VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Regresinė analizė

3 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:

Naudojant kvantilių regresiją įvertinti, kaip priklauso automobilių kainos nuo įvairių transporto priemonę apibūdinančių parametrų.

Uždaviniai:

1. Nuskaityti duomenis ir paruošti juos analizei;
2. Palyginti tiesinę regresiją ir kvantilių regresiją;
3. Nustatyti, kaip keičiasi koeficientai, keičiant kvantilius.

1. DUOMENYS

Duomenų rinkinį pasirinkome iš viešai prieinamo duomenų šaltinio „Data.world“.

Laboratoriniame darbe nagrinėsime duomenis apie automobilius su tokiais kintamaisiais:

- Price – automobilio kaina;
- VehicleType – transporto priemonės tipas;
- YearOfRegistration – metai, kuriais automobilis buvo pirmą kartą užregistruotas;
- Gearbox – pavarų dėžės tipas;
- PowerPS – automobilio galia (PS matavimo skalėje);
- Kilometer – automobilio nuvažiuoti kilometrai;
- FuelType – kuro tipas;

Tyrime naudosime reikšmingumo lygmenį $\alpha = 0,05$.

2. DUOMENŲ PARUOŠIMAS IR PRADINĖ ANALIZĖ

Pirmiausia atsirenkame tik mus dominančius stulpelius, kurie yra susiję su automobiliais. Tyrime nenaudosime kintamųjų, kurie apibūdina skelbimų svetainės ypatumus (skelbimo įvedimo data, kada paskutinį kartą peržiūrėtas skelbimas, nuotraukų skaičius ir t.t.). Papildomai susikuriame stulpelį „amžius“, kurį naudosime vietoj automobilio registracijos metų.

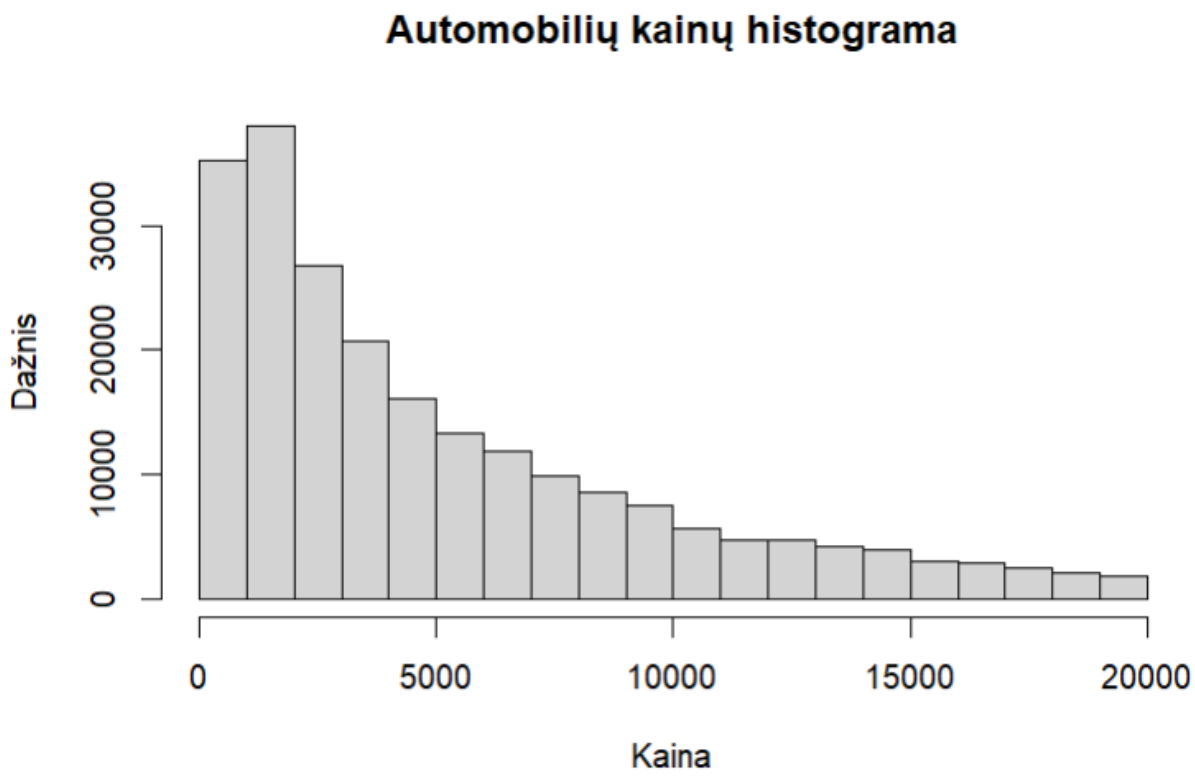
Toliau pasiliekame tik tuos skelbimus, kuriuose ne ieškomas, o parduodamas automobilis ir skelbimas yra įkeltas privataus asmens. Taip pat atmetame skelbimus, kuriuose nurodytos nelogiškos kainos, galios reikšmės ir pašaliname įrašus, kurie turi praleistų stebėjimų.

Tyrimo priklausomas kintamasis – automobilio kaina, todėl pažiūrime jo reikšmių pasiskirstymą.

```
quantile(autos1$price, probs = c(0, 0.5, 0.75, 0.95, 1))
  0%   50%   75%   95%  100%
 101  3499  7950 20499 1234566
```

1 lentelė. Priklausomo kintamojo reikšmių pasiskirstymas

Matome, kad 95% automobilių kainų yra apie 20,5 tūkst., todėl, siekiant išvengti didelių išskirčių, tyrimui pasiliekame tik automobilius, kurių kaina neviršija 20 tūkst.



1 pav. Automobilių kainų histograma

Pradinį duomenų rinkinį sudaro beveik 300 tūkst. stebėjimų, todėl siekiant sumažinti programos vykdymosi laiką, tyrimui atsirinksime atsitiktinę 20 tūkst. dydžio imtį.

Patikriname, ar kategorinių kintamųjų grupės nėra per mažos ir tinka tolimesnei analizei:

```
> table(sample_autos$vehicleType)

    andere        bus     cabrio      coupe kleinwagen      kombi  limousine        suv
       102       1943       1368        988       4736       4179       5911        773
> table(sample_autos$gearbox)

automatik   manuell
     4243     15757
> table(sample_autos$fuelType)

 andere  benzin     cng  diesel elektro  hybrid     lpg
      2   13166      37    6472       1      15     307
> table(sample_autos$notRepairedDamage)

   ja  nein
 2086 17914
```

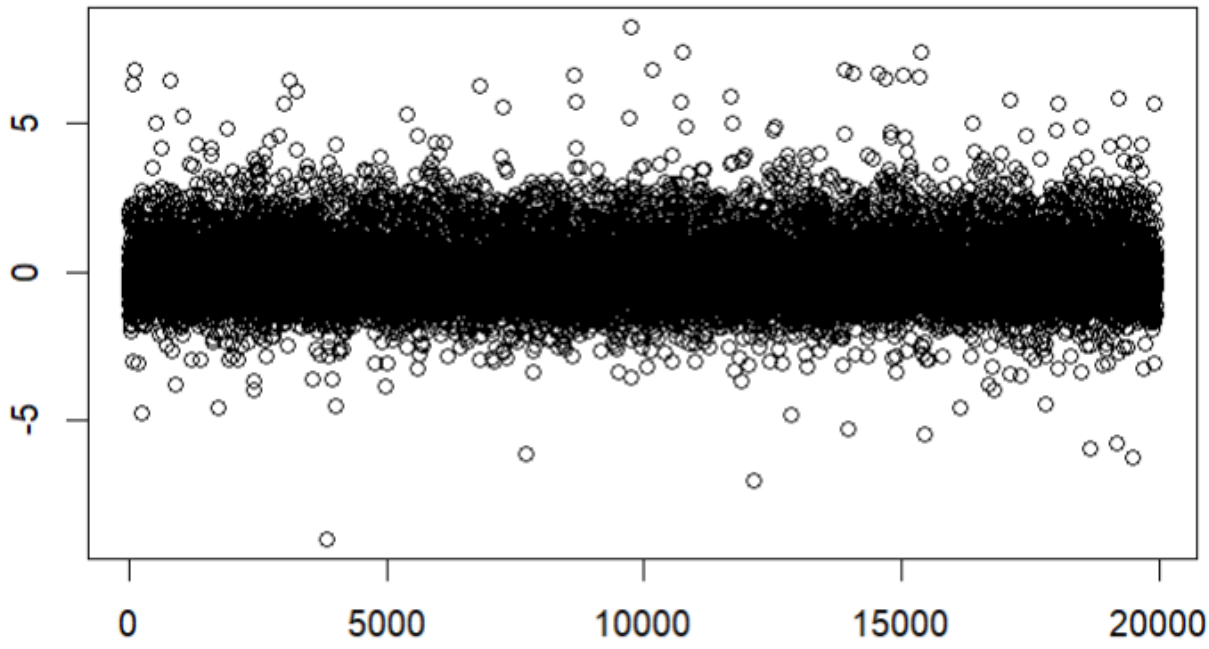
2 lentelė. Kategorinių kintamųjų grupių dydžiai

Matome, kad kuro tipai, tokie kaip „cng“ (suspaustos gamtinės dujos), „elektro“ (elektra varomos), „hybrid“ (elektra ir dar vienu kuru varomos) ir „andere“ (kita) yra labai nepopuliarūs, todėl gali padidinti koeficientų dispersiją. Dėl šios priežasties automobilių su aukščiau išvardintais kuro tipais neimsime.

3. KVANTILIŲ REGRESIJOS MODELIO KŪRIMAS

3.1.Modelio taikymas

Net ir pašalinus dalį stebėjimų, duomenų rinkinyje nemažai išskirčių. Tai matoma pritaikius tiesinės regresijos modelį iš standartizuotų liekanų grafiko:



2 pav. Standartizuotų liekanų grafikas

Dėl išsiskiriančių stebėjimų gausos automobilių kainų prognozavimui naudosime kvantilių regresiją, kuri nėra tokia jautri išskirtims.

Pasirenkame tris mus dominančius kvantilius: 0.25, 0.5 ir 0.75 ir kiekvienam iš jų taikome kvantilių regresiją su visomis kovariantėmis bei pridedame automobilio amžiaus ir nuvažiuotų kilometrų sąveiką, nes ji gali padėti atskirti seną kolekcinį automobilį nuo tiesiog seno. Senas kolekcinis turės daug metų, tačiau jo nuvažiuotas atstumas bus mažas (paprastai jais važinėjama nedaug). Jei atstumas didelis, tai automobilis yra labiau darbinis, kasdieninis, dėl to ir kilometrų nuvažiavęs daugiau.

tau: [1] 0.25

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	6039.81235	465.32467	12.97978	0.00000
vehicleTypebus	4447.24423	459.27292	9.68323	0.00000
vehicleTypecabrio	3017.32018	472.45987	6.38641	0.00000
vehicleTypecoupe	3038.43043	488.97644	6.21386	0.00000
vehicleTypekleinwagen	1125.64540	449.47110	2.50438	0.01227
vehicleTypekombi	5110.81225	462.87790	11.04138	0.00000
vehicleTypelimousine	3885.05398	451.07147	8.61295	0.00000
vehicleTypesuv	6613.40608	501.82776	13.17864	0.00000
gearboxmanuell	-138.46989	36.26803	-3.81796	0.00013
powerPS	23.84332	0.41717	57.15477	0.00000
kilometer	-0.04864	0.00084	-57.71582	0.00000
fuelTypediesel	558.14985	28.84608	19.34924	0.00000
fuelTypelpg	-406.68651	51.84022	-7.84500	0.00000
notRepairedDamagenein	1203.23293	13.34610	90.15617	0.00000
amzius	-372.04832	35.47009	-10.48907	0.00000
vehicleTypebus:amzius	-323.98549	35.45853	-9.13702	0.00000
vehicleTypecabrio:amzius	-152.58139	36.19702	-4.21530	0.00003
vehicleTypecoupe:amzius	-192.42824	35.83791	-5.36940	0.00000
vehicleTypekleinwagen:amzius	-77.52849	35.02320	-2.21363	0.02687
vehicleTypekombi:amzius	-395.20844	35.61100	-11.09793	0.00000
vehicleTypelimousine:amzius	-265.02393	35.03479	-7.56459	0.00000
vehicleTypesuv:amzius	-398.78715	39.82989	-10.01226	0.00000
kilometer:amzius	0.00232	0.00004	53.38133	0.00000

3 lentelė. Pirmas modelis su 0,25 kvantiliu

tau: [1] 0.5

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	5346.74339	711.60418	7.51365	0.00000
vehicleTypebus	5995.65596	703.64315	8.52088	0.00000
vehicleTypecabrio	5187.78807	722.21343	7.18318	0.00000
vehicleTypecoupe	5658.42286	758.69147	7.45813	0.00000
vehicleTypekleinwagen	2813.83925	684.17248	4.11276	0.00004
vehicleTypekombi	7594.15143	692.65014	10.96391	0.00000
vehicleTypelimousine	5901.23575	687.40585	8.58479	0.00000
vehicleTypesuv	8385.82250	733.87413	11.42679	0.00000
gearboxmanuell	-234.46947	58.98239	-3.97525	0.00007
powerPS	34.37693	0.53883	63.79950	0.00000
kilometer	-0.05684	0.00115	-49.28579	0.00000
fuelTypediesel	998.35319	49.68287	20.09451	0.00000
fuelTypelpg	-555.87402	106.47636	-5.22063	0.00000
notRepairedDamagenein	1135.96160	43.94208	25.85134	0.00000
amzius	-219.19220	48.17488	-4.54993	0.00001
vehicleTypebus:amzius	-416.61796	45.53943	-9.14851	0.00000
vehicleTypecabrio:amzius	-264.91936	45.02657	-5.88362	0.00000
vehicleTypecoupe:amzius	-341.81940	47.70221	-7.16569	0.00000
vehicleTypekleinwagen:amzius	-182.07999	43.24653	-4.21028	0.00003
vehicleTypekombi:amzius	-567.14499	43.79727	-12.94932	0.00000
vehicleTypelimousine:amzius	-393.07597	43.55896	-9.02400	0.00000
vehicleTypesuv:amzius	-500.92785	49.76335	-10.06620	0.00000
kilometer:amzius	0.00200	0.00014	14.77084	0.00000

4 lentelė. Antras modelis su 0,5 kvantiliu

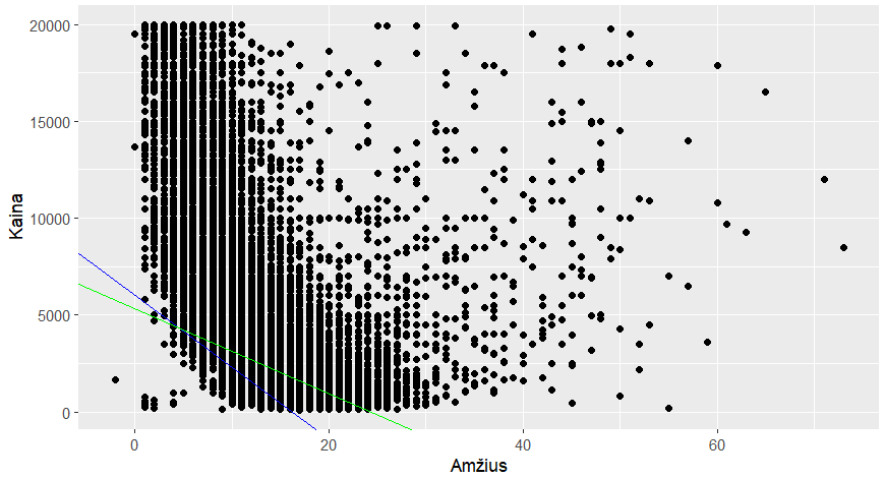
```
tau: [1] 0.75

Coefficients:
                              Value      Std. Error t value    Pr(>|t|)
(Intercept)                8684.23935   819.00475   10.60341    0.00000
vehicleTypebus             2980.81355   860.49687    3.46406    0.00053
vehicleTypecabrio          3341.37677   934.41667    3.57590    0.00035
vehicleTypecoupe           3323.05313   875.61915    3.79509    0.00015
vehicleTypekleinwagen       376.05211   828.60757    0.45384    0.64995
vehicleTypekombi           4959.58996   839.27691    5.90936    0.00000
vehicleTypelimousine       3008.65521   830.86496    3.62111    0.00029
vehicleTypesuv             5171.38499   916.88781    5.64015    0.00000
gearboxmanuell             -629.19647    89.20509   -7.05337    0.00000
powerPS                      43.51401     0.82943   52.46284    0.00000
kilometer                    -0.06189     0.00176  -35.12565    0.00000
fuelTypediesel             1798.45759    73.40541   24.50034    0.00000
fuelTypelpg                -279.85966   218.49556   -1.28085    0.20026
notRepairedDamagenein      1002.02499    62.42104   16.05268    0.00000
vehicleTypeandere:amzius    -94.06690    61.18695   -1.53737    0.12422
vehicleTypebus:amzius      -321.45951    39.22090   -8.19613    0.00000
vehicleTypecabrio:amzius   -237.89509    45.96982   -5.17503    0.00000
vehicleTypecoupe:amzius    -278.53809    42.46090   -6.55987    0.00000
vehicleTypekleinwagen:amzius -157.07778  34.68801   -4.52830    0.00001
vehicleTypekombi:amzius    -526.78438    36.63511  -14.37922    0.00000
vehicleTypelimousine:amzius -336.16714   35.16599   -9.55944    0.00000
vehicleTypesuv:amzius      -389.17928    50.85583   -7.65260    0.00000
kilometer:amzius              0.00046     0.00023    1.95459    0.05065
```

5 lentelė. Trečias modelis su 0,75 kvantiliu

Pirmais dviem atvejais matome, kad nereikšmingų kovariančių nėra. Trečiuoju atveju kovariantės „amžius“ p-reikšmė didesnė už reikšmingumo lygmenį, todėl ją iš modelio pašaliname (sąveikos reikšmingos, todėl jas paliekame).

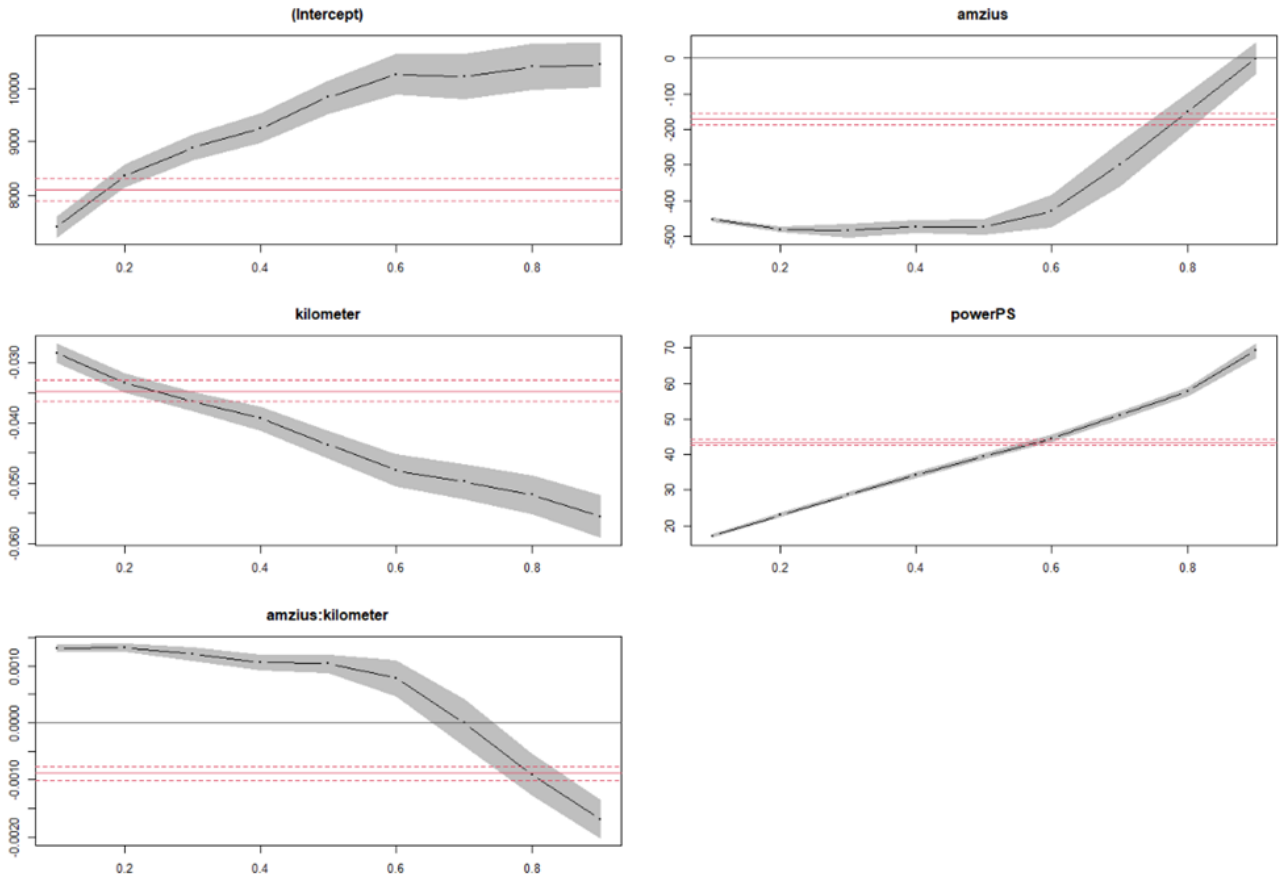
3.2.Rezultatų vizualizavimas



3 pav. 25% ir 50% lygmens kvantilių regresijos tiesės

Pirmame grafike (3 pav.) matome, kad pigiausių automobilių konstanta (β_0) yra didesnė nei vidutinių kainų automobilių. Didžiausia β prie amžiaus yra pigių automobilių (amžius jų kainą mažina labiausiai), tada vidutinių automobilių.

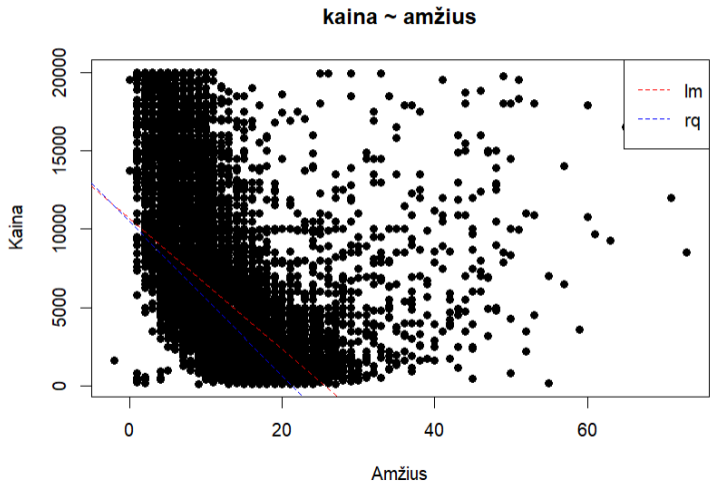
Toliau vizualizuojame, kaip keičiasi modelio koeficientai, keičiantis kvantilių reikšmėms:



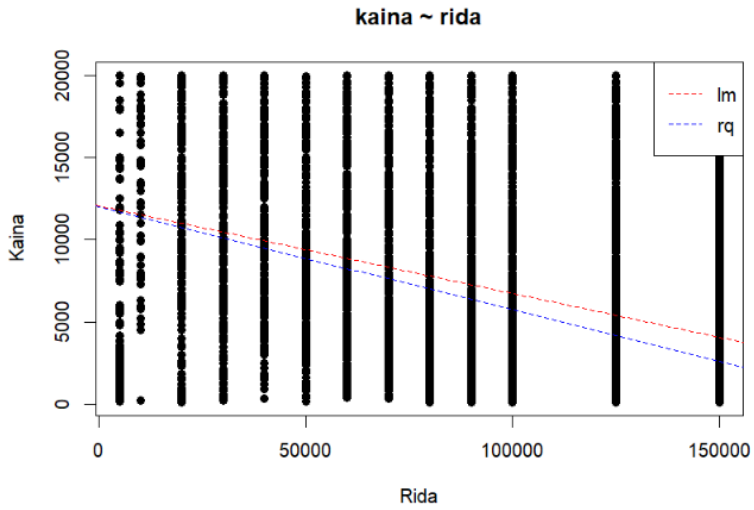
4 pav. Modelio koeficientai, keičiantis kvantilių reikšmėms

Matome tai, ko ir tikėjomės kuriant sąveiką: didėjant kainai β tampa vis labiau neigiama. Tai reiškia, kad brangesniems automobiliams amžiaus ir kainos sąveika yra labiau įtakinga negu pigesniems. Pavyzdžiui jei tirtume 10 % brangiausių automobilių, tai jų kaina nuo sąveikos kristų greičiausiai, nes potencialiai tai mažina automobilio istorinę, kolekcinę verte. Kadangi pigesni automobiliai nėra kolekciniai, jiems sąveikos koeficientas gana pastovus.

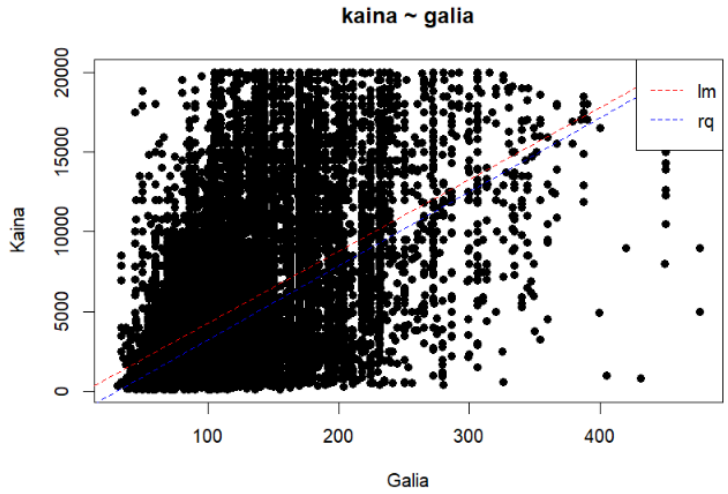
3.3.Tiesinė regresija prieš kvantilių regresiją



5 pav. Sklaidos diagrama, imama kaina ir amžius



6 pav. Sklaidos diagrama, imama kaina ir rida

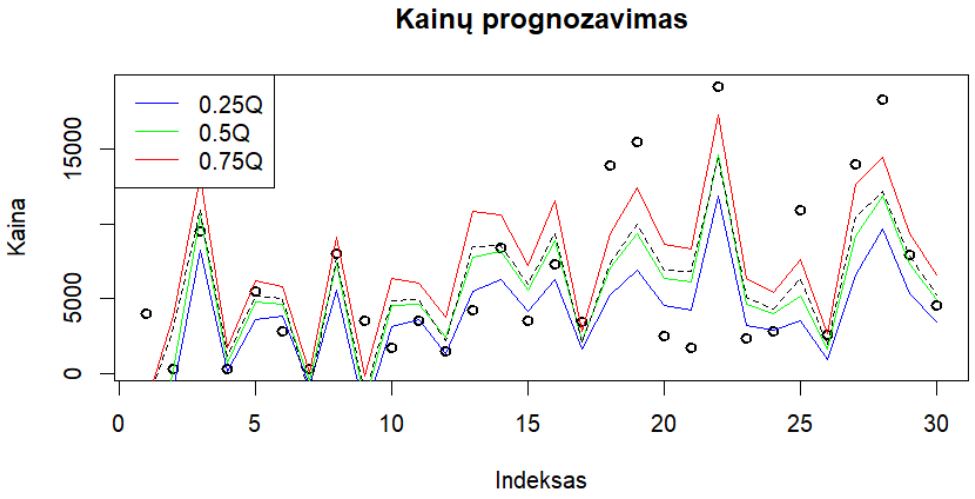


7 pav. Sklaidos diagrama, imama kaina ir galia

Matome, kad tiesinės regresijos linija visada yra aukščiau medianos regresijos, tačiau iš esmės abi tiesės yra šalia viena kitos ir labai panašios.

3.4.Modelio prognozė

Patikriname, kaip gerai mūsų modelis prognozuoja automobilių kainas.

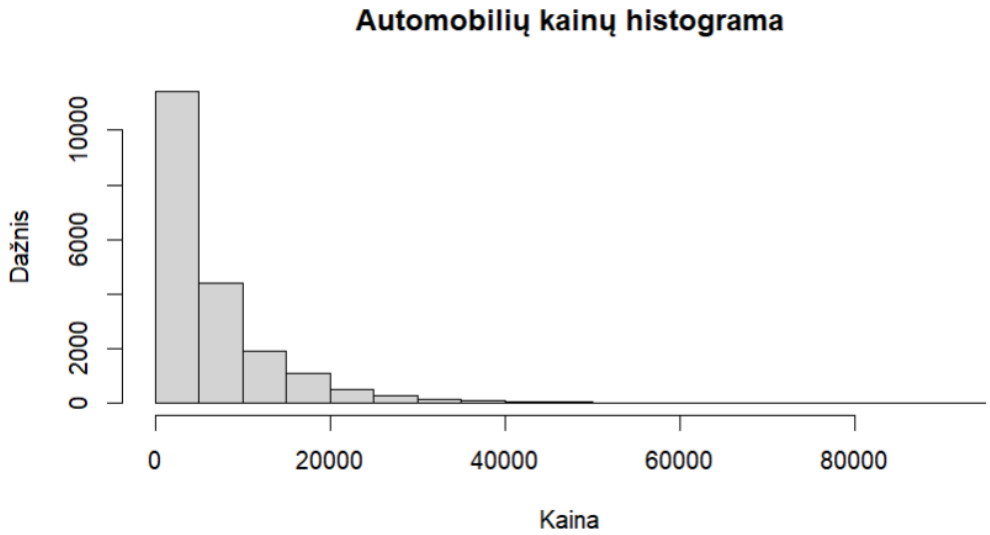


8 pav. Modelio kainų prognozė

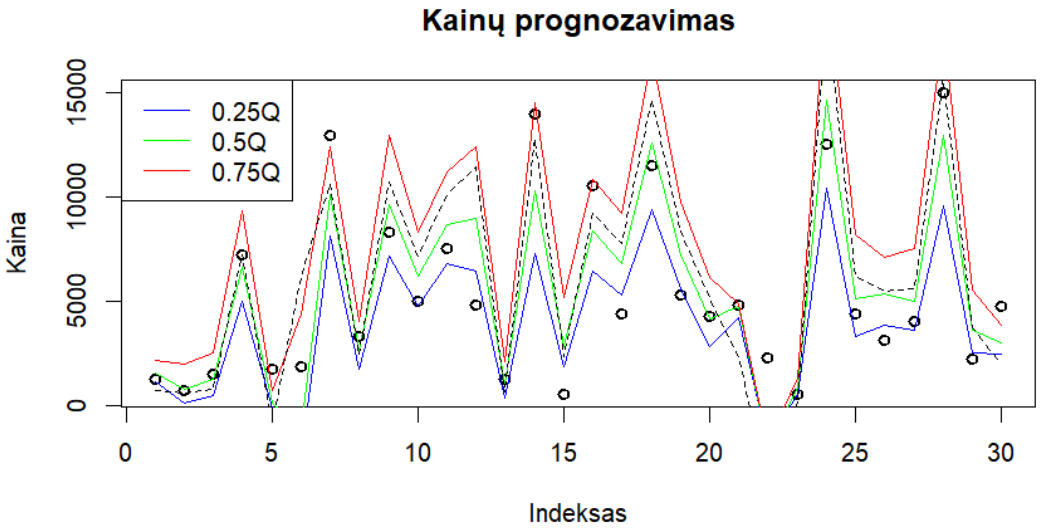
Matome, kad pigesnius automobilius geriau prognozuoja modelis su parinktu 0.25 kvantiliu, vidutinius – su 0.5 kvantiliu ir brangiausius su 0.75 kvantiliu, ko ir buvo tikimasi iš kvantilių regresijos. Tiesinės regresijos prognozė (juoda punktyrinė linija) yra panaši į medianos regresijos prognozę.

3.5.Modelis su brangiais automobiliais

Patikriname, koks gautųsi modelis, jei paliktumėme visus automobilius, neatsižvelgiant į jų kainą.



9 pav. Automobilių kainų histograma su visais automobiliais



10 pav. Modelio kainū prognozavimas su visais automobiliais

Matome, kad paėmus duomenų rinkinį su daugiau išskirčių, kvantilių regresijos prognozė stipriai nepasikeičia.

IŠVADOS

Pritaikę kvantilių regresijos modelį pastebėjome, kad kovariantė „amžius“ nėra svarbi brangesniems automobiliams. Pasirinkę kvantilius matome, kad 0,25 kvantilis geriau prognozuoja pigesnius automobilius, 0,5 kvantilis vidutinius automobilius, o 0,75 kvantilis brangesnius automobilius. Pridėta automobilių amžiaus ir nuvažiuotų kilometrų sąveika buvo reikšminga visuose modeliuose. Tiesinės ir medianos regresijos modelis nedaug skyrėsi savo prognoze, net kai duomenyse ir buvo įtraukta daugiau išskirčių.

ŠALTINIAI

[1] „Data.world“ tinklapis. Tema: Used Cars Dataset. Prieiga per internetą: https://data.world/data-society/used-cars-data?fbclid=IwAR2e3CBH1VWivrRhQualVqWzuJ38ZBDBe5ktYuAKAkh0dq3gOyHokCdTdq4