politica, 58. årg., nr. 4, 77-90

Jonathan Stavnskær Doucette

Udvælgelse og anvendelse af kontrolvariable

Når vi forsøger at identificere den kausale effekt af X på Y med observationelt data, inkluderer vi ofte kontrolvariable. Der mangler dog tilstrækkelig vejledning i udvælgelse af variable. Det kan resultere i bias, både hvis relevante kontroller udelades, og hvis såkaldte "dårlige kontroller" inkluderes. Denne artikel præsenterer en række retningslinjer for udvælgelse og anvendelse af kontrol med det formål at identificere kausale effekter.

Nøgleord: kausal inferens, statistisk kontrol, regression

Forskere, studerende og dataanalytikere, der anvender observationelt data, inkluderer ofte kontrolvariable i deres regression for at fjerne udeladt variabel bias (fx Clarke, 2005; Angrist og Pischke, 2008). Denne type bias kan elimineres ved kontrol for alle variable, som både determinerer X (treatment) og Y (udfald). Denne antagelse er kendt som *conditional independence*, *selection on observables*, *strong ignorability* eller *the backdoor criterion*. Vi kan således komme tættere på at identificere den kausale effekt af X på Y ved at inkludere relevante kontrolvariable.

I praksis eksisterer der en række misforståelser omkring udvælgelsen af kontrolvariable. Et eksempel er, at omfanget af udeladt variabel-bias er tilbøjeligt til at være mindre, når vi inkluderer flere kontrolvariable. Et andet er, at variable, hvis inklusion ændrer på koefficienten for X, bør inkluderes. Et tredje er, at kontrol for en variabel, som er korreleret med en determinant af X og Y, reducerer bias. Der eksisterer en række tilfælde, hvor inklusionen af en ekstra variabel ligefrem kan forstærke bias – såkaldte dårlige kontroller (fx Rosenbaum, 1984; Clarke, 2005; Angrist og Pischke, 2008; Steiner og Kim, 2016; Cinelli, Forney og Pearl, 2024).

Denne artikel præsenterer en række retningslinjer, der kan anvendes til at skelne såkaldte gode kontroller fra dårlige kontroller. Derudover diskuterer den vigtigheden af at approksimere den korrekte funktionelle form, efter at man har udvalgt kontrolvariable. Bemærk ydermere, at kontrol blot er en blandt flere tilgange, såsom difference-in-difference og RDD, der kan anvendes til at minimere bias i analyser af observationelt data (se fx Angrist og Pischke, 2015; Cunningham, 2021; Huntington-Klein, 2021).

Første sektion af artiklen gennemgår formålet med kontrol og diskuterer, hvornår det er relevant at inddrage kontrolvariable. Anden sektion gennemgår en række scenarier, hvor inddragelse af en given variabel vil afhjælpe bias – de gode kontrolvariable. Tredje sektion præsenterer de kontrolvariable, som ikke afhjælper bias eller i de værste tilfælde ligefrem forværrer graden af bias. Dernæst diskuteres, hvorfor den korrekte specifikation af funktionel form også er relevant for korrekt inddragelse af kontrolvariable. Endelig konkluderes artiklen med en kort opsummering.

Hvad er formålet med kontrol

Inklusion af variable i en regression kan tjene flere formål. Vi kan forsøge at forudsige eller prædiktere et udfald. For eksempel kunne vi være interesseret i at vide, hvorvidt socialdemokraterne vinder borgmesterposten i en kommune til næste valg. Her ville vi forsøge at inkludere de variable, som giver os den bedste chance for at ramme rigtigt med vores forudsigelse. Alternativt kunne vi være interesseret i at kende effekten af en variabel på et udfald, for eksempel effekten af et ekstra års uddannelse på livsindkomst. Denne artikel beskæftiger sig kun med overvejelser, der er relevante, når vi forsøger at identificere en kausal effekt.

I en ideel verden ville vi afdække vores kausale spørgsmål via et eksperiment. For eksempel ville vi tilfældigt tildele et ekstra års uddannelse (*treatment*) til en gruppe individer (*treatment-gruppen*) og sammenligne deres senere livsindkomst med den gruppe af individer, der tilfældigvis ikke fik tildelt et ekstra års uddannelse (*kontrolgruppen*). Den tilfældige tildeling af treatment sikrer, at treatment-gruppen og kontrolgruppen er tilnærmelsesvis ens bortset fra treatment. Således vil andre faktorer såsom forældres baggrund ikke kunne forklare en eventuel forskel i livsindkomst mellem grupperne. Desværre er det ofte ikke muligt at afdække kausale spørgsmål via eksperimenter på grund af ressourcemæssige begrænsninger eller etiske overvejelser.

Et alternativt værktøj til at afdække kausale spørgsmål, når eksperimenter ikke er mulige, er regression med kontrol. Her er idéen at 1) sammenligne enheder udsat for treatment med enheder, der ikke er udsat for treatment, og 2) sikre, at enhederne i de to grupper minder om hinanden på observerbare karakteristika. Forhåbningen er, at vi kan minimere såkaldt selektionsbias[1] ved at kontrollere for observerbare karakteristika, som forklarer både, hvorfor nogle enheder modtager treatment, og hvorfor nogle enheder har et andet udfald.

Generelt kan det siges, at en observeret forskel mellem treatment- og kontrolgruppen udgøres af to komponenter. Den gennemsnitlige kausale effekt af treatment (den komponent vi gerne vil kende) og selektionsbias (den komponent vi vil eliminere): Forskel mellem grupper (β) = gnm. kausal effekt (β_k) +

selektionsbias (β_s) (Angrist og Pischke, 2015).[2] Det primære formål med kontrol er at minimere β_s.

Såfremt vi vil identificere en kausal effekt via kontrol, antager vi altså, at vi har inkluderet alle de variable, som kan forklare, hvorfor treatment- og kontrolgruppen ville have haft et forskelligt udfald selv i fravær af treatment. Det svarer til at antage, at på betingelse af vores kontroller, så er treatment *tilfældigt* tildelt. I praksis er dette en antagelse, der ofte er svær at forsvare. Et eksempel kunne være, at vi vil kende effekten af demokrati på økonomisk udvikling. Vi kører en regression af BNP/ind_i på $demokrati_i$ ($BNP/ind_i = \alpha + \beta demokrati_i + \epsilon_i$). Vi finder, at demokratier gennemsnitligt er omtrent $ 15.000 rigere end autokratier ($\widehat{\beta} = 15.000$).[3] Er dette den kausale effekt af demokrati på velstand? Nok ikke.

Denne forskel mellem demokratier og autokratier indeholder både den faktiske effekt af demokrati på velstand (β_k) og selektionsbias (β_s). Hvorfra stammer selektionsbias? I dette tilfælde har det nok mange årsager (se fx Koyama og Rubin, 2022; Miller, 2021). En årsag kunne være geografi. Lande med havadgang og lav sygdomsbyrde har bedre muligheder for handel og økonomisk udvikling (Gallup, Sachs og Mellinger, 1999; Acemoglu, Johnson og Robinson, 2001). Havadgang og sygdomsbyrde har også haft indvirkning på overførslen af prædemokratiske institutioner fra Europa til resten af verdens lande (Acemoglu, Johnson og Robinson, 2001; Hariri, 2012; Gerring et al., 2022). Da lande med gunstig geografi både er mere tilbøjelige til at være demokratiske og velstillede, betyder udeladelsen af geografi i vores regression, at estimatet for forskellen mellem demokratier og autokratier indeholder positiv selektionsbias ($\beta_s > 0$) såvel som effekten af demokrati (β_k). Hvis vi tager de $ 15.000 for gode varer, kommer vi til at overvurdere effekten af demokrati på velstand. Hvis vi skal finde effekten af demokrati på velstand (β_k), er vi derfor nødt til at kontrollere for geografi.

Har vi så identificeret den kausale effekt af demokrati på velstand, hvis vi inkluderer kontrol for geografi? I så fald antager vi, at det er tilfældigt, om et land er blevet demokratisk eller ej, såfremt vi holder forskelle i for eksempel havadgang og gennemsnitstemperatur lige. Den antagelse er svær at forsvare.

Kontrol er altså noget, vi bruger for at eliminere alternative forklaringer på forskelle i treatment og udfald. I praksis er det sjældent simpelt at udvælge de kontrolvariable, der sikrer, at selektionsbias forsvinder. I næste sektion gennemgår artiklen derfor, hvad der kendetegner gode kontrolvariable (såsom geografi i ovenstående tilfælde).

Gode kontroller

Gode kontroller mindsker selektionsbias (β_s). Dette kan indebære, at estimatet for sammenhængen mellem X og Y ($\widehat{\beta}$) styrkes eller svækkes. Det er således ikke utænkeligt, at inklusion af gode kontroller medfører, at sammenhængen forsvinder, hvis den kausale effekt af X er meget lille eller ikkeeksisterende ($\beta_k \approx 0$). For at illustrere hvad der kendetegner gode kontroller, gøres brug af *directed acyclic graphs* (DAG). Afsnittet indledes derfor med en kort og simplificeret introduktion til disse (for en mere dybdegående behandling af emnet se Pearl, 2009; Cunningham, 2021: kap. 3).

DAGs bruges til at visualisere ens kvalitative forståelse af, hvordan data er genereret i den virkelige verden. Store bogstaver, for eksempel X, repræsenterer stokastiske variable.[4] Pile symboliserer en potentiel kausal effekt, $X \rightarrow Y$, af X på Y. Bemærk, at der ikke gøres nogen antagelser om den funktionelle form af sammenhængen.[5] I den følgende diskussion indgår tre slags sammenhænge. Fælles årsager – variable (Z), der påvirker både X og Y – repræsenteres $X \leftarrow Z \rightarrow Y$. Tilstedeværelsen af disse variable inducerer en ikkekausal sammenhæng mellem X og Y. Kontrol for disse blokerer den ikkekausale sti fra X til Y. Disse indgår i diskussionen af gode kontroller.

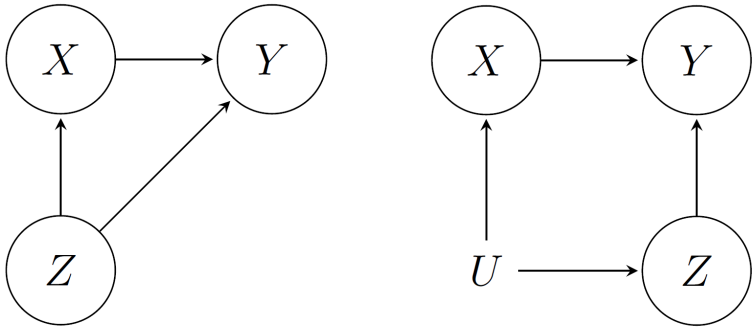
Mediatorer – variable (M), der repræsenterer mekanismer, hvorigennem X har en effekt på Y – repræsenteres $X \rightarrow M \rightarrow Y$. Kontrol for disse blokerer denne sti mellem X og Y. Sammenfaldne effekter (*colliders*) – variable påvirket af både X og Y – repræsenteres $X \rightarrow Z \leftarrow Y$. Tilstedeværelsen af colliders medfører ikke, at der er en sammenhæng mellem X og Y i udgangspunktet. Kontrol for en collider introducerer dog en ikkekausal sammenhæng mellem X og Y (Pearl, 2009; Cinelli, Forney og Pearl, 2024: 1073-1074). Mediatorer og colliders diskuteres derfor yderligere i afsnittet om dårlige kontroller.

Bemærk, at stier kan blokeres ved blot at kontrollere for en variabel på stien. Lad os antage, at vi gerne vil kende effekten af en universitetsuddannelse (X) på indkomst (Y). Vi forestiller os, at medfødte evner (U) påvirker både sandsynligheden for at tage en universitetsuddannelse og indkomst. Vi har intet mål for evner (U), men vi antager, at evner primært påvirker X og Y gennem tidligere akademisk succes[6] målt som karaktersnit ved folkeskolens afgangseksamen (Z). Denne sammenhæng kan udtrykkes $X \leftarrow Z \leftarrow U \rightarrow Y$. Her vil kontrol for Z (karaktersnit) eliminere den ikkekausale del af sammenhængen mellem X og Y, som stammer fra medførte evner (U). (U) repræsenterer generelt variable, som vi ikke har et mål for.

Gode kontroller er således fælles årsager, der har en effekt på både X og Y ($X \leftarrow Z \rightarrow Y$). Alternativt er det variable, som blokerer en sti mellem en fælles årsag og enten X eller Y (fx $X \leftarrow D \rightarrow Z \rightarrow Y$). Figur 1 illustrerer to eksempler

på gode kontrolvariable. Variablene Y, X og Z er observerede, mens U er uobserveret.

Figur 1: Gode kontroller



Scenario (a) er det klassiske eksempel på en god eller relevant kontrol. Z har en effekt på både X og Y. Udeladelsen af kontrol for Z vil derfor inducere en ikkekausal sammenhæng mellem X og Y.

I scenario (b) er Z også en god kontrol. Årsagen er dog en anden. Her påvirker Z ikke både X og Y direkte. I dette tilfælde er U, som er uobserveret, den variabel, som inducerer selektionsbias. Selvom vi ikke har observeret U, kan vi dog tage højde for variablen. Fordi hele U's effekt på Y går gennem Z, kan vi blokere dens sti ved kontrol for Z. Så længe en af stierne mellem U og X eller U og Y er blokeret, kan vi finde den kausale effekt af X på Y. I praksis er det dog ofte svært at forsvare, at hele effekten af U går gennem Z. Bemærk her, at Z ikke behøver at være determineret før X i tid. Inklusion af Z vil i scenario (b) mindske selektionsbias, selv hvis Z er *post-treatment*.

Kontrol kan sommetider afhjælpe selektionsbias, men det er relativt sjældent, at strategien sikrer, at vi faktisk finder den kausale effekt (β_k). Generelt fungerer kontrol bedst i de tilfælde, hvor vi har rigtig god viden om datagenereringsprocessen. Altså skal vi gerne vide så meget som muligt om, hvorfor nogle enheder modtager treatment, mens andre ikke gør. Et eksempel er, når treatment tildeles på baggrund af administrative beslutninger, hvor vi kender og har mål for de kriterier, som anvendes til at træffe beslutninger.

Dale og Krueger (2024) præsenterer et eksempel, hvor kontrol potentielt kan hjælpe os med at finde den kausale effekt. De undersøger effekten af optagelse på et selektivt universitet på senere indkomst i USA. De har indsamlet data på

de kriterier, såsom SAT score, universiteterne anvender til at bestemme, hvilke ansøgere der skal tilbydes en plads. De har således mål for mange af de faktorer, som direkte påvirker optag og højst sandsynligt også senere indkomst (scenario (a)). Derudover kontrollerer de for den gennemsnitlige SAT score for de universiteter, hver ansøger har søgt optag på. Idéen er, at de hermed kan kontrollere for uobserverede karakteristika som for eksempel ambitionsniveau (scenario (b)).

Cornell, Knutsen og Teorell (2020) undersøger effekten af et velfungerende bureaukrati på økonomisk velstand på tværs af lande. Her er der tale om en sammenhæng, hvor det er meget usandsynligt, at vi kan identificere en kausal effekt ved brug af kontrol. I litteraturen om bureaukrati og velstand er der mange bud på årsager til deres fremkomst og stor uenighed om, hvilke årsager der har størst betydning (se fx Tilly, 1975; Ertman, 1997; Acemoglu, Johnson og Robinson, 2001; Abramson og Boix, 2019; Acemoglu et al., 2019). Faktisk er det plausibelt, at vi har flere potentielle årsager, end vi har enheder (lande). Selv hvis vi havde mål for samtlige årsager til både bureaukrati og velstand, er det sandsynligt, at nogle af disse mål ville være, hvad man kalder dårlige kontroller. Cornell, Knutsen og Teorell (2020: 2247 og 2273-2274) er opmærksomme på dette problem og udnytter i stedet blandt andet den tidslige dimension i deres data til at mindske selektionsbias.

Endelig er det værd at nævne, at det nogle gange kan give mening at inkludere variable, der ikke determinerer X, men som er gode til at forudsige Y. Dette mindsker variansen og kan give et mere præcist estimat (mindre standardfejl) for effekten af den eller de variable, vi er interesserede i (Wooldridge, 2020: 200-201). Det er dog vigtigt at sikre sig, at der ikke er tale om en såkaldt *dårlig kontrol.*

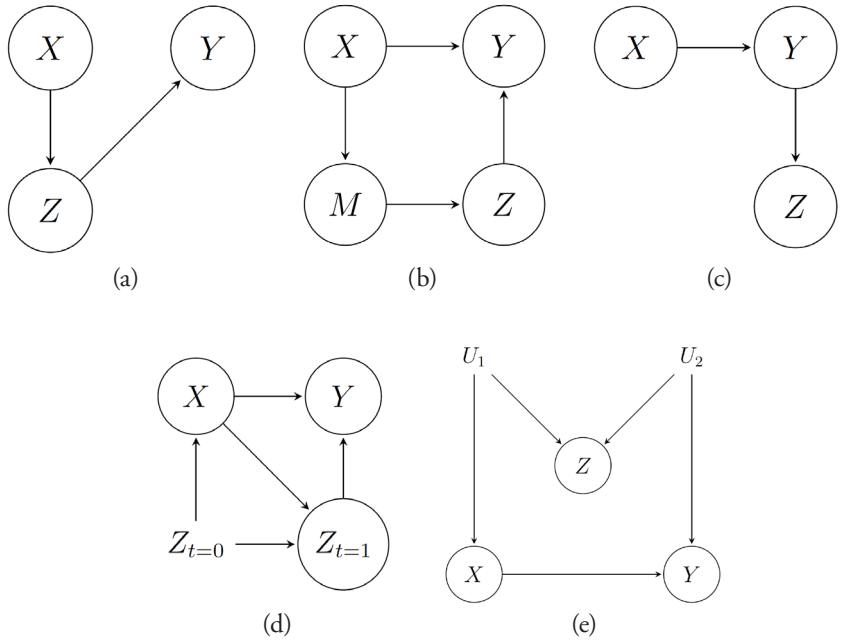
Dårlige kontroller

Udover at dårlige kontroller ikke minimerer selektionsbias, så er der faktisk en risiko for, at deres inklusion medfører yderligere bias. Dette afsnit illustrerer forskellige eksempler på kontrolvariable, hvis inklusion mindsker sandsynligheden for, at en estimeret sammenhæng kan tolkes kausalt.

To relaterede typer af bias, der ofte nævnes, er *post-treatment bias*, dvs. bias der stammer fra kontrol for en konsekvens af X, og *collider bias*, dvs. bias der stammer fra kontrol for en konsekvens af X og Y (Rosenbaum, 1984; Elwert og Winship, 2014). Nogle litteraturer lægger mere vægt på én type dårlig kontrol frem for en anden. Fælles for dem alle er dog, at inklusionen af en kontrol tilføjer bias og derved gør det sværere at estimere den kausale effekt af X. Figur 2 præsenterer fire eksempler på dårlige kontroller. Der eksisterer en række yderli-

gere variationer af disse scenarier, men de har samme udfordringer (se Cinelli, Forney og Pearl, 2024 for flere eksempler).

Figur 2: Dårlige kontroller



Scenario (a) viser det klassiske eksempel på en dårlig kontrol, der forårsager post-treatment bias (Rosenbaum, 1984). Her går X's effekt på Y gennem Z ($X \rightarrow Z \rightarrow Y$). Hvis vi blokerer for Z, vil vi fjerne den faktiske kausale effekt af X på Y. Scenario (b) viser et relateret eksempel. Her går X's effekt på Y ikke direkte gennem Z. Z påvirkes dog af M ($M \rightarrow Z$), som repræsenterer en sti, hvormed X påvirker Y ($X \rightarrow M \rightarrow Z \rightarrow Y$). Ved kontrol for Z vil vi igen fjerne noget af den faktiske effekt af X på Y.

Scenario (c) er et eksempel på det, der kaldes *case-control bias* (Cinelli, Forney og Pearl, 2024). Umiddelbart skulle man mene, at kontrol for Z var harmløs. Z er ikke en konsekvens af X, og Z påvirker heller ikke Y. Vi inducerer dog bias alligevel, da Z er en konsekvens af Y, $Y \rightarrow Z$ (og Y igen er en konsekvens af X, $X \rightarrow Y$). Her bør kontrol for Z udelades.

Scenario (d) er særlig problematisk og desværre ofte forekommende. Her er Z, hvad vi normalt vil kalde en god kontrol (dvs. Z påvirker både Y og X, $X \rightarrow$

$Z \rightarrow Y$). Z er også en konsekvens af X ($X \rightarrow Z$) og derved en dårlig kontrol. Her vil vi enten tillade selektionsbias at influere vores estimat ved at udelade Z eller inducere bias ved at kontrollere for en konsekvens af X, hvis vi inkluderer Z. I begge tilfælde kan vi ikke identificere en kausal effekt, og vi kan ikke vide, hvorvidt inklusion af Z vil føre til mere eller mindre bias. Såfremt man mistænker, at man befinder sig i scenario (d), er en mulighed at køre en model med og uden Z. Hvis koefficienten for X ikke ændrer sig substantielt på tværs af modellerne, kan det øge ens tiltro til, at Z ikke er en vigtig kilde til bias. Omvendt kan en stor ændring i koefficienten indikere, at man har et problem med enten selektionsbias eller post-treatment bias (eller begge).

Scenario (e) er en illustration af *M-bias*. Her er der i udgangspunktet ikke bias i sammenhængen mellem X og Y. Hvis man kontrollerer for Z, inducerer man dog bias ved at koble de uobserverede årsager (U_1, U_2) til henholdsvis X og Y sammen ($X \rightarrow U_1 \rightarrow Z \rightarrow U_2 \rightarrow Y$).

Et andet mere konkret eksempel på scenario (e) kunne være, at vi er interesseret i effekten af perfektionisme (X) på ydmyghed (Y). Vi overvejer at kontrollere for andre personlighedstræk såsom parathed til at tilgive (Z). Vi ved, at skolegang (U_1) påvirker perfektionisme og tilbøjeligheden til at tilgive ($U_1 \rightarrow X$ og $U_1 \rightarrow Z$). Vi har ikke observeret skolegang. Vi ved, at tilstedeværelsem af religion i barndommen (U_2) påvirker tilbøjeligheden til ydmyghed og til at tilgive ($U_2 \rightarrow Y$, $U_2 \rightarrow Z$). Vi har ikke observeret tilstedeværelsen af religion i barndommen (eksemplet er taget fra Bulbulia, 2022). Her vil udeladelsen af kontrol for tilbøjelighed til at tilgive (Z) medføre, at vi kan identificere en kausal effekt. Omvendt vil inklusion af Z medføre bias, da skolegang og tilstedeværelsen af religion i barndommen nu kobles sammen ($X \rightarrow U_1 \rightarrow Z \rightarrow U_2 \rightarrow Y$).

M-bias er en relativt kompliceret størrelse (se fx Pearl, 2009; Sjölander, 2009; Ding og Miratrix, 2015 for yderligere diskussion). Det er dog værd at bemærke, at den relativt sjældent forekommer sammenlignet med scenario (a), (b), (c) og (d). Scenariet er dog værd at huske, da det blandt andet illustrerer, at selv variable, som er *pre-treatment*,[7] kan være dårlige kontroller.

Generelt er der en høj risiko for at inkludere dårlige kontroller i scenarier, hvor der er gensidig påvirkning mellem X, Y og Z og uklar tidsrækkefølge. For eksempel er det blevet teoretiseret, at økonomisk udvikling (X) påvirker demokrati (Y) (Lipset, 1959). Borgerkrig (Z) er ofte forbundet med efterfølgende lavere økonomisk vækst og autokratisering (Kang og Meernik, 2005; Joshi, 2010). Det er dog også sandsynligt, at demokrati påvirker velstand (Acemoglu et al., 2019) og borgerkrig (Cederman, Hug og Krebs, 2010). Her er det plausibelt, at vi er i noget, der minder om scenario (d), hvor borgerkrig (Z_1) både er en relevant kontrol og en dårlig kontrol. Ydermere er vores treatment (økonomisk

udvikling) til dels også en konsekvens af vores Y (demokrati). Vi kunne overveje at kontrollere for niveauet af demokrati tilbage i tid[8] (Z_2) for at tage højde for dette. Det kan dog også ses som en variation af scenario (d). Velstand (X) har blandt andet en effekt på demokratiniveau i dag (Y), fordi det har været med til at sikre fremkomsten af prodemokratiske grupper (Z_2) tilbage i tid ($X \rightarrow Z_2 \rightarrow Y$). I sådanne scenarier vil kontrol derfor sjældent være en brugbar strategi til at identificere kausale effekter.

Bemærk, at inklusionen af en dårlig kontrol kun er en blandt flere måder, hvorpå det at betinge på for eksempel post-treatment faktorer giver bias. Det samme gør sig blandt andet gældende, hvis man smider respondenter ud, der ikke klarer opmærksomhedstjek, laver en interaktion med post-treatment variable eller vælger sit sample på baggrund af treatment.

Funktionel form

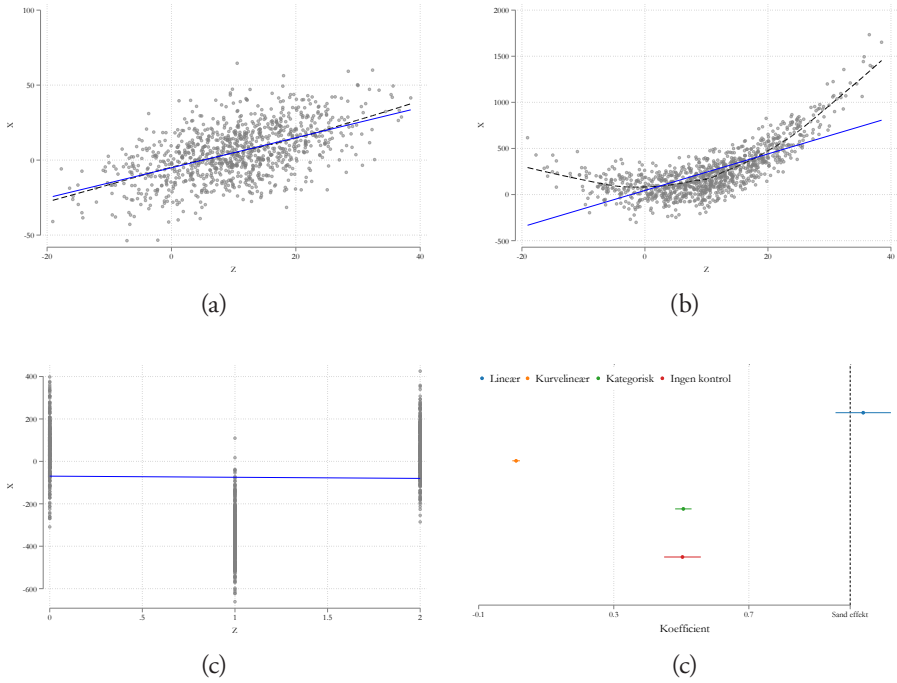
Ovenstående afsnit har gennemgået en række refleksioner, man bør gøre sig, når man påtænker at inkludere kontrolvariable. Det er dog værd at understrege, at man først og fremmest bør overveje, om kontrol faktisk er den bedste strategi til at minimere selektionsbias. Det er sjældent praktisk muligt at inkludere alle relevante kontroller uden samtidig at medtage en dårlig kontrol. Selv med observationelt data kan der ofte anvendes andre tilgange såsom difference-in-difference, RDD og IV (se fx Angrist og Pischke, 2008; Huntington-Klein, 2021; Cunningham, 2021 for en introduktion til disse metoder).

Når det er sagt, er det ikke nok blot at udvælge de rigtige kontrolvariable. Nedenfor gennemgås yderligere overvejelser om funktionel form, man bør gøre sig i forbindelse med anvendelse af kontrol.

En ofte overset detalje er, at det ikke er nok, at man har approksimeret den korrekte funktionelle form mellem X og Y. Hvis kontrol succesfuldt skal eliminere selektionsbias, er det også nødvendigt at tilnærme sig den funktionelle form af sammenhængen mellem Z og X/Y. Husk her, at vores DAG-visualiseringer ikke fortæller os noget om denne.

Graferne (a), (b) og (c) i figur 3 viser tre simulerede eksempler på sammenhængen mellem kontrol (Z) og treatment (X). Graf (d) viser den estimerede effekt af X på Y i de tre scenarier, givet at vi har kontrolleret for Z. I scenario (1), den blå koefficient, er den funktionelle form mellem Z og X/Y lineær, og vi har medtaget Z som en lineær kontrol.[9] Her er vores estimat ($\widehat{\beta}$ = 1,03) meget tæt på den sande effekt i populationen (β = 1). I scenario (2), den orange koefficient, er den funktionelle form mellem Z og X/Y kurvelinear, og vi har medtaget Z som linear kontrol. Her er koefficienten for sammenhængen mellem X og Y meget svag ($\widehat{\beta}$ = 0,02), og koefficientens konfidensinterval overlapper med 0. Dette

Figur 3: Eksempler på funktionel form



sker, selvom vi har inkluderet den korrekte kontrol i vores regression. I scenario (3), den grønne koefficient, er forholdet mellem Z og X/Y forskelligt på tværs af Z's tre kategorier. Vi har igen medtaget Z som lineær kontrol. Her er koefficienten for sammenhængen mellem X og Y halveret ($\widehat{\beta}$ = 0,51) sammenlignet med den sande effekt. Den sidste røde koefficient er vores bud på sammenhængen ($\widehat{\beta}$ = 0,50), når vi undlader at kontrollere for Z (og Z er kategorisk relateret til X/Y). Bemærk, at vi er tættere på den sande effekt end i scenario (2) og (3), selvom vi har undladt en god kontrol.

En forkert specificeret kontrolvariabel kan altså i grelle tilfælde inducere yderligere selektionsbias, selvom det en god kontrol. Dette problem kan dog, i modsætning til udfordringer ved valg af kontrolvariable, løses med standardværktøjer (se fx Wooldridge, 2012: 41-44, 191-200). Anvendelse af kontrol som strategi til at minimere selektionsbias kræver altså både viden om datagenereringsprocessen og korrekt diagnosticering af den funktionelle form.

Konklusion

Kontrol er ofte anvendt som strategi til at minimere selektionsbias. I mange applikationer er det dog uklart, om det faktisk er en succesfuld eller måske ligefrem skadelig strategi, da det ikke er sikkert, at de udvalgte kontroller kun repræsenterer gode og ikke dårlige kontroller. Denne artikel har præsenteret en række retningslinjer, der kan anvendes til at skelne såkaldte gode kontroller fra dårlige kontroller. Det bør dog altid overvejes, om kontrol faktisk er den bedste strategi til at minimere selektionsbias. Det er mest sandsynligt, at kontrol kan afhjælpe selektionsbias i de tilfælde, hvor man har god viden om datagenereringsprocessen, og hvor man har diagnosticeret den funktionelle form korrekt.

Noter

1. Selektionsbias er defineret som forskellen i udfald mellem treatede og ikketreatede enheder i en kontrafaktisk virkelighed, hvor de treatede enheder ikke modtog treatment (Angrist og Pischke, 2015: 10). Vi kunne for eksempel sammenligne den gennemsnitlige livsindkomst hos individer med en universitetsgrad med livsindkomsten hos individer uden en universitetsgrad. Her vil selektionsbias være den forskel i livsindkomst, vi ville observere mellem grupperne, såfremt de universitetsuddannede ikke havde gået på universitetet. Vi kan ikke med sikkerhed sige, hvad den forskel ville være – vi har jo ikke observeret den kontrafaktiske virkelighed. I dette tilfælde ville vi dog med en vis tryghed gætte på, at gruppen nok gennemsnitligt set ville have haft en større livsindkomst alligevel på grund af de forskelle i forældrebaggrund, netværk og lign., som var medvirkende til, at de kom ind på universitetet.
2. Man anvender nogle gange betegnelsen udeladt variabel bias. Dette er også indeholdt i selektionsbias. Navnet antyder dog, at bias bør fjernes ved at inkludere de udeladte variable. I praksis eksisterer der mange andre strategier end kontrol til at fjerne bias. Derfor anvendes selektionsbias fremadrettet.
3. Teksten og eksemplet gennemgår for forståelsens skyld et tilfælde, hvor vi klart kan gruppere enhederne i en treatment- og en kontrolgruppe. Logikken er dog præcis den samme, hvis vi i stedet undersøgte effekten af for eksempel grader af demokrati. Bemærk også, at β har fået en $\hat{}$ på, da vi nu har at gøre med estimat af β baseret på et (fiktiv) datasæt.
4. Tilfældige variable, hvis værdi vi ikke kender på forhånd. Bruges ofte til at repræsentere de faktorer, vi er interesserede i, såsom demokrati i det tidligere eksempel. Termen stokastisk variabel kommer fra statistik og sandsynlighedsteori, og det er ikke et udtryk for, om en variabel er påvirket eksperimentelt.

5. $X \rightarrow Y$ kan blandt andet symbolisere, 1) at højere X medfører højere Y, eller 2) at højere X medfører højere Y ved lave niveauer af X, og at højere X medfører lavere Y ved høje niveauer af X.
6. En urealistisk antagelse i virkelighedens verden.
7. Dvs. determineret før tildeling af treatment.
8. Hvad der kaldes en *lagged dependent variable* i litteraturen ($Y_{i,t-1}$). Denne tilgang er ofte brugt, se fx Cornell, Knutsen og Teorell (2020).
9. I STATA: *reg y x z*.

Litteratur

Abramson, Scott og Carles Boix (2019). Endogenous parliaments: The domestic and international roots of long-term economic growth and executive constraints in Europe. *International Organization* 73 (1): 793-837.

Acemoglu, Daron, Simon Johnson og James Robinson (2001). The colonial origins of comparative development: an empirical investigation. *The American Economic Review* 91 (5): 1369-1401.

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo og James Robinson (2019). Democracy does cause growth. *Journal of Political Economy* 127 (1): 47-100.

Angrist, Joshua D. og Jörn-Steffen Pischke (2008) *Mostly Harmless Econometrics*. Princeton University Press.

Angrist, Joshua D. og Jörn-Steffen Pischke (2015). *Mastering Metrics*. Princeton University Press.

Bulbulia, Joseph (2022). M-bias: Confounding control using three waves of panel data. 22. november.

Cederman, Lars-Erik, Simon Hug og Lutz F. Krebs (2010). Democratization and civil war: empirical evidence. *Journal of Peace Research* 47 (4): 377-394.

Cinelli, Carlos, Andrew Forney og Judea Pearl (2024). A crash course in good and bad controls. *Sociological Methods and Research* 53 (3): 1071-1104.

Clarke, Kevin A. (2005). The phantom menace: omitted variable bias in econometric research. *Conflict Management and Peace Research* 22 (4): 341-352.

Cornell, Agnes, Carl Henrik Knutsen og Jan Teorell (2020). Bureaucracy and growth. *Comparative Political Studies* 53 (14): 2246-2282.

Cunningham, Scott (2021). *Causal Inference: The Mixtape*. Yale University Press.

Dale, Stacy og Alan Krueger (2024). Estimating the effects of college characteristics over the career using administrative earnings data. *The Journal of Human Resources* 49 (2): 323-358.

Ding, Peng og Luke W. Miratrix (2015). To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. *Journal of Causal Inference* 3 (1): 41-57.

Elwert, Felix og Christopher Winship (2014). Endogenous selection bias: the problem of conditioning on a collider variable. *Annual Review of Sociology* 40: 31-53.
Ertman, Thomas (1997). *Birth of the Leviathan*. Cambridge University Press.
Gallup, John, Jeffrey Sachs og Andrew Mellinger (1999). Geography and economic development. *International Regional Science Review* 22 (2).
Gerring, John, Brendan Apfeld, Tore Wig og Andreas Forø Tollefsen (2022). *The Deep Roots of Modern Democracy: Geography and the Diffusion of Political Institutions*. Cambridge University Press.
Hariri, Jacob (2012). The autocratic legacy of early statehood. *American Political Science Review* 106 (3): 471-494.
Huntington-Klein, Nick (2021). *The Effect: An Introduction to Research Design and Causality*. CRC Press.
Joshi, Madhav (2010). Post-civil war democratization: promotion of democracy in post-civil war states, 1946-2005. *Democratization* 17 (5): 826-855.
Kang, Seonjou og James Meernik (2005). Civil war destruction and the prospects for economic growth. *The Journal of Politics* 67 (1).
Koyama, Mark og Jared Rubin (2022). *How the World Became Rich: The Historical Origins of Economic Growth*. Cambridge: Polity.
Lipset, Seymour Martin (1959). Some social requisites of democracy: economic development and political legitimacy. *American Political Science Review* 53 (1): 69-105.
Miller, Michael (2021). *Shock to the System: Coups, Elections, and War on the Road to Democratization*. Princeton University Press.
Pearl, Judea (2009) *Causality: Models, Reasoning, and Inference*, 2. udg. Cambridge University Press.
Rosenbaum, Paul (1984). The consequences of adjustment for a concomitant variable that has been affected by treatment. *Journal of the Royal Statistical Society* 147 (5): 656-666.
Sjölander, Arvid (2009). Propensity scores and M-structures. *Statistics in Medicine* 28 (9): 1416-1420.
Steiner, Peter M. og Yongnam Kim (2016). The mechanics of omitted variable bias: bias amplification and cancellation of offsetting biases. *Journal of Causal Inference* 4 (2).
Tilly, Charles (1975). *The Formation of National States in Western Europe*. Princeton University Press.
Wooldridge, Jeffrey M. (2012). *Introductory Econometrics: A Modern Approach*. South-Western CENGAGE Learning.
Wooldridge, Jeffrey M. (2020). *Introductory Econometrics*. Cengage.

Om forfatteren
Jonathan Stavnskær Doucette, lektor, Institut for Samfund og Politik, Aalborg Universitet, Danmark. E-mail: jostdo@society.aau.dk