BFIN4025 - Big data i eiendomsfinans

Aras Khazal

NTNU Handelshøyskolen 2021

BFIN4025 - Big data i eiendomsfinans

- Lineær regresjonsanalyse (MKM)
- Endogenitet problem
  - Sample selection (Non- random sample)
  - Utelatt variable (Omitted variable bias (OVB))
  - Målefeil: (Measurement error)
  - Samtidighet (Simultaneity bias/reverse causation)

Forrige uka

I dag

Sample Selection Bias

Hoved modell uten å kontrollere for Seleksjon problem:

Labeled-only sample

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 A_i + \beta_3 B_i + \beta_4 C_i + \beta_5 D_i + \beta_6 E_i + \beta_7 F_i + e_i$$

Hoved modell med kontroll for Seleksjon problem:

Heckman model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 A_i + \beta_3 B_i + \beta_4 C_i + \beta_5 D_i + \beta_6 E_i + \beta_7 F_i + \boldsymbol{\beta_8 \hat{\lambda}_i} + e_i$$

Negativ koeffisient til IMR i leiepris modellen?

- Faktorene bak energimerking er korrelert mer lav leieprisen

Positiv koeffisient til IMR i salgspris modellen?

- Faktorene bak energimerking er korrelert mer høy leieprisen

Table A2: Heckman selection models for rental and sales prices

	Rental		Sales	
	Labeled-only sample	Heckman model	Labeled-only sample	Heckman model
A	0.0470*** (22.25)	0.0465*** (21.85)	0.0762*** (3.56)	0.0724*** (3.36)
B	0.0390*** (18.13)	0.0380*** (17.58)	0.0313*** (3.79)	0.0292*** (3.54)
C	0.0356*** (17.70)	0.0353*** (17.46)	0.0374*** (7.95)	0.0363*** (7.72)
D	0.0233*** (11.85)	0.0235*** (11.87)	0.0275*** (8.34)	0.0263*** (8.06)
E	-0.0008 (-0.38)	-0.0012 (-0.57)	0.0147*** (6.24)	0.0139*** (5.97)
F	-0.0094*** (-4.27)	-0.0093*** (-4.20)	0.0080*** (4.49)	0.0075*** (4.26)
G				
Mills ratio		-0.0248** (-2.69)		0.4139*** (19.80)
Observations	108,276	106,715	92,416	92,338
Adjusted R^2	0.760	0.759	0.970	0.970
RMSE	0.165	0.165	0.099	0.099

Note: Table A2 reports the cross-sectional HDFE estimation for the rental data and panel FE estimation for the sales data. The dependent variable is the natural logarithm of the monthly rental price or the sales price. The default for EPC-labels is G-labeled dwellings. Heteroskedasticity robust *t*-statistics are in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$.

Konklusjon

- ❑ Vær forsiktig når du velger utvalget til analyse!
  - ➢ Er det noe gruppe eller område som er eliminert fra samplet?
  - ➢ Hva sier litteraturen om din variabel?
  - ➢ Er samplet nok representativt for populasjon?
- ❑ Dersom man mistenker at én eller flere variable har seleksjonsproblemer:
  - ➢ Koeffisienten(e) kan ikke tolkes som kausalsammenheng

Hva skal man gjøre?

- ❑ Finn én instrument som tilfredsstille de to tidligere nevnt forutsetninger

  → Bruke Heckman Two-Step modell for å håndtere problemet
- ❑ Hvis det er ingen andre kilde for endogenitet (skal gå gjennom neste uka)→ Kan koeffisienten(e) tolkes som kausalsammenheng(er) ☺

Endogenitet

- Sample Selection Bias
- Utelatt variable (Omitted variable bias (OVB))
- Målefeil: (Measurement error)
- Samtidighet (Simultaneity bias/reverse causation)

Husk at forutsetning 5: $\boldsymbol{cov(X, e) = 0}$

Utelatt variable (Omitted variable bias (OVB))

Faktisk populasjonsmodell: $P_i = \beta_0 + \beta_1 Areal_i + \beta_2 Merket_i + \beta_3 Alder_i + e_i$

Men vi estimerer følgende modell: $\widehat{P_i} = \widehat{\beta_0} + \widehat{\beta_1} Areal_i + \widehat{\beta_2} Merket_i + \widehat{e_i}$

Hvis *Merket* og *Alder* er korrelerte: $\boldsymbol{cov}(Merket_i, Alder_i) \neq 0$

Hvorfor inkluderer vi ikke *Alder*?

→ Vil *Merket* og *Residual* bli korrelerte, siden *Alder* **ligger inn i *Residual***:

*Forklaringsvariabelen, merket, er **endogen** fordi* $\boldsymbol{cov}(Merket_i, e_i) \neq 0$ (Brudd på Forutsetning 5)

Den estimerte effekt av *Merket* på *Prisen* vil være biased : $\widehat{\beta_2} = (\beta_2 + \boldsymbol{\beta_3})$

Det vil si at den estimerte effekten er ikke lik faktisk effekt $\widehat{\beta_2} \neq \beta_2$

Husk at forutsetning 5: $\boldsymbol{cov(X, e) = 0}$

Målefeil: (Measurement error)

Faktisk populasjonsmodell: $P_i = \beta_0 + \beta_1 Areal_i + \beta_2 Merket_i + \beta_3 Alder_i + e_i$

Anta at vi har data om *merket* : $merket_i = Merket_i + u_i$

Den faktisk *Merket* er: $Merket_i = (merket_i - u_i)$

Vi ønsker å estimerer : $\widehat{P_i} = \widehat{\beta_0} + \widehat{\beta_1} Areal_i + \widehat{\beta_2} Merket_i + \widehat{\beta_3} Alder_i + \widehat{e_i}$

Men vi faktisk estimerer: $\widehat{P_i} = \widehat{\beta_0} + \widehat{\beta_1} Areal_i + \widehat{\beta_2} (merket_i - u_i) + \widehat{\beta_3} Alder_i + \widehat{e_i}$

$\widehat{P_i} = \widehat{\beta_0} + \widehat{\beta_1} Areal_i + \widehat{\beta_2} merket_i + \widehat{\beta_3} Alder_i + (\widehat{e_i} - \beta_2 u_i)$

Husk at forutsetning 5: $\boldsymbol{cov(X, e) = 0}$

Målefeil: (Measurement error)

Vi har ikke data om u :

$$\widehat{P_i} = \widehat{\beta_0} + \widehat{\beta_1}\, Areal_i + \widehat{\beta_2}\, merket_i + \widehat{\beta_3}\, Alder_i + \underbrace{(\widehat{e_i} - \beta_2 u_i)}$$

$$\widehat{P_i} = \widehat{\beta_0} + \widehat{\beta_1}\, Areal_i + \widehat{\beta_2}\, merket_i + \widehat{\beta_3}\, Alder_i + \boldsymbol{\widehat{\epsilon_i}}$$

$$\boldsymbol{\widehat{\epsilon_i}} = (\widehat{e_i} - \beta_2 u_i)$$

$$\boldsymbol{cov}(merket_i, u_i) \neq 0$$

→ $\boldsymbol{cov}(merket_i, \epsilon_i) \neq 0$ (Brudd på Forutsetning 5)

*Forklaringsvariabelen, merket, er **endogen**!!*

Den estimerte effekt av *merket* på *Prisen* vil være biased: $\widehat{\beta_2} \neq \beta_2$

Husk at forutsetning 5: $\boldsymbol{cov(X, e) = 0}$

Samtidighet (Simultaneity bias/reverse causation)

- Denne typen skjevhet skjer i modellene der variable er samtidige determinerte
- Uavhengig variable påvirker på avhengig variabel
- Men den avhengig variabelen også påvirker på den uavhengige (Reverse Causation)

Regresjon (kausalsammenheng)

Samtidighet (~~kausalsammenheng~~)

Husk at forutsetning 5: $\boldsymbol{cov(X, e) = 0}$

Samtidighet (Simultaneity bias/reverse causation)

Tilbudskurven: $Q_T = \beta\, P_T + e$ (1)

Etterspørselskurven: $P_E = \alpha\, Q_E + u$ (2)

På markedslikevekt: $Q_T = Q_E$ og $P_T = P_E$

Tilbudskurven kan omskrives slik: $Q = \beta\,(\alpha\, Q + u) + e$

$Q = \beta\,\alpha\, Q + \beta\, u + e$ (*)

Vi vet fra tilbudskurven (1) at: $\boldsymbol{cov}(Q, e) \neq 0$

Dette betyr at forutsetning 5 er brudd på modell (*) → estimerte koeffisienten til Q er biased!!!

Endogenitet

- ❑ Utelatt variable (Omitted variable bias (OVB))
- ❑ Målefeil: (Measurement error)
- ❑ Samtidighet (Simultaneity bias/reverse causation)

Løsning?

Instrumental Variable method (IV) / Two Stage Least Squared method (2SLS)

Ved hjelp av **gyldig** Instrument(er) kan vi dekomponere den endogen variabelen i;

- ➢ Én endogen del som kan bli korrelerte med feilleddet
- ➢ Og én eksogen del som er ukorrelerte med feilleddet, som kan bli brukt (erstatte endogen variabel) i hoved modellen for å oppnå uskjevhet i estimerte koeffisient (unbiased)

Two Stage Least Squared method (2SLS)

Regresjonsmodell: $P_i = \beta_0 + \beta_1 X_{1i} + \boldsymbol{\beta_2}\ \boldsymbol{X_{2i}} + e_i$

- Anta at vi har sterk grunn til å tro på at $\boldsymbol{X_2}$ er endogen
- Vi vet at OLS kan ikke brukes fordi den produserer bias estimat ($\boldsymbol{\beta_2}$)
- Alternativt, vi kan bruke 2SLS istedenfor OLS
- For å kunne implementere 2SLS metoden, trenger vi én eller flere instrument(er)
- En instrument er en ny variable som er med å forklare den endogen $\boldsymbol{X_2}$
- Kun én instrument er nok, men dersom man kan få taket i flere så er det bedre

Two Stage Least Squared method (2SLS)

Regresjonsmodell: $P_i = \beta_0 + \beta_1 X_{1i} + \boldsymbol{\beta_2}\ \boldsymbol{X_{2i}} + e_i$ (Strukturellmodell)

- Anta at vi har fått tilgang til nye informasjon om $\boldsymbol{X_2}$
- Dvs. at vi fikk en ny variable som forklarer $\boldsymbol{X_2}$ (korrelert), men som ikke påvirke på $\boldsymbol{P}$
- Vi kaller det instrument $\boldsymbol{Z}$
- I tillegg til de 5 klassiske OLS forutsetninger, trenger vi to nye forutsetninger:
- Relevance $\boldsymbol{cov(X_2, Z) \neq 0}$ (instrument er ***korrelert*** med endogen variabel)
- Exogeneity $\boldsymbol{cov(Z, P) = 0}$ (instrument er ***ukorrelert*** med avhengig variabel)

Two Stage Least Squared method (2SLS)

Regresjonsmodell: $P_i = \beta_0 + \beta_1 X_{1i} + \boldsymbol{\beta_2}\ \boldsymbol{X_{2i}} + e_i$ (Strukturellmodell)

Steg 1: $\boldsymbol{X_{2i}} = \alpha_0 + \alpha_1\ X_{1i} + \alpha_2\ Z_i + e_i$

Del 1 Del 2

Steg 1: Vi inkluderer alle uavhengige variable og instrument(er)

Vi kjører OLS i Steg 1 for å tallfeste koeffisientene → bruk ***Del 1*** til å predikere X_{2i}

$$\widehat{X_{2i}} = \widehat{\alpha_0} + \widehat{\alpha_1}\ X_{1i} + \widehat{\boldsymbol{\alpha_2}}\ \boldsymbol{Z_i}$$

Relevance assumption $\boldsymbol{cov(X_2, Z) \neq 0}$

- At instrument er korrelerte med endogen variable innebærer at → $\alpha_2 \neq 0$
- Én instrument: ***t*-statistisk** til $\alpha_2 > 3$ (Relevance assumption er OK ☺)
- Mer enn én instrument: ***F*-statistisk for** alle instrumenter > 10 (Relevance assumption er OK ☺)

Two Stage Least Squared method (2SLS)

Regresjonsmodell: $P_i = \beta_0 + \beta_1 X_{1i} + \boldsymbol{\beta_2}\ \boldsymbol{X_{2i}} + e_i$ (Strukturellmodell)

Steg 2: $\boldsymbol{P_i} = \beta_0 + \beta_1 X_{1i} + \widehat{\beta_2}\ \widehat{X_{2i}} + e_i$

$\widehat{X_{2i}}$ er den predikert X_{2i} fra Steg 1

Steg 2: Er den samme som strukturellmodellen, men med én forskjell!!

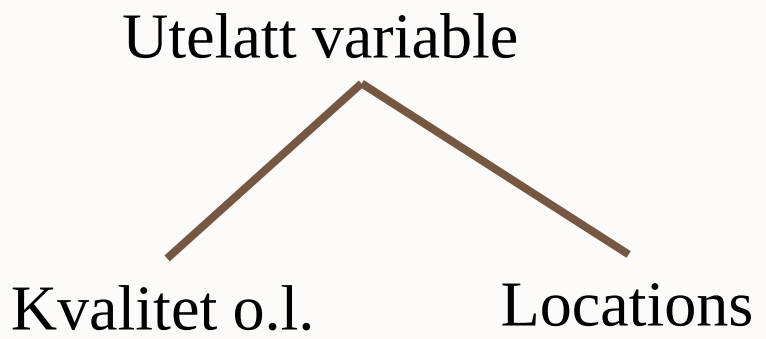
Exogeneity $\boldsymbol{cov(Z, P) = 0}$

- Kan ikke testes hvis vi har kun én instrument
- Kan *delvis* testes hvis vi har mer enn én instrument (*over identification test*)

- Vi kjører OLS i Steg 2 for å tallfeste koeffisientene
- Hvis både *relevance* og *exogeneity* forutsetninger er OK
- Den estimerte effekten av *X* er ***unbiased*** og kan tolkes som kausalsammenheng, ($\widehat{\beta_2} = \boldsymbol{\beta_2}$)

Endogenitetsproblem

Anta at en av **hedoniske** variable er endogen pga. korrelasjon mellom utelatte variable og en (eller flere) forklaringsvariabel(er)



Løsning?

1. Inkluder alle utelatte variable i modellen hvis det er mulig!

- ❑ Hvis utelatte variabel(er) er kvalitet på bolig
  → Inkludere én eller flere variabler som måler kvalitet i regresjonsmodellen som tillegg forklaringsvariabel
- ❑ Hvis utelatte variabel(er) er beliggenhet til bolig
  →Inkluder variable(er) for location eller inkluder ***binære*** variabler for beliggenhet
- ❑ Hvis utelatte variabel(er) er kvalitet på bolig og beliggenhet til bolig
  → Inkludere én eller flere variabler som måler kvalitet og ***binære*** variabler for beliggenhet

Det er viktig å vite om utelatte variable for å kunne bruke riktig tilnærming

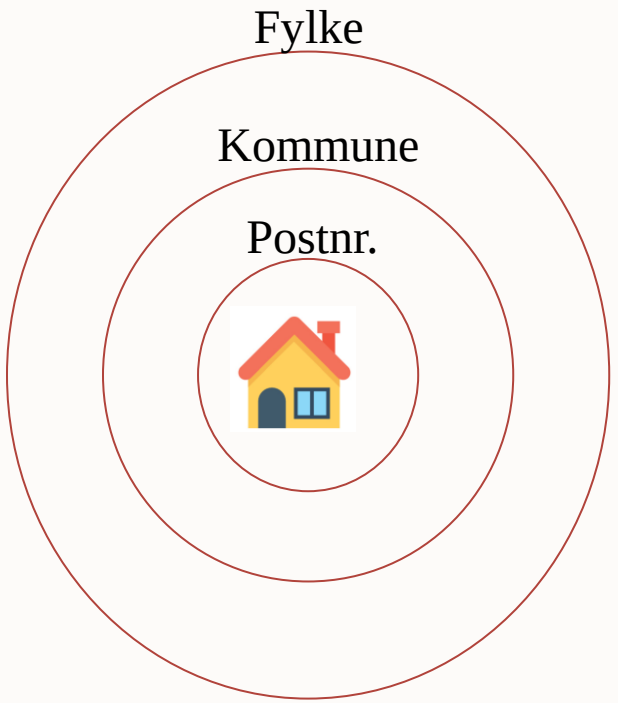
Endogenitetsproblem

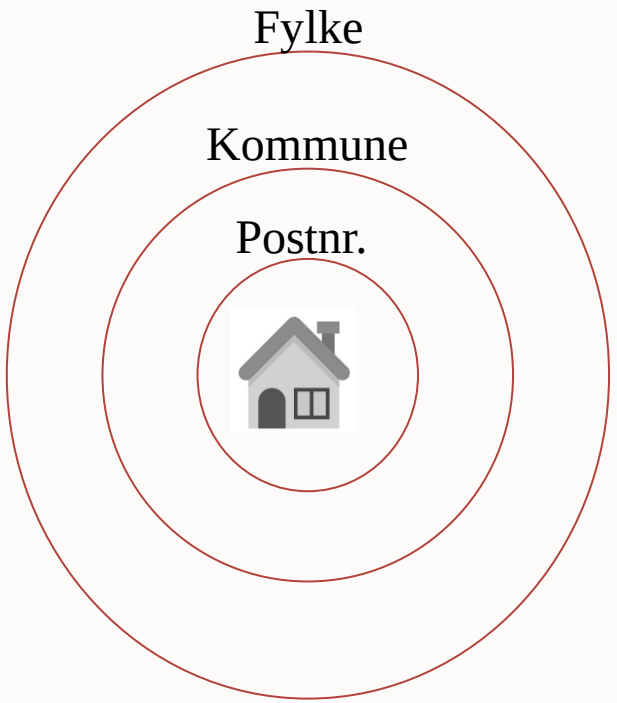
Utelatt variable

Kvalitet o.l.

Locations

- Alder
- Bedre material
- Og lignende …





$$P_i = \beta_0 + \beta_1 X_{1i} + \boldsymbol{\beta_2\ X_{2i}} + (u_i + X_3^{postnr} + X_4^{kommune} + X_5^{fylke} + X_6^{kvalitet})$$

$$P_i = \beta_0 + \beta_1 X_{1i} + \boldsymbol{\beta_2\ X_{2i}} + e_i$$

Korrelerte !! $cov(X_{2_i}, e) \neq 0$ (Foruts. 5 ☹)

Løsning?

Vi inkluderer de variable inn i regresjonsmodellen som følgende:

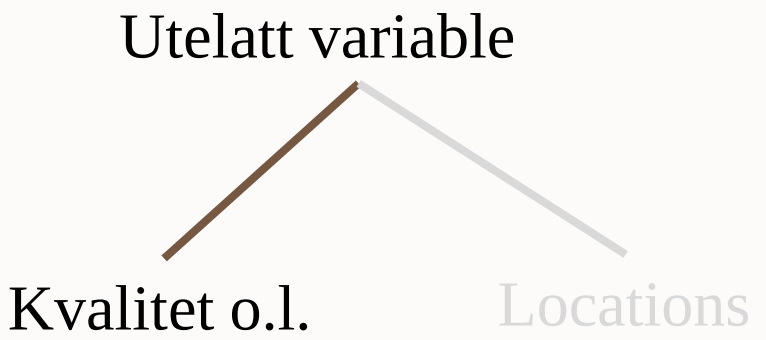
$$P_i = \beta_0 + \beta_1 X_{1i} + \beta_2\ X_{2i} + \beta_3\ X_{3i} + \beta_4\ X_{4i} + \beta_5\ X_{5i} + \beta_6 X_{6i} + e_i$$

Hvis «alle» relevante variabler er inkludert i modellen, blir det ingen korrelasjon mellom forklaringsvariabel(er) og restleddet lenger

→ **Ingen** endogenitet, altså koeffisienten(e) kan tolkes som ***kausale***sammenheng(er)

Endogenitetsproblem

Hva skal vi gjøre hvis vi ikke har tilgang til data om kvalitet (eller location)??



Løsning?

2. Bruk én metode som håndterer dette, som 2SLS tilnærming

Husk at:

Vi trenger én eller flere nye variable (instrumenter) som skal tilfredsstille både

relevance and *exogenity* forutsetninger, i tillegg til OLS klassiske forutsetninger.

Two Stage Least Squared method (2SLS)

Eksempel fra pensum paperet

Strukturellmodell:

$$ln\ price_{icmz} = \tau(Label_{icmz}) + X_{icmz}\,\gamma + \alpha_c + \mu_m + \omega_z + \lambda_t + u_{icmz},$$

- $ln\ price_{icmz}$: Avhengig var.
- $\tau(Label_{icmz})$: Mistenkt endogen var.
- $X_{icmz}\,\gamma$: Andre hedoniske (uavhengige) var.
- $\alpha_c + \mu_m + \omega_z + \lambda_t$: Location var. (binære var. for geografiske beliggenhet)
- u_{icmz}: Feilleddet

At ***Label*** variabelen er endogen betyr: $\boldsymbol{cov(Label, u) \neq 0}$ (brudd i forutsetning 5)

Vi kan ikke bruke OLS direkte på strukturellmodellen fordi vi vet at koeffisienten τ blir ***biased*** ☹

Løsning: Vi bruker 2SLS for å kunne få ***ubiased*** estimat, τ ☺

Two Stage Least Squared method (2SLS)

Strukturellmodell: $$ln\, price_{icmz} = \tau(Label_{icmz}) + X_{icmz}\,\gamma\ + \alpha_c + \mu_m + \omega_z + \lambda_t + u_{icmz},$$

Vi trenger minst én instrument som tilfredsstiller de to forutsetningene til 2SLS

$$cov(Label, Z) \neq 0 \quad \text{og} \quad cov(Z, price) = 0$$

Instrument:

Vi får informasjon om total antall nye energimerket boliger på kommune nivå, vi kaller det *NNC*

Steg 1 (OLS): $$Label_{icmz} = \beta(NCC_m) + X_{icmz}\,\gamma\ + \alpha_c + \mu_m + \omega_z + \lambda_t + e_{icmz},$$

Vi predikere $\widehat{\boldsymbol{Label}_{icmz}}$ for å kunne bruke den i Steg 2

Steg 2 (OLS): $$ln\, price_{icmz} = \tau\ (\widehat{Label}_{icmz}) + X_{icmz}\,\gamma\ + \alpha_c + \mu_m + \omega_z + \lambda_t + u_{icmz},$$

Two Stage Least Squared method (2SLS)

- Koeffisienten fra OLS er overdrevet (positve biased)
- Koeffisienten fra IV (2SLS) uten å kontrollere for locations er feil!!
- HDFE er **OLS** med full kontroll for locations
- IV-HDFE er **2SLS** med full kontroll for locations
- *F*-statistikk er >10 som betyr at **relevance** forutsetningen er OK
- Den er ingen forskjell mellom HDFE og IV-HDFE?
  - Det eneste kilde for Endogenitet er location (og ikke kvalitet)

Table 7: Identifying the unobserved location and quality heterogeneity for rental prices

	OLS	IV	HDFE	IV-HDFE
Label	0.0766***	8.2406***	0.0421***	0.0410***
	(81.28)	(36.21)	(65.84)	(60.35)
Control variables	✓	✓	✓	✓
Fixed location effects	X	X	✓	✓
Excluded instrument	X	*NNC*	X	*NNC*
Observations	669,894	669,366	669,448	665,002
Adjusted R^2	0.403	0.531	0.725	0.724
RMSE	0.286	0.253	0.194	0.194
First stage *F*-statistic		1,147.18		1,104.90

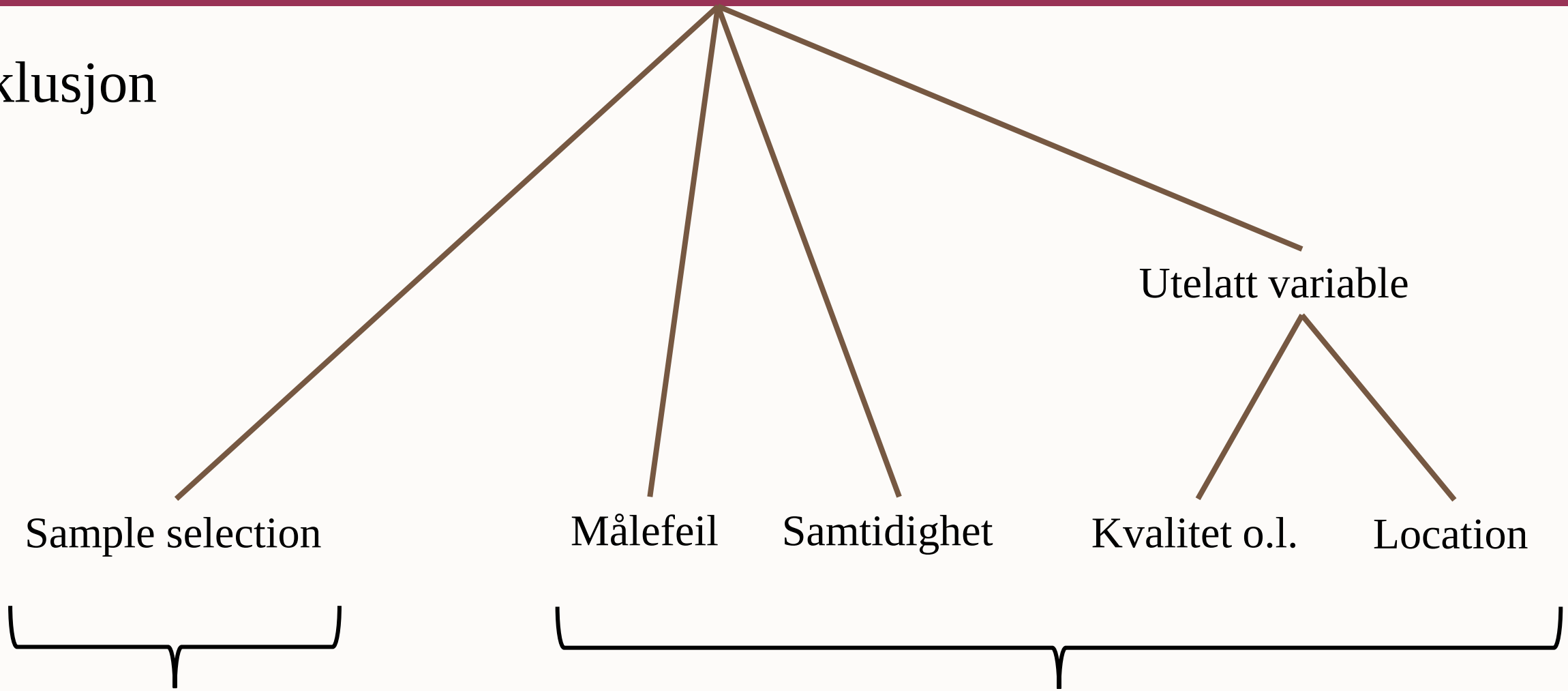
Two Stage Least Squared method (2SLS)

- Koeffisienten fra OLS er overdrevet (positve biased)
- Koeffisienten fra IV uten å kontrollere for locations er feil!!
- *F*-statistikk er >10 som betyr at **relevance** forutsetningen er OK
- Den er ingen forskjell mellom HDFE og IV-HDFE?
  - Det eneste kilde for Endogenitet er location (og ikke kvalitet)

Table 8: Identifying the unobserved location and quality heterogeneity for sales prices

	OLS	IV	HDFE	IV- HDFE
Label	0.0612***	6.657***	0.0303***	0.0297***
	(51.12)	(58.56)	(40.87)	(35.12)
Control	✓	✓	✓	✓
Fixed location effects	X	X	✓	✓
Panel fixed effects	X	X	X	X
Excluded instrument	X	*NNC*	X	*NNC*
Observations	747,387	745,774	747,006	745,247
Adjusted R^2	0.474	0.525	0.817	0.795
RMSE	0.429	0.401	0.253	0.259
Excluded instrument *F*-statistic		1,772.41		691.69

Konklusjon


Utelatt variable
Sample selection
Målefeil
Samtidighet
Kvalitet o.l.
Location
Heckman Two-Step modell
2SLS metode + full kontroll for location