EMPĪRISKĀS TICAMĪBAS METODES IZMANTOŠANA TIEKSMES RĀDĪTĀJA SASKAŅOŠANAS ALGORITMĀ KAUZĀLO EFEKTU NOVĒRTĒŠANAI

BAKALAURA DARBS

RĪGA 2025

ANOTĀCIJA

Bakalaura darbā tiek pētīta empīriskās ticamības metodes izmantošana tieksmes rādītāja novērtēšanai, kas ir būtisks solis tieksmes rādītāja saskaņošanas algoritmā. Klasiski tieksmes rādītāja vērtības novērtē, izmantojot loģistisko regresiju, regresijas koeficientus iegūstot ar maksimālās ticamības metodi. Tomēr parametriska pieeja, kas balstās uz noteiktiem pieņēmumiem, piemēram, lielu izlases apjomu vai log-izredžu linearitāti, var nebūt labākais variants gadījumos, kad viens vai vairāki no šiem nosacījumiem neizpildās. Šajā darbā ir aplūkota arī empīriskās (maksimālās) ticamības pieeja loģistiskās regresijas koeficientu novērtēšanā, salīdzinot to ar ierasto maksimālās ticamības metodi. Aprakstītas teorijas pamatidejas, kas saistītas ar kauzālo efektu novērtēšanu un tieksmes rādītāja saskaņošanas algoritmu, kā arī empīriskās ticamības metodi un tās izmantošanu vispārināto lineāro modeļu (GLM), tai skaitā loģistiskās regresijas, gadījumā. Pēcāk veiktas simulācijas, lai salīdzinātu maksimālās ticamības un empīriskās ticamības rezultātus simulētai datu kopai. Tieksmes rādītāja saskaņošanas algoritms, loģistiskās regresijas koeficientus novērtējot gan ar maksimālo, gan empīrisko ticamības metodi, praktiskā piemērā pielietots ICCS pētījuma datiem, lai salīdzinātu, kā abas metodes strādā reālu datu gadījumā.

Atslēgas vārdi: kauzālo efektu novērtēšana, tieksmes rādītāja saskaņošanas algoritms, empīriskās ticamības metode, loģistiskā regresija.

ABSTRACT

This Bachelor thesis explores the use of the empirical likelihood method to estimate the propensity score, a key step in the propensity score matching algorithm. Typically, the propensity score values are estimated using logistic regression, where regression coefficients are estimated using maximum likelihood method. However, a parametric approach based on certain assumptions, such as large sample size or log-odds linearity, may not be the best option in cases where one or more of these conditions are not met. This work also looks at the empirical (maximum) likelihood approach in estimating the logistic regression coefficients compared to the usual maximum likelihood method. The main theoretical ideas related to the estimation of causal effects and the propensity score matching algorithm, as well as the empirical likelihood method and its use in the case of generalised linear models (GLM), including logistic regression, are described. Simulations have been performed to compare the results of maximum likelihood and empirical likelihood estimation for a simulated dataset. The propensity score matching algorithm, when estimating logistic regression coefficients using both maximum and empirical likelihood methods, has been applied in a practical example to the ICCS study data to compare how well both methods work for real data.

Keywords: causal effect estimation, propensity score matching algorithm, empirical likelihood method, logistic regression.

SATURA RĀDĪTĀJS

Ievads **4**

1. Tieksmes rādītāja saskaņošanas algoritms kauzālo efektu novērtēšanai **6**

1.1. Kauzālo efektu novērtēšana - pamatidejas 6

1.2. Randomizētie eksperimenti 7

1.3. Tieksmes rādītāja saskaņošanas algoritms 9

1.3.1. Svarīgākie teorētiskie rezultāti līdzsvarojošā rādītāja, tai skaitā tieksmes rādītāja, sakarā 9

1.3.2. Tieksmes rādītāja novērtēšana 12

1.3.3. Tieksmes rādītāja pāru saskaņošana 13

1.4. Metodes efekta novērtēšana ar inversās varbūtības svērto novērtētāju (IPW) . . 14

2. Empīriskās ticamības metode **16**

2.1. Empīriskās ticamības metode - pamatidejas 16

2.2. Empīriskā ticamības metode vispārinātajiem lineārajiem modeļiem (GLM) . . . 18

2.3. Empīriskās ticamības metodes iekļaušana tieksmes rādītāja saskaņošanas algoritmā 20

2.4. Simulācijas rezultāti 22

3. ICCS datu analīze **25**

3.1. Datu kopas apraksts 25

3.2. Datu analīze 27

Secinājumi **35**

Izmantotā literatūra un avoti **37**

Pielikumi **39**

A. Programmu kodi **39**

A.1. Simulācijas piemēra kods 39

A.2. Praktiskās daļas kods - ICCS datu analīze 41

B. Papildu izdrukas un grafiki **47**

IEVADS

Ja vēlas pēc iespējas precīzi novērtēt kādas metodes, piemēram, jaunu zāļu vai mācīšanas pieejas, efektu uz populāciju, nepieciešams veikt tā saukto ”randomizēto eksperimentu”, kura ietvaros pēc nejaušības principa atlasa dalībniekus kontroles un testa grupām, pēcāk pielietojot pētīto metodi testa grupai un novērtējot tās rezultātus. Randomizēto eksperimentu gadījumā vidējo metodes efektu jeb ATE var novērtēt, vienkārši aprēķinot starpību kontroles un testa grupu vidējiem iznākumiem, jo šie eksperimenti ir veidoti tā, lai novērstu efekta novērtējuma novirzi, ko varētu radīt citi iznākumu ietekmējošie rādītāji, piemēram, demogrāfiskie rādītāji un sociālekonomiskais stāvoklis. Šī pieeja tiek izmantota klīniskajā pētniecībā, tā ir arī A/B testēšanas pamatā, ko pielieto programmatūru izstrādes procesā, mārketinga izpētē un mediju sfērā.

Lai arī tas tiek saukts par ”zelta standartu” kauzālo efektu novēršanai, randomizētos eksperimentus ne vienmēr ir iespējams veikt ētisku vai praktisku iemeslu, piemēram, pārāk lielu izmaksu, dēļ. Bieži noteiktas metodes kauzālo efektu jānovērtē novērošanas pētījuma (angl.- *”observational study”*) datiem, kur testa un kontroles grupas var būt nevienlīdzīgas pēc citiem iznākumu ietekmējošajiem faktoriem (kovariātiem). Tādā gadījumā, izrēķinot kontroles un testa grupu vidējo iznākumu starpību, efekta novērtējums būs novirzīts. Lai iegūtu nenovirzītu ATE novērtējumu, rodas nepieciešamība līdzsvarot kontroles un testa grupas, lai padarītu tās savstarpēji salīdzināmas. Viens no veidiem, kā to izdarīt, ir, izmantojot tieksmes rādītāja saskaņošanas algoritmu.

Tieksmes rādītāja saskaņošanas algoritma viens no būtiskākajiem aspektiem ir precīza tieksmes rādītāja novērtēšana. Tieksmes rādītājs ir varbūtība izmantot pētīto metodi/saņemt pētīto ārstēšanu, balstoties uz novēroto kovariātu (atkarībā no konteksta var būt, piemēram, dzimums, vecums, izglītība, ienākumi utml.) vērtībām. To visbiežāk novērtē ar loģistisko regresiju, kovariātus izmantojot kā prediktorus.

Parasti loģistiskās regresijas koeficientus novērtē, izmantojot maksimālās ticamības metodi. Tiek izdarīti dažādi parametriski pieņēmumi, piemēram, ka atkarīgā mainīgā (iznākuma) logizredzes ir izsakāmas kā neatkarīgo mainīgo (kovariātu) lineāra kombinācija, ka datos nav izteiktu izlēcēju, ka starp prediktoriem nepastāv multikolinearitāte vai ka izlases apjoms ir pietiekami liels utt. Strādājot ar reāliem datiem, var gadīties, ka viens vai vairāki no šiem nosacījumiem neizpildās, un tad koeficientu novērtēšanai, iespējams, labāk būtu izmantot tādu neparametrisku pieeju kā empīriskās ticamības metode.

Darba mērķis ir izpētīt, vai empīriskās ticamības metodes izmantošana dotu precīzākus loģistiskās regresijas koeficientu novērtējumus nekā maksimālās ticamības metode un, attiecīgi, arī precīzāku tieksmes rādītāja novērtējumu. Mērķa sasniegšanai tika izvirzīti šādi darba uzdevumi:

1. iepazīties ar pieejamo literatūru par kauzālo efektu novērtēšanu, tieksmes rādītāja saskaņošanas algoritmu un empīriskās ticamības metodi (arī vispārināto lineāro modeļu kontekstā);

2. veikt simulācijas, lai salīdzinātu loģistiskās regresijas koeficientu novērtējumus un to kvalitāti maksimālās un empīriskās ticamības metodēm;

3. pielietot tieksmes rādītāja saskaņošanas algoritmu ICCS pētījuma datiem, novērtējot tieksmes rādītāju ar maksimālās un empīriskās ticamības metodēm, salīdzināt rezultātus;

4. izdarīt secinājumus par empīriskās ticamības metodes priekšrocībām un uzlabojumiem, salīdzinājumā ar maksimālās ticamības metodi.

Darbs sastāv no trīs nodaļām, ievada, secinājumiem, izmantotās literatūras saraksta, anotācijas latviešu un angļu valodā un pielikuma. Pirmajā nodaļā ir aprakstītas kauzālo efektu novērtēšanas pamatidejas un definīcijas, randomizētie eksperimenti, tieksmes rādītāja saskaņošanas algoritms, kā arī papildu pieeja - metodes efekta novērtēšana ar inversās varbūtības svērto novērtētāju (IPW). Otrajā nodaļā raksturota empīriskās ticamības metode vispārīgi un vispārināto lineāro modeļu kontekstā, kā arī konkrēti - loģistiskās regresijas gadījumā, apskatīti arī simulāciju rezultāti. Trešajā nodaļā raksturota ICCS pētījuma datu kopa un skatīts praktisks piemērs, kur ICCS datiem pielieto tieksmes rādītāja saskaņošanas algoritmu, novērtējot tieksmes rādītāju ar maksimālās un empīriskās ticamības metodēm un salīdzinot rezultātus.

1. TIEKSMES RĀDĪTĀJA SASKAŅOŠANAS ALGORITMS KAUZĀLO EFEKTU NOVĒRTĒŠANAI

1.1. Kauzālo efektu novērtēšana - pamatidejas

Kad vēlas rast atbildi uz jautājumu, vai kāda rīcība ir cēlonis iegūtajam rezultātam, piemēram, vai noteiktā skolā īstenota jauna mācību metode ir cēlonis augstākiem skolēnu rezultātiem eksāmenā, nepietiek vien atrast sakarību - skola, kurā pielietoja šo metodi, pārbaudījumā vidēji ieguva vairāk punktu nekā citas skolas - un no tā secināt, ka pati metode rada labākus mācību rezultātus. Var būt citi iznākumu ietekmējošie faktori, piemēram, skolas lokācija, skolēnu skaits, skolas vieta reitingā, kas nav saistīti ar pašu mācību metodi. Paļaujoties vien uz to, ka pastāv noteikta sakarība starp mainīgajiem, bet neņemot vērā citus ietekmējošos faktorus, var nonākt pie kļūdainiem cēloņsakarības pieņēmumiem, kas attiecīgi tālāk var novest pie kļūdainiem lēmumiem.

Matemātiski to, ka asociācija ne vienmēr norāda uz kauzalitāti (cēloņsakarību), var parādīt sekojoši. Pieņemsim, ka T apzīmē jaunu mācību metodi, kuras efektu vēlamies novērtēt.

$$T_i = \begin{cases} 1, & \text{ja indivīds } i \text{ izmantoja noteikto metodi;} \\ 0, & \text{pretējā gadījumā.} \end{cases} \tag{1.1}$$

Y_i apzīmē i-tā indivīda iznākumu, Y_{0i} ir potenciālais i-tā indivīda iznākums, neizmantojot iepriekš minēto metodi, bet Y_{1i} - potenciālais attiecīgā indivīda iznākums, izmantojot šo metodi.

1.1. Definīcija. Individuālais metodes T efekts ir starpība $Y_{1i} - Y_{0i}$. [2]

Realitātē individuālais metodes efekts katram izlases elementam nav zināms, jo pieejama informācija tikai par vienu no iznākumiem (Y_{1i} vai Y_{0i}). Vieglāk novērtējami kauzālā efekta mēri ir vidējais metodes efekts (apzīmē - ATE) un vidējais metodes efekts tiem, kas šo metodi pielietoja (apzīmē - ATT).

1.2. Definīcija. $ATE = E[Y_1 - Y_0]$, bet $ATT = E[Y_1 - Y_0|T = 1]$. [2]

1.3. Definīcija. Asociāciju (darbā tiks apzīmēts ar A) var definēt kā iznākumu vidējo vērtību starpību atkarībā no tā, vai metode T tiek pielietota, jeb $A = E[Y|T = 1] - E[Y|T = 0]$. [2]

Asociācijas pierakstā novērotos iznākumus aizvietojot ar potenciālajiem (Y_0 - potenciālais iznākums tiem, kas neizmantoja metodi, Y_1 - tiem, kas izmantoja), iegūst vienādojumu 1.2.

$$A = E[Y|T = 1] - E[Y|T = 0] = E[Y_1|T = 1] - E[Y_0|T = 0] \tag{1.2}$$

Pēcāk vienādojumam 1.2 pieskaita un atņem $E[Y_0|T=1]$ jeb tā saucamo "alternatīvo iznākumu". Šis lielums parāda, kāds būtu bijis iznākums grupai, kas izmantoja jauno metodi, ja tā šo metodi nebūtu izmantojusi. Iegūst izteiksmi 1.3:

$$A = E[Y|T=1] - E[Y|T=0] = E[Y_1|T=1] - E[Y_0|T=0] + E[Y_0|T=1] - E[Y_0|T=1] = \\ = E[Y_1 - Y_0|T=1] + \{E[Y_0|T=1] - E[Y_0|T=0]\} = ATT + \{E[Y_0|T=1] - E[Y_0|T=0]\}, \tag{1.3}$$

kur $E[Y_0|T=1] - E[Y_0|T=0]$ ir kauzālā efekta novērtējuma novirze. [2] Asociācija ir ekvivalenta kauzalitātei vienīgi gadījumā, ja $E[Y_0|T=1] = E[Y_0|T=0]$. Ja $E[Y_0|T=1] > E[Y_0|T=0]$, efekta novērtējums būs novirzīts un tādējādi nesniegs objektīvu informāciju.

Turpmāk darbā aprakstīto metožu mērķis ir līdzsvarot testa un kontroles grupas, lai tās savstarpēji būtu pēc iespējas līdzīgākas un tādējādi tiktu novērsta kauzālā efekta novērtējuma novirze. Ja šīs grupas atšķiras tikai ar to, vai tām ir pielietota metode, kuras efekts tiek pētīts, vai nē, tad $E[Y_0|T=1] = E[Y_0|T=0]$. Attiecīgajā gadījumā kontroles un testa grupu iznākumu vidējās vērtības starpība kļūst par nenovirzītu ATT novērtētāju. Skatīt izvedumu 1.4 [2]:

$$E[Y_1 - Y_0|T=1] = E[Y_1|T=1] - E[Y_0|T=1] = E[Y_1|T=1] - E[Y_0|T=0] = \\ = E[Y|T=1] - E[Y|T=0]. \tag{1.4}$$

Ja testa un kontroles grupas ir savstarpēji salīdzināmas jeb līdzīgas pēc citiem parametriem un atšķiras tikai atkarībā no metodes izmantošanas, tad spēkā ir arī $E[Y_1|T=1] = E[Y_1|T=0]$, no kā izriet

$$E[Y_1 - Y_0|T=1] = E[Y_1 - Y_0|T=0], \tag{1.5}$$

kas, savukārt, dod

$$E[Y|T=1] - E[Y|T=0] = ATT = ATE. \tag{1.6}$$

Šajā gadījumā starpība $E[Y|T=1] - E[Y|T=0]$ kļūst ne tikai par ATT, bet arī ATE jeb vidējā metodes efekta (neatkarīgi no grupas) nenovirzītu novērtētāju.[2]

1.2. Randomizētie eksperimenti

Ja testa un kontroles grupas ir savstarpēji salīdzināmas, tad izpildās vienādība $E[Y|T=1] - E[Y|T=0] = ATT = ATE$ un novirzītība vairs nav šķērslis korektai kauzālā efekta novērtēšanai. Viens no veidiem, kā šo panākt, ir veikt tā sauktos "randomizētos

eksperimentus". Randomizēto eksperimentu ietvaros no populācijas pēc nejaušības principa atlasa dalībniekus, kuri būs kontroles, bet kuri - testa grupā. Proporcijai nav obligāti jābūt 50/50, procentuālais sadalījums grupās var būt arī cits. Randomizētie eksperimenti novērš novirzi, nodrošinot, ka

$$(Y_0,Y_1) \perp\!\!\!\perp T \tag{1.7}$$

jeb - potenciālie iznākumi ir neatkarīgi no metodes izmantošanas. [2] Jāpiemin, ka 1.7 nenozīmē pieņēmumu $Y \perp\!\!\!\perp T$, jo tad reālais iznākums būtu neatkarīgs no metodes izmantošanas un eksperiments zaudētu jēgu.1.7 nozīmē, ka metodes izmantošana ir vienīgais, kas abām grupām ietekmē pētījuma iznākumu, līdz ar to metodes efektu var novērtēt, vienkārši aprēķinot starpību kontroles un testa grupu vidējiem iznākumiem. [2] No 1.7 izriet

$$E[Y_0|T = 0] = E[Y_0|T = 1] = E[Y_0], \tag{1.8}$$

kas, attiecīgi, dod

$$E[Y|T = 1] - E[Y|T = 0] = E[Y_1 - Y_0] = ATE. \tag{1.9}$$

Kā piemērus randomizētajiem eksperimentiem var minēt klīniskos pētījumus un A/B testēšanu.

Klīniskos pētījumus veic, lai noskaidrotu kādas jaunas ārstēšanas metodes (zāļu, ķirurģiskas procedūras, ierīces utml.) ietekmi uz cilvēku veselību. Tas ir nākamais solis pēc testiem laboratorijās. Pētījumos brīvprātīgi var pieteikties visa vecuma cilvēki. Klīniskajiem pētījumiem ir 4 fāzes. 1. fāzē uz nelielu cilvēku grupu pēta ārstēšanas drošumu un iespējamās blakusparādības, bet 2. fāzē tās ārstēšanas metodes, kas pēc 1. fāzes ir apstiprinātas kā drošas, testē uz lielāku cilvēku grupu. 3. fāzē ārstēšanas ietekmi pēta arī dažādās valstīs, bet 4. fāzē ilgāku laika posmu un plašā populācijā notiek tālāka testēšana .[14]

A/B testēšanu (to dēvē arī par tiešsaistes kontrolēto eksperimentēšanu vai nepārtraukto eksperimentēšanu) plaši izmanto uzņēmumos, kas tai skaitā veic programmatūras izstrādi, piemēram, *Google*, *Meta* un *Microsoft*. Bet, protams, to izmanto arī citās sfērās, piemēram, mārketingā un medijos. Tehnoloģiju uzņēmumu gadījumā A/B testēšanas mērķis ir salīdzināt divus (reizēm arī vairāk, tad tā attiecīgi ir, piemēram, A/B/C testēšana) programmatūras variantus no gala lietotāja viedokļa, lai izdarītu secinājumus par iespējamiem uzlabojumiem. Sākotnēji pēc nejaušības principa tiek izvēlēti lietotāji, kuriem eksperimentālos nolūkos noteiktu laiku piedāvā atšķirīgu grafisko izkārtojumu vai kādas papildu funkcijas. Iepriekš tiek izvirzītas arī hipotēzes. Pēcāk analizē tādus rādītājus kā tīmekļa vietnes apmeklētāju klikšķu skaitu, MLV (angl. - *Member Lifetime Value*, rādītājs, lai naudas izteiksmē novērtētu, cik vērts biznesam ir

noteiktais klients, MLV = pakalpojuma abonēšanas mēneša/gada maksa $\times$ mēneši/gadi, cik ilgi bijis klients), kā arī cik bieži lietotājs izmanto vietnes piedāvātos pakalpojumus (publicē tīmekļa ierakstu, veic pirkumu, izveido profilu, nosūta ziņu utml). Tiek veikta hipotēžu pārbaude un izdarīti secinājumi. [18]

1.3. Tieksmes rādītāja saskaņošanas algoritms

Lai arī randomizētie eksperimenti tiek saukti par "zelta standartu" kauzālo efektu novēršanai, tos ne vienmēr ir iespējams veikt ētisku vai loģistisku apsvērumu dēļ (kā, piemēram, šī darba gadījumā, kad tiks apstrādāti iepriekš iegūti dati). Tad nepieciešama metode, ar kuru, pielietojot jau esošos datus, varētu iegūt salīdzināmas testa un kontroles grupas, lai pēcāk korekti novērtētu kauzālo efektu. Šeit noder tieksmes rādītāja saskaņošanas algoritms (angl. - *Propensity score matching*).

Līdzsvarojošais rādītājs (angl. - *balancing score*) $b(x)$, kur x ir kovariāti, ir tāda funkcija, kur pie dota $b(x)$ kovariātu nosacītais sadalījums ir vienāds gan testa $(T = 1)$, gan kontroles $(T = 0)$ grupām jeb

$$x \perp\!\!\!\perp T|b(x). \tag{1.10}$$

Vienkāršākais līdzsvarojošais rādītājs ir $b(x) = x$. Tomēr realitātē būtu gandrīz neiespējami izveidot testa un kontroles grupas, kur katram elementam vienā grupā ir atbilstošs elements otrā grupā tā, lai visu kovariātu vērtības sakristu (it īpaši, ja to ir daudz). Daudz vieglāk šo grupu līdzsvarošanu būtu veikt, nosakot tā saucamo "tieksmes rādītāju" (angl. - *propensity score*), kur būtu jāņem vērā tikai šī viena rādītāja vērtība.

Nosacīto varbūtību $e(x)$, ka pie dotajiem kovariātiem indivīdam tiks pielietota pētītā metode (un attiecīgi tas tiks iekļauts testa grupā), aprēķina

$$e(x) = p(T = 1|x), \tag{1.11}$$

pieņemot, ka

$$p(T_1, \ldots, T_n|x_1, \ldots, x_n) = \prod_{i=1}^{N} e(x_i)^{T_i} \{1 - e(x_i)\}^{1-T_i}. \tag{1.12}$$

Rezultāts $e(x)$ ir iepriekš minētais tieksmes rādītājs.[23]

1.3.1. Svarīgākie teorētiskie rezultāti līdzsvarojošā rādītāja, tai skaitā tieksmes rādītāja, sakarā

1.1. Teorēma. *Ņemot vērā tieksmes rādītāju, metodes izmantošana un novērotie kovariāti ir nosacīti neatkarīgi [23] jeb*

$$x \perp\!\!\!\perp T|e(x). \tag{1.13}$$

No 1.1. Teorēmas var secināt, ka tieksmes rādītājs ir līdzsvarojošais rādītājs.

1.2. Teorēma. *Pieņem, ka b(x) ir funkcija no x. Tad b(x) ir līdzsvarojošais rādītājs jeb*

$$x \perp\!\!\!\perp T|b(x) \tag{1.14}$$

tad un tikai tad, ja b(x) ir smalkāks par e(x) jeb

$$e(x) = f\{b(x)\}, \tag{1.15}$$

kur f - kāda funkcija. [23]

No 1.2. Teorēmas var secināt, ka jebkurš rādītājs, kas ir smalkāks par tieksmes rādītāju, ir līdzsvarojošais rādītājs. Turklāt x ir smalkākais līdzsvarojošais rādītājs, turpretim $e(x)$ - rupjākais. Pierādījumu var skatīt zinātniskā izdevuma "Biometrika" 1983. gada aprīļa numura 44. lappusē (sk. [23]).

Pirms nākamās teorēmas formulēšanas nepieciešams definēt, ko nozīmē tas, ka metodes izmantošana (angl. - "*treatment assignment*") ir stingri ignorējama.

1.4. Definīcija. Metodes izmantošana ir stingri ignorējama [23] pie dotiem kovariātiem x, ja $\forall x$

$$(Y_0,Y_1) \perp\!\!\!\perp T|x,\ 0 < p(T = 1|x) < 1. \tag{1.16}$$

1.3. Teorēma. *Ja metodes izmantošana ir stingri ignorējama pie dota x, tad tā ir stingri ignorējama pie jebkura līdzsvarojošā rādītāja b(x) jeb [23],* $\forall x$ *izpildoties* $(Y_0,Y_1) \perp\!\!\!\perp T|x$ *un* $0 < p(T = 1|x) < 1$, $\forall b(x)$ *ir spēkā* $(Y_0,Y_1) \perp\!\!\!\perp T|b(x)$ *un* $0 < p\{T = 1|b(x)\} < 1$.

1.3. Teorēmas pierādījumu var skatīt zinātniskā izdevuma "Biometrika" 1983. gada aprīļa numura 45. lappusē (sk. [23]).

1.4. Teorēma. *Ja metodes izmantošana ir stingri ignorējama un b(x) ir līdzsvarojošais rādītājs, tad*

$$E[Y_1|b(x),T = 1] - E[Y_0|b(x),T = 0] = E[Y_1 - Y_0|b(x)], \tag{1.17}$$

kur Y_0 *un* Y_1 *šeit ir novērotie iznākumi. [23]*

No 1.4. Teorēmas izriet sekas (1.1-1.3).

1.1. Sekas. *Pieņem, ka metodes izmantošana ir stingri ignorējama. Tāpat pieņem, ka līdzsvarojošā rādītāja b(x) vērtība pēc nejaušības principa tiek atlasīta no populācijas un tad atbilstoši šai b(x) vērtībai tiek izvēlēts viens elements no testa (T=1) un viens no kontroles grupas (T=0).*

Tad šī vienību pāra iznākumu starpības matemātiskā cerība ir vienāda ar vidējo metodes efektu pie noteiktā b(x). Turklāt šī divpakāpju atlases procesā iegūto saskaņoto pāru iznākumu starpību vidējā vērtība ir nenovirzīts novērtētājs vidējam metodes efektam (ATE). [23]

1.2. Sekas. *Pieņem, ka metodes izmantošana ir stingri ignorējama. Tāpat pieņem, ka, izmantojot b(x), no populācijas tiek atlasīta tāda strata, ka b(x) vērtība ir konstanta visiem elementiem attiecīgajā stratā un katrai no iespējamajām situācijām - T=0 (metode netiek izmantota) vai T=1 (metode tiek izmantota) - atbilst vismaz viens stratas elements. Tad šīs stratas elementu vidējo vērtību starpības matemātiskā cerība ir vidējais metodes efekts pie noteiktā b(x). Turklāt šo starpību vidējais svērtais ir nenovirzīts metodes efekta (ATE) novērtētājs, kad svari ir proporcionāli populācijas daļai pie noteiktā b(x). [23]*

1.3. Sekas. *Pieņem, ka metodes izmantošana ir stingri ignorējama tā, lai līdzsvarojošajam rādītājam b(x) izpildītos* $E[Y_t|T = t,b(x)] = E[Y_t|b(x)]$. *Pieņem, ka ir spēkā arī*

$$E[Y_t|T = t,b(x)] = \alpha_t + \beta_t b(x), \text{ kur } (t = 0,1). \tag{1.18}$$

Tad

$$(\hat{\alpha_1} - \hat{\alpha_0}) + (\hat{\beta}_1 - \hat{\beta}_0)b(x) \tag{1.19}$$

ir nosacīti nenovirzīts (pie $b(x_i)$, *kur* $i = 1.\ldots,n$*) metodes efekta novērtētājs, ja* $\hat{\alpha}_t$ *un* $\hat{\beta}_t$ *ir nosacīti nenovirzīti* α_t *un* β_t *novērtētāji (piemēram, OLS novērtētāji). Turklāt*

$$(\hat{\alpha_1} - \hat{\alpha_0}) + (\hat{\beta}_1 - \hat{\beta}_0)\overline{b}, \text{ kur } \overline{b} = n^{-1}\sum_{i=1}^{n} b(x_i), \tag{1.20}$$

ir nenovirzīts vidējā metodes efekta (ATE) novērtētājs, ja datu kopas elementi ir no populācijas atlasīta vienkāršā gadījumizlase. [23]

No 1.4. Teorēmas izrietošās sekas ataino dažādās metodes, ko var pielietot, lai līdzsvarotu kontroles un testa grupas un iegūtu nenovirzītu metodes efekta novērtējumu.

Tieksmes rādītāja vērtības $e(x_i)$ novērtē no pieejamajiem datiem (T_i,x_i), kur $i = 1,\ldots,N$. Ar $prop(A|B)$ apzīmē to nosacījumu B apmierinošo vektoru (T_i,x_i) proporciju, kas apmierina arī nosacījumu A. $prop(A|B)$ nav definēts, ja neviens no vektoriem neapmierina nosacījumu B. Piemēram, $prop(T = 0|x = (1,0,1))$ ir tā elementu proporcija, kas atbilst nosacījumam $T = 0$ starp visiem datu kopas elementiem, kas atbilst nosacījumam $x = (1,0,1)$. $e(x)$ var novērtēt kā $\hat{e}(a) = prop(T = 1|x = a)$. Ja $\hat{e}(a) = 0$ vai 1, tad visi elementi, kas atbilst $x = a$, vai nu ir, vai arī nav pielietojuši pētīto metodi. Visām $\hat{e}(a)$ vērtībām, kas atbilst $0 < \hat{e}(a) < 1$, ir spēkā izlases līdzsvarotība. To parāda 1.5. Teorēma. Jāņem vērā, ka $\hat{e}(x)$ vērtības starp 0 un 1 eksistēs tikai tad, ja x pieņem relatīvi maz vērtību. [23]

1.5. Teorēma. *Pieņem, ka* $0 < \hat{e}(a) < 1$. *Tad [23]*

$$prop\{T = 0,x = a|\hat{e}(x) = \hat{e}(a)\} = prop\{T = 0|\hat{e}(x) = \hat{e}(x)\}prop\{x = a|\hat{e}(T) = \hat{e}(a)\}. \tag{1.21}$$

1.4. Sekas. *Pieņem, ka no bezgalīgas populācijas atlasa N elementus kā vienkāršu gadījumizlasi un ka x šajā populācijā var pieņemt tikai galīgu skaitu vērtību, kā arī katrai šai vērtībai izpildās* $0 < \hat{e}(x) < 1$. *Tad, kad* $N \to \infty$, *stratifikācija pēc* $\hat{e}(x)$ *(skatīt 1.2. Sekas) dod izlases līdzsvarotību (jeb spēkā ir 1.5. Teorēma) ar varbūtību 1. [23]*

1.3.2. Tieksmes rādītāja novērtēšana

Viens no veidiem, kā novērtēt tieksmes rādītāja vērtību, ir, izmantojot loģistisko regresiju. Šo metodi parasti pielieto datiem, kur atkarīgais mainīgais ir binārs (pieņem vērtību 0 vai 1), neatkarīgie mainīgie var būt gan bināri, gan nepārtraukti, un ir nepieciešams novērtēt varbūtību iestāties notikumam 0 vai notikumam 1.

Loģistisko funkciju izsaka vienādojums

$$y - d = \frac{K}{1 + Ce^{rx}} \overset{\text{ja } d=0, C=1, r-\text{neg.}}{\Rightarrow} f(x) = \frac{K}{1 + e^{-rx}}, -\infty < x < \infty, \tag{1.22}$$

kur K ir funkcijas suprēms, r norāda, cik strauji funkcija aug, bet x ir neatkarīgais mainīgais. [19] Funkcijai $f(x)$ ir spēkā

$$lim_{x\to-\infty} f(x) = 0 \text{ un } lim_{x\to+\infty} f(x) = K. \tag{1.23}$$

Loģistiskās funkcijas speciālgadījums, kas ir loģistiskās regresijas pamatā, ir standarta loģistiskā funkcija (angl. - *standard logistic function/the sigmoid function*) [13], kurai $K = 1$ un $r = 1$. Funkcijas vienādojums ir

$$f(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x}, -\infty < x < \infty. \tag{1.24}$$

Loģistiskā funkcija ir inversā funkcija *logit* funkcijai (angl. - *natural logit function*):

$$logit(p) = ln(\frac{p}{1 - p}), \text{ kur } 0 < p < 1, \tag{1.25}$$

$p/(1 - p)$ ir tā saucamā izredžu attiecība. [7]

Loģistiskās regresijas gadījumā log-izredzes novērtē kā lineāru kombināciju

$$ln\left(\frac{e(x_i)}{1 - e(x_i)}\right) = x_i\beta, \tag{1.26}$$

kur β ir regresijas koeficientu vektors, bet $e(x_i)$ - tieksmes rādītāja vērtība. [17] No 1.26, savukārt, var izteikt paša tiekmes rādītāja $e(x_i)$ novērtējumu $\hat{e}(x_i)$ [10] :

$$\hat{e}(x_i) = \frac{e^{x_i\beta}}{1 + e^{x_i\beta}} = \frac{1}{1 + e^{-x_i\beta}}. \tag{1.27}$$

Ja ir vairāk nekā divas iespējas (piemēram, nevis tikai 0 (metode netiek izmantota) vai 1 (tiek izmantota), bet vēl kāds trešais vai ceturtais iespējamais scenārijs), tad $e(x_i)$ novērtēšanai var izmantot multinomiālo loģistisko regresiju, kas ir (binārās) loģistiskās regresijas vispārinājums.[17]

1.3.3. Tieksmes rādītāja pāru saskaņošana

Kad datu kopas elementiem ir novērtētas tieksmes rādītāja vērtības, lai līdzsvarotu testa un kontroles grupas, tādējādi padarot tās salīdzināmas, jāveic pāru saskaņošana. Šim nolūkam ir pieejamas vairākas metodes, bet populārākais ir labi zināmais k-tuvāko kaimiņu algoritms, kura galvenā ideja ir datu kopas elementam atrast tuvāko pēc kādas noteiktas distances metrikas. Tieksmes rādītāja gadījumā skata šī rādītāja absolūtās vērtības starpību jeb

$$D = |e_0 - e_1|, \tag{1.28}$$

kur e_0 ir tieksmes rādītāja vērtība elementam no kontroles grupas, bet e_1 - tieksmes rādītāja vērtība elementam no testa grupas. Šāda distance pēc noklusējuma tiek izmantota arī R pakotnē $MatchIT$, kas tiks izmantota darba praktiskajā daļā. [9] Pāru saskaņošanu var veikt gan veidā 1:1, kur katram elementam no testa grupas pieskaņo tikai vienu elementu no kontroles grupas, bet tādējādi var gadīties, ka tiek zaudēti izlases elementi. Iespējams arī noteiktam elementam no testa grupas pieskaņot vairākus (k) elementus no kontroles grupas, bet tad saskaņošana var nebūt tik precīza.

Izvēlas sākumpunktu - elementu no testa grupas, kuram kontroles grupā meklē tuvāko/tuvākos elementus pēc tieksmes rādītāja vērtības. To var darīt gan nosakot, gan nenosakot maksimālo distanci starp rādītājiem, kāda drīkst būt, lai tos apvienotu pārī. Pirmajā gadījumā (angl. - "*caliper matching*") maksimālo distanci izvēlas, balstoties uz datiem, nereti 0.1 vai 0.2 (ņemot vērā, ka $0 < e(X_i) < 1$), jo, izvēloties pārāk mazu slieksni, var gadīties, ka nav iespējams atrast pāri. [8]

Iespējams izvēlēties, vai pāru saskaņošanu veikt, datu kopas elementus izmantojot atkārtoti, vai arī tikai vienreiz. Otrajā gadījumā, kad izlases elements jau ir apvienots pārī ar kādu elementu, to vairs nevar izmantot saskaņošanai pārī ar citiem elementiem. Pretējā gadījumā, piemēram, vienu kontroles grupas elementu var saskaņot pārī ar vairākiem testa grupas elementiem. [8]

Metodi var pielietot, sakārtojot izlases elementus dažādā secībā - pēc tieksmes rādītāja vērtības (augoši vai dilstoši), pēc distances (pirmos pārī apvieno elementus, starp kuriem distance ir mazākā) vai elementiem esot sakārtotiem pēc nejaušības principa. No dažādajiem variantiem, kā izmantot k-tuvāko kaimiņu algoritmu, vislabākos rezultātus (mazākā novirze un kļūda, bet variācija - nav nozīmīgi lielāka, nekā citiem variantiem) dod :

- pāru saskaņošana, kad elementi ir sakārtoti pēc nejaušības principa un netiek atkārtoti izmantoti, nosakot maksimālo distanci, kāda drīkst būt starp rādītāju vērtībām,
- pāru saskaņošana, kad elementi ir sakārtoti pēc distances un netiek atkārtoti izmantoti, nosakot maksimālo distanci, kāda drīkst būt starp rādītāju vērtībām. [3]

Galvenā problēma, kas rodas, izmantojot k-tuvāko kaimiņu algoritmu, ir tā, ka pāru saskaņošanas rezultāts ir ļoti atkarīgs no sākumpunkta izvēles. Veidojot pāri, algoritms izvēlas mazāko distanci starp diviem elementiem, kas dod labāko rezultātu konkrētajā solī, neapskatot kopumu. Rezultātā kopējā distance var atšķirties atkarībā no izvēlētā sākumpunkta un ne vienmēr iznākums patiešām būs optimālākais.[8] Paul R. Rosenbaum 1989. gada rakstā "*Optimal Matching for Observational Studies*" (skatīt [22]) piedāvā optimālo pāru saskaņošanu, kas testa-kontroles pāru izveidi formulē kā minimālo "izmaksu" plūsmas problēmu (ar izmaksām šeit domājot distanci starp testa un potenciālo kontroles grupas elementu, ar kuru varētu izveidot pāri), tādējādi apvienojot statistikas un operāciju pētīšanas metodes. Šis algoritms ļauj iegūt globāli distanci minimizējošus pārus (jeb algoritms aplūko visus iespējamos veidus, kā varētu veikt pāru saskaņošanu, un izvēlas to, kas dod mazāko kopējo distanču summu). Tomēr Peter C. Austin 2013. gada publikācijā "*A comparison of 12 algorithms for matching on the propensity score*" (skatīt [3]), balstoties uz simulāciju rezultātiem, secina, ka optimālā pāru saskaņošana nedod nozīmīgi labākus rezultātus par k-tuvāko kaimiņu algoritmu.

1.4. Metodes efekta novērtēšana ar inversās varbūtības svērto novērtētāju (IPW)

Metodes efektu iespējams novērtēt kā vidējo vērtību starpību sabalansētajās testa un kontroles grupās, kas iegūtas pēc tieksmes rādītāja saskaņošanas algoritma. Taču cita pieeja ir ATE novērtēšana, izmantojot inversās varbūtības svērto novērtētāju (angl. - *"inverse-probability weighting"*), kas dod nenovirzītu ATE novērtējumu, ja tieksmes rādītāja vērtība noteikta pareizi. To aprēķina

$$\widehat{ATE}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i Y_i}{e(x_i)} - \frac{(1-T_i)Y_i}{1-e(x_i)}\right). \tag{1.29}$$

Tā kā ATE_{IPW} var būt novirzīts gadījumā, kad tieksmes rādītāja novērtējums nav precīzs, literatūrā atrodams arī papildinātais inversās varbūtības svērtais novērtētājs (angl. - *"augmented inverse-probability weighting"*), kur tiek iekļauta arī potenciālo iznākumu matemātisko cerību modelēšana.

$$\widehat{ATE}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i Y_{1i}}{e(x_i)} - \frac{T_i - e(x_i)}{e(x_i)} m_1(x_i,\xi_1) - \frac{(1-T_i)Y_{0i}}{1-e(x_i)} + \frac{T_i - e(x_i)}{1-e(x_i)} m_0(x_i,\xi_0)\right), \tag{1.30}$$

kur $m_1(x_i,\xi_1)$ un $m_0(x_i,\xi_0)$ ir regresijas modeļi, kas modelē attiecīgi $E(Y_1|X = x)$ un $E(Y_0|X = x)$, bet ξ_1 un ξ_0 ir divi papildu parametri. [25] [20] [21]

2. EMPĪRISKĀS TICAMĪBAS METODE

2.1. Empīriskās ticamības metode - pamatidejas

Empīriskās ticamības metode ir neparametriska pieeja dažādu statistisko modeļu parametru novērtēšanai. Viens no galvenajiem empīriskās ticamības metodes pozitīvajiem aspektiem (atšķirībā no parametriskas pieejas) ir tas, ka nav nepieciešams pieņēmums par noteiktu sadalījumu, bet parametru novērtēšana tiek veikta, balstoties uz datiem un izmantojot momentu nosacījumus (angl. - *moment conditions*).

Šajā sadaļā tiks aplūkotas empīriskās ticamības metodes pamatidejas, bet nākamajā - metodes ideju paplašinājums, lai to varētu izmantot vispārināto lineāro modeļu (GLM) gadījumā.

Pieņem, ka $X_1,\ldots,X_n$ ir *iid* novērojumi - diskrēti gadījuma lielumi, kur $X_i \in R^p$, $i = 1,\ldots,n$. F_0 ir tiem atbilstošā sadalījuma funkcija. Tāpat pieņem, ka sadalījuma vidējā vērtība $\mu_0 = \int xdF_0(x)$ jau ir zināma, $\mu_0 \in R^p$. Mērķis ir atrast tādu sadalījuma parametru maksimālās ticamības novērtētāju, kas atbilstu attiecīgajam ierobežojumam par vidējo vērtību. [4]

Varbūtību masas funkcijai P ir spēkā

$$P(X_i) = F(X_i) - F(X_i^-) \overset{apz.}{=} p_i, \tag{2.1}$$

kas nozīmē, ka šajā gadījumā no varbūtību masas funkcijas P izriet arī kumulatīvā sadalījuma funkcija F. [4] Iegūst P ticamības funkciju 2.2

$$L_n(P) = \prod_{i=1}^{n} p_i, \tag{2.2}$$

Izmantojot 2.2, var izteikt empīriskās ticamības "izredžu attiecību" R [11], kur

$$R = \frac{L_n(P)}{\prod_{i=1}^{n} \frac{1}{n}} = \frac{\prod_{i=1}^{n} p_i}{\prod_{i=1}^{n} \frac{1}{n}} = \prod_{i=1}^{n} p_i n. \tag{2.3}$$

Lai vienkāršotu aprēķinus, no 2.2 var izteikt log-ticamības funkciju 2.4

$$l_n(P) = \sum_{i=1}^{n} log\ p_i. \tag{2.4}$$

Empīriskās ticamības metodes mērķis ir atrast tādu P, kas maksimizē 2.4., vienlaikus izpildoties arī $\mu_0 = \int xdF(x) = \sum_{i=1}^{n} p_i X_i$, jeb atrast tādu

$$\widehat{P} = \text{argmax}_{p_1,\ldots,p_n} \sum_{i=1}^{n} log\ p_i, \tag{2.5}$$

ka spēkā ir ierobežojumi

$$\begin{cases} \sum_{i=1}^{n} p_i X_i = \mu_0 \\ \sum_{i=1}^{n} p_i = 1, \quad p_i \geq 0. \end{cases} \tag{2.6}$$

Pieņem, ka eksistē funkcija $g \in \mathbb{R}^p : g(X_i) = X_i - \mu$ un $E[g(X_i)] = 0$, tad empīrisko ticamību var aprēķināt

$$\begin{gathered} \widehat{P} = \text{argmax}_{p_1,\ldots,p_n} \sum_{i=1}^{n} \log p_i \\ \begin{cases} \sum_{i=1}^{n} p_i g(X_i) = 0 \\ \sum_{i=1}^{n} p_i = 1, \quad p_i \geq 0, \end{cases} \end{gathered} \tag{2.7}$$

kur $P = (p_1, \ldots, p_n)$ un $\widehat{P} = (\widehat{p_1}, \ldots, \widehat{p_n})$. [4] Lai atrisinātu optimizācijas problēmu 2.7, izmanto Lagranža reizinātāju metodi :

$$\mathfrak{L}(p_i,\lambda,\mu) = \sum_{i=1}^{n} log\ p_i - \lambda^T \sum_{i=1}^{n} p_i g(X_i) - \mu(\sum_{i=1}^{n} p_i - 1) \underset{p_1,\ldots,p_n,\lambda,\mu}{\rightarrow} min \tag{2.8}$$

Atvasinot 2.8 pēc p_i, iegūst

$$\frac{1}{p_i} = \lambda^T g(X_i) + \mu \Rightarrow p_i = \frac{1}{\mu + \lambda^T g(X_i)}. \tag{2.9}$$

Lai atrastu μ, pietiek secināt, ka 2.9 dod $1 = p_i(\mu + \lambda^T g(X_i))$, ko kā, savukārt, izriet

$$n = \sum_{i=1}^{n} p_i(\mu + \lambda^T g(X_i)) = \mu + \lambda^T \sum_{i=1}^{n} p_i g(X_i) = \mu, \tag{2.10}$$

jo $\sum_{i=1}^{n} p_i g(X_i) = 0$ (skatīt 2.7) [4]. Ievietojot 2.10 vienādojumā 2.9, iegūst

$$p_i = \frac{1}{n + \hat{\lambda}^T g(X_i)}. \tag{2.11}$$

Šeit $\hat{\lambda}$ apmierina nosacījumu

$$0 = \sum_{i=1}^{n} \frac{g(X_i)}{n + \hat{\lambda}^T g(X_i)}. \tag{2.12}$$

Svarīgi atzīmēt, ka n var iekļaut $\hat{\lambda}$ (skatīt avotu [4]), tādējādi 2.11 un 2.12 vietā iegūstot

$$p_i = \frac{1}{1 + \hat{\lambda}^T g(X_i)}, \text{ kur } \sum_{i=1}^{n} \frac{g(X_i)}{1 + \hat{\lambda}^T g(X_i)} = 0. \tag{2.13}$$

Statistisko testu hipotēžu pārbaudei būtisks ir rezultāts, ka, izpildoties iepriekš definētajam regularitātes nosacījumam par funkciju $g(X)$ ($g \in \mathbb{R}^p : E[g(X)] = 0$), R (skatīt 2.3) asimptotiski (pēc sadalījuma) tiecas uz χ^2 sadalījumu jeb $-2logR \underset{n\to\infty}{\rightarrow} \chi^2_p$, kur $p = dim(g(X))$. [15]

Lai noteiktu vidējās vērtības ticamības apgabalu, vispirms definē tā saucamo ”profilēto empīriskās ticamības attiecību” $\Re(\mu)$:

$$\Re(\mu) = \sup\{\prod_{i=1}^{n} np_i | p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i X_i = \mu\}. \tag{2.14}$$

Tā kā $-2log\Re(\mu) \underset{n\to\infty}{\to} \chi_p^2$,, tad μ_0 atbilstošo $(1-\alpha)$- ticamības apgabalu var konstruēt

$$C_{\mu_0} = \{\mu \in R^p | -2log\Re(\mu) \leq c_\alpha\}, \tag{2.15}$$

kur $c_\alpha : P(\chi_p^2 \leq c_\alpha) = 1 - \alpha$. [11]

2.2. Empīriskā ticamības metode vispārinātajiem lineārajiem modeļiem (GLM)

Empīriskās ticamības metodi var izmantot modelēšanā arī vispārināto lineāro modeļu gadījumā (GLM), kā arī tad, ja datu kopas elementi nav *iid* (atšķirībā no iepriekšējā sadaļā formulētā par gadījuma lielumiem X_i). To demonstrē 2.1. Teorēma.

2.1. Teorēma. *Pieņem, ka $Z_{in} \in R^p$, kur $1 \leq i \leq n$ un $p \leq n < \infty$ ir tāds vektoru kopums, ka $Z_{1n}, \ldots, Z_{nn}$ ir neatkarīgi $\forall n$. Tāpat pieņem, ka $E[Z_{in}] = \mu_n, var(Z_{in}) = V_{in}$. Ja spēkā ir nosacījumi 1.-3.:*

1. $P(\mu_n \in ch(\{Z_{1n}, \ldots, Z_{nn}\})) \underset{n\to\infty}{\to} 1$
2. $n^{-2} \sum_{i=1}^{n} E[||Z_{in} - \mu_n||^4 \sigma_{1n}^{-2}] \underset{n\to\infty}{\to} 0$
3. $\exists c > 0 : \forall n \geq p \; \sigma_{pn}/\sigma_{pn} \geq c$,

tad $-2log\Re(\mu_n) \underset{n\to\infty}{\to} \chi_p^2$, (pēc sadalījuma), kur

$$\Re(\mu) = sup\{\prod_{i=1}^{n} np_i | p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i (Z_{in} - \mu) = 0\}. \tag{2.16}$$

2.1. Teorēma pirmo reizi minēta Art Owen 1991. gada rakstā ”*Empirical likelihood for linear models*”, kas publicēts izdevumā ”The Annals of Statistics”. [16] Z_{in} teorēmā reprezentē vienkārši gadījuma lielumus, kuriem nav obligāti jābūt *iid*, taču šo teoriju var attiecināt arī uz tā saucamajām ”novērtēšanas funkcijām” (angl. - *estimating functions*).

Iepriekšējā sadaļā par empīriskās ticamības metodes pamatidejām tika aplūkots gadījums, kad novērtēšanas funkcija $g(X_i)$ ir pieņem formu $g(X_i) = X_i - \mu$, bet tas ir tikai viens noteikts gadījums. Turpmākajā izklāstā novērtēšanas funkcijas tiks aprakstītas vispārinātā formā.

Pieņemsim, ka $\exists(Y_1,X_1),\ldots,(Y_n,X_n) : Y_i \in R$ ir neatkarīgi gadījuma lielumi, bet $X_i \in R^p$ ir fiksēti kovariāti. Tāpat pieņem, ka $f(W_i;\theta)$ ir tādu reālu funkciju vektors, kam izpildās

$$f(W_i;\theta) = \{f_1(W_i;\theta),\ldots,f_p(W_i;\theta)\}^T \text{ un } E[f(W_i;\theta)] = 0. \tag{2.17}$$

Šeit $W_i = (Y_i,X_i)$ un $\theta \in R^p$ ir novērtējamais parametrs. [11] No 2.1. Teorēmas izriet, ka pie $Z_{in} = f(W_{in};\theta)$ un izpildoties $E[f(W_{in};\theta)] = 0$ pie "īstās/patiesās" θ vērtības, $-2log\mathfrak{R}(\theta) \underset{n\to\infty}{\to} \chi^2_p$, (pēc sadalījuma), kur

$$\mathfrak{R}(\theta) = \sup\{\prod_{i=1}^{n} np_i | p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i f(W_{in};\theta) = 0\}. \tag{2.18}$$

Turklāt, ja definē $\Psi_{in} = var\{f(W_{in};\theta)\}$,$\Psi_n = \frac{1}{n}\sum_{i=1}^{n}\Psi_{in}$,$\psi_{1n} = maxeig(\Psi_n)$, $\psi_{pn} = mineig(\Psi_n)$, tad 2.1. Teorēmas 1.-3. nosacījumus var aizstāt ar zemāk esošajiem [11]:

$$P(0 \in ch[\{f(W_{1n};\theta),\ldots,f(W_{nn},\theta)\}]) \underset{n\to\infty}{\to} 1 \tag{2.19}$$

$$n^{-2}\sum_{i=1}^{n} E[||f(W_{in};\theta)||^4 \psi_{1n}^{-2}] \underset{n\to\infty}{\to} 0 \tag{2.20}$$

$$\exists c' > 0; \forall n \geq p, \psi_{pn}/\psi_{1n} \geq c'. \tag{2.21}$$

Ja dotajiem datiem pieņemtajā modelī ir iekļauta novērtēšanas funkciju kopa, kas ir iepriekš aprakstītajā formā (skatīt 2.17) un apmierina nosacījumus 2.19, 2.20 un 2.21, tad empīriskās ticamības metodes ticamības apgabalu izmantošana ir teorētiski pamatota. [11]

Lai apskatītu empīriskās ticamības metodes pielietojumu vispārinātajiem lineārajiem modeļiem, nepieciešams definēt: $E[Y_i,X_i] = \mu_i, G(\mu_i) = X_i^T\theta, \text{Var}(Y_i|X_i) = \phi V(\mu_i)$, kur $\theta = (\theta_0,\ldots,\theta_{p-1})$ ir nezināms p-dimensionāls parametrs, ϕ - dispersijas parametrs, G ir gluda saites funkcija, kas ir dota, un V ir variācijas funkcija, kas arī ir zināma. Ar H apzīmē G inverso jeb $H = G^{-1} : \mu_i = H(X_i^T\theta)$. [11] [5]

Gadījumā, kad ϕ ir fiksēts un zināms, Y_i kvazi-ticamību definē kā

$$Q(\mu_i;Y_i) = \int_{Y_i}^{\mu_i} \frac{Y_i - t}{\phi V(t)} dt, \tag{2.22}$$

bet kvazi-ticamību visai datu kopai, attiecīgi, definē

$$Q(\mu;Y) = \sum_{i=1}^{n} Q(\mu_i,Y_i). \tag{2.23}$$

Pēcāk, atvasinot 2.22 pēc θ, iegūst tā saucamo "kvazi-rādītāju" (angl. - *quasi-score*) [11] [5]:

$$\frac{\partial Q(\mu_i;Y_i)}{\partial\theta} = \frac{Y_i - \mu_i}{\phi V(\mu_i)}\frac{\partial\mu_i}{\partial\theta} = (\frac{H'(X_i^T\theta)(Y_i - H(X_i^T\theta))}{\phi V(H(X_i^T\theta))})X_i \overset{apz.}{=} Q. \tag{2.24}$$

$E[Q] = 0$, tāpēc

$$f(W_i;\theta) = Z_i = Z\{(Y_i,X_i);\theta\} = Q \in R^p. \tag{2.25}$$

Profilētā empīriskās ticamības attiecība ir formā

$$\mathfrak{R}(\theta) = \sup\{\prod_{i=1}^{n} np_i | p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i Q = 0\}. \tag{2.26}$$

Izmantojot 2.1. Teorēmas rezultātu, var secināt, ka maksimālās empīriskās ticamības metodes θ novērtētājs ir

$$\hat{\theta} = arg \max_{\theta \in R^p} \mathfrak{R}(\theta), \tag{2.27}$$

kur θ ticamības apgabala noteikšanai izmanto iepriekš aprakstīto $-2log\mathfrak{R}(\theta) \underset{n\to\infty}{\to} \chi_p^2$ aproksimāciju. [11]

Gadījumā, ja ϕ arī nav zināms, nepieciešama papildu novērtēšanas funkcija. Pieņem, ka $\eta = (\theta,\phi)$. [5] Tad atbilstošā novērtēšanas funkcija būs

$$f(W_i;\eta) = \frac{(Y_i - H(X_i^T\theta))^2}{\phi^2 V(H(X_i^T\theta))} - \frac{1}{\phi}, \tag{2.28}$$

bet profilētā empīriskās ticamības attiecība šoreiz būs formā

$$\mathfrak{R}(\eta) = sup\{\prod_{i=1}^{n} np_i | \sum_{i=1}^{n} p_i Q = 0, \sum_{i=1}^{n} p_i f(W_i;\eta) = 0, p_i \geq 0, \sum_{i=1}^{n} p_i = 1\}. \tag{2.29}$$

Maksimālās empīriskās ticamības metodes ϕ novērtētāju aprēķina

$$\hat{\phi} = arg \max_{\phi \in (0,\infty)} \mathfrak{R}(\eta), \tag{2.30}$$

kur ϕ ticamības intervālu nosaka, izmantojot χ^2 aproksimāciju, kas balstās uz $-2log\mathfrak{R}(\phi) \underset{n\to\infty}{\to} \chi^2_{(1)}$ (pēc sadalījuma), kur $\mathfrak{R}(\phi) = \underset{\theta \in R^p}{sup}\mathfrak{R}(\eta)$. [11]

2.3. Empīriskās ticamības metodes iekļaušana tieksmes rādītāja saskaņošanas algoritmā

Pieņemsim, ka datu kopu veido novērojumi (T_i,X_i), kur $i = 1,\ldots,n$ un T_i ir binārs lielums - metodes pielietojuma indikators:

$$T_i = \begin{cases} 0, \text{ ja pētītā metode netika pielietota,} \\ 1, \text{ ja pētītā metode tika pielietota,} \end{cases} \tag{2.31}$$

bet $X_i \in R^p$ ir kovariāti.

Viens no būtiskākajiem tieksmes rādītāja saskaņošanas algoritma stūrakmeņiem ir pēc iespējas precīza tieksmes rādītāja novērtēšana, lai pēcāk, izmantojot to, varētu līdzsvarot testa

un kontroles grupas. Tieksmes rādītāju (tāpat kā tika parādīts iepriekš - skatīt 1.26 un 1.27) var novērtēt, izmantojot loģistisko regresiju. Šo novērtējumu būtu iespējams uzlabot, regresijas koeficientu β noteikšanai izmantojot empīriskās ticamības metodi ierastās maksimālās ticamības metodes vietā.

Eric D. Kolaczyk raksta "Empirical likelihood for generalized linear models" pielikumā (skatīt [11]) pieejama tabula ar dažādiem novērtējošo funkciju piemēriem. Loģistiskajai regresijai atbilst binomiālā sadalījuma saime $(Bin(m;\mu))$, tāpēc $E[T_i] = \mu, \phi = \frac{1}{m}$, $V(\mu) = \frac{\mu}{1-\mu}, G(\mu) = log(\frac{\mu}{1-\mu}) = X_i\beta$, bet novērtējošā funkcija ir formā

$$Z_i = X_i \left(T_i - m\frac{1}{1+e^{-(X_i^T\beta)}} \right). \tag{2.32}$$

Tā kā loģistiskās regresijas gadījumā negrupētiem datiem (metode tika izmantota ($T_i = 1$) vai nē ($T_i = 0$); grupētu datu gadījumā, piemēram, būtu viena kolonna ar attiecīgās apakšgrupas elementu skaitu, bet otra kolonna - ar to šīs apakšgrupas cilvēku skaitu, kas izmantoja metodi) jāizvēlas $\phi \equiv 1$, tad arī $m = 1$. [12] Attiecīgi 2.32 kļūst par

$$Z_i = X_i \left(T_i - \frac{1}{1+e^{-(X_i^T\beta)}} \right). \tag{2.33}$$

Ar $p_i, i = 1, \ldots, n$, apzīmē izlases varbūtības, bet $\Re(\beta)$ var iegūt, atrisinot optimizācijas uzdevumu

$$\begin{cases} \sum_{i=1}^{n} \log np_i \underset{p_i}{\rightarrow} max \\ \left\{ \begin{array}{l} \sum_{i=1}^{n} p_i Z_i = 0 \\ \sum_{i=1}^{n} p_i = 1, \quad p_i \geq 0, \end{array} \right. \end{cases} \tag{2.34}$$

izmantojot Lagranža reizinātāju metodi līdzīgā veidā, kā tas raksturots 2.1. nodaļā. Iegūst tos $\hat{\beta}$, kas atbilst 2.34. $\hat{\beta}$ ticamības intervālu noteikšanai izmanto 2.2. nodaļā aprakstīto $\chi^2_{(p)}$ aproksimāciju.

Empīriskās ticamības metodi var izmantot ne vien loģistiskās regresijas koeficientu, bet arī metodes efekta (ATE) novērtēšanai. Biao Zhang 2016. gada rakstā, kas publicēts zinātniskajā izdevumā "Econometric Reviews", piedāvāts ATE novērtētājs, kura pamatā arī ir empīriskās ticamības metode.

$$ATE_{EL} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\hat{\lambda}_{5n}^{\tau} g_{(2)}} \left(\frac{T_i Y_{1i}}{e(X_i;\hat{\beta}_n)} - \frac{(1-T_i)Y_{0i}}{1-e(X_i;\hat{\beta}_n)} \right), \tag{2.35}$$

kur $\hat{\beta}_n$ apzīmē ar empīriskās ticamības metodi novērtētos loģistiskās regresijas koeficientus, Y_{1i} un Y_{0i} ir potenciālie iznākumi atkarībā no metodes izmantošanas, $e(X_i;\hat{\beta}_n)$ apzīmē tieksmes

rādītāja vērtību i-tajam izlases elementam, bet $g_{(2)}$ apzīmē novērtēšanas funkciju apvienojumu

$$g_{(2)} = \begin{pmatrix} g_2 \\ g_3 \end{pmatrix} = \frac{T - e(X;\beta)}{e(X;\beta)} h(X;\beta,\xi_1,\xi_0), \text{ kur} \tag{2.36}$$

$$h(X;\beta,\xi_1,\xi_0) = \begin{pmatrix} h_0(X;\beta,\xi_1,\xi_0) \\ e_\beta(X;\beta)/e(X;\beta)(1-e(X;\beta)) \end{pmatrix}, \text{ šeit, savukārt,} \tag{2.37}$$

$$h_0(X;\beta,\xi_1,\xi_0) = \begin{pmatrix} m_1(X;\xi_1)/e(X;\beta) \\ m_0(X;\xi_0)/(1-e(X;\beta)) \end{pmatrix}. \tag{2.38}$$

$m_0(X;\xi_0)$ un $m_1(X;\xi_1)$ nozīmei skatīt 1.4. nodaļu. $\hat{\lambda}_{5n}^{\tau}$ atrod tādu, lai izpildītos

$$\frac{1}{n}\sum_{i=1}^{n} \frac{g_{(2)}}{1+\lambda^\tau g_{(2)}} = 0. \tag{2.39}$$

Rakstā tika minēta arī novērtēšanas funkcija g_1, kur

$$g_1 = \frac{TY_1}{e(X;\beta)} - \frac{(1-T)Y_0}{1-e(X;\beta)} - ATE, \tag{2.40}$$

bet $E[g_1] = 0$, kad tieksmes rādītāja $e(X;\beta)$ vērtības ir precīzi novērtētas. [25]

2.4. Simulācijas rezultāti

Lai labāk izprastu empīriskās ticamības metodes priekšrocības loģistiskās regresijas parametru novērtēšanas kontekstā, tika veikta simulācija. Simulācijas kods pieejams pielikumā (skatīt A.1 pielikumu).

Sākotnēji tikai noteiktas "patiesās" β vērtības un ATE vērtība, kuras pēcāk novērtēt simulētajiem datiem. Viens no koeficientiem - β_3 tika izvēlēts ar vērtību 0. Salīdzināšanai (kā simulācija strādās atkarībā no izlases apjoma) tika ģenerētas izlases apjomā $n = 100, n = 300$ un $n = 500$ ar prediktoru skaitu $p = 3$. Lai radītu situāciju, kur empīriskās ticamības metodei varētu būtu priekšrocības attiecībā pret maksimālās ticamības metodi, tika ģenerēta kovariātu matrica pēc Stjūdenta sadalījuma ar 3 brīvības pakāpēm. Tika aprēķinātas "īstās" tieksmes rādītāja vērtības un, balstoties uz tām, ģenerēti metodes izmantošanas dati T - binārs mainīgais pēc binomiālā sadalījuma. Tālāk tika ģenerēts metodes ietekmētā iznākuma mainīgais Y. Neliels izaicinājums bija saprast, kā ģenerēt datus, kuriem reāli piemīt kāda sakarība, lai novērtētais metodes efekts nebūtu vien nejaušība. Y tika izteikts kā lineāra sakarība, kas apvieno *logit* modeli, metodes efektu un kļūdu. Kļūdas mainīgais arī tika ģenerēts pēc Stjūdenta sadalījuma ar 3 brīvības pakāpēm.

Simulācija katram no izlases apjomiem atkārtota 1000 reizes. Tika izrēķināti vidējie ticamības intervālu platumi katram $\hat{\beta}$ atkarībā no metodes, aprēķināts, cik procentos gadījumu 0 ietilpst $\hat{\beta}_3$ ticamības intervālā katrai no metodēm, kā arī aprēķināts ATE. ATE tika rēķināts kā vidējo vērtību starpība pēc tieksmes rādītāja saskaņošanas sabalansētajām testa un kontroles grupām.

Vidējo ticamības intervālu garumu salīdzinājuma rezultātus skatīt 1. tabulā.

Izlases apjoms	Novērtētais parametrs	Vid_CI_platums_ML	Vid_CI_platums_EL
$n = 100$	$\hat{\beta}_1$	0.8198738	0.8118522
	$\hat{\beta}_2$	0.9579817	0.9533307
	$\hat{\beta}_3$	0.6758378	0.6838774
$n = 300$	$\hat{\beta}_1$	0.4479647	0.4447540
	$\hat{\beta}_2$	0.5165945	0.5147360
	$\hat{\beta}_3$	0.3508678	0.3520528
$n = 500$	$\hat{\beta}_1$	0.3408626	0.3391782
	$\hat{\beta}_2$	0.3951475	0.3934991
	$\hat{\beta}_3$	0.2649423	0.2649452

1. tabula

Ticamības intervālu garumu salīdzinājums maksimālās un empīriskās ticamības metodēm, dažādiem izlases apjomiem

Tabulā Vid_CI_platums_ML apzīmē vidējos ticamības intervālu platumus katram no novērtētajiem parametriem, kas iegūti ar maksimālās ticamības metodi, bet Vid_CI_platums_EL - ar empīriskās (maksimālās) ticamības metodi. Ticamības intervālu platumi abām metodēm samazinās, palielinoties izlases apjomam. Tie ir ļoti līdzīgi, taču empīriskās ticamības metodei $\hat{\beta}_1$ un $\hat{\beta}_2$ ticamības intervālu platumi ir mazāki neatkarīgi no izlases apjoma. $\hat{\beta}_3$ ticamības intervāli ir šaurāki maksimālās ticamības metodei, taču var redzēt, ka, palielinoties izlases apjomam, šī starpība mazinās.

Izrēķinot, cik procentos gadījumu 0 ietilpst $\hat{\beta}_3$ ticamības intervālā katrai no metodēm atkarībā no izlases apjoma, tika iegūti šādi rezultāti: pie izlases apjoma $n = 100$ iegūti 94.7% maksimālās ticamības metodei un 92.1% empīriskās ticamības metodei, pie izlases apjoma $n = 300$ iegūti 94.2% maksimālās ticamības metodei un 93.8% empīriskās ticamības metodei, bet pie apjoma $n = 500$ iegūti 94% maksimālās ticamības metodei un 93.7% empīriskās ticamības metodei. Rezultāti ir tuvu nepieciešamajiem 95%, un, lai arī sākotnēji maksimālās ticamības metode $\hat{\beta}_3$ gadījumā uzrāda labāku rezultātu, līdzīgi ticamības intervālu platumiem, palielino-

ties izlases apjomam, starpība izlīdzinās.

Simulācijas dati tika konstruēti tā, ka "īstais" ATE = 2. ATE novērtējumi ir šādi: pie izlases apjoma $n = 100$ iegūts novērtējums ≈ 2.008 ar maksimālās ticamības metodi un ≈ 1.995 ar empīrisko ticamību, pie apjoma $n = 300$ iegūts ≈ 2.017 ar maksimālās ticamības metodi un ≈ 2.007 ar empīrisko ticamību, bet pie apjoma $n = 500$ iegūts ≈ 2.01 ar maksimālās ticamības metodi un ≈ 2.015 ar empīrisko ticamību. Tātad var secināt, ka simulāciju gadījumā abi novērtējumi ir ļoti tuvu reālajai vērtībai. Tomēr arī ATE abas metodes novērtē līdzīgi.

Kopumā var secināt, ka simulācijas gadījumā empīriskās ticamības metode īpašu uzlabojumu nedod. Datiem tika izmēģināti citi sadalījumi, piemēram, kovariātu matricas ģenerēšanā - normālais un log-normālais, kļūdas ģenerēšanā pie Y - lognormālais, arī χ^2, Gamma un Laplasa sadalījumi, taču visos gadījumos metodes darbojās aptuveni līdzvērtīgi.

Tāpat tika novērots, ka R iebūvētās funkcijas *glm()* (maksimālās ticamības metodei) un *el_glm()* (empīriskās ticamības metodei) izdod identiskus tieksmes rādītāja novērtējumus, taču vēlāk tika secināts, ka abos gadījumos parametru novērtēšanai tiek izmantotas tās pašas funkcijas, tāpēc arī šī rādītāja rezultāti sakrīt.

Lai arī simulēto datu gadījumā nevar teikt, ka empīriskās ticamības metode dotu uzlabojumu, ir vērts to aplūkot darba praktiskajā daļā, kad tiks strādāts ar reāliem datiem, un, iespējams, tad labāk varēs novērtēt empīriskās ticamības metodes priekšrocības.

3. ICCS DATU ANALĪZE

3.1. Datu kopas apraksts

Darba praktiskajā daļā tiks izmantoti ICCS (*International Civic and Citizenship Education Study*) 2022. gada Latvijas dati. ICCS ir IEA (*International Association for the Evaluation of Educational Achievement*) īstenots pētījums, kura mērķis ir noskaidrot 8. klašu skolēnu zināšanu līmeni pilsoniskajā izglītībā un viņu sagatavotību pildīt pilsoņu pienākumus. Lai arī pētījums ir aizsācies 1971. gadā, Latvija tajā līdz šim ir piedalījusies četras reizes - CivED 1999, ICCS 2009, ICCS 2016 un ICCS 2022. [24]

Pētījuma ietvaros ir iegūti 22 valstu dati no skolēniem, skolotājiem un skolu direktoriem, kā arī atsevišķi dati diviem Vācijas reģioniem, kas piedalījās kā salīdzinošās vienības (Ziemeļreina-Vestfālene un Šlēsviga-Holšteina). Papildus anketu dati iegūti no 18 Eiropas valstu (un divu Vācijas salīdzinošo vienību) un 2 Latīņamerikas valstu skolēniem. Ir pieejamas dažādas datu kopas - skolēnu anketu rezultāti (starptautiskie un papildu Eiropas un Latīņamerikas), skolēnu pilsonisko zināšanu testa rezultāti, skolēnu vērtēšanas uzticamības/objektivitātes dati (aktuāli atvērtā tipa jautājumiem pilsonisko zināšanu testā), skolotāju un skolu anketu rezultāti, ar dalībvalstu izglītības sistēmām saistīta informācija, kā arī dati par laiku, ko skolēni veltījuši, atbildot uz pilsonisko zināšanu testa jautājumiem. [1]

No Latvijas datiem darbā tiks pētīta datu kopa ISGLVAC4. Nosaukumā ”ISG” apzīmē no studentu anketu rezultātiem iegūtu izlasi, bet ”LVA” - valsti, savukārt ”C4” nozīmē, ka attiecīgā izlase tiek izmantota starptautiskajā pētījumā. Datu kopa satur 377 mainīgos un 2876 novērojumus (katrs novērojums atbilst vienam skolēnam). Tajā ir ietvertas starptautiskās skolēnu anketas atbildes (piemēram, par skolēnu izcelsmi, attieksmi, pilsonisko iesaistīšanos, pieredzi ar pilsonisko izglītošanu skolā), skolēnu identifikācijas mainīgie (piemēram, valsts, skola, klase, skolēna ID), skolēnu sociāldemogrāfiskie rādītāji (dzimums, vecums, vecāku izglītība utml.), pilsonisko zināšanu testa rezultāti, no anketas atbildēm iegūtie rādītāji (piemēram, S_ELECPART, S_DEMTHRT utt.), kā arī izlases elementu svari (mainīgais TOTWGTS apzīmē kopējo svaru, kas noteikts katram datu kopas elementam - skolēnam -, taču tas sevī apvieno vairākus smalkākus svaru mainīgos).

Lielākā daļa mainīgo datu kopā ir kategoriāli (precīzāk, ordināli pēc Likerta skalas) - tās ir skolēnu atbildes uz anketas jautājumiem noteiktā skalā (piemēram, 1 - 3 vai 1 - 4). Tie kalpo kā indikatormainīgie, kuriem pielieto IRT (angl. - ”*Item response theory*”) svērtās ticamības (angl. - ”*weighted likelihood*”) metodi, lai aprēķinātu rādītājus - datu kopas nepārtrauktos

mainīgos (kā piemēru var skatīt iepriekš minētos S_ELECPART un S_DEMTHRT), kas sniedz visaptverošāku skatījumu par interesējošo tematu, jo apvieno atbilžu rezultātus uz vairākiem attiecīgās tematikas jautājumiem. Šie rādītāji ir standartizēti tā, lai to vidējā vērtība $\mu = 50$, bet standartnovirze $SD = 10$. Tikai tie skolēni, kuriem bija pieejama informācija par vismaz 2 indikatormainīgajiem, tika iekļauti rādītāju aprēķinos. [1]

Pētītā datu kopa ISGLVAC4 ir stratificēta divpakāpju ligzdveida izlase. Pirmajā solī skolu kopums tika sadalīts stratās pēc skolas tipa un valodas (piemēram, ģimnāzijas, vidusskolas, krievu tautības skolas) un tad tika izvēlētas skolas no katras stratas. Pētījumā netika iekļautas skolas, kurās valoda nav latviešu vai krievu, tālmācības un ļoti mazās skolas, skolas bērniem ar īpašām vajadzībām vai ieslodzījuma vietās. Otrajā solī pēc nejaušības principa tika atlasīta viena 8. klase no katras pirmajā solī atlasītās skolas. Būtiski pieminēt, ka bija jāizpildās nosacījumam par klases skolēnu vidējo vecumu - ja tas izvēlētajā 8. klasē nebija vismaz 13.5 gadi, tad izlasē ņēma 9. klasi. Skola skaitās piedalījusies pētījumā, ja vismaz 50% no šīs skolas izlasē iekļautās klases skolēnu ir snieguši atbildes. Pētījumā piedalījās 147 Latvijas skolas. Katram skolēnam datu kopā ir piešķirts noteikts svars (iepriekš minētais mainīgais TOTWGTS), kas tiek aprēķināts kā

$$TOTWGTS = WGTFAC1 \times WGTADJ1S \times WGTFAC2S \times WGTADJ2S \times WGTADJ3S, \tag{3.1}$$

kur WGTFAC1S ir inversais skolas varbūtībai tikt atlasītai, WGTFAC2S - inversais klases varbūtībai tikt atlasītai, bet WGTADJ1S, WGTADJ2S un WGTADJ3S kompensē attiecīgi skolas, klases un skolēnu nerespondenci. Precīzas formulas, kā tiek aprēķināti visi svaru mainīgie, var atrast ICCS 2022 pētījuma tehniskajā atskaitē. [6]

Praktiskās daļas mērķis ir ISGLVAC4 datiem pielietot tieksmes rādītāja saskaņošanas algoritmu (un empīriskās ticamības metodi tieksmes rādītāja novērtēšanai), lai varētu rast atbildi uz jautājumu - cik lielā mērā jauniešu iesaiste skolas vēlēšanās (atbilst mainīgajam IS4G15B) ietekmē viņu potenciālo dalību nacionālajās vēlēšanās nākotnē (atbilst mainīgajam S_ELECPART)? Tā kā ICCS dati nav iegūti, veicot randomizēto eksperimentu, bet gan novērošanas pētījumu (angl. - "*observational study*"), tad korektai kauzālo efektu novērtēšanai ir būtiski izmantot šīs metodes. Rezultāts, kas iegūts, tieksmes rādītāju novērtējot ar empīriskās ticamības metodi, tiks salīdzināts ar rezultātu, ko var iegūt, novērtēšanā pielietojot klasisko maksimālās ticamības metodi.

Kā kovariāti tiks izvēlēti 10 mainīgie: SD_GENDER - skolēna dzimums, S_IMMIG - imigrācijas statuss, S_NISB - sociālekonomiskais stāvoklis, PV1CIV, PV2CIV, PV3CIV, PV4CIV

un PV5CIV - pilsonisko zināšanu testa iespējamie rezultāti (angl. - *plausible values*), šajā gadījumā vienkāršības labad tiks ņemta vidējā vērtība no piecām iespējām (pavisam precīzi būtu veikt datu analīzi piecas reizes, vienu reizi ar katru no vērtībām), S_SCACT - vēlme piedalīties skolas aktivitātēs, S_COMPART - dalība organizācijās (jauniešu, brīvprātīgajās), S_CIVLRN - pilsonisko zināšanu apguve skolā, S_DEMTHRT - izpratne par potenciālajiem draudiem demokrātijai, S_INFDEC - skolēnu pārliecība par viņu ietekmi uz lēmumu pieņemšanu skolā, S_INTRUST - skolēnu uzticēšanās valsts institūcijām, piemēram, valdībai, tiesu sistēmai, policijai. Pētāmajā datu kopā ir pieejams ļoti plašs klāsts mainīgo, taču modeļa vienkāršības un rezultātu skaidrākas interpretācijas labad ir izvēlēti minētie 10, kuri darba autorei šķita ar vislielāko iespējamo ietekmi.

3.2. Datu analīze

Datu analīze tika veikta paketē R (programmas kodu A.2 skatīt pielikumā). Sākotnēji no pilnās datu kopas tika atlasīti nepieciešamie mainīgie (IS4G15B, S_ELECPART un 10 iepriekš minētie kovariāti). Pēcāk veikta datu tīrīšana un transformēšana - tika faktorizēti kategoriālie mainīgie, atmestas rindas ar iztrūkstošām vērtībām. IS4G15B jeb skolēnu dalība skolas vēlēšanās no ordināla mainīgā (pēc Likerta skalas) tika pārveidots par bināru mainīgo ar nosaukumu ”Velesanas”(ja IS4G15B ir 1 vai 2, tad Velesanas = 1 (ir piedalījies), ja IS4G15B = 3, tad Velesanas = 0 (nav piedalījies skolas vēlēšanās)). Atmestas rindas, kur S_IMMIG un SD_GENDER vērtības ir ”9”, kas nozīmē, ka skolēns vai nu nav vēlējies norādīt dzimumu/imigrācijas statusu, vai arī kāda cita iemesla dēļ attiecīgā informācija nav pieejama. Tā kā šādu vērtību ir maz, tas darīts, lai novērstu tālākas novērtējumu neprecizitātes. No PV1CIV, PV2CIV, PV3CIV, PV4CIV un PV5CIV (pilsonisko zināšanu testa iespējamajiem rezultātiem) tika izrēķināta vidējā vērtība, iegūstot kolonnu ar nosaukumu ”Pils_zin”.

Tā kā datu kopā katram izlases elementam (skolēnam) ir piešķirts noteikts svars (mainīgais TOTWGTS), tad tas novērtējumos tiek ņemts vērā. Maksimālās ticamības metodes gadījumā bija iespējams izmantot *survey* pakotnes funkciju *svydesign()*, lai definētu izlases dizainu, un to savienot ar *svyglm()* funkciju loģistiskajai regresijai. Tomēr *melt* pakotnes *el_glm()* funkcijai, ko izmantoja empīriskās maksimālās ticamības metodes gadījumā, šādas sasaistes nebija, līdz ar to novērtēšanā nevarēja pilnā apmērā ņemt vērā izlases dizainu (stratificēta divpakāpju ligzdveida izlase). Tādēļ, lai abu metožu rezultāti būtu salīdzināmi, abos gadījumos ņēma vērā tikai izlases svarus, kas uzskatāms par ierobežojumu.

Vispirms tieksmes rādītājs tika novērtēts, veicot loģistisko regresiju un parametrus novēr-

tējot ar maksimālās ticamības metodi, ņemot vērā arī izlases svarus. Ieguva zemāk redzamo izdruku (2. tabula).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.7793	0.5202	-9.19	< 2e-16
SD_GENDER1	0.1583	0.0919	1.72	0.0850
S_IMMIG2	0.3516	0.2617	1.34	0.1791
S_IMMIG3	-0.0032	0.5136	-0.01	0.9950
S_NISB	0.0135	0.0491	0.27	0.7834
Pils_zin	0.0036	0.0007	4.99	6.47e-07
S_SCACT	0.0216	0.0039	5.48	4.72e-08
S_COMPART	0.0288	0.0049	5.82	6.48e-09
S_CIVLRN	0.0126	0.0049	2.56	0.0106
S_DEMTHRT	-0.0149	0.0054	-2.75	0.0061
S_INFDEC	0.0247	0.0061	4.07	4.75e-05
S_INTRUST	-0.0107	0.0050	-2.16	0.0312

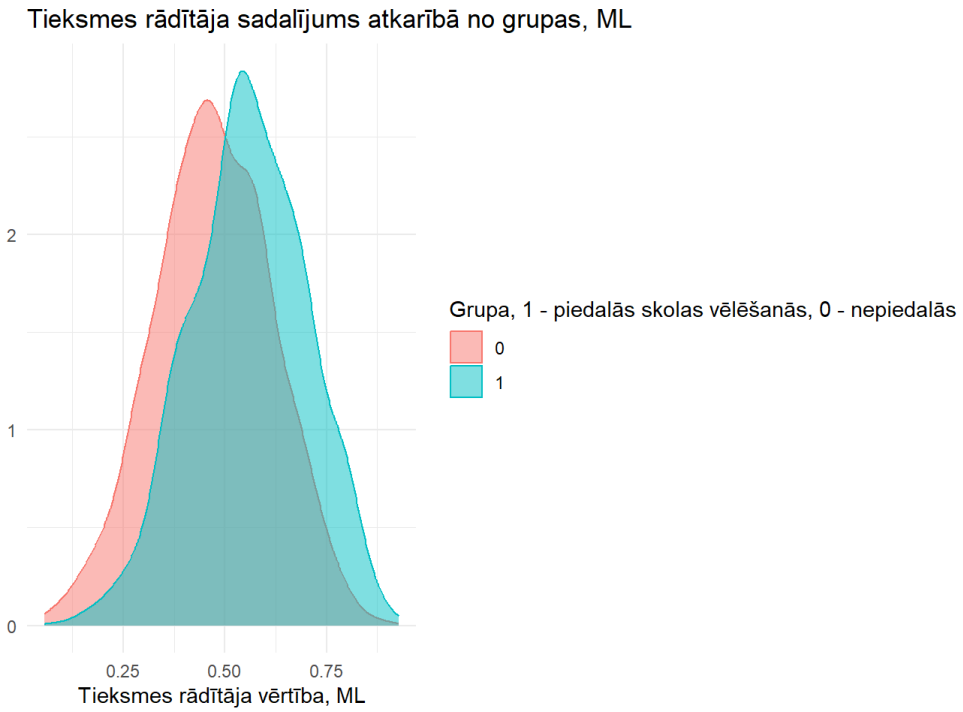
2. tabula

Loģistiskās regresijas parametru novērtējums ar maksimālās ticamības metodi

Šeit koeficientu novērtējumi parāda, kā mainās skolēnu dalības skolas vēlēšanās log-izredzes, ja prediktors mainās par vienu vienību. Piemēram, ja skolēns ir meitene, tad tas palielina log-izredzes piedalīties skolas vēlēšanās par 0.1583, ja salīdzina ar zēniem. Vai arī, ja gan skolēns, gan vecāki ir dzimuši ārzemēs (S_IMMIG3), tad tas samazina log-izredzes piedalīties skolas vēlēšanās par 0.0032, salīdzinot ar gadījumiem, kad skolēns pats vai vismaz viens no skolēna vecākiem ir dzimis Latvijā. Tiesa, aplūkojot p-vērtības, var redzēt, ka imigrācijas statusa rādītāji nav statistiski nozīmīgi. No prediktoriem, kuru datu tips ir nepārtraukti, piemēram, diezgan loģisks ir rezultāts, ka, palielinoties rādītājam S_SCACT par 1 vienību (skolēnu vēlme iesaistīties skolas aktivitātēs), log-izredzes piedalīties skolas vēlēšanās palielinās par 0.0216. No izdrukas var secināt, ka, pretēji sākotnējam uzskatam, imigrācijas statuss (S_IMMIG) un sociālekonomiskais stāvoklis (S_NISB) nav statistiski nozīmīgi prediktori. Arī pie dzimuma (SD_GENDER) p-vērtība ir virs nozīmības līmeņa (0.085 > 0.05), tomēr tā kā šis rezultāts ir ļoti tuvu 0.05 (turklāt pie nozīmības līmeņa 0.1 šis prediktors vēl skaitītos statistiski nozīmīgs), dzimums ir atstājams modelī. Turpretim imigrācijas statusu un sociālekonomisko stāvokli, balstoties uz rezultātiem, var neiekļaut. Regresija tika veikta atkārtoti, neiekļaujot S_IMMIG un S_NISB, un tad visi prediktori bija statistiski nozīmīgi (vai ļoti tuvu tam), skatīt izdruku pieli-

kumā (6. tabula). Šis modelis arī izmantots tālākajos analīzes soļos.

Tālāk tika novērtētas tieksmes rādītāja vērtības un regresijas koeficientu (6. tabula) ticamības intervāli, kā arī vizualizēti tieksmes rādītāja sadalījumi atkarībā no tā, vai ir bijusi dalība skolas vēlēšanās (testa grupa), vai nē (kontroles grupa).



1. att. **Tieksmes rādītāja sadalījums atkarībā no grupas, ar maksimālās ticamības metodi**

1. attēlā var redzēt, ka kontroles un testa grupās tieksmes rādītāja sadalījums atšķiras, un, lai tās līdzsvarotu nenovirzīta kauzālā efekta novērtēšanai, jāpielieto tieksmes rādītāja saskaņošanas algoritms.

Saskaņošanu veica, izmantojot k-tuvāko kaimiņu algoritmu, kad elementi ir sakārtoti pēc nejaušības principa un netiek atkārtoti izmantoti, nosakot 0.1 kā maksimālo distanci, kāda drīkst būt starp rādītāja vērtībām, lai elementi tiktu apvienoti pārī. Arī šajā solī tika ņemti vērā izlases svari. Zemāk redzamie rezultāti (skatīt 3. tabulu) apkopoti no izdrukas.

	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max
Pirms PSM	0.6338	0.9867	0.1713	0.2676
Pēc PSM	0.0212	1.0036	0.0054	0.0205

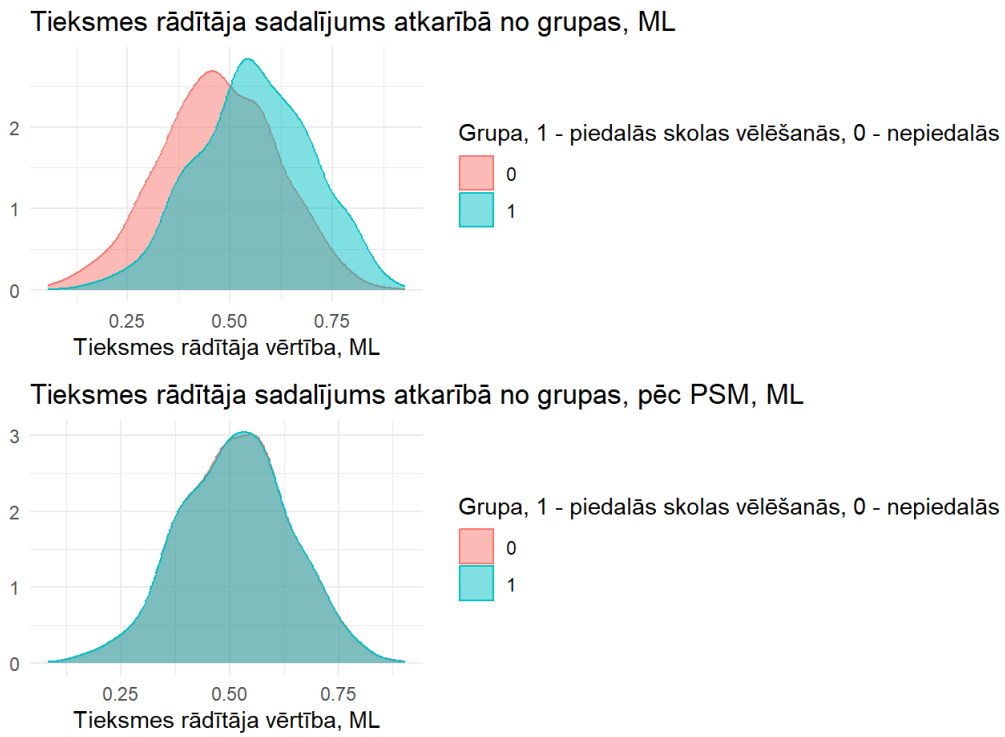
3. tabula

Tieksmes rādītāja saskaņošanas rezultāti, izmantojot k-tuvāko kaimiņu algoritmu

Var redzēt, ka standartizētā vidējā starpība pēc grupu līdzsvarošanas ir tuvu 0. Variācijas

attiecība gandrīz 1, bet empīrisko sadalījuma funkciju atšķirības arī ir tuvu 0. Tas viss norāda uz to, ka testa un kontroles grupas ir veiksmīgi līdzsvarotas. No sākotnēji 1206 vienībām kontroles un 1298 - testa grupā, pēc saskaņošanas algoritma izlasē palika 934 vienības kontroles un 934 vienības testa grupā, kas gan norāda uz vidēji 25% datu zudumu šajā procesā.

To, ka tieksmes rādītāja saskaņošanas algoritms ir bijis veiksmīgs, var secināt arī, aplūkojot zemāk redzamo grafiku (skatīt 2. attēlu), kur kontroles un testa grupu tieksmes rādītāja sadalījumi pārklājas, salīdzinot ar to, kā bija iepriekš.



2. att. **Tieksmes rādītāja sadalījums atkarībā no grupas, ar maksimālās ticamības metodi, pēc tieksmes rādītāja saskaņošanas algoritma**

Turklāt darba autore secināja, ka izlases elementu sakārtošana pēc nejaušības principa patiešām dod labāko rezultātu, jo, piemēram, sakārtojot tos no mazākā uz lielāko vai otrādi, tieksmes rādītāja sadalījumi kontroles un testa grupās pēc saskaņošanas algoritma vēl arvien ir pietiekami atšķirīgi. Kā piemērs pielikumā ievietots 4. attēls, kur elementi tika sakārtoti, sākot no mazākā. Līdzīgu rezultātu deva arī optimālā pāru saskaņošana.

Visbeidzot tika novērtēts dalības skolas vēlēšanās efekts saskaņotajām grupām. Tas tika darīts, izrēķinot vidējo vērtību starpību kontroles un testa grupām (ņemot vērā izlases svarus).

Dalības skolas vēlēšanās efekts uz skolēnu potenciālo dalību nacionālajās vēlēšanās nākotnē, rēķinot to kā vidējo vērtību starpību saskaņotajām kontroles un testa grupām, ņemot vērā arī izlases svarus, ir aptuveni 1.1203, kas nav daudz, ja ņem vērā, ka S_ELECPART vērtības

svārstās robežās $\approx 20 - 60$. Tātad var secināt, ka dalības skolas vēlēšanās efekts uz skolēnu potenciālo dalību nacionālajās vēlēšanās nākotnē ir minimāls.

Tālāk iepriekš aprakstītais process tika atkārtots empīriskās ticamības metodei. Tika veikta loģistiskā regresija, koeficientus novērtējot ar empīrisko maksimālās ticamības metodi, un iegūta zemāk redzamā izdruka (4. attēls).

	Estimate	χ^2	Pr($>\chi^2$)
(Intercept)	-4.779321	1636.434	$< 2e{-}16$
SD_GENDER1	0.158338	3.426	0.06419
S_IMMIG2	0.351624	2.015	0.15573
S_IMMIG3	-0.003250	0.000	0.99435
S_NISB	0.013494	0.103	0.74826
Pils_zin	0.003565	1135.148	$< 2e{-}16$
S_SCACT	0.021599	550.002	$< 2e{-}16$
S_COMPART	0.028803	916.025	$< 2e{-}16$
S_CIVLRN	0.012579	9.632	0.00191
S_DEMTHRT	-0.014809	9.903	0.00165
S_INFDEC	0.024661	725.657	$< 2e{-}16$
S_INTRUST	-0.010676	7.111	0.00766

4. tabula

Loģistiskās regresijas parametru novērtējums ar empīriskās ticamības metodi

Novērtētie parametri sakrīt ar novērtējumiem, kas iegūti ar maksimālās ticamības metodi (skatīt 2. tabulu), atšķiras vien noapaļošana. Tas šķiet loģiski, jo R funkcijas *svyglm()* un *el_glm()* izmanto tās pašas novērtēšanas funkcijas. Tomēr p-vērtības maksimālās ticamības un empīriskās maksimālās ticamības metodēm nedaudz atšķiras - p-vērtības, kas iegūtas empīriskās ticamības metodes ietvaros, izmantojot χ^2 aproksimāciju, ir mazākas par tām, kas iegūtas, veicot *Wald* testu (testa statistiku iegūst kā $\frac{\hat{\beta}-\beta_0}{SE(\hat{\beta})}$, kur β_0 pie H0 parasti ir 0) maksimālās ticamības metodes ietvaros. Iespējams, šāda atšķirība ir tādēļ, ka empīriskās ticamības metodes gadījumā hipotēžu pārbaudei nav nepieciešams novērtēt standartkļūdas, kas var būt problemātiski mazām izlasēm vai sarežģītiem modeļiem. Tādējādi var būt situācijas, kad empīriskā ticamība būtu piemērotāka metode. SD_GENDER p-vērtība ir aptuveni $0.064 < 0.085$ (p-vērtība maksimālās ticamības metodes gadījumā). Lai arī tas vēl arvien ir virs 0.05, rodas jau lielāka pārliecība par mainīgā atstāšanu modelī. Tiesa, S_IMMIG un S_NISB p-vērtības līdzīgi maksimālās ticamības

metodes rezultātam ir krietni virs 0.05. Tas liek domāt, ka šie prediktori nav statistiski nozīmīgi, tādēļ regresija tika veikta atkārtoti, neiekļaujot S_IMMIG un S_NISB, iegūstot modeli, kur visi prediktori ir statistiski nozīmīgi (vai SD_GENDER gadījumā - ļoti tuvu tam). Šis modelis izmantots arī tālākajos analīzes soļos, izdruku var skatīt pielikumā (7. tabula).

Salīdzinot ticamības intervālus abām metodēm (skatīt 5. tabulu), var secināt, ka ar empīriskās ticamības metodi tos iegūst šaurākus. Tabulā CI_platums_ML apzīmē ticamības intervālu platumus, kas iegūti ar maksimālās ticamības metodi, bet CI_platums_EL - ar empīriskās (maksimālās) ticamības metodi, Starpība = CI_platums_ML – CI_platums_EL, un tā parāda, par cik platāki ir ticamības intervāli, kas iegūti ar maksimālās ticamības metodi. Uzlabojums (%) tika aprēķināts (Starpība/CI_platums_ML)*100% . Tā kā ticamības intervālu platumi ir ļoti nelieli, tad arī starpība sākotnēji izskatās niecīga, taču Uzlabojums (%) parāda, ka relatīvais uzlabojums ir vērā ņemams.

Mainīgais	CI_platums_ML	CI_platums_EL	Starpība	Uzlabojums (%)
(Intercept)	1.944563619	1.445541472	0.4990221473	25.7
SD_GENDER1	0.360184014	0.335378101	0.0248059132	6.9
Pils_zin	0.002599242	0.002195525	0.0004037178	15.5
S_SCACT	0.015411361	0.013257967	0.0021533942	14.0
S_COMPART	0.019300730	0.014808093	0.0044926370	23.3
S_CIVLRN	0.019324004	0.015770341	0.0035536629	18.4
S_DEMTHRT	0.021344824	0.018848170	0.0024966533	11.7
S_INFDEC	0.023744415	0.018093581	0.0056508348	23.9
S_INTRUST	0.019330961	0.015404745	0.0039262158	20.3

5. tabula

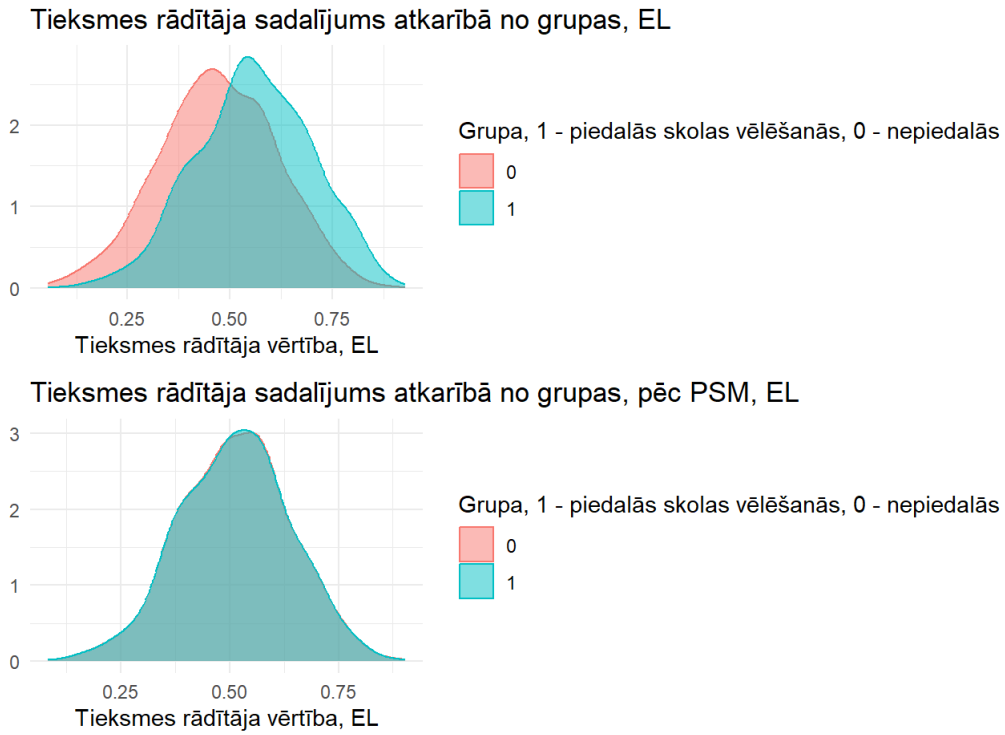
Ticamības intervālu garumu salīdzinājums abām metodēm

Tā kā kovariāti paliek tie paši, kas maksimālās ticamības metodes gadījumā, un loģistiskās regresijas koeficientu novērtējumi, neņemot vērā noapaļošanu, arī sakrīt (skatīt 2. un 4. tabulas), tad loģiski, ka arī tieksmes rādītāja vērtības sakrīt.

Tāpat kā maksimālās ticamības metodes gadījumā, kontroles un testa grupu līdzsvarošanai, veicot pāru saskaņošanu pēc tieksmes rādītāja vērtības, tika izmantots k-tuvāko kaimiņu algoritms, kad elementi ir sakārtoti pēc nejaušības principa un netiek atkārtoti izmantoti, nosakot 0.1 kā maksimālo distanci, kāda drīkst būt starp rādītāja vērtībām. Tā kā tieksmes rādītāja vērtības maksimālās un empīriskās ticamības metodēm sakrīt un tieksmes rādītāja starpība (skatīt 1.28)

elementiem kontroles un testa grupās ir izmantota kā distances metrika, tad loģiski šķiet arī tas, ka k-tuvāko kaimiņu algoritma rezultāts sanāca identisks abām metodēm (arī algoritma rezultātu izdrukā visas vērtības bija identiskas, tādēļ tā netiek atkārtoti iekļauta).

Arī vizuāli aplūkojot tieksmes rādītāja sadalījumus (skatīt 3. attēlu), var redzēt, ka testa un kontroles grupu līdzsvarošana ir bijusi veiksmīga, taču rezultāts ir tāds pats kā maksimālās ticamības metodei (skatīt 2. attēlu).



3. att. **Tieksmes rādītāja sadalījums atkarībā no grupas, ar empīriskās ticamības metodi, pēc tieksmes rādītāja saskaņošanas algoritma**

Pēcāk tika novērtēts metodes efekts saskaņotajām grupām. Arī empīriskās ticamības metodei tas tika darīts, izrēķinot vidējo vērtību starpību saskaņotajām kontroles un testa grupām, ņemot vērā izlases svarus.

Rēķinot testa un kontroles grupu vidējo vērtību starpību, tika iegūts tāds pats rezultāts kā maksimālās ticamības metodes gadījumā - dalības skolas vēlēšanās efekts uz skolēnu potenciālo dalību nacionālajās vēlēšanās nākotnē ir ≈ 1.1203 .

Lai spriestu par ATE novērtējuma statistisko nozīmību, tika pielietota *bootstrap* metode, izmantojot R iebūvēto funkciju *boot()*. Vidējais ATE no 1000 *bootstrap* ATE novērtējumiem ir ≈ 1.1271, kas no oriģinālā, iepriekš pilnai izlasei veiktā ATE novērtējuma atšķiras par ≈ 0.0068. Standartkļūdas novērtējums ir ≈ 0.4662. Izrēķinot attiecību $\frac{\widehat{ATE}}{SE}$ un aprēķinot p-vērtību, iegūst ≈ 0.0162. Tā kā H0 : ATE = 0, tad ar p-vērtību $0.0162 < 0.05$ pie nozīmības līmeņa $\alpha = 0.05$

var noraidīt H0 (turklāt 0 neietilpa arī 95% ticamības intervālā) un apgalvot, ka iegūtais efekta novērtējums ir statistiski nozīmīgs jeb, precīzāk, statistiski nozīmīgi atšķirīgs no 0.

Lai redzētu ATE novērtējuma atšķirību pirms un pēc tieksmes rādītāja saskaņošanas algoritma, ATE tika novērtēts arī sākotnējai datu kopai. Tika iegūts novērtējums ≈ 3.4302, kas ir aptuveni trīsreiz lielāks nekā novērtējums ”saskaņotajai” datu kopai. Taču šis rezultāts neparāda ”tīro efektu”, jo atšķirības testa un kontroles grupās palielina novērtējamā efekta vērtību.

SECINĀJUMI

Bakalaura darba ietvaros tika pētīta empīriskās ticamības metodes iekļaušana tieksmes rādītāja saskaņošanas algoritmā, loģistiskās regresijas, kuru parasti izmanto tieksmes rādītāja novērtēšanai, koeficientus novērtējot ar empīriskās ticamības metodi.

Darba autore iepazinās ar pieejamo literatūru par kauzālo efektu novērtēšanu, tieksmes rādītāja saskaņošanas algoritmu un empīriskās ticamības metodes izmantošanu vispārināto lineāro modeļu gadījumā. Darba autore veica arī simulācijas un pielietoja tieksmes radītāja saskaņošanas algoritmu, loģistiskās regresijas koeficientus novērtējot ar maksimālās un empīriskās ticamības metodēm, praktiskam piemēram ar ICCS pētījuma datiem. Tika meklēta atbilde uz jautājumu - cik daudz skolēnu dalība skolas vēlēšanās ietekmē skolēnu potenciālo dalību nacionālajās vēlēšanās?

Vislabāk kādas metodes (ATE) efektu iespējams novērtēt, veicot randomizēto eksperimentu, taču gadījumos, kad tas kādu iemeslu dēļ nav iespējams, kauzālo efektu jānovērtē novērošanas pētījuma datiem. Lai metodes efekta novērtējums būtu nenovirzīts, nepieciešams samazināt atšķirības kontroles un testa grupās, ko iespējams veikt ar tieksmes rādītāja saskaņošanas algoritmu. Vislabāko rezultātu dod pāru saskaņošana, kad elementi ir sakārtoti pēc nejaušības principa un netiek atkārtoti izmantoti, nosakot maksimālo distanci, kāda drīkst būt starp rādītāju vērtībām. Algoritma rezultātā iegūst sabalansētu izlasi, kas ir pietuvināta datiem, ko iegūtu randomizētā eksperimentā. Līdz ar to ATE var novērtēt kā vidējo vērtību starpību kontroles un testa grupās.

Empīriskās ticamības metodes pamatideja, novērtējot modeļa parametrus, ir maksimizēt empīrisko ticamību (vai ticamības attiecību) pie nosacījumiem, kas iekļauj novērtējošās funkcijas. Lai gan sākotnēji tā tika formulēta lineāriem modeļiem, empīriskās ticamības metodi parametru novērtēšanai iespējams izmantot arī vispārināto lineāro modeļu, tai skaitā loģistiskās regresijas, gadījumā, izvēloties atbilstošu novērtējošo funkciju.

Simulāciju ietvaros tika ģenerētas dažāda izmēra izlases, simulāciju katram apjomam atkārtojot 1000 reizes. Viens no izaicinājumiem bija saprast, kā ģenerēt datus, kuriem varētu būt novērojama reāla "metodes" ietekme, lai novērtētais ATE nebūtu vien nejaušība. Maksimālās un empīriskās ticamības metodēm tika salīdzināti novērtēto parametru ticamības intervālu platumi, tas, cik bieži $\beta_3 = 0$ gadījumā 0 patiešām ietilpst ticamības intervālā, kā arī ATE novērtējumu precizitāte. Lai arī pie parametriem β_1 un β_2 ar empīrisko ticamības metodi iegūtie ticamības intervāli neatkarīgi no izlases apjoma ir šaurāki, kopumā nevar teikt, ka empīriskā ticamības

metode dod kādu īpašu uzlabojumu, jo ar parametru β_3 saistītie rezultāti ir labāki maksimālās ticamības metodei, turklāt ATE novērtējumi būtiski neatšķiras (ne vairāk par 0.01).

ICCS datu praktiskā piemēra gadījumā ar empīrisko ticamības metodi iegūtie ticamības intervāli bija šaurāki, kā arī p-vērtības loģistiskās regresijas koeficientiem bija mazākas. Lai arī dotajā piemērā tas neko īpaši nemainīja, šis novērojums liek secināt, ka rezultāti, kas saistīti ar koeficientu statistisko nozīmību, var atšķirties atkarībā no izvēlētās pieejas - maksimālās vai empīriskās ticamības metodes, un var būt situācijas, kad otrā metode ir piemērotāka. Šis būtu jāņem vērā, novērtējot modeli.

Praktiskā piemēra ietvaros tika rasta atbilde uz iepriekš formulēto jautājumu - skolēnu dalībai skolas vēlēšanās ir neliela ietekme uz skolēnu potenciālo dalību nacionālajās vēlēšanās, ATE novērtējums ir ≈ 1.1203. Metodes sniegtais uzlabojums nav liels, taču tas ir statistiski nozīmīgi atšķirīgs no 0, tātad efekts ir.

Lai arī darbā tas netika darīts praktiski, tika aprakstīta teorija, kas saistīta ar metodes efekta novērtēšanu, izmantojot inversās varbūtības svērto novērtētāju IPW (un arī tā papildināto versiju AIPW), no kuras, savukārt, izriet teorija par empīriskās ticamības ATE novērtētāju. Šajā gadījumā vispirms empīriskās ticamības metodi izmanto tieksmes rādītāja novērtēšanai un pēc tam - arī vidējā metodes efekta novērtēšanai. Tālākajā pētniecībā varētu metodes efektu novērtēt ar IPW (vai AIPW) un ar empīriskās ticamības ATE novērtētāju, salīdzinot rezultātus.

IZMANTOTĀ LITERATŪRA UN AVOTI

[1] Y. Afana, F. Brese, H. Kowolik, D. Cortes, and W. Schulz. ICCS 2022 User guide for the international database, 2024.

[2] M.F. Alves. Causal inference for the brave and true, 2022.

[3] P.C. Austin. A comparison of 12 algorithms for matching on the propensity score. *Statistics in medicine*, 33(6), 2014.

[4] Y.C. Chen. Lecture 5: Empirical likelihood method, 2021. Lekciju pieraksti no University of Washington.

[5] K. Eunseop, S.N. MacEachern, and M. Peruggia. melt: Multiple empirical likelihood tests in R. *Journal of Statistical Software*, 108(5):1–33, 2024.

[6] J. Fraillon, T. Friedman, and W. Schulz. ICCS 2022 technical report, 2024.

[7] D.A. Freedman. *Statistical models: theory and practice*. Cambridge University Press, 2009.

[8] F. Fusi and J. Lecy. Foundations of program evaluation: Regression tools for impact analysis. `https://ds4ps.org/pe4ps-textbook/docs/index.html`, 2020. Skatīts: 15.04.2025.

[9] N. Greifer. Matching methods. `https://cran.r-project.org/web/packages/MatchIt/vignettes/matching-methods.html#optimal-full-matching-method-full`, 2025. Skatīts: 15.04.2025.

[10] A.S. Hadi and S. Chatterjee. *Regression Analysis by Example, 4th Edition*. John Wiley Sons, 2006.

[11] E.D. Kolaczyk. Empirical likelihood for generalized linear models. *Statistica Sinica*, pages 199–218, 1994.

[12] P. McCullagh and J.A. Nelder. *Generalized linear models, Second edition*. Chapman and Hall, 1989.

[13] A.A. Minai and R.D. Williams. On the derivatives of the sigmoid. *Neural Networks*, 6(6):845–853, 1993.

[14] World Health Organization. Clinical trials. https://www.who.int/health-topics/clinical-trials#tab=tab_1. Skatīts: 01.04.2025.

[15] A. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 06 1988.

[16] A. Owen. Empirical likelihood for linear models. *The Annals of Statistics*, pages 1725–1747, 1991.

[17] W. Pan and H. Bai. *Propensity Score Analysis: Fundamentals and Developments*. Guilford Publications, 2015.

[18] F. Quin, D. Weyns, M. Galster, and C.S. Costa. A/B testing: A systematic literature review. *Journal of Systems and Software*, 211, 2024.

[19] L.J. Reed and J. Berkson. The application of the logistic function to experimental data. *The Journal of Physical Chemistry*, 33(5):760–779, 2002.

[20] J.M. Robins, A. Rotnitzky, and L.P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[21] J.M. Robins, A. Rotnitzky, and L.P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121, 1995.

[22] P.R. Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408), 1989.

[23] P.R. Rosenbaum and D.B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[24] Izglītības un zinātnes ministrija. IEA ICCS. https://www.izm.gov.lv/lv/iea-iccs. Skatīts: 16.05.2025.

[25] B. Zhang. Empirical likelihood in causal inference. *Econometric Reviews*, 35(2):201–231, 2016.

PIELIKUMI

A. PROGRAMMU KODI

A.1. Simulācijas piemēra kods

```
### Pakotņu ielāde
library(melt)
library(dplyr)
library(tidyr)
library(xtable)
library(ggplot2)
library(MatchIt)

### Simulācijas funkcija
sim <- function(n=100,p=3,df=3,beta_ists=c(0.5, -0.75, 0), tau=2, beta_0_ists=0){
  #simulē datus
  X <- matrix(rt(n*p, df), nrow = n, ncol = p) #uzģenerē kovariātu matricu
  logit_p <- beta_0_ists + (X %*% beta_ists)
  p_ists <- 1 / (1 + exp(-logit_p)) #"īstās" tieksmes rādītāja vērtības
  T <- rbinom(n, 1, p_ists)#metodes izmantošanas dati (0=neizmanto, 1=izmanto)
  eps_t3 <- rt(n, df)
  Y <- logit_p + tau * T + eps_t3 #simulētais outcome
  sim_dati <- data.frame(T, X, Y)
  #print(sim_dati)

  #novērtē beta vērtības, izmantojot parasto ML un arī EL pieeju, atbilstošos CI un tieksmes rādītāja
  vērtības ar šiem beta
  #izmantojot ML
  beta_nov_ML_mod <- glm(T ~ X1+X2+X3, data = sim_dati, family = binomial())
  beta_nov_ML <- coef(beta_nov_ML_mod)[-1]
  beta_0_ML <- coef(beta_nov_ML_mod)["(Intercept)"]
  #print(beta_0_ML)
  CI_ML <- confint(beta_nov_ML_mod)[-1, ]
  logit_p_ML <- beta_0_ML + (X %*% beta_nov_ML)
  p_ML <- 1 / (1 + exp(-logit_p_ML))
  #izmantojot EL
  beta_nov_EL_mod <- el_glm(T ~ X1+X2+X3, data = sim_dati, family = "binomial")
  beta_nov_EL <- coef(beta_nov_EL_mod)[-1]
  beta_0_EL <- coef(beta_nov_EL_mod)["(Intercept)"]
  #print(beta_0_EL)
  CI_EL <- confint(beta_nov_EL_mod, level = 0.95)[-1, , drop = FALSE]
  logit_p_EL <- beta_0_EL + (X %*% beta_nov_EL)
  p_EL <- 1 / (1 + exp(-logit_p_EL))

  #tieksmes rādītāja saskaņošana, izmantojot k-tuvāko kaimiņu algoritmu
  #ML
  match_obj1 <- matchit(T ~ 1, data = sim_dati, method = "nearest", distance = as.vector(p_ML),
  m.order = "random", ratio = 1, replace = FALSE, caliper = 0.15)
  #print(summary(match_obj1))
```

```
  #EL
  match_obj2 <- matchit(T ~ 1, data = sim_dati, method = "nearest", distance = as.vector(p_EL),
  m.order = "random", ratio = 1, replace = FALSE, caliper = 0.15)
  #print(summary(match_obj2))

  #iegūst matched jeb pēc tieksmes rādītāja saskaņotos datus
  matched_ML <- match.data(match_obj1)
  matched_EL <- match.data(match_obj2)

  #metodes efekta novērtējumi
  ML_efekts <- with(matched_ML, mean(Y[T == 1]) - mean(Y[T == 0]))
  EL_efekts <- with(matched_EL, mean(Y[T == 1]) - mean(Y[T == 0]))

  #saglabā beta novērtējumus un CI
  rez <- data.frame(
    parametrs = paste0("Beta_", 1:p),
    beta_ML = beta_nov_ML,
    beta_EL = beta_nov_EL,
    apaks_CI_ML = CI_ML[, 1],
    augs_CI_ML = CI_ML[, 2],
    apaks_CI_EL = CI_EL[, 1],
    augs_CI_EL = CI_EL[, 2],
    CI_platums_ML = abs(CI_ML[, 2]-CI_ML[, 1]),
    CI_platums_EL = abs(CI_EL[, 2]-CI_EL[, 1])
  )
  #saglabā tieksmes rādītāja vērtības
  tieksmes_rez <- data.frame(p_ML = p_ML,
                             p_EL = p_EL,
                             T = T,
                             Y = Y
  )

  metodes_efekti <- data.frame(ML_efekts,
                               EL_efekts

  )
  return(list(koef_tic=rez, tieksmes_p=tieksmes_rez, efekti=metodes_efekti))
}

### Funkciju atkārto 1000 reizes
set.seed(123)
sim_rez <- replicate(1000, sim(), simplify = FALSE)

### Aplūko vidējo ticamības intervāla platumu atkarībā no metodes katram beta
p <- 3 #parametru skaits
CI_ML_rez <- sapply(sim_rez, function(x) x$koef_tic$CI_platums_ML)
CI_EL_rez <- sapply(sim_rez, function(x) x$koef_tic$CI_platums_EL)

Vid_CI_plat <- data.frame(
  parametrs = paste0("Beta_", 1:p),
  Vid_CI_platums_ML = rowMeans(CI_ML_rez),
```

```
  Vid_CI_platums_EL = rowMeans(CI_EL_rez)
)

print(Vid_CI_plat)

### Aplūko, cik gadījumos (procentuāli) 0 ietilpst beta_3 novērtējuma ticamības intervālā atkarībā no metodes
#ML
apaks_rob_ML <- sapply(sim_rez, function(x) x$koef_tic$apaks_CI_ML[3] <= 0)
augs_rob_ML <- sapply(sim_rez, function(x) x$koef_tic$augs_CI_ML[3] >= 0)
rindas1_check_ML <- sum(mapply(function(x, y) x&y, apaks_rob_ML, augs_rob_ML))
ietilpst_ML <- rindas1_check_ML/length(apaks_rob_ML)
print(paste0(ietilpst_ML * 100, "%"))
#EL
apaks_rob_EL <- sapply(sim_rez, function(x) x$koef_tic$apaks_CI_EL[3] <= 0)
augs_rob_EL <- sapply(sim_rez, function(x) x$koef_tic$augs_CI_EL[3] >= 0)
rindas1_check_EL <- sum(mapply(function(x, y) x&y, apaks_rob_EL, augs_rob_EL))
ietilpst_EL <- rindas1_check_EL/length(apaks_rob_EL)
print(paste0(ietilpst_EL * 100, "%"))

### Aplūko vidējo (no 1000 simulācijām) ATE atkarībā no metodes
ML_rez <- sapply(sim_rez, function(x) x$efekti$ML_efekts)
EL_rez <- sapply(sim_rez, function(x) x$efekti$EL_efekts)

Vid_efekti <- data.frame(
  Vid_efekts_ML = mean(ML_rez),
  Vid_efekts_EL = mean(EL_rez)
)

print(Vid_efekti)
```

A.2. Praktiskās daļas kods - ICCS datu analīze

```
### Datu un pakotņu ielāde
library(MatchIt)
library(haven)
library(dplyr)
library(survey)
library(tidyr)
library(ggplot2)
library(patchwork)
library(melt)

load("ISGLVAC4.RData")
#atlasa vajadzīgos mainīgos
dati_lv <- ISGLVAC4 %>%
  select(
    IDSTUD, #skolēna ID
    S_ELECPART, #Y - ietekmētā iznākuma mainīgais, potenciālā dalība nacionālajās vēlēšanās
    IS4G15B, #T - treatment mainīgais, dalība skolas vēlēšanās
    SD_GENDER, #dzimums
    S_IMMIG, #imigrācijas statuss
```

```
    S_NISB, #sociālekonomiskais stāvoklis
    PV1CIV, PV2CIV, PV3CIV, PV4CIV, PV5CIV, # pilsonisko zināšanu testa iespējamie rezultāti (plausible values)
    S_SCACT, #vēlme piedalīties skolas aktivitātēs
    S_COMPART, #dalība organizācijās
    S_CIVLRN, #pilsonisko zināšanu apguve skolā
    S_DEMTHRT, #izpratne par draudiem demokrātijai
    S_INFDEC, #pārliecība par ietekmi uz lēmumu pieņemšanu skolā
    S_INTRUST, #uzticēšanās valsts institūcijām
    TOTWGTS #skolēnu svari datu kopā
  )

dati_lv_df <- as.data.frame(dati_lv)
#pārbauda kontroles/testa elementu īpatsvaru datu kopā izvēlētajam mainīgajam
# sum(dati_lv_df$IS4G15B == 1)
# sum(dati_lv_df$IS4G15B == 2)
# sum(dati_lv_df$IS4G15B == 3)

### No pilsonisko zināšanu testa iespējamajiem rezultātiem iegūst vidējo vērtību, lai vienkāršotu aprēķinus,
pārveido mainīgo IS4G15B no ordināla uz bināru, faktorizē kategoriālos datus
dati1 <- dati_lv_df %>%
  mutate(
  IDSTUD = as.factor(IDSTUD),
  IS4G15B = as.factor(IS4G15B),
  SD_GENDER = as.factor(SD_GENDER),
  S_IMMIG = as.factor(S_IMMIG),
  Pils_zin = rowMeans(subset(dati_lv_df, select = c(PV1CIV, PV2CIV, PV3CIV, PV4CIV, PV5CIV))),
  Velesanas = as.factor(case_when(
    IS4G15B %in% c(1, 2) ~ 1,
    IS4G15B == 3 ~ 0,
    TRUE ~ NA_real_
  ))
  )

#izņem rindas, kur ir NA vērtības
dati <- na.omit(dati1)

#noņem liekās kolonnas, kas tika izmantotas vien kolonnu Pils_zin un Velesanas aprēķiniem
dati_psm1 <- dati %>%
  mutate(
    PV1CIV = NULL, PV2CIV = NULL, PV3CIV = NULL, PV4CIV = NULL, PV5CIV = NULL,
    IS4G15B = NULL
  )

#lai tās netraucē analīzei, izņem rindas, kurām S_IMMIG vai SD_GENDER bija kategorija "9" jeb "omitted"
dati_psm <- dati_psm1 %>%
  filter(as.character(S_IMMIG) != "9", as.character(SD_GENDER) != "9")

#pārbaude
head(dati_psm)
sum(dati_psm$S_IMMIG == "9")
sum(dati_psm$SD_GENDER == "9")
```

```
any(is.na(dati_psm))

### Nosaka izlases dizaina svarus, ko turpmāk ņemt vērā
dizains <- svydesign(
  ids = ~1,
  weights = ~TOTWGTS,
  data = dati_psm
)
#print(dizains)
#summary(dizains)

### Veic loģistisko regresiju, novērtējot koeficientus ar ML metodi, ņemot vērā arī izlases svarus
ML_mod1 <- svyglm(Velesanas ~ SD_GENDER + S_IMMIG + S_NISB + Pils_zin + S_SCACT + S_COMPART + S_CIVLRN +
S_DEMTHRT + S_INFDEC + S_INTRUST, design=dizains, family=quasibinomial())
summary(ML_mod1)

#veic regresiju bez prediktoriem, kas nav statistiski nozīmīgi
ML_mod <- svyglm(Velesanas ~ SD_GENDER + Pils_zin + S_SCACT + S_COMPART + S_CIVLRN + S_DEMTHRT + S_INFDEC +
S_INTRUST, design=dizains, family=quasibinomial())
summary(ML_mod)

### Novērtē tieksmes rādītāja vērtības ar ML un regresijas koeficientu ticamības intervālus
#izmanto iebūvēto funkciju, lai iegūtu tieksmes rādītāja vērtības
p_ML1 <- predict(ML_mod, newdata = dizains$variables, type = "response")
p_ML <- as.data.frame(p_ML1)
tieksme_ML <- p_ML[[1]]
#loģistiskās regresijas novērtēto koeficientu ticamības intervāli
CI_ML <- confint(ML_mod, level = 0.95)
CI_ML
apaks_CI_ML = CI_ML[, 1]
augs_CI_ML = CI_ML[, 2]
CI_platums_ML = abs(CI_ML[, 2]-CI_ML[, 1])

### Vizualizē tieksmes rādītāja sadalījumu atkarībā no dalības skolas vēlēšanās
tieksmes_df <- data.frame(
  tieksmes_ML = tieksme_ML,
  Skolas_velesanas = dizains$variables$Velesanas
)
print(dim(tieksmes_df))

ggplot(tieksmes_df, aes(x = tieksmes_ML, fill = Skolas_velesanas, color = Skolas_velesanas)) +
  geom_density(alpha = 0.5) +
  labs(
    title = "Tieksmes rādītāja sadalījums atkarībā no grupas, ML",
    x = "Tieksmes rādītāja vērtība, ML",
    y = " ",
    fill = "Grupa, 1 - piedalās skolas vēlēšanās, 0 - nepiedalās",
    color = "Grupa, 1 - piedalās skolas vēlēšanās, 0 - nepiedalās"
  ) +
  theme_minimal()
```

```
### Veic pāru saskaņošanu pēc novērtētā tieksmes rādītāja, vizualizē tieksmes rādītāja sadalījumu atkarībā
no dalības skolas vēlēšanās pēc PSM
set.seed(2)
match_obj <- matchit(Velesanas ~ 1, data = dizains$variables, method = "nearest", m.order = "random",
distance = tieksme_ML, ratio = 1, replace = FALSE, caliper = 0.1, s.weights = weights(dizains))
print(summary(match_obj))
#iegūst matched (saskaņotās) datu kopas
matched_ML <- match.data(match_obj)
tieksmes_df_m <- data.frame(
  Tieksmes_ML_m = matched_ML$distance,
  Skolas_velesanas_m = matched_ML$Velesanas
)
print(dim(tieksmes_df_m))

g1 <- ggplot(tieksmes_df, aes(x = tieksmes_ML, fill = Skolas_velesanas, color = Skolas_velesanas)) +
  geom_density(alpha = 0.5) +
  labs(
    title = "Tieksmes rādītāja sadalījums atkarībā no grupas, ML",
    x = "Tieksmes rādītāja vērtība, ML",
    y = " ",
    fill = "Grupa, 1 - piedalās skolas vēlēšanās, 0 - nepiedalās",
    color = "Grupa, 1 - piedalās skolas vēlēšanās, 0 - nepiedalās"
  ) +
  theme_minimal()

g2 <- ggplot(tieksmes_df_m, aes(x = Tieksmes_ML_m, fill = Skolas_velesanas_m, color = Skolas_velesanas_m)) +
  geom_density(alpha = 0.5) +
  labs(
    title = "Tieksmes rādītāja sadalījums atkarībā no grupas, pēc PSM, ML",
    x = "Tieksmes rādītāja vērtība, ML",
    y = " ",
    fill = "Grupa, 1 - piedalās skolas vēlēšanās, 0 - nepiedalās",
    color = "Grupa, 1 - piedalās skolas vēlēšanās, 0 - nepiedalās"
  ) +
  theme_minimal()

g1/g2

### Novērtē metodes efektu saskaņotajām grupām
#svērtais vidējais
weights <- weights(dizains)
vid_efekts_1 <- with(matched_ML, weighted.mean(S_ELECPART[Velesanas == 1], weights[Velesanas == 1]))
vid_efekts_0 <- with(matched_ML, weighted.mean(S_ELECPART[Velesanas == 0], weights[Velesanas == 0]))
ML_efekts <- vid_efekts_1 - vid_efekts_0
print(ML_efekts)

### Veic loģistisko regresiju, izmantojot EL metodi un iebūvēto funkciju no melt R pakotnes
(arī ņem vērā izlases svarus)
EL_mod1 <- el_glm(Velesanas ~ SD_GENDER + S_IMMIG + S_NISB + Pils_zin + S_SCACT + S_COMPART + S_CIVLRN +
S_DEMTHRT + S_INFDEC + S_INTRUST, data = dati_psm, weights = TOTWGTS, family = "binomial")
summary(EL_mod1)
```

```
#veic regresiju atkārtoti, izmantojot tikai statistiski nozīmīgos prediktorus
EL_mod <- el_glm(Velesanas ~ SD_GENDER + Pils_zin + S_SCACT + S_COMPART + S_CIVLRN + S_DEMTHRT + S_INFDEC +
S_INTRUST, data = dati_psm, weights = TOTWGTS, family = "binomial")
summary(EL_mod)

### Novērtē tieksmes rādītāja vērtības ar EL un regresijas koeficientu ticamības intervālus
#iegūst tieksmes rādītāja vērtības
pred <- model.matrix(~ SD_GENDER + Pils_zin + S_SCACT + S_COMPART +
                  S_CIVLRN + S_DEMTHRT + S_INFDEC + S_INTRUST,
                  data = dati_psm)[, -c(1, 3)]

koef_vec <- coef(EL_mod)
beta_0_EL <- koef_vec[1]
beta_EL <- koef_vec[-1]
beta_EL <- as.matrix(beta_EL)
logit_p_EL <- beta_0_EL + pred %*% beta_EL
tieksme_EL <- 1 / (1 + exp(-logit_p_EL))

#loģistiskās regresijas novērtēto koeficientu ticamības intervāli
CI_EL <- confint(EL_mod, level = 0.95)
CI_EL
apaks_CI_EL = CI_EL[, 1]
augs_CI_EL = CI_EL[, 2]
CI_platums_EL = abs(CI_EL[, 2]-CI_EL[, 1])
starpiba = CI_platums_ML - CI_platums_EL
CI_salidzinajums <- data.frame(
  CI_platums_ML = CI_platums_ML,
  CI_platums_EL = CI_platums_EL,
  Starpiba = starpiba
)
print(CI_salidzinajums)

### Tieksmes rādītāja sadalījums ar EL metodi
tieksmes_df_EL <- data.frame(
  tieksmes_EL = tieksme_EL,
  Skolas_velesanas_EL = dati_psm$Velesanas
)
print(dim(tieksmes_df_EL))

ggplot(tieksmes_df_EL, aes(x = tieksmes_EL, fill = Skolas_velesanas_EL, color = Skolas_velesanas_EL)) +
  geom_density(alpha = 0.5) +
  labs(
    title = "Tieksmes rādītāja sadalījums atkarībā no grupas, EL",
    x = "Tieksmes rādītāja vērtība, EL",
    y = " ",
    fill = "Grupa, 1 - piedalās skolas vēlēšanās, 0 - nepiedalās",
    color = "Grupa, 1 - piedalās skolas vēlēšanās, 0 - nepiedalās"
  ) +
  theme_minimal()
```

```
### Veic pāru saskaņošanu pēc EL novērtētā tieksmes rādītāja, vizualizē tieksmes rādītāja sadalījumu
atkarībā no dalības skolas vēlēšanās pēc PSM
set.seed(2)
match_obj2 <- matchit(Velesanas ~ 1, data = dati_psm, method = "nearest", m.order = "random",
distance = as.vector(tieksme_EL), ratio = 1, replace = FALSE, caliper = 0.1,
s.weights = as.vector(dati_psm$TOTWGTS))
print(summary(match_obj2))
#iegūst matched (saskaņotās) datu kopas
matched_EL <- match.data(match_obj2)
tieksmes_df_EL_m <- data.frame(
  Tieksmes_EL_m = matched_EL$distance,
  Skolas_velesanas_EL_m = matched_EL$Velesanas
)
print(dim(tieksmes_df_EL_m))

g3 <- ggplot(tieksmes_df_EL, aes(x = tieksmes_EL, fill = Skolas_velesanas_EL, color = Skolas_velesanas_EL)) +
  geom_density(alpha = 0.5) +
  labs(
    title = "Tieksmes rādītāja sadalījums atkarībā no grupas, EL",
    x = "Tieksmes rādītāja vērtība, EL",
    y = " ",
    fill = "Grupa, 1 - piedalās skolas vēlēšanās, 0 - nepiedalās",
    color = "Grupa, 1 - piedalās skolas vēlēšanās, 0 - nepiedalās"
  ) +
  theme_minimal()

g4 <- ggplot(tieksmes_df_EL_m, aes(x = Tieksmes_EL_m, fill = Skolas_velesanas_EL_m, color = Skolas_velesanas_EL_m)) +
  geom_density(alpha = 0.5) +
  labs(
    title = "Tieksmes rādītāja sadalījums atkarībā no grupas, pēc PSM, EL",
    x = "Tieksmes rādītāja vērtība, EL",
    y = " ",
    fill = "Grupa, 1 - piedalās skolas vēlēšanās, 0 - nepiedalās",
    color = "Grupa, 1 - piedalās skolas vēlēšanās, 0 - nepiedalās"
  ) +
  theme_minimal()

g3/g4

### Novērtē metodes efektu grupām, kas saskaņotas pēc tieksmes rādītaja, kas novērtēts, izmantojot EL
#svērtais vidējais
weights <- matched_EL$TOTWGTS
Vid_efekts_1 <- with(matched_EL, weighted.mean(S_ELECPART[Velesanas == 1], weights[Velesanas == 1]))
Vid_efekts_0 <- with(matched_EL, weighted.mean(S_ELECPART[Velesanas == 0], weights[Velesanas == 0]))
EL_efekts <- Vid_efekts_1 - Vid_efekts_0
print(EL_efekts)

#bootstrap metode ATE un ATE SE novērtēšanai
ate_fun <- function(data, indices) {
  d <- data[indices, ]
  #weights = weights(dizains)
```

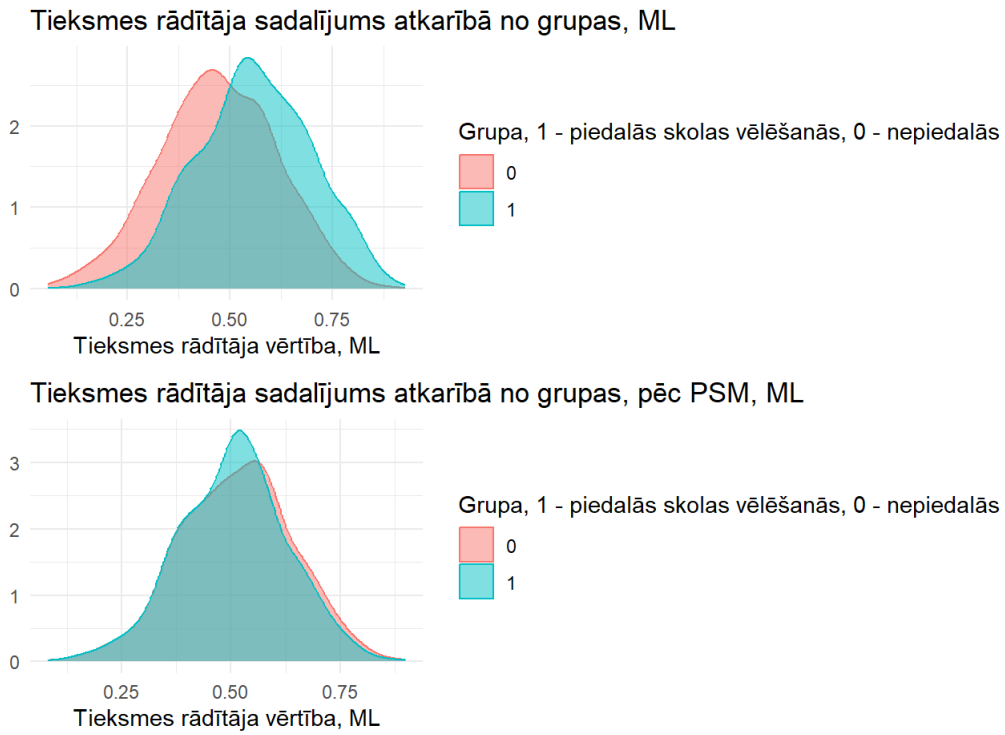
```
  treated_mean <- with(d[d$Velesanas == 1, ], weighted.mean(S_ELECPART, TOTWGTS))
  control_mean <- with(d[d$Velesanas == 0, ], weighted.mean(S_ELECPART, TOTWGTS))
  return(treated_mean - control_mean)
}

set.seed(2)
boot_rez <- boot(matched_EL, statistic = ate_fun, R = 1000)
print(boot_rez)
mean(boot_rez$t)
boot.ci(boot_rez)
#SE un AtE vērtības, iegūtas no izdrukas
SE_EL <- 0.466151
ATE_EL <- 1.120296
stat_EL <- ATE_EL/SE_EL
#stat nozīmība
p_vert_EL <- 2 * pnorm(-abs(stat_EL))
p_vert_EL
#novērtētais ATE datiem pirms matching
boot_rez2 <- boot(dati_psm, statistic = ate_fun, R = 1000)
print(boot_rez2)
mean(boot_rez2$t)
```

B. PAPILDU IZDRUKAS UN GRAFIKI

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.7522	0.4958	-9.58	< 2e-16
SD_GENDER1	0.1598	0.0918	1.74	0.0820
Pils_zin	0.0035	0.0007	5.34	1.03e-07
S_SCACT	0.0216	0.0039	5.49	4.34e-08
S_COMPART	0.0286	0.0049	5.82	6.77e-09
S_CIVLRN	0.0127	0.0049	2.57	0.0102
S_DEMTHRT	-0.0148	0.0054	-2.72	0.0067
S_INFDEC	0.0244	0.0061	4.03	5.65e-05
S_INTRUST	-0.0105	0.0049	-2.13	0.0333

6. tabula

Loģistiskās regresijas parametru novērtējums ar maksimālās ticamības metodi, atkārtoti



4. att. **Tieksmes rādītāja sadalījums atkarībā no grupas, ar maksimālās ticamības metodi, izlases elementi tieksmes rādītāja saskaņošanas algoritmā sakārtoti, sākot no mazākā**

	Estimate	χ^2	Pr($>\chi^2$)
(Intercept)	-4.752225	1635.552	$< 2e{-}16$
SD_GENDER1	0.159819	3.493	0.06163
Pils_zin	0.003537	1127.135	$< 2e{-}16$
S_SCACT	0.021587	549.798	$< 2e{-}16$
S_COMPART	0.028626	909.692	$< 2e{-}16$
S_CIVLRN	0.012674	9.766	0.00178
S_DEMTHRT	-0.014780	9.603	0.00194
S_INFDEC	0.024423	716.042	$< 2e{-}16$
S_INTRUST	-0.010499	6.884	0.00870

7. tabula

Loģistiskās regresijas parametru novērtējums ar empīriskās ticamības metodi, atkārtoti