Jyväskyläläisten kotitalouksien kaupan valinta ja otanta-asetelman vaikutus tuloksiin

Jani Sairanen

Tilastotieteen pro gradu –tutkielma
26. heinäkuuta 2000

Jyväskylän yliopisto
tilastotieteen laitos

Tiivistelmä

Jani Sairanen: *Jyväskyläläisten kotitalouksien kaupan valinta ja otanta–asetelman huomioimisen vaikutus tuloksiin.*
Tilastotieteen pro gradu –tutkielma, Jyväskylän yliopisto, 26 heinäkuuta 2000.
Sivuja 78, liitteitä 4.

Tutkielman tavoitteena on arvioida Jyväskyläläisten kauppojen päivittäistavaramyyntiä sekä syitä, joiden mukaan Jyväskylän ja sen lähialueiden kotitaloudet valitsevat kauppansa. Tutkielmassa tarkastellaan, kuinka eri asuinalueiden asukkaiden arvostus asuinalueensa päivittäistavaroiden tarjontaan vaihtelee. Tutkielmassa vertaillaan myös otanta–asetelman huomioimisen ja rekistereiden käytön vaikutuksia tuloksiin ja tulosten tarkkuuteen.

Tutkielman teoreettiset alueet koskevat otantateoriaa, yleistettyjä lineaarisia malleja ja niistä johdettuja analyysimenetelmiä: logit–malleilla haetaan kaupanvalintaan vaikuttavia syitä, loglineaarisilla malleilla tutkitaan muuttujien välisiä riippuvuuksia, logistista ja lineaarista regressiota, parametrista ja epäparametrista varianssianalyysia käytetään tutkittaessa kotitalouksien käyttäytymistä.

Aineiston käsittelyssä käytetään ositetun otannan teoriaa tekemällä ositus asuinalueen suhteen. Tämä ei kuitenkaan anna eri tuloksia kuin yksinkertaisen satunnaisotannan tilanne. Poikkeus on kauppojen saaman rahamäärän arviointi. Tällöin otantamenetelmien huomioiminen paransi tulosten tarkkuutta. Tilastokeskuksen tekemän rekisterin käyttäminen lisäsi tulosten tarkkuutta edelleen.

Tuloksista käy ilmi, että puhelinluettelosta tehty otanta tuo otokseen liian suuria kotitalouksia jättäen samalla pienet kotitaloudet liian pienelle osuudelle. Kaupan valintaan vaikuttavat eniten asuinalue ja käytetty kulkuväline.

Avainsanoja: Yleistetyt lineaariset mallit, ositettu otanta, loglineaariset mallit, logit–mallit

Sisältö

1 Johdanto

Jyväskyläläinen Tietoykkönen Oy tekee vuosittain *Päivittäistavaratutkimusta Jyväskylän seudulla.* Tämä tutkielma perustuu tämän tutkimuksen aineistoon, joka oli kerätty 25.8–1.9.1998. Aineisto on kerätty puhelinhaastattelulla ja haastatteluja tehtiin 420 kappaletta. Otanta oli tehty alueen puhelinluettelosta ositetulla otannalla siten, että kyseessä oli tasakiintiöity ositettu otanta ja ositteena oli maantieteellinen asuinalue. Ositteen sisällä tapahtunut poiminta suoritettiin yksinkertaisella satunnaisotannalla. Tämän otanta–asetelman käytön idea oli varmistaa, että jokaisesta asuinalueesta poimitaan riittävästi kotitalouksia, jotta eri asuinalueiden välinen vertailu on mahdollista.

Tutkielman tavoitteena oli arvioida Jyväskyläläisten kauppojen päivittäistavaramyyntiä sekä syitä, joiden mukaan kotitaloudet valitsevat kauppansa. Tutkielmassa vertaillaan myös ositetun otannan huomioon ottamisen ja rekistereiden käytön vaikutuksia tuloksiin ja tulosten tarkkuuteen yksinkertaiseen satunnaisotantaan verrattuna.

Tutkimuksessa käytetään yleistettyjen lineaaristen mallien teoriaa. Aineistossa oli paljon luokiteltuja muuttujia, joten aineistoon sovellettiin loglineaarisia ja logit –malleja. Lisäksi tutkielmassa sovelletaan varianssianalyysia, Kruskall–Wallisin yksisuuntaista varianssianalyysia, lineaarista ja logistista regressiota.

Kappaleessa 3 käsitellään eksponentiaalinen jakaumaperhe ja yleistetyt lineaariset mallit, koska aineiston analyysissä käytetyt menetelmät perustuvat niihin lukuunottamatta Keruskall–Wallisin yksisuuntaista varianssianalyysia.

Analyysien tekemiseen on käytetty useita ohjelmistoja. Otanta–asetelman huomioimisen vaativat analyysit on tehty SUDAAN 7.5.2 –ohjelmistolla. Yksinkertaisen satunnaisotannan analyysit on suoritettu SPSS 8.0 – ja SAS – ohjelmistoilla. Lisäksi kauppojen saaman rahamäärän arviointiin on käytetty apuna Quattro Pro 8 –taulukkolaskentaohjelmaa. Tekstinladontaan on käytetty tietenkin LaTeXia.

1.1 Tutkimusaineisto

Tutkimusaineisto muodostuu Jyväskylän kaupungin ja Jyväskylän lähialueiden kotitalouksista. Aineiston on kerännyt Tietoykkönen Oy puhelinhaastatteluilla 25.8.–1.9.1998. Käytetty otanta–asetelma on ositettu otanta, jossa ositus on tehty asuinalueen mukaan ja muistuttaa lähinnä tasakiintiöintiä. Ositteen sisällä tapahtunut poiminta suoritettiin yksinkertaisella satunnaisotannalla. Otanta on tehty alueen puhelinluettelosta.

Otokseen tulleista kotitalouksista haastateltiin päivittäistavaraostoksista vastaavaa henkilöä. Haastatteluja on suoritettu 420 kappaletta. Taulukossa 1 on tutkimuksen aluejako ja otoskoko alueittain.

Taulukko 1. Asuinalueet ja otoskoot

Asuinalue	n
Keskusta	40
Kortepohja / Kypärämäki	30
Keltinmäki / Myllyjärvi	30
Keljo / Keljonkangas / Sarvivuori	30
Kuokkala	30
Aittorinne / Halssila	30
Kangasvuori / Kangaslampi	30
Lohikoski / Mannila / Heinälampi	30
Palokka / Tikkakoski	30
Vaajakoski / Jyskä	40
Muurame	40
Säynätsalo	20
Laukaa	40
Yhteensä	420

Tästä lähtien alueita kutsutaan siksi alueeksi, joka on mainittu ensimmäiseksi.

Jyväskylän kaupunki on tehnyt rekistereihin perustuvan tilaston Jyväskylän

kaupungin ja sen lähiympäristön kotitalouksien määrästä. Tietoykkönen Oy:n ja kaupungin aluejaot vastaavat toisiaan hyvin. Tässä tutkimuksessa tullaan käyttämään hyväksi rekisteritietoja. Rekisteritiedot ovat tarpeen laskettaessa ositetun otannan tapauksessa painoja ositteille. Tiedot ovat Jyväskylän kaupungin rekistereistä, Vaajakosken yhteyspalvelusihteeriltä ja Jyväskylän maalaiskunnan yhteyspalvelusihteeriltä. Taulukossa 2 on vuoden 1998 väestömäärät ja asuntokanta asuinalueittain.

Taulukko 2. Väestömäärä ja asuntokanta asuinalueittain vuonna 1998

Asuinalue	Väestö	Asuntokanta
Keskusta	20139	13580
Kortepohja	9214	5270
Keltinmäki	6675	3400
Keljo	4179	1550
Kuokkala	13233	5620
Aittorinne	5429	2540
Kangasvuori	8614	4410
Lohikoski	4806	2170
Palokka	9800	3920
Vaajakoski	12000	4300
Muurame	7700	2750
Säynätsalo	3514	1500
Laukaa	16542	5907
Yhteensä	121845	56917

Tutkimuksessa tullaan pohtimaan, kuinka ositetun otannan teorian käyttö vaikuttaa tuloksiin. Ositetun otannan käytön mahdollistamiseksi perusjoukon ositteiden koot täytyy tietää. Seuraavassa kappaleessa esitellään ositetun otannan teoria.

2 Otantateoria

Yksinkertainen satunnaisotanta on otannan perusmenetelmä. Se on menetelmistä käytetyin ja helpoin ja voidaan poimintatapansa perusteella luokitella kahteen tyyppiin:
1) poiminta palauttaen
2) poiminta palauttamatta
Teoreettisessa mielessä palauttaen –tyyppinen poiminta on yksinkertaisempi, mutta palauttamatta –tyyppistä poimintaa käytetään useimmin. Estimointi on helppoa yksinkertaisen satunnaisotannan tilanteessa. Minkäänlaisia havaintojen painotuksia ei käytetä. Tilasto–ohjelmistot kuten SPSS ja SAS olettavat, että poiminta on suoritettu yksinkertaisella satunnaisotannalla. Monimutkaisia otanta–asetelmia käytettäessä onkin analyysivaiheessa käytettävä erikoisohjelmistoja kuten SUDAAN, joissa voidaan ottaa huomioon otanta–asetelma.

Ositettu otanta

Ositettu otanta perustuu ennakkotietoon, jolla perusjoukon alkiot voidaan luokitella tai ryhmitellä toisensa poissulkeviin osajoukkoihin. Osajoukkoja sanotaan ositteiksi. Kun ositteet on valittu, jokaisesta ositteesta poimitaan valitulla otantamenetelmällä otos. Ositteiden sisällä voidaan poiminta suorittaa esimerkiksi yksinkertaisella satunnaisotannalla. Lisäksi ositteista poimittujen alkioiden lukumäärä määrätään erikseen. Tätä sanotaan kiintiöimiseksi. Kiintiöimiseen palataan myöhemmin.

Ositettu otanta vaatii huomattavan paljon ennakkotietoa perusjoukosta. Täytyy tietää, kuinka perusjoukko on jakautunut ositteisiin. Lisäksi täytyy tietää perusjoukon koko N. Muuten tarvittavia alkioiden painotuksia ei voida suorittaa ja koko otantamenetelmää ei voida käyttää. Ositetussa otannassa perusjoukon ositus on oltava yhteydessä tutkittavan tulosmuuttujan Y vaihteluun. Esimerkiksi jos tiedetään, että eri ikäryhmissä käytetään rahaa eri tavalla, on järkevää suorittaa ositus ikäryhmittäin. Tarkoituksena on siis se, että ositteiden sisällä olisi mahdollisimman vähän vaihtelua ja vaihtelu olisi ositteiden välillä.

Pahkisen ja Lehtosen (1989) mukaan perusjoukosta valittujen ositteiden välillä on voimassa:

$$\sum_{h=1}^{H} N_h = N,$$

jossa N_h on ositteen h alkioiden lukumäärä. H on ositteiden lukumäärä ja N on perusjoukon alkioiden lukumäärä. Poimimalla jokaisesta ositteesta kokoa n_h oleva yksinkertainen satunnaisotos saadaan tuloksena otantamenettely, jota kutsutaan ositetuksi satunnaisotannaksi. Lehtosen ja Pahkisen (1994) mukaan ositetun otannan käytölle on neljä perussyytä:
- hallinnollisista syistä monet kehysperusjoukot ovat valmiiksi jaettu luonnollisiin osajoukkoihin, joita voidaan käyttää osittamisessa.
- osittaminen mahdollistaa ulkopuolisen informaation käytön ositteen sisällä sekä poiminnassa että estimoinnissa.
- osittaminen voi lisätä estimoinnin tarkkuutta jos jokainen osite on homogeeninen.
- osittaminen voi varmistaa haluttujen osajoukkojen edustuksen otoksessa.
Ositetun otannan tärkeimpiä erikoiskysymyksiä on valitun otoskoon n jakaminen ositteiden kesken. Merkitsemällä ositekohtaista otoskokoa n_h todetaan, että $n_1 + n_2 + \ ... \ + n_h + ... + n_H = n$.
Otoksen jakamista ositteiden kesken nimitetään otoksen kiintiöinniksi. Ositetun otannan poiminta tehdään kahdessa vaiheessa. Ensin määritellään ositteet ja kiintiöidään otos eri ositteiden kesken. Sen jälkeen poimitaan otantayksiköt kustakin ositteesta $h = 1, \ \ldots \ , H$. Ositekohtaisesti ollaan siten tilanteessa, missä kokoa N_h olevasta perusjoukon ositteesta on poimittava kokoa n_h oleva otos. Ositekohtaisesti voidaan soveltaa tilanteen mukaan joko yksinkertaista satunnaisotantaa, systemaattista otantaa tai otantaa otosyksikön koon mukaan. Käytetty poimintamenettely vaikuttaa alkioiden poimintatodennäköisyyksiin ja sitä kautta estimointiin.

Ositetussa otannassa estimaattorit ovat yleensä ositekohtaisten estimaattoreiden painotettuja summia. Painoina ovat ositepainot. Ositepainon laskeminen perustuu siihen otantamenetelmään, jolla alkiot on poimittu ositteesta. Ositetun otannan estimoinnissa on kahdessa tasossa olevia käsitteitä,

joista alempi koskee ositetasoa. Ositetasoa voidaan ajatella alitason perusjoukoksi, jolla on omat parametrit ja estimaattorit. Tarkastellaan seuraavaksi tärkeimpiä ositekohtahtaisia estimaattoreita:
Ositteen $h = 1, 2, \ldots, H$ estimaattoreita:
Keskiarvon estimaattori yksinkertaisen satunnaisotannan tilanteessa:

$$\overline{y}_h = \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h} \tag{1}$$

Varianssin estimaattori yksinkertaisen satunnaisotannan tilanteessa:

$$s_h^2 = \sum_{i=1}^{n_h} (y_{hi} - \overline{y}_h)^2/(n_h - 1) \tag{2}$$

Ositekohtaiset estimaattorit riippuvat alkiotason poiminnasta. Perusjoukon estimaattorit ovat painotettuja summia osite–estimaattoreista. Perusjoukon keskiarvon estimaattori lasketaan kaavalla:

$$\overline{y} = \sum_{h=1}^{H} W_h \overline{y}_h, \tag{3}$$

jossa $W_h = N_h/N$ on ositepaino ja ositteen h keskiarvon estimaattori lasketaan kaavan (1) avulla.
Perusjoukon varianssin estimaattori

$$s^2 = \sum_{i=1}^{H} W_h^2 s_h^2, \tag{4}$$

jossa s_h^2 lasketaaan kaavan (2) avulla ja W_h^2 on ositepaino.

Otoksen kiintiöinti

Ositetussa otannassa on tärkeää se, kuinka paljon jokaiseen ositteeseen poimitaan alkioita ja kuinka kiinteäksi asetettu otoskoko n jaetaan ositteiden $h = 1, 2, \ldots, H$ kesken. Tätä jakamista kutsutaan kiintiöinniksi. Tarkastellaan kolmea kiintiöimismenetelmää.
- tasakiintiöinti
- suhteellinen kiintiöinti
- optimaalinen kiintiöinti

Tutkimuksen aineistossa käytetty kiintiöintimenetelmä muistuttaa lähinnä tasakiintiöintiä. Osituksen toteutus riippuu siitä, minkälaisia tietoja perusjoukosta on käytettävissä. Otos pyritään jakamaan ositteisiin siten, että koko perusjoukkoa koskevan parametrin estimointi tulisi mahdollisimman tehokkaaksi. Pahkisen ja Lehtosen (1989) mukaan *tasakiintiöinnissä* jokaiseen ositteeseen poimitaan yhtä paljon alkioita. Tasakiintiöinti on helppo suorittaa. Menettely hyödyntää ainoastaan perusjoukon ryhmittelyä ositteisiin, eikä lainkaan ositteiden sisäistä informaatiota tutkittavasta muuttujasta. *Suhteellisessa kiintiöinnissä* otetaan huomioon perusjoukon ositteen koko N_h. Tällöin siis suuresta ositteesta poimitaan enemmän alkioita kuin pienestä ositteesta. Suhteellista kiintiöntiä kutsutaan itsepunnitsevaksi, sillä ositepainoja ei tarvita joidenkin tunnuslukujen estimoinnissa. Esimerkiksi keskiarvo on itsepunnitseva, mutta painoja tarvitaan keskiarvon varianssin laskemisessa. Ositteen koon huomioiminen tuottaa painot automaattisesti. Suhteellista kiintiöintiä on syytä käyttää silloin, kun suurissa osittеissa tulosmuuttujalla on suurempi vaihtelu kuin pienissä osittеissa. *Optimaalisessa kiintiöinnissä* päähuomio on valitun estimaattorin varianssin minimointi. Yleensä huomioon otetaan myös kustannukset. Päädytään siis optimointitehtävään, jossa tilanteen mukaan minimoidaan kustannuksia ja estimaattoreiden variansseja.

Otanta–asetelmien tehokkuus

Yksinkertainen satunnaisotanta on yleinen otantamenetelmä ja sitä myös käytetään paljon. Jos halutan puristaa otoksesta vielä parempia ja tarkempia estimaatteja, on käytettävä monimutkaisempia otantamenetelmiä. Monimutkaisemmat otantamenetelmät lisäävät tutkijan työtä, mutta onnistuessaan palkitsevat tutkijan näkemän vaivan. Usein halutaan tietää, kuinka paljon parempi käytetty otantamenetelmä oli verrattuna yksinkertaiseen satunnaisotantaan. Lehtosen ja Pahkisen (1994) mukaan otantamenetelmän tehokkuutta mitataan tunnusluvulla $deft$, tehokkuuskerroin. Tunnusluku $deft$ mittaa käytetyn otantamenetelmän estimaattorin keskihajontaa yksinkertaisen satunnaisotannan vastaavaan keskihajontaan. Tässä tutkittava estimaattori on keskiarvo. Vertailumitta on siis suhdeluku:

$$deft = \frac{d(\hat{y}_{p(s)})}{d(\hat{y}_{SRS})},$$

jossa jakajana on yksinkertaisella satunnaisotannalla poimitun n alkion otoksen estimaattorin keskihajonta. Lisäksi yksinkertaisen satunnaisotannan poiminta on suoritettu palauttaen –tyyppisesti. Jaettavana on vertailtavan otantamenetelmän p(s) estimaattorin keskihajonta laskettuna myös n alkion otoksesta. Saatu suhde kuvaa, kuinka mones osa on vertailtavan otantamenetelmän keskihajonta yksinkertaiseen satunnaisotannan vastaavasta keskihajonnasta. Mitä pienempi suhde sitä tehokkaampi on käytetty otantamenetelmä ollut yksinkertaiseen satunnaisotantaan verrattuna.
$deft$ –suhdeluvulle on voimassa:
- $deft > 1$, käytetty otantamenetelmä on tehottomampi kuin yksinkertainen satunnaisotanta
- $deft = 1$, käytetty otantamenetelmä on yhtä tehokas kuin yksinkertainen satunnaisotanta
- $deft < 1$, käytetty otantamenetelmä on tehokkaampi kuin yksinkertainen satunnaisotanta

Yleensä $deft < 1$. Muutoin monimutkasten otanta-asetelmien käytössä ei olisi järkeä. Tässä tutkimuksessa käytetty otantamenetelmä muistuttaa lähinnä tasakiintiöityä ositettua otantaa. Suhdeluku $deft$ keskiarvolle on tässä tapauksessa:

$$deft = \sqrt{\frac{v_{equ}(\overline{y}_{str})}{v(\overline{y}_{SRS*})}}, \tag{5}$$

jossa

$$v_{equ}(\overline{y}_{str}) = (H/n)\sum_{h=1}^{H} W_h^2 s_h^2 - (1/N)\sum_{h=1}^{H} W_h s_h^2 \tag{6}$$

on keskiarvoestimaatin varianssi tasakiintiöidylle ositetulle otannalle ja

$$v(\overline{y}_{SRS*}) = (1 - n/N)s^2/n \tag{7}$$

on keskiarvoestimaatin varianssi palattamatta –tyyppiselle yksinkertaiselle satunnaisotannalle. Lisäksi perusjoukon varianssin estimaatti s^2 lasketaan kaavalla:

$$s^2 = \sum_{i=1}^{n} \frac{(y_i - \overline{y})^2}{n-1} \tag{8}$$

ja ositetun otannan tilanteessa perusjoukon varianssin estimaatti s_h^2 lasketaan kaavan (2) avulla.

2.1 Kotitalouden koon keskiarvo

Tämä tutkimus on tehty puhelinhaastatteluna ja puhelinnumerot on valittu puhelinluettelosta. Tämä voi tuottaa ongelmia. Matkapuhelimet ovat yleistyneet huomattavasti ja varsinkin nuorten ihmisten keskuudessa matkapuhelin on ainoa puhelin. Heitä on vaikea tai lähes mahdoton saada mukaan tutkimukseen tällä keruumenetelmällä. Jyväskylässä on useita oppilaitoksia, joihin tullaan opiskelemaan hyvinkin kaukaa ja oppilaitoksissa on paljon opiskelijoita. Opiskelijan hankintalistassa ei ole enää lankapuhelin vaan käsipuhelin.

Yritetään tarkentaa tuloksia käyttämällä otantateoriaa. Otantamenetelmänä on käytetty ositettua otantaa. Ositettu otanta suoritetaan siten, että ensin valitaan sopivat ositteet. Tässä tutkimuksessa ositteita ovat asuinalueet. Tällöin siis täytyy olla ennakkotietoa, jonka avulla perusjoukon alkiot voidaan luokitella toisensa poissulkeviin osajoukkoihin. Ositetussa otannassa täytyy myös päättää, kuinka otoskoot ositteiden välillä tullaan määräämään. Tässä tutkimuksessa on käytetty kiintiöintimenetelmä muistuttaa lähinnä tasakiintiöintiä. Keskustasta, Vaajakoskelta ja Muuramesta on kuitenkin poimittu 40 kotitaloutta ja Säynätsalosta vain 20 kotitaloutta. Muutoin jokaisesta ositteesta on poimittu 30 kotitaloutta. Ositteen sisällä kotitaloudet on poimittu otokseen yksinkertaisella satunnaisotannalla.

Lasketaan kotitalouden koon keskiarvo. Tehdään painotus käyttämällä kotitalouksien määrää. Ositekohtaisen keskiarvon estimaatit lasketaan kaavan (1) avulla. Taulukossa 3 on tulokset ja taulukossa 3 oleva paino on kotitalouksien määrästä laskettu ositepaino $W_h = N_h/N$

Taulukko 3. Kotitalouden keskimääräisen koon laskemiseen tarvittavia lukuja

Alue	Otoska.	Paino, W	Estimaatti
Keskusta	2.18	0.2389	0.520133
Kortepohja	2.37	0.0926	0.219441
Keltinmäki	2.60	0.0597	0.155314
Keljo	3.63	0.0272	0.098855
Kuokkala	3.17	0.0987	0.313007
Aittorinne	2.60	0.0446	0.116029
Kangasvuori	2.73	0.0775	0.211524
Lohikoski	3.10	0.0381	0.118190
Palokka	3.33	0.0689	0.229345
Vaajakoski	2.95	0.0756	0.222868
Muurame	3.38	0.0483	0.163308
Säynätsalo	3.05	0.0264	0.080380
Laukaa	3.00	0.1038	0.311348

Kotitalouden koon keskiarvon estimaatiksi saadaan kaavaa (3) käyttämällä 2.76. Keskiarvon estimaatti yksinkertaisen satunnaisotannan tilanteessa on 2.92. Keskiarvot poikkeavat kaupungin rekisteristä saadusta keskiarvosta 2.1 ja tilastokeskuksen ilmoittamasta Väli-Suomen kotitalouden keskikoosta 2.2. Ero voi osaltaan johtua siitä, että otantalähteenä on käytetty puhelinluetteloa, jossa ei käsipuhelimia ole. Keskiarvon estimaatti ositetun otannan tilanteessa antaa tarkemman arvion kuin satunnaisotannan estimaatti.

Lasketaan käytetylle otantamenetelmälle $deft$. Kerroin $deft$ kuvaa, kuinka hyvin käytetty otantamenetelmä on onnistunut. Kerroin $deft$ ilmoittaa keskiarvojen keskihajontojen suhteen. Käyttäen kaavoja (5) – (8) saadaan laskettua $deft$–kerroin tässä tutkimuksessa käytetylle otanta–asetelmalle. Tällöin $deft$–kertoimen arvoksi saadaan 1.08. Tässä tilanteessa ositetun otannan käyttö ei paranna tulosten tarkkuutta muutoin kuin että keskiarvon piste–estimaatti on lähempänä oikeaa arvoa.

Keskiarvon keskivirhe vaikuttaa suoraan keskiarvon luottamusvälin laske-

miseen. Luottamusvälit keskiarvoille ovat:
SRS : 2.92±1.96*0.06300 = [2.797,3.044]
STR : 2.76±1.96*0.06828 = [2.626,2.894]

Molempien otantamenetelmien estimaattien luottamusväli kotitalouden koon keskiarvolle on varsin leveä. Ositetun otannan mukainen luottamusväli on hiukan leveämpi, joka ilmenee myös $deft$–kertoimen arvosta. Ikävä puoli on, että parametri ei kuulu kummankaan estimaatin luottamusväliin. Näinollen estimointi on ollut harhaista. Otokseen on tullut liian suuria kotitalouksia ja otos ei edusta perusjoukkoa tällä perusteella hyvin. Käytettyä otantalähdettä kannattaisi modifioida esimerkiksi siten, että otetaan osa otosalkioista käsipuhelinluettelosta.

Ositetun otannan käytön perusteltu vahvuus tässä tilanteessa on se, että jokaisen asuinalueen edustavuus otoksessa varmistetaan.

3 Eksponentiaalinen jakaumaperhe ja yleistetyt lineaariset mallit

Aineiston analyysissä käytetyt menetelmät perustuvat yleistettyihin lineaarisiin malleihin. Yleistettyjen lineaaristen mallien perusidea on tutkia altisteen X (vaaratekijä, selittäjä, riippumaton muuttuja, syy) vaikutusta vasteeseen Y (selitettävä muuttuja, riippuva muuttuja, seuraus). Yleensä kuitenkin selitettävä muuttuja Y ei noudata normaalijakaumaa, usein vastemuuttuja on jopa luokiteltu. Lisäksi selitettävän muuttujan Y ja selittävien muuttujien X suhde ei välttämättä ole lineaarinen. Tässä kappaleessa tullaan esittelemään eksponentiaalinen jakaumaperhe ja yleistetyt lineaariset mallit. Näiden avulla pystytään analysoimaan aineistoja, jonka muuttujat eivät välttämättä ole jatkuvia ja normaalijakautuneita. Kappaleessa 3 on päälähteenä käytetty Dobsonia (1983).

3.1 Eksponentiaalinen jakaumaperhe

Useimmissa tilastollisissa analysointimenetelmissä oletetaan vastemuuttujan olevan normaalijakautunut. Näin ei kuitenkaan aina ole, koska vastemuuttuja voi olla luokiteltu tai jopa kaksiarvoinen. Yleistetyt lineaariset mallit soveltuvat tilanteeseen, jossa vastemuuttujan jakauma kuuluu eksponentiaaliseen jakaumaperheeseen.

Olkoon vastemuuttuja Y, jonka jakauma riippuu vain yhdestä parametrista θ ja kuuluu eksponentiaaliseen jakaumaperheeseen, jonka tiheysfunktio voidaan Dobsonin (1983) mukaan kirjoittaa muodossa:

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)}, \tag{9}$$

jossa a, b, s ja t ovat tietyt ehdot täyttäviä, tunnettuja funktioita. Kaava (9) voidaan kirjoittaa muodossa:

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)], \tag{10}$$

jossa $s(y) = \exp d(y)$ ja $t(\theta) = \exp c(\theta)$.
Jos $a(y) = y$, niin tiheysfunktion sanotaan olevan kanonista muotoa ja $b(\theta)$

on jakauman luonnollinen parametri.

Monet tunnetut ja paljon käytetyt jakaumat kuuluvat eksponentiaaliseen jakaumaperheeseen. Esimerkiksi Poisson–, normaali– ja binomijakauma voidaan kirjoittaa kanonisessa muodossa:

Poisson–jakauma

$$f(y;m) = \frac{m^y e^{-m}}{y!} = \exp[y \ln m - m - \ln y!],\ y = 0, 1, 2, \ldots$$

Normaalijakauma $Y \sim \mathrm{N}(\mu, \sigma^2)$

$$f(y;\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp[\frac{1}{2\sigma^2}(y-\mu)^2]$$
$$= \exp[-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)]$$

jossa parametri μ on kiinnostuksen kohteena ja σ^2 on tunnettu.
Binomijakauma, $Y \sim \mathrm{Bin}(n, \pi)$

$$f(y;\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-\pi}$$
$$= \exp[y \ln \pi - y \ln(1-\pi) + n \ln(1-\pi) + \ln \binom{n}{y}],$$

jossa $y = 0, 1, \ldots, n$.

Eksponentiaalisen jakaumaperheen eräs ominaisuus on, että suurimman uskottavuuden yhtälöä $\partial l / \partial \underline{\theta} = \underline{0}$ ratkaistaessa maksimi löytyy yksikäsitteisesti.

3.2 Yleistetyt lineaariset mallit

Tässä kappaleessa yleistetään tavallinen lineaarinen malli:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon},$$

jossa $\underline{y}$ on selitettävien muuttujien $(n * 1)$ vektori, X on selittävien muuttujien $(n * p)$ matriisi, $\underline{\beta}$ on estimoitavien parametrien $(p * 1)$ vektori ja $\underline{\epsilon}$ on

satunnaisjäännösten $(n * 1)$ vektori ja $\epsilon_i \sim NID(0, \sigma^2)$.

Olkoot riippumattomat satunnaismuuttujat $Y_1, \ldots, Y_n$ ja jokainen Y_i noudattaa eksponentiaaliseen jakaumaperheeseen kuuluvaa jakaumaa seuraavin oletuksin:
- Y_i:n jakauma on kanonista muotoa ja riippuu ainoastaan yhdestä parametrista θ_i siten, että

$$f(y_i; \theta_i) = \exp[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)]$$

- Kaikkien Y_i, i = 1, ... ,n jakauma on samaa muotoa. Tällöin otoksen $Y_1, \ldots, Y_n$ yhteistiheysfunktio on:

$$f(y_1, \ldots, y_n; \theta_1, \ldots, \theta_n) = \exp[\sum y_i b(\theta_i) + \sum c(\theta_i) + \sum d(y_i)]$$

Mallitustilanteessa parametrit θ_i eivät ole suoran kiinnostuksen kohteena, koska jokaista havaintoa kohden on yksi θ_i. Kiinnostuksen kohteena on pienempi joukko parametreja $\beta_1, \ldots, \beta_p$ (p<n), siten, että parametrien β_p lineaarikombinaatio on odotusarvon μ_i jokin funktio:

$$g(\mu_i) = \underline{x}_i^T \underline{\beta},$$

jossa
- g on monotoninen ja differentoituva funktio, jota kutsutaan linkkifunktioksi.
- $\underline{x_i}$ on (p*1) selittävien muuttujien vektori.
- $\underline{\beta}$ on (p*1) vektori kiinnostuksen kohteena olevista parametreista

$$\underline{\beta} = \begin{bmatrix} \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix}$$

Tällöin yleistetyssä lineaarisessa mallissa on Dobsonin (1983) mukaan kolme komponenttia:
- selitettävien muuttujien $Y_1, \ldots, Y_n$ oletetaan noudattavan samaa eksponentiaaliseen perheeseen kuuluvaa jakaumaa.

- Joukko parametreja $\underline{\beta}$ ja selittäviä muuttujia

$$X = \begin{bmatrix} \underline{x}_1 \\ . \\ . \\ . \\ \underline{x}_n \end{bmatrix}$$

- Linkkifunktio g, siten että

$$g(\mu_i) = \underline{x}_i^T \underline{\beta},$$

jossa $\mu_i = E(Y_i)$.

Mallitustilanteessa on selvitettävä selittävien muuttujien ja selitettävän muuttujan yhteys. Tätä yhteyttä kutsutaan linkkifunktioksi. Linkkifunktion valintaan vaikuttavat riippuvuuden laatu, selittävien muuttujien laatu (jatkuva, luokiteltu, dikotominen muuttuja, laskettavuus, luonnollisuus).
Yleisimpiä linkkifunktioita:
1) identtinen linkki: $g(\mu) = \mu$ (esimerkiksi normaalijakaumavaste)
2) logaritmilinkki: $g(\mu) = \log(\mu)$ (esimerkiksi Poissonvaste)
3) logitlinkki: $g(\mu) = logit(\mu) = \log(\mu/(1-\mu))$ (esimerkiksi binomivaste)
4) probitlinkki $g(\mu) = \Phi^{-1}(\mu)$
5) log–log-linkki $g(\mu) = \log(-\log(1-\mu))$

Esimerkiksi tavallinen lineaarinen malli:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon},$$

jossa $\underline{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^T$ ja $\epsilon_i \sim NID(0, \sigma^2)$, i=1, ... ,n on yleistettyjen lineaaristen mallien erikoistapaus, jossa linkkifunktio g on $g(\mu_i) = \mu_i$. Vektorin $\underline{Y}$ alkiot Y_i ovat riippumattomia ja $Y_i \sim N(\mu_i, \sigma^2)$, jossa $\mu_i = \underline{x}_i^T \underline{\beta}$. Lisäksi normaalijakauma kuuluu eksponentiaaliseen jakaumaperheeseen.

Binomijakauman tilanteessa, $Y_i \sim Bin(n, \pi_i)$, käytetään linkkifunktiona luonnollista logaritmia:

$$ln(\frac{\pi_i}{1-\pi_i}) = \underline{x}_i^T \underline{\beta},$$

jossa

$$\pi_i = \frac{exp(\underline{x}_i^T \underline{\beta})}{1 + exp(\underline{x}_i^T \underline{\beta})}$$

3.3 Estimointi

Tilastollisessa estimoinnissa kaksi käytetyintä menetelmää ovat suurimman uskottavuuden ja pienimmän neliösumman menetelmät. Tärkein ero niiden välillä on, että suurimman uskottavuuden menetelmä nojautuu jakaumaoletukseen. Pienimmän neliösumman menetelmää voidaan käyttää ilman Y:n jakaumaoletusta ja kovarianssirakenteen tuntemusta. Kuitenkin, jos halutaan tietää parametrivektorin $\underline{\beta}$ jakaumasta, esimerkiksi silloin, kun halutaan testata $\underline{\beta}$:hin liittyviä hypoteeseja, täytyy tehdä oletuksia myös Y_i:stä. Käytännössä on vain pieni etu pienimmän neliösumman estimoinnilla ja usein vielä suurimman uskottavuuden estimaattori on helpompi laskea tietokoneella.

Pienimmän neliösumman menetelmä

Olkoon $Y_1, \ldots, Y_n$ satunnaismuuttujia odotusarvoilla

$$E(Y_i) = \mu_i = \mu(\underline{\beta})\, i = 1, \ldots, n,$$

jossa $\underline{\beta} = [\beta_1, \ldots, \beta_p]^T$ $(p<n)$ ovat p kappaletta estimoitavia parametreja. Olkoon yhtälö:

$$Y_i = \mu_i + \epsilon_i,\ i = 1, \ldots, n,$$

jossa μ_i on Y_i:n systemaattinen osa ja ϵ_i on virhetermi. Pienimmän neliösumman menetelmän idea on löytää estimaattorit $\underline{\beta}$, jotka minimoivat virhetermien neliösumman:

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [Y_i - \mu_i(\underline{\beta})]^2.$$

Matriisimerkinnöin tämä voidaan merkitä:

$$S = (\underline{y} - \underline{\mu})^T(\underline{y} - \underline{\mu}),$$

jossa $\underline{y} = [Y_1, \ldots, Y_n]^T$ ja $\underline{\mu} = [\mu_1, \ldots, \mu_n]$. Yleisesti estimaattori $\underline{\beta}$ saadaan derivoimalla S jokaisen β_j:n suhteen ja ratkaisemalla yhtäaikaisesti p kappaletta yhtälöitä:

$$\frac{\partial S}{\partial \beta_j} = 0, \; j = 1, \ldots, p.$$

On tarpeen tarkastaa, että saadut estimaatorit $\underline{\beta}$ ovat globaaleja minimejä siten, että toisten derivaattojen matriisi on positiivisti definiitti.
Käytännössä on usein tarpeen painottaa havaintoja. Olkoon jokaisella Y_i paino w_i, $i = 1, \ldots, n$. Tällöin painotettu pienimmän neliösumman estimaattori on:

$$S_w = \sum_{i=1}^{n} w_i(Y_i - \mu_i)^2,$$

jossa w_i on paino. Esimerkiksi $w_i = [var(Y_i)]^{-1}$ Painona voidaan käyttää ja käytetäänkin yleisesti vektorin $\underline{Y}$ kovarianssimatriisin alkioita. Tällöin painotettu pienimmän neliösumman estimaattori käyttämällä S:n sijasta S_w:tä. saadaan:

$$S_w = (\underline{y} - \underline{\mu})^T V^{-1}(\underline{y} - \underline{\mu}).$$

Suurimman uskottavuuden menetelmä

Olkoon $Y_1, \ldots, Y_n$ n kappaletta satunnaismuuttujia, joiden yhteistiheysfunktio on $f(y_1, \ldots, y_n; \theta_1, \ldots, \theta_n)$, joka riippuu parametreista $\theta_1, \ldots, \theta_n$. Merkitään $[y_1, \ldots, y_n]^T = \underline{y}$ ja vastaavasti $[\theta_1, \ldots, \theta_n]^T = \underline{\theta}$. Uskottavuusfunktio $L(\underline{\theta}; \underline{y})$ on sama kuin $f(\underline{y}; \underline{\theta})$, mutta uskottavuusfunktiossa satunnaismuuttuja on vektori $\underline{\theta}$ ja vektori $\underline{y}$ on kiinteä. Olkoon Ω kaikkien parametrivektorin $\underline{\theta}$ mahdolliset arvot. Tällöin $\underline{\theta}$:n suurimman uskottavuuden estimaattori on vektori $\hat{\underline{\theta}}$ ja sille on voimassa:

$$L(\hat{\underline{\theta}}; \underline{y}) \geqslant L(\underline{\theta}; \underline{y}), \forall \underline{\theta}, \in \Omega$$

Yleensä uskottavuusfunktiosta otetaan logaritmi, koska logaritminen uskottavuusyhtälö on helpompi derivoida ja ratkaista. Olkoon $l(\hat{\underline{\theta}}; \underline{y}) = \ln L(\hat{\underline{\theta}}; \underline{y})$. Tällöin on voimassa:

$$l(\hat{\underline{\theta}}; \underline{y}) \geqslant l(\underline{\theta}; \underline{y}), \forall \theta \in \Omega$$

Tavallisin tapa löytää suurimman uskottavuuden estimaattori on tutkia kaikki $l(\underline{\theta}; \underline{y})$:n paikalliset maksimit. Paikalliset maksimit saadaan derivoimalla logaritminen uskottavuusfunktio kunkin θ_j:n, $j = 1, \ldots, p$ suhteen, asettamalla yhtälöt nollaksi ja ratkaisemalla ne:

$$\frac{\partial l(\underline{\theta}; \underline{y})}{\partial \theta_j} = 0, \quad j = 1, \ldots, p,$$

siten, että $\underline{\theta} \in \Omega$ ja toisten derivaattojen matriisi

$$\frac{\partial^2 l(\underline{\theta}; \underline{y})}{\partial \theta_j \partial \theta_k}$$

on negatiivisesti definiitti. Suurimman uskottavuuden estimaatti $\hat{\underline{\theta}}$ on paikallisista maksimeista suurin. Tärkeä tulos koskien suurimman uskottavuuden estimaattoria on, että jos $\psi(\underline{\theta})$ on jokin parametrin $\underline{\theta}$ funktio, niin suurimman uskottavuuden estimaattori ψ:lle on $\hat{\psi} = \psi(\hat{\underline{\theta}})$.

Olkoon riippumaton otos $Y_i, \ldots, Y_n$ ja estimoitava parametrivektori on $\underline{\beta}$. Logaritminen uskottavuusfunktio on:

$$l(\underline{\theta}; \underline{y}) = \sum y_i b(\theta_i) + \sum c(\theta_i) + \sum d(y_i),$$

jossa Dobsonin (1983) mukaan

$$E(Y_i) = \mu_i = -c'(\theta_i)/b'(\theta_i),$$

ja linkkifunktio g on monotoninen ja differentioituva:

$$g(\mu_i) = \underline{x}_i^T \underline{\beta} = \underline{\eta_i}$$

Käytännössä suurimman uskottavuuden yhtälöiden ratkaiseminen tapahtuu tietokoneella iteratiivisesti. Esitellään tässä kaksi eri menetelmää: Newton–Raphson– ja Fisherin scoring–menetelmä.

Olkoon tulosmuuttuja $\underline{Y} = [Y_1, \ldots, Y_n]^T$ ja estimoitava parametrivektori $\underline{\beta} = [\beta_1, \ldots, \beta_p]^T$. Newton–Raphson –menetelmällä saatava askeleen $k+1$ suurimman uskottavuuden estimaatti saadaan:

$$\underline{\hat{\beta}}_{k+1} = \underline{\hat{\beta}}_k - \left[D^2 l(\underline{\hat{\beta}}_k; \underline{y})\right]^{-1} Dl(\underline{\hat{\beta}}_k; \underline{y}),\ k = 1, 2, \ldots$$

jossa $\underline{\hat{\beta}}_0$ on alkuarvaus. Tämä iterointialgoritmi on hankala toteuttaa ja jos $D^2 l(\underline{\hat{\beta}}_k; \underline{y})$ matriisi ei ole positiivisti definiitti, käänteismatriisia ei voida laskea ja algoritmia ei voida käyttää.

Fisherin scoring–algoritmin idea on korvata Newton–Raphson–algoritmissa $-D^2 l(\underline{\hat{\beta}}; \underline{y})$ informaatiomatriisilla $I(\underline{\hat{\beta}})$. Tällöin käännettävä matriisi ei riipu $\underline{Y}$:stä ja on helpompi kääntää. Informaatiomatriisi on muotoa:

$$I(\underline{\beta}) = E\left[-\frac{\partial^2 l(\underline{\beta}; \underline{y})}{\partial \beta_i \partial \beta_j}\right],\ i, j = 1, \ldots, n$$

Tällöin Fisherin scoring–algoritmi antaa $k+1$ askeleen suurimman uskottavuuden estimaattorin seuraavasti:

$$\underline{\hat{\beta}}_{k+1} = \underline{\hat{\beta}}_k - \left[I(\underline{\hat{\beta}})_k\right]^{-1} Dl(\underline{\hat{\theta}}_k; \underline{y}),\ k = 1, 2, \ldots$$

jossa $\underline{\hat{\beta}}_0$ on alkuarvaus.

Esimerkki.

Olkoon selitettävä muuttuja $Y_i \sim Poisson(\lambda_i)$, $i = 1, \ldots, n$ ja havaintoon Y_i liittyvä kovariaatti x_i. Oletetaan malliksi (log–lineaarinen malli):

$$ln(\lambda_i) = \alpha + \beta x_i.$$

Johdetaan Newton–Raphson– ja Fisherin scoring–algoritmit parametrivektorin $(\underline{\alpha}, \underline{\beta})$ suurimman uskottavuuden estimaattoreiden laskemiseksi.

$$L(\underline{y}; \underline{\lambda}) = \frac{\Pi \lambda_i^{y_i} e^{\lambda_i}}{\sum y_i!},$$

$$l(\underline{y}; \underline{\lambda}) = \sum y_i \ln \lambda_i - \sum \lambda_i + \sum \ln 1/y_i!$$
$$= \sum y_i(\alpha + \beta x_i) - \sum e^{\alpha+\beta x_i} + \sum 1/y_i!$$

$$\frac{\partial l}{\partial \alpha} = \sum y_i - \sum e^{\alpha+\beta x_i}, \quad \frac{\partial^2 l}{(\partial \alpha)^2} = -\sum e^{\alpha+\beta x_i}$$
$$\frac{\partial l}{\partial \beta} = \sum x_i y_i - \sum x_i e^{\alpha+\beta x_i}, \quad \frac{\partial^2 l}{(\partial \beta)^2} = -\sum x_i^2 e^{\alpha+\beta x_i}$$

ja

$$E\left[\frac{\partial^2 l}{(\partial \alpha)^2}\right] = -\sum e^{\alpha+\beta x_i} \quad E\left[\frac{\partial^2 l}{(\partial \beta)^2}\right] = -\sum x_i^2 e^{\alpha+\beta x_i}$$
$$E\left[\frac{\partial^2 l}{\partial \alpha \partial \beta}\right] = -\sum x_i e^{\alpha+\beta x_i} \quad E\left[\frac{\partial^2 l}{\partial \beta \partial \alpha}\right] = -\sum x_i e^{\alpha+\beta x_i}$$

Newton–Raphson–algoritmi:

$$\underline{\hat{\lambda}}_{i_{k+1}} = \underline{\lambda}_{i_k} - \begin{bmatrix} -\sum e^{\alpha+\beta x_i} & -\sum x_i e^{\alpha+\beta x_i} \\ -\sum x_i e^{\alpha+\beta x_i} & -\sum x_i^2 e^{\alpha+\beta x_i} \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i - \sum e^{\alpha+\beta x_i} \\ \sum y_i x_i - \sum x_i e^{\alpha+\beta x_i} \end{bmatrix}$$

Fisherin scoring–algoritmi on tässä tapauksessa sama kuin Newton-Raphson–algoritmi.

4 Analyysimenetelmiä

Tässä luvussa tullaan selvittämään analysointimenetelmiä, joita voidaan käyttää tutkittavalle aineistolle. Tutkimusaineistossa on paljon luokiteltuja muuttujia, joten suurin osa menetelmistä on luokitellun aineiston analyysimenetelmiä. Regressioanalyysi on ainoa, jossa on mukana vain jatkuvia muuttujia. Muita menetelmiä ovat varianssianalyysi, Kruskall–Wallisin yksisuuntainen varianssianalyysi, logistinen regressio, loglineaariset mallit ja logit–mallit.

4.1 Regressioanalyysi

Yksinkertainen regressiomalli on muotoa:

$$\underline{y} = \underline{\beta} X + \underline{\epsilon} \tag{11}$$

jossa selittäviä muuttujia on p kappaletta. Koska $E[\underline{y}] = X\underline{\beta}$, niin linkkifunktio g on $g(\mu) = \mu$.

Oletetaan, että matriisin X aste on p, jotta $p * p$ matriisi $X^T X$ on eisingulaarinen. Tällöin suurimman uskottavuuden estimaattori $\underline{\beta}$:lle saadaan ratkaisemalla yhtälö: $X^T X\underline{b} = X^T\underline{y}$:

$$\underline{b} = (X^T X)^{-1} X^T \underline{y},$$

joka on myös $\underline{\beta}$:n pienimmän neliösumman estimaattori, kun $y_i \perp\!\!\!\perp y_j$, kaikille $i, j = 1, \ldots, n, i \neq j$ ja niillä on sama varianssi. $E[\underline{b}] = \underline{\beta}$ ja $\underline{b}$ on lineaarikombinaatio normaalijakautuneista satunnaismuuttujista Y_i jolloin:

$$\underline{b} \sim N(\underline{\beta}, \sigma^2 (X^T X)^{-1}). \tag{12}$$

Yleistetyissä lineaarisissa malleissa termiä σ^2 kohdellaan vakiona. Harhaton estimaattori σ^2:lle saadaan kaavalla:

$$\hat{\sigma}^2 = \frac{1}{n-p} (\underline{y} - X\underline{\beta})^T (\underline{y} - X\underline{\beta}) \tag{13}$$

Käyttämällä kaavoja (12) ja (13) voidaan johtaa $\underline{\beta}$:n luottamusvälit ja testata $\underline{\beta}$:n liittyviä hypoteeseja.

Jos $E[\underline{y}] = X\underline{\beta}$ ja $E[(\underline{y} - X\underline{\beta})(\underline{y} - X\underline{\beta})]^T = V$, jossa V on tunnettu ja V on diagonaalimatriisi, jonka alkiot ovat $v_{ii} = \sigma_i^2$,$i = 1, \ldots, n$, niin voidaan Dobsonin (1983) mukaan laskea $\underline{\beta}$:n pienimmän neliösumman estimaattori tekemättä mitään oletuksia $\underline{y_i}$, $i = 1, \ldots, n$ jakaumista. Minimoidaan yhtälö:

$$S_w = (\underline{y} - X\underline{\beta})^T V^{-1} (\underline{y} - X\underline{\beta}).$$

Derivoidaan yhtälö, asetetaan se nollaksi ja ratkaistaan yhtälö $\underline{\beta}$:n suhteen:

$$\frac{\partial S_w}{\partial \underline{\beta}} = -2X^T V^{-1}(\underline{y} - X\underline{\beta}) = 0$$

$$\Rightarrow \underline{b} = (X^T V^{-1} X)^{-1} X^T V^{-1} \underline{y}$$

Ositetun otannan tilanteessa matriisi V sisältää alkioiden painot π_h, jossa $\pi_h = n_h/N_h$. Ositteen sisällä kaikkien alkioiden painot ovat samat.

R^2 ja betakertoimet

Olkoon malli $\underline{y} = X\underline{\beta} + \underline{\epsilon}$ ja ϵ_i, $i = 1, \ldots, n$ ovat riippumattomia ja $E[\epsilon_i] = 0$ sekä $var(\epsilon_i) = \sigma^2$ kaikille $i = 1, \ldots, n$. Tällöin pienimmän neliösumman estimaattori on:

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \underline{\epsilon}^T \underline{\epsilon} = (\underline{y} - X\underline{\beta})^T (\underline{y} - X\underline{\beta}).$$

S:n minimiarvo mallille on:

$$\hat{S} = (\underline{y} - X\underline{b})^T (\underline{y} - X\underline{b}) = \underline{y}^T \underline{y} - \underline{b}^T X^T \underline{y}$$

Tätä voidaan käyttää mallin sopivuuden kuvaamiseen. $\hat{S}$:n arvoa voidaan verrata Dobsonin (1983) mukaan yksinkertaisimpaan malliin: $E(y_i) = \mu$ kaikille i. Tämä malli voidaan kirjoittaa yleiseen muotoon: $E(\underline{y}) = X\underline{\beta}$ jos $\underline{\beta} = [\mu]$ ja $\underline{X} = 1$, vektori, joka sisältää ykkösiä. Näin ollen $X^T X = n$, $X^T \underline{y} = \sum y_i$ ja $\underline{b} = \hat{\mu} = \overline{y}$. Näitä vastaava pienimmän neliösumman estimaattori on:

$$S_0 = \underline{y}^T \underline{y} - n\overline{y}^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

S_0 on verrannollinen havaintojen varianssiin ja sitä pidetään pahimpana mahdollisena S:n arvona. Kaikkia $\hat{S}$:n arvoja voidaan siis verrata S_0:n arvoon. Käytetään erotusta:

$$S_0 - \hat{S} = \underline{b}^T X^T \underline{y} - n\overline{y}^2,$$

joka on mallin sopivuuden parannus mallille $E(\underline{y}) = X\underline{\beta}$. Tällöin

$$R^2 = \frac{S_0 - \hat{S}}{S_0} = \frac{\underline{b}^T X^T \underline{y} - n\overline{y}^2}{\underline{y}^T \underline{y} - n\overline{y}^2},$$

joka tulkitaan osuudeksi kokonaisvaihtelusta, jonka malli pystyy selittämään. Jos malli ei pysty selittämään vaihtelua yhtään paremmin kuin yksinkertaisin malli, niin $S_0 = \hat{S}$ ja tällöin $R^2 = 0$. Jos taas malli selittää vaihtelusta kaiken, niin $\underline{b} = \underline{y}$ ja $\underline{b}^T XT\underline{y} = \underline{y}^T \underline{y}$. Tällöin $R^2 = 1$. Yleisesti on voimassa $0 < R^2 < 1$, R^2 kutsutaan selitysasteeksi.

Hypoteesit

Testattaessa hypoteesia:
$H_0 : \hat{\beta}_j = \beta_j$

päädytään seuraavaan testisuureeseen:

$$t = \frac{\hat{\beta}_j - \beta_j}{s\sqrt{(X^T X)^{-1}_{jj}}}$$

ja estimaattoreiden 95 prosentin luottamusväli on
$\hat{\beta}_j = \pm 1.96 s\sqrt{(X^T X)^{-1}_{jj}}$

4.2 Varianssianalyysi

Varianssianalyysia käytetään tutkittaessa, eroavatko k ryhmää toisistaan odotusarvoiltaan. Mallina käytetään lineaarista mallia:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon} \quad \underline{\epsilon} \sim N(\underline{0}, \sigma^2 I),$$

jossa $\underline{y}$ ja $\underline{\epsilon}$ ovat (n*1) satunnaisvektoreita, X on (n*p) matriisi, joka sisältää vakioita, $\underline{\beta}$ on (1*p) vektori, joka sisältää parametreja ja I on yksikkömatriisi.

Tämä malli eroaa esimerkiksi regressiomallista siten, että asetelmamatriisi X sisältää ainoastaan vakioita, joiden avulla asetetaan halutut hypoteesit.

Koska satunnaistermin $\underline{\epsilon}$ oletetaan noudattavan normaalijakaumaa, saadaan uskottavuussuhteen logaritmifunktioksi:

$$l(\underline{\beta}, \underline{y}) = -\frac{1}{2\sigma^2}(\underline{y} - X\underline{\beta})^T(\underline{y} - X\underline{\beta}) - \frac{n}{2}\ln(2\pi\sigma^2)$$

ja suurimman uskottavuuden ja pienimmän neliösumman estimaattori $\underline{b}$ saadaan yhtälöstä:

$$X^T X\underline{b} = X^T\underline{y} \tag{14}$$

ANOVA:ssa on usein enemmän parametreja kuin riippumattomia yhtälöitä mallissa:

$$E(\underline{y}) = X\underline{\beta}$$

Tällöin $X^T X$ on singulaarinen ja ei ole olemassa yksikäsitteistä ratkaisua yhtälölle (14). Tällöin sanotaan, että $\underline{\beta}$ ei ole identifioituva. Jotta saataisiin ratkaisu, käytetään lisäoletuksia:

$$\begin{cases} X^T X\underline{b} = X^T\underline{y} \\ C\underline{b} = \underline{0} \end{cases}$$

Jossa C sisältää rajoituksia. Rajoituksina käytetään esimerkiksi käsittelyjen summan asettamista nollaksi. Olkoon asetelmassa kaksi käsittelyä α_j, $j = 1, \ldots, J$ ja β_k, $k = 1, \ldots, K$. Tällöin voitaisiin asettaa Dobsonin (1983) mukaan esimerkiksi

$$\sum_j^J \alpha_j = 0, \ \sum_k^K \beta_k = 0$$

$$\sum_j^J (\alpha_j\beta_k) = 0, \ k = 1, \ \ldots \ K, \ \sum_k^K (\alpha_j\beta_k) = 0, \ j = 1, \ \ldots \ J$$

Eli kaikkien päävaikutusten summat asetetaan nolliksi ja yhdysvaikutusten summat yli jokaisen käsittelyn vuorotellen asetetaan nollaksi.

Oletukset

Varianssianalyysin oletukset ovat tiukat. Jäännöstermejä koskevat oletukset ovat:
- normaalijakautuneisuus,
- samavarianssisuus ja
- riippumattomuus toisista jäännöksistä ja käsittelyistä.

Hypoteesit

Yksisuuntaisessa varianssianalyysissä hypoteesit koskevat eri ryhmien odotusarvoja. Onko käsittelyllä α_i, $i = 1, \ldots, k$, vaikutusta tulosmuuttujan y arvoon?
H_0: $\mu_1 = \mu_2 = \ldots = \mu_k \Leftrightarrow \alpha_1 = \alpha_2 = \ldots = \alpha_k$.
H_1: $\mu_l \neq \mu_m$ ainakin joillakin l ja m.
H_1: $\alpha_l \neq 0$ ainakin joillakin l.
Nollahypoteesit voidaan myös kirjoittaa matriisimuodossa CBD=0, jossa matriisit C ja D ovat eräänlaisia kontrastimatriiseja, jotka määrittelevät testattavan nollahypoteesin. Matriisin D avulla voidaan asettaa hypoteeseja toistomittausasetelmissa ja muulloin $D = I$. Matriisi B sisältää ryhmien ja mittauskertojen odotusarvot.

Olkoon asetelmassa kaksi ryhmää ja kaksi mittauskertaa, tarkastellaan matriisien C ja D määrittelyjä.
1. Ryhmän päävaikutus

$$H_0 : \begin{cases} \mu_{11} = \mu_{21} \Leftrightarrow \mu_{11} - \mu_{21} = 0 \quad (alkumittaus) \\ \mu_{12} = \mu_{22} \Leftrightarrow \mu_{12} - \mu_{22} = 0 \quad (loppumittaus) \end{cases}$$

Tällöin CBD=0 määriteltäisiin seuraavasti:

$$\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{bmatrix} I = \begin{bmatrix} \mu_{11} - \mu_{21} & \mu_{12} - \mu_{22} \end{bmatrix} = \underline{0}$$

2. Mittaukerran päävaikutus

$$H_0 : \begin{cases} \mu_{11} = \mu_{12} \Leftrightarrow \mu_{11} - \mu_{12} = 0 \\ \mu_{21} = \mu_{22} \Leftrightarrow \mu_{21} - \mu_{22} = 0 \end{cases}$$

Tällöin CBD=0 määriteltäisiin seuraavasti:

$$I \begin{bmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} \mu_{11} - \mu_{12} \\ \mu_{21} - \mu_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

3. Ryhmän ja mittauskerran yhdysvaikutus
$H_0 : \mu_{12} - \mu_{11} = \mu_{22} - \mu_{21} = \mu_{12} - \mu_{11} + \mu_{21} - \mu_{22} = 0$ Tällöin CBD=0 määritellään seuraavasti

$$\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} -\mu_{11} + \mu_{12} \\ -\mu_{21} + \mu_{22} \end{bmatrix} = \begin{bmatrix} -\mu_{11} + \mu_{12} + \mu_{21} - \mu_{22} \end{bmatrix} = 0$$

Matriiseja C ja D muokkaamalla voidaan konstruoida nollahypoteeseja myös useamman ryhmän ja mittauskerran koeasetelmille.

F–testi

Varianssianalyysin idea on jakaa aineiston vaihtelu eri komponentteihin. Yksisuuntaisessa varianssianalyysissä vaihtelu jaetaan ryhmien sisäiseen ja ryhmien väliseen vaihteluun.

$$SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n_h} (y_{ij} - \overline{y}_{i.})^2, \; df_W = n - k$$

$$SS_B = n_h \sum_{i=1}^{k} (\overline{y}_{i.} - \overline{y}_{..})^2, \; df_B = k - 1$$

Aineiston kokonaisvaihtelu $SS_T = SS_W + SS_B$.

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_h} (y_{ij} - \overline{y}_{..})^2, \; df_T = n - 1,$$

joissa n_h on ryhmän h otoskoko, k on ryhmien lukumäärä, $\overline{y}_{i.}$ on käsittelyn i ryhmäkeskiarvo, $\overline{y}_{..}$ on koko aineiston yleiskeskiarvo. Nämä neliösummat jaetaan vapausasteillaan, saadaan keskineliösummat:
$MS_W = SS_W / df_W$

$MS_B = SS_B/df_B$
$MS_T = SS_T/df_T$

$$F = \frac{MS_B}{MS_W} \sim F(k-1, n-k)$$

Suuret testisuuren F arvot johtavat nollahypoteesin hylkäämiseen. Tällöin liian suuri osa vaihtelusta sijaitsee ryhmien välillä, eikä ryhmien sisällä.

Parittaiset vertailut

Varianssianalyysin F–testi kertoo vain sen, onko ryhmien välillä eroja. Monivertailutesteillä saadaan selville ne käsittelyn tasot, joiden välillä esiintyy merkitsevä ero. Monivertailutestejä on kehitetty useita. Esitellään tässä yksi: LSD–menetelmä.
LSD–menetelmä on vanhin ja yksinkertaisin menetelmä. Se perustuu t–testiin. Määritellään niin sanottu pienin merkitsevä ero:

$$LSD = t_{\alpha/2}(n-k)\sqrt{MS_W(\frac{1}{n_i} + \frac{1}{n_j})}$$

jossa MS_W on ryhmien sisäinen keskineliösumma, n_i, n_j ovat ryhmien i ja j havaintojen määrä ja $t_{\alpha/2(n-k)}$ on t-jakauman kriittinen arvo merkitsevyystasolla α ja vapausastein $(n-k)$. Jos ryhmäkeskiarvoille pätee:

$$|\,\overline{y}_{i.} - \overline{y}_{j.}\,| > LSD$$

niin niiden välinen ero katsotaan merkitseväksi merkitsevyystasolla α.

LSD-menetelmän ongelma on merkitsevyystason valinta. Mikäli vertailtavien ryhmien määrä kasvaa, kasvaa myös nollahypoteesin hylkäämistodennäköisyys. Eräs tapa valita α on seuraava:

$$\alpha = \alpha_{F-testi/k}$$

eli varianssianalyysin F-testin merkitsevyystaso jaetaan ryhmien määrällä.

4.3 Kruskal–Wallisin yksisuuntainen varianssianalyysi

Kruskal–Wallisin yksisuuntainen varianssianalyysi järjestysasteikollisille muuttujille on erittäin käyttökelpoinen testi päätettäessä, onko k riippumatonta otosta peräisin samasta perusjoukosta. Testi tutkii otosten mediaanien eroja. Testi olettaa, että tutkittavalla ominaisuudella on perustana sama jatkuva jakauma ja ominaisuutta on mitattu vähintään järjestysasteikollisilla muuttujalla.

Testin suorittaminen

Aineisto esitetään kaksiulotteisessa taulussa siten, että sarakkeina ovat kunkin otoksen tai ryhmän peräkkäiset arvot. Seuraavaksi taulukossa olevat arvot asetetaan suuruusjärjestykseen siten, että pienin havainto saa arvokseen ykkösen, seuraava kakkosen ja suurin arvo saa arvon n, joka on kaikkien ryhmien alkioiden summa. Jos kaksi tai useampi arvo saa saman järjestysnumeron, täytyy näille arvoille laskea näiden järjestysnumeroiden keskiarvo. Kun kullekin otokselle tai ryhmälle on saatu laskettua järjestysnumerot, voidaan laskea jokaiselle ryhmälle järjestysnumeroiden summa. Näiden summien avulla lasketaan kullekin ryhmälle järjestyslukujen keskiarvot. Kruskal-Wallisin yksisuuntainen varianssianalyysi testaakin juuri näiden keskiarvojen eroja. Mikäli keskiarvot poikkeavat toisistaan, niin voidaan sanoa, että ryhmät eivät ole peräisin perusjoukoista, joilla on sama mediaani.

Hypoteesit

Kruskal–Wallisin yksisuuntaisen varianssianalyysin hypoteesit koskevat ryhmien mediaaneja:
$H_0 : \theta_1 = \theta_2 = \ \ldots \ = \theta_k.$
H_1 : Ainakin kaksi ryhmää eroavat toisistaan mediaaneiltaan.

Testisuure

Kruskall–Wallisin yksisuuntaisen varianssianalyysin testisuuretta merkitään KW, joka lasketaan kaavalla:

$$KW = \frac{12}{n(n+1)} \sum_{j=1}^{k} n_j(\overline{R}_j - \overline{R})^2,$$

jossa k on ryhmien tai otosten lukumäärä, n_j = ryhmässä j olevien havaintojen lukumäärä,
n on kaikkien havaintojen lukumäärä,
$\overline{R}_j$ on ryhmän j järjestysnumeroiden keskiarvo,
$\overline{R}$ = kaikkien ryhmien järjestysnumeroiden keskiarvo.

Usein käy niin, että aineistossa on sidoksia ja järjestysluvuille lasketaan keskiarvoja, on myös testisuuretta korjattava. Korjaamisen tarkoituksena on kasvattaa testisuureen arvoa ja siten tehdä tuloksista merkitsevämpiä kuin ne olisivat ilman korjausta. Yleinen testisuure korjatuille tasatulokselle on:

$$KW = \frac{[\frac{12}{n(n+1)} \sum_{j=1}^{k} n_j \overline{R}^2] - 3(n+1)}{1 - [\sum_{i=1}^{g} (t_i^3 - t_i)]/(n^3 - n)},$$

jossa g on tasatulosten lukumäärä, t_i on ryhmän i tasatulosten lukumäärä. Jos ryhmiä on enemmän kuin kolme ja jokaisessa ryhmässä on vähintään viisi havaintoa, testisuure noudattaa χ^2-jakaumaa vapausastein df = k-1, jossa k on ryhmien lukumäärä. Mikäli testisuure ylittää taulukoidun χ^2–arvon valitulla merkitsevyystasolla, H_0 hylätään.

Parittaiset vertailut

Kun testisuureen KW arvo on merkitsevä, niin ainakin yksi ryhmä eroaa muista ryhmistä. Testisuure ei kuitenkaan kerro sitä, mikä ryhmä poikkeaa muista. Tällöin halutaan testata hypoteeseja:
$H_0 : \theta_u = \theta_v$, joillekin ryhmille u ja v $(u \neq v)$.
$H_1 : \theta_u \neq \theta_v$.
Ensin lasketaan kaikkien ryhmien keskiarvojen erotusten itseisarvot:

$$| \overline{R}_u - \overline{R}_v |$$

Seuraavaksi määrätään kriittinen arvo:

$$z_{\alpha/k(k+1)} \sqrt{\frac{n(n+1)}{12} (\frac{1}{n_u} + \frac{1}{n_v})},$$

jossa z:n arvot saadaan normaalijakauman taulukosta ja $n = n_u + n_v$. Mikäli ryhmien keskiarvojen erotuksen itseisarvo on suurempi kuin kriittinen arvo, nollahypoteesi hylätään. Testattujen ryhmien välillä oli merkitsevä ero.

4.4 Logistinen regressio

Olkoon tulosmuuttuja Y dikotominen ja se saa arvot 0 tai 1 ja selittävänä muuttujana on jatkuva muuttuja X. Tällöin regressiomalli olisi:

$$E(Y) = \pi(x) = \alpha + \beta X,$$

jota kutsutaan lineaariseksi todennäköisyysmalliksi. Usein oletetaan, että muuttujalla X on epälineaarinen suhde todennäköisyyteen $\pi(x)$. Esimerkiksi auton oston todennäköisyydessä ostajan rahamäärän lisääntyminen 10000 markalla on huomattavasti merkitsevämpää, kun ostajalla on rahaa ennestään 10000 markaa kuin 200000. Tässä tapauksessa π on siis lähellä arvoa 0, kun $x = 10000$ ja π on lähellä arvoa 1, kun $x = 200000$. Tällaisten ongelmien välttämiseksi käytetään mallina logistista regressiota:

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

Tällöin vedonlyöntisuhdetta ennustetaan kaavalla:

$$\frac{\pi(x)}{1 - \pi(x)} = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}.$$

Vedonlyöntisuhde kasvaa multiplikatiivisesti arvolla e^{β}, kun x:n arvo kasvaa yhdellä yksiköllä. Vedonlyöntisuhteen logaritmilla on siis lineaarinen lauseke:

$$\ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = Q(x) = \alpha + \beta x.$$

Logistinen käyrä saadaan ratkaisemalla edellisestä kaavasta $\pi(x)$:

$$\pi(x) = \frac{1}{1 + \exp(-\alpha - \beta x)},$$

jossa parametri α siirtää käyrää ja parametri β vaikuttaa käyrän jyrkkyyteen ja suuntaan. Yleisessä tilanteessa, jossa selittäviä muuttujia on k kappaletta, logistinen regressiomalli on:

$$\ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \beta_1 x_1 + \ \ldots \ + \beta_k x_k.$$

4.5 Loglineaariset mallit

Loglineaarilla malleilla voidaan tutkia luokiteltujen muuttujien välisiä riippuvuuksia. Kahden luokitellun muuttujan välistä riippuvuutta voidaan tarkastella χ^2–testillä, mutta jos halutaan tarkastella useampaa muuttujaa yhtäaikaa, on syytä käyttää loglineaarisia malleja. Loglineaarisissa malleissa mallinnetaan kontingenssitaulun solujen frekvenssiä ja valitun mallin sopivuutta tarkastellaan mallin antaman odotetun frekvenssin ja havaitun frekvenssin eron suuruudella.

Kahden muuttujan riippumattomuuden tutkiminen tapahtuu yleensä Pearsonin χ^2–testillä. Testi perustuu kontingenssitaulun odotettujen ja havaittujen frekvenssien välisiin eroihin. Odotettu frekvenssi on luku, joka kuvaa kuinka paljon havaintoja kyseiseen soluun pitäisi otoksessa tulla, jos tutkittavat muuttujat olisivat riippumattomia. Testin nollahypoteesi on tutkittavien muuttujien riippumattomuus. Vastahypoteesina on, että muuttujat riippuvat toisistaan. Testin oletuksena on yksinkertaisella satunnaisotannalla poimittu otos. Toinen oletus koskee odotettuja frekvenssejä. Odotetut frekvenssit lasketaan kaavalla:

$$e_{ij} = \frac{f_{i+} f_{+j}}{n}, \tag{15}$$

jossa f_{i+} on rivin i rivisumma ja f_{+j} on sarakkeen j sarakesumma ja n on otoskoko. Oletusten mukaan yksikään odotetuista frekvensseistä ei saa olla pienempi kuin yksi. Lisäksi korkeintaan 20 prosenttia odotetuista frekvensseistä saa olla pienempiä kuin viisi. Jokaiselle solulle lasketaan odotettu frekvenssi, jota verrataan havaittuun frekvenssiin. Mikäli ero on suuri, kasvaa testisuureen arvo, joka voi hylätä nollahypoteesin. Testisuure lasketaan kaavalla:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{16}$$

Testisuure noudattaa $\chi^2(df)$–jakaumaa, jos nollahypoteesi on tosi. Vapausasteet df = (I-1)(J-1). Suuret testisuureen arvot aiheuttavat nollahypoteesin hylkäämisen.

Jos analyysissä on kolme tai useampi muuttuja, niin silloin käytetään loglineaarisia malleja. Toki loglineaarisia malleja voidaan käyttää myös kahden muuttujan riippuvuuksien tarkasteluun. Varsinkin tilanteessa, jossa halutaan tutkia muitakin hypoteeseja kuin pelkästään riippumattomuutta.

Loglineaarisissa malleissa odotettujen frekvenssien oletetaan noudattavan joko Poisson– tai multinomiaalijakaumaa. Poisson–jakaumaa käytetään silloin, kun otoskokoa ei ole ennalta kiinnitetty ja multinomiaalijakaumaa käytetään, kun otoskoko on kiinnitetty. Loglineaaristen mallien yhteydessä käytettäessä suurimman uskottavuuden estimointimenetelmää Poisson–malli ja multinomiaalimalli antavat saman tuloksen. Loglineaarisissa malleissa estimoitavien parametrien määrä kasvaa suureksi. Malliksi valitaankin usein yksinkertaisin malli, joka sopii aineistoon ja jolla on järkevä tulkinta.

Loglineaaristen ja logit–mallien estimointiin käytetään yleensä suurimman uskottavuuden estimointimenetelmää. Tällä menetelmällä on Bishopin (1975) mukaan seuraavia hyviä puolia:
1. SU –estimaatit on suhteellisen helppo laskea loglineaarisille malleille.
2. SU –estimaatit toteuttavat tietyt intuitiiviset marginaalirajoitteet, mikä ei ole ominaista muille estimointimenetelmille.
3. SU –estimointia voidaan soveltaa suoraan multinomiaalijakautuneelle aineistolle, jossa on useita nollasoluja. Näille soluille menetelmä tuottaa miltei aina nollasta poikkeavat estimaatit.
Näistä ominaisuuksista huolimatta Bishopin (1975) mielestä vaihtoehtoiset menetelmät ovat joskus käyttökelpoisia ja ehkä jopa sopivampia.

Vedonlyöntisuhteeksi kaksiarvoiselle muuttujalle sanotaan suhdetta, joka saadaan laskemalla tietyn tapahtuman todennäköisyys jaettuna sen komplementtitapahtuman todennäköisyydellä. Esimerkiksi vedonlyöntisuhde sille, että korttipakasta vedetään ässä on 4 / 52 : 48 / 52 = 1 / 12. 2*2 kontingenssitaulussa vedonlyöntisuhde riville 1 on:

$$\Omega_1 = \frac{\pi_{12}}{\pi_{11}}$$

Vastaavasti vedonlyöntisuhde riville 2 on:

$$\Omega_2 = \frac{\pi_{22}}{\pi_{21}}$$

Yhdistetty vedonlyöntisuhde, jota kutsutaan myös ristitulosuhteeksi, määritellään 2*2 kontingenssitaululle seuraavasti:

$$\theta = \frac{\Omega_2}{\Omega_1} = \frac{\pi_{22}\pi_{21}}{\pi_{12}\pi_{11}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

Loglineaarisissa malleissa assosiaatioparametrien ja ristitulosuhteen välillä on suora suhde. Suhde on yksinkertaisin 2*2 taululle. Täydelle mallille
$\ln = \ln(e_{11}e_{22}/e_{12}e_{21}) = \ln e_{11} + \ln e_{22} - \ln e_{12} - \ln e_{21}$
$= (\mu + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\mu + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY})$
$-(\mu + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) - (\mu + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY})$
$= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$
jossa μ on yleiskeskiarvo, λ_1^X ja λ_2^X ovat muuttujan X päävaikutuksia ja λ_1^Y ja λ_2^Y ovat muuttujan Y päävaikutuksia ja λ_{ij}^{XY} , $i = 1{,}2$ ja $j = 1{,}2$ ovat muuttujien X ja Y yhdysvaikutuksia.

Parametrien asettaminen

Loglineaarisissa malleissa on paljon estimoitavia parametreja. Parametreille täytyy asettaa joitain rajoituksia, jotta estimointi olisi mahdollista. Agrestin (1990) mukaan estimaatteja täydelle mallille voidaan laskea vain niin monta, kuin kontingenssitaulussa on soluja. Esimerkiksi 2*2 kontingenssitaululle voidaan laskea yleistaso μ , rivin ja sarakkeen päävakutukset λ_1^X ja λ_1^Y ja niiden yhdysvaikutus λ_{11}^{XY}. Rajoitteet valitaan joko mallin perusteella suoraan tai rajoitteet voi tehdä myös itse.

Loglineaarisilla malleilla voidaan tutkia useita nollahypoteeseja ja jos tutkittavana on epästandardi hypoteesi, niin tutkijan täytyy asettaa rajoitteet itse. Rajoitteiden asetus vaikuttaa asetelmamatriisiin, jonka avulla lasketaan testisuureita. Asetelmamatriisi määrätään hypoteesien perusteella. Rajoitteita voidaan asettaa useilla eri tavoilla. Agrestin (1990) mukaan parametrit ovat poikkeamia keskiarvosta, joten parametrit summautuvat nolliksi. Esimerkiksi 2*2 kontingenssitaululle Agrestin (1990) rajoitukset olisivat:

$$\sum_{i=1}^{2} \lambda_i^X = \sum_{j=1}^{2} \lambda_j^Y = \sum_{i=1}^{2} \lambda_{ij}^{XY} = \sum_{j=1}^{2} \lambda_{ij}^{XY} = 0$$

Loglineaarinen malli 2*2 kontingenssitaululle

Olkoon loglineaarisessa mallissa kaksi muuttujaa, jotka ovat kaksiluokkaisia. Tällöin odotetun frekvenssin logaritmi $\ln(m_{ij})$ täydellisen mallin tilanteessa rivin i ja sarakkeen j solulle on Agrestin (1990) mukaan

$$\ln(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY},$$

jossa μ on yleiskeskiarvo, λ_i^X on rivimuuttujan rivin i päävaikutus, λ_j^Y on sarakemuuttujan sarakkeen j päävaikutus ja λ_{ij}^{XY} on muuttujien yhdysvaikutus rivillä i ja sarakkeella j. 2*2 kontingenssitaululle täysi malli olisi

$\ln(m_{11}) = \mu + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}$
$\ln(m_{12}) = \mu + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}$
$\ln(m_{21}) = \mu + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}$
$\ln(m_{22}) = \mu + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}$

Mitä suurempi parametrin itseisarvo on, sitä merkitsevämpi se on mallissa. Riippumattomuusmallissa tutkitaan, ovatko muuttujat riippumattomia. Muuttujat ovat riippumattomia, jos solutodennäköisyydet π_{ij} voidaan laskea kertomalla muuttujien reunafrekvenssit keskenään, $\pi_{ij} = \pi_{i+}\pi_{+j}$. Odotetut frekvenssit m_{ij} saadaan jokaiselle solulle kaavalla $m_{ij} = n\pi_{ij}$. Agrestin (1990) mukaan riippumattomuusmalli 2*2 kontingenssitaululle on

$$\ln(m_{ij}) = \ln(n) + \ln(\pi_{ij}) + \ln(\pi_{+j}) = \mu + \lambda_i^X + \lambda_j^Y,$$

jossa

$$\lambda_i^X = \ln(\pi_{i+}) - (\sum_{h=1}^{2} \ln(\pi_{h+}))/I,$$

$$\lambda_j^X = \ln(\pi_{+j}) - (\sum_{h=1}^{2} \ln(\pi_{+h}))/J,$$

$$\mu = \ln(n) + (\sum_{h=1}^{2} \ln(\pi_{h+})/I + (\sum_{h=1}^{2} \ln(\pi_{+h})/J.$$

Edelleen Agrestin (1990) mukaan parametrit λ_i^X ja λ_j^Y summautuvat nolliksi:

$$\sum_{i=1}^{2} \lambda_i^X = \sum_{j=1}^{2} \lambda_j^Y = 0.$$

Edellisen kaavan mukainen oletus on yleinen tapa tehdä parametrit tulkittaviksi, mutta muutkin oletukset ovat mahdollisia. Sama tulos saadaan, jos esimerkiksi johonkin termiin lisätään vakio ja toisesta termistä vähennetään sama vakio.

Kaksiulotteisen taulun riippumattomuushypoteesin testaaminen suurimman uskottavuuden menetelmällä voidaan suorittaa seuraavasti:
Malli on $ln(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y$, i,j=1,2

Tällöin uskottavuussuhteen testi voidaan johtaa multinormaalijakauman uskottavuudesta:

$$f(n_{11}, n_{12}, n_{21}, n_{22}) = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\pi_{11}\pi_{12}\pi_{21}\pi_{22},$$

jossa n_{ij} on havaittu frekvenssi solussa i, j. Rajoittamaton suurimman uskottavuuden estimaattori π_{ij}:lle on $\pi_{ij} = n_{ij}/n$. Tällöin

$$max\ L = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\prod_{i=1}^{2}\prod_{j=1}^{2}\left(\frac{n_{ij}}{n}\right)^{n_{ij}}$$

Riippumattomuushypoteesin $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ vallitessa uskottavuus on:

$$L_0 = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2}$$

Suurimman uskottavuuden estimaattori π_{ij}:lle on $\pi_{ij} = n_{i+}n_{+j}/n^2$. Tällöin

$$max\ L_0 = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}\prod_{i=1}^{2}\prod_{j=1}^{2}\left(\frac{n_{i+}n_{+j}}{n^2}\right)^{n_{ij}}$$

Uskottavuussuhde on tällöin:

$$\lambda = \frac{max\ L_0}{max\ L} = \prod_{i=1}^{2}\prod_{j=1}^{2}\left(\frac{\hat{m}_{ij}}{nij}\right)^{n_{ij}},$$

jossa $\hat{m}_{ij} = n_{i+}n_{+j}/n$, uskottavuussuhteen testisuure G^2 on:

$$G^2 = -2\ln\lambda = 2\sum_{i=1}^{2}\sum_{j=1}^{2}n_{ij}\ln\left(\frac{n_{ij}}{\hat{m}_{ij}}\right)$$

G^2 noudattaa asympoottisesti χ^2 jakaumaa vapausasteella 1, jos H_0 on tosi.

Loglineaarinen malli kolmiulotteiselle kontingenssitaululle

Loglineaaristen mallien mahdollisuudet tulevat esiin useampiulotteisten kontingenssitaulujen tilanteessa. Mallit ovat periaatteeltaan samanlaisia kuin kaksiulotteisen taulun tilanteessa, mutta estimoitavien parametrien määrä kasvaa rajusti. Esimerkiksi $I * J * K$ -ulotteisessa kontingenssitaulussa täysi loglineaarinen malli olisi:

$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{ik}^{YZ} + \lambda_{ijk}^{XYZ}$,
jossa $i = 1, \ldots, I, j = 1, \ldots, J$ ja $k = 1, \ldots, K$.
Täydellisen riippumattomuuden testaaminen tehdään mallilla:
$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$,
jossa

$$m_{ijk} = \frac{n_{i++} n_{+j+} n_{++k}}{n^2},$$

jossa esimerkiksi n_{i++} on i:nnen rivin reunafrekvenssi ja näillä merkinnöillä $n = n_{+++}$.
Estimoitujen ja havaittujen frekvenssien yhteensopivuutta testataan uskottavuussuhteen testisuureella G^2. Testisuure noudattaa likimain χ^2 – jakaumaa vapausastein $df = IJK$ - I - J - K + 2, jossa I, J ja K ovat muuttujien X, Y ja Z luokkien lukumäärät.

Osittainen riippumattomuus

Mikäli täydellisen riippumattomuuden malli
$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$, $i = 1, \ldots, I, j = 1, \ldots, J$ ja $k = 1, \ldots, K$
hylätään, siirrytään testaamaan osittaisen riippumattomuuden hypoteeseja.
Kolmen luokittelevan tekijän tapauksessa mahdolliset nollahypoteesit ovat:
$H_0 : X \perp YZ$ eli muuttuja X on riippumaton tekijöistä Y ja Z,
$H_0 : Y \perp XZ$ eli muuttuja Y on riippumaton tekijöistä X ja Z,
$H_0 : Z \perp XY$ eli muuttuja Z on riippumaton tekijöistä X ja Y.
Vastaavat loglineaariset mallit ovat:
$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$,
$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$,
$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$,
jossa $i = 1, \ldots, I, j = 1, \ldots, J$ ja $k = 1, \ldots, K$.

Ehdollinen riippumattomuus

Ehdollisen riippumattomuuden tilanteessa tutkitaan, ovatko esimerkiksi tekijät X ja Y riippumattomia kullakin tekijän Z arvolla. Nollahypoteesit ehdollisessa riippumattomuudessa ovat:
$H_0 : X \overset{Z}{\perp} Y$: tekijät X ja Y ovat riippumattomia ehdolla $Z = \mathrm{z}$,
$H_0 : X \overset{Y}{\perp} Z$: tekijät X ja Z ovat riippumattomia ehdolla $Y = \mathrm{y}$,
$H_0 : Y \overset{X}{\perp} Z$: tekijät Y ja Z ovat riippumattomia ehdolla $X = \mathrm{x}$.
Loglineaariset mallit ehdollisen riippumattomuuden tilanteessa ovat:
$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$,
$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{kj}^{ZY}$,
$\ln(m_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ji}^{YX} + \lambda_{ki}^{ZX}$,
jossa $i = 1, \ldots, I, j = 1, \ldots, J$ ja $k = 1, \ldots, K$.

Loglineaarinen malli voidaan yleistää minkä tahansa tason moniulotteiselle kontingenssitaululle. Jos verrataan kaksi– ja kolmiulotteisen kontingenssitaulun loglineaarisia malleja, huomataan että kolmiulotteisen taulun mallit ovat monimutkaisia. Tällöin mukaan tulee ehdollisen ja osittaisen riippumattomuuden mallit ja kolmitekijäinen yhdysvaikutus. Kun dimensiot kasvavat, erilaisten assosiaatiomallien ja korkeampiasteisen vuorovaikutusten kuvaavien mallien lukumäärä kasvaa nopeasti. Lisäksi moniulotteisen taulun tarkastelussa solujen määrä kasvaa rajusti. Tällöin nollasolujen määrä kasvaa nopeasti, jos otoskoko ei ole suuri. Jos nollasoluja on paljon, niin tämä vaikuttaa asymptoottiseen χ^2–jakaumaan. Lisäksi mallit, joita saadaan, ovat vaikeita tulkita.

Rajoitteet SPSS–ohjelmistossa

SPSS–ohjelmisto asettaa yksittäisiä parametreja nolliksi. Otetaan esimerkiksi 2*2 kontingenssitaulu. Koska taulussa on vain neljä solua ja parametreja mallissa yhdeksän, asetetaan rajoitteita. Asetetaan esimerkiksi seuraavat parametrien arvot:
$\lambda_2^X = 0, \lambda_2^Y = 0,$
$\lambda_{12}^{XY} = \lambda_{21}^{XY} = \lambda_{22}^{XY} = 0.$
Tällöin estimoitavat parametrit saataisiin kaavoilla:

$\ln(m_{11}) = \mu + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}$
$\ln(m_{12}) = \mu + \lambda_1^X$
$\ln(m_{21}) = \mu + \lambda_1^Y$
$\ln(m_{22}) = \mu$
$\Leftrightarrow$
$\mu = ln(m_{22})$
$\lambda_1^X = \ln(m_{12}) - \ln(m_{22}) = \ln(m_{12}/m_{22})$
$\lambda_1^Y = \ln(m_{21}) - \ln(m_{22}) = \ln(m_{21}/m_{22})$
$\lambda_{11}^{XY} = \ln(m_{11}) - \ln(m_{12}) - (\ln(m_{21}) - \ln(m_{22})) = \ln(m_{11}m_{22}/m_{12}m_{21})$.

Tällöin estimoitavia parametreja on neljä, ja niiden arvot on mahdollista laskea.

Esimerkiksi SPSS–ohjelmiston vaihtoehtoinen määrittely parametreille olisi $\lambda_i^X = \ln(\pi_{i+}) - \ln(\pi_{2+})$ ja $\lambda_j^Y = \ln(\pi_{+j}) - \ln(\pi_{+2})$. Tällöin olisi voimassa $\lambda_2^X = \lambda_2^Y = 0$.
Tulkinta: Jos λ_i^X kasvaa niin i:nnellä X:n tasolla on suhteellisesti enemmän havaintoja.

Olkoon muuttuja X, jolla on viisi tasoa ja muuttujalla Y kuusi tasoa. Täysi loglineaarinen malli olisi:

$$\ln(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY},\ i = 1, \ldots, 5,\ j = 1, \ldots, 6.$$

Tällöin SPSS asettaa parametreja seuraavasti:

$$\lambda_5^X = \lambda_6^Y = 0,$$

$$\lambda_{26}^{XY} = \lambda_{36}^{XY} = \lambda_{46}^{XY} = \lambda_{56}^{XY} = \lambda_{66}^{XY} = 0,$$

$$\lambda_{52}^{XY} = \lambda_{53}^{XY} = \lambda_{54}^{XY} = \lambda_{55}^{XY} = 0.$$

Vastaavasti riippumattomuushypoteesille $\ln(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y$ SPSS asettaa parametrit seuraavasti:

$$\lambda_5^X = \lambda_6^Y = 0.$$

Rajoitteet SAS ja SUDAAN ohjelmistoissa

SAS ja SUDAAN ohjelmistot käyttävät parametrien estimoinnissa Agrestin (1990) esittämää tapaa. Yksittäisiä parametreja ei siis aseteta suoraan nolliksi vaan niiden summat asetetaan nolliksi. Olkoon esimerkiksi 2*2 kontingenssitaulu ja malli on:

$$\ln(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY},$$

jossa i, j=1,2. Tällöin mallissa on yhteensä yhdeksän estimoitavaa parametria, kun vain neljä pystytään estimoimaan. Tällöin asetetaan seuraavat rajoitukset:

$$\sum_{i=1}^{2} \lambda_i^X = 0, \quad \sum_{j=1}^{2} \lambda_j^Y = 0, \quad \sum_{i=1}^{2} \lambda_{ij}^{XY} = \sum_{j=1}^{2} \lambda_{ij}^{XY} = 0$$

4.6 Logit–mallit

Logit–mallit johdetaan loglineaarisista malleista. Logit–malleissa valitaan selitettävä muuttuja, jota selitetään luokitelluilla selittävillä muuttujilla. Mikäli selittävät muuttujat ovat jatkuvia ja vastemuuttuja on dikotominen, niin vastaavaa mallia kutsutaan logistiseksi regressiomalliksi. Loglineaarisissa malleissa tutkitaan riippumattomuutta ja riippuvuussuhteita, mutta logit–malleilla yritetään selittää ja ennustaa selitettävää muuttujaa. Logit–malleissa pyritään ennustamaan selitettävä muuttuja ehdolla, että selitettävien muuttujien arvot tiedetään. Ratkaisu palautuu selitettävän muuttujan ehdollisen jakauman tutkimiseen.

Usein selitettävänä muuttujana Y on dikotominen muuttuja ja olkoon, että kun Y=1, koe onnistuu ja Y=0, kun koe ei onnistu. Tällöin $P(Y = 1) = \pi, P(Y = 0) = 1 - \pi$ ja $\pi = E(Y)$. Kun Y_i noudattaa Bernoullijakaumaa parametrilla π_i , niin Y:n tiheysfunktio on:

$$f(y_i; \pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i} = (1 - \pi_i)exp(y_i ln(\frac{\pi_i}{1 - \pi_i})),$$

kun Y saa arvoja 0 tai 1. Tämä jakaumatyyppi kuuluu exponentiaaliseen jakaumaperheeseen. Luonnollinen parametri $Q(\pi) = ln[\pi/(1 - \pi)]$ on vedonlyöntisuhde sille, että Y=1 sen sijaan, että Y=0. Loglineaarisia malleja, jotka käyttävät tätä logit-linkkiä $Q(\pi)$, kutsutaan logit–malleiksi.

Yleistetty vedonlyöntisuhde

Logit–mallit voidaan yleistää tilanteeseen, jossa tulosmuuttujalla on enemmän kuin kaksi luokkaa. Olkoon luokiteltu tulosmuuttuja Y, jolla on J luokkaa. Tällöin on olemassa $\binom{J}{2}$ paria, joille voidaan laskea vedonlyöntisuhteet. Jos määrätään kiinteä J, niin pareja on enää J-1 kappaletta, loput ovat tarpeettomia. Olkoon selitettävä muuttuja Y kolmiluokkainen ja selittävä muuttuja X neljäluokkainen. Tällöin logit malli olisi:

$$\ln(\frac{m_{ij}}{m_{1j}}) = \alpha_i + \beta_{ij}, i = 2,3\ j = 1, \ \ldots\ , 4, \tag{17}$$

jossa esimerkiksi m_{11} on odotettu frekvenssi selitettävän muuttujan tasolla Y=1 ja selittävän muuttujan tasolla X=1, α_i on perustaso ja β_{ij}, i= 2,3 on muuttujan X vaikutus tasolla j. Tässä vertailuluokaksi on valitty $Y = 1$. Vastaava loglineaarinen malli olisi:

$$\ln(m_{ij}) = \lambda_j^Y + \lambda_i^X + \lambda_{ji}^{YX}, i = 1,2,3,\ j = 1, \ \ldots\ 4. \tag{18}$$

jossa λ_i^X on muuttujan X päävaikutus tasolla i, λ_j^Y on muuttujan Y päävaikutus tasolla j ja λ_{ji}^{YX} on yhdysvaikutusterni. Tällöin

$$\begin{aligned} \ln(\frac{m_{ij}}{m_{1j}}) &= \ln(m_{ij}) - \ln(m_{1j}) \\ &= (\lambda_i^Y + \lambda_j^X + \lambda_{ij}^{YX}) - (\lambda_1^Y + \lambda_j^X + \lambda_{1j}^{YX}) \\ &= (\lambda_i^Y - \lambda_1^Y) + (\lambda_{ij}^{YX} - \lambda_{1j}^{YX}) \end{aligned} \tag{19}$$

Vertaamalla kaavoja (17) ja (19) saadaan
$\alpha_i = \lambda_i^Y - \lambda_1^Y$ ja $\beta_{ij} = \lambda_{ij}^{YX} - \lambda_{1j}^{YX}$

Logit–mallit yleistyvät sekä selittettäville että selittäville muuttujille, joilla on useampiakin luokkia. Lisäksi selitettävän muuttujan vertailuluokkaa voi vaihtaa asetettujen hypoteesien mukaan.

5 Aineiston analyysi

Tutkielmassa on kolme tutkimusongelmaa. Ensin lasketaan arvioita Jyväskyläläisten kauppojen päivittäistavaramyynnille, sitten tarkastellaan asuinalueiden kauppojen palvelujen arvostusta ja viimeiseksi arvioidaan, mitkä syyt vaikuttavat kotitalouksien kaupan valintaan. Ensin esitellään analyyseissä käytettävät muuttujat:

5.1 Muuttujien kuvailu

Tutkielman eräs tavoite oli tarkastella, millä perusteilla kotitaloudet valitsevat kauppansa. Tässä tutkielmassa perehdytään tähän ongelmaan siten, että tarkastellaan valitseeko kotitalous lähikaupan vai suuren kaupan. Kotitalouksilta oli kysytty 17:ssä kaupassa tai kauppaketjussa käymisen määriä. Tutkielmassa kokeiltiin ensin kolmiluokkaista tulosmuuttujaa, jossa oli myös mukana 'lähin suuri kauppa'. Jokaiselle asuinalueelle pystytään määräämään taulukon 4 mukaisesti asuinaluetta lähin tai lähimmät suuret kaupat. Kuitenkin kävi niin, että muuttujan jako kahteen luokkaan oli järkevää, sillä silloin logit–mallien tulkinnat olivat helpompia. Kartta Jyväskylästä ja suurimmista kaupoista on liitteessä 4.

Taulukko 4. Asuinalueita lähinnä olevat kaupat

Asuinalue	Kaupat
Keskusta	Anttila, Minimani, Mestarin herkku, Kymppi
Kortepohja	Länsiväylä
Keltinmäki	Keljonkeskuksen Prisma ja Citymarket
Keljo	Keljonkeskuksen Prisma ja Citymarket
Kuokkala	K-Kotikenttä
Aittorinne	Tourulan Isoetu
Kangasvuori	Seppälän Prisma
Lohikoski	Seppälän Prisma
Palokka	Seppälän Prisma
Vaajakoski	Seppälän Prisma
Muurame	Keljonkeskus Prisma ja Citymarket
Säynätsalo	Keljonkeskus Prisma ja Citymarket
Laukaa	Seppälän Prisma

Dikotomisen tulosmuuttujan 'KAUPPA' jakauma on taulukossa 5:

Taulukko 5. Muuttujan 'KAUPPA' jakauma

Luokiteltu kauppa	Frekvenssi
Lähikauppa	152
Suuret kaupat	237
Ei vastannut	31
Yhteensä	420

Muuttuja kulkuväline oli alunperin viisiluokkainen. Muuttuja luokiteltiin uudestaan siten, että henkilöautolla kaupassa käyvät ovat omassa luokassaan ja muita kulkuvälineitä käyttävät (linja–auto, polkupyörä, kävellen ja muuten) ovat omassa luokassaan. Muuttujan jakauma on taulukossa 6.

Taulukko 6. Muuttujan 'KULKULUO' jakauma

Luokiteltu kulkuväline	Frekvenssi
Auto	296
Muu	124
Yhteensä	420

Kotitalouden käyttämä rahamäärä vuodessa ja kaupassakäyntikerralla. Kotitalouksilta oli kysytty, kuinka paljon he käyttävät rahaa viikossa. Lisäksi kotitaloudelta oli kysytty, kuinka usein he käyvät 17 eri kaupassa tai kauppaketjussa. Tästä voidaan arvioida kotitalouden vuoden aikana tekemien kaupassakäyntien kokonaismäärä. Arvioinnissa käytetyt kertoimet ovat taulukossa 12. Rahankulutus vuodessa saadaan kertomalla kotitalouden viikossa käyttämä rahamäärä 52:lla. Tämä saatu rahamäärä jaetaan kotitalouden vuotuisella kaupassakäyntimäärällä. Kotitalouden vuodessa ja kaupassakäyntikerralle käyttämien rahamäärien jakaumat ovat liitteenä 3. Liitteessä 3 on lisäksi kotitalouden vuodessa kaupassakäyntien määrän jakauma ja rahankulutus asukasta kohden vuodessa sekä jokaisella kaupassakäyntikerralla. Liitteenä 3 on rahankulutuksen ostokerralla muuttujasta logaritmimuunnoksella tehty uusi muuttuja (RAHALN). Rahankulutus vuodessa on jaettu kahteen luokkaan mediaaninsa kohdalta. Dikotomisen rahankulutuksen vuodessa jakauma on taulukossa 7.

Taulukko 7. Muuttujan 'RAHALUO' jakauma

Luokiteltu rahankulutus vuodessa	Frekvenssi
$<$ med	198
$\geq$ med	205
Ei vastannut	17
Yhteensä	420

Talouden henkilömäärä. Kotitaloudelta oli kysytty, kuinka monta henkilöä kuuluu talouteen. Kotitalouden koon jakauma on taulukossa 8.

Taulukko 8. Muuttujan 'KOKO' jakauma

Kotitalouden koko	Frekvenssi
Yksi henki	60
Kaksi henkilöä	128
Kolme henkilöä	76
Neljä henkilöä	97
Viisi tai enemmän	59
Yhteensä	420

Pääasiallinen toiminta. Kotitalouden päivittäistavaroiden ostosta päättävän henkilön toimea yli myös kysytty, muuttujan 'TOIMI' jakauma on taulukossa 9.

Taulukko 9. Muuttujan 'TOIMI' jakauma

Pääasiallinen toiminta	Frekvenssi
Töissä	254
Työtön	17
Opiskelija/kurssilla	56
Kotiäiti/–isä	27
Eläkeläinen	66
Yhteensä	420

Tämä muuttuja luokiteltiin myös uudelleen. Yhdistelemisen perusteena oli, että muuttujan 'KAUPPA' jakauma oli yhditellyissä luokkissa samankaltainen. Uudelleenluokitellun muuttujan jakauma on taulukossa 10.

Taulukko 10. Muuttujan TOIMILUO jakauma

Luokiteltu toimi	Frekvenssi
Töissä, kotiäiti ja opiskelija	337
Työtön	17
Eläkeläinen	66
Yhteensä	420

Kotitalouden päivittäistavaroiden ostosta päättävän henkilön iän jakauma on taulukossa 11.

Taulukko 11. Muuttujan 'IKÄ' jakauma

Ikä	Frekvenssi
15–24	52
25–34	116
35–44	107
45–59	94
60+	51
Yhteensä	420

5.2 Kauppojen saama rahamäärä

Arvioidaan kotitalouksien käyttämää rahamäärää päivittäistavaroihin. Tutkimuksen perusteella tiedetään kuinka monta kertaa vuodessa kotitaloudesta käydään tietyissä kaupoissa. Arvio on tehty taulukossa 12 oleviin kertoimiin perustuen:

Taulukko 12. Kaupassakäyntikertoimet

Vastausvaihtoehto	Käyntikertoja vuodessa
2-3 kertaa viikossa	130
kerran viikossa	52
kerran kahdessa viikossa	26
kerran kuukaudessa	12
kerran kahdessa kuukaudessa	6
harvemmin	3
ei asioi	0

Haastattelussa on kysytty käyntien lukumäärää sekä arkipäiville, että viikonlopuille erikseen. Tämä tapa voi aiheuttaa käyntimäärien yliarvioimista, koska voi käydä niin, että jos vastaaja arvioi käyvänsä jossain kaupassa 2–3 kertaa viikossa, niin hän on jo laskenut siihen viikonloppuna tapahtuvat käynnit. Tällöin viikonloppuna tapahtuvat käynnit voidaan laskea kahteen kertaan.

Lasketaan neljä eri estimaattia jokaisen kaupan tai kaupparyhmän vuotuiselle päivittäistavaramyynnille. Ensimmäinen tapa on laskea kokonaismäärät huomioimatta lainkaan otanta–asetelmaa. Jokaiselta otokseen kuuluvalta kotitaloudelta on kysytty viikossa käytetty rahamäärä päivittäistavaroihin. Tästä saadaan laskettua vuodessa käytetty rahamäärä. Vuodessa käytetyn rahamäärän keskiarvo on 32275 markkaa. Lisäksi tiedetään kunkin kotitalouden kaupassakäyntimäärät vuodessa jokaisessa kaupassa. Tällöin voidaan laskea jokaiseen kauppaan viety rahamäärä jokaisella kaupassakäyntikerralla olettaen, että jokaisella kaupassakäyntikerralla ostetaan samalla rahamäärällä ruokaa. Tästä saadaan laskettua kauppojen vuosimyynnit yksinkertaisen satunnaistannan tilanteessa. Kokonaismäärän estimaattori yksinkertaisen satunnaiso-

tannan tilanteessa on:

$$N\overline{y}_j = N\sum_{i=1}^{n} y_{ji}/n, \tag{20}$$

jossa N on perusjoukon koko ja n on otoskoko, $j = 1, 2, \ldots, 17$ on kaikki tutkimuksessa olevat kaupat.

Toinen tapa on ottaa huomioon Tilastokeskuksen arvio siitä, kuinka paljon kotitaloudet käyttävät keskimäärin rahaa päivittäistavaroihin vuodessa. Tämä arvio on 22734 markkaa. Tilastokeskuksen arvo on huomattavasti pienempi kuin otoksesta laskettu estimaatti. Tähän otokseen on tullut suuria kotitalouksia, jonka jo kotitalouden koon keskiarvon estimointi paljasti. Lisäksi pelkästään päivittäistavaroihin käytetyn rahamäärän arviointi on vaikeaa, sillä usein kaupasta ostetaan muitakin tavaroita samalla käynnillä.

Tilastokeskuksen antama arvio jaetaan kunkin kotitalouden vuosittaisella kaupassakäyntimäärällä. Tämän jälkeen saatu luku kerrotaan kotitalouden arvioimalla kuhunkin kauppaan vuosittain tehtyjen käyntien määrällä. Tällöin saadaan jokaiselle kotitaloudelle arvio siitä, kuinka paljon kotitalous ostaa kustakin kaupasta päivittäistavaroita vuosittain.

Kolmas tapa on ottaa huomioon otanta–asetelma. Kokonaismäärän estimaattori ositetun otannan tilanteessa lasketaan kaavalla

$$N\overline{y}_{j,STR} = \sum_{h=1}^{H} N_h \overline{y}_{jh}, \tag{21}$$

jossa $\overline{y}_{jh}$ lasketaan kaavan:

$$\overline{y}_{jh} = \sum_{i=1}^{n_h} \frac{y_{jhi}}{n_h} \tag{22}$$

avulla, $\overline{y}_{jh}$ on ositteesta h poimitun otoksen keskiarvo kaupalle j, N on perusjoukon koko ja N_h perusjoukon koko ositteessa h.

Neljäs tapa on ottaa huomioon Tilastokeskuksen laskema arvio jokaisen kotitalouden vuodessa käyttämälle rahamäärälle päivittäistavaroihin ja käyttää

ositetun otannan asetelmaa. Tilastokeskuksen mukaan jokainen kotitalous käyttää rahaa 22734 markkaa vuodessa päivittäistavaroihin. Lasketaan tästä arvio jokaiselle kotitaloudelle rahan käyttö jokaisella kaupassakäynnillä.

Taulukossa 13 on kaikkien kauppojen saama rahamäärä yhteensä laskettuna yksinkertaisen satunnaisotannan tilanteessa arvioidulle rahankäytölle (SRS+a), yksinkertaisen satunnaisotannan tilanteessa Tilastokeskuksen laskemalle rahankäytölle (SRS+k), ositetun otannan tilanteessa arvioidulle rahankäytölle (STR+a), ositetun otannan tilanteessa Tilastokeskuksen laskemalle rahankäytölle (STR+k) ja joidenkin kauppojen ilmoittamat oikeat päivittäistavaramyynnit, (oikea).

Taulukko 13. Kauppojen saama rahamäärä vuodessa, miljoonaa markkaa.

Kauppa	SRS+a	SRS+k	STR+a	STR+k	Oikea
Anttila	29.29	49.91	55.25	83.98	62.0
Citymarket	60.25	121.93	51.73	101.24	145.0
Mestarin herkku	36.78	57.20	55.00	86.09	85.0
Minimani	12.75	19.20	26.37	35.72	
Seppälän Prisma	118.03	233.67	116.99	228.87	205.0
Keljonkeskuksen Prisma	73.57	155.27	58.68	123.00	112.0
K-Supermarket Tourula	26.54	45.28	23.97	40.00	
K-Supermarket Länsiväylä	18.18	38.68	24.62	54.76	
K-lähikauppa Kymppi	5.15	6.90	12.59	16.26	
K-Market Kotikenttä	4.78	10.94	6.60	14.95	
Muut K-kaupat	56.89	115.58	52.79	103.84	
S-market/Sale	92.43	172.79	83.85	153.78	
Spar-market	30.03	61.66	29.80	57.95	
Tarmo-lähikaupat	1.79	3.16	2.98	5.93	
Siwa	36.91	65.66	39.79	67.82	
Valintatalo	6.36	13.22	4.24	8.49	
Muut	12.13	27.48	13.31	29.09	

Otanta–asetelmien välillä on huomattavia eroja. Lisäksi estimaatit poikkeavat aivan liikaa oikeista arvoista. Tilastokeskuksen antaman rahankäytön

mukaiset estimaatit ovat suurempia kuin kotitalouden oman arvion mukaiset estimaatit. Kuitenkin otoksesta laskettu estimaatti kotitalouksien keskimääräiselle rahankäytölle oli paljon suurempi kuin Tilastokeskuksen arvio. Haastattelussa on pyydetty kertomaan kotitalouden viikossa käyttämä rahamäärä päivittäistavaroihin. On vaikea muistaa ja tietää, mikä on keskimääräinen rahankulutus. Toisaalta on yhtä vaikea arvioida kussakin kaupassa käymisen määrää.

Ei voida sanoa tässä tapauksessa, että ositetun otannan käyttö arvioidulle rahankäytölle (STR+a) on kovinkaan paljon parempi kuin yksinkertainen satunnaisotanta arvioidun rahankäytön tapauksessa (SRS+a). Mikään arviointikeinoista ei ollut riittävän hyvä, mutta ositetun otannan käyttö kiinteälle rahankäytölle alkaa olla jo kohtuullisen tarkka estimointimenetelmä. Verrataan viittä oikeaa arvoa käytettyihin arviointimenetelmiin:

Menetelmä	Kertaa lähimpänä oikeaa arvoa
SRS+a	0
SRS+k	1
STR+a	1
STR+k	3

Yksinkertainen satunnaisotanta kiinteälle rahankäytölle oli parempi menetelmä Citymarketin tapauksessa. Ositettu otanta arvioidulle rahankäytölle oli parempi menetelmä Anttilan tapauksessa ja ositettu otanta kiinteällä rahankäytöllä oli parempi Mestarin Herkun, Seppälän ja Keljonkeskuksen Prisman tapauksissa.

On selvää, että ositetun otannan teorian käyttö tässä tapauksessa on kannattavaa verrattuna yksinkertaiseen satunnaisotantaan. Alueiden sisällä hakeudutaan samoihin kauppoihin.

Tarkastellaan ensin, miten kotitalouden koko vaikuttaa rahankulutukseen ostokerralla. Rahankulutus ostokerralla saadaan normaalijakautuneeksi muuttujaksi tekemällä sille logaritmimuunnos. Tällöin tähän tarkasteluun voidaan soveltaa regressio– ja varianssianalyysia. Tehdään yksisuuntainen varianssi-

analyysi logaritmimuunnetulle rahankäytölle ostokerralla. Valitaan käsittelyksi kotitalouden koko (KOKO). Analyysista saadaan erittäin merkitsevät tulokset, p=0.000. Testin oletukset ovat voimassa. Monivertailujen mukaan kotitaloudet voidaan jakaa kolmeen osaan: yhden, kahden ja kolmen tai useamman henkilön kotitalouksiin.

Tehdään lineaarinen regressiomalli logaritmimuunnetulle rahankäytölle ostoskerralla. Otetaan selittäväksi muuttujiksi kotitalouden koko (KOKO) ja kotitalouden vuodessa tekemien kaupassakäyntien määrä (YHT). Mallista saadaan erittäin merkitsevä p=0.000 ja kaikki mallissa olevat termit ovat merkittäviä p=0.000. Regressioanalyysin oletukset ovat voimassa. Regressioyhtälöksi saadaan:
Malli: LNRAHA=Vakio+KOKO+YHT
5.141+0.212*KOKO-0.004*YHT. R^2 = 57,9%
Regressiomalli osoittaa, että kotitalouden koon kasvaessa kotitalouden rahankulutus ostokerralla suurenee. Lisäksi mitä useammin kotitalous käy kaupassa, sitä vähemmän päivittäistavaroita ostetaan kerralla.

5.3 Oman asuinalueen kauppojen palvelut

Kotitalouksilta on kysytty arvosanaa useisiin oman asuinalueen kauppapalveluja koskeviin kysymyksiin. Vastaukset on pyydetty antamaan kouluarvosteikolla 4–10. Antavatko eri asuinalueiden asukkaat eri arvosanoja näille ominaisuuksille? Kysymyksiä on kymmenen ja ne käsittelevät seuraavia ominaisuuksia:

- lihat ja lihajalosteet
- leivät ja leipomotuotteet
- maitotuotteet
- hedelmät ja vihannekset
- juomat
- tuotteiden tuoreus ja laatu
- tuotteiden hinnat
- tuotteiden saatavuus
- tuotevalikoimat

Otetaan analyysiin mukaan ainoastaan ne kotitaloudet, jotka ovat vastanneet jokaiseen kymmeneen kohtaan, koska havaintoja ei hylätty liikaa. Tulosmuuttuja ei ole normaalinen, joten käytetään varianssianalyysin sijaan Kruskall–Wallisin yksisuuntaista varianssianalyysiä, joka kertoo, että jokaisen ominaisuuden kohdalla on eroja asuinalueiden välillä. Tehdään summamuuttuja, joka kuvaa yleistä mielipidettä oman asuinalueen kauppojen palveluista ja laadusta. Summamuuttujalle voitaisiin tehdä varianssianalyysi, mutta oletus varianssien yhtäsuuruudesta ei ole voimassa. Käytetään analyysissa edelleen Kruskall–Wallisin yksisuuntaista varianssianalyysia. Taulukossa 14 on testissä tarvittavat arvot.

Taulukko 14. Oman asuinalueen palvelujen arviointi.

ASUINALUE	n	Mean Rank
Keskusta	34	264.18
Laukaa	37	228.03
Kortepohja	27	223.76
Keljo	29	215.48
Aittorinne	24	210.83
Kuokkala	28	206.68
Vaajakoski	38	203.80
Kangasvuori	30	171.23
Palokka	26	152.56
Keltinmäki	24	141.63
Muurame	37	140,65
Säynätsalo	16	126.16
Lohikoski	28	126.11
Total	378	

Asuinalueiden välillä on eroja. Testisuure antaa p–arvon 0.000. Tarkastellaan LSD–menetelmällä, mitkä asuinalueet eroavat toisistaan. Keskustan asukkaat ovat kaikkein tyytyväisimpiä oman asuinalueensa palveluihin. Lohikosken ja Säynätsalon asukkaat ovat kaikkein tyytymättömimpiä. LSD–menetelmä paljastaa eroja Keskustan ja seuraavien asuinalueiden välillä:

- Ryhmä 1: Lohikoski, Säynätsalo, Keltinmäki, Kangasvuori, Palokka, Muurame.

Lohikosken ja Säynätsalon osalta LSD–menetelmä paljastaa eroja seuraavien asuinalueiden väliltä:

- Ryhmä 2: Keskusta, Kortepohja, Keljo, Kuokkala, Aittorinne, Vaajakoski, Laukaa.

Keskusta, Laukaa ja Kortepohja ovat kolme parhaiten kiitosta saaneet asuinalueet ja Säynätsalo, Lohikoski ja Keltinmäki ovat asukkaiden mielestä huonoimmin kysyttyjä palveluita tarjoavat asuinalueet. Näyttää siltä, että etäisyys Jyväskylän keskustasta vaikuttaa mielipiteisiin. Laukaa on piristävä poikkeus. Laukaassa on ilmeisesti palvelut hyvää tasoa. LSD–testin merkitevyydet ovat liitteenä 1. Taulukossa 14 on viivalla eroteltu asuialueet siten, että yläpuolella ovat ryhmän 1 asuinalueet ja alapuolella ryhmän 2 asuinalueet.

Tarkastellaan logistisella regressiolla, miten oman asuinalueen arvostus ja kaupassakäyntimäärä vaikuttavat käytettyyn kulkuvälineeseen:
Malli: KULKULUO=Vakio+SUM+YHT
Malli on huono, selitysaste on 4%. Malli on kuitenkin merkitsevä p=0.0003.
SPSS antaa yhtälöksi -3.651+0.029*SUM+0.002*YHT.
SUDAAN antaa yhtälöksi -3.365+0.031*SUM+0.002*YHT.
Mitä enemmän kotitalous arvostaa oman asuinalueensa palveluita sitä todennäköisemmin kulkuvälineeksi valitaan joku muu kuin auto. Sama vaikutus on myös vuoden aikana tehdyillä kauppoissakäynneillä.
Deft–arvot ovat :

Muuttuja	deft
Vakio	1.36
SUM	1.34
YHT	1.19

5.4 Kotitalouksien kaupan valinta

Tarkastellaan kotitalouden kaupan valintaan vaikuttavia asioita. Analyyseissä käytetään loglineaarisia ja logit–malleja ja logistista regressiota.

Loglineaaristen mallien rajoitteiden käyttö SPSS–ohjelmistossa

Rajoitteet saadaan malliin tekemällä kovariaatti. Mallissa oleva muuttuja koodataan uudelleen rajoitteiden määräämällä tavalla. Esimerkiksi muuttujan X lineaarista vaikutusta muuttujaan Y voitaisiin tutkimalla tekemällä kovariaatti cov X seuraavasti:
X-> cov X
1-> 1
2-> 2
3-> 3
Tällöin sovitettaisiin mallia: vakio+X+Y+covX * Y.
Kovariaatteja voidaan käyttää myös kontingenssitaulujen osittamiseen. Tällöin tutkitaan näiden osataulujen riippumattomuutta. Esimerkiksi olkoon

alkuperäinen kontingenssitaulu neljärivinen. Ositetaan taulu puoliksi asettamalla kovariaatti covX seuraavasti:
X-> cov X
1-> 0
2-> 0
3-> 1
4-> 1
Malliksi asetetaan jälleen vakio+X+Y+covX * Y.
Tällöin malliin liittyvä testisuure on osataulujen (riippumattomuus) testisuureiden summa. Kovariaatteja voidaan käyttää myös kontingenssitaulujen luokkien yhdistelemiseen. Tällöin esimerkiksi edellä tehty kovariaatti asetetaan mallissa faktoriksi ja asetetaan malli:
vakio + covX*Y.
Tällöin saadaan riippumattomuustestisuureen arvo taululle, jossa on yhdistetty rivit 1,2 ja 3,4. Kovariaatteja voidaan tehdä yhdelle tai useammalle kontingenssitaulun muuttujalle yhtäaikaa.

Paras malli muuttujille KOKO, KULKULUO ja RAHALUO, joka sopii on seuraava:
Vakio+KOKO+KULKULUO+RAHALUO
+KOKO*KULKULUO+KOKO*RAHALUO
Uskottavuussuhteen testi antaa p–arvoksi 0.1971. Tämä malli on ehdollisen riippumattomuuden malli. Tämän mallin tulkinta on: Kulkuväline ja rahankäyttö ovat riippumattomia toisistaan kiinteällä kotitalouden koon arvolla. Mikään muu osittaisen tai ehdollisen riippumattomuuden malli ei edes sopinut aineistoon. Kotitaloudeen koon lineaarista vaikutusta yritettiin selvittää tekemällä kovariaatti seuraavasti.
KOKOLUO -> covKOKOLUO
1 -> -1
2 -> 0
3 -> 1

Malliksi sovitetaan:
Vakio+KOKO+KULKULUO+RAHALUO
+covKOKO*KULKULUO+covKOKO*RAHAALUO.
Malli ei sovi ollenkaan, uskottavuussuhteen testi antaa p–arvoksi p=0,002.

Vaihdetaan kotitalouden koko –muuttuja asuinalueeseen. Tällöin ainoa malli, joka sopii aineistoon muutujille ASUU, KULKULUO ja RAHALUO on:
Vakio+ASUU+KULKULUO+RAHALUO+KULKULUO*RAHALUO
Eli Asuinalue ja rahankäyttö ovat riippumattomia kiinteällä kulkuluokan arvolla, p=0.4385.

Kotitalouden kaupan valinnan mallintaminen logit–malleilla

Mallintamisen idea on selvittää, millä perusteilla kotitaloudet valitsevat kaupan, jossa käyvät ostoksilla. Tulosmuuttujaksi valitaan muuttuja 'KAUPPA'. Kotitalouden kuuluminen muuttujan tiettyyn luokkaan riippuu siitä, kuinka usein kotitalous on ilmoittanut käyvänsä eri kaupoissa. Luokka valitaan sillä perusteella, jossa kotitalous ilmoittaa käyvänsä eniten. Lähin suuri kauppa alueittain on taulukossa 4.

Havaintojen pienen lukumäärän johdosta malliin ei saada kaikkia mielenkiintoisia selittäviä muuttujia yhtäaikaa.

Ehdottomasti sopivin malli saadaan, kun selittäviksi muuttujiksi valitaan asuinalue ja kulkuväline. Tulosmuuttuja oli alunperin kolmiluokkainen, jossa luokka 'suuret kaupat' oli jaettu vielä kahteen luokkaan 'lähin suuri kauppa' ja 'muut suuret kaupat'. Tulkinta helpottuu tällä uudella luokittelulla.

Seuraavassa esitellään useita logit–malleja. Jokaisessa mallissa tulosmuuttujana on KAUPPA, joka on kotitalouden luokitus sille, käykö kotitalous enemmän suuressa vai pienessä kaupassa. Malleista esitetään sekä ositetun otannan mukaiset (STR) että yksinkertaisen satunnaisotannan mukaiset logit–mallien beta–kertoimet. Lisäksi on laskettu ositetun otannan mukaiset vedonlyöntisuhteet kaikille mahdollisille muuttujien tasoille. Vedonlyöntisuhde

ilmoittaa, kuinka monta kertaa todennäköisempää on käydä suuressa kaupassa kuin pienessä kaupassa valituilla selittävien muuttujien tasoilla. Jos vedonlyöntisuhde on alle yksi, niin tällöin vedonlyöntisuhteen käänteisluku kertoo kuinka monta kertaa todennäköisempää on käydä pienessä kuin suuressa kaupassa valituilla selitettävien muuttujien tasoilla. Kaikkien mallien vedonlyöntisuhteet ovat liitteessä 2. Mallit tulkitaan ositetun otannan tilanteen mukaan.

Malli: KAUPPA = Vakio+ASUU+KULKULUO.

Kaupan valinnan erot tulevat parhaiten esiin juuri tässä mallissa. Asuinalueiden väliset erot ovat huomattavat ja kulkuvälineen vaikutus on selkeä.
Malliksi saadaan:
$STR : \ln(\frac{m_{1jk}}{m_{0jk}}) = -5.075 + ASUU + 2.415 * KULKULUO$
$SRS : \ln(\frac{m_{1jk}}{m_{0jk}}) = -4.930 + ASUU + 2.268 * KULKULUO$
jossa KULKULUO=1 jos kulkuväline on auto ja KULKULUO=0 jos kulkuväline on jokin muu, tulosmuuttujan arvo 1=suuri kauppa ja 0=lähikauppa ja ASUU arvot ovat seuraavat:

ASUU	beta STR	beta SRS
Keskusta	6.540	6.402
Kortepohja	5.219	5.093
Keltinmäki	4.346	4.297
Keljo	6.450	6.402
Kuokkala	4.096	4.041
Aittorinne	5.516	5.418
Kangasvuori	4.945	4.871
Lohikoski	4.497	4.451
Palokka	3.000	2.998
Vaajakoski	3.344	3.329
Muurame	1.766	1.760
Säynätsalo	2.209	2.205
Laukaa	0.000	0.000

Mitä suurempi asuinalueen parametrin estimaatti on sitä todennäköisemmin asuinalueelta käydään suuressa kaupassa verrattuna Laukaaseen. Autolla kauppaan menevät käyvät $e^{2.415}$=11.19 kertaa todennäköisemmin suuressa kaupassa kuin pienessä kaupassa. Vedonlyötisuhteet ositetun otannan tilanteessa ovat liitteessä 2. Keskustassa, Keljossa ja Aittorinteellä suositaan reilusti suuria kauppoja. Kulkuvälineen vaikutus on todella merkitsevä. Kaukana Keskustasta (Laukaa, Säynätsalo ja Muurame) asuvat suosivat omia lähikauppojaan varsinkin silloin, kun kulkuväline on muu kuin auto.

Malli: KAUPPA = Vakio+ASUU+RAHALUO

Tarkastellaan, kuinka asuinalue ja ostoksiin käytetty rahamäärä vaikuttavat kaupan valintaan.

Malliksi saadaan:
STR: $\ln(\frac{m_{1jk}}{m_{0jk}}) = -2.697 + ASUU - 0.311 * RAHALUO$
SRS: $\ln(\frac{m_{1jk}}{m_{0jk}}) = -2.701 + ASUU - 0.301 * RAHALUO$
jossa RAHALUO=1 jos rahaa käytetään vähän ja RAHALUO=0 jos rahaa käytetään paljon, tulosmuuttujan arvo 1=suuri kauppa ja 0=lähikauppa ja ASUU arvot ovat seuraavat:

ASUU	beta STR	beta SRS
Keskusta	4.593	4.591
Kortepohja	3.061	3.598
Keltinmäki	3.959	3.957
Keljo	6.114	6.114
Kuokkala	3.334	3.334
Aittorinne	4.364	4.363
Kangasvuori	3.980	3.980
Lohikoski	4.075	4.073
Palokka	2.974	2.974
Vaajakoski	3.130	3.129
Muurame	1.579	1.580
Säynätsalo	2.145	2.145
Laukaa	0.000	0.000

Mitä suurempi asuinalueen parametrin estimaati on, sitä todennäköisemmin asuinalueelta käydään suuressa kaupassa kuin pienessä kaupassa. Enemmän rahaa käyttävät hakeutuvat $e^{0.311} = 1.365$ kertaa todennäköisemmin suureen kauppaan kuin pieneen kauppaan. Vedonlyöntisuhteet ositetun otannan tilanteessa ovat liitteessä 2. Enemmän rahaa käyttävät asioivat todennäköisemmin suurissa kaupoissa. Muurame on poikkeus. Keljon asukkaat käyvät paljon suurissa kaupoissa.

Malli: KAUPPA = ASUU+KOKO

Tarkastellaan asuinalueen ja kotitalouden koon vaikutusta kaupan valintaan.

Malliksi saadaan:
STR: $\ln(\frac{m_{1jk}}{m_{0jk}}) = -2.683 + ASUU + KOKO$
SRS: $\ln(\frac{m_{1jk}}{m_{0jk}}) = -2.702 + ASUU + KOKO$
jossa tulosmuuttujan arvo 1=suuri kauppa ja 0=lähikauppa ja ASUU ja KOKO arvot ovat seuraavat:

ASUU	beta STR	beta SRS
Keskusta	4.952	4.889
Kortepohja	3.873	3.846
Keltinmäki	3.373	3.711
Keljo	6.172	6.175
Kuokkala	3.314	3.308
Aittorinne	4.622	4.600
Kangasvuori	4.304	4.269
Lohikoski	4.013	4.022
Palokka	3.042	3.013
Vaajakoski	3.181	3.152
Muurame	1.582	1.586
Säynätsalo	2.093	2.055
Laukaa	0.000	0.000

KOKO	beta STR	beta SRS
Yksi	-1.281	-1.163
Kaksi	0.279	0.332
Kolme	-0.325	-0.422
Neljä	-0.277	-0.181
Viisi ja enemmän	0.000	0.000

Mitä suurempi parametrin estimaatti on, sitä todennäköisemmin asuinalueelta käydään suuressa kaupassa. Sama pätee myös kotitalouden kokoon. Vedonlyöntisuhteet ositetun otannan tilanteessa ovat liitteessä 2. Kahden henkilön kotitaloudesta käydään todennäköisimmin suuressa kaupassa. Keljosta käydään todennäköisimmin suuressa kaupassa. Kaukana keskustasta (Muurame, Säynätsalo ja Laukaa) asuvat suosivat lähikauppoja.

Malli: KAUPPA = Vakio+KOKO+KULKULUO

Tarkastellaan, kuinka kotitalouden koko ja kulkuväline vaikuttavat kaupan valintaan.

Malliksi saadaan:
STR: $\ln(\frac{m_{1jk}}{m_{0jk}}) = -0.006 + KOKO + 0.601 * KULKULUO$
SRS: $\ln(\frac{m_{1jk}}{m_{0jk}}) = -0.295 + KOKO + 0.934 * KULKULUO$
jossa KULKULUO=1, kun kulkuväline on auto ja KULKULUO=0, kun kulkuväline on jokin muu, tulosmuuttujan arvo 1=suuri kauppa ja 0=lähikauppa ja KOKO arvot ovat seuraavat:

KOKO	beta STR	beta SRS
Yksi	0.165	0.110
Kaksi	0.836	0.654
Kolme	-0.294	-0.344
Neljä	-0.057	-0.163
Viisi ja enemmän	0.000	0.000

Mitä suurempi parametrin estimaatti on, sitä todenäköisemmin kotitalous käy suuressa kaupassa. Autolla kaupassa kulkevat käyvät $e^{0.601} = 1.824$ kertaa todennäköisemmin suuressa kuin pienessä kaupassa. Vedonlyöntisuhteet ositetun otannan tilanteessa ovat liitteessä 2. Autoilijat hakeutuvat todennäköisemmin suuriin kauppoihin kuin muita kulkuvälineitä käytävät. Kahden henkilön kotitaloudet hakeutuvat todennäköisemmin suuriin kauppoihin kuin muun kokoiset kotitaloudet riippumatta kulkuvälineestä.

Malli: KAUPPA = Vakio+KOKO+RAHALUO

Tarkastellaan, kuinka kotitalouden koko ja kotitalouden käytämä raha vuodessa vaikuttavat kaupan valintaan.

Malliksi saadaan:
STR: $\ln(\frac{m_{1jk}}{m_{2jk}}) = 0.471 + KOKO - 0.178 * RAHALUO$
SRS: $\ln(\frac{m_{1jk}}{m_{2jk}}) = 0.499 + KOKO - 0.208 * RAHALUO$
jossa RAHALUO=1, kun rahaa käytetään vähän RAHALUO=0, kun rahaa käytetään paljon, tulosmuuttujan arvo 1=suuri kauppa ja 0=lähikauppa ja KOKO arvot ovat seuraavat:

KOKO	beta STR	beta SRS
Yksi	-0.098	-0.354
Kaksi	0.738	0.563
Kolme	-0.218	-0.328
Neljä	0.011	0.069
Viisi ja enemmän	0.000	0.000

Mitä suurempi parametrin estimaatti on, sitä todenäköisemmin kotitalous käy suuressa kaupassa. Enemmän rahaa käyttävät käyvät $e^{0.178} = 1.195$ kertaa todennäköisemmin suuressa kuin pienessä kaupassa. Kahden henkilön kotitaloudesta käydään todennäköisimmin suuressa kaupassa. Kolmen henkilön kotitalouksista käydään vähiten suuressa kaupassa. Vedonlyöntisuhteet ositetun otannan tilanteessa ovat liitteessä 2.

Malli: KAUPPA = Vakio+IKÄ+KULKULUO

Tarkastellaan, kuinka ikä ja kulkuväline vaikuttavat kaupan valintaan.

Malliksi saadaan:
STR: $\ln(\frac{m_{1jk}}{m_{0jk}})$=0.312+IKÄ+0.500*KULKULUO
SRS: $\ln(\frac{m_{1jk}}{m_{0jk}})$=0.032+IKÄ+0.866*KULKULUO
jossa KULKULUO=1, kun kulkuväline on auto ja KULKULUO=0,kun kulkuväline on jokin muu, tulosmuuttujan arvo 1=suuri kauppa ja 0=lähikauppa ja IKÄ arvot ovat seuraavat:

Ikä	beta STR	beta SRS
15–24 vuotta	0.235	0.180
25–34 vuotta	0.524	0.268
35–44 vuotta	-0.732	-0.816
45–59 vuotta	-0.131	-0.214
Yli 60 vuotta	0.000	0.000

Mitä suurempi parametrin estimaatti on, sitä todenäköisemmin kotitalous käy suuressa kaupassa. Autolla kaupassa kulkevat käyvät $e^{0.500} = 1.649$ kertaa todennäköisemmin suuressa kuin pienessä kaupassa. Vedonlyöntisuhteet ositetun otannan tilanteessa ovat liitteessä 2. 25–34 vuotiaiden ryhmässä käydään eniten suurissa kaupoissa. Toisaalta 35–44 vuotiaiden ryhmässä käydään vähiten suurissa kaupoissa. Autoilijat hakeutuvat todenäköisemmin suuriin kauppoihin.

Malli: KAUPPA = Vakio+IKÄ+RAHALUO

Tarkastellaan, kuinka ikä ja rahankäyttö vuodessa vaikuttavat kaupan valintaan.

Malliksi saadaan:
STR: $\ln(\frac{m_{1jk}}{m_{0jk}})$=0.605+IKÄ-0.279*RAHALUO
SRS: $\ln(\frac{m_{1jk}}{m_{0jk}})$=0.548+IKÄ-0.362*RAHALUO
jossa RAHALUO=1, kun rahaa käytetään vähän ja RAHALUO=0, kun rahaa käytetään paljon, tulosmuuttujan arvo 1=suuri kauppa ja 0=lähikauppa ja IKÄ arvot ovat seuraavat:

Ikä	beta STR	beta SRS
15–24 vuotta	0.494	0.643
25–34 vuotta	0.666	0.546
35–44 vuotta	-0.612	-0.598
45–59 vuotta	0.003	0.070
Yli 60 vuotta	0.000	0.000

Enemmän rahaa käyttävät käyvät $e^{0.279} = 1.322$ kertaa todennäköisemmin suuressa kuin pienessä kaupassa. Vedonlyöntisuhteet ositetun otannan tilanteessa ovat liitteessä 2. 25–34 vuotiaat käyvät eniten suuressa kaupassa ja 35–44 vuotiaat käyvät vähiten suuressa kaupassa.

Malli: KAUPPA = Vakio+KULKULUO+RAHALUO

Tarkastellaan, kuinka kulkuväline ja rahankäyttö vuodessa vaikuttavat kaupan valintaan.

Malliksi saadaan:
STR: $\ln(\frac{m_{1jk}}{m_{0jk}}) = 0.256 + 0.479 * KULKULUO + 0.025 * RAHALUO$
SRS: $\ln(\frac{m_{1jk}}{m_{0jk}}) = -0.191 + 0.882 * KULKULUO + 0.049 * RAHALUO$
jossa KULKULUO=1, kun kulkuväline on auto ja KULKULUO=2, kun kulkuväline on jokin muu, RAHAALUO=1, kun rahaa käytetään vähän ja RAHAALUO=0, kun rahaa käytetään paljonja tulosmuuttujan arvo 1=suuri kauppa ja 0=lähikauppa. Vedonlyöntisuhteet ovat seuraavat:

Kulkuväline	Rahaluo $\leq$ med	Rahaluo $>$ med
Auto	$\frac{m_{111}}{m_{211}} = 8.025$	$\frac{m_{112}}{m_{212}} = 8.049$
Muu	$\frac{m_{121}}{m_{221}} = 1.325$	$\frac{m_{122}}{m_{222}} = 1.292$

Autoilijat käyttävät todennäköisemmin suurta kauppaa. Autoilijat siis käyttävät $e^{0.479} = 1.615$ kertaa todennäköisemmin suurta kuin pientä kauppaa. Rahankulutuksella ei ole tässä mallissa kovin huomattavaa merkitystä.

6 Johtopäätökset

Kauppojen päivittäistavaramyynnin arvioiminen on hankalaa. Ihmisten on vaikea muistaa, kuinka usein he käyvät tietyissä kaupoissa ja yhtä hankala on tietää, kuinka paljon yhdellä kauppamatkalla rahaa käytetään. Ulkopuolisen informaation käyttö oli mukana analyysissa. Tässä tapauksessa lisäinformaatiota käytettiin jälkiosituksen tekemiseen ja kauppojen saaman rahamäärän arvioimiseen.

Osittamisen tekeminen ja estimointi ositetun otannan tilanteessa lisäsi tulosten tarkkuutta kauppojen päivittäistavaramyynnin arvioimisen kohdalla. Lisäksi Tilastokeskukselta saatu tieto siitä, kuinka paljon kotitaloudet keskimäärin käyttävät rahaa päivittäistavaroihin paranti tulosten tarkkuutta entisestään. Tulosten mukaan Seppälän Prisma on Jyväskylän paras kauppa päivittäistavamyynnin osalta. Sinne matkustetaan kauempaakin. Tämä tutkimus on tehty ennen Palokan Euromarketin rakentamista. Todennäköistä on, että Seppälän Prisma menettää hiukan myyntiään, kun osa Palokan asukkaista siirtyvät käyttämään Palokan Euromarkettia.

Loglineaaristen mallien käyttö on tarkoituksenmukaista silloin, kun halutaan tietää usean luokitellun muuttujan välisistä riippuvuussuhteista. Loglineaariset mallit yleistyvät muuttujille, joilla on mielivaltainen määrä luokkia. Tässä tutkimuksessa oli vain 420 havaintoa, joten malleissa piti jäädä vain korkeintaan kolmen muuttujan riippuvuuksien tarkasteluun.

Logit–malleilla selvitettiin kotitalouksien kaupan valintaa. Logit–malleja käytetään, kun halutaan selvittää luokiteltua tulosmuuttujaa luokitelluilla selittävillä muuttujilla. Tulosmuuttujaksi valittiin kolmiluokkainen muuttuja, jota yritettiin selittää useilla eri muiden muuttujien malleilla. Tulosmuuttujan luokat yhdistettiin kuitenkin siten, että tulosmuuttuja oli dikotominen. Tämä helpotti tulkintaa. Lopputuloksena voidaan sanoa, että tässä tutkimuksessa parhaat selittävät muuttujat ovat asuinalue ja kulkuväline, kun yritetään selittää kaupan valintaa. Kotitalouden koko –muuttuja oli mielenkiintoinen. Tämän mukaan kahden henkilön kotitaloudet asioivat suurissa kaupoissa use-

ammin kuin yhden tai kolmen tai useamman henkilön kotitaloudet. Lisäksi kolmen henkilön kotitalouksista käytiin vähiten suurissa kaupoissa. 25–34 vuotiaiden ryhmässä suuria kauppoja suosittiin eniten ja 35–44 vuotiaat käyttivät suuria kauppoja vähiten.

Tuloksista ilmeni, että asuinalueiden välillä on huomattavia eroja kaupan valinnan osalta. On selvää, että jokaisella asuinalueella ei olekaan automarkettia, tämä näkyy selvästi tuloksissa. Lisäksi jos omalla asuinalueella, kuten Keljossa on suuria kauppoja, kotitaloudet hakeutuvat herkästi suuriin kauppoihin. Tuloksista kävi myös ilmi, että autolla kauppaan matkustavat hakeutuvat todennäköisemmin suuriin kauppoihin ja keskitasoa enemmän rahaa vuodessa kauppaan käyttävät asioivat myös useimmin suurissa kaupoissa.

Tutkimuksen yksi tavoite oli osoittaa, kuinka oikea otantateorian käyttö vaikuttaa tulosten tarkkuuteen. Yleensä käytäntönä ilmeisesti on, että otanta tehdään hyvilläkin perusteilla ja ihan oikein, mutta analyysivaiheesssa otannasta saatava lisäinformaatio unohdetaan. Tässä tutkimuksessa ei käy selvästi ilmi, että otanta–asetelman huomioiminen parantaa tulosten tarkkuutta. Kuitenkin osittaminen Tietoykkönen Oy:n toimesta on tehty hyvin, mikä on tehty Jyväskylän kaupungin rekisteritietoja vastaavaksi. Tällöin ulkopuolisen informaation käyttö oli vaivatonta ja ositetun otannan vaatimat perusjoukon populaatioiden määrät olivat saatavilla. Tässä tutkimuksessa estimaattien arvot yksinkertaisen satunnasotannan ja ositetun otannan tilanteissa eivät poikenneet toisistaan kovinkaan suuresti. Tehokkuuskertoimen $deft$ arvot olivat yleisesti yli ykkösen, jolloin ositetut otannan estimaatin luottamusväli oli leveämpi verrattuna yksinkertaisen satunnaisotannan vastaavaan. Lisäksi vertailtaessa yksinkertaisen satunnaisotannan analyysien malleja ositetun otannan malleihin huomataan, että mallit ovat lähes samat. Otanta–asetelman huomioiminen ei siis muuta saatuja tuloksia. Kiintiöimisen hyvä puoli on se, että jokainen asuinalue saa varmasti edustuksen otokseen.

Lähteet

Agresti, A. 1984: *Analysis of ordinal categorical data.* John Wiley & Sons, Inc., New York.

Agresti, A. 1990: *Categorical Data Analysis.* Wiley Interscience Publication, John Wiley & Sons, New York.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. 1975: *Discrete Multivariate Analysis: Theory and Practice.* The Massachusetts Institute of Technology, London.

Dobson, A. J. 1983: *An introduction to statistical modelling.* Chapman and Hall, New York.

Lehtonen, R. & Pahkinen, E. 1994: *Practical methods for design and analysis of complex surveys.* John Wiley & Sons, Inc., New York.

Pahkinen, E. & Lehtonen, R. 1989: *Otanta-asetelmat ja tilastollinen analyysi.* Oy Gaudeamus Ab, Helsinki.

Siegel, S., Castellan, N. Jr. 1988: *Nonparametric Statistics for the behavioral sciences.* (Second edition) McGraw Hill, New York

Parittaiset vertailut LSD–menetelmällä, merkitsevyydet

	Keskusta	Kortepohja	Keltinmäki	Keljo	Kuokkala	Aittorinne
Keskusta	1					
Kortepohja	0.747	1				
Keltinmäki	0.011	0.037	1			
Keljo	0.958	0,801	0.020	1		
Kuokkala	0.819	0.930	0.030	0.869	1	
Aittorinne	0.105	0.223	0.386	0.142	0.192	1
Kangasvuori	0.069	0.160	0.496	0.098	0.136	0.852
Lohikoski	0.000	0.000	0.149	0.000	0.000	0.022
Palokka	0.018	0.055	0.869	0.030	0.045	0.482
Vaajakoski	0.309	0.536	0.108	0.374	0.476	0.494
Muurame	0.015	0.054	0.775	0.028	0.043	0.524
Säynätsalo	0.000	0.000	0.026	0.000	0.000	0.003
Laukaa	0.790	0.569	0.005	0.765	0.635	0.062

	Kangasvuori	Lohikoski	Palokka	Vaajakoski	Muurame	Säynätsalo	Laukaa
Keskusta							
Kortepohja							
Keltinmäki							
Keljo							
Kuokkala							
Aittorinne							
Kangasvuori	1						
Lohikoski	0.035	1					
Palokka	0.606	0.109	1				
Vaajakoski	0.337	0.002	0.152	1			
Muurame	0.661	0.700	0.912	0.156	1		
Säynätsalo	0.055	0.352	0.017	0.000	0.010	1	
Laukaa	0.039	0.000	0.009	0.200	0.007	0.000	1

Kaupan valinta logit –malleilla, vedonlyöntisuhteet

Malli: Kävijä2 = Vakio+ASUU+KULKULUO.

Asuinalue	Auto	Muu kulkuväline
Keskusta	$\frac{m_{111}}{m_{211}}$ =48.42	$\frac{m_{112}}{m_{212}}$ = 4.33
Kortepohja	$\frac{m_{121}}{m_{221}}$ =12.92	$\frac{m_{122}}{m_{222}}$ =1.16
Keltinmäki	$\frac{m_{131}}{m_{231}}$ = 5.40	$\frac{m_{132}}{m_{232}}$ = 0.48
Keljo	$\frac{m_{141}}{m_{241}}$ = 44.26	$\frac{m_{142}}{m_{242}}$ = 3.96
Kuokkala	$\frac{m_{151}}{m_{251}}$ = 4.20	$\frac{m_{152}}{m_{252}}$ = 0.38
Aittorinne	$\frac{m_{161}}{m_{261}}$ = 17.39	$\frac{m_{162}}{m_{262}}$ = 1.55
Kangasvuori	$\frac{m_{171}}{m_{271}}$ = 9.83	$\frac{m_{172}}{m_{272}}$ = 0.88
Lohikoski	$\frac{m_{181}}{m_{281}}$ = 6.28	$\frac{m_{182}}{m_{282}}$ = 0.56
Palokka	$\frac{m_{191}}{m_{291}}$ = 1.41	$\frac{m_{192}}{m_{292}}$ = 0.13
Vaajakoski	$\frac{m_{1,10,1}}{m_{2,10,1}}$ = 1.98	$\frac{m_{1,10,2}}{m_{2,10,2}}$ = 0.18
Muurame	$\frac{m_{1,11,1}}{m_{2,11,1}}$ = 0.41	$\frac{m_{1,11,2}}{m_{2,11,2}}$ = 0.04
Säynätsalo	$\frac{m_{1,12,1}}{m_{2,12,1}}$ = 0.64	$\frac{m_{1,12,2}}{m_{2,12,2}}$ = 0.06
Laukaa	$\frac{m_{1,13,1}}{m_{2,13,1}}$ = 0.06	$\frac{m_{1,13,2}}{m_{2,13,2}}$ = 0.006

Kaupan valinta logit –malleilla, vedonlyöntisuhteet

Malli: Kävijä2 = Vakio+ASUU+RAHALUO

Asuinalue	Rahaa < med	Rahaa ≥ med
Keskusta	$\frac{m_{111}}{m_{211}} = 4.879$	$\frac{m_{112}}{m_{212}} = 6.659$
Kortepohja	$\frac{m_{121}}{m_{221}} = 1.809$	$\frac{m_{122}}{m_{222}} = 2.470$
Keltinmäki	$\frac{m_{131}}{m_{231}} = 2.588$	$\frac{m_{132}}{m_{232}} = 3.532$
Keljo	$\frac{m_{141}}{m_{241}} = 22.332$	$\frac{m_{142}}{m_{242}} = 30.478$
Kuokkala	$\frac{m_{151}}{m_{251}} = 1.385$	$\frac{m_{152}}{m_{252}} = 1.891$
Aittorinne	$\frac{m_{161}}{m_{261}} = 3.880$	$\frac{m_{162}}{m_{262}} = 5.296$
Kangasvuori	$\frac{m_{171}}{m_{271}} = 2.643$	$\frac{m_{172}}{m_{272}} = 3.607$
Lohikoski	$\frac{m_{181}}{m_{281}} = 2.907$	$\frac{m_{182}}{m_{282}} = 3.967$
Palokka	$\frac{m_{191}}{m_{291}} = 0.967$	$\frac{m_{192}}{m_{292}} = 1.329$
Vaajakoski	$\frac{m_{1,10,1}}{m_{2,10,1}} = 1.130$	$\frac{m_{1,10,2}}{m_{2,10,2}} = 1.542$
Muurame	$\frac{m_{1,11,1}}{m_{2,11,1}} = 0.240$	$\frac{m_{1,11,2}}{m_{2,11,2}} = 0.327$
Säynätsalo	$\frac{m_{1,12,1}}{m_{2,12,1}} = 0.422$	$\frac{m_{1,12,2}}{m_{2,12,2}} = 0.576$
Laukaa	$\frac{m_{1,13,1}}{m_{2,13,1}} = 0.005$	$\frac{m_{1,13,2}}{m_{2,13,2}} = 0.067$

Kaupan valinta logit –malleilla, vedonlyöntisuhteet

Malli:kävijä2=ASUU+KOKO

Asuinalue	Yksi	Kaksi	Kolme
Keskusta	$\frac{m_{111}}{m_{211}} = 2.69$	$\frac{m_{112}}{m_{212}} = 12.782$	$\frac{m_{113}}{m_{213}} = 6.987$
Kortepohja	$\frac{m_{121}}{m_{221}} = 0.913$	$\frac{m_{122}}{m_{222}} = 4.345$	$\frac{m_{123}}{m_{223}} = 2.375$
Keltinmäki	$\frac{m_{131}}{m_{231}} = 0.554$	$\frac{m_{132}}{m_{232}} = 2.635$	$\frac{m_{133}}{m_{233}} = 1.441$
Keljo	$\frac{m_{141}}{m_{241}} = 9.098$	$\frac{m_{142}}{m_{242}} = 43.293$	$\frac{m_{143}}{m_{243}} = 23.665$
Kuokkala	$\frac{m_{151}}{m_{251}} = 0.522$	$\frac{m_{152}}{m_{252}} = 2.484$	$\frac{m_{153}}{m_{253}} = 1.358$
Aittorinne	$\frac{m_{161}}{m_{261}} = 1.931$	$\frac{m_{162}}{m_{262}} = 9.189$	$\frac{m_{173}}{m_{273}} = 5.023$
Kangasvuori	$\frac{m_{171}}{m_{271}} = 1.405$	$\frac{m_{172}}{m_{272}} = 6.686$	$\frac{m_{173}}{m_{273}} = 3.655$
Lohikoski	$\frac{m_{181}}{m_{281}} = 1.050$	$\frac{m_{182}}{m_{282}} = 5.000$	$\frac{m_{183}}{m_{283}} = 2.732$
Palokka	$\frac{m_{191}}{m_{291}} = 0.398$	$\frac{m_{192}}{m_{292}} = 1.893$	$\frac{m_{193}}{m_{293}} = 1.035$
Vaajakoski	$\frac{m_{1,10,1}}{m_{2,10,1}} = 0.457$	$\frac{m_{1,10,2}}{m_{2,10,2}} = 2.175$	$\frac{m_{1,10,3}}{m_{2,10,3}} = 1.189$
Muurame	$\frac{m_{1,11,1}}{m_{2,11,1}} = 0.019$	$\frac{m_{1,11,2}}{m_{2,11,2}} = 0.440$	$\frac{m_{1,11,3}}{m_{2,11,3}} = 0.240$
Säynätsalo	$\frac{m_{1,12,1}}{m_{2,12,1}} = 0.154$	$\frac{m_{1,12,2}}{m_{2,12,2}} = 0.733$	$\frac{m_{1,12,3}}{m_{2,12,3}} = 0.401$
Laukaa	$\frac{m_{1,13,1}}{m_{2,13,1}} = 0.019$	$\frac{m_{1,13,2}}{m_{2,13,2}} = 0.090$	$\frac{m_{1,13,3}}{m_{2,13,3}} = 0.049$

Kaupan valinta logit –malleilla, vedonlyöntisuhteet

Asuinalue	Neljä	Viisi tai enemmän
Keskusta	$\frac{m_{114}}{m_{214}} = 7.330$	$\frac{m_{115}}{m_{215}} = 9.670$
Kortepohja	$\frac{m_{124}}{m_{224}} = 2.492$	$\frac{m_{125}}{m_{225}} = 3.287$
Keltinmäki	$\frac{m_{134}}{m_{234}} = 1.511$	$\frac{m_{135}}{m_{235}} = 1.994$
Keljo	$\frac{m_{144}}{m_{244}} = 24.829$	$\frac{m_{145}}{m_{245}} = 32.753$
Kuokkala	$\frac{m_{154}}{m_{254}} = 1.425$	$\frac{m_{155}}{m_{255}} = 1.880$
Aittorinne	$\frac{m_{164}}{m_{264}} = 5.270$	$\frac{m_{165}}{m_{265}} = 6.952$
Kangasvuori	$\frac{m_{174}}{m_{274}} = 3.834$	$\frac{m_{175}}{m_{275}} = 5.048$
Lohikoski	$\frac{m_{184}}{m_{284}} = 2.866$	$\frac{m_{185}}{m_{285}} = 3.781$
Palokka	$\frac{m_{194}}{m_{294}} = 1.086$	$\frac{m_{195}}{m_{295}} = 1.432$
Vaajakoski	$\frac{m_{1,10,4}}{m_{2,10,4}} = 1.247$	$\frac{m_{1,10,5}}{m_{2,10,5}} = 1.645$
Muurame	$\frac{m_{1,11,4}}{m_{2,11,4}} = 0.252$	$\frac{m_{1,11,5}}{m_{2,11,5}} = 0.333$
Säynätsalo	$\frac{m_{1,12,4}}{m_{2,12,4}} = 0.420$	$\frac{m_{1,12,5}}{m_{2,12,5}} = 0.554$
Laukaa	$\frac{m_{1,13,4}}{m_{2,13,4}} = 0.052$	$\frac{m_{1,13,5}}{m_{2,13,5}} = 0.068$

Kaupan valinta logit –malleilla, vedonlyöntisuhteet

Malli:Kävijä2 = Vakio+KOKO+KULKULUO

Koko	Auto	Muu kulkuväline
1 henkilö	$\frac{m_{111}}{m_{211}} = 2.138$	$\frac{m_{112}}{m_{212}} = 1.172$
2 henkilöä	$\frac{m_{121}}{m_{221}} = 4.183$	$\frac{m_{122}}{m_{222}} = 2.293$
3 henkilöä	$\frac{m_{131}}{m_{231}} = 1.351$	$\frac{m_{132}}{m_{232}} = 0.741$
4 henkilöä	$\frac{m_{141}}{m_{241}} = 1.713$	$\frac{m_{142}}{m_{242}} = 0.039$
5 tai useampi	$\frac{m_{151}}{m_{251}} = 1.813$	$\frac{m_{152}}{m_{252}} = 0.994$

Malli:kavijalu2 = Vakio+KOKO+RAHALUO

Ikä	Auto	Muu kulkuväline
Yksi henkilö	$\frac{m_{111}}{m_{211}} = 1.215$	$\frac{m_{112}}{m_{212}} = 1.452$
Kaksi henkilöä	$\frac{m_{121}}{m_{221}} = 3.350$	$\frac{m_{122}}{m_{222}} = 2.804$
Kolme henkilöä	$\frac{m_{131}}{m_{231}} = 1.078$	$\frac{m_{132}}{m_{232}} = 1.288$
Neljä henkilöä	$\frac{m_{141}}{m_{241}} = 1.355$	$\frac{m_{142}}{m_{242}} = 1.619$
Viisi tai enemmän	$\frac{m_{151}}{m_{251}} = 1.340$	$\frac{m_{152}}{m_{252}} = 1.602$

Kaupan valinta logit –malleilla, vedonlyöntisuhteet

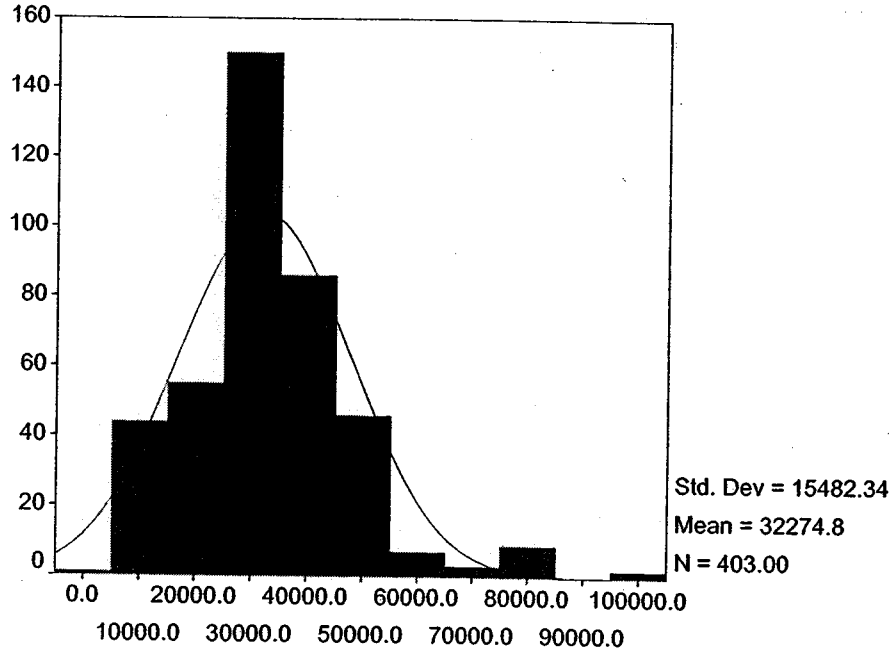
Malli:kavijalu2 = Vakio+IKÄ+KULKULUO

Ikä	Auto	Muu kulkuväline
15–24 vuotta	$\frac{m_{111}}{m_{211}} = 2.849$	$\frac{m_{112}}{m_{212}} = 1.728$
25–34 vuotta	$\frac{m_{121}}{m_{221}} = 3.804$	$\frac{m_{122}}{m_{222}} = 2.307$
35–44 vuotta	$\frac{m_{131}}{m_{231}} = 1.833$	$\frac{m_{132}}{m_{232}} = 0.657$
45–59 vuotta	$\frac{m_{141}}{m_{241}} = 1.976$	$\frac{m_{142}}{m_{242}} = 1.198$
Yli 60 vuotta	$\frac{m_{151}}{m_{251}} = 2.252$	$\frac{m_{152}}{m_{252}} = 1.366$

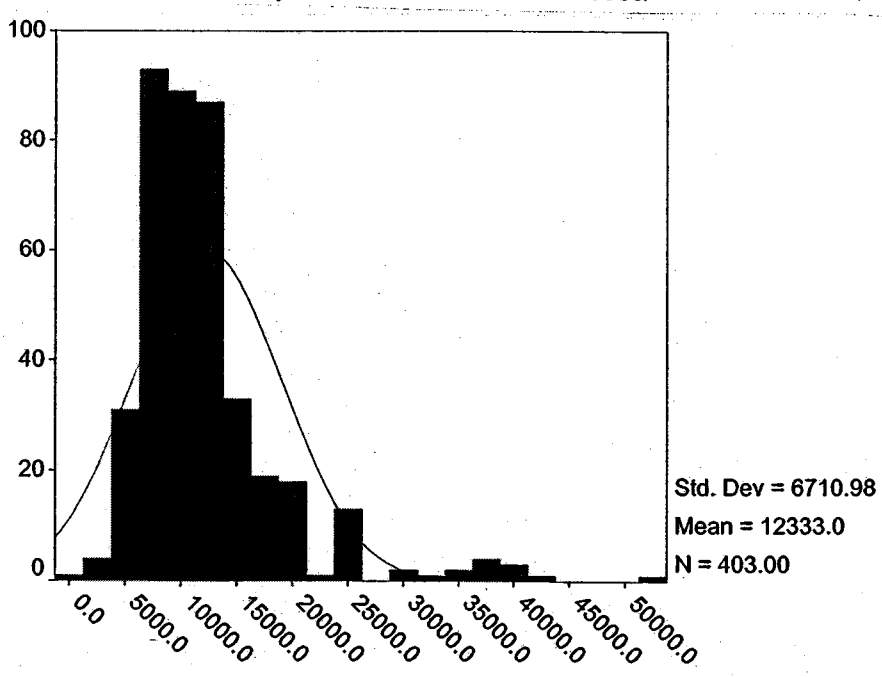
Malli:kavijalu2 = Vakio+IKÄ+RAHAALUO

Ikä	Raha $\leq$ med	Raha $>$ med
15–24 vuotta	$\frac{m_{111}}{m_{211}} = 2.270$	$\frac{m_{112}}{m_{212}} = 3.001$
25–34 vuotta	$\frac{m_{121}}{m_{221}} = 2.697$	$\frac{m_{122}}{m_{222}} = 3.564$
35–44 vuotta	$\frac{m_{131}}{m_{231}} = 0.751$	$\frac{m_{132}}{m_{232}} = 0.993$
45–59 vuotta	$\frac{m_{141}}{m_{241}} = 1.390$	$\frac{m_{142}}{m_{242}} = 1.837$
Yli 60 vuotta	$\frac{m_{151}}{m_{251}} = 1.384$	$\frac{m_{152}}{m_{252}} = 1.831$

Muuttujien jakaumat

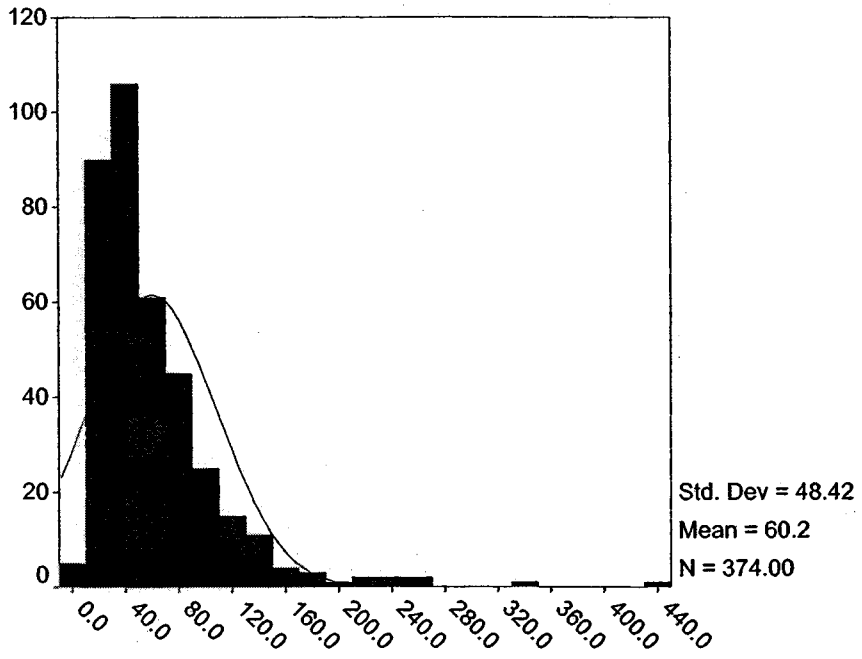


Kotitalouden käyttämä rahamäärä vuodessa

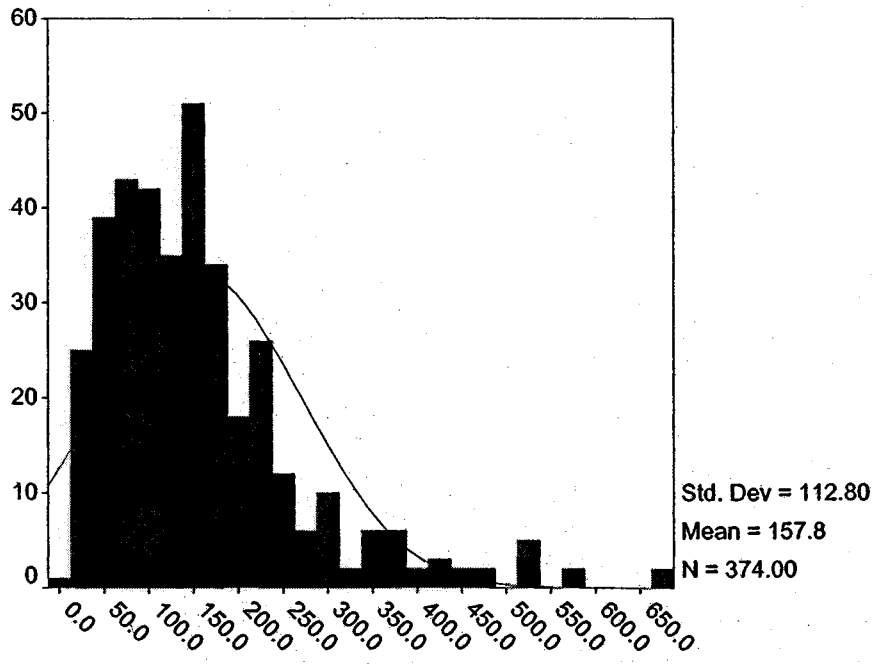


Kotitalouden käyttämä rahamäärä asukasta kohden vuodessa

Muuttujien jakaumat

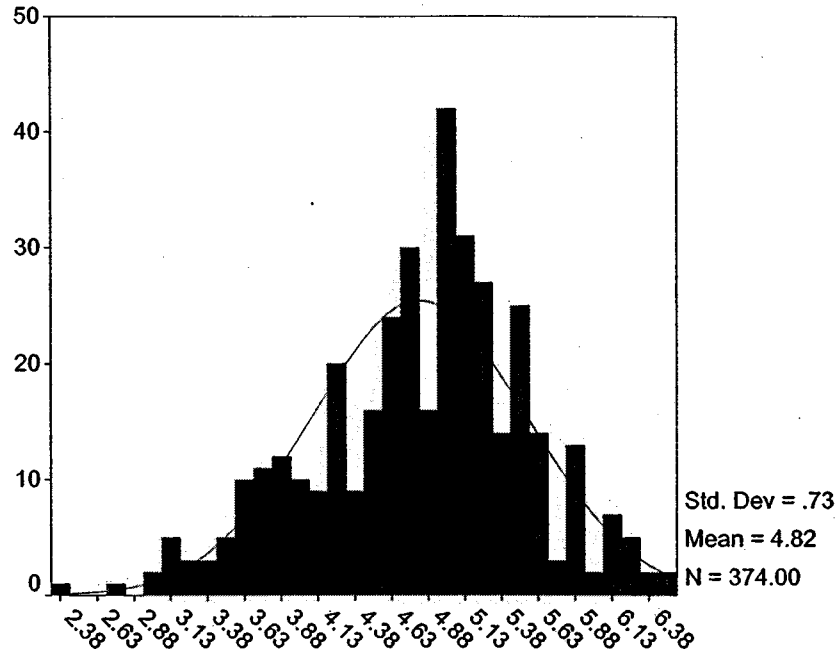


Rahan kulutus ostoskerralla asukasta kohden

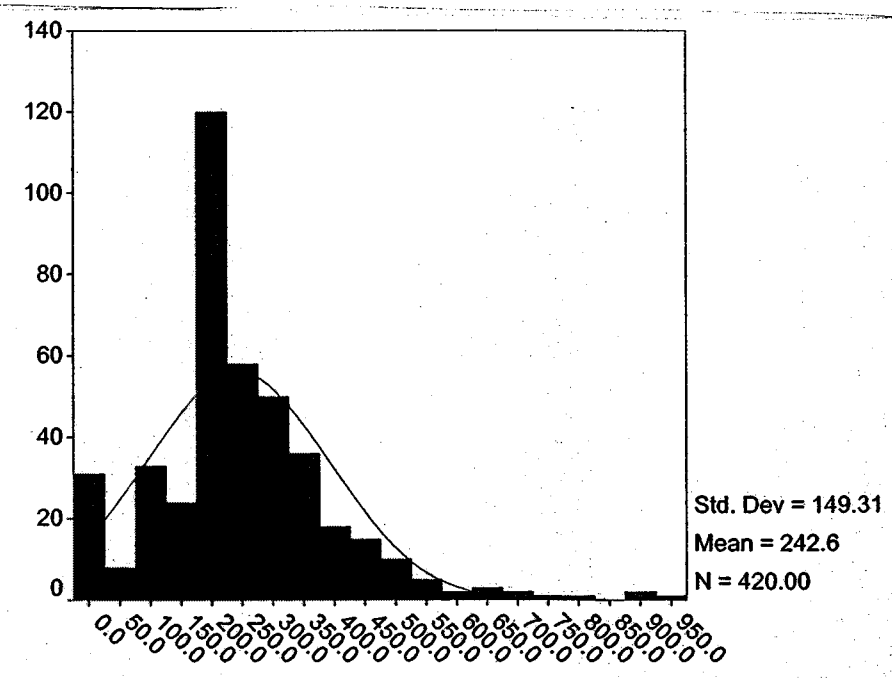


Kotitalouden käyttämä rahamäärä ostoskerralla

Muuttujien jakaumat



Kotitalouden käyttämä rahamäärä ostoskerralla (logaritmimuunnett



Kotitalouden tekemät kaupassakäynnit vuodessa

A	Anttila
B	Citymarket
C	Mestarin Herkku
D	Minimani
E	Seppälän Prisma
F	Keljonkeskuksen Prisma
G	K-Supermarket Tourula
H	K-Supermarket Länsiväylä
I	K-Lähikauppa Kymppi
J	K-Market Kotikenttä