Rettevejledning til Eksamen på Økonomistudiet, sommeren 2021

Økonometri I

Tag-hjem eksamen: 15. juni, 2021, kl.10.00-22.00

Effekten af pensionering på helbredet

Vi tager udgangpunkt i en estimation af en lineær sandsynlighedsmodel, som beskriver sandsynligheden for at dø, inden man er fyldt 70 år:

$$\begin{aligned} dod70 &= \beta_0 + \beta_1 pension66 + \beta_2 \log(indkomst) + \beta_3 kvinde + \\ &\quad \beta_4 faglaert + \beta_5 kortudd + \beta_6 mludd + u, \end{aligned} \tag{1}$$

hvor $faglaert$ er en dummyvariabel for faglært uddannelse, $kortudd$ er dummyvariabel for kort uddannlse og $mludd$ er en dummyvariabel for mellem- og lang uddannelse, log er den naturlige logaritme, og u er fejlleddet.
I denne opgave anvendes et 5 procent signifikansniveau, og hvis ikke andet er angivet, testes mod det dobbeltsidet alternativ. Desuden bruges robuste standardfejl i alle regressioner. I opgaven anvendes datasættet Groupdata0.

Opgave 1 (20%)

a. Regressionsmodel (1) er en lineær sandsynlighedsmodel, som beskriver sandsynligheden for at dø, inden man er fyldt 70. De forklarende variable i modellen er en dummy for pensionering ved 66 år, log indkomst, en dummy for kvinde samt tre uddannelsesdummier. Parameteren β_1 kan fortolkes som ændringen i sandsynligheden for tidlig død, hvis man går på pension som 66-årig. Parameteren β_3 er forskellen i sandsynlighed for tidlig død for en kvinde sammenlignet med en mand.

b. Tabel 1 indeholder en deskriptiv analyse, hvor personerne er opdelt efter deres officielle folkepensionsalder. I tabellen er angivet gennemsnit og standardafvigelse for de relevante variable. Tabellen viser, at andelen, som er pensioneret ved 66 år, er ca. 39 procent for dem med en officiel folkepensionsalder på 67 år, mens andelen er ca. 62 procent for dem med en officiel folkepensionsalder på 65. Desuden ses, at uddannelsesniveauet og indkomsten stiger for de yngre kohorter. For de øvrige variable er der kun små forskelle mellem dem, som har en officiel tilbagetrækningsalder ved 67 og 65 år.

c. Model (1) er estimeret ved OLS og parameterestimater, og robuste standardfejl er angivet i tabel 2, 1. søjle. Her anvendes robuste standardfejl, da modellen er en lineær sandsynlighedsmodel. OLS-estimatoren er formodentlig ikke konsistent, da det er sandsynligt, at der er omvendt kausalitet i modellen. Det er sandsynligt, at dem med dårligt helbred vælger at gå tidligere på pension. Derfor kan estimatet af β_1 ikke fortolkes som den kausale effekt af pensionering.

Opgave 2 (20%)

	Pens. alder 67		Pens. alder 65		Alle	
	mean	sd	mean	sd	mean	sd
død 70	0.144	0.351	0.120	0.325	0.131	0.338
pension 66	0.392	0.489	0.627	0.484	0.512	0.500
indkomst (kr.)	195789	109347	222571	1905028	209530	156862
fødselsår	1938.33	0.469	1939.68	0.467	1939.02	0.823
kvinde	0.497	0.500	0.524	0.500	0.510	0.500
grundskole	0.463	0.499	0.410	0.492	0.436	0.496
faglært	0.383	0.486	0.400	0.490	0.392	0.488
kort udd	0.125	0.331	0.149	0.356	0.137	0.344
mellem/lang udd	0.030	0.169	0.041	0.198	0.035	0.185
no. obs	745		785		1530	

Table 1: Deskriptiv statistik

a. Model (1) estimeres nu ved en IV estimation, hvor dummien for den officielle folkepensionsalder FP anvendes som instrument for *pension*66. For at FP kan anvendes som instrument, skal der gælde, at FP er korreleret med *pension*66. Dette testes i en "first stage regression" ved at teste nulhypotesen om, at parameteren til FP er lig med 0 mod alternativ hypotesen, at parameteren er forskellig fra 0. Parameteren er estimeret til 0.246 med en t-værdi på 10.21 og en p-værdi på 0.00. Det betyder, at vi afviser nulhypotesen og kan konkludere, at instrumentet er korreleret med den endogene variabel *pension*66. For at FP er et gyldigt instrument, skal der gælde, at FP er ukorreleret med fejlleddet. Dette er en rimelig antagelse, idet ændringen af pensionsalderen er gennemført som en reform, hvor personerne berørt af reformen ikke har haft indflydelse på den. Et potentiel problem kunne dog være, at reformen er baseret på fødselsdag. Det betyder, at de berørte personer er yngre fødselskohorter end de ikke berørte personer. Hvis der har været en generel forbedring af helbredet over generationer, kunne dette være et problem. Dette antages dog at være et mindre problem, da der maksimalt er 3 års forskel på personerne i datasættet. Derfor sluttes, at betingelserne for et valid instrument er opfyldt, og IV-estimatoren er konsistent. Estimationsresultaterne er angivet i tabel 2, 2. søjle.

b. Testet for endogenitet bygger på, at man kan opdele variationen i den potentiel endogene variabel i en eksogen del (som er forklaret af instrumentet) og en potentiel endogen del (restvariationen). Hvis restvariationen indgår i modellen, er variablen endogen. Residualerne fra "first stage regression" er et estimat på "restvariationen". Test, for om *pension*66 er en eksogen variabel, udføres derfor ved at inkludere residualerne fra "first stage" regressionen i regressionsmodellen (se tabel 2, søjle 3). Nulhypotesen er, at *pension*66 er en eksogen variabel. Det ses, at residualerne indgår signifikant. Hypotesen om eksogenitet kan afvises, da z-test størrelsen er

$$z = \frac{0.156}{0.075} = 2.09.$$

z-teststørrelsen er asymptotisk normalfordelt og har en kritisk værdi på 1.96, og derfor må hypotesen, om at *pension*66 er eksogen, forkastes. Derfor foretrækkes IV estimationen, og IV estimation benyttes i resten af den empiriske opgaven. IV estimatet for β_1 er -0.067, og vi kan ikke forkaste hypotesen om, at $\beta_1 = 0$. Derfor sluttes, at pensionering ikke har en kausal

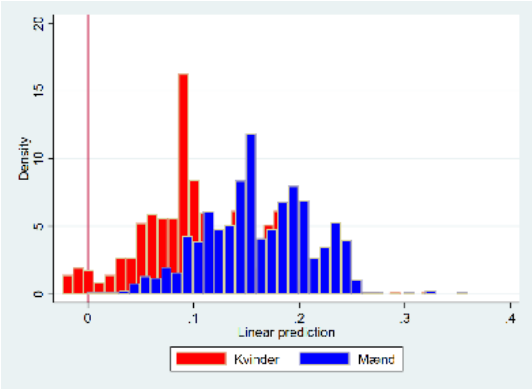


Figure 1: Histogram over prædikterede sandsynligheder

effekt på sandsynligheden for tidlig død. IV estimatet for β_3 er -0.067 og er signifikant forskelligt fra 0. Det betyder, at kvinder sammenlignet med mænd har 6.7 procent point mindre risiko for at dø, inden de fylder 70.

c. Figur 1 viser et histogram over de prædikterede sandsynligheder for tidlig død for hhv. mænd og kvinder. Histogrammet viser, at der for kvinderne er nogle få observationer, hvor den prædikterede sandsynlighed er mindre end 0. Da det er meget få observationer, vurderes at modellen er velspecificeret.

Opgave 3 (20%)

I denne opgave ser vi på et andet mål for helbred. Her benytter vi antallet af lægebesøg som 70-årig. Vi benytter følgende model:

$$\begin{aligned} laege70 &= \gamma_0 + \gamma_1 pension66 + \gamma_2 \log(indkomst) + \gamma_3 kvinde + \\ &\quad \gamma_4 faglaert + \gamma_5 kortudd + \gamma_6 mludd + v, \end{aligned} \tag{2}$$

hvor v er et fejlled.

a. Model (2) er estimeret ved en IV estimation, hvor dummien for den officielle folkepensionsalder FP anvendes som instrument for $pension66$. Det bemærkes, at variable $laege70$ kun er observeret for 1329 observationer, svarende til de personer som stadig er i live ved 70 år. FP er stadig korreleret med $pension66$ i det reducerede datasæt. Estimationsresultaterne er angivet i tabel 2, 4. søjle. Estimationsresultaterne viser, at pensionering ved 66 år sænker antallet af lægebesøg med gennemsnitlig ca. et halvt lægebesøg pr. år. I denne estimation indgår færre observationer, da personer, som dør, inden de fylder 70, udgår. Der kan derfor være endogen selektion, idet kun de sundeste er medtaget i regressionen. Det ikke ser ud til, at pensionering har en effekt på sandsynligheden for at dø inden 70 (jf. opgave 2), og derfor er det et mindre problem.

b. For at teste for om uddannelse har betydning for antallet af lægebesøg formuleres følgende nulhypotese: $H_0 : \gamma_4 = \gamma_5 = \gamma_6 = 0$. Teststørrelsen er beregnet til 11.61 og er asymptotisk

	OLS	IV	IV	IV	IV
afh. var	død 70	død 70	død 70	lægebesøg	lægebesøg
pension 66	0.079***	-0.067	-0.067	-0.567***	-0.201
	(0.018)	(0.072)	(0.070)	(0.151)	(0.229)
log(indkomst)	-0.036*	-0.032	-0.032*	-0.030	-0.031
	(0.016)	(0.016)	(0.016)	(0.033)	(0.031)
kvinde	-0.082***	-0.067***	-0.067***	-0.396***	-0.072
	(0.019)	(0.020)	(0.020)	(0.051)	(0.150)
faglært	-0.016	-0.040	-0.040	-0.035	-0.035
	(0.020)	(0.023)	(0.023)	(0.055)	(0.051)
kort udd.	-0.041	-0.078**	-0.078**	0.118	0.116
	(0.025)	(0.030)	(0.029)	(0.076)	(0.072)
mellem/lang udd	0.013	-0.044	-0.044	0.304*	0.322**
	(0.050)	(0.058)	(0.056)	(0.130)	(0.120)
residual			0.156*		
			(0.075)		
kvindexpen					-0.674*
					(0.304)
konstant	0.578**	0.617**	0.617**	1.542***	1.404***
	(0.194)	(0.200)	(0.194)	(0.405)	(0.389)
obs	1530	1530	1530	1329	1329
R^2	0.030	.	0.033	.	0.075

Table 2: Estimationsresultater

χ^2-fordelt med 3 frihedsgrader og p-værdien er 0.009. Derfor forkastes nulhypotesen, og det konkluderes, at uddannelse har betydning for antallet af lægebesøg.

c. Vi ønsker nu at undersøge, om pensionering har en forskellig effekt på helbredet for mænd og kvinder. Derfor udvides modellen således:

$$\begin{aligned} laege70 &= \gamma_0 + \gamma_1 pension66 + \gamma_2 \log(indkomst) + \gamma_3 kvinde + \\ &\quad \gamma_4 faglaert + \gamma_5 kortudd + \gamma_6 mludd + \gamma_7 pension66 \cdot kvinde + v. \end{aligned} \tag{3}$$

Model (3) estimeres ved en IV estimation, hvor FP og $FP \cdot kvinde$ benyttes som instrument for $pension66$ og $pension66 \cdot kvinde$. First stage regressionerne viser, at instrumenterne er korreleret med de potentielt endogene variable. Estimationsresultaterne er angivet i tabel 2, søjle 5. Vi tester nu, om pensionering har forskellig effekt for mænd og kvinder ved at teste følgende nulhypotese: $H_0 : \gamma_7 = 0$ mod alternativ hypotesen $H_1 : \gamma_7 \neq 0$. Teststørrelsen er en z-teststørrelse og bestemt til -2.22. Teststørrelsen er asymptotisk normaltfordelt, og den kritiske værdi er 1.96. Hypotesen kan derfor forkastes. Det betyder, at der er signifikant forskel på effekten af pensionering for mænd og kvinder. For kvinder betyder pensionering ved 66 år, at antallet af lægebesøg falder med i gennemsnit ca. 0.87 lægebesøg pr. år, og for mænd falder antallet af lægebesøg med ca. 0.2 pr. år ved pensionering, som dog ikke signifikant.

Opgave 4 (20%)

Betragt følgende regressionssmodel:

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

hvor $i = 1, ..., n$. Vi antager, at MLR.1-MLR.3 er opfyldt. Desuden antager vi, at $E(u) = 0$ og $E(xu) = 0.10$.

a. Antag at x er en dummyvariabel, hvor der gælder, at $E(x) = P(x = 1) = 0.15$. Variansen af x kan udregnes som variansen af en binær variabel:

$$Var(x) = P(x = 1) \cdot (1 - P(x = 1)) = 0.15 \cdot 0.85 = 0.1275.$$

Den asymptotiske bias af OLS estimatoren for β_1 er givet ved

$$\begin{aligned} p\lim \hat{\beta}_1 - \beta_1 &= \frac{Cov(u,x)}{Var(x)} = \frac{E(ux) - E(x) \cdot E(u)}{Var(x)} \\ &= \frac{0.10 - 0.15 \cdot 0}{0.15 \cdot 0.85} = 0.784. \end{aligned}$$

b. Vi skal vise, at OLS estimatoren $\hat{\beta}_1$ for β_1, når x er en dummyvariabel, kan skrives som

$$\begin{aligned} \hat{\beta}_1 &= \bar{y}^{x=1} - \bar{y}^{x=0}, \\ \text{hvor } \bar{y}^{x=1} &= \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i}, \bar{y}^{x=0} = \frac{\sum_{i=1}^n y_i (1 - x_i)}{\sum_{i=1}^n (1 - x_i)}. \end{aligned}$$

I de følgende udregninger benyttes, at når x er en dummyvariabel, gælder at $x_i = x_i^2$. OLS estimatoren er givet ved:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n x_i(x_i - \bar{x})} = \frac{\sum_{i=1}^n y_i x_i - \bar{x}\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i} \\
&= \frac{n \cdot \bar{x} \cdot \bar{y}^{x=1} - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i - \bar{x}n\bar{x}} = \frac{n \cdot \bar{x} \cdot \bar{y}^{x=1} - n\bar{x}(\bar{x} \cdot \bar{y}^{x=1} + (1-\bar{x}) \cdot \bar{y}^{x=0})}{n\bar{x} - \bar{x}n\bar{x}} \\
&= \frac{n \cdot \bar{x} \cdot ((1-\bar{x}) \cdot \bar{y}^{x=1} - (1-\bar{x}) \cdot \bar{y}^{x=0})}{n\bar{x}(1-\bar{x})} = \frac{n \cdot \bar{x} \cdot (1-\bar{x}) \cdot (\bar{y}^{x=1} - \bar{y}^{x=0})}{n\bar{x}(1-\bar{x})} \\
&= (\bar{y}^{x=1} - \bar{y}^{x=0})
\end{aligned}$$

c. Vi skal vise, at estimatoren $\tilde{\beta}_1$ er konsistent estimator for β_1

$$\tilde{\beta}_1 = \frac{\bar{y}^{d=1} - \bar{y}^{d=0}}{\bar{x}^{d=1} - \bar{x}^{d=0}}.$$

Dette kan enten gøres ved at indse, at $\tilde{\beta}_1$ er IV estimatoren for β_1, hvor d anvendes som instrument for x. Betingelserne for et validt instrument er opfyldt. Alternativt kan man vise, at estimatoren $\tilde{\beta}_1$ er konsistent, ved at benytte at

$$\begin{aligned}
\bar{y}^{d=1} &= \frac{\sum_{i=1}^n y_i d_i}{\sum_{i=1}^n d_i} = \frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_i + u_i) d_i}{\sum_{i=1}^n d_i} \\
&= \beta_0 + \beta_1 \frac{\sum_{i=1}^n x_i d_i}{\sum_{i=1}^n d_i} + \frac{\sum_{i=1}^n u_i d_i}{\sum_{i=1}^n d_i} \\
&= \beta_0 + \beta_1 \bar{x}^{d=1} + \bar{u}^{d=1}.
\end{aligned}$$

Tilsvarende kan man vise, at $\bar{y}^{d=0} = \beta_0 + \beta_1 \bar{x}^{d=0} + \bar{u}^{d=0}$. Vi kan nu regne på estimatoren:

Trin 1:

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{\bar{y}^{d=1} - \bar{y}^{d=0}}{\bar{x}^{d=1} - \bar{x}^{d=0}} = \frac{\beta_0 + \beta_1 \bar{x}^{d=1} + \bar{u}^{d=1} - (\beta_0 + \beta_1 \bar{x}^{d=0} + \bar{u}^{d=0})}{\bar{x}^{d=1} - \bar{x}^{d=0}} \\
&= \beta_1 + \frac{\bar{u}^{d=1} - \bar{u}^{d=0}}{\bar{x}^{d=1} - \bar{x}^{d=0}} = \beta_1 + \frac{1}{\bar{x}^{d=1} - \bar{x}^{d=0}} \left(\frac{\frac{1}{n}\sum_{i=1}^n u_i d_i}{\frac{1}{n}\sum_{i=1}^n d_i} - \frac{\frac{1}{n}\sum_{i=1}^n u_i (1-d_i)}{\frac{1}{n}\sum_{i=1}^n (1-d_i)} \right).
\end{aligned}$$

Trin 2:

$$\begin{aligned}
p\lim(\tilde{\beta}_1) &= \beta_1 + p\lim\left(\frac{1}{\bar{x}^{d=1} - \bar{x}^{d=0}}\right) \cdot \left(\frac{p\lim(\frac{1}{n}\sum_{i=1}^n u_i d_i)}{p\lim(\frac{1}{n}\sum_{i=1}^n d_i)} - \frac{p\lim(\frac{1}{n}\sum_{i=1}^n u_i (1-d_i))}{p\lim(\frac{1}{n}\sum_{i=1}^n (1-d_i))} \right) \\
&= \beta_1 + p\lim\left(\frac{1}{\bar{x}^{d=1} - \bar{x}^{d=0}}\right) \left(\frac{E(ud)}{p\lim(\frac{1}{n}\sum_{i=1}^n d_i)} - \frac{E(u) - E(ud)}{p\lim(\frac{1}{n}\sum_{i=1}^n (1-d_i))} \right) \\
&= \beta_1 + p\lim\left(\frac{1}{\bar{x}^{d=1} - \bar{x}^{d=0}}\right) \left(\frac{0}{p\lim(\frac{1}{n}\sum_{i=1}^n d_i)} - \frac{0-0}{p\lim(\frac{1}{n}\sum_{i=1}^n (1-d_i))} \right) \\
&= \beta_1.
\end{aligned}$$

Heraf sluttes, at $\tilde{\beta}_1$ er konsistent.

Opgave 5 (20%)

Denne opgave går ud på at illustrere egenskaberne i en lineær sandsynlighedsmodel. Vi generer data fra følgende model:

$$y_i^* = \gamma_0 + \gamma_1 x_i + u_i \tag{4}$$

$$y_i = \begin{cases} 1 \text{ hvis } y_i^* > 0 \\ 0 \text{ ellers} \end{cases}$$

$$\gamma_0 = 3, \gamma_1 = -2,\ x \sim iiN(2,1), u \sim N(0,1)$$

Vi antager nu, at y^* er uobserveret, men y og x er observerede. Vi estimerer følgende regressionsmodel

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

a. I et simulationseksperiment med 1000 replikationer estimeres β_1 ved OLS på en stikprøve med 500 observationer. Her anvendes et seed, som er 117. Deskriptiv statistik er rapporteret i tabel 3, og i figur 2 er histogrammet med OLS estimaterne vist. Simulationseksperimentet viser, at OLS estimatet i gennemsnit er -0.32 med en standardafvigelse på 0.0128.

b. I simulationsstudiet er både de robuste og ikke robuste standardfejl for OLS estimatoren af β_1 beregnet og vist i tabel 3. Det ses, at den gennemsnitlige ikke robuste standardfejl er 0.0153, hvilket er for stor i forhold til standardafvigelsen på estimatoren. Hvis vi sammenligner med de robuste standardfejl, er værdien 0.0131, hvilket er noget tættere på de 0.0128. Hvis man bruger de ikke robuste standardfejl, så vil man anvende en "for stor" standardfejl, og det vil have konsekvenser for teststørrelsen og herved for sandsynligheden for at forkaste en sand nulhypotese.

c. Den teoretiske sandsynlighed for at $y = 1$, når $x = k$, kan beregnes som

$$P(y = 1|x = k) = P(\gamma_0 + \gamma_1 k + u > 0) = 1 - \Phi(-\gamma_0 - \gamma_1 k),$$

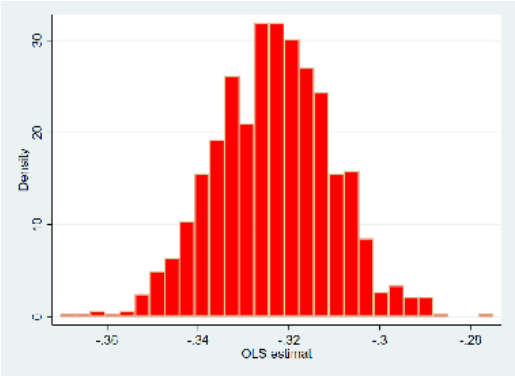
hvor Φ er den kummulative fordelingsfunktion for $N(0,1)$. Den teoretiske ændring i sandsynlighed, når x går fra 0 til 1, er:

$$\begin{aligned} P(y &= 1|x = 1) - P(y = 1|x = 0) = 1 - \Phi(-\gamma_0 - \gamma_1) - (1 - \Phi(-\gamma_0)) \\ &= \Phi(-\gamma_0) - \Phi(-\gamma_0 - \gamma_1) = \Phi(-3) - \Phi(-3 - -2) = -0.157. \end{aligned}$$

Den effekt skal sammenlignes med OLS estimatoren for β_1, som er fundet til -0.32. Forskellen skyldes, at effekten i en OLS estimation er antaget at være en konstant effekt af x på sandsynligheden for $y = 1$, mens i den sande model afhænger effekten af x af niveauet på x.

	mean	sd	min	max
OLS estimat	-0.3233	0.0128	-0.370	-0.275
Std. fejl af OLS	0.0153	0.0006	0.013	0.018
Robuste Std. fejl af OLS	0.0131	0.0009	0.011	0.017
No. replikationer	1000			

Table 3: Resultater fra simulationsstudiet



Fordelingen af OLS estimater af β_1