

Eksamen på Datalogi-Økonomi Blok 1 2022

Økonometri A

Gruppemedlemmernes eksamensnumre (i stigende rækkefølge):

Eksamensnummer: 20

Eksamensnummer: 21

Eksamensnummer: 38

Antal sider i alt (inklusive forsiden): 19

For gruppebesvarelser skal det anføres, hvem der er ansvarlig for hvilke afsnit.

Eksamensnummer 20 er ansvarlig for følgende afsnit: 1A, 2B,3C,4A,5A

Eksamensnummer 21 er ansvarlig for følgende afsnit: 1B,2C,3A,4B,5B

Eksamensnummer 38 er ansvarlig for følgende afsnit: 1C,2A,3B,4C,5C

(Dog er alle opgaver udarbejdet som gruppe)

1 Problem 1

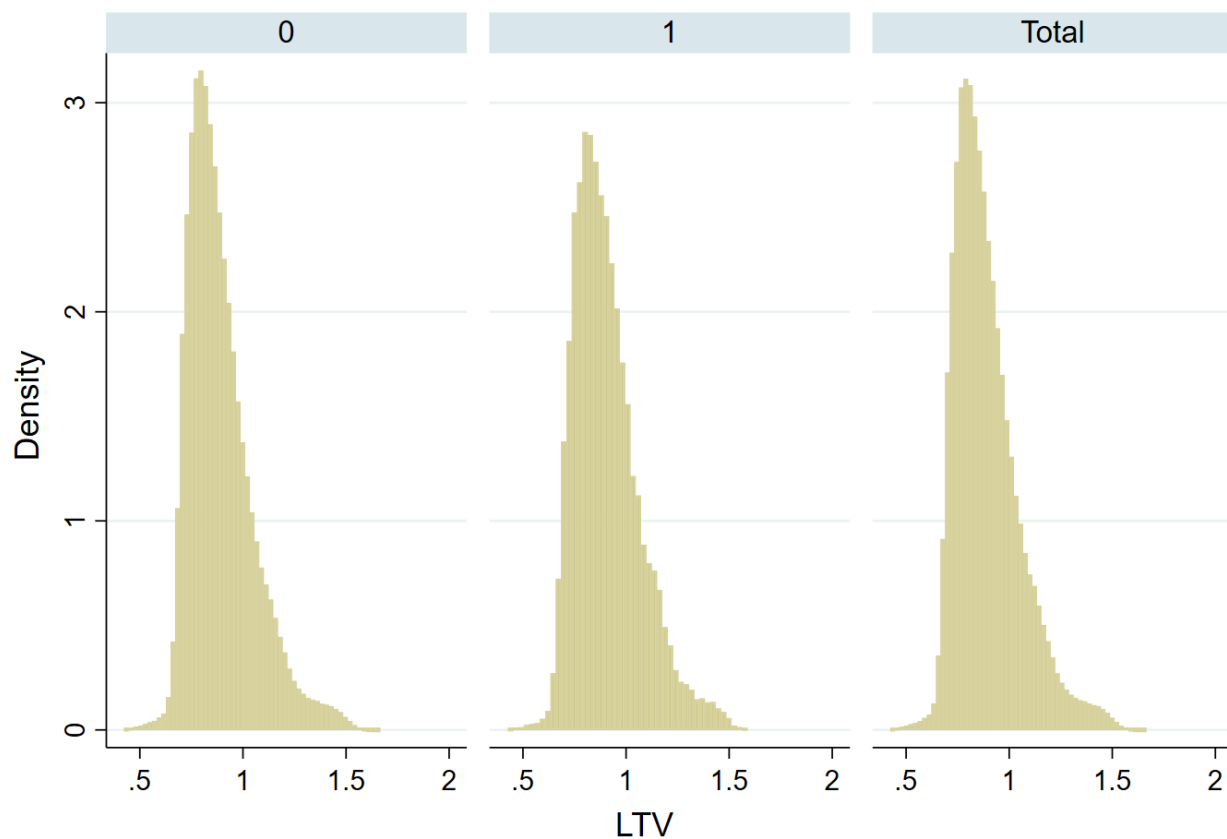
1.1 A

Fra tabellen i Figur 1 ser vi, at LTV er 1,56%-point højere for folk der er flyttet mellem 2000 og 2002. Dem der flytter er i gennemsnit 2 år yngre. Husholdningens størrelse er omtrænt det samme. Endvidere kan vi se, at der er markant flere observationer for folk der ikke flytter end folk der flytter.

Moved	Variable	Mean	Min	Max	N
0	LTV	.8917341	.4230995	1.666715	369516
	Age	41.05667	25	79	369516
	HouseholdSize	1.257166	1	2	369516
1	LTV	.9077262	.4267612	1.58942	32870
	Age	39.05135	25	79	32870
	HouseholdSize	1.237025	1	2	32870

Figur 1: Summarize af datasættet

Fra Figur 2 ser vi, at gennemsnittet for LTV for begge grupper ligger ca. det samme sted: Omkring 0.9. Dem der ikke flytter har dog en større kurtosis, men ca. samme varians. Alle figurene har lidt positiv skævhed.



Figur 2: Histogram af fordelingen af LTV og $move$

1.2 B

Til at starte med estimerer vi følgende simple lineære regressionsmodel:

$$Moved = \beta_0 + \beta_1 LTV + u \quad (1)$$

Linear regression	Number of obs	=	402,386
	F(1, 402384)	=	290.79
	Prob > F	=	0.0000
	R-squared	=	0.0007
	Root MSE	=	.27379

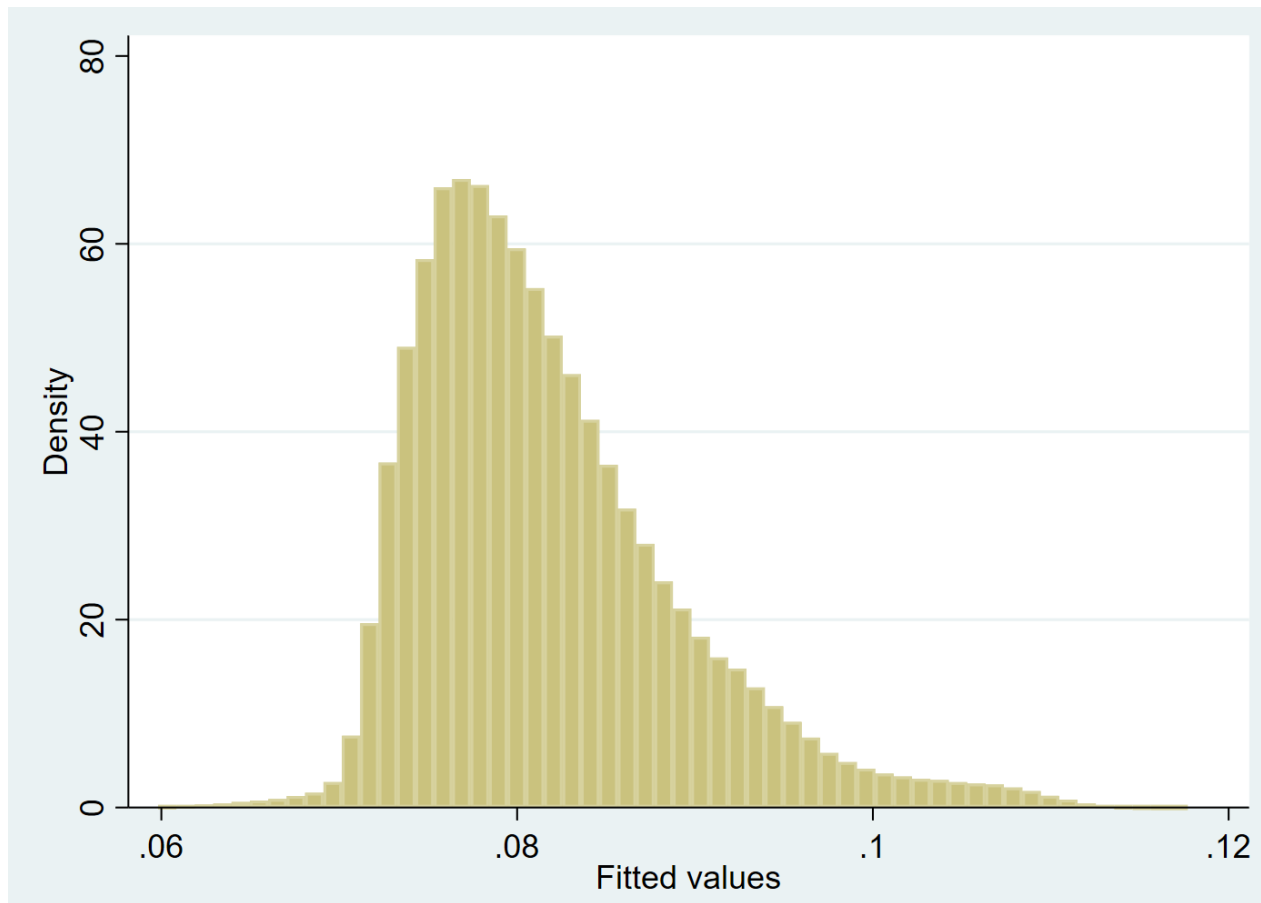
Moved	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]	
LTV	.0465505	.0027298	17.05	0.000	.0412001	.0519008
_cons	.0401163	.0024405	16.44	0.000	.0353331	.0448995

Figur 3: Lineær regression af Model (1)

Model (1) er estimeret med OLS, hvor parameterestimer og de robuste standardfejl er angivet i tabellen på figur 3. Her anvendes robuste standardfejl, da modellen er en lineær sandsynlighedsmodel. Vores parameterestimat for β_1 på 0.04655 indebærer at hvis man har en *LTV* der er 1 højere så har man 4.65%-point øget sandsynlighed for at flytte, med en standardfejl på 0.00273. Vi har derfor en lille statistisk usikkerhed på estimatet. Vi har en meget lav R^2 , hvilket betyder at vi kun kan forklare 0.007% i variationen af y med udelukkende *LTV* som forklarende variabel. Eftersom vi har dummier som outcome variable, altså en lineær sandsynlighedsmodel, ved vi at antagelsen SLR.5 per definition ikke er opfyldt. Da vi ikke kan antage homoskedasticitet, bruger vi robuste standardfejl.

1.3 C

Nu danner vi de prædikterede værdier ud fra estimatet, vi lavede i opgave 1B. De prædikterede værdier kan tolkes som sandsynligheden for at husholdningen flytter i 2000-2002 udelukkende på baggrund af deres *LTV*. For at en prædikteret værdi ikke er meningsfuld, skal den enten være under 0 eller over 1. Som vi kan se på figur 4, har vi en fordeling der kan tilnærmes en normalfordeling, der har positiv skævhed. Ligeledes kan vi se på Figur 5, at der ikke er nogle prædikterede værdier på over 1 eller under 0.



Figur 4: Histogram af de prædikterede værdier

Variable	Obs	Mean	Std. dev.	Min	Max
Moved_est	402,386	.0816877	.0074729	.0598118	.1177026

Figur 5: Summarize af de prædikterede værdier

2 Problem 2

2.1 A

I denne opgave udvider vi modellen så vi får følgende:

$$Moved = \beta_o + \beta_1 LTV + \beta_2 HouseholdSize + \beta_3 Age + \beta_4 Age^2 \quad (2)$$

Linear regression	Number of obs	=	402,386
	F(4, 402381)	=	520.04
	Prob > F	=	0.0000
	R-squared	=	0.0061
	Root MSE	=	.27305

Moved	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]	
LTV	.0369986	.0027692	13.36	0.000	.031571	.0424262
HouseholdSize	-.003194	.0009787	-3.26	0.001	-.0051122	-.0012757
Age	-.0115973	.0003248	-35.70	0.000	-.0122339	-.0109607
Age2	.000116	3.56e-06	32.56	0.000	.000109	.0001229
_cons	.3209933	.0077054	41.66	0.000	.3058909	.3360956

Figur 6: Regression af den udvidede model

Vi ser i figur 6, når vi udvider modellen så falder estimatet af LTV's effekt fra 4.655%-point til 3.7%-point. Dermed har der også været noget bias, som har været indeholdt i enten HouseholdSize eller Age. Vores standardfejl går fra 0.0027298 til 0.0027692, dvs at vores standardfejl er en lille smule højere i denne model, sammenlignet med den fra opgave 1B. Når vi inkluderer flere variable i en regression, får vi ofte et bias-variance tradeoff. Som vi kan se, har vi mindre bias og mere varians. Intuitionen bag dette er, at når vi inkluderer flere variable "tager"de noget af variationen, som øger standardfejlen. Der er dog også en modsatrettet effekt, der mindsker vores standardfejl, da vi fjerner både Age og HouseholdSize fra fejlledet, mindskes vores varians, og derved reduceres standardfejlen. Til sidst kan vi se at vores R^2 er gået fra 0.007% til 0.61%, hvilket er en markant forbedring i forklaringsgraden. Vi ved at R^2 aldrig kan falde ved inklusion af flere variable, men det er ikke garanteret at den stiger så meget.

2.2 B

Vi vælger at gå videre med den udvidede model, da vi som beskrevet i opgave 2.A ved, at den har en markant større R^2 -værdi end vores originale model. Herefter udvider vi modellen

igen med $LTVo1$, som er en dummy for om husstanden er insolvent samt $LTVo1_LTV$, som er en variabel der angiver, at hvis de er insolvente, hvor meget insolvente husstanden de i så fald er.

$$Moved = \beta_o + \beta_1 LTV + \beta_2 HouseholdSize + \beta_3 Age + \beta_4 Age^2 + \delta_1 LTVo1 + \delta_2 LTVo1_LTV \quad (3)$$

Linear regression		Number of obs	=	402,385
		F(6, 402378)	=	361.42
		Prob > F	=	0.0000
		R-squared	=	0.0063
		Root MSE	=	.27303

Moved	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]	
LTV	.0642728	.0054953	11.70	0.000	.0535023	.0750434
HouseholdSize	-.0030888	.0009787	-3.16	0.002	-.005007	-.0011706
Age	-.0115868	.0003248	-35.68	0.000	-.0122234	-.0109503
Age2	.000116	3.56e-06	32.58	0.000	.000109	.000123
LTVo1	.0021596	.001837	1.18	0.240	-.0014409	.0057602
LTVo1_LTV	-.0774826	.0095051	-8.15	0.000	-.0961124	-.0588529
_cons	.2978205	.0086616	34.38	0.000	.280844	.314797

Figur 7: Regression udivdet med $LTVo1$ og $LTVo1_LTV$

Fra Figur 7 ses det, at δ_1 er insignifikant på et 5% signifikansniveau (95% konfidensintervallet er $[-0.00144, 0.00576]$). Derfor kan vi ikke afvise at vores sande parameter er 0, og dermed kan vi ikke tolke på denne parameter. Dog kan det ses, at $\delta_2 = -0.774826$, som angiver at jo mere insolvent en husstand er, desto mindre sandsynlighed er for at de flytter.

2.3 C

Hvis der ikke er et brud af sammenhængen mellem $Moved$ og LTV , når $LTV > 1$, må de følge samme kurve, og der vil derfor kunne bestemmes blot én kurve, som kan bestemme dem begge. Vi vil både undersøge om der er et niveauskifte, samt om der er et hældningsskifte. For at undersøge om der er et niveauskifte laver vi en nulhypotese og alternativ hypotese:

$$H_0 : \delta_1 = 0$$

$$H_A : \delta_1 \neq 0$$

Eftersom vi allerede i opgave 2B konkluderede at δ_0 ikke er signifikant forskellig fra 0, så afkaster vi ikke vores nulhypotese. Vi undersøger, om der er hældningsskifte i de to grupper ved at se på den marginale effekt af LTV. Dermed kan vi opstille vores nulhypotese og alternative hypotese som følgende:

$$H_0 : \delta_2 = 0$$

$$H_A : \delta_2 \neq 0$$

Her kan vi aflæse fra figur 7 at vi har en t-teststørrelse på -8.15 og vi kan hermed på et 5% signifikansniveau forkaste vores nulhypotese og sandsynliggøre vores alternativhypotese. Dermed er der et brud i sammenhængen mellem Moved og LTV ved $LTV > 1$, da der er et hældningsskifte. Dette er blevet løst med en robust t-test. Derfor er der 4 antagelser der skal være opfyldt.

1. LTV skal følge en kontinuert skala
2. Dataen skal være taget fra et repræsentativt tilfældigt udpluk af populationen
3. Dataen skal forme en normalfordeling når plottet. (Dette kan ses i 2 fra opgave 1A)
4. Vi skal have et relativt stort sample size.

Normalt kunne en Chow-test bruges til at undersøge brud på sammenhængen. Dette havde dog krævet antagelse om homoskedasticitet, hvilket vi tidligere har argumenteret imod, da vi har en lineær sandsynlighedsmodel.

3 Problem 3

3.1 A

Vi går videre med samme model, hvor vi laver en 2SLS IV regression med *LocalPriceChange* som instrument for *LTV*.

Instrumental variables 2SLS regression	Number of obs	=	402,385
	Wald chi2(6)	=	2008.33
	Prob > chi2	=	0.0000
	R-squared	=	.
	Root MSE	=	.27403

Moved	Coefficient	Robust std. err.	z	P> z	[95% conf. interval]	
LTV	-.2338646	.0344344	-6.79	0.000	-.3013547	-.1663745
HouseholdSize	-.0005561	.0010234	-0.54	0.587	-.002562	.0014498
Age	-.0118439	.0003277	-36.14	0.000	-.0124862	-.0112016
Age2	.0001165	3.58e-06	32.55	0.000	.0001095	.0001235
LTVo1	.0532871	.0061084	8.72	0.000	.0413149	.0652593
LTVo1_LTV	.2194737	.035188	6.24	0.000	.1505065	.288441
_cons	.550876	.030193	18.25	0.000	.4916989	.6100531

Figur 8: IV Regression 2sls med *LocalPriceChange* som instrument for *LTV*

I tabellen på figur 9 er estimaterne for IV- og OLS regressionen indsat til sammenligning. Her ses det at β_1 falder fra 0.0642728 til -0.23394 . Vi kan dermed se at OLS estimatet er biased opad, faktisk i den grad at fortegnet ændre sig ved IV estimation. Dette kunne være på grund af, at folk der har en høj LTV oftere flytter, eller har andre opsparingstilbøjeligheder og fremtidsplaner. Denne bias bliver fjernet når der bruges IV. Da vores instrument kun bruger den variation der er i lokale prisændringer. Fordi IV estimationen kun bruger denne variation bliver estimationen også mere usikker og derved er standardfejlen større for IV.

	Moved IV	Moved OLS
LTV	-0.2338646 (0.0344344)	0.0642728 (0.0054953)
Householdsize	-0.0005561 (0.0010234)	-0.0030888 (0.0009787)
Age	-0.0118439 (0.0003277)	-0.0115868 (0.0003248)
Age 2	0.0001165 (3.58e-06)	0.000116 (3.56e-06)
LTVo1	0.0532871 (0.0061084)	0.0021596 (0.001837)
LTVo1_LTV	0.2194737 (0.035188)	-0.0774826 (.0095051)

Figur 9: Tabel af estimater fra OLS og IV regression samt deres standardfejl i parentes

Der er to nødvendige antagelser for at vores 2SLS er en konsistent estimator for den kausale effekt af LTV på sandsynligheden af at flytte. De generelle antagelser er følgende:

- Antagelse om exogenitet. Vores instrument(z) må ikke være korreleret med vores fejlede(u).
- Antagelse om relevans. Vores instrument(z) skal være korreleret med vores endogene variable(x).

Antagelsen om relevans kan testes og gøres i opgave 3B. Det er dog en retfærdig antagelse, eftersom det kunne tænkes at, en stigning i lokale boligpriser ville medføre stigning i husets værdi, og derved mindske LTV. Antagelsen om eksogenitet kan ikke testes, da vi ikke ved hvad den sande u er. Derfor må vi her antage at det er gældende. Dog kan man forstille sig, at der er noget i fejlløbet som påvirker de lokale priser. Eksempelvis kunne det tænkes at en stigning i kriminaliteten i området ville trække priserne ned. Dette ville bryde med vores antagelse.

3.2 B

Først kontrolleres om instrumentet (*LocalPriceChange*) er korreleret med den endogene variabel *LTV* ved at køre first stage regressionen, se Figur 10. Her ses det at når *LocalPriceChange* stiger med 1 så falder *LTV* med -0.0227778 . Yderligere ses det også at den er signifikant, da t-teststørrelsen er på -93.63 . Dermed kan vi bekræfte vores antagelse om relevans.

```

Number of obs =    402,385
F(6, 402378)   = 1653715.26
Prob > F       =    0.0000
R-squared      =    0.7661
Adj R-squared  =    0.7661
Root MSE      =    0.0776

```

LTV	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]	
HouseholdSize	.0083786	.0002787	30.06	0.000	.0078323	.0089248
Age	-.0008475	.0000878	-9.65	0.000	-.0010197	-.0006754
Age2	1.31e-06	9.88e-07	1.32	0.185	-6.28e-07	3.25e-06
LTVo1	.1694516	.0001723	983.56	0.000	.1691139	.1697893
LTVo1_LTV	.9931186	.0004156	2389.78	0.000	.9923041	.9939331
LocalPriceChange	-.0227778	.0002433	-93.63	0.000	-.0232546	-.0223009
_cons	.8622562	.001889	456.47	0.000	.8585539	.8659586

Figur 10: First stage af IV regressionen med *LocalPriceChange* med instrument for *LTV*

Herefter laves den reducerede form, hvor outcome variabelen (*Moved*) er regresseret på instrumentet samt de andre forklarende variabler (uden den endogene variabel) i Figur 11.

Linear regression	Number of obs	=	402,385
	F(6, 402378)	=	336.20
	Prob > F	=	0.0000
	R-squared	=	0.0061
	Root MSE	=	.27306

Moved	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]	
LocalPriceChange	.0053269	.0007813	6.82	0.000	.0037956	.0068583
HouseholdSize	-.0025156	.0009782	-2.57	0.010	-.0044328	-.0005983
Age	-.0116457	.0003248	-35.86	0.000	-.0122822	-.0110092
Age2	.0001162	3.56e-06	32.63	0.000	.0001092	.0001232
LTVo1	.0136584	.0015583	8.77	0.000	.0106042	.0167125
LTVo1_LTV	-.0127815	.007777	-1.64	0.100	-.0280242	.0024611
_cons	.3492248	.0072174	48.39	0.000	.335079	.3633706

Figur 11: Reduceret form.

Her ses det, at når LTV stiger med 1 så falder $Moved$ med -0.0053269 . Herfra ses det, at når man dividerer resultatet fra reduceret form med resultatet fra first stage så får vi $\frac{0.0053269}{-0.0227778} = -0.23386$. Hvilket er præcist hvad IV regressionen med 2sls i Figur 8 gav.

3.3 C

For at undersøge om der er forskel i effekten af LTV på $Moved$ afhænger af, hvor mange voksne der er i husstanden, udvider vi modellen fra 3A. Vi udvider modellen med interaktionsledet mellem dummien for om man er en eller to i husstanden og LTV .¹ Da LTV er endogen er interaktionsledet mellem det og størrelsen på husholdningen også endogen, og derfor har vi to endogene variable og behøver de to instrumenter. Vi danner derfor et nyt instrument som er interaktionsledet mellem dummien for husstandsstørrelse og LocalPriceChange.

¹Det kan aflæses på figur 1 at HouseholdSize kun kan tage to værdier, og derfor allerede er en dummy. Dog har vi lavet det om til en dummy med værdien 1, hvis man er to i en husholdning og 0, hvis man er en, for lettere at kunne fortolke estimater.

Instrumental variables 2SLS regression	Number of obs	=	402,385
	Wald chi2(7)	=	2009.08
	Prob > chi2	=	0.0000
	R-squared	=	.
	Root MSE	=	.27402

Moved	Coefficient	Robust std. err.	z	P> z	[95% conf. interval]	
LTV	-.2418583	.0352083	-6.87	0.000	-.3108652	-.1728514
HouseholdSize2_LTV	.0392319	.0428799	0.91	0.360	-.0448111	.1232749
HouseholdSize	-.0361819	.0389817	-0.93	0.353	-.1125846	.0402208
Age	-.0118701	.0003289	-36.09	0.000	-.0125147	-.0112255
Age2	.0001169	3.61e-06	32.41	0.000	.0001099	.000124
LTVo1	.0533992	.0061052	8.75	0.000	.0414333	.0653651
LTVo1_LTV	.2094574	.0372331	5.63	0.000	.136482	.2824329
_cons	.5941585	.0552423	10.76	0.000	.4858856	.7024314

Figur 12: IV regression udvidet med interaktionsled mellem HouseholdSize dummy og LTV som endogen variabel, samt interaktionsled mellem HouseholdSize dummy og LocalPrice-Change som instrument.

Det ses på figur 12 at både HouseholdSize og interaktionsledet mellem dummien for husstandsstørrelsen og LTV er insignifikant på et 5% signifikansniveau. Hvis vi dog skulle se på værdierne alligevel kan vi se at for en husholdning på 2 så vil en forøgelse med 1 på LTV medfører et fald i sandsynligheden på at flytte på $-24.19\% + 3.92\% = -20.27\%$ hvor det for en husstand på en person er et fald på -24.19% . Dog som sagt er denne marginale effekt insignifikant, så vi kan ikke afvise at den er lig 0.

4 Problem 4

4.1 A

For at lave en udledning af IV estimatoren, skal vi først finde β_1 . Det vides man kan bruge z til at finde β_i værdierne i vores model. Vi kan gøre dette eftersom vores instrument z er gyldigt, da den opfylder kravet om instrument eksogenitet $\text{cov}(z, u) = 0$, samt instrument

relevans $\text{cov}(z, x) \neq 0$. Hermed findes β_1 ved følgende:

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

Herefter erstattes populationsparametrene med datamomenterne for kovarians og man får dermed nedenstående:

$$\hat{\beta}_1^{IV} = \frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} = \frac{\frac{\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2}}{\frac{\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2}} = \frac{\frac{\text{cov}(y, z)}{\text{var}(z)}}{\frac{\text{cov}(x, z)}{\text{var}(z)}}$$

I ovenstående udregning har vi forlænget brøken med variansen af z . Nu bruger vi følgende definitioner, der er givet ved formlen for reduceret form ($y = a_0 + a_1 z + \nu$) og first stage ($x = b_0 + b_1 z + e$), som kan skrives som følgende ved datamomenterne:

$$\hat{a}_1 = \frac{\text{cov}(y, z)}{\text{var}(z)} \quad \hat{b}_1 = \frac{\text{cov}(x, z)}{\text{var}(z)}$$

Men heraf ses det at ovenstående er præcist de to brøker som står i den ligning vi har fra vores datamomenter. Dermed får vi udtrykket

$$\hat{\beta}_1^{IV} = \frac{\frac{\text{cov}(y, z)}{\text{var}(z)}}{\frac{\text{cov}(x, z)}{\text{var}(z)}} = \frac{\hat{a}_1}{\hat{b}_1}$$

hvor $\hat{a}_1$ og $\hat{b}_1$ er OLS estimatorne for hhv. de to ligninger for den reducerede form og first stage. Den brøk man har fundet er derfor forskellen mellem to OLS estimater. Dvs. IV estimatet tolkes som effekten af z på y relativt til effekten af z på x .

4.2 B

Vi ved at IV estimatoren for $\hat{\beta}_1$ er givet ved følgende:

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

Ligeledes bliver vi bedt om at antage at z og x er dummier. Dermed ved vi at de eneste udfald der er for disse variable er $\{0, 1\}$. Hermed kan vi opskrive ovenstående formel som følgende:

$$\begin{aligned}\hat{\beta}_1^{IV} &= \frac{\frac{1}{n} [\sum_{z=1} y_i (1 - \bar{z})] + \frac{1}{n} [\sum_{z=0} y_i (0 - \bar{z})]}{\frac{1}{n} [\sum_{z=1} x_i (1 - \bar{z})] + \frac{1}{n} [\sum_{z=0} x_i (0 - \bar{z})]} \\ &= \frac{\frac{n_{x=1}}{n} \frac{1}{n_{x=1}} (1 - \bar{z}) \sum_{z=1} y_i - \frac{n_{x=0}}{n} \frac{1}{n_{x=0}} \bar{z} \sum_{z=1} y_i}{\frac{n_{x=1}}{n} \frac{1}{n_{x=1}} (1 - \bar{z}) \sum_{z=1} x_i - \frac{n_{x=0}}{n} \frac{1}{n_{x=0}} \bar{z} \sum_{z=1} x_i} = \frac{\frac{1}{n_{x=1}} \sum_{z=1} y_i - \frac{1}{n_{x=0}} \sum_{z=1} y_i}{\frac{1}{n_{x=1}} \sum_{z=1} x_i - \frac{1}{n_{x=0}} \sum_{z=1} x_i}\end{aligned}$$

Men de ovenstående udtryk i brøken har præcis samme formel som den empiriske middelværdi:

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Dermed kan det omskrives til følgende:

$$\hat{\beta}_1^{IV} = \frac{\frac{1}{n_{x=1}} \sum_{z=1} y_i - \frac{1}{n_{x=0}} \sum_{z=1} y_i}{\frac{1}{n_{x=1}} \sum_{z=1} x_i - \frac{1}{n_{x=0}} \sum_{z=1} x_i} = \frac{\bar{y}_{z=1} - \bar{y}_{z=0}}{\bar{x}_{z=1} - \bar{x}_{z=0}}$$

Ovenstående udtryk er det samme som der er givet i opgaven. Det vides at når $z = 0$ gælder

$$x = b_0 + e$$

$$y = a_0 + \nu$$

samt når $z = 1$

$$x = b_0 + b_1 z + e$$

$$y = a_0 + a_1 z + \nu$$

Hvis vi indsætter dette i ovenstående formel, får vi det følgende udtryk for IV estimatoren:

$$\frac{a_0 + a_1 z + \nu - (a_0 + \nu)}{b_0 + b_1 z + e - (b_0 + e)} = \frac{a_1 z}{b_1 z} = \frac{a_1}{b_1}$$

Men det ovenstående udtryk er det samme som vi fik førhen i opgave 4A. Det betyder at når vi bruger dummier i IV estimatotion, får vi det ækvivalente som IV estimatotion uden dummier.

4.3 C

Til at besvare denne opgave vil vi vise at β_1 estimatoren er konsistent. Først findes $\bar{y}_{z=1}$ og $\bar{y}_{z=0}$, som dernæst kan indsættes i ligning (6), man fandt i forrige opgave. Vi vil starte med

at udlede for den først nævnte:

$$\begin{aligned}\bar{y}_{z=1} &= \frac{\sum_{i=1}^{n_J} y_i z_i}{\sum_{i=1}^{n_J} z_i} = \frac{\sum_{i=1}^{n_J} (\beta_0^J + \beta_1^J x_i + u_i) z_i}{\sum_{i=1}^{n_J} z_i} = \beta_0^J + \beta_1^J \frac{\sum_{i=1}^{n_J} x_i z_i}{\sum_{i=1}^{n_J} z_i} + \frac{\sum_{i=1}^{n_J} u_i z_i}{\sum_{i=1}^{n_J} z_i} \\ &= \beta_0^J + \beta_1^J \bar{x}_{z=1} + \bar{u}_{z=1}\end{aligned}$$

På samme måde kan man udelede for $\bar{y}_{z=0}$, hvilket giver os $\bar{y}_{z=0} = \beta_0^J + \beta_1^J \bar{x}_{z=0} + \bar{u}_{z=0}$. Nu kan vi indsætte de to ovenstående $\bar{y}$ 'er i vores estimator vi fandt i forrige opgave:

$$\begin{aligned}\hat{\beta}_1^{IV} &= \frac{\bar{y}_{z=1} - \bar{y}_{z=0}}{\bar{x}_{z=1} - \bar{x}_{z=0}} = \frac{\beta_0^J + \beta_1^J \bar{x}_{z=1} + \bar{u}_{z=1} - (\beta_0^J + \beta_1^J \bar{x}_{z=0} + \bar{u}_{z=0})}{\bar{x}_{z=1} - \bar{x}_{z=0}} \\ &= \frac{\beta_1^J \bar{x}_{z=1} + \bar{u}_{z=1} - \beta_1^J \bar{x}_{z=0} - \bar{u}_{z=0}}{\bar{x}_{z=1} - \bar{x}_{z=0}} = \frac{\beta_1^J \bar{x}_{z=1} - \beta_1^J \bar{x}_{z=0} + \bar{u}_{z=1} - \bar{u}_{z=0}}{\bar{x}_{z=1} - \bar{x}_{z=0}} \\ &= \frac{\beta_1^J (\bar{x}_{z=1} - \bar{x}_{z=0})}{\bar{x}_{z=1} - \bar{x}_{z=0}} + \frac{\bar{u}_{z=1} - \bar{u}_{z=0}}{\bar{x}_{z=1} - \bar{x}_{z=0}} = \beta_1^J + \frac{\bar{u}_{z=1} - \bar{u}_{z=0}}{\bar{x}_{z=1} - \bar{x}_{z=0}} = \beta_1^J + \frac{1}{\bar{x}_{z=1} - \bar{x}_{z=0}} (\bar{u}_{z=1} - \bar{u}_{z=0}) \\ &= \beta_1^J + \frac{1}{\bar{x}_{z=1} - \bar{x}_{z=0}} \left(\frac{\frac{1}{n_J} \sum_{i=1}^{n_J} u_i z_i}{\frac{1}{n_J} \sum_{i=1}^{n_J} z_i} - \frac{\frac{1}{n_J} \sum_{i=1}^{n_J} u_i (1 - z_i)}{\frac{1}{n_J} \sum_{i=1}^{n_J} (1 - z_i)} \right)\end{aligned}$$

Herefter indsættes sandsynlighedsgrænser på alle datamomenter:

$$\begin{aligned}p \lim \left(\hat{\beta}_1^{IV} \right) &= \beta_1^J + p \lim \left(\frac{1}{\bar{x}_{z=1} - \bar{x}_{z=0}} \right) \left(\frac{p \lim \left(\frac{\frac{1}{n_J} \sum_{i=1}^{n_J} u_i z_i}{\frac{1}{n_J} \sum_{i=1}^{n_J} z_i} \right)}{p \lim \left(\frac{\frac{1}{n_J} \sum_{i=1}^{n_J} u_i (1 - z_i)}{\frac{1}{n_J} \sum_{i=1}^{n_J} (1 - z_i)} \right)} \right) \\ &= \beta_1^J + p \lim \left(\frac{1}{\bar{x}_{z=1} - \bar{x}_{z=0}} \right) \left(\frac{E(u_i z_1)}{p \lim \left(\frac{1}{n_J} \sum_{i=1}^{n_J} z_i \right)} - \frac{E(u_i) - E(u_i z_0)}{p \lim \left(\frac{1}{n_J} \sum_{i=1}^{n_J} (1 - z_i) \right)} \right) \\ &= \beta_1^J + p \lim \left(\frac{1}{\bar{x}_{z=1} - \bar{x}_{z=0}} \right) \left(\frac{E(u_i z_1)}{E(z_1)} - \frac{E(u_i) - E(u_i z_0)}{E(z_0)} \right)\end{aligned}$$

Eftersom det er givet at $\text{cov}(z, u) = 0$ gælder det at vi kan dele den forventede værdi op som $E(ZU) = E(Z)E(U)$. Dermed får vi:

$$\begin{aligned}p \lim \left(\hat{\beta}_1^{IV} \right) &= \beta_1^J + p \lim \left(\frac{1}{\bar{x}_{z=1} - \bar{x}_{z=0}} \right) \left(\frac{E(u_i z_1)}{E(z_1)} - \frac{E(u_i) - E(u_i z_0)}{E(z_0)} \right) \\ &= \beta_1^J + p \lim \left(\frac{1}{\bar{x}_{z=1} - \bar{x}_{z=0}} \right) \left(\frac{E(u_i)E(z_1)}{E(z_1)} - \frac{E(u_i) - E(u_i)E(z_0)}{E(z_0)} \right)\end{aligned}$$

Da den forventede værdi for fejllenedene er 0, betyder det at alle led i ligning med $E(u) =$ vil erstattes og vi få hermed til sidst:

$$p \lim \left(\hat{\beta}_1^{IV} \right) = \beta_1^J + p \lim \left(\frac{1}{\bar{x}_{z=1} - \bar{x}_{z=0}} \right) \left(\frac{0}{E(z_1)} - \frac{0 - 0}{E(z_0)} \right)$$

$$= \beta_1^J$$

Det vi nu har vist er at estimatoren er konsistent. Det betyder at vores IV estimat for β_1 er asymptotisk efficient, når vi har tilstrækkeligt store stikprøver for både gruppe A og B.

5 Problem 5

5.1 A

Vi sætter vores seed til 38 i Stata, hvorefter der køres blot én simulation ift. den angivende DGP med 500 observationer.

Variable	Obs	Mean	Std. dev.	Min	Max
betaIV	1	1.733011	.	1.733011	1.733011
betase	1	.1881532	.	.1881532	.1881532

Figur 13: 1 simulation

Her ses det på Figur 13 $\beta_1^J = 1.733011$ med en standard fejl på 0.1881532.

5.2 B

Denne gang gemmes hvilken β_1^J værdi der benyttes, og returnerer herefter gennemsnittet af den værdi i *BetaMean*. Kører simulationen med 1000 replikationer.

Variable	Obs	Mean	Std. dev.	Min	Max
betaIV	1,000	1.974066	.157511	1.105959	2.301174
betaMean	1,000	1.49787	.0221232	1.424	1.564
beta_reduced	1,000	.9928869	.2460593	.1766239	2.041497

Figur 14: 1000 replikationer, samt en estimation for den reducerede form.

Her ses det på figur 14, at gennemsnittet af beta-værdierne er ca. 0.5 mindre end vores beta fra IV regressionen.

Dette resultat kan sammenlignes med den analytiske formel fra opgave 4.a

$$\beta^{IV} = \frac{\hat{a}_1}{\hat{b}_1}$$

Vi ved fra opgaven at $E(\alpha_1^J) = 0.5$ og vi får $\beta^{IV} \approx 2$ og derfra vides det at

$$2 = \frac{\hat{a}_1}{0.5} \iff \hat{a}_1 = 1$$

Vi har yderligere i vores simulation lavet den reducerede form, som kan ses i figur 14 at den reducerede form $\beta_{reduced} \approx 1$, som stemmer overens med hvad vi regnede med ovenfor.

5.3 C

Her sættes $\alpha_1^A = 1$ og simulationen køres igen med 1000 replikationer.

Variable	Obs	Mean	Std. dev.	Min	Max
betaIV	1,000	1.48726	.1129747	1.035986	1.809871
betaMean	1,000	1.50073	.0221343	1.426	1.562
beta_reduced	1,000	1.494971	.2350975	.7058429	2.275075

Figur 15: 1000 replikationer med $\alpha_1^A = 1$

Her ses det på figur 15, at de to beta-værdier tilnærmer sig det samme med kun $1.50073 - 1.48726 = 0.01347$ forskel. Relativt set er forskellen kun $\frac{0.01347}{\frac{1.50073 + 1.48726}{2}} \cdot 100\% = 0.90167\%$ hvilket må siges at være en lille forskel.

Resultatet kan på samme måde sammenlignes som gjort i 5B, denne gang blot med $\alpha_1^A = 1$

$$\beta^{IV} = \frac{\hat{a}_1}{1} \iff \hat{a}_1 = 1.5$$

Og ud fra Figur 15 ses det at $\hat{a}_1 \approx 1.5$.

Når $\alpha_1^A = 1$ er vores endogene variabel x korreleret med instrumentet z, hvorimod når $\alpha_1^A = 0$ er den endogene variabel og instrumentet ikke korreleret, hvilket ikke opfylder antagelse om relevans. Og vi får derfor i 4B en positiv bias.