Lietuvos matematikos rinkinys. LMD darbai
52 tomas, 2011, 321–326

ISSN 0132-2818
www.mii.lt/LMR/

Studentų internetinės apklausos duomenų analizė

Gediminas Murauskas[1], **Marijus Radavičius**[1,2]

[1] *Vilniaus Universitetas, Matematikos ir informatikos fakultetas*
Naugarduko 24, LT-03225 Vilnius
[2] *Vilniaus Universitetas, Matematikos ir informatikos institutas*
Akademijos 4, LT-08663 Vilnius
E. paštas: gediminas.murauskas@mif.vu.lt; marijus.radavicius@mii.vu.lt

Santrauka. Straipsnyje nagrinėjama internetinės apklausos rezultatų vertinimo ir neatsakymų sąlygoto poslinkio sumažinimo problema remiantis VU studentų apklausos duomenimis apie VU Informacinių technologijų taikymo centro (ITTC) teikiamų paslaugų kokybę. Greta tradicinių persvėrimo metodų neatsakymų sąlygotam poslinkiui kompensuoti pritaikytas sąlyginės logistinės regresijos modelis.

Raktiniai žodžiai: neatsakymų poslinkio sumažinimas, sąlyginė logistinė regresija, internetinė apklausa, persvėrimas.

Įvadas

Pastaruoju metu populiarėja apklausos, kai anketos pateikiamos interneto svetainėse. Atliekant internetines apklausas keblu įvertinti jų reprezentatyvumą ir juo labiau tikslumą. Internetinio anketavimo problematika yra gana išsamiai aprašyta (žr., pvz., [4, 1]). Situacija pagerėja, jeigu yra vykdoma *registruotų vartotojų* apklausa. Tuomet tiriamoji populiacija yra tiksliai nusakyta. Vis dėlto ir šiuo atveju apklausos rezultatai gali iškreiptai atspindėti populiacijos charakteristikas, kitaip tariant, gali turėti *poslinkį*. Jis gali atsirasti dėl to, kad (dėl internetinės apklausos „savanoriškumo") ne visi registruoti vartotojai dalyvauja apklausoje. Būtent su šia problema ir susiduriame nagrinėjamame studentų nuomonės tyrime, kuriame atsakymų dalis yra labai maža (3–14%).

Pastebėsime, kad registruotų vartotojų atveju internetines apklausas galima traktuoti kaip *visuotinę* (t.y., ištisinę visos populiacijos) apklausą, kurioje yra neatsakymų dėl vartotojų, „atsisakiusių" dalyvauti apklausoje.

Studentų internetinių apklausų atveju paprastai yra žinomi pagalbinių kintamųjų skirstiniai (studentų skaičiai fakultetuose, įvairiose studijų pakopose studijuojančių studentų skaičiai ir t.t.). Šios papildomos informacijos panaudojimo uždavinys yra kompensuoti arba kiek įmanoma sumažinti internetinės apklausos poslinkį. Taigi, šiuo atveju galima taikyti žinomus imčių teorijos metodus (žr., pvz., [6]).

Šiame darbe yra naudojami 2009 m. ir 2010 m. pabaigoje atliktos VU studentų apklausos (išskyrus TVM ir Užsienio kalbų instituto studentus) duomenys [3, 2]). *Darbo tikslas* – pateikti VU studentų ITTC teikiamų IT paslaugų kokybės vertinimų analizę, atsižvelgiant į internetinės apklausos specifiką, aprašyti ir tarpusavyje palyginti įvairius internetinės apklausos rezultatų analizės ir įvertinimo metodus bei

suformuluoti pasiūlymus, kuriuos įgyvendinus, autorių nuomone, studentų internetinis anketavimas būtų efektyvesnis ir duotų patikimesnius rezultatus.

Kitame skyrelyje pateikiamas trumpas atsisakymo dalyvauti apklausoje (trumpai, neatsakymų) modelių aprašymas bei šiame darbe taikytas sąlyginės logistinės regresijos modelis, suteikiantis alternatyvą neatsakymų įnešamam poslinkiui kompensuoti. Trečiame skyrelyje aprašomi naudoti duomenys ir aptariami gauti rezultatai. Pabaigoje pateiktos išvados.

1 Neatsakymų apklausose modeliai

Gana skirtingas neatsakymų procentas įvairiose vartotojų grupėse gali įnešti ženklų poslinkį į apklausos rezultatus. Yra žinomi įvairūs metodai tam poslinkiui sumažinti: persvėrimas įvertinta neatsakymų tikimybe, kalibravimo metodai, naudojantys regresijos modelius ir kiti. 1.1 skyrelyje trumpai aprašomas persvėrimo metodas, kuris kiekvieną atsakymą persveria įvertinta atsakymo į apklausą tikimybe. Pabrėšime, kad pastarasis metodas eliminuoja galimą poslinkį tik tuo atveju, kai į neatsakymo tikimybės statistinį modelį yra įtraukiami visi kintamieji, potencialiai susiję su neatsakymais, ir be to turi būti žinomos tų kintamųjų reikšmės visoje populiacijoje. Praktiškai nėra patikimų būdų ptikrinti, ar ši prielaida yra išpildyta.

Todėl įdomi alternatyva tradiciniams neatsakymų poslinkio kompensavimo metodams yra sąlyginė logistinė regresija, kurios trumpas aprašymas yra pateiktas 1.2 skyrelyje.

1.1 Klasikiniai neatsakymų modeliai

Tegu U, S, ir R žymi atitinkamai visą tiriamą populiaciją, jos imtį ir atsakiusiųjų, t.y., užpildžiusių apklausos anketą, aibę. Pažymėkime atsakymo tikimybę $p_s := \mathbf{P}\{s \in R \mid s \in S\}$. Paprastai šios tikimybės priklausomybei nuo pagalbinių kintamųjų $x = x_s$, $s \in S$, aprašyti taikomas logistinės regresijos modelis:

$$p_s := p(x_s) := \frac{\exp\{\beta^\top x_s\}}{1 + \exp\{\beta^\top x_s\}}.$$

Čia β yra nežinomų parametrų vektorius, kurį reikia įvertinti iš imties. Pagrindinė prielaida, kuria remiamasi taikant didžiausio tikėtinumo metodą ir tikrinant hipotezes formuluojama taip: duotai imčiai S, o tuo pačiu ir papildomo kintamojo x reikšmėms, kiekvienas respondentas atsako ar neatsako nepriklausomai muo likusių imties respondentų.

Gana paprastas, bet praktikoje svarbus šio modelio atskiras atvejis yra AHG (atsakymų homogeniškumo grupių) modelis [6]. Šiame modelyje visą imtį galima suskaidyti į nesikertančias respondentų aibes S_h, $h = 1, \ldots, H$, kuriose respondentų atsakymo/neatsakymo tikimybės yra vienodos, t.y., $p_s \equiv p(h)$, $s \in S_h$, $h = 1, \ldots, H$. Jeigu atsakymo tikimybės p_s yra žinomos, tai pataisyti atsižvelgiant į neatsakymus (tiesiniai) įvertiniai gaunami pradinius imties plano svorius d_s dalinant iš atsakymo tikimybių p_s. Nauji svoriai $w_s = d_s/p_s$, $s \in S$. Kadangi tikimyvės p_s, $s \in S$ nežinomos, jos keičiamos įvertiniais.

Šiame darbe neatsakymams modeliuoti taikomas ir AHG, ir lankstesnis logistinės regresijos modelis, kuris leidžia adaptyviai apjungti panašias ir/arba mažai stebinių

turinčias grupes S_h tokiu būdu sumažinant parametrų skaičių ir padidinant įvertinių stabilumą.

1.2 Sąlyginė logistinė regresija

Žemiau pateiktame sąlyginės logistinės regresijos modelio aprašyme laikoma, kad atsako kintamasis yra binarinis. Nors pradiniai ITTC paslaugų vertinimo apklausos duomenys yra ranginiai, pastaroji prielaida daroma todėl, kad tokiam modeliui yra galimybė skaičiavimuose pasinaudoti atitinkama programine įranga.

Tegu $y_{ij} \in \{0, 1\}$, i yra apklausos anketos klausimo (apie atitinkamą paslaugą) numeris, $i = 1, \ldots, m$, j yra respondento (šiuo atveju, studento) numeris, $j = 1, \ldots, n$. Įvykis $\{y_{ij} = 1\}$ reiškia, kad j-asis tiriamasis teigiamai įvertino i-ąją paslaugą. Tegu $x_{ij} \in \mathbf{R}^k$ žymi su apklausa susijusių *aiškinančiųjų kintamųjų* vektorių ($i = 1, \ldots, m$, $j = 1, \ldots, n$). Sąlyginį y_{ij} skirstinį, kai žinomos aiškinančiųjų kintamųjų reikšmės x_{ij} nusako logistinės regresijos modelis

$$\mathbf{P}\{y_{ij} = 1 \mid x_{ij}\} = \frac{\exp\{\theta_i - b_j + \beta^\top x_{ij}\}}{1 + \exp\{\theta_i - b_j + \beta^\top x_{ij}\}}. \tag{1}$$

Parametras θ_i yra traktuojamas kaip latentinis, kintamasis, kiekybiškai charakterizuojantis i-osios ITTC teikiamos paslaugos kokybę: kuo θ_i yra didesnis, tuo didesnė (kitoms sąlygoms nekintant) tikimybė, kad studentas i-ąją paslaugą įvertins teigiamai. Parametrai b_j, $j = 1, \ldots, n$, taip pat traktuoajmi kaip latentiniai kintamieji. Jie aprašo j-ojo studento „reiklumą", jo kritiškumo lygį ir kinta toje pačioje skalėje kaip ir paslaugų kokybė. Tokiu būdu, tikimybė, kad j-asis studentas teigiamai įvertins i-ąją paslaugą, priklauso nuo skirtumo $\theta_i - b_j$, $i = 1, \ldots, m$, $j = 1, \ldots, n$. Narys $\beta^\top x_{ij}$ aprašo papildomų aiškinančiųjų kintamųjų x_{ij} įtaką apklausos rezultatams. Kadangi x_{ij}, kaip rodo jo indeksai, gali būti susijęs tiek su vartotojo savybėmis, tiek ir su teikiamos paslaugos ypatumais, tai (1) modeliu galima aprašyti ir tam tikras jų sąveikas.

Kadangi j-ajam studentui i-osios ir m-osios paslaugos teigiamo įvertinimo *šansų* $O(i|j)$ ir $O(m|j)$,

$$O(i|j) := \frac{\mathbf{P}\{y_{ij} = 1 \mid x_{ij}\}}{\mathbf{P}\{y_{ij} = 0 \mid x_{ij}\}}, \quad j = 1, \ldots, n,$$

santykis (angl. *odds ratio*)

$$OR_m(i|j) := \frac{O(i|j)}{O(m|j)} = \exp\left\{\theta_i - \theta_m + \beta^\top (x_{ij} - x_{mj})\right\}$$

nepriklauso nuo maišančiojo parametro b_j, tai nuo jų nepriklauso ir sąlyginė tikėtinumo funkcija [5]. Taigi, (1) modelio atveju šansų santykiai $OR_m(i|j)$ priklauso tik nuo klausimo ir stebimų studento savybių x_{ij} bei bendro visai populiacijai nežinomo parametro β. Todėl $OR_m(i|j)$ išlieka toks pats visiems studentams su tom pačiom savybėm x_{ij}, ir vadinasi, nesvarbu kuris iš jų dalyvavo ar nedalyvavo apklausoje.

Kai m yra fiksuotas, maksimizuojant sąlyginio tikėtinumo funkciją gaunami pagrįsti ir asimptotiškai normalieji nežinomų parametrų θ_i, $i = 1, \ldots, m$, ir β įvertiniai.

2 Rezultatai

Tyrime naudojami internetinės apklausos apie ITTC teikiamų IT paslaugų kokybę duomenys. Internetinėje apklausoje dalyvavo 2000 VU studentų: 2009 m. – 759, 2010 m. – 1241 studentai. Tai maža dalis visų VU studentų, kurių 2010.02.01 buvo 21340 (be TVM ir Užsienio kalbų instituto).

Tarp maždaug 30 klausimų buvo klausimai apie šių ITTC teikiamų paslaugų kokybę:

Q1A1 – elektroninio pašto,

Q1A2 – grupinio elektroninio susirašinėjimo paslaugos (elektroninės konferencijos),

Q1A3 – informacinės sistemos,

Q1A4 – prieigos prie interneto (padalinio, kuriame studijuoja, auditorijose),

Q1A5 – EDUROAM, bevielio tinklo (interneto) prieigos,

Q1A6 – TTC bendro naudojimo kompiuterių salės,

Q1A7 – VU VPN, virtualus privatus tinklo (galimybe naudotis užsienio šalių mokslinėmis duomenų bazėmis iš namų ir kita).

Paslaugų vertinimo skalė yra tokia: 1 – nežinau, 2 – nesinaudoju, 3 – kokybė prasta, 4 – kokybė vidutiniška, 5 – kokybė gera.

Kintamieji, charakterizuojantys studentus: Fakultet – fakutetas , Q9A1 – studijų forma, Q10A1 – studijų pakopa/kursas, metai – apklausos metai.

Pirminis uždavinys buvo išsiaiškinti, ar skiriasi skirtingų fakultetų, skirtingų kursų bei studijų formos studentų vertinimai. Taip pat buvo tiriama, ar per metus vertinimai pasikeitė. Naudojant χ^2 požymių nepriklausomumo (homogeniškumo) kriterijų ir lyginant įvairiais pjūviais sudarytas studentų grupes daugeliu atvejų buvo gauti statistiškai reikšmingi skirtumai. Daugelis gautų priklausomybių tarp paslaugų kokybės vertinimų ir papildomų kintamųjų yra nesunkiai paaiškinamos. Pavyzdiui, studentai, pasirinkę ištęstines (neakivaizdines, vakarines) studijas, kaip taisyklė, rečiau naudojasi ITTC paslaugomis, o EDUROM bevielio tinklo paslauga (prienama fakultetuose) yra mažiau aktuali neakivaizdininkams, kurie rečiau būna universitete. Kitas pavyzdys – kadangi bendrojo naudojimo salė yra Saulėtekio alėjoje, tai ja dažniau naudojasi šalia esančių fakultetų studentai, tuo pačiu ir jų šios paslaugos vertinimai skiriasi nuo kitų fakultetų.

Kadangi pagal pagalbinius kintamuosius (Fakultet, Q9A1, Q10A1, metai) sudarytose studentų grupėse anketas užpildžiusių studentų proporcijos, o taip pat ir jų nuomonė apie teikiamų paslaugų kokybę ženkliai skyrėsi, tai galimam poslinkiui kompensuoti buvo taikomi persvėrimo metodai ir sąlyginė logistinė regresija. Pradinis ranginis kokybės vertinimo kintamasis buvo perkoduotas į binarinį: 0, jeigu paslaugą vertina blogai, ja nesinaudoja arba jos nežino, ir 1, jeigu paslaugą vertina gerai arba vidutiniškai.

Pažymėtina, kad sąlyginės logistinės regresijos modelio pagrindu gautus rezultatus palyginti su kitų metodų rezultatais sudėtinga, nes pagal padarytas jame prielaidas nuo neatsakymų pasiskirstymo nepriklauso tik šansų santykiai. Juos ir derėtų naudoti metodų palyginimui, bet kadangi toks palyginimas būtų gremėzdiškas, šiame darbe lyginome įvairiais metodais įvertintas teigiamo paslaugos vertinimo tikimybių procentines reikšmes.

1 lentelė.

Metai	Metodas	Q1A1	Q1A2	Q1A3	Q1A4	Q1A5	Q1A6	Q1A7
2009	BS	69,3	31,4	88,0	55,3	24,5	56,0	37,2
	FS	65,6	24,1	89,3	50,6	22,5	48,4	37,7
	LS	63,9	24,1	89,5	50,7	23,3	49,8	37,6
	SL/BS	68,8	29,1	88,9	53,7	22,1	55,0	36,2
	SL/LS	62,5	22,0	90,4	50,4	20,5	49,8	33,8
2010	BS	74,8	22,9	89,0	51,5	29,3	51,5	41,3
	FS	71,5	20,2	89,0	50,2	26,8	50,8	39,9
	LS	71,8	20,2	89,3	49,9	26,9	50,6	39,6
	SL/BS	74,8	21,1	89,5	49,6	27,0	49,9	39,9
	SL/LS	71,6	18,1	90,5	47,3	25,5	48,3	39,7

1 lentelėje pateiktos 7-ių paslaugų teigiamo vertinimo tikimybės. Santrumpa BS („Be Svorių") žymi atvejį, kai persvėrimas nebuvo atliekamas, FS („Fakultetų Svoriai") – persvėrimui buvo taikomas AHG modelis su homogeniškumo grupėmis, sudarytomis pagal pagalbinius kintamuosius, LS („Logistinės regresijos Svoriai") – persvėrimui buvo taikomas logistinės regresijos modelis. Pažymėjimai SL/BS, SL/FS ir SL/LS reiškia, kad atskiro studento teigiamo vertinimo tikimybės buvo įvertintos naudojant (1) modelį ir po to suvidurkintos su atitinkamais svoriais. Vadinasi, skaičiuojant teigiamo vertinimo tikimybes jau ir šiuo atveju remiamasi atitinkamom metodų BS, FS ir LS prielaidomis.

Apibendrinant gautus rezultatus pastebėsime, kad

(1) studentų skirtingų grupių (fakultetų, studijų pakopų ir formų) neatsakymų proporcijos yra gana skirtingos, o tiek χ^2 testas, tiek ir sąlyginės logistinės regresijos pagrindu gauti rezultatai rodo, kad skirtingų grupių studentai turėjo aplamai skirtingą nuomonę, kuri paslauga yra kokybiškesnė;

(2) matyt, adekvačiausi rezultatai yra gaunami SL/LS metodu, nes jis yra universaliausias ir adaptyviausias iš nagrinėtų metodų, be to SL/LS ir SL/FS metodų rezultatai praktiškai sutampa ir todėl čia nepateikiami;

(3) daugumoje atvejų visų persvėrimo metodų rezultatai yra panašūs (kas dalinai patvirtina gautų rezultatų patikimumą; išimtis yra 7-oji paslauga 2009 m.), bet jie ženkliai skiriasi nuo rezultatų be persvėrimo (BS ir SL/BS metodai) 2010 metais paslaugoms Q1A2, Q1A4 ir Q1A5, o 2009 metais visoms paslaugoms išskyrus Q1A3;

(4) sąlyginės logistinės regresijos pagrindu apskaičiuotos teigiamo paslaugos vertinimo tikimybės paprastai sumažina neatsakymų sąlygotą poslinkį net ir nenaudojant svorių (lyginama SL/BS su BS).

3 Išvados

- Neatsakymų sąlygoto poslinkio sumažinimo metodai remiasi tam tikrom prielaidom, todėl prieš juos taikant yra labai svarbu įvertinti tų prielaidų adekvatumą.

- Atliekant studentų internetines apklausas reiktų įtraukti papildomų klausimų, kurių atsakymai arba bent tų atsakymų proporcijos ar vidurkiai yra žinomi visoje populiacijoje (pvz., lytis, studijų pakopa).

- Sąlyginės logistinės regresijos modelis yra taikytinas tada, kai tyrėją domina ne patys apklausos atsakymai, o jų tikimybių palyginimas, surangavimas. Tuomet jis leidžia gauti nepaslinktus atsakymų palyginimo rezultatus prie gana bendrų sąlygų nenaudojant kitų neatsakymų sąlygoto poslinkio sumažinimo metodų.
- Kartu su internetine apklausa rekomenduotina atlikti ir tiesioginę atsitiktinę apklausą (rezultatų kalibravimui).

Literatūra

[1] J. Bethlehem. *Selection Bias in Web Surveys*. 2010. Available from Internet: http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2010.00112.x/full.

[2] A. Gefenienė, D. Šimkūnienė, M. Jurgutis, M. Girdvainis ir A. Grigonis. *Apklausos apie ITTC teikiamų IT paslaugų kokybę tyrimo ataskaita*. 2011. Adresas internete: http://www.ittc.vu.lt/dokumentai/Dokumentai/Apklausos_apie_ITTC_teikiamas_paslaugas_2010_rezultatai.pdf.

[3] M. Jurgutis, M. Girdvainis ir A. Grigonis. *Apklausos apie ITTC teikiamų IT paslaugų kokybę tyrimo ataskaita*. 2010. Adresas internete: http://www.ittc.vu.lt/dokumentai/Dokumentai/Apklausos_apie_ITTC_teikiamas_paslaugas_2009_rezultatai.pdf.

[4] S. Lee. *Statistical estimation methods in volunteer panel WEB surveys*, 2004. Available from Internet: http://drum.lib.umd.edu/bitstream/1903/2003/1/umi-umd-1957.pdf. Disertation.

[5] M.E. Stokes, C.S. Davis and G.G. Koch. *Categorical Data Analysis Using the SAS System Categorical Data Analysis (2nd ed.)*. SAS Institute Inc., Cary, NC, 2000.

[6] C.-E. Särndal, B. Swensson and J. Wretman. *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1992.

SUMMARY

Analysis of student WEB survey

G. Murauskas, M. Radavičius

In the paper the problem of adjustment for nonresponse in a WEB survey is addressed. The study is based on Vilnius University student survey on quality assessment of information technology services. A conditional logistic regression model is applied along with traditional nonresponse adjustment methods.

Keywords: nonresponse adjustment, conditional logistic regression, WEB survey, reweighting.