VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Regresinė analizė

1 laboratorinis darbas

Atliko: 3 kurso 2 grupės studentai:

Matas Amšiejus

Sandra Macijauskaitė

Salvija Račkauskaitė

Darbo vadovė: doc. dr. Rūta Levulienė

Vilnius, 2022

TURINYS

ĮVADAS

Tikslas:

Taikant binarinės regresijos modelį ištirti kaip galimybė išgyventi „Titaniko“ katastrofą priklauso nuo įvairių parametrų.

Uždaviniai:

1. Nuskaityti duomenis ir paruošti juos analizei;

2. Ištirti sąryšius tarp priklausomo kintamojo ir kovariančių;

3. Taikyti logit ir probit modelius;

4. Išrinkti geriausią modelį.

1. DUOMENYS

Duomenų rinkinį pasirinkome iš viešai prieinamo duomenų šaltinio „Kaggle“.

1.1.Duomenų aprašymas

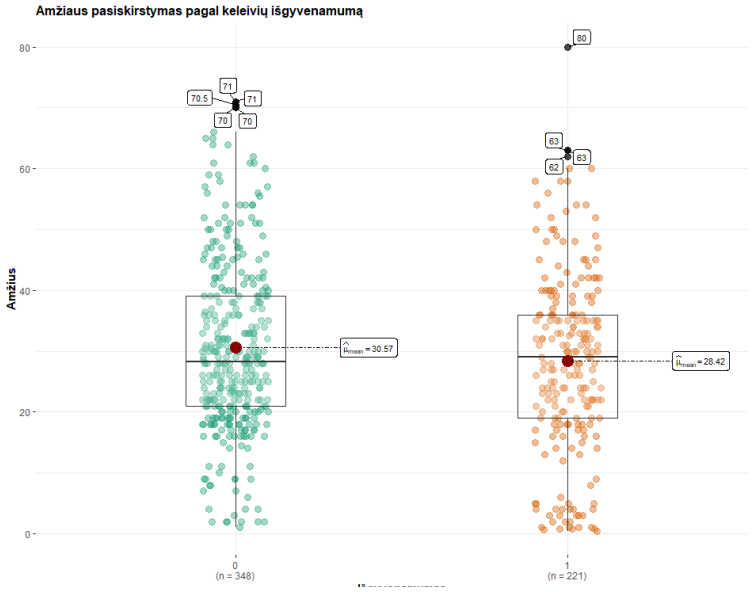
Laboratoriniame darbe naudosime duomenis apie „Titaniko“ katastrofą.

- Survived – ar keleivis išgyveno (0 – ne, 1 - taip);
- Pclass – bilieto klasė;
- Sex – lytis;
- Age – amžius;
- SibSp – brolių ir sesių / sutuoktinių skaičius kelionėje;
- Parch – tėvų / vaikų skaičius kelionėje;
- Ticket – bilieto numeris;
- Fare – bilieto kaina;
- Cabin – kajutės numeris;
- Embarked – įlaipinimo vieta;
- Name – keleivio vardas, pavardė.

Priklausomas kintamasis – *survived*. Tyrime nenaudosime kintamųjų *ticket, cabin, name*.

2. SĄRYŠIAI TARP IŠGYVENAMUMO IR KOVARIANČIŲ

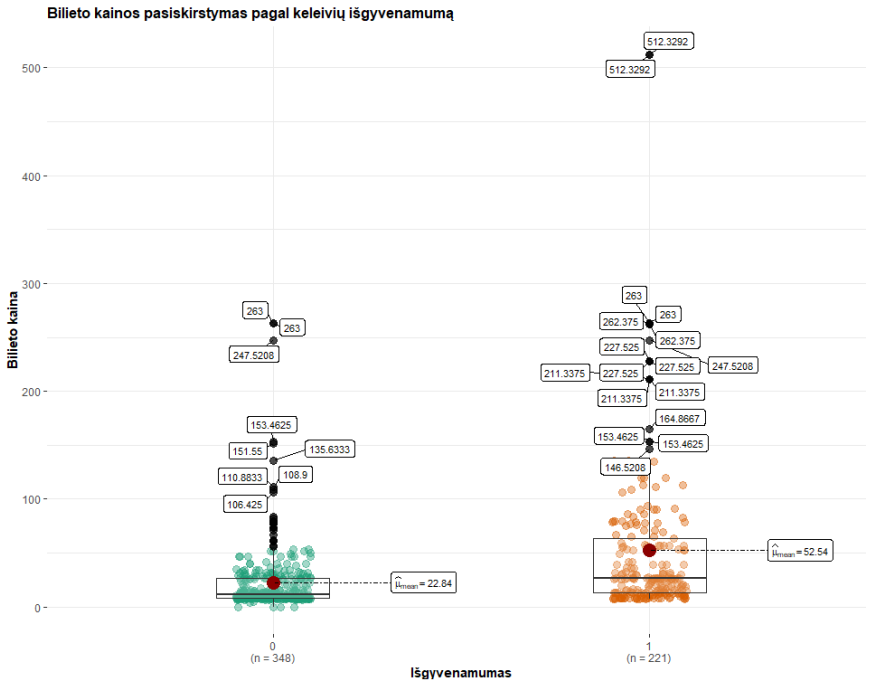
Pirmiausia norėjome ištirti keleivių amžiaus pasiskirstymą atskyrus išgyvenusius ir neišgyvenusius keleivius.



1 pav. Amžiaus pasiskirstymas pagal keleivių išgyvenamumą

Iš stačiakampių diagramų matome, kad vizualiai nėra didelių skirtumų tarp neišgyvenusių ir išgyvenusių amžiaus, tačiau daugiau pastarųjų stebėjimų yra susitelkę prie mažesnio amžiaus.

Toliau patikrinome bilieto kainos pasiskirstymą lyginant abi grupes.



2 pav. Bilieto kainos pasiskirstymas pagal keleivių išgyvenamumą

Matome dideles išskirtis išgyvenusiųjų keleivių grupėje, tačiau patikrinus, ar modelis pagerėja išmetus išskirtis, nustatėme, kad skirtumo nėra.

Taip pat patikrinome, ar visose grupėse yra pakankamai stebėjimų, kad galėtume atlikti regresinę analizę.

Pclass	Survived 0	1
1	0.3673469	0.6326531
2	0.5328467	0.4671533
3	0.7754386	0.2245614

Sex	Survived 0	1
female	0.2535885	0.7464115
male	0.8194444	0.1805556

Parch_c	Survived 0	1
>2	0.7692308	0.2307692
0	0.6626506	0.3373494
1	0.4555556	0.5444444
2	0.4313725	0.5686275

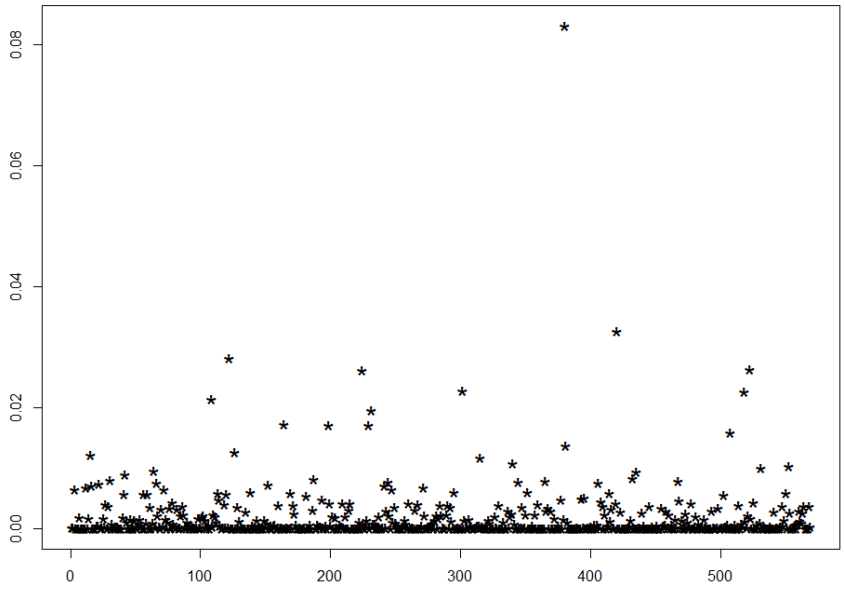
SibSp_c	Survived 0	1
>2	0.7812500	0.2187500
0	0.6612022	0.3387978
1	0.4671053	0.5328947
2	0.5263158	0.4736842

```
         Survived
Embarked         0         1
       C 0.3962264 0.6037736
       Q 0.7500000 0.2500000
       S 0.6560364 0.3439636
```

Pagal visas kovariantes matome, kad duomenų yra pakankamai. Nors pagal lytį išgyvenusių vyrų yra tik 18 %, iš viso imtyje yra 65 stebėjimai.

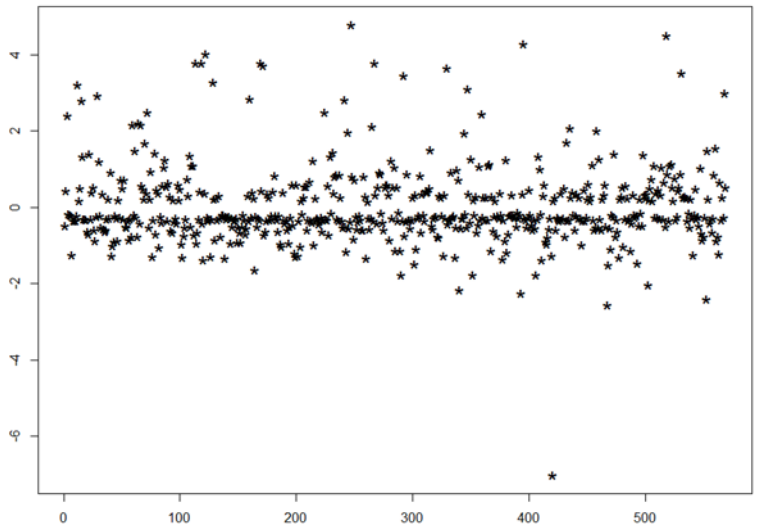
3. REGRESIJOS TAIKYMAS NAUDOJANT LOGIT MODELĮ

Pirma sukuriame modelį su visomis kovariantėmis ir tikriname išskirtis. Pagal Kuką matome, kad išskirčių nėra.



3 pav. Išskirtys pagal Cook

Pagal standartizuotas liekanas matome vieną stipriai išsiskiriančią reikšmę. Pasidomėjus sužinome, kad išskirtis yra 2 metų amžiaus mergaitė iš pirmos klasės. Tai buvo vienintelis neišgyvenęs vaikas iš 1 ir 2 klasių, todėl stebėjimą šaliname.



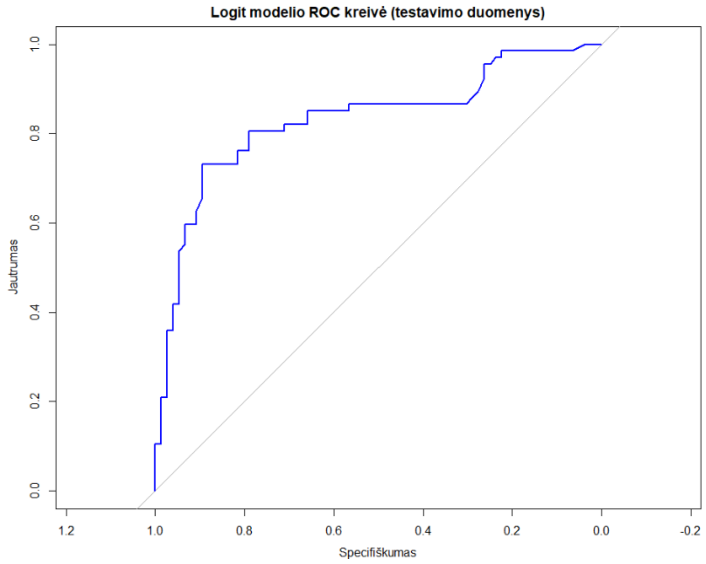
4 pav. Išskirtys pagal Pearson

Toliau tikriname, kurios kovariantės yra reikšmingos. Panaudojus summary funkciją matome, kad yra nereikšmingų kovariančių (tyrimo reikšmingumo lygmuo $\alpha = 0{,}05$). Taikome pažingsninę regresiją. Gauname, kad reikšmingos kovariantės yra Pclass, Sex, Age, SibSp_c.

Tikriname, kaip tiksliai modelis klasifikuoja keleivius su testiniais duomenimis. Iš lentelės ir grafiko matome, kad modelis nuspėja keleivių likimą pakankamai tiksliai.

1 lentelė. Logit modelio klasifikavimo lentelė

	response	
predicted	0	1
0	0.764	0.236
1	0.148	0.852



5 pav. ROC kreivė

3.1. Interpretacija

(Intercept)	Pclass2	Pclass3	Sexmale	Age	SibSp_c0	SibSp_c1	SibSp_c2
14.22108149	0.23472173	0.06327085	0.06046461	0.95345907	5.66780131	5.65828592	2.90385139

1 pav. ~~Koeficientai~~ gal. santykis

Matome, kad jei keleivis yra įsigijęs antros klasės bilietą, jo išgyvenimo galimybė sumažėja maždaug 76,5 %, jei trečios – galimybė sumažėja 93,6 % (lyginant su pirmos klasės bilietą įsigijusiais keleiviais). Jeigu keleivio lytis yra vyras, tai lyginant su moterimis, jo išgyvenimo galimybė sumažėja 94 %. Keleivio amžiui padidėjus vienetu, jo išgyvenimo galimybė sumažėja 4,7 %. Jei keleivis neturi brolių / sesių ir sutuoktinio arba turi tik vieną, tai jo galimybė išgyventi padidėja maždaug 5,6 karto. Jei keleivis turi 2 artimuosius, tai jo galimybė išgyventi padidėja 2,9 karto.

4. REGRESINĖ ANALIZĖ NAUDOJANT PROBIT MODELĮ

Sudarius modelį gavome vienodą išskirtį. Ją pašalinus gavome tas pačias reikšmingas kovariantes Pclass, Sex, Age, SibSp_c. Šiam modeliui sudarę klasifikavimo lentelę matome, kad rezultatai nesiskiria nuo logit modelio.

2 lentelė. Probit modelio klasifikavimo lentelė

	response	
predicted	0	1
0	0.764	0.236
1	0.148	0.852

(Intercept)	Pclass2	Pclass3	Sexmale	Age	SibSp_c0	SibSp_c1	SibSp_c2
0.366547271	-0.202066922	-0.368571382	-0.390041169	-0.006346748	0.236803696	0.233278710	0.141102151

2 pav. Koeficientai

Pagal ženklus iš lentelės matome, kad probit ir logit modelių kovariančių kryptys sutampa.

IŠVADOS

Abiejuose modeliuose gavome vieną išskirtį, ją išanalizavome ir pašalinome. Mūsų modeliuose reikšmingos kovariantės – keleivio amžius, lytis, bilieto klasė, brolių / sesių ir sutuoktinių skaičius. Nustatėme, kad didžiausią išgyvenimo galimybę turėjo moterys, pirmos klasės keleiviai, vaikai bei asmenys, kurie neturėjo daug artimųjų. Gauname, kad abu modeliai yra panašaus tikslumo. Jeigu nenaudosime šių modelių ateities atvejų prognozei, labiau verta rinktis logit modelį dėl aiškesnės interpretacijos.

ŠALTINIAI

[1] „Kaggle“ tinklapis. Tema: Titanic. Prieiga per internetą: https://www.kaggle.com/prkukunoor/TitanicDataset