Eksamen på Økonomistudiet sommer 2020

Økonometri I

Rettevejledning til Tag-hjem eksamen: 19. juni, 2020, kl.10.00-22.00

"Hvilken betydning har arbejdsløshed på kriminalitet"[1]

Opgaven går ud på at undersøge, hvordan arbejdsløshed påvirker kriminaliteten. Data til denne opgave stammmer fra Statistikbanken, Danmarks Statistik. Her er indhentet oplysninger om alle 98 kommuner i Danmark (Christiansø er ikke medtaget). Hver kommune er observeret to gange: i året 2012 og 2017. I analysen fokuseres på mænd i alderen 15-29 år. Datasættet består af 196observationer og 11 variable.

Tabel 1: Variable i 'groupdataX.dta'

Variabelnavn	Indhold
$Krimrate_{it}$	Andelen af 15-29 årige mænd som er dømte i år t i kommune i
$Ledighed\_1524_{it}$	Andelen af 15-24 årige mænd som er arbejdsløse i år t i kommune i
$Ledighed\_2529_{it}$	Andelen af 25-29 årige mænd som er arbejdsløse i år t i kommune i
$Studentrate_{it}$	Andelen af 15-29 årige mænd som er studerende i år t i kommune i
$Koen\_balance_{it}$	Andelen af 15-29 årige mænd i forhold til alle 15-29 årige i år t i kommune i
$Storby_{it}$	Dummy for KBH, FRB, Århus, Odense, Aalborg og Esbjerg
$pop\_mand1529_{it}$	Antallet af 15-29 årige i år t i kommune i
Dz_{it}	Forventet vækstrate i antal job i perioden 2012-2017 for kommune i
aar_{it}	År
kom_nr_{it}	Kommune nr i kommune i
$kommune_{it}$	Kommunenavn i kommune i

Noter: 'groupdataX.dta' indeholder konstruerede data og kan derfor ikke bruges til andre formål end at besvare denne eksamensopgave. Beregninger i denne opgave kan derfor afvige fra beregninger i originalmaterialet.

Variablen for den forventede vækst i antal job i kommunen er konstrueret ud fra de nationale jobværkstrater g_j i perioden 2012 til 2017 inden for hver sektor j. Vi benytter nu andelen af jobs i kommune i i 2012 i sektor j: τ_{ij}. Den forventede vækstrate i job i kommune i er:

$$Dz_{it} = \sum_{i=1}^{J} g_j \tau_{ij}.$$

[1] Denne eksamensopgave er inspireret af Fougère, Denis, et al. "Youth Unemployment and Crime in France." Journal of the European Economic Association, vol. 7, no. 5, 2009, pp. 909–938. I denne opgave anvendes andre data, så resultaterne kan afvige fra artiklen.

	2012		2017		Alle	
Variabelnavn	mean	sd	mean	sd	mean	sd
Andelen af dømte af 15-29 mænd	0.096	0.018	0.077	0.015	0.086	0.019
Ledighed for 15-24 mænd	0.026	0.008	0.013	0.004	0.019	0.009
Ledighed for 25-29 mænd	0.075	0.019	0.055	0.014	0.065	0.020
Andelen af studerende af 15-29 mænd	0.279	0.028	0.256	0.028	0.268	0.030
Andelen af mænd af 15-29 årige	0.523	0.021	0.526	0.018	0.525	0.019
No. obs	98		98		196	

Table 1: Deskriptiv tabel

Opgave 1 (20%)

a Tabel 1 viser en deskriptiv analyse af datasættet groupdata0.dta. I tabellen er gennemsnittet for følgende variable: $krimrate, ledighed\_1524, ledighed\_2529$, $studentrate$ og $koen\_balance$ angivet for årene 2012 og 2017. Tabellen viser, at ledigheden, andelen af studerende og kriminalitetsraten er faldet fra 2012 til 2017. Kriminalitetsraten er faldet med ca. 2 procent point, mens ungdomsledigheden er faldet med ca. 1.3 procent point. For andelen af mænd i de enkelte kommuner ser det ikke ud til, at der er store ændringer.

b Regressionsmodel (1):

$$\begin{aligned} krimrate_{it} &= \beta_0 + \beta_1 ledighed\_1524_{it} + \beta_2 ledighed\_2529_{it} + \beta_3 studentrate_{it} \\ &\quad + \beta_4 koen\_balance_{it} + \beta_5 Storby_{it} + \beta_6 D2017_{it} + v_{it} \qquad t = 2012, 2017 \end{aligned} \tag{1}$$

angiver, hvorledes kriminalitetsraten afhænger af ledigheden for hhv 15-24 og 25-29 årige. Desuden kontrolleres for andelen af studerende, kønsbalancen, om kommunen er en storby og tidseffekter i form af en dummy for 2017. Parametrene β_1 og β_2 skal fortolkes således, at β_1 (β_2) angiver, hvor mange procentpoint kriminalitetsraten ændres, når ledighedsraten for 15-24 (25-29) ændres et procentpoint. Man vil forvente jf. Beckers teori, at fortegnene for β_1 og β_2 er positive. Modellen er estimeret med OLS, og parameterestimater og robuste standardfejl er rapporteret i Tabel 2. OLS estimatet for β_1 kan formodentlig ikke fortolkes som den estimerede kausale effekt af ungdomsledighed på kriminalitetsraten, da MLR.4 ikke er opfyldt. Fejlleddet i modellen vil formodentlig indeholde faktorer som f.eks. uddannelse. Det er sandsynligt, at MLR.4 ikke er opfyldt, da ledighedsvariablene formodentligt er korreleret med den udeladte variabel: uddannelse.

c Vi antager, at variansen på fejlleddet er givet ved

$$Var(v_{it}|X_{it}) = \frac{\sigma^2}{pop\_mand1529_{it}}.$$

Udtrykket angiver, at variansen af fejlleddet er omvendt proportional med antallet af mænd i alderen 15-29 år. Dette er en rimelig antagelse, da den afhængige variabel er et gennemsnit, og gennemsnittet for den enkelte kommune er baseret på observationer $pop\_mand1529_{it}$.[2]

[2]Hvis vi antager, at sandsynligheden for at blive dømt i kommune i til tid t er p_{it}, og der er n_{it} mænd i alderen 15-29, så vil variansen af andelen, som er dømt være $1/(n_{it}(1-p_{it})p_{it})$. Altså omvendt proportional med n_{it}.

Det er derfor sandsynligt, at variansen af gennemsnittet afhænger af populationsstørrelse, som angivet overfor. Modellen er estimeret med WLS, hvor vægtene er

$$h_{it} = \frac{1}{pop\_mand1529_{it}}.$$

Estimationsresultaterne er angivet i Tabel 2. WLS er konsistent, hvis MLR.1-MLR.4 er opfyldt.

Opgave 2 (20%)

a Vi antager nu, at fejlleddet v_{it} kan skrives som $v_{it} = a_i + u_{it}$. Desuden antages, at $ledigheden\_1524_{it}$ og $ledigheden\_2529_{it}$ er korreleret med a_i men ukorreleret med u_{it}. Dette er den klassiske paneldata model. Antagelsen er realistisk, idet man kan forestille sig en række faktorer, som er (næsten) tidsinvariante f.eks. uddannelsessammensætningen i kommunen og som er korreleret med ledigheden. Antagelsen om at ledigheden ikke er korreleret med den tidsvarierende del af fejlleddet kan diskuteres, idet man kan forestille sig at kommuner på nogle tidspunkter prioritere indsatser mod kriminalitet og andre tidspunkter indsatser mod arbejdsløshed. Hvis antagelse nævnt i starten af opgave 2 er opfyldt, kan modellen estimeres konsistent med en FD eller FE estimator (disse to estimatorer vil være identiske, når man har to perioder). FD estimatoren med robuste standardfejl er angivet i tabel 2.

b Vi antager nu, at variansen af u_{it} er givet ved

$$Var(u_{it}|X_{it}) = \frac{\sigma_u^2}{pop\_mand2529_{i2017}} \qquad t = 2012, 2017. \tag{2}$$

Den vægtede FD estimator kan estimeres med udgangspunkt i FD-modellen

$$\begin{aligned} \Delta krimrate_{it} &= \beta_1 \Delta ledighed\_1524_{it} + \beta_2 \Delta ledighed\_2529_{it} + \beta_3 \Delta studentrate_{it} \\ &\quad + \beta_4 \Delta koen\_balance_{it} + \beta_6 + \Delta u_{it} \qquad t = 2017 \end{aligned} \tag{3}$$

hvor vægtene er $h_i = 1/pop\_mand2529_{i2017}$. Estimationsresultaterne er rapporteret i Tabel 2.

c Fordelen ved den vægtede FD estimator er, at den er (mere) efficient, såfremt at antagelsen om variansen er korrekt. Testet for hypotesen, at det kun er ungdomsledigheden ($ledighed\_1524$) og tiden ($D2017$), som påvirker ungdomskriminaliteten, kan formuleres som følgende nulhypotese

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0.$$

Hypotesen testes i den vægtede FD model. Der testes mod det dobbeltside alternativ og anvendes et 5 procent signifikansniveau. Teststørrelsen er givet ved 1.09 og er F-fordelt F(3,93). Testsandsynligheden er 35.7 procent. Konklusionen er, at nulhypotesen ikke kan forkastes. I resten af opgave arbejdes derfor med den restrikterede model.

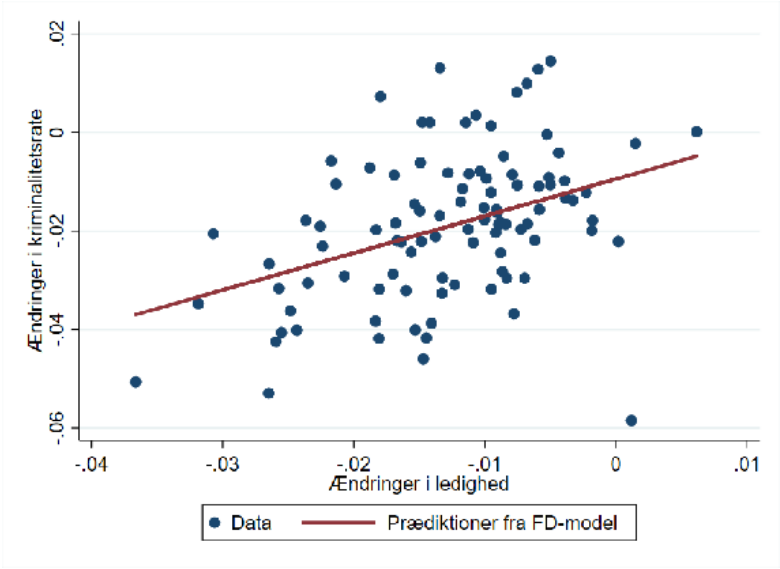

Figure 1: Ændringer i kriminalitet og ungdomsarbejdsløshed 2012-2017

d Figur 1 viser ændringen i kriminalitetsraten fra 2012 til 2017 som funktion af ændringen i ungdomsledighed ($ledighed\_1524$) fra 2012 til 2017. Figuren viser en klar positiv sammenhæng.

Opgave 3 (20%)

Vi tager i denne opgave udgangspunkt i den simpler model

$$krimrate_{it} = \beta_0 + \beta_1 ledighed\_1524_{it} + \beta_6 D2017 + a_i + u_{it}, \tag{4}$$

Vi antager nu at $ledighed\_1524$ potentielt er korreleret med både a_i og u_{it}.

a Model (4) kan omformes til følgende model

$$\begin{aligned} \Delta krimrate_{it} &= \Delta\beta_0 + \beta_1 \Delta ledighed\_1524_{it} + \beta_6 \Delta D2017 + \Delta a_i + \Delta u_{it} \\ \Delta krimrate_{it} &= \beta_1 \Delta ledighed\_1524_{it} + \beta_6 + \Delta u_{it}. \end{aligned}$$

Der gælder, at $\beta_1 = \gamma_1$ og $\beta_6 = \gamma_0$. $\Delta ledighed\_1524_{it}$ er potentielt en endogen variabel og derfor benyttes den forventede vækstrate i antallet af job (Dz_{it}) som instrument for $\Delta ledighed\_1524_{it}$. Dz_{it} er et gyldigt instrument, hvis følgende gælder: $Dzlit$ er korreleret med $\Delta ledighed\_1524_{it}$ og Dz_{it} er ukorreleret med Δu_{it}. Den første del af antagelsen er opfyldt da "First stage" regressionen viser, at instrumentet er korreleret med den endogene variabel. F-testet er 29.69 og F(1,92) med en testsandsynlighed på 0. Altså er instrumentet og den endogene variabel korreleret. Da den forventede jobvækstrate er konstrueret ud fra de nationale sektorvækstrater og jobsammensætningen i kommunen i 2012, vil vækstraten ikke været påvirket af lokale kommunale forhold, som ændre sig mellem 2012 og 2017. Derfor er det sandsynligt at instrumentet er eksogent. IV estimationen er rapporteret i Tabel 2 med robuste standardfejl.

	OLS$^{\#}$	WLS	FD$^{\#}$	FD^{W}	FD^{W}	FDIV$^{\#}$	FDIV^{W}
Ledighed for	0.718**	0.749***	0.945***	0.625**	0.816***	1.754**	1.685**
for 15-24 mænd	(0.233)	(0.205)	(0.188)	(0.201)	(0.155)	(0.539)	(0.539)
Ledighed for	0.232*	0.263***	-0.028	0.109			
25-29 mænd	(0.104)	(0.076)	(0.093)	(0.082)			
Andelen af stud.	-0.196***	-0.233***	0.015	-0.163			
af 15-29 mænd	(0.052)	(0.038)	(0.097)	(0.105)			
Andelen af mænd	-0.059	0.103	0.417*	0.086			
af 15-29 årige	(0.075)	(0.062)	(0.178)	(0.167)			
Dummy for storby	-0.019***	-0.012***					
	(0.004)	(0.003)					
d2017	-0.009**	-0.010***	-0.008*	-0.013***	-0.008***	0.004	0.003
	(0.003)	(0.003)	(0.004)	(0.004)	(0.002)	(0.007)	(0.007)
Constant	0.146***	0.069**					
	(0.037)	(0.026)					
Obs	196	196	98	98	98	98	98
R^2	0.563	0.729	0.719	0.806	0.799		

Note: #: Robuste standard fejl W: Vægtet regression med pop. str. som vægte

Table 2: Estimationsresultater

b For at undersøge om $\Delta ledighed\_1524_{it}$ er en eksogen variabel udføres testet for eksogenitet. Testet udføres ved at estimere følgende hjælperegression:

$$\Delta krimrate_{it} = \beta_1 \Delta ledighed\_1524_{it} + \beta_6 + \delta \hat{e}_{it} + w_{it},$$

hvor $\hat{e}_{it}$ er residualerne fra "first stage" regressionen. Nulhypotesen for eksogenitet er givet ved: $H_0 : \delta = 0$. Teststørrelsen er fundet til –2.29, og hvis man benytter, at teststørrelsen er asymptotisk standard normalfordelt, er den kritiske værdi -1.96. Konklusionen på testet er, at nulhypotesen forkastes, og ledighed betragtes som en endogen variabel.

c Det antages nu, som i spørgsmål 2.b, at variansen af u_{it} er givet ved (2). Den vægtet IV estimation af model (4) kan nu udføres ved at anvende de samme vægte som i spørgsmål 2.b. Estimationsresultatet er angivet i Tabel 2. Når estimationsresultaterne sammenlignes ses, at der kun er mindre forskelle på de vægtede og uvægtede estimationer. Der fokuseres nu på de vægtede resultater. Her ses at OLS estimaterne er generelt mindre end FD estimaterne som er mindre end FD-IV estimaterne. Den forestrukne model er FD-IV estimeret med vægte. Her estimeres den kausale effekt af ungdomsarbejdsløshed på kriminalitiltetsraten til at være 1.7. Det betyder, at hvis ungdomsarbejdsløsheden stiger med et procentpoint, så vil kriminaliteten blandt unge mænd stige med 1.7 procentpoint.

Opgave 4 (20%)

Vi betragter følgende statistiske model:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + u_{it} \tag{5}$$

hvor $i = 1, ..., n$ er antallet af kommuner. Vi observerer hver kommune til tidspunkt 1 og 2. For modellen antages, at MLR.1, MLR.2 og MLR.3 er opfyldt. Desuden antages følgende

$$Cov(x_{it}, a_i) = \delta, Cov(x_{it}, u_{it}) = \rho, Var(x_{it}) = \sigma_x^2 > 0$$

a Den asymptotiske bias for OLS estimatoren til model (5): $p\lim(\hat{\beta}_1^{OLS}) - \beta_1$ kan bregnes idet man kan vise at

$$\hat{\beta}_1^{OLS} = \beta_1 + \frac{\sum_{t=1}^{2}\sum_{i=1}^{n}(x_{it} - \bar{x})(a_i + u_{it})}{\sum_{t=1}^{2}\sum_{i=1}^{n}(x_{it} - \bar{x})^2} = \beta_1 + \frac{\frac{1}{2n}\sum_{t=1}^{2}\sum_{i=1}^{n}(x_{it} - \bar{x})a_i}{\frac{1}{2n}\sum_{t=1}^{2}\sum_{i=1}^{n}(x_{it} - \bar{x})^2} + \frac{\frac{1}{2n}\sum_{t=1}^{2}\sum_{i=1}^{n}(x_{it} - \bar{x})u_{it}}{\frac{1}{2n}\sum_{t=1}^{2}\sum_{i=1}^{n}(x_{it} - \bar{x})^2}$$

Dernæst gælder at

$$p\lim \beta_1^{OLS} = \beta_1 + \frac{Cov(x_{it}, a_i)}{Var(x_{it})} + \frac{Cov(x_{it}, u_{it})}{Var(x_{it})} = \beta_1 + \frac{\delta}{\sigma_x^2} + \frac{\rho}{\sigma_x^2}.$$

Den asymptotiske bias er $(\delta + \rho)/\sigma_x^2$. OLS estimatoren er konsistent, når $\delta = -\rho$.

b Vi antager nu, at vi har en tredje variabel z. Om z gælder der:

$$\begin{aligned} Cov(x_{it}, z_{it}) &= \sigma_{zx} \neq 0 \\ Cov(z_{it}, u_{is}) &= 0, Cov(z_{it}, a_i) = \theta. \end{aligned}$$

Hvis vi anvender z_{it} som instrument for x_{it} i model (5), er IV estimatoren givet ved

$$\hat{\beta}_1^{IV} = \frac{\sum_{t=1}^{2}\sum_{i=1}^{n}(y_{it} - \bar{y})(z_{it} - \bar{z})}{\sum_{t=1}^{2}\sum_{i=1}^{n}(x_{it} - \bar{x})(z_{it} - \bar{z})}.$$

Vi undersøger nu, hvornår IV estimatoren er konsistent ved først at udregne:

$$\hat{\beta}_1^{IV} = \frac{\sum_{t=1}^{2}\sum_{i=1}^{n}(y_{it} - \bar{y})(z_{it} - \bar{z})}{\sum_{t=1}^{2}\sum_{i=1}^{n}(x_{it} - \bar{x})(z_{it} - \bar{z})} = \beta_1 + \frac{\frac{1}{2n}\sum_{t=1}^{2}\sum_{i=1}^{n}(a_i - \bar{a})(z_{it} - \bar{z}) + \frac{1}{2n}\sum_{t=1}^{2}\sum_{i=1}^{n}(u_{ii} - \bar{u})(z_{it} - \bar{z})}{\frac{1}{2n}\sum_{t=1}^{2}\sum_{i=1}^{n}(x_{it} - \bar{x})(z_{it} - \bar{z})}$$

Vi kan nu udregne

$$p\lim \hat{\beta}_1^{IV} = \beta_1 + \frac{Cov(z_{it}, a_i) + Cov(z_{it}, u_{it})}{Cov(x_{it}, z_{it})} = \beta_1 + \frac{\theta}{\sigma_{xz}}.$$

IV estimatoren er konsistent, når $\theta = 0$.

c For at opnå en estimator som er konsistent for alle værdier af δ, ρ og θ, kan man kombinere FD og IV estimatoren. FD-IV estimatoren er baseret på følgende regressionsmodel:

$$\Delta y_{it} = \beta_1 \Delta x_{it} + \Delta u_{it},$$

hvor z_{it} andvendes som instrument for Δx_{it}. Dette kræver dog at $Cov(\Delta x_{it}, z_{it}) \neq 0$. Estimatoren er givet ved

$$\hat{\beta}_1^{FDIV} = \frac{\sum_{i=1}^{n} \Delta y_{i2}(z_{i2} - \bar{z})}{\sum_{i=1}^{n} \Delta x_{i2}(z_{i2} - \bar{z})} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n} \Delta u_{it}(z_{i2} - \bar{z})}{\frac{1}{n}\sum_{i=1}^{n} \Delta x_{i2}(z_{i2} - \bar{z})}$$

Man kan vise at

$$p\lim \hat{\beta}_1^{FDIV} = \beta_1 + \frac{Cov(z_{it}, \Delta u_{it})}{Cov(z_{it}, \Delta x_{it})} = \beta_1.$$

Heraf ses, at estimatoren er konsistent.

Opgave 5 (20%)

Denne opgave går ud på at sammenligne forskellige estimatorer for model (5) ved et simulationsstudie. Vi tager udgangspunkt i modellen fra opgave 4. Betragt følgende datagenererende proces (DGP):

$$\begin{aligned}
y_{it} &= \beta_0 + \beta_1 x_{it} + \alpha_i + u_{it} \\
\alpha_i &\sim N(0,1), u_{it} \sim N(0,1), x_{it}^* \sim N(4,4) \\
x_{it} &= x_{it}^* + \delta a_i + \rho u_{it} \\
z_{it} &= -0.5 x_{it}^* + \theta a_i \\
\beta_0 &= 3,\ \beta_1 = 1, \delta = -0.7, \rho = -0.2, \theta = 0.5
\end{aligned}$$

Det antages at x_{it}^*, α_i og u_{it} er indbyrdes uafhængige.

a OLS estimatoren $\hat{\beta}_1^{OLS}$,FD estimatoren $\hat{\beta}_1^{FD}$ og IV estimatoren $\hat{\beta}_1^{IV}$ evalueres i et simulationseksperiment med en stikprøve på 100 kommuner og to tidsperioder og brug af 1000 replikationer. Simulationsresulaterner er rapporteret i Tabel 3. Simulationsstudiet viser, at alle tre estimatorer er ikke-middelrette, da gennemsnittet af replikationerne er forskelligt fra den sande værdi på 1.

b I spørgsmål 4.a og 4.b blev den asymptotiske bias udregnet for OLS og IV estimatoren. Den asymptotiske bias er givet ved

$$\begin{aligned}
p\lim \hat{\beta}_1^{OLS} - \beta_1 &= \frac{\delta + \rho}{Var(x_{it})} = \frac{\delta + \rho}{Var(x_{it}^*) + \delta^2 Var(a_i) + \rho^2 Var(u_{it})} \\
&= \frac{-0.7 - 0.2}{4 + (-0.7)^2 + (-0.2)^2} = -0.198\,68 \\
p\lim \hat{\beta}_1^{IV} - \beta_1 &= \frac{\theta}{\sigma_{xz}} = \frac{\theta}{Cov(z_{it}, x_{it})} = \frac{\theta}{-0.5 Var(x_{it}^*) + \rho\theta Var(a_i)} \\
&= \frac{\theta}{-0.5 * 4 + \delta\theta} = \frac{0.5}{-0.5 * 4 - 0.7 * 0.5} = -0.21277
\end{aligned}$$

	OLS		FD		IV		FDIV	
	mean	sd	mean	sd	mean	sd	mean	sd
$\hat{\beta}_1$	0.803	0.047	0.947	0.051	0.789	0.049	0.992	0.203
se($\hat{\beta}_1$)t	0.045	0.003	0.050	0.005	0.045	0.003	0.214	0.205
No. repl	1000		1000		1000		1000	

Table 3: Simulationsresultater

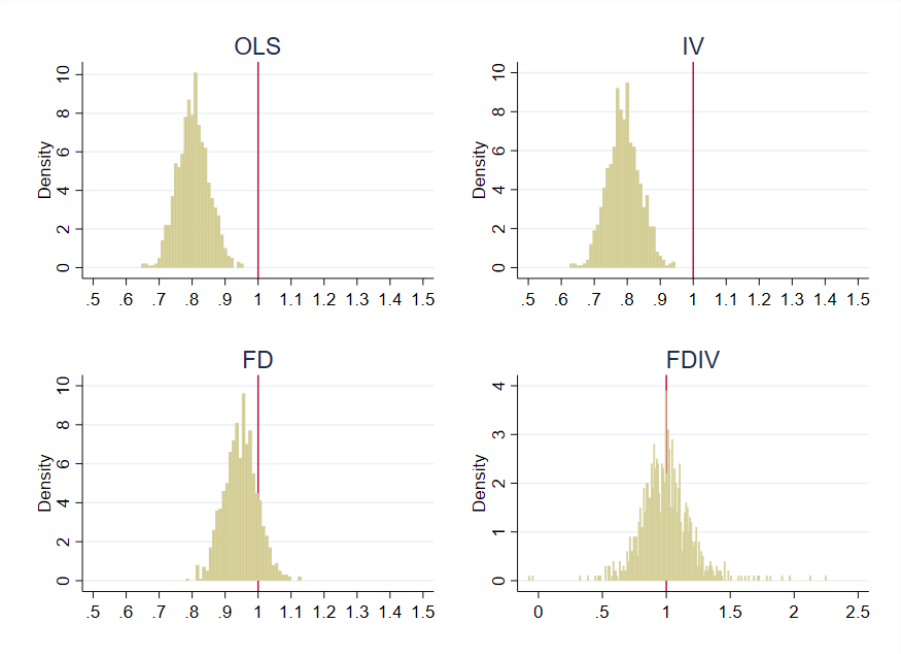

Figure 2: Simulationsresultater for OLS, IV og FDIV estimatoren

Tilsammenligning finder vi i estimationsstudiet en bias i OLS på -0.1970 og bias på IV på -0.21063. Bias fundet ved simulationstudiet ligger meget tæt på den teoretiske asymptotiske bias.

c I simulationsprogrammet er følgende estimator implementeret:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{100} \Delta y_{i2} z_{i2}}{\sum_{i=1}^{100} \Delta x_{i2} z_{i2}}.$$

Estimatoren svarer til en IV estimation på regressionsmodellen i første differenser. I tabel 3 er simulationsresultaterne vist. I figur 2 er fordelingen af de fire estimatorer vist. Tabel 3 og figur 2 viser, at FD-IV estimatoren ikke ser ud til at lide af bias, men tilgengæld har estimatoren en væsentlig større varians. Standardafvigelsen er ca. 4 gange så stor.