

Økonometri I: Ugeseddel 8

Levi van Boekel

Uge 16, 2025

Opgave 1

1.1

Modellen er en lineær sandsynlighedsmodel, og β_1 er en dummy variabel, der angiver om moren har flere eller færre end to børn. Fortolkningen af β_1 er dermed forskellen i sandsynligheden for at arbejde, for mødre med mere end to børn ($morekids = 1$) og mødre med færre end to børn ($morekids = 0$), for fastholdte andre variable. Hvis β_1 betyder det f.eks., at mødre med flere end to børn har en $\beta_1 \cdot 100$ pct. større/mindre sandsynlighed for at arbejde, sammenlignet med kvinder med færre end to børn. Vi forventer jf. opgavebeskrivelsen, at $\beta_1 < 0$, fordi kvinder, der har en høj fertilitet har sværere ved at komme ind på arbejdsmarkedet.

1.2

Både β_4 og β_5 angiver den marginale effekt på sandsynligheden for at være i arbejde af at øge hhv. morens alder og morens alder ved første barn med et år, for fastholdte andre parametre.

1.3

Variable, der kan påvirke både $workm_i$ og $morekids_i$ og dermed gør $morekids_i$ til en endogen variabel er f.eks; (1) arv modtaget (familier med stor arv arbejder mindre og får flere børn?), (2) geografisk placering (nogle udsatte områder har lavere arbejdsudbud og er forbundet med højere fertilitet?) og (3) helbred (bedre helbred tillader flere børn og giver højere indkomst?). Hvis $morekids_i$ er en endogen variabel, betyder det, at **OLS ikke længere er middelret eller konsistent**. I tilfælde af den geografiske placering vil det give en negativ bias, fordi $cov(morekids_i, geo_i) > 0$ og den forventede effekt af geo er negativ.

1.4

For at en variabel, z , kan bruges som et godt instrument skal der gælde, at:

1. Instrumentet er ukorreleret med de uobserverbare faktorer (fejleddet). Denne antagelse kan ikke testes og er baseret på teori.

$$\text{Cov}(u, z) = 0$$

2. Instrumentet er korreleret med den endogene forklarende variabel, morekids_i . Denne antagelse kan testes.

$$\text{Cov}(\text{morekids}_i, z) \neq 0$$

Hvis vi forestiller os, at fejleddet består af ting som arv, geografisk placering og helbred, forekommer det rimeligt at antage, at den første betingelse er opfyldt ved at bruge samesize_i som instrument, da ingen af delene korrelerer med sandsynligheden for at få to af børn af samme køn (selv hvis fejleddet består af andre ting, er det rimeligt at antage ingen korrelation, da ssh. for at få to børn af samme køn ikke påvirkes af noget). Det forekommer ligeledes rimeligt at antage (og er motiveret i opgavebeskrivelsen), at sandsynligheden for at få flere børn er stigende, hvis man har to børn af samme køn, idet man antager, at husholdninger vil have børn af begge køn.

Opgave 2

2.1

Der er nedenfor vist beskrivende statistisk for kvinder med $\text{samesex} = 1$ og $\text{samesex} = 0$. Vi ser, at kvinderne er stort set ens på tværs af alle variable, med undtagelse af nkids og morekids . Konkret ser vi, at kvinder med to børn af samme køn i gennemsnit har ≈ 0.08 flere børn og $\approx 40\text{pct.}$ af kvinder i gruppen $\text{samesex} = 1$ har flere end to børn, sammenlignet med $\approx 33.8\text{pct.}$ i gruppen $\text{samesex} = 0$. Det tyder på, at der er en sammenhæng mellem samesex og morekids , hvilket netop er kravet til instrumentet.

Tabel 1: Beskrivende statistisk

	Mean ($\text{samesex} = 1$)	Mean ($\text{samesex} = 0$)
weekm	26.23948	26.5382
hourm	22.57571	22.80873
lincm	7282.569	7365.05
faminc	34340.2	34437.25
agem	30.44009	30.42597
agem1stkid	21.80262	21.81454
nkids	2.536946	2.452721
morekids	.4017306	.338473

Data: Udleveret

2.2

Vi formoder, at der skal bruges robuste standardfejl, da der oftest forekommer heteroskedasticitet i lineære sandsynlighedsmodeller. For god ordens skyld tester vi dog først med et BP-test, hvor nullhypotesen, at vi har en konstant varians. Teststørrelsen er $\chi^2(8)$ fordelt og i STATA får vi den tilhørende p-værdi til ≈ 0 , svarende til, at sandsynligheden for at få et mere ekstremt udfald er ca. 0 pct. Dermed afvises nullhypotesen til fordel for den alternative hypotese, svarende til, at modellen udviser heteroskedasticitet. OLS regressionen foretages dermed med robuste standardfejl. Resultatet af regressionen er rapporteret i første kolonne af tabel 2. Vi ser, at kvinder med flere end to børn (*morekids* = 1) har en 15.2 pct. mindre sandsynlighed for at være i arbejde, sammenlignet med kvinder med færre end to børn (*morekids* = 0), for fastholdte øvrige parametre. Det er i tråd med teorien fra opgaven. Vi husker dog, at estimatet muligvis har en negativ bias jf. diskussionen i 2.1 Vi ser desuden, at parameterestimatet er robust på et 0.001 signifikansniveau

2.3

Vi starter med at køre first stage regressionen og ser, at *p* værdien hørende til *samesex* (i regressionen på *morekids*) er 0, svarende til, at parameterestimatet er signifikant på et 5 pct. signifikansniveau. Resultaterne af 2SLS ses i anden kolonne af tabel 2. Vi ser, at β_1 nu er -0.0904 , svarende til, at vi ved OLS estimationen havde en negativ bias på ca 0.06. Kvinder med flere end to børn har dog stadigvæk en mindre sandsynlighed for at være i arbejde end kvinder med færre end to børn. Konkret betyder resultatet nemlig, at sandsynligheden for at være i arbejde er 9.04pct. mindre for kvinder, der har flere end to børn, sammenlignet kvinder med færre.

2.4

Resultaterne ses i tredje kolonne af tabel 2. Vi ser, parameterestimatene er stort set ens. Vi ser ligeledes i 'first stage' regressionen, at *twoboys* og *twogirls* begge er stærke instrumenter, da *p* værdien for begge variable er approksimativt lig 0. Vi kan ikke medtage *boy2ndkid*, fordi *twoboys* = (*boy1stkid* & *boy2ndkid*) og der derfor ville være et brud på kravet om perfekt multikollinearitet, hvis vi medtog både *twoboys* og *boy2ndkid*.

2.5

Hvis vi har flere instrumenter, end endogene variable, er modellen overidentificeret, dvs. $g > I$, hvor g er antallet af instrumenter og I er antallet af endogene variable. Det giver os mulighed for at teste g restriktioner. Intuitionen er:

1. Brug de første I instrumenter til at estimere modellen og beregne IV residualerne, $\hat{u}^{IV}$
2. Tjek om de sidste $g - I$ instrumenter er korrelerede med residualen, $\hat{u}^{IV}$

3. Hvis instrumenterne er valide, bør der ikke være nogen korrelation

Vi bruger fremgangsmåden som vist nedenfor: Fremgangsmåden er, at:

1. Vi estimerer modellen med IV, hvor alle instrumenter bruges. IV-residualerne $\hat{u}^{IV}$ gemmes med kommandoen **predict double eIV, resid**
2. Vi estimerer hjælperegressionen med OLS $\hat{u}^{IV} = Z\phi + v$, hvor Z indeholder alle eksogene variable og begge instrumenter
3. Nullhypotesen er, at $\mathcal{H}_0 : \phi = 0$, da dette svarer til, at instrumenterne er gyldige
4. Teststørrelsen bestemmes som $OI = nR^2 \approx 0.139852$ (R^2 fra hjælperegressionen)
5. Teststørrelsen er χ^2 fordelt med $g - I = 1$ frihedsgrader, dvs. $\chi^2(1)$. Det kritiske område er 3.84, og da OI er indenfor dette område, kan vi ikke afvise nullhypotesen om, at begge instrumenter tilsammen opfylder eksogenitetskravet (på et 5 pct. signifikansniveau).

2.6

Regressionsresultaterne er vedhæftet i tabel 3, hvor den første kolonne ved hver variabel angiver OLS regressionen.

Opgave 3

3.1

Vi starter med at udlede IV estimatoren i det simple tilfælde, hvor $y_i = \beta_0 + \beta_1 x_i + u_i$ og z er et relevant og validt instrument for x . I det første skridt tager vi kovariansen mellem z og elementerne i y , dvs:

$$\begin{aligned} \text{cov}(z, y) &= \text{cov}(\beta_0 + \beta_1 x + u, z) \\ \text{cov}(z, y) &= \beta_1 \text{cov}(x, z) + \underbrace{\text{cov}(u, z)}_0 \\ \beta_1 &= \frac{\text{cov}(z, y)}{\text{cov}(x, z)} \end{aligned}$$

Sample ækvivalenten til ovenstående er:

$$\beta_1 = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}$$

3.2

(First stage): Lad k være alle variable, $k - I$ være de variable, der har endogenitetsproblemer og g de overidentificerede instrumenter. Samlet set har vi da $k - I + g = h$ instrumenter, som vi i 'first stage' bruger til at regressere de k variable i X på de h instrumenter i Z altså:

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

OLS estimatoren er givet ved:

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

De prædikterede værdier er:

$$\begin{aligned} \underbrace{\hat{X}}_{n \times k} &= \underbrace{Z}_{n \times h} \cdot \underbrace{\hat{\Pi}}_{h \times k} \\ &= \underbrace{Z(Z'Z)^{-1}Z'}_{P_Z} X = P_Z X \end{aligned}$$

Hvor $Z(Z'Z)^{-1}Z' \equiv P_Z$ er en projektionsmatrix, hvor det gælder, at $P_Z = P_Z'$ og $P_Z P_Z = P_Z$.

(Second stage): Estimer nu følgende regressionsligning med OLS:

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

2SLS estimatoren er nu:

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

Hvor vi undervejs bruger, at $P_Z = P_Z'$ og $P_Z P_Z = P_Z$, samt $\hat{X} = X'P_Z'$. Når $g = I$ har vi eksakt identifikation og dermed et særtilfælde af ovenstående, hvor $\hat{\Pi}$ er en $k \times k$ matrix, hvorfor vi her kan skrive:

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

Hvor vi undervejs bruger, at $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$.

3.3

Instrumentmatricen i det generelle tilfælde, hvor vi har et instrument til en endogen variabel, og ellers blot eksogene variable (x), er givet ved:

$$Z = \begin{bmatrix} 1 & x_{1,1} & x_{k-1,1} & z_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n} & x_{k-1,n} & z_n \end{bmatrix}$$

3.4

Ja, vi får de samme resultater i MATA. Instrumentmatricen er:

$$Z = \begin{bmatrix} 1 & boy1st_{1,1} & agem_{2,1} & age1st_{3,1} & black_{4,1} & hisp_{5,1} & other_{6,1} & boy2nd_{7,1} & z_1(samesex) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & boy1st_{1,n} & agem_{2,n} & age1st_{3,n} & black_{4,n} & hisp_{5,n} & other_{6,n} & boy2nd_{7,n} & z_n(samesex) \end{bmatrix}$$

Tabel 2: Regressionsresultater for model (1) med robuste standardfejl

	(1) OLS, robust	(2) 2SLS, robust	(3) 2SLS, robust
	workm	workm	workm
morekids	-0.152*** (0.00165)	-0.0904*** (0.0241)	-0.0905*** (0.0241)
boy1stkid	-0.00243 (0.00151)	-0.00204 (0.00152)	-0.00205 (0.00152)
boy2ndkid	-0.000492 (0.00151)	-0.000565 (0.00151)	
agem	0.0201*** (0.000244)	0.0186*** (0.000644)	0.0186*** (0.000644)
agem1stkid	-0.0139*** (0.000251)	-0.0116*** (0.000968)	-0.0116*** (0.000968)
black	0.00266 (0.00246)	-0.000825 (0.00282)	-0.000812 (0.00282)
hisp	-0.0989*** (0.00264)	-0.106*** (0.00372)	-0.106*** (0.00372)
other	-0.0602*** (0.00424)	-0.0631*** (0.00439)	-0.0631*** (0.00439)
_cons	0.428*** (0.00746)	0.401*** (0.0131)	0.400*** (0.0131)
N	375000	375000	375000
R^2	0.040	0.037	0.037

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Tabel 3: Regressionsresultater for model (1.a) med robuste standardfejl

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	weekm	weekm	hourm	hourm	lincm	lincm	lfaminc	lfaminc
morekids	-8.624*** (0.0773)	-5.216*** (1.159)	-6.700*** (0.0656)	-4.019*** (0.976)	-3135.2*** (33.06)	-1426.4** (517.5)	-0.0910*** (0.00302)	-0.0841 (0.0436)
boy1stkid	-0.108 (0.0724)	-0.0869 (0.0730)	-0.140* (0.0610)	-0.124* (0.0614)	-28.35 (32.33)	-17.70 (32.62)	0.000414 (0.00274)	0.000459 (0.00275)
boy2ndkid	0.0393 (0.0724)	0.0352 (0.0726)	-0.0479 (0.0610)	-0.0511 (0.0611)	-0.893 (32.31)	-2.925 (32.42)	0.00378 (0.00274)	0.00377 (0.00274)
agem	1.358*** (0.0114)	1.274*** (0.0308)	0.822*** (0.00994)	0.756*** (0.0261)	499.4*** (4.676)	457.1*** (13.60)	0.0397*** (0.000470)	0.0395*** (0.00118)
agem1stkid	-0.733*** (0.0122)	-0.600*** (0.0465)	-0.799*** (0.0102)	-0.695*** (0.0392)	41.51*** (6.198)	107.8*** (20.99)	0.0566*** (0.000455)	0.0569*** (0.00174)
black	0.486*** (0.119)	0.293* (0.136)	2.435*** (0.103)	2.283*** (0.117)	1524.7*** (52.51)	1428.4*** (60.27)	-0.660*** (0.00597)	-0.661*** (0.00646)
hisp	-4.197*** (0.121)	-4.567*** (0.175)	-1.407*** (0.108)	-1.698*** (0.151)	-390.4*** (48.80)	-576.3*** (74.46)	-0.292*** (0.00510)	-0.293*** (0.00691)
other	-2.080*** (0.200)	-2.240*** (0.208)	0.600*** (0.178)	0.475** (0.184)	837.3*** (101.4)	757.5*** (104.6)	-0.118*** (0.00814)	-0.118*** (0.00839)
_cons	4.718*** (0.342)	3.191*** (0.620)	17.55*** (0.298)	16.35*** (0.528)	-7758.9*** (151.6)	-8524.3*** (276.9)	7.834*** (0.0140)	7.831*** (0.0238)
<i>N</i>	375000	375000	375000	375000	375000	375000	370761	370761
<i>R</i> ²	0.061	0.056	0.043	0.039	0.053	0.047	0.186	0.186

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$