Instrumenttimuuttujan käyttö kausaalipäättelyssä

Kalle Nyman

LuK -tutkielma
Maaliskuu 2024

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Kalle Nyman: Instrumenttimuuttujan käyttö kausaalipäättelyssä
LuK -tutkielma, 18s., 4 liites.
Tilastotiede
Maaliskuu 2024

Tutkielman tarkoituksena on antaa lukijalle yleiskatsaus instrumenttimuuttujien käytöstä tilastollisen analyysin työkaluna. Instrumenttimuuttujia käytetään ratkaisemaan endogeenisuuteen liittyviä ongelmia, jotka syntyvät, kun mallin selittävä muuttuja on korreloitunut virhetermin kanssa. Tämä voi johtaa harhaisiin ja epäluotettaviin kausaalipäätelmiin. Instrumenttimuuttujan avulla pyritään löytämään mallin ulkopuolinen muuttuja, joka vaikuttaa vasteeseen ainoastaan analysoitavan selittävän muuttujan kautta, mahdollistaen sekoittavien tekijöiden hallitsemisen ja näin luotettavampien kausaalisuhteiden arvioinnin.

Tutkielmassa käsitellään instrumenttimuuttujan ominaisuuksia, tämän validointiehtoja ja heikon instrumentin ongelmaa sekä vertaillaan tavallista pienimmän neliösumman regressiota (OLS) ja instrumenttianalyysissä käytettyä kaksivaiheista pienimmän neliösumman regressiota (2SLS). Lopuksi esitellään empiirinen esimerkki koulutuksen vaikutuksesta palkkaan ja arvioidaan instrumenttimuuttujamenetelmän soveltuvuutta.

Sanat: instrumenttimuuttuja, instrumenttianalyysi, endogeenisuus, eksogeenisuus, validointiehdot, 2SLS- ja OLS-regressio.

Sisällys

1 Johdanto

Kausaaliväitteiden tutkinta on keskeinen tavoite erityisesti taloustieteen ja sosiaalitieteiden tutkimuksissa, joissa pyritään ymmärtämään syy-seuraussuhteita monimutkaisissa ja usein epätäydellisissä ympäristöissä. Tutkijat kohtaavat usein tilanteen, jossa selittävä muuttuja ja vaste ovat yhteydessä toisiinsa sekoittavien tekijöiden kautta. Tällaiset sekoittavat tekijät (*confounder/confounding factor* [8]) voivat vääristää kausaalipäättelyn tuloksia, mikä tekee syy-seuraussuhteiden luotettavasta arvioinnista haastavaa.

Instrumenttimuuttujamenetelmän käytön motivaatio voidaan ymmärtää tarkastelemalla, miksi perinteinen lineaarinen regressiomalli ei aina sovellu kausaalipäättelyyn. Lineaarisessa regressiossa tehdään keskeinen oletus siitä, että selittävät muuttujat ovat eksogeenisia, eli ne eivät ole yhteydessä mallin virhetermiin. Oletus mahdollistaa luotettavan syy-seuraussuhteiden arvioinnin. Käytännössä tämä kuitenkin usein rikkoutuu esimerkiksi puuttuvien selittävien tekijöiden aiheuttaman harhan vuoksi (*omitted-variable bias*), jolloin tavanomaisesta regressiomallista tulee epäluotettava kausaalivaikutusten arvioinnissa. Tällaisissa tilanteissa endogeenisuuden aiheuttama harha voidaan korjata instrumenttimuuttujan avulla, mahdollistaen tarkemman kausaalipäättelyn.

Sekoittavat tekijät, kuten yksilön motivaatio tai älykkyys, voivat samanaikaisesti vaikuttaa sekä selittävään muuttujaan, kuten koulutukseen, että vasteeseen, kuten palkkaan. Tällöin selittävä muuttuja ja virhetermi ovat keskenään korreloituneita ja mallin tuottamat tulokset ovat harhaisia. Lisäksi käänteinen kausaliteetti, jossa vaikutuksen suunta on epäselvä, voi vääristää tulkintaa entisestään. Mittausvirheet ja puuttuvien tekijöiden aiheuttama harha voivat aiheuttaa ongelmia, jotka rikkovat eksogeenisuusoletuksen ja vaarantavat mallin luotettavuuden.

Instrumenttimuuttujamenetelmä tarjoaa tehokkaan työkalun näiden ongelmien ratkaisemiseksi. Se mahdollistaa kausaalivaikutusten eristämisen käyttämällä eksogeenista eli ulkoista muuttujaa, instrumenttia, joka vaikuttaa selittävään muuttujaan, mutta ei suoraan vasteeseen. Näin menetelmällä voidaan hallita sekoittavien tekijöiden, käänteisen kausaliteetin ja mittausvirheiden aiheuttamia haasteita, mikä tekee siitä keskeisen lähestymistavan monimutkaisissa kausaalianalyyseissä.

Instrumenttianalyysin käyttö ei kuitenkaan ole ongelmatonta. Sen keskeisenä haasteena on löytää sopiva instrumentti, joka täyttää korrelaatio- ja eksogeenisuusehdot. Cunningham (2021) [1] painottaa, että näiden ehtojen täyttäminen ei ole aina yksiselitteistä. Niiden arviointi edellyttää ennakkotietoa aiheesta, teoreettista perustelua sekä empiirisiä testejä.

Aluksi luvussa 2 käsitellään instrumenttimuuttujan luonnetta ja tämän valintaan johtavia ehtoja sekä oletuksia. Samassa esitellään instrumenttimuuttujamenetelmän keskeinen ongelma, heikon instrumentin tapaus, sekä F-testisuureen merkitys korrelaatioehdon varmistamiseksi, jotta heikolta instrumentilta vältytään.

Tämän jälkeen luvussa 3 tutustutaan kahteen tyypilliseen kausaalisuhteita käsittelevään menetelmään, tavalliseen pienimmän neliösumman regressioon (OLS) ja instrumenttianalyysin kaksivaiheiseen pienimmän neliösumman regressioon (2SLS). Menetelmiä vertaillessa esitellään kummankin mallin estimaattien keskivirheet.

Määritelmien ja mallin valinnan jälkeen luvussa 4 käydään läpi havainnollistava empiirinen esimerkki Cunninghamin (2021) [1] teoksesta. Esimerkin avulla havainnollistetaan, kuinka instrumenttimuuttuja (henkilön läheisyys 4-vuotiseen korkeakouluun) voi auttaa ratkaisemaan endogeenisuuteen liittyvän ongelman ja mahdollistaa luotettavamman kausaalipäättelyn, kun tutkitaan koulutuksen vaikutusta palkkaan.

Lopuksi, luvun 5 yhteenvedossa, käydään läpi muuttujan käytön hyödyt, haasteet ja rajoitukset sekä mitä tulee pitää mielessä instrumenttimuuttuja-analyysiä suorittaessa.

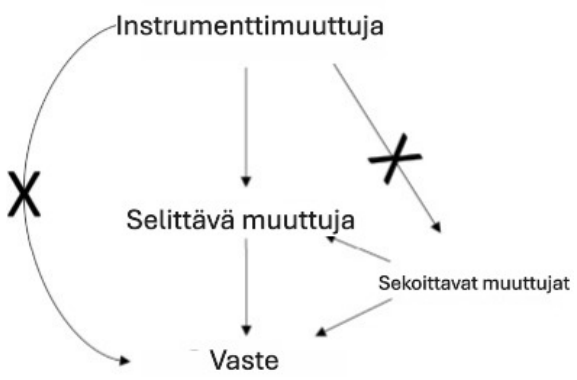
Tutkielmassa käytin pääsääntöisesti seuraavia lähteitä tiedon etsinnässä. Angristin (2009) Mostly harmless econometrics [2] teos oli kaikista kattavin teorian suhteen, kuten myös Wooldridgen (2010) Econometric analysis of cross section and panel data [3] ja Verbeekin (2017) A guide to modern econometrics [4]. Cunninghamin (2021) Causal inference: The Mixtape [1] tarjosi runsaasti esimerkkejä ja pohdintaa eri instrumenteista ja niiden käytöstä. Tiiviinä ja pääpiirteet kattavana johdantona aiheeseen toimi Schuetzen (2009) Economics 499: Honours seminar [5] luentomoniste. Tutkielman empiirinen esimerkki pohjautui Cardin (1995) Using Geographic Variation in College Proximity to Estimate the Return to Schooling [10] tunnettuun aineistoon, joka on yksi keskeisimmistä instrumenttimuuttuja-analyysin sovelluksista.

2 Instrumentin valinta kausaalisuuden tutkimiseksi

Ensimmäisessä luvussa tutustutaan kausaalipäättelyn perusperiaatteisiin ja haasteisiin. Instrumentin käytön intuitiosta on kerrottu pääsääntöisesti kaikissa lähteissä. Cunningham (2021) [1] avaa aihetta käytännönläheisesti esimerkkien kautta. Kausaalipäättelyn suhteita tarkastellaan suhdetaulun 1 avulla, joka on lähtöisin Mossin (2019) [9] puolikokeellisesta tutkimuksesta. Lisäksi käydään läpi instrumenttimuuttujan valintaa ohjaavat validointiehdot ja heikon instrumentin tapaus sekä F-testisuureen käyttö tämän heikon instrumentin välttämiseksi. [[2], [3], [4]]

Kausaaliväitteiden tulkinta on keskeinen tavoite tilastollisessa analyysissä, sillä se mahdollistaa syy-seuraussuhteiden ymmärtämisen pelkkien korrelaatioiden tarkastelun sijaan. Kausaalisuhteiden arviointi on kuitenkin haastavaa, koska selittävät muuttujat voivat olla monimutkaisesti kietoutuneita toisiinsa. Yleinen haaste on sekoittavan tekijän ongelma *(confounding)*, jossa kolmas, ei havaittu tekijä, vaikuttaa sekä syyhyn että seuraukseen. Toinen haaste on käänteinen kausaliteetti (*reverse causality*), jossa selittävän muuttujan (syy) ja vasteen (seuraus) suunta on epäselvä. Ongelmana voi myös olla Schuetzen (2009) [5] mainitsema puuttuvien selittävien tekijöiden aiheuttama harha (*omitted-variable bias*).

Yleensä tutkimuksen tavoitteena on arvioida selittävän muuttujan todellinen vaikutus vasteeseen siten, että vaikutuksen voidaan katsoa olevan kausaalinen. Tämä edellyttää menetelmiä, jotka kontrolloivat sekoittavien muuttujien vaikutuksia ja vähentävät käänteisen kausaliteetin sekä muiden endogeenisuudesta johtuvien harhojen vaikutusta. Tässä yhteydessä instrumenttimuuttujamenetelmä (*IV, instrumental variable*) tarjoaa tehokkaan työkalun kausaalivaikutusten arvioimiseen, erityisesti tilanteissa, joissa selittävä muuttuja on potentiaalisesti korreloitunut virhetermin kanssa.



Kuva 1: Suhdetaulu [9], nuolet kuvaavat kausaalisuhteita, rastilla merkityt nuolet kuvaavat sitä, että muuttujien välillä ei saa esiintyä kausaalisuhdetta.

Kausaalipäättelyn kannalta ongelmallisia ovat suhdetaulussa 1 nuolet jotka kulkevat sekoittavista muuttujista sekä selittävään muuttujaan että vasteeseen, aiheuttaen endogeenisuutta eli yhteyttä selittävän muuttujan ja mallin virhetermin välille.

Tämä yhteys rikkoo tavanomaisen regressioanalyysin eksogeenisuusoletuksen, mikä johtaa harhaisiin ja epäluotettaviin kausaalipäätelmiin.

Instrumenttimuuttuja tarjoaa ratkaisun tähän ongelmaan. Instrumentti vaikuttaa ainoastaan selittävään muuttujaan eikä suoraan vasteeseen tai sekoittaviin muuttujiin, toimien ulkoisena tekijänä. Käyttämällä instrumenttimuuttujaa voidaan eristää selittävän muuttujan vaihtelu, joka ei ole yhteydessä sekoittaviin tekijöihin, mahdollistaen selittävän muuttujan kausaalivaikutuksen estimoinnin vasteeseen. [1]

Tarkastellaan tunnettua yhteiskuntatieteellistä tutkimuskysymystä palkan ja koulutuksen suhteesta Wooldridgen (2010, s. 87) [3] teoksen pohjalta. Esimerkkiä voi havainnollistaa suhdetaulun 1 avulla. Voidaan olettaa, että koulutuksella ja palkalla on vahva positiivinen korrelaatio eli korkeasti koulutetut ansaitsevat usein enemmän. Tämä ei kuitenkaan automaattisesti tarkoita, että koulutus johtaa suoraan korkeampaan palkkaan. Taustalla voi olla muita sekoittavia tekijöitä, kuten kyvykkyys, vanhempien sosioekonominen asema tai motivaatio, jotka vaikuttavat sekä koulutukseen että palkkaan. Nyt pitää löytää koulutukseen liittyvä ulkoinen tekijä, esimerkiksi äidin koulutustausta, joka voi auttaa erottamaan kausaalivaikutuksen muista taustatekijöistä. Äidin koulutustaustan tulee siis vaikuttaa palkkaan pelkästään koulutuksen kautta.

Hyvän instrumentin löytäminen on usein haastavampaa kuin esimerkki antaa olettaa. Cunningham (2021) [1] painottaa, että hyvä instrumentti voi ja usein kuuluukin tuntua oudolta. Valinta vaatii aiempaa tietoa aiheesta, eikä instrumentin ja vasteen välinen yhteys ole aina ilmiselvä.

2.1 Validointiehdot ja oletukset

Instrumenttianalyysin peruslähtökohtana on usein oletus homogeenisista vaikutuksista, jolloin esimerkiksi koulutuksen vaikutus palkkaan oletetaan samaksi kaikille havainnoille. Käytännön aineistoissa vaikutukset voivat kuitenkin olla heterogeenisiä, eli vaihdella populaation sisällä. Tämä on tärkeää huomioida instrumenttien validoinnissa ja oletusten määrittelyssä.

Instrumentin on täytettävä seuraavat ehdot riittävän hyvin, jotta vältytään suurilta keskivirheiltä ja harhaanjohtavilta johtopäätöksiltä, joita heikot instrumenttimuuttujat aiheuttavat.

1. **Korrelaatioehto:** Angrist (2009) [2] on käyttänyt eksluusiorestriktio (*exclusion restriction*) termiä. Instrumentin Z on oltava riittävän vahvasti korreloitunut mallin selittävän endogeenisen muuttujan S kanssa, jotta se olisi relevantti

[1] Kliinisiä tutkimuksia käsittelevässä kirjallisuudessa selittävä endogeeninen muuttuja eli virhetermin kanssa korreloitunut muuttuja on hoito (*treatment*) ja vasteesta käytetään nimikettä lopputulos (*outcome*). [9]

instrumentti, eikä heikko. Matemaattisesti tämä ilmaistaan seuraavasti:

$$\mathrm{Cov}(Z, S) \neq 0. \tag{1}$$

2. **Eksogeenisuus**: Instrumentti Z ei saa korreloida virhetermin ε kanssa. Tämä tarkoittaa, että instrumentin vaikutus vasteeseen tapahtuu ainoastaan selittävän muuttujan kautta. Eksogeenisuus voidaan esittää muodossa:

$$\mathrm{Cov}(Z, \varepsilon) = 0. \tag{2}$$

Eksogeenisuusehtoa voidaan arvioida teoreettisesti esimerkiksi suhdetaulua 1 tarkastelemalla ja aiempaa kirjallisuutta tutkimalla. Eksogeenisuutta voidaan tarkastella myös empiirisesti yli-identifikointitesteillä (*over-identification tests*), jossa testataan, ovatko instrumentit eksogeenisia eli korreloimattomia virhetermin kanssa, kun instrumenttien määrä ylittää endogeenisten muuttujien määrän. [2]

2.2 Heikko instrumentti validoinnin haasteena

Heikko instrumentti viittaa tilanteeseen, jossa instrumentin ja endogeenisen muuttujan välinen yhteys on heikko. Eli korrelaatioehto 1 ei täyty riittävän hyvin. Tämä voi johtaa menetelmän tehottomuuteen, sillä seurauksena on instrumenttimuuttujaanalyysissä käytetyn kaksivaiheisen pienimmän neliösumman menetelmän (2SLS) estimaatin harha.

Instrumentin ollessa liian heikko, se ei pysty riittävästi erottamaan eksogeenista vaihtelua endogeenisestä vaihtelusta. Instrumentti ei tällöin tuota tarpeeksi vaihtelua selittävään muuttujaan, jolloin kausaalisuhteiden arviointi epäonnistuu. Heikko instrumentti johtaa epävakaisiin tuloksiin erityisesti pienissä aineistoissa.

Yksi heikko instrumentti voi johtaa suureen varianssiin ja epäluotettaviin estimaatteihin. Jos instrumentti on täysin satunnainen eikä sisällä informaatiota selittävästä muuttujasta, estimaatti on pelkkää satunnaiskohinaa. Useiden heikkojen instrumenttien tapaukessa estimaatit ovat systemaattisesti harhaisia. Heikot instrumentit eivät sisällä informaatiota, vaan ne tuottavat palkkää satunnais kohinaa regressioon. Harha vetää estimaatit systemaattisesti kohti pienimmän neliösumman menetelmällä saadun tavallisen regressioestimaatin arvoa, tehden instrumenttianalyysistä hyödyttömän. [[6] [7]]

2.3 F-testisuure korrelaation varmistajana

Ensimmäisen vaiheen F-testisuure on tärkeä työkalu instrumentin validiteetin arvioinnissa. Se mittaa instrumentin (Z) ja endogeenisen selittävän muuttujan (S) välistä korrelaatiota. Kahta regressiomallia vertaileva F-testisuure määritellään seuraavasti [4], 5.2.4 ja [3], 5.2.4 mukaan:

$$F = \frac{\widehat{SSR}_r - \widehat{SSR}_{ur}}{q} \Big/ \frac{\widehat{SSR}_{ur}}{N - K},$$

jossa:

- q on vapausasteiden ero rajoitetun ja rajoittamattoman mallin välillä.
- $N - K$ on rajoittamattoman mallin residuaalivapausasteet (*df*), jossa N on havaintojen määrä ja K parametrit rajoittamattomassa mallissa.
- $\widehat{SSR}_r$: Rajoitetun mallin jäännösten neliösumma (*reduced model*):

  $$S_{\text{r,i}} = \hat{\beta}_0 + \hat{\beta}_1 X_i + \epsilon_i,$$

  jossa:

  - $S_{\text{r,i}}$ on rajoitetun mallin havaintoa i vastaava riippuva muuttuja.
  - X_i on havaintoa i vastaava eksogeenisten kontrollimuuttujien vektori (esimerkiksi kokemus, vanhempien etninen tausta).
  - ϵ_i on havaintoa i vastaava virhetermi (esimerkiksi mittausvirheet, satunnaiset häiriöt ja muut muuttujat).
  - Indeksi i viittaa aina yhteen yksittäiseen havaintoon.
- $\widehat{SSR}_{ur}$: Rajoittamattoman mallin jäännösten neliösumma (*unrestricted model*).

  $$S_{\text{ur,i}} = \hat{\beta}_0 + \hat{\beta}_1 Z_i + \hat{\beta}_2 X_i + \epsilon_i,$$

  jossa Z_i on yhden instrumentin tapauksessa skalaari havaintoa i vastaava instrumentti (esimerkiksi vanhempien koulutus tai koulumatkan lyhyys).

Mallit estimoidaan tavanomaisesti pienimmän neliösumman menetelmällä (OLS), joka minimoi jäännösten neliösumman RSS (*residual sum of squares*):

$$RSS = \sum_{i=1}^{N} (S_i - \hat{S}_i)^2,$$

jossa:

- S_i on havaittu endogeeninen muuttuja (esimerkiksi koulutus).
- $\hat{S}_i$ on mallin ennustama arvo instrumentin avulla.

Tämä antaa pienimmän neliösumman estimaatit parametreille $\beta_0, \beta_1, \ldots, \beta_K$.

Testin tarkoituksena on arvioida instrumentin tilastollista merkitsevyyttä kaksivaihesen pienimmän neliösumman (2SLS) menetelmän ensimmäisessä vaiheessa. Suuri F-testisuure viittaa siihen, että instrumentti selittää merkittävän osan kaksivaihesen pienimmän neliösumman menetelmän vasteen (S) vaihtelusta.

Cunninghamin (2021) [1] mukaan F-testisuureen arvon tulisi olla vähintään 10, jotta instrumentti katsotaan riittävän vahvaksi. Mikäli F-testisuure jää alle tämän raja-arvon, instrumentti voi olla liian heikko ja menetelmän tulokset voivat olla harhaisia.

F-testisuure ei ole suoraan sovellettavissa kaksivaiheisen pienimmän neliösumman (2SLS) menetelmän toiseen vaiheeseen. Wooldridgen (2010) [3] mukaan tämä johtuu siitä, että endogeenisen muuttujan arvot ovat ensimmäisessä vaiheessa estimoituja, jolloin ne sisältävät estimointivirhettä. Tämä rikkoo eksogeenisuuden oletuksen, jonka mukaan selittävien muuttujien ei tulisi korreloida virhetermin kanssa. Lisäksi ennustettujen arvojen käyttäminen lisää epävarmuutta, jota tavallinen F-testi ei ota huomioon. Tästä syystä F-testisuure ei enää noudata tunnettua F-jakaumaa, mikä tekee sen käytöstä ongelmallista toisen vaiheen hypoteesitesteissä. Vaihtoehtoisia menetelmiä toisen vaiheen hypoteesien testaamiseen ovat esimerkiksi Wald-testi ja yli-identifikaatiotestit.

3 Instrumenttimuuttuja-analyysin toteutus

Tässä luvussa tarkastellaan kahta pääasiallista menetelmää kausaalisuhteiden arvioimiseksi: tavallinen pienimmän neliösumman regressio (OLS) ja kaksivaiheinen pienimmän neliösumman regressio (2SLS). Luvun menetelmät ja päätelmät on johdettu Angristin (2009) [2] ja Wooldridgen (2010) [3] teosten pohjalta sekä Schuetzen (2009) [5] luontomonisteen pohjalta.

3.1 OLS-regressio: Tavallinen pienimmän neliösumman menetelmä

OLS-regressio (*ordinary least squares regression*) on yksinkertainen ja laajasti käytetty menetelmä, jolla pyritään estimoimaan muuttujien välisten suhteiden suuruutta. OLS-malli voisi olla seuraavanlainen:

$$Y_i = \beta_0 + \beta_1 S_i + \beta_2 X_i + \epsilon_i, \tag{3}$$

jossa:

- Y_i on havaintoa i vastaava vaste (esimerkiksi logaritmoitu palkka).
- S_i on havaintoa i vastaava endogeeninen muuttuja (esimerkiksi koulutus).
- X_i on havaintoa i vastaava eksogeenisten kontrollimuuttujien vektori (esimerkiksi kokemus ja etninen tausta).
- ϵ_i on havaintoa i vastaava virhetermi (esimerkiksi motivaatio ja mittausvirheet).
- Indeksi i viittaa aina yhteen yksittäiseen havaintoon.

OLS-estimaatti määritellään matriisimuodossa seuraavasti:

$$\hat{\beta}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y, \tag{4}$$

jossa:

- $\hat{\beta}_{\text{OLS}}$ on estimaattori, joka antaa arvioidut kertoimet mallille, eli estimaatit parametreille $\beta_0, \beta_1, \ldots, \beta_K$.
- $\mathbf{X}$ on havaintomatriisi, joka sisältää kaikki selittävät muuttujat (endogeeniset ja eksogeeniset). Sen koko on $N \times (K+1)$, jossa N on havaintojen lukumäärä ja $K+1$ selittävien muuttujien lukumäärä vakiotermin kanssa:

$$\mathbf{X} = \begin{bmatrix} 1 & S_1 & X_{11} & X_{12} & \ldots & X_{1K} \\ 1 & S_2 & X_{21} & X_{22} & \ldots & X_{2K} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & S_N & X_{N1} & X_{N2} & \ldots & X_{NK} \end{bmatrix}.$$

- $\mathbf{X}^T$ on $\mathbf{X}$:n transpoosi.

- y on kaikkien havaintojen riippuvan muuttujan arvoista muodostettu vektori:

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}.$$

OLS-estimaatti olettaa, että kaikki selittävät muuttujat X (esimerkiksi koulutus, kokemus jne.) ovat eksogeenisiä, eli ne eivät ole korreloituneet virhetermin ϵ kanssa.

Jos kaikki muuttujat olisivat eksogeenisiä, OLS-regressio olisi toimiva malli. Todellisuudessa tämä oletus ei kuitenkaan aina toteudu ja saamme usein tilanteen, jossa jotkut muuttujat, kuten koulutus, ovat korreloituneet virhetermin kanssa. Tällöin muuttujat ovat endogeenisiä ja OLS antaa harhaisia ja epäluotettavia tuloksia. [[2], luku 3.1 ja [3], luku 4.]

3.2 2SLS-regressio: Kaksivaiheinen pienimmän neliösumman menetelmä

2SLS-regressiota (*two-stage least squares regression*) käytetään kun muuttujat ovat endogeenisiä ja OLS-menetelmä ei ole luotettava. 2SLS-regressio mahdollistaa kausaalisuuden arvioinnin instrumenttimuuttujien avulla. Alaluvun päätelmät on johdettu Angristin (2009, s. 83-91) [2], sekä Wooldridgen (2010, s. 83-94.) [3] teoksista.

2SLS-malli koostuu kahdesta vaiheesta:

Vaihe 1: Instrumentointi Ensimmäisessä vaiheessa estimoidaan endogeeninen muuttuja S instrumenteilla ja eksogeenisilla kontrollimuuttujilla:

$$\widehat{S}_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 X_i, \tag{5}$$

jossa:

- $\widehat{S}_i$ on instrumenttimuuttujan avulla estimoitu arvio endogeenisestä muuttujasta,
- Z_i on instrumenttimuuttuja (esimerkiksi lähellä yliopistoa asuvat) ja
- X_i on eksogeenisten kontrollimuuttujien vektori (esimerkiksi kokemus ja etninen tausta).

Vaihe 2: Regressio estimoidulla muuttujalla Toisessa vaiheessa alkuperäinen regressiomalli estimoidaan käyttämällä ensimmäisessä vaiheessa saatuja ennustettuja arvoja $\widehat{S}_i$ endogeeniselle muuttujalle:

$$Y_i = \beta_0 + \beta_1 \widehat{S}_i + \beta_2 X_i + \delta_i, \tag{6}$$

jossa:

- $\widehat{S}_i$ on instrumenttien avulla estimoitu arvo endogeeniselle muuttujalle (esimerkiksi koulutus). Sen avulla pyritään poistamaan korrelaatio virhetermin kanssa.

- X_i on eksogeenisten kontrollimuuttujien vektori (esimerkiksi kokemus ja ikä).

- δ_i on virhetermi, joka sisältää esimerkiksi mittausvirheitä ja muita selittämättömiä tekijöitä.

Instrumenttimuuttujamenetelmän estimaatti määritellään seuraavasti:

$$\hat{\beta}_{\text{2SLS}} = (\mathbf{Z}^T\mathbf{X})^{-1}\mathbf{Z}^T y, \tag{7}$$

jossa:

- $\hat{\beta}_{\text{SLS}}$ on estimaattori, joka antaa arvioidut kertoimet 2SLS-mallille, eli estimaatit parametreille $\beta_0, \beta_1, \ldots, \beta_K$.

- $\mathbf{Z}$ on instrumenttien havaintomatriisi. Sen koko on $N \times L$, jossa N on havaintojen määrä ja on mahdollista, että $L \geq K + 1$ eli instrumenttien määrä on suurempi kuin selittävien muuttujien määrä.

Empiirisessä esimerkissä on käytetty vain yhtä instrumenttia eli $L = 1$, tämän seurauksena Z on $N \times 1$ -kokoinenen vektori.

Määritelmä kuvaa, kuinka instrumenttien $\mathbf{Z}$ avulla estimoidaan eksogeeninen vaihtelu selittävistä muuttujista $\mathbf{X}$, mikä mahdollistaa kausaalipäättelyn. Angristin (2009) [2] mukaan 2SLS-estimaatti on asymptoottisesti harhaton, mutta pienissä otoksissa se voi olla harhainen erityisesti, jos instrumentti on heikko. Tämä harha kuitenkin vähenee otoskoon kasvaessa, jolloin estimaatti lähestyy todellista parametriarvoa. [[2], luku 4 ja [3], luku 5]

3.3 Estimaattien keskivirheet

Regressiomallien parametriestimaattien luotettavuutta voidaan arvioida niiden keskivirheiden avulla. Keskivirhe (*standard error, SE*) mittaa estimaatin keskihajontaa otoksesta toiseen ja kuvaa sen tarkkuutta. Tarkastellaan keskivirheiden laskentaa kummassakin regressiomenetelmässä. Alaluvun päätelmät on johdettu Wooldridgen (2010, 5.2.2) [3] teoksesta ja Schuetzen (2009, s. 6-12.) luentomonisteesta [5].

3.3.1 OLS-mallin keskivirhe

OLS-mallin estimaatin keskivirhe voidaan laskea matriisimuodossa seuraavasti:

$$\text{Var}(\hat{\beta}_{\text{OLS}}) = \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1},$$

ja yksittäisen estimaatin keskivirhe seuraavasti:

$$\text{SE}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}.$$

Näissä $\hat{\sigma}^2$ on virhetermin varianssin estimaatti:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{\epsilon}_i^2}{N - K}.$$

3.3.2 2SLS-mallin keskivirhe

2SLS-mallissa keskivirheen laskenta on monimutkaisempaa, koska instrumenttimuuttujien käyttö vaikuttaa estimaattorin varianssiin. Matriisimuodossa 2SLS-estimaattorin varianssi on:

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{2SLS}}) = \hat{\sigma}^2(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1},$$

jossa $\hat{\mathbf{X}}$ on instrumenttien avulla estimoidut selittävät muuttujat eli $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{X}$.

Yksittäisen estimaatin keskivirhe on:

$$\text{SE}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2[(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}]_{jj}},$$

jossa $\hat{\sigma}^2$ on jäännösvarianssin estimaatti 2SLS-mallissa:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{\delta}_i^2}{N - K}.$$

Tässä $\hat{\delta}_i = Y_i - \mathbf{X}_i^T\hat{\beta}_{\text{2SLS}}$ on 2SLS-mallin virhetermi havainnolle i.

2SLS-mallin keskivirhe on tyypillisesti suurempi kuin OLS-mallin, koska instrumenttien käyttö lisää monimutkaisuutta, vähentäen mallin tehokkuutta ja lisäten estimaattorin epävarmuutta. Mikäli instrumentti on heikko, keskivirhe kasvaa merkittävästi, mikä voi johtaa hyvin epätarkkoihin estimaatteihin. Tästä syystä voimakkaiden instrumenttien valinta on olennaista 2SLS-mallin luotettavuuden kannalta.

Tämä osoittaa, että vaikka 2SLS korjaa OLS:n endogeenisuudesta johtuvan harhan, se tekee tämän tehokkuuden kustannuksella. Tämä tehokkuuden menetys voidaan kuitenkin hyväksyä, jos 2SLS tarjoaa luotettavamman kausaalipäättelyn kuin OLS tilanteissa, joissa eksogeenisuusolettamus ei pidä paikkaansa.

4 Empiirinen esimerkki: College in the County

Tässä luvussa on seurattu Cunninghamin (2021) [1] teoksen luvun 7.7.1 College in the county esimerkkiä Cardin (1995) [10] tutkimuksen aineistosta. Luvussa suoritetun analyysin R-koodit löytyvät liitteistä 5.

4.1 Esimerkin tausta ja asetelma

Esimerkissä tutkitaan kuinka korkeakoulutuksen saatavuus vaikuttaa yksilön tuloihin. Erityisesti tutkitaan kausaalivaikutusta korkeakoulun läheisyyden perusteella. Tämä asetelma on klassinen esimerkki tilanteesta, jossa pääasiallinen kiinnostuksen kohde on estimoida koulutuksen kausaalivaikutus. Koulutuksen määrää voivat ohjata yksilön ominaisuudet, kuten motivaatio tai kyvykkyys, jotka voivat aiheuttaa endogeenisuutta. Tällöin tavanomaiset regressiomenetelmät voivat antaa harhaisia tuloksia.

Data on Yhdysvalloissa suoritetusta NLS:n kansallisesta nuorten miesten kohortti pitkittäistutkimuksesta. Tiedot kerättiin 14-24 -vuotiailta miehiltä vuosina 1966-1981. Cardin (1995) [10] alkuperäisen aineiston koko on 3010 ja muuttujia on 37. Analyysissa on käytetty taulukossa 1 esitellyt muuttujat.

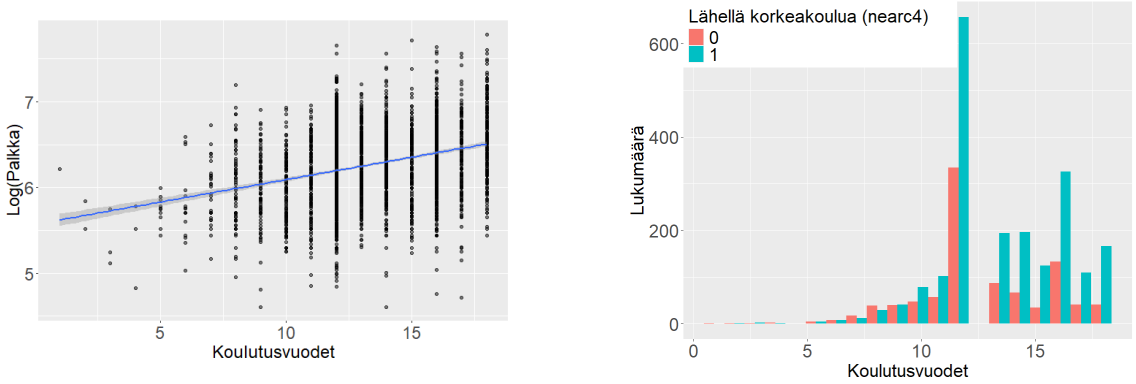
Taulukko 1: Muuttujat ja niiden mitta-asteikko Card (1995) [10]

Koulutuksen ja palkan välinen korrelaatio 2a	0.294
Koulutuksen ja instrumentin välinen korrelaatio 2b	0.136
Muuttuja	**Mitta-asteikko**
lwage: logaritmoitu palkka, senttejä tunnissa	Jatkuva
nearc4: piirikunnassa on 4-vuotinen korkeakoulu	Dummy (0/1)
educ: koulutus vuosina	Jatkuva
exper: kokemus vuosina	Jatkuva
black: tummaihoisuus	Dummy (0/1)
south: etelävaltiosta	Dummy (0/1)
married: naimisissa	Dummy (0/1)
smsa: urbaanisuus	Dummy (0/1)

Dummyt: 1 = Kyllä, 0 = Ei.

4.2 Instrumentin valinta ja validointi

Alkuperäisessä Cardin (1995) [10] tutkimuksessa tarkasteltiin useita muuttujia, mutta 4-vuotisen korkeakoulun olemassa piirikunnassa valikoitui jatko-opintoihin hakeutumisen instrumentiksi. Cunninghamin (2021) [1] mukaan yksi mahdollinen peruste tälle valinnalle on se, että korkeakoulun läheisyys saattoi mahdollistaa opiskelun kotipaikkakunnalta käsin, mikä puolestaan vähensi opiskeluun liittyviä kustannuksia. Lisäksi korkeakoulun sijainti asuinpiirikunnassa saattoi tarjota muita etuja, kuten



(a) Koulutus ja palkka

(b) Koulutus ja instrumentti

Kuva 2: Muuttujien tarkastelua visuaalisesti

mahdollisuuden pysyä lähellä olemassa olevia sosiaalisia verkostoja, kuten perhettä ja ystäviä, jotka voivat toimia opiskelijan tukena päätöksenteossa.

On kuitenkin huomattava, että osa nuorista hakeutuu korkeakoulutukseen riippumatta sen maantieteellisestä läheisyydestä. Toiset taas jättävät hakeutumatta, vaikka lähistöllä olisi oppilaitos. Instrumenttimuuttujamenetelmässä erityisen kiinnostava ryhmä muodostuu niistä yksilöistä, joiden päätökseen jatkokoulutuksesta instrumentti vaikuttaa ratkaisevasti (*compliers*). Tämä compliers-ryhmä koostuu henkilöistä, jotka jatkavat korkeakouluopintoihin ainoastaan siksi, että heidän asuinpiirikunnassaan sijaitsee korkeakoulu. Ilman tätä tekijää he eivät todennäköisesti olisi hakeutuneet jatkokoulutukseen.

Instrumenttimuuttujamenetelmä arvioi LATE-estimaatin (*local average treatment effect*), joka mittaa koulutuksen vaikutusta compliers-ryhmälle, eli niille, joiden koulutusvalinnat muuttuvat instrumenttimuuttujan vaikutuksesta. LATE-estimaatti ei kuitenkaan välttämättä vastaa koko populaation keskimääräistä vaikutusta (*average treatment effect, ATE*), sillä koulutuksen hyödyt voivat vaihdella eri ryhmien kesken. Tämä heterogeenisyys on tärkeä huomioida, sillä compliers-ryhmän saama hyöty voi poiketa esimerkiksi always-takers-ryhmästä, joka hakeutuu korkeakoulutukseen joka tapauksessa.

Sopivaa instrumenttia valitessa on tärkeää varmistaa eksogeenisuusehdon 2 ja korrelaatioehdon 1 toteutuminen. Eksogeenisuusehto edellyttää, että korkeakoulun läheisyys on riippumaton muista tuloihin vaikuttavista tekijöistä. Tätä voidaan arvioida teoreettisesti esimerkiksi suhdetaulun avulla sekä aiempaa kirjallisuutta tarkastelemalla.

Instrumentin validoinnin varmistamiseksi ja heikon instrumentin välttämiseksi tulee korrelaatioehto testata empiirisesti regressioanalyysin F-testillä:

- Nollahypoteesi (H_0): Korkeakoulun läheisyydellä ei ole tilastollisesti merkitsevää yhteyttä koulutustasoon.

- Vastahypoteesi (H_1): Korkeakoulun läheisyydellä on tilastollisesti merkitsevä yhteys koulutustasoon.

F-testisuure mittaa, selittävätkö malliin sisällytetty instrumentti (*nearc4*) koulutustasoa tilastollisesti merkitsevällä tavalla. Jos F-testisuureen arvo on yli 10, hylkäämme H_0 ja tulkitsemme, että korrelaatioehto täyttyy Cunninghamin (2021) [1] mukaan.

4.3 Regressiomallit OLS vs. 2SLS

Cunninghamin (2021) [1] teoksessa käydyn esimerkin mukaan Card (1995) [10] halusi tutkia koulutuksen vaikutusta palkkaan seuraavalla regressioyhtälöllä:

OLS-regressio: estimoi suoran yhteyden koulutuksen ja palkkojen välillä ilman instrumentointia:

$$Y_i = \beta_0 + \beta_1 S_i + \beta_2 X_i + \epsilon_i, \tag{3}$$

jossa:

- Y_i on logaritmoidut tulot (*lwage*),
- S_i koulutusvuodet (*educ*),
- X_i on joukko eksogeenisia kontrollimuuttujia, kuten työkokemus (*exper*), etninen tausta (*black*), alueellinen sijainti (*south*), avioliittotila (*married*) sekä urbaanisuus (*smsa*) ja
- ϵ_i on virhetermi (esimerkiksi motivaatio ja mittausvirheet).

Oletetaan siis, että kontrollimuuttujat eivät ole korreloituneita virhetermin ϵ_i kanssa, joka voi sisältää esimerkiksi yksilön kyvykkyyden tai motivaation kaltaisia tekijöitä.

Kun koulutukseen liittyviä selittämättömiä tekijöitä jää virhetermiin ja ne ovat korreloituneita koulutuksen kanssa, eksogeenisuusehto ei toteudu. Tämä tekee koulutuksen kertoimen estimoinnista harhaista. Tästä syystä Card (1995) [10] ehdotti eksogeenisen instrumenttimuuttuja (*nearc4*) käyttöä, joka osoittaa, asuuko henkilö lähellä nelivuotista korkeakoulua. Tämä instrumentti korreloi koulutuksen, mutta ei suoraan palkkojen kanssa, mikä mahdollistaa koulutuksen kausaalivaikutuksen arvioinnin.

2SLS-regressio: Kaksivaiheisessa mallissa koulutuksen kausaalivaikutus arvioidaan instrumentointia käyttäen.

1. **Ensimmäinen vaihe:** Koulutuksen regressio instrumentilla ja kontrollimuuttujilla:

   $$S_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 X_i + \delta_i, \tag{5}$$

   jossa Z_i on instrumentti (*nearc4*).

2. **Toinen vaihe:** Ensimmäisen vaiheen ennustetun koulutuksen käyttö päämallin estimoinnissa:
$$Y_i = \beta_0 + \beta_1 \widehat{S}_i + \beta_2 X_i + \delta_i, \tag{6}$$
jossa $\widehat{S}_i$ on on instrumentoidut koulutusarvot.

Tämän menetelmän avulla korjataan endogeenisuuden aiheuttama harha, eli pystytään erottamaan eksogeeninen vaihtelu koulutuksessa ja arvioimaan sen kausaalivaikutus palkkoihin. [[2], luku 4 ja [3], luku 5]

4.4 Tulosten raportointi

Tarkistellaan regressiomallien antamia tuloksia taulukosta 2, regressiomallien tulokset ja päätelmät F-tesisuureesta ovat yhteneviä Cunninghamin (2021) [1] kanssa.

OLS-tulokset: OLS-mallin kertoimen arvo (ATE) on 0.071 ja se on tilastollisesti merkitsevä ($p<0.01$). Tämä tarkoittaa, että yksi lisävuosi koulutusta kasvattaa palkkaa keskimäärin 7.1 %. Tulokset voivat kuitenkin olla harhaisia endogeenisuuden vuoksi.

Ensimmäisen vaiheen tulokset: Ensimmäisessä vaiheessa pyritään ennustamaan palkan sijaan koulutuksen määrää instrumentin (nearc4) ja muiden eksogeenisten tekijöiden avulla. Korkeakoulun läheisyys lisää koulutuksen vuosien määrää keskimäärin 0.327 vuotta, kertoimen arvo on tilastollisesti merkitsevä ($p<0.01$).

2SLS-tulokset: 2SLS-mallin kertoimen arvo ($LATE$) on 0.124 ja se on tilastollisesti merkitsevä ($p<0.05$). Tämä tarkoittaa, että yksi lisävuosi koulutusta kasvattaa palkkaa keskimäärin 12.4 %. Instrumentoinnin avulla korjattu tulos viittaa siihen, että OLS-mallin alhaisempi kerroin johtui endogeenisuudesta.

F-testisuure: F-testisuure lasketaan 2SLS-mallin ensimmäisen vaiheen 5 yhteydessä, koska sen avulla voidaan arvioida, kuinka hyvin instrumentti selittää endogeenista muuttujaa. Tässä tapauksessa ensimmäisen vaiheen F-testisuure oli 15.767, joka on suurempi kuin Cunninghamin (2021) [1] käyttämä kynnysarvo 10. Tämä viittaa siihen, että hylkäämme nollahypoteesin (H_0) eli käytetty instrumentti ei ole heikko, jolloin korrelaatioehto 1 toteutuu.

R^2 ja korjattu R^2: OLS-malli selittää noin 30.5 % palkkavaihtelusta, kun taas 2SLS selittää 25.1 %. Pienempi selitysaste on odotettavissa, koska 2SLS käyttää vain instrumentin selittämää vaihtelua.

Jäännösten keskihajonta (RSE): OLS-mallin RSE on 0.370, kun taas 2SLS-mallin RSE on 0.384. Instrumentointi voi lisätä mallin varianssia, mutta kausaalitulkinnan kannalta 2SLS on luotettavampi, kun instrumentti on validi.

Taulukko 2: OLS ja 2SLS tulokset 5

	Vaste: Log Palkka (lwage)		
	OLS (1)	Ensimmäinen vaihe (2)	2SLS (3)
Koulutus (educ)	0.071***		0.124**
	(0.003)		(0.050)
Kokemus (exper)	0.034***	-0.404***	0.056***
	(0.002)	(0.009)	(0.020)
Tummaihoisuus (black)	-0.166***	-0.948***	-0.116**
	(0.018)	(0.091)	(0.051)
Etelävaltiosta (south)	-0.132***	-0.297***	-0.113***
	(0.015)	(0.079)	(0.023)
Naimisissa (married)	-0.036***	-0.073***	-0.032***
	(0.003)	(0.018)	(0.005)
Urbaanisuus (smsa)	0.176***	0.421***	0.148***
	(0.015)	(0.085)	(0.031)
Korkeakoulun läheisyys (nearc4)		0.327***	
		(0.082)	
Vakiotermi (constant)	5.063***	16.831***	4.162***
	(0.064)	(0.131)	(0.850)
Ensimmäisen vaiheen F-testi		15.77	
Havainnot	3,003	3,003	3,003
R^2	0.305	0.477	0.251
Korjattu R^2	0.304	0.476	0.250
Residuaalien keskivirhe (df = 2996)	0.370	1.937	0.384
OLS-mallin F-testisuure	219.153***		

Suluissa olevat arvot kuvaavat kertoimien keskivirheitä. Tähdet osoittavat kertoimien tilastollista merkitsevyyttä: *p<0.1; **p<0.05; ***p<0.01.

5 Yhteenveto

Tutkielmassa tarkastellaan regressiomallien käyttöä kausaalisuhteiden arvioinnissa, erityisesti tilanteissa, joissa endogeenisuus voi vääristää estimaatteja. OLS-malli on yksinkertainen ja tehokas, mutta sen luotettavuus heikkenee endogeenisuuden vaikutuksesta. Tällöin instrumenttimuuttujamenetelmä (2SLS) tarjoaa tavan saada luotettavampia estimaatteja, korjaamalla virhetermiin korreloituneiden muuttujien aiheuttamat harhat. Menetelmän luotettavuus ja tehokkuus riippuvat ratkaisevasti oikean instrumentin valinnasta.

Sopivan instrumenttimuuttujan valinta on 2SLS-menetelmän toimivuudessa kannalta ehdoton. Instrumentin tulee täyttää validointiehdot, kuten korrelaatio 1 - ja eksogeenisuusehdot 2, jotta menetelmä tuottaa luotettavia tuloksia. Heikot instrumentit voivat johtaa suuriin keskivirheisiin ja vääristyneisiin estimaatteihin,

jolloin menetelmästä tulee epäluotettava. Heikkojen instrumenttien tunnistamiseksi käytetään 2SLS-regression ensimmäisen vaiheen F-testisuuretta, jonka kynnysarvoksi on suositeltu 10.

Vaikka 2SLS-menetelmä korjaa OLS:n endogeenisuuden aiheuttamat harhat, se voi kasvattaa estimaattien varianssia ja heikentää tilastollista merkitsevyyttä. Jos instrumentti on heikko, tämä ongelma korostuu entisestään, sillä heikot instrumentit lisäävät mallin monimutkaisuutta ja vähentävät sen tehokkuutta. Näin ollen instrumentin huolellinen valinta on ratkaiseva tekijä, joka määrittää 2SLS-menetelmän luotettavuuden ja tehokkuuden kausaalianalyysissä.

Empiirinen analyysi koulutuksen vaikutuksesta tuloihin havainnollistaa näitä menetelmällisiä haasteita. Instrumentin vahvuus tarkistetaan F-testisuureella, jonka jälkeen tuloksia vertaillaan eri malleilla. OLS-mallissa koulutuksen vaikutus arvioidaan positiiviseksi, mutta endogeenisuus (esimerkiksi kyvykkyyden tai motivaation vaikutus) voi tehdä arvioinnista harhaisen. 2SLS-malli korjaa tämän ongelman, mikä johtaa suurempaan mallin kertoimen arvoon. Tämä viittaa siihen, että OLS aliarvioi koulutuksen todellista vaikutusta, vähentäen harhaa, mutta kasvattaen varianssia. Tämä varianssin kasvu näkyy tilastollisen merkitsevyyden laskemisena taulukon 2 muuttujien kertoimissa.

Yhteenvetona voidaan todeta, että regressiomallien valinta riippuu tutkimuskysymyksestä ja aineiston ominaisuuksista. OLS-malli voi olla käyttökelpoinen, mutta kausaalipäättely voi olla virheellistä, jos estimaatit eivät ole luotettavia muuttujien endogeenisuuden vuoksi. 2SLS tarjoaa tärkeän välineen harhan hallintaan, mutta sen käyttö edellyttää huolellista instrumenttien valintaa. Vaikka 2SLS-menetelmä saattaa johtaa suurempiin estimaattien variansseihin ja siten vähemmän tehokkaisiin tuloksiin, se voi tuottaa luotettavampia tuloksia kausaalisuhteiden arvioinnissa, erityisesti suurilla otoksilla. Menetelmällä voidaan hallita sekoittavien tekijöiden, käänteisen kausaliteetin ja mittausvirheiden aiheuttamia haasteita, mikä tekee siitä keskeisen lähestymistavan monimutkaisissa kausaalianalyyseissä.

Viitteet

[1] Cunningham, S. (2021). Causal inference: The mixtape. Yale university press.

[2] Angrist, J. D., Pischke, J. S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton university press.

[3] Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.

[4] Verbeek, M. (2017). A guide to modern econometrics. John Wiley Sons.

[5] Schuetze, H. (2009). Economics 499: Honours seminar. University of Victoria.

[6] Stock, J. H., Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews J. H. Stock (Eds.), Identification and inference for econometric models: Essays in honor of Thomas Rothenberg (pp. 80–108). Cambridge University Press.

[7] Andrews, I., Stock, J. H., Sun, L. (2018). Weak instruments in IV regression: Theory and practice (Working Paper). Harvard University.

[8] Juha Alho, Elja Arjas, Juha Karvanen, Lasse Leskelä, Esa Läärä ja Pekka Pere (2023). Tilastotieteen sanasto. Verkkoversio 9.4.2023. Suomen Tilastoseura. https://sanasto.tilastoseura.fi/.

[9] Moss, H. A., Melamed, A., Wright, J. D. (2019). Measuring cause-and-effect relationships without randomized clinical trials: Quasi-experimental methods for gynecologic oncology research. Gynecologic Oncology, 153(1), 20–26.

[10] Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In L. N. Christofides, E. K. Grant, R. Swidinsky (Eds.), Aspects of labor market behaviour: Essays in honour of John Vanderkamp. University of Toronto Press.

Liitteet

R-koodiliite sisältää analyysin Cardin (1995) tutkimuksen aineistosta [10]. R-koodi perustui Cunninghamin (2021, 7.7.1) [1] teoksessa jaettuun R-koodiin, ja olen lisännyt siihen täydennyksiä, kuten F-testisuureen manuaalisen laskun, ensimmäisen vaiheen tulokset ja visuaalisia elementtejä. Analyysissä käytin seuraavia R-paketteja: AER (instrumenttimuuttuja analyysin työkaluihin), haven (stata-muotoisen aineiston tuontiin R:ään), tidyverse (aineiston käsittelyyn, muokkaamiseen ja visualisointiin) ja stargazer (regressiotulosten esittämiseen taulukkomuodossa).

```
library(AER)
library(haven)
library(tidyverse)
library(stargazer)

# Tietojen latausfunktio
read_data <- function(df) {
    full_path <-
    paste0("https://github.com/scunning1975/mixtape/raw/master/", df)
    read_dta(full_path)
}

# Aineisto
card_data <- read_data("card.dta")
```

```r
# Maaritellaan muuttujat
Y1 <- card_data$lwage
Y2 <- card_data$educ
X1 <- cbind(card_data$exper, card_data$black, card_data$south,
            card_data$married, card_data$smsa)
Z <- card_data$nearc4

# Korrelaatiot
cor_educ_lwage  <- cor(Y2, Y1, use = "complete.obs", method = "spearman")
cor_nearc4_educ <- cor(Z, Y2, use = "complete.obs", method = "spearman")

# Visuaalisointia: Muuttujien tarkastelua visuaalisesti (kuva 2)
# kuvaaja: Koulutus vs. Log(Palkka)
ggplot(card_data, aes(x = educ, y = lwage)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  labs(x = "Koulutusvuodet", y = "Log(Palkka)") +
  theme(
    text = element_text(size = 25),
    axis.title = element_text(size = 20),
    axis.text = element_text(size = 20),
    plot.title = element_text(size = 20, face = "bold")
  )

# kuvaaja: Histogrammi nearc4:n mukaan
ggplot(card_data, aes(x = educ, fill = factor(nearc4))) +
  geom_histogram(position = "dodge", bins = 20) +
  labs(x = "Koulutusvuodet", y = " Lukumr ", fill = " Lhell  korkeakoulua
    (nearc4)") +
  theme(
    text = element_text(size = 25),
    axis.title = element_text(size = 20),
    axis.text = element_text(size = 20),
    legend.text = element_text(size = 20),
    legend.title = element_text(size = 20),
    plot.title = element_text(size = 20, face = "bold"),
    legend.position = c(0, 1),
    legend.justification = c(0, 1)
  )

# Kausaalisuhteiden arviointi
# OLS regressio
ols_reg <- lm(Y1 ~ Y2 + X1)
ols_summary <- summary(ols_reg)

# 2SLS regression ensimminen vaihe manuaalisesti jotta saadaan laskettua
   f-testisuureen arvo.
# Ensimminen vaihe: Ennustetaan Y2 instrumenteilla
first_stage <- lm(Y2 ~ X1 + Z, data = card_data)
```

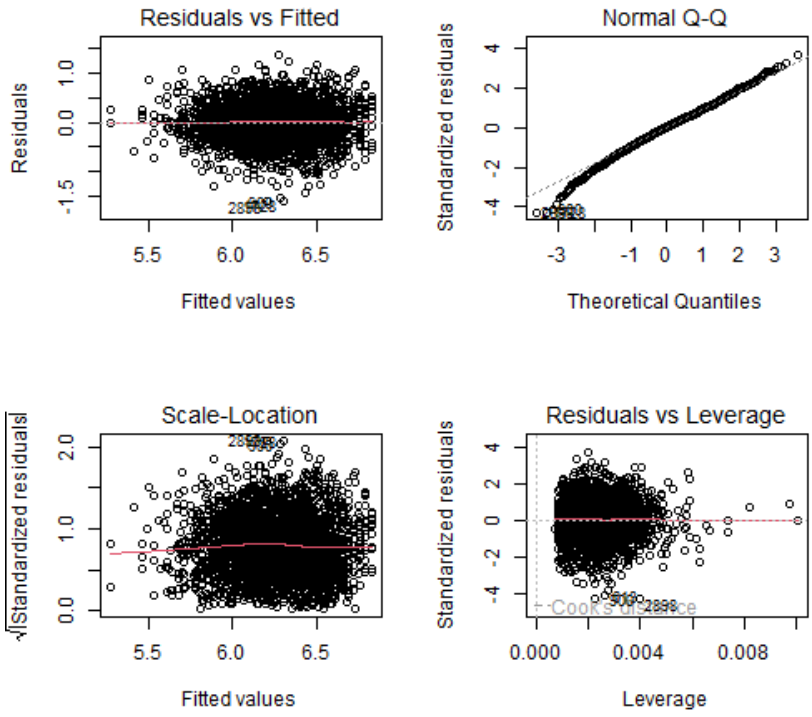
```
reduced_model <- lm(Y2 ~ X1, data = card_data)

# F-testisuure manuaalisesti
f_stat_manual <- ((sum(residuals(reduced_model)^2) -
    sum(residuals(first_stage)^2)) / 1) /
     (sum(residuals(first_stage)^2) / df.residual(first_stage))

# 2SLS regressio
iv_reg <- ivreg(Y1 ~ Y2 + X1 | X1 + Z)
iv_summary <- summary(iv_reg)


# Tulokset
stargazer(ols_reg, first_stage, iv_reg,
          type = "text",
          title = "OLS ja 2SLS tulokset",
          column.labels = c("OLS", "Ensimminen vaihe", "Toinen vaihe"),
          dep.var.labels = "Log Palkka",
          covariate.labels = c("Koulutus", "Kokemus", "Tummaihosuus",
                               "Etelisyys", "Naimisissa", "Urbaanisuus",
    "korkeakoulun lheisyys"),
          add.lines = list(c("Ensimmisen vaiheen F-testi", f_stat_manual)),
          out = "regression_results.txt")


#lispohdintaa
# IV-analyysin (2SLS) kannalta mallin diagnostiikka ei ole yhta keskeista
    kuin tavanomaisessa OLS-regressiossa. Poikkeavat havainnot voivat
    kuitenkin vaikuttaa voimakkaasti 2SLS-tuloksiin, joten ne on
    tarkastettu. Normaalijakautuneisuus on tarkeaa lahinna pienten otosten
    yhteydessa, silla 2SLS-estimaatit ovat asymptoottisesti
    normaalijakautuneita suurissa otoksissa. Multikollineaarisuus ei ole
    2SLS-mallissa yhta kriittista kuin instrumenttien voimakkuus.
par(mfrow = c(2, 2))
plot(ols_reg)
```



Kuva 3: OLS-regression diagnostiikkaa