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)

- Two Stage Least Squares (2SLS)

Test for eksogeneitet og overidentifikation (W15.5)

- Test for eksogeneitet
- Test for overidentifikation

Appendix

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: Militærtjeneste og efterfølgende kriminalitet

- Vi undersøger, om der er en effekt af at aftjene værnepligt på kriminalitet.
- Model:

$$y_{krim} = \beta_0 + \beta_1 D_{militær} + \text{baggrundsvariable} + u$$

- Baggrundsvariable: alder, udd., højde, vægt, IQ, geografi mv.
- Kan OLS estimatet tolkes som en kausal effekt?
- Der kan være udeladte variable.

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:

$$y = b_0 + b_1 z + \nu \tag{1}$$

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

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

- Ligning (1): hvis individet fik højere SU, så stiger lønnen med b_1.
- Ligning (2): hvis individet fik højere SU, så stiger antal års uddannelse med γ_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 b_1/γ_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.

Instrument variabel estimation: Forskningseksempler

Contents lists available at ScienceDirect

Journal of Public Economics

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

Grandchildren and their grandparents' labor supply☆

Peter Rupert[a], Giulio Zanella[b,*]

[a] University of California–Santa Barbara, USA
[b] University of Bologna, Italy

ARTICLE INFO

JEL classification:
D19
J13
J14
J22

Keywords:
Labor supply
Grandparents
Child care

ABSTRACT

Working-age grandparents supply large amounts of child care, an observation that raises the question of how having grandchildren affects grandparents' own labor supply. Exploiting the unique genealogical design of the PSID and the random variation in the timing when the parents of first-born boys and girls become grandparents, we estimate a structural labor supply model and find a negative effect on employed grandmother's hours of work of about 30% that is concentrated near the bottom of the hours distribution, i.e., among women less attached to the labor market. Implications for the evaluation of child care and parental leave policies are discussed.

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.

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"

Egenskaber ved IV estimatoren

Under de betingelse 1 og 2 har IV estimatoren pæne asymptotiske egenskaber:

- Konsistent.
- Asymptotisk normalfordelt.

IV estimatoren er ikke middelret (ikke god med små stikprøver).

- Bias er særlig udtalt, hvis vi har svage instrumenter ($Cov(x, z) \approx 0$).
- IV estimatoren har større varians end OLS.

Konsistens af IV estimatoren: Bevis

Beviset for at IV estimatoren er konsistent svarer til at "gå baglæns" gennem udledningen af estimatoren.

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

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

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

Asymptotisk bias: IV vs OLS

Asymptotisk bias

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

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

- $cov(x,z)$ vil typisk være mindre end $var(x)$.
- Hvis hverken $cov(u,z) \neq 0$ eller $cov(u,x) \neq 0$, kan

$$\left|\frac{cov(u,z)}{cov(x,z)}\right| > \left|\frac{cov(u,x)}{var(x)}\right|$$

Hvis vores antagelser ikke holder, kan IV være mere asympotisk biased end OLS!

Inferens

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

$$Avar(\hat{\beta}_1^{IV}) = \frac{\sigma_u^2}{n\sigma_x^2\rho_{x,z}^2} > \frac{\sigma_u^2}{n\sigma_x^2} = Avar(\hat{\beta}_1^{OLS}) \tag{3}$$

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

Vi kan estimere den asymptotiske varians konsistent som

$$\widehat{Avar}(\hat{\beta}_1^{IV}) = \frac{\frac{1}{n-2}\sum \hat{u}_i^2}{R_{x,z}^2 SST_x} \tag{4}$$

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

Hvorfor er IV estimatoren ikke middelret?

Ihukom tricket da vi beviste at OLS var middelret.

$$E(\hat{\beta}_1^{OLS}|x) = \beta_1 + E\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(u_i)(x_i - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}|x\right) \tag{5}$$

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).

Hvorfor er IV estimatoren ikke middelret?

Vi undersøger fordelingen af IV afhænger af stikprøvestørrelsen (n).

Model

$$x = z - u + \varepsilon$$
$$y = x + u$$
$$z \sim N(0,1), \quad u \sim N(0,1), \quad \varepsilon \sim N(0,1)$$

Simuler modellen 10.000 gange med $n \in \{10, 25, 50, 100, 1000\}$

Se programmet "12 IVsim 1 middelret.do" (Absalon)

Hvorfor er IV estimatoren ikke middelret?

```
summarize

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      ols_n10 |     10,000    .6651013    .1801795  -.3744334   1.810834
      ols_n25 |     10,000     .664992    .1014017   .2239469   1.137489
      ols_n50 |     10,000       .66753    .0683052   .3775977   .9465832
     ols_n100 |     10,000    .6657555    .0482164   .4763262    .850592
    ols_n1000 |     10,000    .6666001    .0149508   .6087326   .7259724
-------------+---------------------------------------------------------
       iv_n10 |     10,000    1.553615    26.66313  -207.3588   2452.405
       iv_n25 |     10,000     1.05976    1.109425   -43.4525   69.82385
       iv_n50 |     10,000    1.025989    .1922662   .5943428   9.128614
      iv_n100 |     10,000     1.01004    .1088066   .6531173   1.605745
     iv_n1000 |     10,000    1.001345    .0322753   .8886861    1.16284
```

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

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

IV som løsning på målefejl: Eksempel

Hvis vi kan finde en variabel, som er korreleret med den sande værdi x^*, men ukorreleret med målefejlen, kan den bruges som instrument.

Forslag: et andet mål z for x^*, som også er observeret med målefejl:

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

Under antagelse af at målefejlene e og v er uafhængige, kan β_1 estimeres konsistent med IV estimation.

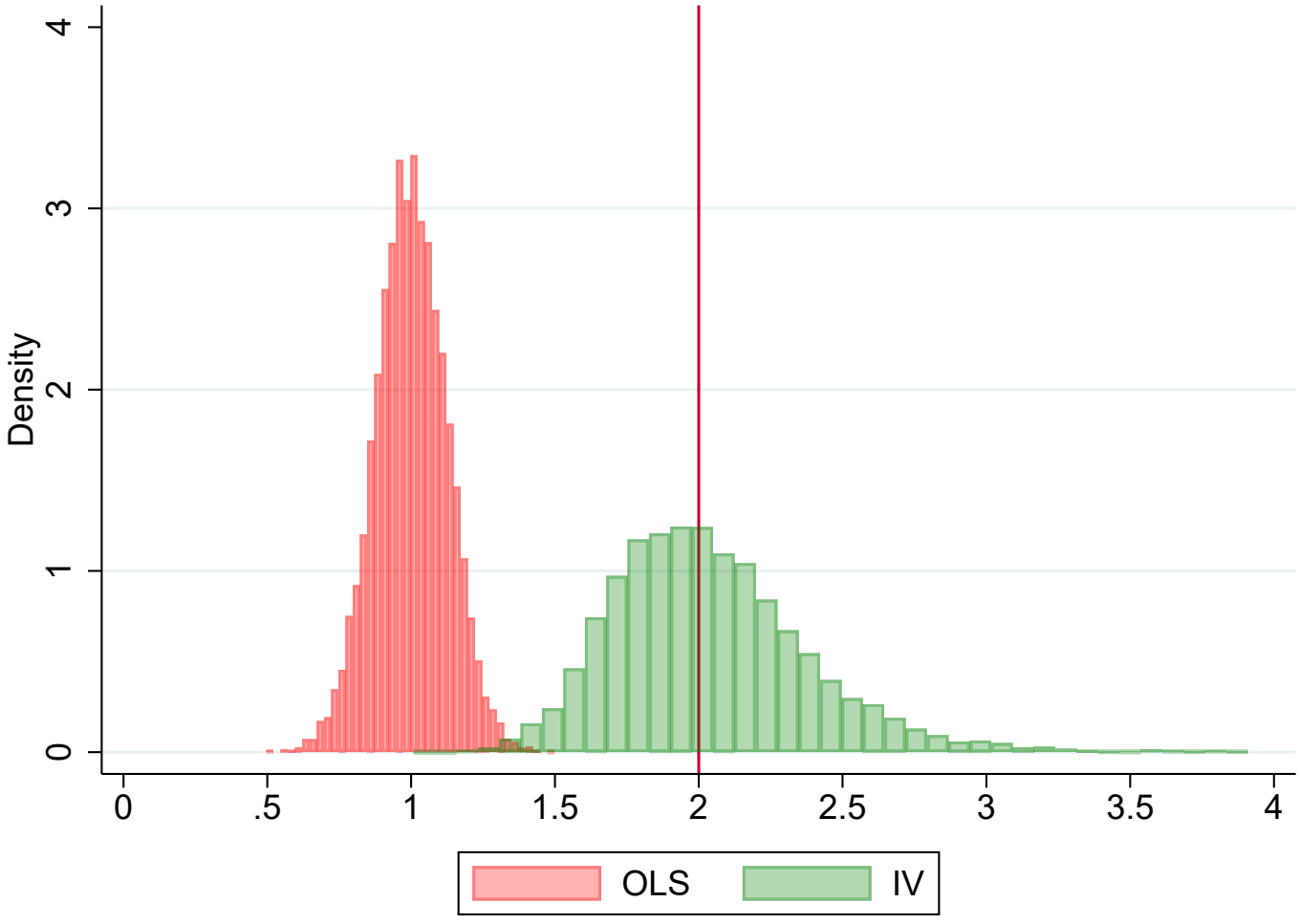
Eksempel: Målefejl i selvrapporteret uddannelse. Spørg en anden, fx kollega eller familiemedlem.[1]

[1] Ashenfelter, Orley, and Alan Krueger. "Estimates of the economic return to schooling from a new sample of twins." The American economic review (1994): 1157-1173.

IV som løsning på målefejl: Eksempel

```
global  var_e = 0.5
program olsdata, rclass
...
*DATA GENERATING PROCESS
generate xstar = 2+1*rnormal()
generate e1 =rnormal(0,${var_e}^0.5)
generate v1 =rnormal(0,${var_e}^0.5)
generate u = rnormal()
generate y = 1+2*xstar+u
generate x = xstar+e1
generate z = xstar+v1

*CALCULATE OLS AND IV ESTIMATES
regress y x
return scalar b_OLS=_b[x]
ivregress 2sls y (x=z)
return scalar b_IV=_b[x]

end
```

IV som løsning på målefejl: Eksempel



Sande parametre: $\beta_1 = 2$. Se programmet "12 IVsim 2 målefejl.do" på Absalon.

IV estimation: MLR

Multipel lineær regressionsmodel med endogene variable

Regressionsmodel med matrixnotation (se SLP afsnit 3.2):

$$\underset{n\times 1}{y} = \underset{n\times k}{X} \cdot \underset{k\times 1}{\beta} + \underset{n\times 1}{u}$$

Vi antager, at vi har $l \leq k$ **endogene** variable og $k - l$ **eksogene** variable:

$$X = (\underbrace{x_1\ x_2 \ldots x_{k-l}}_{k-l \text{ eksogene}}\ \underbrace{x_{k-l+1} \ldots x_k}_{l \text{ endogene}}).$$

Da de sidste l variable er endogene, gælder:

$$p\lim\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}u_i\right) \neq 0 \text{ for } j = k - l + 1, .., k.$$

Dvs. de endogene variable er korreleret med fejlleddet.

Multipel lineær regressionsmodel med endogene variable

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

Eksakt identifikation: Samme antal instrumenter g som endogene variable l:

$$\underset{n\times k}{Z} = (\underbrace{x_1 \; x_2 \ldots x_{k-l}}_{k-l \text{ eksogene}} \quad \underbrace{z_1 \ldots z_l}_{l \text{ instrumenter}} \;)$$

Multipel lineær regressionsmodel med endogene variable

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

Overidentifikation:

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

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

De $k - l$ eksogene variable x_1, $x_2 \ldots x_{k-l}$ er instrumenter for dem selv.

Z indeholder således alle eksogene $h = k - l + g$ variable:

Udledning af IV estimator

For at udlede IV estimatoren antagelser vi følgende:

Instrumenterne er eksogene:

$$p\lim\left(\frac{1}{n}\sum_{i=1}^{n} z_{ij}u_i\right) = 0 \Leftrightarrow p\lim\frac{1}{n}Z'u = \underset{k-l+g\times 1}{0} \tag{6}$$

Instrumenterne er korreleret med de endogene variable:

$$p\lim\frac{1}{n}Z'X = \underset{h\times k}{\Sigma_{ZX}} \text{ har fuld rang} \tag{7}$$

Ingen perfekt multikollinaritet mellem instrumenterne:

$$p\lim\frac{1}{n}Z'Z = \underset{h\times k-l+g}{\Sigma_{ZZ}} \text{ har fuld rang} \tag{8}$$

Udledning af IV estimator: Eksakt identificeret

Eksakt identifikation: $g = l$ instrumenter.

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

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

Denne estimator er konsistent.

Udledning af IV estimator: Eksakt identificeret

Bevis for konsistens: 1) Indsæt for y, reducer og 2) tag $p\lim$.

Two Stage Least Squares (2SLS)

2SLS er et alternativ til IV estimatoren, som også virker, når der er flere instrumenter end endogene variable.

Procedure: Trin 1

- Regresser hver (endogen) variable i X på Z
- Beregn de prædikterede værdier $\hat{x}_j \in \hat{X}$.

Eksempel:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 \tilde{x}_2 + \beta_3 \tilde{x}_3 + u$$

hvor $\tilde{x}$ markerer endogene variable.

Two Stage Least Squares (2SLS)

Procedure: Trin 2

1. Regresser y på $\hat{X}$ (i stedet for X).

Eksempel fortsat:

2SLS på matrix form:

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

hvor $\hat{X}$ er prædiktionerne af x givet z (se appendiks for mere).

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.

Two Stage Least Squares (2SLS): Inferens

Vi kan også udlede variansen på 2SLS:

$$var(\hat{\beta}^{2SLS}|Z, X) =$$

Two Stage Least Squares (2SLS): Inferens

Med homoskedasticitet er variansen på 2SLS givet ved

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

hvor

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

Med heteroskedasticitet

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

hvor

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

Begge dele svarer til OLS med $\hat{X}$ i stedet for X.

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 x væk $\Rightarrow$ Større varians i $\hat{\beta}$.

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

- Standardfejlene fra second stage tager ikke højde for at $\hat{X}$ er en estimeret størrelser og derfor også har en varians.
- $\hat{u}$ skal tages fra den rigtige model med X og $\hat{\beta}^{2SLS}$

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.

Two Stage Least Squares (2SLS): Stata eksempel

STATA kan lave IV/2SLS estimation ved at benytte proceduren **ivregress**

Antag fx at vi ønsker at estimere:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

hvor x_2 er endogen, og vi har et instrument z.

Procedurer:

- Second stage: ivregress 2sls y x1 (x2=z) x3
- First stage: ivregress 2sls y x1 (x2=z) x3, first
- Robuste standardfejl: ivregress 2sls y x1 (x2=z) x3, robust

Two Stage Least Squares (2SLS): Stata eksempel

Effekten af uddannelse for gifte amerikanske kvinder

Model:

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

- 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

Two Stage Least Squares (2SLS): Stata eksempel

```
* Regressions
/* OLS estimation */
regress logwage educ age exper expersq
estimates store OLS
/* IV estimation motheduc instrument */
ivregress 2sls logwage age exper expersq (educ=motheduc), first
estimates store IV_1
/* IV estimation fatheduc instrument */
ivregress 2sls logwage age exper expersq (educ=fatheduc), first
estimates store IV_2
/* IV estimation motheduc and fatheduc instruments */
ivregress 2sls logwage age exper expersq (educ=motheduc fatheduc), first
estimates store IV_3

/* Table with estimation results */
est tab OLS IV_1 IV_2 IV_3, stats(N r2) star(.05 .01 .001) b(%7.3f) varlabel
```

Two Stage Least Squares (2SLS): Stata eksempel

```
First-stage regressions                         Number of obs     =        428
----------------------                          F(   5,    422)   =      22.67
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2118
                                                Adj R-squared     =     0.2024
                                                Root MSE          =     2.0410

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0060299   .0151493     0.40   0.691    -.0237476    .0358073
       exper |   .0451279   .0402916     1.12   0.263    -.0340693     .124325
     expersq |   -.001091    .001222    -0.89   0.372     -.003493     .001311
    motheduc |   .1604449   .0366353     4.38   0.000     .0884344    .2324553
    fatheduc |   .1886367   .0338676     5.57   0.000     .1220664     .255207
       _cons |   8.851148   .7625896    11.61   0.000     7.352201     10.3501
------------------------------------------------------------------------------
```

Two Stage Least Squares (2SLS): Stata eksempel

```
Instrumental variables (2SLS) regression          Number of obs   =        428
                                                  Wald chi2(4)    =      24.63
                                                  Prob > chi2     =     0.0001
                                                  R-squared       =     0.1353
                                                  Root MSE        =     .67169

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0609945   .0313948     1.94   0.052    -.0005381    .1225271
         age |  -.0003542   .0049029    -0.07   0.942    -.0099638    .0092553
       exper |    .044196   .0133736     3.30   0.001     .0179842    .0704078
     expersq |  -.0008947   .0004051    -2.21   0.027    -.0016886   -.0001007
       _cons |   .0667186   .4564205     0.15   0.884    -.8278492    .9612864
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   age exper expersq motheduc fatheduc
```

Two Stage Least Squares (2SLS): Stata eksempel

```
est tab OLS IV_1 IV_2 IV_3, stats(N r2) star(.05 .01 .001) b(%7.3f) varlabel
```

Variable	OLS	IV_1	IV_2	IV_3
Education	0.108***	0.049	0.070*	0.061
Age	0.000	-0.001	-0.000	-0.000
Experience	0.042**	0.045***	0.044**	0.044***
Experience sq.	-0.001*	-0.001*	-0.001*	-0.001*
Constant	-0.533	0.228	-0.051	0.067
N	428	428	428	428
r2	0.157	0.122	0.143	0.135

legend: * p<.05; ** p<.01; *** p<.001

Flere instrumenter: Simulering

Flere instrumenter kan gøre 2SLS estimaterne mere præcise.

- Men kun hvis instrumenterne er stærke
- dvs. signifikant forskellige fra 0 i first stage estimationen.
- Se "12 IVsim 3 flere stærke instrumenter.do" på Absalon.

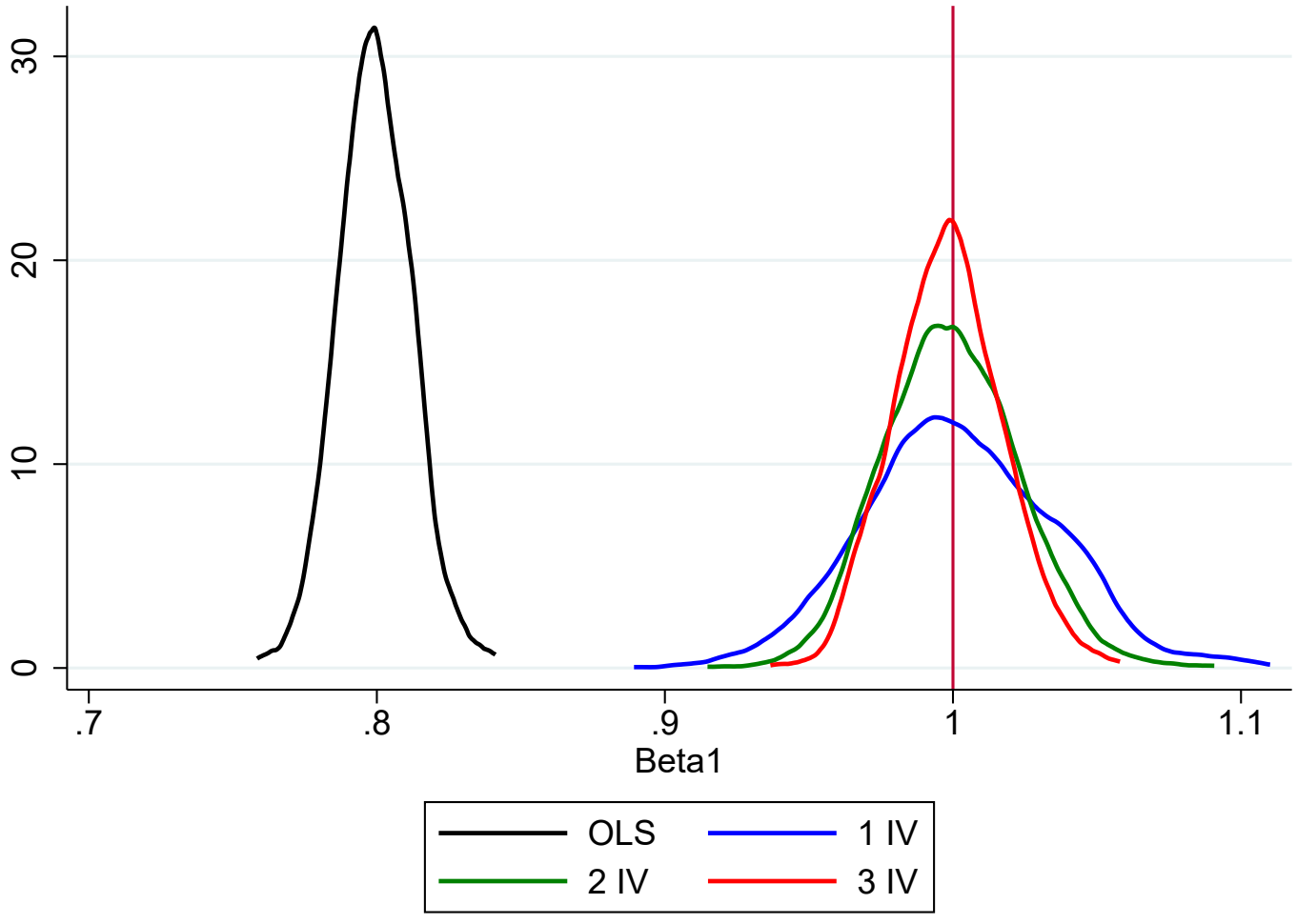
Hvis instrumenterne er svage, er 2SLS biased mod OLS estimatet

- Se "12 IVsim 4 flere svage instrumenter.do" på Absalon.

Flere stærke instrumenter: Simulering

```
*DATA GENERATING PROCESS
generate z1 = rnormal()
generate z2 = rnormal()
generate z3 = rnormal()
generate u =  rnormal()
generate x = z1 + z2 + z3 - u + rnormal()
generate y = x + u

*CALCULATE OLS ESTIMATES
regress y x

*CALCULATE IV ESTIMATES
ivregress 2sls y (x=z1)

ivregress 2sls y (x=z1 z2)

ivregress 2sls y (x=z1 z2 z3)
```

Flere stærke instrumenter: Simulering



Se "12 IVsim 3 flere stærke instrumenter.do" på Absalon.

Flere svage instrumenter: Simulering

```
*DATA GENERATING PROCESS
foreach i of numlist 1/10 {
generate z‘i’ = rnormal()
}
generate u = rnormal()
generate x = - u + rnormal()
generate y = x + u

*CALCULATE OLS ESTIMATES
regress y x

*CALCULATE IV ESTIMATES
ivregress 2sls y (x=z1)

ivregress 2sls y (x=z1 z2)

ivregress 2sls y (x=z1 z2 z3)

ivregress 2sls y (x=z*)
```

Flere svage instrumenter: Simulering

```
summarize, sep(5)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
ols	1,000	.4998717	.0162789	.4505327	.5464192
iv1	1,000	-.1568515	39.67745	-1173.028	382.7818
iv2	1,000	.546548	1.796689	-16.10138	47.21078
iv3	1,000	.5277061	.5362299	-2.70867	7.545203
iv10	1,000	.5012571	.1772621	-.1325141	1.245804

Se "12 IVsim 4 flere svage instrumenter.do" på Absalon.

$\Rightarrow$ Vær opmærksom på First Stage!

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 og overidentifikation

Test for eksogeneitet

Hvis de forklarende variable er endogene, vil OLS være bias, hvorfor vi har brug for et instrument for at opnå konsistente estimator.

- **Men** hvis de forklarende variable er eksogene, vil OLS være mere efficient.
- Det er derfor relevant at teste for om de forklarende variable er eksogene.

Test for eksogenitet:

1. Teste om $\hat{\beta}^{OLS} = \hat{\beta}^{IV}$, fx vha en Hausman test (ikke dækket i Wooldridge).
2. Teste om fejleddene fra first stage er korreleret med fejleddene fra second stage.

Test for eksogeneitet

Procedure 2 bygger på 2SLS proceduren.

Ideen er at **opdele variationen** i den endogene variabel x_i i to dele

$$x = \pi_0 + \pi_1 z_i + e$$

hvor $\hat{x} = \hat{\pi}_0 + \hat{\pi}_1 z_i$ er ukorreleret med u.

x er derfor kun endogen, hvis u og e er korrelerede.

Vi kan teste om u og e er korrelerede ved at estimere

$$u = \rho e + \varepsilon$$

Test for eksogeneitet

Problemet er bare at vi ikke observerer u og e.

Wooldridge foreslår i stedet at erstatter

- e med residualerne first estimationen ($\hat{e}$).
- u med $u = y - \beta_0 - \beta_1 x$

Og derefter estimere

$$y = \beta_0 + \beta_1 x + \rho \hat{e} + \varepsilon$$

Her gælder at hvis

- x er **eksogen** $\Rightarrow \rho = 0$ (H_0).
- x er **endogen** $\Rightarrow \rho \neq 0$ (H_A).

Test for eksogeneitet

Procedure

1. Estimer first stage: $x = \pi_0 + \pi_1 z_i + e$
2. Gem residualerne $\hat{e}$
3. Tilføj resultaterne den "strukturelle model"

$$y = \beta_0 + \beta_1 x + \rho \hat{e} + \varepsilon$$

4. Test om $\rho \neq 0$.

Test for eksogeneitet: Stata eksempel

```
drop if logwage==.

/* IV estimation motheduc and fatheduc instruments */
*Second stage
ivregress 2sls logwage age exper expersq (educ=motheduc fatheduc)
predict uhat, residuals

*First stage
regress educ motheduc fatheduc age exper expersq
predict Ehat, residuals

*Test for exogeneity
regress logwage age exper expersq educ Ehat
```

Test for eksogeneitet: Stata eksempel

```
. regress logwage age exper expersq educ Ehat

      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(5, 422)       =     16.37
       Model |  36.2728196         5  7.25456392   Prob > F        =    0.0000
    Residual |  187.054621       422  .443257396   R-squared       =    0.1624
-------------+----------------------------------   Adj R-squared   =    0.1525
       Total |  223.327441       427  .523015084   Root MSE        =    .66578

------------------------------------------------------------------------------
     logwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0003542   .0048597    -0.07   0.942    -.0099066    .0091981
       exper |    .044196   .0132558     3.33   0.001     .0181403    .0702517
     expersq |  -.0008947   .0004015    -2.23   0.026    -.0016839   -.0001055
        educ |   .0609945   .0311183     1.96   0.051    -.0001716    .1221607
        Ehat |   .0586438   .0349356     1.68   0.094    -.0100257    .1273133
       _cons |   .0667186   .4524011     0.15   0.883    -.8225216    .9559588
------------------------------------------------------------------------------
```

Test for overidentifikation

Hvis vi har flere instrumenter end endogene variable ($g > l$), siger vi at modellen er **overidentificeret**.

Da vi i udgangspunktet kan estimere β konsistent med $g = l$, kan vi danne flere forskellige estimater ved at variere det anvendte sæt af instrumenter.

Det muliggør test af om vores instrumenter er gyldige:

1. Test for signifikante forskelle i $\hat{\beta}^{IV}$'erne, fx vha en Hausman Test (ikke dækket i Wooldridge).
2. Test om $cov(z, u) = 0$.

Test for overidentifikation

Procedure for test 2

- Estimer modellen ved brug af alle g instrumenter.
- Beregne IV residualerne $\hat{u}^{IV}$.
- Estimer hjælperegressionen vha. OLS: $\hat{u}^{IV} = \mathbf{Z}\phi + v$
- Nulhypotese $H_0 : \phi = \mathbf{0}$.
- Teststørrelsen beregnes som $OI = nR^2$ (R^2 fra hjælperegression) og er χ^2-fordelt med $g - l$ frihedsgrader.
- Hvis vi forkaster H_0 er nogle af instrumenterne ugyldige, men vi kan ikke konkludere hvilke.

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

Test for overidentifikation: Stata eksempel

```
/* IV estimation motheduc and fatheduc instruments */
*Second stage
ivregress 2sls logwage age exper expersq (educ=motheduc fatheduc)
predict uhat, residuals

*Test for overidentification
regress uhat motheduc fatheduc age exper expersq

di e(N)*e(r2)
di chi2tail(1,e(N)*e(r2))
```

Test for overidentifikation: Stata eksempel

```
      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(5, 422)       =      0.08
       Model |  .181414327         5  .036282865   Prob > F        =    0.9954
    Residual |  192.918849       422   .45715367   R-squared       =    0.0009
-------------+----------------------------------   Adj R-squared   =   -0.0109
       Total |  193.100263       427  .452225441   Root MSE        =    .67613
------------------------------------------------------------------------------
        uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    motheduc |  -.0069461   .0121364    -0.57   0.567    -.0308013    .0169092
    fatheduc |   .0059875   .0112195     0.53   0.594    -.0160656    .0280405
         age |  -.0004989   .0050186    -0.10   0.921    -.0103634    .0093657
       exper |  -.0000119   .0133476    -0.00   0.999     -.026248    .0262241
     expersq |   7.63e-06   .0004048     0.02   0.985    -.0007881    .0008033
       _cons |   .0315871   .2526267     0.13   0.901    -.4649763    .5281505
------------------------------------------------------------------------------
. di e(N)*e(r2)
.40209853
. di chi2tail(1,e(N)*e(r2))
.52600747
```

Test for overidentifikation

Når vi har eksakt identificeret model:

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

Instrumenterne er således pr definition ukorreleret med fejlleddet, når modellen er eksakt identificeret.

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: $\mathbf{X^b}$ (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.

Lidt historie

First use of an instrument variable occurred in a 1928 book by Philip G. Wright, best known for his excellent description of the production, transport and sale of vegetable and animal oils in the early 1900s in the United States.

Wright attempted to determine the supply and demand for butter using panel data on prices and quantities sold in the United States. The idea was that a regression analysis could produce a demand or supply curve because they are formed by the path between prices and quantities demanded or supplied.

The problem was that the observational data did not form a demand or supply curve as such, but rather a cloud of point observations that took different shapes under varying market conditions. It seemed that making deductions from the data remained elusive.

The problem was that price affected both supply and demand so that a function describing only one of the two could not be constructed directly from the observational data. Wright correctly concluded that he needed a variable that correlated with either demand or supply but not both – that is, an instrumental variable.

After much deliberation, Wright decided to use regional rainfall as his instrumental variable: he concluded that rainfall affected grass production and hence milk production and ultimately butter supply, but not butter demand. In this way he was able to construct a regression equation with only the instrumental variable of price and supply.

Appendix

Two Stage Least Squares (2SLS): Det generelle tilfælde

Trin 1: Regresser de k variable i X på de h instrumenter i Z:

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

Dette er k first stage regressioner kørt i et hug. OLS estimatet er givet ved

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

De prædikterede værdier er:

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

hvor $P_Z = Z(Z'Z)^{-1}Z'$ er en projektionsmatrix med egenskaberne $P_Z = P_Z'$ og $P_Z P_Z = P_Z$.

Two Stage Least Squares (2SLS): Det generelle tilfælde

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

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

2SLS estimatoren

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

Special tilfældet med eksakt identificeret

Når $g = l$ gælder der, at både Z og X har dimensioner $n \times k$. Derfor er Z'X, Z'Z og X'Z alle kvadratiske $k \times k$ matricer.

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

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

Hvilket er IV estimatoren i MLR tilfældet.

Two Stage Least Squares (2SLS): Konsistens

2SLS er konsistent, når IV antagelserne holder.

Ingredienser til bevis:

Antagelser:

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

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

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:

$$\begin{aligned}\hat{\beta}^{2SLS} &= (X'P_ZX)^{-1}(X'P_Zy) \\ &= (X'P_ZX)^{-1}(X'P_ZX\beta) + (X'P_ZX)^{-1}(X'P_Zu) \\ &= \beta + (X'P_ZX)^{-1}X'P_Zu\end{aligned}$$

Two Stage Least Squares (2SLS): Konsistens

Trin 2: Tag Grænsesandsynligheden

$$
\begin{aligned}
&p\lim(\hat{\beta}^{2SLS}) - \beta \\
&\qquad = p\lim\left[(X'P_Z X)^{-1}X'P_Z u\right] \\
&\qquad \text{Husk at } P_Z = Z(Z'Z)^{-1}Z' \\
&\qquad = p\lim\left[(X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'u\right] \\
&\qquad = p\lim\left[(\tfrac{1}{n}X'Z(\tfrac{1}{n}Z'Z)^{-1}\tfrac{1}{n}Z'X)^{-1}\tfrac{1}{n}X'Z(\tfrac{1}{n}Z'Z)^{-1}\tfrac{1}{n}Z'u\right] \\
&\qquad \text{Nu har vi noget, hvor alle elementer har en kendt } p\lim \\
&\qquad = \left[(\Sigma'_{ZX}(\Sigma_{ZZ})^{-1}\Sigma_{ZX})^{-1}\Sigma'_{ZX}(\Sigma_{ZZ})^{-1}0\right] = 0
\end{aligned}
$$

Two Stage Least Squares (2SLS): Inferens

Omskriv 2SLS estimatoren som i trin 1 ovenfor:

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

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

Two Stage Least Squares (2SLS): Inferens

Opsummering

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

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

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

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

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