Instrument Variabel Estimation (IV)

Økonometri A

Bertel Schjerning

Program

IV estimation: SLR (W15.1 + SLP 1-2)

- Endogene variable
- Simpel IV estimation
- Konsistens og inferens

IV estimation: MLR (W15.2-W15.4 + SLP 3)

- Strukturel model ligevægtsmodel
- Two Stage Least Squares (2SLS)

Test for eksogeneitet

Test for overidentifikation (W15.5)

Motivation

Vi kan estimere parametrene i en regressionsmodel med OLS:

$$y = \beta_0 + \beta_1 x + u$$

Men OLS giver ikke altid en kausal fortolkning.

- Kausalitet kræver, at MLR.4 er opfyldt: $E(u|x) = 0$.
- Uden MLR.4 måler vi blot korrelation mellem x og y.
- Korrelation er ofte utilstrækkeligt for policy-analyser.

Overvej dette eksempel – hvad er den kausale fortolkning, og er der problemer?

- Børn med husdyr har mindre allergi.

Motivation

Hvornår kan to variable være (u)korrelerede, uden at man kan tolke $\hat{\beta}_1$ som en kausal effekt af x på y?

- Udeladte variable.
- Målefejl i forklarende variable.
- Omvendt kausalitet.
- Simultanitet

Eksempel 1: Udeladte variable

Eksempel: Lønregression

$$\log(\text{timeløn}) = \beta_0 + \beta_1 \text{uddannelse} + \beta_2 \text{evner} + u,$$

- Hvis vi udelader evner fra regressionen, vil MLR.4 ikke være opfyldt, hvis evner er korreleret med uddannelse.
- Hvis evner er positivt korreleret med både uddannelse og løn ($\beta_2 > 0$), vil OLS estimatet overvurdere den kausale effekt af uddannelse på løn.

$$plim(\hat{\beta}_1^{OLS}) = \beta_1 + \beta_2 \frac{cov(uddannelse, evner)}{var(uddannelse)} > \beta_1$$

Eksempel 2: Målefejl i forklarende variable

Klassisk målefejl i x: $x = x^* + e$, hvor $cov(x^*, e) = 0$.

- OLS estimatoren:

$$p\lim \hat{\beta}_1 = \beta_1 \frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_e}$$

- OLS estimatoren vil ikke give en kausal effekt.
- Målefejl får x og y til at fremstå ukorrelerede, selvom der er en kausal sammenhæng.

Eksempel 3: Omvendt kausalitet

Betragt følgende model for mors og datters højde:

$$y_{mor} = \beta_0 + \beta_1 x_{datter} + u,$$

- y_{mor} er mors højde.
- x_{datter} er datters højde.
- Hvis $\hat{\beta}_1 > 0$, kan vi så slutte, at datterens højde påvirker morens højde?

Omvendt kausalitet: Hvis u er stort, vil $E(x_{datter}|u)$ også være større, fordi højde delvist er arveligt.

- MLR.4 ikke overholdt.

Eksempel 4: Simultanitet

Efterspørgsels- og udbudsmodel for markedsandele og priser:

$$\log\left(\frac{S_{jm}}{S_{0m}}\right) = \beta \mathbf{x}_{jm} - \alpha p_{jm} + \xi_{jm}, \quad \text{(Efterspørgsel)}$$

$$p_{jm} = MC_{jm} + \frac{S_{jm}}{\partial S_{jm}/\partial p_{jm}}, \quad \text{(Udbud)}$$

- S_{jm}: Markedsandel for produkt j.
- p_{jm}: Pris på produkt j.
- ξ_{jm}: Uobserverbare produktkarakteristika.
- MC_{jm}: Marginalomkostning for produkt j.

Endogenitet: Priser fastsættes, hvor udbud møder efterspørgsel:

- Priserne afhænger af uobserverbare produktkarakteristika (ξ_{jm}).
- Dette bryder MLR.4.

IV estimation: SLR

Endogene vs eksogene variable

Indtil nu: **Eksogene** forklarende variable (hvis MLR.4 er opfyldt).

Simpel lineær regresionsmodel:

$$y = \beta_0 + \beta_1 x + u$$

MLR.4: $E(u|x) = 0 \Rightarrow \text{Cov}(u, x) = 0$

Når MLR.1-MLR.4 er opfyldt, er OLS middelret og konsistent.

Nu: **Endogene** forklarende variable.

- Variable hvor $\text{Cov}(u, x) \neq 0 \rightarrow$ MLR.4 ikke opfyldt.
- OLS er ikke længere middelret og ikke konsistent.

Endogene vs eksogene variable

Kan vi estimere kausale sammenhænge med endogene variable?

Eksempel: Lønregression

$$\log(\text{timeløn}) = \beta_0 + \beta_1 \text{uddannelse} + u,$$

hvor *uddannelse* er korreleret med *evner* i u.

Det ideelle eksperiment

- En stikprøve, hvor *uddannelse* er ukorreleret med u.
- **Løsning:** Randomiseret tildeling af uddannelsespladser.
- Randomiserede eksperimenter er guldstandarden, men ofte umulige, uetiske eller dyre i samfundsvidenskab.

Endogene vs eksogene variable

Det næstbedste eksperiment

- Regeringen uddeler uddannelsesstipendier tilfældigt, fx ved lodtrækning om ekstra høj SU.
- MLR.4 er stadig ikke opfyldt. Uddannelse er stadig endogen, så OLS er ikke middelret eller konsistent.

Men nu er noget af variationen i uddannelse tilfældig:

- Højere SU øger sandsynligheden for uddannelse.
- Højere SU er ukorreleret med evner (via lodtrækning).

Kan vi isolere den eksogene variation fra den endogene?

Instrument variabel estimation

Antag vi har en simpel lineær regressionsmodel:

$$y = \beta_0 + \beta_1 x + u,$$

hvor x er en **endogen** variabel: $Cov(u, x) \neq 0$.

Vi har en yderligere variabel z, hvor:

$$Cov(x, z) \neq 0$$
$$Cov(u, z) = 0$$

Vi kalder z for en **instrument vaiable (IV)** for x.

- At finde gyldige instrumenter kan være svært.
- Økonomisk teori bør guide valget af instrumenter. (alternativet er en vildfaren stokastisk klovnebus)

Instrument variabel estimation

Vi kan bruge z til at estimere β'erne i modellen:

$$y = \beta_0 + \beta_1 x + u.$$

Udledning af IV estimatoren:

$$\begin{aligned} cov(u, z) &= 0 \quad \text{(Instrument er eksogent)} \\ cov(y - \beta_1 x, z) &= 0 \\ cov(y, z) - \beta_1 cov(x, z) &= 0 \\ \beta_1 &= \frac{cov(y, z)}{cov(x, z)}. \end{aligned}$$

Erstat populationsmomenter med datamoment (**IV-estimator**):

$$\hat{\beta}_1^{IV} = \frac{\sum z_i y_i}{\sum z_i x_i}.$$

Motivation for IV: Strukturel model og reduceret form

Strukturel model:

$$y = \beta_0 + \beta_1 x + u,$$
$$x = \gamma_0 + \gamma_1 z + \nu$$
$$cov(u, z) = 0, \quad cov(\nu, z) = 0, \quad cov(u, \nu) \neq 0$$

Reduceret form for y:

$$y = \beta_0 + \beta_1(\gamma_0 + \gamma_1 z + \nu) + u,$$
$$y = (\beta_0 + \beta_1\gamma_0) + \beta_1\gamma_1 z + \beta_1\nu + u$$
$$y = b_0 + b_1 z + \xi \quad \text{(Reduceret form for } y\text{)}$$

hvor $b_0 = \beta_0 + \beta_1\gamma_0$, $b_1 = \beta_1\gamma_1$, og $\xi = \beta_1\nu + u$.

Kan vi estimere de strukturelle parametre med reduceret form?

Identifikations problem

Identifikations problem:

Vi kan ikke bestemme β_1 ud fra reducerede form, da $b_1 = \beta_1\gamma_1$.

Estimation strategi:

1. Estimer γ_0 og γ_1.

$$x = \gamma_0 + \gamma_1 z + \nu \quad \text{(first stage regression)}$$

2. Estimer b_0 og b_1

$$y = b_0 + b_1 z + \xi \quad \text{(reduced form for y)}$$

3. Udnyt sammenhængen mellem reduceret form og strukturelle parametre

$$\beta_1 = \frac{b_1}{\gamma_1}, \quad \beta_0 = b_0 - \beta_1\gamma_0$$

Instrument variabel estimation: Intuition

Dette giver netop IV estimatoren:

$$\hat{\beta}_1^{IV} = \frac{\hat{b}_1}{\hat{\gamma}_1} = \frac{\frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2}}{\frac{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2}} = \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})}$$

hvor $\hat{\gamma}_1$ og $\hat{b}_1$ er OLS estimaterne af hhv.

$$y = b_0 + b_1 z + \nu \quad \text{(Reduceret form)}$$
$$x = \gamma_0 + \gamma_1 z + \xi \quad \text{(First stage)}$$

Bemærk: z er eksogen i begge ligninger. Derfor opnår vi en konsistent estimator for β_1, som effekten af z på y relativt til effekten af z på x

Instrument variabel estimation: Python eksempel

Effekten af uddannelse for gifte amerikanske kvinder

Model:

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

- Data fra Mroz
- Endogen variabel: Uddannelse (educ)
- Instrument: Mors og fars uddannelse
- Tre IV estimationer:
  1. Mors uddannelse som instrument
  2. Fars uddannelse som instrument
  3. Både mors og fars uddannelse som instrument (mere om det senere).

Let's Code!

- **Jupyter Notebook:** `11_iv_examples.ipynb`
- **Part 1:** IV estimation med mors uddannelse som instrument

Instrument variabel estimation: Intuition

Eksemplet med uddannelseslotteriet:

$$y = \beta_0 + \beta_1 x + u, \tag{1}$$

$$x = \gamma_0 + \gamma_1 z + \xi \tag{2}$$

$$y = b_0 + b_1 z + \nu \tag{3}$$

hvor y fx er lønnen, x er års uddannelse og z mængden af SU.

- Ligning (1): højere uddannelse og fast $u \Rightarrow$ lønnen stiger med β_1
- Ligning (2): højere SU $\Rightarrow$ antal års uddannelse stiger med γ_1
- Ligning (3): højere SU $\Rightarrow$ lønnen stiger med b_1.

Hvis højere SU (lotteriet) kun påvirker folks løn gennem antal års uddannelse ($cov(z, u) = 0$), så må effekten af et års uddannelse være $\beta_1 = b_1/\gamma_1$.

Instrument variabel estimation: Gyldige instrumenter

Der er to betinget for at z er et gyldigt instrument:

Betingelse 1: $cov(x, z) \neq 0$ (relevans)

- Korreleret med den endogene forklarende variabel x.
- Denne antagelse kan testes: Er z korreleret med x?

Betingelse 2: $cov(u, z) = 0$ (eksogenitet)

- Ukorreleret med fejllleddet u og hermed ukorreleret med de uobserverbare faktorer.
- Denne antagelse kan ikke testes når man kun har et instument. To overvejelser:
  - Er z "så godt som tilfældigt" betinget på de eksogene variable?
  - Påvirker z kun y gennem den endogene variabel (x)?

Instrument variabel estimation: Forskningseksempler

Contents lists available at ScienceDirect

Labour Economics

journal homepage: www.elsevier.com/locate/labeco

Does peacetime military service affect crime? New evidence from Denmark's conscription lotteries

Stéphanie Vincent Lyk-Jensen

VIVE- Danish Centre of Applied Social Science, Herluf Trolles Gade 11, Copenhagen 1052, Denmark

ARTICLE INFO

JEL classification:
K42
H56

Keywords:
Crime
Military service
Draft lottery
Criminal behavior

ABSTRACT

While military service is thought to promote civic values, evidence on its benefits on criminal behavior is mixed. This paper uses the Danish draft lottery to estimate the causal effect of peacetime military service on post-service criminal convictions. The data includes the entire universe of eligible men born 1976–1983. I find that military service does not affect crime in general or any kind of crime in particular, nor does it reduce crime for juvenile offenders. However, I find a temporary disruption in the educational path at age 25, but no impact on the likelihood of being unemployed.

$y = \text{Crime}$, $x = \text{Military}$, $z = \text{Draft lottery number}$

$cov(u, z) = 0$ (lottery), $cov(x, z) \neq 0$ (lottery $\rightarrow$ military)

Instrument variabel estimation: Forskningseksempler

THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. CVI November 1991 Issue 4

DOES COMPULSORY SCHOOL ATTENDANCE AFFECT SCHOOLING AND EARNINGS?*

JOSHUA D. ANGRIST AND ALAN B. KRUEGER

We establish that season of birth is related to educational attainment because of school start age policy and compulsory school attendance laws. Individuals born in the beginning of the year start school at an older age, and can therefore drop out after completing less schooling than individuals born near the end of the year. Roughly 25 percent of potential dropouts remain in school because of compulsory schooling laws. We estimate the impact of compulsory schooling on earnings by using quarter of birth as an instrument for education. The instrumental variables estimate of the return to education is close to the ordinary least squares estimate, suggesting that there is little bias in conventional estimates.

$y =$ løn, $x =$ Udd., $z =$ fødselskvartal

$cov(u, z) = 0$ (z er ret tilfældig), $cov(x, z) \approx 0$ (svagt instrument)

Mulige instrumenter for uddannelse

Quiz

Vurder gyldigheden af følgende instrumenter for uddannelse.

Dvs. om $Cov(u, z) = 0$ og $Cov(udd, z) \neq 0$

1. "Næst-sidste cifre i cpr- nr."
2. "Mors uddannelse"
3. "IQ score"
4. "Afstanden til nærmeste universitet"
5. "Reform af uddannelsessystemet"

Mulige instrumenter for uddannelse

Opfølgining på quiz

Vurder gyldigheden af følgende instrumenter for uddannelse.

Dvs. om $Cov(u, z) = 0$ og $Cov(udd, z) \neq 0$

Instrument	$Cov(u, z)$	$Cov(udd, z)$
"Næst-sidste cifre i cpr- nr."	0	≈ 0
"Mors uddannelse"	$\neq 0$	$\neq 0$
"IQ score"	$\neq 0$	$\neq 0$
"Afstanden til nærmeste universitet"	$\neq 0$	$\neq 0$
"Reform af uddannelsessystemet"	0	$\neq 0$

Egenskaber ved IV estimatoren

Betingelse 1: $cov(x, z) \neq 0$ (relevans)
Betingelse 2: $cov(u, z) = 0$ (eksogenitet)

Under disse betingelser er IV estimatoren:

- Konsistent
- asymptotisk normalfordelt.

IV estimatoren er ikke middelret
(kan være problematisk i små stikprøver):

- Bias øges ved svage instrumenter ($Cov(x, z) \approx 0$).
- Har større varians end OLS.

Konsistens af IV estimatoren: Bevis

$$plim\left(\hat{\beta}_1^{IV}\right) = \frac{plim\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z})\right)}{plim\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})\right)} = \frac{cov(y,z)}{cov(x,z)}$$

Indsæt $y = \beta_0 + \beta_1 x + u$

$$\begin{aligned} plim\left(\hat{\beta}_1^{IV}\right) &= \frac{cov(\beta_0 + \beta_1 x + u, z)}{cov(x,z)} \\ &= \beta_1 \frac{cov(x,z)}{cov(x,z)} + \frac{cov(u,z)}{cov(x,z)} \\ &= \beta_1 \end{aligned}$$

Bemærk: $cov(u,z) = 0$ og $cov(x,z) \neq 0$

Asymptotisk bias: IV vs OLS

Asymptotisk bias:

$$plim\left(\hat{\beta}_1^{IV}\right) - \beta_1 = \frac{cov(u,z)}{cov(x,z)}$$

$$plim\left(\hat{\beta}_1^{OLS}\right) - \beta_1 = \frac{cov(u,x)}{var(x)}$$

- Typisk: $cov(x,z) < var(x)$.
- Hvis $cov(u,z) \neq 0$ og $cov(u,x) \neq 0$, kan det være, at:

$$\text{IV bias} = \left|\frac{cov(u,z)}{cov(x,z)}\right| > \left|\frac{cov(u,x)}{var(x)}\right| = \text{OLS bias}$$

Konsekvens: Hvis eksogenitetsantagelsen ikke holder, kan IV være mere biased end OLS.

Svage instrumenter

Problemer ved svage instrumenter ($cov(x, z) \approx 0$):

- Når $cov(x, z) \approx 0$, er instrumentet næsten irrelevant, og IV estimatoren bliver meget upålidelig.
- Selv hvis $cov(u, z) = 0$, vil der være korrelation mellem u og z i små stikprøver.
- Denne korrelation bliver forstærket, når $cov(x, z) \approx 0$, hvilket øger bias i IV estimatoren.
- Svage instrumenter øger også variansen af IV estimatoren, hvilket gør den mindre præcis end OLS.

Konsekvens: Med svage instrumenter kan IV estimatoren være mere biased og upålidelig end OLS, selvom $cov(u, z) = 0$ i teorien.

Inferens

Under antagelse af homoskedasticitet ($var(u|z) = \sigma^2$), gælder

$$Avar(\hat{\beta}_1^{IV}) = \frac{\sigma_u^2}{n\sigma_x^2\rho_{x,z}^2} > \frac{\sigma_u^2}{n\sigma_x^2} = Avar(\hat{\beta}_1^{OLS}) \tag{4}$$

hvor $0 < \rho_{x,z} < 1$ er korrelationen mellem x og z.

Vi kan estimere den asymptotiske varians konsistent som

$$\widehat{Avar}(\hat{\beta}_1^{IV}) = \frac{\frac{1}{n-2}\sum \hat{u}_i^2}{R_{x,z}^2 SST_x} = \frac{\hat{\sigma}_u}{nR_{x,z}^2\hat{\sigma}_x^2} \tag{5}$$

hvor $R_{x,z}^2$ er R^2 fra en regression af x på z.

I øvrigt gælder at t-test størrelserne er asymptotisk normalfordelte.

Hvorfor er IV estimatoren ikke middelret?

Husk trick da vi beviste at OLS var middelret.

$$E(\hat{\beta}_1^{OLS}|x) = \beta_1 + E\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(u_i)(x_i - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}|x\right) \tag{6}$$

MLR.4 siger at $E(u|x) = 0$ og ved at betinge på x er alt andet end u i parentesen ikke-stokastisk og kan trækkes uden for.

Med et eksogent instrument, gælder $E(u|z) = 0 \Rightarrow cov(u, z) = 0$.

- Men at betinge på z er ikke nok til at gøre alt andet end u i $E(\hat{\beta}_1^{IV}|z)$ ikke-stokastisk.
- For at få det, skal vi betinge på både z og x.
- Men $E(u|x, z) \neq 0$ (generelt).

Let's Code!

- **Jupyter Notebook:** `11_iv_examples.ipynb`
- **Part 2:** Konsistens, bias, asymptotisk fordeling for IV estimatoren

IV som løsning på målefejl: Eksempel

Målefejl i forklarende variable giver problemer for OLS estimatoren. IV estimation kan være løsningen

Model

$$y = \beta_0 + \beta_1 x^* + u$$

hvor x^* er målt med klassisk målefejl ($Cov(x^*, e) = 0$)

$$x = x^* + e$$

Regressionsmodel med målefejl:

$$y = \beta_0 + \beta_1 x + u - \beta_1 e$$

MLR.4: ikke opfyldt, da $cov(u - \beta_1 e|x) = cov(u - \beta_1 e|x^* + e) \neq 0$

IV som løsning på målefejl: Eksempel

Forslag til instrument: et andet mål z for x^* med egen målefejl:

$$z = x^* + \nu$$

Hvis målefejlene e og ν er uafhængige:

Relevant instrument: z korreleret med x^*

Validt instrument: ukorreleret med målefejlene, e og ν.

$\rightarrow$ IV giver et konsistent estimat af β_1.

Eksempel: Målefejl i selvrapporteret uddannelse

- Instrument: Brug kollega eller familiemedlem som reference.[1]

[1] Ashenfelter og Krueger (1994): "Estimates of the economic return to schooling."

Lad os se om vi kan!

- **Jupyter Notebook:** `11_iv_examples.ipynb`
- **Part 3:** IV som løsning på målefejl

IV estimation: MLR

Motivation: Strukturel ligevægtsmodel

To strukturelle relationer (fx efterspørgsels (D) og udbud (S)):

$$\begin{aligned} \text{(D)} \quad & \mathbf{y}^D = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \mathbf{u}, \\ \text{(S)} \quad & \mathbf{y}^S = \mathbf{X_1}\boldsymbol{\delta_1} + \mathbf{X_2}\boldsymbol{\delta_2} + \mathbf{Z_e}\boldsymbol{\delta_3} + \mathbf{v}. \end{aligned}$$

Ligevægt: $\mathbf{y} = \mathbf{y}^D = \mathbf{y}^S$:
$\mathbf{X_2}$ bestemmes endogent som de variable, der bringer (D) og (S) i ligevægt (fx priser eller andre markedsbestemte størrelser).

Dimensioner: $\mathbf{y} \in \mathbb{R}^{n\times 1}$, $\mathbf{X_1} \in \mathbb{R}^{n\times(k-l)}$, $\mathbf{X_2} \in \mathbb{R}^{n\times l}$, $\mathbf{Z_e} \in \mathbb{R}^{n\times g}$.

Endogenitet: Da $\mathbf{X_2}$ bestemmes af uobserverede efterspørgsels- og udbudsstød $(\mathbf{u}, \mathbf{v})$, vil $\mathbf{X_2}$ generelt være korreleret med fejlleddet i (D).

Eksklusiv restriktioner: $\mathbf{Z_e}$ påvirker kun (S)-siden direkte og fungerer som instrumenter.

Fra ligevægt til reduceret form form $\mathbf{X}_2$ ($l = 1$)

Strukturelle ligniner (fx efterspørgsels (D) og udbud (S)):

$$\begin{aligned} \text{(D)} \quad & \mathbf{y}^D = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \mathbf{u}, \\ \text{(S)} \quad & \mathbf{y}^S = \mathbf{X_1}\boldsymbol{\delta_1} + \mathbf{X_2}\boldsymbol{\delta_2} + \mathbf{Z_e}\boldsymbol{\delta_3} + \mathbf{v}. \end{aligned}$$

Ligevægt: Sæt (D) lig (S) og isolér $\mathbf{X_2}$:

$$\mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\beta_2 + \mathbf{u} = \mathbf{X_1}\boldsymbol{\delta_1} + \mathbf{X_2}\delta_2 + \mathbf{Z_e}\boldsymbol{\delta_3} + \mathbf{v}.$$

Med kun en endogen variabel ($l = 1$) kan vi let isolere $\mathbf{X_2}$ hvis $\beta_2 \neq \delta_2$ ($\beta_2 - \delta_2$ er en skalar)

$$\boxed{\mathbf{X_2} = \mathbf{X_1}\frac{\boldsymbol{\delta_1} - \boldsymbol{\beta_1}}{\beta_2 - \delta_2} + \mathbf{Z_e}\frac{\boldsymbol{\delta_3}}{(\beta_2 - \delta_2)} + (\mathbf{v} - \mathbf{u})\frac{1}{\beta_2 - \delta_2} \qquad (\star)}$$

$\mathbf{X_2}$ er endogen i ligevægtsmodellen

Reduceret form for $\mathbf{X_2}$ ($l = 1$):

$$\mathbf{X_2} = \mathbf{X_1}\frac{\boldsymbol{\delta_1} - \boldsymbol{\beta_1}}{\beta_2 - \delta_2} + \mathbf{Z_e}\frac{\boldsymbol{\delta_3}}{(\beta_2 - \delta_2)} + (\mathbf{v} - \mathbf{u})\frac{1}{\beta_2 - \delta_2}$$

$\mathbf{X_2}$ er endogen:

$$\text{Cov}(\mathbf{X_2}, \mathbf{u}) = \text{Cov}\left(\frac{\mathbf{v} - \mathbf{u}}{\beta_2 - \delta_2}, \mathbf{u}\right) = -\frac{\text{Var}(\mathbf{u})}{\beta_2 - \delta_2} + \frac{\text{Cov}(\mathbf{v}, \mathbf{u})}{\beta_2 - \delta_2} \neq 0.$$

Selv hvis $\mathbf{v} \perp \mathbf{u}$, fås $-\text{Var}(\mathbf{u})/(\beta_2 - \delta_2) \neq \mathbf{0}$.

$\Rightarrow$ $\mathbf{X_2}$ er *mekanisk* korreleret med $\mathbf{u}$ pga. ligevægten.

$\Rightarrow$ MLR.4 brydes

$\Rightarrow$ OLS på (D) er inkonsistent.

I ligevægt reagerer $\mathbf{X_2}$ på stød fra både efterspørgsel og udbud $(\mathbf{u}, \mathbf{v})$

$\Rightarrow$ $\mathbf{X_2}$ endogen i begge ligninger.

Reduceret form og endogenitet

Strukturelt system:

$$(D): \ \mathbf{y}^D = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \mathbf{u},$$
$$(S): \ \mathbf{y}^S = \mathbf{X_1}\boldsymbol{\delta_1} + \mathbf{X_2}\boldsymbol{\delta_2} + \mathbf{Z_e}\boldsymbol{\delta_3} + \mathbf{v}.$$

Reduceret form:

$$\mathbf{y} = \mathbf{X_1}\boldsymbol{\pi_1} + \mathbf{Z_e}\boldsymbol{\pi_2} + \mathbf{e_y},$$
$$\mathbf{X_2} = \mathbf{X_1}\boldsymbol{\Pi_1} + \mathbf{Z_e}\boldsymbol{\Pi_2} + \mathbf{e_x}.$$

- $\mathbf{e_y}$ og $\mathbf{e_x}$ er linearkombinationer af $\mathbf{u}$ og $\mathbf{v}$.
- Derfor $\text{Cov}(\mathbf{X_2}, \mathbf{u}) \neq 0 \rightarrow$ OLS inkonsistent i (D).
- Reduceret form identificerer kun $\boldsymbol{\pi}, \boldsymbol{\Pi}$
  (kke de strukturelle parametre.)
- Eksklusiv restriktion: $\mathbf{Z_e}$ påvirker kun $\mathbf{y}^S \Rightarrow$ Identifikation af $\boldsymbol{\beta_2}$
  (kræver $g \geq l$ og rang betingelse).

2SLS (1. trin): Estimation of reduceret form for $\mathbf{X}_2$

Strukturel ligning (outcome):

$$\mathbf{y} = \mathbf{X_1}\beta_1 + \mathbf{X_2}\beta_2 + \mathbf{u}, \quad \mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}) \in \mathbb{R}^{n \times k}.$$

Reduceret form for de endogene (1. trin / first stage):

$$\mathbf{X_2} = \mathbf{Z\Pi} + \mathbf{e}_x, \quad \mathbf{Z} = (\mathbf{X_1}, \mathbf{Z_e}) \in \mathbb{R}^{n \times h}, \; E[\mathbf{Z}'\mathbf{e}_x] = \mathbf{0}.$$

OLS-estimat for 1. trin:

$$\hat{\mathbf{\Pi}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X_2} \Rightarrow \hat{\mathbf{X}}_2 = \mathbf{Z}\hat{\mathbf{\Pi}} = \mathbf{P_Z}\mathbf{X_2},$$

hvor $\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ (projektion).

2SLS (2. trin): Estimation af strukturel ligning

Substituér $\hat{\mathbf{X}}_2$ i den strukturelle ligning:

$$\mathbf{y} = \mathbf{X_1}\beta_1 + \hat{\mathbf{X}}_2\beta_2 + \underbrace{\left(\mathbf{u} + (\mathbf{X_2} - \hat{\mathbf{X}}_2)\beta_2\right)}_{\omega}.$$

Orthogonalitet:

$$\mathbf{Z}'\omega = \underbrace{\mathbf{Z}'\mathbf{u}}_{=\mathbf{0}} + \underbrace{\mathbf{Z}'(\mathbf{I} - \mathbf{P_Z})}_{=\mathbf{0}}\mathbf{X}_2\beta_2 = \mathbf{0}$$

2SLS-estimeringsligning - OLS på $\mathbf{y}$ og $\hat{\mathbf{X}} = (\mathbf{X_1}, \hat{\mathbf{X}}_2)$:

$$\boxed{\hat{\beta}^{2SLS} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} = \left(\mathbf{X}'\mathbf{P_Z}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{P_Z}\mathbf{y}}$$

Bemærk: $\hat{\mathbf{X}} = \mathbf{P_Z}\mathbf{X}$ og $\mathbf{P_Z} = \mathbf{P_Z}' = \mathbf{P_Z}^2$.

MLR med endogene variable - uden strukturel model

Model i matrixnotation (se SLP afsnit 3.2):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} = \mathbf{X_1}\beta_1 + \mathbf{X_2}\beta_2 + \mathbf{u}$$

hvor $\mathbf{X}$ opdeles i:

$$\underset{n\times k}{\mathbf{X}} = (\underset{n\times k-l}{\mathbf{X_1}}, \underset{n\times l}{\mathbf{X_2}}) = \begin{pmatrix} x_1, x_2, \ldots, x_{k-l}, x_{k-l+1}, \ldots, x_k \\ X_1{:}k-l \text{ eksogene} \qquad X_2{:}l \text{ endogene} \end{pmatrix}.$$

$k - l$ **eksogene** variable ($\mathbf{X_1}$) med parametre β_1
l **endogene** variable ($\mathbf{X_2}$) med parametre β_2.

For $\mathbf{X_1}$ og $\mathbf{X_2}$ gælder:

$$p\lim\left(\frac{1}{n}\mathbf{X_1}'\mathbf{u}\right) = 0, \quad p\lim\left(\frac{1}{n}\mathbf{X_2}'\mathbf{u}\right) \neq 0,$$

dvs. de endogene variable i $\mathbf{X_2}$ er korreleret med fejlleddet $\mathbf{u}$.

Eksklusive restriktioner og 2SLS

Eksklusiv restriktion: Nogle variable $\mathbf{Z_e}$ påvirker kun de endogene variable $\mathbf{X_2}$, ikke $\mathbf{y}$ direkte.

$$E[\mathbf{Z_e}'\mathbf{u}] = 0, \quad E[\mathbf{Z_e}'\mathbf{X_2}] \neq 0$$

Identifikation:

- Når $\mathbf{Z_e}$ findes, kan vi bruge variationen i $\mathbf{X_2}$, der drives af $\mathbf{Z_e}$, som **exogen variation**.
- $\mathbf{Z_e}$ fungerer dermed som instrumenter for $\mathbf{X_2}$.

Estimation: Dette leder direkte til **2SLS-estimatoren**:

$$\hat{\beta}^{2SLS} = (\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{P_Z}\mathbf{y})$$

hvor $\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$.

Multipel lineær regressionsmodel med endogene variable

For at kunne estimere modellen med IV, skal vi bruge mindst ét instrument for hver endogen variabel:

Eksakt identifikation: Antallet af instrumenter g er lig med antallet af endogene variable l:

$$\underset{n\times k}{\mathbf{Z}} = \left(\underset{n\times k-l}{\mathbf{X_1}}, \underset{n\times l}{\mathbf{Z_e}}\right) = \left(\underset{\mathbf{X_1}:k-l \text{ eksogene}}{x_1, x_2, \ldots, x_{k-l}}, \underset{\mathbf{Z_e}:l \text{ instrumenter}}{z_1, \ldots, z_l}\right)$$

- $\mathbf{X_1}$ (eksogene variable) fungerer som instrumenter for sig selv.
- $\mathbf{Z_e}$ (eksterne instrumenter) fungerer som $g = l$ instrumenter for de l endogene variable $\mathbf{X_2}$.
- $\mathbf{Z}$ indeholder alle eksogene variable, dvs. $h = k - l + l = k$.

Multipel lineær regressionsmodel med endogene variable

For at kunne estimere modellen med IV, skal vi bruge mindst ét instrument for hver endogen variabel:

Overidentifikation:

Flere instrumenter end endogene variable ($g > l$):

$$\underset{n\times(k-l+g)}{\mathbf{Z}} = \left(\underset{n\times(k-l)}{\mathbf{X_1}}, \underset{n\times g}{\mathbf{Z_e}}\right) = \left(\begin{array}{cc} x_1, x_2, \ldots, x_{k-l}, & z_1, \ldots, z_g \\ \mathbf{X_1}\text{:}k-l \text{ eksogene} & \mathbf{Z_e}\text{:}g \text{ instrumenter} \end{array}\right)$$

- De $k - l$ eksogene variable $\mathbf{X_1}$ er instrumenter for sig selv.
- De g eksterne instrumenter $\mathbf{Z_e}$ er instrumenter for de l endogene variable $\mathbf{X_2}$.
- $\mathbf{Z}$ indeholder alle eksogene variable, dvs. $h = k - l + g > k$.

Udledning af IV estimator

For at udlede IV-estimatoren antager vi følgende:

- **Eksogene instrumenter**:

$$p\lim\left(\frac{1}{n}Z'u\right) = \underset{(k-l+g)\times 1}{0}$$

- **Instrumenter korreleret med de endogene variable**:

$$p\lim\left(\frac{1}{n}Z'X\right) = \underset{(h\times k)}{\mathbf{\Sigma_{ZX}}} \text{ har fuld rang}$$

- **Ingen perfekt multikollinearitet mellem instrumenterne**:

$$p\lim\left(\frac{1}{n}Z'Z\right) = \underset{(h\times h)}{\mathbf{\Sigma_{ZZ}}} \text{ har fuld rang}$$

Udledning af IV estimator: Eksakt identificeret

Eksakt identifikation: $g = l$ instrumenter.

Vi erstatter de teoretiske momenter $p\lim\left(\frac{1}{n}\mathbf{Z}'\mathbf{u}\right) = 0$ med datamomenter:

$$\begin{aligned}
\tfrac{1}{n}\mathbf{Z}'\hat{\mathbf{u}} &= 0 \\
\tfrac{1}{n}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\beta}^{IV}) &= 0 \\
\Rightarrow \mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\hat{\beta}^{IV} &= 0 \\
\Rightarrow \mathbf{Z}'\mathbf{y} &= \mathbf{Z}'\mathbf{X}\hat{\beta}^{IV} \\
\Rightarrow \hat{\beta}^{IV} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}
\end{aligned}$$

Denne estimator er konsistent.

OBS: I tilfældet $g = l$ har $\mathbf{Z}'\mathbf{Z}$ altid fuld rang, når $\mathbf{Z}'\mathbf{X}$ har det.

Konsistens af IV estimator: Eksakt identificeret

Indsæt for $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ i IV estimatoren og reducer:

$$\begin{aligned}\hat{\beta}^{IV} &= (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{y}) \\ &= (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{X}\beta + \mathbf{Z}'\mathbf{u}) \\ &= \beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{u}\end{aligned}$$

Tag grænseværdien:

$$\begin{aligned}\Rightarrow \text{plim}(\hat{\beta}^{IV}) &= \beta + \text{plim}((\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{u}) \\ &= \beta + \text{plim}\left(\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{Z}'\mathbf{u}\right) \\ &= \beta + \left(\text{plim}\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)^{-1}\text{plim}\frac{1}{n}\mathbf{Z}'\mathbf{u} \\ &= \beta + \boldsymbol{\Sigma}_{\mathbf{ZX}}{}^{-1} \cdot 0 = \beta \quad (\hat{\beta}^{IV} \text{er konsistent for } \beta)\end{aligned}$$

IV ved overidentification: GMM estimatoren

Momentbetingelser: $E[\mathbf{Z}'\mathbf{u}] = 0$

Overidentifikation ($g > l$):

$\rightarrow$ Ikke muligt at sætte alle momentbetingelser til nul.

$\rightarrow$ Minimerer en vægtet sum af kvadrerede momentbetingelser.

$\rightarrow$ Generalized Method of Moments (GMM) estimatoren.

GMM estimatoren minimerer:

$$Q(\beta) = \left(\frac{1}{n}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta)\right)' \mathbf{W} \left(\frac{1}{n}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta)\right)$$

GMM estimatoren (på lukket form) opnås ved minimering:

$$\hat{\beta}^{GMM} = \left(\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{y}$$

hvor $\mathbf{W}$ er en positiv semidefinit vægtmatrix.

IV ved overidentifikation: Vægtningsmatrix

Optimal vægtningsmatrix:

- Mere vægt til momentbetingelser med lav varians (som ved GLS).
- Hansen (1982) viser, at den optimale vægtningsmatrix er den inverse af variansen af momentbetingelserne:

$$\mathbf{W} = \left(\text{plim} \frac{1}{n} \mathbf{Z}'\mathbf{u}\mathbf{u}'\mathbf{Z} \right)^{-1}$$

Specialtilfælde:

- Under homoskedasticitet er $\mathbf{W} = (\sigma^2 \mathbf{Z}'\mathbf{Z})^{-1}$
- Denne vægtning giver 2SLS estimatoren.

$$\hat{\beta}^{2SLS} = \left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

Two-Stage Least Squares (2SLS)

GMM med $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$ kaldes Two-Stage Least Squares (2SLS), da den opdeles i to trin:

Two-Stage Least Squares (2SLS)

- **Første trin:** Regressér $\mathbf{X_2}$ på $\mathbf{Z}$ med OLS for at få de predikterede værdier $\hat{\mathbf{X}}_2$.
- **Andet trin:** Regressér $\mathbf{y}$ på $\mathbf{X_1}$ og $\hat{\mathbf{X}}_2$ med OLS.

Nedenfor viser vi, 2-trins proceduren er ækvivalent med

$$\hat{\beta}^{2SLS} = \left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

Bemærk: Regresion af $\mathbf{X}_1$ på $\mathbf{Z}$ er overflødig, da $\mathbf{X}_1$ er inkluderet i $\mathbf{Z}$, og $\hat{\mathbf{X}}_1 = \mathbf{X}_1$.

Two-Stage Least Squares (2SLS): Trin 1

Trin 1: Regresser de k variable i $\mathbf{X}$ på de h instrumenter i $\mathbf{Z}$:

$$\underset{n\times k}{\mathbf{X}} = \underset{n\times h}{\mathbf{Z}} \cdot \underset{h\times k}{\mathbf{\Pi}} + \underset{n\times k}{\mathbf{E}}$$

Denne regression består af k first-stage regressioner kørt samtidigt.
OLS-estimatet er:

$$\underset{h\times k}{\hat{\mathbf{\Pi}}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

De prædikterede værdier er:

$$\underset{n\times k}{\hat{\mathbf{X}}} = \underset{n\times h}{\mathbf{Z}} \cdot \underset{h\times k}{\hat{\mathbf{\Pi}}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{P}_{\mathbf{Z}}\mathbf{X}$$

hvor $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ er en projektionsmatrix med egenskaberne $\mathbf{P}_{\mathbf{Z}} = \mathbf{P}_{\mathbf{Z}}{}'$ og $\mathbf{P}_{\mathbf{Z}}\mathbf{P}_{\mathbf{Z}} = \mathbf{P}_{\mathbf{Z}}$.

Two-Stage Least Squares (2SLS): Trin 2

Trin 2: Estimer følgende regressionsligning med OLS (second stage):

$$\mathbf{y} = \hat{\mathbf{X}}\beta + \mathbf{w}$$

2SLS estimatoren:

$$\begin{aligned}
\hat{\beta}^{2SLS} &= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\
&= (\mathbf{P_Z X})'\mathbf{P_Z X})^{-1}(\mathbf{P_Z X})'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{P_Z P_Z X})^{-1}(\mathbf{X}'\mathbf{P_Z y}) \\
&= (\mathbf{X}'\mathbf{P_Z X})^{-1}(\mathbf{X}'\mathbf{P_Z y}) \\
&= \left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}
\end{aligned}$$

Two Stage Least Squares (2SLS): Intuition

- Trin 1 opdeler variationen i x i to dele, som er hhv. ukorreleret og korreleret med fejlleddet.
- I Trin 2 anvendes kun den del af variationen, som er ukorreleret med fejlleddet.

Specialtilfælde med eksakt identifikation

Når $g = l$, har både $\mathbf{Z}$ og $\mathbf{X}$ dimension $n \times k$. Derfor er $\mathbf{Z}'\mathbf{X}$, $\mathbf{Z}'\mathbf{Z}$ og $\mathbf{X}'\mathbf{Z}$ alle kvadratiske $k \times k$ matricer.

Vi kan derfor bruge regnereglen $(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$.

$$\begin{aligned}
\hat{\beta}^{2SLS} &= (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}) \\
&= (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Z})(\mathbf{X}'\mathbf{Z})^{-1}(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}) \\
&= (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{y})
\end{aligned}$$

Hvilket er IV estimatoren i specialtilfældet med eksakt identifikation.

Two-Stage Least Squares (2SLS): Konsistens

Ingredienser til bevis for konsistens:

Antagelser:

1. $p\lim \frac{1}{n}\mathbf{Z}'\mathbf{u} = 0$
2. $p\lim \frac{1}{n}\mathbf{Z}'\mathbf{X} = \mathbf{\Sigma_{ZX}}$ har fuld rang.
3. $p\lim \frac{1}{n}\mathbf{Z}'\mathbf{Z} = \mathbf{\Sigma_{ZZ}}$ har fuld rang.

2SLS estimatoren:

$$\hat{\beta}^{2SLS} = (\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{P_Z}\mathbf{y})$$

Regneregler:

1. $p\lim(A_n B_n) = AB$ når $p\lim(A_n) = A$ og $p\lim(B_n) = B$.
2. $p\lim(A_n^{-1}) = A^{-1}$ når $p\lim(A_n) = A$.

Two-Stage Least Squares (2SLS): Konsistens

Trin 1: Udtryk estimatoren ved den sande parameter og et led som afhænger af fejlleddet.

2SLS estimatoren:

$$\begin{aligned}\hat{\beta}^{2SLS} &= (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{y}) \\ &= (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X}\beta) + (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{u}) \\ &= \beta + (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{u}\end{aligned}$$

Two-Stage Least Squares (2SLS): Konsistens

Trin 2: Beregn asymptotisk bias

$$p\lim(\hat{\beta}^{2SLS}) - \beta$$

$$= p\lim[(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{u}]$$

Husk at $\mathbf{P}_\mathbf{Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$

$$= p\lim[(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{u})]$$

$$= p\lim[\tfrac{1}{n}\mathbf{X}'\mathbf{Z}(\tfrac{1}{n}\mathbf{Z}'\mathbf{Z})^{-1}\tfrac{1}{n}\mathbf{Z}'\mathbf{X}]^{-1}\tfrac{1}{n}\mathbf{X}'\mathbf{Z}(\tfrac{1}{n}\mathbf{Z}'\mathbf{Z})^{-1}\tfrac{1}{n}\mathbf{Z}'\mathbf{u}$$

Nu har vi noget, hvor alle elementer har en kendt $p\lim$

$$= [(\mathbf{\Sigma}_{\mathbf{zx}}(\mathbf{\Sigma}_{\mathbf{zz}})^{-1}\mathbf{\Sigma}_{\mathbf{zx}})^{-1}\mathbf{\Sigma}_{\mathbf{zx}}(\mathbf{\Sigma}_{\mathbf{zz}})^{-1}0] = 0$$

Two-Stage Least Squares (2SLS): Inferens

Omskriv 2SLS estimatoren som i trin 1:

$$\hat{\beta}^{2SLS} = (\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{P_Z}\mathbf{y}) = \beta + (\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P_Z}\mathbf{u}$$

Beregn variansen af 2SLS estimatoren:

$$\begin{aligned}
\text{var}(\hat{\beta}^{2SLS}|\mathbf{Z},\mathbf{X}) &= \text{var}\left(\beta + (\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P_Z}\mathbf{u}|\mathbf{Z},\mathbf{X}\right) \\
&= 0 + \text{var}\left((\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P_Z}\mathbf{u}|\mathbf{Z},\mathbf{X}\right) \\
&\quad \text{Matrixregneregel for varians} \\
&= (\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P_Z}\text{var}(\mathbf{u}|\mathbf{Z},\mathbf{X})(\mathbf{P_Z}'\mathbf{X})(\mathbf{X}'\mathbf{P_Z}\mathbf{X})^{-1} \\
&\quad \text{Udnytter at } \mathbf{P_Z} = \mathbf{P_Z}' \text{ og } \mathbf{P_Z}\mathbf{P_Z} = \mathbf{P_Z} \\
&= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\text{var}(\mathbf{u}|\mathbf{Z},\mathbf{X})\hat{\mathbf{X}}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}
\end{aligned}$$

Two-Stage Least Squares (2SLS): Inferens

Opsummering:

$$\text{var}(\hat{\beta}^{2SLS}|\mathbf{Z},\mathbf{X}) = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\text{var}(\mathbf{u}|\mathbf{Z},\mathbf{X})\hat{\mathbf{X}}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$$

Vi kan estimere $\mathbf{\Sigma} = \frac{1}{n}\hat{\mathbf{X}}'\text{var}(\mathbf{u}|\mathbf{Z},\mathbf{X})\hat{\mathbf{X}}$ på samme måde som ved heteroskedasticitet i OLS:

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_i \hat{u}_i^2\hat{\mathbf{x}}_i'\hat{\mathbf{x}}_i$$

hvor $\hat{u}_i$ er det i'te element i $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}^{2SLS}$ (Bemærk $\mathbf{X}$ IKKE $\hat{\mathbf{X}}$)

Dvs. vores estimerede varians for $\hat{\beta}^{2SLS}$ bliver:

$$\widehat{\text{var}}(\hat{\beta}^{2SLS}|\mathbf{Z},\mathbf{X}) = n(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{\Sigma}}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$$

Two-Stage Least Squares (2SLS): Inferens

Med homoskedasticitet er variansen på 2SLS givet ved

$$\widehat{\text{var}}(\hat{\beta}^{2SLS}|\mathbf{Z},\mathbf{X}) = \hat{\sigma}^2(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$$

hvor

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_i \hat{u}_i^2$$

Med heteroskedasticitet:

$$\widehat{\text{var}}(\hat{\beta}^{2SLS}|\mathbf{Z},\mathbf{X}) = n(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{\Sigma}}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$$

hvor

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_i \hat{u}_i^2\hat{\mathbf{x}}_i'\hat{\mathbf{x}}_i$$

Svarer (næsten) til variansen for OLS med $\hat{\mathbf{X}}$ i stedet for $\mathbf{X}$

Bemærk dog: $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}^{2SLS}$ (dvs $\hat{\mathbf{u}} \neq \mathbf{y} - \hat{\mathbf{X}}\hat{\beta}^{2SLS}$).

Two-Stage Least Squares (2SLS): Inferens

Vær opmærksom på to ting:

2SLS standardfejlene vil typisk være større end ved OLS.

- Vi "smider" noget af variationen i $\mathbf{x}$ væk $\Rightarrow$ Større varians i $\hat{\beta}$.

Standardfejlene bliver forkerte, hvis man manuelt laver de to trin og bruger $\hat{\mathbf{u}}$ fra second stage.

- Standardfejlene fra second stage tager ikke højde for, at $\hat{\mathbf{X}}$ er en estimeret størrelse og derfor også har en varians.
- $\hat{\mathbf{u}}$ skal beregnes med udgangspunkt i den sande populations model

$$\mathbf{u} = \mathbf{y} - \mathbf{X}\beta \Rightarrow \hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}^{2SLS}$$

Altså med $\mathbf{X}$ (dvs. IKKE $\hat{\mathbf{X}}$) og $\hat{\beta}^{2SLS}$ (konsistent estimator for β)

Two Stage Least Squares (2SLS): Quiz

Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

hvor x_1 er endogen ($Cov(x_1, u) \neq 0$), og x_2 er eksogen ($Cov(x_2, u) = 0$).

Hvordan får vi konsistent estimat af β_1?

1. Anvender x_2 som instrument for x_1 og estimer med 2SLS.
2. Anvender et z som opfylder $Cov(x_1, z) \neq 0$ og $Cov(u, z) = 0$, hvor x_2 og z ikke er perfekt korrelerede.
3. Anvender et z som opfylder $Cov(x_2, z) \neq 0$ og $Cov(u, z) = 0$, hvor x_1 og z ikke er perfekt korrelerede.

Flere instrumenter og styke af instrumenter

Flere instrumenter kan gøre 2SLS estimaterne mere præcise.

- Men kun hvis instrumenterne er stærke
- Dvs. signifikant forskellige fra nul i first stage estimationen.

Hvis instrumenterne er svage, er 2SLS biased mod OLS estimatet

Vigtigt at teste styrken af instrumenterne (F-test for om alle instrumenter er signifikante i first stage).

Two Stage Least Squares (2SLS): Python eksempel

Effekten af uddannelse for gifte amerikanske kvinder

Model:

$$\log(wage) = \beta_0 + \beta_1 age + \beta_2 exp + \beta_3 exp^2 + \beta_4 educ + u$$

$$educ = \pi_0 + \pi_1 age + \pi_2 exp + \pi_3 exp^2 + \pi_4 motheduc + \pi_5 fatheduc + v$$

- Data fra Mroz
- Endogen variabel: Uddannelse (educ)
- Instrument: Mors og fars uddannelse
- Tre IV estimationer:
  1. Mors uddannelse som instrument
  2. Fars uddannelse som instrument
  3. Både mors og fars uddannelse som instrument
- F-test for styrke af instrumenter i first stage ($H_0 : \pi_4 = \pi_5 = 0$)

Lad os se om vi kan!

- **Jupyter Notebook:** `11_iv_examples.ipynb`
- **Part 4a:** Lønligning: OLS og 2SLS med flere instrumenter
- **Part 4b:** Lønligning: Test for styrke af instrumenter (F-test)
- **Part 4c:** Monte Carlo: Flere stærke/svage instrumenter

Hvornår har vi typisk flere instrumenter i praksis?

Det er svært at finde blot ét godt intrument.

Så hvornår har vi den luksus at have flere instrument i praksis?

Et typisk eksempel er brug af interaktionsled:

$$y = \beta_0 + \beta_1 udd + \beta_2 kvinde + \beta_3 udd \times kvinde + u,$$

hvor vi har interagere uddannelse med en dummy for kvinde for at undersøge om der er forskelle i afkastet for mænd og kvinder.

Nu indeholder modellen to endogene variable!

Hvordan får vi to instrumenter hvis vi kun har ét (z)?

Flere instrumenter: Forskningseksempel

THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. 131 August 2016 Issue 3

FIELD OF STUDY, EARNINGS, AND SELF-SELECTION*

LARS J. KIRKEBOEN
EDWIN LEUVEN
MAGNE MOGSTAD

This article examines the labor market payoffs to different types of postsecondary education, including field and institution of study. Instrumental variables (IV) estimation of the payoff to choosing one type of education compared to another is made particularly challenging by individuals choosing between several types of education. Not only does identification require one instrument per alternative, but it is also necessary to deal with the issue that individuals who choose the same education may have different next-best alternatives. We address these difficulties using rich administrative data for Norway's postsecondary education system. A centralized admission process creates credible instruments from discontinuities that effectively randomize applicants near unpredictable admission cutoffs into different institutions and fields of study.

Test for eksogeneitet

Test for eksogeneitet

Hvis de forklarende variable er endogene, er OLS biased. Vi har derfor brug for instrumenter til konsistente estimater.

- **Men** hvis variable er eksogene, er OLS mere efficient.
- Derfor tester vi, om de forklarende variable er eksogene.

Test:

1. Test om $\hat{\beta}^{OLS} = \hat{\beta}^{IV}$, fx vha. et Hausman test
   (ikke dækket i Wooldridge - vent til Econometrics B).
2. Test om residualer fra first stage er korreleret med second stage.

Hvorfor testen for eksogenitet virker (trin 1)

Model og first stage:

$$\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \mathbf{u}, \qquad \mathbf{X_2} = \mathbf{Z}\boldsymbol{\pi} + \mathbf{e},$$

hvor $\mathbf{Z} = (\mathbf{X_1}, \mathbf{Z_e})$ og $\text{Cov}(\mathbf{Z}, \mathbf{u}) = 0$.

Substituér first stage i modellen:

$$\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{Z}\boldsymbol{\pi}\boldsymbol{\beta_2} + \underbrace{(\mathbf{e}\boldsymbol{\beta_2} + \mathbf{u})}_{\text{nyt fejlled}}.$$

Hvis $\text{Cov}(\mathbf{e}, \mathbf{u}) \neq 0$, er $\mathbf{X_2}$ endogen, for så indeholder $\mathbf{e}$ dele af $\mathbf{u}$. Dermed bliver den del af $\mathbf{X_2}$, som ikke kan forklares af $\mathbf{Z}$, korreleret med fejlleddet.

Idé: Test denne korrelation direkte ved at inkludere $\hat{\mathbf{e}}$ i modellen.

Hvorfor testen for eksogenitet virker (trin 2)

Testligning:

$$\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \rho\hat{\mathbf{e}} + \boldsymbol{\varepsilon}.$$

Hypotese:

$$H_0 : \rho = 0 \ (\mathbf{X_2} \text{ eksogen}), \qquad H_1 : \rho \neq 0 \ (\mathbf{X_2} \text{ endogen}).$$

Intuition:

- $\hat{\mathbf{e}}$ repræsenterer variationen i $\mathbf{X_2}$, der ikke kan forklares af $\mathbf{Z}$.
- Hvis $\hat{\mathbf{e}}$ forklarer $\mathbf{y}$, indeholder den noget af $\mathbf{u}$ $\rightarrow$ endogenitet.
- Hvis ρ ikke er signifikant, er der ingen systematisk korrelation $\rightarrow$ $\mathbf{X_2}$ kan behandles som eksogen.

Fortolkning:

$$\hat{\rho} \text{ signifikant } \Rightarrow \text{Cov}(\hat{\mathbf{e}}, \mathbf{u}) \neq 0 \Rightarrow \mathbf{X_2} \text{ er endogen.}$$

Lad os se om vi kan!

- **Jupyter Notebook:** `11_iv_examples.ipynb`
- **Part 5:** Lønligning: Test for eksogeneitet af uddannelse

Test for overidentifikation (W15.5)

Test for overidentifikation

Hvis der er flere instrumenter end endogene variable ($g > l$), er modellen **overidentificeret**.

Relevans af overidentifikation:

- Overidentifikation giver flere IV-estimater ved at variere sæt af instrumenter.
- Residualerne i second stage er ikke nul, hvilket muliggør test af instrumenternes gyldighed.

Test for instrumenternes gyldighed:

1. **Hausman-test**: Sammenligner forskellige $\hat{\beta}^{IV}$-estimater med forskellige instrumenter.
2. **Sargan-test**: LM-test for overidentifikation, hvor $H_0 : cov(\mathbf{Z}, \mathbf{u}) = 0$.

Test for overidentifikation

Sargan-test: Bruger de $g - l$ *overidentificerende restriktioner* til at teste instrumenternes gyldighed, dvs. $H_0 : cov(\mathbf{Z}, \mathbf{u}) = 0$

- Estimer modellen med 2SLS ved brug af alle g instrumenter.
- Beregn IV-residualerne $\hat{\mathbf{u}}^{IV}$.
- Estimer hjælperegression $\hat{\mathbf{u}}^{IV} = \mathbf{Z}\phi + v$ vha. OLS og beregn R^2.
- Test nulhypotesen $H_0 : \phi = \mathbf{0}$, hvilket svarer til $H_0 : cov(\mathbf{Z}, \mathbf{u}) = 0$.
- Beregn LM-teststørrelsen for Sargan-testet:

$$LM = nR^2 \sim \chi^2(g - l)$$

hvor $g - l$ er antallet af overidentificerende restriktioner.

Testet har ofte ikke stor styrke. I små stikprøver risikerer vi ofte at acceptere H_0, selvom den er falsk.

Sargan-test og GMM-kriteriefunktionen

- Vi kan basere testen for overidentifikation på GMM-kriteriefunktionen evalueret i $\hat{\boldsymbol{\beta}}^{IV}$:

$$J = Q(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{u}}^{IV\prime}\mathbf{Z}\mathbf{W}\mathbf{Z}'\hat{\mathbf{u}}^{IV}$$

  hvor $\mathbf{W}$ er vægtmatricen.
- Når $\mathbf{W} = \text{var}(\mathbf{u}^{IV})^{-1}$, er vægtningen optimal, og $J \sim \chi^2(g - l)$, hvor $g - l$ er antallet af overidentificerende restriktioner.
- Under homoskedasticitet er den optimale vægtmatrice $\mathbf{W} = (\sigma^2\mathbf{Z}'\mathbf{Z})^{-1}$, så Sargan-teststørrelsen er identisk med LM-testen nR^2 fra hjælpe regressionen:

$$\begin{aligned} J = Q(\hat{\boldsymbol{\beta}}) &= \hat{\mathbf{u}}^{IV\prime}\mathbf{Z}(\sigma^2\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{u}}^{IV} \\ &= \frac{1}{\sigma^2}\hat{\mathbf{u}}^{IV\prime}\mathbf{P}_{\mathbf{Z}}\hat{\mathbf{u}}^{IV} = \frac{ESS}{\sigma^2} = nR^2 \sim \chi^2(g - l) \end{aligned}$$

Test for overidentifikation

Når modellen er eksakt identificeret ($g = l$):

$$\begin{aligned}
\mathbf{Z}'\hat{\mathbf{u}}^{IV} &= \mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\beta}^{IV}) \\
&= \mathbf{Z}'(\mathbf{y} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{y})) \\
&= (\mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{y})) \\
&= (\mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{y}) = \underset{g\times 1}{\mathbf{0}}
\end{aligned}$$

Teststørrelsen J bliver derfor:

$$\begin{aligned}
J &= \hat{\mathbf{u}}^{IV\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{u}}^{IV} \\
&= \mathbf{0}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{0} = 0
\end{aligned}$$

- Instrumenter er per konstruktion ukorrelerede med fejlleddet $\hat{\mathbf{u}}^{IV}$.
- Test for overidentifikation er derfor irrelevant, når $g = l$.

Lad os se om vi kan!

- **Jupyter Notebook:** `11_iv_examples.ipynb`
- **Part 6:** Lønligning: Test for overidentification (eksogeneitet af instrumenter)

8 trins IV procedure

8 trins IV procedure

Trin 1: Definer en (strukturel) model $y = \mathbf{X}\beta + u$. Bestem hvilke variable, som er potentielt endogene: $\mathbf{X_2}$ (l endogene variable).

Trin 2: Find g instrumenter ($g \geq l$).

Trin 3: Opstil "first stage" regressionen for de l endogene variable ($\tilde{x}$)

$$\tilde{x}_j = \mathbf{Z\Pi} + e_j$$

hvor $\underset{n\times k-l+g}{\mathbf{Z}} = (\underbrace{x_1\ x_2 \ldots x_{k-l}}_{k-l \text{ eksogene}}\ \underbrace{z_1 \ldots z_g}_{g \text{ instrumenter}})$.

Trin 4: Test for om $z_1 \ldots z_g$ er signifikante. Hvis de ikke er signifikante (i det fælles test), har vi svage instrumenter, og vi kan få problemer med IV estimationen. Man bør derfor finde andre instrumenter.

8 trins IV procedure

Trin 5: Gem residualerne fra "first stage" regressionen, $\hat{e}_j$ (l sæt af residualer).

Trin 6: Test om x'erne faktisk er endogene ved at estimere hjælperegression:

$$y = \mathbf{X}\boldsymbol{\beta} + \hat{\mathbf{E}}\boldsymbol{\rho} + \varepsilon,$$

og test hypotesen $H_0 : \boldsymbol{\rho} = \mathbf{0}$. Hvis vi afviser hypotesen, er mindst en af de potentiel endogene variable endogen.

Trin 7: Hvis der er endogene variable, estimer modellen med IV/2SLS. Hvis alle variable er eksogene er det mere efficient at bruge OLS.

Trin 8: Hvis $g > l$ kan man teste for overidentifikation.

Opsummering

Quiz

Model og antagelser:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

$$x_i = \delta u_i + \theta_1 z_{1i} + \theta_2 z_{2i} + e_i, \quad z_{2i} = \rho z_{1i} + w_i,$$

$$w \sim N(0,1),\ u \sim N(0,1),\ e \sim N(0,1),\ z_1 \sim U(-1,1)$$

	A	B	C	D	E
δ	-1	2	0	1	3
θ_1	-1	0	1	2	0
θ_2	1	1	0	0	1
ρ	0	0	1	1	2

- For hvilke sæt af parametreneværdier (A-E) er OLS konsistent?
- I tilfælde D er OLS opad biased, nedad biased eller konsistent?
- For hvilke parameterværdier er z_1 og z_2 et gyldigt instrument?

Opsummering

- **Endogene variable** betyder, at OLS er inkonsistent og ikke middelret.
- Ved endogene variable måler OLS **korrelationer** og ikke **kausale** sammenhæng.
- Konsistent estimation kan opnås ved **IV estimation**.
- Det kræver **gyldige instrumenter**: Korreleret med den endogene variabel og ukorrelateret med fejlleddet.
    - Man kan tjekke empirisk, om instrumentet er korreleret med den endogene variabel.
    - Om instrumentet er ukorreleret med fejlleddet kan ikke umiddelbart tjekkes. Det kræver en (teoretisk) argumentation.

Opsummering

- **2SLS** (Two Stage least Squares): To trins procedure til estimation i den multiple lineære regressionsmodel (med evt. flere endogene variable).
  - **Eksogene variable** kan fungere som instrument for dem selv.
  - **IV antagelserne**: Instrumenterne er ukorrelerede med fejlleddet, instrumenterne er korrelerede med alle de endogene variable, ingen perfekt multikollinaritet mellem instrumenterne.
- **Eksakt identifikation**: Det samme antal instrumenter som endogene variable.
- **Overidentifikation**: Flere instrumenter end endogene variable.
- Vigtigt med **stærke instrumenter** dvs. at instrumenterne er højt korrelerede med de endogene variable, og at de ikke indbyrdes er for højt korrelerede.

Opsummering

- **Eksogenitetstest**: Testet kan bruges til at afgøre, om en potentiel endogen variabel er endogen.
  - Hvis en variable ikke er endogen, er det mere efficient at estimere med OLS.
- **Test for overidentifikation**: Med flere instrumenter end endogene variable, kan vi teste om (nogle af instrumenterne er gyldige). Testet har ofte ikke stor styrke.
- **8 trins proceduren:** Standardprocedure for modeller med potentiel endogene variable.