Hvordan identifisere årsakssammenhenger i ikke-eksperimentelle data?

En ikke-teknisk introduksjon

Henning Finseraas

Institutt for samfunnsforskning

henninfi@samfunnsforskning.no

&

Andreas Kotsadam

Økonomisk institutt, Universitet i Oslo

Andreas.kotsadam@econ.uio.no

Vi ønsker å takke Jon-Ivar Elstad, Niklas Jakobsson, Sofie Kjernli-Wijnen, Ola Listhaug, Frikk Nesje, Viggo Nordvik, Axel West Pedersen, Christiane Marie Ødegård og Politikk, demokrati og sivilsamfunnsgruppa ved ISF for nyttige kommentarer og innspill på et tidligere utkast av denne artikkelen.

Innledning

Reduseres sykefraværet dersom man får gratis trening i arbeidstida? Øker Arbeiderpartiet oppslutningen dersom det blir lettere å stemme ved valg? Økte drapet på den nederlandske filmregissøren Theo van Gogh motstanden mot en liberal innvandringspolitikk? Dette er spørsmål om kausalitet, altså hvorvidt det eksisterer en årsakssammenheng mellom et tiltak (eksempelvis gratis trening i arbeidstida) og en effekt (redusert sykefravær). Innenfor mange forskningsfelt og for mange ulike forskningsspørsmål er det viktig å identifisere årsakssammenhenger. Dette gjelder særlig innenfor evalueringsstudier, men ofte er man også interessert å avdekke kausale sammenhenger mellom sosioøkonomiske variabler (for eksempel utdanning og inntekt) og sosiale og politiske utfallsvariabler. Å identifisere en årsakssammenheng er noe helt annet enn å identifisere en sammenheng mellom to variabler, selv om det dessverre er ganske utbredt blant samfunnsforskere å omtale sine resultater i en kausal språkdrakt uten at man har tilstrekkelig grunnlag for dette.

I empirisk samfunnsvitenskap er det en voksende interesse for å identifisere årsakssammenhenger. Denne trenden har blitt særlig sterk innenfor empirisk mikroøkonomi (Angrist og Pischke 2010), så sterk at det har kommet en motreaksjon (Keane 2010). Etterhvert har interessen for å identifisere årsakssammenhenger fått et sterkt gjennomslag også i (amerikansk) statsvitenskap, hvor et godt eksempel er forskningen på tiltak for å heve valgdeltakelsen (Green og Gerber 2008). To nyere eksempler fra sosiologiske tidsskrifter er Kirks (2009) studie av lokalmiljø og tilbakefall til kriminalitet blant tidligere fanger, og Correll med fleres (2007) studie av hvorfor mødre har lavere lønninger enn kvinner uten barn.[1]

[1] Angrist og Pische (2009), Dunning (2012) og Morgan og Winship (2007) er gode, nylig publiserte introduksjonsbøker fra henholdsvis samfunnsøkonomi, statsvitenskap og sosiologi.

Samfunnsforskning dreier seg om mye mer enn å avdekke årsakssammenhenger. Vi er ikke talspersoner for at teoretisk og beskrivende forskning skal avvikles eller at samfunnsforskere til enhver tid må ha et forskningsdesign som gjør det mulig å si noe om kausalitet. Vårt ønske med artikkelen er at forskere som stiller eller ønsker å stille kausale spørsmål, samt de som evaluerer forskning, skal bli mer bevisst hvilke antagelser som ligger i bunn for at man kan gi resultatene en troverdig kausal tolkning.

Vi skal presentere ulike empiriske strategier for å identifisere årsakssammenhenger i observasjonsdata, altså i ikke-eksperimentelle data hvor enhetenes verdier på den sentrale uavhengige variabelen i utgangspunktet ikke er randomisert. Det er ulike måter å forstå kausalitet på, men etterhvert har det såkalte kontrafaktiske synet på kausalitet blitt dominerende og det er dette perspektivet vi baserer oss på i denne artikkelen. Vi starter neste avsnitt med kort å gjøre rede for det kontrafaktiske synet på kausalitet. Deretter går vi gjennom ulike empiriske strategier for å identifisere en kausal sammenheng mellom to variabler.

Vi vil først vise hvor vanskelig det er å si noe om kausalitet basert på en teknikk man ofte ser brukt i samfunnsvitenskapene, nemlig multivariat regresjon med tverrsnittsdata. Deretter går vi over til en situasjon hvor man har longitudinelle data og forklarer hvorfor man iblant identifiserer troverdige kausale estimat ved å kontrollere for såkalte enhetsfaste effekter (fixed-effects) eller ved å studere forskjeller i trender mellom ulike grupper. Deretter går vi over til det bruker mest plass på i artikkelen, Regression Discontinuity-design (RD) og Instrumentvariabel-metoden (IV), som er to kvasi-eksperimentelle metoder. Gjennom hele artikkelen legger vi vekt på en ikke-teknisk fremstilling med hovedfokus på hvilke antagelser de ulike strategiene bygger på for at et estimat kan gis en troverdig kausal tolkning. For hver teknikk vil vi gi mange eksempler fra publiserte samfunnsvitenskapelige studier. I tillegg har

vi på våre hjemmesider lagt ut forslag og illustrasjoner for hvordan å estimere og tolke modellene vi presenterer ved hjelp av programpakken Stata.

Den kontrafaktiske forståelsen av kausalitet

Det vanligste rammeverket for å forstå hvordan man kan identifisere en kausal effekt er det såkalte kontrafaktiske, også betegnet som potensielle utfallssammenligninger eller Neyman-Rubins forståelse av kausalitet. Dette rammeverket er intuitivt, samt nyttig også fordi det illustrerer begrensningene med observasjonsdata i situasjoner hvor formålet er å si noe om årsakssammenhenger.

Vi vil bruke et enkelt eksempel for å illustrere rammeverket. La oss si at du er interessert i hvorvidt et språkkurs for innvandrere øker sannsynligheten for å få betalt arbeid, og gjennomfører en spørreundersøkelse hvor du spør innvandrere om hvorvidt de har tatt språkkurset i løpet av det siste året og om de er i betalt arbeid. Det er ikke usannsynlig at du finner at færre av de som har tatt kurset er i arbeid enn de som ikke har tatt kurset. Det er i så fall rimelig å tro at dette i så fall skyldes seleksjon, altså at de som har tatt kurset hadde dårligere språkkunnskaper i utgangspunktet, og vi tror få empiriske forskere ville konkludert med at kurset reduserer sannsynligheten for å ta arbeid (selv om dette på kort sikt ikke er helt umulig siden det tar tid å ta kurset, og dette er tid som kunne vært brukt på å finne en jobb). Det vi er interessert i å vite er hva som hadde skjedd hvis de som tok kurset ikke hadde tatt kurset. Studier av betydningen av for eksempel giftemål, utdanning, trening, å motta ulike trygder, å få barn, har ofte den samme problematikken. Vi påstår ikke at dette er ukjent problematikk for empiriske samfunnsforskere, men dessverre er det ikke helt uvanlig at forskere glir over i en kausalitetstolkning av resultater som ikke kan gis denne tolkningen.

La oss igjen ta utgangspunkt i hvorvidt et språkkurs øker sannsynligheten for at innvandrere havner i arbeid. Dette er et kausalt spørsmål og ikke et spørsmål om å beskrive

sammenhengen mellom å ta et språkkurs og å være i arbeid. I det kontrafaktiske rammeverket er man interessert i å sammenligne arbeidsmarkedssituasjon for den samme personen i to ulike situasjoner; én situasjon hvor man tok kurset, og én situasjon hvor man ikke tok kurset. Den kausale effekten av språkkurset er forskjellen i arbeidsmarkedssituasjon i de to situasjonene. Dette er en enkel og intuitiv forståelse. Problemet er det som i kausalitetslitteraturen omtales som det *fundamentale inferensproblemet*: Vi kan bare observere ett av utfallene. Personen tok enten kurset (eller giftet seg, fikk trygd eller barn og så videre) eller ikke, og vi kan aldri vite sikkert hvilken arbeidsmarkedssituasjon personen ville rapportert dersom det motsatte faktisk skjedde. Siden man aldri kan observere begge utfall for den samme personen samtidig så ønsker man å identifisere en kontrollgruppe til de som tok kurset, en gruppe med personer som naturlig nok ikke tok kurset, og deretter sammenligne arbeidsmarkedssituasjon i de to gruppene.

Troverdigheten til en kausal tolkning av forskjellen mellom gruppenes gjennomsnitt er avhengig av hvor like de to gruppene er på alle observerte og ikke-observerte karakteristika. Da er vi tilbake til problemet med seleksjon, som Angrist og Pischke (2009) omtaler som det mest alvorlige problemet for empirisk samfunnsforskning. Med dette i bakhodet er det ikke overraskende at eksperimenter hvor deltakere blir tilfeldig plassert i en av de to gruppene (kontrollgruppa og eksponeringsgruppa) fremstår som gullstandarden for empirisk forskning, da den tilfeldige plasseringen gjør at de to gruppene i gjennomsnitt er identiske på alle observerte *og* ikke-observerte karakteristika.

Men hva hvis man ikke har eksperimentelle data? Hvilke muligheter har man til å si noe om kausalitet da?

Tradisjonelle empiriske strategier

Multivariat regresjon med tverrsnittsdata

En veldig vanlig situasjon for en samfunnsforsker er at hun skal undersøke en sammenheng mellom en uavhengig variabel X og en avhengig variabel Y med data fra ett tidspunkt, hvor enhetenes verdier på X ikke er randomisert. Samtidig kjenner hun til en rekke andre variabler enn X som teori eller tidligere empirisk forskning har identifisert som viktig for Y, og hun estimerer derfor en multivariat regresjonsmodell. Koeffisienten for X angir da den betingede korrelasjonen mellom X og Y, altså den gjennomsnittlige endringen i Y når X endres med én enhet og vi samtidig holder alle de andre variablene *vi har med i regresjonsmodellen* uendret. I de aller fleste tilfeller så er den betingede korrelasjonen et lite troverdig estimat på den kausale sammenhengen mellom X og Y. Det er i hovedsak tre grunner til dette, hvorav to er velkjente, mens ikke alle virker å være like bevisst den tredje hovedgrunnen.

For det første kan det være at Y påvirker X, for eksempel dersom bedrifter med høyt sykefravær i utgangspunktet er mer tilbøyelig til å implementere gratis treningstid, noe som i så fall vil skjule en eventuell kausal effekt av gratis treningstid. Dette er et det samme tilfellet av seleksjon som i eksemplet over. Sykefraværet i gruppen som ikke har innført gratis treningstid er ikke et troverdig kontrafaktisk sykefravær for sykefraværet i gruppen som har innført treningstid.

For det andre kan det være en utelatt variabel som påvirker både Y og X, det vil si at vi har uobserverbar heterogenitet, også kjent som utelatt-variabel-problemet. Dersom vi ikke har observert og kontrollert for alle variabler hvor X-gruppene er ulike og som samtidig påvirker Y, det vil si alle seleksjonsmekanismer som fører til at man implementerer gratis treningstid, så vil ikke den betingede korrelasjonen representere den kausale effekten. Vi vil heller ikke vite retningen på skjevheten, altså om den estimerte effekten er for stor eller for liten sammenlignet med den virkelige kausale effekten.

For det tredje kan det være at kontrollvariabler er delvis påvirket av X slik at man ”kontrollerer bort” deler av den effekten av X på Y som man ønsker å avdekke. For eksempel

kunne man i studien av effekten av gratis treningstid kontrollere for antall personer som spiser lunsj i lunsjrommet, ut fra en tanke om at dette er et proxy på hvor sosialt arbeidsmiljøet er. Et problem vil være at mange med tilgang til gratis trening i arbeidstida kan tenkes å gjennomføre treninga rundt lunsjtider, eller de spiser lunsj sammen fordi de også trener sammen. Med andre ord er personer som spiser lunsj sammen delvis en funksjon av gratis treningstid. Angrist og Pischke (2009) omtaler denne typen kontrollvariabler som "bad controls" fordi dette er variabler som like gjerne kan være en avhengig variabel i studien. Når man setter opp en regresjonsmodell med ønske om å si noe om kausalitet må man være sin egen største kritiker og ikke inkludere en kontrollvariabel som åpner for at en kritisk leser kan hevde at denne kontrollvariabelen kan være en funksjon av X. Vårt inntrykk er at samfunnsforskere er for lite bevisst dette problemet og har en tendens til å inkludere enhver variabel som kan tenkes å være korrelert med Y, uten å tenke grundig nok gjennom forholdet mellom X og de ulike kontrollvariablene. En forklaring på dette fenomenet kan være den lange tradisjonen for å forklare variasjonen i Y i stedet for å identifisere kausale effekter.

Et alternativ til multippel regresjon som vi nedprioriterer i denne artikkelen er såkalte matching-teknikker. Logikken i denne typen teknikker er at man tar hensyn til seleksjonen inn til X gjennom å vekte observasjonene slik at kontroll og eksponeringsgruppa er i gjennomsnitt like på seleksjonsvariablene (se Winship og Morgan 2007 for en introduksjon). Matching har noen potensielle gevinster sammenlignet med multippel regresjon, blant annet kan det gjøre modellen mindre sårbar for antagelser om linearitet, men matching er ikke mindre sårbar enn standard regresjonsteknikker for seleksjon på uobserverte variabler. I og med at seleksjon på uobserverte variabler er hovedbekymringen når man ønsker å tolke koeffisienter kausalt mener vi at matching som regel er en utilfredsstillende teknikk i studier av kausalitet.

Paneldata: Trendanalyser og kontroll for enhets- og tidsfaste effekter[2]

Dersom man har data hvor man har fulgt enhetene over tid har man forbedret mulighetene til å estimere troverdige kausale effekter en god del, også i tilfeller hvor X ikke er randomisert. Hovedårsaken til at estimatene er noe mer troverdig skyldes at med data over tid kan man effektivt fjerne all variasjon (all uobserverbar heterogenitet) i den avhengige variabelen som skyldes faktorer som er konstante over tid. Vi vil først forklare denne strategien med utgangspunkt i en situasjon hvor man har data fra to tidspunkter, og hvor den uavhengige variabelen av interesse endrer verdi for én gruppe av enhetene, men ikke for den andre. Dette kalles å studere forskjeller-i-forskjeller (”Differences-in-differences”). Deretter forklarer vi det man kan kalle en generalisering av denne teknikken for tilfeller hvor man har data fra flere enn to tidspunkter, såkalte analyser med kontroll for enhetsfaste effekter (”unit fixed effects”).

Studier av forskjeller-i-forskjeller har ofte data fra minst to tidspunkt, hvor noen enheter er utsatt for en endring i X i løpet av perioden, mens andre ikke er dette. Dette er en veldig utbredt teknikk innenfor evalueringsstudier. Et klassisk eksempel fra samfunnsøkonomi er Card og Kruegers (1994) studie av hvordan en økning i lovbestemt minstelønn påvirker sysselsettingen. Utgangspunktet for studien er at New Jersey økte nivået på minstelønnen, mens nabostaten Pennsylvania ikke gjorde dette, og Card og Krueger (1994) sammenligner lønninger og sysselsetting før og etter reformen i fast food-restauranter i New Jersey og i østlige deler av Pennsylvania.

Intuitivt så kalkulerer man først forskjellen i Y (sysselsetting) mellom de to tidspunktene for begge gruppene (”within-difference”), og deretter sammenligner man denne forskjellen for de to gruppene (”between-difference”). Med andre ord, har den gruppa som opplevde en endring i X (minstelønnsnivå) en utvikling i Y (sysselsetting) som er

[2] Denne seksjonen er inspirert av Schlotter m. fl. (2009).

vesensforskjellig fra utviklinga i den gruppa som ikke opplevde endring? Under antagelsen at gruppa som opplevde endring i X *ville utviklet seg identisk som gruppa uten endring i X dersom X ikke hadde blitt endret,* så har man identifisert en kausal effekt av X. I Card og Kruegers tilfelle er altså antagelsen at lønninger og sysselsetting i fast food-restauranter ville hatt identiske vekstrater i New Jersey og Pennsylvania uten økningen i minstelønn i New Jersey. Rimeligheten i antagelsen om at trenden i de to gruppene ville vært identisk uten endringen i X er essensiell å diskutere, og vil variere fra studie til studie. Det er likevel åpenbart at man har en mindre restriktiv antagelse enn den antakelsen man må gjøre for å gi en kausal tolkning av estimater fra tverrsnittsdata, som er at *gruppene* (og ikke *trenden*) er identisk på uobserverbare variabler.

Dersom man har data fra flere tidspunkter er studier med såkalt kontroll for enhetsfaste effekter ("unit-specific fixed effects") en generalisering av forskjeller-i-forskjeller. I praksis er det to måter å kontrollere for enhetsfaste effekter. Man kan estimere egne konstanter for hver av enhetene i datasettet,[3] eller man kan sentrere alle variablene i modellen rundt variabelens enhetsspesifikke gjennomsnitt.[4] De to fremgangsmåtene gir identiske koeffisientestimater for X. Med flere enn to perioder og kontroll for enhetsfaste effekter har man en modell som studerer hvordan avvik fra enhetens gjennomsnitt på X er relatert til avvik fra enhetens gjennomsnitt på Y. Variabler som er konstante over tid – for eksempel kjønn i en studie hvor enhetene er individer, sektor i en studie av bedrifter, eller avstand til Oslo i en studie av

---

[3] I praksis gjør man dette ved å inkludere indikatorvariabler for alle minus 1 av enhetene i datasettet, på samme måte som man kontrollerer for andre kategoriske variabler i en regresjonsanalyse.

[4] I statistikkpakken Stata trenger man ikke å gjøre disse omkodingene manuelt, man trenger kun å spesifisere at man ønsker at Stata skal behandle variablene på denne måten.

kommuner – har naturlig nok ingen avvik fra gjennomsnittet og kan heller ikke forklare endring rundt Ys gjennomsnitt.[5]

Det er lett å se at denne typen analyse betraktelig reduserer det som er problem to i forrige seksjon, i og med at vi nå har kontrollert for all uobserverbar heterogenitet som er konstant over tid, og "bare" trenger å bekymre oss for uobserverbar heterogenitet som skyldes utelatte variabler som varierer over tid.[6]

Merk at med tidsseriedata kan man i tillegg til å fjerne uobserverbar heterogenitet som skyldes enhetsfaste effekter, også fjerne uobserverbar heterogenitet som skyldes tidsfaste effekter ved å gi hvert tidspunkt i analysen en egen konstant ("time fixed effects"). Dette fjerner all variasjon i Y som skyldes tidseffekter som er identiske for alle enhetene i analysen, altså fjerner man alle tidsspesifikke sjokk i data som kan gi opphav til spuriøse sammenhenger. For eksempel kan man tenke seg at observasjoner under en global finanskrise er systematisk forskjellige fra observasjoner under mer normale år, og man ønsker å forsikre seg om at resultatene ikke er fullstendig drevet av noen spesielle år i dataene. Dette reduserer åpenbart skjevheter som skyldes uobserverbar heterogenitet ytterligere, siden man nå kun trenger å bekymre seg for utelatte variabler hvor variasjonen er ulik mellom enhetene over tid.

Dessverre er det som regel grunn til å bekymre seg for at systematisk variasjon eksisterer selv etter å ha justert for enhets- og tidsspesifikke effekter, slik at det ikke er rimelig

---

[5] Merk at variablene kun trenger å være konstante over den perioden som studien dekker, det er ikke noe krav at for eksempel individvariabler skal være konstante over hele livsløpet for at man kan hevde at effektene av disse variablene er luket bort ved hjelp av de enhetsfaste konstantene.

[6] Man kan også argumentere for at det første problemet er noe redusert siden man nå ikke fokuserer på nivåforskjeller.

å tro at man estimerer en troverdig kausaleffekt. I noen tilfeller er man så heldig at man kan gi mer troverdige kausale estimat gjennom å benytte forskningsdesign som identifiserer uavhengig variasjon i den uavhengige variabelen man er interessert i. Vi skal nå diskutere to slike forskningsdesign.

Kvasi-eksperimentelle strategier

Regression Discontinuity-design

I løpet av det siste tiåret har det vært en eksplosjon i antallet artikler som benytter seg av såkalte Regression Discontinuity-design (RD). Selve designet går tilbake til Thistlethwaite og Campbell (1960) som var interessert i kausaleffekten av å motta et stipend på senere prestasjoner. Stipendet ble gitt på bakgrunn av en prøve, hvor de studentene som fikk en score over et visst nivå på prøven fikk stipendet. Problemet med å estimere effekten av stipendet ligger i å skille effekten av å motta selve stipendet fra blant annet effekten av kognitive ferdigheter, som er korrelert med både testscore og senere prestasjoner. De som mottok stipendet er sannsynligvis smartere enn de som ikke fikk stipendet, men Thistlethwaite og Campbell (1960) er interessert i motivasjonseffekten av å motta stipendet. En effektiv kontroll for kognitive ferdigheter er imidlertid som regel umulig.

For å estimere effekten av å motta selve stipendet utnytter Thistlethwaite og Campbell (1960) at tildelingen av stipendet strengt følger poengscoren på prøven, slik at en student som er ett poeng under grensa for å få stipendet ikke mottar stipendet, mens en student med kun ett poeng mer på prøven får stipendet. En så liten forskjell mellom to studenter på én enkelt prøve kan med en viss rimelighet antas å være tilfeldig, og man har i så fall identifisert en uavhengig variasjon i tildeling av stipendet. Dette betyr at studenter som er rett under grensa for å få stipendet vil være en troverdig kontrafaktisk sammenligningsgruppe for studenter som er rett over grensa for å få stipendet. Forskjellen i senere prestasjoner mellom disse gruppene

kan da anses å være et troverdig kausalt estimat på det å motta selve stipendet, særlig da man kontrollerer for den kontinuerlige effekten av å oppnå ett poeng mer på prøven.

Muligens er det første man tenker at dette fremstår som en teknikk med begrenset anvendelse, men logikken er gyldig i alle situasjoner hvor enheter allokeres til en status basert på verdien på en kontinuerlig variabel. Som nevnt er det mange eksempler fra de senere årene og vi skal trekke frem tre eksempler som bruker denne logikken på ulike typer data.

Lalive (2007) studerer hvorvidt forlenget tilgang på arbeidsledighetstrygd øker lengden på arbeidsledighetsforløpet. Utgangspunktet er en diskontinuitet i lengden på arbeidsledighetstrygd i Østerrike, hvor de som er over 50 år mottar arbeidsledighetstrygd lengre enn de under 50 år, slik at man kan studere om det er et brudd i lengden på arbeidsledighetsforløpet rundt 50 år. Den nødvendige antagelsen for at et slikt brudd kan gis en kausal tolkning er, analogt til Thistlethwaite og Campbell (1960), at arbeidsledige på 49 år er så å si identiske med arbeidsledige på 50 år på alle andre faktorer enn hvor lang arbeidsledighetstrygd de har krav på.

Eggers og Hainmueller (2009) er interessert i hvilke økonomiske gevinster politikere oppnår ved å bli valgt til nasjonalforsamlingen. De studerer dette ved å samle data på formuene til nylig avdøde tidligere britiske parlamentsmedlemmer og de politikerne som tapte i de respektive valgkretsene. Her er man interessert i effekten å vinne et valg og deretter bli parlamentsmedlem, mens allokeringsvariabelen til statusen som parlamentsmedlem er andel av stemmene i valget. Den underliggende antagelsen for at RD gir kausaleffekten av å være parlamentsmedlem på livsformue er dermed at de som vant med et knapt flertall ikke har vesentlig andre ferdigheter som påvirker livsformuen enn de som så vidt tapte valget.

Tilslutt er vi ubeskjedne nok til å trekke frem et eksempel fra egen forskning. I Finseraas m. fl. (2011) er vi interessert i hvorvidt drapet på den kontroversielle nederlandske filmregissøren Theo van Gogh, som ble drept av en muslimsk ekstremist, medførte en endring

i europeeres holdninger til innvandring. Vi utnytter at European Social Survey (ESS) ble gjennomført i tidsrommet drapet skjedde og at ESS inneholder informasjon om dato for når respondenten ble intervjuet. Dette gjør at vi kan allokere respondentene i to grupper – de som ble intervjuet før drapet skjedde og de som ble intervjuet etter at drapet skjedde. I og med at intervjudatoen er tilfeldig er respondentene i randomisert til en av de to gruppene og gruppene er dermed i gjennomsnitt like på andre relevante variabler. Ved hjelp av RD kan vi estimere den umiddelbare effekten av drapet på holdninger.

Rent praktisk er det enkelt å estimere RD-modeller dersom man har data som gjør at man strengt kan allokere observasjoner til en eksponeringsstatus og en kontrollstatus. Først konstruerer man en indikator for hvilken gruppe observasjonen tilhører, det vil si om observasjonen er utsatt for eksponering eller tilhører kontrollgruppa. Deretter omkoder man allokeringsvariabelen (henholdsvis testscoren, fødselsdatoen, andelen stemmer eller intervjudatoen i de eksemplene vi har diskutert her) slik at verdien 0 angir kuttpunktet for overgangen mellom de to gruppene (i våre eksempler er dette poengscoren som gir rett til stipend, datoen som gir forlenget arbeidsledighetstrygd, 50 prosent av stemmene, og datoen for drapet på van Gogh).

Gitt at utfallsvariabelen er en kontinuerlig variabel estimerer man deretter en OLS regresjon med gruppeindikatoren som den sentrale uavhengige variabelen, samtidig som man kontrollerer for allokeringsvariabelen. I tillegg er det anbefalt å inkludere et samspillsledd mellom gruppeindikatoren og allokeringsvariabelen. Dette er anbefalt fordi man da tillater koeffisienten for allokeringsvariabelen, det vil si trenden i Y, å være forskjellig i de to gruppene. Dette kan ofte gi viktig informasjon, da det i mange tilfeller er interessant hvorvidt effekten av eksponering er stabil i eksponeringsgruppa. Eksempelvis, i studien av effekten av drapet på van Gogh er det interessant om effekten på holdninger holder seg over tid, eller om effekten raskt forsvinner og at opinionen returnerer til holdningene før drapet.

Et viktig valg man må ta når man gjennomfører en RD-analyse er hvor stort det såkalte vinduet («bandwidth») man analyserer skal være, det vil si hvor stort område rundt kuttpunktet man skal inkludere i analysen. I stipendeksempelet, hvor mange testpoeng unna kuttpunktet skal inkluderes i analysen? I arbeidsledighetstrygdeksempelet, hvor mange årskull rundt 50 år skal inkluderes? Avveiningen står mellom å ha tilstrekkelig med observasjoner til å få presise koeffisientestimat og samtidig ha to relativt like grupper. Jo større vindu, jo flere observasjoner, men samtidig blir de to gruppene sannsynligvis mer ulik på andre relevante variabler – og da forsvinner også eksperimentanalogien. I tillegg må man tenke gjennom – og man skal analysere empirisk – hvorvidt trendene i Y før og etter kuttpunktet på X virkelig er lineære, en antagelse som krever et større vindu for å kunne analyseres troverdig. Hvis man antar lineære trender, men disse i virkeligheten er ikke-lineære, står man i fare for å identifisere en eksponeringseffekt som ikke er reell. I utgangspunktet er det lett å undersøke denne antagelsen gjennom å se hva som skjer med koeffisienten når man tillater en mer fleksibel funksjonsform for allokeringsvariabelen ved å inkludere ulike polynomer (et kvadratisk ledd, et kubisk ledd og så videre).

Det finnes tekniske løsninger for å identifisere en optimal størrelse på vinduet, men vi anbefaler en enklere og mer pragmatisk tilnærming (Lee og Lemieux 2009): Gjennomfør analysen med en rekke ulike vinduer og rapporter resultatene fra de ulike vinduene. Man bør forvente at behandlingseffekten er stabil på tvers av de ulike vinduene. Hvis ikke er det grunn til å tro at man ikke har lykkes å modellere trendene riktig.

Som vi har gjentatt til det kjedsommelige er den sentrale antagelsen i RD-analysen at allokering til behandling og kontrollgruppe er tilfeldig innenfor det valgte vinduet, noe som tilsier at de to gruppene skal være identiske på alle observerte og uobserverte variabler. Det er veldig viktig, og et absolutt krav for kredibiliteten til analysen, at man kan vise at gruppene er relativt like på observerte variabler. Det er ulike måter å gjøre dette på, men den vanligste er å

vise at gruppene har like gjennomsnitt og standardavvik på variabler som man vet (fra teori eller tidligere empirisk forskning) er relevante for utfallsvariabelen. Man kan selvfølgelig ikke vite om gruppene er identiske på uobserverte variabler, men dersom gruppene er like på en rekke relevante observerte variabler er dette mer sannsynlig. Det er vanlig å utføre RD også på disse kontrollvariablene, med en forventning om at det ikke skal være et hopp for disse variablene ved kuttpunktet.

Dersom gruppene er helt identiske på et sett med variabler så er det naturlig nok ikke nødvendig å kontrollere for disse variablene, men dersom gruppene ikke er helt identiske så bør man vurdere å inkludere disse variablene som kontrollvariabler for å se om effektkoeffisienten er sensitiv for inkludering av kontrollvariabler. Hvis effektkoeffisienten endrer seg merkbart så indikerer det at RD-analysen ikke fanger den kausale effekten (det kan være at effekten varierer, for eksempel mellom ulike utdanningsgrupper). Det kan også være effektivt rent statistisk å inkludere kontrollvariabler dersom disse er åpenbart eksogene, da inkludering av strengt eksogene variabler kan redusere analysens standardfeil og dermed gi mer presise estimater.

Et potensielt ødeleggende problem i RD-analyser er muligheten for at enhetene har manipulert sin egen verdi på allokeringsvariabelen og blitt selektert inn i eksponerings- eller kontrollgruppa. I eksemplene over er ikke dette så sannsynlig, men man kan se for seg potensiell valgfiksing i Eggers og Hainmueller (2009) og manipulering av egen fødselsdato i Lalive (2007). Et interessant eksempel på mulig seleksjon fra norsk forskning finner vi Cools m. fl. (2011) som studerer effekten av pappapermisjon på blant annet kvinners lønnsutvikling. De utnytter at alle foreldre til barn født etter 1.april 1993 har rett på pappapermisjon, mens ingen født før 1.april har denne rettigheten. Ved hjelp av registerdata studerer de mødre som fikk barn i ukene rundt 1.april 1993. På tross av at denne datoen ble satt så sent at mødrene allerede var gravide da datoen ble satt, viser deres studie at fødselsmønsteret rundt 1.april

1993 er vesensforskjellig fra fordelingen rundt 1. april foregående år. Dette indikerer at enkelte mødre på et eller annet vis, sannsynligvis via utsatt keisersnitt, klarte å selektere seg inn i eksponeringsgruppa. Det er ingen grunn til å tro at de mødrene som klarte dette er identiske med de andre mødrene i utvalget, og dermed har man et seleksjonsproblem.[7] Dette illustrerer at seleksjon kan forekomme selv i situasjoner hvor det virker ekstremt usannsynlig. En måte å studere om resultatene er sensitive for en slik seleksjon er å gjennomføre en sensitivitetstest hvor man ekskluderer observasjoner tett rundt kuttpunktet og så studere om koeffisienten for eksponeringsgruppa forandrer seg mye. Hvis den gjør det, kan det indikere seleksjonsproblemer. I tillegg er det utviklet en test for hvorvidt fordelingens tetthet endres når man krysser kuttpunktet (se McCrary 2008).

I tillegg til å variere størrelsen på vinduet, vurdere ulike funksjonelle former for trendene, studere fordelinger på potensielle kontrollvariabler, og vurdere seleksjonsproblemer, har det blitt standard prosedyre å presentere resultater fra såkalte placebo-analyser. I denne typen analyser manipulerer man allokeringsmekanismen slik at det i virkeligheten ikke er en forskjell mellom eksponerings- og kontrollgruppa. I våre eksempler kan dette være å «flytte» drapet på van Gogh én måned eller ett år, eller å flytte aldersgrensa for utvidet arbeidsledighetstrygd tre år. I disse analysene skal man da naturlig nok ikke finne forskjeller i utfall mellom gruppene fordi eksponerings- og kontrollgruppa er utsatt for det samme miljøet. Finner man forskjeller må man tenke nøye gjennom troverdigheten i forskningsdesignet, da det kan være at man simpelthen fanger opp underliggende, uavhengige trender i dataene, eller andre hendelser eller reformer, og ikke effekten av det man egentlig er ute etter. En annen

[7] Det er mulig at det ble født flere barn rundt 1. april i 1993 av helt tilfeldige grunner. Med et signifikansnivå på 5 prosent kan vi forvente oss at vi gjør en type 1—mistolker et tilfeldig avvik som systematisk—i 5 av 100 tilfeller

vanlig strategi for å øke tilliten til resultatene er å gjøre placebo-analyser ved samme kuttpunkt, men for variabler som man ikke tror skal påvirkes. Finseraas m. fl. (2011) bruker denne strategien da de viser at drapet på van Gogh ikke påvirket holdninger til miljøvern, noe som styrker troverdigheten til studien fordi det indikerer at det ikke skjedde et generelt holdningsskifte, eller noe spesielt med utvalget, rundt drapstidspunktet.

I denne artikkelen har vi konsentrert oss om situasjoner hvor det er et strengt kuttpunkt som definerer observasjonens gruppestatus. RD kan imidlertid brukes også i situasjoner hvor denne allokeringen ikke er deterministisk, men hvor sannsynligheten for å endre status gjør et stort hopp ved et kuttpunkt. Denne typen analyser kalles «fuzzy RD». I eksempelet med utvidet arbeidsledighetstrygd kunne man tenkt seg at noen grupper under 50 år likevel er berettiget utvidet trygd, for eksempel dersom de har barn under en viss alder. Vi diskuterer ikke denne typen analyse her, se for eksempel Lee og Lemieux (2009).

Instrumentvariabel-regresjon

Instrumentvariabel-metoden (IV) er i utgangspunktet en veldig generell metode for å studere årsakssammenhenger som kan brukes både dersom man har kontrollerte eksperimenter og hvis man har ulike typer naturlige eksperimenter. RD-metoden kan forstås som spesiell variant av logikken som ligger bak IV-regresjon. Gevinsten ved å kunne gjennomføre denne typen analyser er stor, men teknikken er teknisk krevende å lære seg. Vi skal konsentrere oss om logikken og intuisjonen bak de ulike testene som metoden krever, men anbefaler alle som skal gjennomføre IV-regresjoner å sette seg inn i de mer tekniske aspektene (se for eksempel Angrist og Pischke 2009). Vi anbefaler også Murray (2006) som en ikke-teknisk introduksjon som går mer i dybden enn det vi gjør her (se også Murray 2010).

Det som gjør IV-metoden veldig nyttig er at den kan fjerne korrelasjon mellom feilleddet i regresjonen og den sentrale uavhengige variabelen. Som kjent gir en slik

korrelasjon skjeve koeffisientestimat. Det er en hel rekke situasjoner som kan skape en korrelasjon mellom feilleddet og en uavhengig variabel, flere av disse diskuterte vi tidlig i artikkelen. En utelatt variabel («omitted variable bias»), simultanitet eller omvendt kausalitet (Y påvirker X) og målefeil i den uavhengige variabelen, er alle situasjoner som medfører en korrelasjon mellom feilleddet og en uavhengig variabel. Konsekvensen av en slik korrelasjon er at koeffisienten ikke representerer en kausaleffekt. Formålet med IV-regresjonen er å fjerne denne korrelasjonen.

La oss anta at du er interessert i hvorvidt X påvirker Y, men at det er grunn til å tro at Y også påvirker X (omvendt kausalitet). Dette er en veldig vanlig situasjon, for eksempel kan inntektsfattigdom påvirke utdanningsvalg, men utdanningsvalg påvirker også fattigdom. IV-metoden kan løse dette problemet dersom du har en variabel Z, den såkalte instrumentvariabelen, som a) er sterkt korrelert med X, men b) samtidig ikke påvirker Y via andre kanaler enn X. Som man kan ane krever det ofte kreativitet, samt teoretisk, historisk eller institusjonell kunnskap for å identifisere gode instrumentvariabler.

Dersom man er så dyktig eller heldig at man har funnet en god instrumentvariabel kan man estimere en kausal effekt av X på Y – som vi skal diskutere nedenfor er det ikke tilfeldig at vi skriver *en* kausal effekt og ikke *den* kausale effekten – ved hjelp av en to-stegs-regresjonsmodell. I det første steget er X en avhengig variabel som forklares av Z og de andre forklaringsvariablene i modellen. I det andre steget forklarer man Y med de samme variablene som i det første steget, bortsett fra Z, men man bruker de *predikerte* X-verdiene fra det første steget som X-verdier istedenfor de faktiske X-verdiene. Intuisjonen er at de predikerte X-verdiene representerer en uavhengig variasjon i X *som kun skyldes Z*. I prinsippet kan man estimere disse to regresjonsmodellene separat da dette gir de riktige koeffisientestimatene, men problemet er at standardfeilene i det andre steget ikke blir korrekte (Angrist og Pischke

2009: 122). De fleste statistikkpakker har imidlertid muligheter for å estimere de to-stegene slik at man får riktige standardfeil.

La oss illustrere logikken ved hjelp av to studier fra samfunnsøkonomi og én studie fra statsvitenskap (se Murray 2006 og 2010 for flere eksempler). Angrist og Krueger (1991) er interessert i sammenhengen mellom utdanning og lønn. De er bekymret for at det vil være en korrelasjon mellom utdanning og feilleddet på grunn av utelatte variabler, siden det er lite troverdig at man kan kontrollere for alle faktorer som både påvirker utdanningsnivå og lønnsnivå. De utnytter imidlertid at noen amerikanske stater har utdanningslover som sier at man må starte på skolen *det året* man fyller 6 år, mens man kan slutte på skolen *den dagen* man fyller 16 år. Da de elever som begynner skolen samme år fyller 16 år ved ulike tidspunkter gir dette en variasjon i utdanningslengde som kun skyldes fødselsdato, og de bruker den variasjonen i utdanningslengde som skyldes fødselsdatoen til å estimere effekten av utdanningslengde på lønn. Instrumentet (Z) i denne studien er i hvilket kvartal i året man er født. Antagelsen er at hvilket kvartal i året man er født ikke påvirker lønn via andre mekanismer enn utdanningslengde (og de variablene de kontrollerer for i regresjonsmodellen). Dette er et eksempel på en IV-regresjon hvor instrumentet er basert på et naturlig eksperiment (utdanningslover) hvor forskerne ikke kontrollerer selve eksperimentet.

Green og Gerber (2000) er et eksempel på en IV-regresjon basert på et felteksperiment hvor forskerne kontrollerer eksperimentet. De er interessert i hvorvidt ulike mobiliseringskampanjer påvirker valgdeltakelsen, særlig hvorvidt personlige besøk hos velgere (X) påvirker tilbøyeligheten til å stemme (Y). De gjennomfører derfor et felteksperiment hvor innbyggere i New Haven, Connecticut, blir delt inn i tilfeldige grupper som fikk oppfordringer om å stemme enten i posten, gjennom oppringning eller ved personlig besøk. Enkelte ble utsatt for flere av disse oppfordringene, mens én gruppe ikke fikk noen form for stimuli. I utgangspunktet kan man tenke at randomiseringen til å motta personlige

besøk gjør at man har et kontrollert eksperiment som gjør at man kun trenger å sammenligne valgdeltakelsen blant de som fikk besøk og kontrollgruppa. Men i og med at ikke alle som havner i gruppen som skal få personlig besøk er hjemme da forskerne kommer på døra, og det er ingen grunn til å tro at det er tilfeldig hvem som er hjemme eller ikke, har man et seleksjonsproblem. Dette løser Green og Gerber (2000) ved å bruke den tilfeldige gruppeinndelingen som instrument (Z) for effekten av å få personlig besøk (X).

Dette eksemplet illustrerer også forskjellen mellom det vi på norsk kan kalle «intensjonseffekten» (intention-to-treat, ITT) og «eksponeringseffekten» (treatment-effect on the treated», TOT). Intensjonseffekten er forskjellen mellom de som allokeres til eksponeringsgruppa og kontrollgruppa, altså forskjellen mellom de man har en intensjon om å eksponere og kontrollgruppa, mens eksponeringseffekten er forskjellen mellom de som allokeres til eksponeringsgruppa og faktisk blir eksponert (dette er de såkalt adlydende enhetene, «compliers» på engelsk) og kontrollgruppa. Avhengig av forskningsspørsmålet kan begge være av interesse, og de kan være veldig forskjellig dersom mange i eksponeringsgruppa unngår eksponering.

I det siste eksemplet har forskerne en situasjon hvor de ikke har kontroll på eksperimentet og hvor det også er mer usikkerhet knyttet til hvorvidt instrumentet skaper eksogen variasjon i den uavhengige variabelen. Acemoglu m.fl. (2001) er interessert i sammenhengen mellom institusjoner som beskytter eiendomsretten og økonomisk utvikling, og er bekymret for at en enkel korrelasjon kan representere omvendt kausalitet – at rike land har «råd» til gode institusjoner. Basert på historisk kunnskap om europeernes koloniseringer argumenterer de for at potensiell dødelighet blant nybyggere på kolonitiden er en variabel som kan identifisere effekten av institusjoner: I områder hvor dødeligheten var høy var europeerne kun interessert i å utnytte områdets ressurser i en kortsiktig jakt på gevinst, mens i områder hvor dødeligheten var lav var de interessert i å bygge gode institusjoner i en mer

langsiktig utviklingsstrategi. Acemoglu m. fl. (2001) samler derfor data for dødeligheten blant nybyggere for mer enn 100 år siden, og bruker denne variabelen som instrumentet som har påvirket hvilke institusjoner disse landene har i dag for så å estimere betydningen av institusjoner for økonomisk utvikling.

Alle gode artikler som benytter IV-teknikker bruker mye tid på å diskutere oppfyllelsen av de nevnte to viktigste kravene til instrumentvariabelen, henholdsvis relevanskriteriet (a) og ekskluderingskriteriet (b). Dersom disse kravene ikke er oppfylt er det ingen grunn til å tro at IV vil gi bedre estimat på kausale effekter enn en standard regresjonsanalyse. Det er relativt uproblematisk å teste om relevanskriteriet er oppfylt. Relevanskriteriet sier at instrumentet Z skal være korrelert med X, noe man kan teste ved å se om koeffisienten for Z er statistisk signifikant i det første steget i to-stegs-modellen. Hvis instrumentet ikke er relevant kan det ikke brukes til å identifisere en årsakssammenheng.

Selv om instrumentet er relevant (signifikant i første seg), kan instrumentet være for svakt korrelert med den variabel man vil instrumentere til at det er et godt instrument. Hvor sterk korrelasjonen er, kan man sjekke ved hjelp av en F-test basert på regresjonsresultatene fra det første steget i to-stegs-modellen. Dersom denne korrelasjonen er for svak har man svake instrumenter («weak instruments») og estimatene fra IV-regresjonen kan være skjeve. En mye sitert tommelfingerregel sier at F-verdien skal være 10 eller større for at man kan ha tiltro til at instrumentene ikke er svake (Staiger and Stock 1997). Hvis man har svake instrumenter finnes det en del andre estimatorer man kan bruke enn to-stegs-modellen (se Murray 2006).

Det er som regel vesentlig mer problematisk å overbevise en kritisk leser om at ekskluderingskriteriet – at instrumentet Z ikke påvirker Y via andre kanaler enn X – er oppfylt. Én grunn til dette er at oppfyllelse av ekskluderingskriteriet ikke utelukkende er et teknisk, statistisk spørsmål, og ekskluderingskriteriet er ikke nødvendigvis oppfylt selv om

variabelverdiene på instrumentvariabelen er randomisert, fordi instrumentet kan påvirke gjennom flere kanaler. Det finnes heller ingen statistisk test som avgjør hvorvidt ekskluderingskriteriet er troverdig. Riktignok finnes det en statistisk test, en såkalt overidentifiseringstest, som er nyttig dersom man har flere instrumentvariabler enn problematiske X-variabler. Denne testen er imidlertid lite troverdig dersom instrumentvariablene er teoretisk beslektet, noe som ofte er tilfellet, særlig fordi det i utgangspunktet er vanskelig å finne gode instrumentvariabler. [8]

Usikkerheten knyttet til overidentifiseringstester, som dessuten bare kan gjennomføres dersom man har flere instrument, gjør at ekskluderingskriteriet må forsvares på andre vis. Murray (2006) diskuterer flere nyttige strategier for å studere om ekskluderingskriteriet er troverdig. På samme måte som man tenker gjennom potensielle utelatte variabler i standard regresjonsanalyser, er det nødvendig å tenke gjennom alternative mekanismer for hvordan instrumentet kan påvirke den avhengige variabelen, og deretter kontrollere for disse mekanismene. Hvis konklusjonene ikke endres av å kontrollere for alternative mekanismer styrkes troverdigheten til ekskluderingskriteriet. Det er også nødvendig å studere resultatene fra det første steget i modellen og forsikre seg om at disse resultatene er plausible og i tråd med de teoretiske forventingene for hvordan Z skal påvirke X. Murray (2006) anbefaler også

[8] La oss si at vi kun har én problematisk X-variabel som vi vil instrumentere med to instrumentvariabler, Z1 og Z2. En overidentifiseringstest tester hvorvidt koeffisienten til for eksempel Z2 er signifikant dersom den inkluderes som en prediktor for Y i en to-stegs-modell med Z1 som instrument. Dersom koeffisienten for Z2 er signifikant svekkes troverdigheten til ekskluderingskriteriet. Svakheten ved testen er når man tester hvorvidt den ene variabelen er et valid instrument, i vårt eksempel Z2, antas det at Z1 er et valid instrument (og motsatt). Hvis begge er invalide så fungerer ikke testen.

å studere den direkte sammenhengen mellom Z og Y i en standard regresjonsmodell, for å forsikre seg om at instrumentet er korrelert med Y på en forventet måte (Husk: ekskluderingskriteriet sier ikke at Y og Z ikke er korrelert, det sier at X er den eneste mekanismen for hvorfor Y og Z er korrelert).

Ekskluderingskriteriet er ofte det svakeste leddet i analyser som bruker instrumentvariabler. Da man aldri kan statistisk vise at kriteriet er oppfylt så brukes ofte mange sider i akademiske artikler på å overbevise leseren om at så er tilfellet. For eksempel har det instrument som brukes av Acemoglu m.fl. (2001), dødelighet hos nybyggere, blitt sterkt kritisert med hensyn på hvorvidt ekskluderingskriteriet er oppfylt. Blant annet argumenterer Glaeser m. fl. (2004) overbevisende for at selv om vi kan vise at dødelighet hos nybyggere er korrelert med rikdom i dag, finnes det ingen måte å være sikker på at det viktigste nybyggerne tok med seg til koloniene var institusjoner. Hvis de tok med seg noe annet som også påvirker rikdom i dag, som for eksempel humankapital, så er ikke ekskluderingskriteriet oppfylt ettersom instrumentet ikke bare påvirker Y (velstand) via X (institusjoner), men også via humankapital.

I tillegg til relevans- og ekskluderingskriteriet, baserer IV-regresjon seg på ytterligere to ikke-testbare antagelser som kan påvirke den kausale tolkningen av resultatene. Den ene antagelsen er at effekten av instrumentet på X er monoton, det vil si at effekten av instrumentet går i samme retning for alle observasjoner. Denne antagelsen sier at de som ikke eksponeres ikke gjør dette *fordi* de er i eksponeringsgruppa. I Angrist og Krugers (1991) studie av drop-out fra videregående hvor fødselskvartal er instrumentet for å droppe ut, sier denne antagelsen at alle som har en fødselsdato som gir de rett til å droppe ut blir enten mer tilbøyelig til å droppe ut eller er upåvirket, ingen blir *mindre* tilbøyelig til å droppe ut fordi de har denne rettigheten. Antagelsen er at man ikke har slike trossere («defiers»).

Den siste antagelsen går vanligvis under navnet SUTVA («stable unit-treatment value assumption»). SUTVA sier at individene i eksponerings- og kontrollgruppa ikke skal påvirke hverandre, eksempelvis skal ikke individer i kontrollgruppa endre adferd som resultat av eksponeringen de i eksponeringsgruppa utsettes for, og at effekten av eksponering ikke er avhengig av hvordan eksponerings- og kontrollgruppa konstrueres, eksempelvis av den relative størrelsen på gruppene. Morgan og Winship (2007: 38-39) gir flere eksempler på situasjoner hvor SUTVA er tvilsom, for eksempel kan betydningen av å gå på én type skole være avhengig av om eksperimentet fundamentalt endret hvor vanlig det er å gå på denne typen skole. Dette betyr ikke nødvendigvis at man ikke estimerer en kausal effekt dersom SUTVA er tvilsom, men man estimerer muligens en annen effekt enn den man i utgangspunktet var interessert i. I skoleeksemplet kan man kanskje si at man studerer effekter av å endre skolesystemet, heller enn å studere effekten av én type pedagogikk. Når det gjelder effekter av påvirkning og spredning mellom individer, er det å identifisere kausalitet i sosiale interaksjoner spørsmål som er helt i forskningsfronten (se for eksempel Dahl et al. 2012).

La oss avslutte med å si litt om tolkningen av resultatene fra IV-regresjoner. I en situasjon hvor vi har et instrument som er «så godt som» randomisert, som påvirker Y kun gjennom X, som er sterkt korrelert med X, og gitt monotoniantagelsen og SUTVA, har vi et forskningsdesign med veldig høy intern validitet. Vi kan være sikre på at vi har estimert en kausaleffekt, men det er viktig å forstå at dette er en kausaleffekt for en subgruppe av populasjonen. Vi har ikke estimert den kausale effekten av X på Y, men vi har estimert den kausale effekten av X på Y *for de som påvirkes av instrumentet*. Dette kalles LATE, local average treatment effect, og det er ofte usikkert hvor representativ LATE er for populasjonen som helhet, det vil si hvor god den eksterne validiteten av studien er. Det er godt mulig at de som påvirkes av instrumentet er en veldig spesiell subgruppe av populasjonen, og at det derfor er tvilsomt at kausaleffekten er den samme for andre deler av populasjonen. Et mye brukt

eksempel på dette er instrumentet med fødselskvartal som brukes av Angrist og Krueger (1991) for å se på effekten av utdanning på lønn som vi diskuterte tidligere. Husk at de som påvirkes av dette naturlige eksperiment er en meget spesiell gruppe; det er de som tvinges til å gå på skole lengere på grunn av at de ikke enda har fylt 16 år og endelig kan slutte på skolen. Det er rimelig å tro at effekten av utdanning for denne subgruppen er annerledes enn effekten av utdanning på befolkningen generelt.

Avslutning

Særlig internasjonalt, men også i Norge, er det en voksende interesse for å identifisere årsakssammenhenger. I denne artikkelen har vi presentert ulike empiriske strategier for å identifisere årsakssammenhenger i observasjonsdata. Disse metodene kan hjelpe forskere til å trekke kausale konklusjoner selv uten å ha tilgang til eksperimentell data. I de tilfeller dette er mulig mener vi at dette ofte gir mer verdifull kunnskap enn beskrivelse av sammenhenger. Disse metodene er imidlertid ikke alltid like lette å forstå, og det finnes mange fallgruver i bruken av dem. I denne artikkelen har vi særlig fokusert på instrumentvariabel- og regression discontinuity-metoden og vi har vist hvordan de fungerer og noen av de vanligste problemene man kan støte på ved bruk av dem. Selv for forskere som ikke har for planer om å bruke disse metodene er det viktig å skjønne hvordan de fungerer og hvilke potensielle problem de har, ikke minst da en voksende andel artikler bruker slike metoder.

Å avdekke effekter av tiltak, å vite om det er årsakssammenheng mellom ulike faktorer, eller å påvise at det ikke er slike sammenhenger er en viktig del av samfunnsvitenskapens raison d'être, og det bevilges store summer til samfunnsforskere for å evaluere ulike offentlige tiltak. Hvis vi visste mer om hvilke offentlig tiltak som virkelig fungerer og hvilke som ikke gjør det vil det være mulig å bygge et enda bedre samfunn med en mer effektiv utnyttelse av samfunnets ressurser.

Å påvise kausale sammenhenger krever at den data man jobber med muliggjør denne typen analyser. I og med at dataene stiller enkelte særegne krav må forskere som er interessert i kausale sammenhenger tenke *ekstra* nøye gjennom prosjektutformingen og datainnsamlingen for å forsikre seg om at de dataene man samler inn er egnet til å svare på det kausale spørsmålet. Av denne grunnen vil det også være nyttig om offentlige myndigheter inkluderer forskere som skal evaluere prosjekter allerede *før* prosjektene starter. Da vil det være mulig å randomisere langt flere offentlig prøvetiltak, noe som gjør at man kan identifisere hvilke tiltak som fungerer etter målsetningen, og hvilke som ikke gjør det. I USA har myndighetene i utstrakt grad innført prøveprosjekter på flere områder, blant annet arbeidsmarkedstiltak, på en slik måte at man kan trekke kausale slutninger om tiltakets effekt i ettertid, før man eventuelt implementerer tiltaket permanent. Denne måten å prøve ut tiltak på har vært mer kontroversiell i Europa, og så klart er det ikke alltid mulig å randomisere tiltak av praktiske og etiske grunner. I slike tilfeller kan man imidlertid i planleggingsfasen likevel legge til rette for den typen analyser vi har beskrevet i denne artikkelen. Ett eksempel kan være å gi bevilgninger ut fra tydelige definerte tildelingskriterier, og deretter dele disse tildelingskriteriene med forskere. En slik prosess vil være både transparent og vil gi gode muligheter for å identifisere kausale effekter.

Vi håper at vår artikkel kan bidra til å inspirere forskere til å stille kausale spørsmål og til samle inn data som gjør det mulig å besvare disse. Men, som vi understreket i innledningen, beskrivende analyser av data er en viktig del av samfunnsforskningen og vi håper at større bevissthet rundt kausalitet ikke gjør at man slutter å stille interessante og viktige spørsmål bare fordi man ikke kan identifisere kausalitet. Vår vurdering er imidlertid at vi er langt unna en slik situasjon i dag.

Referanser

Acemoglu, D., Johnson, S., &Robinson, J. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation, *American Economic Review*, 91(5), 772-793.

Angrist, J. D. & Kruger, A. B. (1991). Does Compulsory Schooling Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, 106(4), 976-1014.

Angrist, J. D. & Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics, *Journal of Economic Perspectives*, 24(2), 3-30.

Card, D. & Krueger, A. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania, *American Economic Review*, 84(4), 772-793.

Cools, S., Fiva, J.H., & Kirkebøen, L.J. (2011). *Causal Effects of Paternity Leave on Children and Parents*. CESifo Working Paper Series 3513.

Correll, S. J., Benard, S., & Paik I. (2007). Getting a job: Is there a motherhood penalty?, *American Journal of Sociology*, 112(1), 297–338.

Dahl, Gordon, Katrine V. Løken og Magne Mogstad (2012). *Peer Effects in Program Participation*. NBER Working Paper 18198.

Dunning, T. (2012). *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge: Cambridge University Press.

Eggers, A., & Hainmueller, J. (2009). MPs For Sale? Returns to Office in Postwar British Politics, *American Political Science Review*, 103(4), 513-533.

Finseraas, H., Jakobsson, N., & Kotsadam, A. (2011). Did the Murder of Theo van Gogh Change Europeans' Immigration Policy Preferences? *Kyklos*, 64(3), 396-409.

Glaeser E, La Porta L., Lopez-de-silanes F., and Schleifer, A. (2004). Do institutions cause growth? *Journal of Economic Growth*, Vol. 9(3): 271-303.

Green, D. P., & Gerber, A. S. (2000). The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment, *American Political Science Review*, 94(3), 653-663.

Green, D. P., & Gerber, A. S. (2008). *Get Out the Vote: How to Increase Voter Turnout*. Washington, DC: Brookings Institution Press.

Keane, M. P. (2010). A Structural Perspective on the Experimentalist School. *Journal of Economic Perspectives*, 24(2), 47–58.

Kirk, D.S. (2009). A Natural Experiment on Residential Change and Recidivism: Lessons from Hurricane Katrina. *American Sociological Review*, 74(3), 484–505.

Lalive, R. (2007). Unemployment Benefits, Unemployment Duration, and Post-Unemployment Jobs: A Regression Discontinuity Approach, *American Economic Review*, 97(2), 108–112.

Lee, D. S. & Lemieux, T. (2009). *Regression Discontinuity Designs in Economics*. NBER Working Paper 14723.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2), 698-714.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Murray, M. P. (2006). Avoiding Invalid Instruments and Coping with Weak Instruments, *Journal of Economic Perspectives*, 20(4), 111-132.

Murray, M. P. (2010). *The Bad, the Weak, and the Ugly: Avoiding the Pitfalls of Instrumental Variables Estimation*. Working Paper.

Schlotter, M., Schwerdt, G., & Woessmann, L. (2009). *Econometric Methods for Causal Evaluation of Education Policies and Practices: A Non-Technical Guide*. IZA DP No. 4725.

Staiger, D. & Stock, J. (1997) Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3), 557-586.

Thistlethwaite, D., & Campbell, D., (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6), 309–317.