Lapsen pituuden selittäminen lineaarisella regressiomallilla

Tuomas Reiterä 013759335
Helsingin yliopisto
Matemaattis-luonnontieteellinen tiedekunta
Matematiikan ja tilastotieteen laitos
Tilastotiede
Kandidaatintutkielma
20. toukokuuta 2015

Sisältö

Luku 1

Johdanto

Tässä kandidaatintutkielmassa esittelen lapsen pituuden selittämistä lineaarisella regressiomallilla vanhempien pituuksien avulla. Lineaarinen regressiomalli on paljon käytetty tilastollinen menetelmä. Menetelmää käytetään esimerkiksi taloustieteessä, joka on itseäni kiinnostava tilastotiedettä soveltava oppiaine. Pituuden selittäminen on yksi varhaisimmista lineaarisen regressiomallin sovelluskohteista.

Aloitan työni esittelemällä lineaarisen mallin yleisessä muodossa sekä lineaarisen regressiomallin. Esittelen mallin muodostamisen ja malliin liittyvät oletukset. Seuraavaksi kerron mallin syntyhistoriasta ja kuinka mallia on aiemmin pituuden selittämisessä käytetty. Mallin keksijänä pidetään Francis Galtonia. Galton kehitti mallin tutkiessaan perinnöllisyyttä. Hän aloitti tutkimalla herneiden siementen kokoa ja siirtyi sitten tutkimaan ihmisten vanhempien ja jälkeläisten pituuksien suhdetta. Biologien mukaan lapsen pituuden tulisi periytyä yhtä lailla sekä isältä että äidiltä, mutta Galtonin mallissa äidin pituuden merkitys on suurempi. Esittelen työssäni eri tutkijoiden näkemyksiä mistä tämä johtuu.

Omassa esimerkissäni selitän lapsen pituutta lineaarisen regressiomallin avulla suomalaisesta pituusaineistosta. Käytössäni oleva aineisto on osa Itä-Suomen yliopiston kasvututkimuksen työryhmän vuonna 2009 Espoosta kokoamaa aineistoa, jossa on 4000 vuonna 1983-2009 syntyneen pituudet ja heidän vanhempiensa pituudet. Osa-aineistoon on valittu 500 poikaa ja 500 tyttöä, jotka ovat parhaiten saavuttaneet lopullisen pituutensa. Valinnan on tehnyt lääketieteen lisensiaatti Antti Saari. Kiitän professori Leo Dunkelia ja Antti Saarta aineiston luovuttamisesta ja Saarta aineiston tekemisestä. Esimerkissäni luon mallin, joka selittää lapsen pituutta vanhempien pituuden avulla ja teen mallin myös standardoiduilla selittäjillä. Lisäksi teen parametreille hypoteesitestin. Hankaluuksia aineiston kanssa tuottavat aineiston lyhyimpien poikien pituudet. Pojat tunnetusti kasvavat tyttöjä myöhemmin ja aineiston nuorin poika on vasta 16,5-vuotias. Kaikki eivät siis vielä välttämättä ole saavuttanut lopullista pituuttaan. Käytän aineistoni käsittelemiseen R-ohjelmistoa.

Luku 2

Lineaarinen malli

Lineaarisessa mallisssa pyritään selittämään jonkun muuttujan vaihtelua jonkun toisen tai useamman eri muuttujien vaihtelun avulla. Mallin muodostamiseen tarvitaan n:n havaintoyksikön aineisto, jossa yksi muuttuja on selitettävä ja loput muuttujat ovat selittäviä. Selitettävän ja selittävien muuttujien riippuvuuden oletetaan olevan lineaarista. Lineaarinen malli on muotoa (Saikkonen 2007, 2-4):

$$Y_i = \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, ..., n. \tag{2.1}$$

$Y_1, ..., Y_n$ ovat satunnaismuuttujia joiden havaitut arvot ovat selitettävän muuttujan arvot $y_1, ..., y_n$. Selittävien muuttujien havaintoarvot ovat x_{ij} ja $\beta_1, ..., \beta_p$ ovat tuntemattomia parametreja. Mallin virhetermiksi kutsuttu ϵ_i on havaintoyksikköön i liittyvä ei-havaittava satunnaismuuttuja. Sen tehtävä on kuvata sitä osaa selitettävän muuttujan vaihtelusta, jota selittävät muuttujat ja niihin liittyvien parametrien muodostama lineaarikombinaatio $\beta_1 x_{i1} + ... + \beta_p x_{ip}$ eivät kykene selittämään. Lineaarikombinaatiota $\beta_1 x_{i1} + ... + \beta_p x_{ip}$ kutsutaan mallin systemaattiseksi osaksi.

Yhtälö 2.1 voidaan kirjoittaa myös matriisimuodossa $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, eli

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Parametrien $\beta_1, ..., \beta_p$ pienimmän neliösumman estimaatit saadaan matriisitoimenpiteillä laskemalla $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, jossa $\mathbf{y}$ on selitettävän muuttujien havaintojen muodostama pystyvektori. Pystyvektori $\hat{\boldsymbol{\beta}}$ sisältää siis estimoidut parametrit $\hat{\beta}_1, ..., \hat{\beta}_p$. Mallin sovite $\hat{\boldsymbol{y}}$ eli estimoitu systemaattinen osa saadaan laskemalla $\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Mallin residuaali eli estimoitu satunnainen osa on tällöin $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. (Saikkonen 2007, 8-9)

Lineaarisen mallin tulee täyttää joukko oletuksia. Selittäviä muuttujia ja niihin liittyviä parametreja koskevat oletukset ovat $\mathbf{X} \in \mathbb{R}^{n \times p}$ ja $\boldsymbol{\beta} \in \mathbb{R}^p$. Lisäksi $\mathbf{X}$:n on oltava täyttä sarakeastetta, eli $r(\mathbf{X}) = p$. Virhetermejen ϵ_i on oltava riippumattomia ja toisinaan malliin liitetään myös oletus virhetermien normaalisuudesta, eli $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$, $\sigma^2 > 0$. (Saikkonen 2007, 1-3)

2.1 Lineaarinen regressiomalli

Tämän tutkielman esimerkissä käytetään kahden selittäjän lineaarista regressiomallia. Usean selittäjän lineaarinen regressiomalli voidaan kirjoittaa muodossa (Saikkonen 2007, 5-6):

$$Y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, ..., n \tag{2.2}$$

Malliin liitetään samat oletukset kuin esiteltyyn lineaarisen mallin yleiseen muotoon. Lineaarisessa regressiomallissa matriisi $\boldsymbol{X}$ on samanlainen kuin lineaarisessa mallissa (2.1) paitsi, että termi $x_{i1} = 1$ kaikilla $i = 1, ...n$. Parametria β_1 kutsutaan mallin vakioksi ja parametreja $\beta_2, ..., \beta_p$ regressiokertoimiksi. Yksittäinen regressiokerroin kertoo kuinka paljon selitettävän muuttujan ehdollinen odotusarvo muuttuu, kun kertoimeen liittyvän selittävän muuttujan arvo muuttuu yhden yksikön, muiden selittäjien pysyessä ennallaan. Mallin yksi käyttötarkoitus on selitettävän muuttujan ja selittävien muuttujien välisen riippuvuuden tiivistetty kuvaaminen.

Luku 3

Regressiomallin ja pituuden selittämisen historiaa

Lapsen pituuden mallintaminen vanhempien pituuksien avulla on yksi vanhimmista lineaarisen regressiomallin käyttötarkoituksista. Mallin isä Francis Galton oli Charles Darwinin koulukuntaa ja omaksui hänen ajatuksiaan evoluutiosta ja luonnonvalintateoriasta. Galton argumentoi, että fyysiset ominaisuudet ja lahjakkuus periytyvät sukupolvelta toiselle ja keräsi tilastoaineiston väitteensä tueksi. Galton käytti ensimmäisenä tutkimuksissaan tietynlaisia herneen siemeniä. Ne pystyvät tuottamaan yksin jälkeläisiä, joten niiden periytyvyys oli mahdollisimman yksinkertaista. Herneiden etuna oli myös niiden pyöreä muoto, jolloin herneen kokoa pystyi kuvaamaan yhdellä mitalla, sen halkaisijalla. (Pere 2014)

Seuraavaksi Galton siirtyi analysoimaan ihmisten ja heidän jälkeläistensä pituuksien suhdetta. Herneaineistoon verrattuna tämä oli monimutkaisempaa, koska pituus periytyi kahdelta vanhemmalta. Galton ratkaisi tämän ongelman luomalla niin sanotun keskivanhemman. Näin sukupuolen vaikutus pituuden periytymiseen voitaisiin sivuuttaa ja aineisto oli myös mahdollisimman samankaltainen yksinkertaisemman herneaineiston kanssa. Keskivanhemman pituuden hän määritteli vanhempien pituuksien painotettuna keskiarvona (Pere 2014):

$$\frac{x_I + 1,08x_A}{2}, \tag{3.1}$$

jossa x_I tarkoittaa isän pituutta ja x_A äidin pituutta. Kertoimen 1,08 merkitystä Galton ei tarkasti selittänyt, mutta myöhemmin aihetta tutkineen Karl Pearsonin mukaan se oli isien ja äitien keskipituuksien suhde. Isät ovat siis keskimäärin 1,08 kertaa äitejä pidempiä. Myös tyttölapsien pituudet kerrottiin 1,08:lla, jotta ne olisivat verrannollisia poikien

pituuksiin. Galtonin keskivanhemmalla tehty regressiomalli voidaan esittää muodossa:

$$Y_i = \beta_1 + \beta_2 \frac{x_{Ii} + 1,08x_{Ai}}{2} + \epsilon_i \qquad i = 1, ..., n, \tag{3.2}$$

jossa Y_i ovat lapsien pituudet, β_1 on mallin vakio ja β_2 on keskivanhemman pituuteen liittyvä regressiokerroin. Regressiokertoimeksi Galton laski $\beta_2 = 2/3$, mutta Pearson kritisoi aineiston laatua ja arveli, että $4/5$ olisi lähempänä totuutta. Galton oivalsi että isien pituutta voi myös selittää poikien pituudella, eli hän irrottautui regression rajaamisesta yhteen suuntaan vain periytyvyyden selittäjänä. (Pere 2014)

3.1 Eroavat regressiokertoimet

Tarkasteltaessa keskivanhemman pituuden määritelmää (3.1) huomataan äidin pituuden vaikuttavan enemmän lapsen pituuteen kuin isän pituuden. Kun äidin pituus kasvaa yksiköllä, vaikuttaa se kertoimen 1,08 takia enemmän lapsen pituuden odotusarvoon kuin isän pituuden yksikön kasvu. Äidin pituuden suurempi merkitys säilyy, vaikka luovuttaisiin keskivanhempi-esitysmuodosta ja esitettäisiin lapsien pituus kahden selittäjän regressiomallina isän ja äidin pituudet selittävinä muuttujina. Yhtälö 3.2 voidaan kirjoittaa näin:

$$Y_i = \beta_1 + \frac{\beta_2 x_{Ii}}{2} + 1,08\frac{\beta_2 x_{Ai}}{2} + \epsilon_i \qquad i = 1, .., n. \tag{3.3}$$

Yhtälöstä nähdään, että kertoimen 1,08 takia äidin pituuteen liittyvä regressiokerroin olisi 1,08 kertaa suurempi kuin isän pituuteen liittyvä regressiokeroin.

Pearson ja Alice Lee perustelivat kertoimen 1,08 järkevyyttä sillä, että pojan pituutta mallinnettaessa yhdellä selittäjällä, joko äidin tai isän pituudella, on regressiokertoimien suhde hyvin lähellä lukua 1,08. Nämä kertoimet näkyvät Marcello Paganon ja Sarah Anoken tekemästä Pearsonin ja Leen tulosten uudelleen taulukoinnista. Taulukko on kopioituna alle:

Poikien mallin regressiosovitteen yhtälö (cm)
85,67+0,516(Isän pituus)
85,47+0,560(Äidin pituus)
35,76+0,409(Isän pituus)+0,430(Äidin pituus)
Tyttöjen mallin regressiosovitteen yhtälö (cm)
77,47+0,493(Isän pituus)
74,37+0,554(Äidin pituus)
27,48+0,386(Isän pituus)+0,431(Äidin pituus)

Taulukko 3.1: Lapsen pituuden eri regressiot (Pagano ja Anoke 2013, tuumat on muunnettu senttimetreiksi)

Mallinnettaessa pojan pituutta sekä isän että äidin pituuden avulla ei regressiokertoimien suhde enää olekaan 1,08 vaan noin 1,05 (0,430/0,409=1,051). Jotta kerroin 1,08 olisi järkevä ja sen voitaisiin ajatella johtuvan vain miesten ja naisten pituuden keskimääräisestä erosta, olisi tyttöjen pituutta mallintaessa regressiokertoimien suhteen myös oltava 1,08. Taulukossa olevien kertoimien suhde on kuitenkin sekä yksien selittäjien, että kahden selittäjän mallissa 1,12. Äidin pituuden vaikutus on siis tyttölapsien kohdalla vielä suurempi isän pituuteen verrattuna kuin poikien tapauksessa. Tämä on Paganon ja Anoken (2013) mukaan ristiriitaista kun tiedetään, että tytöt ovat poikia lyhyempiä.

Periytyvyysteorioiden mukaan pituuden tulisi periytyä yhtä paljon äidin ja isän puolelta. Pagano ja Anoke (2013) pyrkivät osoittamaan, ettei eroa regressiokertoimissa voi kuitata vain sillä, että naiset ovat miehiä lyhyempiä. Heidän selitys äitien suuremmalle merkitykselle on se, että lapsen isä ei oikeasti aina ole lapsen luultu isä. Vaikka lapsi voi esimerkiksi vaihtua synnytysosastolla, on lapsen äiti lähestulkoon aina varmasti tiedossa. Sen sijaan uskottomuutta parisuhteessa on ollut aina. Tämän vuoksi lapsen oikea isä ei aina ole varma. Aihe on ollut erityisesti Galtonin, Pearsonin ja Leen aikakaudella tabu, jonka vuoksi he olisivat saattaneet sivuuttaa sen, vaikka olisivat sen aavistaneet.

Suuria johtopäätöksiä selittävien muuttujien merkityksestä ei välttämättä pitäisi tehdä tarkastelemalla regressiokertoimia. Selittävien muuttujien arvot ja vaihteluvälit vaikuttavat aina regressiokertoimiin. Kertoimia tulisi verrata vain silloin kun selittävien muuttujien vaihteluvälit ovat lähes samanlaisia. Tiedetään, että isät ovat keskimäärin äitejä selvästi pidempiä. Tämän takia regressiokertoimien vertailu on mielekkäämpää silloin, kun vanhempien pituudet ovat standardoitu keskiarvoiltaan ja keskihajoinnoiltaan samoiksi. Selittävien muuttujien standardointiin palataan tarkemmin seuraavassa luvussa.

Luku 4

Esimerkki: Lapsen pituuden mallintaminen suomalaisesta pituusaineistosta

4.1 Aineiston tarkastelu

Tässä luvussa mallinnetaan lapsen pituutta vanhempien pituuksien avulla käyttäen Itä-Suomen yliopiston kasvututkimuksen työryhmän aineistoa, jossa on 500 pojan ja 500 tytön pituudet ja heidän ilmoittamat vanhempiensa pituudet. Aineisto on osa isompaa pituusaineistoa, johon on kerätty 4000 vuosina 1983-2009 syntyneiden henkilöiden pituudet. Tähän osa-aineistoon on valittu pojat ja tytöt, jotka ovat parhaiten saavuttaneet lopullisen pituutensa. Osa-aineiston nuorin poika on iältään 16,5 vuotta ja nuorin tyttö on 17-vuotias. Käytetään aineiston käsittelemiseen r-ohjelmistoa.

Tarkastellaan ensiksi lapsien ja vanhempien pituuksien tunnuslukuja.

	Minimi	**Mediaani**	**Keskiarvo**	**Maksimi**
Pojan pituus	154,4	177,5	177,2	197,0
Isän pituus	160,0	180,0	179,4	198,0
Tytön pituus	149,5	166,5	166,6	184,2
Äidin pituus	145,0	165,0	165,2	182,0

Taulukko 4.1: Lapsien ja vanhempien pituuksien tunnuslukuja (cm)

Poikien pituuden keskiarvo ja mediaani ovat pienempiä kuin isien pituuden. Uusi sukupolvi kasvaa tavallisesti edeltäjäänsä pidemmäksi, joten tulos on hieman yllättävä. Voidaan epäillä ovatko kaikki nuorimmillaan alle 17-vuotiaat pojat saavuttaneet lopullisen

pituutensa. Aineiston lyhin poika on vain 154,4 cm pitkä. Tytöt kasvavat tavallisesti poikia aiemmin ja he ovat aineistossa keskimäärin äitejä pidempiä. Jos aineistossa olisi mukana ikämuuttuja, voisi esimerkiksi kaikki alle 19-vuotiaat pojat poistaa ja tarkastella todennäköisemmin lopullisen pituutensa saavuttaneita.

Tarkastellaan pituuksien välisiä korrelaatioita.

	Pojan pituus	**Tytön pituus**	**Äidin pituus**	**Isän pituus**
Pojan pituus	1,000	-	0,508	0,435
Tytön pituus	-	1,000	0,506	0,551
Äidin pituus	-	-	1,000	0,248
Isän pituus	-	-	-	1,000

Taulukko 4.2: Pituuksien väliset korrelaatiot

Lapsien ja vanhempien pituudet korreloivat keskenään. Myös vanhempien pituudet korreloivat hieman keskenään. Pitkät miehet pariutuvat siis useammin pitkien naisten kanssa ja lyhyemmät ihmiset taas keskenään. Kun tarkastellaan poikien ja isien pituuksista tehtyä hajontakuvaa (ks. liite B: kuva B.1), nähdään pitkillä isillä olevan tavallisesti myös pitkiä poikia. Hajontakuvasta nähdään myös, että poikien joukossa on joitakin pituudeltaan alle 160 cm pitkiä, kun isistä kukaan ei ole näin lyhyt. On jopa yksi havainto, jossa poika on alle 160 cm pitkä, vaikka isä on 190 cm pitkä. Todetaan, ettei aineisto ole poikien kohdalla sopiva lopullisen pituuden selittämiseen. Jatkossa käsitellään vain aineiston tyttöjä, mutta pojilla saadut tulokset ovat myös nähtävissä liitteessä A.

4.2 Lineaarinen regressiomalli lapsen pituudesta

Tehdään seuraavaksi tytöille regressiomallit. Tehdään ne Paganon ja Anoken (2013) tapaan. Ensiksi luodaan yhden selittäjän mallit vain toisen vanhemman pituus selittäjänä ja seuraavaksi kahden selittäjän mallit molempien vanhempien pituudet selittäjinä.

Tyttöjen mallin regressiosovitteen yhtälö (cm)	**Mallin selitysaste**
69,88+0,538(Isän pituus)	30 %
79,83+0,524(Äidin pituus)	26 %
21,64+0,437(Isän pituus)+0,400(Äidin pituus)	44 %

Taulukko 4.3: Tyttöjen pituuden eri regressiot

Paganon ja Anoken (2013) esittelemää ilmiötä, että äidin pituuteen liittyvä regressiokerroin olisi järjestelmällisesti suurempi, ei havaita tässä aineistossa. Tytön pituutta malllinnettaessa isän pituuden regressiokerroin on suurempi kuin äidin pituuden kerroin.

Mallin selitysaste R^2 saadaan vähentämällä residuaalineliösumman $\Sigma_{i=1}^{n}\hat{\epsilon}_i^2$ ja kokonaisneliösumman $\Sigma_{i=1}^{n}(y_i - \bar{y})^2$ osamäärä luvusta yksi (Saikkonen 2007, 10).

$$R^2 = 1 - \frac{\Sigma_{i=1}^{n}\hat{\epsilon}_i^2}{\Sigma_{i=1}^{n}(y_i - \bar{y})^2}$$

Yhtälössä $\hat{\epsilon}$ on residuaali eli estimoitu satunnainen osa, y_i ovat havaitut selitettävän muuttujan arvot, eli tässä tapauksessa lasten pituudet ja $\bar{y}$ on niiden keskiarvo. Selitysaste R^2 saa arvoja nollan ja ykkösen väliltä ja se ilmaistaan yleensä prosentteina. Tyttöjen kahden selittäjän mallin selitysaste on 44 %, eli malli selittää 44 % tyttöjen pituuksien vaihtelusta.

Tarkasteltaessa kahden selittäjän mallin virhetermejä residuaalihajontakuvan avulla (liite B: kuva B.3), nähdään mallissa residuaalien jakaantuvan suhteellisen tasaisesti nollan molemmin puolin. Sekä suuriin että pieniin havaittuihin pituuksiin liittyy itseisarvoltaan sekä suuria että pieniä residuaaleja. Normaalipaperipiirroksesta (liite B: kuva B.5) nähdään, että mallin normaalisuusoletus toteutuu melko hyvin.

4.3 Lineaarinen malli standardoiduilla selittäjillä ja hypoteesitestaus

Pearson kritisoi Galtonin käyttämää keskivanhempaa (määritelmä 3.1) ja hänen mielestään järkevämpää olisi ollut standardoida isän ja äidin pituuksien keskihajonta samaksi (Pere 2014). Standardoitujen selittäjien regressiokertoimia on mielekkäämpää verrata. Kun standardoidaan molempien selittäjien keskihajonnan arvoksi yksi, vanhemman pituuteen liittyvä regressiokerroin kertoo, kuinka paljon lapsen pituuden ehdollinen odotusarvo kasvaa, kun tarkasteltavaan kertoimeen liittyvän vanhemman pituus kasvaa keskihajonnan verran ja toisen vanhemman pituus säilyy ennallaan. Selittävien muuttujien yksikköinä toimii tällöin vanhempien pituuksien keskihajonnat pituusyksiköiden sijaan. Isien suurempi pituus äiteihin verrattuna voidaan tällä tavoin häivyttää.

Selittäjien standardoinnin merkitys avautuu ehkä vielä paremmin, jos selittäjinä toimisivat täysin eriasteikolliset muuttujat. Selitetään esimerkiksi perheen ruokaan käyttämää rahamäärää kuukaudessa perheen kuukauden tuloilla ja perheenjäsenten lukumäärällä. Jos tulojen yksikkönä toimii dollari ja perheen koon yksikkönä henkilömäärä, on tuloihin liittyvä regressiokerroin luonnollisesti pienempi. Yhden dollarin kasvu kuukausituloissa on minimaalinen muutos, kun taas perheen koon kasvaminen yhdellä on iso muutos. Näiden kahden selittäjän vaikutusta selitettävään muuttujaan vertailtaessa on muuttujat standardoitava, jolloin muuttujien yksikköinä toimivat niiden omat keskihajonnat. Standardoinnin jälkeen voidaan tarkastella selittäjien regressiokertoimia. (Schroeder, Sjoquist, Stephan 1986, 31-32)

Standardoitujen muuttujien regressiokertoimien vertailua on kritisoitu siitä, että standardointi riippuu käytössä olevan aineiston muuttujien vaihteluvälistä. Jos kahdella tutkijalla on ainestoa samoilla selittävillä muuttujilla, mutta toisen tutkijan selittäjissä on pieni vaihteluväli ja toisella suuri vaihteluväli, voivat he päätyä regressiokertoimia tarkastellessaan täysin eri lopputuloksiin. (Weisberg, 1985, 186)

Tehdään seuraavaksi lineaarinen malli käyttäen selittäjinä vanhempien pituusmuuttujia, joiden keskiarvo on standardoitu nollaksi ja keskihajonta ykköseksi. Jokaisen isän pituudesta vähennetään isien pituuden keskiarvo, ja erotus jaetaan isien pituuden keskihajonnalla. Nyt isien pituuden keskiarvo on nolla ja keskihajonta on yksi. Vastaavalla tavalla toimitaan äitien pituuksien kanssa, jolloin selittävät muuttujat noudattavat samaa jakaumaa. Standardoimalla vastaavalla tavalla myös selitettävä muuttuja, mallin vakio katoaa sillä mallin sovite kulkee aina origon kautta. Nyt standardoiduilla selittäjillä saadaan tyttöjen standardoitujen pituuksien sovitteen yhtälöiksi:

$$\hat{y}_T^* = 0,448x_I^* + 0,387x_A^*.$$

Yhtälöissä $\hat{y}_T^*$ tarkoittaa tyttöjen mallin sovitetta. x_I^* ja x_A^* tarkoittavat standardoituja vanhempien pituuksia. Selittäjien standardointi ei vaikuta mallin selitysasteeseen (ks. taulukko 4.3). Nyt kun kaikki muuttujat on standardoitu, säilyy selittävien muuttujien parametrien estimaattien suuruusjärjestys edelleen samana. Mallissa isän pituuteen liittyvä estimaatti on suurempi kuin äidin pituuteen liittyvä estimaatti. Parametrien estimaattien arvoista nähdään, että isän pituuden kasvaessa isien pituuden keskihajonnan verran, kasvaa tytön pituuden ehdollinen odotusarvo 0,448 kertaa tytön pituuden keskihajonnan verran, kun äidin pituus säilyy ennallaan. Vastaavasti toisesta estimaatista nähdään tytön pituuden ehdollisessa odotusarvossa tapahtuva muutos, kun äidin pituus muuttuu.

Tehdään seuraavaksi F-testi r-ohjelmistolla John Foxin ja Sanford Weisbergin carpaketin avulla. Tehdään testi hypoteesille, että selittäviin muuttujiin liittyvät parametrit ovat yhtä suuret.

$$H_0 : \beta_I = \beta_A.$$

β_I tarkoittaa isän pituuteen liittyvää parametria ja β_A äidin pituuteen liittyvää parametria. Tehdään testi mallille, jossa selittävät muuttujat ovat standardoitu. F-testissä muodostetaan testisuure F (Saikkonen 2007, 18):

$$F = (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})/qS^2 \sim^H F_{q,n-p}.$$

Matriisi $\mathbf{A}$ ja vektori $\mathbf{c}$ valitaan siten, että nollahypoteesin pätiessä pätee myös $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$. Tässä tapauksessa $\mathbf{A} = \begin{pmatrix} 1 & -1 \end{pmatrix}$ ja $\mathbf{c} = 0$. Tällöin nollahypoteesin $H_0 : \beta_I = \beta_A$ pätiessä pätee myös:

$$\begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_I \\ \beta_A \end{pmatrix} = 0.$$

Vektori $\hat{\boldsymbol{\beta}}$ muodostuu mallin estimoiduista parametrien arvoista. Matriisi $\mathbf{X}$ muodostuu selittävien muuttujien arvoista kuten luvussa 2 on esitelty. Testisuureen yhtälössä oleva q on matriisin $\mathbf{A}$ rivien lukumäärä ja

$$S^2 = (n-p)^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

jossa $\mathbf{Y}$ on selitettävien satunnaismuuttujien $Y_1, ..., Y_n$ muodostama pystyvektori (ks. luku 2) ja n ja p ovat matriisin $\mathbf{X}$ rivien ja sarakkeiden lukumäärät. Testisuureen suuret arvot ovat kriittisiä nollahypoteesin kannalta. Testisuureen p-arvo perustuu tulokseen

$$p = P_H(F(\mathbf{Y}) \geq F(\mathbf{y})) = P(F_{q,n-p} \geq F(\mathbf{y})),$$

jossa $F_{q,n-p}$ on jakaumaa $F_{q,n-p}$ noudattava satunnaismuuttuja (Saikkonen 2007, 18). Tyttöjen pituuden kahden standardoidun selittäjän mallille tehty hypoteesitestaus parametrien yhtäsuuruudesta antaa tulokset $F = 1,233$ ja $p = 0,267$. Nollahypoteesi jää merkitsevyystasolla $\alpha = 0,05$ voimaan. Aineisto ei anna perusteita hylätä nollahypoteesia ja ottaa käyttöön vastahypoteesia $\beta_I \neq \beta_A$.

Luku 5

Mallin toimivuuden tarkastelu ja mallin ulkoisvaikutukset

Lineaarinen regressiomalli on hyvä menetelmä selitettäessä jälkeläisen pituutta vanhempien pituuden avulla. Pituuden selittäminen on myös yksi ensimmäisistä mallin sovelluskohteista. Biologien mukaan pituuden tulisi periytyä lapselle yhtä lailla äidiltä ja isältä. Lapsen pituutta selitettäessä vanhempien pituuksilla on kuitenkin usein huomattu vanhempien pituuteen liittyvien regressiokertoimien olevan erisuuria. Pearsonin ja Leen mukaan äidin pituuteen liittyvä regressiokerroin on suurempi koska äidin kerroin on kerrottava miesten ja naisten välisen keskipituuden suhteella (Pagano ja Anoke 2013). Paganon ja Anoken (2013) johtopäätös äidin pituuden isommasta regressiokertoimesta puolestaan on, että jälkeläiset joskus ilmoittavat isänsä pituudeksi jonkun muun miehen pituuden kuin biologisen isänsä.

Työssä käyttämäni aineisto ei tue väitettä, että äidin pituuteen liittyvä regressiokerroin olisi järjestelmällisesti suurempi kuin isän pituuteen liittyvä kerroin. Tyttöjen pituutta mallinnettaessa isän pituuteen liittyvä regressiokerroin on suurempi kuin äidin pituuteen liittyvä kerroin. Standardoimalla vanhempien pituuksien keskihajonnan samaksi saa selittäjistä vertailukelpoisempia. Standardoiduilla selittäjillä tehdyn mallin hypoteesitestaus osoittaa, ettei tyttöjen mallissa parametrien estimaattien ero riitä kumoamaan nollahypoteesia. Nollahypoteesi on, että parametrt ovat yhtä suuret.

Aineisto sisältää poikkeuksellisen lyhyitä poikia ja aineiston nuorin poika on 16,5-vuotias. Aineiston valintakriteerinä on ollut yli 15 vuoden ikä ja se, että edellisen vuoden pituuskasvu on ollut vähemmän kuin 1 cm. Jotkut pojat saavat kuitenkin kasvupyrähdyksen vasta armeijaikäisenä. Näiden seikkojen takia poikien lopullisen pituuden selittäminen vanhempien pituuksien avulla ei ollut tällä aineistolla mielekästä.

Jälkeläisen pituuteen vaikuttaa perinnöllisyyden lisäksi myös erilaisia ulkoisvaikutuksia. Esimerkiksi sairaus voi estää lasta saavuttamasta koskaan omaa potentiaalista loppu-

pituuttaan. Myös vanhempien kärsimät ulkoisvaikutukset voivat sekoittaa mallia. Jompi kumpi vanhemmista on voinut jäädä esimerkiksi heikon sosioekonomisen aseman takia potentiaalista pituttaan lyhyemmäksi. Ulkoisvaikutuksen tuoma lyhytkasvuisuus ei kuitenkaan periydy, jolloin lapsi kasvaa pidemmäksi kuin olisi vanhempien pituuden perusteella ollut oletettavaa. On tutkitu, että ainakin menneinä vuosikymmeninä, jolloin esimerkiksi ravinnon puute oli yleisempää, vanhempien ja jälkeläisten pituuksien välinen korrelaatio on ollut suurempi ylemmissä yhteiskuntaluokissa kuin alemmissa yhteiskuntaluokissa. (Tanner ym. 1970, 761)

Kirjallisuutta

[1] Pagano M., Anoke S. (2013): Chance 26.3 4-9, Mommy's Baby, Daddy's Maybe: A Closer Look at Regression to the Mean, http://chance.amstat.org/2013/09/1-pagano/ (viitattu 22.11.2014).

[2] Pere P. (2014): Regressioanalyysin historiaa, Helsingin yliopiston sosiaalitieteiden laitos, Tilastotieteen juuret -kurssin oppimateriaali, julkaisematon.

[3] Saikkonen P. (2007) (korj. 2011): Lineaarinen malli, Helsingin yliopiston matematiikan ja tilastotieteen laitos, Lineaariset mallit -kurssin oppimateriaali, julkaisematon.

[4] Schroeder L., Sjoquist D, Stephan P. (1986): Understanding Regression Analysis, Sage Publications, Newbury Park, California.

[5] Tanner J. M., Goldstein H., Whitehouse R. H. (1970): Archives of Disease in Childhood, 45, 755. Standards for Children's Height at Ages 2-9 Years Allowing for Height of Parents.

[6] Weisberg S. (1985): Applied Linear Regression, John Wiley & Sons, New York.

Liite A

Poikien mallien tulosteet

Poikien mallin regressiosovitteen yhtälö (cm)	**Mallin selitysaste**
93,07+0,470(Isän pituus)	19 %
79,90+0,591(Äidin pituus)	26 %
28,67+0,366(Isän pituus)+0,504(Äidin pituus)	37 %

Taulukko A.1: Poikien pituuden eri regressiot

Poikien pituuden kahden selittäjän mallin residuaalihajontakuva ja normaalipaperipiirros ovat liitteessä B: kuvat B.2 ja B.4.

Poikien mallin sovitteen yhtälö standardoiduilla muuttujilla:

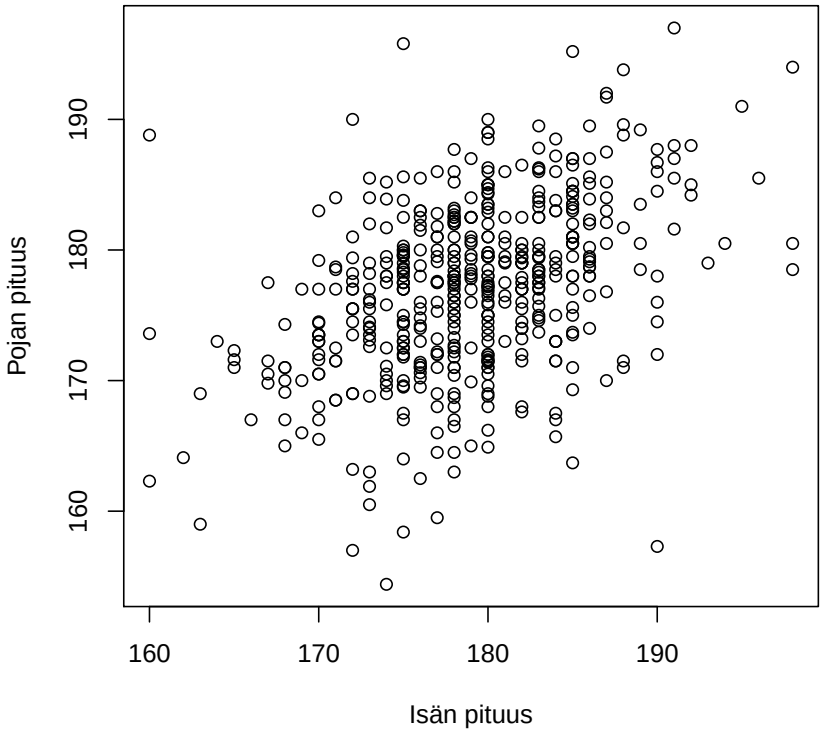
$$\hat{y}_P^* = 0,339x_I^* + 0,434x_A^*,$$

jossa $\hat{y}_P^*$ tarkoittaa poikien standardoidun pituuden sovitetta, x_I^* standardoitua isän pituutta ja x_A^* standardoitua äidin pituutta.

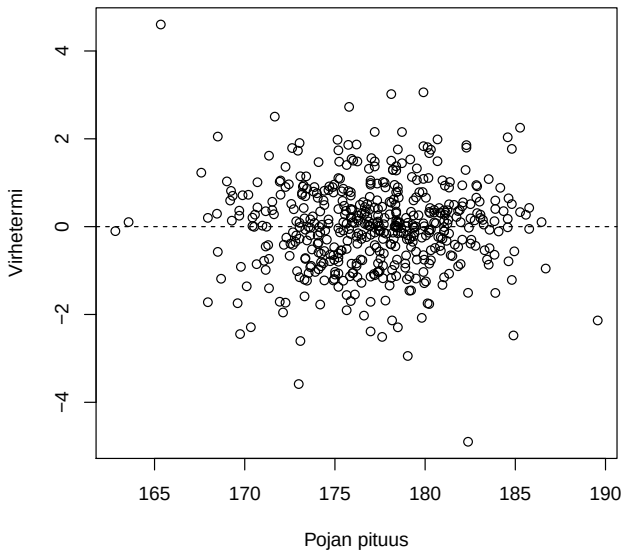
Poikien kahden standardoidun selittäjän mallille tehty F-testi hypoteesille mallin parametrien yhtäsuuruudesta antaa tulokset $F = 2,740$ ja $p = 0,099$.

Liite B

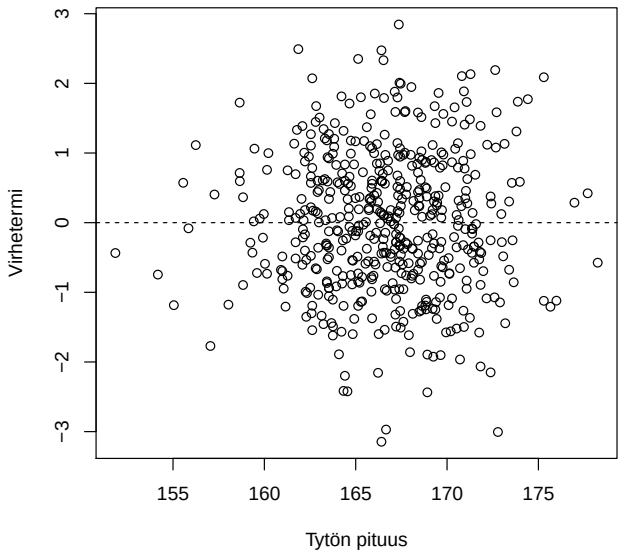
Kuvat



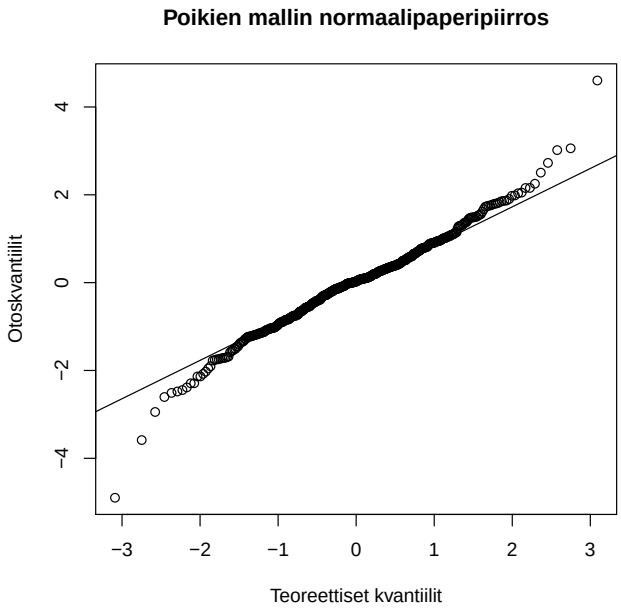
Kuva B.1: Hajontakuva poikien ja isien pituuksista



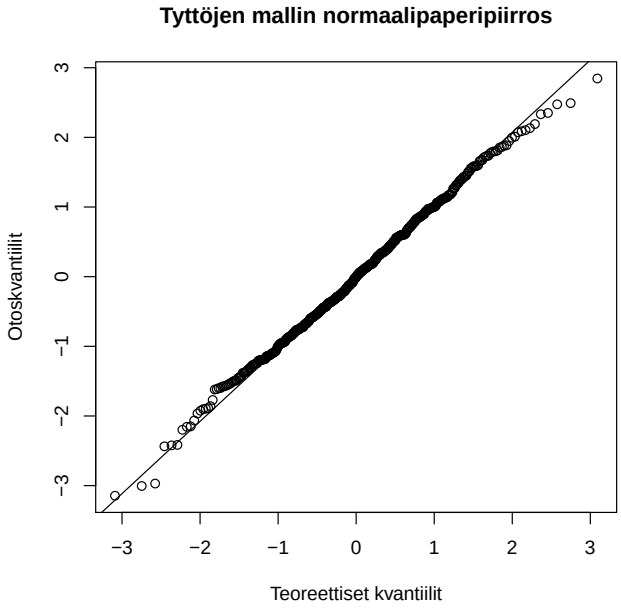
Kuva B.2: Poikien mallin residuaalihajontakuva



Kuva B.3: Tyttöjen mallin residuaalihajontakuva



Kuva B.4: Poikien mallin normaalipaperipiirros



Kuva B.5: Tyttöjen mallin normaalipaperipiirros

Liite C

R-koodit

```
#Data r:n
loppupituus <- read.table("loppupituus.txt", header=TRUE, sep="\t")

#Pituuksien tunnuslukuja
with(loppupituus, summary(Pituus))

#Pituuden tunnusluvut sukupuolittain
with(loppupituus, merge(aggregate(list(mean=Pituus),
 list(Sukupuoli), mean, na.rm=TRUE),
 aggregate(list(sd=Pituus),
 list(Sukupuoli), sd, na.rm=TRUE)))

#Vanhempien pituuksien tunnuslukuja
with(loppupituus, summary(ÄidinPituus))
with(loppupituus, summary(IsänPituus))

#Korrelaatiot
cor(loppupituus[,c("Pituus","ÄidinPituus","IsänPituus")], use="complete.obs")

#Poimitaan naisten pituudet erilleen
loppupituus$Sukupuoli=="F"
F<-loppupituus$Sukupuoli=="F"
Female<-loppupituus[F,]

#Tarkistetaan tunnusluvuilla
with(Female, summary(Pituus))
```

```
#korrelaatioita naisista
cor(Female[,c("Pituus","IsänPituus")], use="complete.obs")
cor(Female[,c("Pituus","ÄidinPituus")], use="complete.obs")

#Regressioanalyysi pelkistä naisten pituuksista
summary(with(Female, lm(Pituus ~ ÄidinPituus + IsänPituus)))
cor(Female[,c("Pituus","ÄidinPituus","IsänPituus")], use="complete.obs")
Female.fit <- with(Female, lm(Pituus ~ ÄidinPituus + IsänPituus))
library(MASS)
plot(fitted(Female.fit), studres(Female.fit), xlab="Tytön pituus",
ylab="Virhetermi")
abline(h=0, lty=2)
qqnorm(studres(Female.fit), main="Tyttöjen mallin normaalipaperipiirros",
xlab="Teoreettiset kvantiilit", ylab="Otoskvantiilit")
qqline(studres(Female.fit))

#Regressioanalyysi pelkistä naisten pituuksista
#(selittäjänä isän pituus)
summary(with(Female, lm(Pituus ~ IsänPituus)))

#Regressioanalyysi pelkistä naisten pituuksista
#(selittäjänä äidin pituus)
summary(with(Female, lm(Pituus ~ ÄidinPituus)))

#Luodaan uudet muuttujat isän ja äidin pituuksille
#vähentämällä pituudet keskiarvolla ja skaalaamalla
#niiden keskihajonnat samaksi
Female$ÄidinPituus2 <- scale(Female$ÄidinPituus, center = TRUE, scale = TRUE)

Female$IsänPituus2 <- scale(Female$IsänPituus, center = TRUE, scale = TRUE)

#Tehdään skaalattu tytön pituus
Female$TytönPituus2 <- scale(Female$Pituus, center = TRUE, scale = TRUE)

#Poimitaan miesten pituudet erilleen
loppupituus$Sukupuoli=="M"
M<-loppupituus$Sukupuoli=="M"
Male<-loppupituus[M,]
```

```
#Tarkistus tunnusluvuilla
with(Male, summary(Pituus))

#korrelaatiot ja hajontakuvat miehistä
cor(Male[,c("Pituus","ÄidinPituus","IsänPituus")], use="complete.obs")
with(Male, plot(IsänPituus, Pituus, xlab="Isän pituus", ylab="Pojan pituus"))

#ja regressioanalyysi
summary(with(Male, lm(Pituus ~ ÄidinPituus + IsänPituus)))
cor(Male[,c("Pituus","ÄidinPituus","IsänPituus")], use="complete.obs")
Male.fit <- with(Male, lm(Pituus ~ ÄidinPituus + IsänPituus))
library(MASS)
plot(fitted(Male.fit), studres(Male.fit), xlab="Pojan pituus",
ylab="Virhetermi")
abline(h=0, lty=2)
qqnorm(studres(Male.fit), main="Poikien mallin normaalipaperipiirros",
xlab="Teoreettiset kvantiilit", ylab="Otoskvantiilit")
qqline(studres(Male.fit))

#ja regressioanalyysi (selittäjänä isän pituus)
summary(with(Male, lm(Pituus ~ IsänPituus)))

#ja regressioanalyysi (selittäjänä äidin pituus)
summary(with(Male, lm(Pituus ~ ÄidinPituus)))

#Luodaan uudet muuttujat isän ja äidin pituuksille
#vähentämällä pituudet keskiarvolla ja skaalaamalla
#niiden keskihajonnat samaksi
Male$ÄidinPituus2 <- scale(Male$ÄidinPituus, center = TRUE, scale = TRUE)
Male$IsänPituus2 <- scale(Male$IsänPituus, center = TRUE, scale = TRUE)

#Tehdään skaalattu pojan pituus
Male$PojanPituus2 <- scale(Male$Pituus, center = TRUE, scale = TRUE)

#Regressioanalyysit skaalatuilla vanhempien pituudella
#Regressioanalyysi pelkistä naisten pituuksista
summary(with(Female, lm(TytönPituus2 ~ ÄidinPituus2 + IsänPituus2)))
cor(Female[,c("TytönPituus2","ÄidinPituus2","IsänPituus2")], use="complete.obs")
```

```
Female.fit <- with(Female, lm(TytönPituus2 ~ ÄidinPituus2 + IsänPituus2))

#ja miesten pituuksista
summary(with(Male, lm(PojanPituus2 ~ ÄidinPituus2 + IsänPituus2)))
cor(Male[,c("PojanPituus2","ÄidinPituus2","IsänPituus2")], use="complete.obs")
Male.fit <- with(Male, lm(PojanPituus2 ~ ÄidinPituus2 + IsänPituus2))

#Car-Packagen lataamiskoodit
local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
if(nchar(pkg)) library(pkg, character.only=TRUE)})
utils:::menuInstallPkgs()
library(car)
require(car)

#Tehdään hypoteesitestaus (äidin pituus = isän pituus)
linear.hypothesis(Female.fit, "ÄidinPituus2 = IsänPituus2", test = "F")
linear.hypothesis(Male.fit, "ÄidinPituus2 = IsänPituus2", test = "F")
```