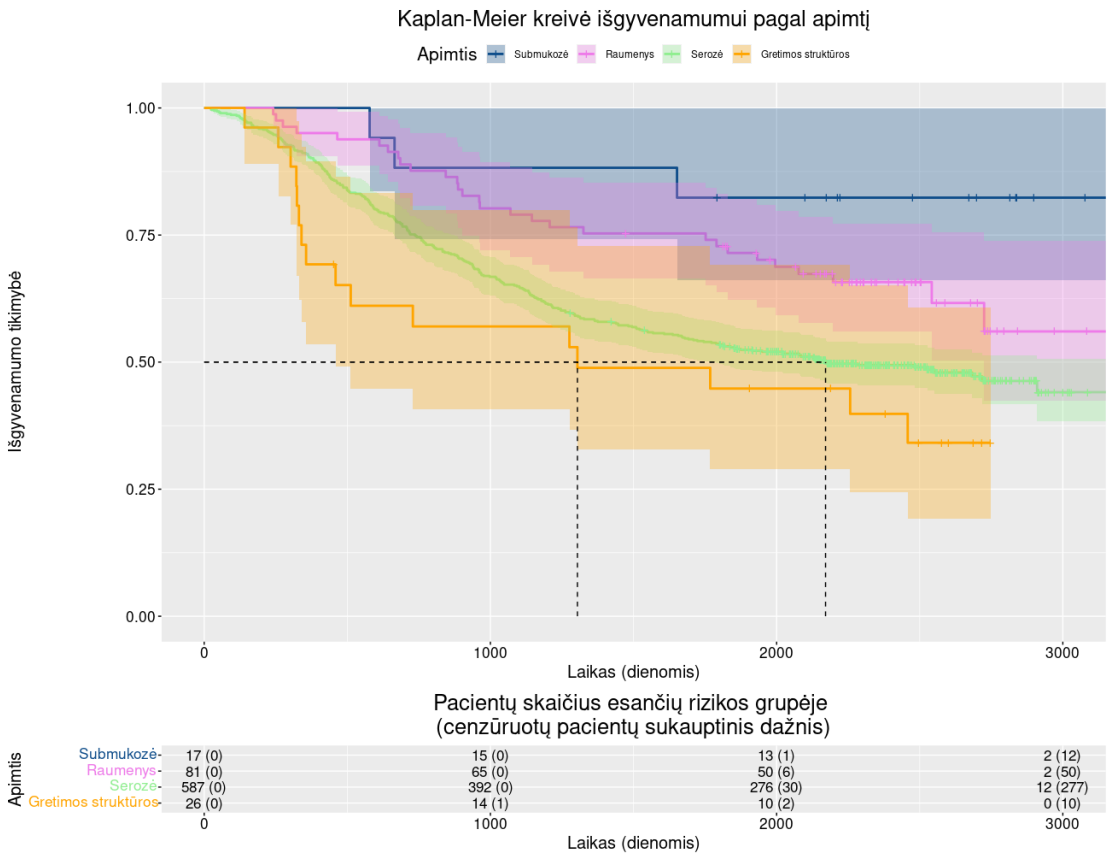
7 pav. Kaplan – Meier kreivė išgyvenamumui pagal prilipimą prie gretimų organų

Iš (7 pav.) matome, jog grupių kreivės nesikerta – dėl to taikysime log ranginį kriterijų grupių palyginimui. Gautas p reikšmė – 0,019, todėl galime teigti, jog yra statistiškai reikšmingų skirtumų tarp grupių. Iš 621 atvejo, kai pacientui nebuvo prilipimo prie gretimų organų, mirė 293, iš 90 atvejų kai prilipimas buvo – mirė 54 pacientai. Tikimybė, jog pacientas išgyvens ilgiau nei 2725 dienas (apie 7,5 m.), kai prilipimo prie gretimų organų nėra, nukrenta iki 50 %, o kai prilipimas yra – tikimybė išgyventi ilgiau nei 1364 dienas (apie 4 m.) nukrenta iki 50 %.



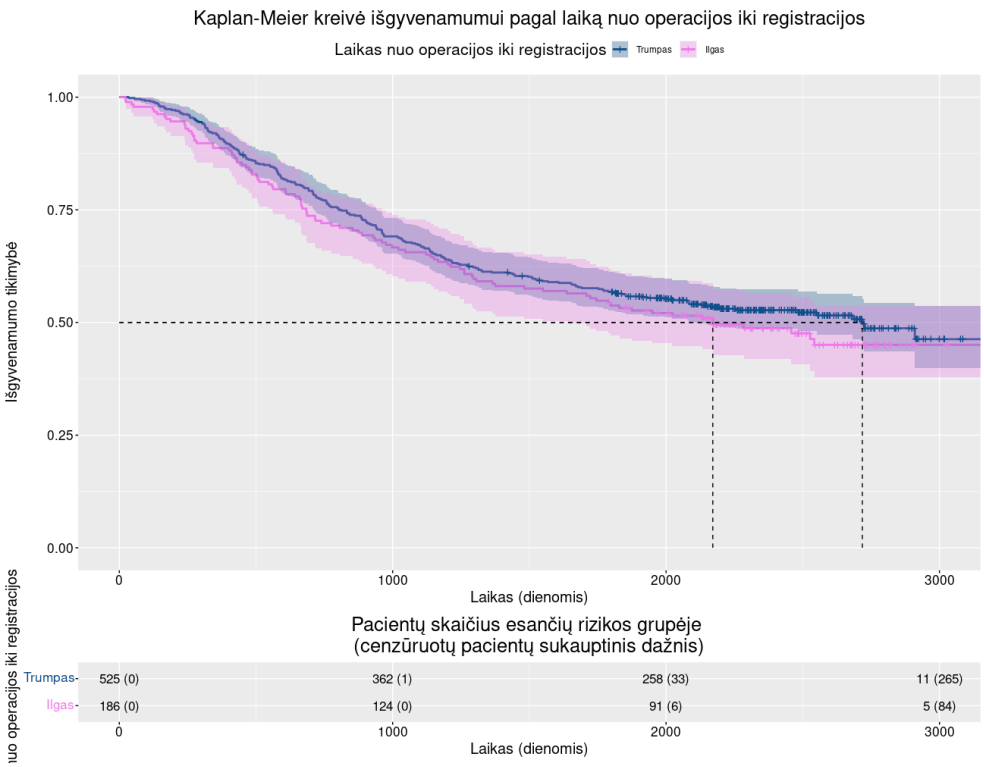
8 pav. Kaplan – Meier kreivė išgyvenamumui pagal diferenciaciją

Iš (8 pav.) matome, jog 36 iš 73 pacientų buvo įvykis, kai naviko diferenciacija buvo gera, 242 iš 528 pacientų buvo įvykis, kai diferenciacija buvo vidutinė, o 69 iš 110 pacientų buvo įvykis, kai diferenciacija buvo bloga. Tikimybė, jog pacientas išgyvens ilgiau nei 2725 dienas (apie 7,5 m.), kai diferenciacija yra vidutinė, nukrenta iki 50 %, o diferenciacija yra bloga – tikimybė išgyventi ilgiau nei 924 dienas (apie 2,5 m.) nukrenta iki 50 %.



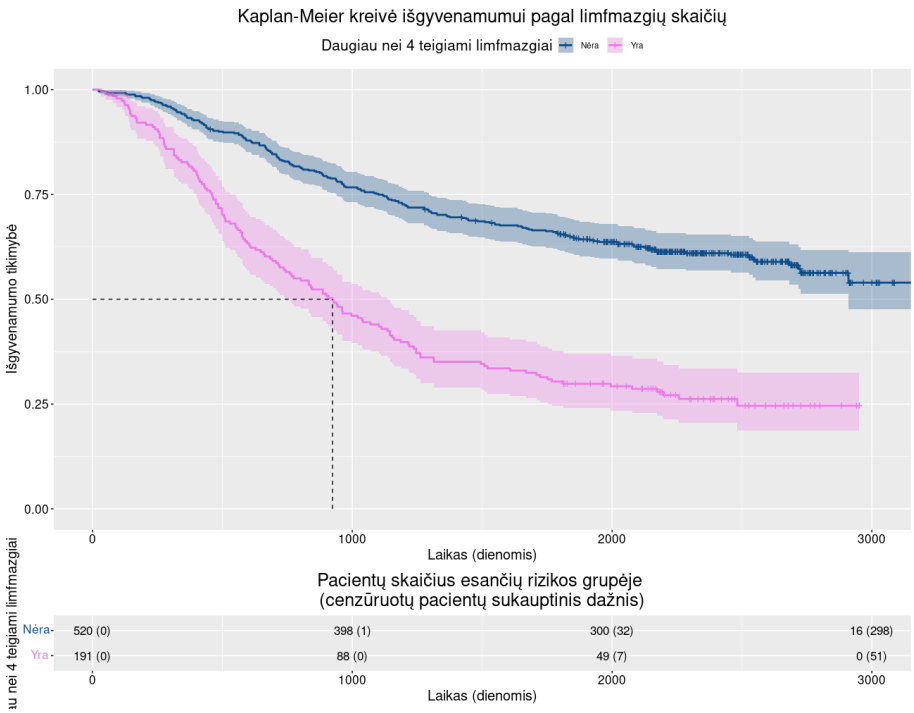
9 pav. Kaplan – Meier kreivė išgyvenamumui pagal apimtį

Iš (9 pav.) matome, jog grupės kertasi, todėl grupės tarpusavyje yra nepalyginamos. 3 iš 17 pacientų buvo įvykis, kai vėžys buvo išplitęs pogleivinėje dalyje, 29 iš 81 pacientų buvo įvykis, kai išplitimas buvo raumenyse, 299 iš 587 pacientų buvo įvykis, kai vėžys buvo išplitęs serozės dalyje, o 16 iš 26 pacientų buvo įvykis, kai išplitimas buvo gretimose struktūrose. Tikimybė, jog pacientas išgyvens ilgiau nei 2171 dienas (apie 6 m.), kai apimtis - serozė, nukrenta iki 50 %, o kai apimtis gretimos struktūros – tikimybė išgyventi ilgiau nei 1304 dienas (apie 3,5 m.) nukrenta iki 50 %.



10 pav. Kaplan – Meier kreivė išgyvenamumui pagal laiką nuo operacijos iki registracijos

Iš (10 pav.) matome, jog grupių kreivės nesikerta – dėl to grupių palyginimui bus naudojamas log ranginis kriterijus. Gavome p reikšmę – 0,33, todėl statistiškai reikšmingų skirtumų tarp grupių nėra. Iš 525 atvejų, kai laikas nuo operacijos iki registracijos buvo trumpas, mirė 250, iš 186 atvejų kai laikas buvo ilgas – mirė 97 pacientai. Tikimybė, jog pacientas išgyvens ilgiau nei 2718 dienas (apie 7,5 m.), kai laikos nuo operacijos iki registracijos buvo trumpas, nukrenta iki 50 %, o kai laikas ilgas – tikimybė išgyventi ilgiau nei 2171 dieną (apie 6 m.) nukrenta iki 50 %.



11 pav. Kaplan – Meier kreivė išgyvenamumui pagal limfmazgių skaičių

Iš (11 pav.) matome, jog grupių kreivės nesikerta – dėl to grupių palyginimui bus naudojamas log ranginis kriterijus. Gavome p reikšmę < 0,001, todėl grupės statistiškai reikšmingai skiriasi. Iš 520 atvejų, kai pas pacientą nebuvo daugiau nei 4 teigiamų limfmazgių, mirė 207, iš 191 atvejų kai pacientas turėjo daugiau nei 4 teigiamus limfmazgius – mirė 140 pacientų. Tikimybė, jog pacientas išgyvens ilgiau nei 924 dienas (apie 2,5 m.), kai limfmazgių teigiamų daugiau nei 4, nukrenta iki 50 %.

2.3 Parametrinio modelio prielaidos – skirstinio parinkimas ir multikolinearumas

Iš pat pradžių pasinaudoję funkcijomis *flexsurvreg* ir *survreg* pasižiūrime AIC – Akaikės informacinio koeficiento reikšmę – labiausiai tinkantys skirstiniai turimiems duomenims turės mažiausią AIC koeficiento reikšmę. Tai darant naudojame modelį su visomis kovariantėmis. Tikrinsime apibendrintąjį gamą, Veibulo, log – normalųjį, eksponentinį, log – logistinį, logistinį, normalųjį skirstinius. Apibendrintąjį gamą naudosime, kad galėtumėme patikrinti tikėtinumo santykiaus kriterijumi, kuris skirstinys: Veibulo, log – normalusis ar eksponentinis geriau tinka duomenims.

```
#Modelių sukurimas
fit.gamma<-flexsurvreg(Surv(time = colon_pilnas_death$time,
                            event = colon_pilnas_death$status) ~ .-id,
                       data=colon_pilnas_death, dist="gengamma")
fit.veibul<-flexsurvreg(Surv(time = colon_pilnas_death$time,
                            event = colon_pilnas_death$status) ~ .-id,
                       data=colon_pilnas_death, dist="weibull")
fit.lognorm<-flexsurvreg(Surv(time = colon_pilnas_death$time,
                             event = colon_pilnas_death$status) ~ .-id,
                       data=colon_pilnas_death, dist="lnorm")
fit.exp<-flexsurvreg(Surv(time = colon_pilnas_death$time,
                              event = colon_pilnas_death$status) ~ .-id,
                       data=colon_pilnas_death, dist="exp")
fit.ll<-flexsurvreg(Surv(time = colon_pilnas_death$time,
                       event = colon_pilnas_death$status) ~ .-id,
                  data=colon_pilnas_death, dist="llogis")
fit.norm<-survreg(Surv(time = colon_pilnas_death$time,
                       event = colon_pilnas_death$status) ~ .-id,
                  data=colon_pilnas_death, dist="gaussian")
fit.logistic<-survreg(Surv(time = colon_pilnas_death$time,
                       event = colon_pilnas_death$status) ~ .-id,
                  data=colon_pilnas_death, dist="logistic")
#AIC
> fit.veibul$AIC
[1] 7750.024
> fit.lognorm$AIC
[1] 7695.359
> fit.exp$AIC
[1] 7750.658
> fit.ll$AIC
[1] 7710.715
> AIC(fit.norm)
[1] 8078.35
> AIC(fit.logistic)
[1] 8116.123
```

Iš gautų rezultatų matome, jog geriausiai turimiems duomenimis turėtų tikti log – logistinis (AIC – 7710,715), log – normalusis (AIC – 7695,359), Veibulo (AIC – 7750,024), tačiau tai nereiškia, kad šie skirstiniai išvis tinka duomenims.

Rezultatų pagrindimui pasinaudosime tikėtinumo santykio kriterijumi, kuris tinka, kai nulinė hipotezė yra siauresnė, o alternatyva platesnė. Atmesime nulinę hipotezę, kai tikėtinumo santykio reikšmė bus didesnė už kritinę reikšmę. Tikėtinumo santykio reikšmė gaunama:

$$X_L = 2 \times \big(LogLik(alternatyvos\ skirstinys) - LogLik(nulinės\ hipotezės\ skirstinys)\big),$$

o kritinė reikšmė gaunama:

$$\chi^2_\alpha(q), \qquad kur\ q = skirstinio\ parametrų\ kiekis\ alternatyvos - skirstinio\ parametrų\ kiekis\ nulinės\ hipotezės.$$

Nulinė hipotezė bus atmesta, kai $X_L > \chi^2_\alpha$.

Tikriname:

$$\begin{cases} H_0: Veibulo, t.y.\ shape = 1 \\ H_1: Apibendrintas\ gama, t.y. shape \neq 1 \end{cases}$$

```
> gamma_loglik<-(fit.gamma$loglik)
> veibul_loglik_0<-(fit.veibul$loglik)
>
> stat<-2*gamma_loglik - 2* veibul_loglik_0
> stat
[1] 61.63435
> critical_value<-qchisq(0.95, df = 1)# nes k0=2, k1=3
> critical_value
[1] 3.841459
> rejected<-(stat > critical_value)
> rejected
[1] TRUE
```

Priimame alternatyvą – *shape* parametras nėra lygus 1.

Tikriname:

$$\begin{cases} H_0: \log - normalusis\ , t.y.\ shape = 0 \\ H_1: Apibendrintas\ gama, t.y. shape \neq 0 \end{cases}$$

```
> lognorm_loglik_0<-(fit.lognorm$loglik)
>
> stat_lognorm<-2*gamma_loglik - 2* lognorm_loglik_0
```

```
> stat_lognorm
[1] 6.969401
> critical_value_lognorm<-qchisq(0.95, df = 1)# nes k0=2, k1=3
> critical_value_lognorm
[1] 3.841459
> rejected_lognorm<-(stat_lognorm > critical_value_lognorm)
> rejected_lognorm
[1] TRUE
```

Priimame alternatyvą – *shape* parametras nėra lygus 0.

Tikriname:

$$\begin{cases} H_0: eksponentinis, t.y.\ shape = scale = 1 \\ H_1: Apibendrintas\ gama, t.y. kažkuri\ lygybė\ nėra\ teisinga \end{cases}$$

```
> exp_loglik_0<-(fit.exp$loglik)
>
> stat_exp<-2*gamma_loglik - 2* exp_loglik_0
> stat_exp
[1] 64.26812
> critical_value_exp<-qchisq(0.95, df = 2)# nes k0=1, k1=3
> critical_value_exp
[1] 5.991465
> rejected_exp<-(stat_exp > critical_value_exp)
> rejected_exp
[1] TRUE
```

Priimame alternatyvą – *shape* parametras nėra lygus *scale* ir jie kartu nėra lygūs 1.

Tikriname:

$$\begin{cases} H_0: eksponentinis, t.y.\ scale = 1 \\ H_1: Veibulas, t.y.\ scale \neq 1 \end{cases}$$

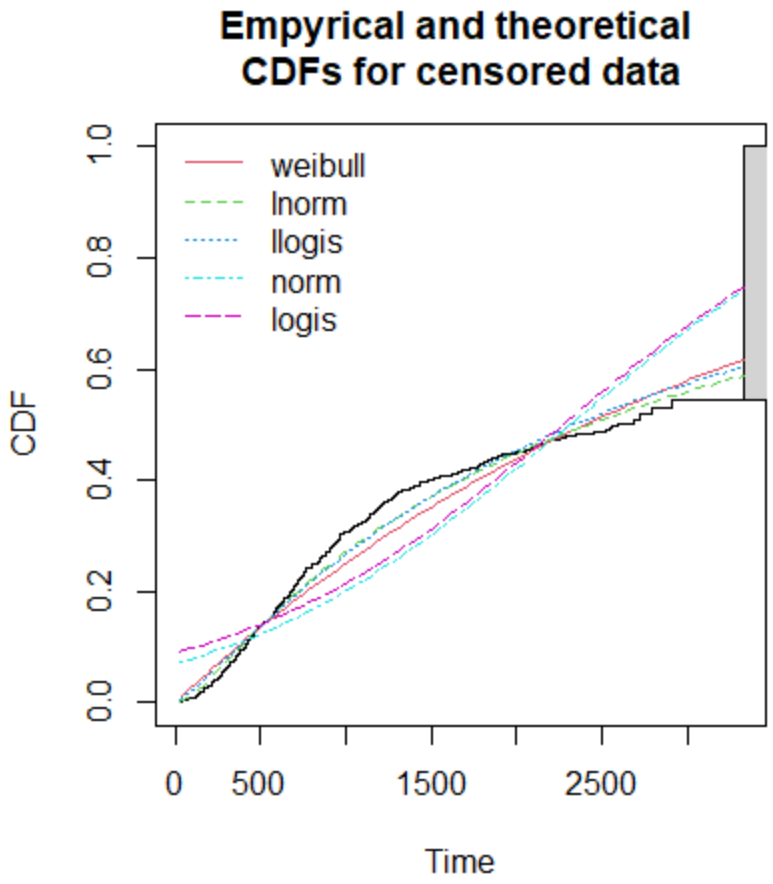
```
> stat_exp_veibulo<-2*veibul_loglik_0 - 2* exp_loglik_0
> stat_exp_veibulo
[1] 2.633764
> critical_value_exp_veibulo<-qchisq(0.95, df = 1)# nes k0=1, k1=2
> critical_value_exp_veibulo
[1] 3.841459
> rejected_exp_veibulo<-(stat_exp_veibulo > critical_value_exp_veibulo)
> rejected_exp_veibulo
[1] FALSE
```

Nulinė hipotezė nėra atmetama.

Iš šių hipotezių tikrinimo geriausiai tinkantis skirstinys duomenims eksponentinis. Taip pat bandėme pritaikyti gamą skirstinį, tačiau neradome pradinių parametrų artinių, kad sukonverguotų, todėl į tyrimą šio skirstinio neįtraukėme, o iš ankstesnių rezultatų galime pastebėti, jog gavome, kad Veibulo skirstinys tinka labiau nei eksponentinis.

Pareisime prie grafinio skirstinių tinkamumo patikrinimo. Pasinaudoję *cdfcompcens* funkcija, kuri nubraižo empirinę sukauptinę pasiskirstymo funkciją cenzūruotiems stebėjimams ir teorines pagal priskirtą skirstinį. Neįtraukėme eksponentinio skirstinio, nes naudojant šią funkciją nerado tinkamų pradinių parametrų reikšmių. Taip pat šiai funkcijai reikėjo specialaus duomenų formato – lentelės, kuriame vienas stulpelis buvo žymimas *right* (įvyko laikas) ir *left* (įvyko laikas, jei įvykis mirtis, priešingu atveju NA). Šis duomenų žymėjimas tinkamas, kai turime duomenis cenzūruotus iš dešinės.

```
fveibul<-fitdistcens(data_tikrinimui, "weibull")

fll<-fitdistcens(data_tikrinimui, "llogis")

fln<-fitdistcens(data_tikrinimui, "lnorm")

fnorm<-fitdistcens(data_tikrinimui, "norm")

fl<-fitdistcens(data_tikrinimui, "logis")

cdfcompcens(list(fveibul, fln, fll, fnorm, fl),
xlim=range(colon_pilnas_death$time), xlegend = "topleft", xlab = "Time",

            main = "Empyrical and theoretical CDFs for censored data")
```



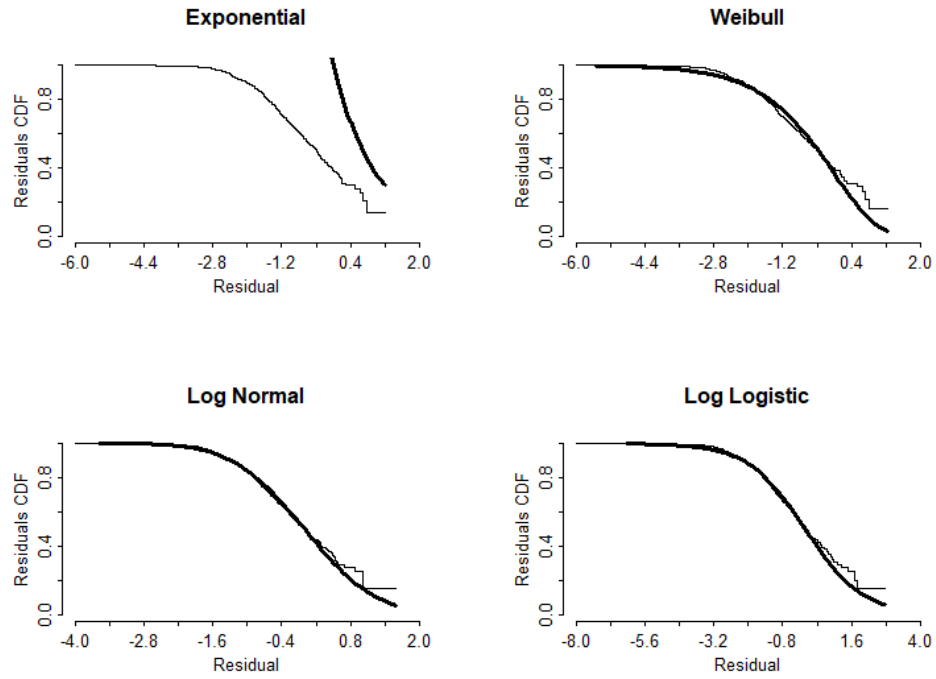
12 pav. Empirinė sukauptinė pasiskirstymo funkcija

Iš (12 pav.) matome, jog geriausiai duomenims tinka Veibulo, log – logistinis bei log – normalusis skirstiniai.

Norėdami grafiškai patikrinti kitu būdu skirstinių tinkamumą, rasime standartizuotas liekanas ir jų išgyvenamumo kreives lyginsime su K – M įverčiu duomenims. Imsime liekanas, nes imant skirtingus liekanų skirstinius gauname skirtingus AFT modelius. Šiuo atveju lyginsime geriausius išrinktus skirstinius – eksponentinį, Veibulo, log – normalųjį, log – logistinį.

```
psmE <- psm(Surv(time = colon_pilnas_death$time,event =
colon_pilnas_death$status)~sex+age+rx+obstruct+perfor+adhere+nodes+differ+ext
ent+surg+node4,dist="exponential",data=colon_pilnas_death)

residE <- residuals(psmE)

psmW <- psm(Surv(time = colon_pilnas_death$time, event =
colon_pilnas_death$status)
~sex+age+rx+obstruct+perfor+adhere+nodes+differ+extent+surg+node4,dist="weibu
ll",data=colon_pilnas_death)
```

```
residW <- residuals(psmW)

psmLN <- psm(Surv(time = colon_pilnas_death$time, event =
colon_pilnas_death$status) ~
sex+age+rx+obstruct+perfor+adhere+nodes+differ+extent+surg+node4,dist="lognor
mal",data=colon_pilnas_death)

residLN <- residuals(psmLN)

psmLL <- psm(Surv(time = colon_pilnas_death$time, event =
colon_pilnas_death$status) ~
sex+age+rx+obstruct+perfor+adhere+nodes+differ+extent+surg+node4,dist="loglog
istic",data=colon_pilnas_death)

residLL <- residuals(psmLL)

par(mfrow=c(2,2))

survplot(residE,main="Exponential",ylab="Complement of residual CDF")

survplot(residW,main="Weibull",ylab="Complement of residual CDF")

survplot(residLN,main="Log Normal",ylab="Complement of residual CDF")

survplot(residLL,main="Log Logistic",ylab="Complement of residual CDF")
```



13 pav. Standartizuotos liekanos ir K-M įverčiai

(13 pav.) storesnė linija žymi teorinį pasiskirstymą, o plonesnė K – M liekanų įvertį iš duomenų. Galime matyti, jog Veibulo, log – normaliojo, log – logistinio skirstiniai gerai tinka nagrinėjamiems duomenims, todėl toliau dirbsime su šiais 3 skristiniais.

Šiems trims modeliams patikrinome, ar nėra multikolinearumo problemos kiekybiniams kintamiesiems (amžiui ir vėžiniams limfmazgiams) – gavome, jog nėra. Tikrinimas vyko – ėmėme modelį su visomis kovariantėmis ir modelį su viena kiekybine kovariante ir žiūrėjome, ar koeficientų ženklai sutapo. Su abejomis kovariantėmis visuose skirstiniuose ženklai sutapo.

2.4 Modelio konstravimas

Toliau atliekame pažingsninę regresiją: rinksime reikšmingas kovariantes „ranka“, remdamiesi AIC rodikliu bei palyginsime gautus rezultatus su „R“ programos funkcija *step()*. Naudosime 3 skirstinius: loglogistinį, lognormalųjį ir Veibulo.

Pradėsime nuo loglogistinio:

Apsirašome modelį su visomis kovariantėmis ir pažiūrime jo AIC. Pirma, nereikšmingas kovariantes šalinsime „ranka“.

```
Call:
survreg(formula = Surv(time, status) ~ rx + sex + age + obstruct +
    perfor + adhere + nodes + differ + extent + surg + node4,
    data = train, dist = "loglogistic")
               Value Std. Error     z       p
(Intercept)  9.47247    0.58838 16.10 < 2e-16
rxLev+5FU    0.21640    0.13873  1.56   0.119
rxObs       -0.08714    0.13038 -0.67   0.504
sex1         0.03135    0.11034  0.28   0.776
age         -0.01056    0.00469 -2.25   0.024
obstruct1   -0.31903    0.14271 -2.24   0.025
perfor1      0.06856    0.31260  0.22   0.826
adhere1     -0.27662    0.16293 -1.70   0.090
nodes       -0.04219    0.02049 -2.06   0.040
differ2      0.25566    0.18210  1.40   0.160
differ3     -0.32903    0.22084 -1.49   0.136
extent2     -0.35555    0.51516 -0.69   0.490
extent3     -0.87160    0.48994 -1.78   0.075
extent4     -1.20592    0.56745 -2.13   0.034
surg1       -0.18505    0.12391 -1.49   0.135
node41      -0.78300    0.17956 -4.36 1.3e-05
Log(scale)  -0.28023    0.04634 -6.05 1.5e-09

Scale= 0.756

Log logistic distribution
Loglik(model)= -3085.8   Loglik(intercept only)= -3156.9
      Chisq= 142.11 on 15 degrees of freedom, p= 8.8e-23
Number of Newton-Raphson Iterations: 4
n= 711
```

AIC = 6205.679

Iš p reikšmių matome, jog nereikšmingiausia kovariantė yra perforacija (*angl. perfor*), ją šaliname.

```
Call:
survreg(formula = Surv(time, status) ~ rx + sex + age + obstruct +
    adhere + nodes + differ + extent + surg + node4, data = train,
    dist = "loglogistic")
              Value Std. Error     z       p
(Intercept)  9.47659    0.58824 16.11 < 2e-16
rxLev+5FU    0.21626    0.13876  1.56   0.119
rxObs       -0.08620    0.13033 -0.66   0.508
sex1         0.03079    0.11033  0.28   0.780
age         -0.01059    0.00469 -2.26   0.024
obstruct1   -0.31579    0.14201 -2.22   0.026
adhere1     -0.26960    0.15980 -1.69   0.092
nodes       -0.04216    0.02049 -2.06   0.040
differ2      0.25324    0.18183  1.39   0.164
differ3     -0.33263    0.22030 -1.51   0.131
extent2     -0.35525    0.51521 -0.69   0.490
extent3     -0.87018    0.48997 -1.78   0.076
extent4     -1.20530    0.56750 -2.12   0.034
surg1       -0.18553    0.12392 -1.50   0.134
node41      -0.78370    0.17956 -4.36 1.3e-05
Log(scale)  -0.28007    0.04633 -6.04 1.5e-09

Scale= 0.756

Log logistic distribution
Loglik(model)= -3085.9   Loglik(intercept only)= -3156.9
      Chisq= 142.06 on 14 degrees of freedom, p= 2.8e-23
Number of Newton-Raphson Iterations: 4
n= 711
```

AIC = 6203.728

Šaliname kovariantę lytis (*angl. sex*).

```
Call:
survreg(formula = Surv(time, status) ~ rx + age + obstruct +
    adhere + nodes + differ + extent + surg + node4, data = train,
    dist = "loglogistic")
              Value Std. Error     z       p
(Intercept)  9.48397    0.58770 16.14 < 2e-16
rxLev+5FU    0.21357    0.13847  1.54   0.123
rxObs       -0.08606    0.13035 -0.66   0.509
age         -0.01059    0.00469 -2.26   0.024
obstruct1   -0.31692    0.14197 -2.23   0.026
adhere1     -0.27190    0.15962 -1.70   0.088
nodes       -0.04204    0.02051 -2.05   0.040
```

```
differ2      0.25590    0.18151  1.41   0.159
differ3     -0.32888    0.21985 -1.50   0.135
extent2     -0.34747    0.51428 -0.68   0.499
extent3     -0.86339    0.48918 -1.76   0.078
extent4     -1.19569    0.56612 -2.11   0.035
surg1       -0.18449    0.12391 -1.49   0.137
node41      -0.78673    0.17932 -4.39 1.1e-05
Log(scale)  -0.27991    0.04633 -6.04 1.5e-09

Scale= 0.756

Log logistic distribution
Loglik(model)= -3085.9   Loglik(intercept only)= -3156.9
      Chisq= 141.98 on 13 degrees of freedom, p= 8.4e-24
Number of Newton-Raphson Iterations: 4
n= 711
```
AIC = 6201.805

Šaliname kovariantę gydymo tipas (*angl. rx*). Pašalinę šią kovariantę, gauname, jog AIC padidėja – be šios kovariantės **AIC = 6202.818.** Todėl kovariantę *rx* paliekame modelyje. Nereikšmingos kovariantės: perforacija ir lytis.

```
Call:
survreg(formula = Surv(time, status) ~ rx + age + obstruct +
    adhere + nodes + differ + extent + surg + node4, data = train,
    dist = "loglogistic")

Coefficients:
(Intercept)    rxLev+5FU         rxObs           age    obstruct1      adhere1
nodes
 9.48397307   0.21357364  -0.08605657  -0.01058565  -0.31692336  -0.27189647  -
0.04203919
    differ2      differ3      extent2      extent3      extent4         surg1
node41
 0.25590171  -0.32887939  -0.34746695  -0.86339085  -1.19569211  -0.18449378  -
0.78672546

Scale= 0.7558512

Loglik(model)= -3085.9   Loglik(intercept only)= -3156.9
      Chisq= 141.98 on 13 degrees of freedom, p= <2e-16
n= 711
```

Toliau patikriname kokį galutinį modelį pateikia funkcija `step(„forward“ ir „backward“)`

```
Step(direction = “forward”)
Start:  AIC=6205.68
```

```
Surv(time, status) ~ rx + sex + age + obstruct + perfor + adhere +
    nodes + differ + extent + surg + node4

Call:
survreg(formula = Surv(time, status) ~ rx + sex + age + obstruct +
    perfor + adhere + nodes + differ + extent + surg + node4,
    data = train, dist = "loglogistic")

Coefficients:
(Intercept)   rxLev+5FU       rxObs        sex1         age   obstruct1
 9.47247459  0.21639887 -0.08714143  0.03135067 -0.01055964 -0.31903278
perfor1
0.06855653
    adhere1       nodes     differ2     differ3     extent2     extent3
-0.27661566 -0.04219212  0.25566493 -0.32902602 -0.35555475 -0.87160354
extent4
-1.20591562
      surg1      node41
-0.18505189 -0.78300325

Scale= 0.755613

Loglik(model)= -3085.8   Loglik(intercept only)= -3156.9
      Chisq= 142.11 on 15 degrees of freedom, p= <2e-16
n= 711
```

Gauname, jog visos kovariantės yra reikšmingos.

Step(direction = "backward")

```
Call:
survreg(formula = Surv(time, status) ~ rx + age + obstruct +
    adhere + nodes + differ + extent + surg + node4, data = train,
    dist = "loglogistic")

Coefficients:
(Intercept)   rxLev+5FU       rxObs         age   obstruct1     adhere1
 9.48397307  0.21357364 -0.08605657 -0.01058565 -0.31692336 -0.27189647
nodes
-0.04203919
    differ2     differ3     extent2     extent3     extent4       surg1
 0.25590171 -0.32887939 -0.34746695 -0.86339085 -1.19569211 -0.18449378
node41
-0.78672546

Scale= 0.7558512

Loglik(model)= -3085.9   Loglik(intercept only)= -3156.9
      Chisq= 141.98 on 13 degrees of freedom, p= <2e-16
n= 711
```

Naudojant *step()* funkciją su nurodymu atlikti atbulinę pažingsninę regresiją, gauname tokias pat reikšmingas kovariantes kaip ir jas renkant „ranka“.

Galutinis modelis su loglogistiniu skirstiniu:

```
Call:
survreg(formula = Surv(time, status) ~ rx + age + obstruct +
    adhere + nodes + differ + extent + surg + node4, data = train,
    dist = "loglogistic")

Coefficients:
(Intercept)   rxLev+5FU       rxObs         age   obstruct1     adhere1
 9.48397307  0.21357364 -0.08605657 -0.01058565 -0.31692336 -0.27189647
Nodes
-0.04203919
    differ2     differ3     extent2     extent3     extent4       surg1
 0.25590171 -0.32887939 -0.34746695 -0.86339085 -1.19569211 -0.18449378
node41
-0.78672546

Scale= 0.7558512

Loglik(model)= -3085.9   Loglik(intercept only)= -3156.9
      Chisq= 141.98 on 13 degrees of freedom, p= <2e-16
n= 711
```

$$S_{\log-logistinis}(t;\eta,\nu) = 1 - \left(\frac{1}{1+\left(\frac{t}{\exp(9{,}484)\times a}\right)^{-\frac{-1}{0{,}756}}}\right)$$

$$= 1 - \left(\frac{1}{1+\left(\frac{t}{13147{,}67\times a}\right)^{-1{,}323}}\right), kur\ a$$

$$= \exp(Levamizolas5 \times 0{,}214 - koservavimas \times 0{,}086 - amžius \times 0{,}011$$
$$- obstrukcija(yra) \times 0{,}317 - prilipimas(yra) \times 0{,}272$$
$$- vėžiniai\ limfmazgiai \times 0{,}042 + diferenciacija(vidutinė) \times 0{,}256$$
$$- diferenciacija(bloga) \times 0{,}329$$
$$- vietinio\ išplitimo\ mastas(raumenys) \times 0{,}347$$
$$- vietinio\ išplitimo\ mastas(serozė) \times 0{,}863$$
$$- vietinio\ išplitimo\ mastas(gretimos\ struktūros) \times 1{,}196$$
$$- laikas\ nuo\ operacijos\ iki\ registracijos(ilgas) \times 0{,}184$$
$$- teigiami\ limfmazgiai(daugiau\ nei\ 4) \times 0{,}787)\,, čia\ \eta = scale$$

$$= \exp(intercept) = mastelio\ parametras = 13147{,}67, \nu = shape = \frac{1}{scale}$$

$$= 1{,}323 = formos\ parametras$$

Toliau ieškosime reikšmingų kovariančių modeliui **su lognormaliuoju skirstiniu**. Kaip ir prieš tai, pirma, kovariantes atrinksime „rankiniu“ būdu, o vėliau pritaikysime funkciją *step()*.

Apsirašome modelį su visomis kovariantėmis ir pažiūrime jo AIC.

```
Call:
survreg(formula = Surv(time, status) ~ rx + sex + age + obstruct +
    perfor + adhere + nodes + differ + extent + surg + node4,
    data = train, dist = "lognormal")
               Value Std. Error     z       p
(Intercept)  9.65165    0.57150 16.89 < 2e-16
rxLev+5FU    0.19554    0.13850  1.41  0.1580
rxObs       -0.05932    0.13419 -0.44  0.6584
sex1         0.02106    0.11202  0.19  0.8509
age         -0.01230    0.00471 -2.61  0.0090
obstruct1   -0.40877    0.14148 -2.89  0.0039
perfor1      0.10653    0.32983  0.32  0.7467
adhere1     -0.25232    0.16617 -1.52  0.1289
nodes       -0.04552    0.02144 -2.12  0.0338
differ2      0.25850    0.18616  1.39  0.1650
differ3     -0.37667    0.22236 -1.69  0.0903
extent2     -0.33830    0.49455 -0.68  0.4939
extent3     -0.88315    0.46678 -1.89  0.0585
extent4     -1.24719    0.54448 -2.29  0.0220
surg1       -0.22345    0.12575 -1.78  0.0756
node41      -0.75969    0.18252 -4.16 3.2e-05
Log(scale)   0.27016    0.04207  6.42 1.4e-10

Scale= 1.31

Log Normal distribution
Loglik(model)= -3079.8   Loglik(intercept only)= -3149.4
      Chisq= 139.25 on 15 degrees of freedom, p= 3.2e-22
Number of Newton-Raphson Iterations: 4
n= 711
AIC = 6193.553
```

Šaliname kovariantę lytis (*angl. sex*).

```
Call:
survreg(formula = Surv(time, status) ~ rx + age + obstruct +
    perfor + adhere + nodes + differ + extent + surg + node4,
    data = train, dist = "lognormal")
               Value Std. Error     z       p
(Intercept)  9.66003    0.57004 16.95 < 2e-16
rxLev+5FU    0.19289    0.13778  1.40  0.1615
rxObs       -0.05965    0.13419 -0.44  0.6566
age         -0.01230    0.00471 -2.61  0.0090
obstruct1   -0.40964    0.14141 -2.90  0.0038
perfor1      0.10571    0.32980  0.32  0.7486
```

```
adhere1     -0.25375    0.16600 -1.53  0.1264
nodes       -0.04545    0.02144 -2.12  0.0340
differ2      0.26019    0.18595  1.40  0.1617
differ3     -0.37415    0.22196 -1.69  0.0919
extent2     -0.33578    0.49456 -0.68  0.4972
extent3     -0.88111    0.46687 -1.89  0.0591
extent4     -1.24337    0.54429 -2.28  0.0224
surg1       -0.22303    0.12573 -1.77  0.0761
node41      -0.76123    0.18235 -4.17 3.0e-05
Log(scale)   0.27018    0.04207  6.42 1.3e-10

Scale= 1.31

Log Normal distribution
Loglik(model)= -3079.8   Loglik(intercept only)= -3149.4
      Chisq= 139.22 on 14 degrees of freedom, p= 1e-22
Number of Newton-Raphson Iterations: 4
n= 711
AIC = 6191.588
```

Šaliname kovariantę perforacija (*angl. perfor*).

```
Call:
survreg(formula = Surv(time, status) ~ rx + age + obstruct +
    adhere + nodes + differ + extent + surg + node4, data = train,
    dist = "lognormal")
              Value Std. Error     z       p
(Intercept)  9.66374    0.57015 16.95 < 2e-16
rxLev+5FU    0.19300    0.13783  1.40  0.1614
rxObs       -0.05847    0.13419 -0.44  0.6630
age         -0.01233    0.00471 -2.62  0.0089
obstruct1   -0.40587    0.14097 -2.88  0.0040
adhere1     -0.24438    0.16346 -1.50  0.1349
nodes       -0.04534    0.02144 -2.11  0.0345
differ2      0.25702    0.18579  1.38  0.1665
differ3     -0.37814    0.22172 -1.71  0.0881
extent2     -0.33485    0.49470 -0.68  0.4985
extent3     -0.87918    0.46695 -1.88  0.0597
extent4     -1.23807    0.54419 -2.28  0.0229
surg1       -0.22276    0.12577 -1.77  0.0765
node41      -0.76298    0.18234 -4.18 2.9e-05
Log(scale)   0.27051    0.04207  6.43 1.3e-10

Scale= 1.31

Log Normal distribution
Loglik(model)= -3079.8   Loglik(intercept only)= -3149.4
      Chisq= 139.11 on 13 degrees of freedom, p= 3.2e-23
Number of Newton-Raphson Iterations: 4
n= 711
AIC = 6189.691
```

Šaliname kovariantę gydymo tipas (*angl. rx*).

```
Call:
survreg(formula = Surv(time, status) ~ age + obstruct + adhere +
    nodes + differ + extent + surg + node4, data = train, dist = "lognormal")
              Value Std. Error     z       p
(Intercept)  9.74300    0.56090 17.37 < 2e-16
age         -0.01246    0.00471 -2.64  0.0082
obstruct1   -0.41528    0.14100 -2.95  0.0032
adhere1     -0.23851    0.16355 -1.46  0.1448
nodes       -0.04783    0.02142 -2.23  0.0256
differ2      0.26875    0.18549  1.45  0.1474
differ3     -0.35720    0.22126 -1.61  0.1064
extent2     -0.36648    0.49384 -0.74  0.4580
extent3     -0.91076    0.46597 -1.95  0.0506
extent4     -1.27314    0.54403 -2.34  0.0193
surg1       -0.23008    0.12575 -1.83  0.0673
node41      -0.75510    0.18238 -4.14 3.5e-05
Log(scale)   0.27200    0.04207  6.46 1.0e-10

Scale= 1.31

Log Normal distribution
Loglik(model)= -3081.7   Loglik(intercept only)= -3149.4
      Chisq= 135.49 on 11 degrees of freedom, p= 1.3e-23
Number of Newton-Raphson Iterations: 4
n= 711
AIC = 6189.311
```

Šaliname kovariantę apimtis (*angl. extent*). Pašalinę šią kovariantę, gauname, jog AIC padidėja – be šios kovariantės **AIC = 6197.595.** Todėl kovariantę *extent* paliekame modelyje. Nereikšmingos kovariantės: lytis, perforacija ir gydymo tipas.

```
Call:
survreg(formula = Surv(time, status) ~ age + obstruct + adhere +
    nodes + differ + extent + surg + node4, data = train, dist = "lognormal")

Coefficients:
(Intercept)         age   obstruct1     adhere1       nodes     differ2
 9.74300270 -0.01245675 -0.41528087 -0.23851482 -0.04783058  0.26875120
differ3
-0.35720094
    extent2     extent3     extent4       surg1      node41
-0.36648364 -0.91076006 -1.27314484 -0.23008207 -0.75510063

Scale= 1.31259

Loglik(model)= -3081.7   Loglik(intercept only)= -3149.4
      Chisq= 135.49 on 11 degrees of freedom, p= <2e-16
n= 711
```

Naudojame *step(direction = "forward")*

```
Start:  AIC=6193.55
Surv(time, status) ~ rx + sex + age + obstruct + perfor + adhere +
    nodes + differ + extent + surg + node4

Call:
survreg(formula = Surv(time, status) ~ rx + sex + age + obstruct +
    perfor + adhere + nodes + differ + extent + surg + node4,
    data = train, dist = "lognormal")

Coefficients:
(Intercept)   rxLev+5FU       rxObs        sex1         age   obstruct1
 9.65164910  0.19553815 -0.05932092  0.02105786 -0.01229961 -0.40876818
perfor1
0.10652883
    adhere1       nodes     differ2     differ3     extent2     extent3
-0.25231618 -0.04551563  0.25850248 -0.37666752 -0.33829893 -0.88315372
extent4
-1.24718618
      surg1      node41
-0.22345308 -0.75969306

Scale= 1.310173

Loglik(model)= -3079.8   Loglik(intercept only)= -3149.4
      Chisq= 139.25 on 15 degrees of freedom, p= <2e-16
n= 711
```

Paliekamos visos kovariantės kaip reikšmingos.

Step(direction = "backward")

```
Call:
survreg(formula = Surv(time, status) ~ age + obstruct + adhere +
    nodes + differ + extent + surg + node4, data = train, dist = "lognormal")

Coefficients:
(Intercept)         age   obstruct1     adhere1       nodes     differ2
 9.74300270 -0.01245675 -0.41528087 -0.23851482 -0.04783058  0.26875120
differ3
-0.35720094
    extent2     extent3     extent4       surg1      node41
-0.36648364 -0.91076006 -1.27314484 -0.23008207 -0.75510063

Scale= 1.31259

Loglik(model)= -3081.7   Loglik(intercept only)= -3149.4
      Chisq= 135.49 on 11 degrees of freedom, p= <2e-16
n= 711
```

Naudojant *step()* funkciją su nurodymu atlikti atbulinę pažingsninę regresiją, gauname tokias pat reikšmingas kovariantes kaip ir jas renkant „ranka“.

Galutinis modelis su lognormaliuoju skirstiniu:

```
Call:
survreg(formula = Surv(time, status) ~ age + obstruct + adhere +
    nodes + differ + extent + surg + node4, data = train, dist = "lognormal")

Coefficients:
(Intercept)         age   obstruct1     adhere1       nodes     differ2
 9.74300270 -0.01245675 -0.41528087 -0.23851482 -0.04783058  0.26875120
differ3
-0.35720094
    extent2     extent3     extent4       surg1      node41
-0.36648364 -0.91076006 -1.27314484 -0.23008207 -0.75510063

Scale= 1.31259

Loglik(model)= -3081.7   Loglik(intercept only)= -3149.4
      Chisq= 135.49 on 11 degrees of freedom, p= <2e-16
n= 711
```

$$S_{log-normalusis}(t;\ \mu,\sigma) = 1 - \ \Phi\left(\frac{ln(t)}{1{,}313 \times a}\right), kur\ a$$

$$= exp(-amžius \times 0{,}012 - obstrukcija(yra) \times 0{,}415$$
$$- prilipimas(yra) \times 0{,}239 - vėžiniai\ limfmazgiai \times 0{,}048$$
$$+ diferenciacija(vidutinė) \times 0{,}269 - diferenciacija(bloga) \times 0{,}357$$
$$- vietinio\ išplitimo\ mastas(raumenys) \times 0{,}366$$
$$- vietinio\ išplitimo\ mastas(serozė) \times 0{,}911$$
$$- vietinio\ išplitimo\ mastas(gretimos\ struktūros) \times 1{,}273$$
$$- laiaks\ nuo\ operacijos\ iki\ registracijos(ilgas) \times 0{,}230$$
$$- teigiami\ limfmazgiai(daugiau\ nei\ 4) \times 0{,}755), \text{čia}\ \sigma = scale$$
$$= mastelio\ parametras = 1{,}313.$$

Analogiškai viską darome **su Veibulo** skirstiniu:

```
Call:
survreg(formula = Surv(time, status) ~ rx + sex + age + obstruct +
    perfor + adhere + nodes + differ + extent + surg + node4,
    data = train, dist = "weibull")
              Value Std. Error     z       p
(Intercept)  9.80854    0.64032 15.32 < 2e-16
```

```
rxLev+5FU    0.28519    0.13218  2.16  0.0310
rxObs       -0.14031    0.12206 -1.15  0.2503
sex1         0.02493    0.10442  0.24  0.8113
age         -0.00933    0.00462 -2.02  0.0433
obstruct1   -0.25116    0.13170 -1.91  0.0565
perfor1     -0.05352    0.29526 -0.18  0.8562
adhere1     -0.24374    0.14817 -1.64  0.1000
nodes       -0.04759    0.01461 -3.26  0.0011
differ2      0.22033    0.17379  1.27  0.2049
differ3     -0.19042    0.20352 -0.94  0.3495
extent2     -0.45386    0.58364 -0.78  0.4368
extent3     -0.87839    0.56064 -1.57  0.1172
extent4     -1.08332    0.60933 -1.78  0.0754
surg1       -0.15024    0.11560 -1.30  0.1937
node41      -0.70430    0.14890 -4.73 2.2e-06
Log(scale)  -0.04572    0.04737 -0.97  0.3345

Scale= 0.955

Weibull distribution
Loglik(model)= -3099.2   Loglik(intercept only)= -3168.8
      Chisq= 139.2 on 15 degrees of freedom, p= 3.3e-22
Number of Newton-Raphson Iterations: 5
n= 711
AIC = 6232.468
```

Iš p reikšmės matome, jog nereikšminga kovariantė yra perforacija (*angl. perfor*), ją šaliname.

```
Call:
survreg(formula = Surv(time, status) ~ rx + sex + age + obstruct +
    adhere + nodes + differ + extent + surg + node4, data = train,
    dist = "weibull")
              Value Std. Error     z       p
(Intercept)  9.80555    0.64005 15.32 < 2e-16
rxLev+5FU    0.28497    0.13218  2.16  0.0311
rxObs       -0.14053    0.12206 -1.15  0.2496
sex1         0.02453    0.10440  0.23  0.8142
age         -0.00929    0.00461 -2.01  0.0439
obstruct1   -0.25466    0.13024 -1.96  0.0505
adhere1     -0.25004    0.14389 -1.74  0.0823
nodes       -0.04758    0.01461 -3.26  0.0011
differ2      0.22148    0.17368  1.28  0.2023
differ3     -0.18976    0.20348 -0.93  0.3510
extent2     -0.45358    0.58367 -0.78  0.4371
extent3     -0.87905    0.56066 -1.57  0.1169
extent4     -1.08691    0.60900 -1.78  0.0743
surg1       -0.15112    0.11549 -1.31  0.1907
node41      -0.70291    0.14870 -4.73 2.3e-06
Log(scale)  -0.04568    0.04737 -0.96  0.3349
```

```
Scale= 0.955

Weibull distribution
Loglik(model)= -3099.3   Loglik(intercept only)= -3168.8
      Chisq= 139.17 on 14 degrees of freedom, p= 1e-22
Number of Newton-Raphson Iterations: 5
n= 711
AIC = 6230.501
```

Šaliname kovariantę lytis (*angl. sex*)

```
Call:
survreg(formula = Surv(time, status) ~ rx + age + obstruct +
    adhere + nodes + differ + extent + surg + node4, data = train,
    dist = "weibull")
               Value Std. Error     z       p
(Intercept)  9.81217    0.63969 15.34 < 2e-16
rxLev+5FU    0.28366    0.13205  2.15  0.0317
rxObs       -0.14142    0.12199 -1.16  0.2463
age         -0.00934    0.00461 -2.03  0.0427
obstruct1   -0.25674    0.12994 -1.98  0.0482
adhere1     -0.25067    0.14388 -1.74  0.0815
nodes       -0.04732    0.01456 -3.25  0.0012
differ2      0.22527    0.17294  1.30  0.1927
differ3     -0.18494    0.20246 -0.91  0.3610
extent2     -0.44590    0.58277 -0.77  0.4442
extent3     -0.87283    0.56003 -1.56  0.1191
extent4     -1.08299    0.60877 -1.78  0.0752
surg1       -0.14955    0.11529 -1.30  0.1946
node41      -0.70568    0.14823 -4.76 1.9e-06
Log(scale)  -0.04566    0.04737 -0.96  0.3352

Scale= 0.955

Weibull distribution
Loglik(model)= -3099.3   Loglik(intercept only)= -3168.8
      Chisq= 139.12 on 13 degrees of freedom, p= 3.2e-23
Number of Newton-Raphson Iterations: 5
n= 711
AIC = 6228.556
```

Šaliname kovariantę apimtis (*angl. extent*). Pašalinę šią kovariantę, gauname, jog AIC padidėja – be šios kovariantės **AIC = 6232.315.** Todėl kovariantę *extent* paliekame modelyje. Nereikšmingos kovariantės: lytis ir perforacija.

Naudojame *step(direction = "forward")*

```
Start:  AIC=6232.47
Surv(time, status) ~ rx + sex + age + obstruct + perfor + adhere +
```

```
    nodes + differ + extent + surg + node4

Call:
survreg(formula = Surv(time, status) ~ rx + sex + age + obstruct +
    perfor + adhere + nodes + differ + extent + surg + node4,
    data = train, dist = "weibull")

Coefficients:
 (Intercept)    rxLev+5FU        rxObs         sex1          age    obstruct1
 9.808535561  0.285185097 -0.140311354  0.024931853 -0.009332824 -0.251157653
perfor1
-0.053515855
     adhere1        nodes      differ2      differ3      extent2      extent3
-0.243735378 -0.047590048  0.220329048 -0.190423569 -0.453856991 -0.878391160
extent4
-1.083324189
       surg1       node41
-0.150244224 -0.704299465

Scale= 0.95531

Loglik(model)= -3099.2   Loglik(intercept only)= -3168.8
      Chisq= 139.2 on 15 degrees of freedom, p= <2e-16
n= 711
```

Visos kovariantės paliekamos kaip reikšmingos.

Step(direction = "backward")

```
Call:
survreg(formula = Surv(time, status) ~ rx + age + obstruct +
    adhere + nodes + differ + extent + node4, data = train, dist = "weibull")

Coefficients:
 (Intercept)    rxLev+5FU        rxObs          age    obstruct1      adhere1
 9.789579076  0.284506548 -0.146285967 -0.009362441 -0.257515668 -0.249309165
nodes
-0.046810087
     differ2      differ3      extent2      extent3      extent4       node41
 0.230389922 -0.185156759 -0.465445094 -0.897384084 -1.094186769 -0.697971367

Scale= 0.9553267

Loglik(model)= -3100.1   Loglik(intercept only)= -3168.8
      Chisq= 137.46 on 12 degrees of freedom, p= <2e-16
n= 711
```

Naudojant step() funkciją su nurodymu atlikti atbulinę pažingsninę regresiją, papildomai pašalinama kovariantė *surg*, pabandę ją pašalinti ir „rankiniu“ būdu, matome, jog AIC dar labiau sumažėjo (**AIC = 6228,207**), tad iš galutinio modelio ją irgi pašaliname. Nereikšmingos kovariantės: lytis, perforacija ir laikas nuo

Galutinis modelis su Veibulo skirstiniu:

```
Call:
survreg(formula = Surv(time, status) ~ rx + age + obstruct +
    adhere + nodes + differ + extent + node4, data = train, dist = "weibull")

Coefficients:
 (Intercept)    rxLev+5FU        rxObs          age    obstruct1      adhere1
 9.789579076  0.284506548 -0.146285967 -0.009362441 -0.257515668 -0.249309165
Nodes
-0.046810087
     differ2      differ3      extent2      extent3      extent4       node41
 0.230389922 -0.185156759 -0.465445094 -0.897384084 -1.094186769 -0.697971367

Scale= 0.9553267

Loglik(model)= -3100.1   Loglik(intercept only)= -3168.8
      Chisq= 137.46 on 12 degrees of freedom, p= <2e-16
n= 711
```

Modelis užsirašo:

$$S_{Veibulas}(t;\eta,\upsilon)$$

$$= \exp\left\{-\left(\frac{t}{\exp(9{,}790)\times a}\right)^{\frac{1}{0{,}955}}\right\}$$

$$= exp\left\{\left(\frac{t}{17854{,}31\times a}\right)^{1{,}047}\right\}, kur\ \text{a} = \exp(Levamizolas5\ \times 0{,}285$$

$$+ konservavimas \times (-0{,}146) - amžius \times 0{,}009$$

$$- obstrukcija(yra) \times 0{,}258 - prilipimas(yra) \times 0{,}249$$

$$- vėžiniai\ limfmazgiai \times 0{,}047 + diferenciacija(vidutinė) \times 0{,}230$$

$$- diferenciacija(bloga) \times 0{,}185$$

$$- vietinio\ išplitimo\ mastas(raumenys) \times 0{,}465$$

$$- vietinio\ išplitimo\ mastas(serozė) \times 0{,}897$$

$$- vietinio\ išplitimo\ mastas(gretimos\ struktūros) \times 1{,}094$$

$$- teigiami\ limfmazgiai(daugiau\ nei\ 4) \times 0{,}698)\text{, čia } \eta$$

$$= mastelio\ parametras = scale = \exp(intercept)\text{, } \upsilon$$

$$= formos\ paarametras = shape = \frac{1}{scale} = 1{,}047.$$

2.5 Modelio gerumo įvertinimas

Modelio tinkamumą vertinsime testinėje aibėje pagal Akaikės ir Bajeso informacinius kriterijus.

5 lentelė. Akaikės ir Bajeso informaciniai kriterijai testavimo imčiai

	Loglogistinis	**Lognormalusis**	**Veibulo**
AIC	1514,343	1509,214	1528,767
BIC	1561,985	1550,504	1573,233

Kaip galime matyti iš 5 lentelės, nors ir didelio skirtumo tarp modelių nėra, tačiau pagal AIC ir BIC kriterijus lognormalusis modelis turi mažiausias reikšmes, o tai rodo, jog jis – geriausias.

2.6 Interpretacija

Kadangi gavome, jog lognormaliojo skirstinio modelis yra geriausias, tai interpretuosime jo koeficientus.

	Value	Std. Error	z	p
(Intercept)	9.74300	0.56090	17.37	< 2e-16
age	-0.01246	0.00471	-2.64	0.0082
obstruct1	-0.41528	0.14100	-2.95	0.0032
adhere1	-0.23851	0.16355	-1.46	0.1448
nodes	-0.04783	0.02142	-2.23	0.0256
differ2	0.26875	0.18549	1.45	0.1474
differ3	-0.35720	0.22126	-1.61	0.1064
extent2	-0.36648	0.49384	-0.74	0.4580
extent3	-0.91076	0.46597	-1.95	0.0506
extent4	-1.27314	0.54403	-2.34	0.0193
surg1	-0.23008	0.12575	-1.83	0.0673
node41	-0.75510	0.18238	-4.14	3.5e-05
Log(scale)	0.27200	0.04207	6.46	1.0e-10

Taip pat galime interpretuoti ir exp (β), matomas 6 lentelėje.

6 lentelė. Eksponentės betos

Amžius	**Obstrukcija (yra)**	**Prilipimas (yra)**	**Vėžiniai limfmazgiai**	**Diferenciacija (vidutinė)**	**Diferenciacija (bloga)**
0,988	0,660	0,788	0,953	1,308	0,700
Vietinis išplitimo mastas (raumenys)	**Vietinis išplitimo mastas (serozė)**	**Vietinis išplitimo mastas (gretimos struktūros)**	**Laikas nuo operacijos iki registracijos (ilgas)**	**Teigiamų limfmazgių skaičius (daugiau nei 4)**	
0,693	0,402	0,280	0,794	0,470	

Galime matyti, jog senstant trumpėja laikas iki mirties. Esant obstrukcijai (obstruct) trumpėja laikas iki įvykio negu nesant. Esant prilipimui prie gretimų organų (adhere) trumpėja laikas iki įvykio. Didėjant limfmazgių skaičiui, kuriuose aptiktas vėžys (nodes) trumpėja laikas iki įvykio. Esant vidutinei diferenciacijai lyginant su gera ilgėja laikas iki įvykio, o esant blogai - trumpėja lyginant su gera diferenciacija. Jei išplitimo mastas yra raumenyse, serozės dalyje ar gretimose struktūrose lyginant su pogleivine dalimi, tai laikas iki įvykio trumpėja. Jei paciento laikas nuo operacijos iki registracijos yra ilgas jam trumpėja laikas iki įvykio. Jei yra daugiau kaip 4 teigiami limfmazgiai, tai laikas iki įvykio trumpėja.

Iš exp(β) koeficientų galime matyti, jog amžiui padidėjus vienetu, vidutinis laikas iki mirties sumažėja 1,2 %, turint obstrukciją – 34 %, turint prilipimą – 21,2 %, kiekvienas limfmazgio skaičiaus padidėjimas, taip pat sumažina vidutinį laiką iki įvykio 4,7 %. Pacientų, kurių diferenciacija yra vidutinė lyginant su gera, vidutinis laikas iki įvykio yra ilgesnis 31 %, o pacientų, kurių diferenciacija yra bloga lyginant su gera diferenciacija, vidutinis laikas iki įvykio sutrumpėja 30 %. Pacientų, kurių vietinis išplitimo mastas yra raumenyse lyginant su pogleivinės dalies išplitimu, vidutinis laikas iki įvykio trumpėja 30 %, atitinkamai pacientų vidutinis išgyvenamumo laikas trumpėja 60 % ir 70 %, kai lyginame išplitimą serozės dalyje su pogleiviniu sluoksniu ir gretimose struktūrose su pogleiviniu sluoksniu. Pacientams vidutinis laikas iki mirties trumpėja 20 % , kai yra ilgas laikotarpis nuo operacijos iki registracijos, lyginant su trumpu laikotarpiu. Taip

pat 53 % trumpėja vidutinis laikas iki įvykio, jei pacientas turi daugiau nei 4 teigiamus limfmazgius.

3. IŠVADOS

Iš viso buvo 929 stebėjimai, pašalinus praleistas reikšmes stebėjimų liko 888. Duomenys buvo padalinti į mokymo ir testavimo aibes santykiu 80 : 20.

Iš pradinės duomenų analizės pastebėta, jog jau mirusiems pacientams dažniausiai priskirtas gydymo tipas – konservavimas, o rečiausias – levamizolas. Daugeliui pacientų nebuvo storosios žarnos obstrukcijos, perforacijos bei prilipimo prie gretimų organų. Taip pat išplitimas pogleivinėje dalyje buvo dažnesnis dar nemirusiems pacientams, o jau mirusiems pacientams dažnesnis išplitimas buvo gretimose struktūrose. Tiek cenzūruotų, tiek mirusių pacientų amžius labai panašus. Taip pat pastebėta, jog mirusiems pacientams limfmazgių skaičius buvo didesnis nei cenzūruotiems.

Nusibraižius Kaplan – Meier kreives ir pritaikius log ranginius arba *twostage* testus gavome, jog statistiškai reikšmingai skiriasi išgyvenamumo tikimybė, priklausanti nuo laiko, tarp lyties bei pacientų, kurie turėjau daugiau nei 4 teigiamus limfmazgius.

Patikrinus įvairius skirstinius nusprendėme dirbti su Veibulo, log – normaliuoju ir log – logistiniu skirstiniais. Taikant modelius, nesusidūrėme su multikolinearumo problema.

Pritaikius pažingsninę regresija buvo gauti modeliai:

$$S_{\log - logistinis}(t;\eta,\nu) = 1 - \left(\frac{1}{1+\left(\frac{t}{\exp(9{,}484)\times a}\right)^{-\frac{-1}{0{,}756}}}\right)$$

$$= 1 - \left(\frac{1}{1+\left(\frac{t}{13147{,}67\times a}\right)^{-1{,}323}}\right), kur\ a$$

$$= \exp(Levamizolas \times 0{,}214 - konservavimas \times 0{,}086 - amžius \times 0{,}011$$
$$- obstrukcija(yra) \times 0{,}317 - prilipimas(yra) \times 0{,}272$$
$$- vėžiniai\ limfmazgiai \times 0{,}042 + diferenciacija(vidutinė) \times 0{,}256$$
$$- diferenciacija(bloga) \times 0{,}329$$
$$- vietinio\ išplitimo\ mastas(raumenys) \times 0{,}347$$
$$- vietinio\ išplitimo\ mastas(serozė) \times 0{,}863$$
$$- vietinio\ išplitimo\ mastas(gretimos\ struktūros) \times 1{,}196$$
$$- laikas\ nuo\ operacijos\ iki\ registracijos(ilgas) \times 0{,}184$$
$$- teigiami\ limfmazgiai(daugiau\ nei\ 4) \times 0{,}787)\text{, čia } \eta = scale$$

$$= \exp(intercept) = mastelio\ parametras = 13147{,}67, \upsilon = shape = \frac{1}{scale}$$

$$= 1{,}323 = formos\ parametras.$$

$$S_{log-normalusis}(t;\ \mu,\sigma) = 1 - \Phi\left(\frac{ln(t)}{1{,}313\times a}\right), kur\ a$$

$$= exp(-amžius \times 0{,}012 - obstrukcija(yra) \times 0{,}415$$
$$- prilipimas(yra) \times 0{,}239 - vėžiniai\ limfmazgiai \times 0{,}048$$
$$+ diferenciacija(vidutinė) \times 0{,}269 - diferenciacija(bloga) \times 0{,}357$$
$$- vietinio\ išplitimo\ mastas(raumenys) \times 0{,}366$$
$$- vietinio\ išplitimo\ mastas(serozė) \times 0{,}911$$
$$- vietinio\ išplitimo\ mastas(gretimos\ struktūros) \times 1{,}273$$
$$- laiaks\ nuo\ operacijos\ iki\ registracijos(ilgas) \times 0{,}230$$
$$- teigiami\ limfmazgiai(daugiau\ nei\ 4) \times 0{,}755), \text{čia } \sigma = scale$$
$$= mastelio\ parametras = 1{,}313.$$

$$S_{Veibulas}(t;\eta,\upsilon)$$

$$= \exp\left\{-\left(\frac{t}{\exp(9{,}790)\times a}\right)^{\frac{1}{0{,}955}}\right\}$$

$$= exp\left\{\left(\frac{t}{17854{,}31\times a}\right)^{1{,}047}\right\}, kur\ \text{a} = \exp(Levamizolas5\ \times 0{,}285$$

$$+ konservavimas \times (-0{,}146) - amžius \times 0{,}009$$

$$- obstrukcija(yra) \times 0{,}258 - prilipimas(yra) \times 0{,}249$$

$$- vėžiniai\ limfmazgiai \times 0{,}047 + diferenciacija(vidutinė) \times 0{,}230$$

$$- diferenciacija(bloga) \times 0{,}185$$

$$- vietinio\ išplitimo\ mastas(raumenys) \times 0{,}465$$

$$- vietinio\ išplitimo\ mastas(serozė) \times 0{,}897$$

$$- vietinio\ išplitimo\ mastas(gretimos\ struktūros) \times 1{,}094$$

$$- teigiami\ limfmazgiai(daugiau\ nei\ 4) \times 0{,}698)\text{, čia } \eta$$

$$= mastelio\ parametras = scale = \exp(intercept)\text{, } \upsilon$$

$$= formos\ paarametras = shape = \frac{1}{scale} = 1{,}047.$$

Pagal AIC ir BIC kriterijus testavimo imtyje log - normalusis modelis turi mažiausias reikšmes, o tai rodo, jog jis – geriausias.

Geriausio modelio interpretacija:

- Senstant trumpėja laikas iki mirties. Esant obstrukcijai (obstruct) trumpėja laikas iki įvykio negu nesant. Esant prilipimui prie gretimų organų (adhere) trumpėja laikas iki įvykio. Didėjant limfmazgių skaičiui, kuriuose aptiktas vėžys (nodes) trumpėja laikas iki įvykio. Esant vidutinei diferenciacijai lyginant su gera ilgėja laikas iki įvykio, o esant blogai - trumpėja lyginant su gera diferenciacija. Jei išplitimo mastas yra raumenyse, serozė ar gretimos struktūros lyginant su pogleivine dalimi, tai laikas iki įvykio trumpėja. Jei paciento laikas nuo operacijos iki registracijos yra ilgas jam trumpėja laikas iki įvykio. Jei yra daugiau kaip 4 teigiami limfmazgiai, tai laikas iki įvykio trumpėja.
- Amžiui padidėjus vienetu, vidutinis laikas iki mirties sumažėja 1,2 %, turint obstrukciją – 34 %, turint prilipimą – 21,2 %, kiekvienas limfmazgio skaičiaus

padidėjimas, taip pat sumažina vidutinį laiką iki įvykio 4,7 %. Pacientų, kurių diferenciacija yra vidutinė lyginant su gera, vidutinis laikas iki įvykio yra ilgesnis 31 %, o pacientų, kurių diferenciacija yra bloga lyginant su gera diferenciacija, vidutinis laikas iki įvykio sutrumpėja 30 %. Pacientų, kurių vietinis išplitimo mastas yra raumenys lyginant su pogleivinės dalies išplitimu, vidutinis laikas iki įvykio trumpėja 30 %, atitinkamai pacientų vidutinis išgyvenamumo laikas trumpėja 60 % ir 70 %, kai lyginame išplitimą seroze su pogleiviniu sluoksniu ir gretimos struktūros su pogleiviniu sluoksniu. Pacientams vidutinis laikas iki mirties trumpėja 20 % , kai yra ilgas laikotarpis nuo operacijos iki registracijos, lyginant su trumpu laikotarpiu. Taip pat 53 % trumpėja vidutinis laikas iki įvykio, jei pacientas turi daugiau nei 4 teigiamus limfmazgius.