VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Regresinė analizė

2 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 gama ir atvirkštinę Gauso regresiją ištirti kaip skiriasi vyno stiprumas nuo įvairių parametrų.

Uždaviniai:

1. Nuskaityti duomenis ir paruošti juos analizei;

2. Rasti tinkamiausią atvirkštinės Gauso ir gama regresijos modelį;

3. Iš jų išrinkti geriausią modelį.

1. DUOMENYS

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

1.1.Duomenų aprašymas

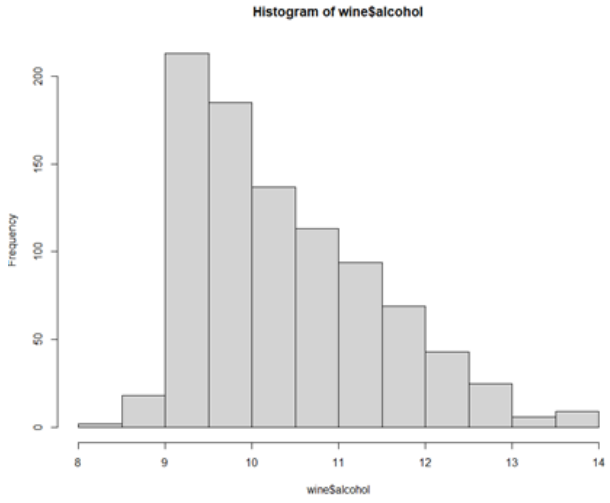
Laboratoriniame darbe nagrinėsime duomenis apie raudoną vyną. Duomenis sudaro:.

- Fixed acidity – fiksuotas vyno rūgštingumas;
- Volatile acidity – lakusis rūgštingumas;
- Citric acid – citrinos rūgštis;
- Residual sugar – liekamasis cukrus;
- Chlorides – druskos kiekis;
- Free sulfur dioxide – laisvasis sieros dioksidas;
- Total sulfur dioxide – visas sieros dioksidas;
- Density – tankis;
- pH – vandenilio jonų rodiklis;
- Sulphates – sulfatai;
- Alcohol – stiprumas;
- Quality – ekspertų įvertinta kokybė (1-10).

Priklausomas kintamasis – *stiprumas*. Tyrime naudosime reikšmingumo lygmenį $\alpha = 0,05$.

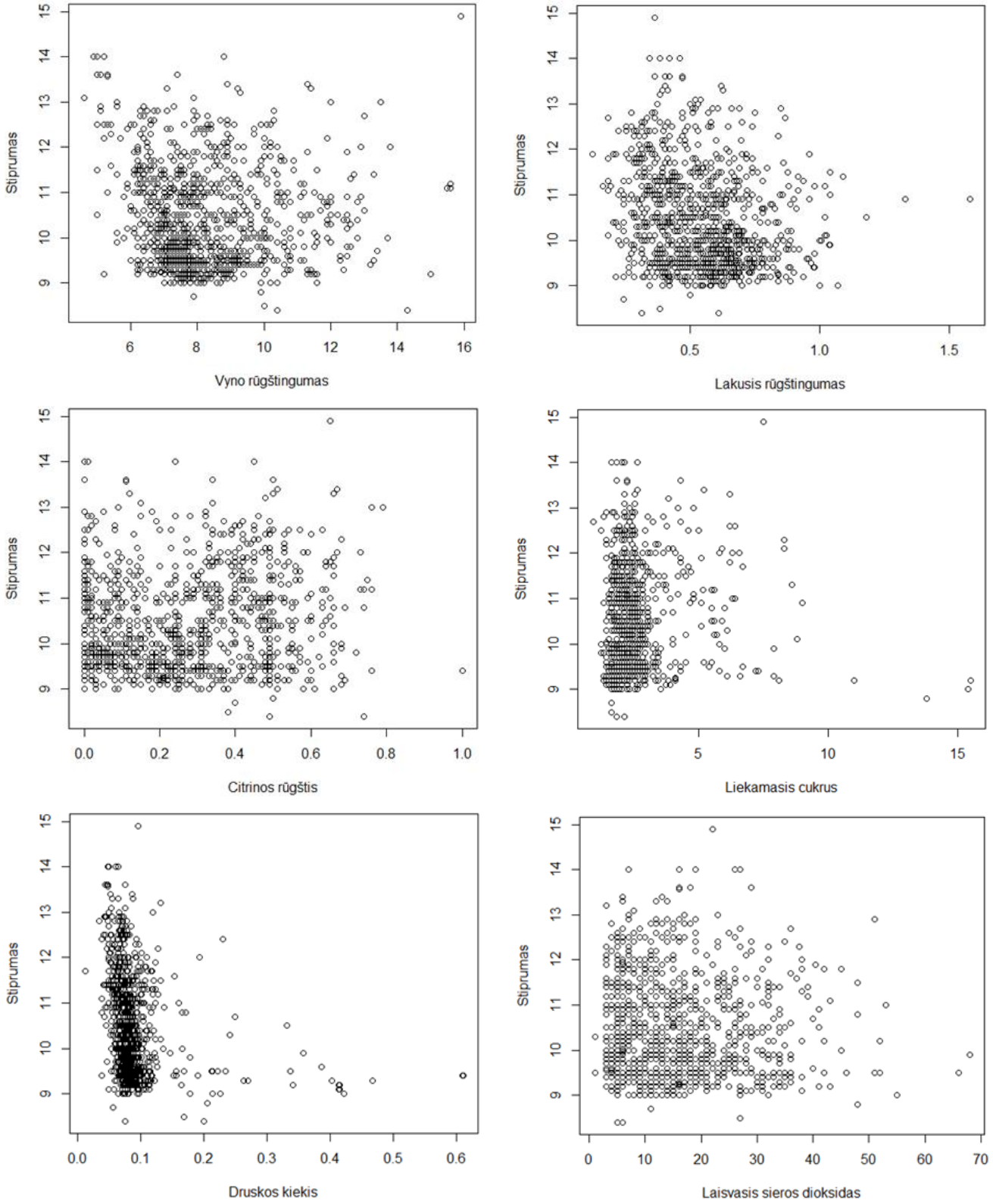
2. SĄRYŠIAI TARP VYNO STIPRUMO IR KOVARIANČIŲ

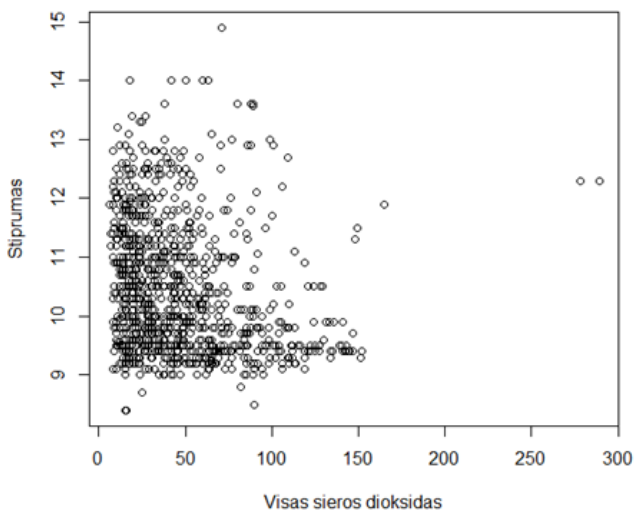
Pirmiausia tikriname priklausomojo kintamojo pasiskirstymą. Iš histogramos matome, kad skirstinys turi dešiniąją asimetriją. Dėl to taikysime gama ir atvirkštinę Gauso regresiją.

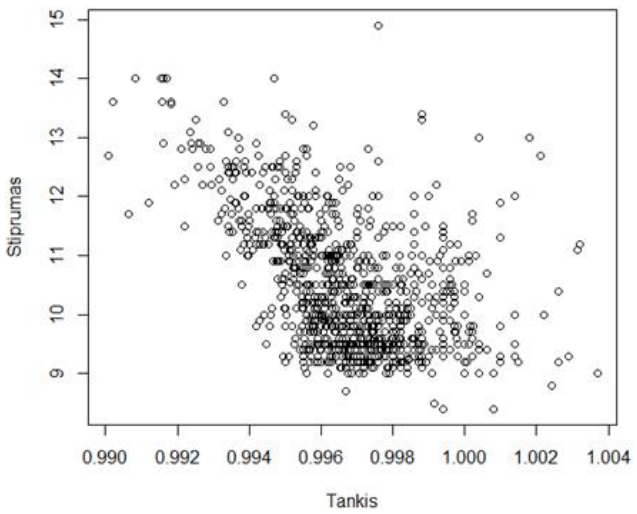


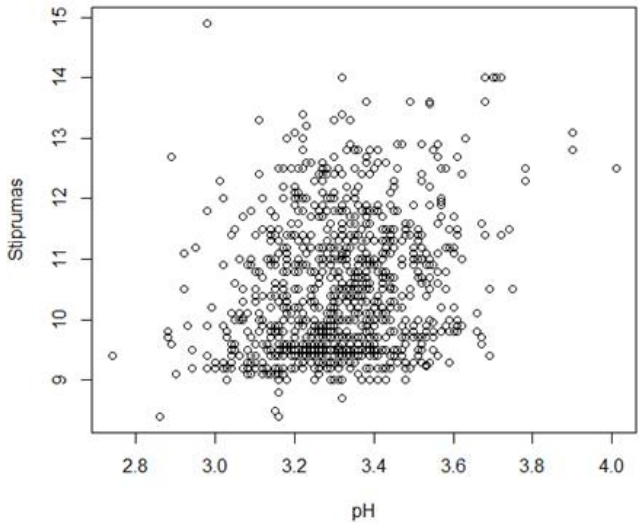
1 pav. Priklausomo kintamojo pasiskirstymo histograma

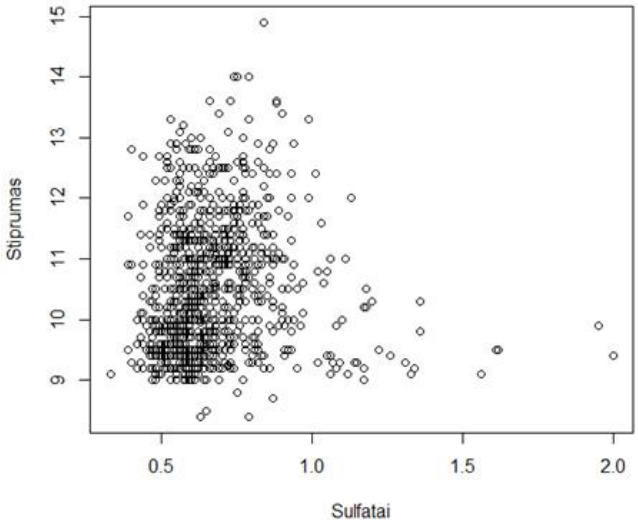
Taip pat norėjome patikrinti, kaip priklauso vyno stiprumas nuo mūsų pasirinktų kovariančių.

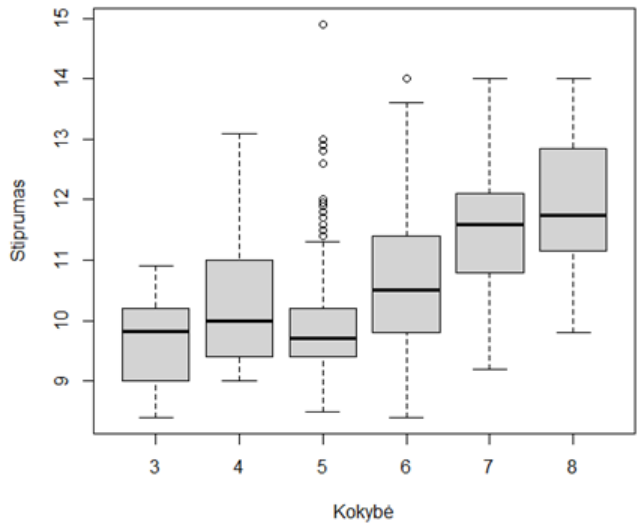












Iš taškinių diagramų atrodo, kad didžiausi sąryšiai yra tarp vyno stiprumo ir tankio, pH ir kokybės įvertinimo. Taip pat yra kelios išskirtys, kurios gali modelyje sukelti problemų.

3. REGRESIJOS TAIKYMAS NAUDOJANT GAMA MODELĮ

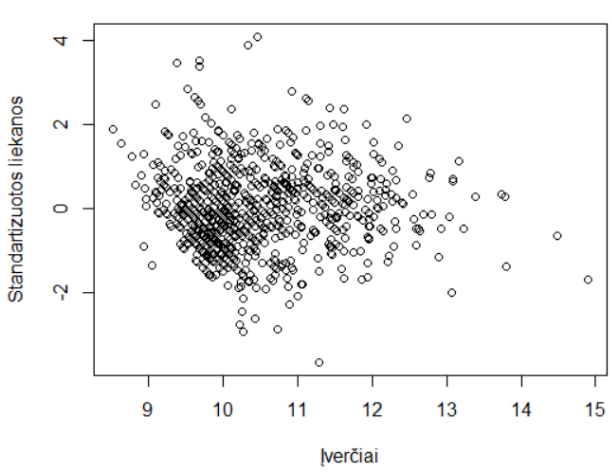
3.1.Jungties funkcija

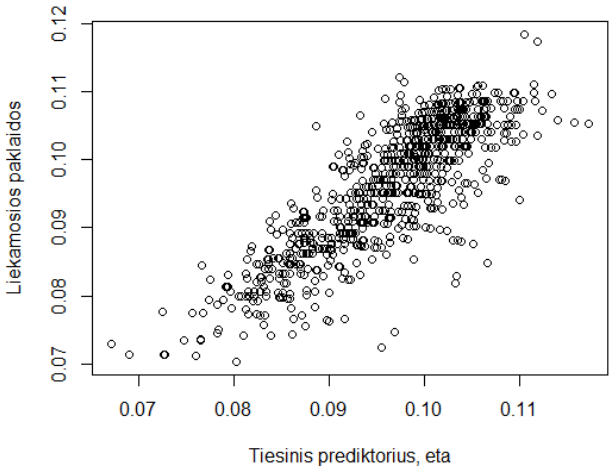
Pirma tiriant gama modelį su visom kovariantėm išbandome skirtingas jungties (angl. *link*) funkcijas. AIC kiekvienai jungties funkcijai:

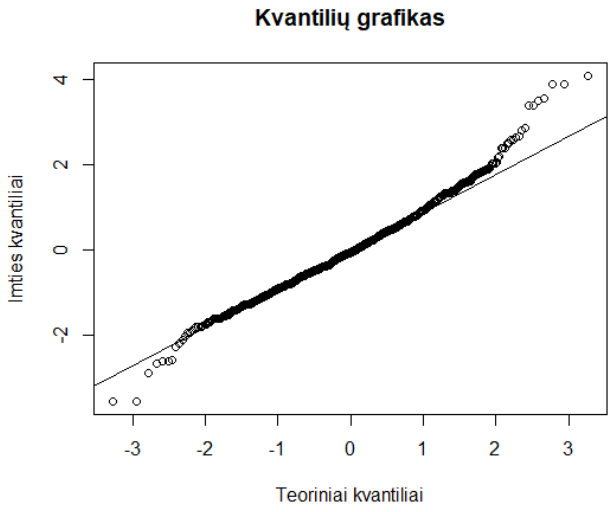
- Log – 1605,3;
- Inverse – 1589,1;
- Identity – 1635,4.

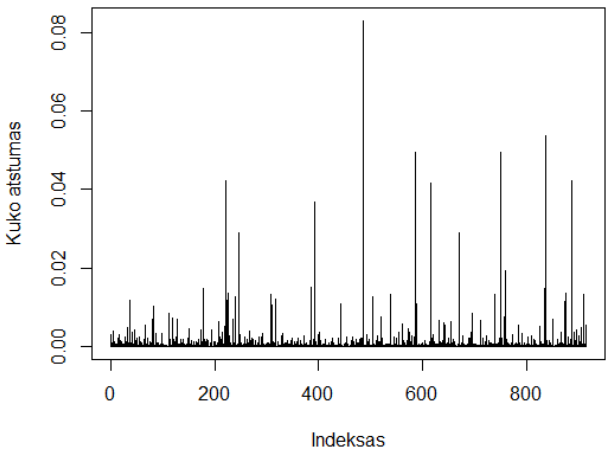
Matome, kad geriausia naudoti atvirkštinę jungties funkciją. Toliau gama modelio tyrimo dalyje naudosime tik ją.

3.2.Modelis su visomis kovariantėmis





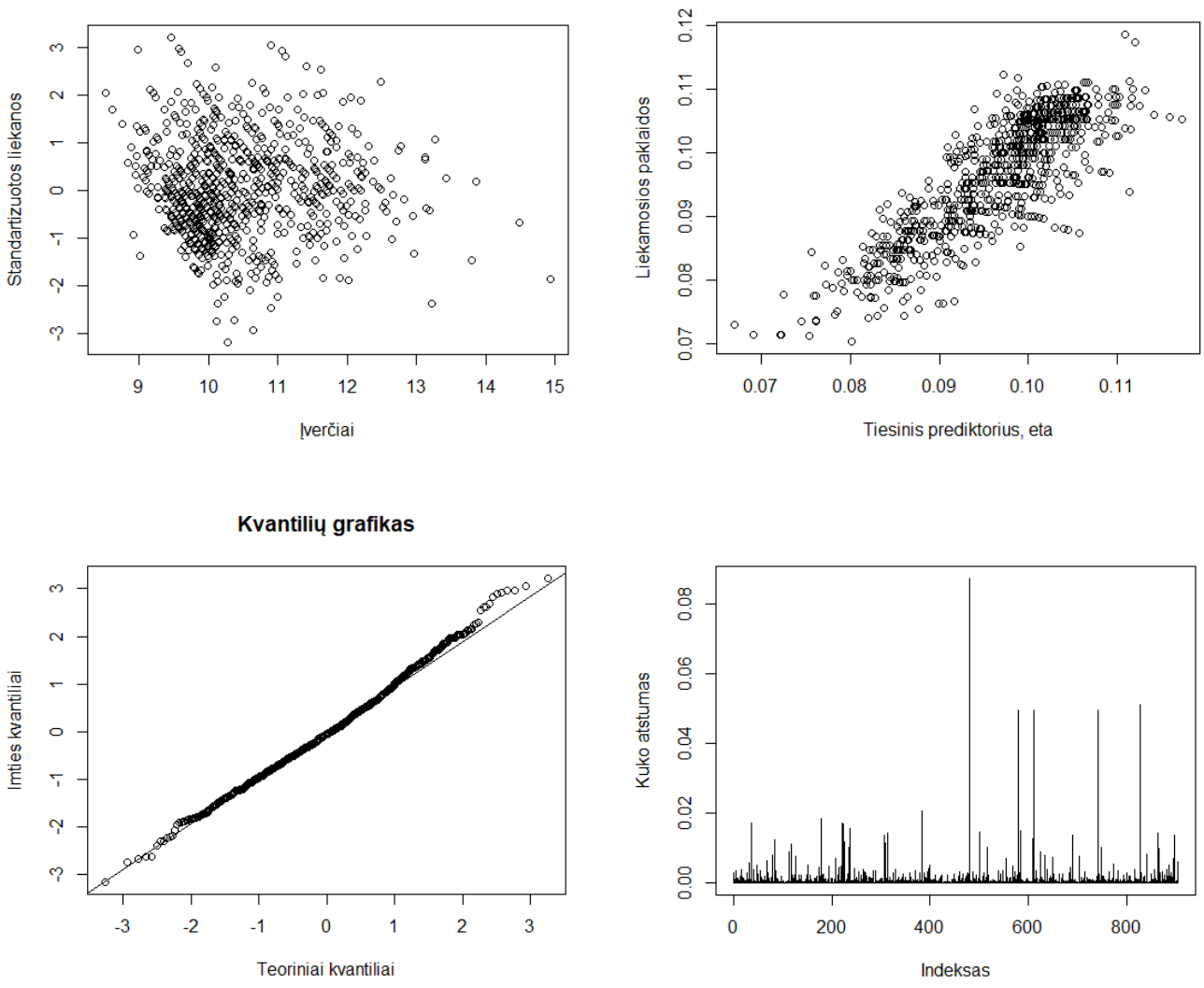




Iš standartizuotų liekanų grafiko matome, kad yra kelios išskirtys, tačiau taškai yra išsidėstę atsitiktinai ir nėra matomos tendencijos. Pagal Kuko matą išskirčių nėra. Taip pat iš kvantilių grafiko sprendžiame, kad modelis nevisai tinka aprašyti turimus duomenis (nukrypimai galuose). Taip pat matome, kad naudodami sąryšio funkciją ir tiesinį prediktorių duomenys aprašomi sąlyginai gerai, matomas tiesiškumas.

3.3.Modelis su pašalintomis išskirtimis

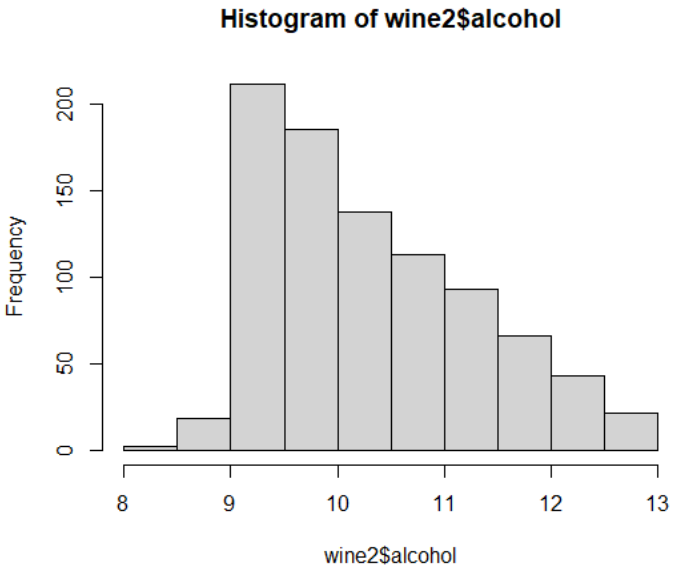
Pasirinkome šalinti išskirtis, kurių standartizuotų liekanų modulis yra didesnis už *3*. Po išskirčių išmetimo modelio AIC nukrito iki 1444,5.

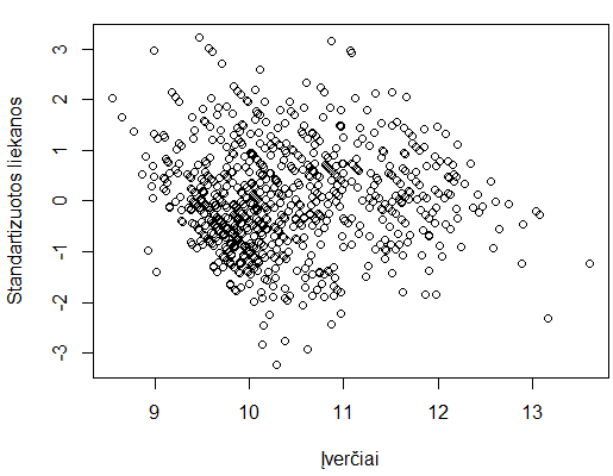


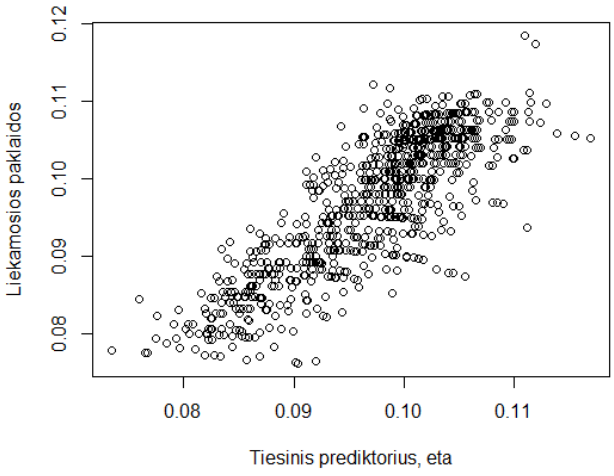
Iš kvantilių grafiko matome, kad mūsų modelis pagerėjo atmetus išskirtis (visi taškai išsidėstę arčiau tiesės). Tiesinis prediktorius smarkiai nepakito.

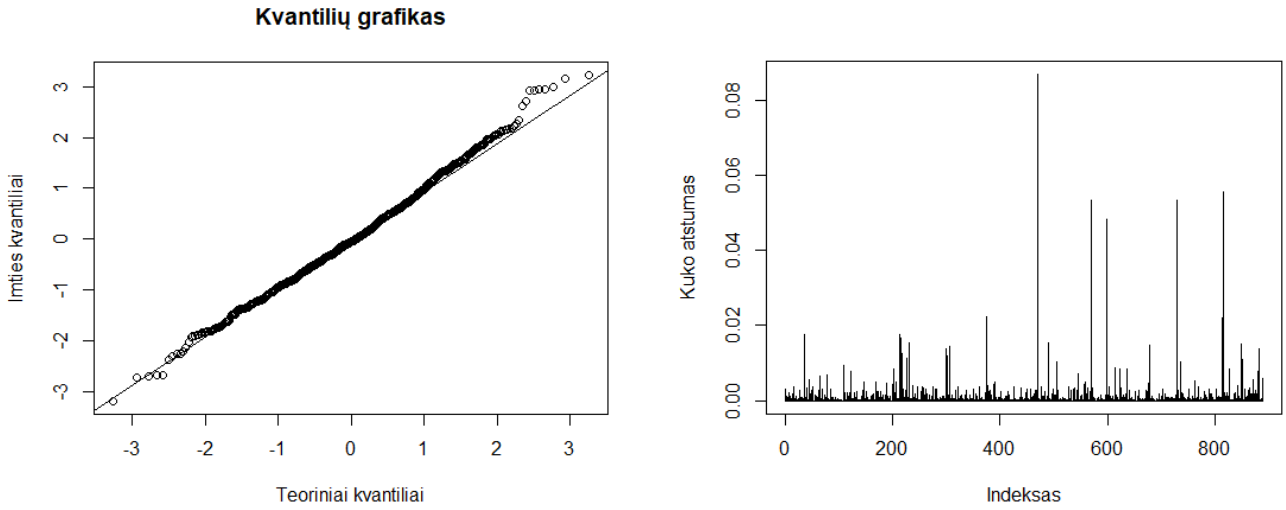
3.4.Modelis su susiaurinta priklausomojo kintamojo sritimi

Norint pagerinti modelio veikimą, bandėme siaurinti y kitimo sritį. Geriausias variantas kurį pavyko rasti – nupjauti visus duomenis, kuriuose vyno stiprumas yra didesnis už 13.









Matome, kad modelis nepagerėjo, todėl priklausomo kintamojo kitimo srities siaurinimo netaikysime.

3.5.Multikolinearumo tikrinimas

Patikrinus, ar modelyje yra multikolinearių kovariančių, gauname:

```
`fixed acidity`      `volatile acidity`          `citric acid`       `residual sugar`
       6.164758                1.905895                3.208533                1.317582
      chlorides  `free sulfur dioxide` `total sulfur dioxide`                 density
       1.572060                1.873937                2.089826                3.211776
             pH               sulphates                 quality
       2.629210                1.402857                1.597694
```

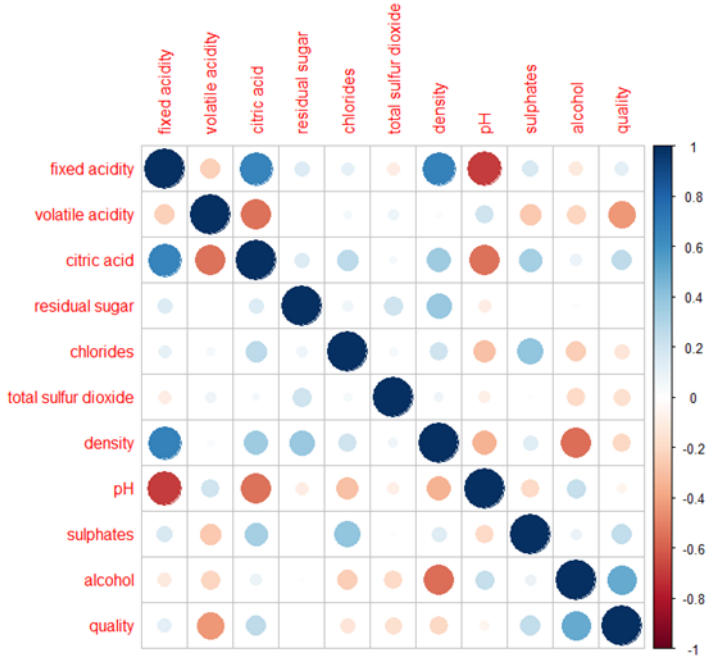
Matome, kad *fixed acidity* yra multikolineari. Iš modelio šaliname nereikšmingą kovariantę (*free sulfur dioxide*) tikėdamiesi, kad problema išsispręs.

```
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -4.762e+00  1.443e-01 -32.990  < 2e-16 ***
`fixed acidity`     -4.266e-03  2.365e-04 -18.039  < 2e-16 ***
`volatile acidity`  -5.186e-03  1.224e-03  -4.237 2.50e-05 ***
`citric acid`       -7.096e-03  1.476e-03  -4.806 1.80e-06 ***
`residual sugar`    -2.252e-03  1.314e-04 -17.138  < 2e-16 ***
chlorides            1.203e-02  4.357e-03   2.761  0.00588 **
`free sulfur dioxide` -1.330e-05  2.191e-05  -0.607  0.54391
`total sulfur dioxide`  3.045e-05  7.205e-06   4.227 2.61e-05 ***
density              5.037e+00  1.477e-01  34.111  < 2e-16 ***
pH                  -3.077e-02  1.689e-03 -18.221  < 2e-16 ***
sulphates           -7.800e-03  1.142e-03  -6.830 1.57e-11 ***
quality             -2.030e-03  2.455e-04  -8.267 4.93e-16 ***
```

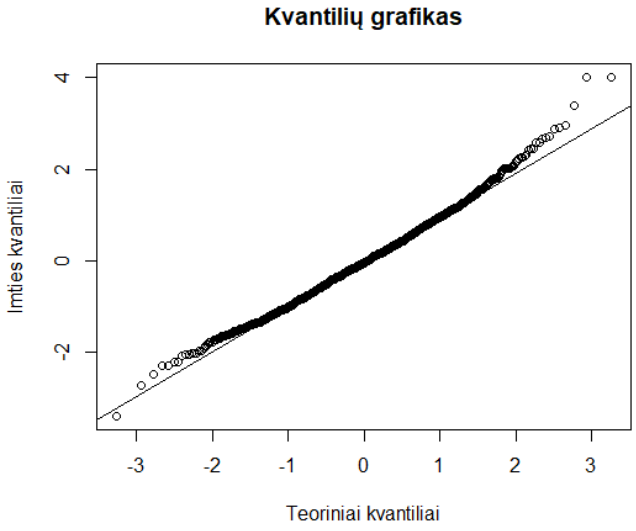
Vėl tikriname multikolinearumą:

```
`fixed acidity`      `volatile acidity`          `citric acid`       `residual sugar`
       6.126896                1.876413                3.146800                1.305863
      chlorides `total sulfur dioxide`                 density                      pH
       1.561643                1.211807                3.193363                2.598575
      sulphates                 quality
       1.401415                1.597427
```

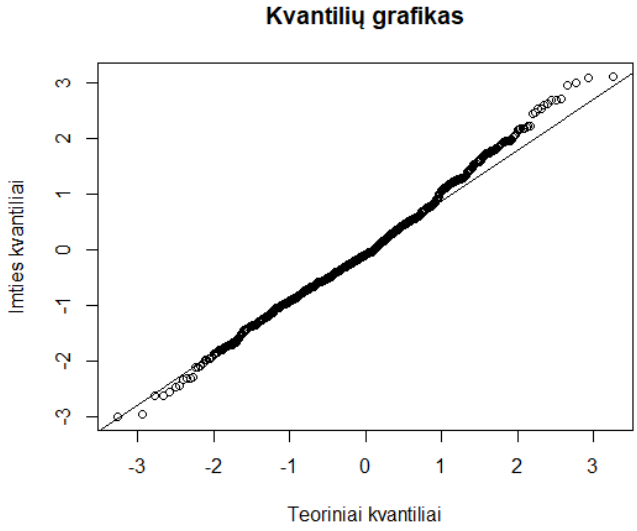
Matome, kad problema neišsisprendė. Kadangi nėra aišku, su kuo labiausiai koreliuoja *fixed acidity*, sukuriame koreliacijų matricą. Koreliacija stipriausia su *citric acid*, *density* ir *pH*.



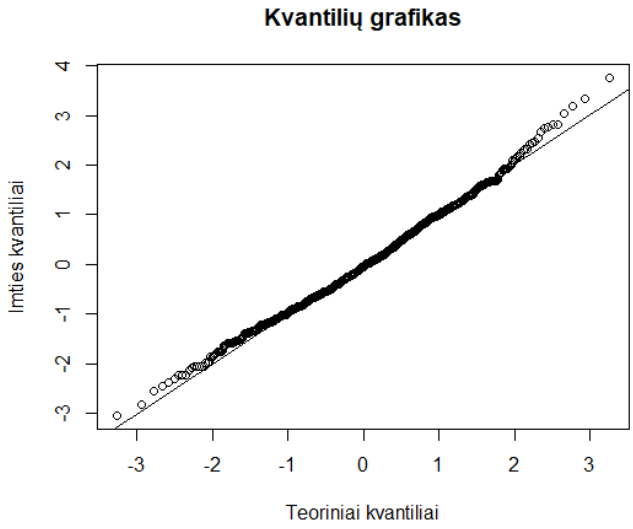
Sukuriame kelis modelius pretendentus. Tikriname kvantilių grafikus.



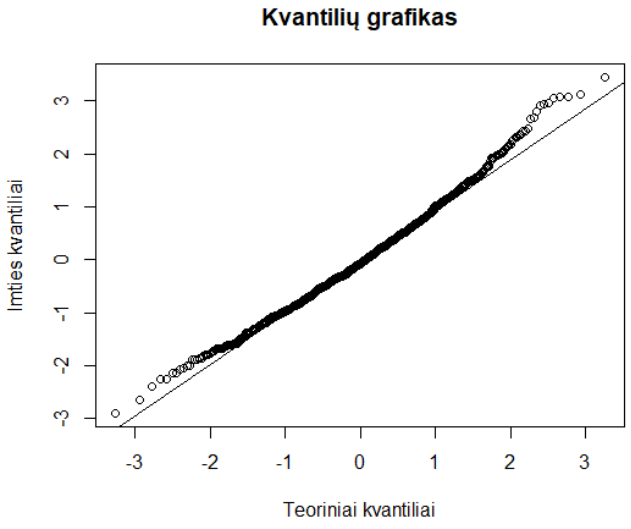
2 pav. Pirmas modelis su atmesta fixed acidity



3 pav. Antras modelis su atmesta citric acid



4 pav. Trečias modelis su atmesta pH

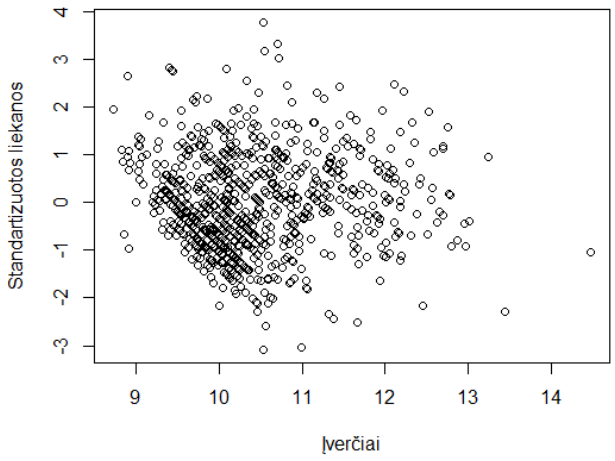


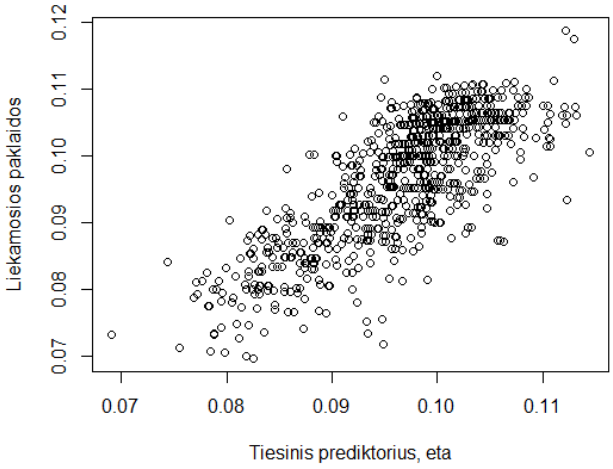
5 pav. Ketvirtas modelis su atmesta density

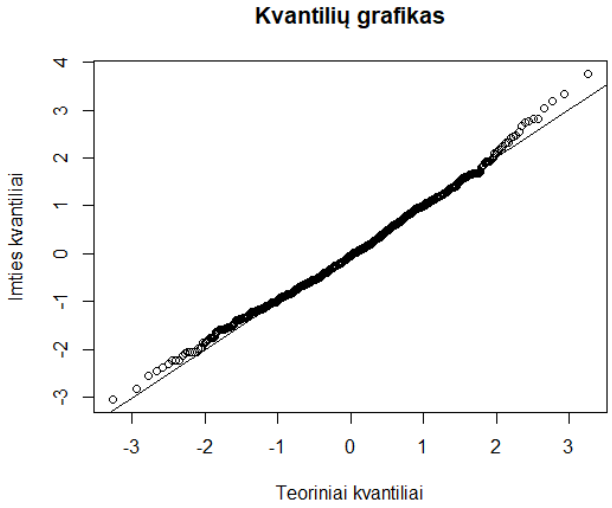
Nors modelio su išmesta *citric acid* kovariante AIC yra didžiausias (1463,8), iš kvantilių grafiko matome, kad jis netinka. Nusprendžiame, kad tinkamiausias modelis yra su išmesta *pH* kovariante.

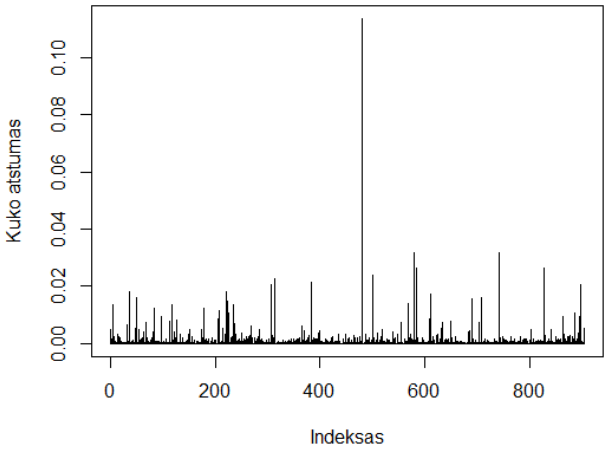
3.6.Galutinis gama modelis

Pasirenkame trečiąjį modelį, kur atmesta pH kovariantė.










```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -4.020e+00  1.631e-01 -24.656  < 2e-16 ***
`fixed acidity`      -1.564e-03  2.164e-04  -7.230 1.03e-12 ***
`volatile acidity`   -6.790e-03  1.433e-03  -4.738 2.50e-06 ***
`citric acid`        -8.347e-03  1.714e-03  -4.869 1.33e-06 ***
`residual sugar`     -2.060e-03  1.526e-04 -13.501  < 2e-16 ***
chlorides             3.763e-02  4.826e-03   7.796 1.77e-14 ***
`total sulfur dioxide`  5.234e-05  6.211e-06   8.428  < 2e-16 ***
density               4.168e+00  1.645e-01  25.331  < 2e-16 ***
sulphates            -8.156e-03  1.313e-03  -6.213 7.94e-10 ***
quality              -2.433e-03  2.874e-04  -8.467  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.003662181)

    Null deviance: 9.0365  on 904  degrees of freedom
Residual deviance: 3.2444  on 895  degrees of freedom
AIC: 1732.1
```

3.7.Interpretacija

(Intercept)	`fixed acidity`	`volatile acidity`	`citric acid`
8.225932e+19	1.017118e+00	1.072197e+00	1.091745e+00
`residual sugar`	chlorides	`total sulfur dioxide`	density
1.022363e+00	6.799951e-01	9.994933e-01	7.286750e-20
sulphates	quality		
1.088302e+00	1.026971e+00		

1 lentelė. Galimybių santykiai

Matome, kad jei keičiasi fiksuotas vyno rūgštingumas, lakusis rūgštingumas, citrinos rūgštis, liekamasis cukrus, sulfatai ir ekspertų įvertinta kokybė, tai vyno stiprumas atitinkamai padidėja 1,7 %, 7,2 %, 9,1 %, 2,2 %, 8,8 % ir 2,6 %. Pasikeitus druskos kiekiui, stiprumas sumažėja 32 %, o keičiantis

visam sieros dioksidui vyno stiprumas sumažėja vos 0,5 %. Keičiantis tankiui vyno stiprumas sumažėja nežymiai.

4. REGRESINĖ TAIKYMAS NAUDOJANT ATVIRKŠTINĮ GAUSO MODELĮ

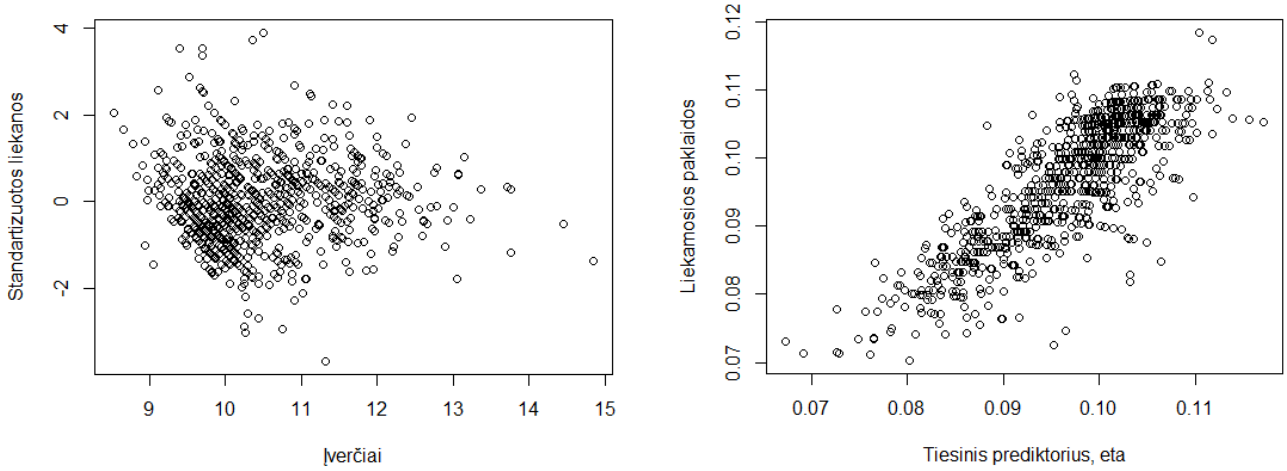
4.1.Jungties funkcija

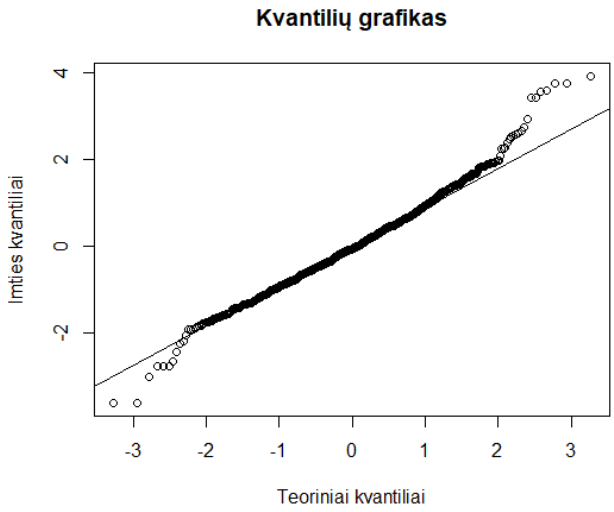
Pirma tiriant atvirkštinį (atv.) Gauso modelį su visom kovariantėm išbandome skirtingas jungties (angl. *link*) funkcijas. AIC kiekvienai jungties funkcijai:

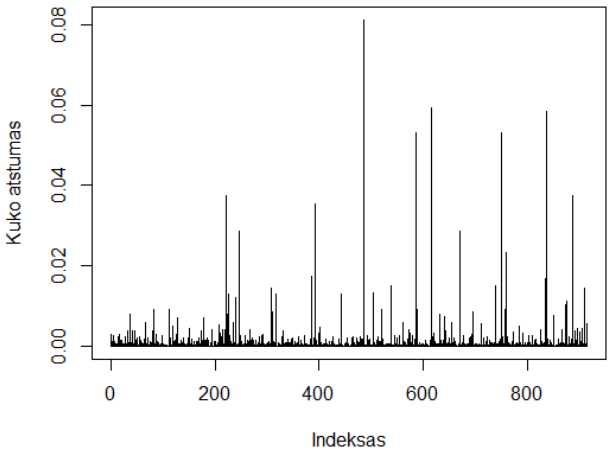
- Log – 1603,8;
- Inverse – 1585,3;
- Identity – 1634,5.

Matome, kad geriausia naudoti atvirkštinę jungties funkciją. Toliau atv. Gauso modelio tyrimo dalyje naudosime tik ją.

4.2.Modelis su visomis kovariantėmis



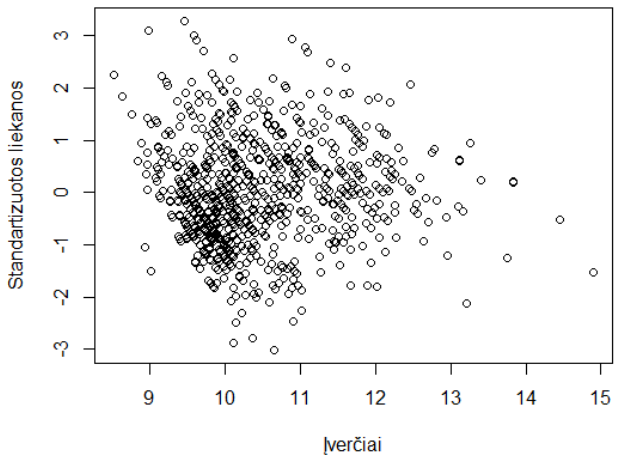


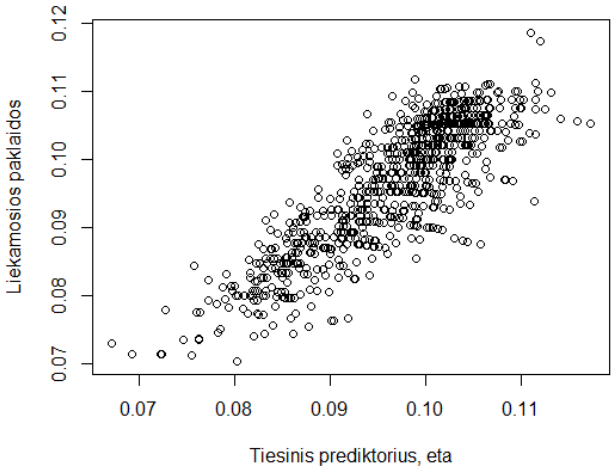


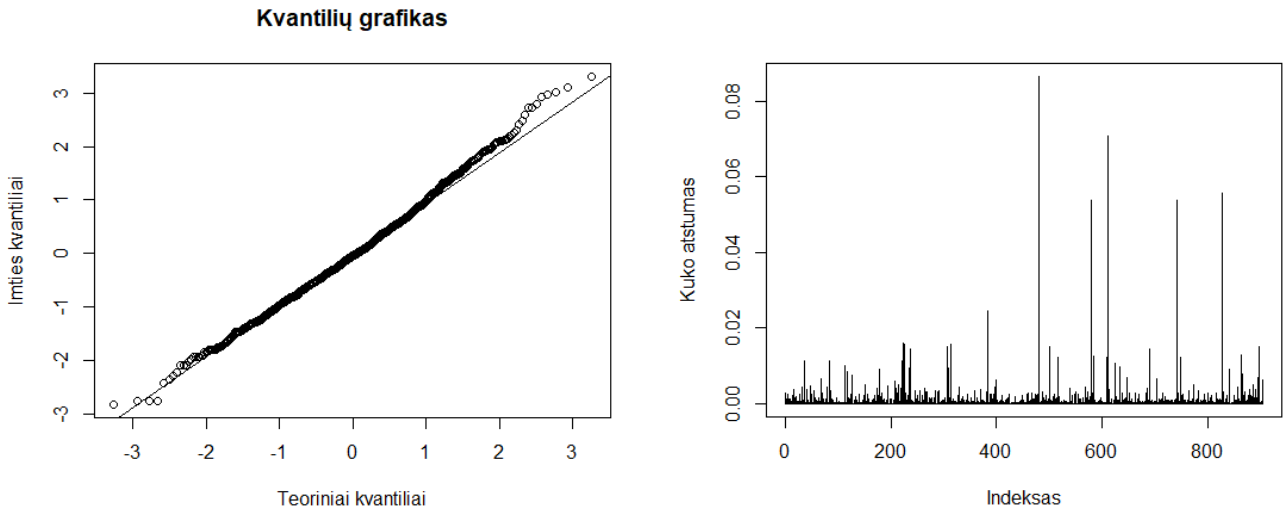
Iš standartizuotų liekanų grafiko matome, kad yra kelios išskirtys, tačiau taškai yra išsidėstę atsitiktinai ir nėra matomos tendencijos. Pagal Kuko matą išskirčių nėra. Taip pat iš kvantilių grafiko sprendžiame, kad modelis nevisai tinka aprašyti turimus duomenis (nukrypimai galuose). Taip pat matome, kad naudodami sąryšio funkciją ir tiesinį prediktorių, duomenys aprašomi sąlyginai gerai, matomas tiesiškumas.

4.3.Modelis su pašalintomis išskirtimis

Pasirinkome šalinti išskirtis, kurių standartizuotų liekanų modulis yra didesnis už *3*. Po išskirčių išmetimo modelio AIC nukrito iki 1431,9.



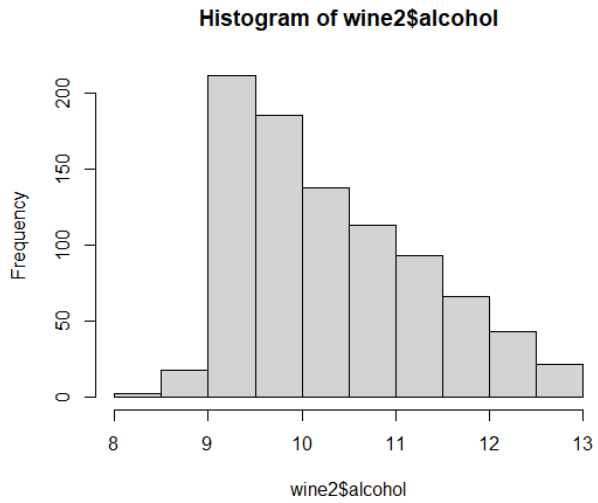


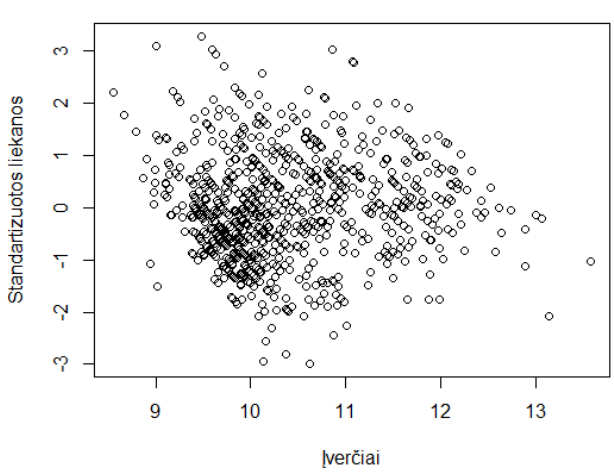


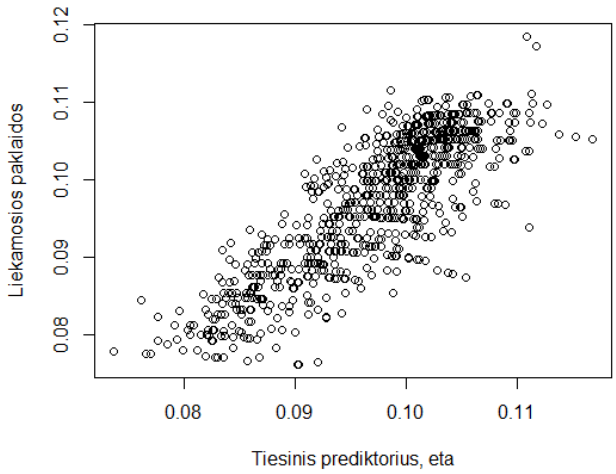
Iš kvantilių grafiko matome, kad mūsų modelis pagerėjo atmetus išskirtis (visi taškai išsidėstę arčiau tiesės). Tiesinis prediktorius smarkiai nepakito.

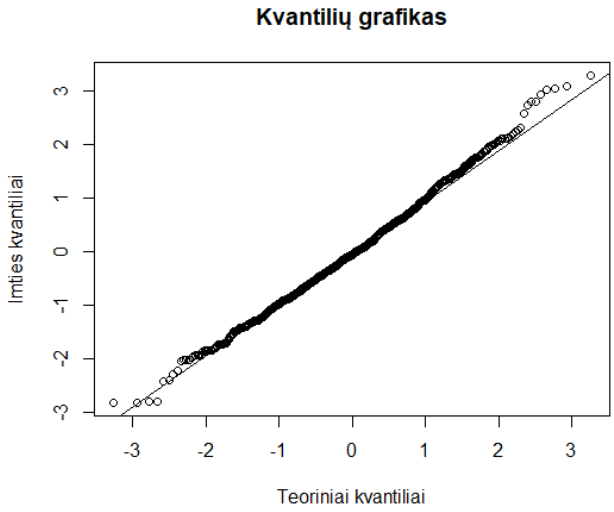
4.4.Modelis su susiaurinta priklausomojo kintamojo sritimi

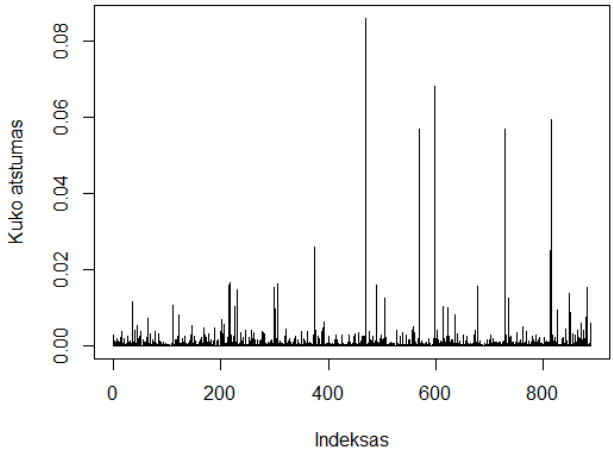
Norint pagerinti modelio veikimą, bandėme siaurinti y kitimo sritį. Geriausias variantas kurį pavyko rasti – nupjauti visus duomenis, kuriuose vyno stiprumas yra didesnis už 13.











Matome, kad modelis žymiai nepagerėjo, todėl priklausomo kintamojo kitimo srities siaurinimo taikyti neverta.

4.5.Multikolinearumo tikrinimas

Patikrinus, ar modelyje yra multikolinearių kovariančių, gauname:

```
`fixed acidity`      `volatile acidity`           `citric acid`       `residual sugar`
       6.086738                1.888715                3.185055                1.329496
      chlorides  `free sulfur dioxide` `total sulfur dioxide`                 density
       1.610369                1.890488                2.105239                3.197110
             pH               sulphates                 quality
       2.607885                1.414905                1.574747
```

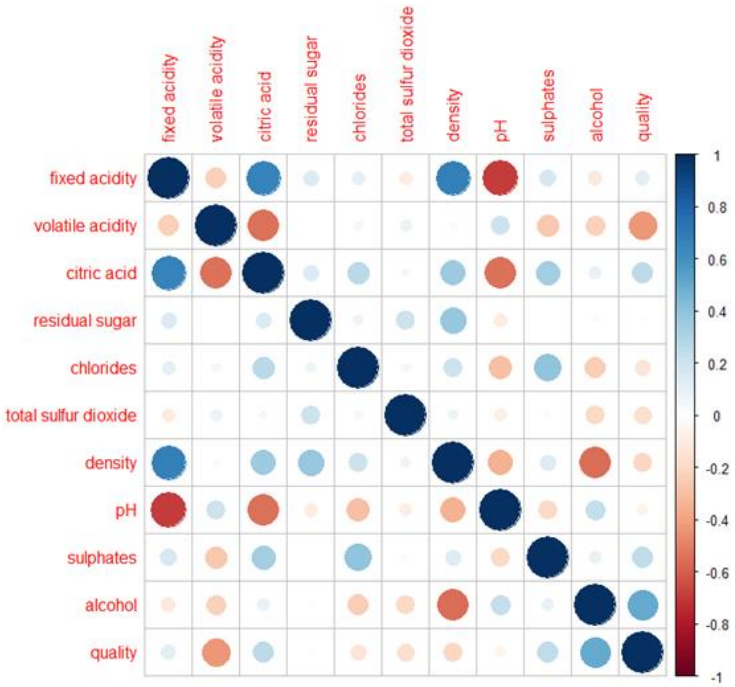
Matome, kad *fixed acidity* yra multikolineari. Iš modelio šaliname nereikšmingą kovariantę (*free sulfur dioxide*) tikėdamiesi, kad problema išsispręs.

```
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -4.728e+00  1.475e-01 -32.059  < 2e-16 ***
`fixed acidity`     -4.253e-03  2.378e-04 -17.882  < 2e-16 ***
`volatile acidity`  -5.062e-03  1.233e-03  -4.106 4.39e-05 ***
`citric acid`       -7.059e-03  1.480e-03  -4.769 2.17e-06 ***
`residual sugar`    -2.239e-03  1.326e-04 -16.884  < 2e-16 ***
chlorides            1.277e-02  4.252e-03   3.004  0.00274 **
`free sulfur dioxide` -1.556e-05  2.201e-05  -0.707  0.47974
`total sulfur dioxide` 3.026e-05  7.264e-06   4.165 3.42e-05 ***
density              5.002e+00  1.509e-01  33.144  < 2e-16 ***
pH                  -3.052e-02  1.703e-03 -17.921  < 2e-16 ***
sulphates           -7.988e-03  1.133e-03  -7.053 3.52e-12 ***
quality             -2.034e-03  2.479e-04  -8.204 8.05e-16 ***
```

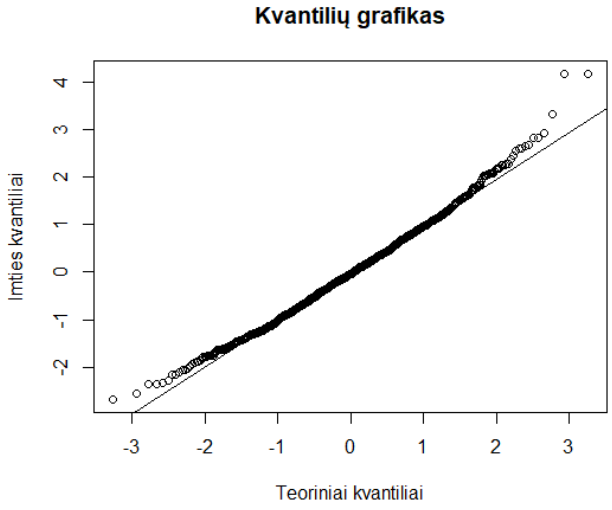
Vėl patikriname multikolinearumą:

```
`fixed acidity`     `volatile acidity`          `citric acid`       `residual sugar`
       6.048657               1.858197               3.126028               1.312800
     chlorides `total sulfur dioxide`                density                     pH
       1.598770               1.213390               3.175315               2.576306
     sulphates                quality
       1.413873               1.574232
```

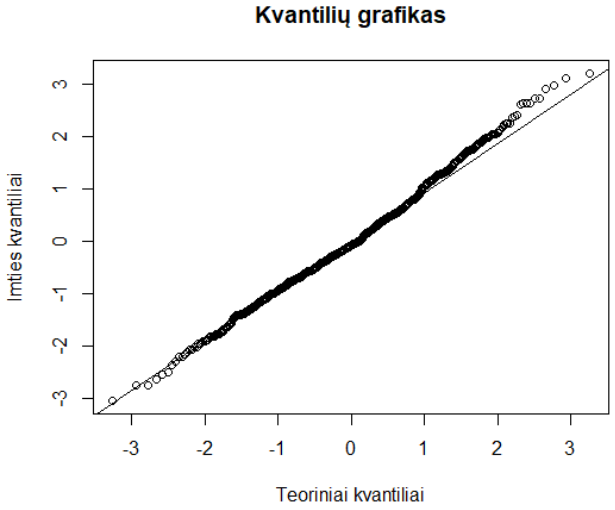
Matome, kad problema neišsisprendė. Kadangi nėra aišku, su kuo labiausiai koreliuoja *fixed acidity*, sukuriame koreliacijų matricą. Koreliacija stipriausia su *citric acid, density* ir *pH.*



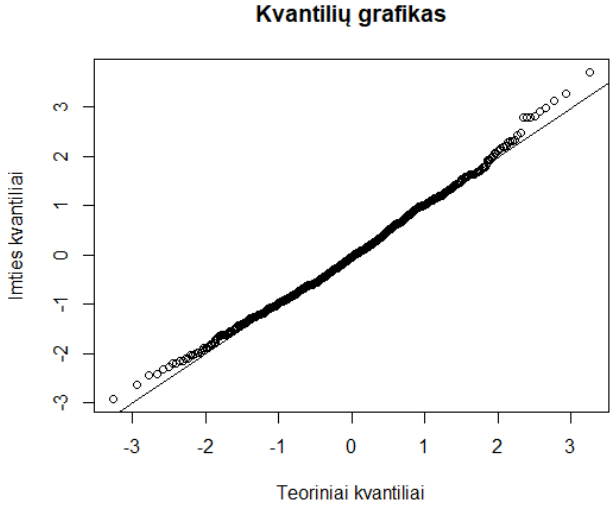
Sukuriame kelis modelius pretendentus. Tikriname kvantilių grafikus.



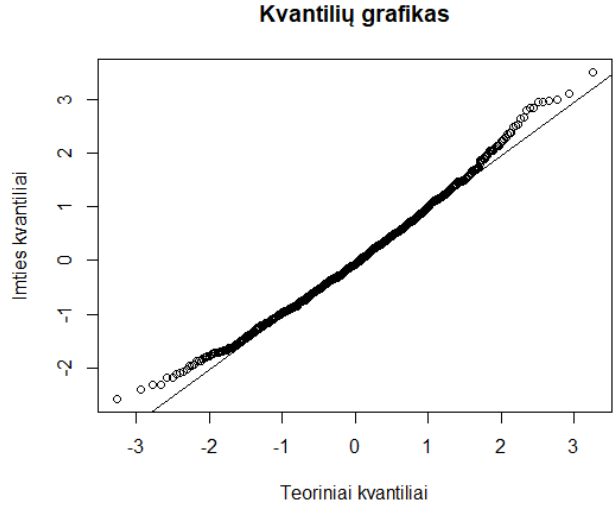
6 pav. Pirmas modelis su atmesta fixed acidity



7 pav. Antras modelis su atmesta citric acid



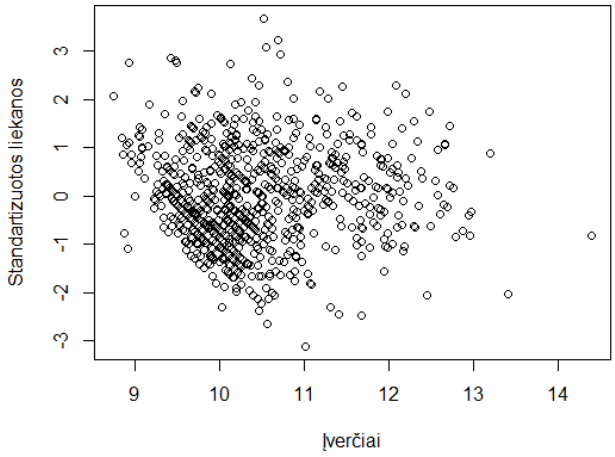
8 pav. Trečias modelis su atmesta pH

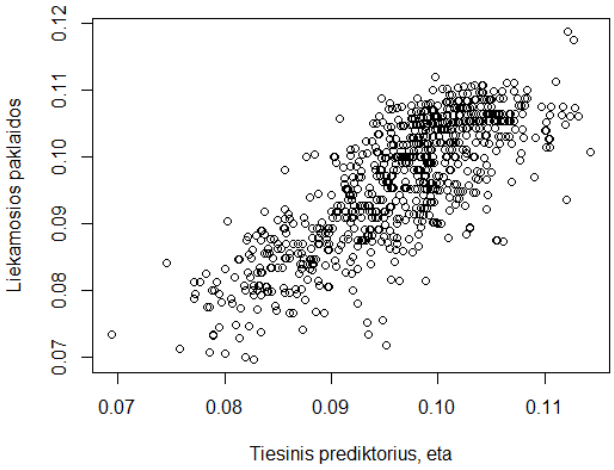


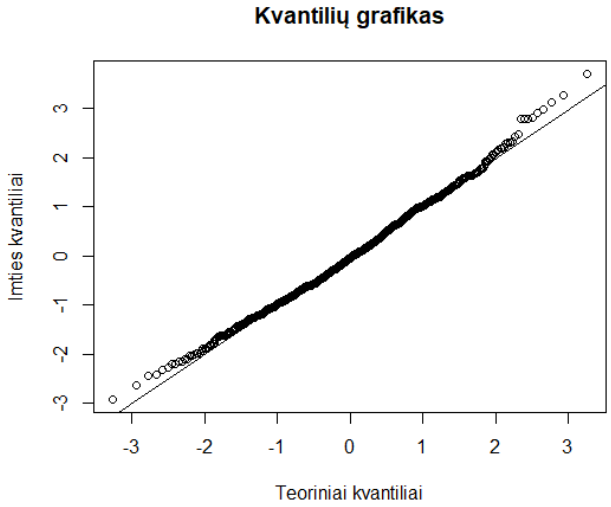
9 pav. Ketvirtas modelis su atmesta density

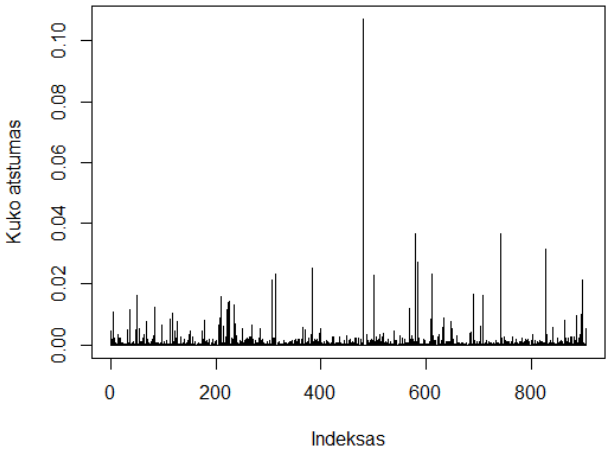
Iš kvantilių grafiko matome, kad pirmas (atmesta *fixed acidity*) ir ketvirtas (atmesta *density*) modeliai nėra tinkami. Antras (atmesta *citric acid*) ir trečias (atmesta *pH*) yra panašūs, tačiau antrame nėra išspręsta multikolinearumo problema. Todėl pasirenkame trečią modelį.

4.6.Galutinis atvirkštinis Gauso modelis









```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -4.020e+00  1.631e-01 -24.656  < 2e-16 ***
`fixed acidity`      -1.564e-03  2.164e-04  -7.230 1.03e-12 ***
`volatile acidity`   -6.790e-03  1.433e-03  -4.738 2.50e-06 ***
`citric acid`        -8.347e-03  1.714e-03  -4.869 1.33e-06 ***
`residual sugar`     -2.060e-03  1.526e-04 -13.501  < 2e-16 ***
chlorides             3.763e-02  4.826e-03   7.796 1.77e-14 ***
`total sulfur dioxide` 5.234e-05  6.211e-06   8.428  < 2e-16 ***
density               4.168e+00  1.645e-01  25.331  < 2e-16 ***
sulphates            -8.156e-03  1.313e-03  -6.213 7.94e-10 ***
quality              -2.433e-03  2.874e-04  -8.467  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.003662181)

    Null deviance: 9.0365  on 904  degrees of freedom
Residual deviance: 3.2444  on 895  degrees of freedom
AIC: 1732.1
```

4.7.Interpretacija

(Intercept)	`fixed acidity`	`volatile acidity`	`citric acid`
7.226035e+19	1.018002e+00	1.062221e+00	1.075953e+00
`residual sugar`	chlorides	`total sulfur dioxide`	density
1.028143e+00	6.830561e-01	9.994821e-01	8.179155e-20
sulphates	quality		
1.099075e+00	1.026104e+00		

2 lentelė. Galimybių santykiai

Matome, kad jei keičiasi fiksuotas vyno rūgštingumas, lakusis rūgštingumas, citrinos rūgštis, liekamasis cukrus, sulfatai ir ekspertų įvertinta kokybė, tai vyno stiprumas atitinkamai padidėja 1,8 %, 6,2 %, 7,5 %, 2,8 %, 9,9 % ir 2,6 %. Pasikeitus druskos kiekiui, stiprumas sumažėja 32 %, o keičiantis visam sieros dioksidui vyno stiprumas sumažėja vos 0,5 %. Keičiantis tankiui vyno stiprumas sumažėja nežymiai.

5. MODELIŲ PALYGINIMAS

Modelius lyginsime pagal AIC ir dispersijos parametro įvertinį $\hat{\phi}$ pagal Pirsono statistiką.

Gamma	IG
1732.106	1713.775

Gamma	IG
0.0036621806	0.0003476197

Pagal AIC matome, kad geresnis yra atv. Gauso modelis. Tą pagrindžia ir mažesnė dispersijos parametro reikšmė pagal dispersijos parametro įvertinį.

IŠVADOS

Abiejų modelių atveju AIC yra mažiausias naudojant atvirkštinę jungties funkciją. Iš kvantilių grafiko matome, kad modeliai nėra visiškai tinkami aprašyti turimus duomenis, o siaurinta priklausomojo kintamojo sritis situacijos nepagerina. Visgi nustatyta, kad atvirkštinis Gauso modelis yra tinkamesnis nei gama. Mūsų modeliuose reikšmingos kovariantės – fiksuotas vyno rūgštingumas, lakusis rūgštingumas, citrinos rūgštis, liekamasis cukrus, sulfatai, ekspertų įvertinta kokybė, druskos kiekis, visas sieros dioksidas ir tankis.

ŠALTINIAI

[1] „Kaggle“ tinklapis. Tema: Red Wine Quality. Prieiga per internetą: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009