

UiO : Department of Mathematics
University of Oslo

Identification of Causal Effects using Instrumental Variables (Angrist, Imbens, Rubin,1996).

Per August Jarval Moen

About the authors

- **Joshua Angrist**: Israeli-American professor of Economics @MIT, born 1960. Expertise: labor economics and econometrics (empirical economics).
- **Guido W. Imbens**: Dutch-American professor of Economics @Stanford, born 1963. Expertise: econometrics (empirical economics).



- **Donald B. Rubin:** American professor of Statistics @Harvard, born 1943. Known for the Rubin Causal Model (the potential outcomes framework).

Some remarks

- Angrist and Imbens will receive the Nobel Prize in Economics in December (together with David Card) for their work in Econometrics
- The last 30-40 years, economists have pioneered in the study of causal relationships in observational data (non-randomized data).
- Have economists beat statisticians at their own game?

- Breiman: statisticians should work more in category B!
- Econometricians: we work in category “super A” and have made huge advances in Causal Inference
- -> Where is your Nobel Prize, Breiman?

The current paper

- Among the papers listed in the Nobel Prize announcement
- The paper transfers tools from Econometrics (the Instrumental Variable) to the Rubin Causal Model
- We give will give a short account of classical concepts in Econometrics before diving into the article.

The Structural Equations Model

- A (linear) *Structural Equations Model* (SEM) is defined by a DAG $\mathcal{G}$ consisting of nodes $\{1, \dots, p\}$ and edges representing causal pathways, and a set of variables $\{X_i\}_{i=1}^{p}$ such that

$$X_i = \sum_{j \in pa(i)} \beta_{j,i} X_j + \epsilon_i \quad (*) \ ,$$

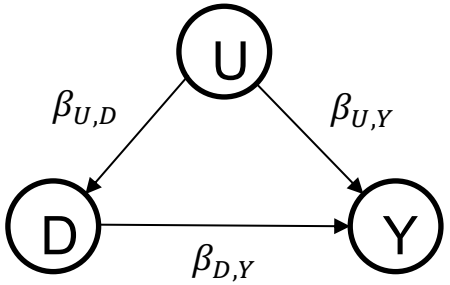
  where the ϵ_i's are independent and $pa(i)$ is the parent set of node i in $\mathcal{G}$.
- What makes the model *structural* is that $(*)$ is assumed to hold even when we intervene on the system and set some of the X_i's to some fixed values.
- The coefficient $\beta_{j,i}$ is interpreted as the causal effect of X_j on X_i.

The Structural Equations Model

- We implicitly assume a structure from which the data is generated, which allows us to define *causal effects* in a meaningful manner.
- A SEM implies a linear model, but the converse is not true
- SEMs have been a centerpiece Econometrics and the Social Sciences
- SEMs date back to Sewall Wright and Trygve Haavelmo (Norwegian Economist, recipient of the Nobel Prize)

Example of a SEM

- $\mathcal{G}$:



- $U = \epsilon_U$,
  D$= \beta_{U,D}U + \epsilon_D$,
  $Y = \beta_{U,Y}U + \beta_{D,Y}D + \epsilon_Y$.

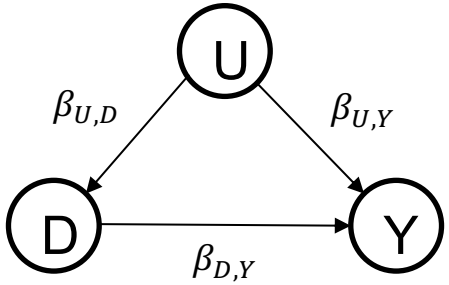
- If we e.g. force $U = u$ and $\mathrm{D} = d$, then $Y = \beta_{U,Y}u + \beta_{D,Y}d + \epsilon_Y$. This is a stronger assumption than your usual linear regression model.

- Note that $Cov(D, Y) = \beta_{D,Y} + \beta_{U,D}\beta_{U,Y} \neq \beta_{D,Y}$, hence the cliché "correlation ≠ causation".

Instrumental Variables (IV)

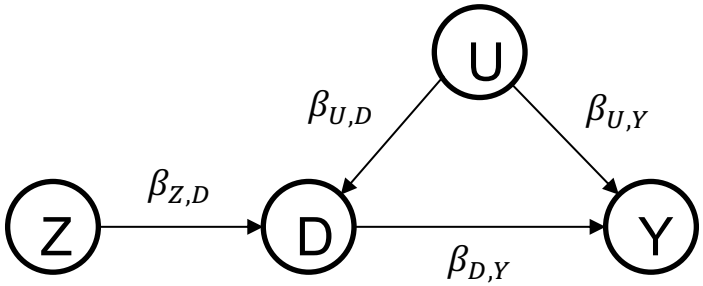
- Suppose we wish to estimate $\beta_{D,Y}$, the causal effect of D on Y, but that we cannot observe U, which affects both D and Y (an *unobserved confounder*).

- Without additional assumptions we cannot estimate the causal effect $\beta_{D,Y}$.



Instrumental variables (IV)

- Assume we can observe a variable Z that *only* causes D:



- Then $Cov(Z,Y) = \beta_{Z,D}\,\beta_{D,Y}$ and $Cov(Z,D) = \beta_{Z,D}$. Hence,

$$\beta_{D,Y} = \frac{Cov(Z,Y)}{Cov(Z,D)},$$

which we can estimate! We use Z as an *instrument* for D.

Motivating scientific question for the article

- Say we want to investigate the *causal* effect of military service ($\mathrm{D} = 1$) on future earnings Y (Objective of Angrist (1990))
- There may be confounders U that affect both military service D and future earnings Y.
- We consider the Vietnam war. Why?
- There was a randomized draft, which we may use as an instrument

Setting

- We consider the Rubin counterfactual framework (not SEM)
- Assume there exist two parallel universes:

$Y_i(0)$,
the future earnings of an individual when he is not sent to war



$Y_i(1)$,
the future earnings of an individual when he is sent to war



Setting

- The causal effect of war participation on earnings is defined by
$$\beta = E\,(Y_i(1) - Y_i(0))$$
- We cannot observe $Y_i(0)$ and $Y_i(1)$ simultaneously
- We only observe $Y_i = Y_i(D_i)$, where D_i is the military service indicator of subject i .
- Importantly, the variables $Y_i(0)$ and $Y_i(1)$ may correlate with D_i, and it can be shown that $E(Y_i \mid D_i = 1) = E\big(Y_i(1)\big)$ is **not** true in general, hence

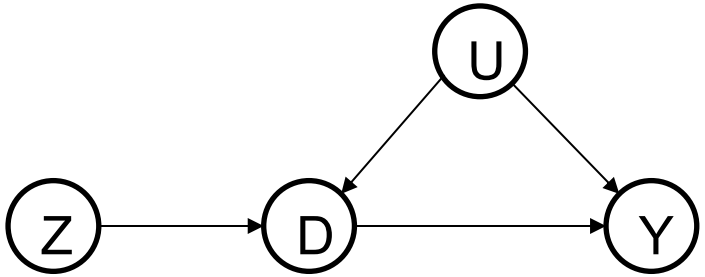
$$\beta \neq E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0)$$



The Instrumental Variable

- Let Z_i denote the indicator function of whether subject i is drafted
- Z_i is randomly assigned and should affect D_i (military participation) but should be independent of confounders $Y_i(0)$, and $Y_i(1)$ due to randomization.

- Informal illustration:



- This is reminiscent of an IV in the SEM framework

Formal model

- Counterfactual for war participation: $D_i(z)$ for $z = 0,1$ indicating draft or no draft. We observe $D_i(Z_i) = D_i$.
- Counterfactual for earnings: $Y_i(z, d)$ for $z, d \in \{0,1\}$. We observe $Y_i = Y_i(Z_i, D_i)$.
- We implicitly assume SUTVA (assumption 1), i.e., D_i only depends on Z_i and not Z_j for $j \neq i$, and Y_i only depends on Z_i and D_i.
- The counterfactuals refer to “parallel universes” as before

Assumptions

- A2: Assignment of $(Z_1, \ldots, Z_n)$ is randomized
- A3: For all i, $Y_i(z, d) = Y_i(z', d)$ for all $z, z', d \in \{0,1\}$. That is, being drafted ($Z_i = 1$) only affects future earnings Y_i through military participation D_i ("exclusion restriction").
- A3 implies that we can write drop the functional dependence of $Y_i(z, d)$ on d, and write $Y_i(d)$.

Assumptions

- A4: For all i, $E\left(D_i(1) - D_i(0)\right) \neq 0$. That is, being drafted has a *causal effect on* war participation.

- A5: For all i, $D_i(1) \geq D_i(0)$ (monotonicity). That is, no person will participate in war if not drafted, but not participate if drafted.

The main result

- **Proposition 1**. Under the given assumptions,

$$\frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(D_i|Z_i = 1) - E(D_i|Z_i = 0)} = E(Y_i(1) - Y_i(0)|D_i(1) = 1, D_i(0) = 0).$$

- The RHS is the causal effect of war participation on earnings for the *sub-population* of individuals who are "compliers", i.e., the ones that serve in war if drafted, but not otherwise. Why only this sub-population? There may be individuals who do not serve even when drafted, and ones that serve regardless of being drafted or not.

- The LHS is an *estimable* quantity of observable variables!

Proof.

- Step 1: The numerator.

$$
\begin{aligned}
& E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0) \\
&= E(Y_i(D_i)|Z_i = 1) - E(Y_i(D_i)|Z_i = 0) && (1) \\
&= E(Y_i(D_i(Z_i))|Z_i = 1) - E\big(Y_i\big(D_i(Z_i)\big)\big|Z_i = 0\big) && (2) \\
&= E(Y_i(D_i(1))|Z_i = 1) - E\big(Y_i\big(D_i(0)\big)\big|Z_i = 0\big) \\
&= E\big[Y_i\big(D_i(1)\big) - Y_i\big(D_i(0)\big)\big]. && (3)
\end{aligned}
$$

- (1) is by definition of $Y_i = Y_i(D_i)$, (2) is by definition of $D_i = D_i(Z_i)$, and (3) is by the random assignment of Z_i.

Step 1 in the proof is based on the lecture notes of Zhiyang Jia in the course ECON4137 at the University of Oslo, spring 2020.

Proof.

- Step 1: continued. Now,

$$
\begin{aligned}
&E\left[Y_i\left(D_i(1)\right) - Y_i\left(D_i(0)\right)\right] \\
&= E\left[Y_i\left(D_i(1)\right) - Y_i\left(D_i(0)\right) \mid D_i(1) = 1, D_i(0) = 0\right] P(D_i(1) = 1, D_i(0) = 0) \\
&+ E\left[Y_i\left(D_i(1)\right) - Y_i\left(D_i(0)\right) \mid D_i(1) = 1, D_i(0) = 1\right] P(D_i(1) = 1, D_i(0) = 1) \\
&+ E\left[Y_i\left(D_i(1)\right) - Y_i\left(D_i(0)\right) \mid D_i(1) = 0, D_i(0) = 0\right] P(D_i(1) = 0, D_i(0) = 0) \\
&+ E\left[Y_i\left(D_i(1)\right) - Y_i\left(D_i(0)\right) \mid D_i(1) = 1, D_i(0) = 0\right] P(D_i(1) = 0, D_i(0) = 1)
\end{aligned}
$$

- Above we condition on subpopulations. Due to monotonicity, $P(D_i(1) = 0, D_i(0) = 1) = 0$. Since we condition on which values $D_i(z)$ takes, we get

$$
\begin{aligned}
&= E[Y_i(1) - Y_i(0) \mid D_i(1) = 1, D_i(0) = 0] P(D_i(1) = 1, D_i(0) = 0) \\
&+ E[Y_i(1) - Y_i(1) \mid D_i(1) = 1, D_i(0) = 1] P(D_i(1) = 1, D_i(0) = 1) \\
&+ E[Y_i(0) - Y_i(0) \mid D_i(1) = 0, D_i(0) = 0] P(D_i(1) = 0, D_i(0) = 0) \\
&= E[Y_i(1) - Y_i(0) \mid D_i(1) = 1, D_i(0) = 0] P(D_i(1) = 1, D_i(0) = 0)
\end{aligned}
$$

Proof.

- Step 2: The denominator. We get that

$$
\begin{aligned}
&\mathrm{E}(D_i|Z_i = 1) - E(D_i|Z_i = 0) \\
&= E(D_i(1)|Z_i = 1) \; - E(D_i(0)|Z_i = 0) && (1)\\
&= E\big(D_i(1) - D_i(0)\big) && (2)\\
&= 1 \; * P(D_i(1) = 1, D_i(0) = 0) - 1 * P(D_i(1) = 0, D_i(0) = 1) \\
&= P(D_i(1) = 1, D_i(0) = 0),
\end{aligned}
$$

where (1) is by definition of $D_i = D_i(Z_i)$ and (2) is due to random assignment of Z_i.

Proof.

- Step 3: Combining step 1 and 2 we obtain

$$\frac{E(Y_i|Z_i = 1) - E(Y_i|Z_i = 0)}{E(D_i|Z_i = 1) - E(D_i|Z_i = 0)}$$

$$= \frac{E[Y_i(1) - Y_i(0) \mid D_i(1) = 1, D_i(0) = 0]P(D_i(1) = 1, D_i(0) = 0)}{P(D_i(1) = 1, D_i(0) = 0)}$$

$$= E[Y_i(1) - Y_i(0) \mid D_i(1) = 1, D_i(0) = 0],$$

as desired.

Results from the study (Angrist 1990)

- $\hat{\beta} = \frac{-435.8}{0.159} = -2741$ (dollars), with SE 1324 for men born in 1950, income year 1981.

- Interpretation: The estimated average causal effect of Vietnam war participation on earnings in 1981 for compliers born 1950 is a reduction of 2741$ per year

Concluding remarks

- The paper shows how causal inferences can be made in an observational study, with transparent assumptions
- Predictive power is not the main objective. We are in category “super A” since we need strong assumptions
- The economists have pioneered in many areas of Causal Inference. Maybe statisticians have something to learn from Economists?

Aftenposten A-magasinet Oslo Sport Meninger Søk Meny

Debatt | Forskning

Nei, det er ikke nødvendigvis en sammenheng mellom parker og dødsfall

Fem økonomer

21. okt. 2021 20:00 | Sist oppdatert i dag 12:52

Ikke tilfeldig hvem som bor hvor

Forskerne sammenligner dødelighet i områder med ulik nærhet til grøntområder, men det er ikke tilfeldig hvem som bor hvor. Det er godt kjent at det som påvirker bosted, også påvirker helse. Det gjelder blant annet utdanning, inntekt, yrke, miljø, trangboddhet, og arbeidsledighet. Å anlegge park vil ikke nødvendigvis endre de underliggende årsakene til helseforskjeller.

Nobelprisen i økonomi, Sveriges Riksbank minnepris, ble i oktober tildelt økonomene David Card, Joshua D. Angrist og Guido W. Imbens. Det er en påminnelse om hvordan man kan måle kausale effekter når kun observasjonelle data er tilgjengelig.

Innlegget er signert av følgende:

- *Sigurd Arntzen, samfunnsøkonom, Folkehelseinstituttet*
- *Hege Gjefsen, samfunnsøkonom PhD., Folkehelseinstituttet*
- *Mari Grøsland, samfunnsøkonom, Folkehelseinstituttet*
- *Ingrid Hjort, samfunnsøkonom PhD., Folkehelseinstituttet*
- *Jonas Minet Kinge, helseøkonom PhD., Folkehelseinstituttet*

Nils sin niese (?)!

https://www.aftenposten.no/meninger/debatt/i/8QR13r/nei-det-er-ikke-noedvendigvis-en-sammenheng-mellom-parker-og-doedsfall

References

- Joshua D. Angrist. “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Se- curity Administrative Records”. In: *The American Economic Review* 80.3 (1990). Publisher: American Economic Association, pp. 313–336. issn: 0002-8282. url: https://www.jstor.org/stable/2006669 (visited on 10/19/2021).

- Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. “Identification of Causal Effects Using In- strumental Variables”. In: *Journal of the American Statistical Association* 91.434 (June 1996). Publisher: Taylor&Francis eprint:https://www.tandfonline.com/doi/pdf/10.1080/01621459.1996.10476902,pp.444–455. issn: 0162-1459. doi: 10.1080/01621459.1996.10476902. url: https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476902 (visited on 10/19/2021).

- Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press, 2015. isbn: 978-0-521-88588-1. doi: 10.1017/CBO9781139025751. url: https://www.cambridge.org/core/books/causal-inference-for- statistics-social-and-biomedical-sciences/71126BE90C58F1A431FE9B2DD07938AB (visited on 10/19/2021).

- Judea Pearl. *Causality*. 2nd ed. Cambridge: Cambridge University Press, 2009. isbn: 978-0-521-89560-6. doi: 10.1017/CBO9780511803161. url: https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B (visited on 10/19/2021).