

Øvelsestime 8

Instrumental Variable

Agenda

- 1) Introduktion til IV-estimation og PS8
- 2) Gruppearbejde: Diskussion
- 3) Opsamling
- 4) Gruppearbejde: Stata
- 5) Opsamling

Tobias Gabel Christiansen
tgc@econ.ku.dk
26-04-2021

KØBENHAVNS UNIVERSITET



0) Planen for semesteret

Dato	Opgave
Uge 7 - 15/02	PS1
Uge 8 - 22/02	PS2
Uge 9 - 01/03	PS3
Uge 10 - 08/03	OO1
Uge 11 - 15/03	PS4
Uge 12 - 22/03	PS5 + gennemgang af OO1
Uge 13 - 29/03	PS6
Uge 14 - 05/04	Påskeferie
Uge 15 - 12/04	OO2
Uge 15 - 19/04	PS7
Uge 16 - 26/04	PS8 + gennemgang af OO2
Uge 17 - 03/05	OO3
Uge 18 - 10/05	PS9
Uge 19 - 17/05	PS10 + gennemgang af OO3
	Spørgetime

 Her er vi :-)

1) Introduktion til IV

- Indtil nu har vi forsøgt at kontrollere os ud af endogenitetsproblemer. Fx ved at inkludere relevante variable som vi mistænker for at samvariere med den variabel vi er interesserede i.
- Nogle gange kan man dog ikke kontrollere sig ud af problemet. Dette er fx tilfældet hvis det som vi ønsker at kontrollerer for er uobserverbart/ikke-måleligt.
- En måde at løse sådanne endogenitetsproblemer er ved at bruge instrumental variable (IV-estimation)

1) Introduktion til IV

- Antag at vi arbejder med følgende model hvor x_i er endogen (samvarierer/korrelerer med u_i):

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad (A)$$

- Det betyder at MLR.4 er brudt og at vores estimatorer vil være biased og inkonsistente. Én måde at undgå dette på er ved at bruge et **instrument** z_i .
- Et **instrument** er en variabel der forklarer/korrelerer med en **eksogen** del af variationen i x_i . Med andre ord skal et instrument z_i opfylde følgende kriterier:

$$\text{cov}(z_i, u_i) = 0 \quad (1)$$

$$\text{cov}(z_i, x_i) \neq 0 \quad (2)$$

- Antagelse (1) er en exclusion restriction. Den angiver at z_i hverken har en direkte effekt på y_i når vi har kontrolleret for x_i og at z_i ikke selv er endogen (korrelerer med andre udeladte faktorer). Når dette er overholdt siger vi at z_i er **eksogen**.
- Antagelse (2) betyder at z_i korrelerer med x_i . Med andre ord er z_i **relevant** da den kan forklare noget af variationen i x_i .
- Når antagelse (1) + (2) er overholdt har vi et **validt** instrument, som kan bruges til at isolere den del af variationen i x_i som er eksogen.

1) Introduktion til IV

Har vi et validt instrument for x_i kan vi finde IV-estimatoren ved at løse følgende moment-betingelser:

$$\begin{aligned} E(u_i) &= E(y_i - \beta_0 - \beta_1 x_i) = 0 \\ E(z_i u_i) &= E(z_i(y_i - \beta_0 - \beta_1 x_i)) = 0 \end{aligned}$$

Løser vi de tilsvarende sample-moments kan vi finde at:

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

IV-estimatoren er altså givet ved forholdet mellem (sample) kovariansen mellem z_i og y_i og (sample) kovariansen mellem z_i og x_i . Ganger vi igennem med $\frac{1}{n}$ og omskriver finder vi:

$$\widehat{\beta}_1^{IV} = \frac{\frac{1}{n} \sum_{i=1}^n (z_i - \bar{z}) y_i}{\frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})} = \frac{\frac{1}{n} \sum_{i=1}^n (z_i - \bar{z}) (\beta_0 + \beta_1 x_i + u_i)}{\frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})} = \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^n (z_i - \bar{z}) u_i}{\frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})} \rightarrow \beta_1 + \frac{\text{cov}(z_i, u_i)}{\text{cov}(z_i, x_i)} = \beta_1$$

Hvor vi bruger både antagelse (1) og (2) fra forrige slide. IV-estimatoren er derfor **konsistent** under disse antagelser. Estimatoren er dog **ikke unbiased** da

$$E(\widehat{\beta}_1^{IV} | z_i, x_i) = \beta_1 + E\left(\frac{\sum_{i=1}^n (z_i - \bar{z}) u_i}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})} \middle| z_i, x_i\right) = \beta_1 + \frac{\sum_{i=1}^n (z_i - \bar{z}) E(u_i | z_i, x_i)}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})} \neq \beta_1$$

Da $E(u_i | z_i, x_i) \neq 0$

1) Introduktion til 2SLS

- IV-estimatoren kan alternativt implementeres gennem 2 trin. Denne metode kaldes Two Stage Least Squares (2SLS) estimation.

Trin 1

- Først regresserer vi den endogene variabel på instrumentet, og vi finder den prædikterede værdi:

$$x_i = \alpha_0 + \alpha_1 z_i + v_i \rightarrow \hat{x}_i = \hat{\alpha}_0 + \hat{\alpha}_1 z_i$$

- Den prædikterede variabel ($\hat{x}_i$) indeholder al den variation i x_i der kan forklares af instrumentet z_i (eksogen variation). Vi ved desuden at $\hat{x}_i$ er ukorreleret med residualet $\hat{v}_i$ som indeholder den resterende endogene variation (mekanisk egenskab for OLS).

Trin 2

- Herefter regresserer vi den afhængige variabel på den prædikterede værdi:

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \eta_i$$

- Vi erstatter altså x_i med $\hat{x}_i$. Da $\hat{x}_i$ består af eksogen variation, er den ukorreleret med fejllidet. $\widehat{\beta}_1^{2sls}$ estimatoren er derfor konsistent.
- 2SLS og IV er sammenfaldene** når vi har **lige mange instrumenter som endogene variable** fx én endogen variabel og ét instrument som i eksemplerne ovenfor (just identified case).

1) Introduktion til 2SLS

- En fordel ved at benytte 2SLS er at metoden er nem at generaliserer, fx til tilfælde hvor vi har **flere instrumenter** end endogene variable (overidentified case). Antag fx igen at vi er interesserede i følgende model:

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad (A)$$

- Hvis vi har **flere instrumenter** for x_i , fx z_{i1} og z_{i2} , kan vi nemt bruge 2SLS:

$$x_{i1} = \alpha_0 + \alpha_1 z_{i1} + \alpha_2 z_{i2} + v_i \quad (First\ stage)$$

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \eta_i \quad (Second\ stage)$$

Huskeregler:

- Hvis vi har flere eksogene variable ($x_{i2}, x_{i3}, \dots, x_{ik}$) i (A) – dvs. variable vi ikke behøver at instrumentere – er det vigtigt at inkludere disse i first stage (ellers er 2SLS estimatoren generelt inkonsistent).
- Hvis 2SLS udføres manuelt (dvs. hvis de to trin udføres manuelt) vil de rapportede standardfejl være forkerte.

Begge ting tages der automatisk højde for når Statas indbyggede kommando bruges (næste slide)

1) Introduktion til 2SLS

Stata-kommando (automatisk)

- I Stata udføres 2SLS ved at skrive følgende:

ivregress 2sls y (x1 = z1) x2

- Når vi har et instrument (z1) og:

ivregress 2sls y (x1 = z1 z2) x2

- Når vi har to (eller flere) instrumenter (z1 og z2).

Eksempler:

```
ivregress 2sls workm (morekids=samesex) boy1stkid boy2ndkid agem agem1stkid black hisp other,
```

```
ivregress 2sls workm (morekids=twoboys twogirls) boy1stkid agem agem1stkid black hisp other
```


1) Introduktion til PS8

- I PS8 skal vi analysere hvordan arbejdsudbuddet for kvinder påvirkes af om de har børn eller ej.
- En naiv løsning vil være at estimere en model hvor den afhængige variabel fx angiver antal timer i arbejde, og den uafhængige variabel er en variabel der angiver antal børn kvinden har.
- Det tager dog ikke højde for at beslutningen om at få børn er endogen. Det kan være kvinder med børn er systematisk forskellige på en lang række relevante faktorer fra kvinder der vælger ikke at få børn
- Hvis dette er tilfældet bryder det med MLR.4 og vi kan derfor ikke troværdigt estimere alt-andet-lige effekter. En ændring i antal børn leder til ændringer i de uobserverede faktorer (fejlledet)
- Vi forsøger at løse dette problem ved brug af instrumental variable (IV)

1) Introduktion til PS8

- Helt konkret ønsker vi i PS8 at estimere følgende model:

$$y_i = \beta_0 + \beta_1 morekids_i + \beta_2 boy1stkid_i + \beta_3 boy2ndkid_i + \beta_4 agem_i + \beta_5 agem1stkid_i + \beta_6 black_i + \beta_7 hisp_i + \beta_8 other_i + u_i \quad (1)$$

- Vi bruger flere mål for hvorvidt kvinden arbejder og hendes indkomst:

$$y_i = \{workm_i, weekm_i, hourm_i, lincm_i, \log(faminc_i)\}$$

- Fortolkningen af parametrene afhænger af hvilken forklarende variabel der benyttes.
 - Fx er $y_i = workm_i$ en dummy der angiver om moderen arbejder eller ej. I det tilfælde er modellen en linear probability model (LPM) der angiver ændringen i sandsynligheden for at moderen er i arbejde ($E[y|x] = P[y = 1|x]$)
- *Morekids* er en dummy der angiver om en kvinde har mere end to børn. Denne er sandsynligvis endogen (antallet af børn er generelt ikke tilfældigt)
 - I PS8 instrumenterer vi denne variabel med en variabel der angiver om de to første børn har samme køn. I det tilfælde kan man argumentere for at forældre generelt er mere tilbøjelige til at få et tredje barn (ideen er at forældre generelt ønsker børn af begge køn)
- Derudover inkluderer vi en række kontroller (se problem set for beskrivelser)

2) Gruppearbejde - Diskussion

- De næste 20 minutter:

Discuss the following questions in groups.

- a. What is the interpretation of β_1 ? What is the expected sign of β_1 ?
- b. What is the interpretation of β_4 and β_5 ?
- c. Provide an example of why and how *morekids* may be an endogenous variable, and discuss how this might affect the properties of the OLS estimator.
- d. Write up the conditions for applying a variable z as an instrument for $morekids_i$. Discuss whether the sex mix of the first two children is a plausible instrument for having a third child.
- e. Write up the IV estimator.
- f. Discuss how the Z matrix of instruments will look like in this case.

- OBS: Brug for hjælp? Skriv Jeres gruppenummer i Padlet (link i chatten)

3) Gruppearbejde - Opsamling

a. What is the interpretation of β_1 ? What is the expected sign of β_1 ?

3) Gruppearbejde - Opsamling

a. What is the interpretation of β_1 ? What is the expected sign of β_1 ?

- Modellen er givet ved:

$$y_i = \beta_0 + \beta_1 \text{morekids}_i + \beta_2 \text{boy1stkid}_i + \beta_3 \text{boy2ndkid}_i + \beta_4 \text{agem}_i + \beta_5 \text{agem1stkid}_i \\ + \beta_6 \text{black}_i + \beta_7 \text{hisp}_i + \beta_8 \text{other}_i + u_i \quad (1)$$

- Hvor

$$y_i = \{\text{workm}_i, \text{weekm}_i, \text{hourm}_i, \text{lincm}_i, \log(\text{faminc}_i)\}$$

- Morekids er en dummy der angiver om moderen har mere end to børn (mødre med to børn er referencekategorien). β_1 angiver derfor den forventede effekt på y_i ved at få et tredje barn (når de inkluderede faktorer holdes fast).
- Alt efter formen på den afhængige variabel har vi at:
 - Level-Level (weekm/hourm/lincm): Den forventede ændring i antal timer / antal uger / løn arbejdet er β_2
 - Log-Level (log(faminc)): Den forventede ændring i familiens indkomst er (approx.) $100 \cdot \beta_2 \%$
 - LPM (workm): Den forventede ændring i sandsynligheden for at moderen er i arbejde er $100 \cdot \beta_2$ pp.
- Det må forventes at β_1 er negativ da flere børn kræver mere pasning og dermed tid uden for arbejdsmarkedet.

3) Gruppearbejde - Opsamling

b. What is the interpretation of β_4 and β_5 ?

3) Gruppearbejde - Opsamling

b. What is the interpretation of β_4 and β_5 ?

- Da $agem_i$ angiver moderens alder tolkes β_4 som den forventede ændring i y_i når moderens alder stiger med et år (når de resterende faktorer holdes fast).
- Da $agem1stkid_i$ angiver moderens alder ved første fødsel tolkes β_5 som den forventede ændring i y_i når alderen ved første fødsel stiger med et år (når de resterende faktorer holdes fast *herunder alder*).

3) Gruppearbejde - Opsamling

- c. Provide an example of why and how *morekids* may be an endogenous variable, and discuss how this might affect the properties of the OLS estimator.

3) Gruppearbejde - Opsamling

c. Provide an example of why and how *morekids* may be an endogenous variable, and discuss how this might affect the properties of the OLS estimator.

- Folk vælger (for det meste) selv hvor mange børn de får og hvor meget de arbejder.
- Hvis der er en tendens til, at folk som vælger at få mange børn, er systematisk forskellige fra andre på en eller flere uobserverbare faktorer som påvirker arbejdsudbuddet (fx præferencer for fritid og arbejde, samt familiestørrelse) vil MLR.4 være brudt ($E(u|X) \neq 0$)
- I det tilfælde er antallet af børn ikke eksogen.
- Fx kan man forestille sig at præferencer for familiestørrelse er:
 - i) Positivt korreleret med antal børn (x_i)
 - ii) Negativt korreleret med hvor meget der arbejdes (y_i)
- Dette ville føre til et negativ bias i OLS estimatoren.

3) Gruppearbejde - Opsamling

- d. Write up the conditions for applying a variable z as an instrument for $morekids_i$. Discuss whether the sex mix of the first two children is a plausible instrument for having a third child.

3) Gruppearbejde - Opsamling

- d. Write up the conditions for applying a variable z as an instrument for $morekids_i$. Discuss whether the sex mix of the first two children is a plausible instrument for having a third child.

- For at et instrument (z_i) er validt i denne sammenhæng skal det være **eksogent**:

$$cov(z_i, u_i) = 0$$

- Det betyder at instrumentet ikke har en direkte effekt på arbejdsudbud og ikke selv er endogent (korrelerer med udeladte faktorer). Denne antagelse kan ikke testes og skal derfor argumenteres for teoretisk.
- Derudover skal instrumentet være **relevant**:

$$cov(z_i, morekids_i) \neq 0$$

- Med andre ord skal instrumentet korrelerer/forklare noget af variationen i hvorvidt mødre får flere end to børn. Denne antagelse kan undersøges ved at regressere $morekids_i$ på z_i og teste om parameteren er signifikant forskellig fra nul.
- Hvis begge antagelser er overholdt er instrumentet validt og IV-estimatoren er konsistent.

3) Gruppearbejde - Opsamling

- d. Write up the conditions for applying a variable z as an instrument for $morekids_i$. Discuss whether the sex mix of the first two children is a plausible instrument for having a third child.
- Man kan argumentere for at mixed af køn ($samesex_i$) er et validt instrument.
 - Argumentation for at det er eksogent [$cov(samesex_i, u_i) = 0$]:
Hvis kønnet på de to første børn er så godt som tilfældigt vil $samesex_i$ være ukorreleret alle ikke-inkluderede faktorer i modellen, og dermed ukorreleret med fejlleddet. Desuden virker det sandsynlig at $samesex_i$ ikke har sin egen direkte effekt på arbejdsudbuddet.
 - Argumentation for at det er relevant [$cov(samesex_i, morekids_i) \neq 0$]:
Hvis folk har præferencer for at have børn af begge køn og fortsætter med at få børn indtil de har opnået det vil $samesex_i$ kunne prædiktere hvorvidt folk får flere end to børn.

3) Gruppearbejde - Opsamling

e. Write up the IV estimator.

3) Gruppearbejde - Opsamling

e. Write up the IV estimator.

- Antag at vi har et validt instrument z_i og at vi estimerer følgende model med IV-estimation:

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

- IV-estimatoren findes ved at løse følgende moment-betingelser:

$$\begin{aligned} E(u_i) &= E(y_i - \beta_0 - \beta_1 x_i) = 0 \\ E(z_i u_i) &= E(z_i (y_i - \beta_0 - \beta_1 x_i)) = 0 \end{aligned}$$

- Løse vi de tilsvarende sample-moments kan vi finde at:

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

- Her kan det hurtigt ses at hvis x_i er eksogen og dermed kan bruges som instrument for sig selv så vil $\widehat{\beta}_1^{IV} = \widehat{\beta}_1^{OLS}$.
- Bruger vi matrix-notation har vi:

$$\widehat{\beta}^{IV} = (Z'X)^{-1}Z'y$$

3) Gruppearbejde - Opsamling

f. Discuss how the Z matrix of instruments will look like in this case.

3) Gruppearbejde - Opsamling

f. Discuss how the Z matrix of instruments will look like in this case.

- I den simple model på forrige slide er X og Z matricerne givet ved:

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad Z = \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{pmatrix}$$

- I modellen hvor vi har $k + 1$ variable hvoraf x_k er endogen og z er et validt instrument har vi:

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,k-1} & x_{1,k} \\ 1 & x_{2,1} & \cdots & x_{2,k-1} & x_{2,k} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k-1} & x_{n,k} \end{pmatrix} \quad Z = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,k-1} & z_1 \\ 1 & x_{2,1} & \cdots & x_{2,k-1} & z_2 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k-1} & z_n \end{pmatrix}$$

- De eskogene variable i X (dvs. $x_1, x_2, \dots, x_{k-1}$) fungerer altså som instrumenter for dem selv, mens vi inkluderer z som instrument for x_k . I vores tilfælde er $z = \text{same_sex}$ og $x_k = \text{more_kids}$.
- Har vi flere endogene variable og/eller instrumenter ekskluderes de endogene variable fra Z og vi tilføjer ekstra kolonner med instrumenter.

4) Gruppearbejde – Stataøvelser

Arbejd på Stataøvelserne indtil 12:30

Hints:

Opg. 1: Brug fx "bysort samesex: su"

Opg. 3: I Stata udføres IV/2SLS-estimation ved:

```
ivregress 2sls workm (morekids=samesex) boy1stkid boy2ndkid agem agem1stkid black hisp other
```

Opg. 4: Start ud med at implementere estimatoren for én af de afhængige variable fx $y_i = workm_i$. Bemærk desuden at $Z'X$ ikke er en symmetrisk matrix. Brug derfor "qrinv($Z'X$)" fremfor "invsym($Z'X$)" til at invertere matricer når i udregner:

$$\hat{\beta}^{IV} = (Z'X)^{-1}Z'y$$

Opg. 6: Se hvordan i tester for overidentifying restrictions på s. 518 i Wooldridge.

OBS: Brug for hjælp? Skriv Jeres gruppenummer i Padlet (link i chatten)

5) Opsamling – Stataøvelser

1. Make a table that compares mothers with *samesex* = 1 to mothers with *samesex* = 0 in terms of relevant characteristics. Comment briefly on the table.

- Datasættet indeholder i alt 375000 observationer af kvinder som alle har mindst to børn. Disse er nogenlunde ligeligt fordelt på tværs af mødre hvor de to første børn har samme køn (189062) og mødre hvor de to første børn har forskelligt køn (185983)
- Andelen af kvinder med flere end to børn er 40,2% for mødre når de to førstefødte er samme køn. For kvinder hvor de to førstefødte er forskellige køn er det 33,8% der har flere end to børn.
- Kvinder med to børn af samme køn ser derfor ud til at få flere børn i gennemsnit.
- De to grupper er derudover tæt på identiske på tværs af vores mål for arbejdsudbud (fem øverste rækker) og på tværs af de resterende kontrolvariable.
- Alt i alt giver det indikationer af at *samesex* kan benyttes som instrument for *morekids*

```
bysort samesex: su
quietly: bysort samesex: outreg2 using Xoutput2, sum(log) tex(frag) eqkeep(N mean) replace
```

	(1)	(2)	(3)	(4)
	samesex 0		samesex 1	
VARIABLES	N	mean	N	mean
workm	185,938	0.670	189,062	0.665
weekm	185,938	26.54	189,062	26.24
hourm	185,938	22.81	189,062	22.58
lincm	185,938	7.365	189,062	7.283
faminc	185,938	34.437	189,062	34.340
agem	185,938	30.43	189,062	30.44
agem1stkid	185,938	21.81	189,062	21.80
nkids	185,938	2.453	189,062	2.537
morekids	185,938	0.338	189,062	0.402
boy1stkid	185,938	0.502	189,062	0.523
boy2ndkid	185,938	0.498	189,062	0.523
twoboys	185,938	0	189,062	0.523
twogirls	185,938	0	189,062	0.477
black	185,938	0.108	189,062	0.110
hisp	185,938	0.104	189,062	0.103
other	185,938	0.0359	189,062	0.0352
lowedu	185,938	0.190	189,062	0.189
hsedu	185,938	0.385	189,062	0.384
highedu	185,938	0.426	189,062	0.427
group	185,938	5.497	189,062	5.503

5) Opsamling – Stataøvelser

2. Estimate the parameters of model (1) by OLS

- På tværs af alle mål for arbejdsmarkedstilknytning finder vi en negativ sammenhæng med *morekids*.
- Fx ser vi fra kolonne (1) at det forventede fald i sandsynligheden for en kvinde er i arbejde er 15,2 pp. lavere hvis hun har flere end to børn (når de resterende kontrolvariablene holdes fast).
- Resultater stemmer overens med vores forventning om at flere børn sænker arbejdsudbuddet.
- Men kan også skyldes at estimatorerne er biased fordi beslutningen om at få flere børn er endogen (som omtalt tidligere).

VARIABLES	(1) workm	(2) weekm	(3) hourm	(4) lincm	(5) lfaminc
morekids	-0.152*** (0.00165)	-8.624*** (0.0773)	-6.700*** (0.0656)	-3,135*** (33.06)	-0.0910*** (0.00302)
boy1stkid	-0.00243 (0.00151)	-0.108 (0.0724)	-0.140** (0.0610)	-28.35 (32.33)	0.000414 (0.00274)
boy2ndkid	-0.000492 (0.00151)	0.0393 (0.0724)	-0.0479 (0.0610)	-0.893 (32.31)	0.00378 (0.00274)
agem	0.0201*** (0.000244)	1.358*** (0.0114)	0.822*** (0.00994)	499.4*** (4.676)	0.0397*** (0.000470)
agem1stkid	-0.0139*** (0.000251)	-0.733*** (0.0122)	-0.799*** (0.0102)	41.51*** (6.198)	0.0566*** (0.000455)
black	0.00266 (0.00246)	0.486*** (0.119)	2.435*** (0.103)	1,525*** (52.51)	-0.660*** (0.00597)
hisp	-0.0989*** (0.00264)	-4.197*** (0.121)	-1.407*** (0.108)	-390.4*** (48.80)	-0.292*** (0.00510)
other	-0.0602*** (0.00424)	-2.080*** (0.200)	0.600*** (0.178)	837.3*** (101.4)	-0.118*** (0.00814)
Constant	0.428*** (0.00746)	4.718*** (0.342)	17.55*** (0.298)	-7,759*** (151.6)	7.834*** (0.0140)
Observations	375,000	375,000	375,000	375,000	370,761
R-squared	0.040	0.061	0.043	0.053	0.186

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

5) Opsamling – Stataøvelser

3. Estimate model (1) by IV using *samesex* as an instrument for *morekids*.

- Før vi benytter IV-estimation bør vi altid diskutere om de to antagelser der gør at instrumentet er validt er overholdt

Eksogent

- Vi kan her argumentere for at hvis kønnet på de to første børn er så godt som tilfældigt vil $samesex_i$ være ukorreleret med alle ikke-inkluderede faktorer i modellen, og dermed ukorreleret med fejllidet. Desuden virker det sandsynlig at $samesex_i$ ikke har sin egen direkte effekt på arbejdsudbuddet når vi kontrollerer for *morekids*. Vi tror derfor at antagelsen $cov(samsex_i, u_i) = 0$ er overholdt.

Relevant

- Vi kan derudover direkte teste om instrumentet er relevant ved at regressere *morekids* på *samesex* og de andre kontroller (dvs. den regression der svarer til vores first-stage i 2SLS). Gør vi dette finder vi at koefficienten er både af praktisk og økonomisk signifikans. Med andre ord er $cov(samsex_i, morekids_i) \neq 0$. Dette stemmer overens med ideen om at folk har præferencer for at have børn af begge køn og dermed fortsætter med at få børn indtil de har opnået det.

VARIABLES	(1) morekids
samesex	0.0627*** (0.00151)
boy1stkid	-0.00759*** (0.00151)
boy2ndkid	-0.000359 (0.00151)
agem	0.0247*** (0.000240)
agem1stkid	-0.0388*** (0.000232)
black	0.0561*** (0.00258)
hisp	0.109*** (0.00257)
other	0.0470*** (0.00404)
Constant	0.418*** (0.00714)
Observations	375,000
R-squared	0.081
Robust standard errors in parentheses	
*** p<0.01, ** p<0.05, * p<0.1	

5) Opsamling – Stataøvelser

3. Estimate model (1) by IV using *samesex* as an instrument for *morekids*.

- Vi finder igen - på tværs af alle mål for arbejdsmarkedstilknytning - en negativ sammenhæng med *morekids*.
- Effekten er dog mindre (absolutte værdier er mindre) hvilket tyder på at OLS estimatorerne var nedad biased.
- Fx ser vi fra kolonne (1) at det forventede fald i sandsynligheden for en kvinde er i arbejde nu er 9 pp. lavere hvis hun har flere end to børn (når de resterende kontrolvariablene holdes fast).
- Resultater stemmer overens med vores forventning om at flere børn sænker arbejdsudbuddet.

VARIABLES	(1) workm	(2) weekm	(3) hourm	(4) lincm	(5) lfaminc
morekids	-0.0904*** (0.0241)	-5.216*** (1.159)	-4.019*** (0.976)	-1.426*** (517.5)	-0.0841* (0.0436)
boy1stkid	-0.00204 (0.00152)	-0.0869 (0.0730)	-0.124** (0.0614)	-17.70 (32.62)	0.000459 (0.00275)
boy2ndkid	-0.000565 (0.00151)	0.0352 (0.0726)	-0.0511 (0.0611)	-2.925 (32.42)	0.00377 (0.00274)
agem	0.0186*** (0.000644)	1.274*** (0.0308)	0.756*** (0.0261)	457.1*** (13.60)	0.0395*** (0.00118)
agem1stkid	-0.0116*** (0.000968)	-0.600*** (0.0465)	-0.695*** (0.0392)	107.8*** (20.99)	0.0569*** (0.00174)
black	-0.000825 (0.00282)	0.293** (0.136)	2.283*** (0.117)	1,428*** (60.27)	-0.661*** (0.00646)
hisp	-0.106*** (0.00372)	-4.567*** (0.175)	-1.698*** (0.151)	-576.3*** (74.46)	-0.293*** (0.00691)
other	-0.0631*** (0.00439)	-2.240*** (0.208)	0.475*** (0.184)	757.5*** (104.6)	-0.118*** (0.00839)
Constant	0.401*** (0.0131)	3.191*** (0.620)	16.35*** (0.528)	-8,524*** (276.9)	7.831*** (0.0238)
Observations	375,000	375,000	375,000	375,000	370,761
R-squared	0.037	0.056	0.039	0.047	0.186

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

5) Opsamling – Stataøvelser

4. Calculate the IV estimates using `mata`

- For at simplificere beregningerne finder vi IV estimaterne for én afhængig variabel af gangen. I eksemplet til højre er $y_i = workm_i$
- Syntaks er som gennemgået i PS5, med den ene undtagelse af vi benytter `qrinv()` til at invertere matrixen $Z'X$ da denne ikke er symmetrisk.
- Den generelle formular for IV-estimatoren er hvis man har flere instrumenter end endogene variable (overidentified).
- Hvis Z og X har samme dimensioner, dvs. hvis vi har lige mange instrumenter som endogene variable (just identified), er $Z'X$ en kvadratrisk matrix der kan inverteres. Det kan bruges til forkorte det generelle udtryk til den simple formel (se IV-note).
- Hvis vi bruger den generelle formel i just identified case finder vi samme resultat som ved brug af den simple formel.
- Vi ser at estimaterne er identiske med dem vi fandt da vi benyttede Statas automatiske kommando.

```

mata
// INPUT DATA //
y = st_data(., "workm")
// y = st_data(., "weekm") //
// y = st_data(., "hourm") //
// y = st_data(., "lincm") //
// y = st_data(., "lfaminc") //

morekids = st_data(., "morekids")
boy1stkid = st_data(., "boy1stkid")
boy2ndkid = st_data(., "boy2ndkid")
agem = st_data(., "agem")
agem1stkid = st_data(., "agem1stkid")
black = st_data(., "black")
hisp = st_data(., "hisp")
other = st_data(., "other")

z = st_data(., "samesex")

// DEFINE CONSTANT AND X AND Z MATRICES //
cons = J(rows(morekids), 1, 1)
X = (cons, morekids, boy1stkid, boy2ndkid, agem, agem1stkid, black, hisp, other)
Z = (cons, z, boy1stkid, boy2ndkid, agem, agem1stkid, black, hisp, other)

// CALCULATE OLS ESTIMATES //
beta_ols = (invsym(X'*X))*(X'*y)

// CALCULATE IV ESTIMATES //
*Just-identified case
beta_iv_simple = (qrinv(Z'*X)*(Z'*y))
*General Case i.e. also works in over-identified case
beta_iv_general = qrinv((X'*Z)*invsym(Z'*Z)*(Z'*X))*(X'*Z)*invsym(Z'*Z)*(Z'*y)

// PRINT RESULTS //
(beta_ols, beta_iv_simple, beta_iv_general)
end

```

	1	2	3
1	.4282369012	.4005784615	.4005784615
2	-.1521075231	-.0903599595	-.0903599595
3	-.0024268317	-.0020419613	-.0020419613
4	-.0004916575	-.0005650923	-.0005650923
5	.02014292	.0186161948	.0186161948
6	-.0139475498	-.0115506842	-.0115506842
7	.0026573484	-.000825318	-.000825318
8	-.098880688	-.1055971596	-.1055971596
9	-.0602193291	-.0631056977	-.0631056977

5) Opsamling – Stataøvelser

5. An advantage of the *samesex*-instrument is that it can be decomposed into two separate instruments: *twoboys* and *twogirls*. Estimate model (1) without the *boy2ndkid_i* dummy variable using *twoboys* and *twogirls* as instruments for *morekids*.

- Vi kan splitte *samesex_i* op i to separate instrumenter.
- Hvis kønnet på de to første børn er så godt som tilfældigt, gælder det altså også at hvorvidt de to første børn er to drenge eller to piger er tilfældigt. Instrumenterne er derfor hver især **eksogene**.
- Vi finder en p-værdi på 0.000 når vi benytter en F-test for samlet signifikans (se også tabel til højre). Når vi splitter instrumentet op i to er de altså stadig samlet set **relevante**.
- Vi kan derfor estimere modellen med 2SLS hvor vi benytter to separate instrumenter. For $y_i = \text{weekm}$ gøres det ved:

```
ivregress 2sls workm (morekids=twoboys twogirls) boy1stkid agem agem1stkid black hisp other, robust
```

- Fordelen ved at benytte to instrumenter er at de potentielt kan forklare en større del af (den eksogene) variation i den endogene variabel, og dermed reducere standardfejlene for 2SLS estimatoren (vi kan danne flere lineære kombinationer med to instrumenter sammenlignet med ét og kan derfor potentielt forklare en større del af den eksogene variation i den *morekids*)
- Bemærk at vi udelader nu *boy2ndkid* som kontrolvariabel for at undgå perfekt multikollinearitet. Da der er 4 måder at få to børn på (dreng/dreng, dreng/pige, pige/dreng, pige/pige) kan vi ikke inkludere 4 dummier samt et konstantled. Drop konstantled eller en dummy.

VARIABLES	(1) morekids
twoboys	0.0623*** (0.00211)
twogirls	0.0630*** (0.00217)
boy1stkid	-0.00723*** (0.00211)
agem	0.0247*** (0.000240)
agem1stkid	-0.0388*** (0.000232)
black	0.0561*** (0.00258)
hisp	0.109*** (0.00257)
other	0.0470*** (0.00404)
Constant	0.418*** (0.00714)
Observations	375,000
R-squared	0.081

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

5) Opsamling – Stataøvelser

5. An advantage of the *samesex*-instrument is that it can be decomposed into two separate instruments: *twoboys* and *twogirls*. Estimate model (1) without the *boy2ndkid_i* dummy variable using *twoboys* and *twogirls* as instruments for *morekids*.

- Estimer og standardfejl er stort set identiske med tidligere hvor vi kun benyttede *samesex* som instrument.
- Indikerer at vi forklarer samme (eskegene) del af variationen i *morekids* med to separate instrumenter (*twogirls*, *twoboys*) som ved et enkelt instrument (*samesex*).
- Med andre ord, mødre med to drenge er hverken mere eller mindre tilbøjelige til at få et tredje barn, ift. mødre med to piger (se også koefficienterne på *twogirls* og *twoboys* i regression på forrige slide – stort set ens). Vi vinder derfor ikke noget på at splitte *samesex* op i to variable i det her tilfælde.

VARIABLES	(1) workm	(2) weekm	(3) hourm	(4) lincm	(5) lfaminc
morekids	-0.0905*** (0.0241)	-5.206*** (1.159)	-4.034*** (0.976)	-1,427*** (517.4)	-0.0830* (0.0436)
boy1stkid	-0.00205 (0.00152)	-0.0866 (0.0730)	-0.124** (0.0614)	-17.73 (32.62)	0.000495 (0.00275)
agem	0.0186*** (0.000644)	1.274*** (0.0308)	0.756*** (0.0261)	457.1*** (13.60)	0.0395*** (0.00118)
agem1stkid	-0.0116*** (0.000968)	-0.600*** (0.0465)	-0.696*** (0.0392)	107.8*** (20.99)	0.0569*** (0.00174)
black	-0.000812 (0.00282)	0.293** (0.136)	2.285*** (0.117)	1,428*** (60.26)	-0.661*** (0.00646)
hisp	-0.106*** (0.00372)	-4.569*** (0.175)	-1.696*** (0.151)	-576.2*** (74.44)	-0.293*** (0.00691)
other	-0.0631*** (0.00439)	-2.240*** (0.208)	0.476*** (0.184)	757.5*** (104.6)	-0.118*** (0.00839)
Constant	0.400*** (0.0131)	3.204*** (0.620)	16.33*** (0.528)	-8,525*** (276.6)	7.832*** (0.0238)
Observations	375,000	375,000	375,000	375,000	370,761
R-squared	0.037	0.056	0.039	0.047	0.186

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

5) Opsamling – Stataøvelser

6. Explain what is meant by overidentifying restrictions and perform a test for it.

- Testen for overidentifying restrictions (OI) er en (partiel) metode til at teste validiteten af instrumenterne når vi har flere instrumenter end endogene variabel (overidentified case).
- Ideen er at hvis alle vores instrumenterne er valide, og vi bruger IV-estimation med dem hver for sig, så bør estimerne være tæt på ens da alle estimatorerne er konsistente i det tilfælde. Forskelle skyldes blot sampling error.
- Vi kan derfor teste om estimerne er signifikant forskellige fra hinanden. Dette kaldes en **Hausmann** test. Finder vi at estimerne er signifikant forskellige fra hinanden tyder det på at mindst et instrument ikke er validt. Vi ved dog ikke hvilket.
- Testen har **en væsentlig mangel** idet vi sagtens kan få to estimer der er tæt på hinanden, på trods af at ingen af instrumenterne valide. Hvis instrumenterne ikke er valide er IV-estimatorerne ikke konsistente, men intet udelukker at de tilfældigvis producere estimer der er tæt på at være ens.
- Det er derfor stadig vigtigt at give teoretiske argumenter for at instrumenter er valide.
- En anden ulempe ved blot at sammenligne estimerne som det gøres i Hausmann testen er at det hurtigt løber op i mange sammenligninger på kryds og tværs hvis vi har flere end to instrumenter.
- Et **alternativ** til Hausmann testen er derfor at bruge residualerne fra 2SLS regressionen. Det er denne fremgangsmåde vi benytter (se næste slide).

5) Opsamling – Stataøvelser

6. Explain what is meant by overidentifying restrictions and perform a test for it.

- OI testen implementeres på følgende vis:

1. Først estimeres modellen med 2SLS, og vi gemmer residualerne fra denne fx:

```
ivregress 2sls workm (morekids=twoboys twogirls) boy1stkid agem agem1stkid black hisp other, robust  
predict workm_hat, residual
```

2. Herefter regresserer vi residualerne på alle instrumenterne og de eksogene variable:

```
regress workm_uhat twoboys twogirls boy1stkid agem agem1stkid black hisp other
```

Ideen er at hvis alle instrumenterne er valide bør disse være ukorrelerede med residualerne fra 2SLS. Testen implementeres som en LM-test givet ved:

$$n \cdot R_{uhat}^2 \sim \chi_q^2$$

Hvor R_{uhat}^2 er forklaringsgraden i regressionen i trin 2. og q angiver antallet af overidentifying restrictions (én i vores tilfælde). Under nulhypotesen om at alle instrumenterne er valide samt antagelse 2SLS.1-5 (se appendix til kap. 15 i Wooldridge) følger test-statistikken er chi-i-anden fordeling med q frihedsgrader. I Stata finder vi LM-statistikken og den kritiske værdi ved:

```
display e(N)*e(r2)  
display invchi2tail(1,0.05)
```

Forkaster vi H_0 er der tegn på at et eller flere af instrumenterne ikke er valide.

5) Opsamling – Stataøvelser

6. Explain what is meant by overidentifying restrictions and perform a test for it.

- Vi kan ikke forkaste nulhypotesen for nogle af vores mål arbejdsudbud.
- Vi konkluderer derfor at de to instrumenter nok er valide (men vi kan ikke være sikre)
- Stadig vigtigt at vi argumenterer teoretisk.

```
OI test for workm
```

```
LM test statistic: .13983442 compared to a critical value of 3.8414588
```

```
OI test for weekm
```

```
LM test statistic: .23553332 compared to a critical value of 3.8414588
```

```
OI test for hourm
```

```
LM test statistic: .69785281 compared to a critical value of 3.8414588
```

```
OI test for lincm
```

```
LM test statistic: .00813691 compared to a critical value of 3.8414588
```

```
OI test for lfaminc
```

```
LM test statistic: 1.89124 compared to a critical value of 3.8414588
```

- Testen fungerer ikke hvis vi kun har ét instrument. I det tilfælde er $R_{\hat{u}}^2 = 0$ og vi vil aldrig forkaste H_0 .

Tak for i dag

- Sidste obligatoriske opgave (OO3) er blevet gjort tilgængelig i dag og skal afleveres 10 maj.
- Næste time er sat af til at arbejde på den.