Vilniaus
universitetas

MATEMATIKOS IR INFORMATIKOS
FAKULTETAS

KOKSO IR BINARINIO ATSAKO REGRESIJOS MODELIAI

Tiriamasis projektas

Atliko: Ugnė Kniukštaitė, Austėja Valeikaitė,
Simona Gelžinytė, Laineda Morkytė,
Urtė Venciūtė
duomenų mokslas 3 k. 2gr.

Vilnius, 2023

TURINYS

1. ĮVADAS

1.1. Tyrimo tikslas

Pritaikyti Kokso ir binarinio atsako regresijos modelius pasirinktiems duomenims.

1.2. Tyrimo uždaviniai

- Atlikti pirminę duomenų analizę;
- Patikrinti modelių prielaidas;
- Sukonstruoti modelius;
- Įvertinti modelių tinkamumą;
- Pateikti gautų modelių interpretacijas;
- Apibendrinti gautus rezultatus, pateikti išvadas.

1.3. Duomenys ir programinė įranga

Pasirinktas duomenų rinkinys apie krūties vėžį. Duomenys surinkti iš Roterdamo navikų banko. Pateikti įvairūs požymiai apie pacientę. Priklausomas kintamasis – dienų skaičius iki pacientės mirties (binarinio atsako regresijos atveju priklausomas kintamasis – paciento statusas – mirė arba išgyveno) ir 10 Kokso regresijos atveju, o binarinio – 11 kovariančių:

- Metai – operacijos metai;
- Paciento amžius kada buvo atlikta operacija;
- Menopauzės statusas (0 – nebuvo / prieš menopauzę, 1 – buvo / po menopauzės);
- Dydis – naviko dydis (≤ 20, 20 - 50, > 50);
- Diferenciacijos laipsnis (2 arba 3);
- Limfmazgiai - teigiamų limfmazgių skaičius, sugrupuoti į 0 (≤ 5) ir į 1 (> 5);
- pgr – progesterono receptoriai (fmol / l);
- er – estrogenų receptoriai (fmol / l);
- Hormoninis gydymas (0 – ne, 1 – taip);
- Chemoterapija – ar taikyta chemoterapija (0 – ne, 1 - taip);
- Dienos iki mirties arba paskutinio stebėjimo.

Iš viso yra 2982 stebėjimai. Tyrimo metu naudota „R“ programinė įranga. Duomenis buvo padalinti į mokymo ir testavimo aibes santykiu 4 : 1.

2. PIRMINĖ ANALIZĖ

Prieš pradėdamos taikyti modelius, susipažinome su duomenimis – kiekybiniams kintamiesiems nusibraižėme stačiakampes diagramas ir pasižiūrėjome koreliacijas, o kategoriniams pasižiūrėjome dažnių lenteles priklausomai nuo paciento statuso, t. y. miręs ar išgyveno.

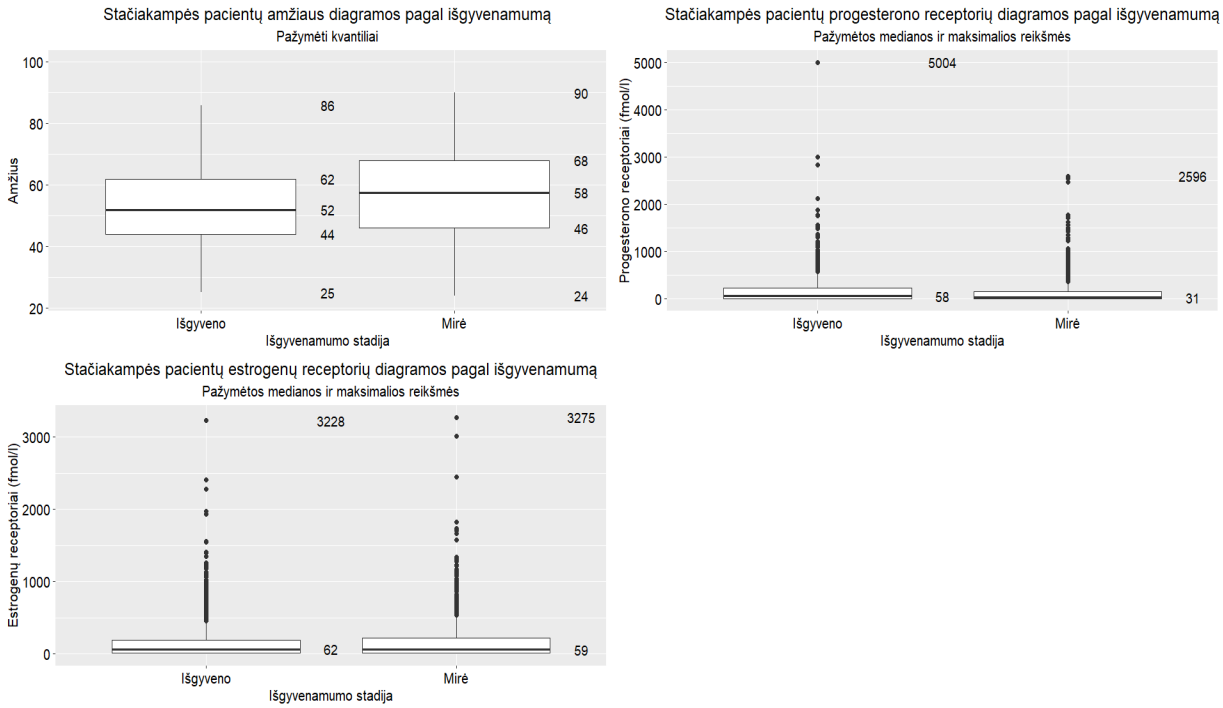
1 lentelė. Kategorinių kintamųjų dažniai

	Menopauzė		**Dydis**			**Diferenciacijos laipsnis**	
Statusas	**Nebuvo**	**Buvo**	**≤ 20**	**20 – 50**	**> 50**	**2**	**3**
0	884	866	973	645	92	532	1178
1	468	804	414	646	212	262	1010

2 lentelė. Kategorinių kintamųjų dažniai

	Hormoninis gydymas		**Chemoterapijos gydymas**	**Teigiamų limfmazgių skaičius**		
Statusas	**Nepritaikytas**	**Pritaikytas**	**Nebuvo**	**Buvo**	**Iki 5 imtinai**	**Daugiau už 5**
0	1530	180	1388	322	1530	180
1	1113	159	1014	258	806	466

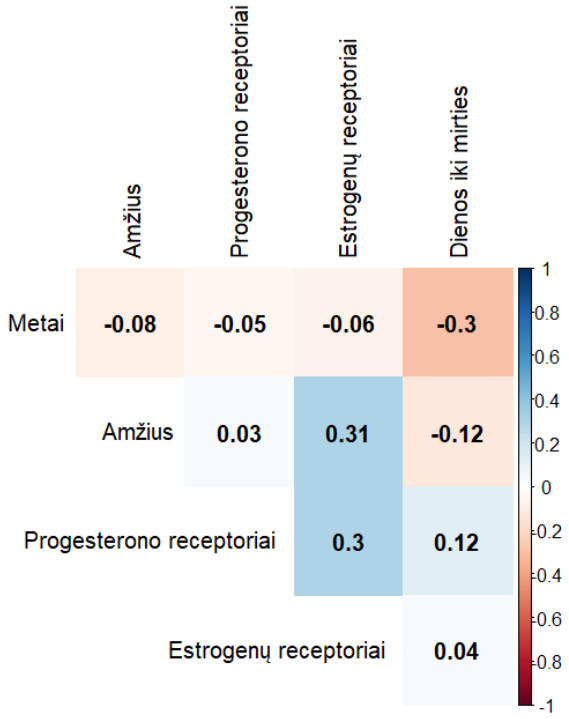
Iš (1 lentelės) matome, jog pacientų pasiskirstymas prieš ir po menopauzę gana panašus (grupės yra po lygiai apie 50 proc.), tačiau pastebėjome, jog taikytame tyrime pacienčių naviko dydis vyrauja ≤ 20 - imtyje tėra tik 92 pacientės > 50. Turinčių diferenciacijos 2 ir 3 laipsnį yra panašus dažnis su išgyvenusiomis ir mirusiomis pacientėmis (apie 50 % išgyvenusių ir neišgyvenusių). Taip pat didžioji dauguma netaiko hormoninio ir chemoterapinio gydymo. Be to, didžioji dalis apie 50 % sveikų pacientų turėjo iki 5 limfmazgių.



1 pav. Stačiakampės diagramos pagal klases

Iš (1 pav.) matome, jog ryškiau atsiskiria išgyvenusių ir mirusių pacientų progesterono diagramos, šio kintamojo medianos reikšmės buvo didesnės išgyvenusiems pacientams. Amžiaus ir estrogenų reikšmės nežymiai skiriasi tarp išgyvenusių ir mirusių pacientų.

Papildomai nubraižytas koreliacijų grafikas (3 pav.) galimai kovariančių koreliacijai numatyti. Galime pastebėti, jog turime silpną koreliaciją tarp amžiaus ir estrogenų receptorių. Tarp likusiųjų kintamųjų koreliacija yra labai silpna.



2 pav. Kovariančių koreliacijos matrica.

3. IŠGYVENAMUMO ANALIZĖ

Išgyvenamumo analizėje braižėme Kaplano – Mejerio kreives, norėdamos įvertinti kategorinių kintamųjų įtaką išgyvenamumo tikimybei. Taip pat pasirinkus semi parametrinį Kokso regresijos modelį perėjome visus modelio parinkimo etapus:

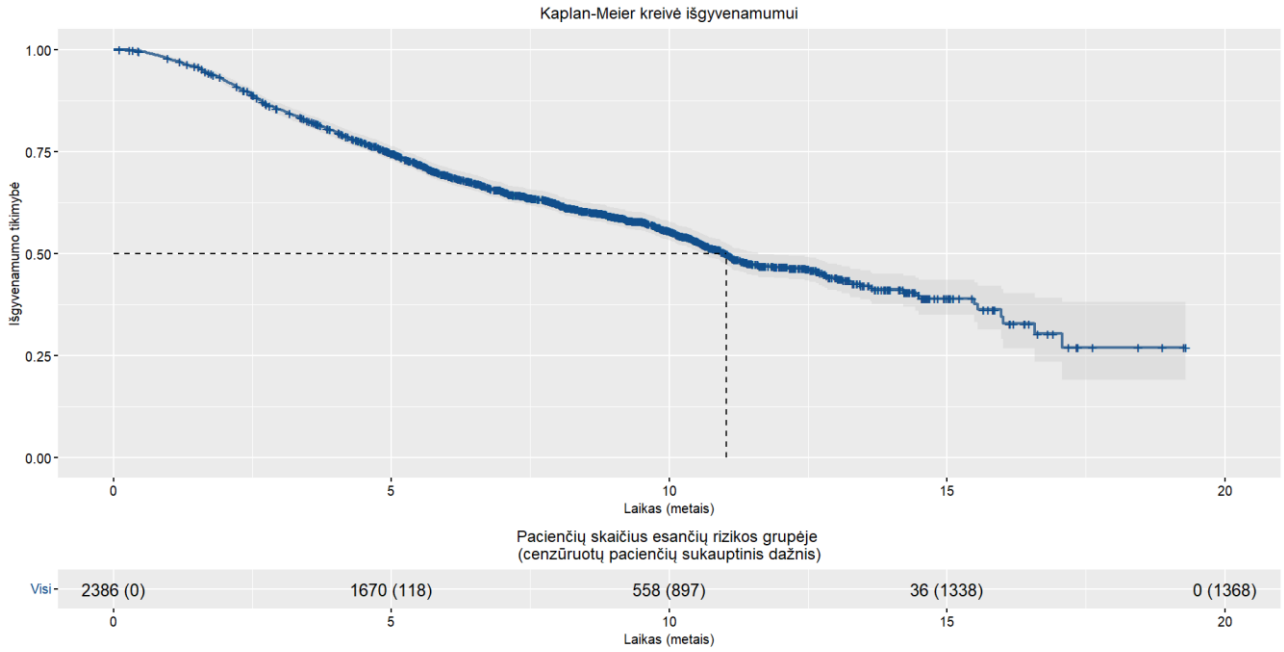
1. Prielaidų tikrinimas – proporcingosios rizikos prielaida, tiesiškumo prielaida kiekybiniams kintamiesiems, išskirčių ir multikolinearumo problema;
2. Reikšmingų kovariančių atranka;
3. Parametrų įvertinimas, interpretacija;
4. Modelio tinkamumo įvertinimas.

Duomenys buvo padalinti į mokymo ir testavimo aibes 80 : 20 santykiu. Pirminis Kokso regresijos modelis atrodo taip:

$$h(t|\boldsymbol{x}) = h_0(t) \cdot e^{\left(\begin{array}{c}\beta_1 \cdot Metai + \beta_2 \cdot Amžius + \beta_3 \cdot Menopauzė + \beta_4 \cdot Dydis + \beta_5 \cdot Diferenciacijos\ laipsnis + \beta_6 \cdot Limfmazgiai + \beta_7 \cdot pgr + \beta_8 \cdot er \\ + \beta_9 \cdot Hormoninis\ gydymas + \beta_{10} \cdot Chemoterapija\end{array}\right)}.$$

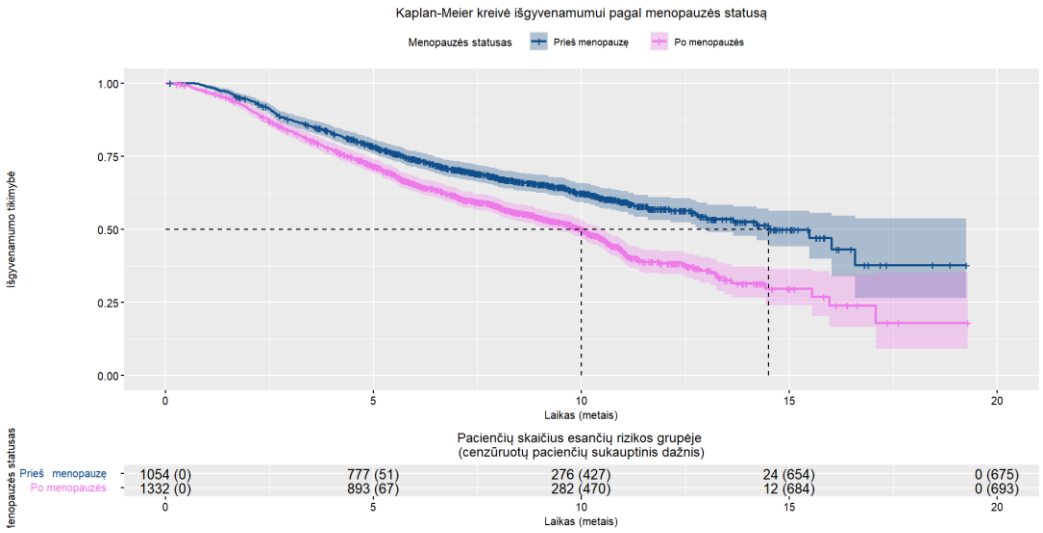
Buvo kuriami vieno kintamojo bei daugialypiai Kokso regresijos modeliai. Pagrindinis šių modelių tikslas atsirinkti kovariantes, kurios būtų statistiškai reikšmingos prognozuojant įvykį / mirtį, atsižvelgiant į laiką.

Pirmiausia, nusibraižėme Kaplan – Meier kreives, kad patikrinti, kaip kinta bendra išgyvenamumo tikimybė bei kaip ji priklauso nuo kategorinių kintamųjų: menopauzės, naviko dydžio, diferenciacijos laipsnio, limfmazgių skaičiaus, hormoninio gydymo ir chemoterapijos.



3 pav. Kaplan-Meier kreivė išgyvenamumui

Iš grafiko (3 pav.), parodančio kaip kinta bendra išgyvenamumo tikimybė matome, jog tikimybė išgyventi ilgiau nei 11 metų nukrenta iki 50 %. Apribotas vidurkio laikas yra 11,2 – tai plotas po išgyvenamumo kreive nuo 0 laiko momento iki paskutinio įvykio momento.

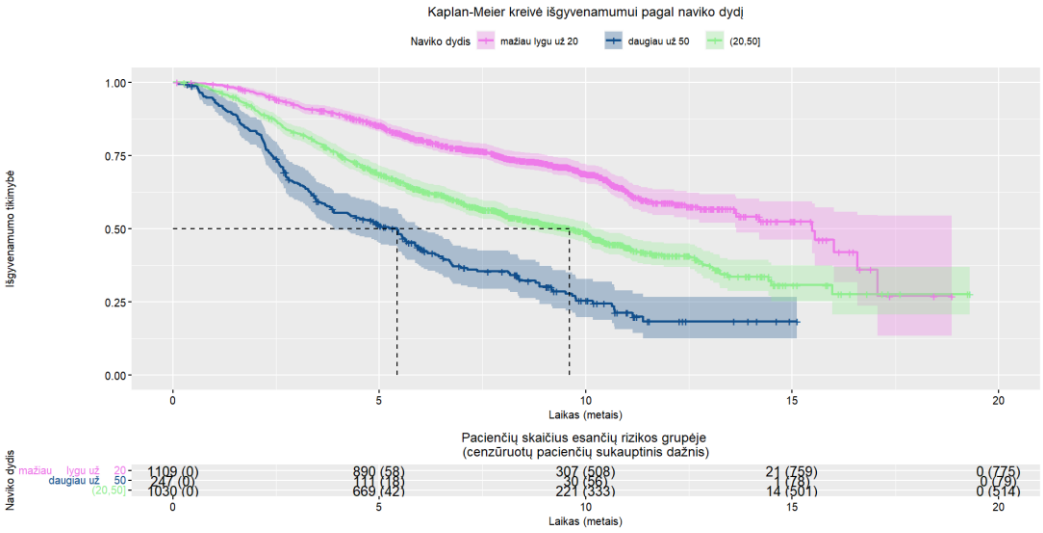


4 pav. Kaplan-Meier kreivė išgyvenamumui pagal menopauzės statusą

Iš išgyvenamumo pagal menopauzės statusą grafiko (4 pav.), matome, jog grupės tarpusavyje atsiskiria, todėl pritaikėme logranginį kriterijų ir tikrinome hipotezę:

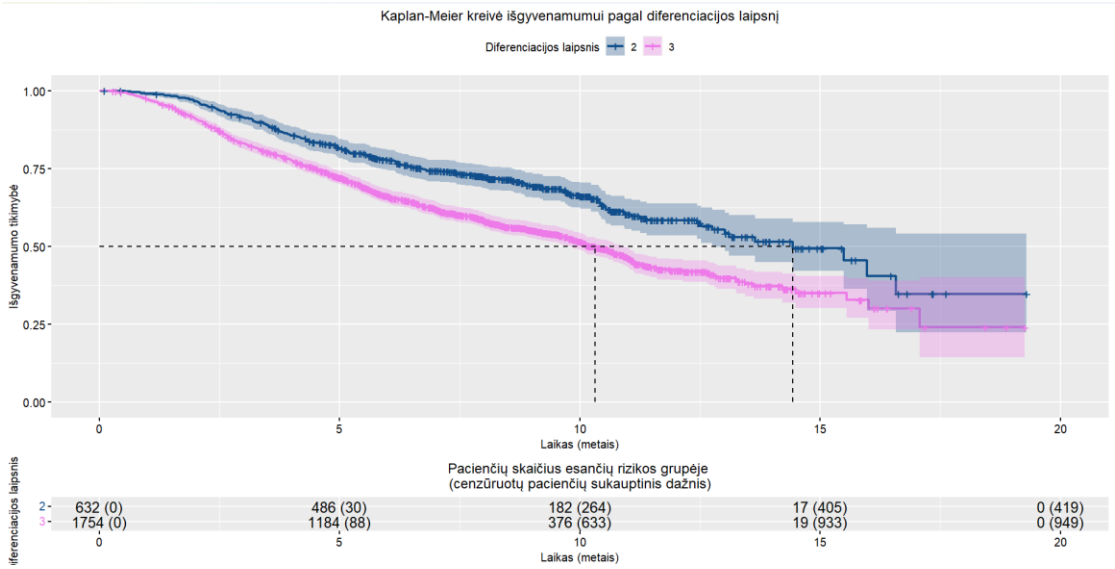
$$\begin{cases} H_0: tarp\ grupių\ nėra\ statistiškai\ reikšmingų\ skirtumų \\ H_A: tarp\ grupių\ yra\ statistiškai\ reikšmingų\ skirtumų \end{cases}.$$

Gavome p reikšmę < 0,001, taigi nulinę hipotezę atmetėme ir priėmėme alternatyvą – tarp menopauzės grupių yra statistiškai reikšmingų skirtumų. Iš 1054 pacienčių, kurioms menopauzė nebuvo įvykusi, 379 mirė, o iš 1332, kurioms menopauzė jau įvyko, mirė 639. Iš grafiko matome, jog pacienčių, kurioms menopauzė neįvyko, išgyvenamumo tikimybė išgyventi ilgiau nei 14 metų nukrenta iki 50 %, o tų, kurioms menopauzė įvyko, išgyvenamumo tikimybė išgyventi ilgiau nei 10 metų nukrenta iki 50 %.



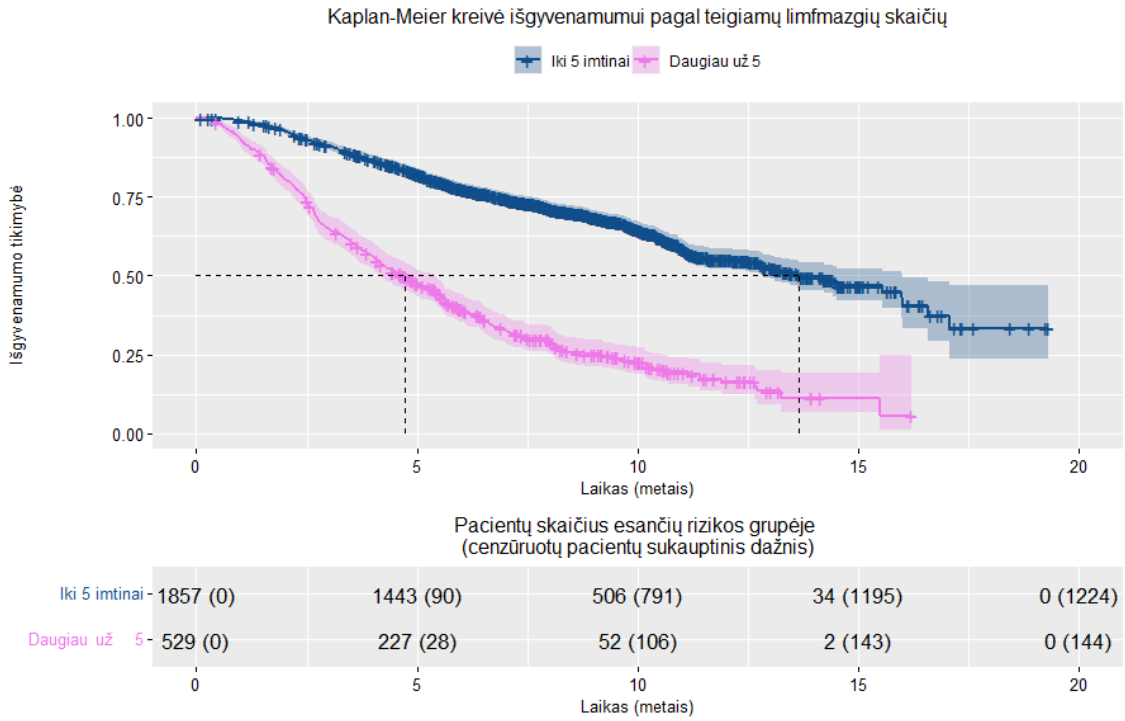
5 pav. Kaplan-Meier kreivė išgyvenamumui pagal naviko dydį

Iš išgyvenamumo pagal naviko dydį grafiko (5 pav.), matome, jog grupės tarpusavyje atsiskiria, todėl pritaikėme daugkartinį logranginį kriterijų su *Bonferroni* pataisa. Gavome, jog visos grupės tarpusavyje statistiškai reikšmingai skiriasi (p reikšmė < reikšmingumo lygmuo 0,05). Iš 1109 pacienčių, kurių naviko dydis buvo mažesnis arba lygus už 20, 334 mirė. Toms, kurių naviko dydis buvo (20, 50], iš 1030 mirė 516, o iš 247 atvejų, kai naviko dydis buvo didesnis nei 50, mirė 168 pacientės. Iš grafiko matome, jog pacienčių, kurių naviko dydis buvo intervale (20, 50], išgyvenamumo tikimybė išgyventi ilgiau nei 9 metus nukrenta iki 50 %, o tų, kurioms naviko dydis buvo didesnis už 50, išgyvenamumo tikimybė išgyventi ilgiau nei į metus nukrenta iki 50 %.



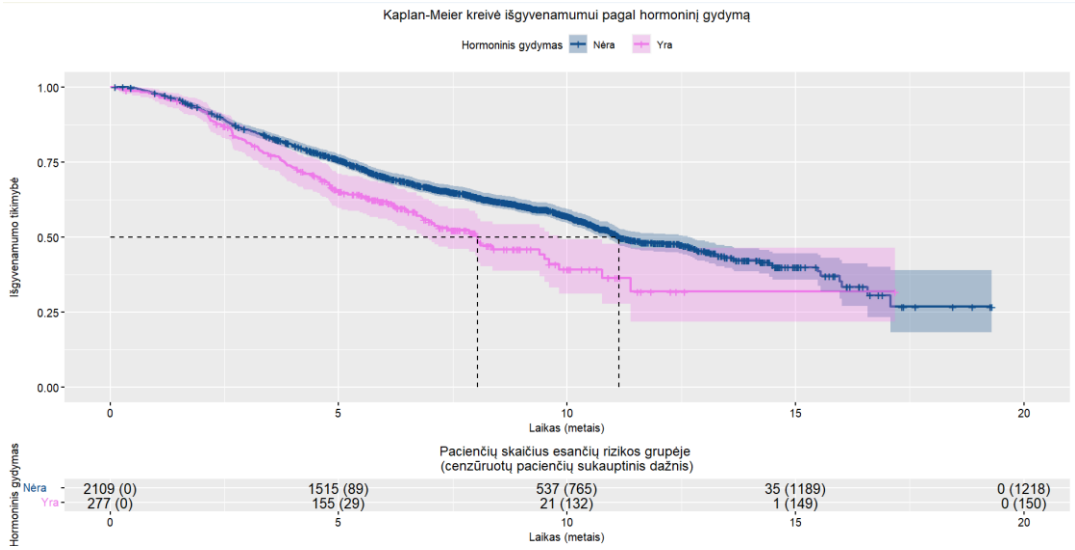
6 pav. Kaplan-Meier kreivė išgyvenamumui pagal diferenciacijos laipsnį

Iš išgyvenamumo pagal diferenciacijos laipsnį grafiko (6 pav.), matome, jog grupės tarpusavyje atsiskiria, todėl pritaikėme logranginį kriterijų. Gavome p reikšmę < 0,001, taiga nulinę hipotezę atmetėme ir priėmėme alternatyvą – tarp diferenciacijos laipsnių grupių yra statistiškai reikšmingų skirtumų. Iš 632 pacienčių, kurių diferenciacijos laipsnis buvo 2, 213 mirė, o iš 1754, kurių diferenciacijos laipsnis buvo 3, mirė 805. Iš grafiko matome, jog pacienčių, kurių diferenciacijos laipsnis buvo 2, išgyvenamumo tikimybė išgyventi ilgiau nei 14 metų nukrenta iki 50 %, o tų, kurių diferenciacijos laipsnis buvo 3, išgyvenamumo tikimybė išgyventi ilgiau nei 10 metų nukrenta iki 50 %.



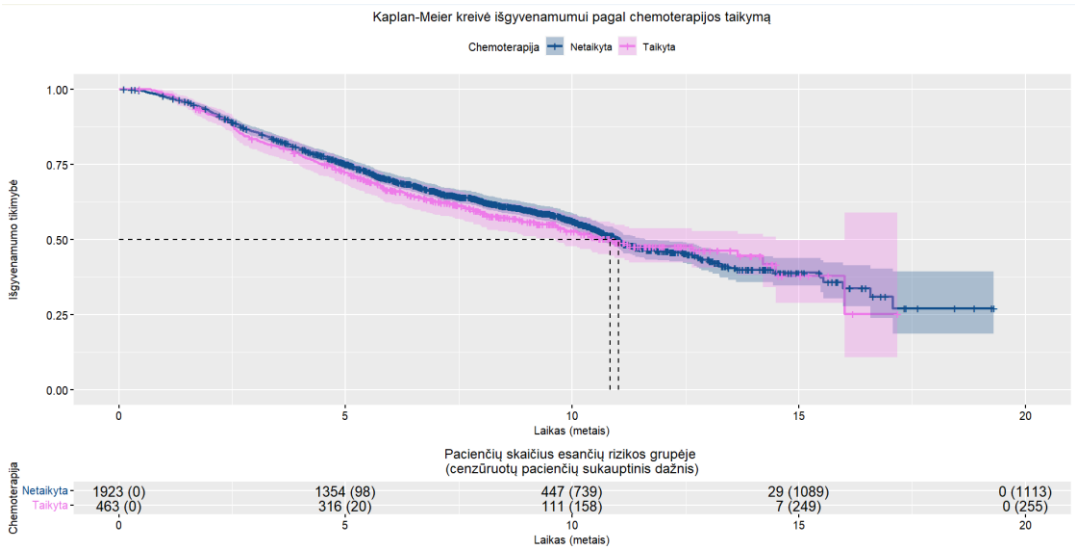
7 pav. Kaplan-Meier kreivė išgyvenamumui pagal teigiamų limfmazgių skaičių

Iš (7 pav.) galime matyti, jog turint daugiau nei 5 limfmazgius ir norint išgyventi ilgiau nei beveik 5 m. išgyvenamumo tikimybė nukrinta iki 50 %, o turint mažiau arba 5 limfmazgius išgyvenamumo tikimybė nukrinta iki 50 %, kai tikimąsi, jog pacientė išgyvens ilgiau nei 13,5 m. Iš grupės „Iki 5 imtinai“ buvo 663 mirtys iš 1857 pacienčių, paskutinis įvykis užfiksuotas praėjus beveik 3 metams, o iš kitos grupės paskutinis įvykis buvo užfiksuotas praėjus 2,5 metų ir iš 529 pacientų 358 mirė. Norint palyginti, ar grupės statistiškai reikšmingai skiriasi, taikėme logranginį kriterijų ir gavome $p < 0{,}0001$, t. y. nulinę hipotezę atmetėme, kuri teigė, jog nėra statistiškai reikšmingų skirtumų tarp grupių.



8 pav. Kaplan-Meier kreivė išgyvenamumui pagal hormoninį gydymą

Iš išgyvenamumo pagal hormoninį gydymą grafiko (8 pav.), matome, jog grupės tarpusavyje atsiskiria, todėl pritaikėme logranginį kriterijų. Gavome p reikšmę < 0,001, taigi nulinę hipotezę atmetėme ir priėmėme alternatyvą – tarp hormoninio gydymo grupių yra statistiškai reikšmingų skirtumų. Iš 2109 pacienčių, kurioms hormoninis gydymas nebuvo taikytas, 891 mirė, o iš 277, kurioms hormoninis gydymas buvo taikytas, mirė 127. Iš grafiko matome, jog pacienčių, kurioms hormoninis gydymas nebuvo taikytas, išgyvenamumo tikimybė išgyventi ilgiau nei 11 metų nukrenta iki 50 %, o tų, kurioms hormoninis gydymas buvo taikytas, išgyvenamumo tikimybė išgyventi ilgiau nei 8 metų nukrenta iki 50 %.

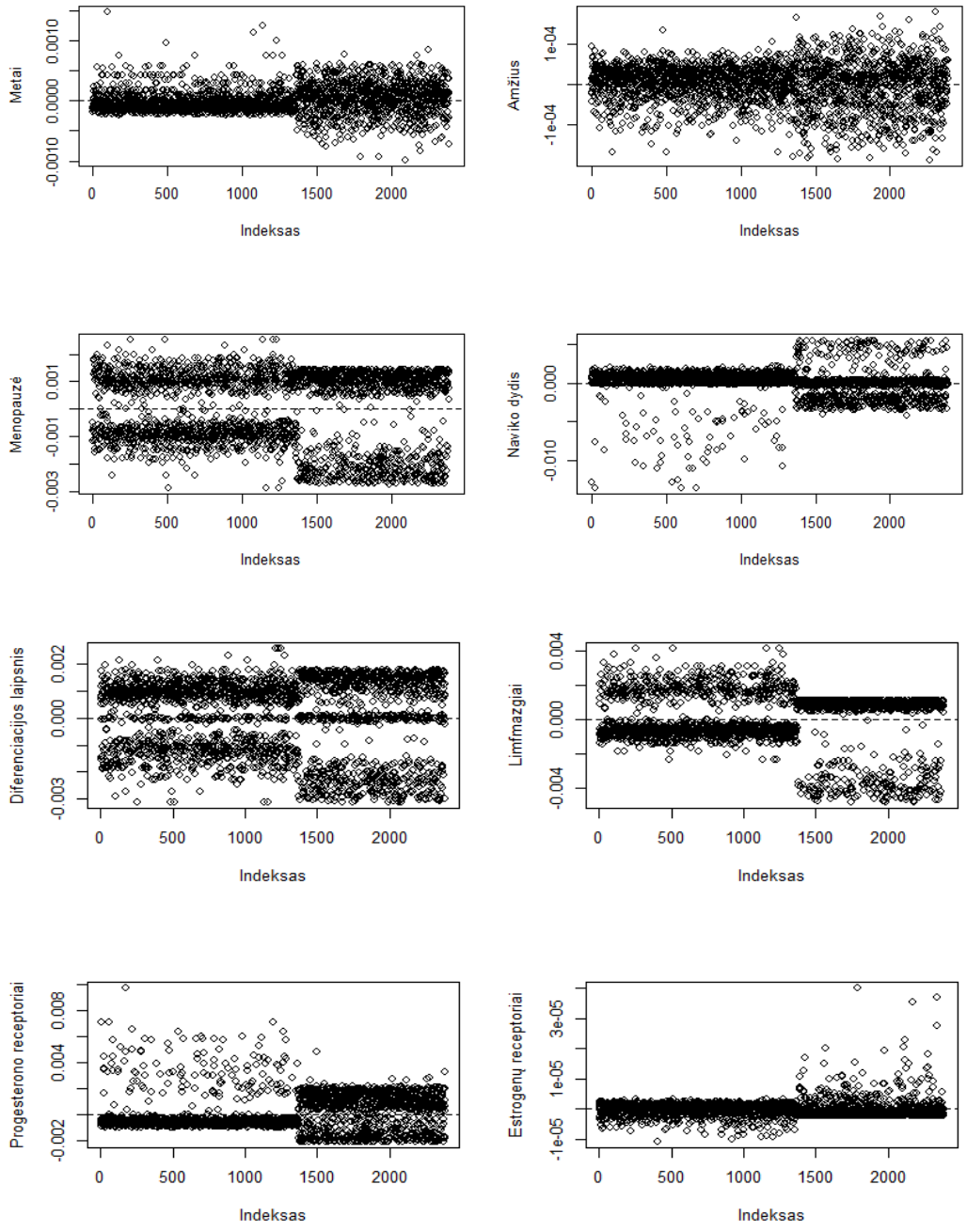


9 pav. Kaplan-Meier kreivė išgyvenamumui pagal chemoterapijos taikymą

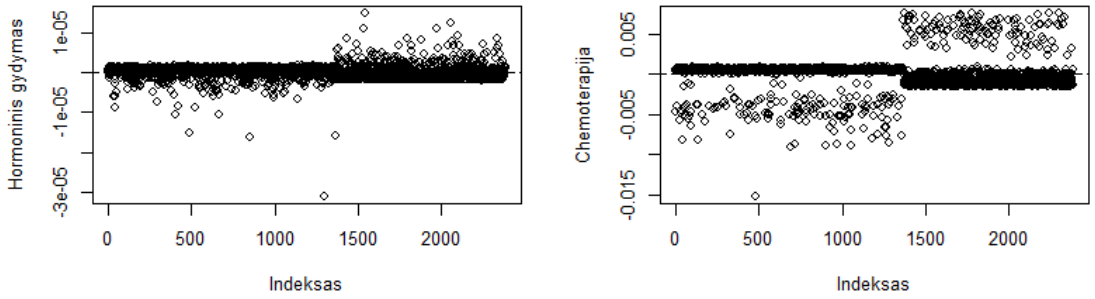
Iš išgyvenamumo pagal chemoterapijos taikymą grafiko (9 pav.), matome, jog grupės tarpusavyje neatsiskiria – kreivės kertasi beveik per vidurį, todėl negalėjome palyginti, ar skirtumai

tarp grupių statistiškai reikšmingi. Iš 1923 pacienčių, kurioms nebuvo taikyta chemoterapija, 810 mirė, o iš 463, kurioms buvo taikyta chemoterapija, mirė 208. Iš grafiko matome, jog pacienčių, išgyvenamumo tikimybė panaši abiem grupėms – išgyventi ilgiau nei maždaug 12 metų tikimybė nukrenta iki 50 %.

Pradėjome nuo vieno kintamojo semi parametrinių Kokso modelių taikymo. Iš pradžių pasitikrinome, ar modeliuose nėra išskirčių, naudojantis DFBetų statistiką. Gavome $[-0{,}04, 0{,}04]$ slenksčius, kuriuos viršijant stebėjimas yra laikomas išskirtimi.



10 pav. Išskirčių tyrimas Kokso modelyje I



11 pav. Išskirčių tyrimas Kokso modelyje II

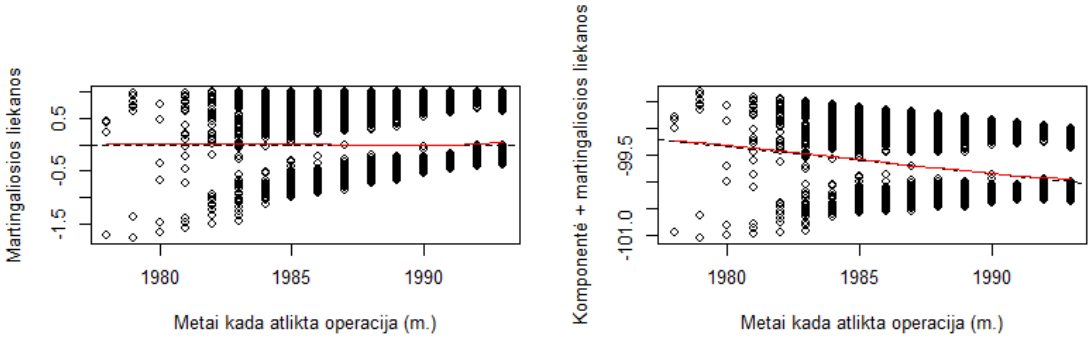
Iš (10 ir 11 pav.) matome, jog kiekviename vieno kintamojo Kokso modelyje nėra išskirčių.

Atsirinkome modelius, kuriuose kovariantė buvo reikšminga ir tenkino proporcingųjų rizikų prielaidą (p reikšmė > reikšmingumo lygmenį).

3 lentelė. Vieno kintamojo statistiškai reikšmingi bei tenkinantys proporcingųjų rizikų prielaidą modeliai

Modelis iš vienos kovariantės	**P reikšmė**	**Proporcingųjų rizikų p reikšmė**
Metai	< 0,0001	0,2000
Menopauzė	< 0,0001	0,0800
Diferencijavimo laipsnis	< 0,0001	0,1100
Hormoninis gydymas	< 0,0001	0,4900

Iš (3 lentelė) matome, jog modeliai, kurie tenkina proporcingųjų rizikų prielaidą bei yra statistiškai reikšmingi, yra metų, menopauzės, diferencijavimo lygio ir hormoninio gydymo kovariantės. Toliau metų kiekybiniam kintamajam patikrinsime tiesiškumo prielaidą.



12 pav. Martingaliosios ir komponentinės liekanos metų kovariantei

Iš (12 pav.) matome, jog metų kovariantė tenkina tiesiškumo prielaidą.

4 lentelė. Vieno kintamojo modelių informacija

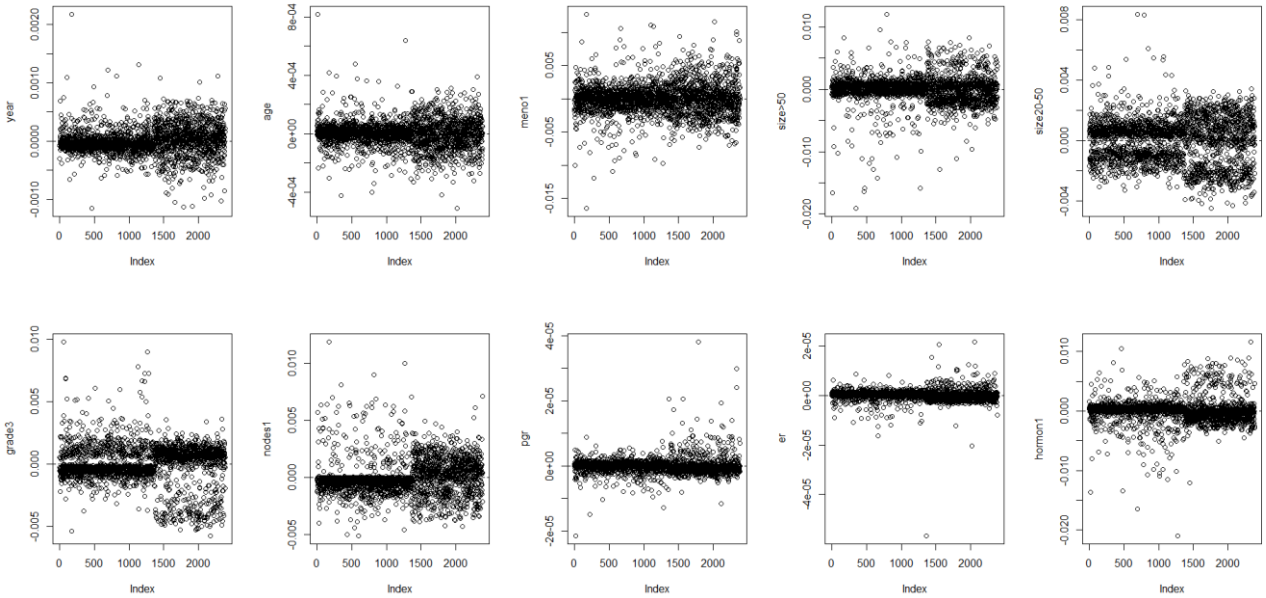
Modelis iš kovariantės	$\boldsymbol{\beta}$	**exp** ($\boldsymbol{\beta}$)	**AIC**	**Harrello konkordacijos indeksas testavimo aibei**
Metai	-0,0502	0,9511	14767,54	0,5128
Menopauzė (1)	0,4095	1,5061	14748,01	0,5571
Diferencijavimo laipsnis (3)	0,4690	1,5984	14748,66	0,5636
Hormoninis gydymas (1)	0,3789	1,4607	14774,55	0,5292

Iš (4 lentelė) galime matyti, jog geriausias modelis iš tirtų pagal Akaikės informacinį indeksą yra vieno kintamojo menopauzės Kokso modelis (14748,01), o pagal Harrello konkordacijos indeksą geriausias iš tirtų modelių yra diferencijavimo laipsnio kovariantės Kokso modelis, nes šis koeficientas didžiausias 0,5636. Taip pat galime matyti, jei 1 metais vėliau daryta operacija, mirties rizika sumažėja 5 %, taip gali būti dėl pažangesnių operavimo strategijų bei instrumentų. Jei buvo menopauzė 1,5 karto, didėja rizika numirti lyginant kai nebuvo menopauzės, be to, jei diferencijavimo laipsnis priklauso 3 kategorijai yra 1,6 karto didesnė rizika numirti nei priklausant 2 kategorijai. Jei buvo taikytas hormoninis gydymas 46 % yra didesnė mirties rizika nei netaikant šio gydymo.

Gauti modeliai atrodo taip:

- Metų $h(t|\boldsymbol{x}) = h_0(t) \cdot e^{(-0{,}0502 \cdot Operacijos\ metai)}$.
- Menopauzės $h(t|\boldsymbol{x}) = h_0(t) \cdot e^{(0{,}4095\ \cdot Menopauzė\ (po))}$.
- Diferiencijavimo lygio $h(t|\boldsymbol{x}) = h_0(t) \cdot e^{(0{,}4690\ \cdot Diferencijavimo\ lygis\ (3))}$.
- Hormoninio gydymo $h(t|\boldsymbol{x}) = h_0(t) \cdot e^{(0{,}3789 \cdot Hormoninis\ gydymas\ (buvo))}$.

Toliau pritaikėme daugialypį Kokso regresijos modelį. Į jį įtraukėme visas kovariantes, išskyrus laiką iki mirties ir mirties statusą. Pirmiausia, patikrinome išskirtis naudojant DFBetų liekanas. Gavome $[-0{,}04, 0{,}04]$ slenksčius, kuriuos viršijant stebėjimas yra laikomas išskirtimi.



13 pav. Požymių išskirtys

Matome, jog nei vienas požymis išskirčių neturi, nes nėra stebėjimų, kurie būtų peržengę anksčiau minėtą slenkstį.

Taip pat patikrinome, ar neturime multikolinearumo problemos.

5 lentelė. Multikolinearumo tikrinimas

Kovariantės	**Metai**	**Amžius**	**Progesterono receptoriai**	**Estrogenų receptoriai**
VIF	1,1119	2,8834	1,1343	1,1924

Iš (5 lentelė) matome, jog visiems kiekybiniams kintamiesiems VIF koeficientas yra mažesnis už 4, t. y. neturime multikolinearumo problemos.

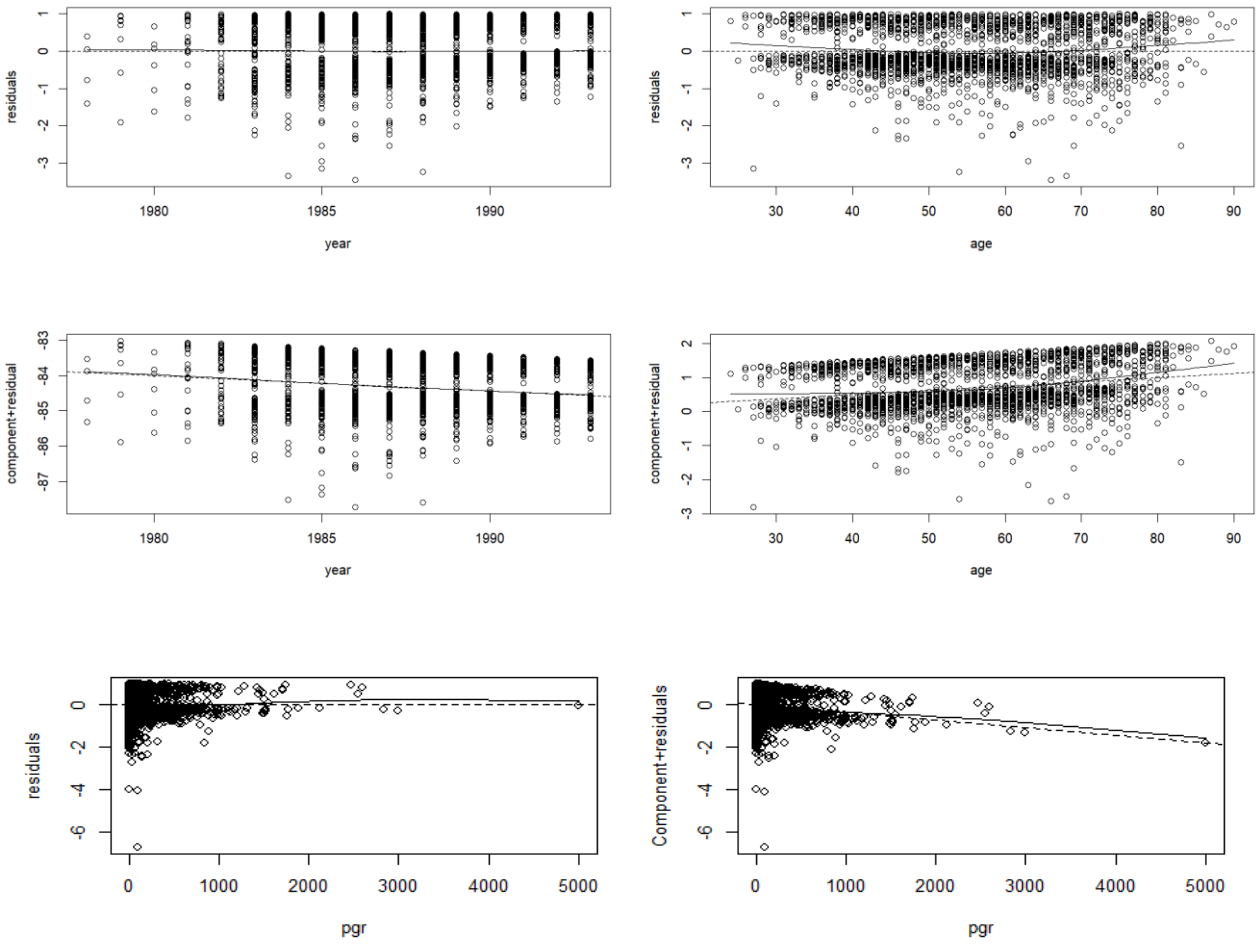
Kadangi su visomis kovariantėmis nebuvo tenkinama proporcingųjų rizikų prielaida, nusprendėme iš pradžių atsirinkti kovariantes. Pritaikėme pažingsninę regresiją ir gavome tokias **reikšmingas** kovariantes: metai, amžius, naviko dydis, diferenciacijos laipsnis, limfmazgiai, progesteronų receptoriai, kai reikšmingumo lygmuo 0,05.

6 lentelė. Po pažingsninės regresijos likusios kovariantės

Kovariantė	$\boldsymbol{\beta}$	**exp** ($\boldsymbol{\beta}$)	**p reikšmė**
Metai	-0,0409	0,9599	0,0002

Amžius	0,0118	1,0118	< 0,0001
Naviko dydis (> 50)	0,7063	2,0265	< 0,0001
Naviko dydis (20 - 50)	0,3888	1,4751	< 0,0001
Diferenciacijos laipsnis (3)	0,2922	1,3393	0,0002
Limfmazgiai (1)	1,0373	2,8217	< 0,0001
Progesteronų receptoriai	-0,0003	0,9996	0,0020

Toliau patikrinome likusių kiekybinių kovariančių tiesiškumą.



14 pav. Kovariančių tiesiškumas

Iš (14 pav.) galime matyti, jog visos kiekybinės kovariantės išpildo tiesiškumo reikalavimus.

7 lentelė. Likusių kovariančių proporcingųjų rizikų prielaidos tikrinimas I

Kovariantė	Proporcingųjų rizikų p reikšmė
Metai	0,1133
Amžius	0,0009
Naviko dydis	0,0117
Diferenciacijos laipsnis	0,1289
Progesteronų receptoriai	< 0,0001
Limfmazgiai	0,0065
Bendras	< 0,0001

Amžiaus, naviko dydžio, progesteronų receptorių bei limfmazgių skaičiaus kovariantės vis dar netenkino proporcingųjų rizikų prielaidos, t. y. jų p reikšmė < reikšmingumo lygmenį = 0,05 (žr. 7 lentelė). Norėdami išspręsti šią problemą sluoksniavome modelį pagal naviko dydžio kovariantę.

8 lentelė. Likusių kovariančių proporcingųjų rizikų prielaidos tikrinimas II

Kovariantė	Proporcingųjų rizikų p reikšmė
Metai	0,1761
Amžius	0,0006
Diferenciacijos laipsnis	0,1864
Progesteronų receptoriai	< 0,0001
Limfmazgiai	0,0461
Bendras	< 0,0001

Iš (8 lentelė) matome, jog sluoksniavus pagal naviko dydžio kovariantę modelyje vis dar liko kovariančių, netenkinančių proporcingosios rizikos, nes jų p reikšmė < reikšmingumo lygmenį = 0,05 (žr. 8 lentelė). Šią problemą sprendėme amžiaus kovariantę suskirstę į grupes (iki 30 m. imtinai – 1 grupė, [31, 50] – 2 grupė, [51, 70] – 3 grupė, nuo 71 m. – 4 grupė) ir sluoksniavę pagal naują kategorinį kintamąjį.

9 lentelė. Likusių kovariančių proporcingųjų rizikų prielaidos tikrinimas III

Kovariantė	Proporcingųjų rizikų p reikšmė
Metai	0,1200
Diferencijavimo laipsnis	0,2580
Progesteronų receptoriai	< 0,0001
Limfmazgiai	0,0540
Bendras	< 0,0001

Iš (9 lentelė) matome, jog limfmazgių kovariantė jau tenkina proporcingųjų rizikų prielaidą, tik progesteronų receptoriai vis dar ne. Norėdami panaikinti šią problemą, progesteronų receptorius suskirstėme į 2 grupes: 0 grupė – iki 100 imtinai, 1 grupė – virš 100 (fmol / l), ir sluoksniavome pagal naujai gautą kintamąjį.

10 lentelė. Likusių kovariančių proporcingųjų rizikų prielaidos tikrinimas IV

Kovariantė	**Proporcingųjų rizikų p reikšmė**
Metai	0,150
Diferencijavimo laipsnis	0,970
Limfmazgiai	0,100
Bendras	0,170

Iš (10 lentelė) galime matyti, jog visos likusios kovariantės tenkina proporcingųjų rizikos prielaidą, nes p reikšmė > reikšmingumo lygmenį = 0,05, tai reiškia, jog rizika nepriklauso nuo laiko.

11 lentelė. Daugialypio modelio kovariančių informacija

Kovariantė	$\boldsymbol{\beta}$	**exp** ($\boldsymbol{\beta}$)	**P reikšmė**
Metai	-0,0451	0,9559	< 0,001
Diferencijavimo laipsnis (3)	0,2997	1,3495	0,0002
Limfmazgiai (1)	1,0544	2,8701	< 0,001

Iš (11 lentelė) galime matyti, jog visos likusios kovariantės modelyje yra reikšmingos. Taip pat galime daryti išvadą, jog didinant 1 metais operacijos atlikimo metus mažėja mirties rizika apie 4 %, tai galime paaiškinti tuo, jog atsirado geresnių technologijų bei pažangesnių technikų atlikti operacijas. Be to, kai diferenciacijos laipsnis yra 3, tai 35 % didėja mirties rizika lyginant su grupe, kur diferenciacijos laipsnis yra 2. Taip pat turint daugiau nei 5 teigiamus limfmazgius mirties rizika padidėja 2,87 karto lyginant, kai pacientas turi 5 arba mažiau teigiamų limfmazgių.

Gautas modelis užrašomas taip:

$$h(t|\boldsymbol{x}) = h_0(t) \times e^{(-0{,}0451 \times Metai + 0{,}2997 \times Diferencijacijos\ laipsnis\ (3) + 1{,}0544 \times Limfmazgiai\ (1))}.$$

Šio modelio akaikės informacinis indeksas yra 8796,672, o Harrello konkordacijos indeksas 0,6442.

12 lentelė. Visų Kokso modelių palyginimas

Modelis	**AIC**	**Harrello konkordacijos indeksas testavimo aibei**

Vieno kintamojo: menopauzė (1)	14748,01	0,5571
Vieno kintamojo: diferencijavimo laipsnis (3)	14748,66	0,5636
Daugialypis	8796,672	0,6442

Visų gautų semi parametrinių Kokso modelių palyginimas pateiktas (12 lentelė). Iš šios lentelės galime matyti, jog geriausias AIC (8796,672) ir Harrello konkordacijos (0,6442) indeksai gauti, taikant daugialypį modelį, sluoksniuotą 3 kartus pagal naviko dydį, paciento amžių ir progesteronų receptorius.

4. BINARINIO ATSAKO REGRESIJA

Binarinio atsako regresijai buvo naudota logistinė jungimo funkcija. Duomenys buvo padalinti į testavimo ir mokymo aibes 20 : 80 santykiu. Šioje tyrimo dalyje limfmazgių 1 grupė reiškia, jog mažiau už 5 arba lygu, o 0 grupė daugiau už 5. Pradinis logistinės regresijos modelis atrodo taip:

$$\ln\left(\frac{P(\text{pacientė mirė})}{P(\text{pacientė nemirė})}\right) = \beta_0 + \beta_1 \cdot metai + \beta_2 \cdot paciento\ amžius + \beta_3 \cdot menopauzės\ statusas\ (turi) + \beta_4 \cdot dydis\ (20-50) + \beta_5 \cdot dydis(>50) + \beta_6 \cdot diferenciacijos\ laipsnis\ (3) + \beta_7 \cdot limfmazgiai\ (\leq 5) + \beta_8 \cdot pgr + \beta_9 \cdot estrogenųreceptoriai + \beta_{10} \cdot hormoninis\ gydymas\ (taip) + \beta_{11} \cdot chemoterapija\ (taip) + \beta_{12} \cdot dienos\ iki\ mirties.$$

4.1. Prielaidų patikrinimas

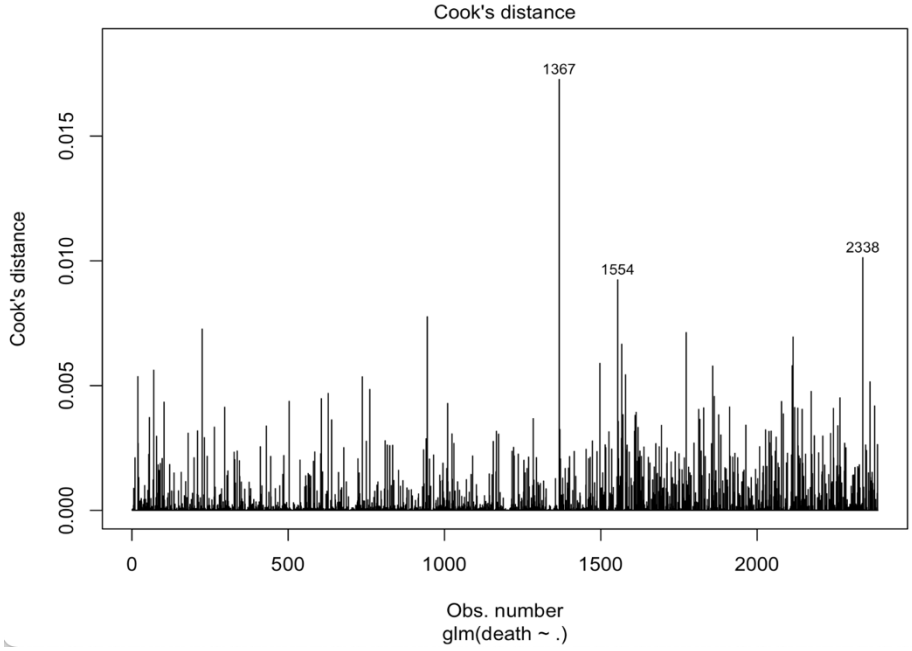
Visų pirma patikriname ar duomenys subalansuoti.

13 lentelė. Duomenų klasių pasiskirstymas

0	**1**
1368	1018

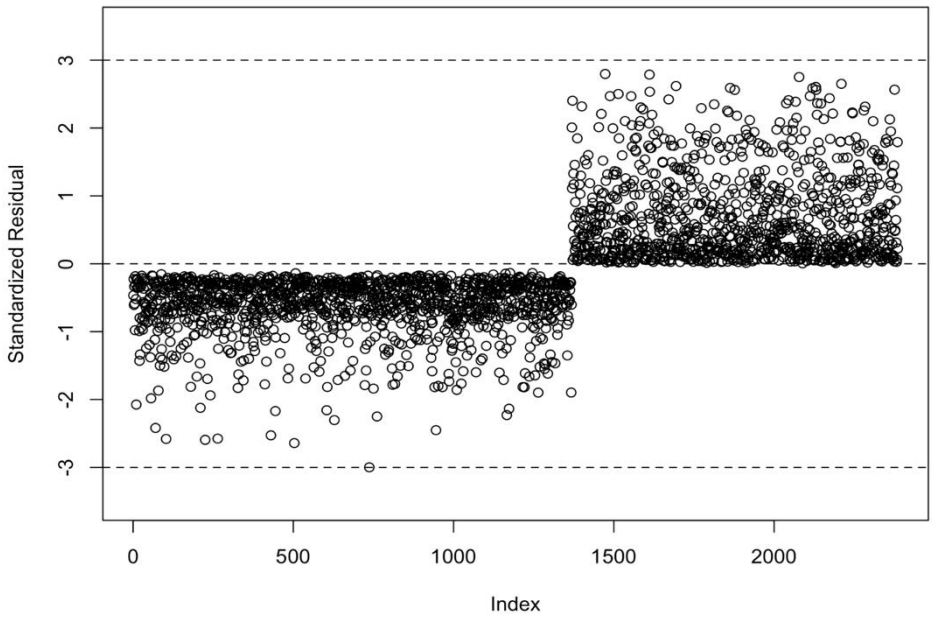
Galime matyti, jog nesubalansuotumo problemos nėra.

Toliau patikrinome išskirtis, tam naudojome Kuko matą bei standartizuotąsias liekanas.



15 pav. Kuko grafikas

Iš Kuko grafiko galime matyti, jog nei vienas stebėjimas neviršija vieneto, o tai reiškia, jog neturime išskirčių. Kad dar kartą įsitikinti, išskirtis patikrinkime su standartizuotomis liekanomis.



16 pav. Liekanų grafikas

Kaip galime matyti, su šiuo metodu taip pat nebuvo rasta išskirčių, stebėjimai neviršija trijų modulio.

Toliau tikrinsime multikolinaerumą kiekybiniams kintamiesiems.

14 lentelė. Multikolinearumo tikrinimas

Kovariantė	Metai	Amžius	Progesterono receptoriai	Estrogenų receptoriai	Dienos iki mirties
VIF	2,34	3,07	1,18	1,28	2,23

Kaip galime matyti, jokios kovariantės VIF neviršija 4, o tai reiškia, jog duomenys neturi multikolinerumo problemos.

Toliau atlikome pažingsninę regresiją.

```
Start:  AIC=1705.83
death ~ year + age + meno + size + grade + nodes + pgr + er +
    hormon + chemo + dtime

         Df Deviance    AIC
- meno    1   1679.8 1703.8
- age     1   1679.9 1703.9
- pgr     1   1680.0 1704.0
- hormon  1   1680.2 1704.2
- chemo   1   1680.4 1704.4
<none>        1679.8 1705.8
- size    2   1684.8 1706.8
- grade   1   1683.5 1707.5
- er      1   1684.0 1708.0
- nodes   1   1716.6 1740.6
- year    1   2253.3 2277.3
- dtime   1   2736.5 2760.5

Step:  AIC=1703.83
death ~ year + age + size + grade + nodes + pgr + er + hormon +
    chemo + dtime

         Df Deviance    AIC
- age     1   1680.0 1702.0
- pgr     1   1680.0 1702.0
- hormon  1   1680.2 1702.2
- chemo   1   1680.4 1702.4
<none>        1679.8 1703.8
- size    2   1684.8 1704.8
- grade   1   1683.5 1705.5
```

```
- er      1   1684.1 1706.1
- nodes   1   1716.6 1738.6
- year    1   2255.3 2277.3
- dtime   1   2736.6 2758.6

Step:  AIC=1701.98
death ~ year + size + grade + nodes + pgr + er + hormon + chemo +
    dtime

         Df Deviance    AIC
- pgr     1   1680.2 1700.2
- hormon  1   1680.5 1700.5
- chemo   1   1680.9 1700.9
<none>        1680.0 1702.0
- size    2   1684.9 1702.9
- grade   1   1683.7 1703.7
- er      1   1684.1 1704.1
- nodes   1   1716.6 1736.6
- year    1   2264.3 2284.3
- dtime   1   2747.3 2767.3

Step:  AIC=1700.15
death ~ year + size + grade + nodes + er + hormon + chemo + dtime

         Df Deviance    AIC
- hormon  1   1680.6 1698.6
- chemo   1   1681.1 1699.1
<none>        1680.2 1700.2
- size    2   1685.1 1701.1
- grade   1   1684.2 1702.2
- er      1   1684.2 1702.2
- nodes   1   1716.8 1734.8
- year    1   2264.7 2282.7
- dtime   1   2754.5 2772.5

Step:  AIC=1698.63
death ~ year + size + grade + nodes + er + chemo + dtime

        Df Deviance    AIC
- chemo  1   1681.8 1697.8
<none>       1680.6 1698.6
```

```
- size   2   1685.4 1699.4
- er     1   1684.5 1700.5
- grade  1   1684.6 1700.6
- nodes  1   1717.5 1733.5
- year   1   2303.7 2319.7
- dtime  1   2754.7 2770.7

Step:  AIC=1697.85
death ~ year + size + grade + nodes + er + dtime

        Df Deviance    AIC
<none>       1681.8 1697.8
- size   2   1686.7 1698.7
- er     1   1685.2 1699.2
- grade  1   1685.6 1699.6
- nodes  1   1720.1 1734.1
- year   1   2304.0 2318.0
- dtime  1   2754.8 2768.8

Call:  glm(formula = death ~ year + size + grade + nodes + er +
dtime,
    family = binomial(link = "logit"), data = df_train)

Coefficients:
(Intercept)          year       size>50    size20-50        grade3
nodes1
  1.238e+03   -6.201e-01    3.943e-01    2.471e-01    2.823e-01   -
9.733e-01
         er        dtime
  4.038e-04   -1.889e-03

Degrees of Freedom: 2385 Total (i.e. Null);  2378 Residual
Null Deviance:         3256
Residual Deviance: 1682  AIC: 1698
```

Pažingsninė regresija paliko operacijos metus, naviko dydį, diferenciacijos laipsnį, teigiamų limfmazgių skaičių, estrogenų receptorius bei dienas iki paskutinio stebėjimo/mirties kaip reikšmingas kovariantes.

Dabar pasižiūrėsime gauto modelio santrauką.

```
Call:
glm(formula = death ~ year + size + grade + nodes + er + dtime,
    family = binomial(link = "logit"), data = df_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0013  -0.5704  -0.2573   0.3491   2.7959

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.238e+03  6.416e+01  19.296  < 2e-16 ***
year        -6.201e-01  3.218e-02 -19.272  < 2e-16 ***
size>50      3.943e-01  2.245e-01   1.757   0.0790 .
size20-50    2.471e-01  1.329e-01   1.859   0.0630 .
grade3       2.823e-01  1.458e-01   1.935   0.0529 .
nodes1      -9.733e-01  1.580e-01  -6.161 7.21e-10 ***
er           4.038e-04  2.218e-04   1.820   0.0687 .
dtime       -1.889e-03  8.258e-05 -22.877  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3256.2  on 2385  degrees of freedom
Residual deviance: 1681.9  on 2378  degrees of freedom
AIC: 1697.9

Number of Fisher Scoring iterations: 6
```

Matome, jog statistiškai reikšmingos kovariantės, kai reikšmingumo lygmuo 0,05, yra operacijos metai, teigiamas linfmazgių skaičius ir dienos iki paskutinio stebėjimo/mirties, tačiau išmetus likusias kovariantes AIC rodiklis padidėja, todėl jas paliksime modelyje.

Kad galėtumėme interpretuoti detaliau kovariantes, kiekvienai kovariančių koeficiento reikšmei suskaičiavome eksponentę.

15 lentelė. Koeficiento eksponenčių reikšmės

Metai	**Amžius**	**Dydis (> 50)**	**Dydis (20 -50)**	**Diferenciacijos laipsnis**	**Limfmazgiai**	**Estrogenų receptoriai**	**Laikas iki mirties**
0,54	1,01	1,48	1,28	1,33	0,38	1,04	0,99

Iš pateiktos lentelės matome, jog pavėlinus operacijos metus per vienetą, labiau tikėtina išgyventi 46 %, padidėjus amžiui 1 metais 1 %. labiau tikėtina, kad žmogus mirs. Būnant 3-ioje diferenciacijos kategorijoje, 33% labiau tikėtina,, kad mirs nei būnant 2-oje. Turint mažiau arba 5 limfmazgius 62 % mažiau tikėtina, kad pacientas mirs nei turint daugiau kaip 5 limfmazgius. Jeigu navikas yra > 50, tai pacientui tikimybė mirti padidėja 48 % lyginant su naviko dydžiu mažesniu už 20, taip pat jeigu naviko dydis 20 - 50, tai tikimybė mirti padidėja 33 % lyginant su naviku dydžio grupe mažiau už 20, estrogenų receptoriui padidėjus per 100 vienetų, tikimybė mirti padidėja 4 %, padidinus per vienetą dienas iki paskutinio stebėjimo / mirties, tikimybė mirti sumažėja 1 %.

Apskaičiavome klasifikavimo lentelę mokymosi duomenims.

16 lentelė. Klasifikavimo lentelė mokymosi duomenims

	1	**0**	
1	788	130	**Jautrumas**: 0,86
0	230	1238	**Specifiškumas**: 0,84
	Tikslumas (precision): 0,77	**Negative predictive value**: 0,9	**Bendras tikslumas**: 0,85

Matome, jog bendras tikslumas yra 0,85. Taip pat gerai atpažįsta nemirusius pacientus, kai iš tikrųjų jie yra gyvi, modelio tikslumas (angl. *precision*) yra 0,77, t. y. kiek iš visų priskirtų mirusių asmenų yra tikrai mirę.

Optimaliausio slenksčio radimui pasinaudojome Youdeno indekso skaičiavimo formule:

$$J = jautrumas + specifiškumas - 1,$$

nes jautrumas ir specifiškumas vienodai svarbūs. Pritaikius Youdeno metodą, gavome, jog optimaliausias slenkstis yra 0,39.

Sudaryta klasifikavimo matrica testavimo duomenims, naudojant slenksčį 0,39.

17 lentelė. Klasifikavimo lentelė testavimo duomenims

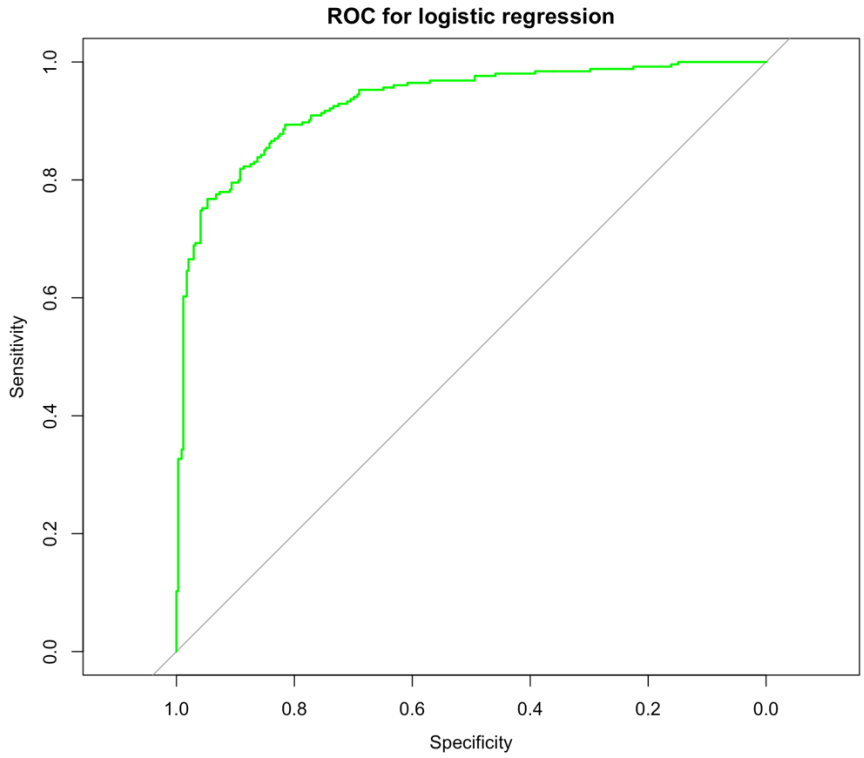
	1	**0**	
1	211	43	**Jautrumas**: 0,83
0	47	295	**Specifiškumas:** 0,86
	Tikslumas (precision): 0,82	**Negative predictive value**: 0,87	**Bendras tikslumas**: 0,85

Iš klasifikavimo matricos matome, jog pagrindinėje įstrižainėje yra teisingai suklasifikuotų pacienčių skaičiai. Matome, jog gerai atpažįsta mirusias ir išgyvenusias pacientes. Bendras tikslumas yra toks pats, kaip naudojant 0,5 slenkstį, radimas optimaliausio slenkščio naudojant Youdeno metodą didelės įtakos neturėjo, kadangi duomenys yra subalansuoti.

F_1 score gavome 0,82, kuris yra randamas pagal formulę:

$$F_\beta = \frac{(1 + \beta^2)(tikslumas \cdot jautrumas)}{(\beta^2 \cdot tikslumas + jautrumas)}, kur\ \beta\ gali\ būti\ 0{,}5, 1, 2.$$

Mes naudojome $\beta = 1$, nes jautrumas ir tikslumas vienodai svarbūs.



17 pav. ROC kreivė

Matome paveikslėlyje ROC kreivę sudarytam modeliui, plotas po kreive yra 0,9316.

Gautas modelis atrodo taip:

$$\ln\left(\frac{P(pacientė\ miręs)}{P(pacientė\ nemirė)}\right)$$
$$= 1238 - 0{,}6201 \cdot metai + 0{,}3943 \cdot dydis\ (>50) + 0{,}2471$$
$$\cdot dydis\ (20-50) + 0{,}2823 \cdot diferenciacijos\ laipsnis\ (3) - 0{,}9733$$
$$\cdot\ limfmazgiai\ (\leq 5) + 0{,}0004038 \cdot estrogenų receptoriai - 0{,}001889$$
$$\cdot\ dienos\ iki\ mirties/paskutinio\ stebėjimo.$$

Jo AIC 1697,9, determinacijos pseudokoeficientas 0,483, kuris gaunamas:

$$R^2 = 1 - \frac{deviacija}{nulinė\ deviacija}.$$

.

5. IŠVADOS

Nubraižius Kaplan – Meier kreivę, parodančią kaip kinta bendra išgyvenamumo tikimybė gavome, jog tikimybė išgyventi ilgiau nei 11 metų nukrenta iki 50 %. Taip pat pritaikius logranginį kriterijų gavome, jog statistiškai reikšmingai skiriasi išgyvenamumo tikimybė, priklausanti nuo laiko tarp menopauzės statuso, skirtingų navikų dydžių, diferenciacijos laipsnio, limfmazgių skaičiaus bei hormoninio gydymo.

Buvo gauti 4 semi parametriniai vieno kintamojo Kokso modeliai, kurie tenkino tiesiškumo, išskirčių nebuvimo ir proporcingųjų rizikų prielaidas. Iš jų geriausias pagal Akaikės informacinį indeksą yra menopauzės modelis (14748,01), o pagal Harrello konkordacijos indeksą testavimo aibei geriausias iš tirtų modelių yra diferenciacijos laipsnio kovariantės Kokso modelis (0,5636). Galima modelių interpretacija: jei buvo menopauzė, tai 1,5 karto didėja rizika numirti lyginant, kai nebuvo menopauzės, be to, jei diferencijavimo laipsnis priklauso 3 kategorijai, yra 1,6 karto didesnė rizika numirti nei priklausant 2 kategorijai.

Pradinis daugialypės regresijos modelis atrodė taip:

$$h(t|\boldsymbol{x}) = h_0(t) \cdot e^{\begin{pmatrix} \beta_1 \cdot Metai + \beta_2 \cdot Amžius + \beta_3 \cdot Menopauzė + \beta_4 \cdot Dydis + \beta_5 \cdot Diferenciacijos\ laipsnis + \beta_6 \cdot Limfmazgiai \\ + \beta_7 \cdot Progesteronų\ receptoriai + \beta_8 \cdot Estrogenų\ receptoriai \\ + \beta_9 \cdot Hormoninis\ gydymas + \beta_{10} \cdot Chemoterapija \end{pmatrix}}$$

Šis modelis neturėjo išskirčių, tačiau nebuvo tenkinama proporcingųjų rizikų prielaida. Su pažingsnine regresija buvo atsirinktos reikšmingos kovariantės, kai reikšmingumo lygmuo 0,3. Šį modelį sudarė 6 kovariantės, iš kurių visos kiekybinės tenkino tiesiškumo prielaidą. Norint išpildyti proporcingųjų rizikų prielaidą, amžiaus ir progesteronų receptorių kovariantės buvo suskirstytos į grupes, atitinkamai {iki 30 m. imtinai – 1 grupė, [31, 50] – 2 grupė, [51, 70] – 3 grupė, nuo 71 m. – 4 grupė} ir {0 grupė – iki 100 imtinai, 1 grupė – virš 100 (fmol / l)}, ir sluoksniuota 3 kartus pagal naviko dydį, sukategorizuotas amžiaus ir progesteronų receptorių kovariantes. Gautas modelis užrašomas taip:

$$h(t|\boldsymbol{x}) = h_0(t) \cdot e^{(-0{,}0451 \cdot Metai + 0{,}2997 \cdot Diferencijacijos\ laipsnis\ (3) + 1{,}0544 \cdot Limfmazgiai\ (1))}.$$

Didinat 1 metais operacijos atlikimo metus mažėja mirties rizika apie 4 %, tai galime paaiškinti tuo, jog atsirado geresnių technologijų bei pažangesnių technikų atlikti operacijas. Be to, kai diferenciacijos laipsnis yra 3, tai 35 % didėja mirties rizika lyginant su grupe, kur diferenciacijos laipsnis yra 2. Taip pat turint daugiau nei 5 teigiamus limfmazgius mirties rizika padidėja 2,87 karto lyginant, kai pacientas turi 5 arba mažiau teigiamų limfmazgių.

Šio modelio Akaikės informacinis indeksas yra 8796,672, o Harrello konkordacijos indeksas testavimo aibei 0,6442. Lyginant su visais gautas semi parametrinio Kokso modeliais, tai geriausias gautas modelis su mažiausiu AIC bei didžiausiu Harrello konkordacijos indeksu.

Pradinis nagrinėtas logistinės regresijos modelis atrodė taip:

$$\ln\left(\frac{P(\text{pacientė mirė})}{P(\text{pacientė nemirė})}\right) = \beta_0 + \beta_1 \cdot metai + \beta_2 \cdot paciento\ amžius + \beta_3 \cdot menopauzės\ statusas\ (turi) + \beta_4 \cdot dydis\ (20-50) + \beta_5 \cdot dydis(>50) + \beta_6 \cdot diferenciacijos\ laipsnis\ (3) + \beta_7 \cdot limfmazgiai\ (\leq 5) + \beta_8 \cdot pgr + \beta_9 \cdot estrogenųreceptoriai + \beta_{10} \cdot hormoninis\ gydymas\ (taip) + \beta_{11} \cdot chemoterapija\ (taip) + \beta_{12} \cdot dienos\ iki\ mirties.$$

Pašalinus nereikšmingas kovariantes galutinis modelis atrodo taip:

$$\ln\left(\frac{P(\text{pacientė mirė})}{P(\text{pacientė nemirė})}\right) = 1238 - 0{,}6201 \cdot metai + 0{,}3943 \cdot dydis\ (>50) + 0{,}2471 \cdot dydis\ (20-50) + 0{,}2823 \cdot diferenciacijos\ laipsnis\ (3) - 0{,}9733 \cdot limfmazgiai\ (\leq 5) + 0{,}0004038 \cdot estrogenųreceptoriai - 0{,}001889 \cdot dienos\ iki\ mirties/paskutinio\ stebėjimo.$$

Jo AIC 1697,9, determinacijos pseudokoeficientas 0,483.

Pavėlinus operacijos metus per vienetą, tikimybė mirti sumažėja 46 %, padidėjus amžiui 1 metais 1 % labiau tikėtina, kad žmogus mirs. Būnant 3-ioje diferenciacijos kategorijoje, 33% labiau tikėtina,, kad mirs nei būnant 2-oje. Padidinus per vienetą teigiamų linfmazgių skaičių, tikimybė mirti sumažėja 62 % Jeigu navikas yra > 50 ir jis padidėja dar vienetu, tikimybė mirti padidėja 48 %, jeigu naviko dydis 20 - 50 ir padidėja per vienetą, tikimybė mirti padidėja 33 %, estrogenų receptoriui padidėjus per 100 vienetų, tikimybė mirti padidėja 4 %, padidinus per vienetą dienas iki paskutinio stebėjimo/mirties, tikimybė mirti sumažėja 1 %.

Buvo rastas optimaliausias slenkstis, pasinaudojus Youden metodu: 0,39, tačiau jo pakeitimas didelės įtakos neturėjo, nes duomenys buvo subalansuoti.

Bendras klasifikavimo tikslumas testavimo duomenims yra 0,85. F_1 balas - 0,82, plotas po ROC kreive 0,9316. Iš visų rodiklių galime teigti, jog logistinės regresijos modelis tinka duomenims.