Imetykseen vaikuttavien tekijöiden etsiminen multinomiaalisella logistisella regressiolla

Satu Jantunen

Pro gradu -tutkielma
Kesäkuu 2022

MATEMATIIKAN JA TILASTOTIETEEN LAITOS
TURUN YLIOPISTO

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

JANTUNEN, SATU: Imetykseen vaikuttavien tekijöiden etsiminen multinomiaalisella logistisella regressiolla
Pro gradu -tutkielma, sivumäärä s. 47, 14.
Sovellettu matematiikka
Kesäkuu 2022

Tässä Pro gradu -tutkielmassa käsitellään logistista regressiota ja sen erikoistapausta, multinomiaalista logistista regressiota. Menetelmää hyödynnetään soveltavassa osiossa mallintamaan sosiaalisen ympäristön vaikutusta lapsen imetykseen.

Hyvän kasvun avaimet (HKA) -tutkimus on Turun yliopiston koordinoima seurantatutkimus, jossa kerätään tietoa lasten ja perheiden sosiaalisesta, psyykkisestä ja fyysisestä hyvinvoinnista. Tämän tiedon avulla pyritään tuottamaan tietoa siitä, miten lasten ja perheiden hyvinvointia voidaan jatkossa tukea. Soveltavan osion materiaali saatiin HKA-tutkimuksesta ja tavoitteena oli selvittää, mitkä sosiaaliset tekijät vaikuttivat lasten imetykseen. Imetykseen vaikuttavia tekijöitä tutkittiin, koska imetyksellä on todettu olevan paljon hyviä vaikutuksia niin lapsen kuin äidinkin hyvinvointiin.

Tutkimuksessa käytetään multinomiaalista logistista regressiota logistisen regression sijaan, koska vastemuuttuja on kolmiluokkainen. Logistinen regressio sopii kaksiluokkaisen vasteen tutkimiselle, mutta kolmen tai useampiluokkaisen vasteen tutkimiseen tarvitaan multinomiaalista logistista regressiota.

Aineiston analysoinnissa käytetään R-ohjelmiston versiota 4.0.3. Tilastollisen merkitsevyyden rajana pidettiin p-arvoa 0.05.

Tutkielman alussa käydään läpi soveltavassa osassa käytettyjen menetelmien teoriaa. Sen jälkeen siirrytään logistiseen regressioon, mistä jatketaan multinomiaaliseen erikoistapaukseen. Tutkielman loppuosiossa esitetään soveltava osio. Liitteissä on esitelty tehty R-koodi.

Asiasanat: logistinen regressio, multinomiaalinen logistinen regressio, HKA-tutkimus, lapsen imetys.

Sisältö

1 Johdanto

Hyvän kasvun avaimet (HKA) -tutkimuksessa kerätään tietoa Varsinais-Suomen sairaanhoitopiirin alueella syntyneistä lapsista. Tavoitteena on selvittää, mitkä tekijät vaikuttavat lasten hyvinvointiin ja kehitykseen sekä miten näitä voitaisiin edistää. [54] Tämä tutkimus keskittyy lasten imetykseen ja sen kestoon. Tarkoituksena on selvittää, mitkä tekijät vaikuttavat imetykseen ja kuinka suurta vaikutus on. Imetystä tutkitaan, koska sillä on todettu olevan terveyttä edistäviä vaikutuksia niin lapselle kuin äidillekin [42–44].

Aineisto jakautuu imetyksen mukaan kolmeen ryhmään. Näin voidaan tutkia lapsia, joita ei ole imetetty, lapsia, joita on imetetty alle neljä kuukautta sekä lapsia, joita on imetetty yli neljä kuukautta. Koska vastemuuttujana oleva imetys on kolmiluokkainen, niin aineistoa tutkitaan multinomiaalisen logistisen regression avulla [26,32]. Tällöin vasteluokat voivat olla kategoriset ja niiden ei tarvitse olla tietyn etäisyyden päässä toisistaan [26, 32].

Tutkielman aluksi käydään läpi menetelmiä, joita on tarvittu multinomiaalisen logistisen regression vaatimusten tutkimiseen. Sitten tutkitaan, ovatko aineisto ja sen muuttujat sopivia tähän menetelmään. Seuraavaksi esitellään, miten t-testi, levenen testi, χ^2 -testi, Cramerin V sekä Hosmerin ja Lemeshowin testi toimivat ja mihin niitä hyödynnetään tässä työssä.

Jotta voitaisiin ymmärtää multinomiaalisen logistisen regression toimintaa, on ensin käytävä läpi, mikä on kerroinsuhde ja miten binäärinen logistinen regressio toimii. Näiden jälkeen on helppo siirtyä multinomiaaliseen logistiseen regressioon.

Työssä tutkitaan eri menetelmin onko malli hyvä eli kuvaako se aineistoa riittävän hyvin. Tämän selvittämiseksi käytetään Hosmerin ja Lemeshowin testiä sekä residuaaleja [24, 26, 31, 40]. Tässä vaiheessa on hyvä lisäksi tarkastella, että mallin käytön kaikki ehdot täyttyvät.

Teoriaosuuksien jälkeen päästään tutkielman soveltavaan osioon. Tässä osassa käytetään HKA-tutkimuksesta saatua aineistoa ja tutkitaan sen sopivuutta multinomiaaliseen logistiseen regressioon. Aluksi tutustutaan aineistoon ja sen alkuperään. Tämän jälkeen muodostetaan multinomiaalisen logistisen regression avulla malli ja lasketaan OR-luvut. Näiden lukujen avulla voidaan tulkita, mitkä tekijät vaikuttavat lapsen imetykseen. Lopuksi vielä varmistetaan, että malli on aineistoon sopiva.

2 Käytetyt menetelmät

Tässä kappaleessa perehdytään tarkemmin soveltavassa osiossa tarvittaviin menetelmiin. Käydään läpi käytettyjen menetelmien teoriaa, jotta ne tulevat tutuiksi. Ensin lähdetään liikkeelle peruskäsitteistä. Tutkittavaa muuttujaa kutsutaan selitettäväksi muuttujaksi eli vasteeksi tai vastemuuttujaksi [1, 2]. Vasteen vaihtelua pyritään selittämään selittävien muuttujien avulla [1–3].

Vastemuuttujan ja selittävän muuttujan yhdistävä linkki on regressiokerroin. Se kertoo selittävän muuttujan tärkeyden vastemuuttujan vaihteluun, kun muut selittävät muuttujat ovat vakioita. Tämä merkitään symbolilla β. Regressiokerroin kuvaa siten selittävien muuttujien suhdetta vastemuuttujaan muiden selittävien muuttujien vaikutuksen ollessa vakiona. Tällöin nähdään, onko juuri kyseessä olevalla selittävällä muuttujalla vaikutusta vastemuuttujaan ja kuinka suurta mahdollinen vaikutus on. [4] Regressiokerrointa voidaan kutsua myös regressioparametriksi [5].

Parametri tarkoittaa teoreettisen jakauman tunnuslukua eli perusjoukon ominaisuutta, jota estimoidaan otoksen avulla [5]. Näitä arvioita merkitään estimaattoreilla. Regression lisäksi käytössä ovat seuraavat parametrit ja niitä vastaavat estimaattorit: [6]

Parametri	Estimaattori
Odotusarvo μ	(Otos)keskiarvo $\bar{x}$
Hajonta σ	(Otos)hajonta s
Varianssi σ^2	(Otos)varianssi s^2 .

Parametrivektori on vektori, joka sisältää parametreja. Tälläinen on esimerkiksi vektori, joka sisältää kaikki regressiokertoimet $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_n)$.

Varianssi kuvaa tilastollista hajontaa. Varianssi kertoo keskimäärin kuinka paljon satunnaismuuttujan arvojen neliöidyt poikkeamat ovat keskiarvosta. Varianssin voi myös ajatella olevan keskihajonnan neliö eli neliö havaintojen sijoittumisesta keskiarvon ympärille [4]

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1},$$

missä s^2 = varianssi, x_i = yksittäisen havainnon arvo, $\bar{x}$ = havaintojen arvojen keskiarvo ja n = havaintojen lukumäärä [4]. Keskihajonta kuvaa aineiston hajautuneisuutta [7]. Se kertoo, miten kaukana keskimäärin havainnot ovat keskiarvosta [7]. Selittäviä muuttujia tarkasteltaessa on ensin tutkittava ovatko muuttujat riippuvaisia toisistaan.

Määritelmä 1. [8] Satunnaismuuttujat x ja y ovat *toisistaan riippumattomia*, jos

$$P(\{x \in A\} \cap \{y \in B\}) = P(x \in A)P(y \in B)$$

kaikilla mitallisilla joukoilla $A \subset \mathbb{R}$ ja $B \subset \mathbb{R}$.

Satunnaismuuttujat x ja y ovat siis riippumattomia, jos tapahtumat ovat riippumattomia kaikilla $A \subset \mathbb{R}$ ja $B \subset \mathbb{R}$. Jos tapahtumat A ja B ovat toisistaan riippumattomat, niin ne eivät vaikuta toistensa tapahtumien todennäköisyyksiin eli A:n tapahtuessa B:n tapahtuminen on yhtä todennäköistä kuin jos A ei tapahdu. Tapahtuman A tapahtuminen ei siis vaikuta tapahtuman B tapahtumiseen ja päinvastoin. [8]

Muuttujien välisiä suhteita tarkastellessa asetetaan jokin hypoteesi, jota lähdetään tutkimaan. Nollahypoteesi tarkoittaa testattavaa oletusta. Tämä on yleensä muotoa ”ei eroa” tai ”ei vaikutusta” eli tutkittavan oletuksen päinvastainen oletus. Esimerkkinä tästä on oletus, ettei ryhmien välillä ole tilastollista eroavaisuutta. Nollahypoteesin merkintä on H_0. Nollahypoteesin lisäksi tehdään toinen hypoteesi, joka on voimassa, jos nollahypoteesi kumoutuu. Vaihtoehtoinen hypoteesi tarkoittaa oletusta, joka tulee voimaan, jos nollahypoteesi kumotaan. Tämä on yleensä tutkimushypoteesin mukainen vaihtoehto ja on yleensä muotoa ”eroa on” tai ”vaikutusta on”. Esimerkkinä tästä on tilanne, että ryhmien välillä on tilastollista eroavaisuutta. Vaihtoehtoisen hypoteesin merkintä on H_1. [4, 9]

Oleellista hypoteeseja tutkittaessa on, että tutkittavien asioiden välillä löytyy tilastollinen merkitsevyys. Tämä tarkoittaa sitä, että tulos ei ole vain sattumaa. Esimerkiksi ryhmien väliset erot tai riippuvuudet voivat olla joko tilastollisesti merkitseviä tai vain sattumasta johtuvia. [4]

Jotta voisimme yleistää otosten tuloksen koskemaan perusjoukkoa, on tuloksen oltava tilastollisesti merkitsevä. Tätä tutkitaan yleensä p-arvon avulla. P-arvo kertoo virheellisen päätelmän todennäköisyyden. Tämä arvo saadaan tilastollisen testin tuloksena. Tilastollisen merkitsevyyden rajana voi pitää esimerkiksi arvoa 0.05, kuten tutkielman soveltavassa osiossa tehdään. [10] P-arvo saadaan yleensä jakaumien taulukoiduista arvoista. Tämän takia käydään läpi seuraavaksi t-jakauma.

2.1 T-testi

T-testi perustuu t-jakaumaan, joka on symmetrinen, normaalijakauman muotoa muistuttava, mutta hieman latteampi jakauma, ja jonka keskikohta on noin kohdassa nolla. Käyrän alle jäävän pinta-alan koko on 1, niin kuin

kaikissa todennäköisyysjakaumissa. Jakauman käyrä ei koskaan tavoita x-akselia eli x-akseli on jakauman asymptootti. Mitä suurempi vaihtelevien muuttujien lukumäärä eli vapausaste on, sitä lähempänä normaalijakaumaa t-jakauma on. Kun keskiarvoja tarvitsee verrata toisiinsa, käytetään usein t-jakaumaa. [11, 12]

T-testi vertaa kahta jatkuvien muuttujien ryhmää toisiinsa. Se kertoo, onko ryhmien välillä tilastollisesti merkitseviä eroavaisuuksia. Nollahypoteesina on, että ryhmien välillä ei ole tilastollista eroavaisuutta. Testin tuloksena saadaan p-arvo, jonka mukaan voidaan määrittää, onko nollahypoteesi voimassa vai kumoutuuko se. Jatkossa käytetään p-arvolle rajaa 0.05. Eli jos p-arvo on pienempi kuin 0.05, niin nollahypoteesi kumoutuu ja kahden ryhmän välillä on tilastollisesti merkitsevää eroavaisuutta. Jos taas p-arvo on 0.05 tai suurempi, nollahypoteesi jää voimaan ja tilastollista eroavaisuutta ei ole. [13]

Tässä tutkielmassa käytetään kahden otoksen t-testiä. Otos on ryhmästä valittu joukko, jota tutkitaan. Ryhmä on tutkimuksen perusjoukko, josta halutaan kerätä tietoa. Ryhmästä valittua otosta käytetään koko ryhmän sijaan, sillä yleensä ryhmä on kooltaan liian iso tutkittavaksi. T-testin oletuksina on, että vaste eli selitettävä muuttuja on numeerinen ja normaalisti jakautunut. Selittävän muuttujan on puolestaan oltava kategorinen ja kaksiluokkainen. Tutkittavien ryhmien tulee olla toisistaan riippumattomia. T-testissä verrataan kahden toisistaan riippumattoman otoksen keskiarvoja. Nollahypoteesi on, että ryhmien keskiarvoissa ei ole eroa eli $H_0 : \mu_1 = \mu_2$ ja vaihtoehtoinen hypoteesi on, että ryhmien välillä on tilastollisesti merkitsevää eroavaisuutta eli $H_1 : \mu_1 \neq \mu_2$. [14]

T-testin lisäoletuksena on molempien ryhmien varianssien yhtäsuuruus. Tätä voidaan testata Levenen testillä. Tuloksen perusteella testisuureen laskutapa valitaan kahdesta mahdollisesta tavasta. [14]

Testimuuttuja on muotoa

$$t = \frac{\text{ero ryhmien välillä}}{\text{ero ryhmien sisällä}},$$

kun ero ryhmien sisällä on erisuurta kuin nolla. Yhtälöstä voidaan päätellä, että kun testimuuttuja t saa suuren arvon, ero on isompaa ryhmien välillä kuin ryhmien sisällä. Tämä tarkoittaa siis, että ryhmien välillä on eroavaisuutta. Jos testimuuttuja t saa pienen arvon, ero ryhmien sisällä on suurempaa kuin ryhmien välillä. Tällöin ryhmien välillä ei siis ole paljon eroa. Kuitenkaan tästä ei voida vielä päätellä, onko eroavaisuus tilastollisesti merkitsevää. [15]

Saatua testimuuttujan t arvoa verrataan liitteen B taulukosta 20 t-jakauman taulukoituihin arvoihin. Taulukosta saadaan p-arvo, kun tiedetään vapausas-

te. Valitaan vapausasteen mukainen rivi ja siltä riviltä etsitään testimuuttujaa lähinnä olevat luvut. Tämän välin avulla saadaan p-arvolle arvioitu väli. Seurataan sarakkeita, joilla testimuuttujaa lähinnä olevat luvut sijaitsevat. Sarakkeiden alaosassa on listattu merkitsevyystasot kaksisuuntaisissa testeissä. Tältä riviltä saadaan sarakkeiden kohdalta p-arvot. Näin saadaan väli, mihin testiarvon p-arvo sijoittuu.

Monet ohjelmistot kertovat testimuuttujan lisäksi suoraan myös sitä vastaavan p-arvon, jolloin saadaan tarkemmat p-arvot. Näin toimii myös R-studio. Saadun p-arvon perusteella tiedetään, onko mahdollinen ryhmien välillä oleva ero tilastollisesti merkitsevä vai ainoastaan sattumasta johtuva ero [10]. P-arvon perusteella voidaan näin ollen joko hyväksyä tai hylätä nollahypoteesi [10].

Ero ryhmien välillä on ryhmien keskiarvojen ero eli $\mu_1 - \mu_2$. Ryhmien keskiarvoja ei kuitenkaan tiedetä, vaan käytetään sen sijaan ryhmän keskiarvoa vastaavia otosten keskiarvoja $\bar{x}_1 - \bar{x}_2$. Ero ryhmien sisällä lasketaan kaavoilla [15]

$$\begin{aligned} SE(y) &= \sqrt{\frac{s^2}{n}} \\ v\bar{a}r(y) &= \frac{s_1^2}{n_1} \\ v\bar{a}r(x-y) &= v\bar{a}r(x) + v\bar{a}r(y), \end{aligned}$$

kun x ja y ovat toisistaan riippumattomia satunnaismuuttujia, SE on keskiarvon keskivirhe eli ero ryhmien sisällä, s^2 on otoksen varianssi, n on otoksen havaintojen lukumäärä ja $v\bar{a}r$ on varianssin keskiarvo. [15]

2.1.1 Erimuotoiset varianssit

Kun ryhmien varianssit ovat erisuuruiset, ero ryhmien sisällä lasketaan seuraavasti:

$$\begin{aligned} v\bar{a}r(x_1 - x_2) &= v\bar{a}r(x_1) + v\bar{a}r(x_2) \\ &= \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}. \end{aligned}$$

Tästä seuraa, että keskiarvon keskivirhe on seuraava

$$SE(x_1 - x_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \tag{1}$$

Näin ollen tästä saadaan, että t-arvo on [16]

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

kun s_1 ja s_2 eivät molemmat ole nollia. Yhtälössä $\bar{x}_1$ ja $\bar{x}_2$ ovat otoksista lasketut keskiarvot, s_1 ja s_2 ovat otosten keskihajonnat, s_1^2 ja s_2^2 ovat otosten varianssit ja n_1 ja n_2 ovat otosten havaintojen lukumäärät. [16, 17]

2.1.2 Yhdenmuotoiset varianssit

Kun ryhmien varianssit ovat samansuuruiset, eron ryhmien sisällä saa laskettua seuraavasti. Koska ryhmien varianssien oletetaan olevan suunnilleen samat, saadaan variansseista yhdistetty keskihajonta laskemalla niiden painotettu keskiarvo eli keskiarvo, jossa otetaan huomioon jokaisen muuttujan painokerroin. [17] Painokerroin on tässä tapauksessa havaintojen lukumäärä miinus yksi eli

$$s_{yhdistetty}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

Yhtälön (1) mukaan saadaan seuraava tulos:

$$\begin{aligned} SE(x_1 - x_2) &= \sqrt{\frac{s_{yhdistetty}^2}{n_1} + \frac{s_{yhdistetty}^2}{n_2}} \\ &= \sqrt{s_{yhdistetty}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \end{aligned}$$

jolloin t-arvoksi saadaan

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

missä $\bar{x}_1$ ja $\bar{x}_2$ ovat otosten keskiarvot, s_1^2 ja s_2^2 ovat varianssit ja n_1 ja n_2 ovat havaintojen lukumäärät [17].

2.2 Levenen testi

Levenen testin avulla tarkistetaan ovatko varianssit yhdenmuotoiset vai erimuotoiset t-testin testimuuttujan laskutavan valintaa varten. Tässä nollahypoteesina on, ettei kahden otoksen varianssien välillä ole eroa. Jos Levenen

testin p-arvo on yli 0.05, nollahypoteesi jää voimaan. Tällöin voidaan todeta, että testi tukee varianssien yhtäsuuruusoletusta. Jos tulos on 0.05 tai sen alle, nollahypoteesi kumotaan ja todetaan, että varianssit ovat erisuuret. [14]

Levenen testi vertailee eroa ryhmien sisäisten varianssien ja ryhmien välisten varianssien välillä. Tämän vertailun avulla voidaan päätellä, onko mahdollinen ryhmien varianssien välillä oleva eroavaisuus merkityksellistä vai ei. Testin tuloksena saatua arvoa verrataan taulukoituihin F-jakauman arvoihin vapausasteilla $1-k$ ja $N-k$, kun k on ryhmien lukumäärä ja N on havaintojen yhteislukumäärä. [18] F-jakauma on oikealle vino ja sen muodon määrittävät nimittäjän ja osoittajan vapausasteet [19]. Tämän takia F-jakaumia on paljon eri huipukkuuksilla eli huipun terävyyksillä [19].

Liitteen B taulukosta 21 saadaan tarvittava p-arvo, katsomalla vapausasteiden mukaisia rivejä ja vertaamalla testimuuttujaa F siinä olevaan arvoon. Taulukkoon on taulukoitu 5 % merkitsevyysasteet. Vertailun perusteella merkityksellisyys voidaan todeta.

Lasketaan jokaisen havainnon absoluuttinen etäisyys ryhmän keskiarvosta:

$$d_{ij} = |x_{ij} - \bar{x}_{i\cdot}|, \text{ kun } i = 1, ..., k \text{ ja } j = 1, ..., n_i,$$

missä x_{ij} on ryhmän i havainto j, $\bar{x}_i$ on ryhmän i keskiarvo, k on ryhmien lukumäärä ja n_i on ryhmän i havaintojen lukumäärä. Testimuuttuja on muotoa

$$F = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} n_i(\bar{d}_i - \bar{d})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (d_{ij} - \bar{d}_{i\cdot})^2},$$

kun ainakin yhden ryhmän havainnon etäisyys ryhmän keskiarvosta on erisuurta kuin nolla. Yhtälössä N on kaikkien ryhmien havaintojen yhteislukumäärä, k on ryhmien lukumäärä, n_i on ryhmän i havaintojen lukumäärä, d_{ij} on ryhmän i havainnon j etäisyys ryhmän i keskiarvosta, $\bar{d}_i$ on ryhmän i keskihajontojen keskiarvo ja $\bar{d}$ on kaikkien ryhmien keskihajontojen keskiarvo. [18]

Levenen testiä hyödynnetään tässä tutkielmassa t-testin laskutavan määrittämiseen, koska laskutapa riippuu siitä, ovatko varianssit yhtä suuria vai eivät.

2.3 χ^2-testi

Kahden kategorisen muuttujien välisten riippuvuuksien tutkimiseen käytetään χ^2-riippumattomuustestiä. Testin nollahypoteesina on oletus, että muuttujat ovat riippumattomia. Myös tässä tapauksessa nollahypoteesi jää voi-

maan, jos testin p-arvo on yli 0.05. Muussa tapauksessa nollahypoteesi kumotaan ja todetaan, että kahden tutkittavan muuttujan välillä vallitsee riippuvuus. [16]

Tässäkin testissä saatua tulosta verrataan vapausasteen avulla taulukoituihin arvoihin, joten vapausasteen selvittäminen on tärkeää. Testin vapausaste muodostuu muuttujien vaihtoehtojen lukumäärästä kaavalla $(r-1)(s-1)$, missä r ja s ovat muuttujien vaihtoehtojen lukumäärät. [16] Esimerkiksi, jos tekijänä on sukupuoli, missä on kaksi vaihtoehtoa: poika ja tyttö, ja toisena tekijänä on, onko lasta imetetty, jossa on kolme vaihtoehtoa: imetetty ainakin 4 kk, imetetty, mutta alle 4 kk ja ei imetetty. Näin ollen vapausaste on (2-1)(3-1) = 2. Luku s kuvaa sarakkeita ja r rivejä taulukossa, joka muodostuu, kun taulukoidaan muuttujien saamat arvot seuraavasti

r	s	1 Imetetty $\geq$ 4 kk	2 Imetetty $<$ 4 kk	3 Ei imetetty	Yhteensä
1	Poika	30	5	50	85
2	Tyttö	15	7	43	65
	Yhteensä	45	12	93	150

Taulukko 1: Esimerkkitaulukko havainnollistamaan vapausasteen muodostumista.

Testimuuttuja lasketaan kaavalla [16, 20]

$$\chi^2 = \sum_{ij} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

missä o on havaittu arvo ja e on arvioitu tai odotettu arvo. Saatua arvoa verrataan liitteen B taulukon 19 taulukoituihin arvoihin vapausasteen mukaan, samalla tavalla kuin t-jakauman kohdalla, mistä saadaan p-arvo. Samoin kuin t-jakauman, myös χ^2-jakauman kohdalla R-studio antaa p-arvon valmiiksi eikä taulukoita tarvitse tutkia. Tämän perusteella tehdään päätös, voidaanko nollahypoteesi hylätä vai ei.

Tähän testiin liittyy kuitenkin rajoituksia, jotka määrittävät milloin testiä voidaan tai ei voida käyttää. Testin χ^2-arvio on pätevä silloin, kun melkein kaikki e_{ij} arvot ovat yli viiden arvoisia ja mikään niistä ei ole alle kolmen. [20] Esimerkiksi seuraavien taulukoiden 2A ja 2B arvot ovat sellaiset, etteivät ne sovi χ^2 tarkasteluun. Taulukossa 2A melkein kaikki arvot ovat alle viiden ja taulukossa 2B on arvoja, jotka ovat alle kolmen.

Taulukko A		
4	4	5
4	4	3
5	3	3

Taulukko B		
6	10	8
15	4	2
3	1	7

Taulukko 2: Esimerkkitaulukot havainnollistavat aineistoja, missä χ^2 arvio ei ole pätevä. Taulukko A havainnollistaa aineistoa, jossa melkein kaikki e_{ij} arvot ovat alle viiden. Taulukko B havainnollistaa aineisto, jossa osa e_{ij} arvoista on alle kolmen.

Riippuvuuksien vahvuuteen χ^2-menetelmä ei kuitenkaan ota kantaa. Se kertoo ainoastaan, onko riippuvuutta ylipäätään havaittavissa. Kuitenkin tietyt riippuvuudet voivat olla niin heikkoja, etteivät ne haittaa analyysiä. [20] Riippuvuuksien vahvuutta voidaan tutkia esimerkiksi Cramerin V -testillä ja tätä testiä on käytetty soveltavassa osiossa.

2.4 Cramerin V

Cramerin V -testi kertoo kahden kategorisen muuttujan välisestä suhteesta. Tätä käytetään, kun halutaan tietää kahden muuttujan välinen suhde, kiinnostavat muuttujat ovat kategorisia ja muuttujien luokissa on kaksi tai useampi arvo. Testimuuttuja lasketaan kaavalla

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, s-1)}},$$

missä χ^2 on χ^2-riippumattomuustestistä saatu arvo, r on rivien lukumäärä, s on sarakkeiden lukumäärä ja n on havaintojen lukumäärä. [21]

Testimuuttuja V saa arvoja välillä [0,1]. Arvo nolla kertoo, ettei muuttujien välillä ole lainkaan yhteyttä. Arvo yksi kertoo, että muuttujien välillä vallitsee täydellinen yhteys. Pienet arvot kertovat näin ollen, että muuttujien välillä on heikko yhteys ja suuret arvot kertovat vahvasta yhteydestä. [22, 23]

2.5 Hosmerin ja Lemeshowin testi

Hosmerin ja Lemeshowin testi kertoo, kuinka hyvin muodostettu malli kuvaa aineistoa eli se on niin sanottu hyvyystesti (Goodness of Fit Test). Nollahypoteesi on, että malli kuvaa aineistoa hyvin. Tämä osoitetaan sillä, ettei löydy todisteita siitä, ettei malli istu aineistoon. [24, 26] Hosmerin ja Lemeshowin testi valittiin tutkielmaan testaamaan mallin hyvyyttä, koska tutkittavana on jatkuva muuttuja (äidin ikä), minkä takia vaihtoehtoisia yhdistelmiä on

lähes tai jopa yhtä monta kuin havaintoja. Yhdistelmillä tarkoitetaan jokaisen muuttujan saamien arvojen kombinaatioita, jotka ovat esiintyneet havainnoissa [24].

Data jaetaan samankokoisiin ryhmiin riskin eli tapahtuman todennäköisyyden perusteella. Yleensä data jaetaan kymmeneen ryhmään, mutta ryhmien lukumäärä voi olla jokin muukin, jos siihen nähdään tarve. Ennustettua arvoa verrataan havaittuun arvoon ja tutkitaan, kuinka tarkkoja tuloksia malli antaa. Testillä verrataan ryhmien onnistumista sen sijaan, että suoritettaisiin vertailu yksittäisille arvoille. [24] Testimuuttuja on muotoa [25, 26]

$$C_g = \sum_{k=1}^{g} \sum_{j=0}^{c-1} \frac{(o_{kj} - e_{kj})^2}{e_{kj}} = \chi^2_{df},$$

missä χ^2_{df} on khiin neliö vapausasteella df, g on luokkien lukumäärä, c on ryhmien lukumäärä, o_{kj} on havaittu arvo luokassa k ryhmässä j, e_{kj} on ennustettu arvo luokassa k ryhmässä j [25, 26]. Vapausaste määräytyy $(g-2)\cdot(c-1)$ lausekkeen mukaan [25, 26].

P-arvo saadaan luettua χ^2 taulukoiduista arvoista. Jos arvo on suurempi kuin 0.05, nollahypoteesi jää voimaan ja malli kuvaa aineistoa hyvin. Muussa tapauksessa nollahypoteesi hylätään ja todetaan, ettei malli kuvaa aineistoa kovinkaan hyvin. Taulukko 3 selventää, mistä lausekkeen arvot tulevat. Taulukossa $c = 2$ ja $g = 10$.

		c			
	Luokka	Ryhmä 1		Ryhmä 2	
		Havaittu	Ennustettu	Havaittu	Ennustettu
g	1	o_{11}	e_{11}	o_{12}	e_{12}
	2	o_{21}	e_{21}	o_{22}	e_{22}
	3	o_{31}	e_{31}	o_{32}	e_{32}
	4	o_{41}	e_{41}	o_{42}	e_{42}
	5	o_{51}	e_{51}	o_{52}	e_{52}
	6	o_{61}	e_{61}	o_{62}	e_{62}
	7	o_{71}	e_{71}	o_{72}	e_{72}
	8	o_{81}	e_{81}	o_{82}	e_{82}
	9	o_{91}	e_{91}	o_{92}	e_{92}
	10	o_{101}	e_{101}	o_{102}	e_{102}

Taulukko 3: Havainnollistetaan mitä Hosmerin ja Lemeshowin testin arvot kuvaavat.

Taulukko 4 havainnollistaa millaiselta taulukko voi näyttää, jos aineistona on 500 havainnon aineisto.

Luokka	Riskin suuruus	Yht.	Ryhmä 1		Ryhmä 2	
			Havaittu	Ennustettu	Havaittu	Ennustettu
1	[0.085 - 0.120]	50	3	3.3	47	46.7
2	(0.120 - 0.240]	50	2	2.6	48	47.4
3	(0.240 - 0.280]	50	1	0.6	49	49.4
4	(0.280 - 0.336]	50	4	3.9	46	46.1
5	(0.336 - 0.399]	51	6	6.7	45	44.3
6	(0.399 - 0.450]	50	6	6.2	44	43.8
7	(0.450 - 0.660]	50	5	4.5	45	45.5
8	(0.660 - 0.691]	50	10	9.1	40	40.9
9	(0.691 - 0.808]	49	13	13.4	36	35.6
10	(0.808 - 0.878]	50	15	16.2	35	33.8

Taulukko 4: Esimerkki, aineiston havaintojen ja mallin ennustamien tulosten jakautumisesta kymmeneen ryhmään Hosmerin ja Lemeshowin testin mukaan.

Taulukosta 4 nähdään, että aineisto on jaettu kymmeneen osaan riskin suuruuden mukaan. Taulukosta näkyy jokaisen osan riskin vaihteluväli. Osat on saatu jakamalla ne siten, että jokaiseen osaan tule noin 50 havaintoa. Havaitut arvot kertovat, kuinka monta aineistosta tehtyä havaintoa kuuluu ryhmään yksi ja kuinka monta havaintoa aineistossa kuuluu ryhmään kaksi. Ennustetut arvot kertovat taas, kuinka monta kappaletta malli on ennustanut olevan ryhmässä yksi tai kaksi. Nämä ovat niitä arvoja, joita Hosmerin ja Lemeshowin testissä tutkitaan. [26] Tämän esimerkin vapausaste on $(10 - 2) \cdot (2 - 1) = 8$. P-arvo saadaan χ^2 taulukoiduista arvoista vapausasteella 8. P-arvon mukaan päätetään, hylätäänkö vai pidetäänkö nollahypoteesi [26].

3 Logistinen regressio

Regressioanalyysi tutkii selittävien muuttujien vaikutusta selitettävään muuttujaan. Selittäviä muuttujia voi olla yksi tai useampia. [27] Analyysin avulla voidaan vastata esimerkiksi kysymykseen, onko synnytystavalla, äidin koulutustasolla tai äidin iällä vaikutusta lapsen imetykseen ja sen kestoon. Mallista nähdään, mitkä selittävät muuttujat vaikuttavat tapahtuman todennäköisyyteen ja kuinka suurta vaikutus on [27]. Regressioanalyysissä voidaan tutkia kaikkien selittävien muuttujien vaikutusta samalla kertaa [27]. Tämä onkin regressioanalyysin suurin etu. Regressioanalyysissä muodostetaan regressiomalli. Malli on yleensä muotoa

$$E(y|\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\beta}),$$

missä y on selitettävä muuttuja, f on mallin muodon ilmaiseva funktio, $\mathbf{x}$ on selittävien muuttujien vektori ja $\boldsymbol{\beta}$ on kerroinvektori [27].

Tavallinen lineaarinen regressioanalyysin malli on tällöin muotoa $y = \alpha+\beta x$. Tämä malli kuvaa kuitenkin hyvin vain lineaariselle suoralla asettuvaa aineistoa eikä siksi sovellu kaikkien aineistojen kuvaamiseen. Jos selitettävä tekijä on kategorinen ja kaksiluokkainen, sen arvot sijoittuvat alueelle [0,1] ja tälläkin alueella havainnot ovat joko kohdassa 0 tai 1. Lineaarinen suora ei rajoitu alueelle [0,1], vaan se kattaa paljon laajemman alan kuin, jolla kaikki havaintoarvot voivat sijaita. Tällöin se ei voi kuvata hyvin aineiston käyttäytymistä. Kun aineisto pysyy alueella [0,1], myös mallin tulee pysyä tällä alueella. [27]

Odotusarvon ja selittävien muuttujien vektorin $\mathbf{x}$ välistä yhteyttä kuvaavia funktioita on useita, mutta tässä tutkielmassa on valittu logistinen regressio sen ominaisuuksien takia. Logistisella funktiolla on nimittäin monia hyviä matemaattisia ominaisuuksia mallin muotoilussa ja estimoinnissa, kuten sen joustavuus ja helppokäyttöisyys [26, 27]. Näin ollen se sopii moniin eri käyttötarkoituksiin [27].

Koska lineaarinen yhtälö on osa logistista regressiota, käydään se seuraavaksi läpi. Lineaarikombinaatio eli lineaariyhdistelmä on joukko termejä, jotka kerrotaan omalla kertoimella ja lasketaan yhteen. Sen avulla voidaan muodostaa lineaarinen yhtälö

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k,$$

missä y on vastemuuttujan arvo, α on vakioarvo, β_i on regressiokerroin ja x_i on havainnon arvo, kun $i = 1, ..., k$. Tällöin havainnolle t yhtälö on seuraa-

vanlainen

$$\begin{aligned} y_t &= \alpha_t + \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} \\ &= \alpha_t + \sum_{i=1}^{k} \beta_i x_{ti} \end{aligned}$$

y_t on selitettävän muuttujan y satunnainen ja havaittu arvo havainnossa t. x_{ti} on selittävän muuttujan x_i , $i = 1, 2, \ldots, k$, kiinteä (ei-satunnainen) ja havaittu arvo havainnossa t, α on vakioselittäjän regressiokerroin, kiinteä (ei-satunnainen) ja tuntematon vakio, β_i on selittäjän x_i , $i = 1, 2, \ldots, k$, regressiokerroin, kiinteä (ei-satunnainen) ja tuntematon vakio. [27, 28]

Logistisessa regressiossa tarvitaan linkkifunktioita, koska lineaarinen malli ei toimi näihin suoraan. Linkkifunktio yhdistää selittävät muuttujat vastemuuttujan y_i odotusarvoon μ [29].

Määritelmä 2. [30] *Yleistetty lineaarinen malli* on muotoa

$$g(\mu) = \beta_0 + \sum_{j=1}^{n} x_j \beta_j,$$

missä funktio g on jokin linkkifunktio, μ on odotusarvo, β_0 on vakio, x_j ovat havaintojen arvot ja β_j ovat regressiokertoimet.

Yleistetyn lineaarisen mallin rakenneosa on lineaarinen ja se on suoraan verrannollinen vastemuuttujan joukon Y odotusarvoon μ jonkin linkkifunktion kautta. Erikoistapaus yleistetystä lineaarisesta mallista on lineaarinen malli, jolloin linkkifunktiota ei tarvita, vaan mallin rakenneosa on jo valmiiksi suoraan verrannollinen odotusarvoon μ. [30]

Logistista regressiota käytetään, kun halutaan ennustaa, millä todennäköisyydellä tarkasteltava asia tapahtuu. Tällöin odotusarvona μ on $\frac{p}{1-p}$ ja kaava on muotoa

$$\log\left(\frac{p}{1-p}\right) = \alpha + \sum_{i=1}^{k} x_i \beta_i, \tag{2}$$

missä p on tapahtuman todennäköisyys, k on selittävien tekijöiden lukumäärä, x_i on havainto ja β_i on regressiokerroin. [27] Vastetapahtuman todennäköisyydelle saadaan johdettua seuraava kaava [31]

$$\begin{aligned} p &= \frac{e^{\alpha+\beta_1 x_1+...+\beta_k x_k}}{1 + e^{\alpha+\beta_1 x_1+...+\beta_k x_k}} \\ &= \frac{1}{1 + e^{-(\alpha+\beta_1 x_1+...+\beta_k x_k)}}. \end{aligned} \tag{3}$$

Logistisessa regressioanalyysissä selitettävän ja selittävien muuttujien suhde ei ole lineaarista vaan muodoltaan enemmänkin niin sanotun s-käyrän mallista. Tätä s-käyrää kutsutaankin logistiseksi käyräksi. Tällöin pienen ja suuren selittävän arvon pieni vaihtelu ei vaikuta selitettävän muuttujan arvoon suuresti vaan todennäköisyys pysyy lähes samana. Jos selittävän muuttujan arvo on keskivaiheilla, tällöin pienellä arvon muutoksella on suuri vaikutus selitettävään muuttujaan ja sen todennäköisyyteen. [32]

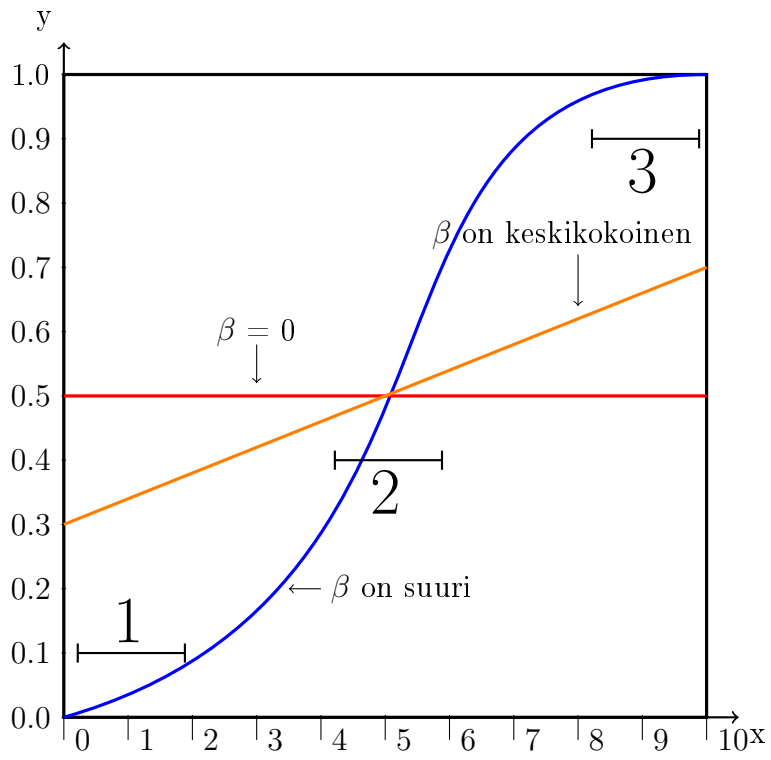
Kuvassa 1 on havainnollistettu yhtälön (2) regressiokertoimen β muutoksia, kun selittäviä tekijöitä on vain yksi eli k on yksi. Tällöin malli on muotoa

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta x. \tag{4}$$

Kun selittävällä muuttujalla ei ole vaikutusta selitettävään muuttujaan, regressiokerroin saa pieniä arvoja. Käyrä on täysin vaakasuora, kun kertoimen arvo on nolla, kuten kuvasta 1 nähdään. Tämä tarkoittaa, että selitettävän muuttujan mittaaman tapahtuman todennäköisyys ei ole riippuvainen tästä selittävästä muuttujasta. [31, 32]

Kun regressiokerroin saa suuria arvoja, käyrästä muodostuu s-käyrä. Tällöin selittävän muuttujan arvon pieni muutos vaihteluvälin ääripäissä ei vaikuta suuresti selitettävän arvon mittaamaan tapahtuman todennäköisyyteen. Kuvassa näitä ovat alueet 1 ja 3. Selittävän muuttujan arvon pieni muutos sen ollessa vaihteluvälin keskivaiheilla aiheuttaa taas suuren muutoksen selitettävän muuttujan kuvaamaan todennäköisyyteen. Kuvassa tämä väli on merkitty väliksi 2. [32]

Regressiokertoimen ollessa keskikokoinen, käyrän muoto on vaakasuoran ja s-käyrän välimuotoa. Kuvassa 1 on vain positiivisen kertoimen mukaan piirretyt kuvaajat. Näiden lisäksi kerroin voi olla myös negatiivinen, jolloin kertoimen vaikutus selitettävän muuttujan kuvaamaan ilmiön todennäköisyyteen on laskeva. Käyrät ovat tällöin samanmuotoiset kuin kuvassa 1, mutta laskevat vasemmalta oikealle. [32]



Kuva 1: Kuva havainnollistaa regressiokertoimen muutoksen vaikutusta käyrän muotoon [32].

Logistisen ja lineaarisen regression suurin ero on, että logistisessa regressiossa selitettävä muuttuja on kategorinen eli diskreetti, kun taas lineaarisessa regressiossa selitettävän muuttujan jakauma on jatkuva. Näin siis logistiset mallit eivät aina ole lineaarisia, minkä takia käytetäänkin linkkifunktioita. Linkkifunktiot muuntavat vastemuuttujien odotusarvot selittävien muuttujien lineaarikombinaatioiksi eli näin ne saadaan lineaariseen muotoon, jolloin niitä on helpompi tutkia. [30, 33] Kuva 2 havainnollistaa tätä muutosta lineaariseen muotoon. Logistisessa regressiossa linkkifunktiona käytetään logit-funktiota

$$\operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right), \qquad \text{kun } p \in (0,1)$$

ja mallissa käytetään selitettävän muuttujan luokkien todennäköisyyksien luonnollista logaritmia sen sijaan, että käytettäisiin suoraan selitettävän muuttujan luokkia [26, 30, 31]. Kun merkitään, että p on yhtäsuuri kuin selitettävän muuttujan tapahtuman todennäköisyys, logistinen malli on kaavassa (2) esitettyä muotoa, missä luonnollisesta logaritmista käytetään merkintää log [30]. Jatkossa käytetään tätä samaa merkintätapaa.

Kuva 2: Miten s-käyrästä muodostuu lineaarinen logitin avulla [34].

Selitettävä muuttuja on kaksiluokkainen, luokittelu- tai järjestysasteikollinen. Joissain tapauksissa logistista regressiota on käytetty jatkuvan selitettävän muuttujan tapauksessa, mutta tällöin selitettävä muuttuja on luokiteltu. [26,31]

Selittävät muuttujat voivat olla kategorisia tai jatkuvia. Kategoriset selittävät muuttujat voivat olla joko kaksi- tai moniluokkaisia. [26,31] Selitettävän kategorisen muuttujan arvojen ei tarvitse olla missään arvosuhteessa toisiinsa nähden, vaan ne voivat olla esimerkiksi värejä, joita ei pystytä laittamaan mihinkään arvojärjestykseen.

Logistisen regression yksinkertaisin muoto on binäärinen logistinen regressio, jossa selitettävä muuttuja on binäärinen. Multinomiaalinen logistinen regressio on logistisen regression yleistys ja sitä käytetään, kun selitettävä muuttuja on moniluokkainen sekä laatuasteikollinen. On olemassa myös muunlaisia logistisen regression yleistyksiä, kuten kumulatiivinen logistinen regressio, jota käytetään moniluokkaiselle järjestysasteikolliselle selitettävälle muuttujalle. [27] Tässä tutkielmassa käydään kuitenkin läpi vain binäärinen ja multinomiaalinen logistinen regressio, sillä ne ovat oleellisia tutkielman

soveltavan osion kannalta.

Logistisen regressioanalyysin mallin tulkinnassa tulee huomioida, ettei ryhmien eroa kerrota suoraan mallin parametrien estimaattia β käyttäen, vaan tulokset esitetään yleensä kerroinsuhdelukuina eli ristitulosuhteina ja niiden luottamusväleinä. [27, 30]

3.1 Kerroinsuhde

Logistisessa regressiossa suhdeluku on olennaisessa asemassa. Suhdeluku tarkoittaa tapahtuman todennäköisyyden suhteuttamista siihen todennäköisyyteen, ettei tapahtuma tapahdu. Suhdeluku on suhde, jossa halutun tapahtuman todennäköisyys jaetaan ei halutun tapahtuman todennäköisyydellä eli

$$\text{Suhdeluku} = \frac{\text{haluttu tapatuma}}{\text{ei haluttu tapahtuma}}.$$

Merkitään vastetapahtuman todennäköisyyttä arvolla p. Tällöin

$$\text{Suhde} = \frac{p}{1-p},$$

missä $1 - p$ kuvaa sitä todennäköisyyttä, että vastetapahtuma ei tapahdu. [27, 30, 35]

Suhdelukujen suhde eli kerroinsuhdeluku (OR-luku) kuvaa vastetapahtuman ja ei-vastetapahtuman suhdetta ryhmien välillä eli sitä, kuinka moninkertainen havaintojen lukumäärä on vastetapahtumalla suhteessa ei-vastetapahtumaan ryhmässä 1 verrattuna ryhmään 2. Vastetapahtumalla tarkoitetaan selitettävän tekijän luokkaa 1 ja ei-vastetapahtumalla tarkoitetaan selitettävän tekijän luokkaa 0. Todennäköisyyksien avulla ilmaistuna OR-luku kuvaa, kuinka moninkertainen vasteluokan 1 todennäköisyys on vasteluokan 0 todennäköisyyteen ryhmässä 1 verrattuna ryhmään 2. Jos taas halutaan tutkia ryhmää 2 verrattuna ryhmään 1, OR-luvusta lasketaan käänteisluku [26, 31]. OR-luvusta voidaan käyttää myös nimitystä ristitulosuhde ja ryhmien kerroinsuhde. Suhdeluku lasketaan seuraavalla tavalla.

	Tapahtuma A	
Tapahtuma B (Vastetapahtuma)	Ryhmä 1 Tapahtui	Ryhmä 2 Ei tapahtunut
Tapahtui (= 1)	n_{11}	n_{12}
Ei tapahtunut (= 0)	n_{21}	n_{22}

Taulukko 5: Taulukko havainnollistaa aineiston jakautumista tapahtumien välille.

Lasketaan ensin tapahtumien todennäköisyydet eli suhdeluvut, kun tapahtuma A on ja ei ole tapahtunut.

$$\text{suhde1} = \frac{n_{11}}{n_{21}} \text{ ja suhde2} = \frac{n_{12}}{n_{22}}.$$

Suhde1 on tapahtuman B todennäköisyys tapahtua, kun tapahtuma A on tapahtunut. Suhde2 on tapahtuman B todennäköisyys tapahtua, kun tapahtuma A ei ole tapahtunut. Näiden suhdelukujen avulla saadaan laskettua OR-luku suhteena $\frac{\text{suhde1}}{\text{suhde2}}$. Suhdeluku kertoo, onko tapahtumalla A vaikutusta tapahtuman B tapahtumisen todennäköisyyteen ja millainen vaikutus on. [36]

Jos suhde on 1, niin tapahtumalla A ei ole vaikutusta tapahtumaan B. Jos suhde on yli 1, niin tapahtuma A vaikuttaa tapahtumaan B kasvavasti. Jos suhde on pienempi kuin yksi, tapahtuma A vaikuttaa tapahtumaan B laskevasti. Suhdeluku kertoo myös kuinka paljon tapahtuma A vaikuttaa tapahtuman B todennäköisyyteen.

Jos halutaan tutkia tapahtuman B tapahtumatta jäämistä eli kiinnostus on ryhmässä "Ei tapahtunut", voidaan suhdeluvusta ottaa käänteisluku. Esimerkiksi jos "tapahtui"vs "ei tapahtunut"OR $= 2$, niin "ei tapahtunut"vs "tapahtui"OR $= 1/2 = 0.5$. [36]

Suhdelukuja käytetään, koska siten saadut arvot vaihtelevat välillä

$$\frac{p}{1-p} \in (0, \infty).$$

Tällöin se vastaa paremmin lineaarista yhtälöä. Pelkän todennäköisyyden mallintaminen lineaariseksi yhtälöksi tuottaa ongelman, koska todennäköisyys sijoittuu aina välille $p \in (0, 1)$. Koska lineaarinen yhtälö voi saada myös negatiivisia ratkaisuja, suhdeluku ei vastaa vielä tarpeeksi hyvin lineaarista yhtälöä. On siis saatava sellainen muoto, joka saa arvoja välillä $(-\infty, \infty)$. [27] Tämän takia käytetään luonnollista logaritmista muunnosta [35]

$$\log\left(\frac{p}{1-p}\right) \in (-\infty, \infty).$$

Tämä johtaakin jo kaavassa (2) esitettyyn yhtälöön [26, 31]. Kaavan (3) perusteella todennäköisyydet ja suhdeluku ovat [31, 33, 37]

$$\begin{aligned} p &= \frac{e^{\alpha+\sum_{i=1}^{k}\beta_i x_i}}{1+e^{\alpha+\sum_{i=1}^{k}\beta_i x_i}} \\ 1-p &= \frac{1}{1+e^{\alpha+\sum_{i=1}^{k}\beta_i x_i}} \\ \frac{p}{1-p} &= e^{\alpha+\sum_{i=1}^{k}\beta_i x_i}. \end{aligned}$$

Seuraavaksi tutkitaan yhden tekijän muutosta kerrallaan ja asetetaan muut selittävät tekijät vakioiksi. Tällöin saadaan OR-luvuksi

$$\begin{aligned}
&\frac{P(y=1|x_1)/P(y=0|x_1)}{P(y=1|x_2)/P(y=0|x_2)} \\
&= \frac{e^{\alpha+\beta_1 x_1+\text{vakio}}}{e^{\alpha+\beta_2 x_2+\text{vakio}}} \\
&= e^{\alpha+\beta_1 x_1+\text{vakio}-(\alpha+\beta_2 x_2+\text{vakio})} \\
&= e^{\beta_1 x_1-\beta_2 x_2} && ||\beta_1=\beta_2=\beta_i \\
&= e^{\beta_i(x_1-x_2)} && ||x_1=1 \text{ ja } x_2=0 \\
&= e^{\beta_i},
\end{aligned}$$

missä β_i on selittävää tekijää x_i vastaava kerroin ja y on vastemuuttujan arvo. OR-luku on siis e^{β_i}. [26, 37, 38]

Tämän perusteella parametrin β avulla voidaan tutkia, onko ryhmien erolla merkitsevyyttä. Tämä saadaan testattua sillä, saako β arvon nolla vai ei. Jos β on nolla, niin OR-luku on yksi ja tämä tarkoittaa ettei selittävällä muuttujalla ole merkitsevää yhteyttä vastemuuttujaan. OR-luvun luottamusväli saadaan laskettua eksponenttimuunnoksen avulla parametrin β luottamusvälistä. [26, 31] Jos arvo 1.0 kuuluu OR-luvun luottamusvälille, ennuste ei ole merkitsevä eli selittävän muuttujan yhteys vastemuuttujaan ei ole merkitsevä [33].

3.2 Binäärinen logistinen regressioanalyysi

Binäärisessä logistisessa regressiossa selitettävä muuttuja on kaksiarvoinen eli binäärinen [26, 27, 31]. Binäärisellä logistisella regressiolla voidaan tutkia esimerkiksi, onko joidenkin ryhmien välillä eroa vastemuuttujan jakaumassa tai onko vastemuuttujan ja selittävän muuttujan välillä riippuvuutta [26, 31]. Esimerkiksi voidaan tutkia, onko lasta imetetty. Lasta joko on tai ei ole imetetty, joten tutkittava muuttuja on binäärinen. Tässä tutkielmassa halutaan selvittää, mitkä tekijät vaikuttavat imettämiseen. Näitä tekijöitä voivat olla esimerkiksi synnytystapa, äidin koulutuksen taso ja äidin ikä.

Yleensä selitettävän muuttujan arvot muutetaan arvoiksi 1 ja 0. Esimerkiksi, jos selitettävä muuttuja voi saada arvot ”lasta on imetetty” tai ”lasta ei ole imetetty”. Jos olemme kiinnostuneita lapsista, joita on imetetty, niin ”lasta on imetetty” muutetaan arvoksi 1 ja ”lasta ei ole imetetty” arvoksi 0. Kiinnostavan tapahtuman tapahtuminen saa siis arvon yksi ja arvo nolla tulee, kun kiinnostava tapahtuma ei tapahdu [27, 32].

Logistiseen regression malliin voidaan ottaa joko yksi tai useampia selittäviä tekijöitä. Ensin käydään läpi tapaukset, joissa on vain yksi selittävä muuttuja. Sen jälkeen laajennetaan tapaus useampaan selittävään tekijään.

3.2.1 Yksi selittävä muuttuja

Selittäviä tekijöitä voi olla kolmea erilaista. Tekijä voi olla joko kategorinen tai jatkuva. Kategorinen tekijä voidaan jakaa vielä kaksiluokkaisen ja useampiluokkaisen muuttujan tapauksiin. Lähdetään liikkeelle kaksiluokkaisesta selittävästä tekijästä, jonka jälkeen laajennetaan tapaus useampiluokkaiseksi. Lopuksi käydään läpi jatkuvan selittävän tekijän tapauksen.

Kaksiluokkainen selittävä muuttuja tarkoittaa sitä, että selittävällä muuttujalla on vain kaksi luokkaa. Esimerkiksi lapsen synnytystapa. Lapsi on syntynyt joko alateitse tai sektiolla. Yhden kaksiluokkaisen selittävän muuttujan tapauksessa logistisesta mallista otetaan luonnollinen logaritmi tapahtuman todennäköisyyksien suhteen eli kaavan (4) tapaan, missä p on tapahtuman todennäköisyys, α on vakiotekijä, β on regressiokerroin ja x on selittävän muuttujan arvo. Tämä on niin sanotusti vastetapahtuman logit(p)-muunnos. [30] Kun merkitään regressiokerrointa symboolilla β_1 ja selittävää muuttujaa symboolilla x_1, saadaan vastetapahtuman todennäköisyydeksi seuraava

$$\begin{aligned} \frac{p}{1-p} &= e^{\alpha+\beta_1 x_1} \\ p &= (1-p)e^{\alpha+\beta_1 x_1} \\ (1+e^{\alpha+\beta_1 x_1})p &= e^{\alpha+\beta_1 x_1} \\ p &= \frac{e^{\alpha+\beta_1 x_1}}{1+e^{\alpha+\beta_1 x_1}}. \end{aligned}$$

Tästä seuraa, että [31]

$$p = \frac{1}{1+e^{-(\alpha+\beta_1 x_1)}}.$$

Yksinkertaisimmillaan logistinen regressiomalli on näin ollen vain regressiomalli, jossa selitettävä muuttuja on suhdeluvun logaritmi [32]. Esimerkissä malliin voisi ottaa synnytystavan, koska synnytystapa on kaksiluokkainen muuttuja, jonka arvot ovat alateitse ja sektiolla syntyneet. Malli voisi olla siis esimerkiksi seuraavanlainen

$$\text{logit}(p) = \log\left(\frac{p(\text{imetetty})}{p(\text{ei imetetty})}\right) = \alpha + \beta \cdot \text{sektio},$$

missä sektio tarkoittaa, että referenssiluokkana on alateitse synnytys ja sektiolla synnytys vaikuttaa kertoimen β verran. OR-luku saadaan laskettua taulukosta 6.

Selittäjä	Vastetapahtuma	
	ON (=1)	EI (=0)
Ryhmä 1	n_{11}	n_{12}
Ryhmä 2	n_{21}	n_{22}

Taulukko 6: Kaksiluokkaisen vasteen havaintojen jakautuminen kaksiluokkaisen selittävän muuttujan ryhmien välille. Tämän taulukon avulla saadaan muodostettua OR-luku.

Taulukosta 6 saadaan, että [26]

$$\text{OR-luku} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}.$$

Tällöin yhtälön (3) perusteella OR-luku on e^{β_1} [26].

Moniluokkaisen selittävän muuttujan tapauksessa tutkitaan, onko kolmen tai useamman ryhmän välillä eroa selitettävän muuttujan jakaumaan. Malli toimii muuten samoin kuin kaksiluokkaisen selittävän muuttujan tapauksessa lukuunottamatta sitä, että selittävällä muuttujalla on enemmän kuin kaksi luokkaa ja täten siinä estimoidaan useampia parametreja. Esimerkkitapauksessa tällainen selittävä muuttuja on äidin koulutuksen taso. Äidin korkein koulutus voi olla peruskoulu, toisen asteen koulutus tai korkeakoulu. Tällöin luokkien vertailussa yksi luokka on vertailuluokka, johon muita luokkia verrataan. Tätä luokkaa kutsutaan referenssiluokaksi [27]. Parametreja β muodostuu yksi vähemmän kuin mitä luokkia on. [30] Esimerkiksi neljän ryhmän tapauksessa parametreja on kolme. Kun luokka yksi valitaan referenssiluokaksi, parametrit jaottuvat seuraavasti: β_1 vastaa luokkaa 2 verrattuna luokkaan 1, β_2 vastaa luokkaa 3 verrattuna luokkaan 1 ja β_3 vastaa luokkaa 4 verrattuna luokkaan 1. Moniluokkaisen selittävän muuttujan logistinen muunnos on

$$\text{logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_h x_h,$$

missä $h =$ selittävän muuttujan luokkien lukumäärä miinus yksi [30]. Vastetapahtuman todennäköisyydeksi saadaan seuraava

$$p = \frac{1}{1 + e^{-(\alpha + \sum_{i=1}^{h} \beta_i x_i)}}.$$

Myös OR-lukuja muodostuu saman verran kuin mitä parametreja β.

Selittäjä	Vastetapahtuma	
	ON (=1)	EI (=0)
Ryhmä 1	n_{11}	n_{12}
Ryhmä 2	n_{21}	n_{22}
Ryhmä 3	n_{31}	n_{32}

Taulukko 7: Kaksiluokkaisen vasteen havaintojen jakautuminen kaksiluokkaisen selittävän muuttujan ryhmien välille. Tämän taulukon avulla saadaan muodostettua OR-luku. [26]

Taulukosta 7 saadaan OR-luvut vastaavasti kuten aikaisemmin taulukosta 6 saatiin OR-luku. Näin ollen ryhmää kaksi koskeva

$$\text{OR-luku} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}}$$

ja ryhmän kolme vastaavasti

$$\text{OR-luku} = \frac{n_{11}/n_{12}}{n_{31}/n_{32}} = \frac{n_{11} \cdot n_{32}}{n_{12} \cdot n_{31}}.$$

Esimerkkitapauksessa malli on seuraavanlainen, jos peruskoulua käytetään referenssiluokkana

$$\text{logit}(p) = \alpha + \beta_1 \cdot \text{toisen asteen koulutus} + \beta_2 \cdot \text{korkeakoulu}.$$

Näin ollen kolmiluokkaisen selittävän tekijän mallissa parametreja β muodostuu kaksi ja OR-lukuja muodostuu tällöin myöskin kaksi.

Numeerisen jatkuvan selittävän muuttujan tapauksessa halutaan selvittää, onko muuttujalla yhteyttä kaksiluokkaiseen selitettävään muuttujaan. Tässä tapauksessa ei voida hyödyntää ristiintaulukointia, koska tästä tulisi laajuudeltaan liian suuri ja nollafrekvenssejä olisi paljon. Jatkuvia numeerisia muuttujia joudutaan siten tutkimaan eri tavalla kuin aiempia kategorisia selittäviä muuttujia.

Selitettävä muuttuja y saa arvoja 0 ja 1. Tutkitaan, onko numeerisella muuttujalla vaikutusta selitettävän muuttujan todennäköisyyteen, että $y = 1$. Halutaan selvittää, onko vastemuuttuja riippuvainen tästä muuttujasta siten, että numeerisen muuttujan arvon kasvaessa, vastemuuttujan jakauma kasvaa toisen luokan esiintymisen puoleen. Tällöin on kyse monotonisesta trendistä eli siitä, että muuttujan arvot joko pelkästään kasvavat

tai vähenevät numeerisen muuttujan muuttuessa. Tätä voidaan tarkastella seuraavan mallin avulla

$$\text{logit}(p|x^*) = \alpha + \beta x,$$

missä $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$, $p|x^*$ kuvaa todennäköisyyttä, että y saa arvon yksi, silloin kun $x = x^*$. Tästä saadaan seuraava kaava parametrille β

$$\beta = \text{logit}(p|x^* + 1) - \text{logit}(p|x^*).$$

Koska OR-luku on $e^{\beta(x_1 - x_2)}$ ja asetetaan $x_1 = x^* + 1$ ja $x_2 = x^*$, yhden suuruista eroa muuttujan x arvoissa vastaava muuttujan y jakaumaeroa kuvaava OR-luku on $e^{\beta(x^*+1-x^*)} = e^{\beta}$. [31]

Esimerkkitapauksessa äidin ikä on jatkuva muuttuja. Malli muodostuu siten seuraavasti

$$\text{logit}(p) = \alpha + \beta \cdot \text{äidin ikä}.$$

3.2.2 Usea selittävä muuttuja

Usean selittävän muuttujan tapauksessa yhtälö (4) johtaa yhtälön (2) muotoon, missä α on vakiotermi ja $\beta_1, ..., \beta_k$ ovat parametrien estimaatteja eli regressiokertoimia. [26, 31] Arvo k saadaan, kun lasketaan selittävien muuttujien vaihtoehdoista summa seuraavalla tavalla. Jos selittävä muuttuja on kaksiluokkainen tai jatkuva, sitä kuvaa yksi arvo. Jos selittävä muuttuja on useampiluokkainen, sitä kuvaamaan tarvitaan selittävän tekijän luokkien lukumäärä miinus yksi arvoa. Tällöin

$$k = A + B + \sum_{l=1}^{C}(D_l - 1),$$

missä A on kaksiluokkaisten selittävien tekijöiden lkm, B on jatkuvien selittävien tekijöiden lkm, C on useampiluokkaisten selittävien tekijöiden lkm ja D on selittävän tekijän luokkien lkm.

Mallin tulkinta tapahtuu OR-luvun avulla samoin kuin yhden selittäjän malleissa. Mallia ei siis tulkita suoraan kertoimista β_i. Muiden selittävien muuttujien vaikutukset on poistettu eli muut selittävät muuttujat on asetettu vakioiksi. OR-lukujen tulkinta tapahtuu siten samoin kuin yhden selittävän muuttujan mallissa eli OR-luku on e^{β_i} ja näitä lukuja muodostuu yhtä monta kuin kertoimia β.

Esimerkki sisältää kolme eri selittävää tekijää: synnytystapa, äidin koulutuksen taso ja äidin ikä. Näistä synnytystapa on kaksiluokkainen, äidin koulutuksen taso kolmeluokkainen ja äidin ikä jatkuva muuttuja. Tällöin arvo $k = 1 + (3 - 1) + 1 = 4$. Saamme seuraavanlaisen mallin

$$\log(p) = \alpha + \beta_1 \cdot \text{synnytystapa} + \beta_2 \cdot \text{toisen asteen koulutus} \\ + \beta_3 \cdot \text{korkeakoulu} + \beta_4 \cdot \text{äidin ikä}.$$

Malli voidaan muodostaa joko täydellisenä mallina tai selittäviä muuttujia karsimalla. Täydellinen malli koostuu kaikista selittävistä muuttujista, jopa niistä, jotka eivät ole merkitseviä. Täydellistä mallia käytetään soveltavassa osiossa. Malliin lisättiin näin ollen kaikki alussa päätetyt tekijät eikä niitä karsittu pois merkitsevyyden mukaan. [27]

Jos selittäviä tekijöitä halutaan karsia, karsinta suoritetaan tekijän merkitsevyyden mukaan. Karsinta voidaan suorittaa eri valintamenetelmillä. Esimerkkeinä näistä menetelmistä ovat etenevä, takautuva ja askeltava menetelmä. Etenevässä menetelmässä lähdetään liikkeelle lisäämällä malliin selittävistä tekijöistä parhaiten selitettävän tekijän kanssa korreloiva tekijä. Tämän jälkeen tutkitaan, lisääkö joku ei mallissa olevista selittävistä tekijöistä mallin yhteiskorrelaatiota, kun huomioidaan mallissa jo olevat tekijät. Jos malli paranee, korrelaatiota eniten lisäävä tekijä lisätään malliin ja sama toistetaan uudestaan. Jos malli ei enää parane, se on valmis ja loput tekijät jätetään mallista pois. [27]

Takautuvassa karsinnassa lähdetään liikkeelle siitä, että kaikki selittävät tekijät lisätään malliin. Selittäviä tekijöitä lähdetään tiputtamaan pois mallista huonoimmin selittävästä tekijästä lähtien. Tekijöiden poistaminen lopetetaan, kun malli ei enää oleellisesti parane. [27]

Askeltava menetelmä on etenevän ja takautuvan menetelmän yhdistelmä. Askeltavassa menetelmässä edetään aina yksi askel kerrallaan. Kullakin askeleella testataan parantaako jokin mallista puuttuva selittävä tekijä mallia. Malliin lisätään se selittävä tekijä, jonka p-arvo on pienin ja alittaa valitun rajan. Sen jälkeen testataan, onko mallin selittävät tekijät kaikki edelleen merkityksellisiä. Jos löydetään ei-merkitseviä muuttujia, niistä suurimman p-arvon omaava jätetään pois mallista. Askellusta jatketaan, kunnes mikään mallista puuttuva selittävä tekijä ei enää parantaisi mallia. [26, 27]

Selittävien tekijöiden valintaan vaikuttavat myös muuttujien keskinäiset suhteet. Malliin ei voida ottaa selittäjiä, jotka korreloivat voimakkaasti keskenään eli joilla on multikollineaarisuutta. Tämä voisi vaikuttaa mallien parametrien estimointiin vääristävästi. Esimerkiksi multikollineaarisuus voi aiheuttaa regressiokertoimille väärän etumerkin tai poikkeuksellisen suuret keskivirheet. Multikollineaarisuus voidaan todeta tarkastelemalla selittävien

tekijöiden keskinäisiä suhteita ja riippuvuuksia. Toisistaan riippuvista selittäjistä vain toinen voidaan valita malliin, jottei edellä mainittuja ongelmia muodostu. [31] Esimerkissä selittävänä tekijänä on äidin ikä. Tämän kanssa malliin ei voisi valita esimerkiksi äidin syntymävuotta, koska ikä ja syntymävuosi ovat selkeästi toisistaan riippuvia tekijöitä, joten niillä on suuri multikollineaarisuus.

Jos selitettävä muuttuja on useampi kuin kaksiarvoinen, niin binäärinen logistinen regressio ei toimi. Tällöin pitää siirtyä käyttämään multinomiaalista logistista regressiota.

3.3 Multinomiaalinen logistinen regressioanalyysi

Multinomiaalinen logistinen regressio on logistisen regression laajempi muoto. Myös tässä käytetään linkkifunktiona yleistettyä logit-funktiota, mutta kaksiarvoisen selitettävän muuttujan sijaan vastemuuttujan jakauma on oltava multinomiaalinen eli useampiluokkainen. Erona binääriseen logistiseen regressioon on se, että vastemuuttujalla on kolme tai useampi luokkaa. Selitettävän muuttujan arvojen ei tarvitse olla missään arvosuhteessa toisiinsa nähden eli vastemuuttuja voi olla nominaaliasteikollinen. [26,32] Vaihtoehtoina voi olla esimerkiksi kolme eri väriä: keltainen, punainen ja sininen. Näitä värejä ei voida muuntaa numeroiksi suoraan asettamalla 1 = keltainen, 2 = punainen ja 3 = sininen, koska keltaisen ja punaisen etäisyys toisistaan ei välttämättä ole sama kuin keltaisen ja sinisen etäisyys. Myös värien järjestykseen asettaminen osoittautuu vääristäväksi, koska punainen ei välttämättä sijaitse keltaisen ja sinisen välissä. Värit voitaisiin asettaa myös toiseen järjestykseen. Tämän ongelman välttämiseksi käytetään multinomiaalista logistista regressiota [26, 32].

Multinomiaalista logistista regressiota voidaan soveltaa myös järjestysasteikollisille vastemuuttujille, joilla on suuruusjärjestys, mutta arvojen etäisyydet toisistaan eivät ole vakiot. Jos etäisyydet ovat vakiot, silloin käytetään kumulatiivista logistista regressiota. [26]

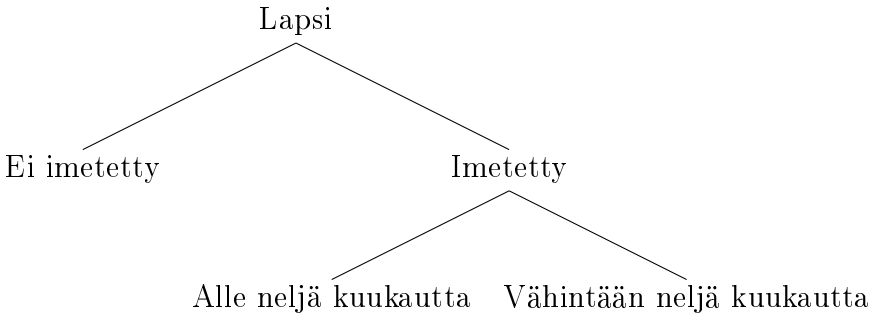
Multinomiaalisessa logistisessa regressiossa tutkitaan, mitkä tekijät ovat vaikuttaneet siihen, että esimerkiksi vaihtoehto keltainen on valittu vaihtoehtojen punainen ja sininen sijaan, eli mitkä tekijät vaikuttavat siihen, että jokin vaihtoehto valitaan muiden sijaan. Selittäviä tekijöitä multinomiaalisessa logistisessa regressiossa voi olla yksi tai useita. Ne voivat olla joko kategorisia tai jatkuvia. Mallille muodostetaan logit-funktioita yksi vähemmän kuin mallissa olevia vastemuuttujien luokkia on. Kolmiluokkaiselle vastemuuttujalle muodostetaan näin ollen kaksi logit-funktioita. [26, 39]

Multinomiaalisen logistisen regression selitettävän tekijän luokkien vertailua voidaan tehdä erilaisin tavoin. Vertailu voidaan suorittaa osituksena,

käyttämällä yhtä referenssiluokkaa tai käyttämällä muuttuvaa referenssiluokkaa. [26, 39] Seuraavaksi käydään nämä tavat läpi lähtien osituksesta ja siirtyen yhden referenssiluokan kautta muuttuvaan referenssiluokkaan.

3.3.1 Ositus

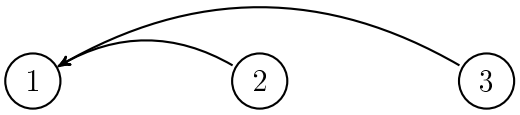
Osituksessa lähdetään liikkeelle siitä, että yksi luokka eroaa selkeästi muista [39]. Esimerkiksi, onko lasta imetetty. Lasta joko on tai ei ole imetetty. Seuraavaksi jaotellaan imetetyt lapset kahteen osioon, onko lasta imetetty vähintään neljä kuukautta vai ei. Ositusta voidaan jatkaa niin pitkälle kuin on tarve [39]. Kuva 3 havainnollista ositusta.



Kuva 3: Miten ositus tapahtuu.

3.3.2 Yksi referenssiluokka

Toinen tapa vertailla luokkia on asettaa yksi luokista vertailuluokaksi. Tätä tapaa käytetään tutkielman soveltavassa osiossa. Vertailuluokaksi asetetaan yleensä luokka, jota on eniten ja muita luokkia verrataan tähän luokkaan. Jos on selkeä ajatus, mikä luokka on kiinnostuksen kohteena, tällöin se asetetaan vertailuluokaksi. [39] Kuva 4 havainnollistaa tätä vertailua.



Kuva 4: Asetetaan luokka 1 vertailuluokaksi.

Kun kaikissa logit-funktiossa käytetään samaa vertailuluokkaa, todennäköisyydet muodostuvat seuraavalla tavalla [38]. Esimerkiksi kolmiluokkaiselle vastemuuttujalle muodostetaan kaksi logit-funktioita, joissa luokkaa 1 pide-

tään vertailuluokkana:

$$
\begin{aligned}
&p_1 = \text{todennäköisyys, että vastemuuttuja saa arvon 1}\\
&p_2 = \text{todennäköisyys, että vastemuuttuja saa arvon 2}\\
&p_3 = \text{todennäköisyys, että vastemuuttuja saa arvon 3}\\
&p_1 + p_2 + p_3 = 1
\end{aligned}
$$

$$
\begin{aligned}
\text{logit}(p_2) &= \alpha_1 + \beta_1 x\\
\text{logit}(p_3) &= \alpha_2 + \beta_2 x.
\end{aligned}
$$

Parametrit α_1, α_2, β_1 ja β_2 ovat tuntemattomia. Ne saadaan estimoitua havaintoaineiston avulla. Huomioidaan, että $p_1 + p_2 + p_3 = 1$. [38] Tällöin saadaan yhtälöryhmät

$$
\begin{cases}
\dfrac{p_2}{p_1} = e^{\alpha_1+\beta_1 x}\\
\dfrac{p_3}{p_1} = e^{\alpha_2+\beta_2 x}\\
p_1 + p_2 + p_3 = 1
\end{cases}
\quad \text{eli} \quad
\begin{cases}
p_2 = e^{\alpha_1+\beta_1 x} p_1\\
p_3 = e^{\alpha_2+\beta_2 x} p_1\\
p_1 + p_2 + p_3 = 1
\end{cases}
.
$$

Näistä saadaan sijoittamalla seuraava yhtälö

$$
\begin{aligned}
p_1 + e^{\alpha_1+\beta_1 x} p_1 + e^{\alpha_2+\beta_2 x} p_1 &= 1\\
(1 + e^{\alpha_1+\beta_1 x} + e^{\alpha_2+\beta_2 x}) p_1 &= 1\\
p_1 &= \frac{1}{1 + e^{\alpha_1+\beta_1 x} + e^{\alpha_2+\beta_2 x}}.
\end{aligned}
$$

Tällöin mallin ennustetodennäköisyydet ovat seuraavat [26]

$$
\begin{aligned}
p_1 &= \frac{1}{1 + e^{\alpha_1+\beta_1 x} + e^{\alpha_2+\beta_2 x}}\\
p_2 &= \frac{e^{\alpha_1+\beta_1 x}}{1 + e^{\alpha_1+\beta_1 x} + e^{\alpha_2+\beta_2 x}}\\
p_3 &= \frac{e^{\alpha_2+\beta_2 x}}{1 + e^{\alpha_1+\beta_1 x} + e^{\alpha_2+\beta_2 x}}.
\end{aligned}
$$

Koska selitettävä muuttuja on kolmiluokkainen, parametreja β muodostuu kaksi kertaa enemmän kuin binäärisen selittäjän tapauksessa. Näin ollen myös OR-lukuja muodostuu kaksi kertaa enemmän kuvaamaan luokkien välisiä eroja. Esimerkiksi, jos vaste on kolmiluokkainen ja selittäviä tekijöitä on yksi kaksiluokkainen tekjä, saadaan frekvenssitaulukko 8.

Selittäjä	Vasteluokat		
	1	2	3
Ryhmä 1	n_{11}	n_{12}	n_{13}
Ryhmä 2	n_{21}	n_{22}	n_{23}

Taulukko 8: Havaintojen lukumäärät eroteltuna luokkien ja ryhmien välillä, kun vaste on kolmiluokkainen ja selittäviä tekijöitä on vain yksi kaksiluokkainen tekijä.

OR-lukujen suhteen arvot saadaan taulukosta 8, mikä muistuttaakin paljon taulukkoa 6, mutta siinä vasteluokkia on kolme kahden luokan sijaan. Vasteluokka 2 verrattuna luokkaan 1

$$\text{OR-lukujen suhde} = e^{\beta_1} = \frac{n_{12} \cdot n_{21}}{n_{11} \cdot n_{22}}.$$

Vasteluokka 3 verrattuna luokkaan 1

$$\text{OR-lukujen suhde} = e^{\beta_2} = \frac{n_{13} \cdot n_{21}}{n_{11} \cdot n_{23}}.$$

OR-lukujen luottamusvälit saadaan laskettua samoin kuin binäärisen logistisen regression kohdalla. [26]

Tutkitaan vielä tapausta, jossa on useampi selittävä muuttuja. Vasteena on kolmiluokkainen imetysmuuttuja: imetetty vähintään 4 kk, imetetty, mutta alle 4 kk ja ei imetetty. Selittävinä tekijöinä ovat kaksiluokkainen sukupuoli muuttuja: poika ja tyttö, sekä kolmiluokkainen äidin koulutus muuttuja: peruskoulu, toisen asteen koulutus ja korkeakoulu. Frekvenssitaulukko 9 näyttää seuraavalta.

Selittäjä		Vasteluokat		
		1	2	3
Sukupuoli	Tyttö	n_{11}	n_{12}	n_{13}
	Poika	n_{21}	n_{22}	n_{23}
Äidin koulutus	Peruskoulu	n_{31}	n_{32}	n_{33}
	Toisen asteen koulutus	n_{41}	n_{42}	n_{43}
	Korkeakoulu	n_{51}	n_{52}	n_{53}

Taulukko 9: Havaintojen lukumäärät eroteltuna luokkien ja ryhmien välillä, kun vaste on kolmiluokkainen ja selittäviä tekijöitä on kaksi.

Kun käytetään vasteluokkaa yksi referenssiluokkana, tästä saadaan sukupuoleen liittyvät OR-luvut

$$e^{\beta_1} = \frac{n_{12} \cdot n_{21}}{n_{11} \cdot n_{22}}$$

ja

$$e^{\beta_2} = \frac{n_{13} \cdot n_{21}}{n_{11} \cdot n_{23}}$$

samoin kuin edellä. Koska sukupuoli ei ole ainoa selittävä tekijä, OR-lukuja tulee useampi kuin kaksi. Vasteluokka 2 verrattuna luokkaan 1 verrattaessa peruskoulua toisen asteen koulutukseen OR-luku on

$$e^{\beta_3} = \frac{n_{32} \cdot n_{41}}{n_{31} \cdot n_{42}}.$$

Jos vasteluokkien vertailu pysyy samana, mutta verrataan peruskoulua korkeakouluun, OR-luku on

$$e^{\beta_4} = \frac{n_{32} \cdot n_{51}}{n_{31} \cdot n_{52}}.$$

Vastaavasti verrattaessa luokkaa 3 luokkaan 1 saadaan seuraavat OR-luvut

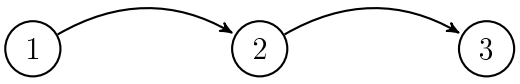
$$e^{\beta_5} = \frac{n_{33} \cdot n_{41}}{n_{31} \cdot n_{43}}$$

ja

$$e^{\beta_6} = \frac{n_{33} \cdot n_{51}}{n_{31} \cdot n_{53}}.$$

3.3.3 Muuttuva referenssiluokka

Kolmas tapa vertailla luokkia on muuttaa aina vertailuluokkaa kuvan 5 osoittamalla tavalla. [39]



Kuva 5: Vaihdetaan vertailuluokkaa.

Tällöin saadaan

$$\text{logit}(p_1) = \log\left(\frac{p_1}{p_2}\right) = \alpha_1 + \beta_{11}x_1 + \beta_{12}x_2 + ... + \beta_{1k}x_k$$
$$\text{logit}(p_2) = \log\left(\frac{p_2}{p_3}\right) = \alpha_2 + \beta_{21}x_1 + \beta_{22}x_2 + ... + \beta_{2k}x_k$$
$$\text{logit}(p_{r-1}) = \log\left(\frac{p_{r-1}}{p_r}\right) = \alpha_{(r-1)} + \beta_{(r-1)1}x_1 + \beta_{(r-1)2}x_2 + ... + \beta_{(r-1)k}x_k.$$

Asetetaan β_{i1} arvot yhtä suuriksi, koska muuttujan x_1 vaikutus pitää olla aina samansuuruinen, ja merkitään niitä symbolilla β_1. [39] Tehdään sama kaikille β arvoille. Tällöin saadaan [39]

$$\text{logit}(p_1) = \log\left(\frac{p_1}{p_2}\right) = \alpha_1 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$
$$\text{logit}(p_2) = \log\left(\frac{p_2}{p_3}\right) = \alpha_2 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$
$$\text{logit}(p_{r-1}) = \log\left(\frac{p_{r-1}}{p_r}\right) = \alpha_{(r-1)} + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k.$$

Kun asetetaan muut muuttujat vakioiksi ja kasvatetaan muuttujaa x_1 yhdellä yksiköllä saadaan OR-luku $= e^{\beta_1}$. Samalla tavalla menetellään muiden muuttujien kohdalla. [39] Tätä ei ole tarpeen käsitellä syvällisemmin, koska tämä menee kumulatiivisen logistisen regression puolelle, joka ei ole oleellista soveltavan osion kannalta.

Kun malli on saatu muodostettua, on tärkeää tarkastella, onko malli yhteensopiva aineiston kanssa ja kuinka hyvin se kuvaa aineistoa.

4 Mallin tarkastelu

4.1 Mallin yhteensopivuus

Mallin hyvyydellä tarkoitetaan sen yhteensopivuutta aineiston kanssa eli kuvaako malli aineistoa riittävän hyvin. Tätä voidaan tarkastella erilaisten testien avulla. Jos mallissa ei ole numeerisia selittäviä muuttujia, hyvyyttä voidaan tutkia Pearsonin ja devianssin yhteensopivuustestien avulla. Jos mallissa on numeerisia selittäviä muuttujia, hyvyyden tutkimiseen voidaan käyttää Hosmer-Lemeshowin yhteensopivuustestiä. [26, 31] Soveltavassa osiossa käytetään Hosmer-Lemeshowin yhteensopivuustestiä, koska äidin ikä on jatkuva numeerinen muuttuja. Tämän testin teoriaa on esitelty tutkielman kappaleessa kolme. Hyvyyden tutkimisessa tarkastellaan yhteensopivuustestin p-arvoa [26, 31]. Jos testin tuloksena saatu p-arvo on tilastollisesti merkitsevä eli alle 0.05, malli ei ole yhteensopiva aineiston kanssa [26, 31]. Muussa tapauksessa malli on yhteensopiva aineiston kanssa [26, 31].

4.2 Diagnostiset tarkastelut

Diagnostisessa tarkastelussa on hyvä tarkistaa, täyttyvätkö mallin vaatimat oletukset. Multinomiaalisella logistisella regressiolla näitä ovat vasteen moniluokkaisuus, selittävien muuttujien välinen korrelaatio, täydellisten rivien käyttö, lineaarisuuden oletus ja se, että poikkeavia arvoja ei ole. [26, 40] Osa oletuksista on yksikertaista tutkia aineistoa analysoimalla. Näin toimitaan vasteen moniluokkaisuuden ja täydellisten rivien oletusten kanssa. Selittävien tekijöiden välisten korrelaatioiden tutkiminen tehdään χ^2- ja t-testin avulla [13, 16]. Lineaarista suhdetta tutkittaessa analysoidaan mallin yhteensopivuutta aineiston kanssa Hosmeri-Lemeshowin yhteensopivuustestillä [25].

Aineistosta poikkeavien havaintojen tutkimiseen on kehitetty erilaisia menetelmiä. Logistisen mallin ennustamien ja havaittujen arvojen eroja tarkastellaan Pearsonin, devianssin ja kvantiiliresiduaalien (Quantile Residuals) avulla. Nämä kertovat, kuinka hyvin logistinen malli pystyy ennustamaan havaintoarvoja. Poikkeavaa havaintoarvoa on syytä epäillä, jos standardoitu Pearson, devianssi tai kvantiiliresiduaalin itseisarvo on suurempi kuin kolme. Poikkeavat arvot voidaan tunnistaa helposti residuaalien kuvasta. [26, 31, 40]

Pearsonin ja devianssin residuaalit jakautuvat suunnilleen normaalisti. Diskreetit jakaumat eivät kuitenkaan aina noudata tätä sääntöä ja tällöin vaihtoehtona on käyttää kvantiiliresiduaaleja. Näiden residuaalien etuna on, että ne ovat jatkuvia myös diskreeteille jakaumille toisin kuin devianssin ja Pearsonin residuaalit. Tämän takia soveltavassa osiossa käytetään kvantiiliresiduaaleja. [26, 31, 40]

5 Soveltaminen aineistoon

Tutkielman sovelluskohde saatiin Hyvän kasvun avaimet -tutkimuksesta. Tutkimuskysymyksenä on "sosiaalisen elinympäristön yhteys imetykseen" eli onko sosiaalisella ympäristöllä vaikutusta siihen, imetetäänkö lasta ja jos imetetään, jatkuuko imetys yli neljä kuukautta. Sosiaalinen elinympäristö tarkoittaa sosiaalisista suhteista kuten perheestä, työstä, ystävistä ja yhteisöistä muodostuvaa ympäristöä [41].

Tutkielmassa tutkitaan imetystä ja sen kestoa, koska tällä on todettu olevan terveydellisiä hyötyjä sekä lyhyellä että pidemmällä aikavälillä niin lapselle kuin äidille. Imetys muun muassa vähentää lapsen ja äidin riskiä sairastua erilaisiin kroonisiin sairauksiin, kuten diabetekseen, sydän- ja verisuonitauteihin, astmaan sekä erilaisiin syöpiin. Esimerkki lyhyellä aikavälillä saatavasta hyödystä on lapsen pienentynyt riski saada suolistotulehdus tai vakavia alahengitysteiden infektioita. [42–44]

Tutkittaviksi tekijöiksi valikoituivat lapsen sukupuoli, synnytystapa, äidin koulutus, äidin painoindeksi, asuinalueen huono-osaisuus ja äidin ikä. Sukupuoli valikoitui tutkittavaksi tekijäksi, koska lääketieteessä sukupuoli on yleensä mielekäs tekijä tutkittavaksi. Muut tekijät valikoituivat jo olemassa olevan tiedon perusteella. Useissa eri tutkimuksissa on huomattu, että äidin koulutuksella on vaikutusta imetykseen [45–47]. Esimerkiksi kalifornialaisten äitien koulutuksen ja imetyksen väliltä on löydetty yhteys [48]. Eräässä imetyksen lopettamiseen vaikuttavia tekijöitä kartoittaneessa tutkimuksessa nousi esille, että äidin koulutuksen lisäksi myös äidin ikä ja äidin ylipaino ovat imetykseen vaikuttavia tekijöitä [45]. Myös muut tutkimukset tukevat tätä tulosta [46]. Lisäksi eräässä Yhdysvalloissa tehdyssä tutkimuksessa imetykseen vaikuttaviksi tekijöiksi nousivat synnystapa ja asuinalue [47, 49].

Tässä aineistossa asuinalue on otettu tutkittavaksi perheen asuinalueen huono-osaisuutta kuvaavan sosioekonominen status -tekijän (SES) avulla. Tämä on koottu kolmen eri tekijän pohjalta kuvaamaan asuinalueen sosioekonomista huono-osaisuutta. Sosioekonominen asema tarkoittaa henkilön asemaa yhteiskunnassa. Asema määräytyy pääsääntöisesti henkilön työn perusteella. Luokat jakautuvat ammatissa toimivien ja ammatissa toimimattomien välille ja niitä on esimerkiksi yrittäjät, palkansaajat, opiskelijat ja työttömät. Työn lisäksi myös koulutus vaikuttaa sosioekonomiseen asemaan. [50] Samoilla kriteereillä voidaan määritellä asuinalueen sosioekonominen asema.

Huono-osaisuudelle on monia eri määritelmiä, mutta tässä työssä sillä tarkoitetaan pieniä tuloja, alhaista koulutustasoa ja korkeaa työttömyysastetta eli elinolojen ja hyvinvoinnin puutetilojen kasaantuminen samalle alueelle. Asuinalueen sosioekonomista huono-osaisuutta kuvaavat tekijät ovat kotitalouksien mediaanitulot (käännetty niin päin, että pienituloiset saavat kor-

keamman arvon), alhainen koulutustaso (yli 18 vuotiaiden osuus, joiden korkein koulutustaso oli peruskoulu) ja työttömyysaste. Näin saatiin tekijä, joka saa arvoja välillä (-2.13, 3.86). Korkeat arvot kertovat, että huono-osaisuus on suurta eli asuinalue on huono. Matalat arvot taas kertovat, että huono-osaisuus on pientä eli asuinalue on hyvä. Tässä tutkielmassa on käytetty asuinalueen kokona 250 m x 250 m. [51, 52]

5.1 Aineisto

Aineisto on saatu Turun yliopiston koordinoimasta (HKA) Hyvän kasvun avaimet -tutkimuksesta. HKA -seurantatutkimus on monitieteellinen tutkimus, jossa on kerätty tietoa varsinaissuomalaisten perheiden fyysisestä, psyykkisestä ja sosiaalisesta terveydestä. Varsinais-Suomen sairaanhoitopiirissä synnyttäneet suomen- ja ruotsinkieliset äidit ($n = 9\ 811$) lapsineen ($n = 9\ 936$) vuosina 2008 - 2010 on alun perin kuuluneet tutkimuskohorttiin. [54] Tutkimuskohortti tarkoittaa ryhmää, jota tarkkaillaan tiettynä ajanjaksona [53]. Tutkimuskohortti ei voi kasvaa jälkikäteen [53]. Hyvän kasvun avaimet -tutkimuksen seurantaosioon (The STEPS Study), osallistui 1 797 äitiä ja 1 658 puolisoa [54]. Lapsia tutkimukseen osallistui 1 805 [54]. Taulukossa 10 on kuvailtu aineiston kaikki muuttujat.

Muuttujan nimi	Määritelmä	Mistä saatu
nro	Lapsen henkilökohtainen id	-
perheennro	Perheen id	-
SP	Lapsen sukupuoli	Syntyneiden lasten rekisteri
kuollut	Kertoo onko lapsi kuollut	Digi- ja väestötietovirasto
kuollut_paiva_lkm	Kuoleman jälkeisten päivien lukumäärä päivissä, jos lapsi on kuollut	Digi- ja väestötietovirasto
taysimetys_laskennallinen	Täysimetyksen kesto kuukausissa	Seurantakirja
kokonaisimetys	Imetyksen kokonaiskesto kuukausissa	Seurantakirja
lisaruokienAloitus	Ikä, jolloin lisäruokinta on aloitettu kuukausissa	Seurantakirja
osittaisimetys	Täysimetyksen jälkeen jatkuneen imetyksen kesto kuukausissa	Seurantakirja
korvikkeen_aloitus_ika	Ikä, joilloin korvikeen käyttö on aloitettu kuukausissa	Seurantakirja
eiImetetty	Kertoo onko lasta imetetty	Seurantakirja
SYNNYTYSTAPA_2luok	Synnytystapa	Syntyneiden lasten rekisteri
BMI_aiti	Äidin painoindeksi	Syntyneiden lasten rekisteri
Aidinika	Äidin ikä synnytyshetkellä vuosissa	Syntyneiden lasten rekisteri
apkoul	Äidin peruskoulutus	Kyselylomake
aamkoul	Äidin ammatillinen koulutus	Kyselylomake
disadv	Lapsen syntymähetken asuinalueen huono-osaisuus aluekoolla 250m x 250m	Tilastokeskuksen ruututietotkannasta
disadv_2	Lapsen syntymähetken asuinalueen huono-osaisuus aluekoolla 750m x 750m	Tilastokeskuksen ruututietotkannasta

Taulukko 10: Saadun aineiston muuttujien kuvailu ja tieto, mistä ne on saatu [54].

Tämän tutkimuksen aineisto rajattiin niihin lapsiin, joilta oli saatavilla kaikki tutkimukseen tarvittavat tiedot eli tiedot imetyksestä, lapsen sukupuolesta, synnytystavasta, äidin koulutuksesta, äidin painoindeksistä, asuinalueen huono-osaisuudesta sekä äidin iästä. Imetystiedot olivat puutteelliset 696 lapsella, äidin koulutuksen tieto puuttui 26 lapselta, tieto äidin painoindeksistä puuttui seitsemältä lapselta ja tieto asuinalueen huono-osaisuudesta puuttui 157 lapselta, joten nämä lapset rajautuvat pois tämän tutkimuksen aineistosta. Aineistosta karsiutui lisäksi pois 45 lasta, koska nämä eivät sopineet mihinkään aineiston jaottelussa valittuun imetysluokkaan puutteellisten tietojen vuoksi. Esimerkiksi tieto kokonaisimetyksestä oli saatavilla, mutta sen jakautumisesta täysimetyksen ja osittaisimetyksen välille ei ole tietoja. Tällöin ei voida tietää, kuuluuko lapsi imetysluokkaan yksi vai kaksi. Nämä luokat määritellään aineiston karsinnan jälkeen.

Aineistossa esiintyy 32 kaksosta ja 47 lasta, jotka ovat syntyneet ennen

raskausviikkoa 37. Myös nämä jätettiin aineiston ulkopuolelle, koska imetys voi erota selvästi sekä kaksosilla että keskosilla verrattuna muihin lapsiin. Aineistossa kaksi lasta menehtyi aineiston keräämisen aikana. Lapset kuitenkin elivät niin pitkään, että heidän imetystään voidaan tutkia, joten heidät pidetään mukana aineistossa. Tällöin lopullinen otoskoko oli 872 eli 48% alkuperäisestä aineistosta.

Koska tutkimuskysymyksenä on ”sosiaalisen elinympäristön yhteys imetykseen”, aineiston jaottelu tehdään imetyksen mukaan. Aineistossa on kuusi muuttujaa, jotka kuvaavat imetystä. Aineisto jaoteltiin kolmeen ryhmään neljän imetystä kuvaavan muuttujan avulla. Muuttujat ovat ”kokonaisimetys”, ”osittaisimetys”, ”taysimetys_laskennallinen” ja ”eiImetetty”. Ryhmään yksi (1) kuuluvat lapset, joita oli täysimetetty neljä kuukautta tai pidempään. Ryhmään kaksi (2) kuuluvat lapset, joita oli täysimetetty alle neljä kuukautta, mutta sen jälkeen osittaisimetetty, niin että imetyksen kokonaiskesto oli vähintään kuukauden. Ryhmään kolme (3) kuuluvat lapset, joita ei ollut imetetty tai imetys oli kestänyt alle kuukauden. Taulukossa 11 kuvataan, miten aineisto on jakautunut kolmeen imetysluokkaan.

	Imetysluokka			
	1	2	3	Yhteensä
Määrä (kpl)	333	500	39	872
Prosenttiosuus (%)	38	57	4	

Taulukko 11: Käytetyn aineiston jakautuminen imetysluokkiin ilmoitettuna havaintojen määrinä ja prosentteina.

Ollaan kiinnostuneita seuraavien muuttujien vaikutuksesta varhaisiän ravitsemukseen: lapsen sukupuoli, synnytystapa, äidin koulutus, äidin painoindeksi (BMI), lapsen syntymähetken asuinalueen huono-osaisuus (SES) ja äidin ikä. Tutkitaan näitä muuttujia tarkemmin. Lapsen sukupuoli ja synnytystapa ovat kaksiluokkaisia muuttujia (0,1). Äidin koulutus -muuttujassa on luokat 1-9. Koska on kiinnostavaa tutkia korkeasti koulutettujen ja matalasti koulutettujen äitien eroja imetyksessä, tämä skaalattiin kaksiluokkaiseksi koulutusmuuttujaksi. Skaalaus tapahtui siten, että luokat 1-4 arvioitiin matalaksi koulutusasteeksi ja luokat 5-8 korkeaksi koulutusasteeksi. Luokka 9 tarkoittaa luokkaa muu, joten ne katsottiin tapauskohtaisesti.

Äidin painoindeksiä kuvaava muuttuja BMI saa arvoja välillä (17-52). Koska ylipainolla oli löydetty olevan merkitystä imetyksen kanssa, skaalataan painoindeksi kahteen luokkaan: ylipainoiset ja normaalipainoiset äidit. Arvot skaalattiin uudeksi muuttujaksi BMI_raja siten, että ne, joilla BMI-arvo on yli 25 ovat ylipainoisia. Tätä merkataan arvolla 1. Ne, joilla BMI ei

ollut puuttuva ja se oli 25 tai alle, saivat arvon 0. Lapsen syntymähetken asuinalueen huono-osaisuutta kuvaava muuttuja disadv saa arvoja välillä (-2.22, 2.61). Asuinalueen huono-osaisuus skaalattiin kaksiluokkaiseksi muuttujaksi SES sen mukaan, oliko asuinalueen huono-osaisuus matala vai korkea. Jos arvo oli yli nollan, huono-osaisuus oli korkea, joten uuden muuttujan arvoksi asetettiin 1. Jos arvo oli nolla tai alle sen, alueen huono-osaisuus oli matala ja arvoksi asetetaan 0. Äidin ikä on jatkuva muuttuja, joka saa arvoja välillä (18-41). Tälle ei tarvitse tehdä skaalauksia, joten se pysyy ainoa jatkuvana muuttujana.

5.2 Tutkimuksen suunnitelma

Tutkimuksen analyysit tehtiin R-ohjelmiston versiolla 4.0.3. Tilastollisen merkitsevyyden rajana käytettiin p-arvo 0.05. Soveltavan osion R-koodi on esitetty liitteenä A.

Tutkimuksen tarkoituksena on selvittää, onko sosiaalisella elinympäristöllä yhteys imetykseen. Kiinnostavia tekijöitä ovat näin ollen asuinalueen sosioekonominen huono-osaisuus ja äidin koulutus, joka kuvaa myös äidin sosioekonomista asemaa. Lisäksi tutkitaan, vaikuttaako lapsen sukupuoli tai synnytystapa imetykseen. Tutkitaan myös, onko äidin taustalla vaikutusta lapsen imetykseen. Äidin taustoja ovat ikä ja painoindeksi.

Aineiston frekvenssit ovat taulukon 12 mukaiset. Taulukosta 13 nähdään, miten aineisto on jakautunut imetysluokkien välille ja millaiset frekvenssit ovat imetysluokkien sisällä.

		Yhteensä	
		Määrä	Prosentti
Sukupuoli	Poika	455	52
	Tyttö	417	48
Synnytystapa	Alateitse	776	89
	Sektio	96	11
Äidin koulutus	Matala	286	33
	Korkea	586	67
Äidin BMI	Normaalipainoinen	619	71
	Ylipainoinen	253	29
Asuinalueen huono-osaisuus	Matala	619	71
	Korkea	253	29
Äidin ikä	Mediaani	30.9	
	Kvartaalit	28.2, 33,6	
Yhteensä		872	

Taulukko 12: Tässä tutkimuksessa käytetyn aineiston jakautuminen sukupuolen, synnytystavan, äidin koulutuksen, äidin painoindeksin, lapsen syntymähetken asuinalueen huono-osaisuuden ja äidin iän mukaan. Jakautuminen esitetty havaintojen määrinä ja prosentteina.

		Imetysluokka					
		1		2		3	
		Määrä	Prosentti	Määrä	Prosentti	Määrä	Prosentti
Sukupuoli	Poika	162	49	273	55	20	51
	Tyttö	171	51	227	45	19	49
Synnytystapa	Alateitse	303	91	446	89	27	69
	Sektio	30	9	54	11	12	31
Äidin koulutus	Matala	96	29	164	33	26	67
	Korkea	237	71	336	67	13	33
Äidin BMI	Normaalipainoinen	249	75	349	70	21	54
	Ylipainoinen	84	25	151	30	18	46
Asuinalueen huono-osaisuus	Matala	238	71	357	71	24	62
	Korkea	95	29	143	29	15	38
Äidin ikä	Mediaani	30.6		31.0		30.2	
	Kvartaalit	28.3,33.6		28.1,33.6		27.4,34.6	
Kaikki	872	333	38	500	57	39	4

Taulukko 13: Tässä tutkimuksessa käytetyn aineiston jakautuminen imetysluokkien sisällä sukupuolen, synnytystavan, äidin koulutuksen, äidin painoindeksin, lapsen syntymähetken asuinalueen huono-osaisuuden ja äidin iän mukaan. Jakautuminen esitetty havaintojen määrinä ja prosentteina.

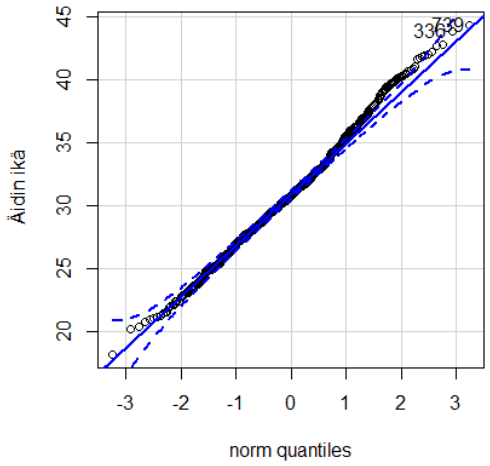
Koska lapset on jaettu kolmeen luokkaan imetyksen perusteella, aineistoa lähdettiin tutkimaan multinomiaalisen logistisen regression avulla. Selittävät tekijät ovat lapsen sukupuoli, synnytystapa, äidin koulutus, ikä ja painoindeksi sekä asuinalueen huono-osaisuus. Käytetään vertailuluokkana lapsia, joita on täysimetetty vähintään neljä kuukautta eli imetysluokkaa yksi.

5.3 Toteutus

Selittävien muuttujien välillä ei saa olla vahvoja riippuvuussuhteita, koska tällöin multinomiaalinen regressio ei toimi [29]. Tutkitaan suhteita t-testin ja χ^2-testin avulla. T-testin avulla voidaan tutkia jatkuvan muuttujan suhdetta muihin muuttujiin, kun taas χ^2 -testin avulla tutkitaan kahden kategorisen muuttujan välistä suhdetta [13, 16]. Ainoa jatkuva muuttuja on äidin ikä. Kuvasta 6 nähdään, että äidin ikä -muuttuja on lähellä normaalijakaumaa. Tämän muuttujan suhdetta muihin selittäviin muuttujiin voidaan tutkia t-testin avulla [13]. Ensin tutkitaan Levenen testin avulla, käytetäänkö t-testissä varianssien yhdenmuotoisuutta vai erimuotoisuutta [14]. Sen jälkeen tutkitaan muuttujien väliset suhteet t-testillä. Siitä saadut p-arvot näkyvät taulukossa 14.

Aineiston jakautumista kuvaavasta taulukosta 14 nähdään, että asuinalueen huono-osaisuudella ja äidin iän välillä on riippuvuussuhde, koska p-arvo on pieni (<0.001). Tutkitaan tätä suhdetta tarkemmin kuvan 7 avulla. Kuvasta nähdään, että muuttujien välinen riippuvuus ei ole suurta, joten se ei ole haitaksi multinomiaaliselle logistiselle regressiolle ja molemmat muuttujat voidaan säilyttää mallissa.

Koska vain äidin ikä on jatkuva muuttuja, muiden selittävien muuttujien väliset suhteet tutkitaan χ^2-testin avulla. Saadut p-arvot näkyvät taulukossa 15. Taulukosta nähdään, että riippuvuuksia on asuinalueen huono-osaisuuden ja äidin painoindeksin välillä (p-arvo ≈ 0.017), asuinalueen huono-osaisuuden ja äidin koulutuksen välillä (p-arvo < 0.001) sekä äidin painoindeksin ja koulutuksen välillä (p-arvo < 0.001). Tutkitaan näiden muuttujien välisiä suhteita tarkemmin Cramerin V:n avulla. Taulukosta 15 nähdään, että Cramerin V:n kaikki arvot ovat pieniä. Näin ollen muuttujien väliset riippuvuudet eivät ole suuria, joten ne eivät ole haitaksi multinomiaaliselle logistiselle regressiolle. Näin ollen kaikki muuttujat voidaan säilyttää mallissa.



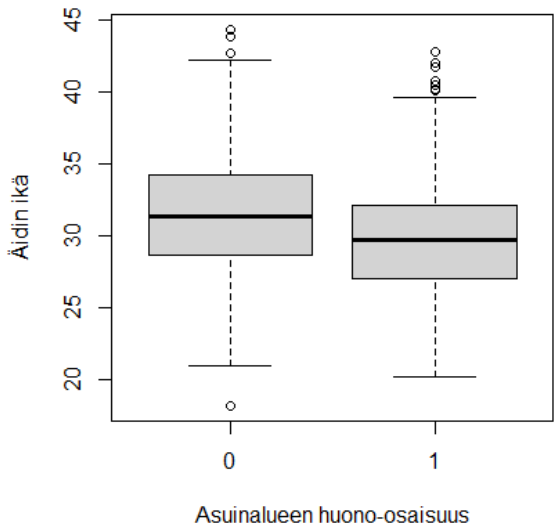
Kuva 6: Äitien iän normaalisti jakautumisen testaus.

t-testi	Asuinalueen huono-osaisuus	BMI	Koulutus	Sukupuoli	Synnytystapa
p-arvo	< 0.001	0.309	0.756	0.762	0.764

Taulukko 14: Äidin iän ja muiden muuttujien välisten suhteiden tarkastelu t-testin avulla. Taulukossa esitetty t-testillä saadut p-arvot.

χ^2 -testin p-arvot	BMI	Koulutus	Lapsen sukupuoli	Synnytystapa
Asuinalueen huono-osaisuus	0.008	< 0.001	0.204	0.747
Cramerin V:n kerroin	0.092	0.140		
BMI		< 0.001	0.939	0.384
Cramerin V:n kerroin		0.156		
Koulutus			0.207	0.355
Lapsen sukupuoli				0.898

Taulukko 15: Kategoristen muuttujien välisten suhteiden tarkastelu khiin neliötestin avulla. Taulukossa esitetään testin avulla saadut p-arvot. Tilastollisesti merkitsevät suhteet on tarkasteltu Cramerin V:n avulla ja nämä tulokset on esitetty myös taulukossa.



Kuva 7: Äitien ikien sijoittuminen asuinalueen huono-osaisuuden mukaan.

5.4 Diagnostiikka

Mallin hyvyyttä tarkastellaan diagnostiikan avulla. Tutkitaan täyttääkö malli kaikki oletukset, mitä multinomiaalisessa logistisessa regressiossa on. Kuten aikaisemmin jo todettiin, oletukset ovat seuraavanlaiset: multinomiaalisessa logistisessa regressiossa vasteen on oltava moniluokkainen kategorinen muuttuja, selittävien muuttujien välillä ei saa olla keskenään suurta korrelaatiota, käytetään vain täydellisiä rivejä, oletus lineaarisuudesta ja poikkeavia arvoja ei saa olla. Lineaarisuuden oletus tarkoittaa, että selittävillä muuttujilla on lineaarinen suhde vasteen logitiin. [55]

Vasteen moniluokkaisuuden oletus täyttyy, koska vasteena on kolmeluokkainen imetysmuuttuja. Muuttujien välillä ei ole multikolineaarisuutta eli muuttujat eivät korreloi keskenään voimakkaasti. Tämä on tutkittu t-testin ja khiin neliötestin avulla ja tulokseksi saatiin, ettei muuttujien välillä ole voimakkaita vuorovaikutuksia. Alussa aineistosta karsittiin epätäydelliset rivit eli rivit, jotka sisältävät puuttuvia arvoja, joten rivien täydellisyyssääntö täyttyy.

Logistisen regression yksi vaatimuksista on, että muuttujien ja logaritmin välisen suhteen on oltava lineaarinen. Tätä suhdetta tutkittiin Hosmerin ja Lemeshowin testin avulla. Oletuksena on, että suhde on lineaarinen, joten jos tulokseksi saadaan tilastollisesti ei merkitsevä p-arvo, malli on tältä osin

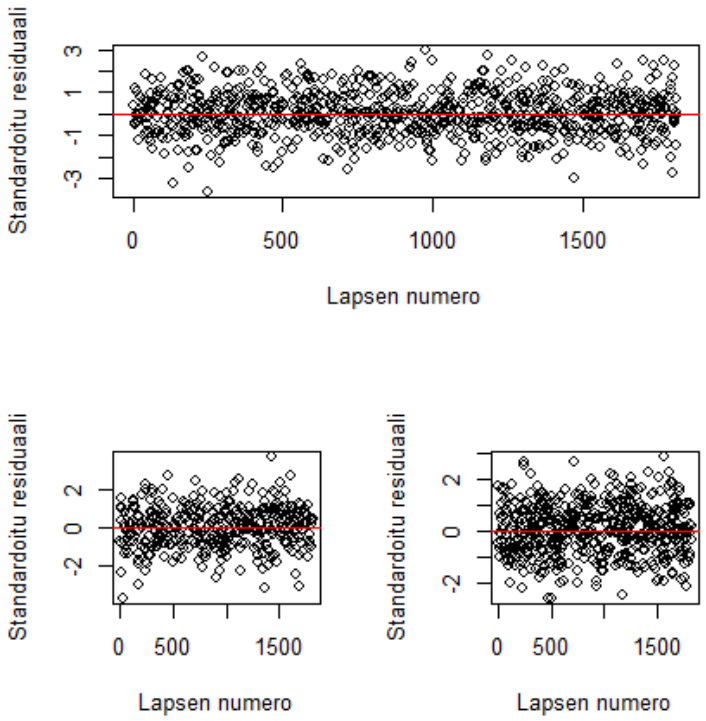
sopiva. Tulokseksi saatiin p-arvon 0.963, joka ei ole tilastollisesti merkitsevä, joten mallin logaritmin ja muuttujien välinen suhde on lineaarinen ja tämä oletus täyttyy. [56]

Osa	Riskin suuruus	Yht.	Ryhmä 1		Ryhmä 2		Ryhmä 2	
			Havainto	Ennuste	Havainto	Ennuste	Havainto	Ennuste
1	[0.539 - 0.544]	88	40	40.3	47	46.3	1	1.4
2	(0.544 - 0.545]	87	40	39.6	47	46.1	0	1.2
3	(0.545 - 0.601]	87	34	35.3	49	49.5	4	2.1
4	(0.601 - 0.602]	87	35	34.6	48	48.0	4	4.4
5	(0.602 - 0.604]	87	34	34.4	50	51.0	3	1.5
6	(0.604 - 0.605]	87	35	34.4	51	51.3	1	1.3
7	(0.605 - 0.658]	87	32	31.7	51	51.5	4	3.9
8	(0.658 - 0.660]	87	32	29.7	50	53.9	5	3.4
9	(0.660 - 0.691]	87	31	29.1	51	51.0	5	6.8
10	(0.691 - 0.808]	88	20	23.7	56	51.3	12	12.9

Taulukko 16: Aineiston havaintojen ja mallin ennustamien tulosten jakautumisesta kymmeneen ryhmään Hosmerin ja Lemeshowin testin mukaan. $\chi^2_{16} = 7.464$

Mallin standardoidut residuaalit laskettiin jokaiselle logistiselle regressiolle erikseen. Koska vaste ei ole jatkuva vaan selkeästi kaksiluokkainen, ovat residuaalit jakautuneet selkeästi kahteen eri luokkaan. Tämän takia standardoitujen residuaalien laskemiseen käytettiin kvantiiliresiduaaleja. Kvantiiliresiduaalit ovat jatkuvia myös diskreeteille jakaumille. [40]

Residuaalikuvista 8 nähdään, ettei jakaumissa ole poikkeavia arvoja. Multinomiaalisen logistisen regression ehtona oli, ettei poikkeavia arvoja saa olla. Näin ollen malli toteuttaa myös tämän oletuksen.



Kuva 8: Mallin standardoidut residuaalit jokaiselle logistiselle regressiolle lapsen numeron mukaan esitettynä. Logistisia regressioita on siis kolme, koska toisiinsa verrattavia luokkia on kolme. Ensimmäisessä kuvassa logistinen regressio on tehty imetysluokan yksi ja kaksi välille. Toisessa kuvassa regressio on imetysluokan yksi ja kolme välillä ja viimeisessä kuvassa imetysluokan kaksi ja kolme välillä. Residuaalien laskemiseen on käytetty kvantiiliresiduaaleja (Quantile Residuals). Kuvista nähdään, onko jakaumissa poikkeavia arvoja.

5.5 Tulokset

Taulukoista 12 ja 13 nähdään, että suunnilleen puolet aineistosta on sukupuoleltaan poikia (52%) ja puolet tyttöjä. Myös imetysluokkien sisällä tyttöjen ja poikien suhde pysyy samana. Suurin osa lapsista on syntynyt alateitse (89%) ja tämä toistuu myös kaikissa imetysluokissa. Imetysluokassa kolme vaikuttaisi kuitenkin muihin luokkiin verrattuna olevan poikkeavan paljon sektiolla syntyneitä lapsia (31%). Suurin osa äideistä on korkeasti koulutettuja (67%). Luokassa yksi ja kaksi korkeasti koulutettuja äitejä on enemmän kuin matalasti koulutettuja äitejä. Luokka kolme eroaa taas muista luokista, sillä siinä matalasti koulutettuja äitejä (67%) on enemmän kuin korkeasti koulutettuja. Painoindeksin mukaan normaalipainoisia äitejä on suurin osa aineistosta (71%). Tämä pätee myös luokissa yksi ja kaksi. Luokassa kolme sen sijaan on normaalipainoisia ja ylipainoisia äitejä melkein yhtä paljon.

Suurin osa perheistä asuu matalan asuinalueen huono-osaisuuden alueella (71%). Tämä pätee kaikissa imetysluokissa, mutta imetysluokassa kolme on havaittavissa pientä eroavaisuutta muihin luokkiin verrattuna. Siinä matalan huono-osaisuuden alueella asuvia perheitä on vain 62%. Koko aineistossa äitien keskimääräinen ikä on 30,9 vuotta. Kaikkien imetysluokkien äitien keskimääräinen ikä on suunnilleen sama, joten iän suhteen ei taulukoita vertailemalla havaita eroavaisuuksia.

Taulukoiden 12 ja 13 pohjalta näyttää näin ollen, että imetysluokassa kolme on selkeitä eroavaisuuksia imetysluokkaan yksi ja kaksi verrattuna. Imetysluokkien yksi ja kaksi välillä ei havaita merkittäviä eroja. Suurimmat erot löytyvät niiden lasten väliltä, joita on ylipäätään imetetty yli kuukauden ja alle kuukauden tai ei ollenkaan imetettyjen väliltä. Selkein ero taulukoiden perusteella on äidin koulutuksen osalta. Imetysluokassa kolme on enemmän matalasti koulutettuja äitejä kuin korkeasti koulutettuja. Asuinalueen huono-osaisuudessa, synnytystavassa ja äidin painoindeksissä on myös eroavaisuutta muihin luokkiin verrattuna. Alle kuukauden tai ei ollenkaan imetettyjen luokassa sektiolla syntyneiden osuus on isompi kuin muissa imetysluokissa. Myös ylipainoisten äitien ja korkean asuinalueen huono-osaisuuden osuudes ovat näissä luokissa isompia kuin muissa imetysluokissa.

Multinomiaalisesta logistisesta regressiosta saadaan seuraavanlaiset yhtälöt:

$$\begin{aligned}
\log\left(\frac{p(\text{imetysluokka}=2)}{p(\text{imetysluokka}=1)}\right) = \quad & 0.580 - 0.248 \cdot \text{tyttö} \\
& + 0.196 \cdot \text{sektio} - 0.170 \cdot \text{korkea koulutus} \\
& + 0.226 \cdot \text{ylipainoinen} \\
& - 0.024 \cdot \text{korkea asuinalueen huono-osaisuus} \\
& - 0.000 \cdot \text{äidin ikä}
\end{aligned} \tag{5}$$

$$\begin{aligned}
\log\left(\frac{p(\text{imetysluokka}=3)}{p(\text{imetysluokka}=1)}\right) = & - 1.834 - 0.179 \cdot \text{tyttö} \\
& + 1.446 \cdot \text{sektio} - 1.459 \cdot \text{korkea koulutus} \\
& + 0.640 \cdot \text{ylipainoinen} \\
& + 0.196 \cdot \text{korkea asuinalueen huono-osaisuus} \\
& - 0.000 \cdot \text{äidin ikä}.
\end{aligned} \tag{6}$$

Taulukkoon 17 on kerätty multinomiaalisen logistisen regression p-arvot imetysluokkien mukaan. Taulukosta nähdään, että synnytystapa ja koulutus ovat tilastollisesti merkitseviä tekijöitä varhaisravitsemuksessa. Sukupuolel-

la, painoindeksillä, asuinalueen huono-osaisuudella ja äidin iällä ei ole tilastollisesti merkitsevää yhteyttä varhaisravitsemukseen.

Yhtälöistä (5) ja (6) saadaan α ja $\boldsymbol{\beta}$ parametrit. Näiden avulla saadaan laskettua OR-luku kaavalla OR $= e^{\boldsymbol{\beta}}$. Taulukossa 18 on kuvattuna nämä suhdeluvut ja niiden 95 % luottamusvälit.

Verrattaessa imetysluokkaa kaksi imetysluokkaan yksi, mikään selittävistä tekijöistä ei ole tilastollisesti merkitsevä. Näiden kahden luokan välillä ei siten ole selkeää eroa, kun taas verrattaessa imetysluokkaa kolme imetysluokkaan yksi, tilastollisesti merkitseviä selittäviä tekijöitä ovat synnytystapa ja äidin koulutus. Näiltä selittäviltä tekijöiltä saadaan seuraavia tuloksia. Verrattuna alateitse syntyneisiin lapsiin, sektiolla syntyneillä lapsilla on 4.2 kertainen riski kuulua imetysluokkaan kolme kuin imetysluokkaan yksi. Tässä kuitenkin on suuri luottamusväli (1.908 - 9.460), joten tulos ei ole kovinkaan tarkka. Tuloksen huono tarkkuus voi johtua imetysluokan kolme pienuudesta. Verrattuna matalasti koulutettujen äitien lapsiin korkeasti koulutettujen äitien lapsilla on 0.2 kertainen riski kuulua imetysluokkaan kolme verrattuna imetysluokkaan yksi. Kiinnostavan muuttujan, asuinalueen huono-osaisuuden, p-arvot ovat 0.883 ja 0.603, joten se ei ole merkitsevä tekijä kummassakaan imetysluokassa.

Multinomiaalisen logistisen regression antamat tulokset ovat samassa linjassa taulukoista 12 ja 13 tehtyjen päätelmien kanssa.

p-arvot	Sukupuoli ref. poika	Synnytystapa ref. alatie	Koulutus ref. matala	BMI ref. normaali	Asuinalueen huono-osaisuus ref. matala	Äidin ikä
Imetysluokka 2	0.081	0.416	0.279	0.163	0.883	0.983
Imetysluokka 3	0.609	< 0.001	< 0.001	0.076	0.603	0.997

Taulukko 17: Multinomiaalisen logistisen regression kertoimien p-arvot esitettynä imetysluokkien mukaan. Referenssiluokkana käytetty imetysluokkaa 1 eli yli neljä kuukautta imetettyjä. Painoindeksi on tässä lyhennetty BMI.

	Imetysluokka 2		Imetysluokka 3	
	OR-luku	95% luottamusväli	OR-luku	95% luottamusväli
Sukupuoli ref. poika	0.780	0.590 - 1.031	0.836	0.422 - 1.658
Synnytystapa ref. alatie	1.216	0.759 - 1.949	4.248	1.908 - 9.460
Koulutus ref. matala	0.843	0.620 - 1.148	0.233	0.112 - 0.482
Painoindeksi ref. normaali	1.254	0.913 - 1.722	1.896	0.936 - 3.841
Asuinalueen huono-osaisuus ref. matala	0.976	0.711 - 1.340	1.216	0.582 - 2.541
Äidin ikä	1.000	0.967 - 1.033	1.000	0.929 - 1.076

Taulukko 18: Multinomiaalisen logistisen regression kerroinsuhdeluvut ja niiden 95% luottamusvälit esitettynä imetysluokkien mukaan. Referenssiluokkana käytetty imetysluokkaa yksi eli yli neljä kuukautta imetettyjä.

6 Yhteenveto

Tämän tutkielman tutkimuskysymyksenä on sosiaalisen elinympäristön yhteys imetykseen. Tutkielmassa tutkitaan, onko lapsen sukupuolella, synnytystavalla, äidin koulutuksella tai painoindeksillä, asuinalueen huono-osaisuudella tai äidin iällä vaikutusta imetykseen ja sen kestoon. Tutkimus suoritettiin multinomiaalisella logistisella regressiolla ja analyysit tehtiin R-ohjelmistolla pitäen merkitsevyyden rajana p-arvoa 0.05.

Tutkielman alussa käytiin läpi erilaisia menetelmiä, joita tarvittiin multinomiaalisen logistisen regression oletusten tarkastamiseen. Yksi keskeisimpiä käsitteitä logistisessa regressiossa on kerroinsuhde, joten se käsiteltiin erillisenä alakohtana. Logistisesta regressiosta käytiin läpi binäärinen logistinen regressio ja siitä laajennettiin multinomiaaliseen logistiseen regressioon. Mallille tehtiin diagnostinen tarkastelu Hosmerin ja Lemeshowin menetelmällä ja kaikki mallin vaatimukset täyttyivät. Malli siis kuvaa aineistoa hyvin ja on sopiva käytettäväksi.

Tutkimuksen tuloksina selvisi, että suurimmat eroavaisuudet ilmenivät ei imetettyjen ja imetettyjen välillä. Näiden ryhmien välillä merkitseviä tekijöitä olivat synnytystapa sekä äidin koulutus. Sektiolla syntyneet lapset kuuluvat todennäköisemmin ei imetettyjen luokkaan kuin alateitse syntyneet lapset. Tulos ei ole kuitenkaan kovinkaan tarkka. Matalasti koulutettujen äitien lapset kuuluvat todennäköisemmin ei imetettyjen luokkaan kuin korkeasti koulutettujen äitien lapset.

Muissa imetykseen liittyvissä tutkimuksissa on päädytty samankaltaisiin tuloksiin. Tutkimuksissa äidin koulutus on noussut esille merkittävänä imetykseen vaikuttavana tekijänä samoin kuin tässä tutkielmassa havainnoiduin tavoin eli korkeammin koulutetut äidit imettävät lapsiaan todennäköisemmin kuin matalasti koulutetut äidit [45–48, 57, 58]. Myös synnytystavalla on todettu olevan vaikutusta imetykseen, sillä alateitse synnytettyjä lapsia imetetään todennäköisemmin kuin sektiolla synnytettyjä [47, 58]. Äidin iästä ja painoindeksistä on sen sijaan saatu vaihtelevia tuloksia. Joissain tutkimuksissa imetyksen ja äidin iän välillä on löydetty vaikutusta, mutta toisissa tutkimuksissa ei vaikutusta ole havaittu [45–48, 57, 58]. Asuinalueen huono-osaisuudella on löydetty muissa tutkimuksissa olevan vaikutusta lapsen imetykseen [47,48,57]. Samoissa tutkimuksissa on lisäksi todettu, että äidin koulutuksella on ollut vahvempi vaikutus imetykseen kuin alueen huono-osaisuudella [47, 48, 57]. Asuinalueen huono-osaisuus on vaikuttanut imetykseen negatiivisesti eli suurilla asuinalueen huono-osaisuusarvoilla on todennäköisempää imettää vähemmän kuin matalilla asuinalueen huono-osaisuuden alueilla [47, 48, 57]. Tämän tutkielman tulokset ovat näin ollen linjassa aikaisemmin tehtyjen tutkimusten kanssa.

Koska imetyksellä on todettu olevan paljon terveyttä edistäviä vaikutuksia niin lapselle kuin äidillekin, on mielekästä tutkia imetykseen vaikuttavia tekijöitä. Tällaisella imetystä koskevalla tutkimustiedolla voidaan edistää imettämistä, koska tunnistettujen imetykseen vaikuttavien tekijöiden avulla voidaan puuttua havaittuihin epäkohtiin. Synnytystavalla ja koulutuksella oli tilastollista merkitystä imetykseen. Koska sektiolla syntyneillä lapsilla on suurempi riski siihen, ettei heitä imetetty, jatkossa voitaisiin tutkia sitä, mistä tämä johtuu ja miten asiaan voitaisiin reagoida. Tutkimuksissa todetun mukaisesti korkeammin koulutetut äidit imettävät lapsiaan todennäköisemmin kuin matalammin koulutetut, joten jatkossa voitaisiin keskittyä informoimaan matalammin koulutettuja äitejä imetyksen merkityksestä.

Multinomiaalinen logistinen regressio valittiin menetelmäksi tutkia imetysmuuttujia ja niihin vaikuttavia tekijöitä, koska sen avulla voidaan tutkia kategorisia muuttujia, joiden vastearvot eivät tarvitse olla tietyn etäisyyden päässä toisistaan. Muita imetykseen liittyviä tutkimuksia analysoitaessa nousi esiin, että myös muutamista niistä oli päädytty käyttämään logistista regressiota [46, 48, 57].

Tästä tutkielmasta voisi jatkaa vaikka tarkastelemalla muissa tutkimuksissa esiin nousseita imetykseen vaikuttavia tekijöitä kuten äidin tupakointi. Suurin ero on havaittu olevan imetettyjen ja ei imetettyjen välillä. Näiden välisiä eroja olisi kiinnostava tutkia jatkossa. Muissa tutkimuksissa nousi esiin myös erilainen jaottelu imetettyjen välille. Esimerkiksi imetettyjen välinen luokittelu voitaisiin tehdä alle kuusi kuukautta imetettyjen ja sen yli imetettyjen välille neljän kuukauden sijaan. Tässä voisi nousta esiin suurempia eroavaisuuksia kuin mitä nyt nousi yli neljä kuukautta imetettyjen ja yli kuukauden, mutta alle neljä kuukautta imetettyjen välille.

Lähteet

[1] *Riippuva muuttuja.* Tilastokeskus, https://www.stat.fi/meta/kas/riippuva_muuttu.html. (Luettu 3.5.2022).

[2] Tilastotieteen ja todennäköisyyslaskennan englantilais–suomalainen sanasto, http://www.math.helsinki.fi/petrin/sanasto/tilastosanasto.html. (Luettu 17.05.2021).

[3] *Riippumaton muuttuja.* Tilastokeskus, https://www.stat.fi/meta/kas/riippumaton_muu.html. (Luettu 3.5.2022).

[4] *Käsitteet.* Tilastokeskus, https://www.stat.fi/meta/kas/index.html. (Luettu 15.05.2021).

[5] *Parametri.* Tilastokeskus, https://www.stat.fi/meta/kas/parametri.html. (Luettu 10.12.2021).

[6] Vehkalahti, Kimmo: *Tilastotieteen johdantokurssi, Teema 8: Parametrien estimointi ja luottamusvälit.* Helsingin yliopisto, 2009.

[7] *Tilastojen ABC, Hajonnan kuvaaminen.* Tilastokeskus, https://tilastokoulu.stat.fi/verkkokoulu_v2.xql?page_type=sisalto&course_id=tkoulu_tlkt&lesson_id=4&subject_id=5. (Luettu 8.12.2021).

[8] Liski, Erkki: *Matemaattinen tilastotiede.* Matematiikan, Tilastotieteen ja Filosofian Laitos, Tampereen Yliopisto, 2005.

[9] Vehkalahti, Kimmo: *Teema 9: Tilastollinen merkitsevyystestaus.* Tilastotieteen johdantokurssi, Helsingin yliopisto, 2009. https://wiki.helsinki.fi/download/attachments/314379Ft/teema9.pdf. (Luettu 8.9.2021).

[10] *Hypoteesin testaus.* KvantiMOTV, https://www.fsd.tuni.fi/menetelmaopetus/hypoteesi/testaus.html. (Luettu 7.12.2021).

[11] Gibson, Jason: *The Student t-Distribution.* SAGE Research Methods Video, julkaistu 2014. https://methods.sagepub.com/video/the-student-t-distribution. (Luettu 13.12.2021).

[12] Metsämuuronen, Jari: *Tilastollisen päättelyn perusteet.* Helsinki: Methelp, 2000.

[13] Knapp, Herschel: *An Introduction to the t-Test.* SAGE Research Methods Video, julkaistu 2017. https://methods.sagepub.com/video/an-introduction-to-the-t-test. (Luettu 15.12.2021).

[14] Hurme, Saija: *Tilastolliset perustestit.* Turun yliopisto, 2015.

[15] Chieh, Chew Jian: *Making sense of the two-sample t-test.* iSixSigma, https://www.isixsigma.com/tools-templates/hypothesis-testing/making-sense-two-sample-t-test/. (Luettu 25.2.2022).

[16] Holopainen, Martti ja Pulkkinen, Pekka: *Tilastolliset menetelmät.* WSOY Oppimateriaalit Oy, Porvoo, 2008.

[17] Roterman-Konieczna, Irena: *Statistics by Prescription.* Jagiellonian University Press, 2009.

[18] Gastwirth, Joseph L., Gel, Yulia R. and Miao, Weiwen: *The Impact of Levene's Test of Equalityof Variances on Statistical Theory and Practice.* Statistical Science 2009. Vol. 24, No. 3, s: 343–360.

[19] Math and Science: *Lesson 1 - What is the F-Distribution in Statistics?.* Youtube-videopalvelu, julkaistu 15.2.2017. https://www.youtube.com/watch?v=S8VzUYJjBmw. (Luettu 18.5.2022).

[20] Heiberger, Richard M. and Burt. Holland: *Statistical Analysis and Data Display: An Intermediate Course with Examples in S-Plus, R and SAS.* New York: Springer, 2004.

[21] McHugh, Mary L.: *The Chi-square test of independence.* Biochemia Medica, 2013. Vol. 23 , No. 2, s: 143–149.

[22] Dytham, Calvin: *Choosing and Using Statistics : A Biologist's Guide.* John Wiley & Sons, Incorporated, 2011.

[23] Frey, Bruce B.: *The Sage Encyclopedia of Educational Research, Measurement, and Evaluation.* Thousand Oaks, California: SAGE Publications, Inc., 2018.

[24] Kleinbaum, David: *Class 11: Goodness of Fit- Deviance, Hosmer-Lemeshow statistic.* Youtube-videopalvelu, julkaistu 22.11.2017. https://www.youtube.com/watch?v=cH0Hj69eT-8. (Luettu 18.12.2021).

[25] Fagerland, Morten W. and Hosmer, David W.: *A generalized Hosmer–Lemeshow goodness-of-fit test for multinomial logistic regression models*. The Stata Journal, 2012. Vol. 12, No. 3, s: 447–453.

[26] Hosmer Jr., David W. and Lemeshow, Stanley and Sturdivant, Rodney X.: *Applied Logistic Regression*. John Wiley & Sons, Incorporated, Hoboken, New Jersey, 2013.

[27] Sarna, Seppo: *Kliinisen biostatistiikan jatkokurssi*. Helsingin yliopisto, 2012.

[28] Koppinen, Markku: *Lineaarialgebra: Osa 1*. Turun yliopisto, 2006.

[29] McCullagh, P. and Nelder, J. A.: *Generalized Linear Models*. CRC Press LLC, 1989.

[30] Nyblom, Jukka: *Yleistetyt lineaariset mallit*. Jyväskylän yliopisto, Matematiikan ja tilastotieteen laitos, 2015.

[31] Agresti, Alan: *Categorical Data Analysis*. John Wiley & Sons, Incorporated, Hoboken, New Jersey, 2012.

[32] *Logistinen regressio*. KvantiMOTV, `https://www.fsd.tuni.fi/menetelmaopetus/logregressio/logistinen.html`. (Luettu 18.11.2020).

[33] Hilbe, Joseph M.: *Logistic Regression Models*. Boca Raton, Florida: CRC Press, 2009.

[34] Rowe, Philip: *Essential Statistics for the Pharmaceutical Sciences*. John Wiley & Sons, Incorporated, 2015.

[35] The Data Science Show: *What is a Logistic Regression Model? The Mathematical Principle*. Youtube-videopalvelu, julkaistu 26.8.2017. `https://www.youtube.com/watch?v=8BoTI1UV_O8`. (Luettu 13.12.2021).

[36] The NCCMT: *Odds Ratios*. Youtube-videopalvelu, julkaistu 5.7.2016. `https://www.youtube.com/watch?v=5zPSD_e_N04`. (Luettu 25.11.2021).

[37] Hoffman, Julien: *Biostatistics for Medical and Biomedical Practitioners*. Elsevier Science & Technology, 2015.

[38] Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd edition. Stanford, California, 2008.

[39] Linear Algebra: *Analysis of Discrete Data Lesson 8: Multinomial Logistic Regression Part 1*. Youtube-videopalvelu, julkaistu 26.2.2016. `https://www.youtube.com/watch?v=7w_xf1uUQP4`. (Luettu 24.11.2021).

[40] Dunn, Peter K. and Smyth, Gordon K.: *Generalized Linear Models With Examples in R*, Springer Texts in Statistics. Springer, New York, 2018.

[41] Sairinen, Rauno, Manninen, Rikhard, Peltonen, Lasse ja Wiik, Maarit: *Ympäristöterveys yhdyskuntasuunnittelussa, Näkökulmia hyvinvointia edistävään elinympäristöön*. Suomen ympäristö, Ympäristöministeriö, 2006. Vol. 13, s: 1-71.

[42] Salone, Lindsey Rennick, Vann Jr., William F. and Dee, Deborah L.: *Breastfeeding: An overview of oral and general health benefits*. The Journal of the American Dental Association, 2013. Vol. 144, No. 2, s: 143-151.

[43] Labbok, Miriam H.: *Effects of Breastfeeding on the Mother*. Pediatric Clinics of North America, 2001. Vol. 48, No. 1, s: 143-158.

[44] Binns, Colin, Lee, MiKyung and Low, Wah Yun: *The Long-Term Public Health Benefits of Breastfeeding*. Asia-Pacific Journal of Public Health, 2016. Vol. 28, No. 1, s: 7-14.

[45] Kehler, Heather L., Chaput, Katie H. and Tough, Suzanne C.: *Risk Factors for Cessation of Breastfeeding Prior to Six Months Postpartum among a Community Sample of Women in Calgary, Alberta*. Can J Public Health, 2009. Vol. 100, No. 5, s: 376–380.

[46] Evers, Susan, Doran, Lori and Schellenberg, Kathryn: *Influences on Breastfeeding Rates in Low Income Communities in Ontario*. Canadian Journal of Public Health, 1998. Vol. 89, s: 203–207.

[47] Smith, D.P.: *Breastfeeding in the United States*. Soc Biol. Spring-Summer, 1985. Vol. 32, No. 1-2, s: 53-60.

[48] Heck, Katherine E., Braveman, Paula, Cubbin, Catherine, Chávez, Gilberto F. and Kiely, John L.: *Socioeconomic Status and Breastfeeding Initiation among California Mothers*. Public Health Reports, 2006. Vol. 121, No. 1, s: 51-59.

[49] Anstey, Erica H., Chen, Jian, Elam-Evans, Laurie D. and Perrine, Cria G.: *Racial and Geographic Differences in Breastfeeding — United States, 2011–2015*. US Department of Health and Human Services/Centers for Disease Control and Prevention, Morbidity and Mortality Weekly Report, 2017. Vol. 66 , No. 27, s: 723 - 727.

[50] *Sosioekonominen asema.* Tilastokeskus, https://www.tilastokeskus.fi/meta/kas/sosioekon_asema.html#tab3 (Luettu 13.12.2021).

[51] Lagström, Hanna, Halonen, Jaana I., Kawachi, Ichiro, Stenholm, Sari, Pentti, Jaana, Suominen, Sakari, Kivimäki, Mika ja Vahtera, Jussi: *Neighborhood socioeconomic status and adherence to dietary recommendations among Finnish adults: A retrospective follow-up study.* Health & Place, 2019. Vol. 55, s: 43-50.

[52] Halonen, Jaana I, Pulakka, Anna, Pentti, Jaana,Kallio, Minna, Koskela, Sofia, Kivimäki, Mika, Kawachi, Ichiro, Vahtera, Jussi ja Stenholm, Sari: *Cross-sectional associations of neighbourhood socioeconomic disadvantage and greenness with accelerometer-measured leisure-time physical activity in a cohort of ageing workers.* BMJ Open, 2020. Vol. 10, No. 8, s: 1-9.

[53] *Cohort.* National Cancer Institute, https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cohort (Luettu 10.12.2021).

[54] Lagström, Hanna, Rautava, Päivi, Kaljonen, Anne, Räihä, Hannele, Pihlaja, Päivi, Korpilahti, Pirjo, Peltola, Ville, Rautakoski, Pirkko, Österbacka, Eva, Simell, Olli ja Niemi, Pekka: *Cohort Profile: Steps to the Healthy Development and Well-being of Children (the STEPS Study).* International Journal of Epidemiology, 2013. Vol 42, No.5. s: 1273–1284.

[55] Osborne, Jason: *Best practices in logistic regression.* SAGE Publications, 2015.

[56] Prabasaj, Paul P., Pennell, M.L. and Lemeshow, S.: *Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets.* Statistics in Medicine, 2012. Vol. 32, s: 67-80.

[57] Skafida, V.: *The relative importance of social class and maternal education for breast-feeding initiation.* Public Health Nutrition, 2009. Vol. 12, No. 12, s: 2285-2292.

[58] Cohen, Sarah S., Alexander, Dominik D., Krebs, Nancy F., Young, Bridget E., Cabana, Michael D., Erdmann, Peter, Hays, Nicholas P., Bezold, Carla P., Levin-Sparenberg, Elizabeth, Turini, Marco and Saavedra, Jose M.: *Factors Associated with Breastfeeding Initiation and Continuation: A Meta-Analysis.* The Journal of Pediatrics, 2018. Vol. 203, s: 190-196.

[59] Mellin, Ilkka: *Tilastolliset taulukot*. https://math.aalto.fi/opetus/ms-a0502/luennot14/Ilkka-Mellinin-tilastolliset-taulukot.pdf. (Luettu 11.5.2022).

Liite A Soveltavan osion R-koodi

```
#Käytetyt kirjastot
library(car)
library(rcompanion)
library(dplyr)
library(nnet)
library("generalhoslem")
library(statmod)
#-----------------------------------------------------------------------
#Haetaan aineisto, karsitaan siitä puutteeliset rivit pois
ja tehdään aineiston jaottelu kolmeen luokkaan.
#Haetaan data
kokeilu <- read.csv("C:/Users/Satu/Documents/Opinnot/Gradu/
Aineisto/Jantunen_10_5_2021_piste.csv", header=TRUE)

#Otetaan pois ne rivit, joilla ei ole mitään tietoja imetyksestä
oma_data <- subset(kokeilu, !(kokeilu$kokonaisimetys ==" " &
kokeilu$osittaisimetys==" " & kokeilu$taysimetys_laskennallinen==" "
& kokeilu$eiImetetty ==" "))

#Jaotellaan aineisto kolmeen imetysluokkaan
for(i in 1:length(oma_data$nro)){
  #Luokkaan 1 ne lapset, joilla täysimetys on kestänyt 4kk tai yli
  if(oma_data$taysimetys_laskennallinen[i] >= 4){
    oma_data$imetys_muuttuja[i] <- 1
  }
  #Luokkaaan 2 ne lapset, joita on täysimetetty ja kokonaisimetys on
  kestänyt yli kuukauden
  else if(oma_data$taysimetys_laskennallinen[i] < 4 &
   oma_data$taysimetys_laskennallinen[i] !=" "
   & (oma_data$kokonaisimetys[i] > 1 ||
   oma_data$taysimetys_laskennallinen[i] > 1)){
    oma_data$imetys_muuttuja[i] <- 2
  }
  #Luokkaan 3 ne lapset, joita ei ole imetetty tai imetys on kestänyt
  alle 1 kk
  else if( oma_data$eiImetetty[i] == 1 ||
  (oma_data$kokonaisimetys[i] <= 1 & oma_data$kokonaisimetys[i] !=" ")){
    oma_data$imetys_muuttuja[i] <- 3
  }
```

```
  #Katsotaan vielä että kaikki lapset päätyvät johonkin ryhmään.
  Eli arvoja 0 ei pitäisi taulusta löytyä
  else {
    oma_data$imetys_muuttuja[i] <- 0
  }
}

#---------------------------------------------------------------------

#Tutkitaan aineistoa ja poistetaan rivit, jotka eivät täytä vaatimuksia

#Piiraskaavio imetysluokista
mytable <- table(oma_data$imetys_muuttuja)
lbls <- paste(names(mytable), "\n", mytable, sep="")
pie(mytable, labels = lbls, main="Piiraskaavio")

#Tulostetaan näkyviin rivit, jotka kuuluvat luokkaan 0, jotta saadaan
katsottua jäikö sinne rivejä joita olisi voitu laittaa edeltäviin
luokkiin
oma_data[oma_data$imetys_muuttuja == 0, 1:19]

#Otetaan valmiiseen dataan vain ne rivit, jotka kuuluvat luokkaan
1,2 tai 3
aineisto <- subset(oma_data, (oma_data$imetys_muuttuja != 0))

#Poistetaan aineistosta kaksoset
aineisto <- subset(aineisto, (aineisto$nro <= 12212))

#Poistetaan aineistosta keskoset
aineisto <- subset(aineisto, (aineisto$Raskkesto > 258))

#Tulostetaan näkyviin ne rivit, joissa sarake "kuollut" on yksi
aineisto[aineisto$kuollut == 1, 1:19]

#Piiraskaavio imetysluokista ilman luokkaa 0
taulukko <- table(aineisto$imetys_muuttuja)
luvut <- paste(names(taulukko), "\n", taulukko, sep="")
pie(taulukko,labels = luvut, main = "Imetysmuuttujat")
#---------------------------------------------------------------------

#Muokataan muuttujat muotoon (0,1)
```

```
#Muutetaan vielä kaikki tekijät muodosta (1,2) muotoon (0,1)
for(i in 1:length(aineisto$nro)){
  if(aineisto$SP[i] == 1){
    aineisto$SP[i] <- 0
  } else{
    aineisto$SP[i] <- 1
  }
}

for(i in 1:length(aineisto$nro)){
  if(aineisto$SYNNYTYSTAPA_2luok[i] == 1){
    aineisto$SYNNYTYSTAPA_2luok[i] <- 0
  }
  else{
    aineisto$SYNNYTYSTAPA_2luok[i] <- 1
  }
}

#-----------------------------------------------------------------------

#Muodostetaan skaalatatut muuttujat

#Luodaan 20. sarake, joka kertoo onko koulutuksen taso matala vai korkea
for(i in 1:length(aineisto$nro)){
  if(aineisto$aammkoul[i] <= 4 & aineisto$aammkoul[i]!="."){
    aineisto$koulutus[i] <- 0
  }
  else if(aineisto$aammkoul[i] >= 5 & aineisto$aammkoul[i] < 9){
    aineisto$koulutus[i] <- 1
  }
  #Muutetaan äidin koulutus luokka 9 numerot oikeiksi
  else if(aineisto$aammkoul[i] == 9){
    if(aineisto$nro[i] == 217 || aineisto$nro[i] == 750 ||
        aineisto$nro[i] == 751 || aineisto$nro[i] == 1626){
      aineisto$koulutus[i] <- 1
    }
    else if(aineisto$nro[i] == 302 || aineisto$nro[i] == 797 ||
            aineisto$nro[i] == 1110 || aineisto$nro[i] == 1283 ||
            aineisto$nro[i] == 1323 || aineisto$nro[i] == 1347){
      aineisto$koulutus[i] <- 0
```

```
    }
    else{
      aineisto$koulutus[i] <- "."
    }
  }
  else{
    aineisto$koulutus[i] <- "."
  }
}

#Luodaan 21. sarake, joka kertoo onko äiti BMI:n mukaan ylipainoinen
 vai ei
for(j in 1:length(aineisto$nro)){
  if(aineisto$BMI_aiti[j] > 25){
    aineisto$BMI_raja[j] <- 1
  }
  else if(aineisto$BMI_aiti[j] == "."){
    aineisto$BMI_raja[j] <- "."
  }
  else{
    aineisto$BMI_raja[j] <- 0
  }
}

#Luodaan 22. sarake, joka kertoo kuuluuko perhe disadv alueelle low
vai high
for (k in 1:length(aineisto$nro)) {
  if(aineisto$disadv[k] > 0){
    aineisto$SES[k] <- 1
  }
  else if(aineisto$disadv[k] <= 0 & aineisto$disadv[k] != "."){
    aineisto$SES[k] <- 0
  }
  else{
    aineisto$SES[k] <- "."
  }
}

#-----------------------------------------------------------------------

#Tutkitaan muuttujien jakaumia
```

```
#Katsotaan aineistosta sukupuolten jakauma
addmargins(table(aineisto$SP, aineisto$imetys_muuttuja))
typeof(aineisto$SP)
#Katsotaan aineistosta synnytystavan jakauma
addmargins(table(aineisto$SYNNYTYSTAPA_2luok, aineisto$imetys_muuttuja))
typeof(aineisto$SYNNYTYSTAPA_2luok)
#Katsotaan aineistosta koulutuksen jakauma
addmargins(table(aineisto$koulutus, aineisto$imetys_muuttuja))
typeof(aineisto$koulutus)
#Katsotaan aineistosta BMI_raja jakauma
addmargins(table(aineisto$BMI_raja, aineisto$imetys_muuttuja))
typeof(aineisto$BMI_raja)
#Katsotaan aineistosta SES rajan jakauma
addmargins(table(aineisto$SES, aineisto$imetys_muuttuja))
typeof(aineisto$SES)

#Tutkitaan jatkuvaa muuttujaa äidin ikä
nrow(aineisto[aineisto$Aidinika != ".", 1:2])
quantile(as.numeric(aineisto$Aidinika[aineisto$imetys_muuttuja == 1]))
hist(as.numeric(aineisto$Aidinika[aineisto$imetys_muuttuja == 3]))

#Koko aineiston jakauma imetysmuutujan suhteen
addmargins(table(aineisto$imetys_muuttuja))
typeof(aineisto$imetys_muuttuja)


#-----------------------------------------------------------------------
#Tallennetaan aineisto

#Tallennetaan data
write.csv(aineisto, "C:/Users/Satu/Documents/Opinnot/Gradu/
                        Aineisto/Jantunen_20_8_21.csv")

#-----------------------------------------------------------------------
#Alustetaan riippuvuuksien tarkastelua

#Otetaan aineistosta vain halutut muuttujat
val <- subset(aineisto, select=c("SP", "SYNNYTYSTAPA_2luok",
              "imetys_muuttuja", "koulutus", "BMI_raja",
              "SES", "Aidinika"))
```

```
#Tiputetaan rivit, joissa puuttuvia arvoja
valittu_aineisto <- val[!(val$SP=="." | val$SYNNYTYSTAPA_2luok=="." |
                           val$imetys_muuttuja=="." | val$koulutus=="." |
                           val$BMI_raja=="." | val$SES=="." |
                           val$Aidinika=="."), ]


#Muutetaan muuttujien muoto factoriksi
valittu_aineisto$SP <- factor(valittu_aineisto$SP)
valittu_aineisto$SYNNYTYSTAPA_2luok <- factor(valittu_aineisto
                                              $SYNNYTYSTAPA_2luok)
valittu_aineisto$imetys_muuttuja <- factor(valittu_aineisto
                                           $imetys_muuttuja)
valittu_aineisto$koulutus <- factor(valittu_aineisto$koulutus)
valittu_aineisto$BMI_raja <- factor(valittu_aineisto$BMI_raja)
valittu_aineisto$SES <- factor(valittu_aineisto$SES)
options(digits = 10)
valittu_aineisto$Aidinika <- as.numeric(valittu_aineisto$Aidinika)


#Piirretään kaikista jatkuvista muuttujista histogrammit, jotta nähdään
onko ne normaalijakautuneita
hist(as.numeric(valittu_aineisto$imetys_muuttuja))
hist(valittu_aineisto$Aidinika) #Vaikuttaa normaalisti jakautuneelta
qqPlot(valittu_aineisto$Aidinika, ylab = " Äidin ikä")

#-----------------------------------------------------------------------
#Tutkitaan jatkuvan muuttujan ja kategorisen muuttujan riippuvuus
t-testillä

#Levene's testi korrelaatiot äidin ikään ja t-testit
#Jos p arvo on > 0.05, niin var samat, jos p-arvo <, niin var eri suuret
leveneTest(valittu_aineisto$Aidinika ~ valittu_aineisto$SES,
           valittu_aineisto)
t.test(valittu_aineisto$Aidinika ~ valittu_aineisto$SES,
       var.equal = TRUE)

leveneTest(valittu_aineisto$Aidinika ~ valittu_aineisto$BMI_raja,
           valittu_aineisto)
```

```
t.test(valittu_aineisto$Aidinika ~ valittu_aineisto$BMI_raja,
       var.equal = TRUE)

leveneTest(valittu_aineisto$Aidinika ~ valittu_aineisto$koulutus,
           valittu_aineisto)
#Koska edellisen p-arvo < 0.05, niin var.equal=FALSE
t.test(valittu_aineisto$Aidinika ~ valittu_aineisto$koulutus,
       var.equal = FALSE)

leveneTest(valittu_aineisto$Aidinika ~ valittu_aineisto$SP,
           valittu_aineisto)
t.test(valittu_aineisto$Aidinika ~ valittu_aineisto$SP, var.equal = TRUE)

leveneTest(valittu_aineisto$Aidinika ~ valittu_aineisto$SYNNYTYSTAPA_2luok,
           valittu_aineisto)
t.test(valittu_aineisto$Aidinika ~ valittu_aineisto$SYNNYTYSTAPA_2luok,
       var.equal = TRUE)


#-------------------------------------------------------------------------
#Tutkitaan kategoristen muuttujien välisiä suhteita khii toiseen -testillä

#Khii toiseen -testillä korrelaatiot muihin
#Tutkitaan taulukoimalla arvot, ettei niissä ole nollasoluja
table(valittu_aineisto$SES, valittu_aineisto$BMI_raja)
chisq.test(valittu_aineisto$SES, valittu_aineisto$BMI_raja)
table(valittu_aineisto$SES, valittu_aineisto$koulutus)
chisq.test(valittu_aineisto$SES, valittu_aineisto$koulutus)
table(valittu_aineisto$SES, valittu_aineisto$SP)
chisq.test(valittu_aineisto$SES, valittu_aineisto$SP)
table(valittu_aineisto$SES, valittu_aineisto$SYNNYTYSTAPA_2luok)
chisq.test(valittu_aineisto$SES, valittu_aineisto$SYNNYTYSTAPA_2luok)

table(valittu_aineisto$BMI_raja, valittu_aineisto$koulutus)
chisq.test(valittu_aineisto$BMI_raja, valittu_aineisto$koulutus)
table(valittu_aineisto$BMI_raja, valittu_aineisto$SP)
chisq.test(valittu_aineisto$BMI_raja, valittu_aineisto$SP)
table(valittu_aineisto$BMI_raja, valittu_aineisto$SYNNYTYSTAPA_2luok)
chisq.test(valittu_aineisto$BMI_raja, valittu_aineisto$SYNNYTYSTAPA_2luok)

table(valittu_aineisto$koulutus, valittu_aineisto$SP)
```

```
chisq.test(valittu_aineisto$koulutus, valittu_aineisto$SP)
table(valittu_aineisto$koulutus, valittu_aineisto$SYNNYTYSTAPA_2luok)
chisq.test(valittu_aineisto$koulutus, valittu_aineisto
                                      $SYNNYTYSTAPA_2luok)

table(valittu_aineisto$SP, valittu_aineisto$SYNNYTYSTAPA_2luok)
chisq.test(valittu_aineisto$SP, valittu_aineisto$SYNNYTYSTAPA_2luok)


#---------------------------------------------------------------------
#Tarkistetaan saatujen riippuvuuksien suuruudet

#Tarkistetaan cramerin v:llä, arvo < 0.7
cramerV(valittu_aineisto$SES, valittu_aineisto$BMI_raja)
cramerV(valittu_aineisto$SES, valittu_aineisto$koulutus)
cramerV(valittu_aineisto$BMI_raja, valittu_aineisto$koulutus)



#Katso histogrammilla iän ja sessin eroavaisuudet
plot(valittu_aineisto$SES, valittu_aineisto$Aidinika,
     xlab = "Asuinalueen huono-osaisuus", ylab = "Äidin ikä")

#---------------------------------------------------------------------
#Multinomiaalinen logistinen regressio


#Frekvenssit
addmargins(table(valittu_aineisto$SP, valittu_aineisto$imetys))
addmargins(table(valittu_aineisto$SYNNYTYSTAPA_2luok,
                 valittu_aineisto$imetys))
addmargins(table(valittu_aineisto$koulutus, valittu_aineisto$imetys))
addmargins(table(valittu_aineisto$BMI_raja, valittu_aineisto$imetys))
addmargins(table(valittu_aineisto$SES, valittu_aineisto$imetys))
quantile(valittu_aineisto$Aidinika[valittu_aineisto$imetys == 1])
quantile(valittu_aineisto$Aidinika[valittu_aineisto$imetys == 2])
quantile(valittu_aineisto$Aidinika[valittu_aineisto$imetys == 3])
quantile(valittu_aineisto$Aidinika)
table(valittu_aineisto$imetys)
```

```
#Multinomiaalinen logistinen regressio
#Vertailuryhmäksi valitaan ryhmä 1
valittu_aineisto$imetys2 <- relevel(valittu_aineisto$imetys, ref = "1")
malli <- multinom(imetys2 ~ SP + SYNNYTYSTAPA_2luok + koulutus + BMI_raja
                  + SES + Aidinika, data = valittu_aineisto)

#Mallin tarkastelu
summary(malli)

#Kaksisuuntainen Z-testi
z <- summary(malli)$coefficients/summary(malli)$standard.errors
z
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p

#Taulukoidaan saadut tulokset
output <- summary(malli)
Pimetys2 <- rbind(output$coefficients[1,],output$standard.errors[1,],
                  z[1,],p[1,])
rownames(Pimetys2) <- c("Coefficient","Std. Errors","z stat","p value")
knitr::kable(Pimetys2)

Pimetys3 <- rbind(output$coefficients[2,],output$standard.errors[2,],
                  z[2,],p[2,])
rownames(Pimetys3) <- c("Coefficient","Std. Errors","z stat","p value")
knitr::kable(Pimetys3)

#Kerroinsuhde
exp(output$coefficients[1,])
exp(output$coefficients[2,])


#Kerroinsuhteen 95% luottamusväli
exp(confint(malli, level = 0.95))

#-------------------------------------------------------------------------
#Diangnostiikka

# 1) STANDARDOIDUT RESIDUAALIT
#Tutkitaan standardoidut residuaalit erikseen 1 vs. 2, 1 vs. 3 ja 2 vs. 3
#Tehdään logistiset regressiot kaikista malleista ja tutkitaan niiden
```

```
diagnostiikka. Näin toimitaan, koska multinomiaalisesta mallista on
erittäin hankalaa tutkia diagnostiikkaa

#Valitaan aineistoon vain ne sarakkeet mitä halutaan tutkia
valittu1 <- subset(aineisto, select=c("nro", "SP", "SYNNYTYSTAPA_2luok",
                    "imetys_muuttuja", "koulutus", "BMI_raja", "SES",
                     "Aidinika"))
#Poistetaan rivit, joissa puuttuvia arvoja
valittu <- valittu1[!(valittu1$SP=="." | valittu1$SYNNYTYSTAPA_2luok=="." |
                      valittu1$imetys_muuttuja=="." | valittu1$koulutus=="."
                      | valittu1$BMI_raja=="." | valittu1$SES="." |
                        valittu1$Aidinika="."), ]
#Muodostetaan muuttujista factorit
valittu$SP <- factor(valittu$SP)
valittu$SYNNYTYSTAPA_2luok <- factor(valittu$SYNNYTYSTAPA_2luok)
valittu$imetys_muuttuja <- factor(valittu$imetys_muuttuja)
valittu$koulutus <- factor(valittu$koulutus)
valittu$SES <- factor(valittu$BMI_raja)
options(digits = 10)
valittu$Aidinika <- as.numeric(valittu$Aidinika)


#Imetysmuuttuja 1 vs 2
apuAineisto1 <- valittu[valittu$imetys_muuttuja == 1 |
                        valittu$imetys_muuttuja == 2,]
table(apuAineisto1$imetys_muuttuja)
logreg1 <- glm(imetys_muuttuja ~ SP + SYNNYTYSTAPA_2luok + koulutus +
               BMI_raja + SES + Aidinika, data = apuAineisto1,
               family=binomial)

#Standardoidut residuaalit
#Käytetään quantile residuals, koska tämä on jatkuva, vaikka funktio
olisi diskreetti
layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
rQ <- qresid( logreg1 )
rQ.std <- rQ / sqrt( 1 - hatvalues(logreg1) )
plot(apuAineisto1$nro, rQ.std, xlab = "Lapsen numero",
     ylab = "Standardoitu residuaali")
     + abline(h = 0, col = "red")
```

```
#Imetysmuuttuja 1 vs 3
apuAineisto2 <- valittu[!(valittu$imetys_muuttuja == "2"),]
table(apuAineisto2$imetys_muuttuja)
logreg2 <- glm(imetys_muuttuja ~ SP + SYNNYTYSTAPA_2luok + koulutus
               + BMI_raja + SES + Aidinika, data = apuAineisto2,
                family = binomial)
summary(logreg2)

#Standardoidut residuaalit
rQ2 <- qresid( logreg2 )
rQ2.std <- rQ2 / sqrt( 1 - hatvalues(logreg2) )
plot(apuAineisto2$nro, rQ2.std, xlab = "Lapsen numero",
     ylab = "Standardoitu residuaali")
     + abline(h = 0, col = "red")


#Imetysmuuttuja 2 vs 3
apuAineisto3 <- valittu[!(valittu$imetys_muuttuja == "1"),]
table(apuAineisto3$imetys_muuttuja)
logreg3 <- glm(imetys_muuttuja ~ SP + SYNNYTYSTAPA_2luok + koulutus +
               BMI_raja + SES + Aidinika, data = apuAineisto3,
               family = binomial)

summary(logreg3)

#Standardoidut residuaalit
rQ3 <- qresid( logreg3 )
rQ3.std <- rQ3 / sqrt( 1 - hatvalues(logreg3) )
plot(apuAineisto3$nro, rQ3.std, xlab = "Lapsen numero",
     ylab = "Standardoitu residuaali")
     + abline(h = 0, col = "red")

layout(matrix(c(1,1,1,1), 1, 1, byrow = TRUE))


# 2) LINEAARINEN SUHDE
#Tarkistetaan muuttujien ja logistisen mallin tuloksen lineaarinen suhde.
Suhde lineaarinen, jos p-arvo > 0.05
hosmer <- logitgof(valittu_aineisto$imetys_muuttuja, fitted(malli))
hosmer
x <- hosmer$observed
```

```
y <- hosmer$expected
cbind(x,y[2:4])
```

Liite B Jakaumien taulukoidut arvot

$\chi^2(df)$-JAKAUMA
$\chi^2(df)$-DISTRIBUTION

Kriittisiä arvoja / Critical values

	Merkitsevyystaso 1-suuntaisissa testeissä / Sigficance level in 1-sided tests							
df	0.999	0.99	0.95	0.9	0.1	0.05	0.01	0.001
1	0.000	0.000	0.004	0.016	2.706	3.841	6.635	10.828
2	0.002	0.020	0.103	0.211	4.605	5.991	9.210	13.816
3	0.024	0.115	0.352	0.584	6.251	7.815	11.345	16.266
4	0.091	0.297	0.711	1.064	7.779	9.488	13.277	18.467
5	0.210	0.554	1.145	1.610	9.236	11.070	15.086	20.515
6	0.381	0.872	1.635	2.204	10.645	12.592	16.812	22.458
7	0.598	1.239	2.167	2.833	12.017	14.067	18.475	24.322
8	0.857	1.646	2.733	3.490	13.362	15.507	20.090	26.124
9	1.152	2.088	3.325	4.168	14.684	16.919	21.666	27.877
10	1.479	2.558	3.940	4.865	15.987	18.307	23.209	29.588
11	1.834	3.053	4.575	5.578	17.275	19.675	24.725	31.264
12	2.214	3.571	5.226	6.304	18.549	21.026	26.217	32.909
13	2.617	4.107	5.892	7.042	19.812	22.362	27.688	34.528
14	3.041	4.660	6.571	7.790	21.064	23.685	29.141	36.123
15	3.483	5.229	7.261	8.547	22.307	24.996	30.578	37.697
16	3.942	5.812	7.962	9.312	23.542	26.296	32.000	39.252
17	4.416	6.408	8.672	10.085	24.769	27.587	33.409	40.790
18	4.905	7.015	9.390	10.865	25.989	28.869	34.805	42.312
19	5.407	7.633	10.117	11.651	27.204	30.144	36.191	43.820
20	5.921	8.260	10.851	12.443	28.412	31.410	37.566	45.315
21	6.447	8.897	11.591	13.240	29.615	32.671	38.932	46.797
22	6.983	9.542	12.338	14.041	30.813	33.924	40.289	48.268
23	7.529	10.196	13.091	14.848	32.007	35.172	41.638	49.728
24	8.085	10.856	13.848	15.659	33.196	36.415	42.980	51.179
25	8.649	11.524	14.611	16.473	34.382	37.652	44.314	52.620
26	9.222	12.198	15.379	17.292	35.563	38.885	45.642	54.052
27	9.803	12.879	16.151	18.114	36.741	40.113	46.963	55.476
28	10.391	13.565	16.928	18.939	37.916	41.337	48.278	56.892
29	10.986	14.256	17.708	19.768	39.087	42.557	49.588	58.301
30	11.588	14.953	18.493	20.599	40.256	43.773	50.892	59.703
35	14.688	18.509	22.465	24.797	46.059	49.802	57.342	66.619
40	17.916	22.164	26.509	29.051	51.805	55.758	63.691	73.402
45	21.251	25.901	30.612	33.350	57.505	61.656	69.957	80.077
50	24.674	29.707	34.764	37.689	63.167	67.505	76.154	86.661
55	28.173	33.570	38.958	42.060	68.796	73.311	82.292	93.168
60	31.738	37.485	43.188	46.459	74.397	79.082	88.379	99.607
70	39.036	45.442	51.739	55.329	85.527	90.531	100.425	112.317
80	46.520	53.540	60.391	64.278	96.578	101.879	112.329	124.839
90	54.155	61.754	69.126	73.291	107.565	113.145	124.116	137.208
100	61.918	70.065	77.929	82.358	118.498	124.342	135.807	149.449
200	143.843	156.432	168.279	174.835	226.021	233.994	249.445	267.541
500	407.947	429.388	449.147	459.926	540.930	553.127	576.493	603.446

Esimerkki / Example:

Jos $\alpha = 0.01$ ja $df = 11$, niin $\Pr(\chi^2 > 24.725) = 0.01$
If $\alpha = 0.01$ and $df = 11$, then $\Pr(\chi^2 > 24.725) = 0.01$

Taulukko 19: χ^2-jakauman taulukoidut arvot ja niihin lasketut p-arvot [59].

t-DISTRIBUTION $t(df)$
t-JAKAUMA $t(df)$

Kriittisiä arvoja / Critical values

	Merkitsevyystaso 1-suuntaisissa testeissä / Sigficance level in 1-sided tests									
df	**0.4**	**0.3**	**0.2**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646
35	0.255	0.529	0.852	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704	3.307	3.551
45	0.255	0.528	0.850	1.301	1.679	2.014	2.412	2.690	3.281	3.520
50	0.255	0.528	0.849	1.299	1.676	2.009	2.403	2.678	3.261	3.496
55	0.255	0.527	0.848	1.297	1.673	2.004	2.396	2.668	3.245	3.476
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660	3.232	3.460
70	0.254	0.527	0.847	1.294	1.667	1.994	2.381	2.648	3.211	3.435
80	0.254	0.526	0.846	1.292	1.664	1.990	2.374	2.639	3.195	3.416
90	0.254	0.526	0.846	1.291	1.662	1.987	2.368	2.632	3.183	3.402
100	0.254	0.526	0.845	1.290	1.660	1.984	2.364	2.626	3.174	3.390
200	0.254	0.525	0.843	1.286	1.653	1.972	2.345	2.601	3.131	3.340
500	0.253	0.525	0.842	1.283	1.648	1.965	2.334	2.586	3.107	3.310
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576	3.090	3.291
df	**0.8**	**0.6**	**0.4**	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
	Merkitsevyystaso 2-suuntaisissa testeissä / Sigficance level in 2-sided tests									

Esimerkki / Example:

Jos α = 0.01 ja df = 11, niin Pr(t > 2.718) = 0.01
If α = 0.01 and df = 11, then Pr(t > 2.718) = 0.01

Taulukko 20: T-jakauman taulukoidut arvot ja niihin lasketut p-arvot [59].

F-JAKAUMA / F-DISTRIBUTION $F(df_1, df_2)$

Kriittisiä arvoja 5 %:n merkitsevyystasolle / Critical values at the 5 % level of significance.

0.05	df_1									
df_2	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
35	4.121	3.267	2.874	2.641	2.485	2.372	2.285	2.217	2.161	2.114
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
45	4.057	3.204	2.812	2.579	2.422	2.308	2.221	2.152	2.096	2.049
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
55	4.016	3.165	2.773	2.540	2.383	2.269	2.181	2.112	2.055	2.008
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951
90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927
200	3.888	3.041	2.650	2.417	2.259	2.144	2.056	1.985	1.927	1.878
500	3.860	3.014	2.623	2.390	2.232	2.117	2.028	1.957	1.899	1.850
∞	3.841	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831

Esimerkki / Example:

Jos $df_1 = 5$ ja $df_2 = 8$, niin $\Pr(F > 3.687) = 0.05$.
If $df_1 = 5$ and $df_2 = 8$, then $\Pr(F > 3.687) = 0.05$.

Taulukko 21: F-jakauman taulukoidut arvot ja niihin lasketut p-arvot [59].