VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

DUOMENŲ MOKSLAS

Tiriamasis darbas

Įmonės darbuotojų analizė

Darbą ruošė:
Matas Amšiejus, Sandra Macijauskaitė,
Salvija Račkauskaitė, Jekaterina Sergejeva,
Iveta Silkauskaitė
Duomenų mokslas III kursas

VILNIUS 2022 m.

TURINYS

Įvadas

Tikslas – ištirti darbuotojų atlyginimų pasiskirstymą pagal įvairias asmens savybes bei įvertinti kokie asmenys yra labiau linkę išeiti ar būti išmesti iš darbo.

Uždaviniai:

1. Pasirinkti duomenų rinkinį.
2. Atlikti pirminę duomenų analizę.
3. Pasirinkti regresijos modelius, kurie bus taikomi pasirinktai duomenų aibei.
4. Pritaikyti regresijos modelius, patikrinti modelių prielaidas.
5. Pateikti išvadas ir interpretacijas.

Duomenų aprašymas

Duomenys paimti iš kaggle internetinės svetainės. Prieiga internete:

https://www.kaggle.com/datasets/rhuebner/human-resources-data-set/code?select=HRDataset_v14.csv

Kintamųjų aprašymas:

Employee Name – darbuotojo vardas ir pavardė, kategorinis.

EmpID – darbuotojo unikalus identifikacijos numeris.

MarriedID – ar asmuo vedęs (1 – taip, 0 – ne).

MaritalStatusID – vedybinio statuso kodas, kiekybinis.

EmpStatusID – įdarbinimo statuso kodas, kiekybinis.

DeptID – departamento identifikacinis kodas, kiekybinis.

PerfScoreID – veiklos įvertinimo kodas, kiekybinis.

Salary – metinė alga, kiekybinis.

Termd – ar darbuotojas buvo atleistas, (1 – taip, 0 – ne).

PositionID – asmens darbo pozicijos kodas, kiekybinis.

Position – darbo pozicijos pavadinimas, kategorinis.

State – valstija, kurioje žmogus gyvena, kategorinis.

Zip – pašto kodas, kategorinis.

DOB – darbuotojo gimimo data.

Sex – lytis, kategorinis (M – vyras, F – moteris).

MaritalDesc – vedybinis statusas (išsiskyręs, vienišas, našlys,

CitizenDesc – ar asmuo yra pilietis, kategorinis.

HispanicLatino – ar asmuo iš Lotynų Amerikos regiono, kategorinis (taip, ne).

RaceDesc – asmens rasė, kategorinis.

DateofHire – asmens įdarbinimo data.

DateofTermination – asmens atleidimo iš darbo data.

TermReason – kodėl žmogus atleistas iš darbo, kategorinis.

EmploymentStatus – įdarbinimo statusas, kategorinis.

Department – departamento, kuriame dirba asmuo, pavadinimas.

ManagerName – asmens tiesioginio vadovo vardas ir pavardė.

ManagerID – unikalus kiekvieno vadovo identifikacinis numeris.

RecruitmentSource – atrankos šaltinis, per kurį darbuotojas buvo atrinktas.

PerformanceScore – veiklos įvertinimas, kiekybinis.

EngagementSurvey – darbuotojo įsitraukimo apklausos rezultatai, kiekybinis.

EmpSatisfaction – darbuotojo pasitenkinimas skalėje 1-5.

SpecialProjectsCount – specialių projektų, su kuriais darbuotojas dirbo per pastaruosius 6 mėnesius, skaičius.

LastPerformanceReviewDate – paskutinė asmens veiklos vertinimo data.

DaysLateLast30 – dienų skaičius, kai darbuotojas vėlavo į darbą per paskutines 30 dienų, kiekybinis.

Absences – skaičius, kiek kartų darbuotojas nebuvo darbe, kiekybinis.

Atlikti duomenų pertvarkymai tų kintamųjų, kurie žymi tam tikrą datą. Pvz. iš gimimo datos gautas asmens amžius, iš darbo pradžios ir pabaigos datų gautas laikas, kai žmogus dirba įmonėje.

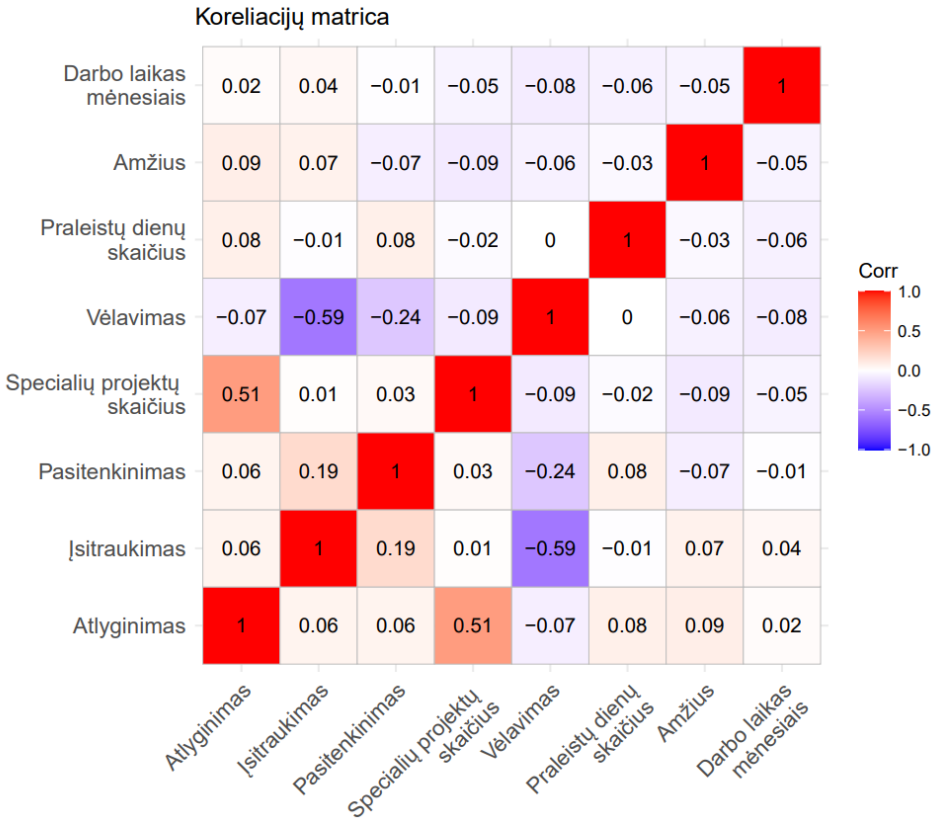
Pirminė duomenų analizė

1 lentelė. Aprašomosios duomenų statistikos.

Požymis	Min	Q1	Mediana	Vidurkis	Q3	Max
Atlyginimas	45046	55502	62810	69021	72036	250000
Įsitraukimas	1.12	3.69	4.28	4.11	4.70	5.00
Pasitenkinimas	1.00	3.00	4.00	3.89	5.00	5.00
Spec. projektai	0.00	0.00	0.00	1.219	0.00	8.00
Vėlavimas	0.00	0.00	0.00	0.41	0.00	6.00
Neatvykimas	1.00	5.00	10.00	10.24	15.00	20.00
Amžius	29.00	35.50	41.00	42.81	48.00	71.00
Darbo stažas	0.00	57.00	90.00	83.23	106.00	196.00

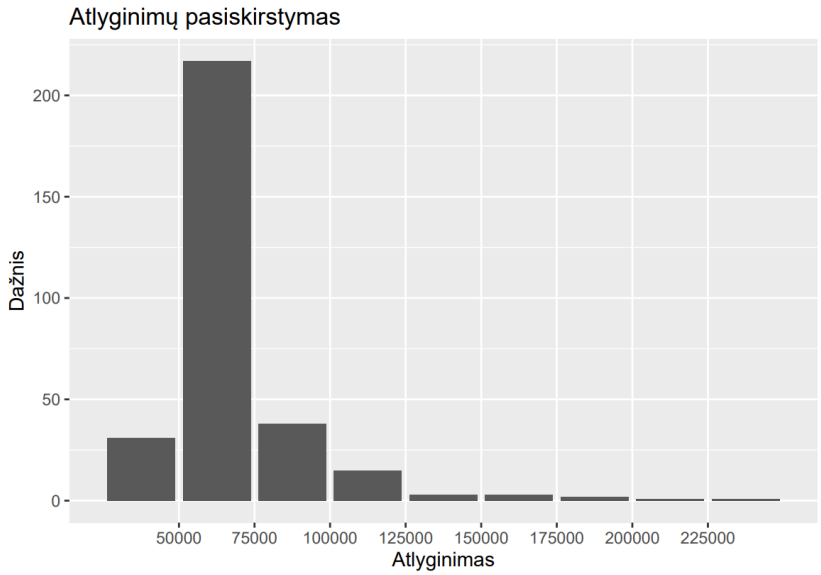
Iš Pirminė duomenų analizė

1 lentelės matome, kad stebėtų asmenų atlyginimas pasiskirsto nuo 45 iki 250 tūkst. dolerių per metus. Darbuotojai yra įsitraukę į įmonės veiklą, įvertinimo, kuris yra pasiskirstęs intervale nuo 1 iki 5, vidurkis lygus 4.11. Pastebime, kad mažai asmenų turi specialiųjų projektų, trečiojo kvartilio reikšmė dar vis yra 0, daugiausiai vienam asmeniui tenkančių projektų skaičius – 8. Darbuotojų amžius pasiskirstęs nuo 29 iki 71 metų. Taip pat įmonėje yra ir naujų darbuotojų, kurie nedirba nei mėnesio, ilgiausiai įmonėję dirbęs asmuo čia dirba jau daugiau nei 16 metų (196 mėnesius).



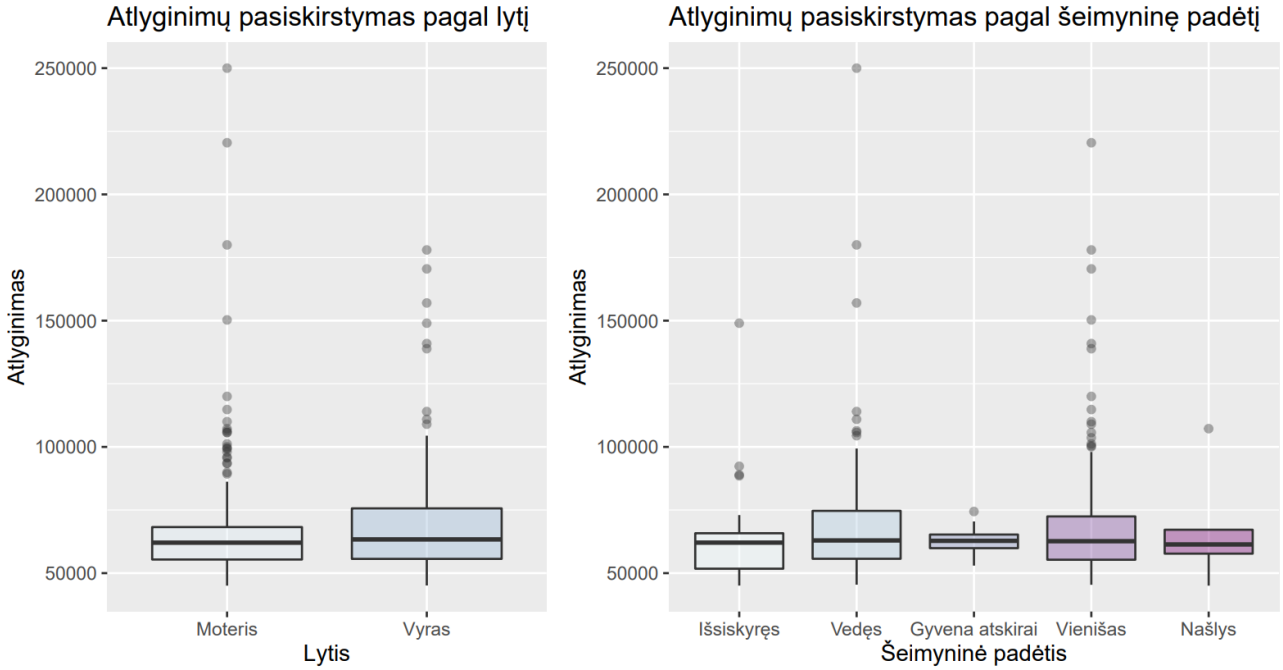
1 pav. Koreliacijų matrica.

Iš koreliacijų matricos (žr. 1 pav.) matome, kad didžiausią koreliaciją turi darbuotojo įsitraukimas ir vėlavimas, vidutinė neigiama koreliacija. Vėluojantys žmonės nebūna tiek įsitraukę į įmonės veiklą. Taip pat vidutinė koreliacija pastebima tarp kintamųjų specialių projektų skaičius ir atlyginimas. Šiuo atveju fiksuojama teigiama koreliacija – žmogus turintis daugiau specialių projektų uždirba daugiau. Tarp visų kitų kintamųjų pastebima tik labai silpna ar silna koreliacija.



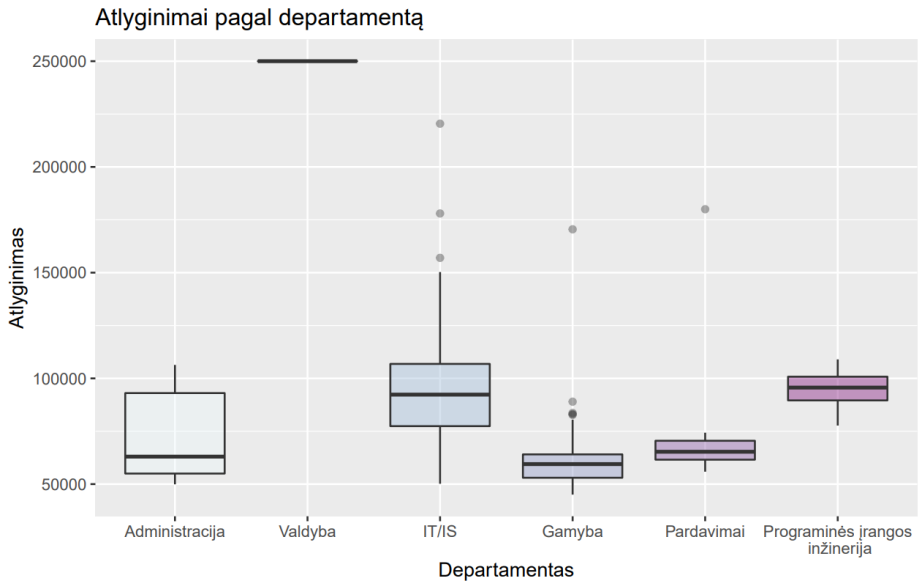
2 pav. Atlyginimų histograma.

Pagal atlyginimų histogramą (žr. 2 pav.) matome, kad egzistuoja dešininė asimetrija, daugiausia atlyginimų yra intervale nuo 50 iki 75 tūkst. dolerių per metus.



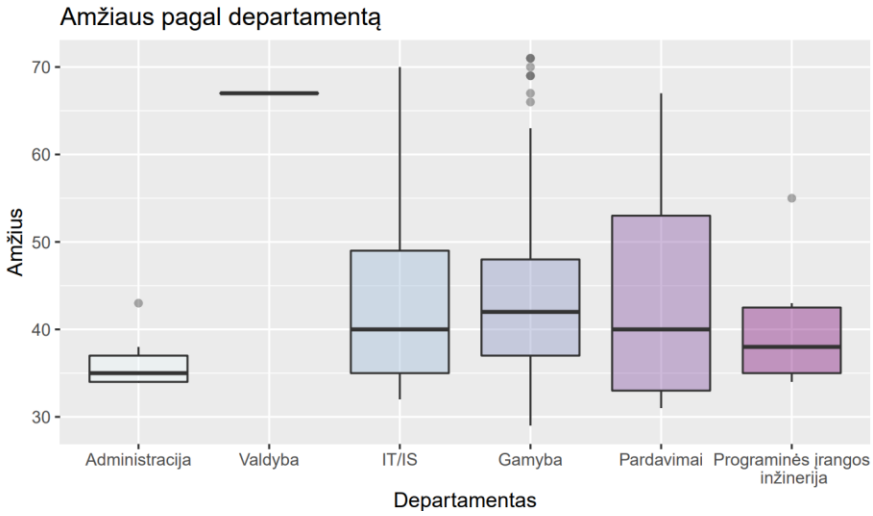
3 pav. Atlyginimai pagal lytį ir šeimyninę padėtį.

Iš 3 pav. galime pastebėti, kad vyrų atlginimų pasiskirstymas yra didesnis, visgi vidurkis skiriasi nežymiai, vyrų atlyginimų vidurkis didesnis apie 3 tūkst. dolerių per metus. Visgi matome ryškių išskirčių, moterų atlyginimas siekiantis net 250 tūkst. dolerių per metus. Žvelgiant į tai, kaip pasiskirsto atlyginimai pagal asmenų šeimyninę padėtį pastebime, kad išsiskyrę ir našliai uždirba kiek mažiau, daugiausia uždirba vieniši ir vedę asmenys.



4 pav. Atlyginimai pagal departamentą.

Kaip galime matyti iš 4 pav. labiausiai išsiskiria aukščiausias pareigas užimančio vadovo atlyginimas siekiantis 250 tūkst. dolerių per metus, toliau aukščiausiais atlyginimais gali pasigirti informacinių technologijų/informacinių sistemų skyrius, bei programinės įrangos inžinierių skyrius. Žemiausi atlyginimai gamybos departamente.



5 pav. Amžiaus pasiskirstymas pagal departamentą.

Kaip matome iš 5 pav., jauniausias departamentas – administracijos, seniausias – aukščiausias pareigas užimančio pareigūno.

2 lentelė. Darbuotojų skaičius pagal departamentą.

Administracija	Valdyba	IT/IS	Gamyba	Pardavimai	Programinės įrangos inžinerija

9	1	50	209	31	11

Kaip matome iš 2 lentelės, gamybos departamente dirba daugiausiai žmonių – 209.

Vėliau buvo apjungti kategoriniai kintamieji tokie kaip valstija, kadangi pastebėta, jog beveik visi stebėti asmenys gyvena Masačusetso valstijoje, kitos reikšmės buvo apjungtos į kategoriją „Other“ (Kita).

Žmogaus pilietybės kategorijos susiaurintos iki dviejų – Jav pilietis ir ne JAV pilietis.

Apjungti įdarbinimo šaltiniai ir asmens rasės pagal panašumus.

Dėl daug ir įvairių atleidimo iš darbo priežasčių bei mažo stebėjimų skaičiaus kai kuriose iš jų, kategorijų skaičius buvo sumažintas iki 5, kitos priežastys pridėtos į kategoriją „Other“ (Kita).

Darbo pozicijos pagal atlyginimo ir kitus panašumus buvo suskirstytos į kategorijas „Junior“, „Mid-level“, „Senior“, „Manager“, „Director“.

Binarinio atsako modelis

Pirmu modeliu buvo pasirinktas binarinio atsako logit modelis. Atsakas – stulpelio „Termd“ reikšmės, kur 1 žymi atvejį, kai darbuotojas buvo atleistas iš darbo, o 0 – ne,

Pradžioje nuskaitome duomenis ir padaliname juos į mokymosi ir testinę aibes.

Pasižiūrėjus, kiek kiekvienoje grupėje buvo stebėjimų, gauname tokius rezultatus:

3 lentelė. Atleistų ir dirbančių asmenų pasiskirstymas duomenyse.

0 (nebuvo atleistas)	1 (buvo atleistas)
206	104

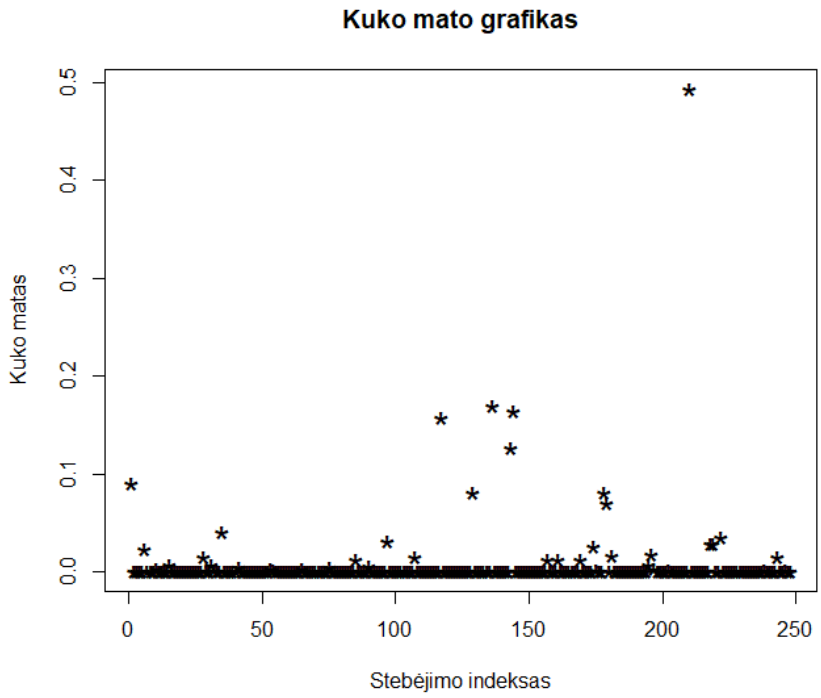
Matome, kad nesubalansuotų grupių problemos neturėsime, nes kiekvienoje grupėje yra daugiau nei 20% nuo visų stebėjimų.

Pabandę sudaryti modelį su visais stulpeliais iš sutvarkytų duomenų failo, gauname, kad modelis nesukonverguoja. Po vieną išmetus stulpelius TermReason, Department, RecruitmentSource (nes jie yra kategoriniai, įgyja nemažai reikšmių), gauname modelį, kuriame neįsivertina „Position“ stulpelio parametrai. Toliau po vieną šalinant nereikšmingas kovariantes ir lyginant modelius pagal AIC kriterijų, buvo gautas sekantis modelis, kurį ir toliau naudosime darbe:

```
Call:
glm(formula = Termd ~ Salary + Sex + CitizenDesc + HispanicLatino +
    PerformanceScore + EngagementSurvey + EmpSatisfaction + SpecialProjectsCount +
    DaysLateLast30 + Absences + Age + dirbo_men, family = binomial("logit"),
    data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5058  -0.1202  -0.0083   0.0053   3.7732

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    1.268e+01  6.314e+00   2.008  0.04460 *
Salary                        -4.516e-05  3.668e-05  -1.231  0.21820
SexM                          -7.690e-02  7.875e-01  -0.098  0.92220
CitizenDescUS                 -4.620e+00  2.341e+00  -1.974  0.04840 *
HispanicLatinoNo              -1.098e+00  1.344e+00  -0.817  0.41401
PerformanceScoreFully Meets    2.784e+00  1.731e+00   1.609  0.10771
PerformanceScoreNeeds Improvement  8.943e+00  3.521e+00   2.539  0.01110 *
PerformanceScorePIP            1.204e+01  4.416e+00   2.726  0.00642 **
EngagementSurvey               1.060e+00  6.115e-01   1.733  0.08317 .
EmpSatisfaction                5.091e-01  4.481e-01   1.136  0.25586
SpecialProjectsCount          -6.532e-01  2.932e-01  -2.228  0.02590 *
DaysLateLast30                -1.114e+00  6.273e-01  -1.775  0.07584 .
Absences                       1.466e-01  7.466e-02   1.963  0.04961 *
Age                            1.905e-02  4.812e-02   0.396  0.69212
dirbo_men                     -2.108e-01  4.358e-02  -4.837 1.32e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 310.382  on 247  degrees of freedom
Residual deviance:  55.399  on 233  degrees of freedom
AIC: 85.399

Number of Fisher Scoring iterations: 9
```

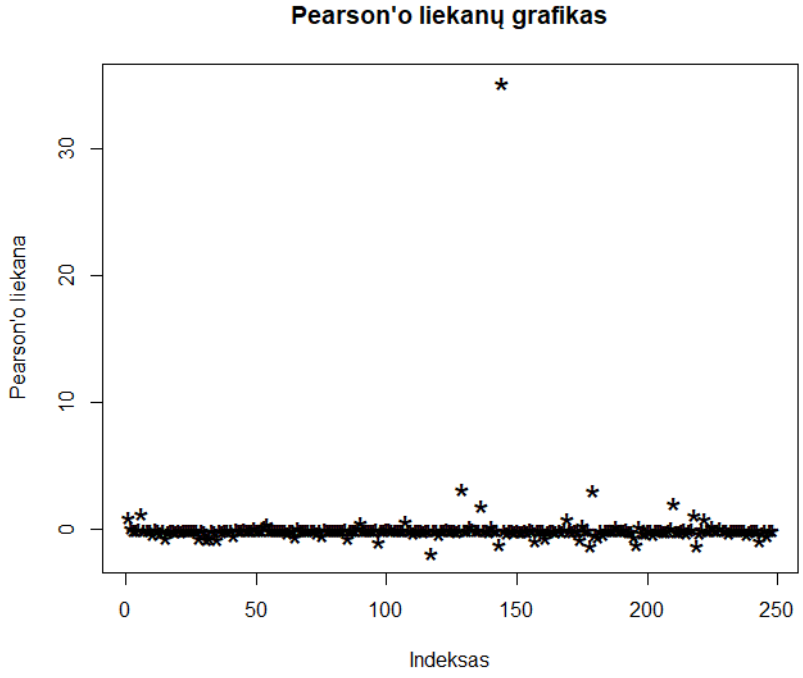
Iš rezultatų matosi, kad ne visos į modelį įtrauktos kovariantės yra reikšmingos, bet kol kas jų nešaliname.

Toliau tiriame išskirtis, tam buvo apskaičiuotas Kuko matas ir Pearsono liekanos.



6 pav. Kuko mato grafikas išskirtims.

Iš čia matome, jog nors nei viena reikšmė neviršija 1, vieno stebėjimo Kuko matas yra žymiai didesnis už visų stebėjimų atitinkamas reikšmes.



7 pav. Perasono liekanų grafikas išskirtims rasti.

Pasižiūrėjus Pearson'o liekanas (žr. 7 pav.) ir vėl matosi išsiskirianti reikšmė, jos indeksas yra kitoks nei tos išskirties, kurią matėme Kuko mato grafike. Išskirtis buvo pašalinta, daugiau išskirčių nerasta.

Toliau buvo patikrinta multikolinearumo prielaida. Buvo rastas multikolinearumas tarp kovariančių „PerformaceScore" (darbuotojo veiklos įvertinimas) ir „DaysLateLast30" (kiek kartų per paskutines 30 dienų darbuotojas pavėlavo į darbą).

```
> vif(modelis5)
                           GVIF Df GVIF^(1/(2*Df))
Salary                 3.397269  1        1.843168
Sex                    1.271797  1        1.127740
CitizenDesc            1.231548  1        1.109751
HispanicLatino         1.233035  1        1.110421
PerformanceScore      21.084321  3        1.662111
EngagementSurvey       2.677577  1        1.636330
EmpSatisfaction        1.753477  1        1.324189
SpecialProjectsCount   2.868147  1        1.693561
DaysLateLast30        10.047842  1        3.169833
Absences               1.625526  1        1.274961
Age                    1.215050  1        1.102293
dirbo_men              3.693933  1        1.921961
```

Išbandome du modelius, vieną be kovariantės „DaysLateLast30", kitą – be „PerformanceScore". Visas kitas kovariantes paliekame.

```
model1 <- glm(formula=Termd~Salary+Sex+CitizenDesc+HispanicLatino+PerformanceScore+EngagementSurvey+
                EmpSatisfaction+SpecialProjectsCount+Absences+Age+dirbo_men, data=data_train, family=binomial("logit"))
summary(model1)

model2 <- glm(formula=Termd~Salary+Sex+CitizenDesc+HispanicLatino+EngagementSurvey+EmpSatisfaction+
                SpecialProjectsCount+DaysLateLast30+Absences+Age+dirbo_men, data=data_train, family=binomial("logit"))
summary(model2)
```

Palyginus modelius pagal AIC reikšmes, geresnis gavosi pirmas modelis su kovariante „PerformanceScore" (be „DaysLateLast30"). Tačiau ir šiame modelyje buvo rasta nereikšmingų stulpelių, atliekame pažingsninę regresiją naudodami būtent šį modelį. Gauti rezultatai:

```
Call:  glm(formula = Termd ~ CitizenDesc + PerformanceScore + EngagementSurvey +
    SpecialProjectsCount + Absences + dirbo_men, family = binomial("logit"),
    data = data_train)

Coefficients:
                  (Intercept)                   CitizenDescUS       PerformanceScoreFully Meets  PerformanceScoreNeeds Improvement
                      13.4365                         -6.7121                          3.2984                          5.7759
        PerformanceScorePIP                EngagementSurvey             SpecialProjectsCount                        Absences
                       8.3013                          1.4546                         -1.3155                          0.1573
                  dirbo_men
                      -0.2404

Degrees of Freedom: 245 Total (i.e. Null);  237 Residual
Null Deviance:      305.8
Residual Deviance: 40.63        AIC: 58.63
```

Tai ir bus mūsų galutinis modelis. Atlikus pažingsninę regresiją, buvo išmesti stulpeliai „Salary", „Sex", „HispanicLatino", „EmpSatisfaction", „Age".

Taip pat buvo ištirti modeliai su sąveikomis PerformanceScore*EngagementSurvey, EngagementSurvey*SpecialProjectsCount ir PerformanceScore*SpecialProjectsCount. Tačiau nei vieno iš šių modelių rezultatai nelenkė modelio, kurį parinko pažingsninė regresija, todėl pasilikome jį.

Pereiname prie modelio tinkamumo vertinimo. Nors ir nesusidūrėme su nesubalansuotų grupių problema, vis tiek parinkome geriausią slenkstį, gavome 0.4082802.

Lentelėje galime matyti, kad modelio specifiškumas ir jautrumas viršija 97%, tai yra labai geras rezultatas.

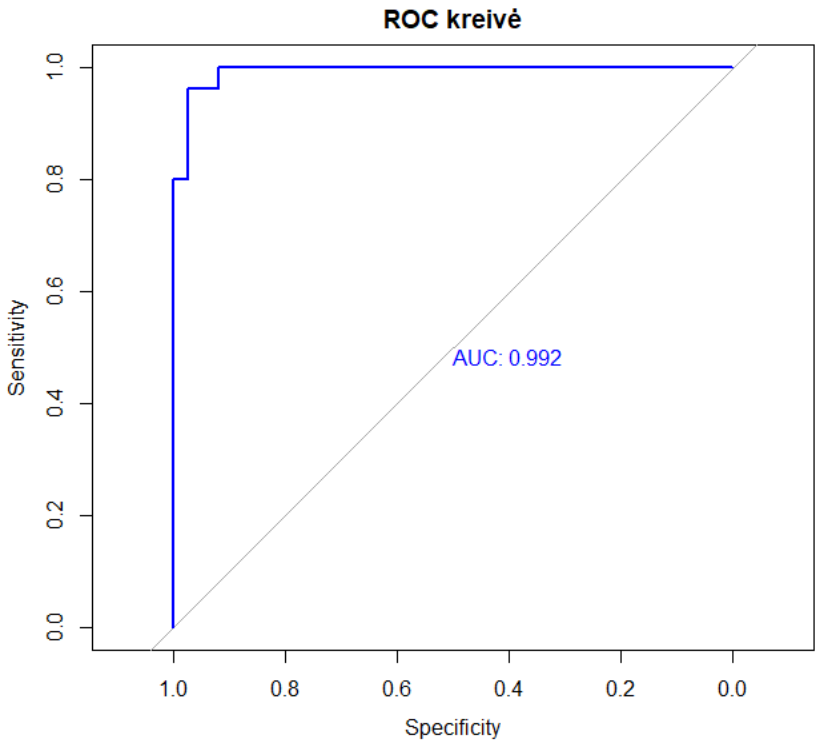
```
> ClassLog(final_model, data_train$Termd, cut=c$threshold)
$rawtab
       resp
          0   1
  FALSE 165   2
  TRUE    4  75

$classtab
       resp
                  0          1
  FALSE 0.97633136 0.02597403
  TRUE  0.02366864 0.97402597

$overall
[1] 0.9756098

$mcFadden
[1] 0.8671155
```

Iš pateiktos klasifikavimo lentelės (čia modelį testavome su testinės aibės duomenimis) matome, jog modelis gan tiksliai klasifikuoja ar žmogus buvo atleistas iš darbo, ar ne. Specifiškumas siekia 92%, o jautrumas – beveik 96%.

```
> (class_table_logit_test <- xtabs(~ predicted + response, data = classDF))
         response
predicted  0  1
        0 36  3
        1  1 22
> round(prop.table(class_table_logit_test, 1),3)
         response
predicted     0     1
        0 0.923 0.077
        1 0.043 0.957
```

Taigi pasižiūrėjus į klasifikavimo lenteles, matome, kad modelis tikrai gerai klasifikuoja darbuotojus, kurie buvo atleisti iš darbo ir tuos, kurie nebuvo.

Nusibraižome ROC kreivė:



8 pav. ROC kreivė

Kreivė yra toli nuo 45 laipsnių tiesės, einančios per grafiko įstrižainę, o plotas po ja yra 0.992, todėl galime teigti, kad modelis yra geras.

Koeficientai:

```
round(exp(coef(final_model)),4)
                  (Intercept)                  CitizenDescUS      PerformanceScoreFully Meets PerformanceScoreNeeds Improvement
                    684570.2125                       0.0012                         27.0690                         322.4351
           PerformanceScorePIP                EngagementSurvey                SpecialProjectsCount                        Absences
                     4029.0558                       4.2830                         0.2683                          1.1704
                     dirbo_men
                        0.7863
```

Matome, kad jeigu darbuotojas yra JAV pilietis, tai tikimybė būti atleistam sumažėja, palyginus su kitų šalių piliečiais. Taip pat, kuo blogesnis yra darbuotojo veiklos įvertinimas, tuo greičiau didėja tikimybė, kad jį atleis iš darbo. Pvz., gavusiems įvertinimą PIP (angl. Performance improvement plan), tikimybė būti atleistiems didėja net 4029 kartais. Taip pat pastebime, kad darbuotojo įsitraukimo apklausos rezultatui padidėjus vienetu, tikimybė būti atleistam padidėja 4,283 karto. Taip pat, kuo daugiau yra specialių projektų, kuriuose dalyvauja darbuotojas ir kuo ilgiau jis dirba įmonėje, tuo mažesnė tikimybė jam būti atleistam. O su kiekvienu neatvykimu į darbą, tikimybė būti atleistam padidėja 1,17 karto.

GLM modeliai atlyginimo dydžiams nustatyti

Iš pradinės duomenų analizės pastebėjome (2 pav. Atlyginimų histograma.), kad atlyginimų histograma turi dešiniąją asimetriją, todėl galime įtarti, kad tiks gama ir atvirkštinio Gauso regresijos modeliai.

Pirma tiriame gama modelį su visomis kovariantėmis išskyrus TermReason (per mažai įrašų grupėse). Išbandome skirtingas jungties funkcijas. Tikriname kiekvienos jungties funkcijos AIC:

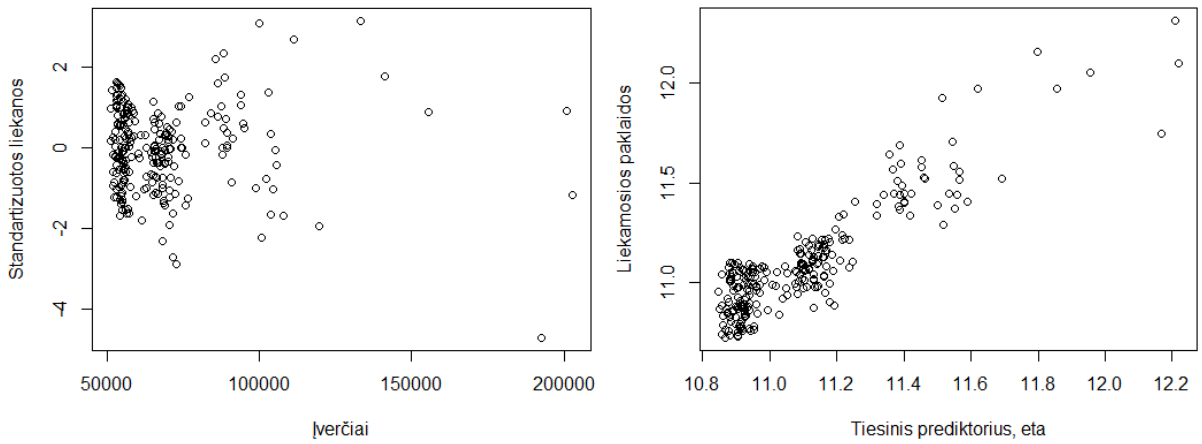
- Log – 5233,8;
- Inverse – 5242,1;
- Identity – 5235,9.

Matome, kad geriausia naudoti log jungties funkciją. Toliau gama modelio analizėje ją ir naudosime.

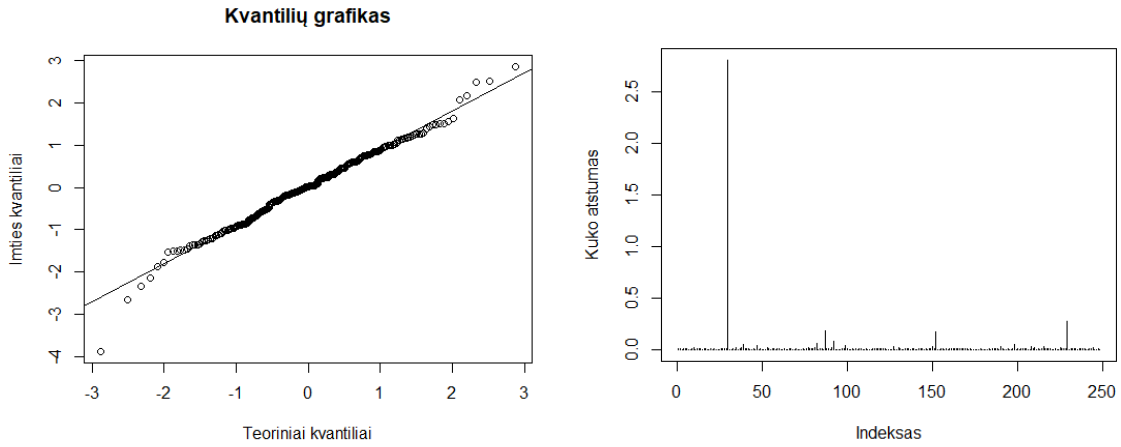
Atliekame Šapiro-Vilko testą patikrinti, kad paklaidos yra pasiskirsčiusios pagal normalųjį skirstinį:

```
        Shapiro-Wilk normality test

data:  qresid(gamma.log)
W = 0.98779, p-value = 0.03345
```

Patikrinus gauname, kad p reikšmė yra mažesnė už reikšmingumo lygmenį, todėl nulinę hipotezę atmetame. Paklaidos nėra pasiskirsčiusios pagal normalųjį skirstinį.



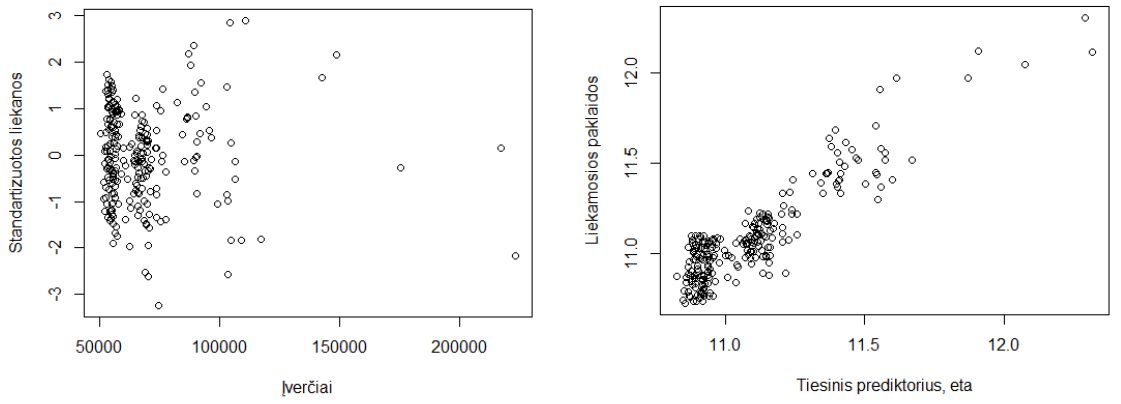
pav. 9



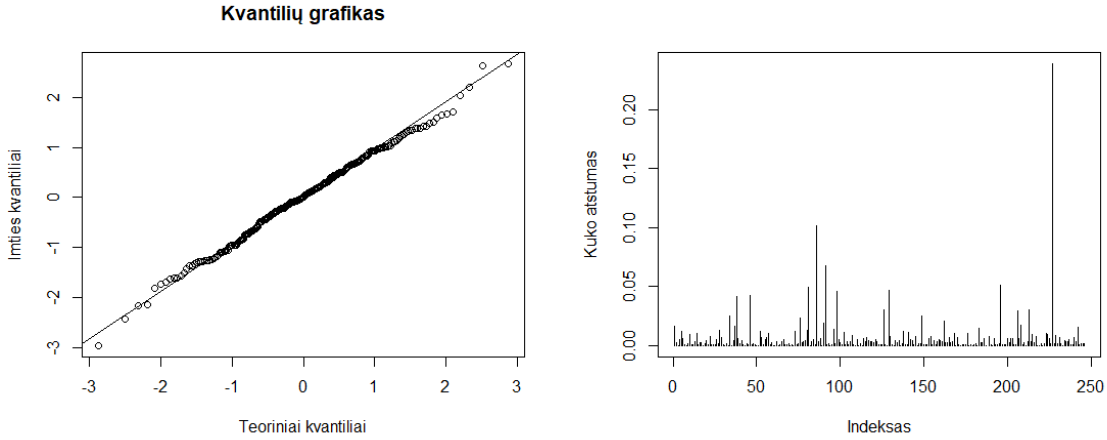
10 pav. Gama modelio išskirčių, tiesinio prediktoriaus ir kvantilių grafikai

Iš standartizuotų liekanų grafiko matome, kad yra kelios išskirtys, o pagal Kuko matą – viena. Iš kvantilių grafiko sprendžiame, kad modelis nevisai tinka aprašyti turimus duomenis, yra nukrypimų galuose. Taip pat matome, kad naudodami sąryšio funkciją ir tiesinį prediktorių duomenys aprašomi gerai, yra matomas tiesiškumas.

Pašalinus didžiausią išskirtį pagal Kuko matą ir Stjudentizuotas liekanas kuriame modelį be išskirčių. Po išskirčių išmetimo modelio AIC nukrito iki 5166,1.



pav. 11



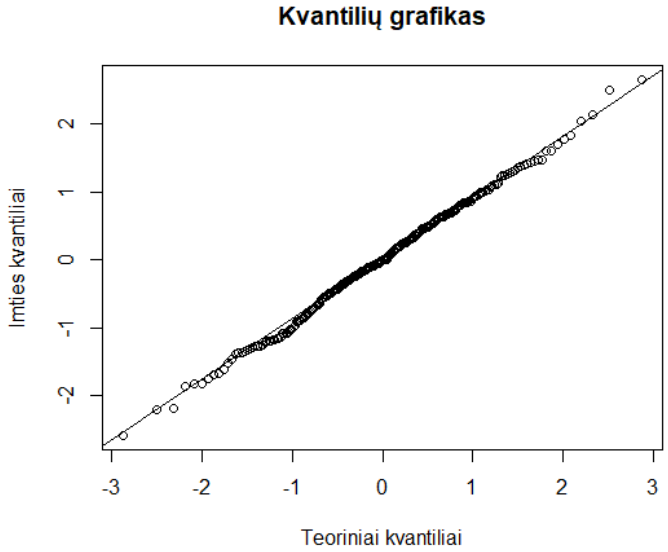
12 pav. Gama modelio (su pašalintom 2 išskirtim) išskirčių, tiesinio prediktoriaus ir kvantilių grafikai

Kvantilių grafikas pagerėjo, susitvarkė vieno galo nukrypimai. Matome, kad iš standartizuotų liekanų grafiko yra dar viena išskirtis, o pagal Kuko matą išskirčių nėra. Tiesinis prediktorius smarkiai nepakito. Taip pat atliekame Šapiro – Vilko testą:

```
        Shapiro-Wilk normality test

data:  qresid(gamma.log2)
W = 0.99618, p-value = 0.812
```

Gauname, kad p reikšmė pasikeitė ir yra didesnė už reikšmingumo lygmenį, todėl nulinės hipotezės neatmetame. Paklaidos yra pasiskirsčiusios pagal normalųjį skirstinį.

Išmetus paskutinę išskirtį pagal standartizuotas liekanas gauname modelį, kurio AIC nukrito iki 5134,7, susitvarkė kvantilių grafiko nukrypimai galuose.



13 pav. Gama modelio (be išskirčių) kvantilių grafikas

Taikant pažingsninę regresiją atrenkame reikšmingas kovariantes. Modelio AIC nukrito iki 5102. Atrinktos reikšmingos kovariantės buvo CitizenDescUS (ar asmuo yra JAV pilietis), Department (departamentas, kuriame asmuo dirba), PerformanceScore (jo veiklos įvertinimas), SpecialProjectsCount (kiek projektų yra priskirta asmeniui), Position_merged (pozicijos lygis) bei PerformanceScore ir SpecialProjectsCount sąveika.

```
Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                                         12.175335   0.100560 121.075  < 2e-16 ***
CitizenDescUS                                       -0.055759   0.036767  -1.517 0.130772
DepartmentIT/IS                                      0.171337   0.054997   3.115 0.002074 **
DepartmentProduction                                -0.100291   0.064612  -1.552 0.122006
DepartmentSales                                     -0.237546   0.072769  -3.264 0.001266 **
DepartmentSoftware Engineering                       0.210237   0.060108   3.498 0.000565 ***
PerformanceScoreFully Meets                         -0.010336   0.027731  -0.373 0.709691
PerformanceScoreNeeds Improvement                   -0.002566   0.042091  -0.061 0.951444
PerformanceScorePIP                                 -0.049785   0.049457  -1.007 0.315188
SpecialProjectsCount                                -0.008587   0.016792  -0.511 0.609584
Absences                                             0.003032   0.001341   2.261 0.024706 *
Position_mergedJunior                               -1.129119   0.067237 -16.793  < 2e-16 ***
Position_mergedManager                              -0.810712   0.069550 -11.656  < 2e-16 ***
Position_mergedMid-level                            -0.927628   0.067220 -13.800  < 2e-16 ***
Position_mergedSenior                               -0.773007   0.078041  -9.905  < 2e-16 ***
PerformanceScoreFully Meets:SpecialProjectsCount     0.012050   0.011211   1.075 0.283594
PerformanceScoreNeeds Improvement:SpecialProjectsCount  0.058782   0.021071   2.790 0.005723 **
PerformanceScorePIP:SpecialProjectsCount             0.174523   0.127771   1.366 0.173323
---
```

Patikrinus, ar modelyje yra multikolinearių kovariančių gauname:

```
                                        GVIF Df GVIF^(1/(2*Df))
CitizenDesc                         1.104693  1        1.051044
Department                         41.755629  4        1.594371
PerformanceScore                    1.821912  3        1.105150
SpecialProjectsCount               25.549748  1        5.054676
Absences                            1.079597  1        1.039037
Position_merged                     4.681821  4        1.212835
PerformanceScore:SpecialProjectsCount 15.270566  3        1.575104
```

Matome, kad SpecialProjectsCount yra multikolineari kovariantė. Tačiau taip yra todėl, kad yra įtraukta šios ir PerformanceScore kovariančių sąveika. Patikrinę modelio beta koeficientų pasikliovimo intervalus pastebėjome, kad intervalai nėra nelogiškai dideli, todėl teigėme, kad tai neiškraipė modelio interpretacijos.

```
                                        (Intercept)
                                          193946.0269
                                       DepartmentIT/IS
                                                1.1869
                                       DepartmentSales
                                                0.7886
                                PerformanceScoreFully Meets
                                                0.9897
                                      PerformanceScorePIP
                                                0.9514
                                             Absences
                                                1.0030
                                  Position_mergedManager
                                                0.4445
                                   Position_mergedSenior
                                                0.4616
PerformanceScoreNeeds Improvement:SpecialProjectsCount
                                                1.0605
```

```
                                        CitizenDescUS
                                                0.9458
                                  DepartmentProduction
                                                0.9046
                              DepartmentSoftware Engineering
                                                1.2340
                          PerformanceScoreNeeds Improvement
                                                0.9974
                                   SpecialProjectsCount
                                                0.9914
                                  Position_mergedJunior
                                                0.3233
                                Position_mergedMid-level
                                                0.3955
PerformanceScoreFully Meets:SpecialProjectsCount
                                                1.0121
                                PerformanceScorePIP:SpecialProjectsCount
                                                1.1907
```

Galiausiai atliekame modelio interpretaciją. Matome, kad jei žmogus yra JAV pilietis, tai jo alga sumažėja 5,4 %. Jei asmuo dirba IT/IS, programinės įrangos inžinerijos departamentuose, tai jo

atlyginimas (lyginant su administracijos departamentu) atitinkamai padidėja 18,7 % ir 23,4 %. Kita vertus, darbas gamybos ir pardavimų departamentuose metinę algą atitinkamai sumažina 9,5 % ir 21,1 %. Kitas gana žymus algos sumažėjimas (4,9 %) įvyksta, kai asmuo yra PIP (angl. Performance improvement plan) veiklos įvertinimo grupėje (lyginant su tais, kurių veikla viršija darbdavio lūkesčius). Kai darbuotojas yra jaunesnysis, vidutinis, vyresnysis ekspertas, tai jo alga (lyginant su direktoriais) atitinkamai sumažėja 67,7 %, 60,4 %, 53,8 %. Kai darbuotojas yra vadybininkas, tai jo alga sumažėja 55,5 %. Kai asmens praleistas darbo dienų skaičius padidėja vienetu, tai jo atlyginimas padidėja 0,3 %. Ši išvada gali pasirodyti keista, tačiau taip yra todėl, kad didžioji dalis į darbą nevaikštančių asmenų yra iš aukštesnių pozicijų (vyr. ekspertas, vadybininkas, direktorius). Kai darbuotojas yra PIP veiklos įvertinimo grupėje ir jo spec. projektų skaičius padidėja vienetu, tai jo atlyginimas padidėja 19,1 % (lyginant su viršijančiu lūkesčius). Taip gali būti todėl, nes PIP yra planas darbuotojus skatinti gerinti savo darbo kokybę, tad vienas iš potencialių paskatinimų yra pakelta alga.

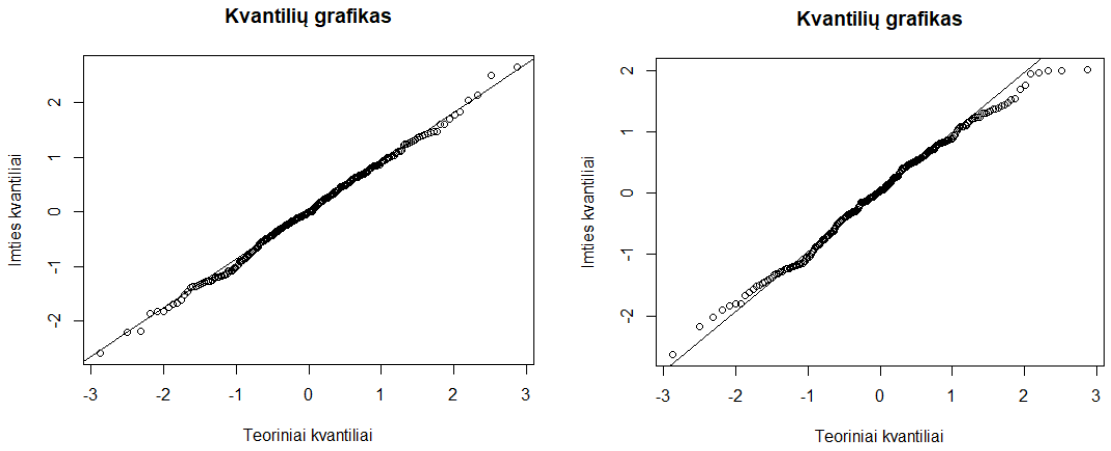
Tą pačią analizę atlikome ir taikant atvirkštinę Gauso regresiją. Tačiau jungties funkcija, išskirtys, reikšmingos kovariantės sutapo su gama modeliu, todėl apraše šios informacijos neįtraukiame. Šio modelio gauti koeficientai:

(Intercept)	CitizenDescUS
195022.6086	0.9465
DepartmentIT/IS	DepartmentProduction
1.1981	0.9193
DepartmentSales	DepartmentSoftware Engineering
0.7941	1.2671
PerformanceScoreFully Meets	PerformanceScoreNeeds Improvement
0.9893	0.9984
PerformanceScorePIP	SpecialProjectsCount
0.9542	0.9930
Absences	Position_mergedJunior
1.0029	0.3179
Position_mergedManager	Position_mergedMid-level
0.4372	0.3857
Position_mergedSenior	PerformanceScoreFully Meets:SpecialProjectsCount
0.4594	1.0101
PerformanceScoreNeeds Improvement:SpecialProjectsCount	PerformanceScorePIP:SpecialProjectsCount
1.0565	1.1829

Matome, kad jei žmogus yra JAV pilietis, tai jo alga sumažėja 5,3 % . Jei asmuo dirba IT/IS, programinės įrangos inžinerijos departamentuose, tai jo atlyginimas (lyginant su administracijos departamentu) atitinkamai padidėja 19,8 % (18,7) ir 26,7 % (23,4). Kita vertus, darbas gamybos ir pardavimų departamentuose metinę algą atitinkamai sumažina 8,1 % (9,5) ir 20,6 % (21,1). Kitas gana žymus algos sumažėjimas (4,6 % (4,9)) įvyksta, kai asmuo yra PIP (angl. Performance improvement plan) veiklos įvertinimo grupėje (lyginant su tais, kurių veikla viršija darbdavio lūkesčius). Kai darbuotojas yra jaunesnysis, vidutinis, vyresnysis ekspertas, tai jo alga (lyginant su direktoriais) atitinkamai sumažėja 68,2 % (67,7), 61,4 % (60,4), 54,1 % (53,8). Kai darbuotojas yra vadybininkas, tai jo alga sumažėja 56,3 % (55,5). Kai asmens praleistas darbo dienų skaičius padidėja vienetu, tai jo atlyginimas padidėja 0,3 % (0,3). Kai darbuotojas yra PIP veiklos įvertinimo grupėje ir jo spec. projektų skaičius padidėja vienetu, tai jo atlyginimas padidėja 18,3 % (19,1) (lyginant su viršijančiu lūkesčius).

Pastaba: skliaustuose buvo **gama** regresijos koeficientų rezultatai.

Galiausiai atlikome modelių palyginimą. Tam naudojome kelis kriterijus. Pirma tikrinome kvantilių grafikus:



14 pav. Gama (kairėje) ir atvirkštinio Gauso modelių kvantilių grafikai

Matome, kad gama modelio kvantilių grafikas yra arčiau tiesės su mažesniais nukrypimais galuose. Taip pat vertinome pagal AIC:

- Gamos modelios AIC – 5102.919
- Atvirkštinio gauso regresijos modelio AIC – 5094,919

Šį kartą geriau įvertinamas yra atvirkštinis Gauso modelis (tačiau nežymiai). Galiausiai tikriname pagal testavimo duomenų vidutinę absoliučiąją paklaidą (MAE) ir vidutinę kvadratinę paklaidą (RMSE):

- MAE gama modelis – 7759,806;
- MAE atvirkštinis gauso modelis – 7797,175;
- RMSE gama modelis – 11936,76;
- RMSE atvirkštinis gauso modelis – 12027,78.

Matome, kad mažesnės paklaidos yra gama modelyje. Iš šių rezultatų kaip tinkamiausią modelį pasirinktume gama.

Išgyvenamumo analizė

Pirmiausia sumažiname dviejų kategorinių kintamųjų skirtingų grupių skaičių, kadangi tęsiant analizę be šių pakeitimų, atsiranda pilno atskyrimo problema – tam tikrose grupėse lieka tik tokie darbuotojai, kurie neišėjo ar nebuvo išmesti iš darbo. Kintamasis „RecruitmentSource“ apjungiamas paliekant 3 dažniausiai pasitaikiusias reikšmes, o kitas perkeliant į grupę „Other“. Toks pat apjungimas atliekamas ir su „MaritalDesc”.

Toliau taikome modelį su visomis kovariantėmis:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
Salary	1.964e-06	1.000e+00	6.237e-06	0.315	0.75288	
SexM	-9.298e-03	9.907e-01	2.139e-01	-0.043	0.96533	
Marital_joinedMarried	-4.458e-01	6.403e-01	3.087e-01	-1.444	0.14869	
Marital_joinedSingle	-1.044e+00	3.522e-01	3.232e-01	-3.229	0.00124	**
Marital_joinedOther	-9.712e-01	3.786e-01	5.227e-01	-1.858	0.06319	.
CitizenDescUS	-1.743e-01	8.401e-01	4.136e-01	-0.421	0.67351	
HispanicLatinoYes	9.116e-02	1.095e+00	3.663e-01	0.249	0.80347	
RaceDescBlack or African American	4.995e-01	1.648e+00	4.187e-01	1.193	0.23290	
RaceDescTwo or more races	-7.553e-02	9.272e-01	6.910e-01	-0.109	0.91296	
RaceDescWhite	2.458e-01	1.279e+00	3.757e-01	0.654	0.51286	
DepartmentIT/IS	1.029e+00	2.798e+00	1.007e+00	1.022	0.30661	
DepartmentProduction	-2.635e-01	7.684e-01	8.678e-01	-0.304	0.76144	
DepartmentSales	-1.214e+00	2.971e-01	9.604e-01	-1.264	0.20632	
DepartmentSoftware Engineering	1.005e+00	2.732e+00	9.132e-01	1.101	0.27105	
RSource_joinedIndeed	-8.208e-01	4.401e-01	3.116e-01	-2.634	0.00843	**
RSource_joinedLinkedIn	-1.007e+00	3.655e-01	3.128e-01	-3.218	0.00129	**
RSource_joinedOther	-6.485e-01	5.228e-01	2.969e-01	-2.184	0.02897	*
PerformanceScoreFully Meets	5.795e-01	1.785e+00	3.848e-01	1.506	0.13207	
PerformanceScoreNeeds Improvement	-3.098e-01	7.336e-01	1.067e+00	-0.290	0.77166	
PerformanceScorePIP	-4.485e-01	6.386e-01	1.380e+00	-0.325	0.74517	
EngagementSurvey	1.024e-01	1.108e+00	1.614e-01	0.635	0.52572	
EmpSatisfaction	2.209e-02	1.022e+00	1.203e-01	0.184	0.85430	
SpecialProjectsCount	-3.184e-01	7.273e-01	1.730e-01	-1.840	0.06580	.
DaysLateLast30	3.719e-01	1.451e+00	2.279e-01	1.632	0.10273	
Absences	3.053e-02	1.031e+00	1.796e-02	1.700	0.08908	.
Age	1.827e-02	1.018e+00	1.176e-02	1.553	0.12035	

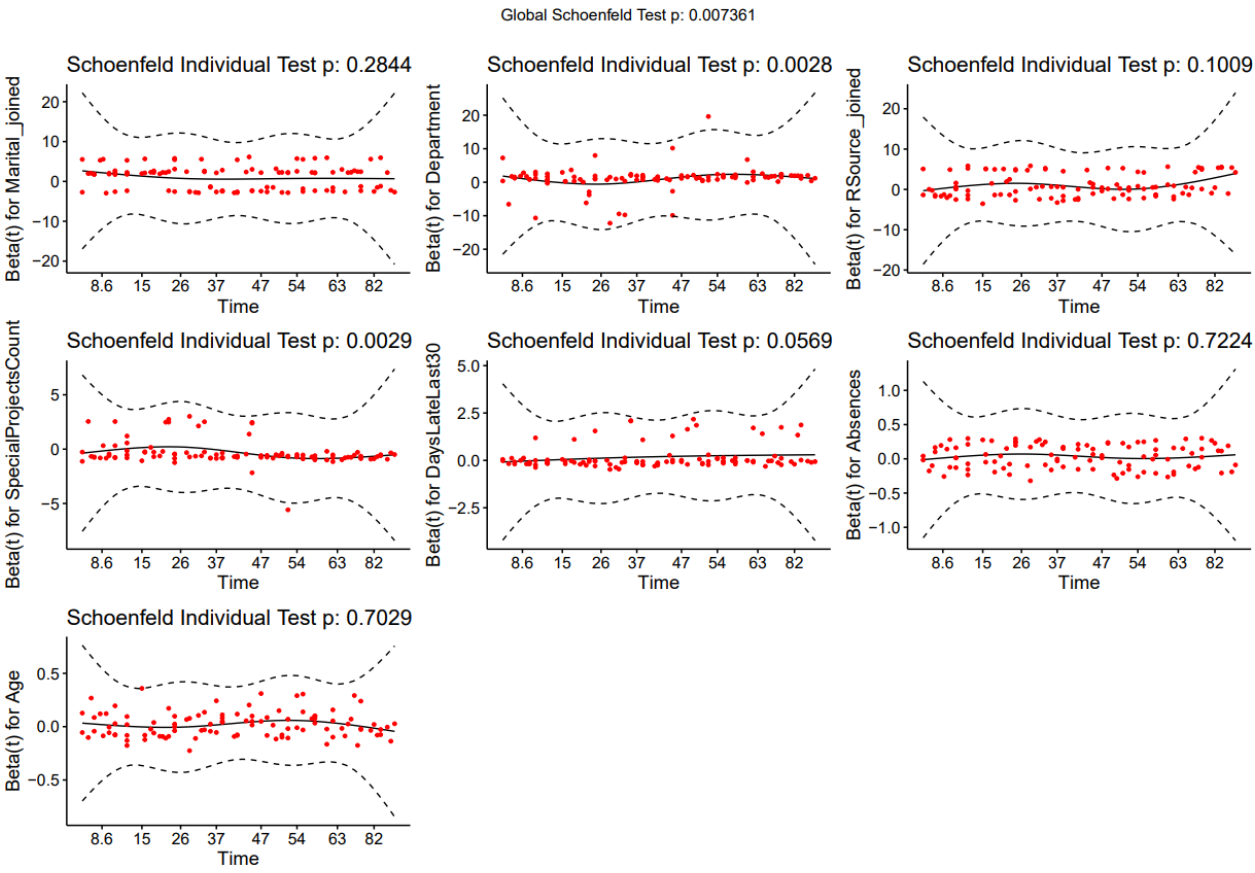
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Matome, kad ne visos kovariantės reikšmingos, todėl taikome pažingsninę regresiją ir gauname modelį su septyniomis reikšmingomis kovariantėmis: šeimyninė padėtis, departamentas, atrankos šaltinis, per kurį darbuotojas buvo rastas, specialiųjų projektų skaičius, vėlavimų skaičius per paskutines 30 dienų, neatvykimai į darbą ir darbuotojo amžius.

Tikriname proporcingosios rizikos prielaidą:

	chisq	df	p
Marital_joined	3.796	3	0.2844
Department	16.175	4	0.0028
RSource_joined	6.231	3	0.1009
SpecialProjectsCount	8.893	1	0.0029
DaysLateLast30	3.625	1	0.0569
Absences	0.126	1	0.7224
Age	0.146	1	0.7029
GLOBAL	30.114	14	0.0074

Tiek iš testo p-reikšmių, tiek iš grafikų (žr. 15 13 pav.) matome, kad kovariantės „Department“ ir „Special Projects Count“ netenkina šios prielaidos – rizikos funkcijų santykis kinta laike.



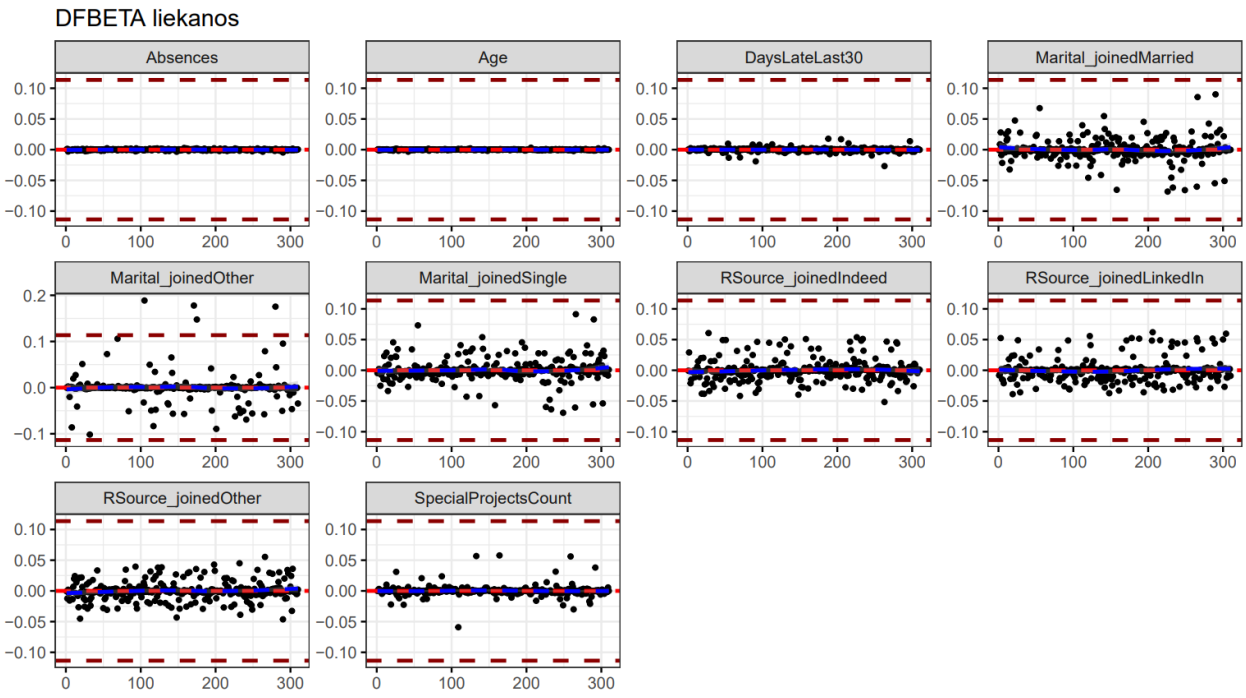
15 13 pav. Schoenfeld liekanos proporcingos rizikos prielaidai tikrinti.

Kad išspręstume šią problemą ir patenkintume proporcingosios rizikos prielaidą naudojame „strata“ funkciją, t.y. skirstome duomenis į sluoksnius. Pirmiausia tai padarome „Department“ kovariantei ir tada pakartotinai patikriname proporcingosios rizikos prielaidą.

```
                      chisq df     p
Marital_joined      3.06741  3 0.381
RSource_joined      6.93512  3 0.074
SpecialProjectsCount 0.45572  1 0.500
DaysLateLast30      2.81315  1 0.093
Absences            0.01139  1 0.915
Age                 0.00408  1 0.949
GLOBAL             13.68472 10 0.188
```

Kaip galime matyti, problema susitvarkė, proporcingosios rizikos prielaida tenkinama – rizikos funkcijų santykis nekinta laike.

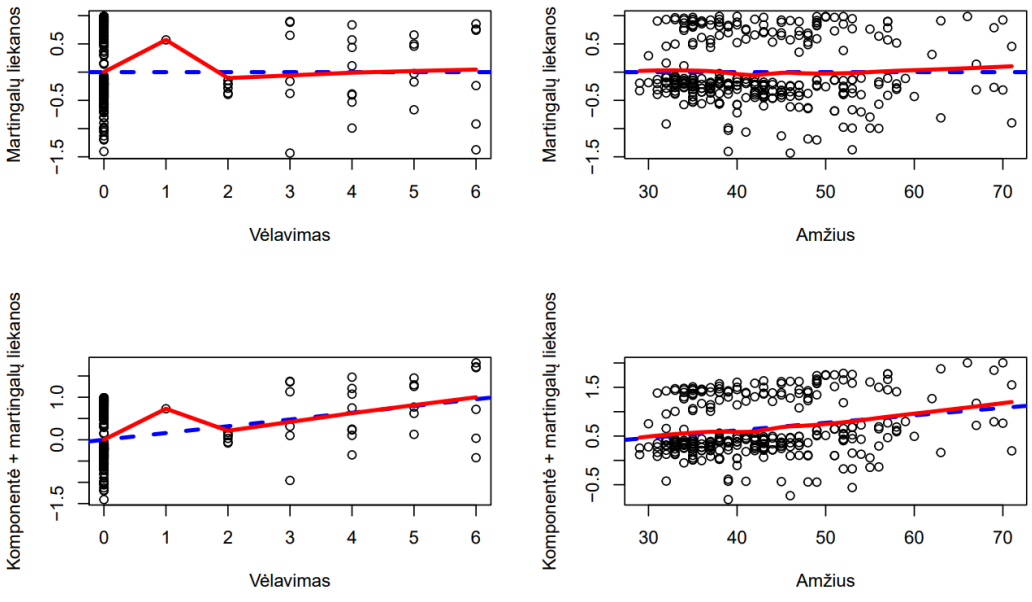
Pereiname prie išskirčių tikrinimo. Išskirčių riba apskaičiuojama pagal formulę 2/sqrt(n), kur n – stebėjimų skaičius. Gauname, kad stebėjimai turi pakliūti į intervalą [-0.1136; 0.1136].



1416 pav. DFBETA liekanos

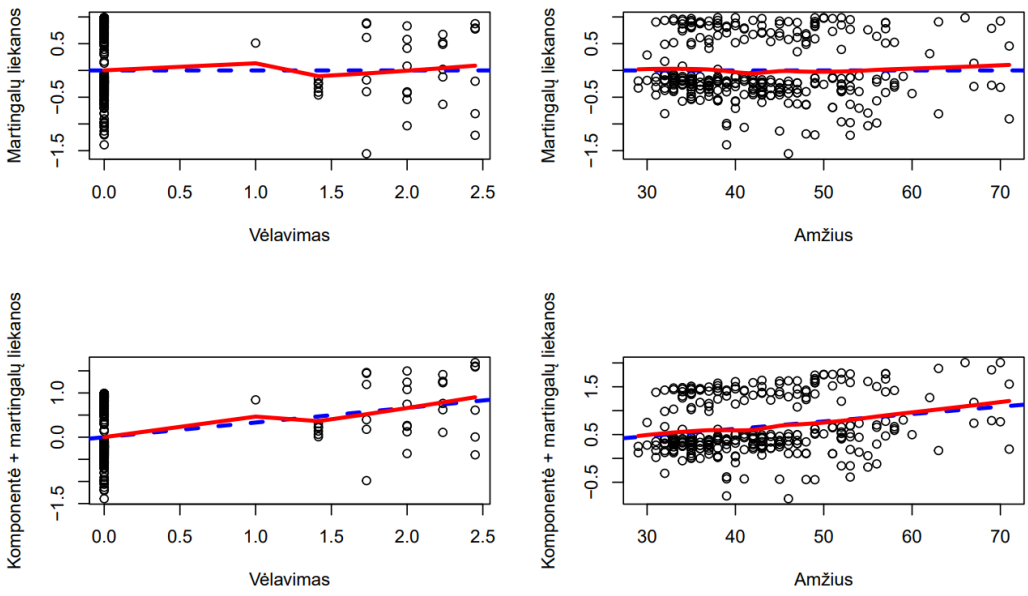
Iš grafiko (žr. 1416 pav.) matome, kad yra keturios išskirtys. Stebėjimai išsiskiria, kadangi pagal kintamąjį „Marital_joined“ grupėje „Other“ yra tik keturi darbuotojai, kurie išėjo iš darbo – jie rodomi kaip išskirtys. Visgi, pašalinus šiuos stebėjimus vėl susidurtume su pilno atskyrimo problema, todėl nusprendžiame juos palikti.

Toliau pagal martingalų liekanas kiekybiniams kintamiesiems tikriname tiesiškumą:



1517 pav. Tiesiškumo tikrinimas

Iš grafiko matome, kad yra tiesinis sąryšis su kintamuoju – „Amžius“, tačiau kovariantę „Vėlavimas“ reiktų koreguoti. Pabandome šį kintamąjį transformuoti panaudojant kvadratinę šaknį.



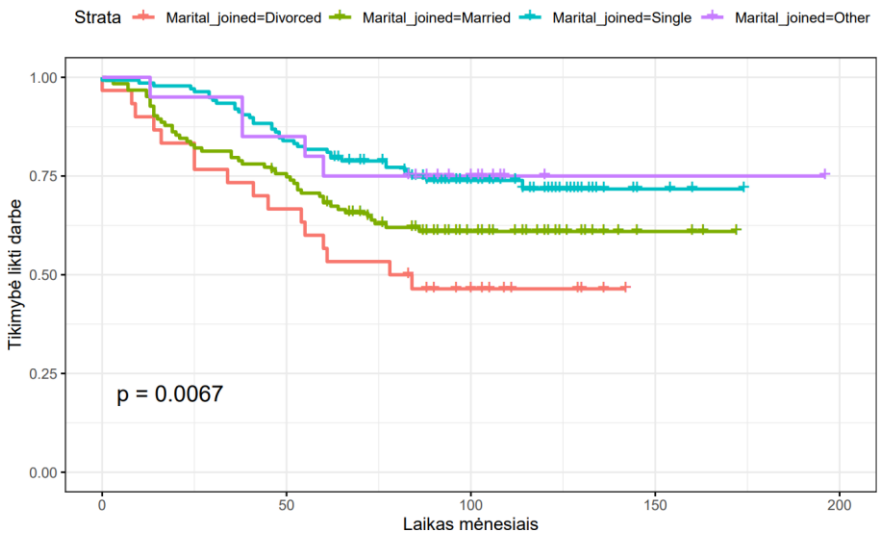
16 pav. Tiesiškumo tikrinimas transformavus kintamąjį Amžius.

Matome, kad rezultatai pagerėjo (žr. 1517 pav.), todėl modelyje paliekam transformuotą kintamąjį „Vėlavimas“.

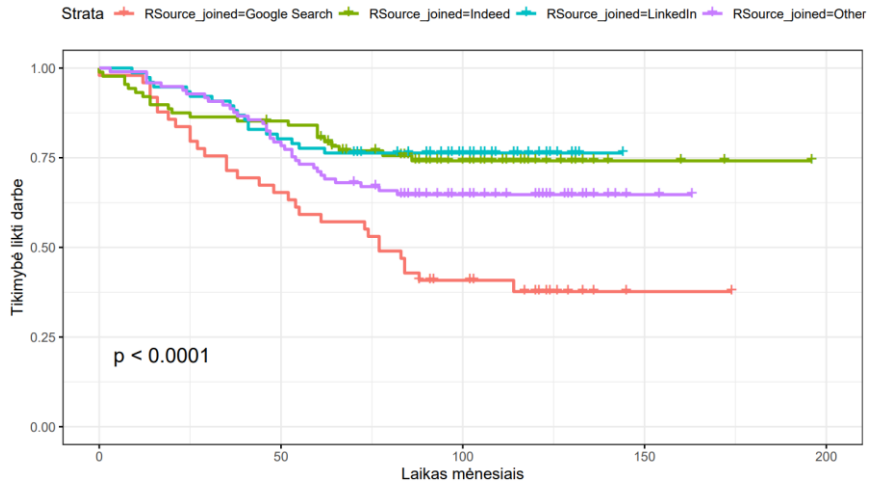
Galutinis modelis:

```
                          coef exp(coef) se(coef)      z Pr(>|z|)
Marital_joinedMarried  -0.29748   0.74268  0.29525 -1.008  0.31365
Marital_joinedSingle   -0.88330   0.41342  0.31243 -2.827  0.00470 **
Marital_joinedOther    -0.88097   0.41438  0.51415 -1.713  0.08663 .
RSource_joinedIndeed   -0.86983   0.41902  0.28994 -3.000  0.00270 **
RSource_joinedLinkedIn -0.98693   0.37272  0.30281 -3.259  0.00112 **
RSource_joinedOther    -0.57021   0.56541  0.26615 -2.142  0.03216 *
DaysLateLast30_mod      0.33202   1.39378  0.13779  2.410  0.01597 *
Age                     0.01548   1.01560  0.01097  1.411  0.15821
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Kad pažiūrėtume, kaip likimo darbe tikimybė atrodo tarp skirtingų grupių, kiekvienam kategoriniam kintamajam nusibraižome grafikus:



17 pav. Tikimybė likti darbe pagal šeimyninę padėtį.



18 pav. Tikimybė likti darbe pagal įdarbinimo būdą.

Matome, kad mažiausia tikimybė likti darbe yra išsiskyrusiems ir įsidarbinusiems „Google Search“ pagalba. Mažiausia rizika – grupėje „Other“ esantiems ir įsidarbinusiems per LinkedIn platformą.

Kadangi nepašalinome išskirčių, tam, kad neturėtume pilno atskyrimo problemos, kadangi kai kuriose šeimyninę padėtį nusakančiose grupėse buvo labai mažai žmonių išėjusių ar išmestų iš darbo, o visi šie stebėjimai priskirti išskirtims, bandome taikyti ir modelį, kuriame nebūtų įtraukta kovariantė – šeimyninė padėtis.

Pirminiame modelyje naudojame tokius kintamuosius: „Department“, „RSource_joined“, „DaysLateLast30“, „Age“.

```
                                  coef exp(coef) se(coef)      z Pr(>|z|)
DepartmentIT/IS                -0.10046   0.90442  0.78156 -0.129 0.897727
DepartmentProduction            0.37681   1.45763  0.73411  0.513 0.607752
DepartmentSales                -0.60121   0.54815  0.84685 -0.710 0.477739
DepartmentSoftware Engineering  0.28581   1.33083  0.87621  0.326 0.744286
RSource_joinedIndeed           -0.80519   0.44700  0.28993 -2.777 0.005483 **
RSource_joinedLinkedIn         -1.00523   0.36596  0.30252 -3.323 0.000891 ***
RSource_joinedOther            -0.47805   0.61999  0.26158 -1.828 0.067616 .
DaysLateLast30                  0.13455   1.14402  0.06140  2.192 0.028413 *
Age                             0.01504   1.01515  0.01091  1.378 0.168112
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

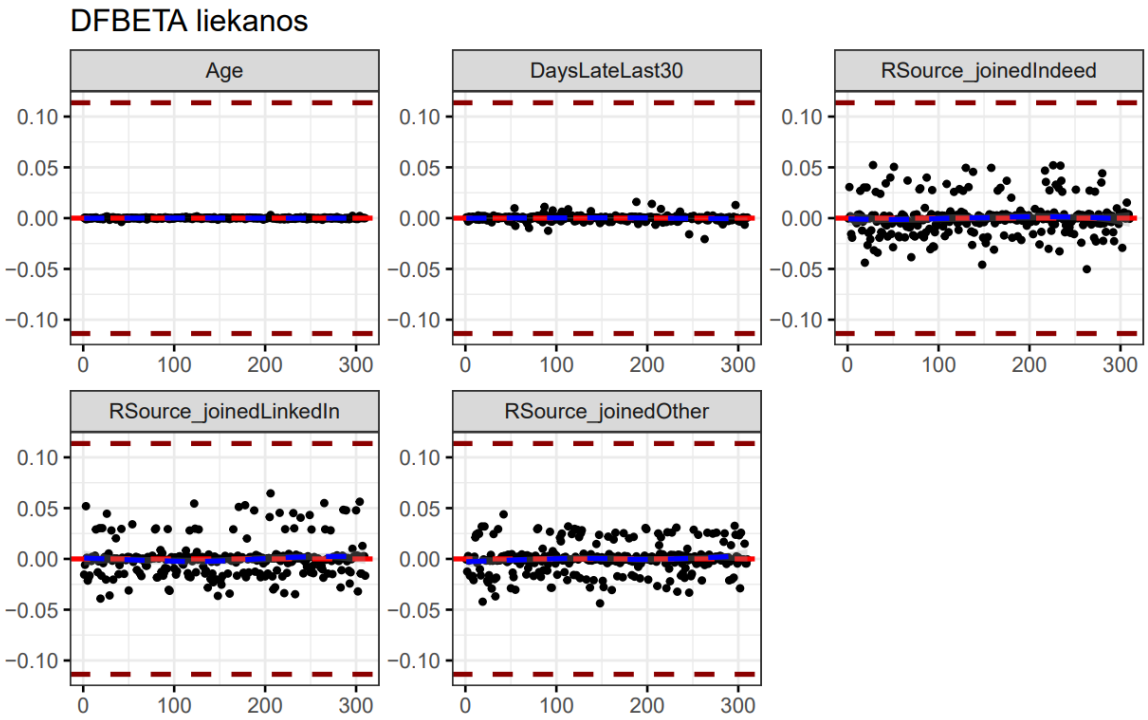
Partikriname proporcingosios rizikos prielaidą:

```
                chisq df      p
Department     15.399  4 0.0039
RSource_joined  5.556  3 0.1354
DaysLateLast30  4.309  1 0.0379
Age             0.188  1 0.6646
GLOBAL         25.359  9 0.0026
```

Gauname, kad kovariantės „Department“ ir „DaysLateLast30“ prielaidos netenkina, naudojame funkciją strata, kad išspręstume problemą. Pirmiausia ją pritaikome „Department“ kovariantei ir patikriname prielaidą dar kartą.

```
                chisq df     p
RSource_joined 6.0121  3 0.111
DaysLateLast30 3.4034  1 0.065
Age            0.0117  1 0.914
GLOBAL         9.7940  5 0.081
```

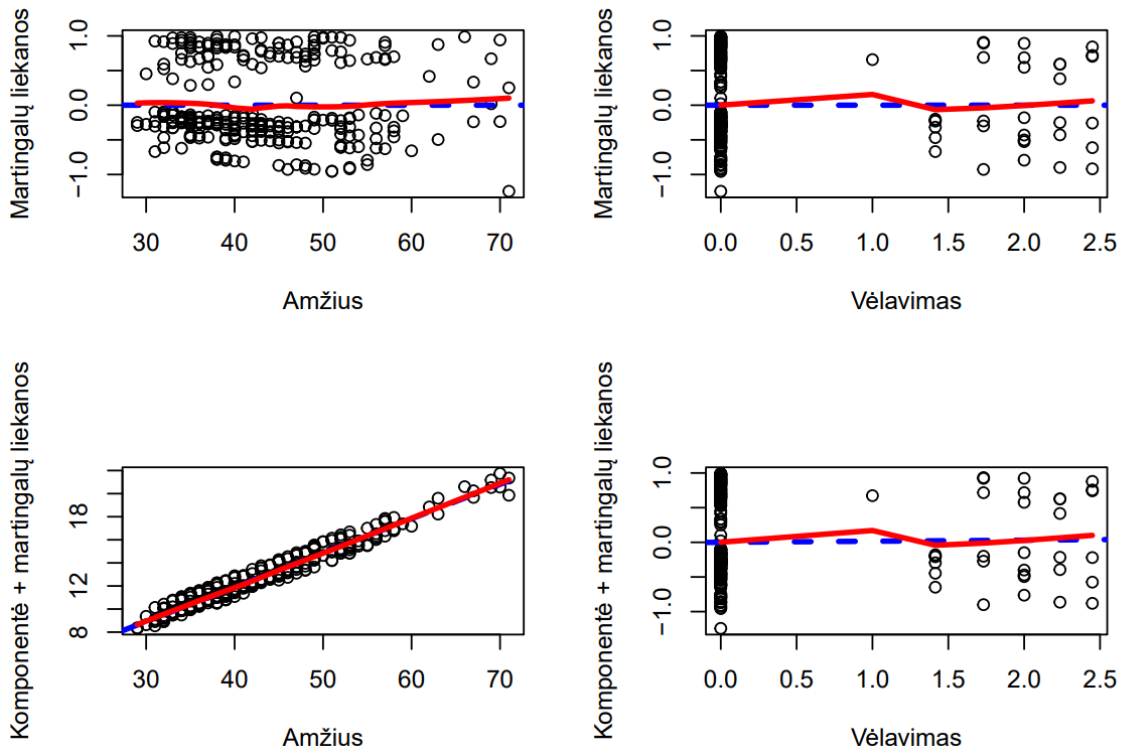
Problema išsisprendžia. Toliau tikriname išskirtis. Išskirčių riba vėl apskaičiuojama pagal formulę 2/sqrt(n). Kaip ir prieš tai – stebėjimai turi pakliūti į intervalą [-0.1136; 0.1136].



19 pav. DFBETA liekanos

Išskirčių nėra, liekanos pasiskirsčiusios apie 0.

Tam, kad išlaikytume tiesiškumą vėl naudojame transformuotą kintamojo „DaysLateLast30“ reikšmę – panaudojame kvadratinę šaknį.



20 pav. Tiesiškumo tikrinimas.

Kaip ir prieš tai buvusiame modelyje matome, kad tiesiškumas yra išlaikomas.

Galutinio modelio koeficientai:

```
                          coef exp(coef) se(coef)      z Pr(>|z|)
RSource_joinedIndeed   -0.79623   0.45103  0.28876 -2.757 0.005825 **
RSource_joinedLinkedIn -1.00886   0.36464  0.30267 -3.333 0.000859 ***
RSource_joinedOther    -0.46615   0.62742  0.26050 -1.789 0.073546 .
DaysLateLast30_mod      0.29720   1.34608  0.13535  2.196 0.028110 *
Age                     0.01538   1.01550  0.01099  1.399 0.161704
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                       exp(coef) exp(-coef) lower .95 upper .95
RSource_joinedIndeed      0.4510     2.2172    0.2561    0.7943
RSource_joinedLinkedIn    0.3646     2.7425    0.2015    0.6599
RSource_joinedOther       0.6274     1.5938    0.3765    1.0454
DaysLateLast30_mod        1.3461     0.7429    1.0324    1.7550
Age                       1.0155     0.9847    0.9939    1.0376
```

Iš gautų rezultatų galime teigti, kad rizika palikti darbo vietą didėja su amžiumi, tai yra suprantama, nes galiausiai būna išeinama į pensiją ar panašiai. Taip pat didesnę riziką išeiti ar būti išmestam iš darbo turi ir tie, kurie yra linkę dažnai vėluoti. Didžiausia riziką turi darbuotojai, kurie buvo įdarbinti su „Google Search“ pagalba (tai lyginamoji grupė), kadangi visų kitų grupių koeficientai yra mažesni už 1.

Išvados

Iš binarinio atsako modelio pastebėjome, kad tikimybė būti atleistam yra didesnė asmenims, kurie yra ne JAV piliečiai, taip pat darbuotojų, kurių veiklos įvertinimas yra blogiausias net 4029 karto padidina galimybę būti išmestam iš darbo. Darbuotojo vykdomų specialiųjų projektų skaičius bei ilgesnis darbo laikas mėnesiais sumažina tikimybę būti atleistam, o su kiekvienu neatvykimu į darbą, tikimybė būti atleistam padidėja 1,17 karto.

Taikant glm modelius darbuotojų atlyginimams suskaičiuoti gavome, kad geriausiai veikia gama modelis su kovariantėmis CitizenDescUS (ar asmuo yra JAV pilietis), Department (departamentas, kuriame asmuo dirba), PerformanceScore (jo veiklos įvertinimas), SpecialProjectsCount (kiek projektų yra priskirta asmeniui), Position_merged (pozicijos lygis) bei PerformanceScore ir SpecialProjectsCount sąveika. Jei asmuo dirba IT sferoje, tai jo atlyginimas (lyginant su administracijos departamentu) didėja. Kita vertus, darbas gamybos ir pardavimų departamentuose metinę algą mažina. Pagal darbo ekspertizę algos eina (didėjančiai) jaunesniųjų, vidutinių ekspertų, vadybininkų, vyresniųjų ekspertų, direktorių.

Iš išgyvenimo analizės gautų rezultatų galime teigti, kad rizika palikti darbo vietą didėja su amžiumi, tai yra suprantama, nes galiausiai būna išeinama į pensiją ar panašiai. Taip pat didesnę riziką išeiti ar būti išmestam iš darbo turi ir tie, kurie yra linkę dažnai vėluoti, tai galėjome pastebėti ir taikant binarinio atsako modelį. Didžiausią riziką turi darbuotojai, kurie buvo įdarbinti su „Google Search“ pagalba. Per LinkedIn įsisdarbinę asmenys turi mažiausią riziką būti pašalinti iš darbo.