Univerza v Ljubljani

Interdisciplinarni študijski program Uporabna statistika

Maruša Kustec

Primerjava porazdelitev skupin pri podatkih z visokim deležem nakopičenih vrednosti

Magistrsko delo

Mentor: izr. prof. dr. Lara Lusa

Ljubljana, 2016

Zahvala

Magistrsko delo je zaključno delo študijskega programa Uporabna statistika. Kot del prve generacije študentov se zahvaljujem vsem profesorjem in ostalim sodelavcem, ki so omogočili nastanek programa in se vseskozi trudili za njegovo dobro izvedbo. Študij mi je omogočil tudi pridobitev praktičnih izkušenj in konkurenčnost v tujini.

Kot mikrobiologinja se še posebno zahvaljujem profesorjem, ki so me naučili ključnih veščin obdelave bioloških podatkov v sklopu biostatističnega modula študija. Hvaležna sem doc. dr. Damijani Kastelec in prof. dr. Katarini Košmelj iz Biotehniške fakultete, ki sta mi omogočili, da sem se od njiju učila podajati znanje pri vajah statistike in utrdila osnove le-te.

Pri nastajanju magistrskega dela se zahvaljujem za pomoč mentorici prof. dr. Lari Lusa, od katere sem se ogromno naučila že v času študija. Doc. dr. Nataši Kejžar se zahvaljujem za pomoč in pomembne popravke besedila. Prof. dr. Maji Pohar Perme se zahvaljujem za izvrstno podajanje znanja v času študija in za ostalo podporo.

Hvaležna sem tudi svojim bližnjim. Sestri Irena in Urška sta mi predstavljali dober zgled na celotni študijski poti. Staršem se zahvaljujem za vsesplošno podporo, zaupanje in potrpežljivost. Fantu Maticu se zahvaljujem za razumevanje in ljubezen, Emi in Kaji pa za tolažbo v stresnih trenutkih.

Vsebina

Seznam slik

Seznam tabel

Notacija in oznake

V magistrskem delu so uporabljene naslednja notacija in oznake:

	Notacija
X	pojasnjevalna spremenljivka
X_i	i-ta komponenta pojasnjevalne spremenljivke
x_i	i-ta komponenta opazovane vrednosti spremenljivke X
Y	odvisna spremenljivka, izid
Y_i	i-ta komponenta odvisne spremenljivke, izida
y_i	i-ta komponenta opazovane vrednosti izida Y
n	število statističnih enot
P	verjetnost
α, β	komponenti modela
$\hat{\alpha}$, $\hat{\beta}$	ocenjeni komponenti modela α in β
ϵ	neodvisna napaka v modelu
j	vrednost meje med kategorijami izida Y v modelu sorazmernih obetov
SE	standardna napaka

Tabela 1: Notacija in oznake

Natančnejši pomen oznak je pojasnjen v spremljajočem besedilu, kjer je oznaka uporabljena.

Povzetek

V tej magistrski nalogi preučujemo pristope za analizo posebne vrste podatkov, ki se pogosto pojavijo pri raziskavah v genomiki. Primer takšnih podatkov so obravnavani podatki o virusih, kjer nas zanima primerjava koncentracije posameznega virusa v vzorcu med skupinami otrok, ki so bili predhodno različno diagnosticirani. Spremenljivke, ki opisujejo koncentracijo virusov, imajo del vrednosti nakopičen pri eni točki, kar predstavlja problem pri analizi podatkov. Takšne spremenljivke poimenujemo spremenljivke z nakopičenimi vrednostmi.

Za ugotavljanje povezanosti skupin otrok s koncentracijo virusa, smo izbrali štiri metode: model sorazmernih obetov, model Tobit, kombiniran pristop logistične in linearne regresije (model Log+Lin) in Mann-Whitneyev test. Prve tri metode omogočajo vključitev dodatnih pojasnjevalnih spremenljivk v analizo. Spremenljivka z nakopičenimi vrednostmi je obravnavana kot odvisna spremenljivka.

S simulacijami smo preučevali delovanje izbranih metod v različnih situacijah. Izkaže se, da imata velik vpliv na delovanje metod skladnost razlik in sorazmernost v podatkih. Model sorazmernih obetov, model Tobit in Mann-Whitneyev test imajo primerljive moči v večini situacij, le model Tobit pa ohrani ustrezno velikost testa v vseh situacijah. Edini obravnavani dvodelni pristop, model Log+Lin, ima bistveno prednost pred omenjenimi enodelnimi pristopi ob prisotnosti neskladnih razlik in nesorazmerij. Ker v podatkih o virusih pričakujemo oboje, dvodelni pristop prepoznamo kot najbolj primeren pristop za analizo.

Dodatno preučimo še delovanje testa, ki preverja veljavnost predpostavke o

sorazmernih obetih. Test je anti-konzervativen in ima majhno moč pri majhnem vzorcu.

Ključne besede: spremenljivke z nakopičenimi vrednostmi, primerjava skupin, model sorazmernih obetov, model Tobit, Mann-Whitneyev test, kombiniran pristop logistične in linearne regresije, sorazmerni obeti, skladnost razlik

Abstract

This thesis evaluates the literature suggested data analysis approaches for a special type of data that arises in genomic experiments. In our data, we are interested in comparing the viral concentration of four groups of children. Variables that describe viral concentrations have a high proportion of values clumped at one point, which poses a problem when analysing this data. We denote such variables as variables with a point-mass.

To analyse the relationship between groups and viral concentrations we used four methods: the proportional odds model, the Tobit model, a combination of logistic regression and a linear model (Log+Lin model), and the Mann-Whitney test. The first three methods allow us to include additional explanatory variables in the analysis. Variables with a point-mass are handled as outcome variables.

We assessed the performance of the selected methods in different situations through a series of simulation studies. Our results indicate that consonant and dissonant effects, as well as the property of proportionality in the data, have a big impact on method performance. The proportional odds model, the Tobit model and the Mann-Whitney test have comparable power to identify differences between groups in most situations, but only the Tobit model keeps an adequate type I error in all situations. The only considered two-part test in the study, the Log+Lin model, has an advantage over the mentioned one-part tests when dissonant differences and nonproportionality are present in the data. Because we expect to find both properties in our data, we recommend the use of the two-part approach.

Additionally, we assessed the performance of a test that validates the proportional odds assumption. The test proved to be anti-conservative and loses power when the sample size is small.

Key words: variables with a point mass, group comparison, proportional odds model, Tobit model, Mann-Whitney test, a combination of logistic regression and a linear model, proportional odds, consonant and dissonant effects

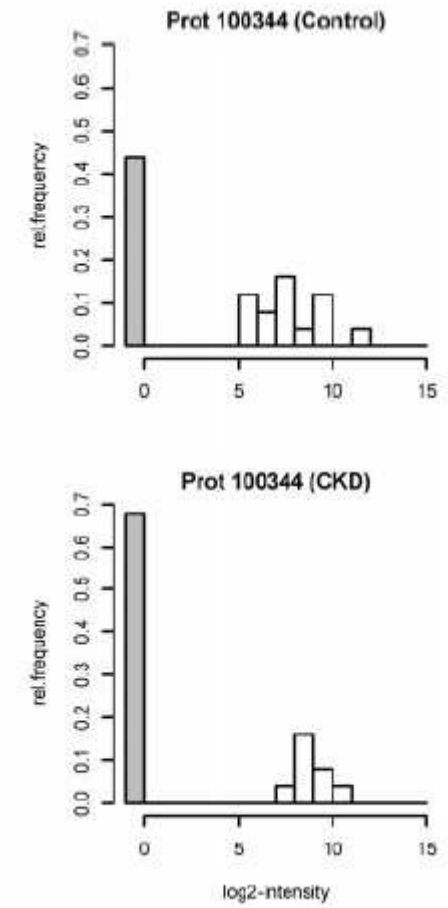
1 Uvod

V molekularni biologiji in genetiki, raziskovalci pogosto naletijo na spremenljivke, kjer je en del vrednosti porazdeljen po zvezni porazdelitvi, drugi del vrednosti pa je nakopičen pri eni točki (največkrat pri nuli, ni pa nujno) [2]. Delež nakopičenih vrednosti je lahko zelo velik, kar v nadaljevanju predstavlja težave pri analizi podatkov. Omenjene spremenljivke so v literaturi poimenovane spremenljivke z nakopičenimi vrednostmi (ang. point mass values, inflated intensity values) [1]. Primer takšne spremenljivke je prikazan na sliki 1.1.

Spremenljivke z nakopičenimi vrednostmi so značilne za podatke iz raziskav v genomiki, kjer merijo koncentracijo nukleinskih kislin, proteinov ali metabolitov. Pojavijo se tudi pri poizkusih z uporabo DNA mikromrež, kjer merijo izražanje genov, v proteomiki pri eksperimentih, ki uporabljajo masno spektrometrijo in v študijah, kjer preučujejo DNA metilacijo [3].

Do pojava nakopičenih vrednosti v eksperimentih, kjer merijo koncentracijo neke spojine, lahko pride zaradi dveh razlogov. Lahko gre za resnično odsotnost spojine v vzorcu (biološke nakopičene vrednosti, ang. biological point mass values) ali pa je vsebnost spojine v vzorcu tako majhna, da je ta pod mejo detekcije (tehnične nakopičene vrednosti, ang. technical point mass values) [1]. Za biološke nakopičene vrednosti je vrednost spremenljivke zares enaka 0, v primeru tehničnih nakopičenih vrednosti pa prave vrednosti spremenljivke ne poznamo - je večja od 0 in manjša od meje detekcije [1]. Za tehnične nakopičene vrednosti se v literaturi pojavi tudi izraz "nondetects"[4] in izraz "truncated values"[3]. Za biološke nakopičene vrednosti se v literaturi pojavi tudi izraz "true zeros"[3].

Glede na naravo nakopičenih vrednosti v spremenljivki, so določene statistične metode bolj oz. manj primerne [1].



Slika 1.1: Primer spremenljivke z nakopičenimi vrednostmi, prikaz za dve eksperimentalni skupini [1].

V podatkih iz številnih bioloških eksperimentov so hkrati prisotne tako biološke, kot tehnične nakopičene vrednosti, kar pa še oteži obravnavo je, da raziskovalci ne vejo za katere nakopičene vrednosti gre. Ne glede na vrsto nakopičenih vrednosti, pa velja, da oba dela spremenljivke, tako zvezni del, kot tudi del nakopičenih vrednosti, vsebujeta neko informacijo in zato moramo pri statistični analizi upoštevati oba dela [1].

V magistrski nalogi bomo za primer podatkov predstavili možne pristope za primerjavo porazdelitev z nakopičenimi vrednostmi med skupinami in preko simulacij preučili njihovo delovanje. Spremenljivko z nakopičenimi vrednostmi bomo obravnavali v vlogi odvisne spremenljivke, skupine pa kot opisno pojasnjevalno spremenljivko. Osredotočili se bomo na metode, ki omogočajo vključitev dodatnih pojasnjevalnih spremenljivk v analizo, vendar bomo v magistrski nalogi ostali v okviru analize z eno pojasnjevalno spremenljivko.

1.1 Predstavitev podatkov

V nalogi bomo ponovno analizirali podatke, ki so bili zbrani na Kliniki za infekcijske bolezni in vročinska stanja Univerzitetnega kliničnega centra Ljubljana, z namenom raziskave povezanosti določenih virusnih okužb s pojavom vročinskih krčev pri otrocih.

Vročinski krči so krči pri otroku, ki se pojavijo ob vročini, niso posledica okužbe osrednjega živčevja in ne ustrezajo drugim akutnim simptomatskim krčem. Pojavijo se pri 2,5% vseh otrok. Delimo jih na enostavne in kompleksne, kjer enostavni vročinski krči trajajo manj časa in prenehajo spontano, medtem ko kompleksni trajajo dalj časa, spremljajo pa jih nevrološki znaki. Vročinski krči so najpogostejši vzrok vročinskega epileptičnega statusa pri otrocih [5].

Najpogostejši vzrok vročine pri otrocih z vročinskimi krči so okužbe z virusi. Če bi uspeli povezati pojav vročinskih krčev z določenimi virusi, bi s preprečevanjem teh virusnih okužb pri dovzetnih otrocih lahko pojavnost vročinskih krčev pomembno zmanjšali [5].

Vzorci so bili pridobljeni z brisom nosno-žrelnega prostora (za preučevanje virusov dihal) in z odvzemom blata (za preučevanje virusov prebavil) [5]. V magistrski nalogi obravnavamo samo viruse, ki so bili pridobljeni z brisom nosno-žrelnega prostora in povzročajo okužbo dihal: respiratorni sincicijski virus (RSV), človeški metapneumovirus (hMPV), virus gripe (InfV), človeški koronavirus (HCoV), človeški bokavirus (HBoV), človeški rinovirus (hRV), virus parainfluence (PIV) in adenovirus (AdV).

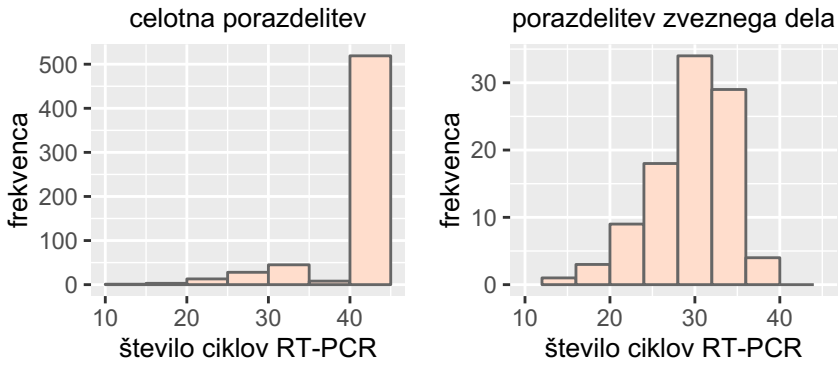
Za detekcijo virusov je bila uporabljena metoda verižne reakcije s polimerazo v realnem času (ang. real time polymerase chain reaction, okr. RT-PCR), kjer s pomnoževanjem nukleinske kisline virusa privedemo do takšne koncentracije, ki jo lahko zaznamo in tako potrdimo prisotnost virusa v vzorcu. Več ciklov RT-PCR kot potrebujemo, da zaznamo virus, manjša je njegova koncentracija v vzorcu. Če po 40 ciklih še vedno ni moč zaznati virusa, potem zaključimo, da virus v

vzorcu ni prisoten in da posameznik z virusom ni okužen.

V raziskavo so bile poleg otrok z vročinskimi krči (FS) vključene zdrave kontrole (C) [5]. Sočasno so bili zbrani še podatki za dve skupini otrok: otroci z akutnim bronhiolitisom (AB) in akutnim gastroenteritisom (AGE), ki jih bomo v nalogi prav tako vključili v analizo. Skupino otrok opisuje spremenljivka *diagnoza*.

1.2 Opredelitev problema

Koncentracijo virusov v podatkih opiše spremenljivka *ct*, ki meri število ciklov RT-PCR, potrebnih za zaznavo virusa v vzorcu. Vrednosti od 0-40 označujejo pozitivne posameznike, vrednost 41 pa označuje negativne posameznike. Vrednost 41 predstavlja nakopičene vrednosti. Za vse viruse je delež nakopičenih vrednosti v spremenljivki zelo velik (večji od 0,8), medtem ko je pozitivnih posameznikov relativno malo. Porazdelitev spremenljivke *ct* za enega od obravnavanih virusov je prikazana na sliki 1.2.



Slika 1.2: Porazdelitev števila ciklov RT-PCR za človeški rinovirus (hRV): prikaz vseh vrednosti (levo) in prikaz vrednosti za pozitivne posameznike (desno).

V preteklih študijah so že dokazali viruse pri otrocih z vročinskimi krči [6, 7, 8, 9, 10, 11]. Danes nam nove molekularne metode omogočajo večji vpogled v vlogo virusov pri pojavu vročinskih krčev, vendar pa zaradi večje občutljivosti

metod, viruse dokažemo tudi pri otrocih, ki nimajo bolezenskih znakov. Za ugotavljanje povezanosti okuženosti z virusi s pojavom vročinskih krčev je zato ključna primerjava s skupino zdravih otrok [5].

Med skupinami otrok (spremenljivka *diagnoza* z možnimi vrednostmi: FS, C, AB, AGE) primerjamo okuženost z virusom preko razlike v porazdelitvi spremenljivke *ct*. Razliko lahko izrazimo na več načinov; kot razliko med deležem nakopičenih vrednosti, razliko med povprečji zveznih delov spremenljivke ali oboje hkrati [3].

Standardne statistične metode se osredotočajo na eno vrsto obravnave (primerjava deležev nakopičenih vrednosti ali primerjava porazdelitev) in ne uspejo pravilno opisati razlike med eksperimentalnimi skupinami. Npr. binomski test proučuje razliko v deležih nakopičenih vrednosti med eksperimentalnima skupinama, vendar ne uporabi informacije o vrednostih v zveznem delu. V nasprotju pa testi, ki ugotavljajo razliko med povprečji, obravnavajo nakopičene vrednosti kot del celotne porazdelitve, kar večinoma krši predpostavko o porazdelitvi testne statistike (predpostavka o normalnosti v primeru testa t) in ob večjem deležu nakopičenih vrednosti vodijo do napačnih rezultatov [3].

Testi, ki ne zahtevajo specifične porazdelitve (neparametrični testi), so primerni, vendar ker ne upoštevajo ločenih prispevkov posameznih delov k razliki, je lahko ne zaznajo. To je problematično v primeru neskladnih razlik (ang. dissonant effect), ko je opažena obratna smer vpliva v posameznih delih porazdelitve (npr. večji delež nakopičenih vrednosti pri nuli in večje vrednosti v zveznem delu skupine, v primerjavi z drugo skupino) [3]. V primeru skladnih razlik (ang. consonant effect) neparametrični testi delujejo dobro [12]. Med njimi se za primerjavo dveh skupin pogosto uporablja Mann-Whitneyev test [1, 4].

Specifične metode za primerjavo spremenljivk z nakopičenimi vrednostmi delimo na enodelne in dvodelne. Enodelne metode obravnavajo nakopičene vrednosti kot del celotne porazdelitve in so bolj primerne, kadar so prisotne tehnične nakopičene vrednosti [1]. Nakopičene vrednosti obravnavajo kot krnjeni del po-

razdelitve. V literaturi so za analizo uporabljene različne prilagoditve testa t (npr. imputacija nakopičenih vrednosti pred uporabo testa t) ali model Tobit (parametrični model, ki predpostavlja levo-krnjeno normalno porazdelitev v ozadju)[1, 4].

Dvodelne metode posebej obravnavajo nakopičene vrednosti in zvezni del porazdelitve. Vsak obravnavani del prispeva k skupni testni statistiki, eden meri razlike v deležu nakopičenih vrednosti med skupinami, drugi pa gleda razlike v zveznem delu skupin. Dvodelni testi so bolj primerni za spremenljivke, kjer so prisotne biološke nakopičene vrednosti [1]. Na odločitev o izbiri testa vpliva tudi porazdelitev zveznega dela. Tako za ugotavljanje razlik v zveznih delih uporabimo parametričen oz. neparametričen test. Glede na dobro delovanje in enostavnost se je izkazal dvodelni test, ki kombinira binomski test za primerjavo deležev nakopičenih vrednosti med dvema skupinama, in test t (ali njegova neparametrična različica) za primerjavo zveznih delov [1, 3].

V raziskavi so poleg števila ciklov RT-PCR in diagnoze beležili še druge podatke o vključenih posameznikih (npr. starost in spol), ki prav tako lahko vplivajo na koncentracijo virusa. Čeprav se v magistrski nalogi ne bomo osredotočili na multiplo regresijsko analizo, nas bo bolj zanimalo delovanje metod, ki to omogočajo. Spremenljivko števila ciklov RT-PCR bomo obravnavali kot odvisno spremenljivko, spremenljivka diagnoze pa bo nastopala v vlogi pojasnjevalne spremenljivke.

Za analizo vpliva več pojasnjevalnih spremenljivk na odvisno spremenljivko je pogosto uporabljena metoda linearne regresije. Vendar je uporaba modela linearne regresije utežena, ko imamo odvisno spremenljivko z visokim deležem nakopičenih vrednosti, saj njena porazdelitev ni blizu nobeni bolj znani porazdelitvi (npr. normalni ali lognormalni) [2, 13]. Možna rešitev bi bila transformacija odvisne spremenljivke pred prileganjem linearnega modela, vendar ta ni nujno primerna ob prisotnosti ničelnih opazovanj (primer logaritemske transformacije) [13, 14]. Prav tako je pristop težaven zaradi kasnejše povratne transformacije ocenjenih koeficientov, ki so zaradi aproksimacije lahko pristranski. Iz praktičnega

vidika je tako bolj zaželjena uporaba primernega modela, pri katerem transformacija odvisne spremenljivke ni potrebna [14].

2 Predstavitev izbranih pristopov za analizo podatkov

Številni pristopi so bili predlagani za analizo podatkov ob prisotnosti nakopičenih vrednosti v odvisni spremenljivki. V tem poglavju bomo predstavili pristope, na katere se bomo osredotočili v nalogi.

2.1 Izbrani pristopi

V nalogi preučujemo lastnosti modela sorazmernih obetov, modela Tobit, dvodelnega pristopa, ki kombinira logistično in linearno regresijo in Mann-Whitneyevega testa za primerjavo dveh skupin ob prisotnosti nakopičenih vrednosti. Izbrane metode se razlikujejo v predpostavkah in značilnostih.

Ob prisotnosti nakopičenih vrednosti v odvisni spremenljivki Chang in Pocock [2] predlagata uporabo modela sorazmernih obetov (ang. proportional odds model), pri katerem odvisno spremenljivko obravnavamo kot kategorično, nakopičene vrednosti pa predstavljajo svojo kategorijo. Model sorazmernih obetov je uporaben zaradi svoje preprostosti in ugodne interpretacije [15], vendar je njegovo delovanje vprašljivo, ko ni izpolnjena ključna predpostavka modela - predpostavka o sorazmernih obetih. V tem primeru se priporoča uporaba drugih pristopov [2].

V primeru prisotnosti tehničnih nakopičenih vrednosti v odvisni spremenljivki, Zhang s sodelavci [4] predlaga uporabo modela Tobit, ki se uporablja za analizo podatkov ob prisotnosti krnjenih opazovanj. Model se pogosto uporablja

v ekonomiji [16]. V primeru bioloških nakopičenih vrednosti Chang in Pocock [2] priporočata dvodelni pristop - kombinacijo logistične in linearne regresije, kjer z logistično regresijo modeliramo verjetnost nakopičene vrednosti v odvisni spremenljivki, z linearno regresijo pa obravnavamo zvezni del odvisne spremenljivke [2].

Kot alternativo regresijskim pristopom bomo preverili še delovanje Mann-Whitneyevega testa, ki je zaradi enostavnosti in odsotnosti predpostavk o porazdelitvi pogosto uporabljen pri univariatni obravnavi spremenljivke z nakopičenimi vrednostmi. Poročajo, da deluje enako dobro v primerjavi z bolj kompleksnimi metodami [1, 4].

V naslednjih podpoglavjih je podano teoretično ozadje izbranim štirim metodam. Metode predstavimo za primer prisotnosti ene pojasnjevalne spremenljivke. Pred modelom sorazmernih obetov predstavljamo model logistične regresije, ki je njegova osnova. V nadaljevanju opišemo pristop, ki kombinira logistično in linearno regresijo in na koncu še Mann-Whitneyev test in model Tobit.

2.2 Model logistične regresije

Recimo, da v vsaki ponovitvi poskusa lahko dosežemo enega od dveh izidov, ki jih označimo z 1 (uspeh ali dogodek) in 0 (neuspeh ali nedogodek). V vsaki ponovitvi predpostavljamo, da je ena ali več pojasnjevalnih spremenljivk povezana z verjetnostjo posameznega izida. Odnos med verjetnostjo uspeha in eno pojasnjevalno spremenljivko X opišemo preko logističnega modela

$$P(Y = 1|x) = p(x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)}, \tag{2.1}$$

kjer je $0 \leq p(x) \leq 1$. Krivulja ima značilno S-obliko. Z logaritmom obetov (logitom) model zapišemo kot

$$logit[p(x)] = log\frac{p(x)}{1 - p(x)} = \alpha + \beta x, \tag{2.2}$$

kar izraža linearno zvezo med pojasnjevalno spremenljivko in logaritmom obetov [17, 18, 19].

Levi del enačbe omogoča, da rezultate interpretiramo kot odvisnost obeta za uspeh od pojasnjevalne spremenljivke X. e^{β} je obet za uspeh za osebe z dano vrednostjo pojasnjevalne spremenljivke X, glede na tiste, ki imajo vrednost X za eno enoto nižjo.

Za lažjo predstavo si poglejmo primer, kjer so ugotavljali ali je smrčanje dejavnik tveganja za bolezen srca [18]. Za bolezen srca sta možni dve vrednosti, prisotnost ali odsotnost bolezni. Z modelom logistične regresije modeliramo obet za prisotno bolezen srca v odvisnosti od prisotnosti smrčanja:

$$logit[p(x)] = -3\ 87 + 0\ 40x. \tag{2.3}$$

Ocenjen obet za prisotnost bolezni srca je zaradi pozitivne vrednosti $\hat{\beta} = 0\ 40$ večji pri prisotnosti smrčanja. Model oceni, da imajo pacienti, ki smrčijo, 1.5-krat večji obet za srčno bolezen v primerjavi s tistimi, ki ne smrčijo [18].

2.3 Model sorazmernih obetov

V primeru modela logistične regresije smo obravnavali dva možna odziva (prisotnost bolezni srca, odsotnost bolezni srca) in ocenjevali obet za prisotnost bolezni srca glede na pojasnjevalno spremenljivko X (prisotnost smrčanja).

Kadar imamo več kot dva možna odziva, model logistične regresije ni več primeren. McCullagh je v ta namen opisal model sorazmernih obetov [15].

Model sorazmernih obetov se uporablja za modeliranje kategorične odvisne spremenljivke, ko je število kategorij te spremenljivke večje od 2 (več kot dva možna odziva) in jih lahko uredimo po velikosti [20]. V predhodno opisanem primeru bolezni srca, bi to pomenilo, da imamo poleg prisotnosti in odsotnosti bolezni, še vmesne stopnje, npr. stopnjo "prisotni znaki bolezni". Odzive bi lahko uredili po vrstem redu: bolezen ni prisotna < prisotni znaki bolezni < bolezen je prisotna.

Naj bo J število kategorij slučajne spremenljivke Y. Slučajna spremenljivka Y tako lahko zavzame vrednosti kategorije $1, ..., J$. Z x označimo vrednost pojasnjevalne spremenljivke X. Model predpostavlja, da velja:

$$log\frac{P(Y \leq j|x)}{P(Y > j|x)} = logit[P(Y \leq j|x)] = \alpha_j + \beta x, j = 1, ..., J-1, \tag{2.4}$$

kjer so $\alpha_1, ..., \alpha_{J-1}$ in β neznani parametri, z j pa označimo meje med kategorijami odvisne spremenljivke Y. Vseh mej je za ena manj kot kategorij J [18].

V enačbi modela (2.4) ocenjujemo kumulativne logite (ang. cumulative logits), ki so definirani preko kumulativnih verjetnosti (ang. cumulative probabilities):

$$\begin{aligned} logit[P(Y \leq j|x)]) &= log\frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)} \\ &= log\frac{p_1(x) + \cdots + p_j(x)}{p_{j+1}(x) + \cdots + p_J(x)}, j = 1, ..., J-1. \end{aligned} \quad (2.5)$$

V izračun vsakega kumulativnega logita so vključene vse kategorije spremenljivke Y [18].

Poenostavljeno; s kumulativnim logitom ocenjujemo obet, da je pri določeni vrednosti spremenljivke X, vrednost izida manjša ali enaka od izbrane meje j. V omenjenem primeru bi z enim logitom torej ocenjevali obet skupaj za odsotnost bolezni in prisotnost le znakov bolezni, glede na prisotnost smrčanja.

Model sorazmernih obetov (2.4) istočasno ocenjuje vse možne kumulativne logite, torej posamezne obete za izid manjši od neke meje j. Vsak kumulativni logit ima svojo začetno vrednost (ang. intercept). Začetne vrednosti α_j se večajo z večanjem vrednosti meje j, saj velja, da se izraz $P(Y \leq j|x)$ veča z večanjem j pri fiksnem x in je tako logit naraščajoča funkcija te verjetnosti [18]. Poenostavljeno; ko večamo mejo j, je verjetnost vrednosti odziva pod to mejo vedno večja.

Kumulativnost v pristopu se kaže v tem, da za $\beta > 0$, večja vrednost X vodi v povečanje verjetnosti za vse odzive, ki so manjši ali enaki meji j, ne samo njej enaki. Ime sorazmerni obeti pa izhaja iz lastnosti, ki velja za dve vrednosti pojasnjevalne spremenljivke x_1 in x_2:

$$log\frac{obet(Y \leq j|x_1)}{obet(Y \leq j|x_2)} = \beta(x_1 - x_2), \quad (2.6)$$

saj desna stran enačbe ni odvisna od j [20].

Obet odziva $\leq j$ za opazovanja z $x = x_1$ je enako $e^{\beta(x_1 - x_2)}$ -kratnemu obetu za opazovanja z $x = x_2$. Enaka lastnost sorazmerja velja tudi za vsak logit [18].

2.3.1 Ocenjevanje parametrov in statistično sklepanje

Ocene koeficientov so v modelu logistične regresije in modelu sorazmernih obetov ocenjene preko metode največjega verjetja, ki za izbrano verjetnostno porazdelitev podatkov obravnava funkcijo verjetja v odvisnosti od neznanega parametra. Ocena največjega verjetja je tista vrednost parametra, ki maksimizira to funkcijo. Dobljeni podatki so najbolj verjetni pri tej vrednosti parametra. Cenilka pridobljena po metodi največjega verjetja je nepristranska, asimptotsko normalno porazdeljena in asimptotsko učinkovita [18].

Za statistično sklepanje lahko uporabimo Waldov test, s katerim testiramo domnevo $H_0 : \beta = \beta_0$. Z neničelno standardno napako ($SE_{\hat{\beta}}$) ocenjene $\hat{\beta}$, testna statistika

$$z = \frac{(\hat{\beta} - \beta_0)}{SE_{\hat{\beta}}}, \tag{2.7}$$

sledi standardizirani normalni porazdelitvi, ko je $\beta = \beta_0$. Ta lastnost nam omogoča pridobitev vrednosti p. Za dvostransko alternativno domnevo ima z^2 porazdelitev pod ničelno domnevo χ_1^2 [18].

Globalno, vpliv pojasnjevalne spremenljivke v modelu preverjamo preko testa razmerja verjetij. Testna statistika je pridobljena preko funkcije verjetja, ki primerja maksimizirano funkcijo verjetja v prostoru parametrov pod H_0 (L_0), z maksimizirano funkcijo verjetja v prostoru parametrov pod H_A (L_1):

$$-2log\Lambda = -2(L_0 - L_1) \tag{2.8}$$

in asimptotsko sledi χ^2-porazdelitvi s tolikimi stopinjami prostosti, kot je razlika med ocenjenimi parametri v L_1 in L_0. Za ocenjevanje prispevka pojasnjevalne spremenljivke v modelu s testom razmerja verjetij primerjamo model brez pojasnjevalne spremenljivke z modelom, ki spremenljivko vsebuje. Statistično značilen test kaže na vpliv pojasnjevalne spremenljivke na izid [18].

2.3.2 Predpostavka o sorazmernih obetih

Ključna predpostavka modela (2.4) pravi, da je vpliv pojasnjevalnih spremenljivk na kumulativne verjetnosti izidov enak za vsako mejno vrednost. Resničnost predpostavke lahko preverimo s statističnimi testi (Waldov test, test razmerja verjetij, Score test), ali z grafičnimi testi [2].

Ena od možnosti preverjanja predpostavke o sorazmernih obetih je s testom razmerja verjetij, kjer preko razlike v deviancah primerjamo model (2.4) z modelom, ki ne predpostavlja sorazmernih obetov

$$log\frac{P(Y \leq j|x)}{P(Y > j|x)} = logit[P(Y \leq j|x)] = \alpha_j + \beta_j x, j = 1, ..., J-1, \tag{2.9}$$

kjer so $\beta_1, ..., \beta_{J-1}$ različni, neznani parametri. Model (2.9) imenujemo polinomski model in omogoča različen vpliv pojasnjevalne spremenljivke X na kumulativno verjetnost izida za različne mejne vrednosti [19, 20].

Test razmerja verjetij lahko primerja maksimizirano funkcijo verjetja za prilegan model z maksimizirano funkcijo verjetja za polni model. Tej razliki pravimo tudi devianca (ang. deviance):

$$\begin{aligned} D(y, \hat{\mu}) &= -2log\frac{\text{največje verjetje za prilegan model}}{\text{največje verjetje za polni model}} \\ &= -2[L(\hat{\mu}; y) - L(y; y)] \end{aligned} \tag{2.10}$$

Polni model ima za vsako opazovanje svoj parameter in se tako podatkom povsem prilega. Bližje kot je naš model polnemu modelu, bolje opišemo z njim podatke - boljše je njegovo prileganje [18].

Pri primerjavi dveh gnezdenih modelov preko testa razmerja verjetij, je testna statistika enaka razliki med deviancama za ta dva modela:

$$\begin{aligned} & -2[L(\hat{\mu_0}; y) - L(\hat{\mu_1}; y)] \\ = \; & -2[L(\hat{\mu_0}; y) - L(y; y)] - [-2[L(\hat{\mu_1}; y) - L(y; y)]] \\ = \; & D(y, \hat{\mu_0}) - D(y, \hat{\mu_1}). \end{aligned} \tag{2.11}$$

Testna statistika ima veliko vrednost, ko se prvi model slabše prilega v primerjavi z drugim modelom. Ničelna porazdelitev testne statistike je χ^2-porazdelitev s tolikimi stopinjami prostosti, kot je razlika med številom ocenjenih parametrov med obema modeloma [18].

Ničelna domneva v našem primeru pravi, da se polinomski model nič bolje ne prilega podatkom kot model sorazmernih obetov. Majhna vrednost p nakazuje na boljše prileganje polinomskega modela podatkom v primerjavi z modelom sorazmernih obetov, kar kaže na to, da sorazmerni obeti ne veljajo [20].

2.3.3 Uporaba na podatkih z nakopičenimi vrednostmi

V primeru nakopičenih vrednosti odvisna spremenljivka ni kategorična, zato se priporoča razporeditev vrednosti v zveznem delu v urejenostne kategorije, medtem ko nakopičene vrednosti tvorijo svojo kategorijo [2]. Kljub pretvorbi številskih vrednosti v kategorične, pa je Green [21] pokazal, da tudi razporeditev vrednosti zveznega dela v malo skupin privede k zelo podobnim ocenam regresijskih koeficientov in njihovih standardnih napak, kot analiza na nespremenjenih vrednostih zveznega dela.

Pred uporabo modela sorazmernih obetov je potrebno preveriti veljavnost predpostavke o sorazmernih obetih. Thomas [20] poroča, da test, opisan v prejšnjem poglavju, deluje dobro, kadar imamo dovolj veliko opazovanj v vsaki od kategorij odvisne spremenljivke. Kadar pa so opazovanja porazdeljena v kategorije neenakomerno in je posledično njihovo število v nekaterih kategorijah majhno, testna statistika ni porazdeljena po predpostavljeni χ^2-porazdelitvi, kar vodi v povečanje napake 1.vrste in torej v anti-konzervativne rezultate. Prav

tako Score test in Waldov test za preverjanje predpostavke o sorazmernih obetih delujeta anti-konzervativno, ko je število opazovanj v nekaterih kategorijah zelo majhno [22].

Značilnost testa za preverjanje predpostavke tako ne pomeni nujno, da uporaba modela sorazmernih obetov ni primerna za podatke, ki jih želimo analizirati [23]. Da se izognemo neveljavnemu testu, je priporočljivo razporediti vrednosti zveznega dela v kategorije tako, da je število opazovanj v vsaki kategoriji dovolj veliko [2]. Dodatna težava bi bila lahko še določitev števila kategorij, čeprav izgleda, da pridemo do podobnih rezultatov z uporabo različnega števila. Vendar pa optimalna izbira števila kategorij in mejnih vrednosti (še) ni znana [2].

2.4 Kombinacija logistične in linearne regresije

Lachenbruch [24] je predlagal kombinacijo logistične in linearne regresije na podlagi koncepta, da imajo morda pojasnjevalne spremenljivke drugačen vpliv na vrednost odziva 0 (v primeru nakopičene vrednosti pri 0) kot pa na spreminjanje jakosti odziva, ko je ta pozitiven. Testna statistika

$$V^2 = B^2 + T^2, \tag{2.12}$$

združi binomski test (B) za razliko v deležu nakopičenih vrednosti med dvema skupinama, s parametričnim testom t (T), ki testira razliko v povprečjih zveznih delov obeh skupin. Ker sta B in T neodvisni in asimptotsko normalni, vsota njunih kvadratov asimptotsko sledi χ^2 porazdelitvi z dvema stopinjama prostosti [3, 25].

Vrednost testne statistike iz testa t lahko nadomestimo s testno statistiko, ki smo jo pridobili za primerjavo zveznega dela porazdelitve (npr., Mann-Whitneyev test ali Kolmogorov-Smirnov test). Podobno lahko nadomestimo testno statistiko iz binomskega testa s testno statistiko χ^2-testa za deleže [12].

Problem dveh skupin in ene pojasnjevalne spremenljivke se da razširiti tudi na situacije, kjer želimo uporabiti modele multiple regresije, s katerimi preverjamo vpliv pojasnjevalne spremenljivke ob prisotnosti dodatnih pojasnjevalnih spremenljivk. V tem primeru lahko uporabimo model logistične regresije za modeliranje verjetnosti za nakopičeno vrednost in model linearne regresije za modeliranje pozitivnih vrednosti [2, 25].

2.4.1 Uporaba modela logistične regresije

Model logistične regresije služi za primerjavo dveh izidov, kjer v primeru kombiniranega pristopa en izid predstavljajo nakopičene vrednosti, drugi izid pa vrednosti iz zveznega dela porazdelitve.

Da ugotovimo, ali ima pojasnjevalna spremenljivka X vpliv na obet za nakopičeno vrednost, s testom razmerja verjetij primerjamo model brez spremenljivke X z modelom s spremenljivko X. Testna statistika je porazdeljena po χ^2 porazdelitvi s $k-1$ stopinjami prostosti, kjer je k število ravni opisne pojasnjevalne spremenljivke X.

2.4.2 Model linearne regresije

Model enostavne linearne regresije

$$Y = \alpha + \beta X + \epsilon, \tag{2.13}$$

določa, da je opazovana vrednost Y skupek linearne odvisnosti od pojasnjevalne spremenljivke X in naključne napake ϵ. Napake so neodvisne in normalno porazdeljene, s pričakovano vrednostjo $E(\epsilon) = 0$ in varianco $Var(\epsilon) = \sigma^2$ [26]. Model opiše podatke z regresijsko premico

$$y = \hat{\alpha} + \hat{\beta} x, \tag{2.14}$$

kjer sta vrednosti $\hat{\alpha}$ in $\hat{\beta}$ oceni za α in β iz enačbe 2.13. Vrednost $\hat{\alpha}$ nam pove kje ocenjena premica seka ordinatno os in tako predstavlja njeno začetno vrednost. Vrednost $\hat{\beta}$ opisuje naklon ocenjene premice. Ko se vrednost x spremeni za eno enoto, se povprečje Y spremeni za $\hat{\beta}$ enot [27].

Oceni za α in β sta izračunani preko metode najmanjših kvadratov, tako da se minimizira izraz

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2, \tag{2.15}$$

kar pomeni, da je vsota kvadriranih odklonov od premice minimalna [26].

Ker so napake ϵ v enačbi 2.13 neodvisne in normalno porazdeljene slučajne spremenljivke, potem sta oceni za α in β linearni kombinaciji neodvisnih, nor-

malno porazdeljenih slučajnih spremenljivk, iz česar sledi, da sta tudi sami približno normalno porazdeljeni [26]. Ta lastnost ocen nam omogoča določitev intervalov zaupanja in testiranje posameznih domnev o statistični značilnosti ocen ($H_0 : \beta = 0$), saj velja

$$\frac{\hat{\beta} - \beta}{SE_{\hat{\beta}}} \sim t_{n-2}. \tag{2.16}$$

Globalno, vpliv pojasnjevalne spremenljivke v modelu linearne regresije preverjamo z analizo variance [26]. Ničelno domnevo $H_0 : \beta = 0$ preverjamo preko F statistike

$$F = \frac{SS_R/(k)}{SS_E/(n-k-1)}, \tag{2.17}$$

kjer je SS_R vsota kvadriranih odmikov ocenjenih vrednosti izida od povprečja vseh ocenjenih vrednosti, SS_E je vsota kvadriranih ostankov, k so stopinje prostosti modela in n velikost vzorca. Števec enačbe 2.17 izraža pojasnjeni del variabilnosti modela, medtem ko imenovalec izraža variabilnost zaradi napak. F statistika je porazdeljena po $F_{k,n-k-1}$ porazdelitvi. Pri visoki vrednosti F statistike zavrnemo ničelno domnevo in zaključimo, da β ni enaka 0 [26].

2.4.3 Združitev obeh testnih statistik

Če želimo združiti rezultate iz logistične in linearne regresije za primer ene pojasnjevalne spremenljivke z 2 ravnema, B^2 iz enačbe 2.12 nadomestimo s testno statistiko testa razmerja verjetij, ki je porazdeljena po χ^2_1-porazdelitvi. Sočasno, T^2 nadomestimo z F statistiko iz testa analize variance, ki je porazdeljena po $F_{1,n-2}$, kar lahko zapišemo kot:

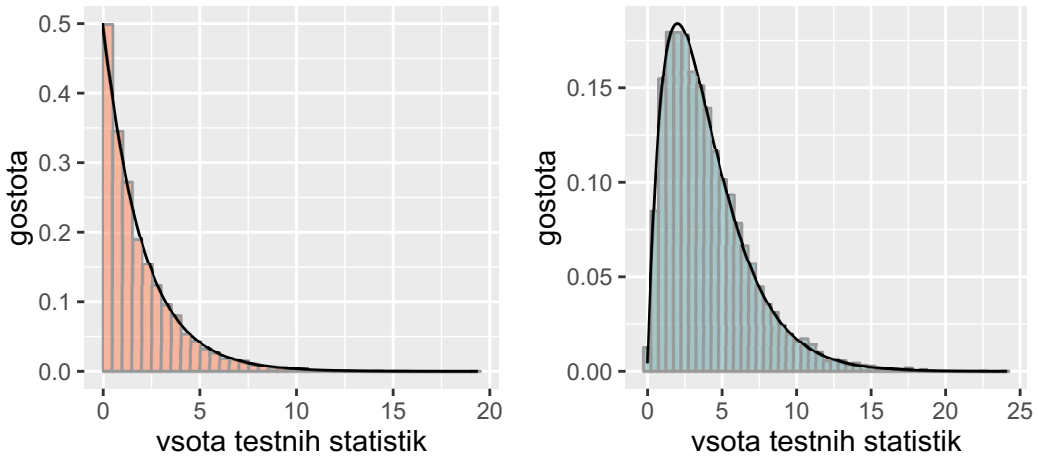
$$F_{1,n-2} = \frac{\chi^2_1}{\chi^2_{n-2}/(n-2)}. \tag{2.18}$$

Ko gre $n \to \infty$, gre imenovalec proti 1 in F se bliža χ^2_1-porazdelitvi [28].

Naj bo $V \sim \chi^2_n$ in $U \sim \chi^2_m$. Če sta U in V neodvisni, sledi [26]

$$U + V \sim \chi^2_{m+n}, \tag{2.19}$$

kar za naš primer univariatne analize pomeni, da je vsota posameznih testnih statistik porazdeljena po χ^2_2 (slika 2.1, levo).



Slika 2.1: Porazdelitev vsote testnih statistik iz simulacije, za primer dveh ravni (levo) in za primer treh ravni opisne pojasnjevalne spremenljivke (desno).

Teoretični rezultat smo pokazali tudi s simulacijo. Ogledali smo si primer, kjer smo v model vključili eno opisno pojasnjevalno spremenljivko z dvema ravnema in primer, kjer ima spremenljivka tri ravni. Skupna testna statistika se v prvem primeru porazdeljuje po χ^2_2 porazdelitvi (slika 2.1, levo). V drugem primeru je statistika testa razmerja verjetij (U) porazdeljena po χ^2_2 porazdelitvi, enačba 2.18 za ta primer pa je enaka

$$F_{2,n-3} = \frac{\chi^2_2/2}{\chi^2_{n-3}/(n-3)}. \tag{2.20}$$

Ko gre $n \to \infty$, se dvakratnik F statistike (V) bliža χ^2_2 porazdelitvi, vsota $U + V$ pa je porazdeljena po χ^2_4 porazdelitvi (slika 2.1, desno). Na sliki 2.1

je prikazano, da se rezultati pridobljeni s simulacijami ujemajo s teoretičnimi rezultati (črna krivulja).

2.5 Mann-Whitneyev test

Mann-Whitnejev test (poimenovan tudi Wilcoxonov test vsot rangov) je neparametričen test za primerjavo porazdelitev dveh skupin, ki ga uporabljamo kot alternativo testu t, ko ne želimo predpostaviti, da sta spremenljivki v populaciji normalno porazdeljeni. S pomočjo Mann-Whitneyevega testa določimo ali skupini izhajata iz iste oz. različne populacije. Preverja se ničelna domneva, ki pravi, da imata populaciji isto porazdelitev [26, 29].

Mann-Whitneyev test priredi rang vsaki vrednosti obeh skupin in nato primerja porazdelitev rangov obeh skupin.

Naj bo n_x število opazovanj v prvi skupini in n_y število opazovanj v drugi skupini. Vrednosti iz obeh skupin so razporejene po vrstnem redu od najmanjše do največje. Najmanjši vrednosti je prirejen rang 1, naslednji rang 2, največji vrednosti pa rang $n_x + n_y$ [29].

Vsoto rangov prve skupine zapišemo kot:

$$R_x = \sum_{i=1}^{n_x} (rX_i). \tag{2.21}$$

Če je ta vsota premajhna ali prevelika, bomo ničelno domnevo zavrnili [26].

Testna statistika Mann-Whitneyevega testa temelji na izračunu U statistike:

$$U = n_x n_y + \frac{n_x(n_x + 1)}{2} - R_x, \tag{2.22}$$

ki nam pove število primerov, ko je bila vrednost v drugi skupini večja od vrednosti v prvi skupini, za vse možne primerjave. Ob veljavni ničelni domnevi pričakujemo, da bo za polovico primerjav vrednost v drugi skupini večja, za

polovico primerjav pa bo vrednost v prvi skupini večja. Testna statistika Mann-Whitneyevega testa primerja dejansko vrednost na podatkih (U) s pričakovano vrednostjo, če H_0 velja ($\mu_W = (n_x n_y)/2$):

$$W = \frac{U - \mu_W}{\sigma_W}, \tag{2.23}$$

kjer je $\sigma_W = \sqrt{n_x n_y (n_x + n_y + 1)/12}$. Testna statistika W je pod H_0 porazdeljena po standardizirani normalni porazdelitvi [1].

V primeru, da želimo primerjati porazdelitev več kot dveh skupin in ne želimo predpostavljati normalne porazdelitve spremenljivk v populaciji, uporabimo Kruskall-Wallisov test (ang. Kruskal-Wallis test) [30], ki je razširitev Mann-Whitneyevega testa za primerjavo porazdelitev več skupin in je neparametrična alternativa enosmerni analizi variance [31]. Ničelna domneva pravi, da vse skupine prihajajo iz iste populacije. Sklep ob zavrnitvi ničelne domneve je, da to ne velja za vsaj eno od obravnavanih skupin. Da izvemo med katerimi skupinami prihaja do razlik v porazdelitvi, v nadaljevanju uporabimo Dunnov test (ang. Dunn's test) mnogoterih primerjav [32].

2.6 Model Tobit

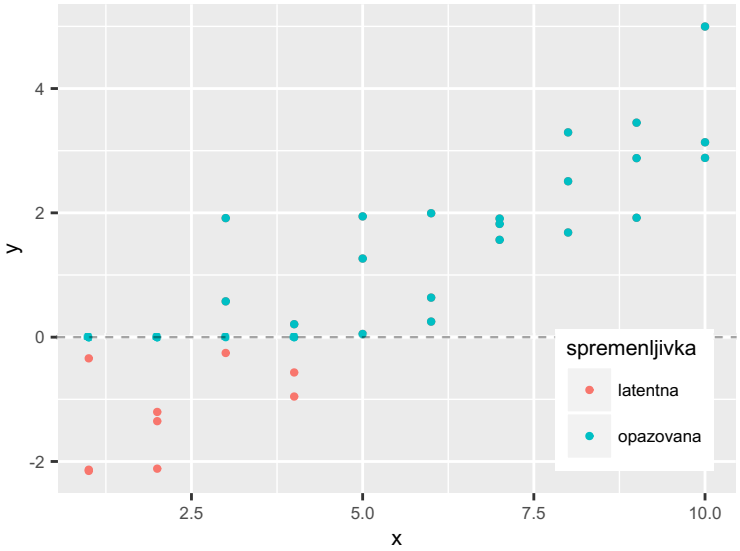
Model Tobit je parametrični regresijski model, ki predpostavlja levo-krnjeno normalno porazdelitev podatkov in za pridobitev ocene regresijske premice uporabi vsa opazovanja, tudi tista, ki so nakopičena pri neki mejni vrednosti [1, 16].

Model predpostavlja, da v ozadju porazdelitve spremenljivke y, ki jo vidimo, obstaja nekakšna skrita - latentna spremenljivka $Y*$ (ang. latent variable) [30], od katere mi opazimo samo vrednosti večje od 0. Latentno spremenljivko opišemo preko enačbe

$$Y* = \alpha + \beta X + \epsilon, \tag{2.24}$$

kjer je X pojasnjevalna spremenljivka, ϵ pa označuje neodvisno napako, za katero predpostavljamo, da je porazdeljena po normalni porazdelitvi s povprečjem 0 in varianco σ^2 [16].

Opazovana vrednost y je enaka vrednosti latentne spremenljivke $Y*$, ko je ta večja od 0. V primeru, da je vrednost latentne spremenljivke manjša ali enaka 0, je opazovana vrednost y enaka 0, kar prikazuje slika 2.2.



Slika 2.2: Prikaz vrednosti latentne spremenljivke $Y*$ in opazovanih vrednosti y.

Opazovano vrednost y lahko zapišemo tudi z enačbo

$$y = \begin{cases} Y*, & \text{če } Y* > 0 \\ 0, & \text{če } Y* \leq 0. \end{cases}$$

Model Tobit oceni parametre latentne spremenljivke preko metode največjega verjetja, kar daje ocenam ugodne lastnosti za statistično sklepanje. Pozorni moramo biti na to, da z modelom ocenimo latentno spremenljivko $Y*$ in ne opazovane spremenljivke y [1, 16].

3 Metode dela

3.1 Cilji in hipoteze

Namen naloge je ugotoviti, ali je model sorazmernih obetov primeren za analizo podatkov z nakopičenimi vrednostmi, kadar je delež nakopičenih vrednosti zelo velik in delovanje modela primerjati s tremi metodami, ki se uporabljajo za analizo podatkov z nakopičenimi vrednostmi in imajo vsaka svoje prednosti v določenih situacijah:

- kombinacija logistične in linearne regresije (v nadaljevanju bomo za to metodo uporabljali izraz *model Log+Lin*)
- Mann-Whitneyev test
- model Tobit

Za model sorazmernih obetov, Mann-Whitneyev test in model Tobit bomo v nalogi uporabljali tudi izraz *enodelni testi*, saj nakopičene vrednosti obravnavajo kot del celotne porazdelitve. Za model Log+Lin pa bomo uporabljali tudi izraz *dvodelni test*, ker posebej obravnava del nakopičenih vrednosti in zvezni del porazdelitve.

V nalogi več pozornosti namenimo modelu sorazmernih obetov zaradi dveh razlogov. Njegova uporaba je v primeru podatkov z nakopičenimi vrednostmi zaželjena zaradi ugodne interpretacije preko razmerja obetov. Poleg tega, pa njegovo delovanje še ni bilo zelo raziskano v tem kontekstu.

S simulacijami smo želeli preveriti, kakšno je delovanje modela sorazmernih obetov v primerjavi z ostalimi testiranimi metodami v primeru, ko je domneva o sorazmernih obetih veljavna (v nadaljevanju: *sorazmerni obeti veljajo, sorazmernost*) in v primeru, ko domneva ni veljavna (v nadaljevanju: *sorazmerni obeti ne veljajo, nesorazmernost*).

Dodatno smo želeli ugotoviti kako delež nakopičenih vrednosti vpliva na moč testov in kakšne so značilnosti pristopov v različnih situacijah.

Pod ničelno domnevo pričakujemo ustrezno velikost vseh testov. Pri večjem deležu nakopičenih vrednosti pričakujemo, da se bo velikost vseh testov začela razlikovati od ustrezne, zaradi manjšega števila vrednosti v zveznem delu porazdelitve izida. Večji delež nakopičenih vrednosti v modelu Tobit namreč pomeni več krnjenih opazovanj, pri modelu Log+Lin majhno velikost vzorca za del "Linškupne testne statistike, pri modelu sorazmernih obetov bolj neenakomerno porazdelitev opazovanj med kategorijami izida in pri Mann-Whitneyevem testu več vezanih rangov (ang. tied rank) [30].

Pod alternativno domnevo pričakujemo, da bo imel model sorazmernih obetov rahlo večjo moč od ostalih dveh enodelnih testov, medtem ko se bo model Log+Lin odrezal slabše.

V primeru, da sorazmerni obeti ne veljajo, pričakujemo veliko zmanjšanje moči modela sorazmernih obetov, medtem ko ne pričakujemo zmanjšanja moči za ostale tri metode. Pričakujemo bolj primerljivo moč modela Log+Lin z močjo Mann-Whitneyevega testa in modela Tobit.

Dodatno bomo preverili tudi delovanje testa razmerja verjetij, s katerim se odločamo ali so v podatkih prisotni sorazmerni obeti ali ne. Pričakujemo, da se bo pri majhnih vzorcih in velikem deležu nakopičenih vrednosti, test obnašal anti-konzervativno.

3.2 Opis simulacij

Vse simulacije in statistične analize so bile izvedene s programskim jezikom R, verzija 3.2.3 [33].

V prvem delu simulacij smo podatke generirali preko latentne spremenljivke. Ko smo preverjali delovanje metod, ko sorazmerni obeti veljajo, smo simulacije ponovili 10000-krat, medtem ko smo v delu simulacij, ko sorazmerni obeti ne veljajo, simulacije ponovili le 1000-krat. V prvem delu smo namreč preučevali delovanje metod ob veljavni ničelni domnevi, kjer že manjše odstopanje med vrednostmi ocenjenimi s simulacijami in pravimi vrednostmi lahko vpliva na sklepe, medtem ko smo v drugem delu preučevali delovanje metod pod alternativno domnevo, kjer ima natančnost pri ocenjeni moči manjši vpliv na tvorjenje zaključkov.

V drugem delu simulacij smo podatke generirali še iz logaritemsko normalne porazdelitve. Za dodatni pristop generiranja podatkov smo se odločili zaradi razlik med generiranimi podatki za različne situacije v prvem delu simulacij, kar otežuje preučevanje lastnosti, ki nas zanimajo (razlike med situacijami so opisane v podpoglavju 3.3, v razdelkih 3.3.1.5 in 3.3.2.1). Pri dodatnih simulacijah način uvajanja nakopičenih vrednosti ustvari bolj primerljive situacije. Z dodatnimi simulacijami želimo preveriti predvsem vpliv skladnih oz. neskladnih razlik med ravnema pojasnjevalne spremenljivke na delovanje izbranih metod (pojav skladnih in neskladnih razlik je opisan v podpoglavju 1.2). Tudi te simulacije smo ponovili 1000-krat. Pri vseh simulacijah smo za mejo statistične značilnosti uporabili stopnjo značilnosti $\alpha = 0{,}05$.

3.3 Generiranje preko latentne spremenljivke

Za kategorično odvisno spremenljivko Y predpostavimo, da se pod njo skriva zvezna številska spremenljivka. Regresijski model, ki opisuje to številsko spremenljivko, omogoča skupen vpliv β za različne meje j v modelu sorazmernih obetov. Latentno spremenljivko označimo z Y^*. Če velja

$$Y^* = \beta X + \epsilon, \tag{3.1}$$

kjer napaka ϵ izhaja iz logistične porazdelitve, potem bo ne glede na to, kako razrežemo spremenljivko Y v kategorije, parameter β, ki opiše vpliv na odvisno spremenljivko Y, enak za vse kategorije. Rezultat so sorazmerni obeti pri generiranih podatkih [18].

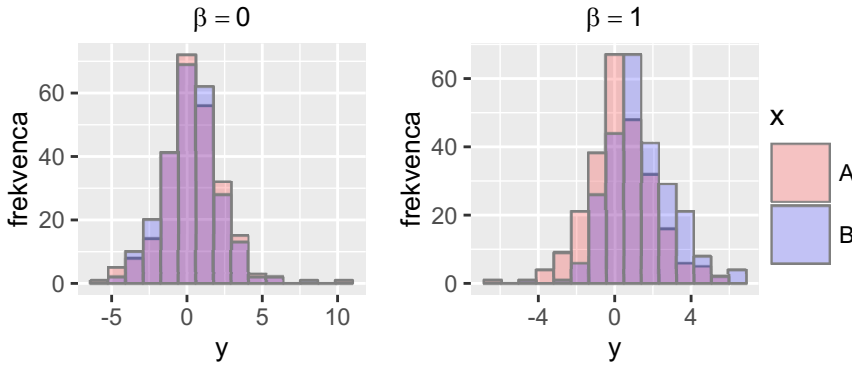
3.3.1 Prvi del simulacij: sorazmerni obeti veljajo

Vse vrednosti izida Y generiramo preko enačbe:

$$y = \alpha + \beta x + \epsilon, \tag{3.2}$$

kjer ima α v vseh situacijah vrednost 0,3, x je vrednost opisne spremenljivke X z dvema ravnema (A in B), β opisuje jakost povezanosti med pojasnjevalno spremenljivko X in izidom Y, ϵ pa predstavlja napako, generirano iz logistične porazdelitve. Ob veljavni ničelni domnevi predpostavimo, da X in Y v populaciji nista povezana in generiramo izid z vrednostjo $\beta = 0$. Posledično odpade člen βx pri izračunu vrednosti izida Y za obe ravni spremenljivke X.

Pod alternativno domnevo generiramo izid z vrednostjo $\beta = 0{,}5$ za manjšo povezanost X in Y (oz. manjšo razliko med ravnema spremenljivke X) in z vrednostjo $\beta = 1$ za večjo povezanost X in Y (oz. večjo razliko med ravnema). Porazdelitvi generiranih izidov z $\beta = 0$ in $\beta = 1$ sta prikazani na sliki 3.1.



Slika 3.1: Porazdelitev generiranega izida Y za primer, ko ni razlik med ravnema spremenljivke X (levo) in za primer, ko med njima obstaja razlika (desno).

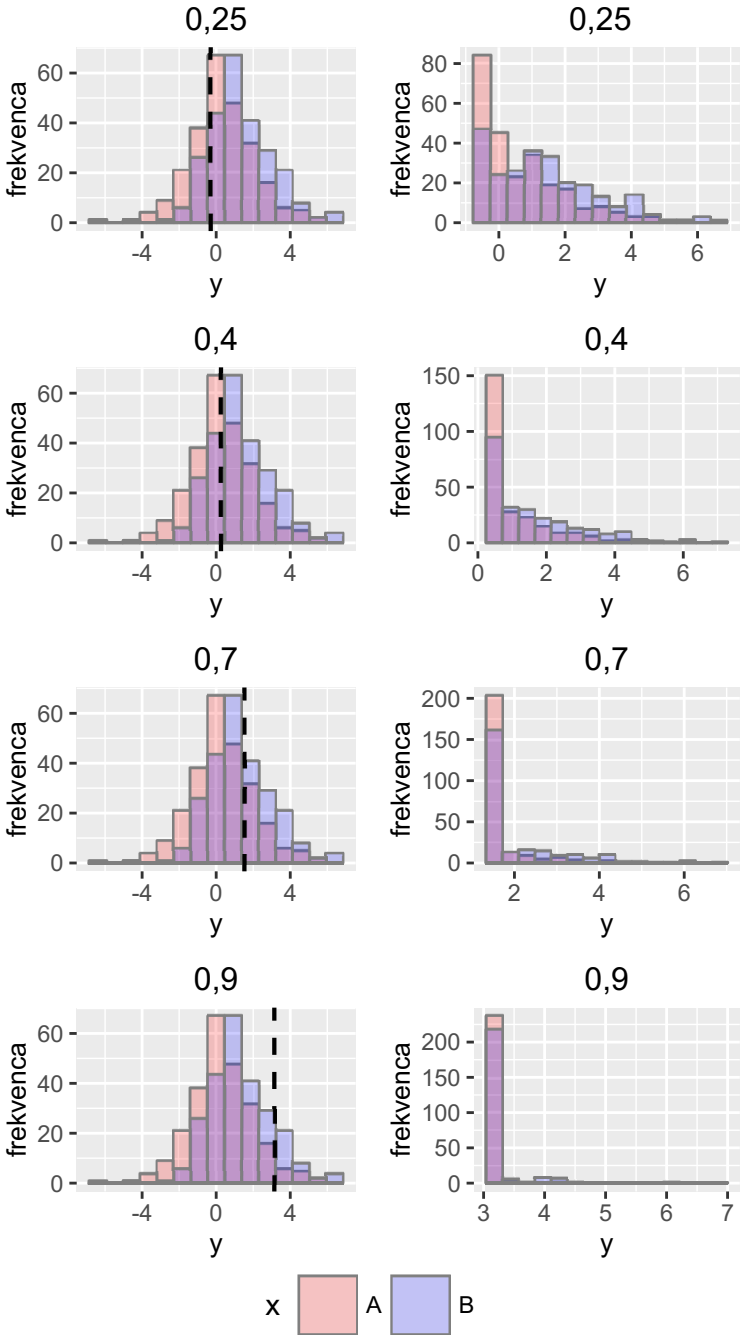
Generirali smo podatke velikosti n= 100, 300 in 500 (isto število vrednosti za raven A in B spremenljivke X) in izidu Y priredili nakopičene vrednosti tako, da je bil njihov delež enak 0,25, 0,4, 0,7 in 0,9. Pri izbiri deležev nakopičenih vrednosti in velikosti vzorca smo upoštevali lastnosti obravnavanih podatkov. Simulirane situacije so povzete v tabeli 3.1.

Delež nakopičenih vrednosti	n=100		n=300		n=500	
	nakopičeni del	zvezni del	nakopičeni del	zvezni del	nakopičeni del	zvezni del
0,25	25	75	75	225	125	375
0,4	40	60	120	180	200	300
0,7	70	30	210	90	350	150
0,9	90	10	270	30	450	50

Tabela 3.1: Velikost posameznega dela porazdelitve generiranega izida Y za različne velikosti vzorca in deleže nakopičenih vrednosti.

Nakopičene vrednosti priredimo tako, da najmanjše vrednosti generiranega Y nastavimo na isto vrednost - tisto, ki je med njimi največja. Število nakopičenih vrednosti je odvisno od predhodno izbranega deleža. Slika 3.2 prikazuje postopek uvajanja nakopičenih vrednosti, med katere uvrstimo vrednosti levo od črne

črtkane črte.



Slika 3.2: Prikaz postopka uvajanja nakopičenih vrednosti generiranemu izidu Y za različne deleže ob predpostavljenih sorazmernih obetih. Prikaz porazdelitve izida pred (levo) in po uvedbi nakopičenih vrednosti (desno).

3.3.1.1 Model sorazmernih obetov

Za model sorazmernih obetov najprej generirani izid Y kategoriziramo. Odločili smo se za razdelitev v štiri kategorije, kjer prva kategorija predstavlja nakopičene vrednosti, ostale tri kategorije pa predstavljajo vrednosti zveznega dela izida Y. Vrednosti Y razporedimo v kategorije 2, 3 in 4 tako, da je število vrednosti v vsaki od kategorij primerljivo (preko tercilov zveznega dela Y). Kategorizirani izid Y shranimo v spremenljivko *y.cat*.

Za prileganje modela sorazmernih obetov smo uporabili funkcijo `polr` (kratica za "proportional odds logistic regression") iz paketa *MASS*. V programu R uporabimo kodo:

```
modPO = polr(y.cat ~ x, Hess=TRUE)
```

Argument `"Hess"` nam ob izbiri `TRUE` omogoči pridobitev informacijske matrike iz procesa optimizacije in s tem izpis standardnih napak ocenjenih parametrov [34]. Oznaka PO v imenu modela je kratica za "proportional odds", ki jo uporabljamo tudi v nadaljevanju.

Preverjanje predpostavke o sorazmernih obetih

Za prileganje polinomskega modela smo uporabili funkcijo `multinom` iz paketa *nnet* [35]. Vrednost p smo pridobili preko razlike v deviancah obeh modelov in pripadajočo ploščino pod krivuljo χ^2 porazdelitve z `df` stopinjami prostosti. Koda v jeziku R s katero smo določili vrednost p je:

```
modM = multinom(y.cat~x)
df = modM$edf - modPO$edf
p_prop.odds = 1-pchisq(deviance(modPO)-deviance(modM),df)
```

3.3.1.2 Model Log+Lin

Pri tem pristopu najprej ločeno podatkom prilegamo model logistične regresije s funkcijo `glm` in model linearne regresije preko funkcije `lm`. V programu R

uporabimo kodo:

```
mod_Log = glm(y01~x, family=binomial)
mod_Lin = lm(yL~xL)
```

`y01` je binomska spremenljivka, kjer smo nakopičenim vrednostim priredili vrednost 1 in vrednostim zveznega dela izida Y vrednost 0. Model linearne regresije izvedemo le na vrednostih zveznega dela izida Y. Skupna testna statistika `s2` je sestavljena iz posameznih testnih statistik iz obeh modelov, ki ju pridobimo preko ukaza `anova`. V programu R uporabimo kodo:

```
chi = anova(mod_Log,test="Chisq")$Deviance[2]
f = anova(mod_Lin,test="Chisq")$F[1]
s2 = chi + f
```

3.3.1.3 Mann-Whitneyev test

Mann-Whitneyev test smo izvedli z uporabo funkcije `wilcox.test`. V programu R uporabimo kodo:

```
w_test = wilcox.test(y~x)
```

3.3.1.4 Model Tobit

Za prileganje modela Tobit smo uporabili funkcijo `vglm` iz paketa *VGAM* [36]. V programu R uporabimo kodo:

```
mod_Tob <- vglm(y~x, tobit(Lower = quantile(y,delez_nicel)))
```

Funkcija `tobit` znotraj narekuje prileganje modela Tobit, argument `Lower` pa vse vrednosti Y, ki so manjše, krni [36].

3.3.1.5 Lastnosti generiranih podatkov

Kljub temu, da smo za vse deleže nakopičenih vrednosti podatke generirali preko iste enačbe, po pretvarjanju različnega deleža izida Y v nakopičene vrednosti, se preostali zvezni deli porazdelitve močno razlikujejo glede na delež nakopičenih vrednosti in tako situacije niso povsem primerljive. Razlike med generiranimi podatki za različne deleže nakopičenih vrednosti moramo upoštevati, saj lahko vplivajo na rezultate.

V simulaciji smo 1000-krat generirali podatke za vsak delež nakopičenih vrednosti. V tabeli 3.2 so prikazane povprečne vrednosti iz simulacij za naslednje lastnosti podatkov: delež nakopičenih vrednosti za ravni A in B pojasnjevalne spremenljivke X, razlika med deležema, povprečje zveznega dela za raven A in B in njuna razlika. V zadnjem stolpcu tabele je podan še odstotek skladnih razlik (razlika med dvema skupinama je skladna, ko ima skupina z večjim deležem nakopičenih vrednosti pri 0, manjše povprečje zveznega dela porazdelitve).

	n=500						
	delež nakopičenih vrednosti			povprečje zveznega dela			skladnost
Delež	A	B	A-B	A	B	B-A	(%)
0,25	0,34	0,16	0,18	2,3	2,8	0,6	100,0
0,4	0,52	0,28	0,23	2,7	3,1	0,4	99,7
0,7	0,80	0,60	0,20	3,8	4,0	0,2	81,8
0,9	0,94	0,86	0,08	5,1	5,2	0,0	55,7
	n=100						
0,25	0,34	0,16	0,18	2,3	2,8	0,6	96,5
0,4	0,52	0,28	0,23	2,7	3,1	0,4	90,1
0,7	0,80	0,60	0,21	3,8	4,0	0,2	66,3
0,9	0,94	0,86	0,08	5,1	5,2	0,1	51,8

Tabela 3.2: Lastnosti generiranih podatkov z $\beta = 1$ za velikost vzorca 100 in 500; pridobljeno iz simulacij s 1000 ponovitvami.

Opazimo, da se z večanjem deleža nakopičenih vrednosti manjša razlika med ravnema A in B v zveznem delu porazdelitve, kar bo vplivalo na zmanjšanje moči

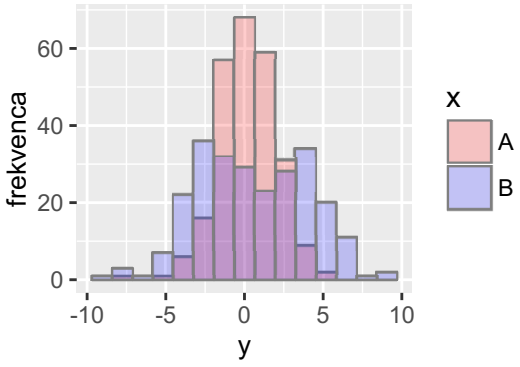
preučevanih testov. Učinek ni tako izrazit za manjše predpostavljene razlike med ravnema A in B, za katere so lastnosti predstavljene v dodatku A.

3.3.2 Drugi del simulacij: sorazmerni obeti ne veljajo

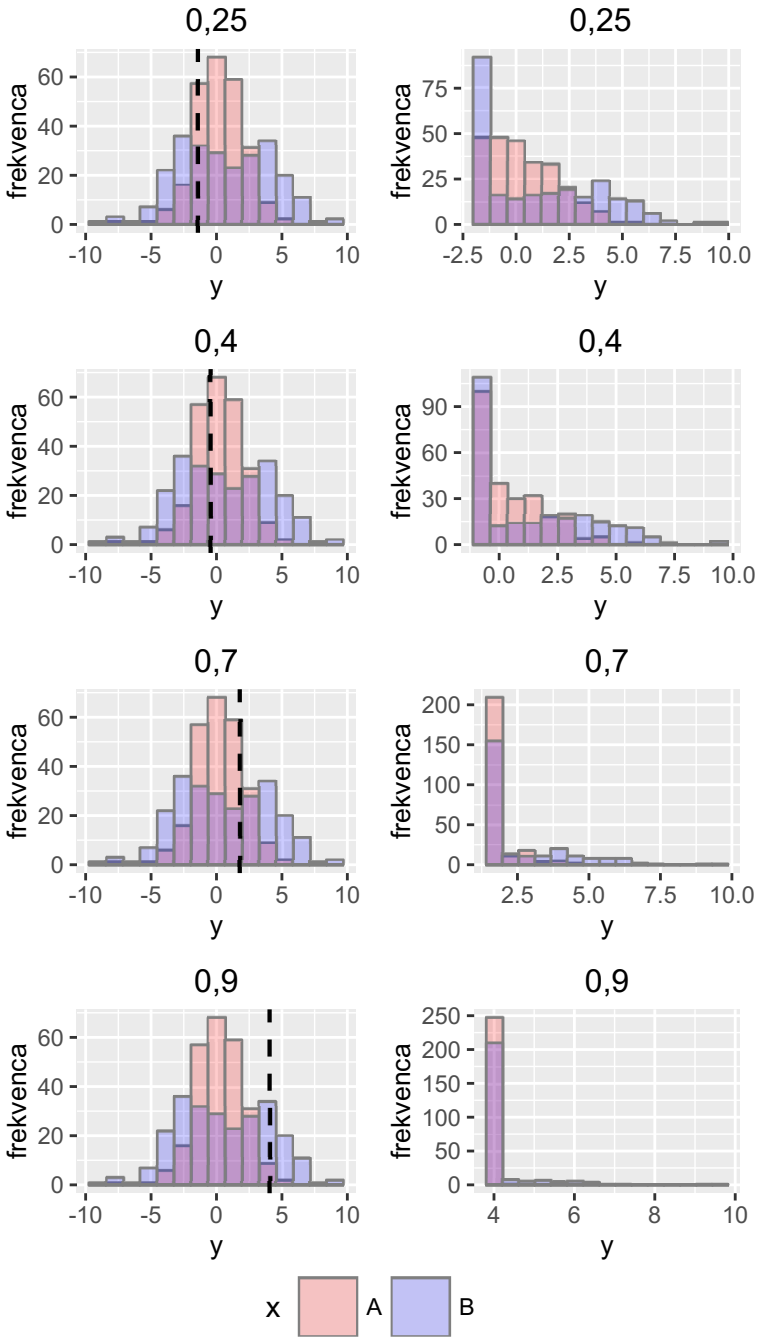
Ko sorazmerni obeti veljajo, je vpliv spremenljivke X na izid Y konstanten, ne glede na izbrano mejno vrednost med kategorijami. V primeru, da sorazmerni obeti ne veljajo, izid Y generiramo tako, da za vsako mejno vrednost med kategorijami ustvarimo drugačen vpliv spremenljivke X na izid Y. Vsako kategorijo Y generiramo iz svoje enačbe 3.1, tako da je β za vsak j drugačna, nato pa vse izide združimo v skupen Y.

$$y_j = \alpha + \beta_j x + \epsilon, \tag{3.3}$$

Y generiramo preko dveh različnih kombinacij β, kjer enkrat predpostavimo manjše razlike med ravnema spremenljivke X, drugič pa večje. Vrednost izida za manjše razlike generiramo z vrednostmi $\beta_1 = 1$, $\beta_2 = -1$, $\beta_3 = 2$, $\beta_4 = -1{,}5$, za večje razlike so te vrednosti dvakrat večje. Porazdelitev generiranega Y za večje predpostavljene razlike je prikazana na sliki 3.3, slika 3.4 pa prikazuje porazdelitev izida Y pred in po uvedbi nakopičenih vrednosti za vsak delež.



Slika 3.3: Porazdelitev izida Y za primer večje predpostavljene razlike med ravnema pojasnjevalne spremenljivke X.



Slika 3.4: Prikaz postopka uvajanja nakopičenih vrednosti generiranemu izidu Y za različne deleže ob kršeni predpostavki o sorazmernih obetih. Prikaz porazdelitve izida pred (levo) in po uvedbi nakopičenih vrednosti (desno).

3.3.2.1 Lastnosti generiranih podatkov

Lastnosti generiranih podatkov za primer, ko predpostavimo, da sorazmerni obeti ne veljajo, povzamemo v tabeli 3.3.

	n=500						
	delež nakopičenih vrednosti			povprečje zveznega dela			skladnost
Delež	A	B	A-B	A	B	B-A	(%)
0,25	0,15	0,35	-0,20	0,8	2,5	1,7	0,0
0,4	0,35	0,45	-0,10	1,3	3,1	1,8	0,0
0,7	0,80	0,60	0,20	2,8	4,0	1,2	100,0
0,9	0,98	0,82	0,15	5,1	5,4	0,4	81,6
	n=100						
0,25	0,15	0,35	-0,21	0,8	2,4	1,6	0,2
0,4	0,34	0,46	-0,12	1,3	3,1	1,8	2,5
0,7	0,79	0,61	0,18	2,7	4,0	1,2	98,3
0,9	0,98	0,82	0,15	5,0	5,4	0,4	73,2

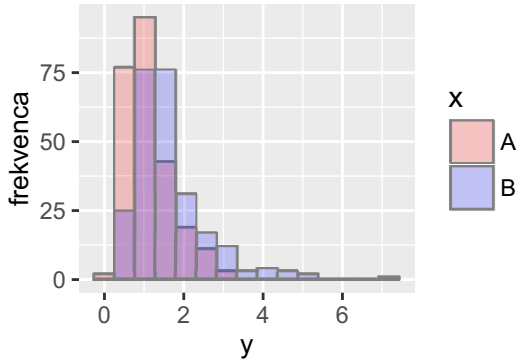
Tabela 3.3: Lastnosti generiranih podatkov za večje predpostavljene razlike med ravnema pojasnjevalne spremenljivke, za velikost vzorca 100 in 500; pridobljeno iz simulacij s 1000 ponovitvami.

Generirani podatki se med različnimi deleži nakopičenih vrednosti še bolj razlikujejo med seboj kot v prvem delu simulacij. Z večanjem deleža nakopičenih vrednosti se manjša razlika v povprečjih zveznih delov (učinek se izgubi pri generiranju manjših razlik med ravnema A in B), največja razlika med situacijami pa je v skladnosti. Pri deležih nakopičenih vrednosti 0,25 in 0,4 so razlike med ravnema pojasnjevalne spremenljivke X večinoma neskladne, medtem ko pri višjih dveh deležih nakopičenih vrednosti prevladujejo skladne razlike (viden enak učinek za manjše generirane razlike). Iz Slike 3.4 je razvidno, da se obrne smer razlike v posameznih deležih nakopičenih vrednosti med ravnema, kar ob večanju skupnega deleža nakopičenih vrednosti pomeni prehod neskladnih razlik v skladne. Lastnosti za manjše predpostavljene razlike so predstavljene v dodatku A.

3.4 Generiranje iz logaritemsko normalne porazdelitve

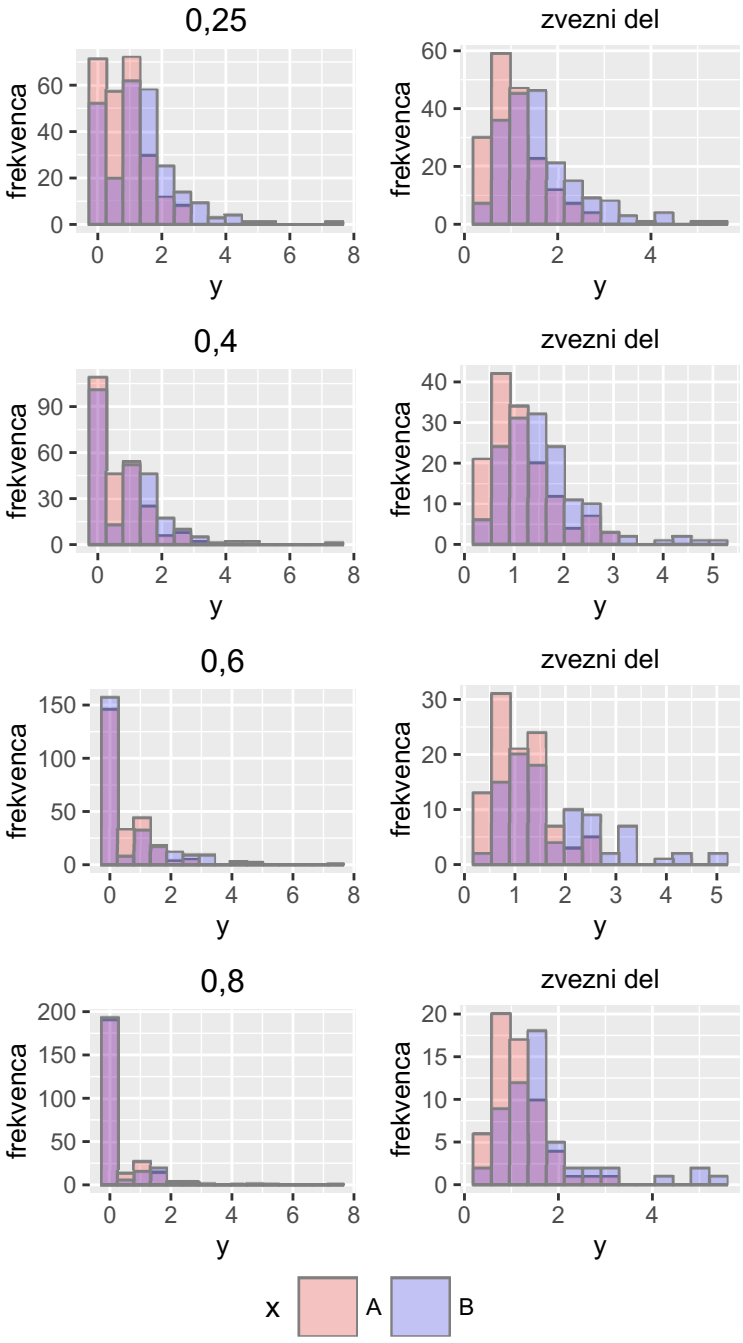
Pri dodatnih simulacijah izid Y generiramo iz logaritemsko normalne porazdelitve, vrednosti izida pa pretvorimo v nakopičene vrednosti na naključnih mestih porazdelitve. Tako lažje nadzorujemo razlike med ravnema pojasnjevalne spremenljivke, ki ostanejo enake, ne glede na delež nakopičenih vrednosti. Simuliramo na dva različna načina: enkrat generiramo skladne razlike med ravnema pojasnjevalne spremenljivke X, drugič pa neskladne.

Vrednosti izida Y za vsako raven posebej generiramo iz logaritemsko normalne porazdelitve $Y_i \sim LogNorm(\mu_i, \sigma)$, kjer izid za raven A generiramo s povprečno vrednostjo $\mu_A = 0$, izid za raven B pa z vrednostjo $\mu_B = 0{,}3$. Standardni odklon nastavimo enak za obe ravni, $\sigma = 0{,}5$. Porazdelitev skupnega izida Y za ravni A in B je prikazana na sliki 3.5.



Slika 3.5: Porazdelitev generiranega izida Y iz logaritemsko normalne porazdelitve.

Za vsako raven spremenljivke X iz binomske porazdelitve generiramo indikator za nakopičeno vrednost, $I \sim Bin(n, p)$, kjer p predstavlja verjetnost za nakopičeno vrednost. Vrednosti izida Y, za katere velja $I = 1$, nastavimo na 0. V primeru skladnih razlik je vrednost $p_A > p_B$, v primeru neskladnih razlik pa velja obratno. Primer generiranih podatkov iz logaritemsko normalne porazdelitve za primer skladnih razlik je prikazan na sliki 3.6.



Slika 3.6: Prikaz porazdelitve generiranega izida Y za primer skladnih razlik: prikaz porazdelitve z nakopičenimi vrednostmi (levo) in prikaz zveznega dela porazdelitve (desno).

3.4.1 Lastnosti generiranih podatkov

Pri tako generiranih podatkih so razlike v deležih nakopičenih vrednosti in povprečjih zveznih delov med ravnema pojasnjevalne spremenljivke X primerljive za različne velikosti vzorca in izbrane deleže nakopičenih vrednosti. Pri generiranju skladnih razlik, skladnost pada z manjšanjem velikosti vzorca, podobno se dogaja z neskladnostjo pri generiranju neskladnih razlik. Lastnosti podatkov prikazuje tabela 3.4.

	skladne razlike, n=500						
	delež nakopičenih vrednosti			povprečje zveznega dela			skladnost
Delež	A	B	A-B	A	B	B-A	(%)
0,25	0,28	0,23	0,05	1,0	1,3	0,3	86,8
0,4	0,43	0,38	0,05	1,0	1,3	0,3	84,1
0,6	0,62	0,58	0,05	1,0	1,3	0,3	84,0
0,8	0,83	0,78	0,05	1,0	1,3	0,3	89,6
			n=100				
0,25	0,28	0,24	0,04	1,0	1,3	0,3	62,1
0,4	0,43	0,39	0,04	1,0	1,3	0,3	60,4
0,6	0,63	0,59	0,03	1,0	1,3	0,3	60,4
0,8	0,83	0,79	0,04	1,0	1,3	0,3	56,5
	neskladne razlike, n=500						
0,25	0,23	0,28	-0,05	1,0	1,3	0,3	7,4
0,4	0,38	0,43	-0,05	1,0	1,3	0,3	11,3
0,6	0,57	0,63	-0,05	1,0	1,3	0,3	10,1
0,8	0,77	0,83	-0,05	1,0	1,3	0,3	7,6
			n=100				
0,25	0,23	0,29	-0,06	1,0	1,3	0,3	20,5
0,4	0,38	0,44	-0,06	1,0	1,3	0,3	27,9
0,6	0,58	0,64	-0,06	1,0	1,3	0,3	27,2
0,8	0,78	0,84	-0,06	1,0	1,3	0,3	25,4

Tabela 3.4: Lastnosti generiranih podatkov ob predpostavljenih skladnih oz. neskladnih razlikah med ravnema pojasnjevalne spremenljivke X za velikost vzorca 100 in 500; pridobljeno iz simulacij s 1000 ponovitvami.

Dodatno smo preverili ali pri načinu generiranja iz logaritemsko normalne porazdelitve lahko sklepamo o prisotnosti sorazmernih obetov v podatkih. Po definiciji modela sorazmernih obetov (glej podpoglavje 2.3) preverimo, ali je razmerje kumulativnih obetov za raven A in B enako za vse kategorije j izida Y:

$$\frac{obet(Y \leq j|x = A)}{obet(Y \leq j|x = B)} = e^{\beta}. \tag{3.4}$$

Enačba 3.4 v primeru sorazmernih obetov vrne enak rezultat, ne glede na izbrano mejo j. Pri izračunih smo uporabili iste meje, na podlagi katerih smo razvrstili vrednosti izida Y v kategorije pred prileganjem modela sorazmernih obetov. Vrednost $j_1 = 0$, vrednosti j_2 in j_3 pa predstavljata prvi in drugi tercil pozitivnih vrednosti izida Y. Za vsako mejo j izračunamo cOR (okr. cumulative odds ratio):

$$cOR_j = \frac{P(Y \leq j|x = A)}{1 - P(Y \leq j|x = A)} \Big/ \frac{P(Y \leq j|x = B)}{1 - P(Y \leq j|x = B)}. \tag{3.5}$$

Na podlagi izračunanih vrednosti cOR sklepamo, da sorazmerni obeti za tako generirane podatke ne veljajo, saj se rezultati za izbrane tri meje bistveno razlikujejo, česar vsaj ob predpostavljenih skladnih razlikah nismo pričakovali. Izračuni za vsak delež nakopičenih vrednosti so podani v tabeli 3.5. Izračune smo potrdili tudi s simulacijo, v kateri ocenimo razmerje kumulativnih obetov za raven A in B iz enačbe 3.5 na generiranih podatkih. Generiramo veliko vrednosti izida Y ($n = 10000$), ga kategoriziramo in izračunamo cOR_j preko števila vrednosti Y, ki so manjše od posamezne meje j za vsako raven spremenljivke X.

	skladne razlike			neskladne razlike		
Delež	cOR1	cOR2	cOR3	cOR1	cOR2	cOR3
0,25	1,3	2,2	2,6	0,8	1,7	2,2
0,4	1,2	2,0	2,5	0,8	1,5	2,0
0,6	1,2	1,9	2,5	0,8	1,3	1,8
0,8	1,4	2,0	2,7	0,7	1,1	1,5

Tabela 3.5: Razmerja kumulativnih obetov ravni A in B za tri meje j in različne deleže nakopičenih vrednosti za generiranje iz logaritemsko normalne porazdelitve.

S simulacijami, kjer podatke generiramo na opisan način iz logaritemsko normalne porazdelitve, torej preverjamo delovanje izbranih metod za primer skladnih oz. neskladnih razlik na podatkih, kjer so prisotna nesorazmerja.

Zgoraj opisan način preverjanja enakosti razmerja kumulativnih obetov za ravni A in B, nam z izenačitvijo enačbe 3.5 za j_1 in j_2, omogoča izpeljavo potrebne povprečne vrednosti izida Y za raven B, da sorazmerni obeti veljajo. Za primeren razpon moči obravnavanih testov povečamo razliko v deležih nakopičenih vrednosti za raven A in B. Za povprečno vrednost ravni B, $\mu_B = 0{,}078$, so v tabeli 3.6 prikazana razmerja kumulativnih obetov za predpostavljene skladne razlike med ravnema A in B.

Delež	cOR1	cOR2	cOR3
0,25	1,7	1,6	1,5
0,4	1,5	1,5	1,5
0,6	1,5	1,6	1,6
0,8	1,9	2,0	2,1

Tabela 3.6: Razmerja kumulativnih obetov ravni A in B za tri meje j in različne deleže nakopičenih vrednosti za generiranje sorazmernih obetov iz logaritemsko normalne porazdelitve.

Dodatno simuliramo še podatke, za katere sorazmerni obeti veljajo in predpostavljamo skladne razlike med ravnema pojasnjevalne spremenljivke X. Lastnosti generiranih podatkov za vsak delež nakopičenih vrednosti so prikazane v tabeli

3.7. Iz tabele 3.6 sklepamo, da so razlike med ravnema A in B pojasnjevalne spremenljivke večje za delež nakopičenih vrednosti 0,8.

	n=500						
	delež nakopičenih vrednosti			povprečje zveznega dela			skladnost
Delež	A	B	A-B	A	B	B-A	(%)
0,25	0,30	0,20	0,10	1,0	1,1	0,1	87,8
0,4	0,45	0,35	0,10	1,0	1,1	0,1	84,0
0,6	0,65	0,55	0,10	1,0	1,1	0,1	80,0
0,8	0,85	0,75	0,10	1,0	1,1	0,1	72,2
	n=100						
0,25	0,31	0,21	0,10	1,0	1,1	0,1	62,3
0,4	0,46	0,36	0,10	1,0	1,1	0,1	61,6
0,6	0,66	0,57	0,09	1,0	1,1	0,1	57,0
0,8	0,86	0,76	0,09	1,0	1,1	0,1	54,7

Tabela 3.7: Lastnosti generiranih podatkov, za katere predpostavljamo skladne razlike in sorazmernost v podatkih.

3.5 Primerjava generiranih podatkov iz obeh načinov

V prvem delu simulacij preko latentne spremenljivke generiramo tako, da sorazmerni obeti veljajo, v drugem delu simulacij, kjer generiramo preko različnih vrednosti β, pa predpostavljamo, da sorazmerni obeti v podatkih ne veljajo. Sorazmernost za omenjena primera lahko preverimo s simulacijami, opisanimi v razdelku 3.4.1, če podatke znotraj simulacije generiramo preko latentne spremenljivke z enakimi parametri, kot v preučevanih simulacijah. Rezultati za predpostavljene manjše razlike med ravnema pojasnjevalne spremenljivke so podani v tabeli 3.8.

Iz tabele 3.8 je razvidno, da smo v prvem delu simulacij preko latentne spremenljivke zares generirali podatke, v katerih sorazmerni obeti veljajo. Za drugi del simulacij sklepamo, da lastnost nesorazmernosti izhaja predvsem iz cOR_3, vendar ob primerjavi s tabelo 3.5 nesorazmernost ni tako velika.

	predpostavljeni sorazmerni obeti			niso predpostavljeni sorazmerni obeti		
Delež	cOR1	cOR2	cOR3	cOR1	cOR2	cOR3
0,25	1,7	1,7	1,7	3,4	3,3	3,7
0,4	1,7	1,7	1,7	3,3	3,4	3,8
0,7	1,7	1,7	1,7	3,6	3,8	4,2
0,9	1,7	1,7	1,7	4,2	4,4	4,6

Tabela 3.8: Razmerja kumulativnih obetov ravni A in B za tri meje j in različne deleže nakopičenih vrednosti, za primer generiranja preko latentne spremenljivke, ko predpostavljamo sorazmerne obete (levo) in ko teh ne predpostavljamo (desno).

Generirane razlike med ravnema pojasnjevalne spremenljivke X so približno enake za generiranje iz logaritemsko normalne porazdelitve in za generiranje preko latentne spremenljivke, ko predpostavljamo manjše razlike med A in B. Lastnosti generiranih podatkov preko latentne spremenljivke, ko so predpostavljene manjše razlike so predstavljene v Dodatku A.

Od generiranih razlik iz logaritemsko normalne porazdelitve so malo večje generirane razlike preko latentne spremenljivke, ko predpostavljamo, da sorazmerni obeti ne veljajo, vendar pa je večja tudi razpršenost podatkov.

3.6 Opis analize pravih podatkov

Pri analizi podatkov o virusih nas zanima povezanost predhodno postavljene diagnoze otroka s koncentracijo virusa. Diagnozo otroka v podatkih opiše spremenljivka *diagnoza*, ki ima štiri možne vrednosti: FS, AB, AGE in C. Koncentracijo virusa v podatkih opiše spremenljivka *ct*, ki označuje število potrebnih ciklov RT-PCR za zaznavo virusa v vzorcu. Po 41 ciklih je bila reakcija prekinjena in sklenili smo, da virus v vzorcu ni prisoten. Podatki o virusih so podrobneje opisani v podpoglavju 1.1.

Podatke smo analizirali s programom R, verzija 3.2.3 [33]. Pred analizo z obravnavanimi metodami smo podatke najprej pripravili, tako da se je identifikacija okuženosti z virusom (vrednosti "Pos"in "Neg") ujemala z vrednostjo spremenljivke *ct*. Ob negativni identifikaciji virusa je vrednost *ct* enaka 41, medtem ko je ob pozitivni identifikaciji, vrednost *ct* med 0 in 41.

Ker nas pri analizi zanima primerjava več skupin diagnoze, Mann-Whitneyevega testa ne moremo uporabiti, saj omogoča le primerjavo porazdelitev dveh skupin. Za analizo uporabimo njegovo različico za primerjavo porazdelitev več skupin - Kruskal-Wallisov test.

Z vsako obravnavano metodo preverimo globalno statistično značilnost povezanosti diagnoze s koncentracijo virusa. Za regresijska pristopa, model Tobit in model sorazmernih obetov, v ta namen uporabimo test razmerja verjetij, ki primerja model brez spremenljivke *diagnoza* z modelom, ki vsebuje spremenljivko *diagnoza*. Pri modelu Log+Lin za vsoto testnih statistik predpostavljamo χ^2_6 porazdelitev. V primerih virusov hMPV in InfV za kontrolno skupino ni pozitivnih posameznikov, zato predpostavljamo, da je vsota posameznih testnih statistik porazdeljena po χ^2_5 porazdelitvi.

Statistično značilnost primerjav posameznih parov skupin diagnoze pridobimo iz modelov. Pri Kruskal-Wallisovem testu v ta namen uporabimo Dunnov test mnogoterih primerjav.

4 Rezultati

V prvem delu poglavja predstavljamo rezultate simulacij, s katerimi preučujemo značilnosti izbranih pristopov za analizo podatkov ob prisotnosti nakopičenih vrednosti. Najprej predstavimo rezultate simulacij, kjer smo podatke generirali preko latentne spremenljivke, za primera veljavnih in neveljavnih sorazmernih obetov. Ker način generiranja ne omogoča, da bi povsem nadzorovali lastnosti generiranih podatkov, smo v dodatnih simulacijah generirali še iz logaritemsko normalne porazdelitve. Z dodatnimi simulacijami preverimo rezultate iz prvih simulacij, omogočajo pa tudi bolj natančno preučevanje lastnosti, ki nas zanimajo. Rezultate predstavimo za primera prisotnosti skladnih in neskladnih razlik. Izkaže se, da za tako generirane podatke sorazmerni obeti ne veljajo. Rezultate primerjamo še z rezultati iz simulacije, v kateri smo generirali s takšnimi parametri, da sorazmerni obeti veljajo. Za ta primer predpostavljamo skladne razlike med ravnema pojasnjevalne spremenljivke. Podrobnosti o načinu simuliranja podatkov so podane v poglavju 3.

Dodatno predstavimo še rezultate delovanja testa, s katerim preverjamo veljavnost predpostavke o sorazmernih obetih. Ničelna domneva, ki jo preverjamo s tem testom pravi, da sorazmerni obeti v podatkih veljajo. Za ta test simuliramo ob veljavni ničelni domnevi pri prvem delu generiranja preko latentne spremenljivke, ko sorazmerni obeti veljajo.

V drugem delu poglavja predstavljamo rezultate analize pravih podatkov o prisotnosti virusov pri različno diagnosticiranih otrocih in rezultate primerjamo z rezultati simulacij.

4.1 Rezultati simulacij

S simulacijami preučujemo delovanje metod v dveh situacijah. V prvi situaciji smo podatke generirali tako, da ni povezanosti med pojasnjevalno spremenljivko in izidom (torej ob veljavni ničelni domnevi), v drugi situaciji pa smo predpostavili povezanost med pojasnjevalno spremenljivko in izidom (generiramo pod alternativno domnevo).

Bolj formalno lahko zapišemo ničelno in alternativno domnevo kot:

H_0 : Spremenljivka X ni povezana z izidom Y v populaciji.

H_A : Obstaja povezanost med spremenljivko X in izidom Y v populaciji.

Delovanje metod preverjamo preko ocenjenega deleža zavrnjenih ničelnih domnev iz simulacij. S simulacijami ob veljavni ničelni domnevi ocenjujemo velikost testov, ki je ob pravilnem delovanju metode približno enaka predpostavljeni stopnji značilnosti $\alpha = 0{,}05$. S simulacijami, kjer generiramo podatke pod alternativno domnevo, pa ocenjujemo moč testa, ki kaže na zmožnost metode, da zaznava razlike, ko le te obstajajo. Večja moč pomeni večjo verjetnost zavrnitve ničelne domneve v situacijah, kjer ta ne velja.

4.1.1 Generiranje preko latentne spremenljivke

Preko latentne spremenljivke izid generiramo enkrat na način, da sorazmerni obeti veljajo (generiramo ob veljavni ničelni domnevi in pod alternativno domnevo), drugič pa na način, ko sorazmerni obeti ne veljajo. Ker nesorazmerja izvirajo iz razlik med ravnema opisne spremenljivke X, v tem primeru simuliramo le pod alternativno domnevo. Sledi predstavitev velikosti in moči testov, ko sorazmerni obeti veljajo, nato pa še moči testov, ko sorazmerni obeti ne veljajo. Številski rezultati so predstavljeni v dodatku B.

4.1.1.1 Sorazmerni obeti veljajo

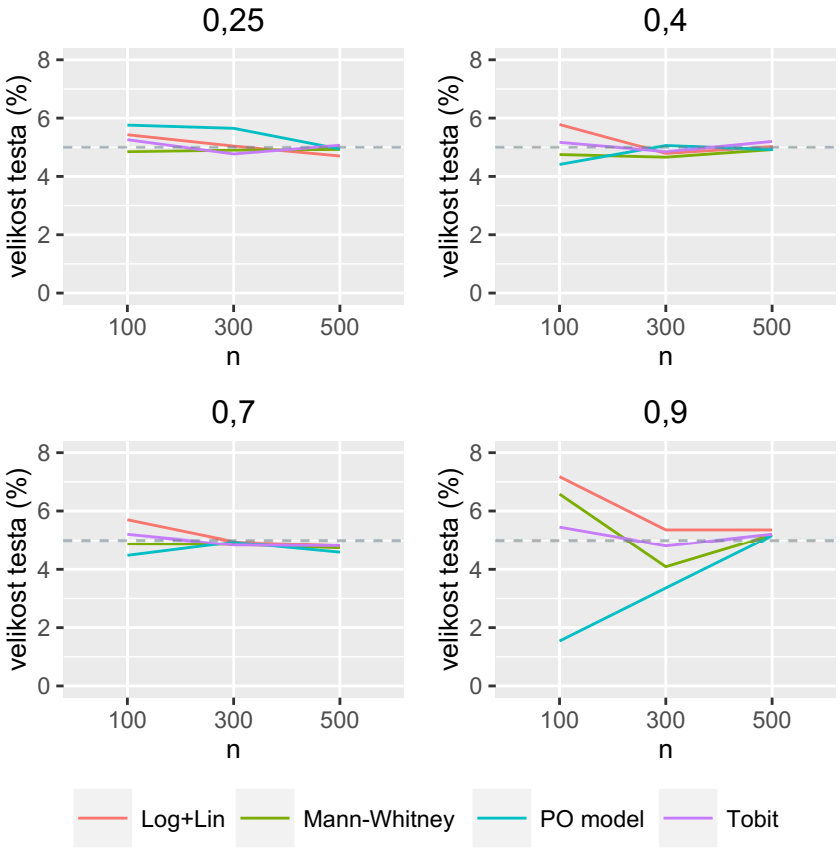
V primeru, ko simuliramo tako, da sorazmerni obeti veljajo, smo vse simulacije ponovili 10000-krat. 95% interval zaupanja za ocenjeno velikost testa ob veljavni ničelni domnevi izračunamo preko formule za delež p:

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{k}}, \tag{4.1}$$

kjer je $\hat{p}$ ocenjen delež zavrnjenih ničelnih domnev iz simulacij, k število ponovitev simulacij, z pa kvantil iz standardizirane normalne porazdelitve. Ocenjeni velikosti testa $\hat{p}$ priredimo interval zaupanja. Če predpostavljena $\alpha = 0{,}05$ ne leži znotraj intervala, metoda v dani situaciji ne deluje dobro. V primeru, ko je celoten interval manjši od α, označimo metodo za konzervativno. Delovanje metode označimo za anti-konzervativno, ko je celoten interval zaupanja za ocenjeno velikost testa večji od vrednosti α. Slednje pomeni, da metoda prevečkrat zavrne ničelno domnevo, ko ta velja in je tveganje za napako 1. vrste v resnici večje, kot pa ga predpostavljamo.

Velikosti testov pod ničelno domnevo

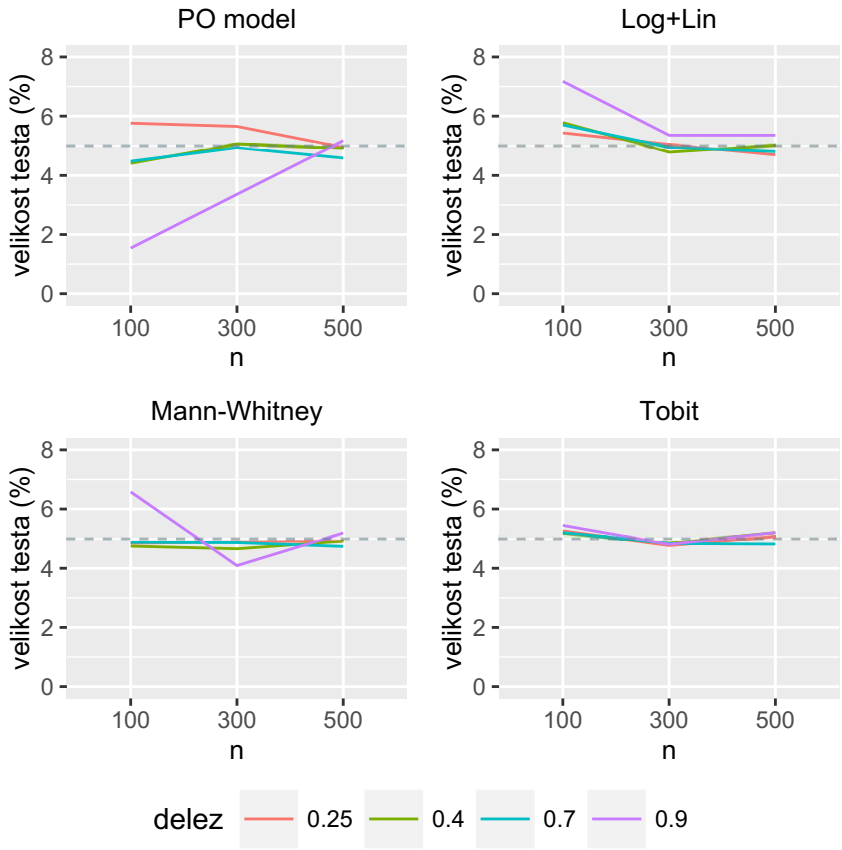
Velikosti testov za vsak delež nakopičenih vrednosti prikazuje slika 4.1. Za vse deleže nakopičenih vrednosti se velikost testov bliža vrednosti 0,05 z večanjem velikosti vzorca. Velikosti vseh testov so približno enake 0,05 za $n = 500$. Najslabšo velikost testi dosegajo pri visokem deležu nakopičenih vrednosti (delež 0,9), kjer za vse velikosti vzorca ustrezno velikost ohrani le model Tobit.



Slika 4.1: Velikost testov, ko ni razlik med ravnema pojasnjevalne spremenljivke X, za vsak delež nakopičenih vrednosti.

Ob veljavni ničelni domnevi le model Tobit ohrani ustrezno velikost v vseh preizkušenih situacijah. Mann-Whitneyev test deluje slabše pri deležu nakopičenih vrednosti 0,9, kjer se obnaša nepredvidljivo, medtem ko ima v ostalih situacijah

ustrezno - rahlo konzervativno delovanje. Model Log+Lin ima slabše delovanje pri manjši velikosti vzorca (n=100), pri deležu nakopičenih vrednosti 0,9 pa se obnaša anti-konzervativno. Od vseh metod se model sorazmernih obetov obnaša najbolj nepredvidljivo. Pri deležu nakopičenih vrednosti 0,25 je njegovo delovanje anti-konzervativno (velikost testa pri n=100 znaša 0,058 in pri n=300 0,057), pri deležu nakopičenih vrednosti 0,9 pa postane močno konzervativen (velikost testa pri n=100 znaša 0,015 in pri n=300 0,034). Velikosti testov za vsak test posebej prikazuje slika 4.2.

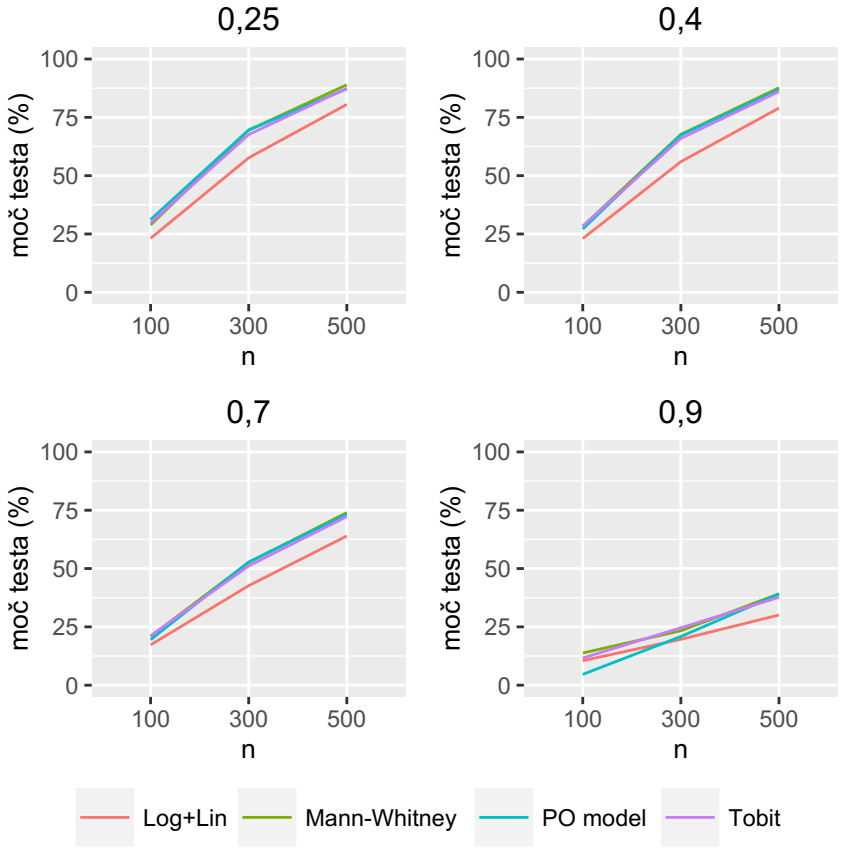


Slika 4.2: Velikost testov, ko ni razlik med ravnema opisne spremenljivke X, za vsako od preučevanih metod.

Moči testov pod alternativno domnevo

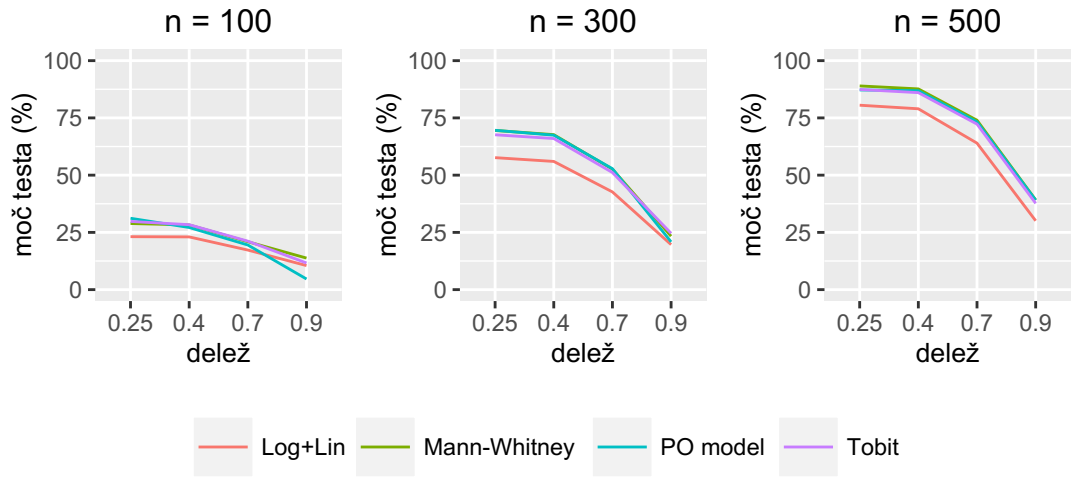
Pod alternativno domnevo smo simulirali manjše in večje razlike med ravnema pojasnjevalne spremenljivke X in rezultati se ujemajo za obe razliki. Pri večjih predpostavljenih razlikah so moči testov pričakovano večje, a hitreje padajo z večanjem deleža nakopičenih vrednosti, kar je posledica bolj izrazitega zmanjševanja generiranih razlik v zveznem delu za ta primer.

Predstavljeni so rezultati simulacij, kjer smo predpostavili manjše razlike med ravnema A in B. Moči testov ob predpostavljenih sorazmerjih prikazuje slika 4.3.



Slika 4.3: Moč testov za podatke generirane preko latentne spremenljivke, predpostavljeni sorazmerni obeti in manjše razlike, za različne deleže nakopičenih vrednosti.

Pri vseh deležih nakopičenih vrednosti vsem testom moč narašča z večjim vzorcem. Enodelni testi imajo v vseh situacijah višjo moč kot model Log+Lin. Zaradi močne konzervativnosti modela sorazmernih obetov pri deležu nakopičenih vrednosti 0,9 in velikosti vzorca n=100 ob veljavni ničelni domnevi, ima test tudi pod alternativno domnevo zelo nizko moč v tej situaciji - celo nižjo kot model Log+Lin.



Slika 4.4: Moč testov za podatke generirane preko latentne spremenljivke, predpostavljeni sorazmerni obeti in manjše razlike, za posamezne velikosti vzorca.

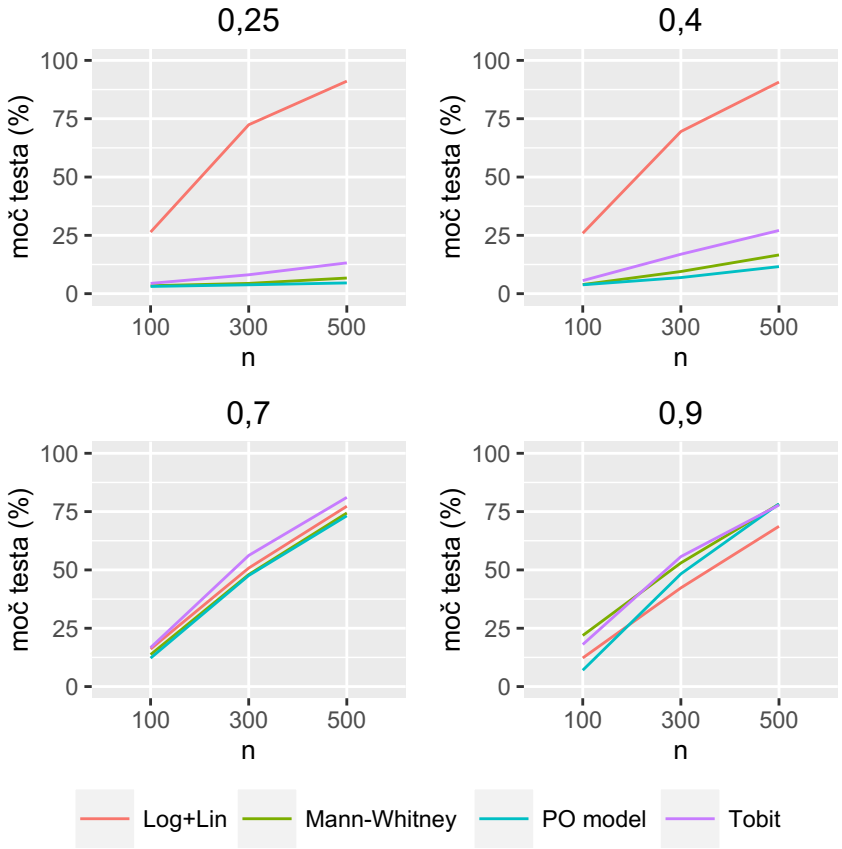
Z večanjem deleža nakopičenih vrednosti moč vsem testom pada, kar je posledica manjšanja števila vrednosti v zveznem delu porazdelitve izida. Poleg tega pa je manjša moč lahko tudi posledica načina generiranja, pri katerem se z večanjem deleža nakopičenih vrednosti zmanjšujejo razlike med ravnema pojasnjevalne spremenljivke X, za oba dela porazdelitve. Razlike so v vseh situacijah večinoma skladne, razen pri večjih deležih nakopičenih vrednosti in manjši velikosti vzorca, kar bi lahko bil razlog za bolj primerljivo delovanje modela Log+Lin z ostalimi metodami v teh situacijah.

Ko pogledamo delovanje posameznih komponent modela Log+Lin, ima del "Lin"v vseh situacijah manjšo moč od dela "Log", tako za manjše kot za večje predpostavljene razlike med ravnema pojasnjevalne spremenljivke. Pri deležih

nakopičenih vrednosti 0,7 in 0,9 del "Lin"praktično ne zaznava več razlik, kar povzroči, da je moč kombinacije Log+Lin nižja kot pa posamezna moč dela "Log".

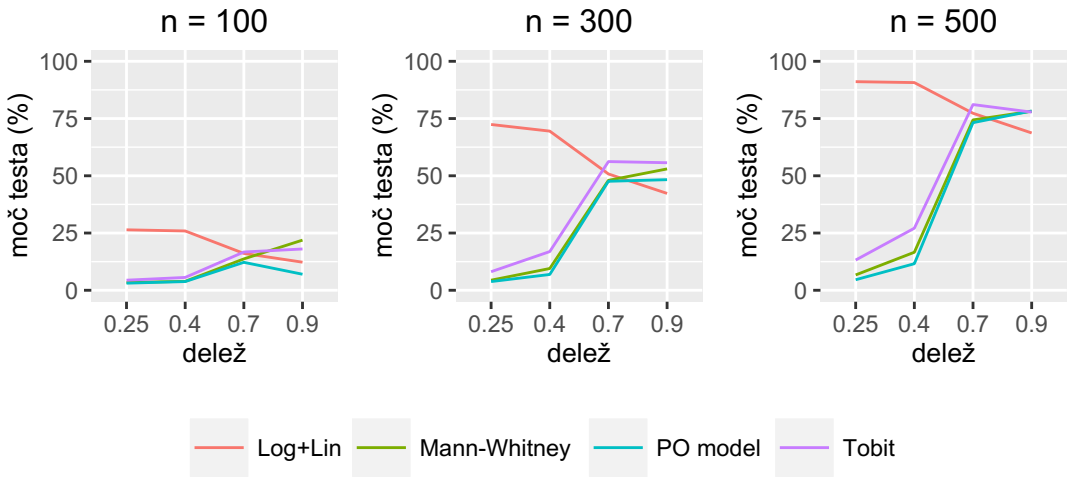
4.1.1.2 Sorazmerni obeti ne veljajo

V tem razdelku so predstavljeni rezultati simulacij preko latentne spremenljivke, kjer podatke generiramo pod alternativno domnevo. V podatkih so prisotna manjša nesorazmerja. Moči testov so prikazane na sliki 4.5.



Slika 4.5: Moč testov za podatke generirane preko latentne spremenljivke, predpostavljeni nesorazmerni obeti in manjše razlike, za različne deleže nakopičenih vrednosti.

Generirani podatki za ta primer se precej razlikujejo v lastnostih med različnimi deleži nakopičenih vrednosti. Pri deležih 0,25 in 0,4 so razlike med ravnema opisne spremenljivke X izrazito neskladne. Za omenjena deleža nakopičenih vrednosti ima model Log+Lin zelo veliko moč, medtem ko je moč enodelnih testov zelo nizka. Izmed enodelnih testov ima model Tobit najvišjo moč, model sorazmernih obetov pa najnižjo. Pri deležih nakopičenih vrednosti 0,7 in 0,9 postanejo generirane razlike med ravnema pojasnjevalne spremenljivke X skladne in enodelnim testom se moč poveča, medtem ko se moč modela Log+Lin zmanjša. Pojav prikazuje slika 4.6.



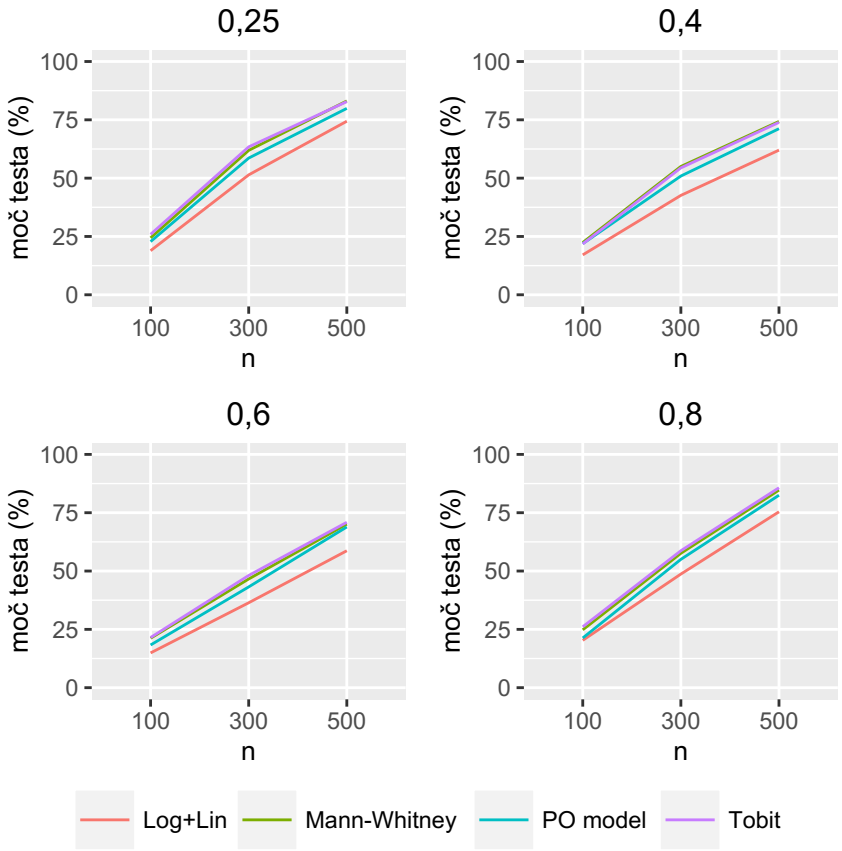
Slika 4.6: Moč testov za podatke generirane iz latentne spremenljivke, predpostavljeni nesorazmerni obeti in manjše razlike, za posamezne velikosti vzorca.

Ocenjena razmerja kumulativnih obetov (tabela 3.8) ne nakazujejo večjih razlik v sorazmerjih za različne deleže nakopičenih vrednosti za ta primer generiranja podatkov. Večja razlika v delovanju metod je torej posledica prehoda neskladnih razlik v skladne pri tem načinu generiranja.

4.1.2 Generiranje iz logaritemsko normalne porazdelitve

Iz logaritemsko normalne porazdelitve generiramo podatke, kjer sorazmerni obeti ne veljajo, za predpostavljene skladne in neskladne razlike med ravnema pojasnjevalne spremenljivke X. Dodatno generiramo še podatke, kjer predpostavljamo sorazmernost in skladne razlike. Tako lahko ocenimo vpliv skladnosti in sorazmernosti na delovanje obravnavanih metod. Številski rezultati so predstavljeni v dodatku C.

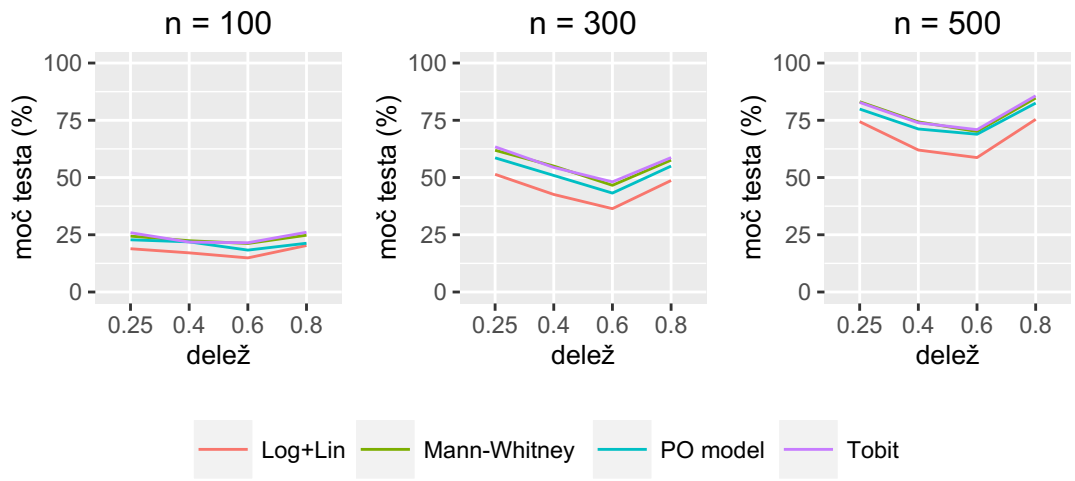
4.1.2.1 Sorazmerni obeti veljajo



Slika 4.7: Moč testov za podatke generirane iz logaritemsko normalne porazdelitve, predpostavljeni sorazmerni obeti in skladne razlike, za različne deleže nakopičenih vrednosti.

Ko generiramo iz logaritemsko normalne porazdelitve tako, da sorazmerni obeti veljajo, vidimo podobno sliko moči (slika 4.7) kot v primeru simuliranja preko latentne spremenljivke, ko sorazmerni obeti veljajo. V obeh primerih je moč v vseh situacijah nižja za model Log+Lin, medtem ko je moč modela sorazmernih obetov bližja močem ostalih dveh enodelnih testov za primer generiranja podatkov preko latentne spremenljivke.

Na sliki 4.8 vidimo nenadno povišanje moči pri deležu nakopičenih vrednosti 0,8, ki je za $n = 500$ celo večja od moči testov pri deležu nakopičenih vrednosti 0,25. Povišanje moči je posledica večjih generiranih razlik med ravnema pojasnjevalne spremenljivke X za to situacijo.

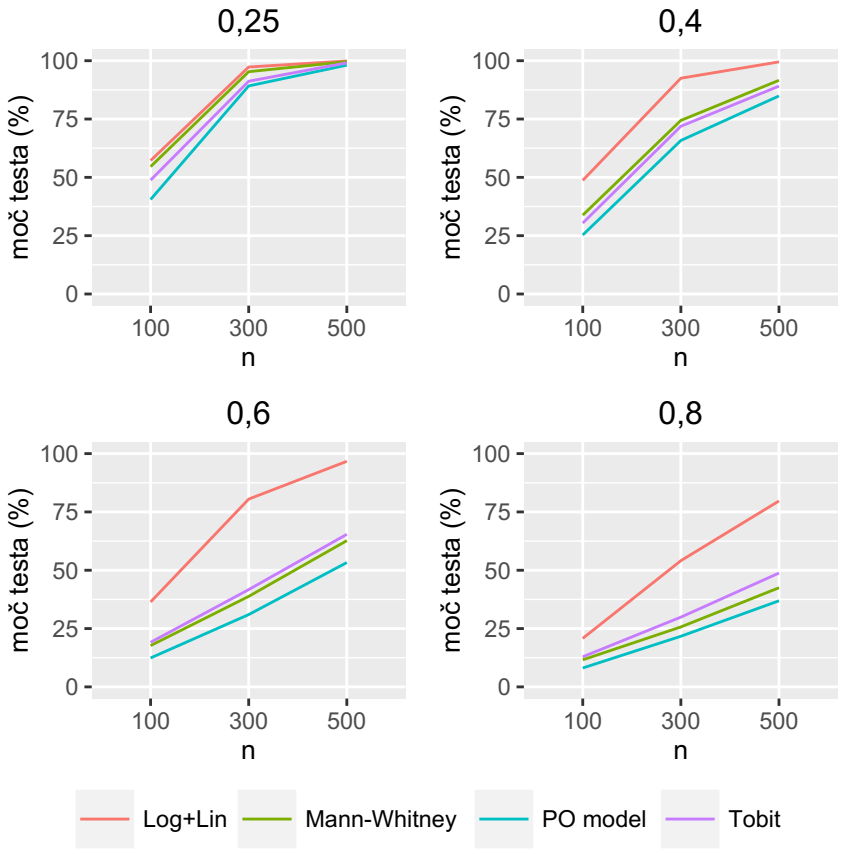


Slika 4.8: Moč testov za podatke generirane iz logaritemsko normalne porazdelitve, predpostavljeni sorazmerni obeti in skladne razlike, za velikosti vzorca.

4.1.2.2 Sorazmerni obeti ne veljajo

Moči testov, ko simuliramo skladne razlike

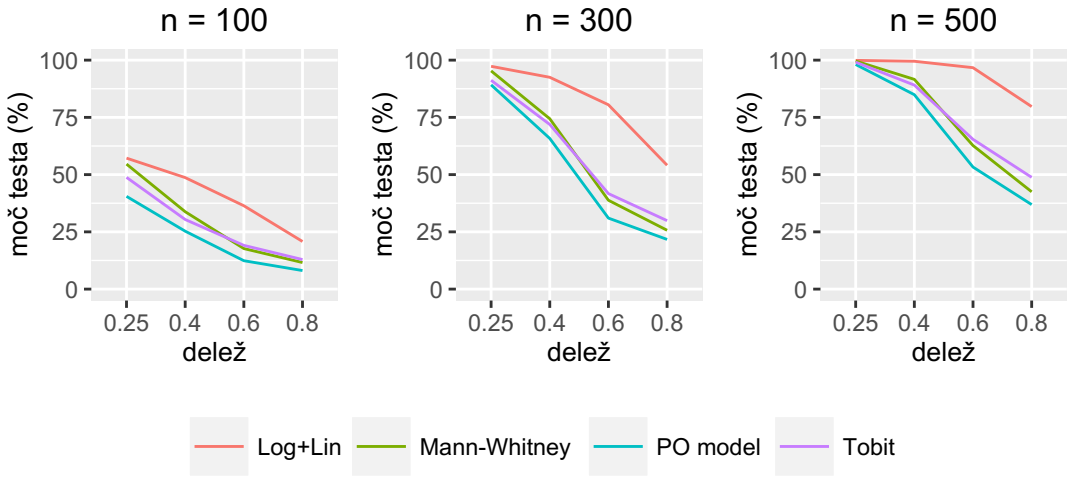
Moči testov so predstavljene na sliki 4.9. Za vse situacije je moč modela Log+Lin največja. Enodelni testi so konkurenčni pri deležu nakopičenih vrednosti 0,25, z večanjem deleža pa se razlika med modelom Log+Lin veča. Sklepamo, da je to posledica pojava nesorazmernih obetov v podatkih.



Slika 4.9: Moč testov za podatke generirane iz logaritemsko normalne porazdelitve, predpostavljeni nesorazmerni obeti in skladne razlike, za različne deleže nakopičenih vrednosti.

Najmanjšo moč pri vseh deležih nakopičenih vrednosti ima model sorazmernih obetov. Mann-Whitneyev test in model Tobit imata približno enako moč za vse situacije, pri čemer ima rahlo prednost pri nižjih dveh deležih nakopičenih vrednosti Mann-Whitneyev test, pri višjih dveh deležih pa model Tobit.

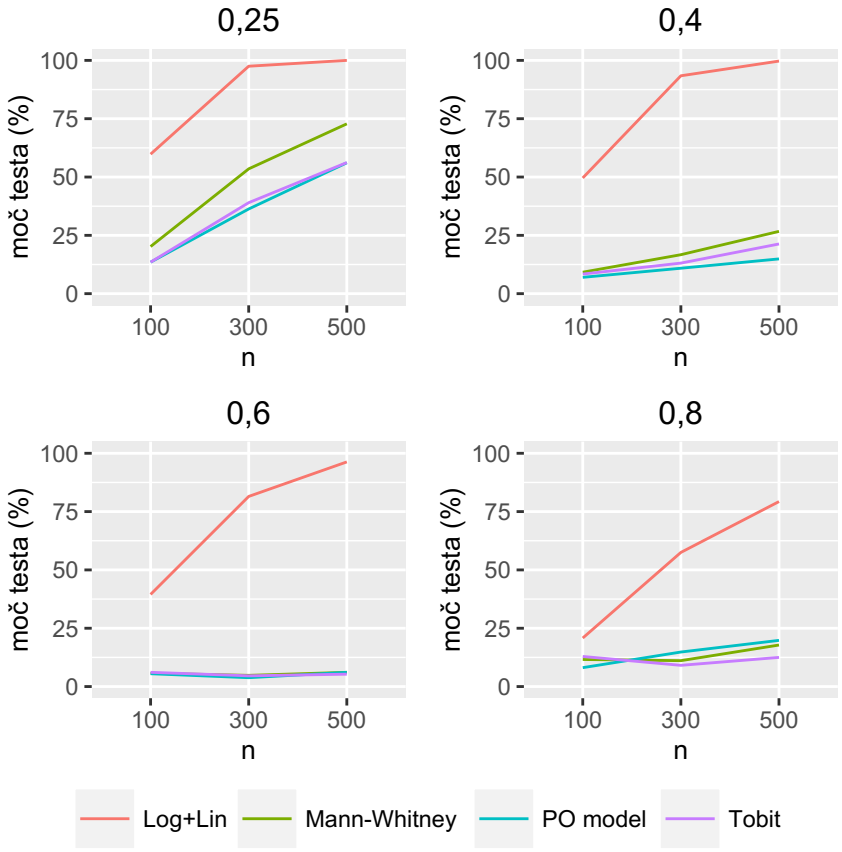
Moč vsem testom pada z večanjem deleža nakopičenih vrednosti. Moč enodelnih testov hitreje pade pri deležu nakopičenih vrednosti 0,4 in 0,6 kot modelu Log+Lin, medtem ko moč modela Log+Lin v primerjavi z ostalimi metodami hitreje pade pri deležu 0,8. Pojav je prikazan na sliki 4.10.



Slika 4.10: Moč testov za podatke generirane iz logaritemsko normalne porazdelitve, predpostavljeni nesorazmerni obeti in skladne razlike, za različne velikosti vzorca.

Moči testov, ko simuliramo neskladne razlike

Moč modela Log+Lin se v primerjavi s situacijo, kjer generiramo skladne razlike, ohrani, medtem ko se moč ostalih treh testov močno zniža in razlike med modelom Log+Lin in enodelnimi testi se še povečajo. Najslabšo moč ima v večini situacij model sorazmernih obetov, kar prikazuje slika 4.11. Izmed enodelnih testov ima najvišjo moč Mann-Whitneyev test.

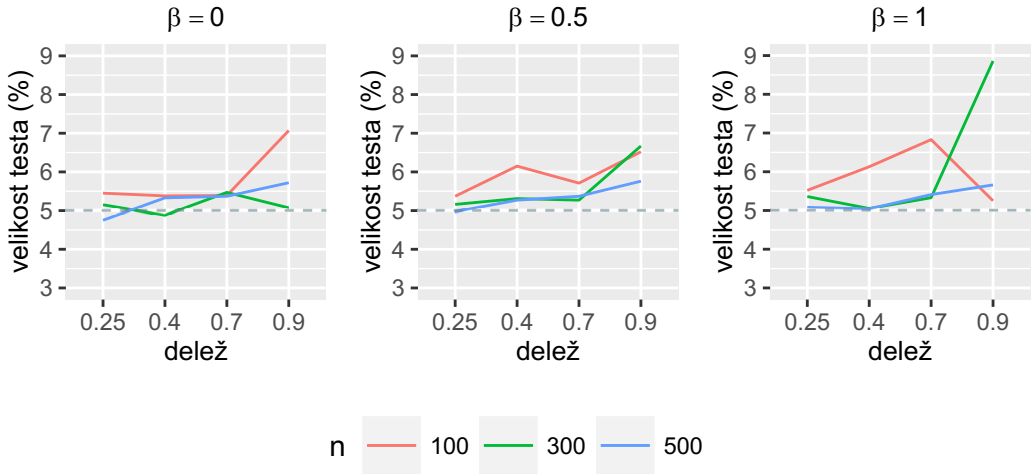


Slika 4.11: Moč testov za podatke generirane iz logaritemsko normalne porazdelitve, predpostavljeni nesorazmerni obeti in neskladne razlike, za različne deleže nakopičenih vrednosti.

4.1.3 Test za preverjanje predpostavke o sorazmernih obetih

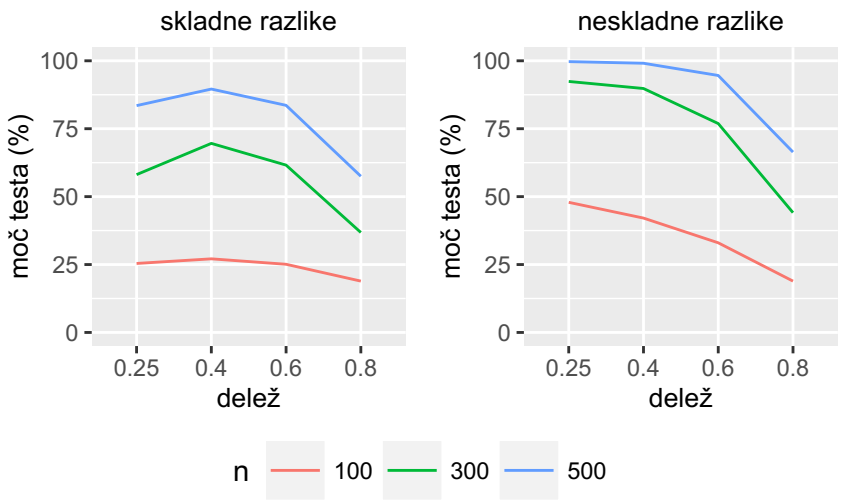
Test preverja ničelno domnevo, ki pravi, da sorazmerni obeti v podatkih veljajo. Velikosti testa smo pridobili iz rezultatov simuliranja preko latentne spremenljivke, ko smo predpostavili, da sorazmerni obeti veljajo. Moči testa pa preučujemo preko simulacij, kjer generiramo iz logaritemsko normalne porazdelitve, saj smo za ta primer bolj prepričani, da generiramo pod alternativno domnevo - da sorazmerni obeti v podatkih ne veljajo. Številski rezultati so predstavljeni v dodatku D.

Velikost testa pod ničelno domnevo smo preverili za tri situacije: ko ni razlik med ravnema pojasnjevalne spremenljivke X in ko razlike obstajajo (manjše in večje razlike). Rezultati so prikazani na sliki 4.12. Test ima za manjše deleže nakopičenih vrednosti (0,25, 0,4 in 0,7) in večje velikosti vzorca (n=300, 500) ustrezno velikost, čeprav rahlo anti-konzervativno. Pri velikosti vzorca n=100 je delovanje testa še bolj anti-konzervativno. Pri deležu nakopičenih vrednosti 0,9 test deluje nezanesljivo.



Slika 4.12: Velikost testa, s katerim preverjamo veljavnost predpostavke o sorazmernih obetih, pod ničelno domnevo, ko med ravnema opisne spremenljivke X ni razlike (levo) in ko predpostavljamo manjše (sredina) in večje razlike (desno).

Moči testa so prikazane na sliki 4.13. Leva slika prikazuje moči testa, ko generiramo skladne razlike med ravnema spremenljivke X, desna slika pa moči testa, ko generiramo neskladne razlike. Moči so večje za slednje, kar bi lahko pomenilo, da z neskladnostjo v podatke vnašamo tudi nesorazmernost. Moč je večja pri večji velikosti vzorca in se počasi manjša z večanjem deleža nakopičenih vrednosti, za deleža 0,25 in 0,4 je moč približno enaka.



Slika 4.13: Moč testa za ugotavljanje sorazmernih obetov, ko predpostavimo skladne (levo) in neskladne (desno) razlike med ravnema opisne spremenljivke X.

Za večjo velikost vzorca je moč testa dobra, vendar se za večji delež nakopičenih vrednosti precej zmanjša. Pri majhni velikosti vzorca obstaja nevarnost sklepanja, da predpostavka o sorazmernih obetih ni kršena, čeprav sorazmerni obeti na podatkih ne veljajo.

4.2 Analiza pravih podatkov

Pri analizi pravih podatkov ugotavljamo povezanost predhodno postavljene diagnoze otroka s koncentracijo virusa. Koncentracijo virusa v podatkih opisuje spremenljivka *ct*, ki predstavlja število potrebnih ciklov RT-PCR, da zaznamo virus v vzorcu. Manjše kot je število ciklov, večja je izraženost virusa v vzorcu in obratno. Možne so 4 različne predhodno določene skupine diagnoz: otroci z akutnim bronhiolitisom (AB), otroci z akutnim gastroenteritisom (AGE) , otroci z vročinskimi krči (FS) in kontrolna skupina (C). Osredotočili se bomo na primerjavo otrok z vročinskimi krči z ostalimi skupinami.

V podatkih imamo 617 opazovanj ($n = 617$). Tabela 4.1 prikazuje število pozitivnih opazovanj za vsak virus (vrednosti spremenljivke *ct* manjše od 41) po skupinah diagnoze. V zadnjih dveh stolpcih tabele sta prikazana skupno število pozitivnih opazovanj in odstotek nakopičenih vrednosti v spremenljivki *ct*.

virus	AB (n=155)	AGE (n=114)	FS (n=192)	C (n=156)	otroci z dokazanim virusom (n=617)	skupen % nakopičenih vrednosti
RSV	82	13	21	1	117	81
hMPV	8	4	4	0	16	97
InfV	3	5	28	0	36	94
HCoV	12	11	19	3	45	93
HBoV	21	12	20	9	62	90
hRV	37	18	26	17	98	84
PIV	7	7	20	5	39	94
AdV	18	22	30	15	85	86

Tabela 4.1: Število pozitivnih opazovanj po skupinah za vsak virus, skupno število pozitivnih opazovanj in skupen odstotek nakopičenih vrednosti.

Iz tabele 4.1 vidimo, da je pozitivnih opazovanj v kontrolni skupini (C) za nekatere viruse zelo majhno. Pri virusih RSV, HCoV in PIV je v kontrolni skupini ≤ 5 opazovanj, medtem ko pri hMPV in InfV ni pozitivnih opazovanj. Slednja imata tudi skupno najmanj pozitivnih opazovanj (hMPV 16 in InfV 36). Število

pozitivnih opazovanj je manjše od 50 tudi za virusa HCoV in PIV. Odstotek nakopičenih vrednosti v spremenljivki *ct* pa za vse štiri omenjane viruse presega 90% in je največji za hMPV (97%).

Virusi HBoV, hRV in AdV imajo več pozitivnih opazovanj (odstotek nakopičenih vrednosti je med 80 in 90%), ki so precej enakomerno porazdeljena med skupinami diagnoze.

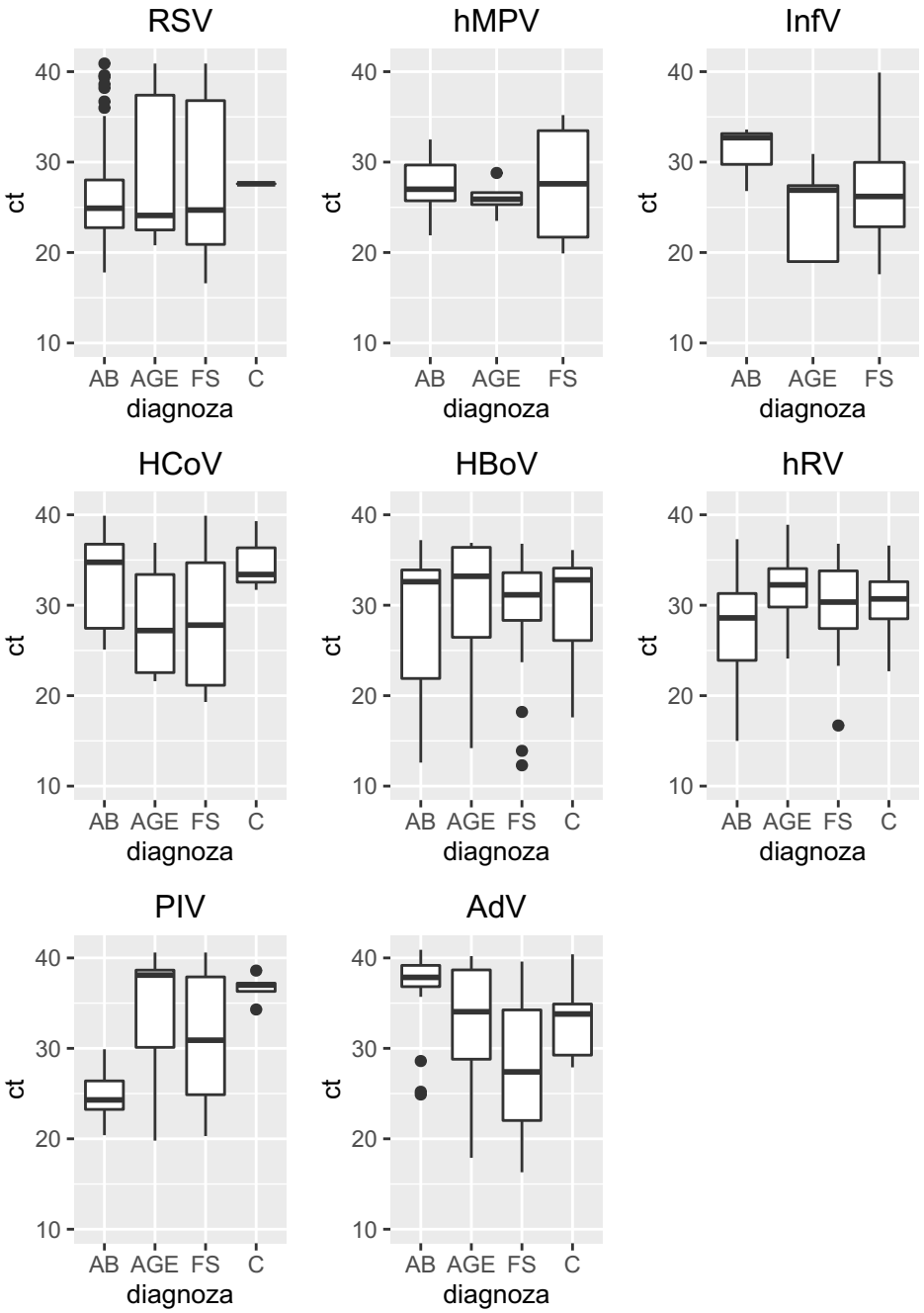
V tabeli 4.2 podajamo za vsak virus deleže nakopičenih vrednosti in povprečja pozitivnih opazovanj po skupinah. Slika 4.14 prikazuje porazdelitev spremenljivke *ct* za pozitivna opazovanja po skupinah preko okvirjev z ročaji. O skladnosti razlik med skupinami diagnoze je zelo težko sklepati, saj so ponekod razlike v deležih nakopičenih vrednosti zelo majhne, zaradi majhne velikosti vzorca za izračun povprečij pozitivnih opazovanj pa sklepamo, da so njihove ocene precej nenatančne.

	% nakopičenih vrednosti				povprečje pozitivnih opazovanj			
virus	AB	AGE	FS	C	AB	AGE	FS	C
RSV	47	89	89	99	26,0	29,3	27,8	27,6
hMPV	95	96	98	100	27,4	26,0	27,6	-
InfV	98	96	85	100	31,0	24,6	26,8	-
HCoV	92	90	90	98	32,9	28,3	28,1	34,8
HBoV	86	89	90	94	28,4	30,9	29,1	29,8
hRV	76	84	86	89	27,8	32,0	30,0	30,3
PIV	95	94	90	97	24,8	33,7	31,0	36,7
AdV	88	81	84	90	36,4	33,1	28,0	33,1

Tabela 4.2: Odstotek nakopičenih vrednosti in povprečna vrednost pozitivnih opazovanj po skupinah za vsak virus.

Glede na naravo podatkov bi pričakovali skladne razlike, torej višjo koncentracijo virusa v skupini, kjer je tudi delež pozitivnih opazovanj večji. Vendar, če pogledamo primer za PIV virus, velja za skupini AB (5%, 24,8) in FS (10%, 31,0), kjer prva vrednost v oklepaju pomeni %, druga pa povprečni *ct* pozitivnih opazovanj. V skupini AB je manj posameznikov pozitivnih, vendar imajo v pov-

prečju višjo koncentracijo virusa od pozitivnih posameznikov v FS, ki jih je več. Pojava neskladnih razlik tako ne moremo povsem izključiti.



Slika 4.14: Okvirji z ročaji za spremenljivko *ct* pozitivnih posameznikov po skupinah diagnoze za posamezne viruse.

Za vsak virus smo analizirali podatke z vsemi štirimi obravnavanimi metodami. Preverjamo ničelno domnevo, ki pravi, da ni povezave med skupinami diagnoze in okuženostjo z virusom. Rezultati globalnih testov (vrednosti p) so prikazani v tabeli 4.3. V prvem stolpcu je prikazana vrednost p za test, ki preverja predpostavko o sorazmernih obetih. Namesto Mann-Whitneyevega testa smo izvedli Kruskal-Wallisov (KW test), ki je njegova različica za primerjavo porazdelitev več kot dveh skupin. Za model Log+Lin so podani tudi rezultati globalnih testov posameznih delov. Tabela 4.4 prikazuje, v katerih primerih ničelno domnevo zavrnemo in kako močna je statistična značilnost.

virus	predpostavka PO	PO model	Log+Lin	Log	Lin	KW test	Tobit
RSV	0,131	<0,001	<0,001	<0,001	0,253	<0,001	<0,001
hMPV	0,694	0,008	0,032	0,008	0,862	0,033	NA
InfV	0,814	<0,001	<0,001	<0,001	0,296	<0,001	NA
HCoV	0,558	0,007	0,005	0,008	0,105	0,020	0,004
HBoV	0,546	0,128	0,352	0,130	0,796	0,146	0,140
hRV	0,148	0,007	0,001	0,014	0,015	0,005	0,004
PIV	0,039	0,035	0,002	0,034	0,015	0,032	0,072
AdV	<0,001	0,077	<0,001	0,097	<0,001	0,072	0,020

Tabela 4.3: Vrednosti p globalnih testov za vsak virus.

virus	predpostavka PO	PO model	Log+Lin	Log	Lin	KW test	Tobit
RSV	×	***	***	***	×	***	***
hMPV	×	**	*	**	×	*	NA
InfV	×	***	***	***	×	***	NA
HCoV	×	**	**	**	×	*	**
HBoV	×	×	×	×	×	×	×
hRV	×	**	**	*	*	**	**
PIV	*	*	**	*	*	*	×
AdV	***	×	***	×	***	×	*

Tabela 4.4: Prikaz vrednosti p za značilne razlike, $0{,}01 < p < 0{,}05$ označimo z *, $0{,}001 < p < 0{,}01$ z ** in $p < 0{,}001$ s ***.

Vse metode dajo enake rezultate v primerih virusov HBoV in hRV. Za oba virusa predpostavke o sorazmernih obetih ne zavrnemo. Za virus HBoV nobena metoda ne najde razlik med skupinami diagnoze, medtem ko za virus hRV vse metode najdejo razlike. Tudi ugotovljene razlike posameznih testov se ujemajo. Rezultati posameznih testov so podani v dodatku E.

Predpostavko o sorazmernih obetih zavrnemo pri dveh virusih, pri PIV in AdV. Pri AdV je vrednost p testa zelo majhna ($p < 0{,}001$), kar kaže na močno kršenje predpostavke. Modela Log+Lin in Tobit v tem primeru zaznata razliko med skupinami diagnoze, medtem ko model sorazmernih obetov in Kruskal-Wallisov test razlike ne zaznata. Model Log+Lin ima najnižjo globalno vrednost p. Sklep za ta virus bi bil tako precej različen, saj modela Log+Lin in Tobit zaznata posamezni razliki med FS-AB in FS-C, model Log+Lin pa najde še razliko v deležih med FS-AGE.

Pri PIV je vrednost p testa, ki preverja veljavnost predpostavke o sorazmernih obetih enaka 0,039. Za ta virus razlike ne zazna le model Tobit, najmanjšo vrednost p pa ima model Log+Lin. Globalni vrednosti p za model sorazmernih obetov in Kruskal-Wallisov test sta blizu 0,05 (0,035 in 0,032). Test sorazmernih obetov posamezno najde le razliko FS-C, medtem ko Kruskal-Wallisov test in model Log+Lin najdeta še razliko FS-AB.

Izkaže se, da sta obe razliki, ki ju model Log+Lin dodatno najde pri virusih AdV in PIV, neskladni.

Za viruse v zgornji polovici tabele 4.4 (RSV, hMPV, InfV, HCov) vidimo enak vzorec. Predpostavke o sorazmernih obetih ne zavrnemo, vse metode najdejo razlike med skupinami diagnoze, razen ko ena skupina nima pozitivnih opazovanj in je skupno število pozitivnih opazovanj zelo majhno (hMPV, InfV). Za ta dva primera model Tobit ne more oceniti vrednosti p (v tabeli označeno z NA - not available), pri modelu Log+Lin pa se vrednost testne statistike dela "Linželo poveča, zaradi majhne velikosti vzorca za ta del, kar vodi v povečanje skupne testne statistike in blizu neznačilnosti celotnega testa, čeprav del "Log"najde razlike. Za

oba primera model sorazmernih obetov in Kruskal-Wallisov test najdeta razlike med skupinami diagnoz, čeprav pri hMPV, model sorazmernih obetov dokaže le razliko C-FS (skupina C nima pozitivnih opazovanj).

Za virusa RSV in HCoV vse metode najdejo iste posamezne razlike, čeprav model Log+Lin te razlike najde le med deleži.

4.2.0.1 Primerjava z rezultati iz simulacij

V pravih podatkih odstotek nakopičenih vrednosti pri spremenljivki *ct* variira od $81 - 97\%$ za različne viruse dihal, kar ustreza najvišjemu obravnavanemu deležu nakopičenih vrednosti v simulacijah. Velikost vzorca za vse viruse je 617. Od 8 virusov so le trije takšni (HBoV, hRV, AdV), da je velikost vzorca za vsako skupino zadostna ($\geq$ 10). Te situacije bi lahko primerjali s simulacijami pri največjem deležu nakopičenih vrednosti in največji velikosti vzorca.

Za enega od omenjenih treh virusov (AdV) smo zavrnili predpostavko o sorazmernih obetih ($p < 0{,}001$). V tej situaciji ima test razmerja verjetij, s katerim preverjamo predpostavko, ustrezno napako prve vrste, torej zelo verjetno sorazmerni obeti pri virusu AdV ne veljajo. Modela Log+Lin in Tobit dokažeta povezanost koncentracije virusa in diagnoze, medtem ko Mann-Whitneyev test in model sorazmernih obetov nista statistično značilna. Rezultati se ujemajo s tem kar smo pokazali v simulacijah, ko smo generirali nesorazmerne obete. Izkaže se še, da so skupne dokazane razlike modelov Log+Lin in Tobit skladne, medtem ko model Log+Lin še dodatno zazna razliko med skupinama FS-AGE, ki je neskladna. S simulacijami smo pokazali, da moč modela Tobit v primerjavi z modelom Log+Lin močno pade, ko generiramo neskladne razlike med skupinama v situaciji, kjer sorazmerni obeti ne veljajo.

Za dva od treh virusov, za katere so velikosti posameznih skupin dovolj velike (HBoV in hRV), predpostavke o sorazmernih obetih ne zavrnemo, vrednosti p znašata 0,15 in 0,55. Test razmerja verjetij, s katerim preverjamo predpostavko, ima dobro moč v tej situaciji, zato zelo verjetno sorazmerni obeti niso

kršeni. Za oba primera obravnavane metode pridejo do istih zaključkov. Za virus HBoV nobena metoda ne dokaže povezanosti med diagnozo in koncentracijo virusa, medtem ko za virus hRV vse metode dokažejo povezanost. Dokazane razlike med skupinami diagnoz so skladne. Tudi s simulacijami smo pokazali, da je to situacija, kjer je delovanje metod najbolj primerljivo.

Pet izmed osmih obravnavanih virusov ima vsaj v eni skupini diagnoze 5 ali manj opazovanj (RSV, hMPV, InfV, HCoV, PIV), štirje izmed njih pa imajo skupno manj kot 50 pozitivnih opazovanj (izjema je RSV), kar je ekstremna situacija za analizo podatkov. Število pozitivnih opazovanj v posamezni skupini se ujema s simulacijami, kjer smo preučevali delovanje metod za največji delež nakopičenih vrednosti pri velikosti vzorca $n = 100$. Za to situacijo metode delujejo slabše, ko ne predpostavljamo razlik med skupinami, ob predpostavljenih razlikah pa je moč metod zelo nizka.

Izmed omenjenih petih virusov, predpostavko o sorazmernih obetih zavrnemo pri enem virusu (PIV), vrednost p testa razmerja verjetij pa znaša 0,04, kar je zelo blizu mejne vrednosti za statistično značilnost. Za to situacijo smo s simulacijami potrdili, da se test obnaša zelo anti-konzervativno, kar pomeni, da morda predpostavka o sorazmernih obetih na podatkih ni kršena. Čeprav ima najmanjšo vrednost p model Log+Lin, pa povezanost dokažemo tudi z modelom sorazmernih obetov in Kruskal-Wallisovim testom, ki sicer v simulacijah kažeta zelo nizko moč v primeru neveljavnih sorazmernih obetov.

Pri ostalih štirih obravnavanih virusih (RSV, hMPV, InfV, HCoV) predpostavke o sorazmernih obetih ne zavrnemo, vrednosti p so visoke. Model sorazmernih obetov in Kruskal-Wallisov test za vse viruse dokažeta povezanost diagnoze s koncentracijo virusa v vzorcu. Vse najdene razlike so skladne. Zelo majhno število pozitivnih opazovanj vpliva na model Log+Lin, saj za majhen vzorec del "Lin"nima moči za zaznavanje razlik, kar lahko vodi v večjo vrednost p skupno. Za dva virusa so podatki bolj ekstremni, saj v kontrolni skupini ni pozitivnih opazovanj. V tej situaciji model Tobit ne oceni vrednosti p.

5 Razprava

V magistrski nalogi smo s simulacijsko študijo preučevali delovanje modela sorazmernih obetov, kombinacije logistične in linearne regresije (model Log+Lin), modela Tobit in Mann-Whitneyevega testa za primerjavo porazdelitev dveh skupin ob prisotnosti nakopičenih vrednosti. Delovanje metod smo preučevali za različne velikosti vzorca in deleže nakopičenih vrednosti, preko velikosti testov - ob predpostavki, da razlik med skupinama ni in preko moči testov - ob predpostavljenih razlikah med skupinama. Preučevane metode smo nato uporabili pri analizi podatkov o virusih, kjer nas je zanimala razlika v okuženosti s posameznim virusom med štirimi skupinami otrok, ki so bili predhodno različno diagnosticirani. Na koncu smo rezultate analize še primerjali z rezultati iz simulacij.

S simulacijami smo za vse deleže nakopičenih vrednosti pokazali ustrezno delovanje preučevanih metod za velikost vzorca $n = 500$, ko ne predpostavljamo razlik med skupinama. Manjšanje velikosti vzorca vpliva na delovanje modela Log+Lin, ki se začne obnašati anti-konzervativno in na delovanje modela sorazmernih obetov, ki postane konzervativen. Učinek je najbolj izrazit pri največjem preučevanem deležu nakopičenih vrednosti (0,9), kjer je uporabna velikost vzorca najmanjša. Izmed vseh preučevanih metod edino model Tobit ohrani ustrezno napako prve vrste za vse kombinacije nakopičenih vrednosti in velikosti vzorca.

Mann-Whitneyev test deluje nezanesljivo za velik delež nakopičenih vrednosti, za katerega ima ob predpostavljeni povezanosti skupin z izidom tudi manjšo moč. Manjša moč je posledica velikega števila vezanih rangov, kar je ugotovil tudi Lachenbruch [37], Hallstrom [38] pa je v ta namen predlagal pristop, kjer

pred uporabo Mann-Whitneyevega testa odstranimo največje skupno število nakopičenih vrednosti v obeh skupinah. Gleiss s sodelavci [1] pokaže, da pristop deluje bolje za primer bioloških nakopičenih vrednosti, a slabše ob prisotnosti tehničnih nakopičenih vrednosti. Zhang s sodelavci [4] preizkusi različici Mann-Whitneyevega testa za obravnavo krnjenih opazovanj in vezanih rangov (Gehan test in Peto-Peto test), ki imata ob prisotnosti le-teh v podatkih višjo moč.

Delovanje metod ob predpostavljeni povezanosti skupin z izidom smo preučevali v različnih situacijah. Predvsem nas je zanimalo delovanje modela sorazmernih obetov v primerjavi z ostalimi metodami v primerih, ko sorazmerni obeti veljajo in ko ne veljajo. Ko sorazmerni obeti na podatkih veljajo, ima model sorazmernih obetov približno enako veliko moč kot model Tobit in Mann-Whitneyev test, medtem ko model Log+Lin deluje slabše. Ko sorazmerni obeti ne veljajo, ima model Log+Lin prednost pred ostalimi testi. Med njimi ima model sorazmernih obetov najmanjšo moč. Za model Tobit in Mann-Whitneyev test je moč manjša kot smo pričakovali, saj zanju v tem primeru nismo predpostavili padca moči.

Ko namesto skladnih razlik med skupinama predpostavimo neskladne razlike, moč dvodelnega testa ohrani enako velikost, medtem ko moč ostalim trem enodelnim metodam močno pade. Velik učinek neskladnih razlik je zaznal že Lachenbruch [25], ki je v [12] opozoril, da večji delež nakopičenih vrednosti pri nuli, za skupino z večjim povprečjem zveznega dela (neskladna razlika med skupinama), zmanjša to povprečje in enodelni testi razlike lahko ne zaznajo. Ker dvodelni testi vsak del porazdelitve obravnavajo posebej, razliko zaznajo ne glede na skladnost. Prednost modela Log+Lin se še poveča, ko sorazmerni obeti ne veljajo, saj enodelni testi skoraj povsem izgubijo na moči.

Za model sorazmernih obetov smo pokazali, da v nobeni situaciji nima prednosti pred ostalimi preučevanimi metodami, v situacijah kjer sorazmerni obeti ne veljajo, pa ima pričakovano najslabšo moč. Dobro moč ohrani le v situacijah, ko sorazmerni obeti veljajo, razlike med skupinama pa so skladne. Če smo

prepričani, da te dve lastnosti veljata za obravnavane podatke, potem bi priporočali uporabo modela sorazmernih obetov, ker omogoča ugodno interpretacijo preko razmerja obetov. Chang in Pocock [2] sta primerjala delovanje modela sorazmernih obetov in modela Log+Lin, ob prisotnosti nakopičenih vrednosti v odvisni spremenljivki, in zaključila, da je ob kršeni predpostavki sorazmernih obetov priporočljiva uporaba modela Log+Lin, drugače pa enako dobro delujeta obe metodi.

O tem, ali sorazmerni obeti v podatkih veljajo ali ne, se največkrat odločamo preko testa razmerja verjetij. S simulacijami smo preverili njegovo delovanje in ugotovili, da ob veljavnih sorazmernih obetih test ohranja pravo velikost pri dovolj velikem vzorcu, medtem ko je njegovo delovanje pri majhnem vzorcu antikonzervativno. Ugotovitev se ujema z ugotovitvijo Thomasa [20]. Pri večjem deležu nakopičenih vrednosti je delovanje testa nezanesljivo. Ob predpostavljenih nesorazmernih obetih v podatkih, na moč testa vpliva predvsem velikost vzorca.

S simulacijami smo preučevali delovanje metod v primerih, ko sorazmerni obeti veljajo in ko ne veljajo, za skladne in neskladne razlike. Lastnosti v podatkih lahko do neke mere predvidimo glede na predpostavljeno vrsto nakopičenih vrednosti. Za prisotne tehnične nakopičene vrednosti so značilne skladne razlike. V primeru bioloških nakopičenih vrednosti, pa so možne tako skladne, kot neskladne razlike med skupinama [1].

Do sedaj smo lastnosti skladnosti in sorazmernosti obravnavali ločeno, vendar ju lahko povežemo. Neskladne razlike med skupinami zagotovo pomenijo tudi nesorazmerja. Vendar pa smo v nalogi odkrili nesorazmernost tudi v podatkih, ko generiramo skladne razlike med skupinama. Sklepamo, da nesorazmernost lahko izvira tudi iz neskladnosti med kategorijami znotraj zveznega dela porazdelitve. To smo v simulacijah ustvarili nenačrtno, ko smo generirali iz logaritemsko normalne porazdelitve, nakopičene vrednosti pa naključno izbirali iz le-te. Gleiss s sodelavci [1] omenja, da takšen način generiranja ustreza biološkim nakopičenim vrednostim.

Ob prisotnosti bioloških nakopičenih vrednosti tako sklepamo, da imamo v podatkih, poleg neskladnih razlik, do neke mere prisotna tudi nesorazmerja. S simulacijami smo pokazali, da je oboje razlog slabšega delovanja enodelnih metod v teh situacijah. Gleiss s sodelavci [1] poroča, da dvodelne metode dajejo boljše rezultate v primeru bioloških nakopičenih vrednosti in enodelne metode v primeru tehničnih nakopičenih vrednosti, vendar ne pojasni zakaj. Razlog bi bil lahko v pojavu nesorazmernosti pri bioloških nakopičenih vrednosti, kjer imajo enodelne metode slabšo moč. Pri tehničnih nakopičenih vrednosti nesorazmernosti ne pričakujemo.

V primeru tehničnih nakopičenih vrednosti sklepamo na prisotnost skladnih razlik med skupinama in na sorazmerja v podatkih, zato bi priporočili uporabo enodelnih metod. V primeru bioloških nakopičenih vrednosti, pa zaradi možnih neskladnih razlik in nesorazmerij v podatkih, priporočamo uporabo dvodelnih metod. Pri pravih podatkih se lahko hkrati pojavijo tako tehnične, kot tudi biološke nakopičene vrednosti, kar hkrati pomeni prisotnost vseh omenjenih pojavov skladnosti in sorazmernosti. Za te primere, Gleiss s sodelovci [1] ocenjuje, da imajo dvodelne metode na splošno večjo moč od enodelnih metod.

Za podatke o virusih smo glede na naravo podatkov sklepali, da so razlike med obravnavanimi skupinami skladne. To pomeni, da imajo pozitivni otroci v skupini z večjim deležem okuženih, tudi višjo koncentracijo virusa v vzorcu, kar bi se ujemalo z definicijo tehničnih nakopičenih vrednosti. Verjetno pa vse nakopičene vrednosti ne prihajajo iz porazdelitve, ki je nadaljevanje zveznega dela, ki ga vidimo, ampak obstaja večji del otrok, ki z virusom niso okuženi in je dejanska koncentracija virusa v vzorcu teh otrok 0. To ustreza definiciji bioloških nakopičenih vrednosti. Za analizirane podatke tako sklepamo tudi o prisotnosti neskladnih razlik in nesorazmerij - kar v podatkih tudi vidimo. Pri vseh virusih, kjer dokažemo povezanost diagnoze s koncentracijo virusa v vzorcu, je najnižja vrednost p rezultat modela Log+Lin.

Za analizo izbranih podatkov bi tako priporočili uporabo dvodelnega pristopa,

ki ohrani dobro moč tudi v situacijah, kjer so prisotne neskladne razlike in nesorazmerja. Model Log+Lin prav tako ne zaostaja veliko v situacijah, kjer veljajo sorazmerni obeti in so prisotne le skladne razlike. Lachenbruch [12] je sklenil, da imajo enodelni testi zares prednost, ko preučujemo skladne razlike, vendar ta prednost ni velika v primerjavi s prednostjo dvodelnih testov ob neskladnih razlikah. Pristop omogoča raziskovalcu tudi vpogled v razlike znotraj posameznih delov porazdelitve in ločeno interpretacijo teh razlik. Vendar pa metoda ni primerna ob izjemno majhnih vzorcih, zaradi slabe moči dela "Lin". V teh primerih bi priporočili uporabo univariatne metode in za del testa, ki preučuje razlike v zveznem delu, uporabo npr. Mann-Whitneyevega testa, ki ni tako občutljiv na majhne vzorce. Gleiss s sodelavci [1] pokaže, da dvodelni Mann-Whitneyev test deluje dobro v situacijah z obema vrstama nakopičenih vrednosti, njegova uporaba pa je priporočljiva tudi zaradi odsotnosti predpostavke o porazdelitvi.

6 Zaključki

1. Model sorazmernih obetov je primeren za analizo podatkov z nakopičenimi vrednostmi le, kadar je vzorec za dani delež nakopičenih vrednosti dovolj velik in lahko za analizirane podatke predpostavimo veljavnost sorazmernih obetov ter skladnost razlik.

2. Model Tobit in Mann-Whitneyev test sta pokazala rahlo prednost v primerjavi z modelom sorazmernih obetov, predvsem v situacijah, ko sorazmerni obeti v podatkih ne veljajo.

3. Model Log+Lin ima veliko prednost pred ostalimi preučevanimi metodami v situacijah, kjer sorazmerni obeti ne veljajo ali so prisotne neskladne razlike.

4. Test razmerja verjetij, s katerim preverjamo veljavnost predpostavke o sorazmernih obetih, se je izkazal za ustreznega pri dovolj velikem vzorcu. Pri majhnem vzorcu se obnaša anti-konzervativno, pod alternativno domnevo pa ima zelo majhno moč.

5. Nesorazmernost lahko opišemo kot neskladnost v zveznem delu porazdelitve in jo povežemo s prisotnostjo bioloških nakopičenih vrednosti. Za analizo podatkov z biološkimi nakopičenimi vrednostmi so tako bolj primerne dvodelne metode.

Literatura

[1] A. Gleiss, M. Dakna, H. Mischak in G. Heinze, "Two-group comparisons of zero-inflated intensity values: the choice of test statistic matters," *Bioinformatics*, str. 154, 2015.

[2] B.-H. Chang in S. Pocock, "Analyzing data with clumping at zero: an example demonstration," *Journal of Clinical Epidemiology*, vol. 53, no. 10, str. 1036–1043, 2000.

[3] S. Taylor in K. Pollard, "Hypothesis tests for point-mass mixture data with application toomics data with many zero values," *Statistical applications in genetics and molecular biology*, vol. 8, no. 1, str. 1–43, 2009.

[4] D. Zhang, C. Fan, J. Zhang in C.-H. Zhang, "Nonparametric methods for measurements below detection limit," *Statistics in medicine*, vol. 28, no. 4, str. 700, 2009.

[5] M. Pokorn, M. Jevšnik, M. Petrovec, A. Steyer, T. Mrvič, Š. Grosek, L. Lusa in F. Strle, "Respiratory and enteric virus detection in children a prospective study comparing children with febrile seizures and healthy controls," *Journal of Child Neurology*, 2016.

[6] B. Chung in V. Wong, "Relationship between five common viruses and febrile seizure in children," *Archives of disease in childhood*, vol. 92, no. 7, str. 589–593, 2007.

[7] H. Rantala, M. Uhari in H. Tuokko, "Viral infections and recurrences of febrile convulsions," *The Journal of pediatrics*, vol. 116, no. 2, str. 195–199, 1990.

[8] H. M. Lewis, J. V. Parry, R. P. Parry, H. A. Davies, P. Sanderson, D. Tyrrell in H. Valman, "Role of viruses in febrile convulsions.," *Archives of disease in childhood*, vol. 54, no. 11, str. 869–876, 1979.

[9] M. Stokes, M. Downham, J. Webb, J. McQuillin in P. Gardner, "Viruses and febrile convulsions.," *Archives of disease in childhood*, vol. 52, no. 2, str. 129–133, 1977.

[10] S. J. Wallace in H. Zealley, "Neurological, electroencephalographic, and virological findings in febrile children," *Archives of disease in childhood*, vol. 45, no. 243, str. 611–623, 1970.

[11] J. G. Millichap in J. J. Millichap, "Role of viral infections in the etiology of febrile seizures," *Pediatric neurology*, vol. 35, no. 3, str. 165–172, 2006.

[12] P. A. Lachenbruch, "Comparisons of two-part models with competitors," *Statistics in medicine*, vol. 20, no. 8, str. 1215–1234, 2001.

[13] A. Saei, J. Ward in C. McGilchrist, "Threshold models in a methadone programme evaluation," *Statistics in Medicine*, vol. 15, no. 20, str. 2253–2260, 1996.

[14] M. Rodrigues-Motta, D. M. Galvis Soto, V. H. Lachos, F. Vilca, V. T. Baltar, E. V. Junior, R. M. Fisberg in D. M. Lobo Marchioni, "A mixed-effect model for positive responses augmented by zeros," *Statistics in medicine*, vol. 34, no. 10, str. 1761–1778, 2015.

[15] P. McCullagh, "Regression models for ordinal data," *Journal of the royal statistical society. Series B (Methodological)*, str. 109–142, 1980.

[16] J. F. McDonald in R. A. Moffitt, "The uses of tobit analysis," *The review of economics and statistics*, str. 318–321, 1980.

[17] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society. Series B (Methodological)*, str. 215–242, 1958.

[18] A. Agresti in M. Kateri, *Categorical data analysis*. Springer, 2011.

[19] G. A. Seber, *Statistical models for proportions and probabilities*. Springer, 2013.

[20] A. M. Thomas, "The proportional odds model: Simulations studies and predictive accuracy," 2014.

[21] P. J. Green, "Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives," *Journal of the Royal Statistical Society. Series B (Methodological)*, str. 149–192, 1984.

[22] B. Peterson in F. E. Harrell Jr, "Partial proportional odds models for ordinal response variables," *Applied statistics*, str. 205–217, 1990.

[23] J.-H. Kim, "Assessing practical significance of the proportional odds assumption," *Statistics & probability letters*, vol. 65, no. 3, str. 233–239, 2003.

[24] P. A. Lachenbruch, "Analysis of data with clumping at zero," *Biometrische Zeitschrift*, vol. 18, no. 5, str. 351–356, 1976.

[25] P. A. Lachenbruch, "Analysis of data with excess zeros," *Statistical methods in medical research*, vol. 11, no. 4, str. 297–302, 2002.

[26] A. John, *Mathematical statistics and data analysis*. Wadsworth & Brooks/Cole, 1988.

[27] K. Košmelj, *Uporabna statistika*. Biotehniška fakulteta, 2001.

[28] D. A. Kenny, *Statistics for the social and behavioral sciences*. Little, Brown, 1987.

[29] R. J. Tallarida in R. B. Murray, "Mann-whitney test," v *Manual of Pharmacologic Calculations*, str. 149–153, Springer, 1987.

[30] B. Košmelj, "Statistični terminološki slovar [elektronski vir]," 2014.

[31] P. E. McKight in J. Najab, "Kruskal-wallis test," *Corsini Encyclopedia of Psychology*, 2010.

[32] O. J. Dunn, "Multiple comparisons using rank sums," *Technometrics*, vol. 6, no. 3, str. 241–252, 1964.

[33] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[34] W. N. Venables in B. D. Ripley, *Modern Applied Statistics with S*. New York: Springer, fourth izd., 2002. ISBN 0-387-95457-0.

[35] W. N. Venables in B. D. Ripley, *Modern Applied Statistics with S*. New York: Springer, fourth izd., 2002. ISBN 0-387-95457-0.

[36] T. W. Yee, *Vector Generalized Linear and Additive Models: With an Implementation in R*. New York, USA: Springer, 2015.

[37] P. A. Lachenbruch, "Power and sample size requirements for two-part models," *Statistics in medicine*, vol. 20, no. 8, str. 1235–1238, 2001.

[38] A. P. Hallstrom, "A modified wilcoxon test for non-negative distributions with a clump of zeros," *Statistics in medicine*, vol. 29, no. 3, str. 391–400, 2010.

Dodatek

A Lastnosti generiranih podatkov

Simuliranje preko latentne spremenljivke - sorazmerni obeti veljajo

	n=500						
	delež nakopičenih vrednosti			povprečje zveznega dela			skladnost
Delež	A	B	A-B	A	B	B-A	(%)
0,25	0,30	0,20	0,09	2,2	2,4	0,3	97,2
0,4	0,46	0,34	0,12	2,6	2,8	0,2	92,2
0,7	0,75	0,65	0,10	3,6	3,6	0,1	67,9
0,9	0,92	0,88	0,04	4,8	4,8	0,0	53,0
	n=100						
0,25	0,30	0,20	0,09	2,2	2,4	0,3	74,6
0,4	0,46	0,34	0,12	2,6	2,8	0,2	65,0
0,7	0,75	0,65	0,11	3,6	3,6	0,1	54,1
0,9	0,92	0,88	0,04	4,8	4,8	0,0	42,7

Tabela A.1: Lastnosti generiranih podatkov z $\beta = 0{,}5$ za velikost vzorca 100 in 500; pridobljeno iz simulacij s 1000 ponovitvami.

Simuliranje preko latentne spremenljivke - sorazmerni obeti ne veljajo

	n=500						
	delež nakopičenih vrednosti			povprečje zveznega dela			skladnost
Delež	A	B	A-B	A	B	B-A	(%)
0,25	0,22	0,28	-0,06	1,0	1,5	0,5	4,4
0,4	0,39	0,41	-0,02	1,4	1,9	0,6	22,4
0,7	0,74	0,66	0,09	2,5	2,9	0,4	95,7
0,9	0,93	0,87	0,07	4,0	4,1	0,1	64,0
	n=100						
0,25	0,21	0,29	-0,07	1,0	1,4	0,5	21,3
0,4	0,38	0,42	-0,04	1,4	1,9	0,5	25,2
0,7	0,74	0,66	0,07	2,5	2,9	0,4	60,7
0,9	0,93	0,87	0,06	3,9	4,1	0,1	50,0

Tabela A.2: Lastnosti generiranih podatkov za manjše predpostavljene razlike med ravnema pojasnjevalne spremenljivke, za velikost vzorca 100 in 500; pridobljeno iz simulacij s 1000 ponovitvami.

B Rezultati: simuliranje preko latentne spremenljivke

- Log+Lin = model Log+Lin
- MW test = Mann-Whitneyev test
- PO model = model sorazmernih obetov
- Tobit = model Tobit

delež	n	Log+Lin	MW test	PO model	Tobit
0,25	100	5,43	4,85	5,76	5,26
	300	5,04	4,90	5,65	4,77
	500	4,70	4,93	4,94	5,07
0,4	100	5,78	4,75	4,41	5,17
	300	4,79	4,66	5,06	4,85
	500	5,02	4,92	4,92	5,20
0,7	100	5,70	4,88	4,48	5,20
	300	4,95	4,87	4,94	4,84
	500	4,81	4,74	4,59	4,82
0,9	100	7,18	6,58	1,54	5,45
	300	5,35	4,09	3,36	4,81
	500	5,35	5,19	5,17	5,20

Tabela B.1: Velikost testov (%) ob veljavni ničelni domnevi, sorazmerni obeti veljajo.

sorazmerni obeti veljajo

delež	n	Log+Lin	MW test	PO model	Tobit
0,25	100	23,1	28,9	31,2	29,8
	300	57,6	69,5	69,6	67,6
	500	80,5	89,0	87,2	87,4
0,4	100	23,0	28,3	27,1	28,4
	300	56,0	67,7	67,5	66,0
	500	79,0	87,7	86,9	86,0
0,7	100	17,3	20,9	19,5	21,2
	300	42,7	52,6	52,8	51,1
	500	64,0	74,0	73,3	72,3
0,9	100	10,4	13,8	4,6	11,6
	300	19,7	23,4	20,9	24,6
	500	30,1	39,2	39,1	37,7

sorazmerni obeti ne veljajo

0,25	100	26,4	3,4	3,1	4,4
	300	72,4	4,4	3,8	8,1
	500	91,1	6,7	4,6	13,2
0,4	100	25,9	3,9	3,8	5,6
	300	69,5	9,5	6,9	16,9
	500	90,7	16,6	11,6	27,1
0,7	100	16,1	13,7	12,2	16,7
	300	50,8	48,0	47,6	56,2
	500	77,3	74,4	73,2	81,1
0,9	100	12,3	21,9	7,0	18,0
	300	42,3	53,0	48,3	55,7
	500	68,7	78,2	78,3	77,8

Tabela B.2: Moč testov (%) pod alternativno domnevo za manjše predpostavljene razlike med ravnema pojasnjevalne spremenljivke X.

C Rezultati: simuliranje iz logaritemsko normalne porazdelitve

- Log+Lin = model Log+Lin
- MW test = Mann-Whitneyev test
- PO model = model sorazmernih obetov
- Tobit = model Tobit

delež	n	Log+Lin	MW test	PO model	Tobit
0,25	100	18,9	24,4	22,8	25,9
	300	51,4	61,9	58,6	63,4
	500	74,4	83,1	79,9	82,8
0,4	100	17,1	22,4	21,9	21,8
	300	42,6	55,0	50,9	54,4
	500	62,0	74,3	71,2	73,9
0,7	100	14,9	21,2	18,3	21,5
	300	36,4	46,6	43,2	48,1
	500	58,7	70,1	68,9	70,9
0,9	100	20,3	24,8	21,3	26,1
	300	48,7	57,6	55,0	58,7
	500	75,4	84,7	82,5	85,7

Tabela C.1: Moč testov (%) pod alternativno domnevo, sorazmerni obeti veljajo.

skladne razlike

delež	n	Log+Lin	MW test	PO model	Tobit
0,25	100	57,2	54,6	40,5	48,8
	300	97,3	95,3	89,2	91,2
	500	99,9	99,6	98,1	99,1
0,4	100	48,7	33,8	25,3	30,4
	300	92,5	74,4	65,8	71,9
	500	99,5	91,6	84,9	89,1
0,7	100	36,4	17,7	12,4	19,1
	300	80,5	38,8	31,0	41,7
	500	96,7	62,7	53,3	65,4
0,9	100	20,8	11,6	8,1	12,9
	300	54,1	25,7	21,7	29,9
	500	79,7	42,5	36,9	48,8

neskladne razlike

0,25	100	59,8	20,2	13,5	13,5
	300	97,5	53,5	36,3	39,0
	500	100,0	72,8	56,1	56,2
0,4	100	49,6	9,2	7,0	8,4
	300	93,4	16,7	10,9	13,1
	500	99,7	26,7	14,9	21,3
0,7	100	39,5	5,7	5,5	6,1
	300	81,5	4,8	3,8	4,6
	500	96,3	6,1	6,1	5,3
0,9	100	20,8	11,6	8,1	12,9
	300	57,5	11,1	14,8	9,1
	500	79,3	17,8	19,8	12,5

Tabela C.2: Moč testov (%) pod alternativno domnevo, sorazmerni obeti ne veljajo.

D Rezultati: test razmerja verjetij za preverjanje predpostavke o sorazmernih obetih

delež	n	$\beta = 0$	$\beta = 0{,}5$	$\beta = 1$
0,25	100	5,45	5,37	5,52
	300	5,15	5,16	5,36
	500	4,75	4,97	5,08
0,4	100	5,38	6,15	6,13
	300	4,88	5,31	5,05
	500	5,33	5,27	5,04
0,7	100	5,39	5,71	6,83
	300	5,47	5,27	5,33
	500	5,37	5,37	5,41
0,9	100	7,07	6,52	5,25
	300	5,07	6,67	8,86
	500	5,72	5,76	5,66

Tabela D.1: Velikost testa (%) ob veljavni ničelni domnevi za tri situacije: ni razlik med skupinama pojasnjevalne spremenljivke X ($\beta = 0$), manjše ($\beta = 0{,}5$) in večje razlike ($\beta = 1$).

delež	n	skladne razlike	neskladne razlike
0,25	100	25,4	47,9
	300	58,1	92,4
	500	83,5	99,7
0,4	100	27,1	42,1
	300	69,6	89,8
	500	89,6	99,1
0,6	100	25,1	33,0
	300	61,6	76,9
	500	83,6	94,6
0,8	100	18,9	18,9
	300	36,8	44,1
	500	57,5	66,4

Tabela D.2: Moč testa (%) pod alternativno domnevo za dve situaciji: generiranje skladnih in neskladnih razlik

E Rezultati analize pravih podatkov: posamezne primerjave

del Log

virus	AB-AGE	AB-FS	AB-C	C-AGE	C-FS	FS-AGE
RSV	<0,001	<0,001	<0,001	0,004	0,004	0,900
hMPV	0,519	0,131	0,990	0,990	0,991	0,455
InfV	0,255	<0,001	0,986	0,985	0,984	0,009
HCoV	0,581	0,485	0,027	0,011	0,006	0,944
hRV	0,106	0,014	0,003	0,240	0,457	0,588
PIV	0,555	0,047	0,550	0,256	0,014	0,207

del Lin

virus	AB-AGE	AB-FS	AB-C	C-AGE	C-FS	FS-AGE
hRV	0,002	0,071	0,066	0,286	0,801	0,153
PIV	0,013	0,032	0,003	0,428	0,082	0,338
AdV	0,093	<0,001	0,128	0,992	0,011	0,005

Tabela E.1: Vrednosti p za posamezne primerjave modela Log+Lin, prikaz le za globalno značilne viruse

Model sorazmernih obetov

virus	AB-AGE	AB-FS	AB-C	C-AGE	C-FS	FS-AGE
RSV	<0,001	<0,001	<0,001	0,004	0,004	0,896
hMPV	0,523	0,132	<0,001	<0,001	<0,001	0,454
InfV	0,246	<0,001	0,986	<0,001	<0,001	0,009
HCoV	0,550	0,445	0,028	0,010	0,006	0,934
hRV	0,051	0,008	0,002	0,285	0,431	0,705
PIV	0,598	0,054	0,511	0,255	0,014	0,200

Kruskal-Wallisov test

virus	AB-AGE	AB-FS	AB-C	C-AGE	C-FS	FS-AGE
RSV	<0,001	<0,001	<0,001	0,015	0,008	0,481
hMPV	0,204	0,037	0,002	0,036	0,112	0,221
InfV	0,190	<0,001	0,240	0,063	<0,001	<0,001
HCoV	0,252	0,196	0,026	0,007	0,002	0,466
hRV	0,017	0,002	<0,001	0,174	0,251	0,357
PIV	0,323	0,014	0,290	0,166	0,003	0,065

model Tobit

virus	AB-AGE	AB-FS	AB-C	C-AGE	C-FS	FS-AGE
RSV	<0,001	<0,001	<0,001	0,001	0,001	1,000
HCoV	0,466	0,367	0,030	0,008	0,004	0,940
hRV	0,052	0,010	0,002	0,339	0,461	0,750
AdV	0,034	0,032	0,976	0,036	0,034	0,831

Tabela E.2: Vrednosti p za posamezne primerjave modela sorazmernih obetov, Kruskal-Wallisovega testa in modela Tobit, prikaz le za globalno značilne viruse

F Slovarček uporabljene terminologije

angleški pojem	**slovenski pojem**
point mass values	nakopičene vrednosti
biological point mass values	biološke nakopičene vrednosti
technical point mass values	tehnične nakopičene vrednosti
proportional odds model	model sorazmernih obetov
consonant effect	skladna razlika
dissonant effect	neskladna razlika