UNIVERZA V LJUBLJANI
FAKULTETA Z MATEMATIKO IN FIZIKO

Finančna matematika – 2. stopnja

Miha Petrič

Uporaba posplošenih linearnih modelov pri modeliranju višine škod v zavarovalništvu

Magistrsko delo

Mentor: izred. prof. dr. Janez Bernik
Somentor: asist. dr. Aleš Toman

Ljubljana, 2017

Zahvala

Hvala vsem, ki ste mi stali ob strani pri pisanju magistrskega dela. Predvsem bi se rad zahvalil Tei za njeno pomoč in potrpežljivost, staršem in Martini za spodbude, izr. prof. dr Janezu Berniku za podporo, posebna zahvala pa gre asist. dr. Alešu Tomanu za vso pomoč. Brez njega bi naloga v taki obliki težko nastala.

Kazalo

Program dela

V magistrskem delu obravnavajte možnosti uporabe linearne regresije in posplošenih linearnih modelov v zavarovalništvu. Analizirajte prednosti in slabosti posameznega modela, ter na konkretnih podatkih predstavite vse korake statističnega modeliranja, in sicer: pregled podatkov, izbiro modela, ocenjevanje modelskih parametrov, interpretacije dobljenih ocen in validacijo modela.

Za osnovno literaturo uporabite deli

- D. Anderson et all, *A practicioner's guide to generalized linear models*,Tower watson, 2007,
- A. Agresti, *Foundations of linear and generalized linear models*, John Wiley & Sons,2015.

Ljubljana, 2017 izr. prof. dr. Janez Bernik

Povzetek

Zavarovalnice že od nekdaj iščejo načine, kako bi modelirale spremenljivke s pomočjo drugih, znanih, pojasnjevalnih spremenljivk. Širše poznana in uporabljena je metoda linearne regresije, vendar ima kar nekaj slabosti in pomanjkljivosti, saj v osnovni obliki zahteva normalno porazdelitev spremenljivk, ki jih modeliramo. Posplošeni linearni model (ang. generalized linear model, v nadaljevanju GLM) je posplošitev linearne regresije, ki le to razširi do te mere, da z njim lahko modeliramo skoraj vsako vrsto spremenljivk, ne glede na porazdelitev, ki jo določa. GLM izhaja iz ideje linearne regresije, nato pa s pomočjo povezovalne funkcije spremeni aditivnost faktorjev pri pojasnjevalnih spremenljivkah v neko drugo funkcijo, ki bolj ustreza lastnostim naših podatkov. Širok nabor možnosti, ki jih imamo pri GLM-u, je tako prednost kot breme, saj je težko identificirati prave izbire. V modeliranju smo se poskusili tudi sami. S pomočjo GLM-a smo na testnih podatkih poskusili ugotoviti vpliv različnih pojasnjevalnih spremenljivk na pričakovano višino škode. Na koncu naloge pa smo preverili, kako uporabljen je GLM v slovenskih zavarovalnicah.

Abstract

Insurance companies are always looking for a way to fit data as best as possible. The most widely used method for data fitting is linear regression, which is simple to use and understand, but has it's own limitations and flaws. GLM or generalized linear model is a generalization of the linear regression model. It broadens the possibility to fit almost any kind of data regardless the distibutions by which the variable of interest is distributed. GLM takes the core of the linear regression model and then changes the additive factoring model by using a link function to transform it into something we can fit our data better with. The abundance of choices when dealing with GLM is a big advantage, but a disadvantage at the same time. We need to be carefull to make the right decisions when modelling. We modeled some sample insurance data ourselves. We tried to determine how different variables affect expected claims. We also tried to find out if GLM is used at all in Slovenian insurance companies.

Math. Subj. Class. (2010): 62J12, 97M30, 62J05, 97K80

Ključne besede: zavarovalnica, posplošen linearni model, regresijski model, linearna regresija, modeliranje podatkov

Keywords: insurance company, generalized linear model, regression model, linear regression, data modelling

1. Uvod

Živimo v digitalnem svetu, svetu, ki si ga brez tehnologije ne znamo predstavljati. Vsak dan vsak od nas, bodisi zasebno bodisi službeno, ustvari velike količine podatkov, ki vsebujejo informacije o skoraj vsem, kar se dogaja v in okoli nas. Zavedanje, da lahko razpolagamo z veliko količino podatkov, vedno znova odpira nove ideje o tem, kako preiskovati zveze med različnimi pojavi in kako lahko uporabimo pretekle podatke, da bi nam pomagali razumeti prihodnje dogajanje na najrazličnejših področjih. Znanost, ki poskuša oblikovati nova spoznanja na podlagi zbranih podatkov, je statistika.

Statistika je postala zelo pomembna na vseh področjih, izredno veliko se uporablja tudi na področju financ, kjer vsak dan nastajajo novi modeli za analize podatkov. Banke, zavarovalnice in druge finančne institucije konstantno iščejo nova znanja na tem področju, saj si tako pridobijo različne finančne koristi. Za zavarovalnice je zelo pomembno postalo predvem napovedovanje najrazličnejših kazalcev, od višine donosov, višine škod do njihove frekvence itd. Za napovedovanje podatkov iz obstoječih podatkov se najpogosteje uporablja najbolj intuitivna metoda na tem področju, to je linearna regresija.

Linearna regresija je priljubljena zaradi svoje preprostosti, hkrati pa v določenih primerih dobro napovedoveduje vrednosti, ki smo jih želeli dobiti iz poznavanja pojasnjevalnih spremenljivk. Metoda linearne regresije za odvisno spremenljivko predpostavlja normalno porazdelitev. Normalna porazdelitev je, podobno kot linearna regresija, najbolj preučevana porazdelitev, saj omogoča lažje izračune. Vendar vse spremenljivke niso vedno porazdeljene normalno in za takšne primere linearna regresija ni ustrezen način modeliranja.

Te omejitve so matematike in statistike gnale k razvoju načinov za modeliranje, ki bi bili bolj splošni in bi se jih dalo uporabiti na mnogo različnih področjih. Ta raziskovanja so privedla do posplošenega linernega modela (ang. generalized linear model, v nadaljevanju GLM). GLM se na začetku ni najbolje uveljavil. Večina analitikov in znanstvenikov v razširitvi linearnega regresijskega modela ni videla dovolj prednosti, da bi se naučili uporabe GLM-a ter ga uporabljali za napovedovanje.

Globalna finančna in nato gospodarska kriza leta 2008 je spremenila miselnost pri preučevanju in modeliranju podatkov. Stare metode so se umaknile novim. GLM je oživel in danes postaja alternativa za linearno regresijo, saj so podjetja spoznala, da so dobri podatkovni modeli pomembni za dobro poslovanje družb. Hkrati so pri napovedovanju podjetja od krize dalje postala veliko bolj previdna pri uporabi modelov, ki privzemajo normalne porazdelitve, saj se je izkazalo, da je to lahko zelo nevarno početje v primeru, ko preučevane spremenljivke niso porazdeljene približno normalno. GLM je postal model, ki je dobra kombinacija intuitivnosti in preprostosti za razumevanje, hkrati pa lahko z njim modeliramo dovolj širok nabor različnih pojavov, ne da bi za vse predpostavili samo eno porazdelitev.

Primer podjetij, za katere je GLM zelo uporabna metoda za modeliranje, so zavarovalnice. Točne napovedi prihodnjih pojavov so za zavrovalnice zelo pomembne, saj morajo imeti na voljo dovolj kapitala, da pokrijejo možne izgube, ki se lahko zgodijo. Prav tako bi zavarovalnice rade modelirale veliko pojavov, kjer podatki

očitno niso normalno porazdeljeni. Dober primer je modeliranje višine škodnih zahtevkov, saj so škodni zahtevki navzgor neomejeni in nenegativni. To pomeni, da so po opisu bolj podobni neki gama porazdeljeni slučajni spremenljivki, kot normalno porazdeljeni slučajni spremenljivki. Tak primer podatkov dokaj očitno zavrača predpostavke linearne regresije. Ustrezen GLM se zdi naravna izbira.

Zelo pomemben del magistrskega dela je uporaba regresijskih modelov na podatkih, ki jih bomo uporabljali za določitev pričakovane višine škode. Zato se bosta obravnava podatkov in teorija, ki obravnavo podpira, prepletali in dopolnjevali. Tako bomo vedno lahko videli, kako se teorijo uporablja tudi v praksi.

Na začetku je najbolj pomembno, da se s podatki spoznamo. Poznati želimo vse vidike, ki bi nam lahko pomagali pri kasnejšem modeliranju. Zato bomo najprej podatke opisali, povedali, kakšne so pojasnjevalne spremenljivke in kakšna je odvisna spremenljivka. Za lažjo predstavo bomo podatke prikazali tudi grafično. Na tem mestu bomo tudi že izvedli prve statistične teste, da bomo podatke že pred modeliranjem uredili ter se odločili, kateri so za modeliranje bolj in kateri manj pomembni.

Nato bomo spoznali oz. obnovili znanje linearne regresije. Dobro poznavanje linearne regresije je osnova za modeliranje z GLM, hkrati pa se veliko pojmov, ki so pomembni pri linearni regresiji, ponovi tudi pri GLM. Linearno regresijo bomo uporabili na testnih podatkih, nato pa bomo poskusili še oceniti kvaliteto linearnega regresijskega modela. Na koncu razdelka bomo nekaj povedali še o prednostih in slabostih linearne regresije.

Potem bomo spoznali teorijo, ki je potrebna za razumevanje samega GLM-a ter modeliranje podatkov z njim. Ker lahko z GLM modeliramo spremenljivke, ki so porazdeljene s porazdelitvami iz t.i. eksponentne družine porazdelitev, bomo to družino podrobneje predstavili. Nato bomo s pomočjo povezovalne funkcije in funkcije variance spoznali, kako linearni regresijski model posplošimo v GLM, kako GLM izračunava ocene za modelske parametre ter se naučili ocenjevati primernost GLM-a. Vse to bomo seveda nato preizkusili tudi v praksi, kjer bomo najprej ocenili kvaliteto modela in ga nato poskušali izboljšati. Na koncu bomo še ugotavljali kako model modelira še neznane podatke, ter ocenjene vrednosti primerjali s pravimi. Za zaključek bomo izvedeli, kako je z uporabo GLM-a v zavarovalnicah v Sloveniji.

2. Opis podatkov

Najpomembnejši del magistrskega dela je uporaba GLM na podatkih, s katerimi se pogosto srečamo v zavarovalništvu. V ta namen smo si izbrali podatke o višini škod avtomobilskih zavarovanj. Pri delu s podatki je zelo pomebno, da jih pred analizo čim bolj podrobno spoznamo, kar nam zelo olajša delo kasneje. V tem razdelku magistrskega dela bomo torej podrobneje spoznali podatke, s katerimi bomo delali v nadaljevanju magistrskega dela.

V Sloveniji so premoženjska zavarovanja leta 2016 obsegala 76,2 % obračunane premije. Velik delež te premije (30 odstotkov) je predstavljala premija za zavarovanje motornih vozil in avtomobilske odgovornosti [9], kar pomeni, da je modeliranje škodnih zahtevkov za zavarovanja motornih vozil za zavarovalnice zelo uporabno.

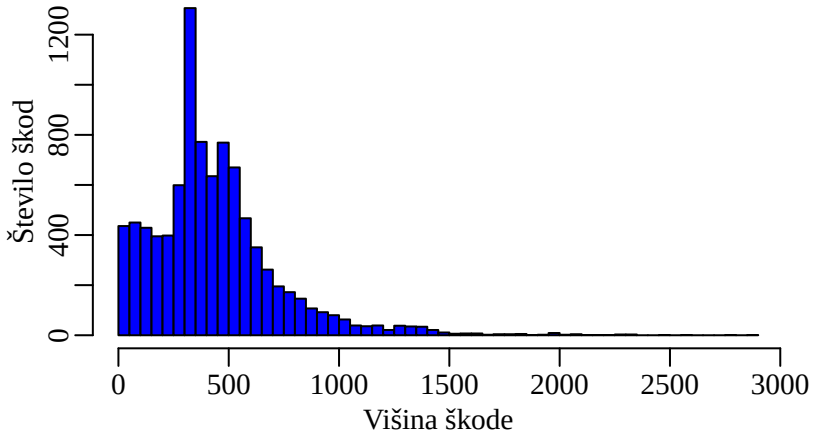
Slovenskih podatkov na žalost nismo dobili za uporabo, zato smo na internetu poiskali podatke, ki bi bili za nas uporabni. Priročne podatke smo tako našli na strani programa za analizo podatkov Emcien scan [8], na kateri se nahaja tudi nekaj podatkovnih paketov, s katerimi bi lahko preiskusili njihov program. Tako smo dobili CSV datoteko s podatki o višini škod avtomobilskih zavarovanj iz ZDA, natančneje srednjega zahoda ZDA. Podatkovni paket vsebuje 9314 škodnih zahtevkov. Vsak škodni zahtevek ima napisano višino škodnega zahtevka, poleg tega pa so zraven še različni podatki o zavarovalni polici, zavarovalcu, vozilu ipd. Vse podatke smo najprej poslovenili, nato pa podatke še uredili, tako da so kategorije pri faktorjih urejene v ustreznem vrstnem redu. Tako lahko višino škodnih zahtevkov modeliramo s pomočjo podatkov, ki jih imamo o zavarovanju, ko je le to sklenjeno. Faktorji, ki jih lahko uporabimo pri modeliranju, so:

- zvezna država,
- vrsta kritja,
- izobrazba zavarovalca,
- stanje zaposlitve,
- spol,
- tip naselja,
- poročni stan,
- prodajna pot,
- velikost vozila,
- tip vozila,

Uporabljene spremenljivke so različnih tipov. Odvisna spremenljivka višina škode je zvezna (razmernostna) slučajna spremenljivka, faktorji pa so diskretni. Nekateri so dihotomni (spol), drugi nominalni (poročni stan), tretji pa ordinalni (izobrazba).

Vsaka izmed pojasnjevalnih spremenljivk ima več različnih kategorij, ki jih bomo podrobneje spoznali v nadaljevanju razdelka. Najprej si lahko pogledamo brezpogojno porazdelitev višine škod, ki nam bo kasneje pomagala pri izbiri GLM-a, saj lahko nekaj lastnosti naših podatkov že takoj razberemo iz slike. Slika 1 prikazuje histogram višin škod.

Pri podatkih o višini škod lahko iz slike 1 razberemo dve glavni lastnosti, ki sta, če govorimo o višini škod, zelo jasni. Višina škode je nenegativna, poleg tega pa



Slika 1. Porazdelitev višine škode.

je višina škodnega zahtevka navzgor neomejena oz. se lahko pojavijo zelo visoke vrednosti. Za modeliranje je uporabno, da poskušamo podatke približati neki verjetnostni porazdelitvi, ki nam pomaga pri modeliranju podatkov in pojasnjevanju napak modela. Lastnosti, ki smo ju navedli zgoraj, nakazujeta, da bi bila za nas zanimiva gama porazdelitev. To porazdelitev bomo bolj podrobno opisali v nadaljevanju dela.

2.1. Opis pojasnjevalnih spremenljivk

Poleg pregleda spremenljivke, ki jo bomo modelirali in napovedovali, je nujno raziskati tudi spremenljivke s katerimi si bomo pomagali pri ocenjevanju vrednosti višine škod. Zanima nas, če imajo različne vrednosti posamezne pojasnjevalne spremenljivke kakšen vpliv na višino škode. Pomembno je tudi število opazovanj pri posameznih kategorijah, ter ali so vrednosti približno simetrično porazdeljene okoli povprečij višine škode pri posemeznem faktorju, kar bi lahko nakazovalo na normalnost, ali pa imamo opravka s kakšnimi drugimi porazdelitvami, ki bi potrebovale posebno pozornost.

Ker imamo na voljo kar veliko različnih pojasnjevalnih sprememljivk, bi najprej radi iz modeliranja izločili tiste, pri katerih imamo na različnih ravneh pojasnjevalne spremenljivke približno enake višine škodnih zahtevkov, in se nam torej ne zdijo pomembni pri modeliranju. Statistični test, ki ga lahko opravimo za ta namen, je enosmerna analiza variance oz. enosmerni ANOVA. Tega bomo malo bolj podrobno spoznali v razdelku, kjer bomo govorili o linearni regresiji. Za nas je bolj pomemben opis, enosmernega ANOVA testa.

Enosmerni ANOVA test se pogosto uporablja, ko imamo vzorčne in populacijske enote razvrščene v različne skupine. V našem primeru so enote razvrščene v skupine glede na vrednosti na posameznih faktorjih.

Enosmerni ANOVA test predpostavlja, da je preučevana spremenljivka porazdeljena normalno na vsaki enoti, ter da ima v vseh skupinah enako varianco. V našem

primeru to zelo verjetno ni res, je pa enosmerni ANOVA test kljub temu dovolj robusten. Tako bomo dobili vsaj okvirno sliko, ki nam bo povedala, katere pojasnjevalne spremenljivke se nam splača vzeti pod drobnogled.

Ničelna hipoteza za enosmerni ANOVA test je, da imajo povprečja preučevane spremenljivke v vseh skupinah enako vrednost. Enosmerni ANOVA test vrne F-statistiko, nato pa lahko s pomočjo F-porazdelitve izračunamo p-vrednost, ki nam nato pove ali pri izbrani stopnji značilnosti sprejmemo ali zavrnemo ničelno hipotezo. Naše teste bomo izvajali s 5% stopnjo značilnosti.

Enosmerni ANOVA test lahko dokaj preprosto izvedemo v programu R.

```
oneway.test(Skoda ~ Zvezna.Drzava, var.equal=TRUE, data=podatki)
```

V testu smo primerjali višino škode po posameznih zveznih državah. Dobimo naslednji rezultat.

```
One-way analysis of means

data:  Skoda and Zvezna.Drzava
F = 0.59175, num df = 4, denom df = 9129, p-value = 0.6686
```

Vrednost F nam pove vrednost F statistike, zraven imamo še navedene prostostne stopnje njene testne porazdelitve. p-vrednost nam pove, če ničelno hipotezo sprejmemo ali zavrnemo. Ker je p-vrednost višja od $0,05$, pomeni, da ničelne hipoteze ne moremo zavrniti. Zaključimo, da zvezna država, iz katere prihaja zavarovanje, nima vpliva na višino škode.

Tabela 1 vsebuje enosmerne ANOVA teste za vse pojasnjevalne spremenljivke. Na

Tabela 1. Enosmerni ANOVA test.

Pojasnjevalna spremenljivka	p-vrednost
Zvezna država	0,67
Kritje	$2{,}2 \cdot 10^{-16}$
Izobrazba	$2{,}2 \cdot 10^{-16}$
Zaposlitveni status	$2{,}2 \cdot 10^{-16}$
Spol	$9{,}65 \cdot 10^{-13}$
Tip naselja	$2{,}2 \cdot 10^{-16}$
Poročni stan	$2{,}2 \cdot 10^{-16}$
Tip police	0,91
Prodajna pot	0,66
Tip vozila	$2{,}2 \cdot 10^{-16}$
Velikost vozila	$2{,}2 \cdot 10^{-16}$

podlagi rezultatov iz tabele 1 bomo iz nadaljnje analize podatkov izločili tri pojasnjevalne spremenljivke, pri katerih nismo mogli zavrniti ničelne hipoteze enosmernega

ANOVA testa. To so zvezna država, tip police in prodajna pot. Vse ostale pojasnjevalne spremenljivke pa bomo porobneje preučili.

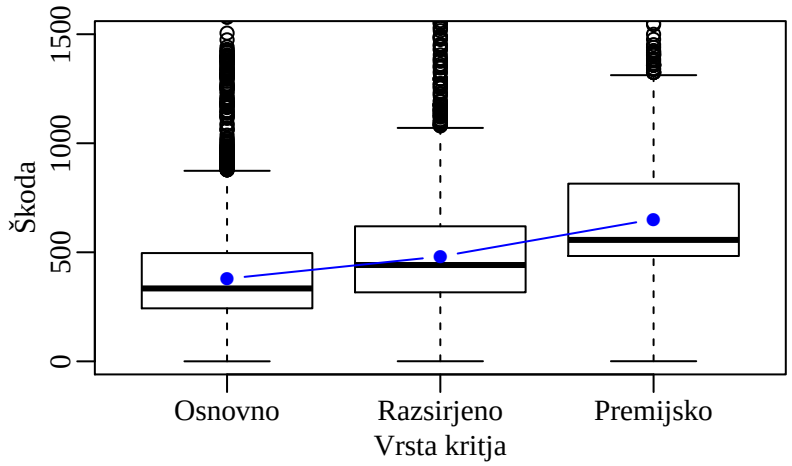
Na tem mestu nas bosta zanimali predvsem dve stvari. Za vsako pojasnjevalno spremenljivko bomo narisali grafikon kvantilov, to je graf, ki nam nakaže porazdelitev, ki bi jo lahko imela višina škode pri posamezni skupini vsake pojasnjevalne spremenljivke. Če je črta, ki označuje mediano, na sredini pravokotnika, bi to lahko pomenilo, da imamo opravka z normalno porazdeljenimi podatki, vidnejša odstopanja pa nakazujejo drugačne porazdelitve.

Poleg tega bomo po skupinah prikazali tudi povprečje višine škod vsake posamezne skupine. Statistični test za razlike med skupinami smo že naredili. Na tem mestu bomo preverili, med katerimi skupinami so razlike največje in kako velike so. Na koncu bomo še prešteli število opazovanih vrednosti po skupinah. Če bosta imeli dve vsebinsko sorodni skupini podatkov zelo podobno povprečno višini škode, hkratii pa opazovanih vrednosti ne bo veliko, lahko razmišljamo o združitvi obeh skupin v eno večjo skupino.

Kritje. Podatek o kritju na polici je pomemben, saj je od tega odvisno kaj vse bo krila zavarovalnica. Boljše, kot je kritje, več lahko od zavarovalnice zahtevamo, če se nam kaj zgodi. V naših podatkih imamo za kritje tri skupine: osnovno, razširjeno in premijsko kritje.

Tabela 2. Razvrstitev enot glede na vrsto kritja.

Skupina	Osnovno	Razširjeno	Premijsko
Število enot	5568	2742	824



Slika 2. Višina škode glede na vrsto kritja.

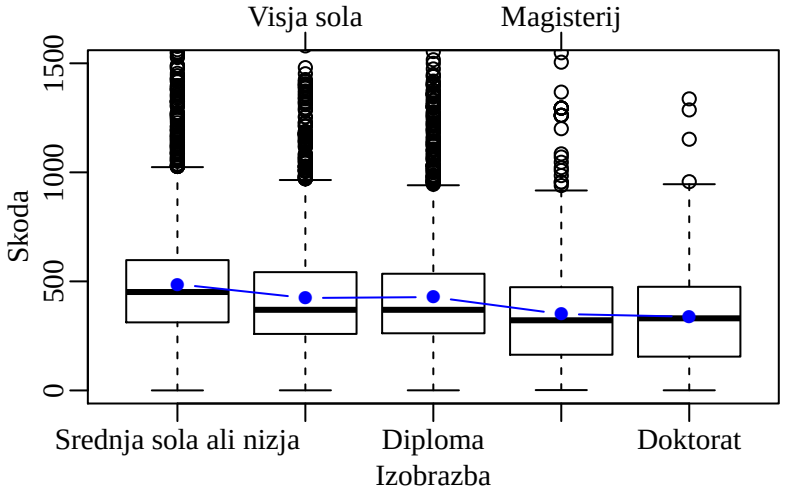
Kakor lahko vidimo na sliki 2, se višina škode za različne vrste kritja razlikuje. Opazimo tudi, da je za vsako vrsto kritja mediana pod povrečjem škode, vsi osamelci pa so nad grafom, kar nakazuje, da nimamo normalno porazdeljenih vrednosti. Vsi

osamelci na tem grafu niso prikazani, da lažje opazimo razlike v osrednjem delu porazdelitve. Za primer vrste kritja se nam skupin ne splača združevati.

Izobrazba. Druga pojasnjevalna spremenljivka, ki jo bomo pogledali, je izobrazba. Ta spremenljivka ima 5 različnih skupin. Pri sklepanju zavarovanj je zelo pogosto, da nas vprašajo o najvišji doseženi stopnji izobrazbe. Tudi s pomočjo tega podatka bomo modelirali višino škode.

TABELA 3. Razvrstitev enot glede na stopnjo izobrazbe.

Skupina	Sr. šola ali nižja	Višja šola	Diploma	Magisterij	Doktorat
Število enot	2622	2681	2748	741	342



SLIKA 3. Višina škode glede na stopnjo izobrazbe.

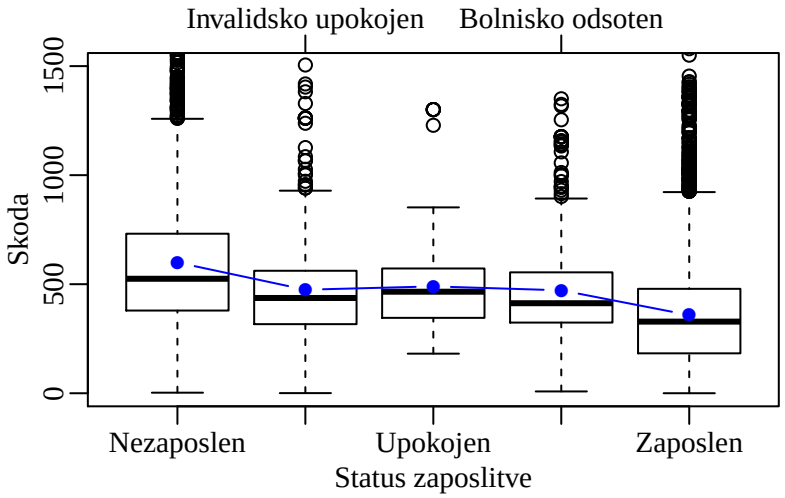
Na sliki 3 vidimo, da so tokrat mediane približno na sredini pravokotnikov, kar nakazuje, da bi bili podatki o višini škode znotraj teh skupin lahko normalno porazdeljeni. Iz tabele 3 preberemo, da imata skupini magisterij in doktorat manjše število enot, kot ostale skupine, hkrati pa sta tudi njuna grafa podobna. Odločimo se, da bomo skupini združili v eno skupino: podiplomska.

Status zaposlitve. Tudi pri tej pojasnjevalni spremenljivki imamo pet različnih skupin. Zavarovalca ob sklenitvi zavarovanja vprašajo, kakšen je njegov zaposlitveni status.

TABELA 4. Razvrstitev enot glede na status zaposlitve

Skupina	Nezaposlen	Inv. upokojen	Upokojen	Bolniška	Zaposlen
Število opažanj	2317	405	282	432	5698

Pri tej pojasnjevalni spremenljivki najprej opazimo, da imamo dve veliki skupini (zaposleni in nezaposleni), in tri dokaj majhne. Razmislili bi lahko o združitvi treh



Slika 4. Višina škode glede na status zaposlitve.

manjših skupin. Tudi tokrat bomo te skupine združili v eno večjo skupino, saj iz slike 4 vidimo, da so tudi njihovi grafi dokaj podobni. Naredimo torej skupino delavno neaktivnih. V njej so združeni zavarovalci, ki zaradi starosti in bolezni ne delajo.

Tip vozila. Ta spremenljivka pove tip vozila, ki je zavarovalni objekt police. Manjša in manj vredna vozila imajo ponavadi nižje škodne zahtevke, kot bolj vredna vozila, kot so luksuzni avtomobili in terenski avtomobili. Za začetek imamo šest različnih skupin.

Tabela 5. Število enot glede na tip vozila.

Skupina	Dvovratno	Štirivratno	Športno	Športni terenec	Luksuzno	Luksuzni terenec
Število enot	1886	4621	484	1796	163	184

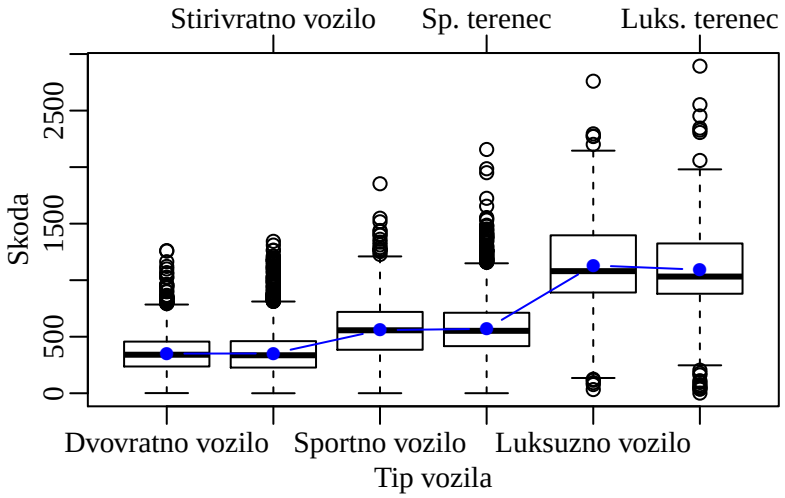
Iz tabele 5 in slike 5 je razvidno, da imamo 3 različne skupine vozil s tremi različnimi povprečnimi višinami škodnih zahtevkov. V kategoriji luksuznih vozil je enot zelo malo, tako da je smiselno luksuzna vozila in luksuzne terence združiti v eno skupino luksuznih vozil. Prav tako lahko združimo športna vozila in športne terence v eno skupino športnih vozil. Skupini dvovratnih in štirivratnih vozil sta si zelo podobni, ampak sta zelo veliki in ju bomo zato zaenkrat pustili ločeni.

Poročni stan. Zanimivo bo videti, kakšen vpliv ima na višino škode poročni stan zavarovalca ob sklenitvi zavarovanja

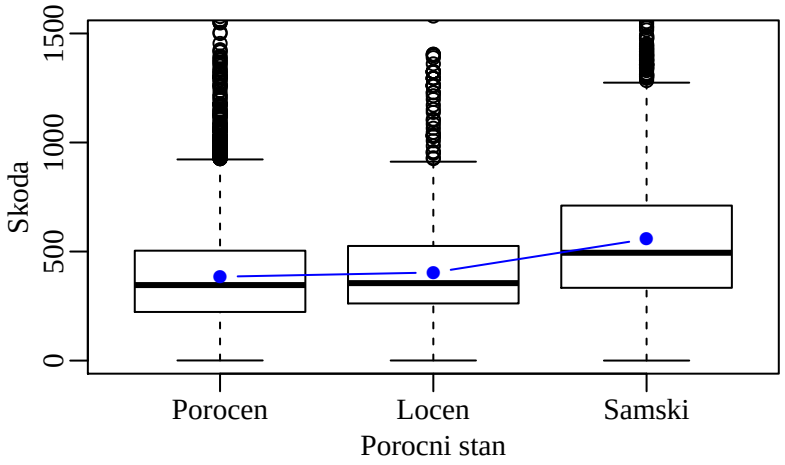
Tabela 6. Število enot glede na poročni stan.

Skupina	Poročen	Ločen	Samski
Število enot	5298	1369	2467

Na sliki 6 in v tabeli 6 vidimo, da imamo tri dokaj velike skupine, ki imajo tudi različna povprečja višin škodnih zahtevkov. Opazimo lahko, da so mediane na



Slika 5. Višina škode glede na tip vozila.



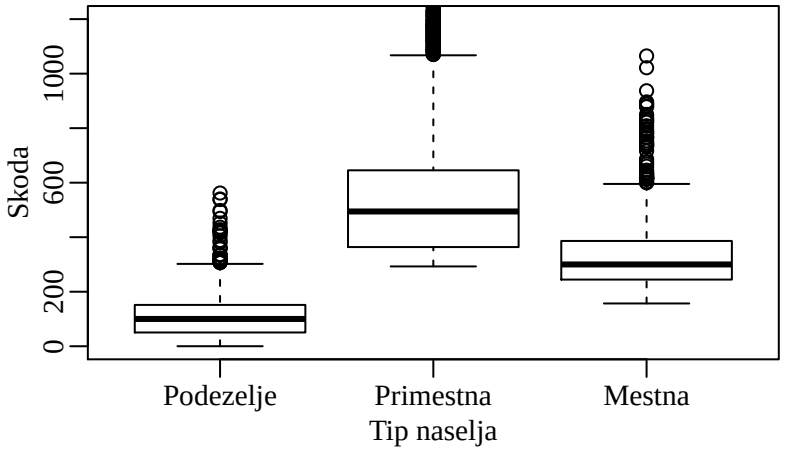
Slika 6. Višina škode glede na poročni stan.

sredini grafov, kar lahko pri tej spremenljivki nakazuje pogojno normalnost. Pri tej pojasnjevalni spremenljivki kategorij ne bomo združevali v večje skupine, saj se nam to na začetku ne zdi potrebno.

Tip naselja. S to spremenljivko opišemo tip naselja zavarovalčevega doma. Dinamika vožnje, število drugih vozil in verjetnost, da se zgodi nesreča, je lahko v mestih precej drugačna kot na podeželju. Ta spremenljivka nam ta vpliv pomaga razumeti.

Tabela 7. Število enot glede na tip naselja.

Skupina	Podeželje	Primestna	Mesto
Število enot	1773	5779	1582



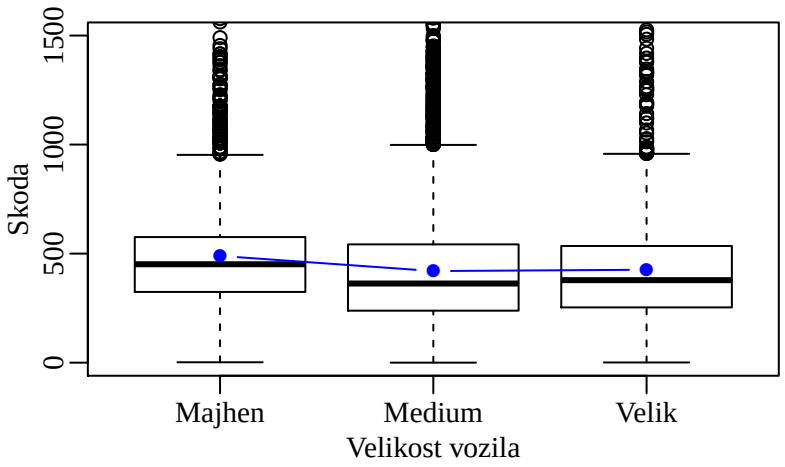
Slika 7. Višina škode glede na tip naselja.

Pri tej spremenljivki lahko vidimo, da ima vsaka skupina zelo velik vpliv na povprečje. Poleg tega so skupine tudi todaj velike, kar pomeni, da tudi tukaj ne bomo združevali skupin. Opazimo lahko tudi, da so grafi nesimetrični.

Velikost vozila. V podatkih imamo poleg tipa vozila na voljo tudi podatek o velikosti vozila. Manjša vozila so ponavadi bolj poceni, kar lahko nakazuje nižje škodne zahtevke.

Tabela 8. Število enot glede na velikost vozila.

Skupina	Majhno	Medium	Veliko
Število enot	1764	6424	946



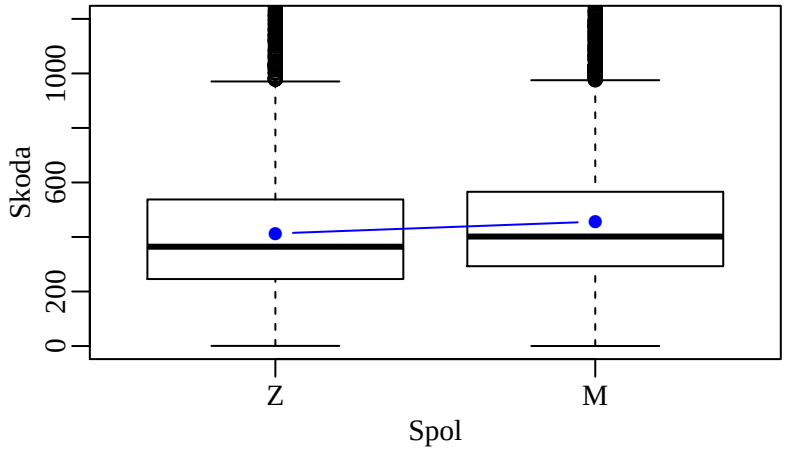
Slika 8. Višina škode glede na velikost vozila.

Tudi pri tej pojasnjevalni spremenljivki imamo tri velike skupine z različnimi povprečji, tako da se nam združevanje ne izplača.

Spol. Zadnja spremenljivka, ki jo bomo obravnavali, je spol zavarovalca.

Skupina	Ženska	Moški
Število opažanj	4658	4476

TABELA 9. Število enot glede na spol zavarovalca.



SLIKA 9. Višina škode glede na spol zavarovalca.

Iz tabele 9 vidimo, da so ženske in moški zastopani v približno enakih deležih. Na sliki 9 vidimo, da sta povrečna škodna zahtevka različnih višin, kar pomeni, da neke razlike med spoloma vendarle obstajajo. Več bodo pokazali testi v naslednjih razdelkih.

Do sedaj smo si večinoma ogledovali vpliv posamezne spremenljivke na višino škodnega zahtevka. Izločili smo spremenljivke, kjer se nam je zdelo, da vpliv ne bo bistveno vplival na višino škode. Nato smo vse spremenljivke predstavili vizualno ter po potrebi skupine združili v večje skupine, ki so za nas bolj smiselne.

Naslednja stvar, ki nas pri pregledu podatkov zanima, pa je, ali so spremenljivke med sabo kako povezane. Če bi na primer vsi na podeželju vozili velike avtomobile, vsi v mestih pa majhne, lahko eno izmed spremenljivk opustimo, saj je nepotrebna. Pogledali si bomo, če med spremenljivkami obstaja interkorelacija. To bomo naredili s pomočjo statistike, ki ji rečemo Cramerjeva V-statistika.

2.2. **Cramerjeva V-statistika.**

Cramerjeva V-statistika je statistična mera povezanosti oz. interkorelacije med dvema sprmenljivkama, ki sta kategorični, torej imata več kategorij. Cramerjeva V-statistika je osnovana na Pearsonovi χ^2-statistiki, s katero testiramo neodvisnost spremenljivk.

Za izračun Cramerjeve V-statistike med dvema spremenljivkama pripravimo kombinacijsko tabelo, kjer vrstice in stolpci predstavljajo kategorije, v celicah tabele pa so frekvence enot z ustrezno kombinacijo lastnosti. Za izračun Cramerjeve V-statistike lahko uporabimo funkcijo `cramersV`, ki jo najdemo v knjižnici `lsr` v programu R. Funkcija sprejme podatke v obliki, ki smo jo opisali zgoraj, in jo v programu R dobimo s pomočjo ukaza `table`. Najprej si poglejmo primer pripravljenih podatkov. Kombinacijska tabela za spremenljivki vrsta kritja in tip naselja je prikazana spodaj.

```
            Podezelje Primestna Mestna
Osnovno          1100      3448   1020
Razsirjeno        526      1747    469
Premijsko         147       584     93
```

Definicija 2.1. Cramerjeva V-statistika [2, 110]. Cramerjeva V-statistika je mera za korelacijo med dvema skupinama kategoriziranih pojasnjevalnih spremenljivk, ki jo definiramo kot

$$\sqrt{\frac{\sum_{i,j} \frac{(n_{ij}-e_{ij})^2}{e_{ij}}}{n \cdot \min\{a-1, b-1\}}},$$

kjer je

- a enak številu kategorij spremenljivke ena,
- b enak številu kategorij spremenljivke dva,
- n_{ij} število enot v i-ti skupini spremenljivke 1 in j-ti skupini spremenljivke 2,
- $n = \sum_{ij}(n_{ij})$,
- $e_{ij} = \sum_i(n_{ij}) \sum_j(n_{ij})/n$.

V-statistika zavzame vrednosti med 0 in 1. Vrednost 0 pomeni, da poznavanje enega faktorja ne vpliva na poznavanje drugega faktorja. Obratno, vrednost 1 pomeni, da s poznavanjem enega faktorja poznamo tudi vrednost drugega faktorja.

Na teh podatkih nato uporabimo `cramersV`. Ker je statistika simetrična, se splača zgraditi matriko vrednosti.

Tabela 10. Cramerjeva V statistika.

	Izobrazba	*Zap. status*	*Poročni stan*	*Tip naselja*	*Spol*	*Vrsta kritja*	*Tip Vozila*	*Velikost vozila*
Izobrazba		0,080	0,077	0,126	0,014	0,027	0,022	0,022
Zap. status	0,080		0,290	0,305	0,024	0,015	0,026	0,020
Poročni stan	0,077	0,290		0,152	0,044	0,009	0,033	0,022
Tip naselja	0,126	0,305	0,152		0,087	0,041	0,046	0,126
Spol	0,014	0,024	0,044	0,087		0,018	0,030	0,023
Vrsta kritja	0,027	0,015	0,009	0,041	0,018		0,030	0,015
Tip vozila	0,021	0,026	0,033	0,046	0,030	0,030		0,034
Velikost vozila	0,022	0,020	0,022	0,126	0,022	0,015	0,034	

V tabeli 10 lahko vidimo, da je velika večina vrednosti zelo blizu 0, kar pomeni, da podatki med seboj verjetno niso preveč frekvenčno korelirani. Edina izjema je zaposlitveni status, ki ima malo višji vrednosti pri poročnem stanu in lokaciji, vendar pa so vrednosti Cramerjeve V-statistike še vedno dovolj nizke, da nas nebi smelo preveč skrbeti.

3. Linearna regresija

V statistiki si želimo raziskati povezave med različnimi spremenljivkami. Za raziskovanje odnosov med t.i. odvisnimi spremenljivkami in neodvisnimi oz. pojasnjevalnimi spremenljivkami lahko uporabimo različne regresijske modele. Regresijski modeli delujejo tako, da poskušajo z znanimi vrednostmi pojasnjevalnih spremenljivk napovedati vrednost odvisne spremenljivke. Če smo pri tem uspešni, lahko z modelom analiziramo zveze med spremenljivkami in napovemo vrednosti odvisnih spremenljivk pri tistih enotah, kjer ta zaenkrat še ni znana.

GLM model je nadgradnja linearne regresije, zato je smiselno, da začnemo z definicijo in lastnostmi le te. Linearna regresija je najbolj enostavna za razumevanje in zato najpogosteje uporabljena metoda za modeliranje odvisnosti odvisne spremenljivke od ene ali več neodvisnih spremenljivk. Predpostavlja normalno porazdelitev odvisne spremenljivke.

Preden lahko začnemo govoriti o regresijih modelih, moramo definirati oznake, ki jih bomo uporabljali pri modeliranju. Imejmo zaporedje k-razsežnih slučajnih vektorjev $(y_i, x_{2i}, x_{3i}, \ldots, x_{ki})$, kjer je $i = 1, 2, \ldots, n$ in je n velikost vzorca. y_i je odvisna slučajna spremenljivka na i-ti enoti, vse ostale pa skupaj s konstanto 1 združimo v k-razsežen slučajni vektor $x_i = (1, x_{2i}, x_{3i}, \ldots, x_{ki})'$ pojasnjevalnih spremenljivk. V primeru, da imamo poleg konstante le eno pojasnjevalno spremenljivko, modelu rečemo enostaven regresijski model. Če je teh spremenljivk več, se model imenuje multivariatni regresijski model [3, 1-4].

V splošnem ima regresijski model tri glavne elemente, ki ga določajo. To so:

(1) *Enačba regresijskega modela.* Enačba regresijskega modela nam pove, kako lahko modeliramo vrednost y_i s pomočjo vrednosti x_i. Vrednost y_i lahko modeliramo kot
$$y_i = f(x_i) + \epsilon_i,$$
kjer je $f(x_i)$ regresijska funkcija, ϵ_i pa šum, s katerimi modeliramo slučajno napako modela na vsaki enoti.

(2) *Vzorčne predpostavke.* Pri vzorčnih predpostavkah moramo paziti predvsem na to, da je naš vzorec dovolj velik in reprezentativen, torej da enakomerno predstavlja celotno populacijo oz. nabor podatkov.

(3) *Porazdelitvene predpostavke.* Te predpostavke določajo porazdelitve vseh slučajnih porazdelitev, ki nastopajo v modelu. V praksi se večinoma za vse predpostavi normalna porazdelitev.

Pri regresijskih modelih nas zanimata predvsem prva dva momenta pogojne porazdelitve, namesto celotne (pogojne) porazdelitvene funkcije. Z regresijo želimo torej poiskati čim bolj natančno oceno za pogojno upanje $\mathbb{E}(y_i|x_i)$ in pogojno varianco $\mathrm{Var}(y_i|x_i)$. V splošnem sta ta dva momenta lahko nelinearni funkciji x, vendar pa bomo v rezdelku linearne regresije obravnavali primer kjer je $\mathbb{E}(y_i|x_i)$ linearna funkcija x_i, $\mathrm{Var}(y_i|x_i)$ pa bo konstanta.

Linearna regresija je regresijski model z naslednjimi predpostavkami:

(1) Regresijska funkcija je linearna, torej
$$f(x_i) = x_i'\beta,$$

kjer je $\beta = (\beta_1, \beta_2, \ldots, \beta_k)'$ vektor regresijskih parametrov. Zaradi linearnosti se splača vpeljati še oznake $y = (y_1, y_2, \ldots, y_n)'$,

$$X = \begin{bmatrix} x_{11} & x_{21} & x_{31} & \ldots & x_{k1} \\ x_{12} & x_{22} & x_{32} & \ldots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & x_{3n} & \ldots & x_{kn} \end{bmatrix} = \begin{bmatrix} 1 & x_{21} & x_{31} & \ldots & x_{k1} \\ 1 & x_{22} & x_{32} & \ldots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{2n} & x_{3n} & \ldots & x_{kn} \end{bmatrix}$$

in $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$. Linearno regresijo lahko potem v matrični obliki zapišemo kot

$$y = X\beta + \epsilon. \tag{1}$$

(2) Matrika X je matrika konstant s polnim stolpčnim rangom. Torej v analizi pojasnjevalne spremenljivke niso več slučajne, ampak je celoten model izpeljan na konkretnih realizacijah teh spremenljivk.
(3) $\mathbb{E}[\epsilon_i] = 0$. Upanje napake je enako nič, kar pomeni, da nimamo sistematične napake, ki bi nas motila pri modeliranju.
(4) $\mathrm{Var}[\epsilon_i] = \mathbb{E}[\epsilon_i^2] = \sigma^2$. Zahtevamo, da se varianca napak ne spreminja, kar pomeni, da imamo izpolnjen pogoj homoskedastičnosti.
(5) $\mathbb{E}[\epsilon_i \epsilon_j] = 0$ za vse $i \neq j$. S tem pogojem zagotovimo, da so realizacije med seboj nekorelirane.
(6) Za normalen regresijski model predpostavimo še normalno porazdelitev napak. Velja

$$\epsilon_i \sim N(0, \sigma^2).$$

Tedaj so tudi odvisne spremenljivke y_i porazdeljene normalno

$$y_i \sim N(x_i'\beta, \sigma^2), \tag{2}$$

iz česar pa sledi, da je tudi y porazdeljen multivariatno normalno

$$y \sim N(X\beta, \sigma^2 I). \tag{3}$$

3.1. Metoda najmanjših kvadratov

Ocena za regresijske parametre β po metodi najmanjših kvadratov je tista vrednost β, ki minimizira vsoto kvadratov ostankov modela. Ostanki modela so razlike med ocenjeno vrednostjo $\hat{y} = X\hat{\beta}$ in dejansko vrednostji y. Naj bo

$$\begin{aligned} S(\beta) &= (y - X\beta)'(y - X\beta) \\ &= y'y - 2y'X\beta + \beta'X'X\beta. \end{aligned}$$

Sedaj odvajajmo $S(\beta)$ po β in odvod enačimo z 0.

$$\frac{\partial S}{\partial \beta} = -2X'y + 2X'X\beta = 0,$$

Ker ima matrika X poln stolpčni rang, je matrika $X'X$ obrnljiva in lahko izračunamo

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{4}$$

To je cenilka regresijskih parametrov β po metodi najmanjših kvadratov. Če enačbo (1) vstavimo v (4) dobimo

$$\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\epsilon.$$

Od tu vidimo, da je $\mathbb{E}(\hat{\beta}) = \beta$, torej je cenilka nepristranska. Iz enačbe zgoraj dobimo še variančno kovariančno matriko

$$\mathrm{Var}(\hat{\beta}) = \mathbb{E}((\hat{\beta} - \beta)(\hat{\beta} - \beta)') = \mathbb{E}((X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}) = \sigma^2(X'X)^{-1}.$$

Sedaj, ko imamo optimalni $\hat{\beta}$, lahko zapišemo

$$\hat{\epsilon} = y - X\hat{\beta}, \tag{5}$$

in to poimenujemo vektor ostankov metode najmanjših kvadratov. S pomočjo $\hat{\epsilon}$ lahko ocenimo σ^2 s

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n}. \tag{6}$$

Izkaže se, da je ta cenilka za σ^2 pristranska. Vpeljimo še nepristransko cenilko za σ^2.

$$s^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-k}.$$

Izraz v števcu je vsota kvadratov ostankov in jo bomo označili z $RSS = \hat{\epsilon}'\hat{\epsilon}$ [3, 4-5].

3.2. Metoda največjega verjetja

Če imamo model zapisan kot v enačbi (3), lahko zapišemo funkcijo verjetja kot

$$\ell(\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{\frac{(y-X\beta)'(y-X\beta)}{2\sigma^2}}.$$

Funkcijo verjetja ℓ logaritmiramo, hkrati pa opustimo dele enačbe, ki niso vezani na β ali σ^2. Dobljeno funkcijo označimo z $\log L$. Dobimo

$$\log L = -\frac{n}{2} log\sigma^2 - \frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}. \tag{7}$$

Očitno je maksimizacija $\log U(\beta)$ enaka minimizaciji $(y - X\beta)'(y - X\beta)$, kar pa pomeni, da je optimalna ocena za β enaka $\hat{\beta}$ enaka, kot smo jo izračunali pri metodi najmanjših kvadratov. Ugotovitev preverimo s parcialnim odvajanjem.

$$\frac{\partial logL}{\partial \beta} = \frac{1}{2\sigma^2}(-2X'y + 2X'X\beta) = 0.$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

Sedaj vstavimo to oceno $\hat{\beta}$ v (7). Izračunajmo odvod $logL$ po σ^2 in izenačimo z nič.

$$\log L = -\frac{n}{2}\log\sigma 2 - \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{2\sigma^2}.$$

Upoštevamo $y - X\hat{\beta} = \hat{\epsilon}$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\hat{\epsilon}'\hat{\epsilon}}{2\sigma^4} = 0,$$

$$n = \frac{\hat{\epsilon}'\hat{\epsilon}}{\sigma^2}.$$

Dobimo cenilko največjega verjetja σ^2

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n},$$

kjer je $\hat{\epsilon} = y - X\hat{\beta}$, kar je zopet enako kot pri metodi najmanjših kvadratov, le da je dobljena cenilka pristranska. V praksi vedno uporabimo nepristransko cenilko [3, 13-14].

3.3. Sklepanje na osnovi linearne regresije

Potem, ko smo izračunali ocene za parametre β in smo izračunali oceno za σ^2, lahko ocenimo še variančno kovariančno matriko $\hat{\beta}$, kot $\widehat{\mathrm{Var}}(\hat{\beta}) = s^2(X'X)^{-1}$. Z zbranimi informacijami lahko preverimo, katere spremenljivke imajo statistično značilen vpliv na odvisno spremenljivko.

V najenostavnejši obliki t-test testira hipotezo, ali je izbrani regresijski parameter enak 0. Če hipotezo zavrnemo, rečemo, da ustrezna pojasnjevalna spremenljivka statistično značilno vpliva na odvisno spremenljivko.

Z najenostavnejšim F-testom pa testiramo hipotezo, da je več regresijskih parametrov hkrati enakih 0. S pomočjo tega testa določimo, če je vsaj en izmed vseh obravnavanih parametrov ob določeni stopnji značilnosti statistično značilno različen od 0.

Za uporabo linearne regresije je predvsem pomembno, kako oba testa delujeta, zato se v verjetnostne podrobnosti, ki omogočajo testiranje, v delu ne bomo preveč spuščali. Opisa t-testa in F-testa najdemo v [4, 35-44].

3.3.1. t-test

S t-testom testiramo posamezne regresijske parametre. Vpliv i-te pojasnjevalne spremenljivke testiramo z naslednjo hipotezo:

$$H_0 : \beta_i = 0$$

proti alternativni hipotezi $H_1 : \beta_i \neq 0$.

Izrek 3.1. *V normalni linearni regresiji je pod pogojem ničelne hipoteze $H_0 : \beta_i = 0$ testna statistika*

$$t_j = \frac{\hat{\beta}_j}{\sqrt{s^2((X'X)^{-1})_{jj}}},$$

porazdeljena po $t(n-k)$ oz. po studentovi porazdelitvi z $(n-k)$ prostostnimi stopnjami.

Dokaz. Dokaz je na strani 36 vira [4] □

Izvedba t-testa pri stopnji značilnosti α je tako preprosta.

(1) S pomočjo formule zgoraj izračunamo t-statistiko.
(2) V tabeli s kritičnimi vrednostmi studentove porazdelitve pri pravi prostostni stopnji poiščemo vrednost $t_{\alpha/2}(n-k)$. Ta vrednost je vrednost pri kateri desno na grafu leži $\alpha/2$ vrednosti te porazdelitve.
(3) H_0 zavrnemo, če je $|t_j| \geq t_{\alpha/2}(n-k)$.

Računalnik nam ob implementaciji t-testa poleg te statistike vrne tudi njeno p-vrednost. Velja, da H_0 zavrnemo, če je $p < \alpha$.

S tem testom torej pri določeni stopnji zaupanja vidimo, če določena pojasnjevalna spremenljivka kaj prispeva k linearnemu modelu. Če ne, je njen prispevek blizu ničle in bi jo lahko izključili iz modela.

3.3.2. **F-test**

F-test za razliko od t-testa testira več parametrov hkrati. Ničelna hipoteza pravi, da je podmnožica parametrov β enaka nič, alternativna hipoteza pa pravi, da je vsaj eden izmed teh parametrov različen od nič. Če torej zavrnemo ničelno hipotezo, vemo, da vsaj ena pojasnjevalna spremenljivka nekaj pojasni. Pojasnjevalne spremenljivke lahko preuredimo tako, da testiramo skupen vpliv zadnjih k_2 spremenljivk. Vektor β tako razdelimo na prvih k_1 in zadnjih k_2 komponent. Velja $k_1 + k_2 = k$.

Skupen vpliv zadnjih k_2 pojasnjevalnih spremenljivk testiramo kot naslednjo hipotezo

$$H_0 : \beta_{k_1+1} = \beta_{k_1+2} = \ldots = \beta_k = 0$$

$$H_1 : \beta_j \neq 0 \text{ za vsaj en } j \in \{k_1 + 1, k_1 + 2, \ldots, k\}$$

Za implementacijo testa najprej ocenimo linearno regresijo s prvimi k_1 pojasnjevalnimi spremenljivkami in izračunamo pripadajočo vsoto kvadratov ostankov RSS_1, nato pa še celoten regresijski model in pripadajočo vsoto kvadratov ostankov RSS_2.

Izrek 3.2. *V normalni linearni regresiji je pod pogojem ničelne hipoteze*

$$H_0 : \beta_{k_1+1} = \beta_{k_1+2} = \ldots = \beta_k = 0$$

testna statistika

$$F = \frac{(RSS_1 - RSS_2)/k_2}{RSS_2/(n-k)},$$

porazdeljena po $F_{(k_2,n-k)}$ oziroma po Fisher-Snedecorjevi porazdelitvi z k_2 in $(n-k)$ prostostnimi stopnjami.

Dokaz. Dokaz je na strani 41 vira [4] □

Izvedba F-testa pri stopnji značilnosti α je tako preprosta.

(1) S pomočjo formule zgoraj izračunamo F-statistiko.
(2) V tabeli s kritičnimi vrednostmi Fisher-Snedecorjevi porazdelitve pri pravih prostostnih stopnjah poiščemo vrednost $F_\alpha k_2, n-k$. Ta vrednost je vrednost, pri kateri desno na grafu leži α vrednosti te porazdelitve.
(3) H_0 zavrnemo, če je $F \geq F_\alpha k_2, n-k$.

Računalnik nam ob implementaciji F-testa poleg te statistike vrne tudi njeno p-vrednost. Velja, da H_0 zavrnemo, če je $p < \alpha$.

3.4. Obravnava kategoričnih pojasnjevalnih spremenljivk

V prejšnjem razdelku smo opisali ocenjevanje in statistično sklepanje v linearni regresiji. Med predpostavkami modela smo navedli, da je matrika X matrika števil s polnim stolpičnim rangom. Spremenljivke, s katerimi bomo v nadaljevanju modelirali škodne zahtevke, so kategorične in jih v takšni obliki ne moremo vključiti v matriko X. Problem rešimo z uvedbo slamnatih spremenljivk. Pri vsaki kategorični spremenljivki izberemo osnovno skupino in nato v model vpeljemo indikatorske spremenljivke ostalih skupin.

Oglejmo si postopek v primeru, ko želimo v model vključiti pojasnjevalno spremenljivko izobrazba. Ta spremenljivka ima ravni srednja šola ali manj, višja šola, diploma in podiplomska. Pripadajoča linearna regresija ima obliko

$$y_i = \beta_1 + \beta_2 1_{\text{višja šola},i} + \beta_3 1_{\text{diploma},i} + \beta_4 1_{\text{podiplomska},i} + \epsilon_i$$

Parameter β_1 ustreza ravni izobrazbe osnovne skupine, ostale β_j pa primerjajo ostale ravni z osnovno skupino.

Statističnega vpliva kategorične pojasnjevalne spremenljivke nikoli ne testiramo s testi posameznih parametrov β_j s t-testi, ampak vselej testiramo vse parametre ob indikatorskih spremenljivkah hkrati z F-testom. Testu hipoteze $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ proti alternativi, da je vsaj en od teh parametrov različen od 0, rečemo enosmerna analiza variance. Enosmerno analizo variance škodnih zahtevkov po posameznih kategoričnih spremenljivkah smo že naredili v poglavju z opisom podatkov.

3.5. Linearna regresija na primeru

Sedaj lahko uporabimo linearno regresijo tudi na našem primeru podatkov.

V programu R lahko linearno regresijo delamo s pomočjo ukaza `lm`, ki nam nato izračuna ocene za posamezne parametre, ter t in F statistike. Pojasnjevalne spremenljivke, ki jih uporabljamo, so kategorične, torej bomo uporabili linearno regresijo s pomočjo slamnatih spremenljivk. Tako bo prva skupina osnova, drugi faktorji pa bodo zavzeli vrednosti 0 in 1. Poglejmo kaj vrne model.

```
Residuals:
    Min      1Q  Median      3Q     Max
-712.02  -78.30  -18.60   64.67 1719.94
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
   (Intercept)             84.4220     7.7826  10.847  < 2e-16
Izobrazba (SŠ ali nižja)
   Višja sola             -17.8094     3.9649  -4.492 7.15e-06
   Diploma                -11.0585     3.9469  -2.802 0.005092
   Podiplomska            -20.0742     5.2946  -3.791 0.000151
Tip Vozila (Dvovratno vozilo)
   Štirivratno vozilo      -0.8254     3.9316  -0.210 0.833715
   Športno vozilo         203.5401     4.4820  45.412  < 2e-16
   Luksuzno vozilo        695.1513     8.4193  82.566  < 2e-16
Status Zaposlitve (Nezaposlen)
   Delavno neaktiven      -70.1608     5.4574 -12.856  < 2e-16
   Zaposlen               -84.7014     4.1219 -20.549  < 2e-16
```

```
Poročni Stan (Poročen)
   Ločen                    4.0458      4.3852   0.923 0.356246
   Samski                  67.4969      3.8327  17.611  < 2e-16
Tip Naselja (Podeželje)
   Primestna              379.6370      4.3059  88.168  < 2e-16
   Mestna                 220.3154      4.9751  44.284  < 2e-16
Spol (Ženski)
   M                       12.4041      3.0256   4.100 4.17e-05
Kritje (Osnovno)
   Razširjeno              92.6248     3.3568  27.593  < 2e-16
   Premijsko              238.7048     5.3815  44.356  < 2e-16
Velikost.Vozila (Majhen)
   Medium                 -9.7243      3.9426  -2.466 0.013663
   Velik                  -5.1165      5.8307  -0.878 0.380235


Residual standard error: 143.7 on 9116 degrees of freedom
```

Da si lažje predstavljamo, lahko naredimo primer. Prvi koeficient nam pove, kakšna je ocenjena višina škode, če imamo pri vseh pojasnjevalnih spremenljivkah referenčne ravni. Model višino škode na polici, kjer je zavarovalčeva stopnja izobrazbe srednješolska ali nižja, vozilo dvovratno, zavarovalec nezaposlen in poročen, živi na podeželju, je ženska, ima osnovno kritje ter vozi majhen avto, ocenjuje na 84,422.

Osnova za spol je ženska. Če imamo torej polico, kjer so vse druge kategorije enake, spol pa je moški, bo povprečna škoda na tej polici za 12,40 višja od prve police na kateri je zavarovalka ženska. Podobno lahko iz primera za zavarovalčevo izobrazbo razberemo, da je ocenjena škoda v primerjavi z zavarovalcem, ki je srednješolsko ali slabše izobražen, nižja za 17,8, če je le ta višješolsko izobražen, 11,05 če ima diplomo, ter 20,07 če ima podiplomsko izobrazbo.

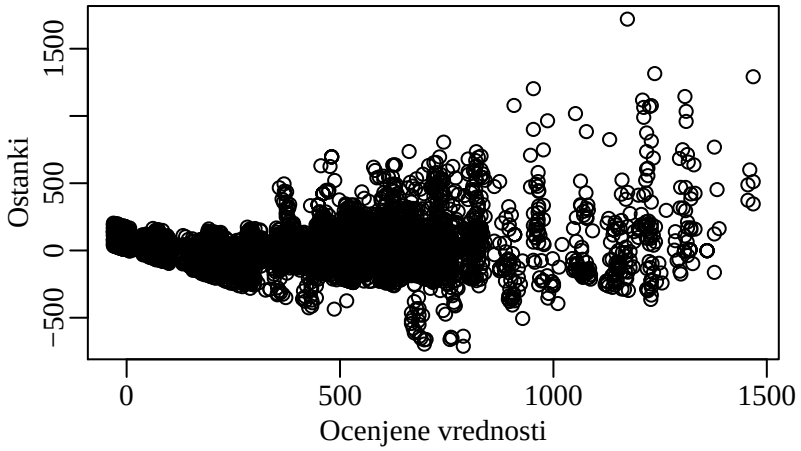
Zanima nas tudi kvaliteta modela. V ta namen lahko narišemo nekaj grafov, ki nam pomagajo pri oceni našega modela.

Na sliki 10 smo na x-os narisali ocenjene vrednosti, na y-os pa ostanke pri modelu linearne regresije. Slika, ki bi pokazala primeren model, bi imela ostanke enakomerno razporejene okoli nič za vse vrednosti. Na sliki 10 pa lahko vidimo, da se graf širi, kar nakazuje, da naš model ni najboljši. Očitno je kršena predpostavka o homoskedastičnosti. Opazimo tudi, da so pri zelo nizkih ocenjenih vrednostih ostanki tipično pozitivni, pri ocenjeni vrednosti okoli 250 pa bolj negativni. Takšna sistematičnost v ostankih kaže na to, da zveza med višino škode in pojasnjevalnimi spremenljivkami ni linearna.

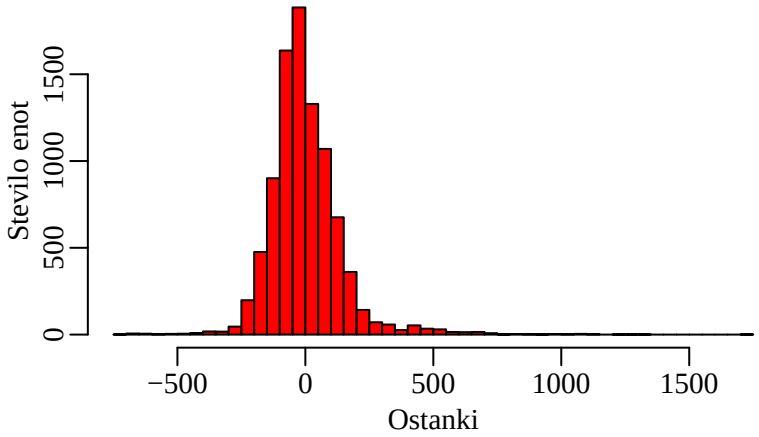
Za diagnozo modela lahko narišemo tudi histogram ostankov.

Tudi slika 11 bi v primeru primernosti modela nakazovala normalno porazdelitev. Na prvi pogled lahko sicer opazimo, da imamo zvončasto obliko, ki je značilna za normalno porazdelitev. Če pogledamo bolj natančno, lahko na desni strani opazimo debelejši rep. To pomeni, da naš model s svojo oceno večkrat močno podcenjuje višino škode. Pri modeliranju višine škode je to lahko zelo nevarno.

Tretji graf, ki ga bomo narisali, je t.i. Q-Q-graf, ki testira normalnost modela.



Slika 10. Ocenjene vrednosti in ostanki pri linearni regresiji.

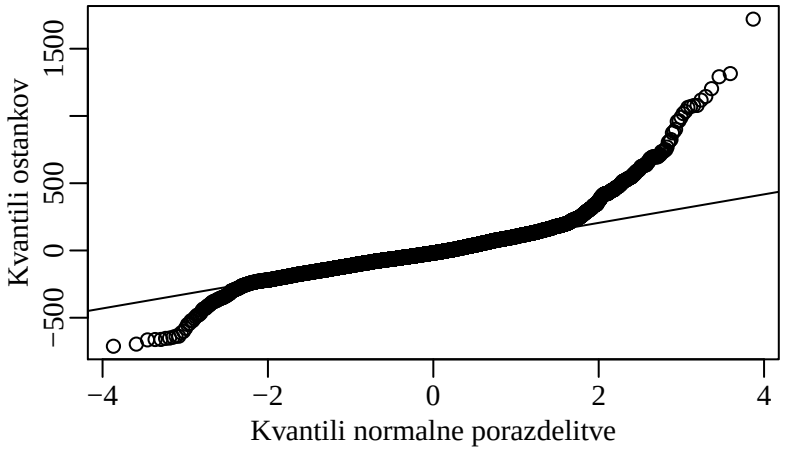


Slika 11. Histogram ostankov pri linearni regresiji.

Če bi bili ostanki modela porazdeljeni približno normalno, bi večina vrednosti ležala na premici, ki jo vidimo na sliki 12. Podobno kot na sliki 11 lahko opazimo večje odstopanje pri višjih vrednostih, kar pomeni, da slabo ocenjujemo visoke škode.

Linearna regresija za modeliranje višin škod na našem primeru ni najbolj primerna oblika modela.

Kot smo opisali v razdelku, kjer smo opisali obravnavo kategoričnih pojasnjevalnih spremenljivk, lahko naredimo še F-test za vsako pojasnjevalno spremenljivko posebej. S pomočjo tega testa bi se lahko za kakšno izmed pojasnjevalnih spremenljivk odločili za odstranitev iz modela, saj pri naši stopnji značilnosti ne spremeni ocene za višino škode.



SLIKA 12. Q-Q grafikon linearne regresije.

V programu R bomo za vsako pojasnjevalno spremenljivko naredili model, ki te spremenljivke ne vključuje. Nato bomo s pomočjo ukaza `ANOVA()` izvedli F-test, ki bo izračunal, ali ima spremenljivka, ki smo jo izpustili, sistematičen doprinos k oceni, ki jo vrne model. Ukaz

```
anova(lin_mod, lin_mod_iz),
```

kjer je `lin_mod` linearna regresija, ki smo jo naredili zgoraj, `lin_mod_iz` pa linearna regresija za višino škode, kjer smo izpustili pojasnjevalno spremenljivko izobrazba. Dobimo naslednji rezultat

```
Analysis of Variance Table

Model 1: Skoda ~ Izobrazba + Tip.Vozila + Status.Zaposlitve
    + Porocni.Stan + Tip.Naselja + Spol + Kritje + Velikost.Vozila
Model 2: Skoda ~ Tip.Vozila + Status.Zaposlitve
    + Porocni.Stan + Tip.Naselja +
    Spol + Kritje + Velikost.Vozila
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1   9116 188196820
2   9119 188713678 -3   -516858 8.3453 1.542e-05
---
```

Ker je p-vrednost za ta test manjša od 5%, to pomeni, da lahko ničelno hipotezo, da nobena kategorija pojasnjevalne spremenljivke izobrazba k oceni modela za višino škode ne prispeva ničesar, zavrnemo. Naredimo ta test za vse pojasnjevalne spremenljivke.

V tabeli 11 lahko preberemo, da večina spremenljivk sistematično prispeva k oceni za višino škode. Na meji je vrednost pri pojasnjevalni spremenljivki za velikost vozila.

TABELA 11. F-test za posamezne pojasnjevalne spremenljivke pri linearni regresiji.

Pojasnjevalna spremenljivka	p-vrednost
Kritje	$2{,}2 \cdot 10^{-16}$
Izobrazba	$1{,}5 \cdot 10^{-5}$
Zaposlitveni status	$2{,}2 \cdot 10^{-16}$
Spol	$4{,}1 \cdot 10^{-5}$
Tip naselja	$2{,}2 \cdot 10^{-16}$
Poročni stan	$2{,}2 \cdot 10^{-16}$
Tip vozila	$2{,}2 \cdot 10^{-16}$
Velikost vozila	0,04207

3.6. **Prednosti in slabosti linearne regresije**

Linearna regresija ima veliko prednosti, vendar pa ima tudi nekaj slabosti, ki pa so lahko zelo usodne. Zaradi pomankljivosti linearne regresije je obstajala potreba, da se najde regresijski model, ki bo presegel slabosti linearne regresije in bil bolj splošen. Ena izmed alternativ je tudi GLM. Poglejmo si glavne prednosti in slabosti linearne regresije.

3.6.1. **Prednosti**

(1) Odkril jo je Gauss že daljnega leta 1795 in je zgodovinsko prva metoda za modeliranje odvisnosti med slučajnimi spremenljivkami.

(2) Je analitično rešljiv problem.

(3) Matematična analiza linearne regresije je enostavna in lahko razumljiva.

(4) Tudi za ljudi, ki se ne ukvarjajo izključno z matematiko, je linearna regresija zelo intuitivna in lahko predstavljiva, njene rešitve pa si lahko preprosto interpretiramo.

(5) Pri določenih problemih, kjer je pojasnjevalna funkcija res linearna, in so napake porazdeljene normalno z vedno enako varianco, je linearna regresija resnično najboljša metoda za napovedovanje podatkov.

3.6.2. **Slabosti**

(1) Osamelci. Ti imajo zelo velik vpliv na rezultate linearne regresije. Ker linearna regresija minimizira vsoto kvadratov ostankov, že en sam osamelec k vrednosti napak prispeva zelo veliko. Ocena parametrov se lahko zaradi enega samega osamelca zelo spremeni.

(2) Nelinearnost modelov. Že ime linearna regresija kaže na to, da bi lahko model imel velike težave pri ocenjevanju nelinearnih sistemov. Veliko problemov, ki se pojavljajo v praksi, ni linearnih in jih bo zato linearna regresija zelo slabo modelirala, če nismo ustrezno obravnavali nelinearnosti.

(3) Odvisnost med spremenljivkami. Z linearno regresijo dobimo slabe rezultate, če so pojasnjevalne spremenljivke med seboj korelirane. V primeru dveh popolnoma koreliranih pojasnjevalnih spremenljivk množica rešitev ni enolična.

(4) Heteroskedastičnost. Linearna regresija slabo modelira sisteme, kjer imamo heteroskedastičnost, torej podatke, pri katerih se varianca spreminja skupaj z vrednostmi pojasnjevalnih spremenljivk.

(5) Predpostavka normalnosti. Pri linearni regresiji smo predpostavili, da je odvisna spremenljivka normalno porazdeljena. Če opazimo, da temu ni tako, sta F-test in t-test neveljavna.

4. Posplošeni linearni model

Leta 1972 sta J. A. Nelder in R. W. M. Wedderburn objavila članek [6], kjer sta opisala razširitev linearne regresije in jo poimenovala GLM oz. posplošeni linearni model. Družina GLM modelov obsega širok nabor različnih regresijskih modelov. Zbirka različnih modelov v posebnem primeru vsebuje tudi linearne regresijske modele. Omejujoče predpostavke normalne porazdelitve, konstantne variance in aditivnosti linearnega modela se pri GLM sprostijo. Edina predpostavka GLM-a je pogoj, da porazdelitev odvisne spremenljivke spada v eksponentno družino porazdelitev, ki jo bomo podrobneje spoznali v nadaljevanju. Ta razdelek je večinoma povzet po [2].

GLM torej lahko predstavlja zelo širok nabor modelov za modeliranje podatkov. V splošnem lahko vse te modele združimo. GLM ima naslednje predpostavke. Večino oznak se je ohranilo iz razdelka, kjer smo govorili o linearni regresiji.

(1) Odvisne spremenljivke y_i, $i = 1, 2, \ldots, n$, so med seboj neodvisne, porazdeljene pa so po eni izmed porazdelitev, ki pripada eksponentni družini porazdelitev. Parametri porazdelitev se lahko med spremenljivkami razlikujejo. Naj bo
$$\mathbb{E}(y_i) = \mu_i,$$
in $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$.

(2) Z vektorjem pojasnjevalnih spremenljivk x_i dobimo linearno napovedovalno funkcijo η_i, ki napove vrednost sistematične komponente GLM-a μ_i. Ta predstavlja rezultat linearne napovedovalne funkcije
$$\eta_i = x_i'\beta.$$
Vse skupaj lahko zapišemo v matrični obliki. Označimo $\eta = (\eta_1, \eta_2, \ldots, \eta_n)'$. Linearna napovedovalna funkcija izgleda podobno kot pri linearni regresiji.
$$\eta = X\beta.$$

(3) Povezavo med sistematično komponento μ_i iz točke (1) in linearno napovedovalno funkcijo η_i iz točke (2) podamo s povezovalno funkcijo g. Za funkcijo g zahtevamo odvedljivost in monotonost na celotnem definicijskem območju. Velja:
$$\eta_i = g(\mu_i)$$

oziroma

$$\mathbb{E}[y_i] = \mu_i = g^{-1}(\eta_i) = g^{(-1)}x_i'\beta.$$

Zapišimo še v matrični obliki

$$\mu = g^{-1}(\eta) = g^{-1}(X\beta).$$

Tu smo funkcijo g^{-1} uporabili po komponentah.

(4) *Vzorčne predpostavke.* Pri vzorčnih predpostavkah moramo paziti predvsem na to, da je naš vzorec dovolj velik in reprezentativen, torej da enakomerno predstavlja celotno populacijo oz. nabor podatkov.

4.1. Eksponentna družina porazdelitev

Na tem mestu je smiselno nekaj povedati o eksponentni družini porazdelitev, saj so te porazdelitve za GLM zelo pomembne. Eksponentno družino porazdelitev lahko opišemo z eno samo enačbo za gostoto ali funkcijo verjetnosti z dvema parametroma, ki sta za posamezne porazdelitve različna.

V eksponentno družino porazdelitev lahko uvrstimo nekaj zelo pomembnih porazdelitev. V to družino med drugim spadajo: normalna porazdelitev, gama porazdelitev, eksponentna porazdelitev, beta porazdelitev, Poissonova porazdelitev, Bernoullijeva porazdelitev, binomska porazdelitev in druge. Za naš problem sta še posebej pomembni normalna in gama porazdelitev. Pri zveznih slučajnih spremenljivkah je definicijsko območje gostote verjetnostni interval, pri diskretnih pa funkcijo verjetnosti opazujemo le v številno mnogo točkah.

Naj bo Y slučajna spremenljivka porazdeljena po eni izmed zveznih porazdelitev iz eksponentne družine porazdelitev. Gostoto verjetnosti za slučajno spremenljivko Y lahko zapišemo kot

$$f_Y(y;\theta,\phi) = e^{\frac{y\theta - B(\theta)}{\phi} + C(y,\phi)}, \tag{8}$$

kjer sta $B(\cdot)$ in $C(\cdot,\cdot)$ znani funkciji. Pri parametrizaciji, ki smo jo navedli zgoraj, θ predstavlja kanonični parameter, ϕ pa parameter razpršenosti. Družino porazdelitev lahko parametriziramo tudi s pomočjo upanja $\mathbb{E}[Y] = \mu$, tako da velja $\theta = h(\mu)$, za neko funkcijo h. Funkcijo h v tem primeru poimenujemo kanonična povezovalna funkcija.

Za slučajno spremenljivko Y, ki smo jo opisali zgoraj, veljata dve zelo zanimivi lastnosti. Najprej velja

$$\mathbb{E}[Y] = \mu = B'(\theta).$$

Upanje slučajne spremenljivke tako dobimo z odvajanjem funkcije B po θ. Nato pa še

$$\mathrm{Var}[Y] = B''(\theta)\phi = V(\mu)\phi. \tag{9}$$

Varianco slučajne spremenljivke Y tako dobimo s pomočjo drugega odvoda funkcije B po θ. Funkcijo $V(\mu)$ pa imenujemo funkcija variance. Za porazdelitve iz eksponentne družine porazdelitev je varianca lahko odvisna od upanja.

V tabeli 12 smo navedli funkcije variance za nekaj porazdelitev, ki spadajo v eksponentno družino porazdelitev in se pogosto uporabljajo pri GLM modeliranju.

TABELA 12. Funkcije variance za različne porazdelitve.

Porazdelitev	$V(x)$
Normalna porazdelitev	1
Poissonova porazdelitev	x
Gama porazdelitev	x^2
Inverzna Gaussova porazdelitev	x^3

Iz tabele 12 prav tako vidimo, da pri normalni porazdelitvi varianca ni odvisna od upanja, kar smo že spoznali kot slabost linearne regresije.

Iz enačbe 9 vidimo, da poleg funkcije variance na varianco vsake realizacije vpliva še parameter ϕ, ki varianco ustrezno skalira. Pri GLM modelu bi lahko k funkciji variance dodali še uteži, s katerimi bi modelirali pomembnost posameznih opažanj oz. enot. V primeru naših podatkov tega ne bomo potrebovali in bomo zato uteži izpustili.

4.1.1. **Normalna porazdelitev**

Pri linearni regresiji je normalna porazdelitev zelo pomembna. Poglejmo, kako izgleda normalna porazdelitev, če jo parametriziramo na način opisan v enačbi (8). Naj bo Y slučajna spremenljivka porazdeljena normalno $Y \sim N(\mu, \sigma^2)$. Gostoto porazdelitve lahko zapišemo

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}},$$
$$f_Y(y) = e^{-\frac{1}{2}\log(2\pi\sigma^2)} e^{-\frac{(y^2-2y\mu+\mu^2)}{2\sigma^2}},$$
$$f_Y(y) = e^{\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{1}{2}(y^2/\sigma^2 + \log(2\pi\sigma^2)}.$$

Iz zgoraj zapisane enačbe lahko preberemo kanonični parameter $\theta = \mu$ in $B(\mu) = \mu^2/2$, parameter razpršenosti $\psi = \sigma^2$ in $C(y, \sigma) = -\frac{1}{2}(y^2/\sigma^2 + \log(2\pi\sigma^2)$. Ker je $\theta = \mu$ kanonična povezovalna funkcija identiteta $h(x) = x$. Vidimo, da je $B'(\mu) = \mu$, kar je res upanje spremenljivke Y. če izračunamo še drugi odvod $B''(\mu)\sigma^2 = \sigma^2$, dobimo varianco $\mathrm{Var}[Y] = \sigma^2$.

4.1.2. **Gama porazdelitev**

Gama porazdelitev je ime za družino zveznih verjetnostnih porazdelitev, ki prav tako spadajo v eksponentno družino porazdelitev. Družino gama porazdelitev parametriziramo z dvema parametroma. Navadno jo parametriziramo s pomočjo parametra za obliko $a > 0$ in parametra za obseg $\rho > 0$. Naj bo Y gama porazdeljena slučajna spremenljivka s parametroma $a > 0$ in $\rho > 0$. Gostota verjetnosti slučajne spremenljivke je

$$f_Y(y) = \frac{1}{\Gamma(a)\rho^a} y^{a-1} e^{-\frac{y}{\rho}} \text{ za } y > 0.$$

Velja:

- $\mathbb{E}[Y] = a\rho$,
- $\mathrm{Var}[Y] = a\rho^2$.

Parametra lahko spremenimo tako, da en predstavlja upanje porazdelitve, drugi pa meri razpršenost, kar nam bo pomagalo v nadaljevanju magistrskega dela. Naj bo

Y gama porazdeljena slučajna spremenljivka z upanjem $\mu = a\rho > 0$, in parametrom razpršenosti $\phi = \frac{1}{a} > 0$.

Gostota verjetnosti slučajne spremenljivke je

$$f_Y(y) = \frac{1}{\Gamma(\frac{1}{\phi})(\mu\phi)^{\frac{1}{\phi}}} y^{(\frac{1}{\phi}-1)} e^{\frac{y(-\frac{1}{\mu})}{\phi}} \text{ za } y > 0.$$

Preuredimo zgornjo enačbo v obliko, ki smo jo napisali v enačbi (8).

$$f_Y(y) = e^{\frac{y(-\frac{1}{\mu})-\log(\mu)}{\phi} - \log(\phi)/\phi + (1/\phi - 1)\log y - \log\Gamma(\frac{1}{\phi})}.$$

Iz enačbe zgoraj lahko razberemo kanonični parameter $\theta = -\frac{1}{\mu}$, torej je kanonična povezovalna funkcija $h(x) = -\frac{1}{x}$. Nadalje opazimo $B(\theta) = \log(-\frac{1}{\theta})$. Odvajajmo

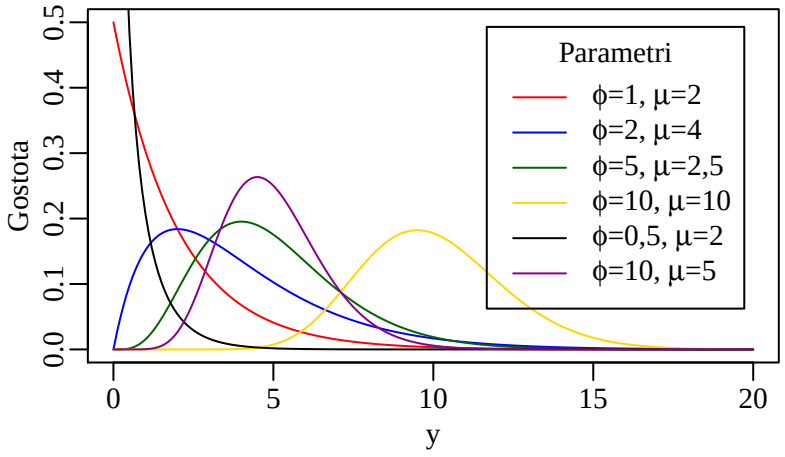
$$B'(\theta) = -\frac{1}{-1/\theta}(-1/\theta^2) = -\frac{1}{\theta} = \mu.$$

Vidimo, da lastnost, s katero lahko izračunamo upanje porazdelitev iz eksponentne družine, res drži. Poglejmo še drugi odvod.

$$B''(\theta) = (-\frac{1}{\theta})' = \frac{1}{\theta^2} = \mu^2,$$

kar pomeni, da je $\mathrm{Var}[Y] = \mu^2\phi$, funkcija variance pa je $V(\mu) = \mu^2$.

Na sliki 13 lahko vidimo, kako izgledajo gostote verjetnosti gama porazdelitev pri različnih parametrih μ in ϕ. Vidimo, da se lahko dokaj približamo obliki porazdelitve škod, ki smo jo narisali na sliki 1. Porazdelitev gama je nesimetrična, levi rep je navzdol omejen z 0, desni rep pa je navzgor neomejen.



SLIKA 13. Primerjava gama porazdelitev z različnimi parametri.

4.2. **Povezovalna funkcija**

Pri linearni regresiji je zelo pomembna linearnost upanja v pojasnjevalnih spremenljivkah, kar nam zmanjša nabor podatkov, za katere lahko linearno regresijo uporabimo brez skrbi. Povezovalna funkcija v linearni regresiji je identiteta. Pri GLM-u te zahteve ni. Za potrebe našega modela lahko namesto zahteve, da je upanje odvisne spremenljivke y_i linearna kombinacija pojasnjevalnih spremenljivk x_i, postavimo zahtevo, da je neka transformacija upanja odvisne spremenljivke y_i, ki ga zapišemo kot $g(\mu_i)$, linearna kombinacija pojasnjevalnih spremenljivk x_i. Torej $\mu_i = \mathbb{E}[y_i]$ gledamo kot funkcijo linearnega prediktorja η_i, ki nam jo vrne linearni model. Zanima nas torej inverz povezovalne funkcije $g(x)$

$$\mu_i = g^{-1}(\eta_i).$$

Teoretično bi za vsako realizacijo lahko uporabili različno povezovalno funkcijo. Za to funkcijo imamo le dva pogoja. Mora biti odvedljiva in monotona. V tabeli 13 so zapisane povezovalne funkcije, ki se pogosteje uporabljajo v praksi. [2, 20]

TABELA 13. Pogoste povezovalne funkcije.

Ime	$g(x)$	$g^{-1}(x)$
Identiteta	x	x
Logaritem	$\log(x)$	e^x
Logit funkcija	$\log(x/(1-x))$	$e^x/(1+e^x)$

Vsaka porazdelitev iz eksponentne družine ima svojo naravno pripadajočo povezovalno funkcijo, vendar pa te kombinacije za uporabo niso obvezne. Pri gama porazdelitvi je naravna izbira recipročna povezovalna funkcija. Ker dopušča negativne vrednosti, je v praksi redko uporabljena. GLM model omogoča, da lahko poljubno izbiramo med vsemi porazdelitvami in povezovalnimi funkcijami, če le te zadostujejo pogojem in lastnostim podatkov. Tako imamo zelo širok nabor izbir za regresijske modele, s katerimi kar najbolje možno aproksimiramo podatke, ki jih opazujemo.

V tabeli 14 si lahko ogledamo pogoste kombinacije porazdelitev in povezovalnih funkcij za podatke, ki jih v zavarovalništvu pogosto modeliramo.

Opomba: Izpostavljenost je zavarovalniška mera, ki meri časovno izpostavljenost police v določenem obdobju. Če imamo polico, ki traja od 1.1.2015 do 1.1.2016, in gledamo njeno izpostavljenost v času od 1.1.2015 do 1.7.2016, je ta izpostavljenost 1/2. Če bi gledali njeno izpostavljenost v obdobju veljavnosti je izpostavljenost enaka 1.

Izbira povezovalne funkcije vpliva na interpretacijo parametrov β. Zanimiv primer je logaritemska povezovalna funkcija. Poglejmo, kaj se zgodi, ko jo uporabimo na modelu.

$$\mu_i = g^{-1}(x_i'\beta) = g^{-1}(\beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}) = e^{\beta_1} \cdot \ldots \cdot e^{\beta_k x_{ki}},$$

kar pomeni, da smo namesto klasične aditivne povezave med pojasnjevalnimi spremenljivkami uporabili multiplikativno povezavo.

TABELA 14. Pogoste kombinacije funkcij variance in povezovalnih funkcij za modeliranje pogostih zavarovalniških podatkov.

	Frekvenca škod	Število škod	Višina škode	Verjetnost obnove
Povezovalna funkcija $g(x)$	$\ln(x)$	$\ln(x)$	$\ln(x)$	$\ln(x/(1-x))$
Porazdelitev	Poissonova	Poissonova	Gama	Bernoullijeva
Parameter razpršenosti ϕ	1	1	ocenjen	1
Funkcija variance $V(x)$	x	x	x^2	$x(1-x)$
Predhodne uteži	Izpostavljenost	1	Število škod	1

4.3. Metoda največjega verjetja

Ocene za parametre β zopet izračunamo s pomočjo metode največjega verjetja. Ta metoda nam vrne ocene za parametre, pri katerih so opazovani podatki najbolj verjetni. Funkcija verjetja je sestavljena tako, da zmnožimo gostote verjetnosti odvisne spremenljivke pri vsaki realizaciji. Pogosto to funkcijo logaritmiramo in nato maksimiziramo vsoto namesto produkta. Iščemo torej parametre β, ki maksimizirajo logaritemsko funkcijo verjetja.

Pri enostavnih primerih porazdelitvene in povezovalne funkcije lahko te parametre najdemo analitično, kot smo pokazali v razdelku linearne regresije, ki je poseben primer GLM-a. V splošnem to ne gre. Če gostote verjetnosti iz eksponentne družine pomnožimo za vsako opazovanje i, vse skupaj še logaritmiramo in razčlenimo, dobimo logaritemsko funkcijo verjetja

$$L = \sum_i \left(\frac{y_i\theta_i - B(\theta_i)}{\phi} + C(y_i, \phi) \right).$$

To funkcijo maksimiziramo tako, da za vsak $j; j = 1, \ldots, k$, izračunamo parcialni odvod po β_j in ga izenačimo z 0.

$$\frac{\partial L}{\partial \beta_j} = 0; j = 1, \ldots, k.$$

Takšno računanje lahko hitro postane zelo zahtevno, zato raje verižno odvajajmo.

$$0 = \frac{\partial L}{\partial \beta_j} = \sum_i \frac{\partial}{\partial \theta_i} \left(\frac{y_i\theta_i - B(\theta_i)}{\phi} + C(y_i, \phi) \right) \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

Na tem mestu se moramo spomniti relacij med funkcijami:

$$\mu_i = B'(\theta_i) \Rightarrow \frac{\partial \mu_i}{\partial \theta_i} = B''(\theta_i) \Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{B''(\theta_i)},$$

$$\eta_i = g(\mu_i) \Rightarrow \frac{\partial \eta_i}{\partial \mu_i} = g'(\mu_i) \Rightarrow \frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)},$$

$$\eta_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} \Rightarrow \frac{\partial \eta_i}{\partial \beta_j} = x_{ji} \text{ (ali 1, če } j = 1)$$

Vse to vstavimo v prvotni odvod

$$\frac{\partial L}{\partial \beta_j} = \sum_i \frac{(y_i - \mu_i)}{\phi} \frac{1}{B''(\theta_i)} \frac{1}{g'(\mu_i)} x_{ji}$$
$$= \sum_i \frac{(y_i - \mu_i) x_{ji}}{\phi V(\mu_i) g'(\mu_i)} \; ; \; j = 1, \ldots, k.$$

Zgornje odvode moramo enačiti z 0 za vsak$j$. Tako dobimo sistem enačb, če bi hoteli z metodo največjega verjetja dobiti ocene za parametre β. Kljub temu da je sistem enačb, ki jih moramo rešiti za maksimizacijo funkcije verjetja, teoretično lahko zapisati, je poiskati rešitev teh enačb veliko bolj zapleteno. V praksi je zaradi velikega števila podatkov in parametrov pogosto nemogoče dobiti analitično rešitev. Namesto tega uporabimo numerične rešitve [2, 99-100].

Iščemo torej vrednosti parametrov β, ki so rešitve sistema enačb, ki ga dobimo z enačenjem parcialnih odvodov logaritemske funkcije verjetja z 0. V ta namen se pogosto uporabi t.i. Newton-Raphson iteracija, ki je dana s formulo

$$\beta_{(n+1)} = \beta_{(n)} - H_{(n)}^{-1} s_{(n)},$$

kjer je $\beta_{(n)}$ n-ta iteracija za oceno parametrov β, $s_{(n)}$ je gradient, torej vektor prvih parcialnih odvodov logaritemske funkcije verjetja, $H_{(n)}$ pa je Hessejeva matrika, torej matrika velikost $k \times k$, ki vsebuje druge parcialne odvode logaritemske funkcije verjetja. Gradient in Hessejevo matriko izvrednostimo v $\beta_{(n)}$.

Proces lahko začnemo s poljubnimi ocenami za β in končamo, ko sta zaporedni vrednosti $\beta_{(n+1)}$ in $\beta_{(n)}$ dovolj blizu skupaj. Dobre začetne ocene pospešijo iskanje optimalne rešitve in povišajo njihovo zanesljivost v zelo zapletenih modelih.

Alternativa Newton-Raphsonovi iteraciji je Fisherjeva iteracija. Ta namesto opazovane Hessejeve matrike $H_{(n)}$ uporabi njeno pričakovano vrednost $J_{(n)}$ prav tako izvrednoteno v $\beta_{(n)}$. Če v GLM uporabimo kanonično povezovalno funkcijo, sta matriki enaki.

Ker je ta algoritem že dobro implementiran v različna orodja za statistiko (primeri v tej nalogi so izdelani s programskim jezikom R, imajo pa ga tudi drugi programi za statistično analizo podatkov), se v podrobnosti algoritma v magistrskem delu ne bomo spuščali.

Podrobneje si oglejmo še, kako ocenimo parameter razpršenosti ϕ. V splošnem primeru parameter ϕ ni znan vnaprej, zato ga moramo oceniti iz podatkov. Lahko bi ga obravnavali kot še enega izmed parametrov ter ga poskušali oceniti z metodo največjega verjetja, vendar pa je ta način običajno zelo zamuden in zapleten. Namesto tega lahko ta parameter ocenimo s pomočjo cenilke za drugi moment

$$\hat{\phi} = \frac{1}{n-k} \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

Za izračun $\hat{\phi}$ moramo najprej poznati $\hat{\beta}$, da lahko izračunamo ocene

$$\hat{\eta}_i = x_i' \hat{\beta}$$

in

$$\hat{\mu}_i = g^{(-1)}(\hat{\eta}_i).$$

4.4. Sklepanje na osnovi GLM

4.4.1. Asimptotske lastnosti cenilke po metodi največjega verjetja

Lastnosti teh cenilk po metodi največjega verjetja v končnih vzorcih razen v posebnih primerih, kot je linearna regresija, niso znane, zato nas zanimajo asimptotske lastnosti cenilke po metodi največjega verjetja. Zaradi lastnosti metode največjega verjetja ob standardnih pogojih regularnosti velja, da je za velike n cenilka po metodi največjega verjetja $\hat{\beta}$ za parameter β nepristranska, najboljša in je porazdeljena asimptotsko normalno. Njena kovarianccna matrika je inverz Fisherjeve informacijske matrike J. Elemente J lahko zapišemo kot $\mathbb{E}(-\partial^2 L/\partial\beta_h\partial\beta_j)$. Da poiščemo kovariančno matriko, lahko uporabimo naslednji rezultat.

$$\mathbb{E}\left[\frac{-\partial^2 L}{\partial\beta_h\partial\beta_j}\right] = \mathbb{E}\left[\frac{\partial L}{\partial\beta_h}\cdot\frac{\partial L}{\partial\beta_j}\right],$$

kar drži za eksponetno družino porazdelitev. Lahko razčlenimo

$$\mathbb{E}\left[\frac{-\partial^2 L}{\partial\beta_h\partial\beta_j}\right] = \sum_i \mathbb{E}\left[\frac{(y_i-\mu_i)x_{hi}}{\phi V(\mu_i)g'(\mu_i)}\cdot\frac{(y_i-\mu_i)x_{ji}}{V(y_i)}\frac{\partial\mu_i}{\partial\eta_i}\right].$$

Ker je $\phi V(\mu_i) = \mathrm{Var}[y_i] = \mathbb{E}[(y_i-\mu_i)^2]$ lahko to poenostavimo v

$$\sum_i \frac{x_{hi}x_{ji}}{\mathrm{Var}[y_i]g'(\mu_i)^2}.$$

Naj bo W diagonalna matrika z diagonalnimi elementi

$$w_i = \frac{1}{\mathrm{Var}[y_i]g'(\mu_i)^2}.$$

Potem lahko, če posplošimo tipični element informacijske matrike na celotno matriko, dobimo informacijsko matriko

$$J = X'WX.$$

Izrek 4.1. *Cenilka $\hat{\beta}$ parametra β po metodi največjega verjetja je asimptotsko porazdeljena z $N(\beta, (X'WX)^{-1})$ porazdelitvijo, kjer je W diagonalna matrika z elementi $w_i = \frac{1}{\mathrm{Var}[y_i]g'(\mu_i)^2}$.*

Asimptotska kovariančna matrika je ocenjena z $\widehat{\mathrm{Var}}[\hat{\beta}] = (X'\hat{W}X)^{-1}$, kjer je $\hat{W}$ enaka W z ocenama $\hat{\beta}$ in $\hat{\phi}$ [1, 125-127].

4.4.2. Waldov test za posamezen parameter

Podobno kot pri linearni regresiji bomo tudi pri GLM opisali nekaj statističnih testov, s katerimi lahko presojamo vpliv pojasnjevalnih spremenljivk na odvisno spremenljivko. Prvi test, ki ga lahko uporabimo pri GLM, je t.i. Waldov test, ki testira posamezne parametre β_i.

Ocenjene standardne napake $\hat{SE}(\hat{\beta}_i)$ ocen parametrov β_i dobimo iz diagonalnih elementov kovariančne matrike asimptotske porazdelitve $\hat{\beta}$. Želimo testirati hipotezo

$$H_0 : \beta_i = 0.$$

Testno statistiko t, ki jo uporabimo, izračunamo kot

$$t = \hat{\beta}/SE(\hat{\beta}_i),$$

in jo poimenujemo Waldova statistika. Ima asimptotsko standardno normalno porazdelitev, pri čemer pa lahko omenimo, da program R upošteva $t(n-k)$ porazdelitev. Pri velikih n je razlika zanemarljiva [1, 129-130].

Izvedba Waldovega testa pri stopnji značilnosti α.

(1) S pomočjo formule zgoraj izračunamo t-statistiko.
(2) V tabeli s kritičnimi vrednostmi studentove porazdelitve pri pravi prostostni stopnji poiščemo vrednost $t_{\alpha/2}(n-k)$. Ta vrednost je vrednost pri kateri desno na grafu leži $\alpha/2$ vrednosti te porazdelitve.
(3) H_0 zavrnemo, če je $|t_j| \geq t_{\alpha/2}(n-k)$.

Računalnik nam ob implementaciji Waldovega testa poleg te statistike vrne tudi njeno p-vrednost. Velja, da H_0 zavrnemo, če je $p < \alpha$.

S tem testom torej pri določeni stopnji zaupanja vidimo, če določena pojasnjevalna spremenljivka kaj prispeva k posplošenemu linearnemu modelu. Če ne, je njen prispevek blizu ničle in bi jo lahko izključili iz modela.

4.4.3. **Test s kvocientom verjetij za več parametrov hkrati**

Če je bil Waldov test ekvivalent t-testa pri linearni regresiji, je test s kvocientom verjetij za več parametrov hkrati ekvivalent F-testa pri linearni regresiji. Test s kvocientom verjetij primerja dva modela. Model 1 je brez ene ali več pojasnjevalnih spremenljivk, ki jih testiramo, model 2 pa ima te pojasnjevalne spremenljivke vključene. Ničelna hipoteza pravi

$$H_0 : \beta_{k_1+1} = \beta_{k_1+2} = \ldots = \beta_k = 0,$$

kjer so $\beta_{k_1+1}, \beta_{k_1+2}, \ldots, \beta_k$ parametri pri pojasnjevalnih spremenljivkah, ki jih v prvem modelu nimamo. Naj bo $k_1 + k_2 = k$, torej H_0 zadnjih k_2 parametrov postavi na 0. Test kvocienta verjetij uporablja maksimuma funkcij verjetja, in sicer ℓ_1 pri prvem modelu in ℓ_2 pri drugem modelu. Kvocient $\Lambda = \ell_1/\ell_2 \leq 1$, saj je vrednost ℓ_1 v vsakem primeru manjša ali enaka ℓ_2. Testno statistiko tega testa definiramo z

$$-2\log\Lambda = -2\log(\ell_1/\ell_2) = -2(L_1 - L_2),$$

kjer sta L_1 in L_2 maksimuma logaritemske funkcije verjetja. Pod pravimi pogoji je zgornji rezultat porazdeljen s hi-kvadrat porazdelitvijo z k_2 prostostnimi stopnjami [1, 128-129].

Izvedba testa s kvocientom verjetij testa pri stopnji značilnosti α.

(1) S pomočjo formule zgoraj izračunamo testno statistiko $-2\log\Lambda$.
(2) V tabeli s kritičnimi vrednostmi hi-kvadrat porazdelitve pri pravi prostostni stopnji poiščemo vrednost $\chi^2_\alpha(k_2)$. Ta vrednost je vrednost, pri kateri desno na grafu leži α vrednosti te porazdelitve.
(3) H_0 zavrnemo, če je $-2\log\Lambda \geq \chi^2_\alpha(k_2)$.

4.4.4. **Devianca**

Naj bo $L(\mu)$ logaritemska funkcija verjetja, parametrizirana s pomočjo vektorja pričakovanih vrednosti $\mu = (\mu_1, \ldots, \mu_n)$. Naj bo $L(\hat{\mu})$ maksimum logaritemske funkcije verjetja za GLM, ki ga obravnavamo. Najvišji možni maksimum, ki ga lahko dosežemo, bi bil $L(y)$. To bi se zgodilo pri modelu, ki bi popolno modeliral odvisno spremenljivko y in bi imel za vsako enoto v vzorcu svojo pojasnjevalno spremenljivko. Tak model imenujemo nasičeni model. Tak model pojasni vso variabilnost odvisne spremenljivke modela. Kljub temu, da se na prvi pogled to sliši zelo v redu, pa velja, da nasičen model ni model, ki pojasni povezavo med odvisno in neodvisnimi spremenljivkami, saj se preveč prilega podatkom, ki smo jih modelirali in torej ne pojasni splošne povezave, ki jo želimo. Pogosto pa tak model velja za osnovo pri primerjanju z ostalimi modeli.

S testno statistiko

$$-2\log\left(\frac{\ell(\hat{\mu})}{\ell(y)}\right) = -2(L(\hat{\mu}) - L(y))$$

na osnovi kvocienta verjetij lahko testiramo hipotezo H_0, da naš model zadovoljivo opiše podatke, proti alternativi H_1, da je model povsem neustrezen. Testna porazdelitev je $\chi^2(n-k)$. Če H_0 zavrnemo, lahko zaključimo, da smo v model pozabili vključiti vsaj eno zelo pomembno pojasnjevalno spremenljivko. V nasprotnem primeru (H_0 ne zavrnemo), pa to ne pomeni, da je model zares ustrezen. Lahko je slab v primerjavi z modelom, ki vključuje le kakšno pojasnjevalno spremenljivko več.

Naj bodo $\hat{\theta}$ ocene parametrov θ v našem modelu, in $\tilde{\theta}$ ocene parametrov θ v nasičenem modelu. Zgornjo testno statistiko lahko zapišemo kot

$$\begin{aligned}
&-2(L(\hat{\mu}) - L(y)) = \\
&= 2\sum_i \frac{y_i\tilde{\theta}_i - B(\tilde{\theta}_i)}{\phi} - 2\sum_i \frac{y_i\hat{\theta}_i - B(\hat{\theta}_i)}{\phi} = \\
&= \frac{2}{\phi}\sum_i (y_i(\tilde{\theta}_i - \hat{\theta}_i) - B(\tilde{\theta}_i) + B(\hat{\theta}_i)) \\
&= \frac{D(\hat{\mu})}{\phi}.
\end{aligned}$$

Statistiko $\frac{D(\hat{\mu})}{\phi} = D^*$ imenujemo skalirana devianca modela, statistiko $D(\hat{\mu})$ pa devianca modela. Iz lastnosti $L(\hat{\mu}) \leq L(y)$ sledi, da večja, kot je devianca, slabše je prileganje modela podatkom.

Če je parameter razpršenosti ϕ znan, lahko skalirano devianco uporabimo za testiranje modela. Z njim namreč testiramo obravnavani model proti alternativi nasičenega modela. Obravnavani model zavrnemo, če vzorčna skalirana devianca presega $(1-\alpha)$-ti kvantil $\chi^2(n-k)$ porazdelitve. Če razpršitveni parameter ϕ ni znan, si lahko pomagamo z njegovo oceno $\hat{\phi}$ [1, 132-134].

Za model linearne regresije je devianca točno vsota kvadratov ostankov. Ker velja $\phi = \sigma^2$, pomeni, da je skalirana devianca enaka $\sum_i (y_i - \hat{\mu}_i)^2/\sigma^2$.

Izračunajmo še devianco za primer gama GLM-a. Spomnimo se, da v primeru gama GLM-a velja $\theta = -\frac{1}{\mu}$ in $B(\theta) = \log(-\frac{1}{\theta}) = \log(\mu)$. Prav tako pri zasičenem modelu velja $y_i = \mu_i$. Devianca za eno opažanje je tako

$$2(y_i(-1/y_i + 1/\mu_i) - \log(y_i) + \log(\mu_i)) = -2(1 - \frac{y_i}{\mu_i} + \log(\frac{y_i}{\mu_i})).$$

Iz rezultata lahko opazimo, da je, ko je $y_i = \mu_i$, vrednost deviance enaka 0.

Devianco nam računalniški programi ponujajo kot del standardnega izpisa rezultatov. S pomočjo D^* lahko naredimo kar nekaj statističnih testov. Eden takih je na primer test za dva gnezdena modela. Namesto, da bi vsakega posebej primerjali z nasičenim, ju lahko primerjamo med seboj. V prvi model vključimo vse pojasnjevalne spremenljivke, nato pa nekatere izmed pojasnjevalnih spremenljivk izključimo. Tako dobimo drugi model. Razlika $D_2^* - D_1^*$ med skaliranima deviancama obeh modelov je porazdeljena po χ^2 z razliko prostostnih stopenj $df_2 - df_1$ obeh modelov. Prostostna stopnja modela (df) je definirana kot razlika med številom realizacij in številom parametrov $(n - k)$. S pomočjo tega lahko testiramo statistično značilnost parametrov, ki jih ni v drugem modelu. V praksi lahko s tem testiramo izboljšave modela z dodajanjem ali odstranjanjem različnih pojasnjevalnih spremenljivk.

Težava lahko nastane pri porazdelitvah, kjer ne poznamo vrednosti parametra razpršenosti ϕ. Primer take porazdelitve je gama porazdelitev. V tem primeru ne moremo izvesti testov, ki smo jih opisali zgoraj, velja pa, da je testna statistika

$$F = \frac{(D_2 - D_1)/(df_2 - df_1)}{D_1/df_1}.$$

Porazdeljena po F-porazdelitvi z $(df_2 - df_1)$ in df_1 prostostnimi stopnjami.

Izvedba F-testa pri stopnji značilnosti α je tako preprosta.

(1) S pomočjo formule zgoraj izračunamo F-statistiko.
(2) V tabeli s kritičnimi vrednostmi F-porazdelitve pri pravih prostostnih stopnjah poiščemo kritično vrednost c. Ta vrednost je vrednost, pri kateri desno na grafu leži α vrednosti te porazdelitve.
(3) H_0 zavrnemo, če je $F \geq c$.

4.5. Ostanki

Razlike med z modelom ocenjenimi vrednostmi $\hat{\mu}_i$ in dejanskimi realizacijami y_i lahko merimo na več načinov. Za potrebe tega dela bomo definirali deviančni ostanek r_i^D kot

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{2(y_i(\tilde{\theta}_i - \hat{\theta}_i) - B(\tilde{\theta}_i) + B(\hat{\theta}_i))}.$$

Deviančni ostanek je pravzaprav enak korenu prispevka i-te enote k devianci modela, ki smo jo definirali že prej, in je pomnožen z 1 ali z -1, glede na to, ali je realizacija večja ali manjša od napovedi. Ti deviančni ostanki imajo kar nekaj lastnosti, ki nam pomagajo pri potrjevanju modela. Za zvezne porazdelitve so ti ostanki bolj normalno porazdeljeni kot bi bili porazdeljeni navadni ostanki. Deviančne ostanke lahko torej testiramo za normalnost. Če test ugotovi, da normalnosti verjetno ni, je to znak za to, da naš model ni dober.

Težava deviančnih ostankov je, da imajo različne variance. Zato lahko deviančne ostanke standardiziramo. S standardiziranjem dosežemo to, da ima varianca teh ostankov vrednost 1. Definirajmo torej standardiziran deviančni ostanek r_i^{DS}

$$r_i^{DS} = \frac{\text{sign}(y_i - \hat{\mu}_i)}{\sqrt{\hat{\phi}(1-\hat{h}_i)}} \sqrt{2(y_i(\tilde{\theta}_i - \hat{\theta}_i) - B(\tilde{\theta}_i) + B(\hat{\theta}_i))},$$

kjer je $\hat{h}_i$ t.i. vzvod. Vzvod meri vpliv realizacije same na svojo napovedano vrednost. Formalno je definicija vzvoda zapletena, intuitivno pa nam pove, za koliko se spremeni napovedana vrednost, če spremenimo realizacijo. Vedno zavzame vrednosti med 0 in 1. Vzvod, ki je blizu 1, pomeni, da sta sprememba opazovane vrednosti in sprememba napovedane vrednosti popolnoma korelirani. Če se realizacija spremeni za neko vrednost, se bo tudi napovedana vrednost spremenila za enako vrednost. Zaradi tega je ostanek pri tej realizaciji zelo majhen. Če deviančni ostanek normiramo s $(1-\hat{h}_i)$ to pomeni, da deviančni ostanek na tem mestu povečamo za primeren faktor [2, 51-55].

Ta nam pomaga, da lahko primerjamo realizacije, ki imajo različna povprečja. Če so standardizirani deviančni ostanki porazdeljeni po $N(0,1)$ to pomeni, da je naš model v redu.

Zanimiva mera, ki jo lahko grafično prikažemo je tudi vzvod. Če imamo nekaj realizacij, kjer je vzvod veliko večji od vzvoda ostalih opazovanj, lahko to pomeni, da imajo te realizacije nepotreben vpliv na model. S pomočjo vzvoda jih lahko opazimo, nato pa podrobneje pogledamo, katere realizacije to so, ter se na podlagi lastne presoje odločimo, če jih v modelu obdržimo ali ne.

5. Uporaba GLM na podatkih

Ker ne poznamo pravega GLM modela, bomo z GLM najprej modelirali samo polovico vseh podatkov, ki smo jih naključno izbrali. Nato bomo na osnovi rezultatov prvega modela le tega skušali izboljšati. S tem tvegamo, da bo naš končni model preveč sledil podatkom in ne bo izražal splošnih zakonitosti zavarovalniškega trga. Ker imamo na voljo zelo velik vzorec, bomo slučajno izbrali polovico enot in jih analizirali. Na drugi polovici enot bomo nato preverili ali lahko posplošimo naše ugotovitve.

Na tem mestu lahko s pomočjo nabora slučajnih spremenljivk, ki smo jih uporabili že pri linearni regresiji, začnemo modelirati višine škod z GLM. Podobno kot prej bomo za modeliranje uporabljali program R, kjer lahko podatke preprosto modeliramo z GLM. V programu R obstaja ukaz `glm()`, s katerim lahko podatke modeliramo na podoben način, kot pri linearni regresiji.

Ker modeliramo višino škode, ki ima samo nenegativne vrednosti in je navzgor neomejena, bomo pri GLM-u uporabili gama porazdelitev. Zaradi nenegativnosti in enostavne interpretacije regresijskih parametrov izberemo logaritemsko povezovalno funkcijo. Izvedimo ukaz v programu R.

```
gama_glm <- glm(Skoda ~ Izobrazba + Tip.Vozila
                      + Status.Zaposlitve + Porocni.Stan
                      + Tip.Naselja + Spol
```

```
                    + Kritje + Velikost.Vozila
                , family = Gamma(link = log), data = podatki_model)
```

V programu R dobimo rezultat, ki izgleda zelo podobno, kot pri linearni regresiji.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1650  -0.1923  -0.0527   0.1380   1.0916

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
   (Intercept)               4.587316   0.025091 182.827  < 2e-16
Izobrazba (SŠ ali nižja)
   Višja šola               -0.042631   0.012982  -3.284 0.001031
   Diploma                  -0.019773   0.012808  -1.544 0.122710
   Podiplomska              -0.043056   0.017236  -2.498 0.012523
Tip Vozila (Dvovratno vozilo)
   Štirivratno vozilo        0.018173   0.012716   1.429 0.153043
   Športno vozilo            0.471888   0.014512  32.516  < 2e-16
   Luksuzno vozilo           1.016063   0.027762  36.599  < 2e-16
Status Zaposlitve (Nezaposlen)
   Delavno neaktiven        -0.117599   0.017869  -6.581 5.19e-11
   Zaposlen                 -0.129570   0.013414  -9.659  < 2e-16
Poročni Stan (Poročen)
   Ločen                     0.008247   0.014275   0.578 0.563508
   Samski                    0.091750   0.012530   7.322 2.87e-13
Tip Naselja (Podeželje)
   Primestna                 1.478770   0.014041 105.322  < 2e-16
   Mestna                    1.085608   0.016326  66.495  < 2e-16
Spol (Ženski)
   M                         0.018152   0.009871   1.839 0.065988
Kritje (Osnovno)
   Razširjeno                0.203631   0.010930  18.630  < 2e-16
   Premijsko                 0.480499   0.017551  27.377  < 2e-16
Velikost Vozila (Majhen)
   Medium                   -0.031113   0.012816  -2.428 0.015231
   Velik                    -0.073381   0.019045  -3.853 0.000118

Dispersion parameter for Gamma family taken to be 0.1097885

Deviance:  701.64  on 4549  degrees of freedom
AIC: 57485
```

Podobno kot pri linearni regresiji je tudi tu začetna vrednost (intercept) predstavlja pričakovano višino škode pri škodnem zahtevku, ki se pri vseh pojasnjevalnih spremenljivkah uvršča v osnovno raven. V našem primeru je to polica, pri kateri ima zavarovalec ob sklenitvi srednješolsko ali nižjo izobrazbo, vozi dvovratno vozila, je nezaposlen, poročen, živi na podeželju, je ženska, polica ima osnovno kritje in vozi majhen avto.

Vrednost 4,587, ki je ocena za začetno vrednost, pa ne predstavlja modelirane višine škode, kot smo navajeni pri linearni regresiji. Ker imamo logaritemsko povezovalno funkcijo je to logaritem pričakovane višine škode. Da dobimo višino škode, ki jo je za zavarovalca, opisanega zgoraj, napovedal model, moramo izračunati inverz logaritemske funkcije. Izračunati moramo $e^{4,587} = 98{,}199$, kar pa je ocenjena višina pričakovane škode.

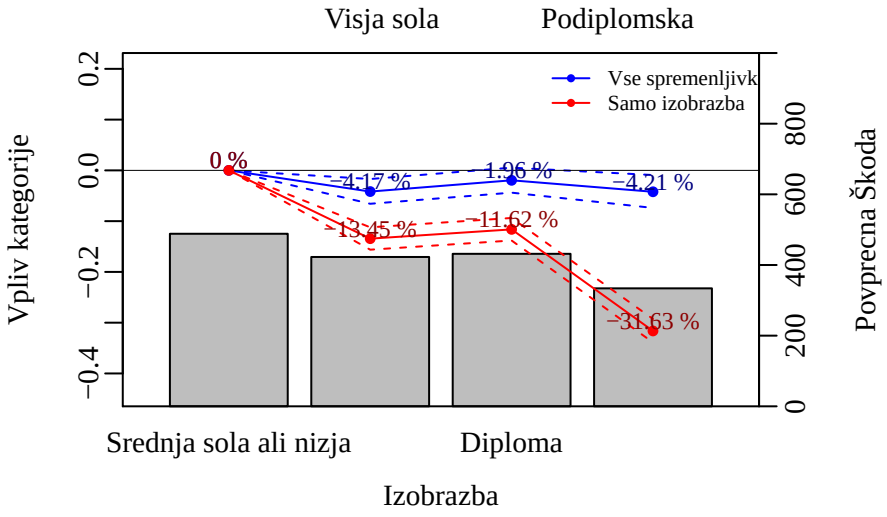
Podobno si moramo razlagati tudi druge ocene za parametre β. Poglejmo na primer pojasnjevalno spremenljivko kritje. Opisali smo že, kaj je za nas osnovni nivo. Zanima nas, kako bo na napovedano višino škode vplivala sprememba kritja iz osnovnega na razširjeno, pri tem pa vse ostale skupine pojasnjevalnih spremenljivk ostanejo enake. Zopet moramo izračunati inverz povezovalne funkcije za vrednost ocene za parameter β pri razširjenem kritju. $e^{0,204} = 1{,}2262$. To pomeni, da se ocena škode poveča za 1,2262-krat, oz. za 22,62%, če ravni vseh ostalih pojasnjevalnih spremenljivk ostanejo iste, spremeni pa se kritje in sicer iz stopnje osnovno na stopnje razširjeno. Pri tem ni pomembno, kakšne so dejanske ravni pri ostalih pojasnjevalnih spremenljivkah, važno je le, da kljub spremembi kritja ostanejo nespremenjene. Tudi vse ostale ocene za parametre β, si lahko razlagamo na enak način. če je ocena parametra β zelo majhna, uporabimo aproksimacijo $e^{\beta} \simeq 1 + \beta$ in vrednost kar direktno interpretiramo kot odstotno spremembo.

Poglejmo še primer zavarovalca, ki ima diplomsko izobrazbo, je samski zaposleni moški, ki ima osnovno kritje in vozi medium športno vozilo ter živi na podeželju. Za takega zavarovalca bo ocenjena škoda $e^{4,587} \cdot e^{-0,02} \cdot e^{0,09} \cdot e^{-0,129} \cdot e^{0,018} \cdot e^{0} \cdot e^{0,471} \cdot e^{0} \cdot e^{-0,03} = 122{,}297$.

Pri modeliranju GLM nam veliko pove tudi grafični prikaz modela. Na tem mestu bomo za vsako pojasnjevalno spremenljivko narisali sliko, ki prikazuje njen vpliv na pričakovano škodo. Narisali bomo povprečne vrednosti višine škod po vseh ravneh posamezne pojasnjevalne spremenljivke, nato pa bomo na graf narisali še oceno za prispevek posamezne ravni k oceni za višino škode, ter interval zaupanja te ocene s 95% stopnjo zaupanja. Te dobimo z normalno aproksimacijo. Modre črte bodo predstavljale ocene za model z vsemi pojasnjevalnimi spremenljivkami, rdeča črta pa bo predstavljala model, ki višino škode modelira samo z opazovano pojasnjevalno spremenljivko.

Na sliki 14 smo torej prikazali rezultate GLM modela za izobrazbo. V stolpcih je za vsako skupino pojasnjevalne spremenljivke izobrazba prikazana povprečna višina škode, kar nam da predstavo, kakšna je škoda. Nato smo s pomočjo modrih in rdečih črt poskušali prikazati rezultate GLM-a. Modre črte prikazujejo GLM, ki vsebuje vse pojasnjevalne spremenljivke, rdeče črte pa GLM, kjer višino škode modeliramo samo s pojasnjevalno spremenljivko izobrazbe. Prva skupina je osnova, nato pa za vsako skupino vrnemo procentualno vrednost, za koliko se spremeni napovedana pričakovana višina škode, če imamo namesto osnovne skupine opazovano skupino. Črtkane črte nam nakazujejo 95% interval zaupanja okoli teh ocen.

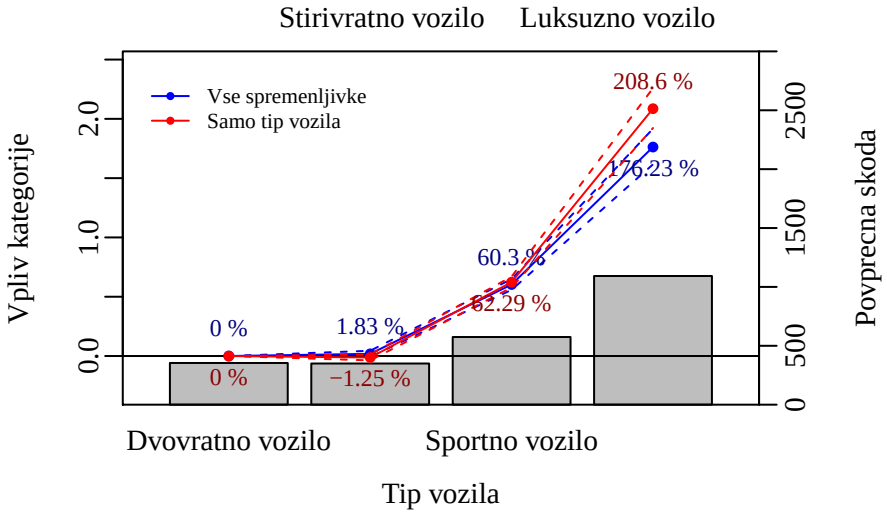
Iz slike 14 lahko razberemo, da je pojasnjevalna spremenljivka izobrazbe pri modelu z vsemi pojasnjevalnimi spremenljivkami na meji značilnosti, saj je interval zaupanja za vse skupine zelo blizu 0, oz. ničlo tudi vsebuje. Značilnost posamezne pojasnjevalne spremenljivke bomo kasneje testirali še s testi deviance. Morda bomo



Slika 14. GLM modeliranje višine škode z izobrazbo.

ravno pojasnjevalno spremenljivko izobrazbe iz modela odstranili. Ker so ravni izobrazbe urejene od najnižje do najvišje nas moti tudi to, da vpliv teh ravni ni monoton. Po drugi strani, je pojasnjevalna spremenljivka izobrazbe na pogled značilna pri modelu, ki vsebuje samo to pojasnjevalno spremenljivko.

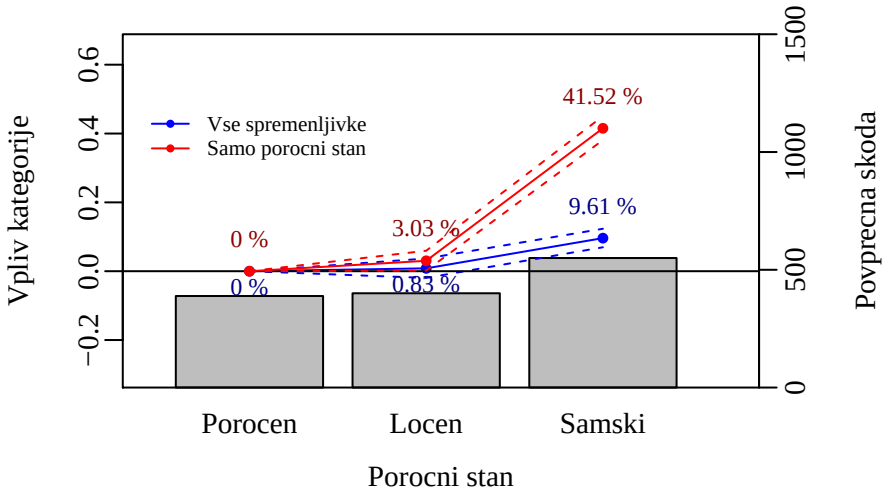
Takšno razhajanje med vplivoma izobrazbe v dveh modelih nakazuje povezanost med pojasnjevalnimi spremenljivkami. Hkrati vidimo, kako napačne so lahko interpretacije modela in posledične odločitve, če v model ne vključimo vseh pomembnih pojasnjevalnih spremenljivk.



Slika 15. GLM modeliranje višine škode s tipom vozila.

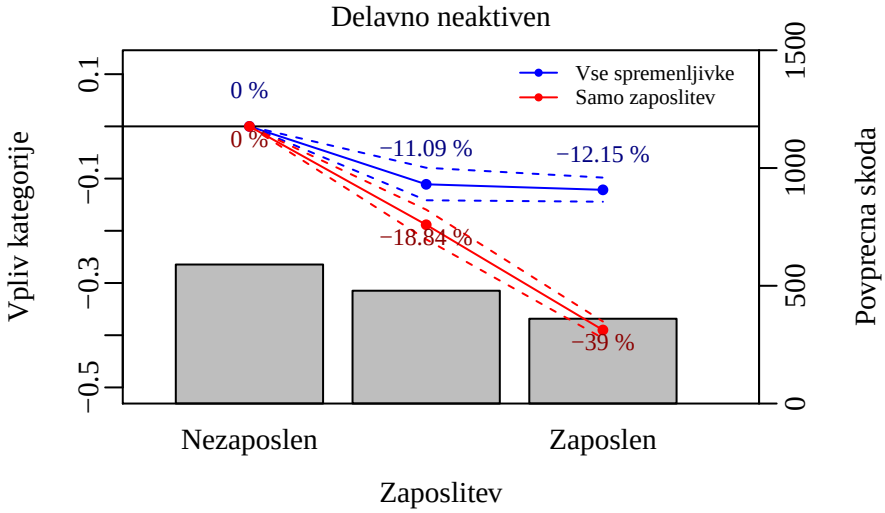
Slika 15 prikazuje tip vozila v GLM-u. Za razliko od slike 14 izgleda, da imamo zelo značilne ocene za parametre pri posameznih pojasnjevalnih spremenljivkah v obeh modelih. Ozki intervali zaupanja nakazujejo, da so ocene dokaj dobre. Zdi se nam, da sta skupini dvovratnih in štirivratnih vozil zelo podobni, torej bi bilo morda

smiselno zopet razmisliti o združevanju. Vpliv vozila je zelo podoben v modelih z vsemi ali samo to pojasnjevalno spremenljivko. To pomeni, da ta spremenljivka vsebuje informacije, ki jih ni mogoče nadomestiti z uporabo drugih spremenljivk



Slika 16. GLM modeliranje višine škode s poročnim stanom.

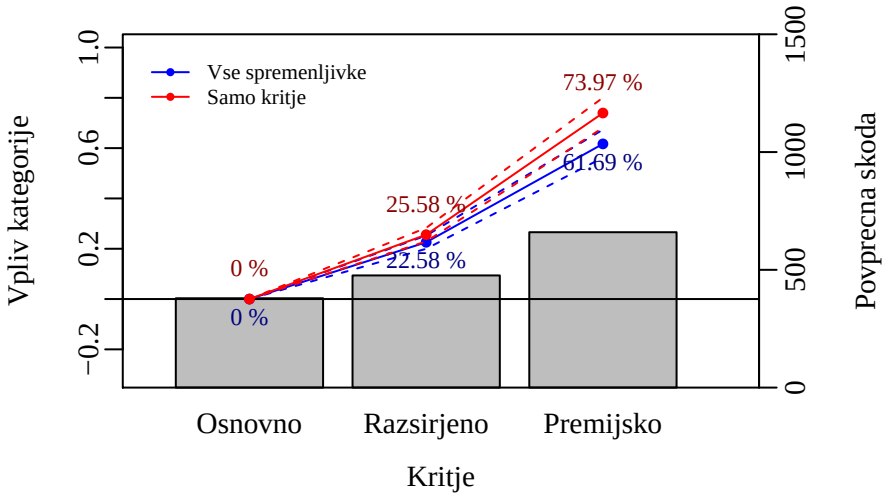
Vpliv pojasnjevalne spremenljivke poročni stan na višino škode pri GLM je prikazan na sliki 16. Opazimo, da noben model ne prikaže značilne razlike med škodo ločenih in poročenih. Po drugi strani so škode samskih očitno višje. Zanimivo je, da kljub temu model z vsemi spremenljivkami skupini samskih škode ne poveča toliko, kot jo model, kjer je edina pojasnjevalna spremenljivka poročni stan. Razlog za to je najverjetneje v povezavi med poročnim stanom in ostalimi spremenljivkami, kar smo ugotovili, ko smo računali Cramerjevo V-statistiko.



Slika 17. GLM modeliranje višine škode s statusom zaposlitve.

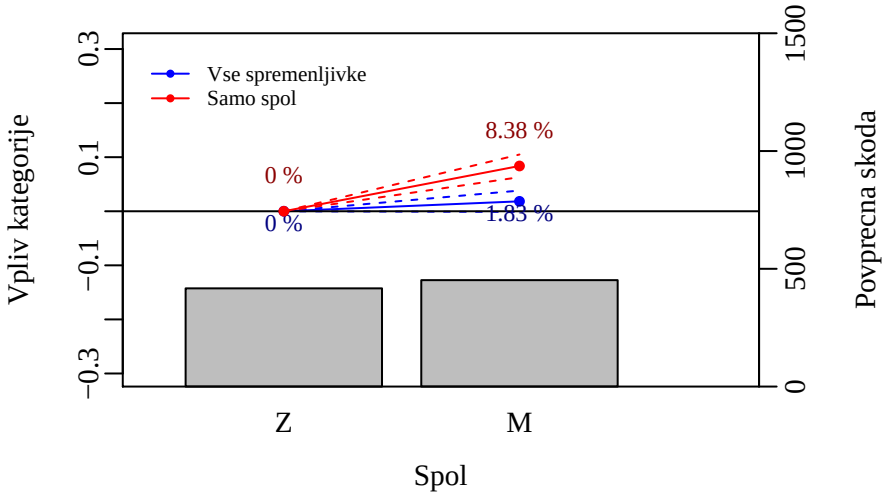
Pri pojasnjevalni spremenljivki status zaposlitve lahko iz slike 17 sklepamo, da imajo nezaposleni višjo škodo kot ostali, kar modelira tudi GLM. Zanimivo je, da

model z vsemi spremenljivkami zaposlenim in delavno neaktivnim pripisuje približno enako škodo, model, ki ima samo pojasnjevalno spremenljivko status zaposlitve pa modelira zelo velike razlike. Razlog se morda zopet skriva v povezavi med statusom zaposlitve in poročnim stanom. Cramerjeva V-statistika za ti dve spremenljivki je bila najvišja, kar lahko vidimo v tabeli 10.



Slika 18. GLM modeliranje višine škode z vrsto kritja.

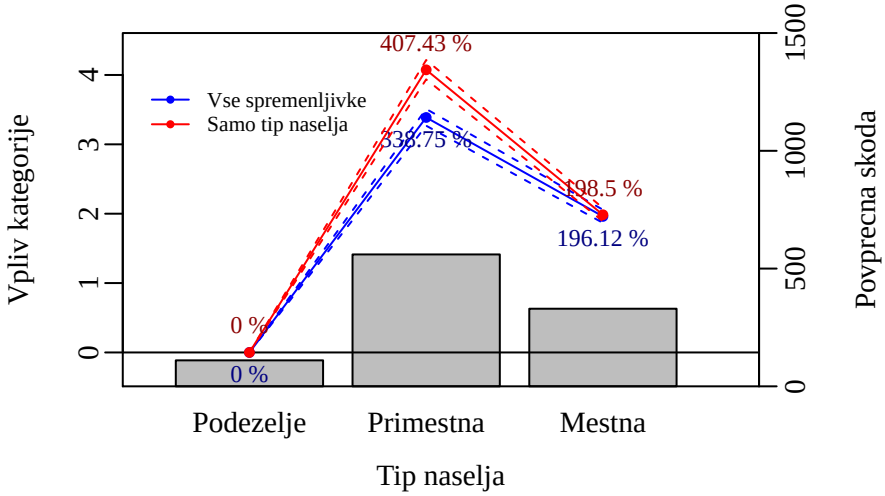
Spremenljivka iz slike 18 je najbolj očitno že vnaprej dobro izbrana. Tudi če ne bi modelirali GLM modela, bi vedeli, da je višina škode, ki nam jo krije zavarovalnica, odvisna od tega, katero kritje vzamemo. Boljše kot je kritje, več krije zavarovalnica. Prav to se pokaže tudi pri obeh modelih.



Slika 19. GLM modeliranje višine škode s spolom.

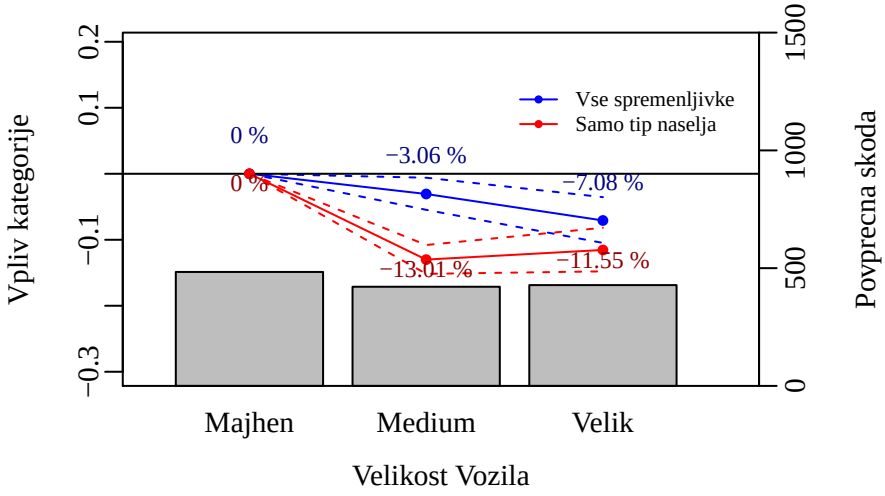
Moški imajo malo višjo povprečno škodo kot ženske, kar se vidi iz slike 19, vendar pa model z vsemi spremenljivkami statistično značilne razlike med obema skupinama

ne pokaže. Večjo razliko spet pokaže GLM, kjer imamo za pojasnjevalno spremenljivko le spol.



Slika 20. GLM modeliranje višine škode s tipom naselja.

Naslednja spremenljivka je tip naselja. Po podatkih iz slike 20 lahko vidimo, da oba modela tipu naselja pripisujeta zelo velik vpliv na napovedano višino škode. Sploh je opazno, da so v mestih in primestjih škode občutno višje kot na podeželju. Razlog bi lahko bil v večjem številu in boljših vozilih v mestih in primestnih naseljih. Ozki intervali zaupanja nakazujejo na dobre ocene za parametre β.



Slika 21. GLM modeliranje višine škode z velikostjo vozila.

Zadnja pojasnjevalna spremenljivka, ki jo pogledamo, je velikost vozila. Iz slike 21 vidimo, da imajo manjši avtomobili višje škode kot veliki in srednje veliki avtomobili. To je ravno obratno od tega, kar smo pričakovali. Možno je, da so pri nesrečah manjši avtomobili bolj poškodovani, in imajo torej višje škodne zahtevke. Druga možnost, na katero pomislimo je ta, da so manjši avtomobili manj varni, zaradi tega pa lahko

hitreje pride to telesnih poškodb, škode, ki nastanejo pri telesnih poškodbah pa so ponavadi veliko višje, kot če je nastala samo materialna škoda.

5.1. Diagnostika GLM-a

Zanima nas, kako dober je naš model. Za začetek bomo s pomočjo deviance naredili F-test na vseh spremenljivkah ter tako poskusili ugotoviti, če lahko model izboljšamo s tem, da kakšno izmed pojasnjevalnih spremenljivk odstranimo. F-test poteka zelo podobno, kot pri linearni regresiji. Modelirali bomo dva GLM-a, kjer bodo v prvem vse spremenljivke, v drugem pa bomo eno odstranili. F-test bo nato preverjal ničelno hipotezo, ki pravi, da so vsi parametri β pri pojasnjevalni spremenljivki, ki smo jo iz modela odstranili, enaki nič. Če ničelno hipotezo zavrnemo, pomeni, da vsaj ena skupina v pojasnjevalni spremenljivki nekaj prispeva k oceni za višino škode.

V programu R lahko tudi za F-test na GLM uporabimo funkcijo `anova()`. Izvedemo ukaz

```
anova(gama_glm, gama_glm1, test = "F"),
```

kjer je `gama_glm` model, ki ima vse pojasnjevalne spremenljivke, `gama_glm1` pa model, ki smo mu odstranili pojasnjevalno spremenljivko izobrazba. Dobimo naslednji rezultat.

```
Analysis of Deviance Table

Model 1: Skoda ~ Izobrazba + Tip.Vozila + Status.Zaposlitve
   + Porocni.Stan +  Tip.Naselja + Spol + Kritje + Velikost.Vozila
Model 2: Skoda ~ Tip.Vozila + Status.Zaposlitve
    + Porocni.Stan + Tip.Naselja +
    Spol + Kritje + Velikost.Vozila

  Resid. Df Resid. Dev Df Deviance      F   Pr(>F)
1      4549     739.14
2      4552     740.56 -3  -1.4134 4.1301 0.006203
---
```

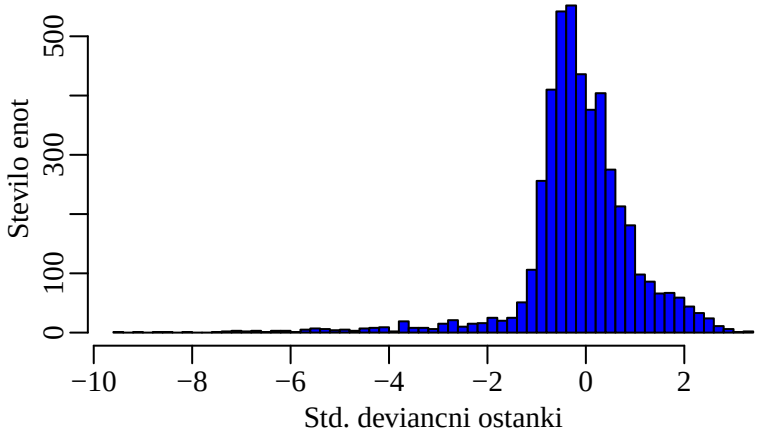
p-vrednost tega testa je 0,0062, kar pomeni, da lahko s 5% stopnjo značilnosti zavrnemo ničelno hipotezo. To pomeni, da vsaj ena raven pojasnjevalne spremenljivke izobrazbe nekaj prispeva k oceni za višino škode. Naredimo tak test za vse pojasnjevalne spremenljivke.

Kakor vidimo v tabeli 15, vse pojasnjevalne spremenljivke k modelu prispevajo vsaj nekaj značilnega. Če upoštevamo rezultate tega testa, potem iz modela ne bi odstranili nobene spremenljivke, da bi izboljšali model.

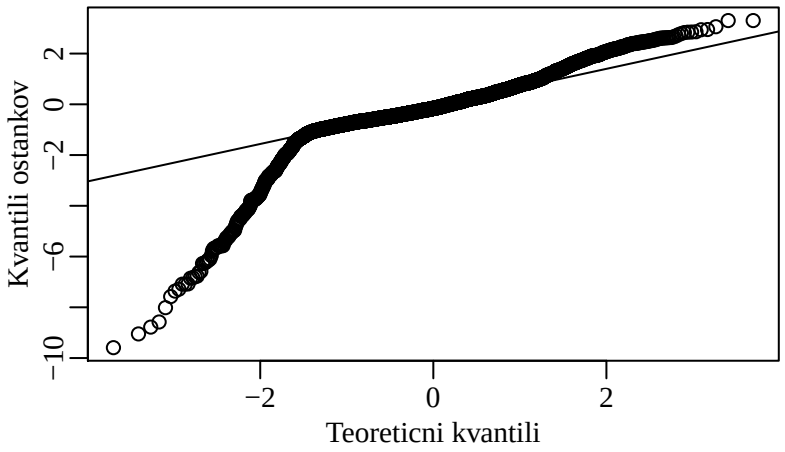
Slika 22 prikazuje histogram standardiziranih deviančnih ostankov. Vidimo, da ima porazdelitev krajši in debelejši rep v desno ter tanjši in zelo dolg rep v levo. To pomeni, da nekatere škode modeliramo precej višje, kot so v resnici. Iz poslovnega vidika to ni slabo, saj pomeni, da nas slab škodni rezultat težje preseneti. Vsekakor je napoved boljša kot pri linearni regresiji. Bi si pa seveda želeli imeti histogram deviančnih ostankov, ki bi bolj spominjal na normalno porazdelitev. Izrazit levi rep porazdelitve nakazuje tudi Q-Q grafikon na sliki 23.

TABELA 15. F-test za posamezne pojasnjevalne spremenljivke pri GLM.

Pojasnjevalna spremenljivka	p-vrednost
Kritje	$2{,}2 \cdot 10^{-16}$
Izobrazba	0,0062
Zaposlitveni status	$2{,}2 \cdot 10^{-16}$
Spol	0,0001
Tip naselja	$2{,}2 \cdot 10^{-16}$
Poročni stan	$1{,}3 \cdot 10^{-13}$
Tip vozila	$2{,}2 \cdot 10^{-16}$
Velikost vozila	$6{,}2 \cdot 10^{-5}$

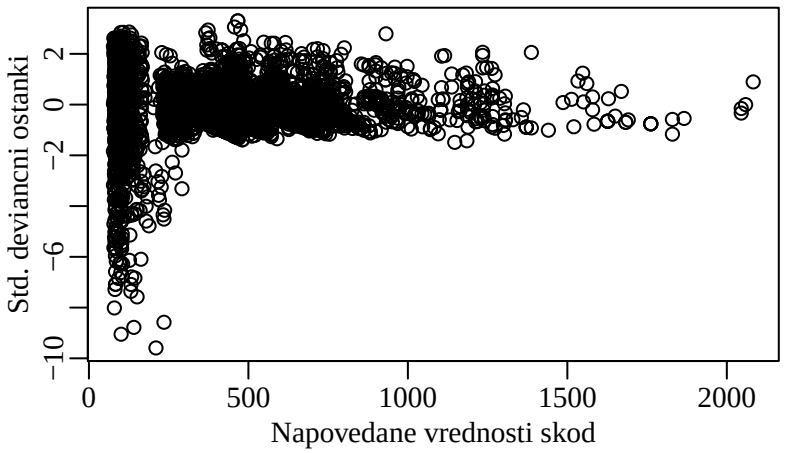


SLIKA 22. Histogram standardiziranih deviančnih ostankov GLM-a.



SLIKA 23. Q-Q grafikon standardiziranih deviančnih ostankov GLM-a.

Slika 24 prikazuje razsevni grafikon ocenjenih pričakovanih vrednosti škod in standardiziranih deviančnih ostankov. V ustreznem modelu bi morali biti ostanki neodvisni od napovedane vrednosti. Grafikon bi moral izgledati kot simetričen oblak točk brez kakršnegakoli vzorca. Opazimo, da so pri nizkih napovedih nekateri ostanki izrazito negativni.



Slika 24. Razsevni grafikon standardiziranih deviančnih ostankov in napovedanih vrednosti GLM.

Če se vrnemo na sliko 1, lahko vidimo, da imamo na začetku histograma ravno črto. Torej imamo za majhne škode približno enako število enot. Vsebinskega razloga za to ne poznamo, saj v podatkih ni opisa, lahko pa bi šlo na primer za odbitno franšizo ali pa kaj podobnega.

Model bomo v naslednjem razdelku s pomočjo zbranih ugotovitev poskušali še malce izboljšati.

5.2. **Iteracija modela**

V splošnem iz enega samega GLM-a ni mogoče določiti, katere vse pojasnjevalne spremenljivke so pomembne za naš model, saj se z vključevanjem in izključevanjem posameznih pojasnjevalnih spremenljivk v model spreminjajo parametri in statistične značilnosti pri ostalih pojasnjevalnih spremenljivkah. S tem v mislih je velikokrat dobro narediti več različnih modelov, ki si sledijo v seriji, da določimo optimalni nabor pojasnjevalnih spremenljivk.

Poleg tega se nam splača, da začnemo modelirati s čisto vsemi pojasnjevalnimi spremenljivkami, ki jih imamo na voljo. Bolj smiselno je najprej dodati vse pojasnjevalne spremenljivke, za katere domnevamo, da so pomembne, potem pa počasi odstranjujemo še posamezne pojasnjevalne spremenljivke ter računamo teste. Najbolj značilne obdržimo v modelu. Vse to je najbolje početi ročno, saj s tem opazimo največ. Če je pojasnjevalnih spremenljivk zelo veliko, lahko ta proces avtomatiziramo, vendar se taka avtomatizacija velikokrat izkaže za slabo, saj je težko sprogramirati vse možnosti, na podlagi katerih izločimo ali pa vključimo kakšno

od pojasnjevalnih spremenljivk. Naš model ima samo devet pojasnjevalnih spremenljivk, zato bomo poskusili z izločanjem.

Na tem mestu je zelo pomembno poudariti, da zelo pomembno vlogo pri modeliranju GLM tudi naša intuicija in vsebinsko poznavanje problema. Razmisliti moramo, katere pojasnjevalne spremenljivke bi v modelu imeli radi in katerih ne bi imeli. Pogosto je več modelov med seboj zelo podobnih in se potem sami odločamo, katere pojasnjevalne spremenljivke so za nas bolj pomembne ali pa bolj dostopne. Tudi zaradi tega je iteracije GLM-ev težko avtomatizirati.

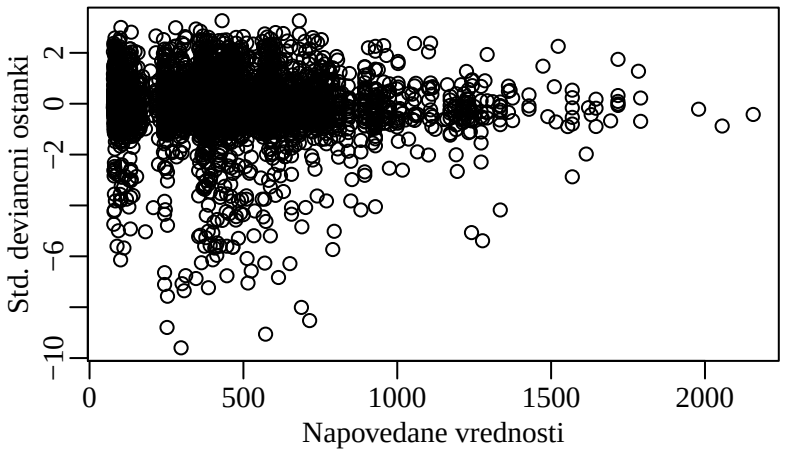
Ko smo v tabeli 15 izvedli F-test, nobena izmed pojasnjevalnih spremenljivk statistično ni upravičila izključitve iz modela, opazimo pa lahko, da k skupni oceni pričakovane višine škode podatka o izobrazbi in spolu ne pripomoreta veliko. Iz slik 14 in 19 lahko vidimo, da je prispevek k oceni škode za katerokoli raven nižji od 5%. Odločimo se lahko, da to za nas poslovno nima prevelikega pomena. Poleg tega je podatek o spolu problematičen tudi z etične in pravne strani, saj v Sloveniji pri postavljanju cene zavarovalnih produktov ne smemo delati razlik med spoloma. Na podlagi zgoraj navedenih argumentov bomo tako iz modela odstranili omenjeni pojasnjevalni spremenljivki ter nato zopet pogledali grafikone, s katerimi ocenjujemo model. Poglejmo kaj nam tokrat vrne model.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1722  -0.1920  -0.0537   0.1390   1.0803

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept                4.569144   0.023320 195.934  < 2e-16
Tip Vozila (Dvovratno vozilo)
   Štirivratno vozilo     0.016064   0.012712   1.264  0.20640
   Športno vozilo         0.470855   0.014504  32.464  < 2e-16
   Luksuzno vozilo        1.013794   0.027783  36.489  < 2e-16
Status Zaposlitve (Nezaposlen)
   Delavno neaktiven     -0.121118   0.017860  -6.781 1.34e-11
   Zaposlen              -0.130433   0.013405  -9.730  < 2e-16
Poročni stan (Poročen)
   Ločen                  0.006951   0.014261   0.487  0.62601
   Samski                 0.093727   0.012528   7.481 8.78e-14
Tip naselja (Podeželje)
   Primestna              1.485922   0.013882 107.037  < 2e-16
   Mestna                 1.086541   0.016327  66.550  < 2e-16
Kritje (Osnovno)
   Razširjeno             0.203904   0.010932  18.652  < 2e-16
   Premijsko              0.481758   0.017557  27.440  < 2e-16
Velikost vozila (Majhno)
   Medium                -0.029924   0.012816  -2.335  0.01959
   Velik                 -0.071412   0.019049  -3.749  0.00018
```

```
Dispersion parameter for Gamma family taken to be 0.109976

Deviance:  703.4  on 4553  degrees of freedom
AIC: 57489
```

Če primerjamo ocene parametrov β v polnem in poenostavljenem modelu, opazimo, da izločitev dveh pojasnjevalnih spremenljivk ni vplivala na ocene ostalih parametrov. Ti so ohranili svoje predznake, velikostne rede in statistične značilnosti. To pomeni, da interpretacije, ki smo jih izpisali ob polnem modelu še vedno veljajo.

Ne vemo še, ali smo izboljšali porazdelitve deviančnih residualov. Zopet narišimo slike, ki nam bodo pokazale bolj jasno sliko, s katero si bomo pomagali pri oceni modela.

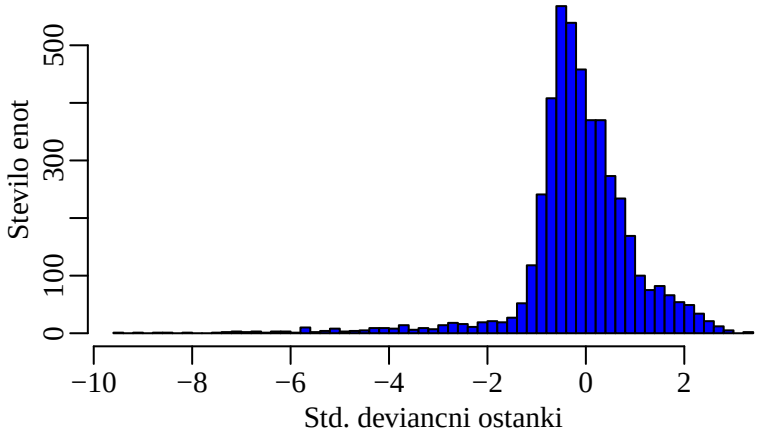


Slika 25. Razsevni grafikon standardiziranih deviančnih ostankov in napovedanih vrednosti drugega GLM-a.

Iz slike 25 vidimo, da razsevnega grafikona deviančnih ostankov in napovedanih vrednosti nismo vidno izboljšali. Res je, da smo nižje vrednosti škod sedaj opisali malo boljše, saj anomalija na grafikonu ne izstopa več tako kot na sliki 24, vendar pa se je zato povečal standardni odklon ostankov za vse vrednosti. Retultati so neneavadni. Če smo pri sliki 24 dobili občutek, da najbolj negativni ostanki sistematično nastopajo pri enotah, za katere z modelom napovemo škodo, so veliki negativni ostanki na sliki 25 bolj enakomerno razpršeni glede na napovedano pričakovano škodo. Možno je, da pojasnjevalni spremenljivki, ki smo ju v iteraciji modela izpustili, pojasnita nekatere izrazito negativne ostanke.

Na sliki 26 lahko še vedno opazimo debelejši rep v desno stran, ter tanjši in zelo dolg rep v levo stran. Naš model torej še vedno ni najboljši in bomo tako nadaljevali z iteracijami modela.

Za odstranitev drugih pojasnjevalnih spremenljivk nimamo ne statističnega ne poslovnega razloga, vendar pa še imamo nekaj možnosti, da poskušamo izboljšati



Slika 26. Histogram standardiziranih deviančnih ostankov drugega GLM-a.

model. Najprej bomo preverili, če med kakšnimi pojasnjevalnimi spremenljivkami obstajajo interakcije, nato pa bomo poskuili poiskati vrednosti, zaradi katerih se na histogramu deviančnih ostankov pojavi rep v levo, ter na njih poiskati kakšno skupno lastnost, ter jih na podlagi te lastnosti obravnavati ločeno od drugih podatkov.

Interakcija med dvema pojasnjevalnima spremenljivkama pomeni, da ocenjena vrednost, ki nas zanima, ni odvisna le od vrednosti vsake posamezne pojasnjevalne spremenljivke, ampak je povezana z vrednostjo kombinacije teh dveh spremenljivk. Če med dvema pojasnjevalnima spremenljivkama obstaja interakcija, moramo v model vključiti tudi kombinacijo obeh. Da si bomo lažje predstavljali interakcije med dvema spremenljivkama, si poglejmo preprost primer.

Primer 5.1. Modelirajmo pričakovano višino škodnega zahtevka s pomočjo dveh pojasnjevalnih spremenljivk. Prva pojasnjevalna spremenljivka je spol, druga pa je lokacija, kjer imamo dve ravni. Mestna lokacija in podeželje.

Tabela 16. Primer spremenljivk brez interakcije.

Višina škode	Mesto	Podeželje
Moški	300	100
Ženske	600	200

V tabeli 16 vidimo, da imajo moški dvakrat nižjo višino škode kot ženske ne glede na to, ali so iz mesta ali podeželja, v mestih pa je višina škode trikrat višja kot na podeželju, ne glede na to, kakšnega spola je oseba. Višino škode lahko v tem primeru preprosto modeliramo samo z obema pojasnjevalnima spremenljivkama, saj ne vplivata druga na drugo.

V tabeli 17 pa imamo primer podatkov, kjer interakcija med dvema pojasnjevalnima spremenljivkama obstaja. Ženske imajo v mestih dvakrat nižjo škodo kot moški, na podeželju pa je škoda za ženske trikrat nižja od škode, ki jo doživijo moški. Prav tako je škoda na podeželju trikrat višja v kot v mestu za moške, za ženske pa je

Tabela 17. Primer z interakcijo.

Višina škode	Mesto	Podeželje
Moški	200	600
Ženske	100	200

škoda na podeželju le dvakrat višja kot v mestu. V tem primeru neka interakcija obstaja. V regresijskem modelu bi interakcijo modelirali tako, da bi med pojasnjevalne spremenljivke dodali še skupno pojasnjevalno spremenljivko, ki bi imela vrednosti moški v mestih, moški na podeželju, ženske v mestih in ženske na podeželju. Če je osnovna raven za lokacijo mesto in osnovna raven za spol moški, potem dodatna slamnata spremenljivka predstavlja skupino žensk s podeželja.

Interakcije v našem modelu bomo preverili tako, da bomo v model iz prve iteracije (torej brez izobrazbe in spola) dodali še kombinacijo dveh pojasnjevalnih spremenljivk. Nato bomo na obeh modelih naredili F-test podobno, kot smo ga naredili v tabeli 15. Če bo katera izmed kombinacij, ki jih bomo dodali v model, statistično značilno vplivala na model, jo bomo v model dodali.

Tabela 18. F-test za interakcije.

p-vrednost	*Zap. status*	*Poročni stan*	*Lokacija*	*Vrsta kritja*	*Tip vozila*	*Velikost vozila*
Zap. status		$2 \cdot 10^{-16}$	0,037	0,931	0,336	0,297
Poročni stan	$2 \cdot 10^{-16}$		$6 \cdot 10^{-10}$	0,472	0,126	0,004
Lokacija	0,037	$6 \cdot 10^{-10}$		$5 \cdot 10^{-6}$	0,007	$2 \cdot 10^{-5}$
Vrsta kritja	0,931	0,472	$5 \cdot 10^{-6}$		0,015	0,057
Tip vozila	0,336	0,126	0,007	0,015		0,011
Velikost vozila	0,297	0,004	$2 \cdot 10^{-5}$	0,057	0,011	

Iz tabele 18 lahko vidimo, da F-test potrdi značilnost kar nekaj kombinacij pojasnjevalnih spremenljik iz modela. Posebej izstopajo kombinacije poročni stan in zaposlitveni status, poročni stan in lokacija ter lokacija in velikost vozila. Poleg teh jepri 5% stopnji značilnosti značilnih še kar nekaj drugih kombinacij spremenljivk, vendar se moramo vedno vprašati, ali znamo značilno interakcijo tudi interpretirati. Ko smo v model poskusili dodati še kakšno kombinacijo spremenljivk, se standardizirani deviančni ostanki niso izboljšali. Zaradi tega, in tudi zato ker bi s tem pridobili veliko novih kombinacijskih ravni pojasnjevalnih spremenljivk, ki jih je težko interpretirati, se odločimo, da kombinacij v model ne bomo dodali.

Druga metoda za izboljšave modela, ki smo se jo poslužili, je pregled vrednosti, ki se nahajajo v levem repu histograma iz slike 26. Pregledali smo enote, kjer je standardiziran deviančni ostanek manjši od -3. Takih enot je 120. Opazili smo, da je bilo 64 ot teh enot takih, kjer je bil zavarovalec poročen, s podeželja ter je vozil medium vozilo. Če iz podatkov odstranimo te enote potem dobimo histogram deviančnih ostankov, ki ima obliko bolj podobno normalni porazdelitvi kot na sliki 26. Kljub temu take odstranitve podatkov ne moremo pravilno utemeljiti ne statistično ne poslovno, saj ne poznamo pravega izvora podatkov. Boljša bi bila vpeljava dodatne slamnate spremenljivke, ki bi imela vrednost ena za točno to skupino ljudi in vrednost 0 za vse ostale enote. Vendar bi zopet naleteli na težave, kako interpretirati pripadajoči parameter.

Pri modeliranju naših podatkov imamo predvsem težavo z nižjimi pričakovanimi škodami, kar smo opisali že preden smo začeli z iteracijo modela. Ker pojava ne moremo pojasniti poslovno, saj imamo o vsebini podatkov bolj skope informacije lahko na tem mestu zaključimo, da z modelom tega pojava ne znamo dobro pojasniti.

5.3. Validacija modela na novih podatkih

Sedaj se bomo osredotočili na drugo polovico podatkov, ki smo jo pri oblikovanju našega GLM modela v prejšnjem poglavju izpustili. V tem razdelku bomo s pomočjo te druge polovice podatkov poskusili oceniti, kako dober je naš model. Zanima nas, če je model, ki smo ga postavili v prejšnjem razdelku, dovolj splošen, da res pojasni in kar se da natančno oceni pričakovano višino škode, ali pa se preveč prilega prvi polovici podatkov in torej pričakovane višine škode v splošnem ne pojasni dobro.

5.3.1. Ocena modela

Prvi način za ocenjevanje modela je dokaj preprost. Na drugi polovici podatkov bomo zopet ocenili GLM model, ki bo imel enake pojasnjevalne spremenljivke kot naš originalni GLM. Zanimale nas bodo predvsem ocene za parametre β. Zanimalo nas bo, ali sta prametra pri isti ravni iste pojasnjevalne spremenljivke v enakem velikostnem razredu, ali imata enka predznak in ali sta v obeh primerih statistično značilna. V kolikor bomo opazili kakšna večja odstopanja, to lahko pomeni, da te spremenljivke nismo najbolje opisali, saj ima očitno kakšna konkretna vrednost na našo oceno zelo velik vpliv. Poglejmo si torej, kaj nam vrne program R, ko ocenimo GLM z enakimi pojasnjevalnimi spremenljivkami še na drugi polovici podatkov.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1674  -0.1921  -0.0545   0.1402   1.0579

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
   (Intercept)            4.566213   0.024249 188.308  < 2e-16
Tip Vozila (Dvovratno vozilo)
   Štirivratno vozilo     0.007725   0.013163   0.587 0.557329
   Športno vozilo         0.465188   0.015001  31.011  < 2e-16
   Luksuzno vozilo        1.022915   0.028498  35.894  < 2e-16
Status Zaposlitve (Nezaposlen)
   Delavno neaktiven     -0.111582   0.018052  -6.181 6.92e-10
   Zaposlen              -0.132687   0.013710  -9.678  < 2e-16
Poročni Stan (Poročen)
   Ločen                  0.012505   0.014772   0.847 0.397291
   Samski                 0.114685   0.012682   9.043  < 2e-16
Tip Naselja (Podeželje)
   Primestna              1.493863   0.014107 105.898  < 2e-16
   Mestna                 1.096104   0.016677  65.727  < 2e-16
Kritje (Osnovno)
   Razširjeno             0.213940   0.011172  19.150  < 2e-16
   Premijsko              0.471934   0.017749  26.590  < 2e-16
```

```
Velikost Vozila (Majhen)
   Medium               -0.036104   0.013314  -2.712 0.006718
   Velik                -0.074829   0.019667  -3.805 0.000144

Dispersion parameter for Gamma family taken to be 0.1145145

Deviance:  713.25  on 4553  degrees of freedom
AIC: 57512
```

Najprej lahko pogledamo oceno za razpršitveni parameter. V originalnem modelu je bila ta ocena enaka 0,110, v modelu na drugi polovici podatkov pa je ocena za razpršitveni parameter enaka 0,114. Obe oceni imata enak velikostni razred, torej tu ne opazimo posebnih razlik.

Ko pogledamo ocene za parametre, vidimo, da tudi tu ni velikih razlik med modelom iz prve polovice podatkov in modelom iz druge polovice podatkov. Ocena za višino škode pri osnovni ravni je skoraj enaka (4,569 v prvem modelu proti 4,566 v drugem). Tudi pri drugih ocenah za parametre ne opazimo večjih razlik. Vse ocene imajo pri obeh modelih isti predznak. Še največjo razliko opazimo pri oceni parametra za skupino ločenih ljudi pri pojasnjevalni spremenljivki poročni stan. V prvem modelu je ta ocena 0,0069, pri drugem modelu pa je 0,0125. Če to preračunamo v procente, to pomeni, da je v prvem modelu pričakovana višina škode za ločene 0,7% višja kot pri poročenih, v drugem modelu pa je pričakovana višina škode za ločene višja za približno 1,3%. Razlika med modeloma torej znaša 0,6% kar se nam ne zdi razlika, ki bi pomenila, da sta oceni pri obeh modelih zelo različni.

Tudi ko pogledamo statistično značilnost ocen, ne opazimo prevelikih razlik. V obeh modelih sta statistično neznačilni le oceni za parametra pri štirivratnem vozilu in ločenih osebah. Vse druge ocene pa so v obeh modelih statistično značilne. Med modeloma torej nismo opazili bistvenih razlik. Zaključimo lahko, da vse ugotovitve, do katerih smo prišli na prvi polovici podatkov, veljajo splošno.

5.3.2. **Točkovne napovedi**

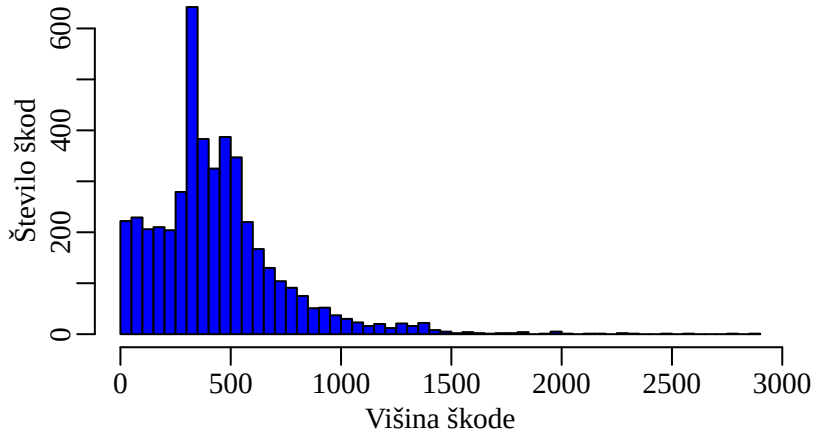
Drugi način, ki ga bomo uporabili za oceno kvalitete našega modela, je s pomočjo točkovnih napovedi. Naslednja dva razdelka sta povzeta po viru [7]. Naj bo x_0 vektor vrednosti pojasnjevalnih spremenljivk za novo enoto iz druge polovice vzorca in y_0 pripadajoča škoda. Točkovno napoved za y_0 določimo tako, da minimiziramo pričakovano razliko med dejansko in napovedano vrednostjo. To velja za $\hat{\mu}_0 = \exp(x_0^T\hat{\beta})$, kjer je $\hat{\beta}$ ocena , pridobljena s prvo polovico podatkov.. To je asimptotsko nepristranska cenilka za vrednost $\mathbb{E}(y_0|x_0)$.Velja, da ima $x_0^T\hat{\beta}$ asimptotsko normalno porazdelitev $N(x_0^T\hat{\beta}, \phi x_0^T(X^TX)^{-1}x_0)$. Za to točkovno napoved je pogosto uporabljen naslednji interval s 95% stopnjo zaupanja

$$\exp\left(x_0^T\hat{\beta} \pm 1{,}96\sqrt{\hat{\phi}x_0^T(X^TX)^{-1}x_0}\right).$$

Za bolj robustne intervale zaupanja, ki v oceno vključujejo tudi ocenjevanje $\hat{\phi}$, lahko kvantil normalne porazdelitve zamenjamo s $t_{0,975}(n-k)$. Za velike n bo razlika med obema intervaloma zaupanja majhna. Zgoraj zapisani interval zaupanja je

asimetričen okoli ocenjenega upanja. Obstaja tudi simetričen interval, ki ga bomo predstavili v nadaljevanju.

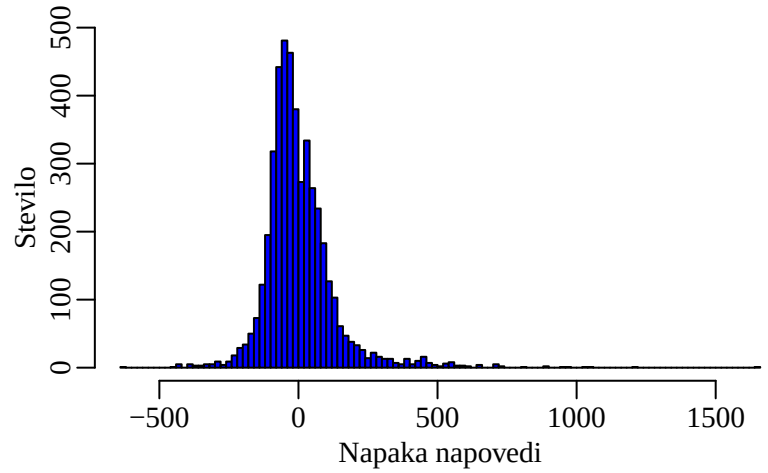
Najprej bomo narisali histogram višine škode na drugi polovici podatkov, ter le tega primerjali s tistim, ki smo ga narisali na vseh podatkih.



Slika 27. Histogram višine škod na drugi polovici podatkov.

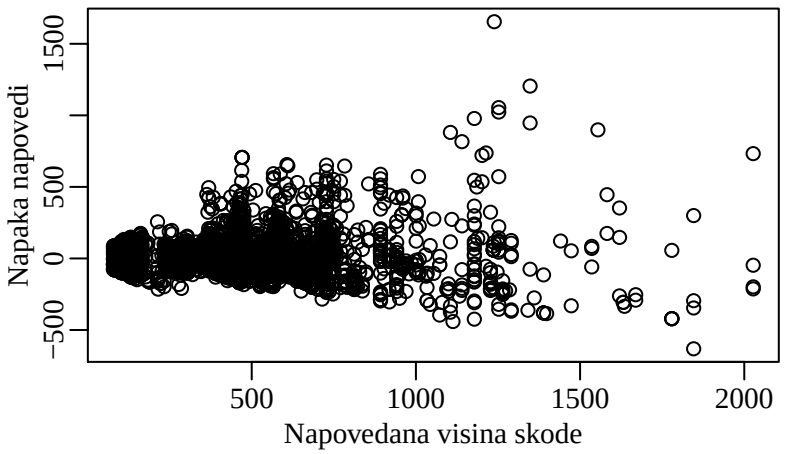
Opazimo lahko, da sta si sliki 27 in 1 zelo podobni. Na obeh slikah opazimo, da so frekvence pri nizkih škodah približno enake, nato pa se frekvenca močno poviša in nato počasi pada, ko se višina škode povečuje. Tudi frekvence, ki jih opazimo na sliki 27, so približno polovice frekvenc s slike 1.

Sedaj bomo uporabili model s prve polovice podatkov in z njim napovedali pričakovane vrednosti višine škode za enote iz druge polovice podatkov. Dobljenim vrednostim bomo rekli točkovne napovedi $\hat{\mu}_i$, razlikam med dejanskimi vrednostmi y_i in napovedanimi pa napake napovedi $y_i - \hat{\mu}_i$.



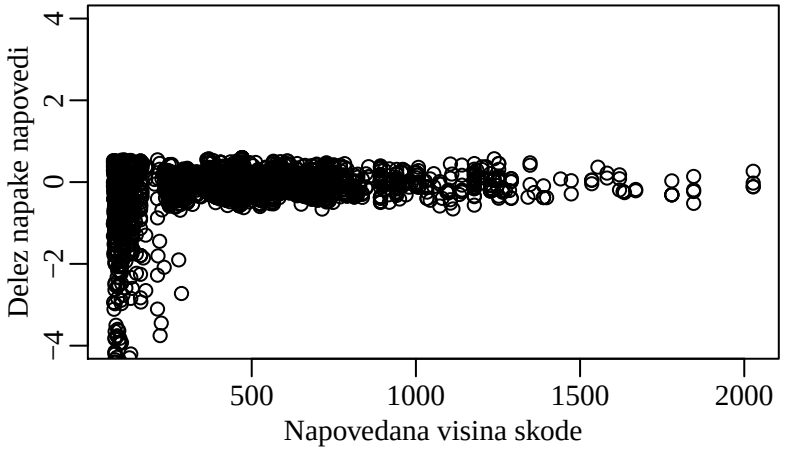
Slika 28. Histogram napak napovedi na drugi polovici podatkov.

Na sliki 28 lahko vidimo, da so napake napovedi skoncentrirane okoli vrednosti 0, kar pomeni, da model večinoma dobro napove višino škode.Opazimo, da je večina napak napovedi pod negativnih, kar pomeni, da večino škod na drugi polovici podatkov napovemo malo višje od dejanske realizacije. Poslovno je to dobro. Po drugi strani pa imamo dolg tanek rep na desni, kar pomeni, da smo nekaj škod močno podcenili. Narišimo še razsevni grafikon napovedanih vrednosti in napak napovedi.



Slika 29. Razsevni grafikon napak napovedi in višine škode.

Na sliki 29 lahko opazimo, da se napake napovedi povečujejo z višanjem škode. Na prvi pogled izgleda, kot da ima model pri ocenah za višje pričakovane škode težave, kar je ravno obratno od tega, kar smo mislili do sedaj. Opazimo pa tudi, da so za vsako višino realizacije napovedi približno enakomerno porazdeljene okoli 0, kar lahko pomeni, da v povprečju dobimo pravo oceno za pričakovano škodo. Zanima nas tudi odstotna napaka napovedi, ki jo bomo izračunali s $\frac{y_i - \hat{\mu}_i}{y_i}$. Narišimo še raszevni grafikon teh vrednosti.



Slika 30. Razsevni grafikon deležev napak napovedi in višine škode.

Slika 30 nam potrdi, da imamo težave pri ocenjevanju nizkih pričakovanih škod. Zanimivo je, da je slika 29 lahko malo zavajujoča, saj bi lahko sklepali, da imamo težave pri napovedih za višje škode. Ko izračunamo deleže napak napovedi opazimo, da se največji deleži napak pojavijo ravno pri najnižjih pričakovanih škodah.

5.4. Napovedni intervali

Točkovna napoved ocenjuje upanje višine škodnega zahtevka. Dejanski škodni zahtevek je realizacija slučajne spremenljivke, ki seveda lahko odstopa od svoje pričakovane vrednosti. Napovedni intervali napovedujejo realizacije slučajnih spremenljivk, ki še niso znane. Zgrajeni so torej tako, da poskušajo vsebovati neko neznano slučajno vrednost. Intervali zaupanja v nasprotnem primeru poskušajo vsebovati neko neznano fiksno vrednost na primer upanje te slučajne spremenljivke. Za regresijske modele so napovedni intervali zelo zanimivi, vendar pa jih pri statistiki redko obravnavamo. Izjema je klasična linearne regresija V tem razdelku bomo razvili preprosto metodo za računanje napovednih intervalov za GLM z gama porazdelitvijo, ki so dokaj natančni, njihova natančnost pa se s številom opazovanj tudi izboljšuje. Splošnih napovednih intervalov za GLM ni, v viru [7] iz leta 2016, pa so poskušali skonstruirati prav napovedne intervale za naš problem. Ta razdelek je povzet po tem viru.

Naj bo y_0 vrednost, za katero bi radi izračunali napovedni interval. Vemo, da je y_0 gama porazdeljena ter da se da opisati z enačbo $y_0 \sim \text{Gamma}(\mu_0, \phi)$, kjer je $\mu_0 = \exp(x_0'\beta)$ in x_0 vektor vnaprej poznanih pojasnjevalnih spremenljivk. Če želimo skontruirati tak interval, da bo dejanska vrednost y_0 vsebovana v intervalu v 95% primerih, moramo najprej opaziti, da velja, da je $2y_0/(\exp(x_0^T\beta)\phi)$ porazdeljena s hi-kvardat porazdelitvijo z $2/\phi$ prostostnimi stopnjami $\chi^2(2/\phi)$. Prav tako velja, da je cenilka $\hat{\mu}_0 = \exp(x_0^T\hat{\beta})$ asimptotsko normalno porazdeljena s porazdelitvijo $N(\exp(x_0^T\beta), \exp(2x_0^T\beta)x_0^T(X^TX)^{-1}x_0\phi)$. Naravno se nam ponuja naslednji simetrični interval zaupanja za $\mu_0 = \mathbb{E}(y_0|x_0)$

$$\exp(x_0^T\hat{\beta})\left(1 \pm 1{,}96\sqrt{\hat{\phi}x_0^T(X^TX)^{-1}x_0}\right),$$

kot asimptotski interval zaupanja okoli upanja s 95% stopnjo značilnosti. Iz tega sledi, da je $\hat{\mu}_0/\exp(x_0^T\beta)$ asimptotsko porazdeljen po $N(1, x_0^T(X^TX)^{-1}x_0\phi)$. Ker sta cenilki za ϕ in β asimptotsko neodvisni, je $\hat{\mu_0}/\exp(x_0^T\beta)$ približno porazdeljen po $t(n-p)\sqrt{x_0^T(X^TX)^{-1}x_0\hat{\phi}} + 1$. Vpeljimo slučajno spremenljivko

$$G = \frac{\chi^2(2/\phi)}{t(n-k)\sqrt{x_0^T(X^TX)^{-1}x_0\hat{\phi}} + 1},$$

kjer sta v številcu in imenovalcu neodvisni χ^2 in studentovo porazdeljeni slučajni spremenljivki. Potem velja

$$\text{P}\left(G_{0,025} \leq \frac{2y_0}{\exp(x_0^T\beta)\phi} : \frac{\hat{\mu}_0}{\exp(x_0^T\beta)} \leq G_{0,975}\right) =$$

$$= \text{P}\left(\frac{G_{0,025}\hat{\mu}_0\phi}{2} \leq y_0 \leq \frac{G_{0,975}\hat{\mu}_0\phi}{2}\right).$$

Zgornja verjetnost je asimptotsko enaka 0,95, saj sta $\hat{\mu}_0$ in y_0 neodvisni slučajni spremenljivki, $G_{0,025}$ in $G_{0,975}$ pa sta 2,5% in 97,5% kvantila G porazdelitve. Te kvantile dobimo s simulacijo. Tako lahko zapišemo asimptotski napovedni interval za y_0.

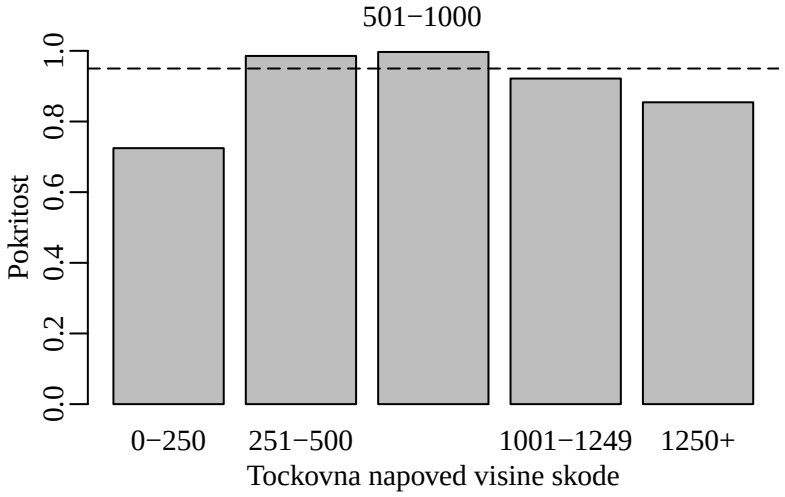
$$\left(\frac{\hat{G}_{0,025}\mu_0\hat{\phi}}{2}, \frac{\hat{G}_{0,975}\mu_0\hat{\phi}}{2} \right),$$

kjer je $\hat{G}$ enak kot G, kjer smo ϕ zamenjali z $\hat{\phi}$.

Kot že omenjeno zgoraj, bomo v programu R za vsako enoto druge polovice podatkov simulirali porazdelitev G, nato bomo izračunali napovedni interval s 95% stopnjo zaupanja ter na koncu pogledali, če je dejanska višina škode vsebovana v tem intervalu. Pri določanju napovednega intervala bomo uporabili ocene parametrov, dobljene na prvi polovici podatkov.

V programu R smo naredili funkcijo, ki za vsako enoto izračuna napovedni interval s 95% stopnjo zaupanja. Nato smo preverili, če je dejanska višina škode vsebovana v tem napovednem intervalu. Če je bila vsebovana, smo to enoto prišteli k številu enot, ki so bile že vsebovane v svojem napovednem intervalu. Vseh enot na polovici podatkov, ki smo jo obravnavali, je bilo 4567. Število enot, kjer smo imeli škodo, ki je padla v napovedni interval je bilo 4195 oz 91,9% vseh vrednosti, kar je manj od 95%, ki smo jih pričakovali.

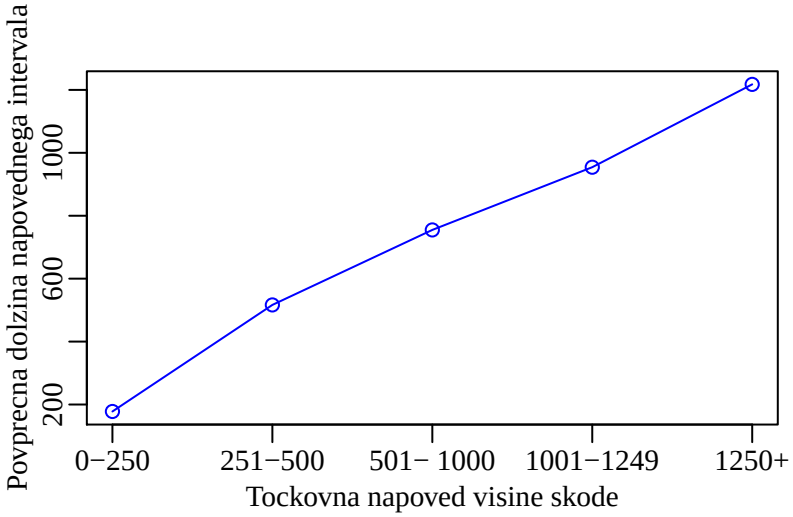
Zgornji rezultat pomeni, da končni model verjetno ne razloži čisto vseh vplivov na višino škodnega zahtevka. Tak rezultat smo pričakovali, potem ko smo videli, kako izgledajonapake napovedi. Na žalost smo delali s takšnimi podatki, kjer za nižje škode obstaja nek vpliv, ki ga iz samih podatkov ne znamo razložiti. Za ta namen bomo narisali še graf, kjer bomo pokritost prikazali po posameznih intervalih višine škode.



Slika 31. Delež škod vsebovanih v intervalu napovedi po točkovni napovedi višine škode.

Na sliki 31 se potrdi naša teza, saj so napovedni intervali za nizke škode manj natančni. Opazimo tudi, da se tudi pri višjih škodah napovedni intervali vedno manj natančni, vendar pa odstotek ne pada hitro

Po istih kategorijah višin škod bomo narisali še dolžine napovednih intervalov. Tako bomo ugotovili, kako uporabni ti intervali sploh so. Če so napovedni intervali zelo široki, nam o sami vrednosti škode ne povedo veliko.



Slika 32. Povprečna širina napovednega intervala po višini škode.

Na sliki 32 torej vidimo, da se dolžine napovednih intervalov z višino škode povečujejo. Na splošno so dolžine intervalov kar velike, kar pomeni, da sam napovedni interval ni nekaj, na kar bi se lahko zanašali, in ga je torej potrebno obravnavati skupaj z ostalimi metodami za ocenjevanje modela. Tudi rezultat iz slike 31 si lahko razložimo s tem, da so napovedni intervali za nižje škode najožji, s tem pa je torej verjetnost, da višina škode pade izven napovednega intervala, višja.

Kljub temu pa je rezultat, da je skoraj 92% vseh enot znotraj napovednega intervala pomeni, da model nekatere vplive na višino škode dobro razloži. Poleg tega model ocenjuje malo višje pričakovane vrednosti škod, kot imamo podakte, kar vsebinsko ni napačno. Zavarovalnice želijo višine škod modelirati rajše malo višje kot nižje, saj to pomeni, da so bolje pripravljene na kakšne večje in nepričakovane škodne dogodke.

Zaključimo lahko, da model ni popoln, hkrati pa nam je dal kar nekaj vpogleda v podatke in pričakovano višino škode, prav to pa je glavni namen regresijskih modelov.

6. Uporaba GLM v Slovenskih zavarovalnicah

Omenili smo že, da zavarovalnice po Evropi in tudi izven nje uporabljajo GLM za ocenjevanje splošne kvalitete zavarovalnega portfelja. Ob tem se je pojavilo naravno vprašanje, kako je z uporabo GLM-a v zavarovalnicah v Sloveniji.

Konec leta 2016 je na slovenskem trgu poslovalo štirinajst zavarovalnic. Štiri izmed njih se ukvarjajo izključno s premoženjskimi zavarovanji, tri pa so takšne, da pretežno opravljajo posle življenjskega zavarovanja. Ostale zavarovalnice opravljajo tako posle premoženjskega kot življenjskega zavarovanja ([9], seznam zavarovalnih subjektov).

Vsem tem zavarovalnicam smo poslali elektronsko sporočilo, v katerem smo jih povprašali o uporabi GLM modelov za modeliranje raznih spremenljivk. Poslana so bila naslednja vprašanja:

(1) Ali v zavarovalnici na kakršenkoli način uporabljate GLM metodo? Kako dolgo že? Če je odgovor da, me zanima še, zakaj ste se odločili za uporabo te metode?
(2) Prosil bi vas, če vsaj približno opišete, pri čem uporabljate GLM. Ali uporabljate GLM pri življenjskih ali neživljenjskih zavarovanjih? Ali s to metodo modelirate škode, exposure, premijo, popuste ali pa kaj drugega?
(3) Kako ste implementirali GLM? Ste se ga naučili sami iz knjig, ste za to najeli zunanje svetovalce, ali pa bili na kakšnem izobraževanju?
(4) Se vam zdi GLM dobra metoda za modeliranje? Je GLM pokazal izboljšanje v primerjavi z regresijskim modelom, ki ste ga uporabljali prej?

Nazaj smo prejeli odgovore s strani sedmih zavarovalnic.
Iz ene izmed večjih zavarovalnic v Sloveniji so nam odgovorili, da GLM uporabljajo že od leta 2007 dalje. Za GLM so se odločili, ker je GLM za modeliranje zavarovalniških predpostavk bolj primeren kot navadna linearna regresija ali pa t.i. enostranske (ang. one-way) analize. Sam GLM uporabljajo za določanje višine premije (modelirajo višino popustov in provizij, ter določajo pravilno segmentacijo rizikov), ter obnovljivost zavarovanj. Modelirajo torej povprečne škode, škodne pogostosti, škodne rezultate ipd. z namenom določanja višine premije na avtomobilskih zavarovanjih. GLM uporabljajo tudi za tržne namene, torej za pospeševanje prodaje. O GLM-ju so se naučili iz knjig in interneta, obiskali pa so tudi kakšen seminar v tujini. Trdijo, da je GLM izboljšal modele, ki jih uporabljajo, hkrati pa so omenili, da že preizkušajo kakšne še naprednejše načine modeliranja. Modela so se naučili sami s pomočjo orodja SAS ter literature. GLM vidijo bolj kot orodje za kvalitativno, kot kvantitativno analizo. Zdi se jim, da GLM ponuja večjo fleksibilnost pri modeliranju, kot jim jo je ponujal star način, in prav zato se jim zdi zelo uporaben.

Tudi v sektorju življenjskih zavarovanj že uporabljajo GLM. Uporabljati so ga začeli leta 2016 z namenom, da bi izboljšali razlago trajnosti portfelja življenjskih zavarovanj, kar so do nedavnega počeli z uporabo Coxovega modela. GLM metodo uporabljajo pri analizi trajnosti polic življenjskega zavarovanja. Verjetnost, da polica preneha s plačevanjem premije (kapitalizacija police ali odstop od zavarovanja), skušajo modelirati s pomočjo nekaterih karakteristik police (spol zavarovanca, zavarovalna vrsta, prodajna pot, starostna skupina, regija ...), za kar uporabljajo GLM. S to metodo modelirajo tudi število odstopov glede na izpostavljenost, ki se meri v enoti polica-leto.

Odgovor smo dobili tudi od ene izmed večjih tujih zavarovalnic, ki posluje v Sloveniji. Povedali so, da GLM uporabljajo približno eno leto. GLM je že ustaljena praksa pri njihovi zavarovalnici v tujini, od koder je prišlo tudi navodilo, da ga začnejo uporabljati tudi v Soveniji. GLM uporabljajo za modeliranje frekvence in višine škod. V prihodnosti planirajo tudi modeliranje frekvence obnov. Iz GLM-a dobijo faktor rizičnosti, ki naprej določa višino premije. Za modeliranje uporabljajo orodje Emblem, ki ga je razvilo podjetje Tower Watson. Zdi se jim, da je, kljub uporabi orodja Emblem, potrebno poznati teorijo, ki se skriva v ozadju GLM-ev. GLM-a so se naučili sami iz literature, zavarovalniška skupina pa je organizirala

tudi nekaj predavanj na to temo. Če potrebujejo pomoč imajo na voljo sodelavce iz tujine, na katere se lahko obrnejo. GLM se jim zdi dobra izbira, saj presega okvire, ki jih postavlja linearne regresija.

Od preostalih petih zavarovalnic smo prejeli odgovor, da GLM-ja še ne uporabljajo.

Lahko zaključimo, da manjši procent zavarovalnic že uporablja GLM pri svojem delu, njegovo uporabo pa do se naučile samoiniciativno. Glede na te rezultate lahko sklepamo, da je uporaba GLM-a na slovenskem zavarovalniškem trgu prednost, saj večina igralcev na tem trgu metode GLM še ne pozna ali pa je ne uporablja.

7. Zaključek

Kaj lahko na koncu povemo o GLM-u? V primerjavi z linearno regresijo, ki je trenutno najbolj poznana in najbolj uporabljena regresijska metoda, ponuja predvsem večjo fleksibilnost pri prilagajanju podatkom, s tem pa omogoča izboljšane napovedi. Omejitve linearne regresije na normalno porazdeljeno odvisno spremenljivko ter aditivni način modeliranja s pojasnjevalnimi spremenljivkami, preseže s pomočjo uporabe mnogih različnih povezovalnih funkcij in porazdelitev. Ta širina je glavna prednost pri uporabi GLM-a

V zavarovalništvu se potreba po GLM modelu pojavi predvsem zato, ker večina podatkov, iz katerih bi zavarovalnice rade napovedovale prihodnja gibanja, ni porazdeljenih normalno. Višina škod je po navadi porazdeljena z gama porazdelitvijo, kjer je večina vrednosti blizu ničle, rep pa je dolg in debel. Po drugi strani je frekvenca škod porazdeljena Poissonovo, kjer modeliramo naravna števila. Na posamezni polici se je škoda se je bodisi zgodila, ali pa je ni bilo, kar nakazuje Bernoulli porazdeljeno slučajno spremenljivko. Zaradi zelo različnih potreb pri modeliranju je GLM popoln za uporabo v zavarovalnicah.

Kljub temu, da je GLM model bolj kompleksen kot linearna regresija, je še vseeno dovolj preprost, da tudi nepoznavalcu statistike ni preveč težko razumeti, kaj povedo parametri pri pojasnjevalnih spremenljivkah. V nalogi smo naredili primer GLM-a na podatkih višine škod. Pri modeliranju GLM-a smo si pomagali s programskim jezikom R, ki omogoča hitro delo in lahek dostop do diagnostičnih vrednosti, s katerimi ovrednotimo naš model. Prednost GLM-a je hkrati tudi njegova slabost, saj zaradi pestre izbire različnih možnosti pri izbiri povezovalne funkcije in porazdelitve, pa tudi zaradi številnih možnosti vključevanja in izključevanja pojasnjevalnih spremenljivk v model, lahko hitro zaidemo s prave poti. To nam prinese dodatno delo, saj je lahko potrebno veliko število iteracij modela, da pridemo do izboljšanja napovedi.

Čeprav smo imeli na voljo podatke, ki smo jih težko modelirali, smo se veliko naučili o modeliranju z GLM. Opazili smo, da je veliko število izbir, ki jih imamo na voljo pri modeliranju, lahko tudi breme, saj pogosto ne vemo, za kaj se je najbolje odločiti. Sumimo, da je v ozadju težav pri modeliranju z našimi podatki nek vsebinski razlog, ki ga zaradi skopega vsebinskega opisa podatkov nismo mogli določiti in torej izločiti. Vseeno smo s konstruiranjem napovednih intervalov, ter ocenjevanjem modela na drugi polovici podatkov pokazali, da so ocene pričakovane višine škode

dokaj dobre, zaključili pa smo lahko, da je za zavarovalnico bolje, da pričakovano višino škode modelira malo višje, kot je v resnici, saj tako poskrbi za bolj konzervativno oceno prihodnjh izgub. Iz poslovnega vidika torej naš model niti ni tako zelo slab.

Za konec smo nato še ugotovili, da je v slovenskih zavarovalnicah uporaba GLM-a še v povojih. Zaradi tega lahko GLM trenutno prestavlja veliko prednost za tiste zavarovalnice, ki bodo vložile trud in znanje v ta način modeliranja, saj bodo tako korak pred konkurenco. Vsaj do takrat, ko bodo zavarovalnice začele z bolj množično uporabo tega modela, kar pričakujemo, da se bo zgodilo kmalu v prihodnosti.

Literatura

[1] A. Agresti, *Foundations of linear and generalized linear models*, John Wiley & Sons,2015

[2] D. Anderson et all, *A practicioner's guide to generalized linear models*, Tower watson, 2007

[3] T. Amemiya, *Advanced Econometrics*, Harvard University Press, Cambridge, Massachusetts, 1985

[4] F. Hayashi, *Econometrics*, Princeton University Press, Princeton, New Jersey, 2000

[5] G. G. Roussas, *A Course in Mathematical Statistics*, Academic Press, Davis, California, 1997

[6] J. A. Nelder, R. W. M. Wedderburn, *Journal of the Royal Statistical Society. Series A (General)*, Vol. 135, No. 3 (1972), 370-384

[7] M. W. Hattab, *A derivation of prediction intervals for gamma regression*, Journal of Statistical Computation and Simulation 86 (2016), 3512-3526

[8] Sample data sets, [ogled 3.7.2017], dostopno na
`http://www.scan-support.com/help/sample-data-sets`

[9] Spletna stran AZN, [ogled 12.10.2016], dostopno na
`https://www.a-zn.si/`