VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
MATEMATINĖS STATISTIKOS KATEDRA

Jurgita Utovkaitė

PRALEISTŲ DUOMENŲ ĮRAŠYMO METODAI BAIGTINIŲ POPULIACIJŲ STATISTIKOJE

Magistro baigiamasis darbas

Vadovas
Doc. Aleksandras Ernestas Plikusas

VILNIUS 2006

TURINYS

I. ĮVADAS

Statistinių tyrimų būtinybė kyla tuomet, kai valstybinėms institucijoms ar kitiems vartotojams prireikia informacijos apie socialinius ar ekonominius veiksnius ir egzistuojantys duomenų šaltiniai šio poreikio nepatenkina.

Netgi tobuliausiai suplanuotame tyrime atsiranda įvairių rūšių klaidų, dėl kurių gali būti gauti nepatikimi ar nepakankamai tikslūs tyrimo rezultatai, taigi labai svarbu kiek įmanoma labiau sumažinti tų klaidų įtaką tyrimo rezultatams – sumų, vidurkių, santykių įvertiniams.

Vienas iš galimų statistinio tyrimo klaidų tipų yra *klaidos dėl neatsakymo į apklausą*. Jos atsiranda tuomet, kai atsakytojas neatsako į vieną ar kelis klausimyno klausimus. Neatsakymai tyrimuose pasitaiko dėl įvairių priežasčių. Jie iššaukia standartinių įvertinių, kuriuose neatsižvelgiama į neatsakymus, nuokrypį nuo tikrųjų mus dominančių reikšmių, o taip pat šių įvertinių dispersijos padidėjimą. Dabartinėje praktikoje neatsakymai į apklausą nagrinėjami dviem požiūriais: visų pirma bandoma išvengti arba sumažinti neatsakymų lygį. Yra nemažai literatūros ir metodologinės medžiagos tyrinėjančios neatsakymų priežastis bei pateikiančios rekomendacijas kaip sumažinti neatsakymų lygį, tačiau, kai tyrime jau yra neatsakymų, dominančius įvertinius reikia sukonstruoti taip, kad tyrimo rezultatai būtų kuo tikslesni.

Neatsakymų sukeliamiems tyrimo rezultatų nuokrypiams sumažinti naudojami įvairūs būdai. Vienas tokių metodų yra *praleistų reikšmių įrašymas*. Įrašymas – tai trūkstamų duomenų užpildymo būdas, kuris yra labai naudingas analizuojant nepilnas duomenų sekas. Jis išsprendžia duomenų trūkumo problemą duomenų analizės pradžioje.

Praleistų reikšmių įrašymo metodika šiuo metu sparčiai vystosi, galima rasti nemažai straipsnių šia tema. Praleistų duomenų įrašymas sparčiai populiarėja tyrimų praktikoje, todėl labai svarbu ištirti duomenų įrašymo pritaikymo galimybes šalyje atliekamiems tyrimams ir pritaikyti turimiems duomenims.

Šiame darbe aptariami galimi neatsakymo į apklausą klaidos mažinimo būdai ir standartinių įvertinių taikymas, esant neatsakymams. Išsamiau išnagrinėti praktikoje dažniausiai taikomi praleistų reikšmių įrašymo metodai. Šių metodų veiksmingumas patikrintas turimai eksperimentinei imčiai sumodeliuotai pagal vieną Statistikos departamento prie Lietuvos Respublikos Vyriausybės (toliau – Statistikos departamento) atliekamą tyrimą, įrašymo metodai sulyginami tarpusavyje, pateikiami rezultatai bei išvados.

II. TEORINĖ DALIS

II.1. Paklaidos, atsiradusios dėl neatsakymo į apklausą

Statistiniai tyrimai gali būti įvairūs:

- Duomenys gaunami iš registrų, tokių tyrimų informacija dažnai yra ribota;
- Surašymai. Informacija renkama apie kiekvieną populiacijos individą. Šie tyrimai yra gana brangūs, atliekami kas kelis metus;
- Tyrimai, sudarant imtis, t.y. informacija gaunama renkant duomenis tik apie kai kuriuos populiacijos individus, rezultatus apibendrinant ir pritaikant visai populiacijai.

Pirmas tyrimo planavimo žingsnis yra tiksliai nustatyti tyrimo tikslus. Sekantys žingsniai – sudaryti imties planą, parengti duomenų redagavimo ir statistinių įverčių skaičiavimo metodologiją.

Visuose tyrimuose atsiranda klaidų, joms sumažinti naudojami įvairūs būdai, tačiau dalis jų vis dėl to lieka. Tyrimuose pasitaikančios klaidos dažniausiai skirstomos į du tipus. Pirmasis, tai ėmimo paklaida, ji atsiranda vien jau dėl tos priežasties, kad apklausiama ne visa populiacija, o tik jos dalis.

Antrasis tipas vadinamas su ėmimu nesusijusia paklaida. Ji apima visų kitų rūšių klaidas, dėl kurių gauti įvertiniai gali skirtis nuo tikrųjų vertinamų reikšmių. Su ėmimu nesusijusios paklaidos gali būti įvairių tipų ir dažniausiai yra klasifikuojamos:

Aprėpties paklaidos, atsiradusias dėl tyrimo ir tikslo populiacijų nesutapimo;

Tyrimo imčiai sudaryti Statistikos departamentas naudojasi subjektų registro duomenimis. Subjektų registras yra nuolatos atnaujinamas, jame saugoma informacija apie visų šalies įmonių veiklą, darbuotojų skaičių, metines pajamas, taip pat kontaktinė informacija – adresas, telefono numeris, atsakingų asmenų duomenys. Tačiau, kad ir kaip greitai bei produktyviai informacija subjektų registre atnaujinama – klaidos nėra išvengiamos, tokiu būdu ir atsiranda aprėpties paklaidos.

Matavimo paklaidos, atsirandančios dėl atsakytojų nesugebėjimo pateikti tikruosius atsakymus;

Duomenų apdorojimo klaidos, sukeltos bet kokių veiksmų, dėl kurių gauname prastus parametrų įvertinius;

Bei *klaidos dėl neatsakymo į apklausą*, atsiradusios dėl įvairiausių priežasčių, dėl kurių prarandame duomenis.

Atsakytojų, kurie neatsako į pateiktus anketos ar klausimyno klausimus pasitaiko kiekviename tyrime. Taigi turime neatsakymus. Tyrime 20 procentų neatsakymų lygis yra gana normalus reiškinys ir daugumoje statistinių tyrimų jis kur kas didesnis. Neatsakymų lygio analizė rodo, kad socialiniuose tyrimuose žemesnis atsakymų lygis pastebimas miestiečių, vienišų žmonių, bevaikių šeimų narių, vyresnių, išsiskyrusių žmonių, našlių tarpe, taip pat tarp žemesnio išsilavinimo žmonių bei įmonių savininkų tarpe. Įmonių tyrimuose jis priklauso nuo įmonės veiklos, jos dydžio ir kitų priežasčių.

Galimi du neatsakymų tipai:

Elemento (vieneto) neatsakymas arba neatsakymas į klausimyną – kai atsakytojas neatsako nė į viena anketos ar klausimyno klausimą.

Neatsakymas į klausimą – kai trūksta vieno ar kelių anketos, ar klausimyno atsakymų. Šį atvejį taip pat galime vadinti daliniu atsakymu. Pastarasis praktikoje pasirodo daug rečiau nei elemento neatsakymas.

Geras neatsakymų įtaką mažinantis veiksnys yra *papildoma informacija*, ji padeda sumažinti neatsakymų sukeliamą vertinamų parametrų nuokrypį bei dispersiją.

II.2. Papildoma informacija

Papildomas kintamasis – rodiklis, kurio reikšmės yra žinomos kiekvienam populiacijos ir/arba į imtį patekusiam elementui, nesvarbu ar jis atsakė ar ne.

Papildomas vektorius yra sudarytas iš vieno ar daugiau papildomųjų kintamųjų.

Papildomą informaciją turime tuomet, kai žinome papildomą vektorių, o tuo pačiu ir bendras sumines papildomo vektoriaus kintamųjų reikšmes visai populiacijai.

Papildoma informacija galime laikyti administracinius duomenis, registrų duomenis, kitų tyrimų duomenis ar to paties, tik prieš kurį laiką atlikto tyrimo duomenis. Jie naudojami kartu su tyrimo duomenimis padėdami pagerinti tyrimo kokybę, sumažinti tyrimo kaštus.

II.3. Pagrindiniai žymenys

Tegu $U = (1,...,k,...,N)$ – mus dominančios populiacijos vienetų aibė;

Tarkime stebimas tyrimo kintamasis *y*, kurio baigtinės populiacijos elementų reikšmės yra $y_1,...,y_N$.

y_k – dominanti *k*-tojo elemento stebėjimo reikšmė;

Mūsų tikslas įvertinti sumą $t_y = \sum_{k=1}^{N} y_k$.

n – efektyvusis imties dydis arba skirtingų elementų skaičius imtyje;

$\mathbf{i} = (i_1,...,i_n)$– imties vienetų (elementų) aibė, išrinkta su tikimybe *p*(**i**);

π_k – *k*-tojo populiacijos elemento priklausymo imčiai tikimybė;

$\pi_k = \sum_{\mathbf{i}:k\in\mathbf{i}} p(\mathbf{i})$, $\pi_k > 0$;

d_k – elemento *k* svoris, $d_k = 1/\pi_k$;

π_{kl} – *k*-tojo ir *l*-tojo populiacijos elementų tikimybė kartu priklausyti imčiai arba antros eilės priklausymo imčiai tikimybė;

$\pi_{kl} = \sum_{\mathbf{i}:(k,l)\subset\mathbf{i}} p(\mathbf{i})$, $k \in U$, $l \in U$, jei $k = l$, tai $\pi_{kk} = \pi_k$;

d_{kl} – antros eilės patekimo į imtį svoris, $d_{kl} = 1/\pi_{kl}$;

Kai mus domina įvairios populiacijos sritys (sluoksniai), turime:

$U_1,...,U_d,...,U_D$ – populiacijos sritys (dengia visą populiaciją, tarpusavyje nesikerta);

$Y_1,...,Y_d,...,Y_D$ – dominančios sumų reikšmės;

$Y_d = \sum_{k\in U_d} y_k$, $d = 1,\ldots,D$ arba $Y_d = \sum_{k\in U} y_{d,k}$, $y_{d,k} = \begin{cases} y_k, & kai\ k \in U_d \\ 0, & kai\ k \notin U_d \end{cases}$

II.4. Įvertiniai

II.4.1. Horvico – Tompsono įvertinys

Šis įvertinys yra universalus populiacijos sumos įvertinys, tinkantis bet kokiam imties planui. Šis įvertinys yra nepaslinktas, kai $\pi_k > 0$ visiems elementams *k*.

Nepaslinktas populiacijos sumos įvertinys:

$$\hat{t}_\pi = \sum_{\mathbf{i}:k\in\mathbf{i}} \frac{y_k}{\pi_k},$$

Šio įvertinio dispersija:

$$\mathrm{D}\hat{t}_\pi = \sum_{k=1}^{N}\left(\frac{1-\pi_k}{\pi_k}\right)y_k^2 + \sum_{k=1}^{N}\sum_{\substack{l=1\\ l\neq k}}^{N}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}y_k y_l = \sum_{k=1}^{N}\sum_{l=1}^{N}(\pi_{kl}-\pi_k\pi_l)\frac{y_k y_l}{\pi_k\pi_l},$$ jei pažymime $\pi_{kk} = \pi_k$.

Nepaslinktas dispersijos įvertinys:

$$\hat{\mathrm{D}}\hat{t}_\pi = \sum_{k\in\mathbf{i}}\left(\frac{1-\pi_k}{\pi_k^2}\right)y_k^2 + \sum_{k\in\mathbf{i}}\sum_{\substack{l\in\mathbf{i}\\ l\neq k}}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\frac{y_k y_l}{\pi_{kl}} = \sum_{k\in\mathbf{i}}\left(\frac{1}{\pi_k^2}-\frac{1}{\pi_{ki}}\right)y_k^2 + \sum_{k\in\mathbf{i}}\sum_{\substack{l\in\mathbf{i}\\ l\neq k}}\left(\frac{1}{\pi_k\pi_l}-\frac{1}{\pi_{ki}}\right)y_k y_l =$$

$$= \sum_{k\in\mathbf{i}}\sum_{l\in\mathbf{i}}\frac{(\pi_{kl}-\pi_k\pi_l)}{\pi_{kl}}\frac{y_k y_l}{\pi_k\pi_l}.$$

Horvico – Tompsono sumos įvertinio dispersijos įvertinio pagrindinis trūkumas yra tai, kad jis kartais gali įgyti ir neigiamas reikšmes, nors pati dispersija visada yra neneigiama.

II.4.2. Regresinis įvertinys

Regresiniame sumos įvertinyje naudojame papildomą kintamąjį su žinoma suma visai populiacijai. Ji tam tikromis sąlygomis padeda tiksliau įvertinti tiriamo kintamojo sumą arba vidurkį.

Vertinant parametrus šiuo atveju naudojame papildomą kiekybinį kintamąjį *x*. Tariame, kad žinome šio kintamojo imties elementų reikšmes ir šio kintamojo populiacijos sumą t_x.

Taigi stebime tyrimo kintamąjį *y*, nagrinėjame paprastąją atsitiktinę *n* dydžio imtį iš *N* dydžio baigtinės populiacijos. Turime papildomąjį kintamąjį su žinomomis imties elementų reikšmėmis x_k, $k \in \mathbf{i}$ ir žinoma suma t_x. Turėdami duomenis (x_k, y_k), $k \in \mathbf{i}$, galime išvesti kintamojo *y* regresijos tiesę *x* atžvilgiu. Ji išreiškiama lygtimi:

$\hat{y}_k = \hat{A} + \hat{B}x_k$, $k \in$ i, kur $\hat{A}$ ir $\hat{B}$ yra mažiausių kvadratų metodu gaunami koeficientai minimizuojantys sumą $\sum_{k\in\mathbf{i}}(y_k - A - Bx_k)^2$:

$$\hat{A} = \bar{y} - \hat{B}\bar{x}\ ,\quad \bar{y} = \frac{1}{n}\sum_{k\in\mathbf{i}} y_k\ ,\quad \bar{x} = \frac{1}{n}\sum_{k\in\mathbf{i}} x_k\ ,\quad \hat{B} = \frac{\sum_{k\in\mathbf{i}}(x_k-\bar{x})(y_k-\bar{y})}{\sum_{k\in\mathbf{i}}(x_k-\bar{x})^2}.$$

Visų populiacijos elementų kintamojo *y* reikšmių įvertinių suma vadinama regresiniu sumos t_y įvertiniu, jis skaičiuojamas:

$$\hat{t}_{yreg} = \sum_{k=1}^{N}\hat{y}_k = \hat{t}_y + \hat{B}(t_x - \hat{t}_x),$$ čia $\hat{t}_y = \sum_{k\in\mathbf{i}} y_k = n\bar{y}$, $\hat{t}_x = \sum_{k\in\mathbf{i}} x_k$.

Paprastosios atsitiktinės imties atveju regresinio sumos įvertinio apytikslė dispersija yra:

$$\mathrm{AD}\hat{t}_{yreg} = N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}(1-\rho^2) = N^2\left(1-\frac{n}{N}\right)\frac{S_{yreg}^2}{n}, \text{ kur}$$

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i-\mu_y\right)^2,$$

$$S_{yreg}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i-\left(\mu_y+B\left(x_i-\mu_x\right)\right)\right)^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i-\left(A+Bx_i\right)\right)^2, \text{ čia}$$

$$A=\mu_y-B\mu_x,\ \ B=\frac{\sum_{k=1}^{N}\left(x_k-\mu_x\right)\left(y_k-\mu_y\right)}{\sum_{k=1}^{N}\left(x_k-\mu_x\right)^2}=\frac{S_{xy}}{S_x^2},$$

$$\mu_x=\frac{1}{N}\sum_{i=1}^{N}x_i,\quad \mu_y=\frac{1}{N}\sum_{i=1}^{N}y_i,$$

$\rho=\frac{S_{xy}}{S_xS_y}$ - kintamųjų *y* ir *x* populiacijos koreliacijos koeficientas.

Regresinio sumos įvertinio $\hat{t}_{yreg}$ dispersija gali būti vertinama:

$$\hat{\mathrm{D}}\hat{t}_{yreg} = N^2\left(1-\frac{n}{N}\right)\frac{\hat{S}_{yreg}^2}{n}, \text{ čia}$$

$$\hat{S}_{yreg}^2 = \frac{1}{n-2}\sum_{i\in\mathbf{i}}\left(y_i-\left(\bar{y}+\hat{B}\left(x_i-\bar{x}\right)\right)\right)^2 = \frac{1}{n-2}\sum_{i\in\mathbf{i}}\left(y_i-\left(\hat{A}+\hat{B}x_i\right)\right)^2,$$

$$\hat{A}=\bar{y}-\hat{B}\bar{x},\quad \hat{B}=\frac{\sum_{k\in\mathbf{i}}\left(x_k-\bar{x}\right)\left(y_k-\bar{y}\right)}{\sum_{k\in\mathbf{i}}\left(x_k-\bar{x}\right)^2}.$$

Populiacijos vidurkio μ_y regresinis įvertinys paprastosios atsitiktinės imties atveju yra

$$\hat{\mu}_{yreg}=\frac{1}{N}\hat{t}_{yreg}=\frac{1}{N}\left(\hat{A}+\hat{B}\hat{t}_x\right)=\hat{\mu}_y+\hat{B}\left(\mu_x-\hat{\mu}_x\right),$$

Jo apytikslė dispersija gaunama, sumos įvertinio $\hat{t}_{xreg}$ dispersiją padalijus iš N^2.

II.4.2.1. Regresiniai įvertiniai bet kokio imties plano atveju

Bet kokio imties plano atveju sumai vertinti taikysime tokio pat pavidalo regresinį sumos įvertinį, kaip ir paprastosios atsitiktinės imties atveju:

$$\hat{t}_{yreg} = \hat{t}_{y\pi} + \hat{B}\left(t_x - \hat{t}_{x\pi}\right),$$

tačiau čia $\hat{t}_x$ ir $\hat{t}_y$ yra kintamųjų *y* ir *x* sumų t_x ir t_x Horvico – Tompsono įvertiniai, t.y.

$$\hat{t}_{y\pi} = \sum_{k\in\mathbf{i}} \frac{y_k}{\pi_k}, \quad \hat{t}_{\pi\pi} = \sum_{k\in\mathbf{i}} \frac{x_k}{\pi_k}.$$

π_k – *k*-tojo populiacijos elemento priklausymo imčiai tikimybė.

Regresijos koeficiento *B* įvertinį $\hat{B}$ galime perrašyti taip:

$$\hat{B} = \frac{\hat{S}_{xy}}{\hat{S}_x^2} = \frac{\sum_{k\in\mathbf{i}} (x_k - \bar{x})(y_k - \bar{y})}{\sum_{k\in\mathbf{i}} (x_k - \bar{x})^2} = \frac{\sum x_k y_k - \bar{y}\sum x_k - \bar{x}\sum y_k + n\bar{x}\bar{y}}{\sum x_k^2 - 2\bar{x}\sum x_k + n\bar{x}^2} = \frac{\sum x_k y_k - n\bar{x}\bar{y}}{\sum x_k^2 - n\bar{x}^2} =$$

$$= \frac{\frac{N}{n}\sum_{k\in\mathbf{i}} x_k y_k - \frac{1}{N}\hat{t}_x\hat{t}_y}{\frac{N}{n}\sum_{k\in\mathbf{i}} x_k^2 - \frac{1}{N}\hat{t}_x^2}.$$

Kadangi paprastojo atsitiktinio ėmimo atveju elemento priklausymo imčiai tikimybė yra $\pi_k = \frac{n}{N}$, tai galime rašyti:

$$\hat{B} = \frac{\sum_{k\in\mathbf{i}} \frac{x_k y_k}{\pi_k} - \frac{1}{\hat{N}}\hat{t}_{x\pi}\hat{t}_{y\pi}}{\sum_{k\in\mathbf{i}} \frac{x_k^2}{\pi_k} - \frac{1}{\hat{N}}\hat{t}_{x\pi}^2}.$$

Regresinio sumos įvertinio $\hat{t}_{yreg} = \hat{t}_{y\pi} + \hat{B}\left(t_x - \hat{t}_{x\pi}\right)$ apytikslė dispersija yra

$$\mathrm{AD}\hat{t}_{yreg} = \sum_{k=1}^{N}\sum_{l=1}^{N} (\pi_{kl} - \pi_k\pi_l) \frac{y_k - A - Bx_k}{\pi_k} \frac{y_l - A - Bx_l}{\pi_l}, \text{ čia}$$

$A = \mu_y - B\mu_x$, $B = \frac{S_{xy}}{S_x^2}$, π_{kl} – *k*-tojo ir *l*-tojo populiacijos elementų tikimybė kartu priklausyti bent vienai iš galimų imčių.

Rekomenduojamas dispersijos įvertinys:

$$\hat{\mathrm{D}}\hat{t}_{yreg} = \sum_{k\in\mathbf{i}}\sum_{l\in\mathbf{i}} \left(1 - \frac{\pi_k\pi_l}{\pi_{kl}}\right) \frac{y_k - \hat{A} - \hat{B}x_k}{\pi_k} \frac{y_l - \hat{A} - \hat{B}x_l}{\pi_l},$$

$$\hat{A}=\frac{1}{\hat{N}}\left(\sum_{k\in\mathbf{i}}\frac{y_k}{\pi_k}-\hat{B}\sum_{k\in\mathbf{i}}\frac{x_k}{\pi_k}\right),\ \hat{B}=\frac{\sum_{k\in\mathbf{i}}\frac{x_k y_k}{\pi_k}-\frac{1}{\hat{N}}\hat{t}_{x\pi}\hat{t}_{y\pi}}{\sum_{k\in\mathbf{i}}\frac{x_k^2}{\pi_k}-\frac{1}{\hat{N}}\hat{t}_{x\pi}^2}.$$

II.4.3. Kalibruotas sumos įvertinys

Kalibruotų įvertinių atveju taip pat naudojamės papildomų kiekybinių kintamųjų informacija. Tarkime, kad turime baigtinę populiaciją $U=(1,...,k,...,N)$, kurioje apibrėžtas tyrimo kintamasis *y* su reikšmėmis $y_1,...,y_N$. Tarkime, kad populiacijoje turime dar *J* papildomų kintamųjų $x^{(1)},...,x^{(J)}$, kintamojo $x^{(j)}$ reikšmes žymėkime x_{jk}, *k=1,…,N*. Taigi turime *J*x*N* papildomos informacijos matricą. Pažymėkime papildomų kintamųjų $x^{(1)},...,x^{(J)}$ sumas:

$$t_{x1}=\sum_{k=1}^{N}x_{1k}\ ,\ \ldots\ ,\ t_{xJ}=\sum_{k=1}^{N}x_{Jk}\ ,$$

šių kintamųjų vektoriaus $\mathbf{x}=\left(x^{(1)},...,x^{(J)}\right)'$ populiacijos reikšmių sumą pažymėkime

$$\mathrm{t}_x=\sum_{k=1}^{N}\mathrm{x}_k=\left(t_{x1},...,t_{xJ}\right)'.$$

Šią sumą laikome žinoma.

Kintamojo *y* sumos t_y kalibruotu įvertiniu vadinamas toks įvertinys: $\hat{t}_{yw}=\sum_{k\in\mathrm{i}}w_k y_k$.

Čia svoriai w_k parenkami taip, kad kaip galima mažiau skirtūsi nuo imties plano svorių $d_k=\frac{1}{\pi_k}$, $\pi_k=P\{k\in\mathbf{i}\}$, tenkintų kalibracijos lygtį

$\hat{t}_{xw}=\sum_{k\in\mathbf{i}}w_k\mathrm{x}_k=\mathrm{t}_x$, čia

$$\hat{t}_{xw}=\left(\hat{t}_{xw1},...,\hat{t}_{xwJ}\right)',\quad \hat{t}_{x1}=\sum_{k\in\mathbf{i}}w_k x_{1k}\ ,\ \ldots\ ,\ \hat{t}_{xJ}=\sum_{k\in\mathbf{i}}w_k x_{Jk}\ ,$$

bei minimizuotų funkciją

$$L\left(w_k,d_k,k\in\mathbf{i}\right)=E\sum_{k\in\mathbf{i}}\frac{(w_k-d_k)^2}{d_k q_k}\rightarrow\min.$$

Čia q_k - laisvai pasirenkami svoriai. Dažniausiai laikoma, kad q_k = 1.

Minimizuodami *L* funkciją gauname tokio pavidalo kalibruotus svorius:

$$w_k = d_k\left(1+q_k\left(\mathrm{t}'_x-\sum_{k\in\mathbf{i}} d_k\mathrm{x}'_k\right)\left(\sum_{i\in\mathbf{i}} d_i q_i \mathrm{x}_i\mathrm{x}'_i\right)^{-1}\mathrm{x}_k\right).$$

Šio įvertinio apytikrė dispersija yra

$$\mathrm{AD}\hat{t}_{yw}=\sum_{k,l=1}^{N}(\pi_{kl}-\pi_k\pi_l)w_k w_l(y_k-\mathrm{x}'_k\,\hat{\mathrm{b}}_w)(y_l-\mathrm{x}'_l\,\hat{\mathrm{b}}_w),$$

čia vektorius $\hat{\mathrm{b}}_w$ yra lygties $\left(\sum_{k\in\mathbf{i}} w_k q_k \mathrm{x}_k\mathrm{x}'_k\right)\hat{\mathrm{b}}_w=\sum_{k\in\mathbf{i}} w_k q_k \mathrm{x}_k y_k$ sprendinys.

Siūlomas kalibruoto sumos įvertinio dispersijos įvertinys

$$\hat{\mathrm{D}}\hat{t}_{yw}=\sum_{k,l\in\mathbf{i}}\left(1-\frac{\pi_k\pi_l}{\pi_{kl}}\right)w_k w_l(y_k-\mathrm{x}'_k\,\hat{\mathrm{b}}_w)(y_l-\mathrm{x}'_l\,\hat{\mathrm{b}}_w).$$

Galime pasirinkti ir kitokias *L* funkcijas bei konstruoti daug skirtingų įvertinių.

II.5. Neatsakymai

Esant neatsakymų į apklausą, bandoma įvairiais būdais įvertinti neatsakymo tikimybę. Bandoma įrašinėti praleistas reikšmes, rezultatus vertinti pagal turimus duomenis ir panašiai. Vienas iš būdų populiacijos sumai įvertinti, kai yra neatsakymų į apklausą, yra imties plano svorių kalibravimas.

II.5.1. Imties plano svorių kalibravimas, esant neatsakymams į apklausas

Tai viena dažniausiai taikomų neatsakymų įtakos tyrimui mažinimo priemonių. Naudojant šį būdą būtina turėti tinkamos papildomos informacijos, kuri padėtų suformuoti kalibruotus svorius. Kalibracijos metodas papildoma informacija naudojasi imties ir neatsakymų paklaidos sumažinimui. Metodas patogus tuo, kad gali būti pritaikytas beveik visiems imčių sudarymo būdams.

Kalibruoti įvertiniai esant neatsakymams į apklausas gali būti pritaikyti įvairioms situacijoms.

Pažymėkime $\mathbf{i}^{(a)}$ į apklausą atsakiusių elementų aibę $\mathbf{i}^{(a)}\subset\mathbf{i}\subset U$. Kintamojo *y* sumai vertinti sudarome kalibruotą įvertinį $\hat{t}_w^{(a)}=\sum_{k\in\mathbf{i}^{(a)}} w_k y_k$.

Tarkime žinome papildomų kintamųjų vektoriaus **x** į apklausą atsakiusių imties elementų reikšmes

$\mathrm{x}_k = \left(x_{1k},...,x_{Jk}\right)'$, $k \in \mathbf{i}^{(a)}$ ir visų šio vektoriaus populiacijos reikšmių sumą $\mathrm{t}_x = \mathrm{x}_1 + ... + \mathrm{x}_N$, tuomet kalibruoto sumos įvertinio $\hat{t}_{yw}^{(a)}$ svoriai yra

$$w_k = d_k\left(1 + q_k\left(t'_x - \sum_{k\in\mathbf{i}^{(a)}} d_k \mathrm{x}'_k\right)\left(\sum_{i\in\mathbf{i}^{(a)}} d_i q_i \mathrm{x}_i \mathrm{x}'_i\right)^{-1}\mathrm{x}_k\right),$$

jie tenkina kalibracijos lygtį $\hat{\mathrm{t}}_{xw} = \sum_{k\in\mathbf{i}^{(a)}} w_k \mathrm{x}_k = \mathrm{t}_x$ ir minimizuoja funkciją

$$L\left(w_k, d_k, k \in \mathbf{i}^{(a)}\right) = \sum_{k\in\mathbf{i}^{(a)}} \frac{(w_k - d_k)^2}{d_k q_k} \to \min.$$

Jei papildomo vektoriaus t_x reikšmių visai populiacijai nežinotume, tai kalibruoto sumos įvertinio $\hat{t}_{yw}^{(a)}$ svoriai būtų lygūs

$$w_k = d_k\left(1 + q_k\left(\sum_{k\in\mathbf{i}} d_k \mathrm{x}'_k - \left(\sum_{k\in\mathbf{i}^{(a)}} d_k \mathrm{x}'_k\right)\right)\left(\sum_{i\in\mathbf{i}^{(a)}} d_i q_i \mathrm{x}_i \mathrm{x}'_i\right)^{-1}\mathrm{x}_k\right),\ k \in \mathbf{i}^{(a)},$$

tenkinantys kalibracijos lygtį $\sum_{k\in\mathbf{i}^{(a)}} w_k \mathrm{x}_k = \sum_{k\in\mathbf{i}} d_k \mathrm{x}_k$.

Kitas gana populiarus būdas neatsakymų įtakos tyrimui sumažinti yra praleistų stebėjimų įrašymas.

II.5.2. Įrašymas

Įrašymas – procedūra, kurios metu vienas ar keli studijuojami kintamieji yra užpildomi pakaitalais, konstruojamais remiantis kažkokiomis taisyklėmis arba stebėtomis reikšmėmis kitiems elementams, būtinai patekusiems į imtį ir atsakiusiems.

Įrašymo paklaida panaši į matavimo klaidą tuomet, kai atsakytojas pateikia neteisingą informaciją klausėjams ir tikroji reikšmė nėra užfiksuojama.

Įrašytos reikšmės gali būti klasifikuojamos:

1) reikšmės sukonstruotos statistinės prognozės metodais;
2) reikšmės stebėtos ne neatsakiusiems elementams, bet vienarūšiams atsakiusiems elementams;
3) reikšmės sukonstruotos eksperto arba „geriausias galimas įvertinimas“;

Įrašymas naudojantis statistinio prognozavimo taisykle priskiriamas statistiniam įrašymui.

Tyrime duomenys dažniausiai išdėstomi stačiakampio forma. T.y., kiekvienam imties elementui, kurių yra *n*, turime po *J* tiriamų kintamųjų. Tokiu atveju viso turime *n*x*J* duomenų matricą, kurioje yra kažkoks skaičius „skylių“ – nežinomų reikšmių. Dažniausiai naudojami 2 požiūriai į įrašymą:

- Kai įrašomos reikšmės tiems *m* respondentų, kurie praleido bent vieną ar kelis klausimyno klausimus, o kitas grafas užpildė. Šiuo atveju turime *m*x*J* duomenų matricą.
- Kai įrašomos reikšmės ir neatsakiusiems į kelis klausimyno klausimus, ir visiškai klausimyno neužpildžiusiems elementams. Šiuo atveju turime *n*x*J* duomenų matricą, kur *n* yra imties dydis.

Yra nemažai įrašymo metodų, o kiekvienas metodas turi skirtingų pritaikymo būdų.

Tegu

$U = (1,...,k,...,N)$ – mus dominančios populiacijos vienetų aibė,

$\mathbf{i} = (i_1,...,i_n)$ – imties vienetų (elementų) aibė,

$\mathbf{i}^{(a)}$ į apklausą atsakiusių elementų aibę $\mathbf{i}^{(a)} \subset \mathbf{i} \subset U$,

$\mathrm{i}^{(n)}$ - į apklausą neatsakiusių elementų aibė.

Baigtine įrašyto kintamojo duomenų seka vadinsime reikšmių seką $\{y_{\bullet k} : k \in \mathbf{i}\}$, kur

$$y_{\bullet k} = \begin{cases} y_k, k \in \mathbf{i}^{(a)} \\ \hat{y}_k, k \in \mathbf{i}^{(n)} \end{cases}$$

T.y. $y_{\bullet k}$ yra lygi stebėtai reikšmei y_k, kai elementas *k* į klausimą atsakė ir $\hat{y}_k$, kai elementas *k* neatsakė į klausimą arba nepildė klausimyno.

II.5.2.1. Įvertiniai, kai naudojamas įrašymas

Norėdami įvertinti populiacijos sumą $t_y = \sum_{i=1}^{N} y_k$, tarsime, kad įrašomos reikšmės yra tokios pat geros, kaip ir tikrosios stebėtos $y_k, k \in \mathbf{i}^{(a)}$ reikšmės. Tokiu būdu įvertinių skaičiavimui galėsime naudoti tas pačias įvertinių skaičiavimo formules, kaip ir 100% atsakymų atveju. Čia elementas *k* turės tokį patį svorį, nepaisant to, ar y_k buvo tikra stebėta reikšmė, ar tai yra įrašoma reikšmė $y_{\bullet k}$.

Taigi Horvico – Tompsono įvertinys dabar turės tokį pavidalą:

$$\hat{t}_{\pi I} = \sum_{k\in\mathbf{i}} \frac{y_{\bullet k}}{\pi_k} = \sum_{k\in\mathbf{i}^{(a)}} \frac{y_k}{\pi_k} + \sum_{k\in\mathbf{i}^{(n)}} \frac{\hat{y}_k}{\pi_k}.$$

Regresinis įvertinys šiuo atveju atrodys

$\hat{t}_{yregI} = \hat{t}_{\bullet y\pi} + \hat{B}\left(t_x - \hat{t}_{x\pi}\right)$, čia

$$\hat{t}_{\bullet y\pi} = \sum_{k\in\mathbf{i}} \frac{y_{\bullet k}}{\pi_k} = \sum_{k\in\mathbf{i}^{(a)}} \frac{y_k}{\pi_k} + \sum_{k\in\mathbf{i}^{(n)}} \frac{\hat{y}_k}{\pi_k}, \quad \hat{t}_{x\pi} = \sum_{k\in\mathbf{i}} \frac{x_k}{\pi_k}.$$

Dabartinėje praktikoje įvertiniai, skaičiuojami taikant įrašymą yra gana paprasti, nes svoriai nesikeičia, tačiau dispersijos įvertinys yra gana sudėtingas ir kiekvienu atveju vis kitoks.

II.5.2.2. Statistinėmis taisyklėmis paremti įrašymo būdai

Praktikoje dažniausiai naudojami šie įrašymo būdai:

- Santykiu pagrįstas įrašymas;
- Daugiamatis regresijos įrašymas;
- Artimiausio kaimyno įrašymas;
- Šiltų duomenų įrašymas;
- Vidurkio įrašymas;

Pirmiems trims būdams reikia papildomos informacijos. Papildomą vektorių nuo šiol vadinsime įrašymo vektoriumi ir žymėsim **Z**. Trečias ir ketvirtas – donorais paremti būdai, t.y. įrašome reikšmes, kurios buvo stebėtos iš tiesų, tačiau ne neatsakiusiems elementams, o kitiems, atsakiusiems elementams. Penktas būdas – determinuotas, ketvirtas – atsitiktinis.

Praktikai praleistus stebėjimus dažniausiai įrašo naudodamiesi *metodų hierarchija*, t.y. iš pradžių reikšmes įrašo naudodamiesi stipresniais metodais, kurie duoda tikslesnes reikšmes, tuomet, jei neturi papildomos informacijos, naudoja kitus – silpnesnius būdus.

Prielaida: visai imčiai **i** naudojame tik vieną įrašymo būdą.

II.5.2.3. Įvertinių skaičiavimas, kai naudojamas įrašymas

II.5.2.3.1. Daugiamatis regresijos įrašymas

Regresijos įrašymas, kurios vienas iš atskirų atvejų yra santykiu pagrįstas įrašymas, yra dažnai naudojamas būdas neatsakymų sukeltam nuokrypiui tyrimuose sumažinti.

Šiuo atveju turime turėti papildomos informacijos vektorių z_k, $k \in \mathbf{i}$. Daroma prielaida, kad papildomos informacijos vektorius ***z*** yra koreliuotas su *y*. Įrašoma *k* elemento reikšmė yra

$\hat{y}_k = \mathrm{z}'_k \hat{\beta}$.

Čia $\hat{\beta}$ yra (y_k, z_k) regresijos koeficientų vektorius, kai $k \in \mathbf{i}^{(a)}$. $\hat{\beta}$ randamas sprendžiant lygtį: $\sum_{k \in \mathbf{i}^{(a)}} q_k (y_k - \mathrm{z}'_k \beta) \mathrm{z}_k = 0$.

$$\hat{\beta} = \left(\sum_{k \in \mathbf{i}^{(a)}} q_k \mathrm{z}_k \mathrm{z}'_k \right)^{-1} \sum_{k \in \mathbf{i}^{(a)}} q_k \mathrm{z}_k y_k .$$

Specialiu atveju, kai $\mathrm{z}_k = (1, z_k)'$, įrašomos reikšmės yra pavidalo $\hat{y}_k = \hat{\alpha} + \hat{\beta} \mathrm{z}_k$, t.y. turi tiesinės regresijos pavidalą.

Determinuoti įrašymo metodai, tokie, kaip daugiamatės regresijos ar santykiu pagrįstas įrašymas, duoda tą pačią reikšmę, jei yra kartojami. Jie gali būti padaryti stochastiniais pridedant *atsitiktinai parinktą liekaną*. Tokios liekanos pridėjimas duomenų seką padaro realesne ir labiau varijuojančia.

Daugiamatės regresijos atveju įrašoma elemento *k* reikšmė su atsitiktinai pasirinkta liekana yra $\hat{y}_k = z'_k \hat{\beta} + e_k^*$, kur

$\hat{\beta} = \left(\sum_{k \in \mathbf{i}^{(a)}} q_k \mathrm{z}_k \mathrm{z}'_k \right)^{-1} \sum_{k \in \mathbf{i}^{(a)}} q_k \mathrm{z}_k y_k$, o e_k^* yra atsitiktinai parinkta liekana iš apskaičiuotos liekanų aibės $\{e_k : k \in \mathbf{i}^{(a)}\}$, čia $e_k = y_k - z'_k \hat{\beta}$.

Atsitiktinai parinktos liekanos gali būti naudojamos įvertinių skaičiavimui, dispersijos skaičiavimui ar abiems. Atsitiktinės liekanos pridėjimas naudojant šį įrašymo metodą yra rekomenduojamas dispersijos įvertinio skaičiavimui, nes kitaip duomenys per mažai varijuoja.

II.5.2.3.2. Santykiu pagrįstas įrašymas

Tai yra atskiras daugiamatės regresijos įrašymo atvejis. Tarkime papildomas vektorius $z_k = z_k$ yra vienmatis kintamasis ir pažymėkime $q_k = 1/z_k$, tuomet daugiamatės regresijos įrašoma kintamojo reikšmė apskaičiuojama $\hat{y}_k = z_k\hat{\beta}$, kur $\hat{\beta} = \frac{\sum_{k\in\mathbf{i}^{(a)}} y_k}{\sum_{k\in\mathbf{i}^{(a)}} z_k}$.

Šis būdas dažniausiai naudojamas, kai tas pats kintamasis yra matuojamas kas kažkiek laiko pakartotinai.

Santykiu pagrįsto įrašymo atveju įrašoma elemento *k* reikšmė su atsitiktinai pasirinkta liekana yra $\hat{y}_k = z_k\,\hat{\beta} + e_k^*$, kur $\hat{\beta} = \frac{\sum_{k\in\mathbf{i}^{(a)}} y_k}{\sum_{k\in\mathbf{i}^{(a)}} z_k}$, o $\{e_k : k \in \mathbf{i}^{(a)}\}$, kur $e_k = y_k - z_k\hat{\beta}$.

Čia taip pat patartina pridėti atsitiktinai parinktą liekaną, tam, kad dispersija būtų įvertinta tiksliau.

II.5.2.3.3. Artimiausio kaimyno įrašymas

Naudojant šį įrašymo būdą, taip pat reikia turėti papildomos informacijos vektorių z_k. Daroma prielaida, kad panašių įmonių, su panašiomis papildomų rodiklių z_k reikšmėmis ir kitų rodiklių reikšmės turėtų būti panašios. Todėl įrašoma reikšmė $\hat{y}_k = y_{l(k)}$, kur *l(k)* yra *elementas donoras* neatsakančiam elementui. Elementas donoras yra surandamas randant mažiausiai nuo įrašomo elemento besiskiriantį elementą, t.y. minimizuojant kokį nors atstumą.

Jei z_k yra vienmatis, tai tinkamiausias elemento *k* donoras *l* surandamas $\min_{l\in\mathbf{i}^{(a)}} D_{lk} = |z_l - z_k|$.

Jei z_k yra daugiamatis, tai artimiausio kaimyno galime ieškoti minimizuojant atstumą $D_{lk} = \left(\sum_{j=1}^{J} h_j\left(z_{jl} - z_{jk}\right)^2\right)^{\frac{1}{2}}$.

II.5.2.3.4. Respondentų vidurkio įrašymas

Kai neturime papildomos informacijos, tai galime į nežinomus rodiklius įrašyti atsakiusių elementų vidurkius, t.y. $\hat{y}_k = \bar{y}_{\mathbf{i}^{(a)}}$, $k \in \mathbf{i}^{(n)}$, čia $\bar{y}_{\mathbf{i}^{(a)}} = \frac{\sum_{k \in \mathbf{i}^{(a)}} y_k}{m}$, jei turime paprastąją imtį, kur m – atsakiusių elementų skaičius.

Kadangi visos įrašytos reikšmės bus vienodos, tai galutinių duomenų skirstinys bus nenatūralus.

II.5.2.3.5. Šiltų duomenų įrašymas

Šiuo atveju tariame, kad irgi neturime papildomos informacijos. Tai donoru paremtas atsitiktinio įrašymo būdas. Įrašyta reikšmė yra $\hat{y}_k = y_{l(k)}$, kur *l(k)* yra atsitiktinai parinktas donoras iš visų galimų elementų donorų $l \in \mathbf{i}^{(a)}$. Šiuo būdu gautos duomenų sekos skirstinys atrodys gan natūraliai, tačiau vis tiek skirsis nuo tikrosios duomenų sekos, nes kiekvienas donoras yra atsakęs elementas, o atsakiusieji ir neatsakiusieji elementai gali reikšmingai skirtis.

II.5.2.4. Įrašymo grupės

Įrašymas dažniausiai atliekamas nesikertančiose *įrašymo grupėse* $\mathbf{i}_g, g = 1,...,G$, t.y. sluoksniuose, kurių sąjunga yra visa imtis i. Kiekvienoje įrašymo grupėje reikšmės yra įrašomos tuo pačiu būdu. $\mathbf{i}$, $\mathbf{i}^{(a)}$ bei $\mathbf{i}^{(n)}$ yra pakeičiami $\mathbf{i}_g, \mathbf{i}_g^{(a)}$ ir $\mathbf{i}_g^{(n)}$.

Pagrindinės skirtingų įrašymo metodų naudojimo grupėse priežastys:

- Skirtingose imties grupėse stebimi skirtingi elementų ryšiai;
- Skirtingoms imties grupėms prieinama skirtinga papildoma informacija.

II.5.2.5. Dispersijos įvertinio skaičiavimas, kai naudojamas įrašymas

Dispersijos vertinimas, kai praleistų stebėjimų vertinimui naudojame įrašymą, yra sudėtinga statistinė problema. Neatsakymai padidina normalią imties dispersiją, ir įrašytas reikšmes naudoti dispersijos skaičiavime nėra gerai:

- Standartinė dispersijos skaičiavimo formulė, pritaikyta pilnai duomenų imčiai duoda paslinktą dispersijos įvertinį;
- Niekaip neįvertinama neatsakymų sukelta dispersiją.

Norint tiksliai įvertinti dispersiją būtina išnagrinėti įrašytų rodiklių statistines charakteristikas (nuokrypį, dispersiją).

$\hat{t}_I$ - tai dominančios sumos įvertinys paskaičiuotas įrašytiems duomenims $\hat{t}_I = \sum_{k \in \mathbf{i}} \frac{y_{\bullet k}}{\pi_k}$

Standartinė $\hat{t}_I$ dispersija yra užrašoma

$V_{pq}\left(\hat{t}_I\right) = E_p E_q \left(\hat{t}_I - t\right)^2$.

Čia E_p ir E_q yra tikėtini vidurkiai, kai turime imties planą *p* ir atsitiktinį atsakymų mechanizmą *q*.

$\hat{t}_I - t = \left(\hat{t} - t\right) + \left(\hat{t}_I - \hat{t}\right)$.

Čia $\left(\hat{t} - t\right)$ yra imties paklaida, o $\left(\hat{t}_I - \hat{t}\right)$ sudaro įrašymo paklaidą, taigi

$V_{pq}\left(\hat{t}_I\right) = E_p E_q \left(\hat{t} - t\right)^2 + E_p E_q \left(\hat{t}_I - \hat{t}\right)^2 + E_p E_q \left(\hat{t} - t\right)\left(\hat{t}_I - \hat{t}\right) = V_{SAM} + V_{IMP} + 2V_{MIX} =$

$= V_{SAM} + V_{NR}$,

čia V_{SAM} yra imties dispersija, o V_{NR} neatsakymų dispersija.

Dispersijos įvertinys $\hat{V}_{pq}\left(\hat{t}_I\right) = \hat{V}_{SAM} + \hat{V}_{NR}$.

Praktikoje labai svarbu įvertinti kiekvieną iš šių dviejų dalių atskirai.

$\hat{V}_{SAM}$ gali būti skaičiuojama dviem būdais:

1. Naudojantis tik prieinama stebėta informacija y_k, $k \in \mathbf{i}^{(a)}$

2. Naudojantis baigtine duomenų seka $y_{\bullet k}$, $k \in \mathbf{i}$, susidedančia iš stebėtų ir įrašytų reikšmių.

Antras būdas yra patrauklesnis, nes įrašytos reikšmės pateiks įdomios papildomos informacijos, lyginant su pirmu atveju. Nors įrašytos reikšmės ir yra netikros, jos dažniausiai yra geresnės, nei duomenų trūkumas. Tačiau $\hat{V}_{SAM}$ apskaičiavimui standartinių (tokių, kaip 100% atsakymų atveju) dispersijų skaičiavimo formulių naudoti nerekomenduojama, nes gautas

dispersijos įvertinys bus mažesnis už tikrąją dispersiją. Literatūra siūlo du šios problemos sprendimo būdus:

1) Naudoti standartinę dispersijos formulę ir pridėti tinkamą korekciją:

$\hat{V}_{SAM} = \hat{V}_{ORD} + \hat{V}_{DIF}$, $\hat{V}_{ORD}$ ir $\hat{V}_{DIF}$ priklauso nuo imties išrinkimo būdo, naudojamo įrašymo metodo.

2) Dispersiją skaičiuoti naudojantis standartinėmis dispersijų skaičiavimo formulėmis, naudojantis duomenų seka

$y_{\bullet k}$, $k \in \mathbf{i}$, kur $y_{\bullet k} = \begin{cases} y_k, k \in \mathbf{i}^{(a)} \\ \hat{y}_k, k \in \mathbf{i}^{(n)} \end{cases}$,

įrašytas reikšmes papildant jau minėtomis atsitiktinai parinktomis liekanomis.

II.5.2.5.1. Neatsakymų dispersijos įvertinio skaičiavimas, kai naudojamas įrašymas

$\hat{V}_{NR}$ formulė priklauso nuo naudojamo įrašymo metodo, nes neatsakiusių elementų skaičius daro įtaką neatsakymų dispersijai: kuo daugiau reikšmių įrašome, tuo dispersija būna didesnė.

$\hat{V}_{NR}$ vertinimui yra naudojami du pagrindiniai metodai:

- Dviejų fazių ėmimo metodas;
- Pasitelkiantis modelį metodas;

Dviejų fazių metodas priklauso nuo dviejų skirstinių: imties skirstinio ir nežinomo atsakymų mechanizmo skirstinio.

Pasitelkiančio modelį metodo požiūris taip pat priklauso nuo dviejų skirstinių: imties skirstinio *p* ir įrašymo modelio skirstinio *q*. Pastarasis yra tiriamo rodiklio *y* ir papildomos informacijos vektoriaus **z** ryšys.

Kiekvienam konkrečiam modeliui $\hat{V}_{NR}$ yra skaičiuojama kitaip.

III. PRAKTINĖ DALIS

III.1. Tyrimo duomenys

Šiame darbe praleistoms reikšmėms įrašyti naudojami duomenys sumodeliuoti pagal tam tikrų ekonominių veiklų įmonių, atsiskaitančių Įmonės veiklos ataskaitą F-01, skirstinį. Šį tyrimą atlieka Statistikos departamento Įmonių statistikos skyrius (toliau – Įmonių statistikos skyrius).

Tyrimo metu apklausiama didžioji dalis šalyje veikiančių nefinansinių įmonių. Šiame tyrime atsiskaitančių įmonių sąrašams sudaryti taikomi ne tikimybiniai imčių sudarymo būdai. Bendras šalies įmonių sąrašas sudaromas iš bendro šalyje veikiančių įmonių sąrašo, įmones išrūšiavus pagal pajamas skirtingose veiklose (naudojamas Ekonominės veiklos rūšių klasifikatoriuje (toliau – EVRK)) trijų ženklų lygyje ir atrinkus 80% pajamų kiekviename sluoksnyje sudarančias įmones. Taigi tyrime dažniausiai atsiskaito didžiosios šalies įmonės, tačiau kai kurios jų, dėl lėto registro atnaujinimo į sąrašą nepatenka, be to kai kurios mažosios įmonės atsiskaito savanoriškai ir jų duomenys yra įtraukiami skaičiuojant rezultatus.

Mažųjų, ataskaitos nepildančių įmonių rodikliai vertinami naudojantis mokesčių inspekcijos pateiktais duomenimis. Mokesčių inspekcija pateikia tik kai kuriuos tyrimui reikalingus rodiklius, taigi trūkstami duomenys yra vertinami naudojantis įvairiais duomenų įrašymo ir vertinimo būdais. Trūkstamiems duomenims įrašyti Įmonių statistikos skyriaus specialistai sukūrė gan sudėtingą programą, tačiau regresiniai duomenų įrašymo metodai joje netaikomi. Taigi šio tyrimo tikslas – panaudoti duomenų įrašymo metodus vieno tyrimo rodiklio įrašymui, palyginti alternatyvių metodų trukumus bei pranašumus, nuspręsti, kuris būdas yra geriausias mus dominančio rodiklio vertinimui.

Tyrime naudojami ataskaitos lentelės „Darbuotojai“ duomenys (1 lentelė).

Vieni iš mokesčių inspekcijos ataskaitos nepildančioms įmonėms, pateikiamų rodiklių yra vidutinis darbuotojų skaičius (100), algos ir atlyginimai (104) bei socialinio draudimo įmokos (105). Kintamieji vidutinis dirbančių ne visą darbo dieną arba savaitę darbuotojų skaičius (101), vidutinis darbuotojų, dirbančių ne visą darbo dieną, perskaičiuotų į dirbančius visą darbo dieną, darbuotojų skaičius (102) bei dirbtų valandų skaičius (žmogaus valandos) (103) ataskaitos nepildančioms įmonėms nėra žinomi.

1 lentelė. F-01 ataskaitos lentelė „Darbuotojai“.

Kintamojo pavadinimas	*Kintamojo pažymėjimas ataskaitoje*	*Kintamojo pažymėjimas darbe*
Vidutinis darbuotojų skaičius	100	$y^{(0)}$
iš jų dirbančių ne visą darbo dieną arba savaitę	101	$y^{(1)}$
darbuotojų, dirbančių ne visą darbo dieną, perskaičiuotų į dirbančius visą darbo dieną, skaičius	102	$y^{(2)}$
Dirbtų valandų skaičius (žmogaus valandos)	103	$y^{(3)}$
Algos ir atlyginimai	104	$y^{(4)}$
Darbdavio socialinio draudimo įmokos	105	$y^{(5)}$

Įrašymui buvo pasirinktas rodiklis, kurį pateikia ataskaitą pildančios įmonės ir, kuris yra nežinomas mažosioms, ataskaitos nepildančioms įmonėms:

- Dirbtų valandų skaičius (žmogaus valandos), kintamasis $y^{(3)}$.

Papildomais kintamaisiais įrašymams pasirinkti:

- Vidutinis darbuotojų skaičius, kintamasis $y^{(0)}$,
- Algos ir atlyginimai, kintamasis $y^{(4)}$.

Kintamieji $y^{(4)}$ ir $y^{(5)}$ yra labai stipriai tarpusavyje susiję, todėl kintamojo $y^{(5)}$ naudojimas papildomos naudos neatneštų. Tyrime šio rodiklio duomenimis nesinaudojame.

Duomenų įrašymui buvo pasirinktos 2 veiklos, kuriose yra didelis įmonių skaičius, (darbe veiklos vadinamos sluoksniais):

- Statybos (EVRK klasifikuojama 45),
- Didmeninė ir mažmeninė prekyba, išskyrus variklinių transporto priemonių pardavimą (EVRK klasifikuojamos 51 ir 52).

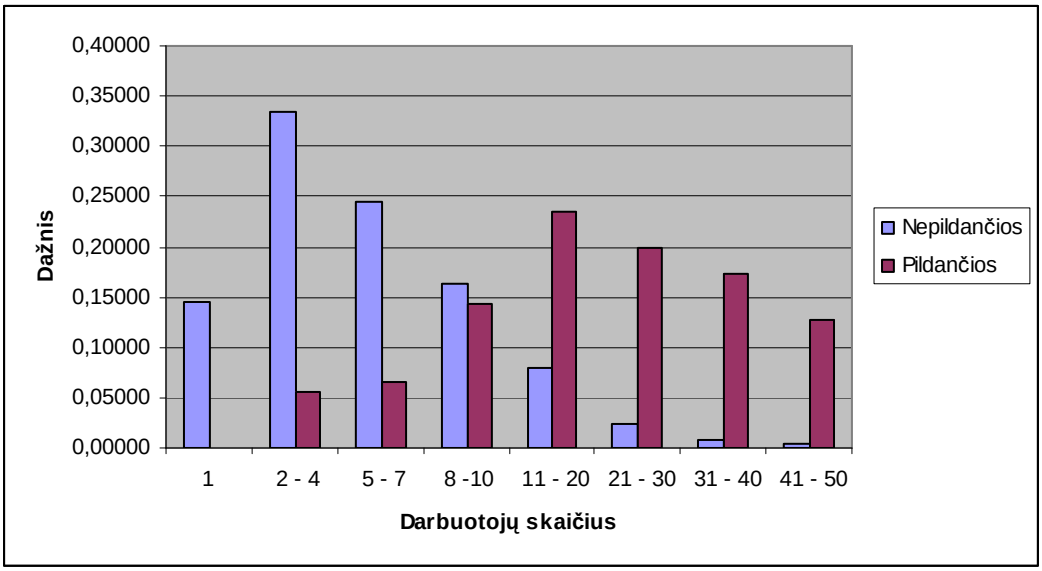
III.2. Turimų duomenų analizė

Aptarsime nagrinėjamus duomenis.

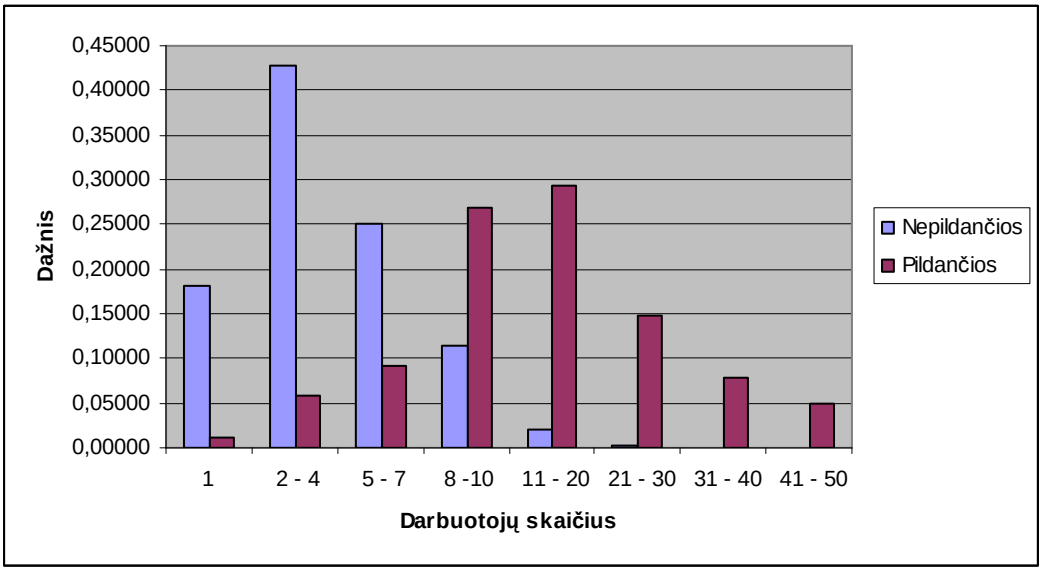
2 lentelė. Įmonių pasiskirstymas pagal darbuotojų skaičių

	Ataskaitos nepildančios įmonės				Ataskaitą pildančios įmonės			
Darbuotojų skaičus	**Statybos**	**Dažnis**	**Prekyba**	**Dažnis**	**Statybos**	**Dažnis**	**Prekyba**	**Dažnis**
1	54	0,04	404	0,07	0	0,00	3	0,00
2 - 4	371	0,28	2853	0,46	10	0,01	44	0,02
5 - 7	272	0,21	1676	0,27	12	0,01	69	0,03
8 -10	182	0,14	765	0,12	26	0,03	202	0,10
11 - 20	294	0,22	425	0,07	142	0,16	731	0,35
21 - 30	85	0,06	63	0,01	121	0,13	368	0,17
31 - 40	27	0,02	11	0,00	105	0,12	198	0,09
41 - 50	13	0,01	6	0,00	77	0,09	121	0,06
> 50	15	0,01	10	0,00	410	0,45	370	0,18
Suma	1313	1	6213	1	903	1,00	2106	1,00

Įmonių pasiskirstymų pagal darbuotojų skaičių palyginimas veiklose (histogramos).



1 pav. Įmonių, kurių darbuotojų skaičius neviršija 50, pasiskirstymas pagal darbuotojų skaičių statybų veikloje.



2 pav. Įmonių, kurių darbuotojų skaičius neviršija 50, pasiskirstymas pagal darbuotojų skaičių prekybos veikloje.

3 lentelė. Įmonių, kurių darbuotojų skaičius neviršija 50, populiacijos kintamųjų pagrindinės statistinės charakteristikos

Sluoksnis *d* = 1 (Veikla = 45)					
	Nepildančios		**Pildančios**		
Įmonių skaičius N_1	1298		493		
Kintamasis	$y^{(0)}$	$y^{(4)}$	$y^{(0)}$	$y^{(4)}$	$y^{(3)}$
Suma $t_1^{(j)}$	12328	83793037	1291 2	143083618	22099947
Vidurkis $\mu_1^{(j)}$	9,5	64555,5	26,2	290230,5	44827,5
Minimumas	1	0	2	6890	960
Maksimumas	50	1261573	50	2058972	102956
Dispersija $s_1^{2(j)}$	68,8	6360899369	153, 9	6178802290 6	596115599
Standartinis nuokrypis $s_1^{(j)}$	8,3	79755,2	12,4	248572	24415,5
Sluoksnis *d* = 2 (Veikla = 51, 52)					
	Nepildančios		**Pildančios**		
Įmonių skaičius N_2	6203		1736		
Kintamasis	$y^{(0)}$	$y^{(4)}$	$y^{(0)}$	$y^{(4)}$	$y^{(3)}$
Suma $t_2^{(j)}$	33116	272557389	3512 6	466563094	61519124
Vidurkis $\mu_2^{(j)}$	5,3	43939,6	20,2	268757,5	35437,3
Minimumas	1	0	1	1116	168
Maksimumas	47	1625650	50	3719905	956997
Dispersija $s_2^{2(j)}$	18	5153859240	122, 8	9711576950 8	1689523126

Standartinis nuokrypis $s_2^{(j)}$	4,2	71790,4	11,1	311634	41103,8

Pažymėjimai:

N_d - įmonių skaičius sluoksnyje (veikloje) d = 1,2.

$y_{d,k}^{(j)}$ - k – toji kintamojo $y^{(j)}$ reikšmė sluoksnyje d = 1,2, j = 1,2,…,5.

$t_d^{(j)}$ - kintamojo $y^{(j)}$ suma sluoksnyje d, $t_d^{(j)} = \sum_{k=1}^{N_d} y_{d,k}^{(j)}$, d = 1,2, j = 1,2,…,5.

$\mu_d^{(j)}$ - kintamojo $y^{(j)}$ vidurkis sluoksnyje d, $\mu_d^{(j)} = \frac{1}{N_d}\sum_{k=1}^{N_d} y_{d,k}^{(j)}$, d = 1,2, j = 1,2,…,5.

$S_d^{2(j)}$ - kintamojo $y^{(j)}$ dispersija sluoksnyje d, $S_d^{2(j)} = \frac{1}{N_d - 1}\sum_{k=1}^{N_d}\left(y_{d,k}^{(j)} - \mu_d^{(j)}\right)^2$, d = 1,2, j = 1,2,…,5.

Kaip matome ataskaitą pildančių ir jos nepildančių įmonių skirstiniai skiriasi, tačiau ataskaitą pildančių įmonių tarpe yra pakankamai mūsų kriterijus atitinkančių, t.y. nedidelių įmonių, todėl tyrimui naudosimės ataskaitą pildančių įmonių duomenimis. Tyrimui pasirinktos įmonės, kurių skaičius neviršija 50 darbuotojų.

III.3. Tyrimo atlikimo schema

Modeliavimą atliekame pagal tokią schemą:

1. Iš ataskaitą pildančių įmonių (darbuotojų skaičius neviršija 50) atsitiktinai išrenkamos įmonės, kurios laikomos neužpildžiusiomis įrašomų rodiklių, pašalinama pasirinkta dalis duomenų;
2. Gautai įmonių imčiai taikomi visi toliau išvardyti duomenų įrašymo metodai;
3. Suskaičiuojamos šiais įrašymo būdais gautos mus dominančios įrašomojo kintamojo statistinės charakteristikos: sumos, absoliutinės bei standartinės santykinės paklaidos, dispersijos ir pasikliautinieji intervalai;
4. Vėl grįžtama į pirmą punktą ir rezultatai skaičiuojami naujai pašalintų duomenų imčiai. Punktai 1 – 3 kartojami pasirinktą skaičių kartų;
5. Skirtingi įrašymo metodai sulyginami tarpusavyje, sulyginami absoliutinių santykinių paklaidų vidurkių bei dispersijų įvertiniai, pateikiamos išvados;

Norint patikrinti kiekvieno įrašymo metodo įrašytų duomenų tikslumą, įrašymo metu gauti rezultatai apibendrinami, skaičiuojamos šios charakteristikos:

- Kintamojo dirbtų valandų skaičius (žmogaus valandos) tikroji suma sluoksniuose $t_d^{(3)}$, d = 1,2;
- Įrašomojo rodiklio dirbtų valandų skaičius (žmogaus valandos) suma sluoksniuose $\hat{t}_d^{(3)}$, d = 1,2,

Čia $\hat{t}_d^{(3)} = \sum_{k=1}^{N_d} y_{d,\bullet k}^{(3)}$, $y_{d,\bullet k}^{(3)} = \begin{cases} y_{d,k}^{(3)}, k \in \mathrm{i}^{(a)} \\ \hat{y}_{d,k}^{(3)}, k \in \mathrm{i}^{(n)} \end{cases}$,

$\mathrm{i}^{(a)}$ - į apklausą atsakiusių elementų aibė,

$\mathrm{i}^{(n)}$ - į apklausą neatsakiusių elementų aibė,

$\hat{y}_{d,k}^{(3)}, k \in \mathrm{i}^{(n)}$ - aprašytais duomenų įrašymo metodais įrašytos nežinomos kintamojo reikšmės.

- Santykinės absoliutinės įrašytų ir tikrųjų kintamųjų sumų paklaidos sluoksniuose:

$$\left| \frac{\sum_{k=1}^{N_d} y_{d,k}^{(3)} - \sum_{k=1}^{N_d} y_{d,\bullet k}^{(3)}}{\sum_{k=1}^{N_d} y_{d,\bullet k}^{(3)}} \right| = \left| \frac{t_d^{(3)} - \hat{t}_d^{(3)}}{\hat{t}_d^{(3)}} \right|, \; d = 1,2.$$

- Standartinės santykinės paklaidos sluoksniuose:

$$\frac{\sqrt{\hat{\mathrm{D}}\hat{t}_d^{(3)} - (\hat{t}_d^{(3)} - t_d^{(3)})^2}}{\hat{t}_d^{(3)}}, \; d = 1,2.$$

- Kintamojo dirbtų valandų skaičius (žmogaus valandos) sumų dispersijų įvertiniai sluoksniuose $\hat{\mathrm{D}}\hat{t}_d^{(3)}$, jų skaičiavimas aptartas III.5. skyrelyje.
- Kintamojo dirbtų valandų skaičius (žmogaus valandos) pasikliautinieji 95% rėžiai skaičiuojami:

$$\left(\hat{t}_d^{(3)} - 1{,}96 * \sqrt{\hat{\mathrm{D}}\hat{t}_d^{(3)}}, \hat{t}_d^{(3)} + 1{,}96 * \sqrt{\hat{\mathrm{D}}\hat{t}_d^{(3)}} \right).$$

III.4. Praleistų reikšmių įrašymo metodų taikymas

Šiame tyrime turime papildomos informacijos visiems įrašomojo kintamojo elementams, todėl praleistų reikšmių įrašymams galime iš karto naudotis stipresniais įrašymo metodais. Taigi respondentų vidurkio bei šiltų duomenų įrašymo metodai netaikomi.

Praleistų reikšmių įrašymams pasirinkti 20% (kitų tyrimų praktikoje gan įprastas) ir tikrasis, pagal turimus duomenis apskaičiuotas (62% pirmame sluoksnyje ir 72% antrame sluoksnyje) neatsakymų lygiai.

SAS programų kodai, naudoti praleistų reikšmių įrašymui pateikti priede nr. 1, bandymų rezultatai – statistinės įrašymų charakteristikos skirtingiems neatsakymų lygiams pateiktos prieduose nr. 2, 3.

III.4.1. Artimiausio kaimyno įrašymo metodas

Duomenis įrašant artimiausio kaimyno metodu tinkamiausiam elemento *k* donorui *l* surasti buvo taikomas atstumas $\min_{l\in\mathbf{i}^{(a)}} D_{lk} = \min\left|y_{d,l}^{(0)} - y_{d,k}^{(0)}\right|$. Buvo sumodeliuoti abu atvejai, vienu jų mažiausiam atstumui surasti naudojamas vidutinis darbuotojų skaičius įmonėje $y^{(0)}$. Šis rodiklis neblogai koreliuoja su mūsų įrašomu rodikliu – dirbtų valandų skaičiumi $y^{(3)}$ (4 lentelė). Kitu – algos ir atlyginimai $y^{(4)}$, šis rodiklis su įrašomu koreliuoja silpniau (5 lentelė).

III.4.3. Santykiu pagrįsto įrašymo metodas

Įrašydami tyrimo duomenis šiuo būdu taip pat naudojamės tik vienu papildomu kintamuoju – vidutiniu darbuotojų skaičiumi $y^{(0)}$.

Santykiu pagrįstu įrašymo būdu įrašomos kintamojo reikšmės yra $\hat{y}_{d,k}^{(3)} = y_{d,k}^{(0)}\hat{\beta}_d$.

Čia $\hat{\beta}_d = \dfrac{\sum_{k\in\mathbf{i}^{(a)}} y_{d,k}^{(3)}}{\sum_{k\in\mathbf{i}^{(a)}} y_{d,k}^{(0)}}$, *d*=1,2.

4 lentelė. Kintamųjų $y^{(0)}$ ir $y^{(3)}$ koreliacijos koeficientai

Statybos	Didmeninė ir mažmeninė prekyba, išskyrus variklinių transporto priemonių pardavimą
$y^{(0)}$ ir $y^{(3)}$ koreliacijos koeficientai	
ρ = 0,892, N_1 = 493	ρ = 0,503, N_2 =1736

III.4.4. Regresinis įrašymo metodas

Regresiniam įrašymui taip pat naudojamas tas pats papildomas kintamasis – vidutinis darbuotojų skaičius $y^{(0)}$. Tokiu būdu įrašytų duomenų statistinės charakteristikos bus lengviau sulyginamos su santykiu pagrįsto metodo duomenų įrašymu.

Įrašomos kintamojo reikšmės apskaičiuojamos:

$$\hat{y}_{d,k}^{(3)} = \hat{\alpha}_d + y_{d,k}^{(0)}\hat{\beta}_d,$$

$$\hat{\beta}_d = \frac{\frac{1}{N_d^{(a)}}\sum_{k\in\mathbf{i}^{(a)}} y_{d,k}^{(0)} y_{d,k}^{(3)} - \frac{1}{N_d^{(a)}}\sum_{k\in\mathbf{i}^{(a)}} y_{d,k}^{(0)} * \frac{1}{N_d^{(a)}}\sum_{k\in\mathbf{i}^{(a)}} y_{d,k}^{(3)}}{\frac{1}{N_d^{(a)}}\sum_{k\in\mathbf{i}^{(a)}} \left(y_{d,k}^{(0)}\right)^2 - \frac{1}{N_d^{(a)}}\left(\sum_{k\in\mathbf{i}^{(a)}} y_{d,k}^{(0)}\right)^2},$$

$\hat{\alpha}_d = \frac{1}{N_d^{(a)}}\sum_{k\in\mathbf{i}^{(a)}} y^{(3)}_{d,k} - \hat{\beta}_d * \frac{1}{N_d^{(a)}}\sum_{k\in\mathbf{i}^{(a)}} y^{(0)}_{d,k}$, $N_d^{(a)}$ - atsakiusių elementų skaičius sluoksnyje, *d*=1,2.

III.4.5. Regresinis įrašymas, kai naudojami du papildomi kintamieji

Įrašant trūkstamus rodiklius šiuo būdu naudojami du papildomi kintamieji – vidutiniu darbuotojų skaičius $y^{(0)}$ bei algos ir atlyginimai $y^{(4)}$.

5 lentelė. Kintamųjų $y^{(4)}$ ir $y^{(3)}$ koreliacijos koeficientai

Statybos	Didmeninė ir mažmeninė prekyba, išskyrus variklinių transporto priemonių pardavimą
$y^{(4)}$ ir $y^{(3)}$ koreliacijos koeficientai	
ρ = 0,628, N_1 = 493	ρ = 0,296, N_2 = 1736

6 lentelė. Kintamųjų $y^{(0)}$ ir $y^{(4)}$ koreliacijos koeficientai

Statybos	Didmeninė ir mažmeninė prekyba, išskyrus variklinių transporto priemonių pardavimą
$y^{(0)}$ ir $y^{(4)}$ koreliacijos koeficientai	
ρ = 0,622, N_1 = 493	ρ = 0,484, N_2 = 1736

Pirmame sluoksnyje antrasis papildomas kintamasis $y^{(4)}$ su įrašomuoju kintamuoju $y^{(3)}$ šiek tiek koreliuoja, antrame sluoksnyje koreliacijos visiškai nėra, be to papildomieji kintamieji tarpusavyje koreliuoja stipriau nei algos ir atlyginimai koreliuoja su įrašomu rodikliu (dirbtų valandų skaičiumi), o tai nėra geros sąlygos šiam įrašymo būdui taikyti. Tačiau šio įrašymo rezultatai vis tiek yra įdomūs.

Įrašomos kintamojo reikšmės apskaičiuojamos:

$\hat{y}_{d,k}^{(3)} = \hat{\alpha}_d + y_{d,k}^{(0)}\hat{\beta}_{d,1} + y_{d,k}^{(4)}\hat{\beta}_{d,2}$, kur $\hat{\alpha}_d$, $\hat{\beta}_{d,1}$, $\hat{\beta}_{d,2}$ - regresijos koeficientų įvertiniai sluoksniuose.

III.4.6. Regresinis įrašymas, kai naudojami trys papildomi kintamieji

Kadangi turimoje duomenų aibėje turime du su įrašomuoju kintamuoju koreliuojančius papildomus kintamuosius, trečią papildomąjį kintamąjį susigeneruojame patys. Kintamasis darbe bus žymimas $y^{(6)}$. Šis įrašymas naudos tikrajam tyrimui neatneš, tačiau yra įdomus analizuojant regresinį įrašymą.

7 lentelė. Kintamųjų $y^{(6)}$ ir $y^{(3)}$ koreliacijos koeficientai

Statybos	Didmeninė ir mažmeninė prekyba, išskyrus variklinių transporto priemonių pardavimą
$y^{(6)}$ ir $y^{(3)}$ koreliacijos koeficientai	
ρ = 0,872, N_1 = 493	ρ = 0,824, N_2 = 1736

Naujas kintamasis $y^{(6)}$ gan neblogai koreliuoja su įrašomu kintamuoju – dirbtų valandų skaičiumi $y^{(3)}$ abiejuose sluoksniuose ir beveik nekoreliuoja su kitais papildomais kintamaisiais – vidutiniu darbuotojų skaičiumi $y^{(0)}$ bei algomis ir atlyginimais $y^{(4)}$, todėl turėtų pagerinti įrašomojo parametro sumos įvertinius.

8 lentelė. Kintamųjų $y^{(6)}$ ir $y^{(0)}$ bei $y^{(4)}$ koreliacijos koeficientai

Statybos	Didmeninė ir mažmeninė prekyba, išskyrus variklinių transporto priemonių pardavimą
$y^{(6)}$ ir $y^{(0)}$ koreliacijos koeficientai	
ρ = 0, 773, N_1 = 493	ρ = 0,472, N_2 = 1736
$y^{(6)}$ ir $y^{(4)}$ koreliacijos koeficientai	
ρ = 0,514, N_1 = 493	ρ = 0,237, N_2 = 1736

Įrašomos kintamojo reikšmės apskaičiuojamos:

$\hat{y}_{d,k}^{(3)} = \hat{\alpha}_d + y_{d,k}^{(0)}\hat{\beta}_{d,1} + y_{d,k}^{(4)}\hat{\beta}_{d,2} + y_{d,k}^{(6)}\hat{\beta}_{d,3}$, kur $\hat{\alpha}_d$, $\hat{\beta}_{d,1}$, $\hat{\beta}_{d,2}$, $\hat{\beta}_{d,3}$ - regresijos koeficientų įvertiniai sluoksniuose.

III.5. Įrašytų duomenų dispersijų vertinimas

Neatsakymų dispersijai įvertinti praktikoje siūlomi skirtingi sprendimai ir nė vienas jų nėra visiškai optimalus. Nauji dispersijos vertinimo būdai, kai yra neatsakymų, gali būti pasiūlyti artimiausiu metu. Šiame darbe siūlomi vienas galimas neatsakymų sukeliamos kintamojo sumos dispersijos įvertinio skaičiavimo būdas.

Įmonių statistikos skyriaus atliekamame tyrime bandome įvertinti bendrą dirbtų valandų skaičiaus sumą. Atsiskaitymų lygis tyrime laikomas 100% t.y. visos pagal tam tikrus kriterijus atrinktos ir ataskaitą F-01 gavusios įmonės atsiskaito, o neatsiskaito tik ataskaitos negavusios įmonės. Taigi, šiuo atveju padarę tokias prielaidas:

1. Tariame, jog turime įvertinti kintamųjų reikšmes įmonėms, kurių darbuotojų skaičius neviršija 50 ir turime tikruosius šių įmonių sąrašus abiejose veiklose, t.y. abiejuose sluoksniuose;
2. Tariame, jog turime paprastąją atsitiktinę sluoksninę imtį, t.y. kiekviename sluoksnyje atskirai išrenkame paprastąją atsitiktinę įmonių imtį ir šioms įmonėms siunčiama ataskaita;
3. Įmonės, kurioms ataskaita nesiunčiama neatsiskaito. Šioms įmonėms dominančius rodiklius įrašome ir bandome įvertinti dominančio kintamojo sumos dispersiją $\mathrm{D}\hat{t}_d^{(3)}$, t.y. suskaičiuoti $\hat{\mathrm{D}}\hat{t}_d^{(3)}$.

Šiuo atveju sumos dispersiją galime skaičiuoti:

$$\hat{\mathrm{D}}\hat{t}_d^{(3)} = N_d^2\left(1-\frac{N_d^{(a)}}{N_d}\right)\frac{\left(\hat{S}_d^{(3)}\right)^2}{N_d^{(a)}},\ \left(\hat{S}_d^{(3)}\right)^2 = \frac{1}{N_d^{(a)}-1}\sum_{k=1}^{N_d^{(a)}}\left(y_{d,k}^{(3)} - \hat{\mu}_d^{(3)}\right)^2 \text{ arba}$$

$\hat{\mathrm{D}}\hat{t}_d^{(3)} = N_d^2\left(1-\frac{N_d^{(a)}}{N_d}\right)\frac{\left(\hat{S}_{\bullet d}^{(3)}\right)^2}{N_d^{(a)}}$, čia

$\left(\hat{S}_{d\bullet}^{(3)}\right)^2 = \frac{1}{N_d-1}\sum_{k=1}^{N_d}\left(y_{d,\bullet k}^{(3)} - \hat{\mu}_{d\bullet}^{(3)}\right)^2$ ir $\hat{\mu}_{d\bullet}^{(3)} = \frac{1}{N_d}\sum_{k=1}^{N_d} y_{d,\bullet k}^{(3)}$, d=1,2.

Deja, praktikoje įmonių, kurioms siunčiama ataskaitą, imtis realiai nėra atsitiktinė, ataskaitas pildo didesnes pajamas gaunančios įmonės, todėl šitokia dispersijos įvertinio skaičiavimo metodika nėra visiškai korektiška. Čia remiamės prielaida, kad įmonės atsako į apklausą su lygiomis tikimybėmis ir nepriklausomai. Darbe buvo skaičiuojami abu dispersijų įvertiniai, pasikliautinieji intervalai abiem dispersijos skaičiavimo būdams visais atvejais padengė tikrąsias sumines kintamojo $y^{(3)}$ reikšmes, taigi toks dispersijos vertinimo būdas gali būti taikomas.

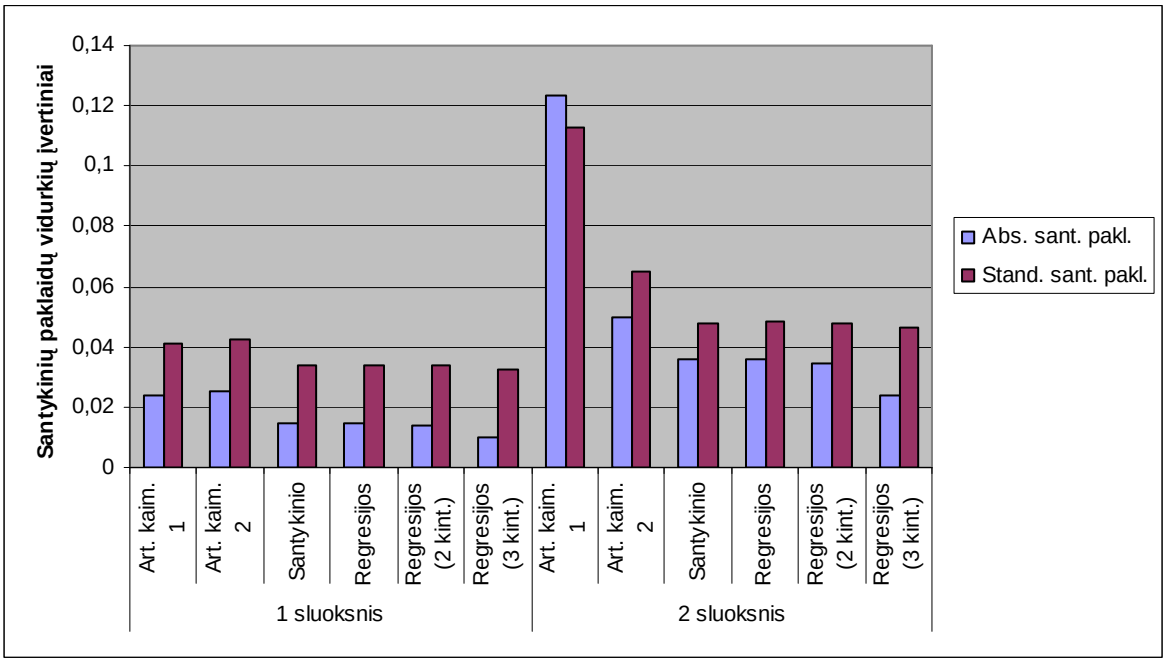
III.6. Gautų rezultatų komentarai

Sulyginus artimiausio kaimyno ir regresinius įrašymo būdus tarpusavyje galima teigti, kad abu įrašymo metodai gan neblogai įvertina praleistus stebėjimus. Tai yra logiška, nes turime neblogą papildomąjį kintamąjį – vidutinį darbuotojų skaičių ir ryšys tarp įrašomojo ir papildomųjų kintamųjų yra tiesinis.

Apskaičiuotosios santykinės paklaidos pirmajame sluoksnyje mažesnės nei antrajame sluoksnyje abiems neatsakymų lygiams. Taip yra todėl, kad antrajame sluoksnyje turime didesnį įmonių skaičių ir tuo pačiu, atlikdami bandymus pašaliname daugiau įrašomojo kintamojo duomenų (3 pav.).

Įrašymus pakartojus 10 kartų 20% neatsakymų lygiui ir 20 kartų tikrajam neatsakymų lygiui bei sulyginus santykinių paklaidų statistines charakteristikas pastebime, jog regresiniai praleistų reikšmių įrašymo metodai yra geresni už artimiausio kaimyno įrašymo metodus. Gauti absoliutinių bei standartinių paklaidų vidurkių ir dispersijų įvertiniai mažesni regresiniams duomenų įrašymo būdams (2, 3 priedai). Standartinėms paklaidoms skirtingiems įrašymams sulyginti buvo panaudotas t-kriterijus (t.y. tikrinamos hipotezės, kad skirtumų tarp skirtingų įrašymo metodų nėra), skirtumai tarp regresinių bei artimiausio kaimyno įrašymo metodų akivaizdūs didesniam neatsakymų lygiui ir didesniam bandymų skaičiui. 20% neatsakymų lygiui hipotezės apie standartinių paklaidų lygybę skirtingiems įrašymo metodams, su reikšmingumo lygmeniu $\alpha = 0{,}05$, nėra atmetamos. Didesniam nei 50% neatsakymo lygiui regresiniai įrašymo

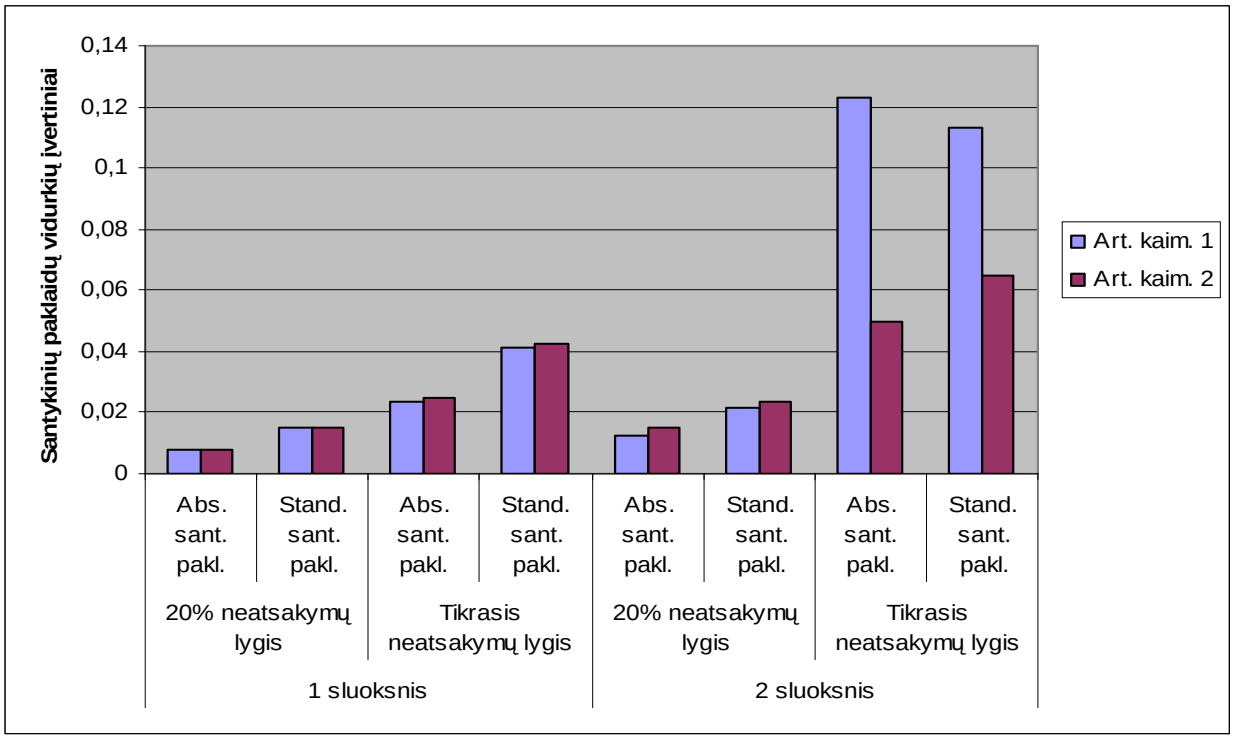
metodai pranoko artimiausio kaimyno įrašymo metodus abiejuose sluoksniuose, t.y. gautos *p-reikšmės* buvo mažesnės už mūsų pasirinktą reikšmingumo lygmenį.



3 pav. Standartinių ir absoliutinių santykinių paklaidų vidurkiai sluoksniuose tikrajam neatsakymų lygiui, atlikus 20 bandymų.

Palyginus artimiausio kaimyno įrašymo metodus paaiškėjo, kad abu metodai dirbtų valandų skaičių vertina labai panašiai, ryškesnis skirtumas pastebėtas antrajame sluoksnyje (prekybos veikloje) didesniam neatsakymų lygiui, todėl šiame sluoksnyje duomenis įrašant artimiausio kaimyno metodu elementui – donorui išrinkti geriau būtų naudoti algas ir atlyginimus, o ne vidutinį darbuotojų skaičių (4 pav., 2 priedas), *p-reikšmės* hipotezėms apie absoliutinių ir standartinių paklaidų lygybę buvo 0.049 absoliutinei ir 0.025 – santykinei standartinei paklaidai. Taigi skirtinguose sluoksniuose duomenų struktūra yra skirtinga ir skirtingose veiklose duomenis įrašinėti reikia atsižvelgiant į duomenų skirstinį šiose veiklose.

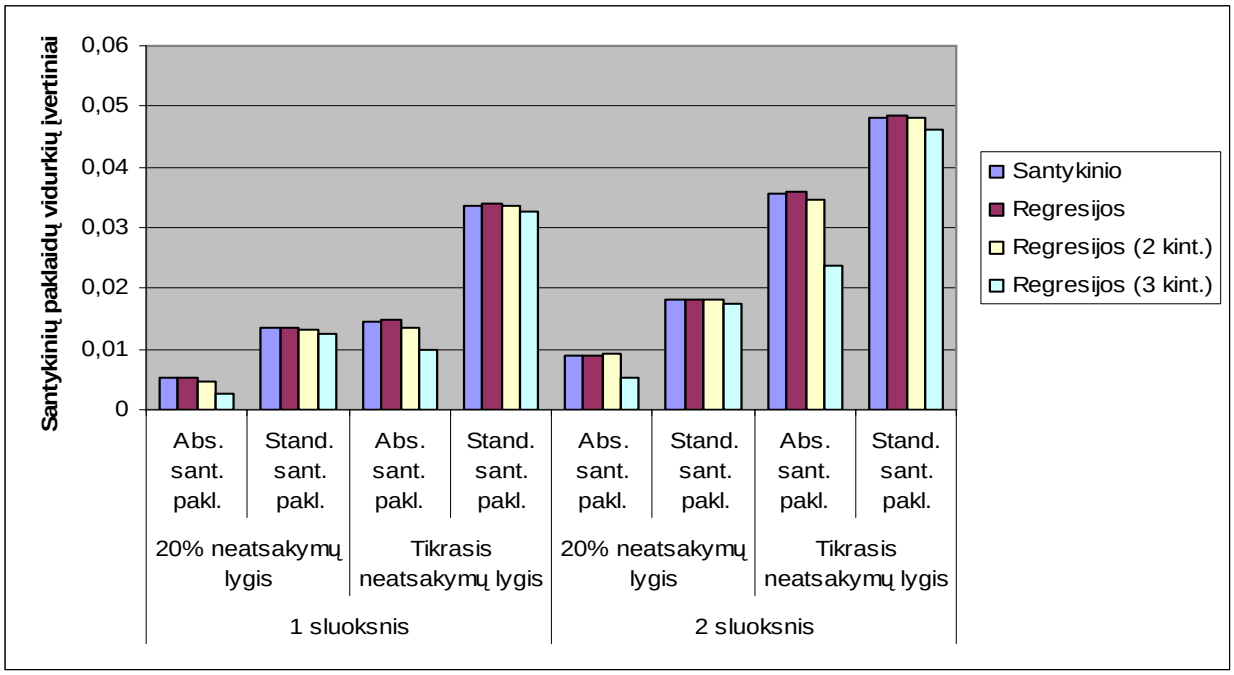
Sulyginus regresinius duomenų įrašymo metodus pastebėta, kad skirtingi regresiniai duomenų įrašymo būdai pateikia labai panašius apibendrintus rezultatus (absoliutines santykines paklaidas bei dispersijas). Tai galima paaiškinti tuo, kad ryšiai tarp įrašomojo ir papildomųjų kintamųjų yra tiesiniai, be to regresinė tiesė eina per koordinačių pradžią.



4 pav. Standartinių ir absoliutinių paklaidų vidurkių bei standartinių nuokrypių įvertiniai artimiausio kaimyno įrašymo metodams visiems neatsakymų lygiams.

Remiantis bandymų rezultatais galima teigti, jog duomenims įrašyti tikrajame tyrime geriau taikyti santykinį arba paprastą regresinį įrašymo būdą, nes regresinis ryšys tarp dirbtų valandų skaičiaus $y^{(3)}$ bei algų ir atlyginimų $y^{(4)}$ pirmame sluoksnyje yra labai silpnas, o antrame sluoksnyje jo visiškai neturime. Taigi regresinis metodas, nuo dviejų papildomųjų kintamųjų ($y^{(0)}$, $y^{(4)}$) nėra tinkamas dirbtų valandų skaičiaus įrašymui nei santykinis ar paprastas regresinis įrašymas.

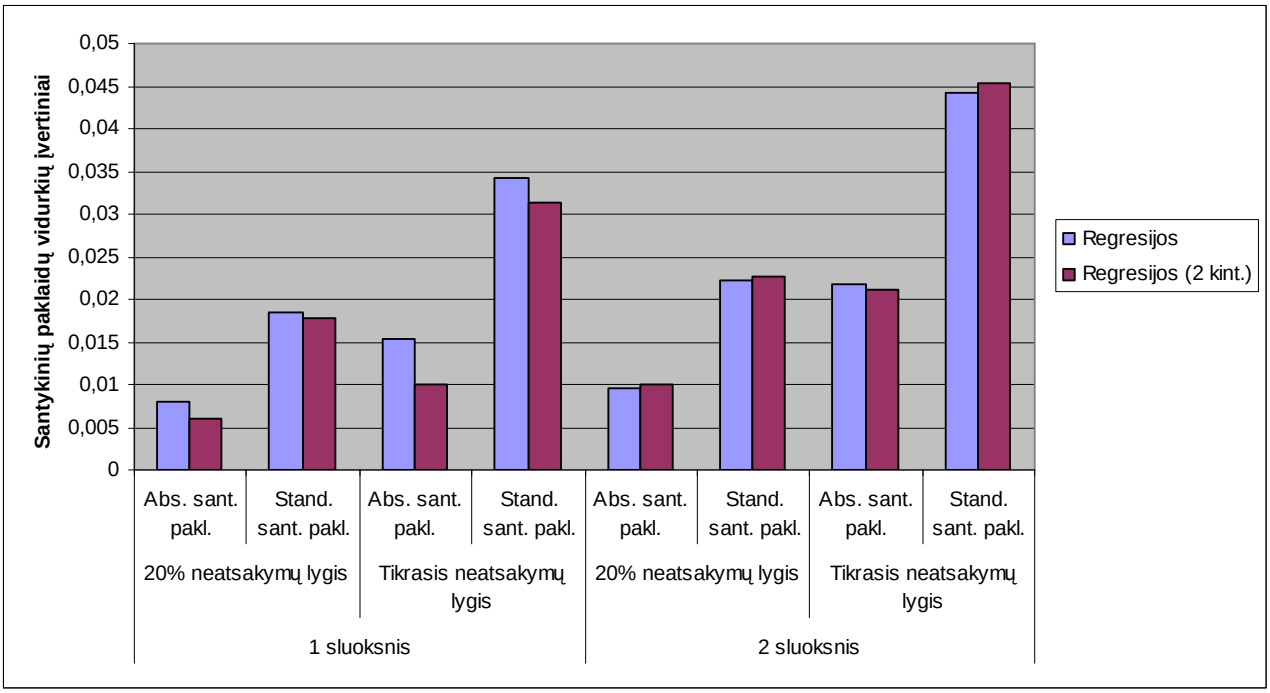
Pridėjus trečią – gan stipriai koreliuotą sumodeliuotą kintamąjį $y^{(6)}$ pastebime, kad santykinės paklaidos sumažėjo abiejuose sluoksniuose abiems neatsakymų lygiams (5 pav.), tačiau patikrinus ar šis įrašymo metodas yra geresnis už mažiau papildomųjų kintamųjų naudojančius metodus t-testu su reikšmingumo lygmeniu $\alpha = 0.05$ galime teigti, jog šis metodas paprastus regresinius metodus pranoksta tik antrajame sluoksnyje didesniam neatsakymų lygiui. Taip atsitiko todėl, kad tyrime turimi papildomieji kintamieji antrame sluoksnyje su įrašomuoju kintamuoju koreliuoja prastai, o sumodeliuotasis kintamasis koreliuoja gan stipriai.



5 pav. Standartinių ir absoliutinių paklaidų vidurkių įvertiniai regresiniams įrašymo metodams visiems neatsakymų lygiams.

Siekiant ištirti ar stiprios koreliacijos atveju daugiamatis regresinis duomenų įrašymo metodas padeda tiksliau įvertinti standartinius įvertinius buvo atlikta smulkesnė regresinio įrašymo analizė. Duomenys į dominantį kintamąjį $y^{(3)}$ buvo įrašyti regresijos metodu naudojantis sugeneruotu papildomu kintamuoju $y^{(6)}$ ir naudojantis dviem stipriau koreliuotais papildomais kintamaisiais - $y^{(6)}$ir $y^{(0)}$, 20% ir tikrajam neatsakymų lygiams. Kad rezultatai būtų tikslesni abiems neatsakymų lygiams buvo atlikta po 30 bandymų (6. pav.).

Iš 6 pav. galima pastebėti, kad didesnių skirtumų tarp šių įrašymo metodų neturime abiems neatsakymų lygiams. T-testas parodė, kad regresinis duomenų įrašymas, kai naudojamas didesnis papildomųjų kintamųjų skaičius yra geresnis tik tuomet, kai turimas didelis neatsakymų lygis ir abu papildomi kintamieji stipriai koreliuoja su įrašomuoju (1 sluoksnis), jei bent vienas kintamasis su įrašomuoju kintamuoju koreliuoja prastai – gaunami prastesni įvertiniai, taip ir atsitinka antrajame sluoksnyje, kur koreliacijos koeficientas tarp $y^{(3)}$ ir $y^{(0)}$ kintamųjų yra lygus 0,503.



6 pav. Standartinių ir absoliutinių paklaidų vidurkių įvertiniai regresiniams įrašymo metodams, tiesinei regresijai $\hat{y}_{d,k}^{(3)} = \hat{\alpha}_d + \hat{\beta}_d * y_{d,k}^{(6)}$ *ir regresijai nuo dviejų kintamųjų* $\hat{y}_{d,k}^{(3)} = \hat{\alpha}_d + \hat{\beta}_{d,1} * y_{d,k}^{(6)} + \hat{\beta}_{d,2} * y_{d,k}^{(0)}$*, 20% ir tikriems neatsakymų lygiams, atlikus 30 bandymų.*

III.7. Praleistų reikšmių įrašymas į tikrąją duomenų aibę

Įrašant duomenis tikriesiems tyrimo praleistiems stebėjimams tikimės, kad ataskaitos nepildančioms įmonėms galioja tie patys dėsniai, kaip ir ataskaitą pildančioms įmonėms. Iš tiesų koreliacijos koeficientai tarp papildomųjų kintamųjų $y^{(0)}$ bei $y^{(4)}$ ataskaitos nepildančioms įmonėms yra labai panašūs į ataskaitą pildančių įmonių koreliacijos koeficientus, taigi natūralu tikėtis, kad panašūs ryšiai egzistuoja ir tarp įrašomojo kintamojo $y^{(3)}$ bei papildomųjų kintamųjų $y^{(0)}$ ir $y^{(4)}$.

9 lentelė. Kintamųjų $y^{(0)}$ *ir* $y^{(4)}$ *koreliacijos koeficientai ataskaitos nepildančioms įmonėms*

Statybos	Didmeninė ir mažmeninė prekyba, išskyrus variklinių transporto priemonių pardavimą
$y^{(0)}$ ir $y^{(4)}$ koreliacijos koeficientai	
ρ = 0,749, N_1 = 1298	ρ = 0,566, N_2 = 6203

Kadangi papildomasis kintamasis $y^{(6)}$ yra sugeneruotas ir tikrajame tyrime jo neturime, tai tikrajame tyrime jo naudoti negalime. Šiuo atveju duomenis įmonėms, kuriose dirba mažiau nei 50 darbuotojų geriausia įrašinėti santykio metodu ar paprastos regresijos būdu, papildomuoju kintamuoju naudojant $y^{(0)}$ abiejuose sluoksniuose. Įmonėms, kurių darbuotojų skaičius viršija 50, duomenis įrašinėti galima artimiausio kaimyno metodu. Papildomojo kintamojo pasirinkimas šiuo atveju turėtų priklausyti nuo įmonių pasiskirstymo veikloje.

Šiame tyrime naudota praleistų reikšmių įrašymo programa artimiausiu metu bus pakoreguota ir pritaikyta duomenims į tikrąją praleistų stebėjimų aibę įrašyti.

IV. IŠVADOS

Praleistų reikšmių įrašymo efektyvumui patikrinti atliktas tyrimas. Tyrimui naudoti duomenys sumodeliuoti pagal tam tikrų ekonominių veiklų įmonių skirstinį, naudojantis vieno Statistikos departamente atliekamo tyrimo duomenimis. Tyrimo metu buvo bandoma išrinkti geriausią praleistų reikšmių įrašymo metodą įmonių darbuotojų bendram dirbtų valandų skaičiui įvertinti, kai žinomi vidutiniai įmonių darbuotojų skaičiai bei darbuotojams išmokėtos algos ir atlyginimai. Praleistų reikšmių įrašymas buvo atliekamas dviejuose sluoksniuose, t.y. dviejų ekonominių veiklų įmonėms. Programa duomenų įrašymui buvo parašyta Statistical Analysis System (SAS) paketu.

Remiantis tyrimo rezultatais galima teigti, kad skirtingiems duomenims įrašyti turi būti taikomi skirtingi įrašymo būdai. Įrašymo metodo pasirinkimas priklauso nuo duomenų struktūros, turimos papildomos informacijos, ryšių tarp įrašomųjų ir papildomųjų kintamųjų, ir nuo neatsakymų lygio.

Tyrimo neatsakymų lygis yra labai svarbus renkant geriausią praleistų reikšmių įrašymo metodą. Jei neatsakymų lygis yra nedidelis, tai skirtumai tarp praleistų reikšmių įrašymo būdų nėra dideli ir išrinkti geriausiai praleistus stebėjimus įvertinantį metodą yra sudėtinga. Vienų metodų pranašumas prieš kitus išryškėja tik turint didelį neatsakymų lygį.

Turima papildoma informacija yra labai efektyvus neatsakymų nuokrypį mažinantis veiksnys. Įrašymai naudojantys papildomą informaciją yra geresni už determinuotus ar atsitiktinius įrašymo metodus, ypač, jei populiacijos elementai nėra vienalyčiai.

Artimiausio kaimyno įrašymo metodas gali būti taikomas, kai tarp įrašomojo ir papildomųjų kintamųjų egzistuoja tam tikras ryšys. Jei toks ryšys neegzistuoja, šiuo būdu įrašomos reikšmės nėra tinkamos įrašymui ir gaunamos gan didelės standartinių įvertinių santykinės paklaidos.

Jei Statistikos departamente atliekamo tyrimo duomenims įrašyti būtų nuspręsta taikyti artimiausio kaimyno įrašymą, tai papildomu kintamuoju vienoje veikloje (pirmajame sluoksnyje) turėtų būti naudojamas vidutinis darbuotojų skaičius, o kitoje (antrajame sluoksnyje) algos bei atlyginimai. Šis tyrimo rezultatas patvirtina, kad geriausias praleistų reikšmių įrašymo metodas turi būti kruopščiai parinktas, atsižvelgiant į duomenų struktūrą ir ypatybes, šiuo atveju įmonės veiklą.

Tyrimo metu paaiškėjo, kad regresiniai praleistų reikšmių įrašymo metodai yra efektyvūs, kai turime koreliuotus papildomuosius kintamuosius. Daugiamatės regresijos įrašymo metodas, kai naudojame didesnį papildomųjų kintamųjų skaičių, geresnis nei paprastos regresijos

įrašymas tik tuomet, kai abu kintamieji gana stipriai koreliuoja su įrašomuoju. Jei nors vienas papildomas kintamasis su įrašomuoju koreliuoja silpnai – gaunami prastesni standartiniai įvertiniai. Tokių kintamųjų praleistų reikšmių įrašymui geriau nenaudoti.

Iš eksperimentiniame tyrime naudotų įrašymo metodų tikrajam tyrimui siūloma naudoti santykio arba paprastos tiesinės regresijos įrašymo metodą. Turimam duomenų skirstiniui abiem įrašymais gauti rezultatai buvo labai panašūs ir neprastesni nei rezultatai, gauti duomenis įrašant daugiamačiu regresiniu būdu. Artimiausiu metu šio darbo rekomendacijas planuojama panaudoti duomenims į tikrąją praleistų duomenų aibę įrašyti.

V. SUMMARY

Nonresponse has been a matter of concern for several decades in survey theory and practice. The problem can be viewed from two different angles: the prevention or avoidance of nonresponse before it occurs, and the special estimation techniques when nonresponse has occurred. These problems are examined in this work called „Missing data imputation in finite population statistics“.

The objective of this work is to describe main methods of estimation when nonresponse occurs. Special attention is drawn on one nonresponse estimation method – imputation.

Imputation is the procedure when missing values for one or more study variables are “filled in” with substitutes constructed according to some rules, or observed values for elements other than nonrespondents.

In this work imputation methods based on some of the more commonly used statistical rules are considered. Some of them are tested on data set having the same distribution as the data of the real survey taken in Statistics Lithuania. The imputation methods are compared with each other and the best imputation method for this data set is picked up. Special attention is paid on regression imputation.

Data imputation was made with Statistical Analysis System SAS, SAS code for data imputation is given in appendix 1 and in CD enclosed.

VI. LITERATŪRA

- Sixten Lundstrom, Carl-Erik Sarndal, Estimation in the presence of Nonresponse and Frame Imperfections, Statistics Sweden 2002.
- Wiliam G. Cochran, Sampling Techniques, 1957.
- Jean-Francois Beaumont, Statistics Kanada, Household Survey Methods Division publication On regression imputation in the presence of nonignorable nonresponce, Statistics Kanada, Ottawa (Ontario), Canada K1A 0T6.
- John Wiley & Sons, Practical Methods for Design and Analysis of Complex Surveys, 2003.
- Julius Kruopis, Matematinė statistika, Mokslo ir enciklopedijų leidykla, Vilnius 1993.
- Paul A. Herzberg, How SAS Works.
- Internetiniai puslapiai:

  http://v8doc.sas.com

  http://www.stat.jyu.fi/mpss/VLISS/index.php

VII. PRIEDAI

- Duomenų įrašymui naudotas SAS kodas, atspausdintas variantas (priedas nr. 1).
- 10 bandymų $y^{(3)}$ kintamąjam įrašyti absoliutinių ir standartinių paklaidų pagrindinės statistinės charakteristikos turint 20% neatsakymų lygį (priedas nr. 2).
- 20 bandymų $y^{(3)}$ kintamąjam įrašyti absoliutinių ir standartinių paklaidų pagrindinės statistinės charakteristikos turint tikruosius neatsakymų lygius (priedas nr. 3).
- Kompaktinis diskas su informacija (priedas nr. 4):
    - Magistrinio medžiaga, Word'inis failas.
    - Duomenų įrašymui naudotas SAS kodas.