VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Bakalaurinis darbas

Ne gyvybės draudimo veiklos modeliavimas ir prognozavimas

Modeling and Forecasting of Non-Life Insurance Activity

Mindaugas Stackevičius

VILNIUS 2016

MATEMATIKOS IR INFORMATIKOS FAKULTETAS
EKONOMETRINĖS ANALIZĖS KATEDRA

Darbo vadovas: prof. Remigijus Leipus _____

Darbas apgintas 2016.06.10

Registravimo NR. _____

Turinys

Ne gyvybės draudimo veiklos modeliavimas ir prognozavimas

ANOTACIJA

Bakalaurinio darbo tikslas – taikant matematinius, statistinius bei ekonometrinius metodus atlikti išsamią analizę bei surasti tinkamą modelį draudimo išmokų sumai, draudimo išmokų skaičiui bei rizikai prognozuoti, turint draudimo įmonės duomenis. Šiam tikslui pasiekti taikomas apibendrintasis tiesinis modelis (angl. Generalized Linear Model). Teorinėje dalyje apžvelgiama pastarojo modelio specifika ir rezultatų analizė bei geriausio modelio parinkimo metodika. Praktinėje dalyje įgytas žinias nagrinėjant pritaikysime GLM modelius ir ieškosime geriausio varianto turimiems duomenims.

Raktiniai žodžiai: ne gyvybės draudimas, GLM modelis, Puasono skirstinys, Gama skirstinys.

Modeling and Forecasting of Non-Life Insurance Activity

SUMMARY

The goal of this bachelor thesis - complete thorough analysis and find best fitting model for sum of claims paid, number of claims and risk forecasting by using mathematical, statistical and econometrical methods for given insurance company‘s data. In order to achieve this goal, Generalized Linear Model will be used. In theoretical part, model specifics, result’s analysis and best model gaining methods will be reviewed. After that, gained knowledge for GLM model analysis will be practically used to achieve best possible model for available data.

Key words: Non-life insurance, GLM model, Poisson distribution, Gamma distribution.

1 ĮVADAS

Draudimo įmonės, norėdamos tikslingai įvertinti turimų klientų srautą bei galimas rizikas, apie kiekvieną esamą kontrakto savininką surenka duomenis, kuriuos kaupia duomenų bazėse. Vėliau iš esamų duomenų kuriami matematiniai-statistiniai modeliai, kurių pagalba bandoma įvertinti būsimų klientų riziką ir atitinkamą kainodarą, siekiant maksimizuoti pelną. Kadangi turimi duomenys gali būti kategoriniai, o taip pat juose yra ir diskrečių dydžių, tai, tikėtina, duomenys nėra normalūs, todėl įprasta duomenų analizė su tiesinėmis regresijomis nebus tinkama dėl griežtų modelio sąlygų. Laikui bėgant modeliavimo specifika kinta dėl duomenų kiekio gausos, todėl 1972 m. Nelder ir Wedderburn straipsnyje buvo pasiūlytas Apibendrintasis tiesinis modelis.

Apibendrintasis Tiesinis Modelis (toliau tekste bus vartojamas jo angliškas trumpinys – GLM) yra naudojamas draudimo industrijoje pagrįsti reikšmingus ar net kritinius sprendimus. Šio modelio kūrimo paskirtis - nustatyti ryšius tarp kintamųjų. GLM modelis apibendrina klasikinį tiesinį modelį, supaprastindamas keletą griežtų reikalavimų, taip suteikiant galimybę atlikti normalumo prielaidų netenkinančių duomenų analizę.

Draudimo duomenys, su kuriais bus dirbama šiame bakalauriniame darbe, susideda iš daugybės kategorinių kintamųjų, kurie yra kliento amžius, patirtis, asmenų skaičius, besinaudojantis transporto priemone, transporto priemonės amžius, transporto grupė, vairavimo teritorija, transporto nuosavybė, transporto variklio tūris, kliento rūšis, registravimo teritorija, o taip pat tokie dydžiai kaip pasirašytos draudimo įmokos, uždirbtos draudimo įmokos, draudimo išmokos, patirtų žalų skaičius ir draudimo sutarties trukmė, kai buvo surinkti duomenys. Kadangi duomenys yra konfidencialūs, tai visi kategoriniai dydžiai bus pateikiami numeruota išraiška, o skaičiai neženkliai pakeisti nuo tikrųjų. Duomenų rinkinį su daro tik ne gyvybės draudimas. Detalesnės draudimo įmonių vartojamų sąvokų žodynas su paaiškinimais pateikiamas kitame skyriuje.

Darbo su šiais duomenimis tikslas yra nustatyti kainodarą, pagal kurią, įvertinus asmens riziką, būtų prasminga drausti asmenį, kad žalos išmokėjimo atveju, įmonei tai būtų pelninga.

2 SĄVOKŲ ŽODYNAS[1]

1. **Ne gyvybės draudimas** (angl. Non-life Insurance) - viena iš draudimo šakų (taip pat žr. gyvybės draudimas). Ne gyvybės draudimo šakai priskiriamos šios grupės (pagal Valstybinės draudimo priežiūros tarnybos prie LR finansų ministerijos klasifikaciją):
   1) draudimas nuo nelaimingų atsitikimų;
   2) draudimas ligos atveju;
   3) sausumos transporto priemonių, išskyrus geležinkelio transporto priemones, draudimas;
   4) geležinkelio transporto priemonių draudimas;
   5) skraidymo aparatų draudimas;
   6) laivų (jūrų, ežerų, upių ir kanalų) draudimas;
   7) vežamų krovinių draudimas;
   8) turto (išskyrus 3, 4, 5, 6, 7 punktus) draudimas nuo gaisro ar gamtinių jėgų;
   9) turto draudimas nuo kitų žalų (išskyrus 8 punktą);
   10) sausumos transporto priemonių civilinės atsakomybės draudimas;
   11) skraidymo aparatų civilinės atsakomybės draudimas;
   12) laivų (jūrų, ežerų, upių ir kanalų) civilinės atsakomybės draudimas;
   13) bendrosios civilinės atsakomybės draudimas;
   14) kredito draudimas;
   15) laidavimo draudimas;
   16) finansinių nuostolių draudimas;
   17) teismo išlaidų draudimas;
   18) pagalbos draudimas.
2. **Pasirašytos draudimo įmokos** (angl. Gross Written Premium) – įmokos, kurias draudėjas moka draudikui, pirkdamas tam tikrą draudimo produktą, t. y. pasirašydamas draudimo sutartį. Įmokų dydis nurodytas draudimo polise. Sutarties galiojimo pradžioje pasirašytos įmokos dydis gali nesutapti su gauta draudimo įmoka, pavyzdžiui, jei draudimo įmokos mokamos ne iš karto, o kas mėnesį arba kas ketvirtį.
3. **Uždirbtos draudimo įmokos** (angl. Earned Premium) – įmokos, kurias draudikas, taikydamas tam tikras apskaičiavimo metodikas, gali pripažinti savo pajamomis.

[1] Lietuvos Respublikos Seimas. Draudimo įstatymas. 2003 m.

4. **Draudimo išmoka** (angl. Claim Paid) - išmoka, numatyta atitinkamose draudimo rūšies taisyklėse, kurią, atsitikus draudimo sutartyje nustatytam draudiminiam įvykiui, draudimo įmonė, remdamasi šį įvykį patvirtinančiais oficialiais dokumentais, privalo mokėti draudėjui, apdraustajam, naudos gavėjui, tretiesiems asmenims.
5. **Draudimo sutartis** (angl. Insurance policy) - sutartis, kuria viena šalis (draudikas) įsipareigoja sutartyje nustatytą draudimo įmoką (premiją) sumokėti kitai šaliai (draudėjui) arba trečiajam asmeniui, kurio naudai sudaryta sutartis, draudimo sutartyje nustatytą draudimo išmoką, apskaičiuotą draudimo sutartyje nustatyta tvarka, jeigu įvyksta draudimo sutartyje nustatytas draudiminis įvykis.

3 TEORINĖ DALIS

3.1. Apibendrintasis Tiesinis Modelis[2]

Apibendrintasis tiesinis modelis yra skirtas nustatyti ir įvertinti ryšį tarp priklausomo ir nepriklausomo kintamųjų. Tikslesnis apibrėžimas pateikiamas V. Čekanavičiaus ir G. Murausko *Statistika ir jos taikymai. III dalis* knygoje: „GLM – tai tiesinis modelis, apibūdinantis priklausomojo kintamojo vidurkio funkciją:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K.“$$

Iš tikro, nesunku pastebėti, kad šis modelis yra Tiesinio Normaliojo Modelio apibendrinimas, tačiau jis turi esminių skirtumų, dėl kurių jis yra plačiai naudojamas:

1. Priklausomo kintamojo skirstinys yra parenkamas iš eksponentinių skirstinių šeimos, be to, priklausomas kintamasis neturi būti normalusis ar arti normaliojo atsitiktinio dydžio, bet priešingai, gali visiškai nepasižymėti normaliojo dydžio charakteristikomis (atvirkščiai, negu tiesiniame modelyje).
2. Transformuoto priklausomo kintamojo vidurkio ryšys su nepriklausomais kintamaisiais yra tiesinis, t.y. funkcija g turi būti tolydi ir jai turi egzistuoti atvirkštinė funkcija g^{-1}.

Eksponentinių skirstinių taikymas modelyje dažniausiai padeda išvengti homoskedastiškumo problemos. Be to, dispersija kinta tik su nepriklausomais kintamaisiais. Tai yra dar viena priešingybė normaliajai tiesinei regresijai.

„Skirstinys priklauso eksponentinių skirstinių šeimai, jeigu jo tankis (diskrečiuoju atveju – tikimybė) yra žemiau nurodytos struktūros:

$$f(y) = \exp\{\frac{y\theta - b\theta}{a(\phi)} + c(y, \theta)\}$$

Čia θ, ϕ – parametrai, o $a(\theta), b(\theta), c(y, \phi)$ – žinomos funkcijos“[3]. Kaip matome, lygtyje esanti eksponentinė išraiška nurodo, kad skirstinys bus eksponentinis. Pastaroji, vidurkio funkcija tuomet teigia, kad vidurkio transformacija yra tiesiška su nepriklausomais kintamaisiais. $a(\phi)$ parinkimas nurodo atsako skirtinį. Funkcija $g(\mu)$, vadinama *jungiamąja funkcija* arba tiesiog *jungtimi*, nulemia kaip vidurkis bus siejamas su nepriklausomais kintamaisiais. GLM modelyje g turi būti monotoniška funkcija (tokios kaip logaritminės ar laipsninės funkcijos).

[2] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 64
[3] V. Čekanavičius, G. Murauskas. *Statistika ir jos taikymai III dalis,* Vilnius: TEV, 2009, p. 47

3.2. Skirstinio parinkimas

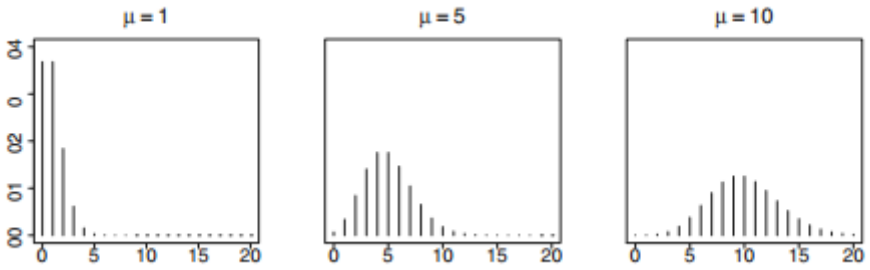
Eksponentinių skirstinių šeimą sudaro tokie diskretūs skirtiniai kaip: Bernulio, binominis, Puasono, neigiamas binominis bei tolydūs skirstiniai: normalusis, χ^2 ir gama bei atvirkštinis Gauso skirstinys. Šie skirstiniai taip pat yra skirstomi į diskrečiuosius ir tolydžiuosius skirstinius. Dėmesį skirsime tik tiems skirstiniams, kurie bus naudojami praktinėje dalyje kaip labiausiai atitinkantys mūsų duomenis.

3.2.1. Puasono skirstinys[4]

Tarkime, kad binominio skirstinio kintamųjų skaičius n yra didelis, tuo tarpu „sėkmės“ tikimybė π tampa maža tokiu būdu, kad $\mu = n\pi$ tampa konstanta. Tuomet tikimybės funkcija yra išreiškiama kaip:

$$f(y) = \frac{e^{-\mu}\mu^y}{y!}, \qquad y = 0,1,2, \dots$$

Ši funkcija apibrėžia Puasono skirstinį, kuris yra užrašomas $y \sim P(\mu)$. Puosono dydis yra diskretus dydis. Tikimybės funkcija priklauso nuo vienintelio parametro μ ir turi $E(y) = \mu$ bei $Var(y) = \mu$. Žemiau esantis grafikas pateikia teorines Puasono skirstinio realizacijas su skirtingomis μ reikšmėmis.



1 pav. Puasono teorinis pasiskirstymas

3.2.2. χ^2 ir Gama skirstinys[5]

χ^2_υ skirstinys susideda iš nepriklausomų atsitiktinių dydžių $\upsilon \sim N(0,1)$ kvadratų sumos ir yra žymimas $y \sim \chi^2_\upsilon$. Parametras υ yra vadinamas laisvės laipsniu. χ^2_υ atsitiktiniai dydžiai yra neneigiami ir jų pasiskirstymas yra asimetriškas į dešinę pusę. Vidurkis ir dispersija yra υ ir 2υ atitinkamai. Didelėms υ reikšmėms, y yra apytiksliai normalusis dydis.

Padaugindami $\chi^2_{2\upsilon}$ atsitiktinį dydį iš $\frac{\mu}{2\upsilon}$ gauname Gama atsitiktinį dydį su parametrais μ ir υ bei žymime $G(\mu, \upsilon)$. Gama atsitiktiniai dydžiai yra tolydūs, neneigiami ir asimetriški į dešinę pusę su galimai didelėmis reikšmėmis uodegoje.

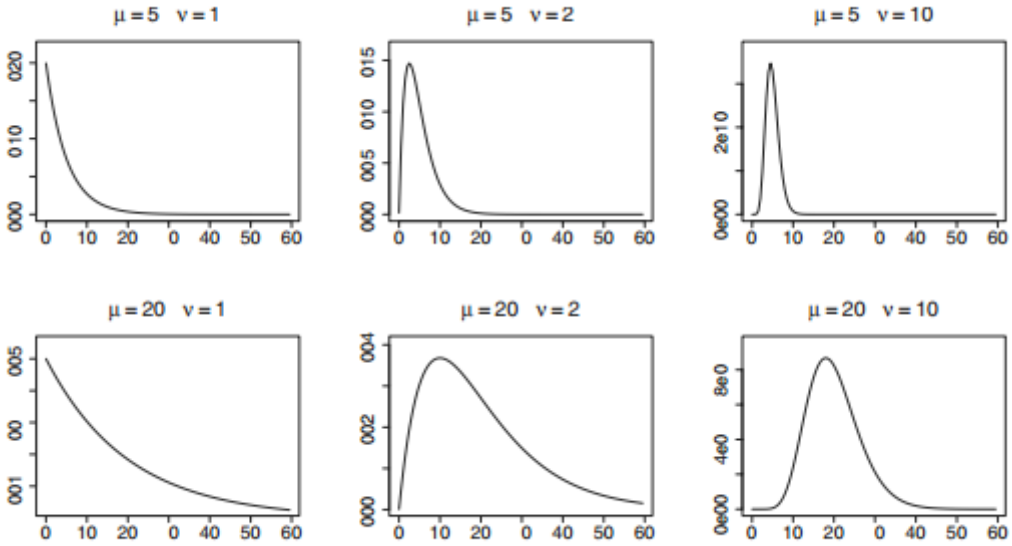
[4] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 23
[5] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 27

Gama tankio funkcija yra:

$$f(y) = \frac{y^{-1}}{\Gamma(\upsilon)}\left(\frac{y\upsilon}{\mu}\right)^{\upsilon} e^{-\frac{y\upsilon}{\mu}}, \qquad y > 0.$$

Gama skirstinio vidurkis yra $E(y) = \mu$, o dispersija $Var(y) = \mu^2/\upsilon$.

Gama skirstinio teorinė realizacija su skirtingais parametrais žemiau pateikiama grafine išraiška:



2 pav. Gama skirstinys

3.3. Nepriklausomų kintamųjų analizė

Duomenų analizė, taikant tinkamas grafines išraiškas, yra pirmasis žingsnis konstruojant „jautrius“ modelius. Tai padeda suprasti ryšius tarp kintamųjų, leidžia atlikti pradinį pagrįstumo patikrinimą individualiems dydžiams. Duomenų patikrinimas taip pat yra naudojamas nustatyti:

Ryšius tarp priklausomo kintamojo ir potencialių nepriklausomų kintamųjų;

Ryšius tarp pačių nepriklausomų kintamųjų.

Pirmojo punkto radimas padeda surasti tinkamus kintamuosius ir nustatyti galimą jų įtaką nepriklausomam kintamajam. Kitas punktas padeda nustatyti, kaip dydžiai tarpusavyje yra susiję. Šis supratimas yra būtinas, norint sukonstruoti praktišką modelį. Stipriai susijusių kintamųjų įtraukimas į modelį turi būti rimtai apgalvotas.

Vis dėlto, duomenų analizė skiriasi fundamentaliai, priklausomai nuo kintamojo tipo, t.y. ar jis yra tolydus (diskretus) dydis, ar kategorinis.

3.3.1. Tolydaus kintamojo lyginimas su tolydžiuoju kintamuoju

Analizuoti šių dviejų kintamųjų ryšį yra geriausia naudojant sklaidos diagramą. Naudojant skirtingas spalvas ar simbolius, kartais į sklaidos diagramas galima įtraukti ir trečiąjį, kategorinį kintamąjį. Šių diagramų tikslas yra parodyti galimą netiesinio modelio judėjimą, taip padedant nustatyti, kokio laipsnio kintamąjį būtų prasminga analizuoti modelyje. Vis dėlto, šie modeliai neparodo galimo ryšio stiprumo ar statistinio reikšmingumo.

3.3.2. Kategorinio kintamojo lyginimas su kategoriniu kintamuoju

Šiems dydžiams nagrinėti dažniausiai yra naudojama dažnių lentelė. Taip pat prasminga nagrinėti juos su mozaikos grafiku. Mozaikos grafiką sudaro daugybė skirtingo dydžio stačiakampių, tai kiekvieno iš jų plotas yra proporcingas tolydaus kintamojo dažniui, o stulpelio plotis yra proporcingas kvadratinei kiekvieno dažnių lentelės stulpelio reikšmes šakniai. Verta paminėti, kad mozaikos grafikai nėra labai efektyvūs tada, kai kategorinio kintamojo indikatorių skaičius yra didelis. Tuomet vertėtų duomenis grupuoti, taip mažinant indikatorių skaičių.

3.3.3. Tolydaus kintamojo lyginimas su kategoriniu

Tinkamiausias būdas analizuoti pavadinime minimų kintamųjų ryšį yra naudojant stačiakampes diagramas. Remiantis V. Čekanavičiaus ir G. Murausko knyga, *Statistika ir jos taikymai. I dalis*, „Stačiakampė diagrama parodo grafinį penkiaskaitės suvestinės vaizdą (min, Q_1, Md, Q_3, max). Stačiakampėje diagramoje yra „dėžė“ – stačiakampis, braižomas nuo pirmojo kvartilio Q_1 iki trečiojo kvartilio Q_3, padalytas brūkšniu į dvi dalis ties mediana Md. [...] Nuo stačiakampio šono brėžiami „ūsai“, besitęsiantys iki paskutinės neišsiskiriančios duomenų aibės reikšmės ir didžiausios neišsiskiriančios duomenų aibės reikšmės.“ Taigi, su šia diagrama mes galime matyti kiekvieno kategorinio kintamojo pasiskirstymą pagal turimas tolydžiojo kintamojo reikšmes taip nustatydami, kokie dydžiai turės didžiausios įtakos kintamajam, ar jų bus daug ir koks yra duomenų asimetriškumas.

3.4. Modelio konstravimo etapai[6]

Modelio kūrimo etapai yra tokie:

1. Parenkamas $f(y)$, o kartu ir $a(\theta)$ eksponentinis skirstinys.
2. Parenkama jungiamoji funkcija $g(\mu)$. Šis pasirinkimas yra supaprastinas tiesiog pasirenkant „kanoninę“ jungiamąją funkciją nurodytam skirstiniui. Detalesnis paaiškinimas bus pateikiamas žemiau.

[6] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 65

3. Nepriklausomų kintamųjų parinkimas su kuriais kursime modelį. Šis etapas neskiria nuo tiesinės regresijos kūrimo etapų.
4. Suderintųjų reikšmių (angl. fitted values) apskaičiavimas pagal išmatuotas β reikšmes ir, jei nežinomas, φ. Ši procedūra dažniausiai atliekama taikant didžiausio tikėtinumo metodą, kurį aptarsime vėliau.

Galiausiai patikriname mūsų gautąsias suderintąsias reikšmes arba prognozines reikšmes su faktinėmis reikšmėmis, sužinant modelio prognozės ir turimų dydžių paklaidą. Taip pat atliekame kitus testus bei analizę, nustatant mūsų modelio tinkamumą ir pačių nepriklausomų kintamųjų svarbą. Šis eiliškumas nėra būtinas, ypač dirbant su draudimo duomenimis, nes, norint nuspręsti, kokia turėtų būti modelio specifika, geriausia būtų atlikti duomenų analizę.

3.5. Jungiamoji funkcija[7]

Jungiamosios funkcijos, išskyrus logaritminę funkciją, yra laipsninės, t.y. $g(\mu) = \mu^p$. Tuo tarpu $(y^p - 1)/p$ yra taikomas logaritminei funkcijai. Jeigu $g(\mu) = \theta$ tada g yra vadinama kanonine jungtimi suderinta su $a(\phi)$. Tokiu atveju $\theta = X'\beta$. Pasirenkant kanoninį ryšį, suderintą su f funkcijos skirstiniu palengvina įverčių nustatymą. Žemiau esančioje lentelėje[8] yra pateikiami kanoniniai ryšiai tarp konkrečių skirstinių. Konstantos, bendru atveju, nėra rašomos:

Skirstinys	Funkcijos pavadinimas	Jungties funkcija	Modelis	Vidurkio įvertinys
Normalusis	Identiška	$g(\mu) = \mu$	$\mu = X\beta$	$\hat{\mu} = X\hat{\beta}$
Puasono	Log	$g(\mu) = \ln \mu$	$\mu^{-1} = X\beta$	$\hat{\mu} = \left(X\hat{\beta}\right)^{-1}$
Eksponentinis	Atvirkštinė	$g(\mu) = \mu^p$	$\ln \mu = X\beta$	$\hat{\mu} = \exp\{X\hat{\beta}\}$
Binominis	Logit	$g(\mu) = \ln \frac{\mu}{1-\mu}$	$\ln \frac{\mu}{1-\mu} = X\beta$	$\hat{\mu} = \frac{\exp\{X\hat{\beta}\}}{1+\exp\{X\hat{\beta}\}}$

1 Lentelė. Kanoninės jungties funkcijos

3.6. Ofsetas[9]

Modeliuojant skaičiavimus, tokius kaip žalų skaičius ar mirties rizika, kažkokioje grupėje galime tikėtis korekcijos. Jeigu μ yra skaičiaus y vidurkis, tada tikslumo dažnis μ/n ir $g\left(\frac{\mu}{n}\right) = X''\beta$, kur n yra visų kintamųjų eilės suma.

Kada logaritmuojame funkciją g, gauname:

$$\ln\left(\frac{\mu}{n}\right) = X'\beta \Rightarrow \ln \mu = \ln n + X'\beta.$$

[7] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 66
[8] V. Čekanavičius, G. Murauskas. *Statistika ir jos taikymai III dalis*, Vilnius: TEV, 2009, p. 49
[9] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 66

Čia $\ln n$ yra vadinamas ofsetu. Ofsetai yra efektyvūs papildomi aiškinamieji kintamieji regresijoje su koeficientu β_i lygiu 1. GLM modelio lygtis su ofsetu įgauna reikšmes:

$$\mu = ne^{x'\beta}.$$

Ofsetai yra naudojami, norint pakoreguoti grupės dydį arba skirtingo laiko periodų stebėjimus.

3.7. Kategoriniai duomenys modelyje[10]

Kada potencialus nepriklausomas kintamasis yra kategorinis, tada jis yra perdaromas kaip žymimasis kintamasis (angl. dummy variable), kad būtų galima jį įtraukti į modelį. Kategorinis kintamasis turi keletą indikatorių, bet, naudojant modelyje pastarąjį dydį, mes išmetame vieną iš indikatorių, kuris yra vadinamas baziniu lygiu, pagal kurį yra matuojami skirtumai tarp kitų indikatorių.

3.7.1. Bazinio lygio pasirinkimas[11]

Kategorinis bazinis kintamasis turėtų būti vienas iš dažniausiai pasirodančių indikatorių reikšmių, jeigu kategorinį kintamąjį pateiksime kaip dažnio lentelę. Norėdami tai paaiškinti kaip atrodo praktikoje, tarkime, kad vienas kategorinis kintamasis turi r lygių ir paskutinysis lygis yra pasirenkamas kaip bazinis. Tada modelio matricoje visiems likusiems lygiams suteikiama reikšmė 1, o baziniam lygiui suteikiama reikšmė 0. Tuomet bazinio lygmens koeficientas, teigiama, kad yra lygus 1.

Vadinasi, bet koks lygis, kuris turi didelį stebėjimų skaičių lyginant su kitais lygiais, gali būti baziniu lygiu. Kadangi β_j yra skirtumas tarp aiškinamojo kintamojo visų lygių lyginant su baziniu lygiu, yra patogu pasirinkti bazinį dydį, kaip „normalų“ arba „dažną“ lygį, lyginant su kitais indikatoriais.

3.8. Didžiausio tikėtinumo metodas[12]

GLM įverčiai $\hat{\beta}$ ir $\hat{\phi}$ yra gaunami maksimizuojant logaritminę tikėtinumo funkciją, kuri yra apibrėžiama:

$$l(\beta, \phi) \equiv \sum_{i=1}^{n} \ln f(y_i; \beta, \phi) = \sum_{i=1}^{n} \{ \ln c(y_i, \phi) + \frac{y_i\theta_i - a(\theta_i)}{\phi} \},$$

t.y. kiekvienai y_i realizacijai $f(y)$ yra sukuriama tikėtinumo funkcija $f(y_i)$. Tikimybė tuomet priklauso nuo θ ir, jei galima, ϕ.

Tada sudarome tikėtinumo funkciją $f(y_i; \theta, \phi)$. Jeigu y_i yra nepriklausomi, tada galime parašyti jungtinę tikėtinumo funkciją, kuri yra:

[10] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 51
[11] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 52
[12] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 67

$$f(y;\theta,\phi) = \prod_{i=1}^{n} f(y_i;\theta,\phi).$$

Šiai funkcijai taikome prieš tai parašytą logaritminę išraišką.

GLM modelio atveju logaritminė tikėtinumo funkcija atrodytų taip:

$$l(\beta,\phi) = \sum_{i=1}^{n} \{ \ln c(y_i,\phi) + \frac{y_i\theta_i - a(\theta_i)}{\phi} \}.$$

Tarkime, mes norime apskaičiuoti β_j naudodami didžiausio tikėtinumo metodą. Kad rastume maksimumą, $l(\beta,\phi)$ yra diferencijuojama pagal β_j:

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j}, \text{ kur } \frac{\partial l}{\partial \theta_i} = \frac{y_i - \dot{a}\theta_i}{\phi} = \frac{y_i - \mu_i}{\phi} \text{ ir } \frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i} x_{ij}.$$

Čia $\eta_i = x'\beta$ ir x_{ij} yra komponentė i iš x_j. Tariant, kad $\frac{\partial l}{\partial \beta_j} = 0$ gauname pirmos eilės sąlygą didžiausiam tikėtinumui:

$$\sum_{i=1}^{n} \frac{\partial \theta_i}{\partial \eta_i} x_{ij}(y_i - \mu_i) = 0 \iff X'D(y-\mu) = 0,$$

kur D yra diagonalinė matrica su elementais $\frac{\partial \theta_i}{\partial \eta_i}$,

$$\left(\frac{\partial \theta_i}{\partial \eta_i}\right)^{-1} = \frac{\partial \eta_i}{\partial \theta_i} = \frac{\partial \eta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \theta_i} = \mathrm{g}(\mu_i)\alpha(\phi_i) = \mathrm{g}(\mu_i)V(\mu_i)$$

Taigi D yra diagonalė su reikšmėmis $\{\mathrm{g}(\mu_i)V(\mu_i)\}^{-1}$. Lygybė, turinti ekvivalentumo ženklą, dažnai yra vadinama β įverčio lygtimi. Vėlgi, priderinant mūsų turimą informaciją GLM modeliui, apibrėžkime diagonalines matricas G ir W su reikšmėmis atitinkamai $\dot{g}(\mu_i)$ ir $[\{\mathrm{g}(\mu_i)\}^2 V(\mu_i)]^{-1}$. Tada $D = WG$ ir lygtis su ekvivalentumo ženklu būtų ekvivalenti:

$$X'WG(y-\mu) = 0.$$

3.9. Nepriklausomų kintamųjų reikšmingumo tikrinimas[13]

GLM modelio kintamųjų reikšmingumo tikrinimas yra panašus į tiesinių modelių reikšmingumo tikrinimą. Mes tikriname nulinę hipotezę $H_o: C\beta = r$, kur C yra žinoma matrica, kartais dar vadinama hipotezės matrica, o r yra duotosios reikšmės.

Yra trys pagrindiniai siūlymai, kaip tikrinti hipotezę. Kiekvienas iš šių testų naudoja tikėtinumo arba logaritminę tikėtinumo funkcijas. Tarkime, kad $\hat{\beta}$ yra neapribotas didžiausio tikėtinumo β įvertis ir pažymėkime, kad $\tilde{\beta}$ yra pastarasis minėtasis didžiausio tikėtinumo įvertis, kai l yra maksimizuotas subjektas pagal apribojimus $C\beta = r$. Nuo šios vietos metodika yra panaši į tiesinio modelio. Priskirkime reikšmę $\hat{l}$, kaip dydžio $l, \hat{\beta}$ įverčiu ir $\tilde{l}$, dydžio l, $\tilde{\beta}$ įverčiu. Tuomet $\hat{l} \geq \tilde{l}$.

[13] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 74

3.9.1. Tikėtinumų santykio testas[14]

Su šiuo testu mes lyginame tarpusavyje $\hat{l}$ ir $\tilde{l}$. Jeigu $\hat{l}$ yra daug didesnis už $\tilde{l}$, tai parodo, kad modelis nėra apribotas. Šiam testui yra reikalingi abu $\hat{\beta}$ ir $\tilde{\beta}$ įverčiai. Tikėtinumo santykis yra apibrėžiamas kaip $\lambda = \hat{L}/\tilde{L}$, kur pastarieji dydžiai atitinkamai yra neapriboti ir apriboti modeliai. Šio testo statistika yra $2\ln\lambda = 2(\hat{l} - \tilde{l})$. Šis dydis yra visada neneigiamas ir turi χ^2_q pasiskirstymą, jeigu $C\beta = r$, kur q yra C eilučių skaičius. Jeigu $2\ln\lambda$ yra mažas (arti nulio) skaičius, tuomet apribotas modelis yra taip pat geras kaip ir neapribotas.

Tikėtinumo santykio statistikos χ^2_q pasiskirstymas įtraukia dispersijos parametrą ϕ, kuris dažniausiai yra nežinomas, t.y. jeigu turime Puasono arba binominį skirstinį, tuomet šis dydis yra duotas, kitais atvejais, jis turi būti išmatuojamas. Šis kriterijus taip pat gali būti išreiškiamas kaip skirtumas tarp neapriboto ir apriboto modelių dispersijų, tada, kai ϕ kriterijus yra naudojamas abiejose logaritmuotose tikėtinumo funkcijose.

3.9.2. Wald testas [15]

Šis testas matuoja, kaip toli $C\hat{\beta}$ yra nutolęs nuo r. Šiam testui būtinos apskaičiuotos $\hat{\beta}$ reikšmės. Jeigu $C\beta = r$ tada žinome, kad $\hat{\beta} \sim N\{\beta, \phi(X'WX)^{-1}\}$, vadinasi $C\hat{\beta} - r \sim N\{0, \phi C(X'WX)^{-1}C'\}$.

Šis rezultatas mums padeda sukonstruoti Wald statistiką, kuri yra:

$$\left(C\hat{\beta} - r\right)'\{\phi C(X'WX)^{-1}C'\}^{-1}\left(C\hat{\beta} - r\right) \sim \chi^2_q.$$

Taigi Wald statistika yra kvadratinis statistinis atstumas nuo $C\hat{\beta}$ iki r, remiantis kovariacijų matrica $C\hat{\beta}$.

3.9.3. Kiekvieno koeficiento tikrinimas [14]

Kada tikriname, ar $\beta_j = r$, matricoje C atsiranda nulinių reikšmių eilutė išskyrus $j - tąją$ eilutę, kurios vietoje atsiranda 1. Tuomet teigiama, jog visi kiti nepriklausomi kintamieji, naudojami modelio lygtyje, jau yra įtraukti į modelį. Tuomet Wald statistika sumažėja iki vienos diagonalinės reikšmės $\phi(X'WX)^{-1}$, dispersija patampa $\widehat{\beta}_j$, kuri yra žymima kaip $\phi\psi_j$. Tada Wald statistika pakeičiama į tokią išraišką:

$$\frac{(\widehat{\beta}_j - r)^2}{\phi\psi_j} \sim \chi^2_1.$$

Kada ϕ yra nežinomas, tada jis yra pakeičiamas savo įvertiniu.

[14] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 74
[15] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 75

3.9.4. Visų koeficientų tikrinimas [14]

Globalus testas visiems koeficientams, kuris tikrina, kad visi koeficientai, išskyrus laisvąjį narį, yra nuliai, tuomet C yra $(p-1)$ x p matrica:

$$C = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

3.9.5. „Score“ testas [16]

Šio testo pagrindas yra l išvestinė arba pasvirimas nuo $\tilde{\beta}$, kuris yra vadinamas „rezultatu“ (angl. score). Didelės „rezultatų“ reikšmės nurodo, kad galime pagerinti modelį, sumažindami jo apribojimus. GLM modeliams rezultatas yra $\dot{l}(\beta) = \phi^{-1}X'WG(y-\mu)$. Toliau, $E\{\dot{l}(\beta)\} = 0$, $var\{\dot{l}(\beta)\} = E\{\dot{l}(\beta)\,\dot{l}'(\beta)\} = \phi^{-1}X'WX$. Jeigu $C\beta = r, \dot{l}(\beta)$ turėtų būti netoli nulinės reikšmės. Taigi gauname statistiką:

$$\dot{l}'(\tilde{\beta})\left[Var\{\dot{l}(\beta)\}\right]^{-1}\ \dot{l}(\tilde{\beta})$$

Čia $\dot{l}(\tilde{\beta}) = \phi^{-1}X'WG(y-\tilde{\mu}), Var\{\dot{l}(\tilde{\beta})\} = \phi^{-1}(X'WX)$ ir $\tilde{\mu}$ yra tikimybinis vidurkis apskaičiuotas pagal $\tilde{\beta}$.

Ši rezultato statistika apytiksliai atitinka χ^2_q pasiskirstymą. Praktikoje W ir ϕ yra pakeičiami įvertinais, kurie χ^2 skirstinio aproksimaciją padaro nebe tokią tikslią. Verta pastebėti, kad X yra apibrėžiama kaip pilna matrica, įtraukianti visus aiškinamuosius kintamuosius, kurių koeficientai yra 0 nulinėje hipotezėje $C\beta = r$.

3.9.6. Testų apibendrinimas praktiniam taikymui [17]

1. Visiems skirstiniams, išskyrus Gama skirstinį, „mastelio“ parametras yra apibrėžiamas kaip $\sqrt{\phi}$, kur ϕ yra dispersijos parametras. Gama skirstiniui jis yra apibrėžiamas kaip $1/\phi$.
2. Pirmojo tipo testai yra nuoseklūs testai. Kiekvienas aiškinamasis kintamasis yra pridedamas į seką taip, kaip yra nurodyta modelio lygtyje ir kiekvieno reikšmingumas yra tikrinamas su prielaida, kad anksčiau pridėti kintamieji jau yra modelyje. Jeigu nepriklausomas kintamasis yra kategorinis su r lygiu, tada visi $r-1$ indikatoriniai kintamieji priklausantys tam nepriklausomam kintamajam yra tikrinami kartu. Visas dėmesys turi būti skiriamas statistikų interpretacijai, kai skirtingi rezultatai gali būti gauti pakeičiant nepriklausomų kintamųjų išsidėstymą.

[16] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 76
[17] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 77

3. Trečiojo tipo testai tikrina kiekvieno kintamojo reikšmingumą su prielaida, kad visi kiti kintamieji yra įtraukiami į modelį. Šie testai yra nepriklausomi nuo kintamųjų išsidėstymo tvarkos.
4. Wald apibrėžimas yra tapatinamas su trečio tipo ANOVA testu.

3.10. Išskirtys [18]

Dirbdami su labai dideliais duomenų rinkiniais, tikėtina, susidursime su duomenimis, kurie labai išsiskirs iš viso duomenų rinkinio, t.y. kai kurios reikšmės bus ženkliai didesnės ar mažesnės nei kitos. Tokių duomenų egzistavimas turi reikšmingos įtakos modelio koeficientams, todėl modelio suderintosios reikšmės gali stipriai nesutapti su mūsų stebimomis reikšmėmis. Vadinasi, mūsų tikslas būtų atsikratyti šių išskirčių, siekiant optimalesnio modelio.

Kaip tai padaryti? Tam atlikti geriausiai tinka svertų sistema. Duomenys, kurie sudaro neįprastas kombinacijas su aiškinamuoju kintamuoju, turės didelę svertinę reikšmę. Svertas yra $\hat{y} = Hy$, vadinasi, svertinės reikšmės yra gaunamos iš tiesinės kombinacijos $\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{in}y_n$, kur $(h_{i1}, h_{i2}, \ldots, h_{in})$ yra matricos H i-toji eilutė, įgyjanti reikšmes nuo 0 iki 1. Suma šios matricos diagonalės elementų yra modelio *p* parametras, todėl kiekvienas atvejis turi atitinkamai $\frac{p}{n}$ suderintąją reikšmę. Jeigu vienas h_{ii} yra daug didesnis nei $\frac{p}{n}$, tuomet šis atvejis yra laikomas „izoliuotu arba savo paties įverčiu“ [Lindsey, 1997]. Matricos elementai matuoja atstumą nuo centro iki reikšmės x_i, kaip tai matome iš tiesinio modelio su vienu aiškinamuoju kintamuoju x_i:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}.$$

Vadinasi, kuo toliau aiškinamojo kintamojo reikšmė nuo vidurkio, tuo didesnė h_{ii} reikšmė. Šios reikšmės ir yra vadinamos svertinėmis reikšmėmis. Jeigu $h_{ii} > \frac{2p}{n}$, tai mums pasako, kad egzistuoja neįprastai didelis „atstumas“ tarp x_i reikšmės ir nepriklausomo kintamojo.

Kadangi mes naudojame ne tiesinį modelį, o GLM, tai šio modelio svertinėms reikšmėms nustatyti yra naudojama tokia matricinė išraiška:

$$H = \sqrt{W}X(X'WX)^{-1}X'\sqrt{W}.$$

Ši išraiška yra tokia pati kaip ir tiesinė regresijos lygtis, tik vietoj X mes pakeičiame į $\sqrt{W}X$, kur W yra diagonalinė matrica.

[18] P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008, p. 61

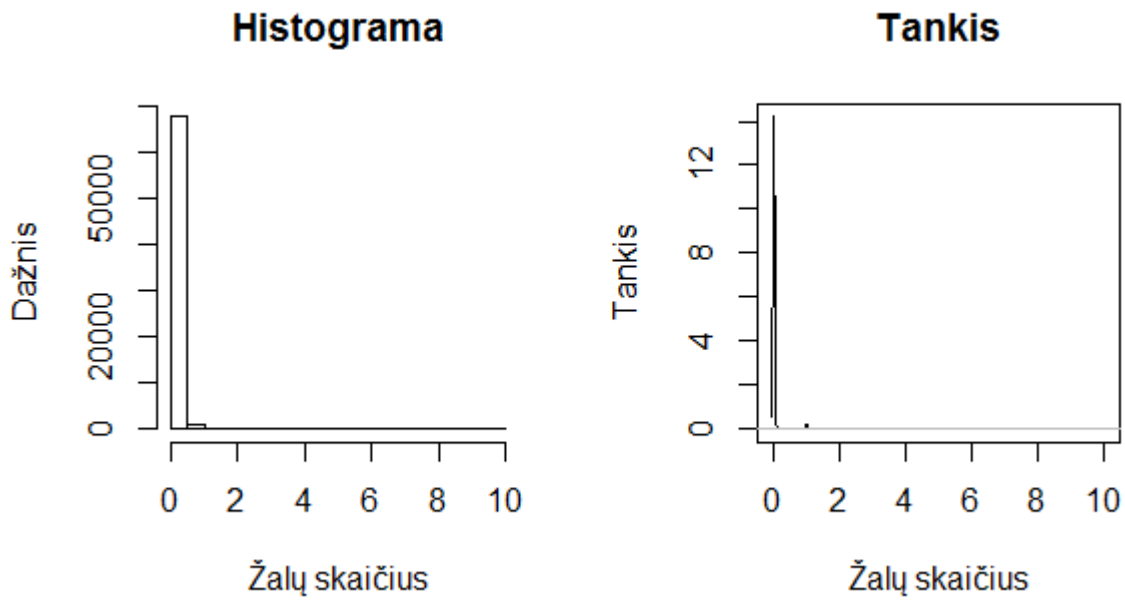
4 PRAKTINĖ DALIS

Norėdami sukurti žalų sumos prognozę, optimaliausia būtų konstruoti du modelius. Vienas iš šių modelių būtų skirtas prognozuoti žalų skaičiui, o kitas modelis žalų aritmetiniam vidurkiui, nes žalų sumą gausime žalų dažnį padauginus iš vidutinės žalos dydžio. Analizę pradėsime nuo abiejų anksčiau paminėtų priklausomų kintamųjų analizės, vėliau pereisime prie aiškinamųjų arba nepriklausomų kintamųjų analizės su priklausomu kintamuoju. Galiausiai sukonstruosime modelį, geriausiai atspindintį mūsų turimus duomenis ir patikrinsime jo tinkamumą.

4.1. Žalų skaičiaus modelis

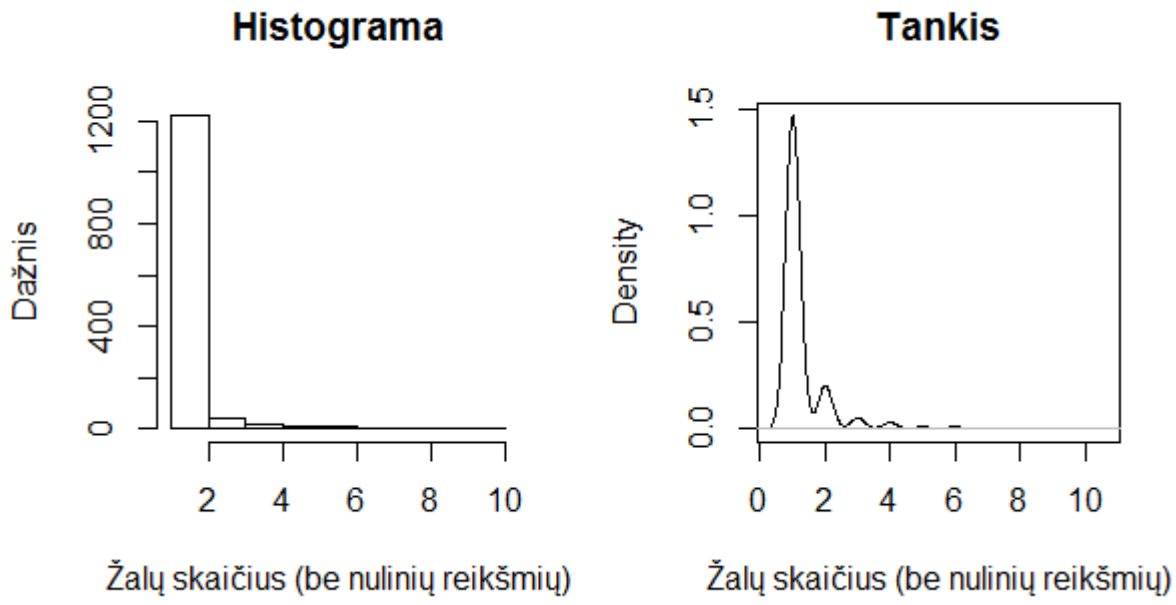
4.1.1. Skirstinio nustatymas

Skirtinio nustatymui naudosime Cullen ir Frey grafikus, kurie pagal suskaičiuotus duomenų parametrus, kurie yra vidurkis, mediana, standartinis nuokrypis, asimetrija ir „uodegos“ ilgumas, pateiks pasiskirstymus, kuriuos vertėtų patikrinti, taip atrandant geriausiai duomenis atspindintį skirstinį. Verta paminėti, kad mūsų žalų skaičius yra diskretus dydis, nes gali įgyti reikšmes 0 (žalos nėra) arba 1 (yra viena žala) ar net 1> (yra padaryta daugiau nei viena žala per kontrakto galiojimo laikotarpį). Iš viso mes nagrinėjame 1308 turimas žalas. Žalų skaičiaus histograma atrodytų taip:



3 pav. Pradinių duomenų histograma ir tankis

Vis dėlto, susiduriame su problema, kad didžiąją mūsų duomenų dalį sudaro nulinių žalų skaičiaus grupė, todėl nuspręsta daryti modelį tik su tais duomenimis, kuriuose buvo fiksuota bent viena žala. Išmetus visas nulines eilutes gauname tokią duomenų sklaidą:



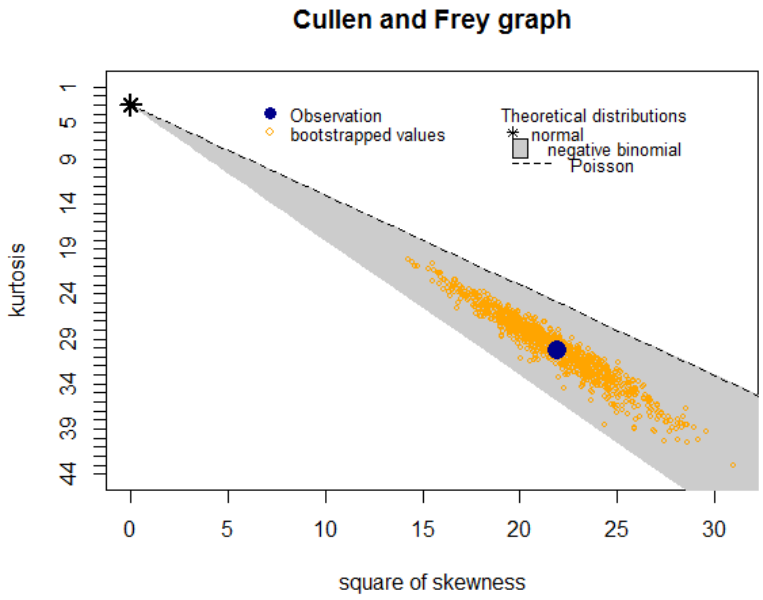
4 pav. Pakoreguotų duomenų histograma ir tankis

Vėlgi matome, kad daugiausia reikšmių turi vienos žalos grupė. Aiškumo dėlei pateiksime duomenų dažnio lentelę:

Žalų skaičius	1	2	3	4	5	6	7	8	9	10
Dažnis	1071	148	38	22	8	8	5	2	2	4

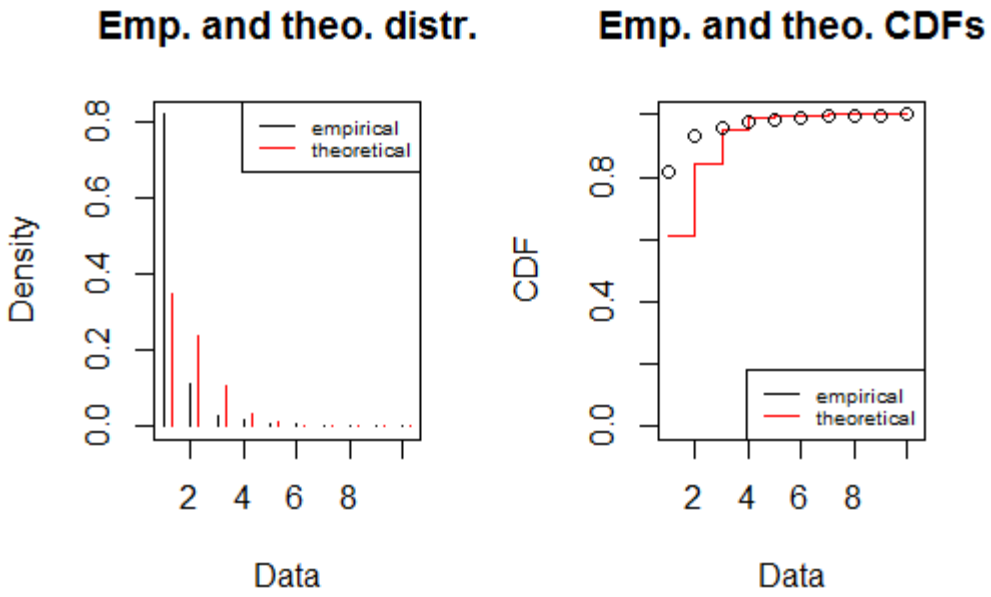
2 Lentelė. Žalų skaičiaus dažnio lentelė

Žvelgiant į duomenų išsidėstymą ir tankio pasiskirstymą matome, kad mūsų duomenys labiausiai panašūs į neigiamo binominio skirstinio pasiskirstymą ir Puasono skirstinį. Taigi pažiūrėkime, ką mums siūlo Cullen ir Frey grafikas:



5 pav. Cullen ir Frey grafikas žalų skaičiui

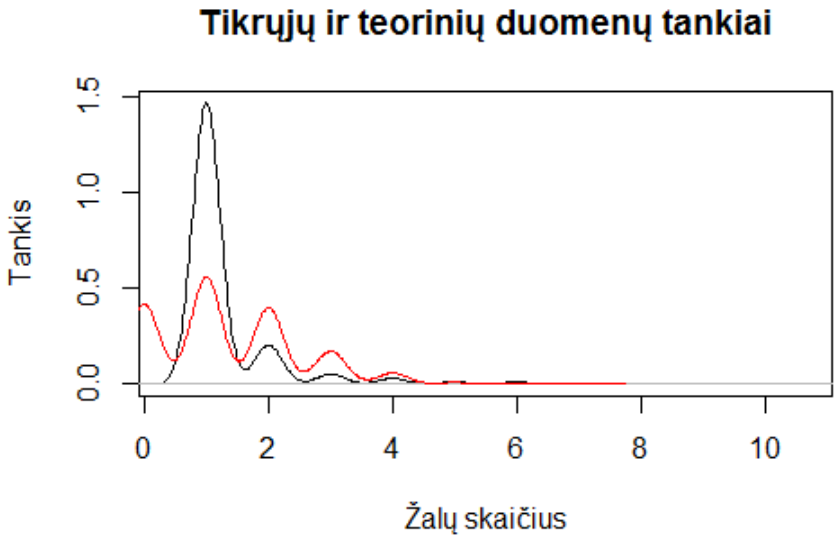
Mūsų prielaida pasitvirtino ir matome, kad labiausiai tikėtini yra neigiamas binominis ir Puasono skirstiniai. Patikrinsime, kaip Puasono skirstinys atitinka mūsų turimus duomenis. Galime teigti, kad Puasono skirstinys nėra pats geriausias mūsų turimiems duomenims:



6 pav. Puasono skirstinio patikrinimas turimiems duomenims

Deja, bet neigiamo binominio skirstinio patikrinti nepavyko, nes duomenys turi tenkinti Bernulio „sėkmės“ ir „nesėkmės“ pasiskirstymą, o dėl prieš tai padaryto sprendimo to nebegalime padaryti. Taigi prieiname prie išvados, kad mūsų modelis bus kuriamas remiantis Puasono skirstiniu.

Vis dėlto, patikrinkime, kaip Puasono skirstinys atitinka mūsų duomenis su χ^2 testu. Tam mums reikia sugeneruoti Puasono teorinį skirstinį ir palyginti su mūsų turimomis reikšmėmis. Prieš tai atliktoje analizėje rekomenduojama λ reikšmė yra 1,35. Grafiškai Puasono teorinis skirstinys su turimais duomenimis atrodo taip:

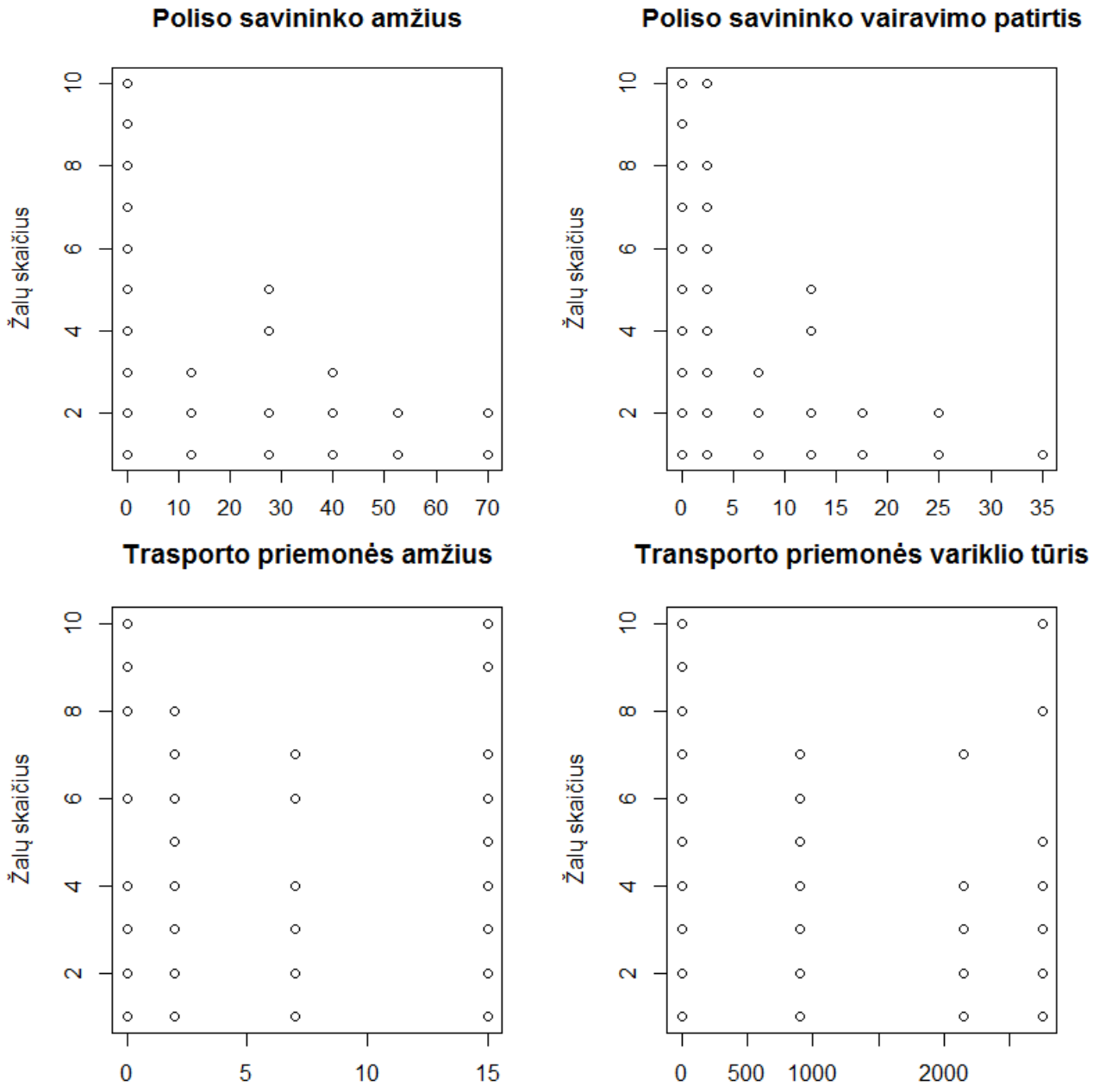


7 pav. Teorinio ir faktinio Puasono skirstinio tankių grafikas

Matome, kad tankiai iš tikro nelabai sutampa. χ^2 testo p-reikšmė taip pat yra didesnė už 0,05, todėl tai reikštų, kad šis skirstinys neatitinka mūsų duomenų. Vis dėlto, neturime kito pasirinkimo ir kursime GLM modelį naudodami Puasono skirstinį.

4.1.2. Potencialių aiškinamųjų kintamųjų analizė

Teorinėje dalyje rekomenduojama tolydžiuosius kintamuosius, lyginant su kategoriniais kintamaisiais, analizuoti naudojant stačiakampes diagramas, tačiau mūsų reikšmės yra diskrečios, todėl grafikai yra labai neinformatyvūs. Pateiksime tik tų reikšmių grafikus, kur kategoriniams kintamiesiems galėjome išskirti skaitines reikšmes. Pagal nurodytą teoriją pateiksime tik sklaidos grafikus.



8 pav. Žalų skaičiaus sklaidos diagramos turimiems duomenims

Deja, bet grafikai nėra labai aiškūs, nors iš transporto amžiaus ir variklio tūrio grafikų galime matyti, kad duomenims tiktų parabolės formos tiesė, todėl vertėtų modelyje naudoti kvadratinius narius.

4.1.3. Modelio kūrimas

Visų pirma, parodysime kaip modeliai, remiantis prieš tai aprašyta teorija, yra analizuojami. Verta paminėti, kad modelyje naudosime ofsetą, kuris yra sutarčių skaičius. Kadangi jis yra naudojamas kaip dar vienas aiškinamasis kintamasis, tai dydis turi būti išreikštas logaritmine išraiška dėl korektiškumo. Tarkime sukuriame pradinį GLM modelį, kurio išraiška yra:

$$Žalų\ skaičius = Amžius\ (kategorinis)$$

Peržiūrėkime, kokie yra gautojo modelio koeficientai.

Parametras	Koeficientas	Standartinis Nuokrypis	Įvertinys	P-reikšmė
Laisvasis narys	-0,55	0,08	0,58	<0,01
Amžius (1 kategorija)	0,34	0,15	1,4	0,02
Amžius (2 kategorija)	0,04	0,1	1,04	0,72
Amžius (3 kategorija)	-0,02	0,1	0,98	0,85
Amžius (4 kategorija)	0	-	1	-
Amžius (5 kategorija)	-0,3	0,17	0,74	0,07
Amžius (6 kategorija)	-3	0,1	0,05	<0,01

3 Lentelė. Pirmojo modelio rezultatai

Matome, kad turime keletą ir nereikšmingų kintamųjų, tačiau tai yra pradinis modelis, todėl tai nestebina. Šio modelio Akaike Informacijos kriterijus yra 5027,5, o Schwarz Informacinis kriterijus yra 5054,8. Dabar atlikime pirmojo ir trečiojo tipo testus. Taip pat, vertėtų atlikti Wald testą, tačiau šio testo rezultatai nesiskiria nuo pirmojo tipo testo rezultatų, todėl naudosime tik pirmojo tipo testą. Pirmojo tipo testas tikrina kiekvieną aiškinamąjį kintamąjį iš eilės su prielaida, kad ankstesni kintamieji yra įtraukti į modelį, tuo tarpu trečio tipo testas tikrina kiekvieną kintamąjį su prielaida, kad visi likusieji kintamieji jau yra modelyje. Pirmojo tipo testui įtakos turi kintamųjų skirstymo tvarka, tuo tarpu trečiajam testui, tai įtakos neturi. Abiejų testų rezultatus, paprastumo dėlei, pateiksime vienoje lentelėje:

	Pirmojo tipo testas			Trečiojo tipo testas	
	Standartinis nuokrypis	χ^2	P-reikšmė	χ^2	P-reikšmė
Laisvasis narys	3476,2				
Amžius	2565,2	1666,6	<0,01	1666,6	<0,01

4 Lentelė. Pirmojo ir trečiojo testų rezultatai

Gavome, kad kintamasis „Amžius“ yra reikšmingas, o modelio pradinis standartinis nuokrypis sumažėjo nuo 3476,2 iki 2565,2.

Dabar pabandykime į modelį įdėti mūsų išskirtąjį kintamąjį, kuris nebėra kategorinis. Tokiu būdu mes galėsime patikrinti, ar yra prasminga modelyje taikyti polinominę kintamojo išraišką, tačiau čia galime susidurti su viena problema. Jei pasirenkamas polinomo laipsnis yra didelės eilės (neigiamas ar teigiamas), tada susiduriame su perpildymo problema, o dėl to $\hat{\beta}$ yra paslinktas. To galima išvengti, jeigu modelyje aiškinamasis kintamasis yra naudojamas standartizuota išraiška. Tam atlikti taikome tiesinę transformaciją:

$$x^* = \frac{x - \frac{x_{\max} - x_{min}}{2}}{\frac{x_{\max} - x_{min}}{2}}$$

Taigi, norėdami išvengti prieš tai minėtos problemos, atliksime tai mūsų išskirtam dydžiui, kuris yra gautas iš „amžiaus“ kategorinių duomenų. Šį kartą pateiksime tik pirmojo ir trečiojo tipo testų rezultatus. Dabar mūsų modelio išraiška atrodo taip:

$$Žalų\ skaičius = Amžius + Amžius^2 + Amžius^3 + Amžius^4$$

	Pirmojo tipo testas			Trečiojo tipo testas	
	Standartinis nuokrypis	χ^2	P-reikšmė	χ^2	P-reikšmė
Laisvasis narys	3476,2				
Amžius	2283,5	1102,32	<0,01	27	<0,01
Amžius^2	2042,6	417,8	<0,01	19,8	<0,01
Amžius^3	1785,2	77,2	<0,01	111,2	<0,01
Amžius^4	1706,5	59,62	<0,01	59,6	<0,01

5 Lentelė. Pirmojo ir trečiojo tipų testo rezultatai

Matome, kad visi kintamieji yra reikšmingi ir standartinį nuokrypį sumažinome iki 1706,5. Modelio Akaike Informacinis kriterijus yra 5035,2 (didesnis nei ankstesnio modelio), o Schwarz Informacinis kriterijus yra 5058 (taip pat didesnis, nei ankstesnio modelio).

Dėl tikrinamų modelių gausos pateiksime tik galutinius rezultatus lentelės forma, neįtraukiant modelių su nereikšmingais kintamaisiais.

Nr.	Modelis	Standartinis nuokrypis	Akaike reikšmė	Schwarz reikšmė
1.	Amžius (kategorinis) + Transporto grupė + Transporto nuosavybė + Registracijos vieta + Kliento tipas + Patirts^2 + Transporto priemonės amžius^2 + Transporto priemonės tūris	1476,5	4337,6	4435,92
2.	Amžius (kategorinis) + Transporto grupė + Transporto nuosavybė + Kliento tipas + Patirts^2 + Transporto priemonės amžius^2 + Transporto priemonės tūris	1486,6	4343,7	4431,71
3.	Amžius (kategorinis) + Transporto grupė + Transporto nuosavybė + Regionas + Kliento tipas + Patirts^2 + Amžius^2 + Transporto priemonės tūris + Transporto priemonės tūris^2	1474,8	4337,9	4441,44

4.	Amžius (kategorinis) + Transporto grupė + Vairavimo teritorija + Transporto nuosavybė + Kliento tipas + Patirts^2 + Transporto amžius^2 + Transporto priemonės tūris	1484,4	4343,5	4436,68
5.	Amžius (kategorinis) + Patirtis (kategorinis) + Transporto grupė + Transporto nuosavybė + Registracijos vieta + Kliento tipas + Transporto amžius + Transporto amžius^2 + Transporto variklio tūris	1465,5	4338,6	4467,99

6 Lentelė. Geriausių modelių sąrašas

Geriausias modelis, remiantis Akaike Informacijos kriterijumi yra pirmas, tuo tarpu remiantis Schwarz Informacijos kriterijumi antras, bet, kadangi pirmojo modelio standartinis nuokrypis yra mažesnis, pasirinksime jį. Vis dėlto, vertėtų patikrinti, ar šių modelių skirtumas yra reikšmingas. Tam naudosime rezultato ir pirmojo tipo testais. Abiejų testų P-reikšmė yra mažesnė nei 0,05, vadinasi skirtumas tarp šių dviejų modelių yra statistiškai reikšmingas. Pažiūrėkime, kokie yra modelio koeficientai bei koks jų statistinis reikšmingumas:

Parametras	Koeficientas	Standartinis Nuokrypis	Įvertinys	P-reikšmė
Laisvasis narys	-0,28	0,11	0,76	0,27
Amžius (1 kategorija)	-0,24	0,17	0,79	0,37
Amžius (2 kategorija)	-0,24	0,11	0,79	0,92
Amžius (3 kategorija)	-0,11	0,11	0,89	0,73
Amžius (4 kategorija)	0	-	1	-
Amžius (5 kategorija)	-0,22	0,17	0,79	0,29
Amžius (6 kategorija)	-3,46	0,17	0,03	<0,01
Transporto Grupė (1 kategorija)	0	-	1	-
Transporto Grupė (2 kategorija)	0,08	0,07	1,09	0,15
Transporto Grupė (3 kategorija)	0,82	0,10	2,27	<0,01
Transporto Grupė (4 kategorija)	0,42	0,31	1,53	<0,01
Transporto Grupė (5 kategorija)	0,45	0,26	1,57	<0,01
Transporto Nuosavybė (1 kategorija)	0	-	1	-
Transporto Nuosavybė (2 kategorija)	1,09	0,24	2,96	<0,01
Transporto Nuosavybė (3 kategorija)	0,47	0,08	1,6	<0,01
Transporto Nuosavybė (4 kategorija)	0,67	0,12	1,95	<0,01
Registracijos vieta (1 kategorija)	0,21	0,08	1,23	0,04
Registracijos vieta (2 kategorija)	0	-	1	-
Registracijos vieta (3 kategorija)	0,00	0,08	1,00	0,05
Kliento tipas (1 kategorija)	0,95	0,09	2,58	<0,01
Kliento tipas (2 kategorija)	0	-	1	-
Patirtis^2	0,96	0,15	2,63	<0,01
Transporto priemonės amžius^2	-0,74	0,07	0,48	<0,01
Transporto variklio tūris	0,17	0,05	1,19	<0,01

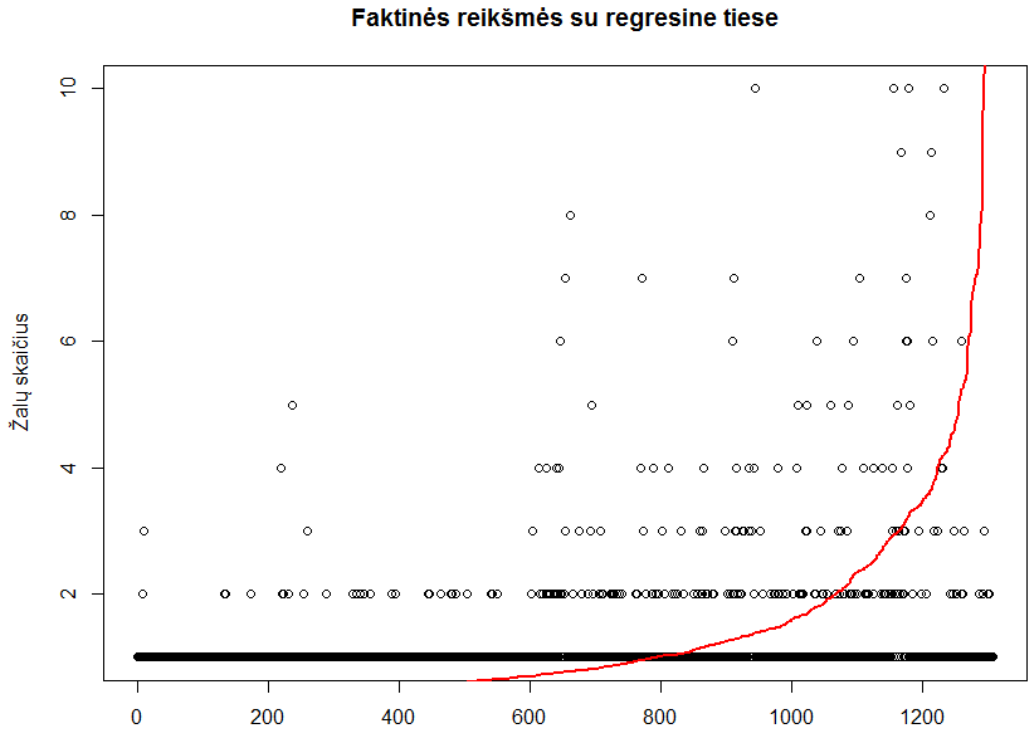
7 Lentelė. Geriausio modelio koeficientų reikšmingumas

Tarkime, reikšmingumo lygmuo yra $\alpha = 0{,}05$. Tuomet turime 6 indikatorius, kurie nėra statistiškai reikšmingi ir 11 reikšmingų kintamųjų (neįtraukiant bazinių dydžių). Dabar patikrinkime su visais teorijoje minėtais testais. Visų pirma, pateikime pirmojo ir trečiojo tipų testų rezultatus:

	Pirmojo tipo testas			Trečiojo tipo testas	
	Standartinis nuokrypis	χ^2	P-reikšmė	χ^2	P-reikšmė
Laisvasis Narys	3476,2	1391,26	<0,01	467,24	<0,01
Amžius	2085,0	237,55	<0,01	56,43	<0,01
Transporto Grupė	1847,4	90,19	<0,01	65,79	<0,01
Transporto Nuosavybė	1757,2	12,06	<0,01	10,14	<0,01
Kliento tipas	1619,2	134,93	<0,01	103,93	<0,01
Patirtis	1599,9	21,08	<0,01	37,98	<0,01
Transporto Amžius	1501,5	97,99	<0,01	104,56	<0,01
Transporto Variklio tūris	1486,6	14,04	<0,01	14,1	<0,01

8 Lentelė. Pirmojo ir trečiojo tipų testų patikrinimas geriausiam modeliui

Vėlgi matome, kad visi mūsų kintamieji yra reikšmingi, o standartinis nuokrypis sumažėjo nuo 3476,2 iki 1486,6. Dabar patikrinkime, kaip atrodo duomenų grafikas su regresijos tiese:



9 pav. Žalų skaičiaus duomenų sklaidos diagrama su regresine geriausiojo modelio kreive

5.1.4. Išskirtys

Dabar atlikime teorijoje aprašytą išskirčių patikrinimą. Atlikus testą gauname, kad 155 reikšmių yra išskirtys. Pažiūrėkime, kaip pasikeitė koeficientai:

Parametras	Koeficientas	Standartinis Nuokrypis	Įvertinys	P-reikšmė
Laisvasis narys	-0,13	0,12	0,88	0,27
Amžius (1 kategorija)	0,15	0,18	1,17	0,37
Amžius (2 kategorija)	-0,01	0,11	0,99	0,92
Amžius (3 kategorija)	-0,04	0,11	0,96	0,73
Amžius (4 kategorija)	0	-	1	-
Amžius (5 kategorija)	0,20	0,20	1,23	0,29
Amžius (6 kategorija)	-2,98	0,18	0,05	<0,01
Transporto Grupė (1 kategorija)	0	-	1	-
Transporto Grupė (2 kategorija)	0,11	0,08	1,12	0,15
Transporto Grupė (3 kategorija)	1,21	0,12	3,34	<0,01
Transporto Grupė (4 kategorija)	0,11	1,00	14,7	<0,01
Transporto Grupė (5 kategorija)	1,77	0,45	5,86	<0,01
Transporto Nuosavybė (1 kategorija)	0	-	1	-
Transporto Nuosavybė (2 kategorija)	2,11	0,38	8,24	<0,01
Transporto Nuosavybė (3 kategorija)	0,51	0,09	1,67	<0,01
Transporto Nuosavybė (4 kategorija)	1,03	0,13	2,81	<0,01
Registracijos vieta (1 kategorija)	0,12	0,09	1,13	0,16
Registracijos vieta (2 kategorija)	0	-	1	-
Registracijos vieta (3 kategorija)	-0,12	0,09	0,89	0,19
Kliento tipas (1 kategorija)	0,84	0,11	2,32	<0,01
Kliento tipas (2 kategorija)	0	-	1	-
Patirtis^2	-0,68	0,07	0,50	<0,01
Transporto priemonės amžius^2	0,68	0,15	1,99	<0,01
Transporto variklio tūris	0,22	0,05	1,24	<0,01

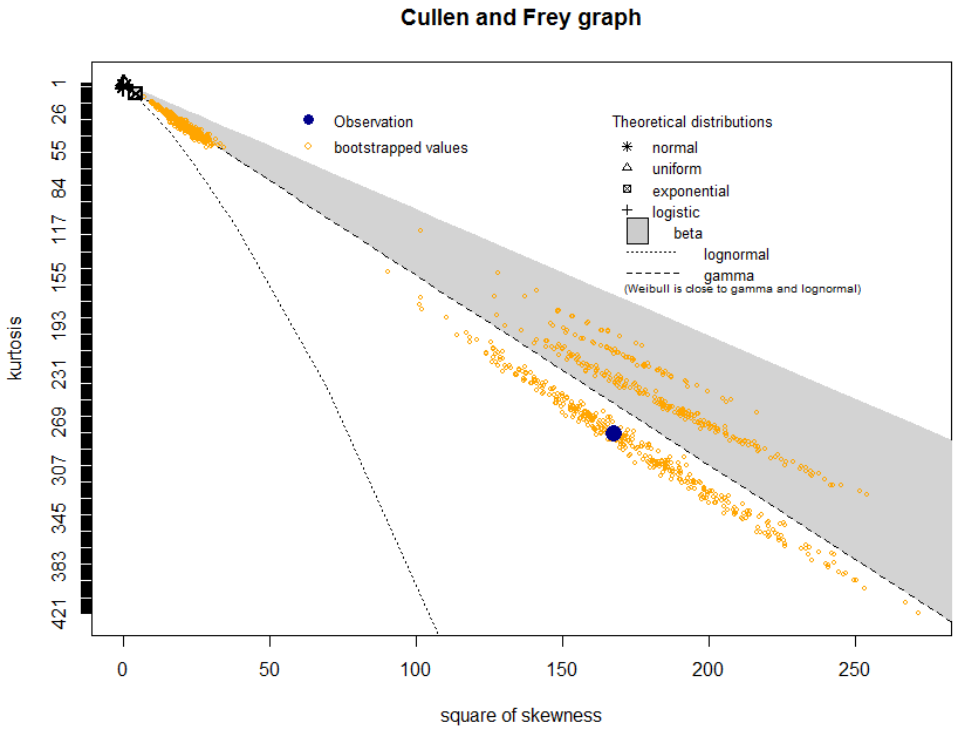
9 Lentelė. Geriausiojo modelio koeficientų sąrašas be išskirčių

Matome, kad situacija pasikeitė iš esmės. Dabar statistiškai nereikšmingų indikatorių turime 8, o reikšmingų 9. Vis dėlto, pagal teoriją pastarasis modelis turėtų būti geresnis.

4.2. Vidutinės žalų sumos modelis

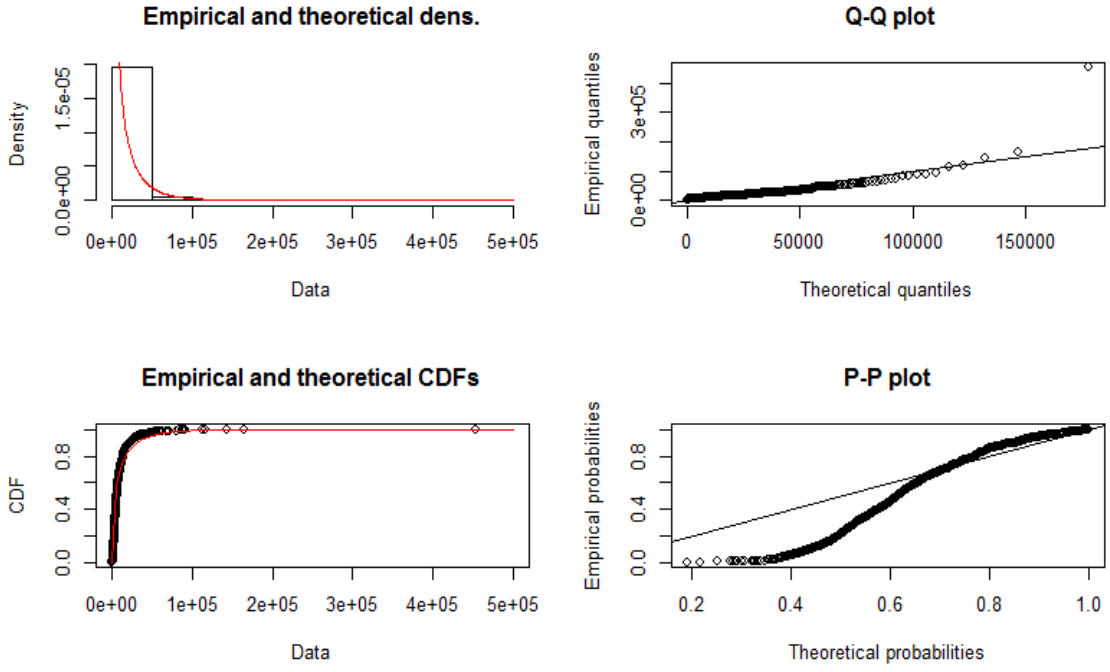
4.2.1. Skirstinio analizė

Analogiškai patikrinsime, koks skirstinys būtų geriausias mūsų turimiems duomenims. Pagal teoriją, šitam dydžiui geriausiai turėtų tikri Gama skirstinys. Patikrinkime, ką siūlo Cullen ir Frey grafikas:



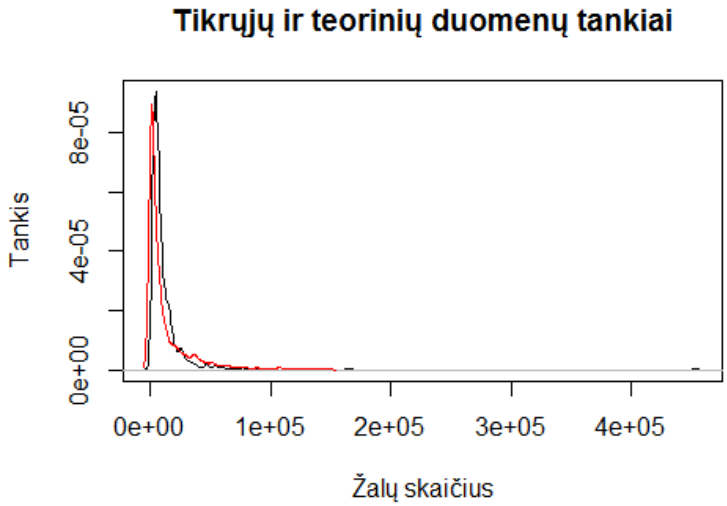
10 pav. Cullen ir Frey grafikas

Matome, kad iš tiesų geriausiai apibūdina Gama skirstinys. Pažiūrėkime, kaip grafiškai šis skirstinys tenkina duomenis:



11 pav. Gama skirstinio patikrinimas turimiems duomenims

Iš tiesų, rezultatai džiuginantys. Vėlgi pabandykime patikrinti, ar tankiai iš tiesų yra panašūs:

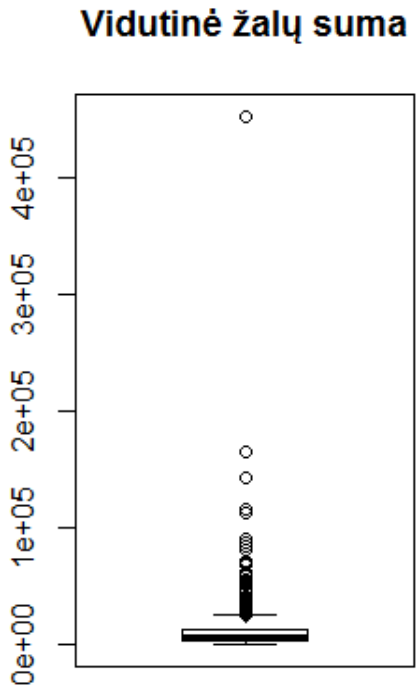


12 pav. Gama skirstinio patikrinimas turimiems duomenims

Matome, kad tikrai taip. Dabar patikrinkime, ar χ^2 suderinamumo testas mums suteiks tokius pačius rezultatus. Vis dėlto, testo statistika yra didesnė nei 0,05, todėl mūsų skirstinys nėra pats geriausias, bet tinkamesnio neturime.

4.2.2. Potencialių aiškinamųjų kintamųjų analizė

Vėlgi, kaip ir prieš tai susiduriame su problema - norint nagrinėti duomenis su stačiakampėmis diagramomis mes negalime to padaryti prasmingai, nes duomenys yra neinformatyvūs. Šį kartą pateiksime pavyzdį, kaip tai atrodo:



13 pav. Vidutinės žalų sumos stačiakampis grafikas

Nepaisant duomenų informatyvumo stokos galime pastebėti, kad remiantis teorija, visi duomenys, esantys už „ūsų“, yra laikomi išskirtimis. Vadinasi, duomenys apima labai didelę amplitudę, dėl kurios modelis gali būti paslinktas.

Taip pat nėra prasmės nagrinėti ir sklaidos grafikus, nes pakartotinai susidursime su pastarąja problema.

4.2.3. Modelio kūrimas

Šio modelio kūrimo specifika šiek tiek skiriasi nuo pastarojo. Šiame modelyje mes naudosime svorius, kurie atitinkamai bus žalų skaičius. Stengdamiesi pernelyg neišsiplėsti, šį kartą pateiksime tik geriausio modelio rezultatus ir modelio analizę. Kadangi duomenys nėra labai korektiški, tai rezultatai yra stebėtinai neįprasti. Geriausio modelio išraiška yra:

$$Vidutinė\ žalų\ suma = Patirtis^2 + Transporto\ variklio\ tūris^2$$

Kombinacijos su kitais kintamaisiais yra paprasčiausiai nereikšmingos. Peržiūrėkime, kaip atrodo reikšmingumo lentelė:

Parametras	Koeficientas	Standartinis Nuokrypis	Įvertinys	P-reikšmė
Laisvasis narys	9,24	0,09	10288	<0,01
Patirtis	0,37	0,11	1,44	<0,01
Transporto variklio tūris	-0,27	0,12	0,76	0.02

10 Lentelė. Geriausiojo modelio koeficientų sąrašas

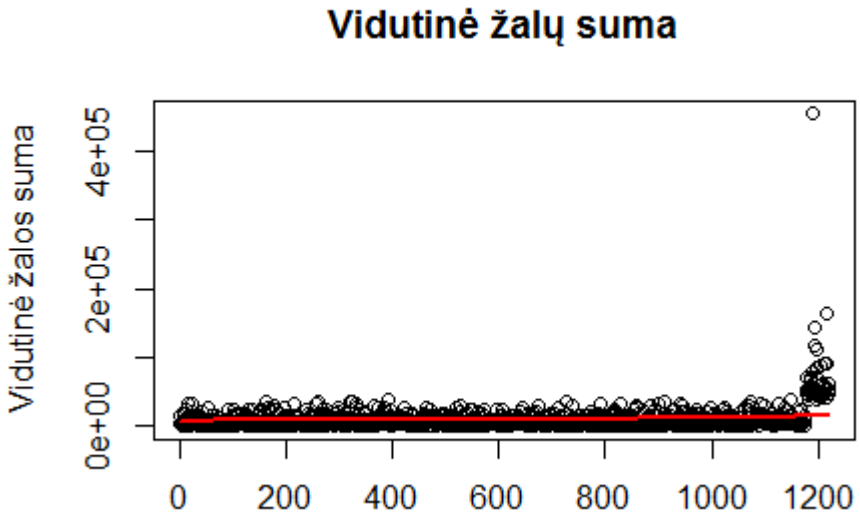
Šio modelio Akaike Informacinis kriterijus yra 34 394, o Schwarz Informacinis kriterijus yra 34 414 (mažiausias iš visų tikrintų modelių). Stebėtina tai, kad modeliai, ieškant mažiausio Schwarz mažiausio skaičiaus, buvo tikslesni ir reikšmingesni, nei modeliai remiantis Akaike.

Dabar patikrinkime, ar modelis tenkina pirmojo ir trečiojo tipų testus:

	Pirmojo tipo testas			Trečiojo tipo testas	
	Standartinis nuokrypis	χ^2	P-reikšmė	χ^2	P-reikšmė
Laisvasis narys	1446,3			1446,3	
Patirtis	1422,8	23,5	<0,01	11,8	<0,01
Transporto variklio tūris	1408,0	14,8	0,02	5,6	0,02

11 Lentelė. Geriausiojo modelio pirmojo ir trečiojo tipų rezultatai

Vėlgi matome, kad visi kintamieji, esantys modelyje, yra reikšmingi. Vis dėlto, galime teigti, kad modelio nesugebėjome reikšmingai pakeisti, nes standartinis nuokrypis nuo 1446,3 sumažėjo iki 1408, t.y. labai nežymiai, todėl galime manyti, jog bus didelė duomenų paklaida.



14 pav. Vidutinės žalų sumos duomenų sklaida su regresine kreive

4.2.4. Išskirtys

Atlikime išskirčių analizę mūsų turimam modeliui ir pažiūrėkime, ar kintamieji išliks reikšmingi ir, ar modelis pasidarys tikslesnis. Remiantis mūsų taikomu metodu, šiuo metu modelis turi 84 išskirtis. Pabandykime jas išmesti ir patikrinti naujųjų duomenų suderinamumą su modeliu:

Parametras	Koeficientas	Standartinis Nuokrypis	Įvertinys	P-reikšmė
Laisvasis narys	9,23	0,10	10288	<0,01
Patirtis	0,42	0,13	1,44	<0,01
Transporto variklio tūris	-0,27	0,13	0,76	0.04

12 Lentelė. Geriausiojo modelio be išskirčių koeficientų sąrašas

Matome, kad modelis nežymiai pasikeitė, bet modelio kintamieji ir toliau išliko reikšmingi. Dabar patikrinkime testų rezultatus:

	Pirmojo tipo testas			Trečiojo tipo testas	
	Standartinis nuokrypis	χ^2	P-reikšmė	χ^2	P-reikšmė
Laisvasis narys	1296,2			1296,2	
Patirtis	1271,5	24,7	0,02	11,9	<0,01
Transporto variklio tūris	1260,6	10,9	0,04	4,234	0,04

13 Lentelė. Geriausiojo modelio be išskirčių pirmojo ir trečiojo tipų rezultatai

4.3. Modelio prognozinių reikšmių patikrinimas su faktinėmis reikšmėmis

Visų pirma patikrinkime, kaip skirsis mūsų geriausių duomenų (su išskirtimis) prognozinės reikšmės nuo tikrųjų. Faktinių duomenų išmokų suma yra 17 945 361, tuo tarpu modelio prognozuojama suma yra 19 366 385, vadinasi mūsų padaryta paklaida yra ~7,3 %.

Dabar patikrinkime, kokia yra paklaida modeliuose be išskirčių. Kadangi ištrynėme tuos duomenis, kurie generavo pernelyg dideles reikšmes, tai išmokų suma sumažėjo iki 14 553 581. Mūsų modelių sugeneruotų prognozių sandaugos suma yra 14 475 068. Vadinasi, mūsų paklaida nesiekia nei vieno procento, t.y. 0,5 %. Taigi, modeliai be išskirčių yra kur kas tikslesni ir geresni.

5 IŠVADOS

1. Jeigu GLM modelyje priklausomas kintamasis yra diskretus dydis, tuomet imtis neturi būti didelė, norint gauti tinkamą naudojimui modelį, tuo tarpu, jeigu priklausomas kintamasis yra tolydus dydis, tokiu atveju, kuo didesnė imtis, tuo didesnė galimybė surasti statistiškai reikšmingus modelio koeficientus.
2. GLM modelis, iš esmės, skiriasi nuo tiesinės regresijos tuo, kad modelyje nėra būtinos normaliojo skirstinio sąlygos, tačiau dėl šio fakto visi potencialūs nepriklausomi kintamieji yra galimai tinkami modeliui. Taigi didelis kintamųjų skaičius skatina patikrinti itin didelį modelių skaičių, todėl, nagrinėjant gausų galimų aiškinamųjų kintamųjų rinkinį, patartina taikyti formules-algoritmus. Kitu atveju, tenka kliautis atsitiktinumu bei nuojauta ir daryti tai „rankiniu“ būdu. Vis dėlto, taikant formules-algoritmus, naudojami visi kompiuterio resursai, iš ko išplaukia, kad galingos sudėties kompiuteriai yra būtinybė, o rankiniu būdu atrinkti kintamieji negarantuoja, kad atrastas modelis yra tinkamiausias. Be kita ko, taikant formules-algoritmus, tokius kaip $glmulti$, mes negalime patikrinti modelio kintamųjų reikšmingumo, o galime tik nurodyti pagal kokį informacinį kriterijų mes ieškome praktiškiausio modelio. Tai nesuteikia garantijos, kad atlikus išsamią modelio analizę, visi nepriklausomi kintamieji bus reikšmingi.

6 LITERATŪRA

1. V. Čekanavičius, G. Murauskas. *Statistika ir jos taikymai III dalis,* Vilnius: TEV, 2009 m.

2. V. Čekanavičius, G. Murauskas. *Statistika ir jos taikymai I dalis,* Vilnius: TEV, 2001 m.

3. P. de Jong, G. Z. Heller. *Generalized Linear Models for Insurance Data*, 2008 m.

4. R. Lapinskas, *Practical Econometrics I. Regression Models. Lecture Notes,* Vilnius, 2013m.

5. Lietuvos Respublikos Seimas. Draudimo įstatymas, 2003 m. http://www3.lrs.lt/pls/inter3/dokpaieska.showdoc_l?p_id=494450&p_tr2=2