VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Regresinė analizė

4 laboratorinis darbas

Atliko: 3 kurso 2 grupės studentai:

Matas Amšiejus

Sandra Macijauskaitė

Salvija Račkauskaitė

Darbo vadovė: doc. dr. Rūta Levulienė

Vilnius, 2022

TURINYS

ĮVADAS....4
1. DUOMENYS....5
1.1.Duomenų aprašymas....5
2. PRADINĖ ANALIZĖ....5
3. MODELIO PARINKIMAS....7
3.1.Parametrinio ar semiparametrinio modelio rinkimas....7
4. MODELIO REZULTATAI....11
4.1. Galutinis modelis....11
4.2. Modelio vizualizavimas....11
5. IŠVADOS....13
ŠALTINIAI....14

ĮVADAS

Tikslas:

Taikant išgyvenamumo analizės regresiją ištirti storosios žarnos vėžio pasikartotinumą.

Uždaviniai:

1. Nuskaityti duomenis ir paruošti juos analizei;

2. Atlikti pradinę analizę;

3. Patikrinti, ar tinkamas parametrinis išgyvenamumo analizės modelis;

4. Sudaryti išgyvenamumo analizės regresijos modelį.

1. DUOMENYS

1.1.Duomenų aprašymas

Laboratoriniame darbe nagrinėsime duomenis apie storosiosios žarnos vėžio pasikartojimą pacientams po auglio šalinimo operacijos, paimtus iš „Kaggle“ (Real Colorectal Cancer Datasets | Kaggle). Duomenis sudaro:

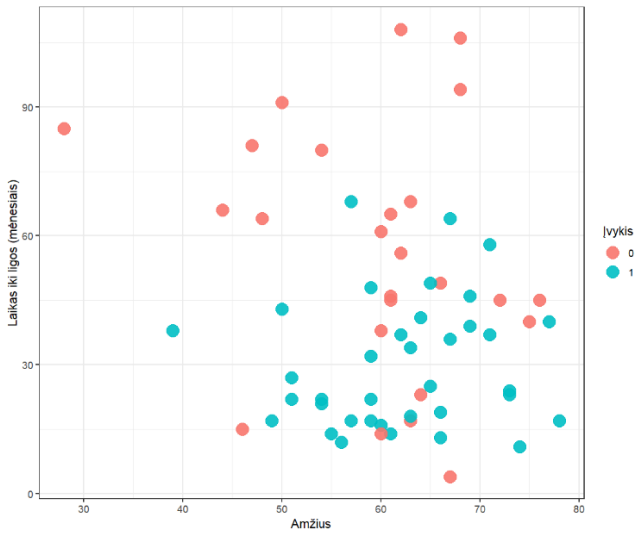
- DFS – laikas mėnesiais iki ligos pasikartojimo;
- Event – dvejetainis kintamasis, kuris nurodo ar liga pasikartojo (0 – ne, 1 – taip);
- Age – amžius, kai buvo nustatytas vėžys;
- Įvairių genų raiškos lygis.

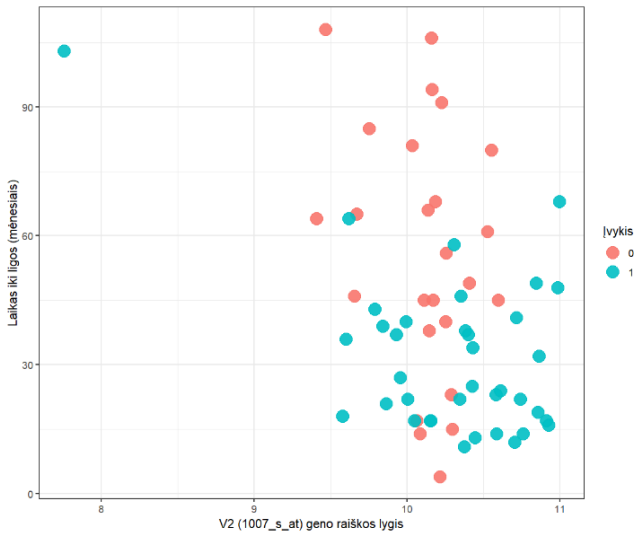
Priklausomas kintamasis – *DFS (laikas iki vėžio sugrįžimo)*. Tyrime naudosime reikšmingumo lygmenį $\alpha = 0{,}05$.

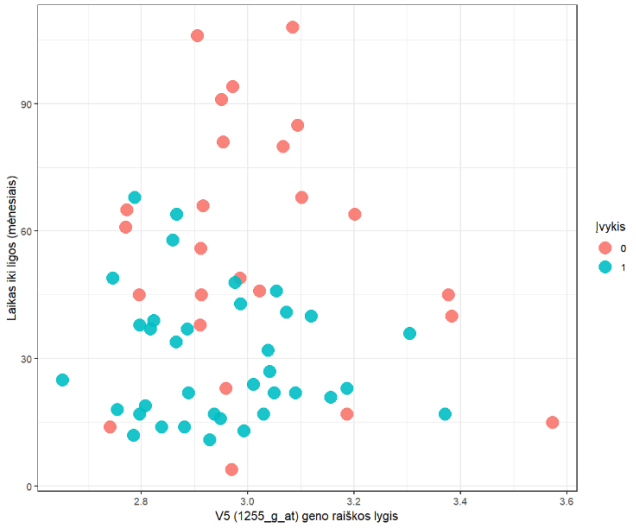
Pastaba: duomenyse buvo apie 2000 skirtingų genų informacijos, tačiau mes analizei atsitiktinai atrinkome 10 genų.

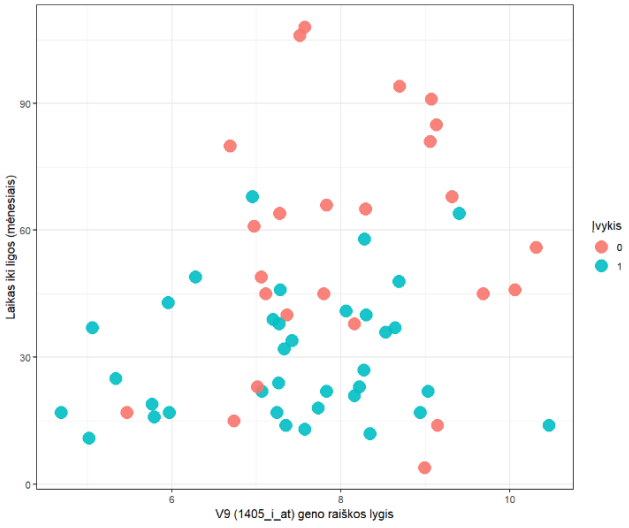
2. PRADINĖ ANALIZĖ

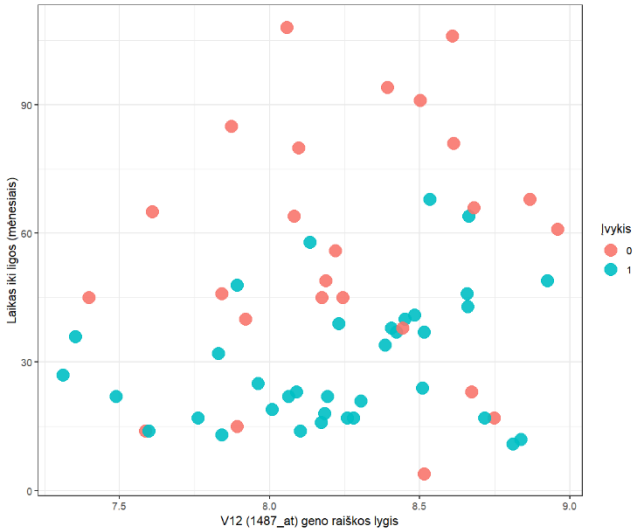
Pirmiausia nuskaitome reikalingus stulpelius. Tada braižome sklaidos diagramas, pažiūrime kaip yra pasiskirstę kintamieji pagal priklausomą kintamąjį.

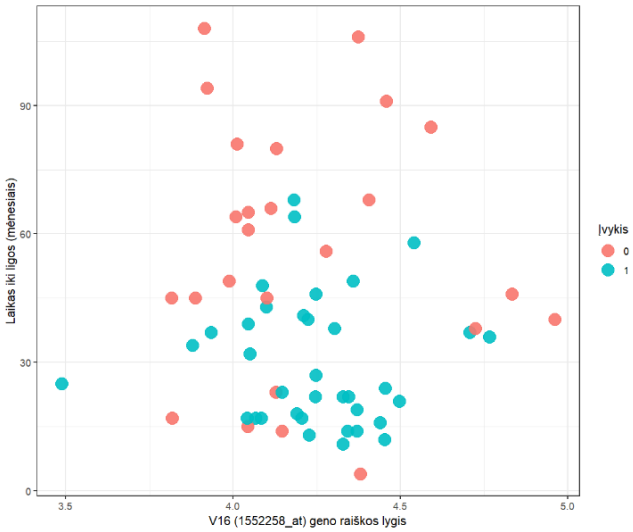


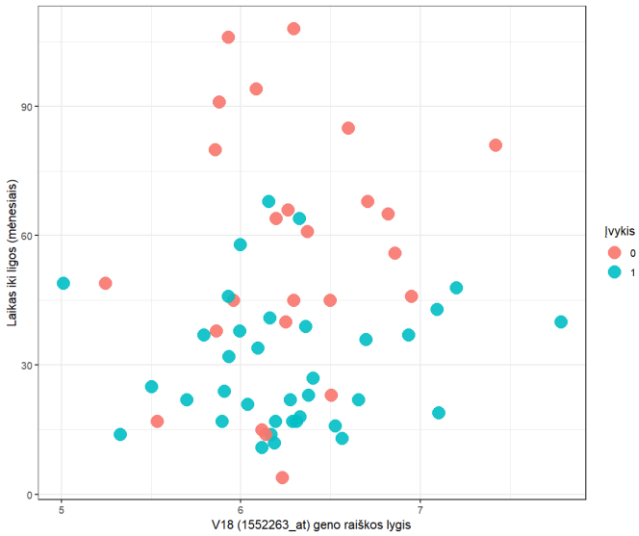


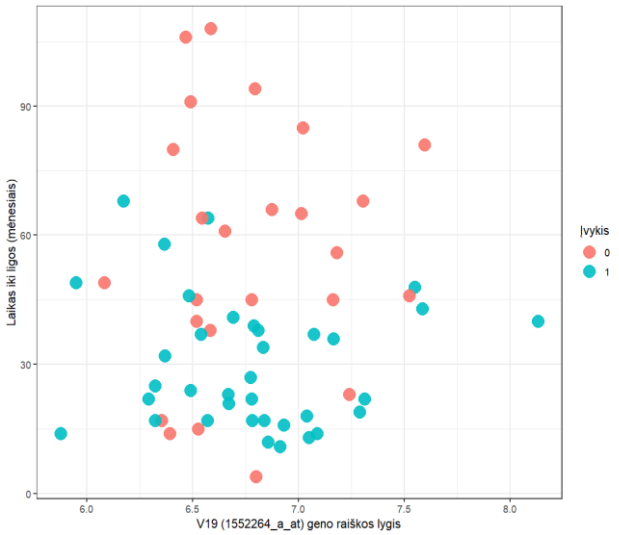

Laikas iki ligos (mėnesiais)
V5 (1255_g_at) geno raiškos lygis
Įvykis
0
1

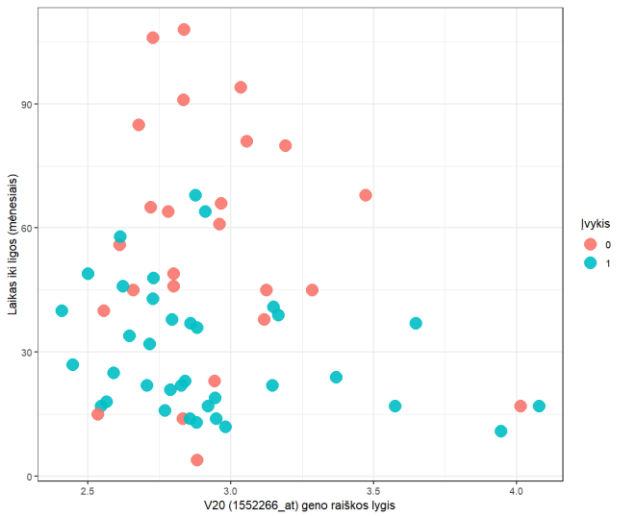

Laikas iki ligos (mėnesiais)
V9 (1405_i_at) geno raiškos lygis
Įvykis
0
1


Laikas iki ligos (mėnesiais)
V12 (1487_at) geno raiškos lygis
Įvykis
0
1


Laikas iki ligos (mėnesiais)
V16 (1552258_at) geno raiškos lygis
Įvykis
0
1


Laikas iki ligos (mėnesiais)
V18 (1552263_at) geno raiškos lygis
Įvykis
0
1


Laikas iki ligos (mėnesiais)
V19 (1552264_a_at) geno raiškos lygis
Įvykis
0
1



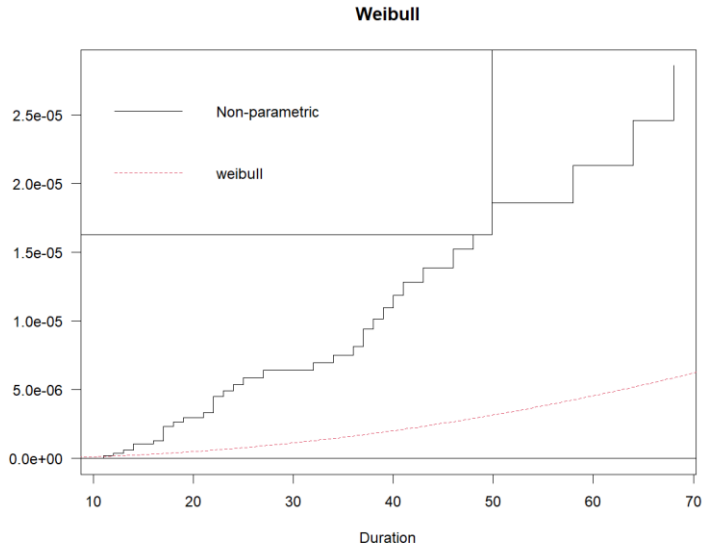
1 pav. Kovariančių sklaidos diagramos.

Iš sklaidos diagramų matome, kad turime vieną išskirtį V2 kovariantėje, ją šaliname.

3. MODELIO PARINKIMAS

3.1.Parametrinio ar semiparametrinio modelio rinkimas

Pirma tikriname, ar parametrinis modelis yra tinkamas šiam uždaviniui. Tam atliekame parametrinio ir semiparametrinio modelių įverčių palyginimą.



2 pav. Įverčių palyginimas

Iš grafiko matome, kad parametrinio ir semiparametrinio skirstinių įverčiai nesutampa, tikriname, kuris iš tų modelių yra geresnis. Tam naudojame AIC.

```
AIC(fit.cr)
] 252.8865
AIC(fit.wc)
] 365.1683
```

1 lentelė. AIC.

Semiparametrinio modelio AIC mažesnis, todėl toliau naudosime jį.

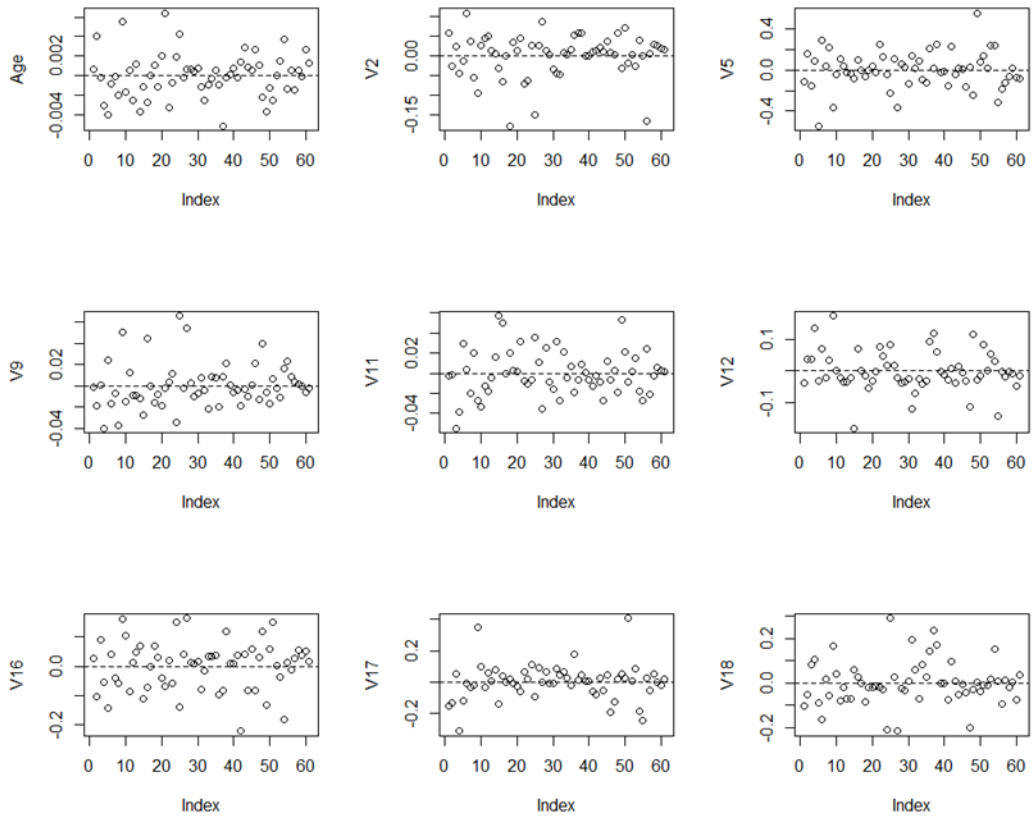
Toliau tikriname, ar modelis tenkina proporcingosios rizikos prielaidą.

	chisq	df	p
Age	2.87018	1	0.09
V2	0.10026	1	0.75
V5	0.41816	1	0.52
V9	1.11124	1	0.29
V11	0.69411	1	0.40
V12	0.09467	1	0.76
V16	0.09863	1	0.75
V17	0.00252	1	0.96
V18	0.73171	1	0.39
V19	0.03487	1	0.85
V20	2.55721	1	0.11
GLOBAL	14.08634	11	0.23

2 lentelė. Rizikos prielaidą.

Gauname, kad p-reikšmės prie visų kovariančių yra didesnės už reikšmingumo lygmenį, todėl nulinės hipotezės neatmetame.

Toliau tikriname išskirtis. Gauta išskirčių riba yra 0.256, todėl tikriname, ar yra stebėjimų, kurie viršija šią reikšmę. Iš grafikų matome, kad yra stebėjimų, kuriuos reikia pašalinti - iš viso 6 stebėjimai.



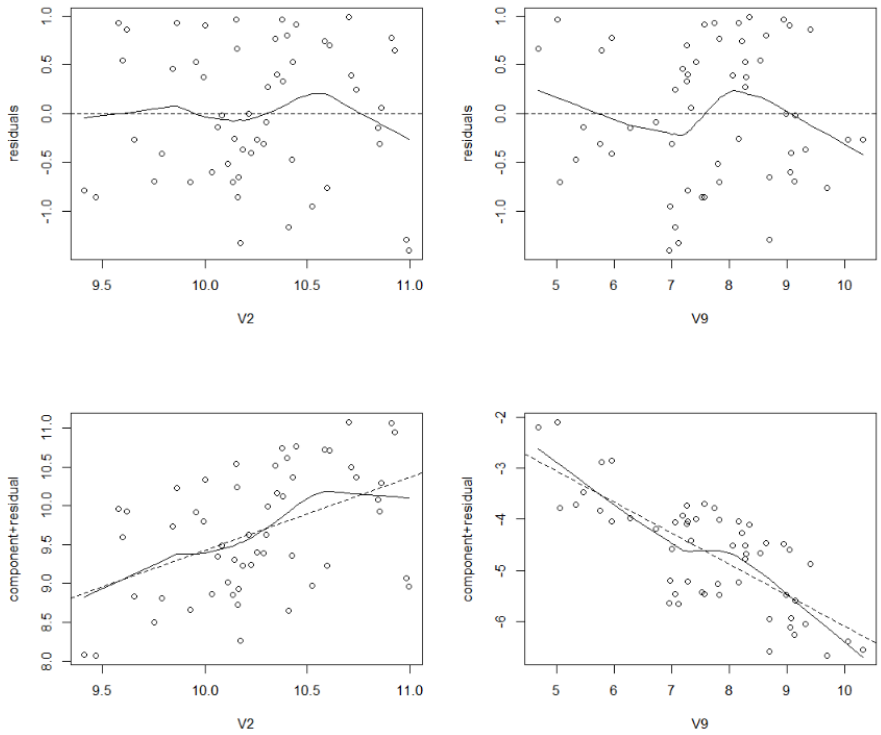
3 pav. Išskirčių tikrinimas

Su duomenimis be išskirčių dar vėl taikome pažingsninę regresiją, galiausiai modelyje lieka keturios reikšmingos kovariantės: V2, V9, V12, V19.

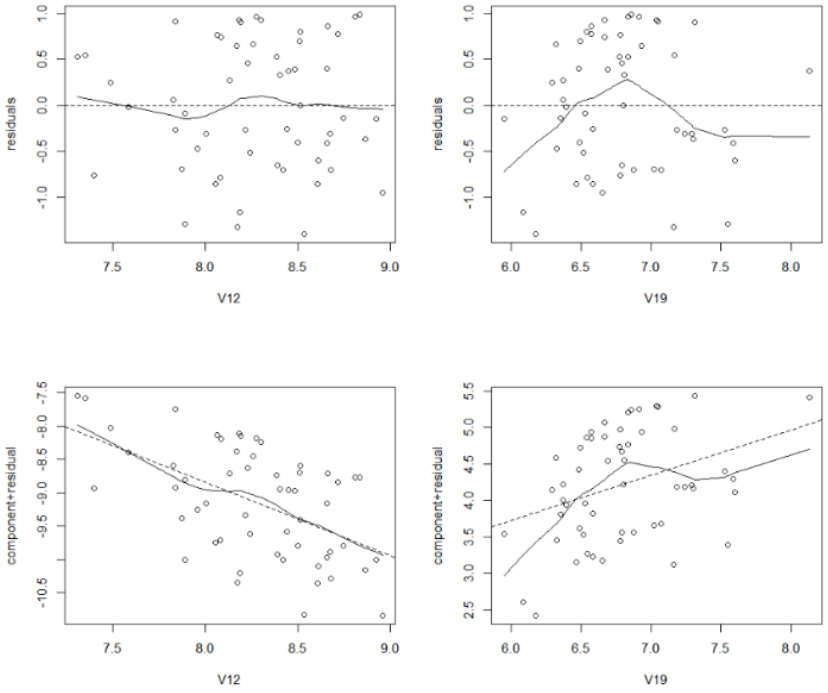
	coef	exp(coef)	se(coef)	z	p
V2	1.1506	3.1602	0.4446	2.588	0.00965
V9	-0.6946	0.4993	0.1691	-4.107	4.01e-05
V12	-1.3059	0.2709	0.4980	-2.622	0.00874
V16	1.2371	3.4457	0.6642	1.862	0.06254
V18	-1.2671	0.2817	0.8553	-1.481	0.13850
V19	2.0740	7.9569	1.1096	1.869	0.06161

3 lentelė. Kovariančių reikšmingumas ir koeficientų įverčiai.

Tuomet tikriname tiesiškumą – braižome martingalų liekanų ir kovariančių reikšmių sklaidos diagramas.



4 pav. martingalų liekanų ir kovariančių reikšmių sklaidos diagrama.



5 pav. martingalų liekanų ir kovariančių reikšmių sklaidos diagrama.

4. MODELIO REZULTATAI

4.1. Galutinis modelis

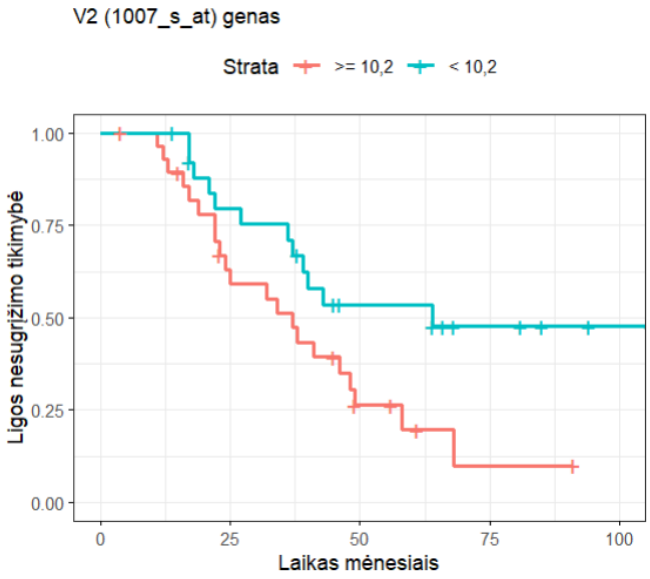
	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
V2	0.9426	2.5666	0.4032	2.338	0.019406	*
V9	-0.6096	0.5436	0.1612	-3.783	0.000155	***
V12	-1.1041	0.3315	0.4921	-2.244	0.024860	*
V19	0.6202	1.8593	0.3750	1.654	0.098121	.

4 lentelė. Galutinio modelio įverčiai.

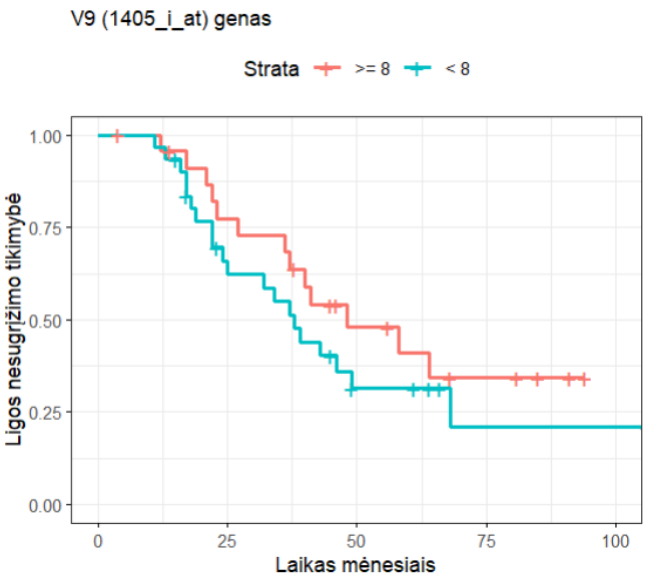
Iš lentelės matome, kad padidėjus V2 arba V19 geno reikšmėms vienetu, rizika, kad liga pasikartos, padidėja 2,57 karto ir 1,86 karto atitinkamai. Padidėjus V9 arba V12 vienetu, ligos rizika sumažėja 0,46 ir 0,67 karto atitinkamai.

4.2. Modelio vizualizavimas

Kad galėtume vizualiai pateikti rezultatus, visus kiekybinius kintamuosius pasiverčiame kategoriniais, dalindami juos maždaug per pusę pagal įgyjamas reikšmes.

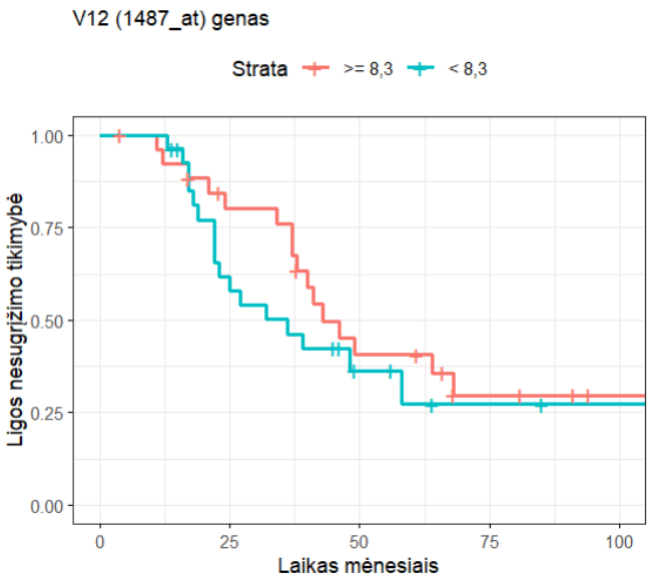


6 pav. Homogeniškumo hipotezės tikrinimas.

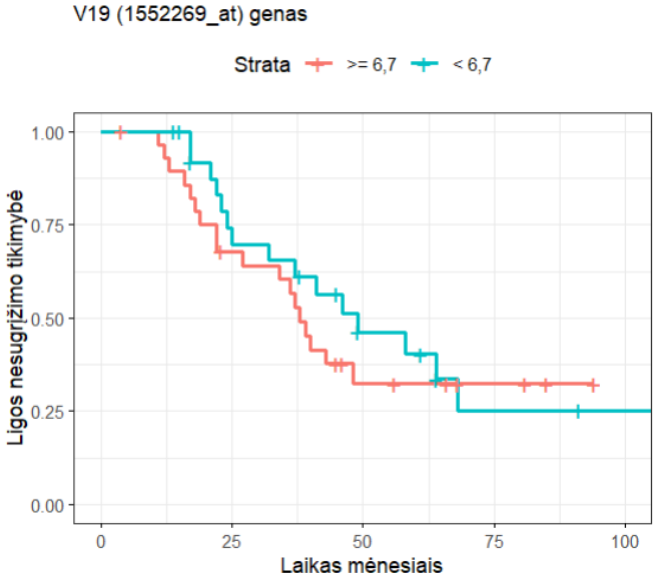


7 pav. Homogeniškumo hipotezės tikrinimas.

Matome, kad didesnis geno 1007_s_at (V2) kiekis ligos nepasikartojimo tikimybę mažina labiau, o praėjus 50 mėnesių po ligos, nepasikartojimo tikimybė sumažėja net iki 0.25. Jei geno 1007_s_at kiekis yra mažesnis nei 10.2, tai maždaug ties 60 mėnesiu tikimybė nukrenta mažiau 0.5, tačiau bėgant laikui išlieka tokia pati. Geno 1405_i_at (V9) ligos nesugrįžimo tikimybė didesnė, kai geno kiekis viršija 8. Praėjus daugiau nei 60 mėn. tikimybė nukrenta iki maždaug 0.3, o jei geno kiekis mažiau už 8 – lieka apie 0.2.



8 pav. Homogeniškumo hipotezės tikrinimas.



9 pav. Homogeniškumo hipotezės tikrinimas.

1487_at (V12) geno kiekis, mažesnis nei 8.3 pirmais 50 mėnesių labiau mažina ligos nepasikartojimo tikimybę, tačiau vėliau maždaug susilygina su tais pacientais, kurių šio geno kiekis ne mažesnis nei 8.3. Visgi, nežymiai didesnė tikimybė nesusirgti pakartotinai išlieka didesnį kiekį šio geno turintiems pacientams. 1552269_at (V19) geno kiekis, mažesnis už 6.7 pirmais mėnesiais ligos pasikartojimo tikimybę mažina mažiau, tačiau praėjus daugiau nei 60 mėnesių – didesnį kiekį nei 6.7 turintys pacientai turi didesnę tikimybę, kad liga nepasikartos.

5. IŠVADOS

Pritaikius semiparametrinį kokso regresijos modelį pastebėta, kad įvairių genų raiška yra reikšminga prognozuojant, ar liga pasikartos ateityje, ar ne. Atlikus analizę nustatyta, kad reikšmingi genai yra 1007_s_at (V2), 1405_i_at (V9), 1487_at (V12) ir 1552269_at (V19). Svarbu paminėti, kad analizėje buvo naudojami tik 10 atsitiktinai atrinktų genų.

ŠALTINIAI

[1] „Kaggle“ tinklapis. Tema: Real Colorectal Cancer Datasets. Prieiga per internetą: https://www.kaggle.com/datasets/amandam1/colorectal-cancer-patients?select=Colorectal+Cancer+Patient+Data.csv