VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

STATISTIKOS BAKALAURO STUDIJŲ PROGRAMA

Šarūnas Bagdonas

Jungtinių Amerikos Valstijų namų ūkių išlaidų transportui tyrimas

Analysis of US household expenditure on transportation

Bakalauro baigiamasis darbas

Darbo vadovas dr. prof. Vytautas Kazakevičius

Vilnius
2020

Jungtinių Amerikos Valstijų namų ūkių išlaidų transportui tyrimas

Santrauka

Šiame darbe nagrinėjame pajamų dydžio, demografinių bei kitų kintamųjų įtaką su transportu susijusioms išlaidoms. Šiems sąryšiams nagrinėti naudojamos, laisvai prieinamos, JAV vartotojų išlaidų apklausos. Taip pat šiam tikslui pristatomi bei pritaikomi trys statistiniai modeliai: tiesinės regresijos, Tobit bei Heckman. Šie modeliai palyginami tarpusavyje ir pateikiamos jų statistinės išvados.

Raktiniai žodžiai : Tobit, Heckman modelis, išlaidos, tiesinė regresija, transporto išlaidos

Analysis of US household expenditure on transportation

Abstract

In this paper we analyse transportation expenditure dependence on income, demographic and other variables. To achieve this goal Consumer Expenditure Survey data is used. We present and fit three different models: linear regression, Tobit and Heckman two-step procedure. These models are compared, and their statistical insights are presented.

Key words : Tobit, Heckman model, expenditure, linear regression, transportation expenditure

Turinys

1 Įžanga

Išlaidos transportui yra viena didžiausių išlaidų kategorijų, sudaranti apie 15% visų namų ūkių išlaidų. Ją lenkia tik išlaidos būstui. Lyginant skirtingus namų ūkius, galima pastebėti, kad šios išlaidos ženkliai skiriasi – kai kurie namų ūkiai, transportui skiria labai nedidelę (arba netgi nulinę) dalį savo pajamų, o kiti išleidžia labai daug. Natūraliai kyla klausimas, kaip paaiškinti tokį išsiskyrimą ir kokie namų ūkio požymiai galėtų indikuoti didesnes ar mažesnes išlaidas transportui. Pavyzdžiui logiška, kad namų ūkiai turintys daugiau transporto priemonių, šiai išlaidų kategorijai paskirs daugiau pinigų, nei turintys mažiau. Bet ar turi šioms išlaidoms įtakos namų ūkio vietovė, pajamų dydis, šeimos dydis, amžius, sudėtis ir panašūs kintamieji? Tokie ir panašūs klausimai pastaraisiais metais literatūroje susilaukė nemažai dėmesio [3, 16, 10, 17]. Šiame darbe, pasinaudodami laisvai prieinamais JAV vartotojų išlaidų apklausos (*angl. Consumer Expenditure Survey*) duomenimis, į tokį klausimą ir bandysime atsakyti. Šiam tikslui pasitelksime tiesinės regresijos, bei regresijos cenzūruotiems duomenims modelius.

Iš pradžių, antroje dalyje, trumpai pristatome duomenų šaltinį bei duomenų surinkimo ir apdorojimo metodus. Tada, trečioje dalyje, pristatome taikomus statistinius modelius. Ketvirtoje dalyje nagrinėjame duomenis, aptariame jų paruošimą analizei ir suskaičiuojame tiriamų kintamųjų skaitines charakteristikas. Penktojoje dalyje pateikiame taikytų modelių parametrų įverčius bei statistines išvadas. Galiausiai, šeštame skyriuje, pateikiame šio darbo išvadas bei kryptis tolimesniems tyrimams.

2 Duomenų šaltinis

Norint atsakyti į išsikeltus klausimus pasirinktas gerai žinomas ir plačiai literatūroje nagrinėjamas duomenų šaltinis - JAV vartotojų išlaidų apklausa. Vartotojų išlaidų apklausa (*angl. Consumer Expenditure Survey*) atliekama JAV darbo statistikos biuro (*angl. U.S. Bureau of Labor Statistics (BLS)*). Šio tyrimo pagrindinis tikslas - apskaičiuoti vartotojų kainų indeksui (*angl. Consumer Price Index (CPI)*) reikiamo prekių krepšelio svorius.

Pirmoji tokia apklausa buvo atlikta 1888 - 1891 m., kai buvo apklausta vos 3000 dirbančių žmonių, o apklausa buvo vykdoma siekiant ištirti dirbančiųjų išlaidų įvairioms prekių kategorijoms sąryšį su tų prekių gamybos kaštais. Antroji apklausa buvo atlikta 1901 m., siekiant geriau įvertinti maisto produktų kainų infliaciją. Trečioji apklausa buvo atlikta 1917 - 1919 m., siekiant pateikti svorius pragyvenimo išlaidų indeksui (*angl. cost-of-living index*), dabar žinomam kaip vartotojų kainų indeksas. Ketvirtoji ir penktoji apklausa buvo atliktos Didžiosios depresijos metu. Jų tikslas buvo ištirti žmonių dirbančių miestuose arba biuruose gaunamas pajamas, siekiant peržiūrėti socialines ir ekonomines sąlygas. Šeštoji apklausa atlikta 1950 m., kurios metu buvo tiriami tik miestų gyventojai. Septintoji apklausa atlikta 1960 - 1961 m.. Tada buvo tiriami miestų ir kaimų gyventojai, o medžiaga pateikta platesnei ekonominei, socialinei ir rinkos analizei. Aštunta apklausa buvo atlikta 1972 - 1973 m. Tuo metu įvesti keli svarbūs pakeitimai. Pirmas pakeitimas - dienoraščio tipo tyrimo įvedimas. Dienoraštyje respondentai apskaito dalies prekių kategorijų savo namų ūkio kasdienines išlaidas. Antras pakeitimas - perėjimas nuo interviu vykdomo kas metus prie interviu vykdomo kas ketvirtį. Dabartinė apklausos versija pradėta vykdyti 1980 m., kai ji pradėta vykdyti testiniu būdu su rotacijos principu atnaujinama namų ūkių imtimi. Be didesnių pasikeitimų tokio tipo apklausa vykdoma iki šių dienų. Tokia ilga šios apklausos istorija leidžia vykdyti daug įdomių tyrimų susijusių su vartotojų išlaidų pokyčiais.

Apklausos duomenų detalumas išties įspūdingas. Pateikiama ne tik stambių vartojimo kategorijų (pvz. apranga, maistas, transportas ar sveikatos apsauga) praėjusio ketvirčio išlaidų duomenys, bet ir kiekvienos iš stambiųjų kategorijų išskaidymas į labai smulkias. Netgi iki tokio detalumo kaip pavyzdžiui "Automobilių degalų bakų remontas ir keitimas". Iš viso išlaidos suskirstytos į daugiau nei 1200 skirtingų kategorijų. Apklausos duomenyse pateikiama, ne tik vartojimo duomenys. Taip pat ten galime rasti detalius duomenis apie pajamas iš skirtingų šaltinių, duomenis apie turimą turtą bei demografinius rodiklius.

2.1 Apklausos programa

Kaip minėjome anksčiau vartotojų išlaidų apklausa susideda iš dviejų dalių: apklausos už praėjusio ketvirčio išlaidas, vadinamos interviu ir dienoraščio tipo apklausos. Interviu apklausoje apklausiamos išlaidos susijusio su stambiomis ir retai pasitaikančiomis išlaidų kategorijomis. Šios išlaidos sudaro apie du trečdalius visų išlaidų (2018 metų duomenimis). Apklausai išrenkama apie 12000 namų ūkių, iš kurių sulaukiama apie 6900 atsakymų. Kiekvienas išrinktas namų ūkis apklausiamas keturis ketvirčius iš eilės, tačiau namų ūkių imtis keičiasi kiekvieną ketvirtį, nes ketvirtadalis namų ūkių yra apklausiami pirmą kartą. Kita, nesusijusi su pirmąja, apklausos dalis - dienoraščio pildymas. Apklausai atrenkama apie 12000 namų ūkių, iš kurių sulaukiama apie 6900 atsakymų. Respondentai dvi savaites iš eilės pildo dienraštį apie savo išlaidas. Dviejų savaičių laikotarpiai yra išskirstomi per metus tolygiai. Šiame dienoraštyje prašoma pildyti, tik smulkesnių ir/ar dažniau pasikartojančių pirkinių duomenis, pavyzdžiui išlaidas įvairiems maisto produktams. Dienoraštis apima apie trečdalį visų išlaidų (2018 m. duomenimis).

Kadangi interviu ir dienoraščio tipo apklausoms namų ūkiai atrenkami nepriklausomai, demografinių ar kitų duomenų įtakos bendram suvartojimui tyrimas tampa problemiškas. Laimei didžioji dalis su transportu susijusių duomenų yra gaunama vykdant interviu tipo apklausą. Į ją patenka tokios stambios transporto išlaidų kategorijos kaip išlaidos degalams, naujų/padėvėtų automobilių pirkimo ir nuomos kaštai bei viešasis transportas. Į dienoraščio apklausą patenka tik kelios smulkios kategorijos, pavyzdžiui "Automobilių plovimo paslaugos". Detalus transporto išlaidų kategorijų sąrašas pateikiamas priede 2 lentelėje. Šiame darbe mes apsiribojame, tik transporto išlaidų duomenimis, kurie yra gaunami interviu metu.

2.2 Imties sudarymas

Imtis sudarome taip, kad reprezentuotų visą Jungtinių Amerikos Valstijų namų ūkių populiaciją. Į ją patenka žmonės gyvenantys nuosavuose namuose, kotedžuose, butuose ir bendruomeniniuose namuose, pavyzdžiui studentų bendrabučiuose. Į apklausą nepatenka kariuomenės personalas gyvenantis kitose šalyse arba karinėse bazėse, taip pat slaugos namų gyventojai, bei žmonės apklausos metu esantys kalėjimuose. Apklausos populiaciją sudaro daugiau nei 98% visos JAV populiacijos.

2.3 Imties elementų svorių nustatymas

Kiekvienas į apklausą įtrauktas namų ūkis reprezentuoja tam tikrą dalį JAV populiacijos, kas yra laikoma visos apklausos populiacija. Ta dalis nusakoma įtraukiant svorius. Svoriai skaičiuojami siekiant patikslinti imties elemento indėlį, taip, kad jie būtų atvirkščiai proporcingi patekimo į imtį tikimybei. Taip pat svoriai yra kalibruojami atsižvelgiant į neatsakymų į apklausą dažnį bei iš kitų šaltinių žinomas populiacijos charakteristikas.

2.4 Duomenų cenzūravimas

Kai kurie apklausoje pateikiami duomenys galimai leistų identifikuoti konkrečių respondentų tapatybę. Siekiant to išvengti, kai kurie duomenys yra pakeičiami (cenzūruojami). Tai daroma laikantis principo, kad viešai prieinami duomenys neleistų identifikuoti nei vieno apklausos respondento. Duomenims taikomi du cenzūravimo tipai: viršutinis cenzūravimas (*angl. topcoding*), jei imties elemento charakteristikos reikšmė viršija nustatytą slenkstį, ir apatinis cenzūravimas (*angl. bottom coding*), jei reikšmė yra mažesnė nei nurodytas slenkstis. Cenzūruojami laukai pakeičiami visų tos charakteristikos cenzūruotų reikšmių vidurkiu. Kiekvienas pateikiamas laukas turi indikatorių nurodantį ar tas laukas buvo cenzūruotas ar ne.

Daugiau apie apklausos sudarymo principus ir duomenų apdorojimą galima rasti JAV darbo statistikos biuro vadove [12].

3 Statistiniai metodai

3.1 Tobit modelis cenzūruotiems duomenims

Tobit modeliais vadinama plati klasė regresijos modelių kurių stebimas priklausomas kintamasis yra kokia nors forma cenzūruotas. Pirmą kartą tokį modelį pasiūlė James Tobin. Straipsnyje [18] jis šį modelį pasiūlė nulinių namų ūkių išlaidų problemai spręsti. Paprasčiausią modelį galime apibrėžti kaip

$$\begin{aligned} y_i^* &= X_i\beta + \epsilon_i \\ y_i &= \begin{cases} y_i^*, & \text{jei} \quad y_i^* \geq y_L, \\ y_L, & \text{jei} \quad y_i^* < y_L, \end{cases} \end{aligned} \tag{3.1}$$

kur $\epsilon_i \sim N(0, \sigma^2)$, o y_L žymi apatinį cenzūravimo lygį (modelis su viršutiniu cenzūravimo lygiu formuojamas analogiškai). Tobit modelis plačiai naudojamas ekonominiuose tyrimuose, problemoms susijusioms su duomenų cenzūravimu spręsti. Pavyzdžiui, vedusių moterų atlyginimų dydžių tyrimui arba jau minėtam namų ūkių išlaidų tyrimui. Tobit modelio formuluotėje didžiausių kvadratų metodas duoda nesuderintus parametrų įverčius todėl naudojamas didžiausio tikėtinumo metodas [1]. Didžiausio tikėtinumo funkcija Tobit modeliui yra

$$L(\beta, \sigma) = \prod_{j=1}^{n} \left(\frac{1}{\sigma} \phi \left(\frac{y_j - X_j\beta}{\sigma} \right) \right)^{I(y_j)} \left(1 - \Phi \left(\frac{X_j\beta - y_L}{\sigma} \right) \right)^{1-I(y_j)}, \tag{3.2}$$

kur

$$I(y) = \begin{cases} 0, & \text{jei } y < y_L, \\ 1, & \text{jei } y \geq y_L, \end{cases}$$

o Φ ir ϕ yra standartinio normaliojo atsitiktinio dydžio pasiskirstymo ir tankio funkcijos.

3.2 Heckman dviejų žingsnių procedūra

Heckman modelis [6, 7], kartais vadinamas Heckit modeliu, yra procedūra leidžianti naudoti regresijos modelius šališkai sudarytoms imtims. Heckman metode priklausomas kintamasis yra stebimas tik daliai imties. Klasikinis šališkos imties sudarymo pavyzdys - moterų darbo atlyginimo regresijos lygtis. Moters atlyginimo dydis yra stebimas tik tada jei ji pasirenka dalyvauti darbo rinkoje ir yra nestebimas jei ji to nepasirenka. Heckman savo straipsnyje [7] pristatydamas modelį būtent ir naudojo šį pavyzdį.

Modelis nusakomas dviem lygtimis

$$\begin{aligned} y_i &= X_i\beta + \epsilon_i \\ z_i^* &= W_i\gamma + u_i \end{aligned} \tag{3.3}$$

kur

$$z_i = \begin{cases} 1, & \text{jei } z_i^* > 0, \\ 0, & \text{jei } z_i^* \leq 0, \end{cases} \tag{3.4}$$

o z_i^* yra nestebimas (latentinis) kintamasis, z_i yra binarus kintamasis, y_i yra stebimas tik tada kai $z_i = 1$. ϵ_i ir u_i – paklaidos, kurių skirstinys yra dvimatis normalusis atsitiktinis dydis

$$\begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix} \right) \tag{3.5}$$

kur σ yra skalės parametras, o ρ – koreliacijos. u_i dispersija normalizuojama vienetu, nes variacija šitame modelyje nėra vienareikšmiškai apibrėžta. Pirmoji lygtis vadinama atsakymų lygtimi, su priklausomu kintamuoju y_i, o antroji - pasirinkimų lygtimi. Ji nusako ar y_i yra stebimas ar ne. Verta paminėti, kad antroji 3.3 lygtis kartu su 3.4 yra probit modelis [9] ir gali būti užrašyta kaip

$$P(z_i = 1|W_i) = \Phi(W_i\gamma). \tag{3.6}$$

Heckman dviejų žingsnių procedūra remiasi pastebėjimu, kad

$$E(y_i|z_i = 1) = X_i\beta + \rho\sigma\lambda_i(W_i\gamma)$$

kur $\lambda(X) = \frac{\phi(X)}{\Phi(x)}$ yra vadinamas atvirkštiniu Mills santykiu [5], o ϕ ir Φ yra, atitinkamai, standartinio normaliojo atsitiktinio dydžio tankio ir pasiskirstymo funkcijos. Pirmu žingsniu įvertinami antrosios 3.3 lygties parametrai $\hat{\gamma}$ naudojant probit regresiją. Tada suskaičiuojamas $\lambda_i(W_i\hat{\gamma}_i)$ ir tiesinės regresijos parametrų įverčiai lygčiai

$$y_i = X_i\beta + \rho\sigma\hat{\lambda}_i + v_i.$$

Tokiu būdu gauname suderintus β ir $\theta = \rho\sigma$ įverčius. Parametrų įverčių kovariacijų matrica gaunama apskaičiuojant

$$\hat{\Omega} = \hat{\sigma}^2(X^{*\prime}X^*)^{-1}(X^{*\prime}(I - \hat{\rho}^2\hat{\Delta})X^* + Q)(X^{*\prime}X^*)^{-1}$$

čia $X^* = (X_i', \hat{\lambda_i}')$, $\hat{\Delta}$ – diagonalinė matrica su diagonalės elementais $\delta_i = \hat{\lambda}_i(\hat{\lambda}_i - W_i\hat{\gamma})$, I –

vienetinė matrica, o

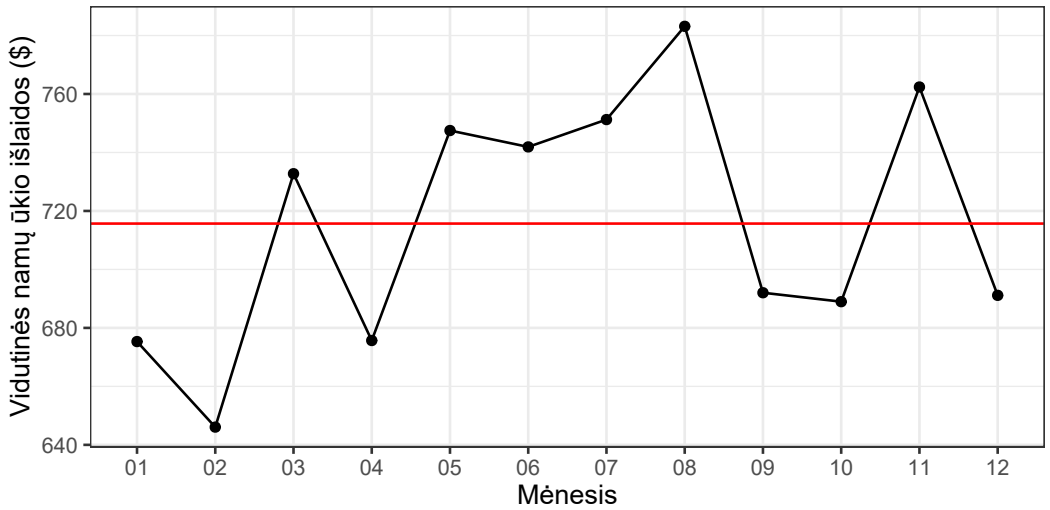
$$Q = \hat{\sigma}^2(X^{*\prime}\hat{\Delta}W)\hat{V}(X^{*\prime}\hat{\Delta}W)',$$

kur $\hat{V}$ yra probit lygties koeficientų matricos įvertis.

4 Duomenų paruošimas ir pirminė analizė

4.1 Sezoniškumas ir duomenų agregavimas

Yra gerai dokumentuota, kad įvairias išlaidų kategorijas veikia sezoniniai efektai [13, 19, 14, 8]. Tai įtakoja įvairūs faktoriai, pavyzdžiui metų pabaigos premijos, šventiniai išpardavimai ir vasaros/žiemos atostogų sezonai ir panašiai. Logiška tikėtis, kad transporto išlaidos taip pat yra ne išimtis. Iš tikrųjų mūsų pasirinktuose duomenyse stebimas sezoniškumas. Pažvelgus į 1 pav. matome, kad išlaidos transportui tipiškai padidėja vasarą ir sumažėja žiemą. Tai taip pat patvirtinama ir literatūroje, pavyzdžiui [10].



1 pav.: Vidutinės namų ūkių išlaidos transportui pagal kalendorinį mėnesį

Norint supaprastinti analizę ir nepaisyti sezoniškumo buvo pasirinkta nagrinėti vienų metų laikotarpį. Kadangi interviu pobūdžio apklausos vykdomos ištisus metus ir klausiama apie praėjusių trijų kalendorinių mėnesių išlaidas pasirinkome nagrinėti 2018 metais vykdytų apklausų laikotarpį. Toks pasirinkimas lemia, kad realūs vartojimo duomenys yra tarp 2017 m. rugsėjo ir 2018 m. lapkričio, tačiau visų kalendorinių mėnesių įnašas yra vienodas. JAV darbo statistikos biuras skaičiuodamas kalendorių metų suminius ir vidurkinius įverčius elgiasi kitaip - naudojami penkių ketvirčių duomenys (2018K1, 2018K2, 2018K3, 2018K4, 2019K1), o pateikti elementų svoriai pirmam ir paskutiniam

ketvirčiui yra perskaičiuojami atsižvelgiant į tai kiek apklausos mėnesių pateko į kalendorinius metus. Mūsų pasirinkimas leidžia neperskaičiuoti pateiktų svorių ir yra labiau tinkamas ieškant sąryšių ir priklausomybių tarp skirtingų duomenų. Šio pasirinkimo minusas - negalime teigti, kad mūsų vykdytas tyrimas apima tik 2018 kalendorinius metus, tačiau sezoniškumo pašalinimui jis yra pakankamas.

4.2 Priklausomi ir nepriklausomi kintamieji

Tolimesnei analizei iš turimų apklausos duomenų paruošėme tokius kintamuosius. Priklausomas kintamasis:

- `EXP_TRANS` Namų ūkio išlaidos transportui (įskaitant PVM). Tik ta dalis kurią apima interviu būdu vykdoma apklausa. Detalų sąrašą galima rasti 2 lentelėje.

Demografiniai kintamieji:

- `I_FEMALE` Ar pagrindinis namų ūkio savininkas yra moteris?
- `I_BLACK` Ar pagrindinis namų ūkio savininkas yra juodaodis?
- `I_HISP` Ar pagrindinis namų ūkio savininkas yra ispanakalbis?
- `I_HIGHSCHOOL` Ar pagrindinis namų ūkio savininkas nebaigė vidurinės mokyklos?
- `I_COLLEGE` Ar pagrindinis namų ūkio savininkas turi koledžo diplomą arba aukštesnį išsilavinimą?

Vietovės kintamieji:

- `I_URBAN` Ar namų ūkis yra miesto teritorijoje?
- `REGION` JAV regionas (*Midwest*, *Northeast*, *South* arba *West*)

Ekonominiai kintamieji:

- `INCOME` Namų ūkio praėjusių metų pajamos (neatskaičiavus mokesčių).
- `I_TENURE` Ar pagrindinis būstas nėra nuosavas (yra nuomojamas)?
- `N_EARNERS` Žmonių, gaunančių pajamas, skaičius.
- `N_VEHQ` Namų ūkio turimų transporto priemonių skaičius.

Šeimos sudėties kintamieji:

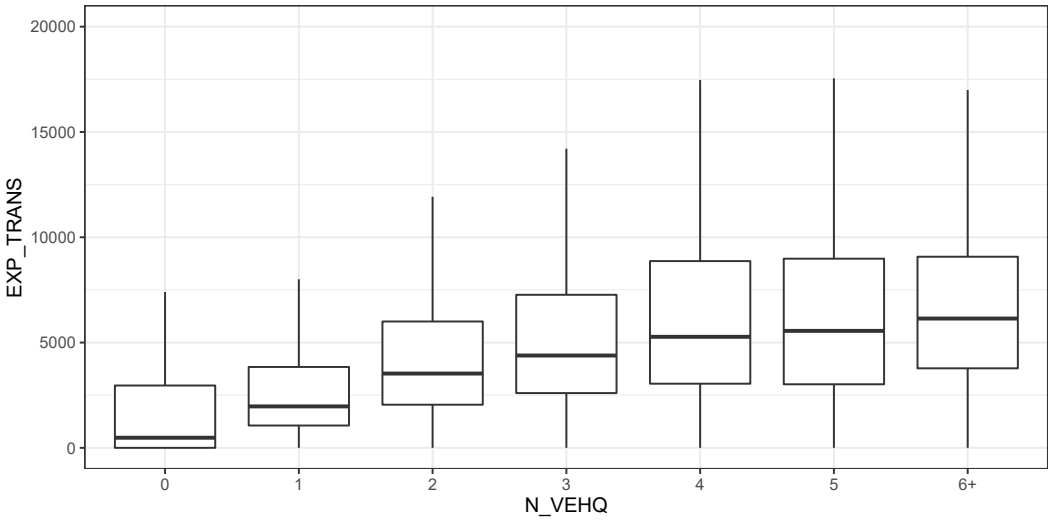
- `I_SINGLEPARENT` Ar pagrindinis namų ūkio savininkas yra vieniša motina ar vienišas tėvas?
- `N_SIZE` Namų ūkiui priskiriamų žmonių skaičius.
- `N_P64` Žmonių, kurių amžius yra didesnis nei 64 metai, skaičius.
- `N_P18` Žmonių, kurių amžius yra mažesnis nei 18 metų, skaičius.
- `N_AGE` Pagrindinio namų ūkio savininko amžius metais.

4.3 Pirminė analizė

Nagrinėjamu periodu vidutinės vieno namų ūkio išlaidos transportui buvo $8605. Tai sudarė apie 14% visų vidutinio namų ūkio išlaidų. Iš šios sumos $1918 (arba 22%) sudarė išlaidos benzinui, $1318 (arba 15%) išlaidos naudotiems džipams ir pikapams, $1166 (arba 14%) naujiems džipams ir pikapams. Išlaidų standartinis nuokrypis – $23044, mediana – $3088. Net 6.9% apklaustųjų atsakė, kad neturėjo jokių išlaidų susijusių su transportu. Kyla klausimas ar šių namų ūkių išlaidos buvo klaidingai priskirtos kitai kategorijai ar jie iš tikrųjų neturėjo jokių išlaidų transportui. Palyginus nulinių išlaidų namų ūkius su likusiais matome aiškias tendencijas. Lentelėje 1 pateikiame nagrinėjamų rodiklių vidurkius nulinių ir didesnių už nulį išlaidų grupėms. Pavyzdžiui nulinių išlaidų transportui namų ūkiai vidutiniškai turi ženkliai mažesnes pajamas ($27480 prieš $81311) bei turi ženkliai mažiau transporto priemonių (0.32 prieš 1.93). Taip pat daug didesnė tikimybė, kad jie nėra baigią vidurinės mokyklos (0.26 prieš 0.09) ir yra vyresnio amžiaus (55.66 prieš 50.79). Pažvelgus į išlaidų dydį pagal turimų transporto priemonių skaičių, kaip ir buvo galima laukti, pastebimas akivaizdus teigiamas sąryšis. Jis pavaizduotas 2 paveiksle. Pavyzdžiui namų ūkių neturinčių transporto priemonių vidutinės išlaidos transportui per metus yra $2576, turinčių vieną transporto priemonę – $4603, kai tuo tarpu turinčių daugiau nei 3 transporto priemones – $20689.

4.4 Normalumo tyrimas

Daugelis standartinių modelių sukurti laikantis normalinių paklaidų prielaidos. Stiprus nukrypimas nuo šios prielaidos gali turėti nelauktų pasekmių, bet gali leisti daryti klaidingas išvadas. 3 paveiksle pavaizduota svorinė `EXP_TRANS` histograma, bei kvantilių-kvantilių grafikas. Šio grafiko x ašyje atidedamas teoriniai normalinio skirtinio kvantiliai, o y ašyje - imties. Jei imties elementai atitinka normalinį dėsnį taškai turėtų gulėti ant $y = x$ tiesės. Kaip matome iš minėto grafiko sunku tikėtis, kad `EXP_TRANS`



2 pav.: Vidutinės namų ūkių išlaidos transportui pagal turimų transporto priemonių kiekį

pasiskirstęs pagal normalinį dėsnį - matoma stipri asimetrija bei sunki dešinė uodegos. Literatūroje yra pasiūlytas ne vienas statistinis testas normalumo prielaidos tikrinimui. Pavyzdžiui Kolmogorov-Smirnov[11], Shapiro-Wilk [15], Anderson-Darling [2] ir daugelis kitų. Atlikus Anderson-Darling testą, taikant pasikliovimo lygmenį $\alpha = 0.05$, nulinė hipotezė apie kintamojo normalumą atmetama ($p < 0.001$).

Ši problema tyrimuose pasitaiko gan dažnai ir vienas iš būdų kovojant su ja yra kintamojo transformacija. Viena dažniausiai taikomų transformacijų yra logaritmavimas. 3 paveiksle pavaizduota naujo kintamojo, `log_EXP_TRANS`, kuriuo žymėsime `EXP_TRANS` logaritmą svorinė histograma bei kvantilių-kvantilių grafikas. Kaip minėjome apie 7% namų ūkių atsakė, kad neturėjo išlaidų transportui. Kadangi logaritmas nulio taške neapibrėžtas jis šiame paveiksle nepavaizduotas. Atlikus Anderson-Darling testą, taikant pasikliovimo lygmenį $\alpha = 0.05$, nulinė hipotezė apie kintamojo normalumą atmetama ($p < 0.001$). Tai galima paaiškinti dideliu imties dydžiu ($n = 23159$). Turint didelę imtį, net menkiausi nukrypimai nuo normalumos prielaidos turi didelę įtaką testo rezultatams [4]. Toliau analizėje naudosime logaritmuotas išlaidas transportui, laikydami, kad jų skirstinys yra pakankamai artimas normaliajam. Namų ūkio pajamas taip pat logaritmuojame, kadangi jos taip pat pasižymi sunkia dešine uodega. Be to patogiau analizuoti ir interpretuoti modeliavimo rezultatus kai išlaidos ir pajamos yra toje pačioje skalėje.

5 Tyrimo rezultatai

Šiame skyrelyje bandysime atsakyti į klausimas kas ir kaip įtakojo namų ūkio išlaidų transportui dydį. Toliau pasiūlysime trys modelius kuriuose priklausomas kintamasis yra `log_EXP_TRANS` – logaritminės metinės išlaidos transportui. Į visus tris modelius įtraukiami demografiniai, vietovės, ekonominiai ir šeimos sudėties kintamieji. Bendrai visi modeliai gali būti užrašyti tokiu pavidalu:

$$\texttt{log\_EXP\_TRANS}_i = f(X_i) + \epsilon_i,$$

kur X_i – aukščiau išvardinti kintamieji, o ϵ_i – modelio paklaidos.

5.1 Tiesinės regresijos modelis

Vienas geriausiai žinomų ir plačiausiai naudojamų modelių yra tiesinės regresijos modelis. Mūsų atveju jį taikyti nėra paprasta, kadangi dalis apklaustųjų atsakė, kad nepatyrė su transportu susijusių išlaidų. Kaip minėjome praeitame skyrelyje `log_EXP_TRANS` kintamasis šiems namų ūkiams nėra apibrėžtas. Pradinei analizei taikysime tiesinės regresijos modelį tik tiems namų ūkiams kurie patyrė teigiamas išlaidas transportui. Tai nepilnai atsako į mūsų išsikeltą klausimą apie transporto išlaidų priklausomybę nuo demografinių ir kitų faktorių, bet leidžia greitai ir gan paprastai pajausti egzistuojančius sąryšius. Regresijos lygtį galime užrašyti taip:

$$\texttt{log\_EXP\_TRANS}_i = \beta_0 + \beta_1 X_i + \epsilon_i. \tag{5.1}$$

Čia X_i – nepriklausomi kintamieji, β_0, β_1 – modelio parametrai, o ϵ_i – modelio paklaidos. Parametrus randame mažiausių kvadratų metodu. Verta paminėti, kad `REGION` yra kategorinis kintamasis todėl į regresiją įtraukėme 3 fiktyvius kintamuosius - `Midwest` (vidurio vakarų regionas), `South` (pietų regionas), `West` (vakarų regionas), kurie įgyja reikšmę 0 arba 1. Jeigu visi šie kintamieji įgyja 0, laikome, kad respondentas priklauso `Northeast` (šiaurės rytų) regionui. 3 lentelėje pateikta parametrų įverčius, standartinių nuokrypių įverčius ir parametrų reikšmingumą. Dauguma pasirinktų kintamųjų yra reikšmingi. Pavyzdžiui, didesnes pajamas gaunantys namų ūkiai, transportui išleidžia reikšmingai daugiau (parametras prie `log_INCOME` yra teigiamas ir lygus 0.186). Šis sąryšis ko gero nestebina, nes gaunant didesnes pajamas, didesnę sumą galima skirti kelionėms, gali įsigyti prabangesnį automobilį ir panašiai. Taip pat gan intuityvus `N_VEHQ` ir `log_EXP_TRANS` (parametro reikšmė 0.22). Buvo galima tikėtis, kad turint daugiau transporto priemonių, transportui išleis daugiau. Įdomus išlaidų transportui sąryšis su

I_SINGLEPARENT. Gauname, kad vieniši tėvai transportui išleidžia statistiškai reikšmingai daugiau (parametras 0.187). Modelis nerodo reikšmingos išlaidų transportui priklausomybės nuo to ar šeimos galva yra juodaodis (kintamasis I_BLACK), ar šeimos būstas yra miesto teritorijoje (kintamasis I_URBAN), kiek šeimoje yra žmonių vyresnių nei 64 metai (kintamasis N_P64), kiek šeimoje yra asmenų jaunesnių nei 18 metų, bei vakarų regiono fiktyvaus kintamojo. Pastarasis rodo, kad nėra statistiškai reikšmingo skirtumo tarp šiaurės rytų ir vakarų regiono, tačiau pastebimas statistiškai reikšmingas skirtumas tarp šiaurės rytų ir vidurio vakarų bei pietų regionų. Modelio paaiškinama variacijos dalis $R^2 = 0.25$. Paklaidų skirstinys artimas normaliajam ir pateiktas 5 paveiksle.

5.2 Tobit modelis

Kadangi netoli 7% procentų apklaustųjų atsakė, kad nepatyrė jokių išlaidų transportui, tolesnei ir tikslesnei analizei pasirinkome Tobit modelį [18]. Kaip minėjome anksčiau jis naudojamas cenzūruotiems duomenims. Šiuo atvejų laikome, kad log_EXP_TRANS cenzūruotas iš apačios nuliu. Modelis užrašomas tokia lygčių sistema:

$$\begin{aligned} \texttt{log\_EXP\_TRANS}_i^* &= \beta_0 + \beta_1 X_i + \epsilon_i \\ \texttt{log\_EXP\_TRANS}_i &= \begin{cases} \texttt{log\_EXP\_TRANS}_i^*, & \texttt{log\_EXP\_TRANS}_i^* > 0 \\ 0, & \texttt{log\_EXP\_TRANS}_i^* \leq 0. \end{cases} \end{aligned} \tag{5.2}$$

Čia log_EXP_TRANS* – nestebimas (latentinis) kintamasis. Jį interpretuojame kaip norimas namų ūkio išlaidas transportui ir todėl galimai įgyjančias neigiamas reikšmes. log_EXP_TRANS sutampa su log_EXP_TRANS* kai log_EXP_TRANS* yra didesnis už nulį, tačiau įgyja nulį kai log_EXP_TRANS* yra mažesnis už nulį. Tai yra kai namų ūkis pateikia nulines išlaidas, galima daryti prielaidą, kad arba namų ūkis išties turi nulines išlaidas arba jis siekia "pasipelnyti" iš esamos transporto paslaugų paklausos, neleisdamas pinigų šiai kategorijai ir vietoje jos pasirinkdamas kitas. Tokį elgesį interpretuojame kaip neigiamas išlaidas. Modelio parametrai rasti naudojant didžiausio tikėtinumo metodą ir pateikti 3 lentelėje. Lyginant su tiesinės regresijos modeliu matome tam tikrų skirtumų. Pirmiausia beveik vis kintamieji yra reikšmingi (išskyrus I_SFEMALE ir Midwest). Taip pat nemažai parametrų įverčių gaunasi gan skirtingi, pavyzdžiui, South parametras Tobit atveju yra 0.121, o tiesinės regresijos – -0.1. Modelio paklaidos artimos normaliosioms. Taip pat neaptikome požymių indikuojančių, kad jos būtų heteroskedastiškos.

5.3 Heckman modelis

Tobit modelyje cenzūravimas atsiranda, kai priklausomas kintamas `log_EXP_TRANS`* yra neigiamas. Šiame skyrelyje pateikiame rezultatus gautus naudojant Heckman dviejų žingsnių procedūrą. Ji yra bendresnė nei Tobit modelis, nes leidžia modeliuoti tiek išlaidų priklausomybę nuo pasirinktų kintamųjų, tiek pasirinkimo leisti pinigus transportui tikimybę. Toks modelis užrašomas dviem lygtimis kurias vadinsime išlaidų ir pasirinkimo lygtimis. Pasirinkimo lygtis nusako, kada išlaidos yra stebimos, o kada ne (kada jos yra cenzūruotos). Matematiškai tokį modelį galime užrašyti taip:

$$\begin{aligned} \texttt{log\_EXP\_TRANS}_i^* &= \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \\ \texttt{D\_EXP\_TRANS}_i^* &= \gamma_0 + \gamma_1 Z_i + u_i, \quad u_i \sim N(0, 1) \end{aligned} \tag{5.3}$$

Antroji lygtis yra probit modelis [9], o Z_i yra priklausomų kintamųjų vektorius, kuris gali sutapti (arba ne) su X_i. Pirmosios lygties kintamasis $\texttt{log\_EXP\_TRANS}_i^*$ yra stebimas tik tada kai $\texttt{D\_EXP\_TRANS}_i^* > 0$. Pirmosios iteracijos metu, išlaidų lygčiai nepriklausomus kintamuosius pasirinkome tokius pačius kaip ir tiesinės regresijos bei Tobit atveju. Pasirinkimo lygčiai naudojame tuos pačius kintamuosius išskyrus `N_VEHQ`. Šio kintamojo įtraukimas nebūtų labai prasmingas, nes jis parodo kiek transporto priemonių namų ūkis turi apklausos metu. Pažvelgus į 3 lentelę matome, kad daugelis parametrų yra reikšmingi. Išlaidų lygties parametrų įverčiai yra artimi tiesinio modelio parametrų įverčiams. Prisiminkime, kad išlaidų lygtis ir yra tiesinė regresiją, o jos kintamieji stebimi tik tada, kai tą indikuoja pasirinkimo lygtis. Jeigu sugebėtume idealiai prognozuoti kada namų ūkis pasirinks leisti pinigus transportui ir kada ne, gautume, kad Heckman parametrų įverčiai sutaptų su tiesinės regresijos. Žvelgiant į pasirinkimų lygtį matome, kad kai kurie kintamieji kurie buvo nereikšmingi tiesinės regresijos atveju, šioje lygtyje yra reikšmingi. Pavyzdžiui `N_P18` kintamasis nurodantis kiek šeimoje yra individų kuriems mažiau nei 18 metų. Šio kintamojo parametro įvertis yra neigiamas (reikšmė -0.178) ir reikšmingas. Tai indikuoja, kad daugiau vaikų turinčios šeimos mažiau yra linkusios pradėti leisti pinigus transportui.

5.4 Galutinis modelis

Apžvelgėme tris skirtingus modeliavimo būdus. Nors Heckerman modelio išlaidų lygties parametrų įverčiai gaunasi panašūs į mūsų pradinio modelio – tiesinės regresijos, įverčius, šis modelis turi esminį pranašumą lyginant su tiesine regresija. Jis leidžia į tą patį modelį įtraukti ir pasirinkimo leisti transportui lygtį. Todėl ir siūlome naudoti šį modelį. Iš modelio išmetame nereikšmingus kintamuosius, taip supaprastindami tiek išlaidų tiek pasirinkimo

lygtis. Galutinio modelio parametrų įverčiai pateikiami 4 lentelėje.

6 Išvados

Šiame darbe pritaikėme tris statistinius modelius, pateikėme jų parametrų įverčius bei įvertinome statistinį reikšmingumą. Palyginę modelius radome daug panašumų, bet ir tam tikrų skirtumų. Geriausiai nagrinėjamą problemą atitiko Heckman modelio formuluotė todėl būtent tokį modelį ir siūlome naudoti. Radome, kad dauguma nagrinėtų kintamųjų (pvz. pajamų dydis, išsilavinimas, amžius, dirbančiųjų skaičius) turi reikšmingą įtaką išlaidų transportui dydžiui.

Nors atsakėme į nemažai išsikeltų klausimų transporto išlaidų tyrimą galima tęsti toliau. Pavyzdžiui galima įvesti laiko dimensiją, pasinaudojant, tuo, kad JAV namų ūkių išlaidų apklausų duomenys yra prieinami ir turi ilgą istoriją (nuo 1980 m.). Tada, taikant laiko eilučių metodus, būtų galima nagrinėti kaip keitėsi išlaidų transportui tendencijos, kokią įtaką išlaidoms ir išlaidų komponentėms turėjo besikeičiančios degalų kainos ir panašiai. Kitas kelias būtų tirti kitas išlaidų kategorijas ir tų kategorijų sąryšius. Pavyzdžiui, jei namų ūkiai pasirenka neleisti pinigų transportui, kokioms prekėms jie tuos pinigus išleidžia?

7 Literatūra

[1] Takeshi Amemiya. Tobit models. *Advanced Econometrics*, pages 360–411, 1985.

[2] Theodore W Anderson and Donald A Darling. A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769, 1954.

[3] Nazneen Ferdous, Abdul Rawoof Pinjari, Chandra R Bhat, and Ram M Pendyala. A comprehensive analysis of household transportation expenditures relative to other goods and services: an application to united states consumer expenditure data. *Transportation*, 37(3):363–390, 2010.

[4] Asghar Ghasemi and Saleh Zahediasl. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, 10(2):486, 2012.

[5] William H Greene. *Econometric analysis*. Pearson Education India, 2003.

[6] James Heckman. J., 1979. sample selection bias as a specification error. *Econometrica*, 47(1):153161, 1979.

[7] James J Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pages 475–492. NBER, 1976.

[8] Masahiro Hori and Satoshi Shimizutani. The response of household expenditure to anticipated income changes: Bonus payments and the seasonality of consumption in japan. *The BE Journal of Macroeconomics*, 9(1), 2009.

[9] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[10] Yihua Liao. *Patterns of Transportation-Related Household Expenditures: The Effects of Income and Housing Subsidies*. PhD thesis, University of Illinois at Chicago, 2003.

[11] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[12] Bureau of Labor Statistics. Consumer expenditures and income. *BLS Handbook of Methods*, 2006.

[13] Denise R. Osborn. Seasonality and habit persistence in a life cycle model of consumption. *Journal of Applied Econometrics*, 3(4):255–266, 1988.

[14] Christina H. Paxson. Consumption and income seasonality in thailand. *Journal of Political Economy*, 101(1):39–72, 1993.

[15] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[16] Piyushimita Thakuriah and Yihua Liao. Transportation expenditures and ability to pay. In *Evidence from the Consumer Expenditure Survey. Presentation made at the 50th Annual Meetings of the North American Regional Science Association. Also UIC-FTA Working Paper WP-3B-04*, 2003.

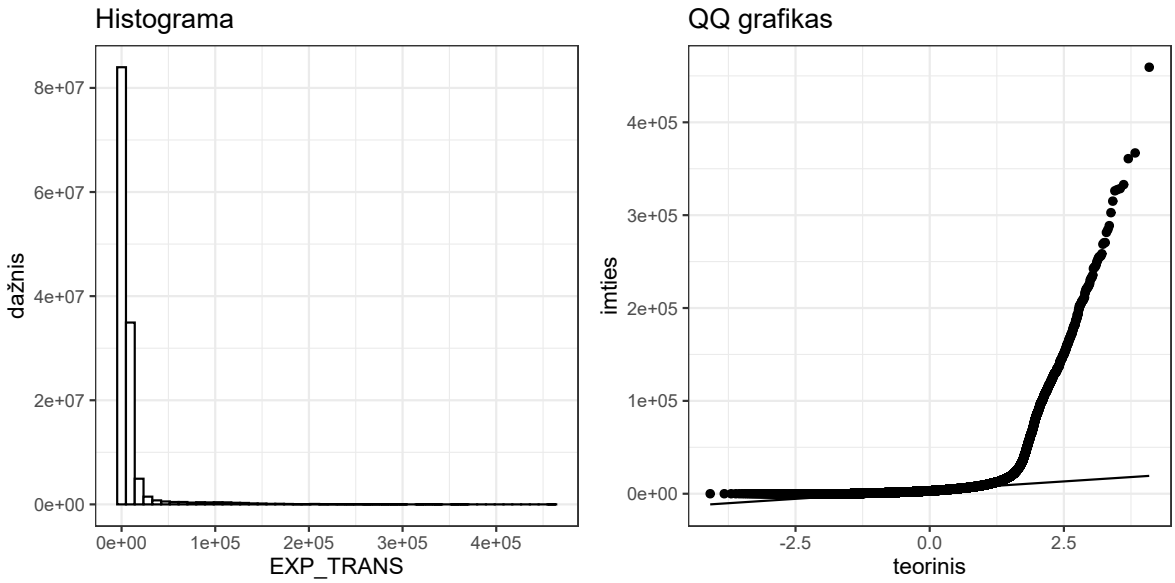
[17] Piyushimita Thakuriah and Yihua Liao. Analysis of variations in vehicle ownership expenditures. *Transportation Research Record*, 1926(1):1–9, 2005.

[18] James Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26(1):24–36, 1958.

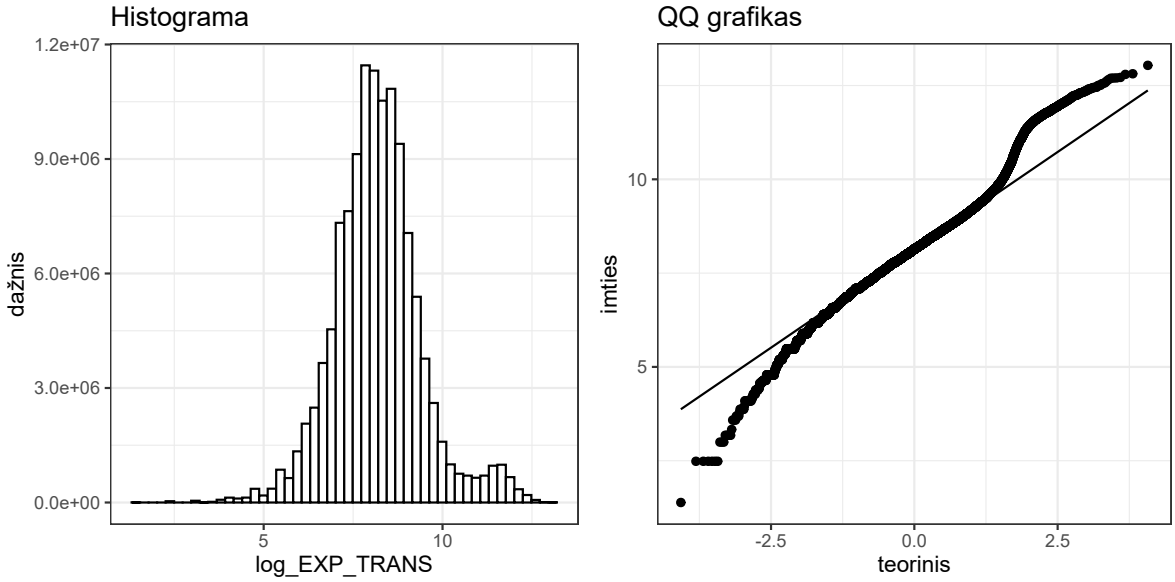
[19] Janet Wagner and Manouchehr Mokhtari. The moderating effect of seasonality on household apparel expenditure. *Journal of Consumer Affairs*, 34(2):314–329, 2000.

8 Priedai

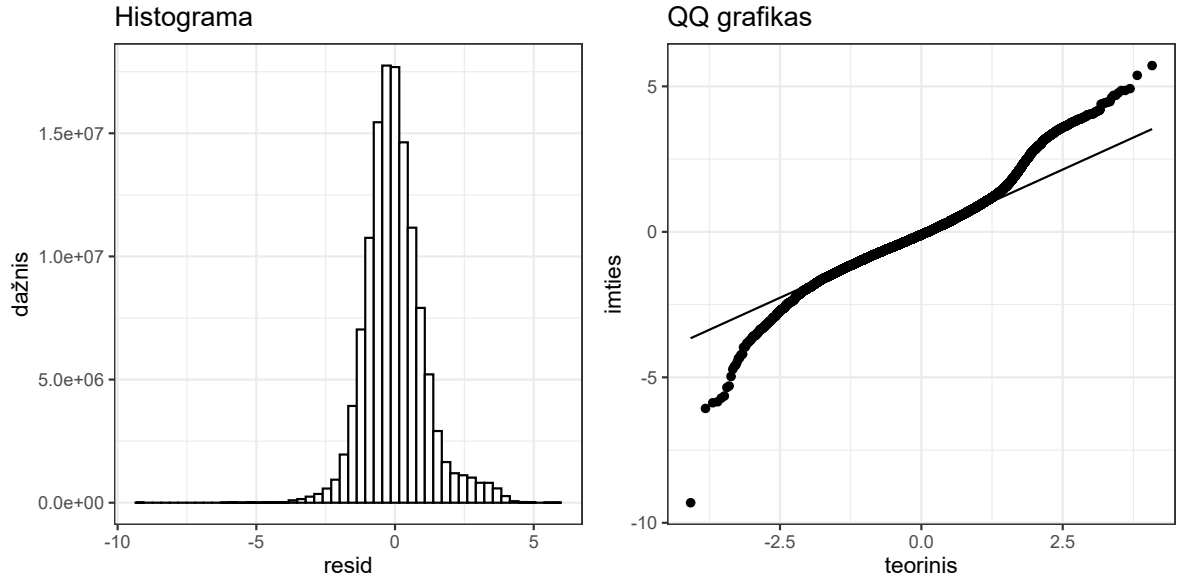
8.1 Grafikai



3 pav.: Vidutinių namų ūkių išlaidų transportui histograma ir normaliojo a.d. kvantilių-kvantilių grafikas



4 pav.: Vidutinių logaritmuotų namų ūkių išlaidų transportui histograma ir normaliojo a.d. kvantilių-kvantilių grafikas



5 pav.: Tiesinės regresijos modelio paklaidų histograma ir normaliojo a.d. kvantilių-kvantilių grafikas

8.2 Lentelės

EXP_TRANS	INCOME	I_FEMALE	I_BLACK	I_HISP	I_HIGHSCHOOL
> 0	81310.89	0.53	0.12	0.13	0.09
$= 0$	27479.89	0.59	0.25	0.14	0.26

EXP_TRANS	I_COLLEGE	I_URBAN	I_TENURE	I_SINGLEPARENT
> 0	0.48	0.93	0.34	0.05
$= 0$	0.21	0.91	0.69	0.06

EXP_TRANS	N_VEHQ	N_SIZE	N_P64	N_P18	N_EARNERS	N_AGE
> 0	1.93	2.52	0.39	0.60	1.35	50.79
$= 0$	0.32	1.73	0.45	0.32	0.64	55.66

1 lentelė: Nulinių išlaidų namų ūkių ir didesnių už nulį išlaidų palyginimas

Description	UCC	Source
Transportation	TRANS	G
Vehicle purchases (net outlay)	VEHPURCH	G
Cars and trucks, new	NEWCARS	G
New cars	450110	I
New trucks	450210	I
Cars and trucks, used	USEDCARS	G
Used cars	460110	I
Used trucks	460901	I
Other vehicles	OTHVEHCL	G
New motorcycles	450220	I
New aircraft	450900	I
Used motorcycles	460902	I
Used aircraft	460903	I
Gasoline and motor oil	GASOIL	G
Gasoline	470111	I
Diesel fuel	470112	I
Gasoline on out-of-town trips	470113	I
Motor oil	470211	I
Motor oil on out-of-town trips	470212	I
Electric vehicle charging	470311	I
Other vehicle expenses	VEHOTHXP	G
Vehicle finance charges	VEHFINCH	G
Automobile finance charges	510110	I
Truck finance charges	510901	I
Motorcycle and plane finance charges	510902	I
Other vehicle finance charges	850300	I
Maintenance and repairs	CAREPAIR	G
Coolant, brake fluid, transmission fluid, and other additives	470220	I
Tires - purchased, replaced, installed	480110	I
Parts, equipment, and accessories	480213	I
Vehicle products and cleaning services	480212	D
Misc. auto repair, servicing	490000	D
Body work and painting	490110	I
Vehicle or engine repairs (new UCC Q20132)	490300	I
Motor tune-up	490311	I
Lube, oil change, and oil filters	490312	I
Front-end alignment, wheel balance and rotation	490313	I
Shock absorber replacement	490314	I
Gas tank repair, replacement	490316	D
Repair tires and other repair work	490318	I
Auto repair service policy	490900	I
Vehicle insurance	500110	D
Vehicle rental, leases, licenses, and other charges	VEHRNTLC	G
Leased and rented vehicles	VEHRENT	G
Rented vehicles	RENTVEH	G
Auto/truck rental (new UCC Q20132)	520516	I
Auto/truck rental, out-of-town trips (new UCC Q20132)	520517	I
Motorcycle rental	520902	I
Motorcycle rental, out-of-town trips	520905	I
Aircraft rental	520903	I
Aircraft rental, out-of-town trips	520906	I
Leased vehicles	LEASVEH	G
Car/truck lease payments (new UCC Q20132)	450350	I
Extra fees for car/truck lease (new UCC Q20132)	450351	I
Trade in allowance for car/truck lease (new UCC Q20132)	450352	I
Cash downpayment car/truck lease (new UCC Q20132)	450353	I
Termination fee for car/truck lease (new UCC Q20132)	450354	I
Vehicle registration state (new UCC Q20012)	520111	I
Vehicle registration local (new UCC Q20012)	520112	I
Drivers" license	520310	I
Vehicle inspection	520410	I
Parking fees	PARKING	G
Parking fees in home city, excluding residence	520531	I
Parking fees, out-of-town trips	520532	I
Tolls or electronic toll passes	520541	I
Tolls on out-of-town trips	520542	I
Towing charges	520550	I
Automobile service clubs and GPS services (new UCC Q20132)	620114	I
Public transportation	PUBTRANS	G
Airline fares	530110	I
Intercity bus fares	530210	I
Intracity mass transit fares	530311	I
Local trans. on out-of-town trips	530312	I
Taxi fares and limousine services on trips	530411	I
Taxi fares and limousine services	530412	D
Intercity train fares	530510	I
Ship fares	530901	I
School bus	530902	I

Pastabos: G - grupė, I - interviu apklausa, D - dienoraščio tipo apklausa

2 lentelė: Išlaidų kategorijų, susijusius su transportu, kodai ir apklausos šaltiniai (originalo kalba)

	Dependent variable:			
	log_EXP_TRANS			D_EXP_TRANS
	OLS	*Tobit*	*Heckman*	
			Outcome	Selection
	(1)	(2)	(3)	(4)
log_INCOME	0.186^{***}	0.352^{***}	0.167^{***}	0.162^{***}
	(0.007)	(0.014)	(0.010)	(0.011)
I_FEMALE	-0.049^{***}	-0.050^{*}	-0.046^{***}	-0.064^{**}
	(0.015)	(0.030)	(0.015)	(0.030)
I_BLACK	0.006	-0.290^{***}	0.033	-0.312^{***}
	(0.024)	(0.048)	(0.028)	(0.038)
I_HISP	0.094^{***}	0.277^{***}	0.089^{***}	0.017
	(0.024)	(0.047)	(0.025)	(0.046)
I_HIGHSCHOOL	-0.117^{***}	-0.656^{***}	-0.077^{**}	-0.374^{***}
	(0.028)	(0.053)	(0.032)	(0.041)
I_COLLEGE	0.190^{***}	0.331^{***}	0.177^{***}	0.290^{***}
	(0.016)	(0.032)	(0.017)	(0.034)
I_URBAN	0.029	0.141^{**}	0.021	0.153^{***}
	(0.030)	(0.058)	(0.030)	(0.053)
I_TENURE	-0.210^{***}	-0.601^{***}	-0.170^{***}	-0.582^{***}
	(0.019)	(0.036)	(0.024)	(0.034)
I_SINGLEPARENT	0.187^{***}	0.569^{***}	0.160^{***}	0.320^{***}
	(0.036)	(0.073)	(0.039)	(0.070)
N_VEHQ	0.220^{***}	0.487^{***}	0.220^{***}	
	(0.006)	(0.012)	(0.006)	
N_SIZE	0.086^{***}	0.109^{***}	0.074^{***}	0.192^{***}
	(0.013)	(0.027)	(0.014)	(0.028)
N_P64	-0.021	0.133^{***}	-0.028^{*}	0.019
	(0.016)	(0.032)	(0.017)	(0.034)
N_P18	-0.031^{*}	-0.076^{**}	-0.020	-0.178^{***}
	(0.016)	(0.033)	(0.018)	(0.036)
N_EARNERS	0.089^{***}	0.110^{***}	0.088^{***}	0.089^{***}
	(0.013)	(0.026)	(0.013)	(0.026)
N_AGE	-0.007^{***}	-0.022^{***}	-0.006^{***}	-0.009^{***}
	(0.001)	(0.001)	(0.001)	(0.001)
Midwest	-0.099^{***}	0.024	-0.117^{***}	0.229^{***}
	(0.025)	(0.047)	(0.025)	(0.045)
South	-0.100^{***}	0.121^{***}	-0.123^{***}	0.299^{***}
	(0.022)	(0.043)	(0.024)	(0.039)
West	-0.020	0.162^{***}	-0.041	0.266^{***}
	(0.024)	(0.045)	(0.025)	(0.045)
Constant	5.803^{***}	3.525^{***}	6.042^{***}	-0.198
	(0.090)	(0.171)	(0.127)	(0.135)
Observations	20,750	22,291	22,291	22,291
R^2	0.252			
Adjusted R^2	0.251			
Log Likelihood		$-47{,}835.090$		
Akaike Inf. Crit.		95,710.180		
Bayesian Inf. Crit.		95,870.420		
ρ			-0.352	-0.352
Inverse Mills Ratio			-0.382^{***} (0.142)	-0.382^{***} (0.142)
Residual Std. Error	1.070 (df = 20731)			
F Statistic	388.161^{***} (df = 18; 20731)			
Note:				*p<0.1; **p<0.05; ***p<0.01

3 lentelė: Modelių parametrų įverčiai, standartinių nuokrypių įverčiai ir reikšmingumas

	Dependent variable:	
	log_EXP_TRANS	D_EXP_TRANS
	(1)	(2)
log_INCOME	0.174***	0.164***
	(0.009)	(0.011)
I_FEMALE	−0.049***	
	(0.015)	
I_HISP	0.086***	
	(0.024)	
I_BLACK		−0.318***
		(0.037)
I_HIGHSCHOOL	−0.088***	−0.372***
	(0.031)	(0.040)
I_COLLEGE	0.181***	0.290***
	(0.017)	(0.034)
I_URBAN		0.150***
		(0.052)
I_TENURE	−0.183***	−0.581***
	(0.023)	(0.034)
I_SINGLEPARENT	0.165***	0.303***
	(0.036)	(0.070)
N_VEHQ	0.219***	
	(0.006)	
N_SIZE	0.060***	0.199***
	(0.007)	(0.026)
N_P18		−0.188***
		(0.035)
N_EARNERS	0.103***	0.083***
	(0.011)	(0.025)
N_AGE	−0.006***	−0.008***
	(0.001)	(0.001)
Midwest	−0.094***	0.229***
	(0.020)	(0.045)
South	−0.095***	0.299***
	(0.017)	(0.039)
West		0.268***
		(0.045)
Constant	5.987***	−0.250*
	(0.111)	(0.129)
Observations	22,291	22,291
ρ	−0.223	−0.223
Inverse Mills Ratio	−0.240** (0.119)	−0.240** (0.119)
Note:		*p<0.1; **p<0.05; ***p<0.01

4 lentelė: Galutinio modelio parametrų įverčiai, standartinių nuokrypių įverčiai bei parametrų reikšmingumas

8.3 R kodas

```
library(haven)
library(data.table)
library(sampleSelection)
library(VGAM)

## Parameters ------------------------------------------
year = 2018

## Funkcijos -------------------------------------------
stubfile = as.data.table(
  read.fwf(
    paste0("data//CE-HG-Integ-", year,".txt"),
    width = c( 1 , -2 , 1 , -2 , 60 , -3 , 6 , -7 , 1 , -2 , 1, -2, 7 ) ,
    col.names = c("type", "level", "title", "ucc", "survey", "multiplier", "group")
  )
)

get_uccs <- function(category) {
  stubfile[,{
    l <- level[ucc == category];
    i <- which(level == l);
    k <- which(ucc[i] == category);
    uccs <- ucc[i[k]:(i[k+1]-1)];
    list(uccs[grepl("^[0-9]{1,}$", uccs)])
  }]$V1
}
## Duomenu paruosimas ----------------------------------
year_str = sprintf("%02d", year %% 100)
year_next = sprintf("%02d", (year+1) %% 100)

fmli = rbindlist(list(
  read_dta(paste0("data/intrvw", year_str, "/fmli", year_str, "1x.dta")),
  read_dta(paste0("data/intrvw", year_str, "//fmli", year_str, "2.dta")),
  read_dta(paste0("data/intrvw", year_str, "//fmli", year_str, "3.dta")),
  read_dta(paste0("data/intrvw", year_str, "//fmli", year_str, "4.dta")),
  read_dta(paste0("data/intrvw", year_str, "//fmli", year_next, "1.dta"))
))

mtbi = rbindlist(list(
  read_dta(paste0("data/intrvw", year_str, "//mtbi", year_str, "1x.dta")),
  read_dta(paste0("data/intrvw", year_str, "//mtbi", year_str, "2.dta")),
  read_dta(paste0("data/intrvw", year_str, "//mtbi", year_str, "3.dta")),
  read_dta(paste0("data/intrvw", year_str, "//mtbi", year_str, "4.dta")),
  read_dta(paste0("data/intrvw", year_str, "//mtbi", year_next, "1.dta"))
))

mult = stubfile[, list(multiplier = first(multiplier)), by = "ucc"]
mtbi = merge(mtbi, mult, by = "ucc", all.x = TRUE)

# UCC kodai tranporto sektoriuj
uccs = get_uccs("TRANS_")

# Sujungiame duomenu masyvus
exp_mtbi = merge(
  mtbi[ucc %in% uccs & pubflag == 2, list(exp_total = sum(cost*multiplier)), by = "newid"],
  fmli, by = "newid", all = TRUE)
exp_mtbi[is.na(exp_total), exp_total := 0]

# Paruosiame kintamuosiu
d = exp_mtbi[qintrvyr == year & region != "", list(
  EXP_TRANS = exp_total*4,
  log_EXP_TRANS = log(exp_total*4),
  D_EXP_TRANS = exp_total*4 > 0,
  w = finlwt21/4,
  w_N = finlwt21/4/sum(finlwt21/4)*.N,
  INCOME = fincbtxm,
  log_INCOME = log(pmax(1,fincbtxm)),
  I_FEMALE = sex_ref == 2,
  I_BLACK = ref_race == 2,
  I_HISP = hisp_ref == 1,
  I_HIGHSCHOOL = as.numeric(educ_ref) %in% c(1,2,11,0,7,10),
  I_COLLEGE = as.numeric(educ_ref) %in% c(14,15,5,16,17),
```

```
  I_URBAN = bls_urbn == 1,
  I_TENURE =  cutenure > 3,
  I_SINGLEPARENT = fam_type %in% c(6,7),
  N_VEHQ = vehq,
  N_SIZE = fam_size,
  N_P64 = persot64,
  N_P18 = perslt18,
  N_EARNERS = no_earnr,
  N_AGE = age_ref,
  REGION = factor(region, levels = c(as.character(1:4)), labels = c("Northeast", "Midwest", "South", "West"))
)]

d[, Northeast := (REGION == "Northeast")]
d[, Midwest := (REGION == "Midwest")]
d[, South := (REGION == "South")]
d[, West := (REGION == "West")]
d[EXP_TRANS == 0, log_EXP_TRANS := 0]

Pop = sum(d$w)

## Modeliai ----------------------------
X =  d[, -c("w", "w_N","INCOME", "D_EXP_TRANS", "EXP_TRANS", "REGION", "Northeast")]
w = d$w_N
filter = d$EXP_TRANS > 0

# Tiesine regresija
model.lm = lm(log_EXP_TRANS ~ ., data = X[filter], weights = w[filter])
summary(model.lm)

# Tobit
model.tobit = vglm(formula = log_EXP_TRANS ~ ., family = tobit(Lower = 0), data = X, weights = w)
summary(model.tobit)

# Heckman
X =  d[, -c("w", "w_N","INCOME", "EXP_TRANS", "REGION", "Northeast")]
model.heckman = selection(D_EXP_TRANS ~ log_INCOME + I_FEMALE + I_BLACK + I_HISP + I_HIGHSCHOOL + I_COLLEGE +
                            I_URBAN + I_TENURE + I_SINGLEPARENT + N_SIZE + N_P64 + N_P18 + N_EARNERS + N_AGE +
                            Midwest + South + West,
                          log_EXP_TRANS ~ log_INCOME + I_FEMALE + I_BLACK + I_HISP + I_HIGHSCHOOL + I_COLLEGE +
                            I_URBAN + I_TENURE + I_SINGLEPARENT + N_VEHQ + N_SIZE + N_P64 + N_P18 + N_EARNERS +
                            N_AGE + Midwest + South + West, X, weights = w, method = "2step")
summary(model.heckman)

# Galutinis Heckman
model.final = selection(D_EXP_TRANS ~ log_INCOME + I_BLACK + I_HIGHSCHOOL + I_COLLEGE + I_URBAN + I_TENURE +
                          I_SINGLEPARENT + N_SIZE + N_P18 + N_EARNERS + N_AGE + Midwest + South + West,
                        log_EXP_TRANS ~ log_INCOME + I_FEMALE + I_HISP + I_HIGHSCHOOL + I_COLLEGE + I_TENURE +
                          I_SINGLEPARENT + N_VEHQ + N_SIZE + N_EARNERS + N_AGE + Midwest + South, X, weights = w,
                        method = "2step")

summary(model.final)
```