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)

I dag

Neste uka

Lineær regresjonsanalyse

Analyse mellom to eller flere variabler; der den ene er definert som respons og en eller flere andre er definert som forklaringsvariabler

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + e_i$$

Respons/avhengig variabel

Konstantledd

Forklaringsvariabler/ uavhengige variabler

Restledd

Lineær regresjonsanalyse

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + e_i$$

Vi ønsker å tallfeste (estimere) koeffisientene β_1, β_2, ….

β_1 er effekten av X_1 på Y

β_2 er effekten av X_2 på Y

Lineær regresjonsanalyse

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

Vi bruker Minste Kvadraters Metode MKM (Ordinary Least Squares, OLS)

→ for å tallfeste (estimere) koeffisientene β_1, β_2

Estimering $Y_i = \widehat{\beta_0} + \widehat{\beta_1} X_{1i} + \widehat{\beta_2} X_{2i} + e_i$

Prediksjon $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_{1i} + \widehat{\beta_2} X_{2i}$

Residual $e_i = Y_i - \widehat{Y_i}$

OLS er den linjen som minimerer summen av kvadrerte residualer (SSR)

$$\sum e_i^2 = (Y_i - \widehat{Y_i})^2$$

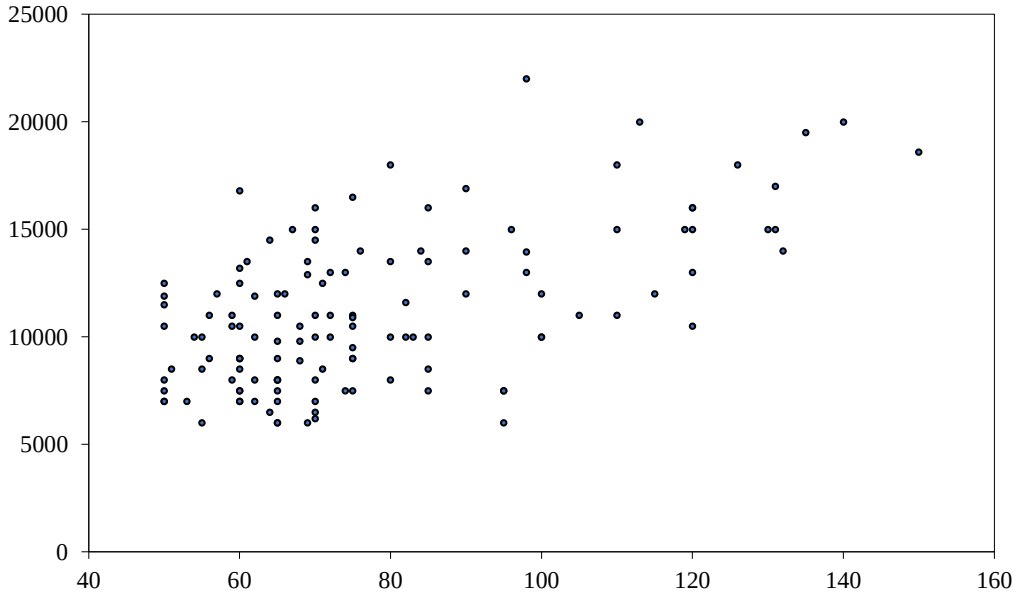
$$Pris_i = \beta_0 + \beta_1(Areal_i) + \beta_2(Alder_i) + \cdots + Feilledd_i$$

ID	Dato	Pris	Antall rom	Areal	Alder	Fylke	Kommune	Post no	Adress
1	01.01.2011	10000	2	80	15	Oslo	Oslo	320	Kirkegate 52
2	01.01.2011	15000	4	130	25	Trondelag	Trondheim	7093	Sandmovegen 7
3	01.01.2011	8000	1	55	32	Viken	Nøtterøy	3032	Teieveien 2A
4	02.01.2011	11000	2	70	22	Agder	Arendal	4203	Agdergate 11
5	02.01.2011	15000	2	100	15	Oslo	Oslo	320	Juni Plassen 1

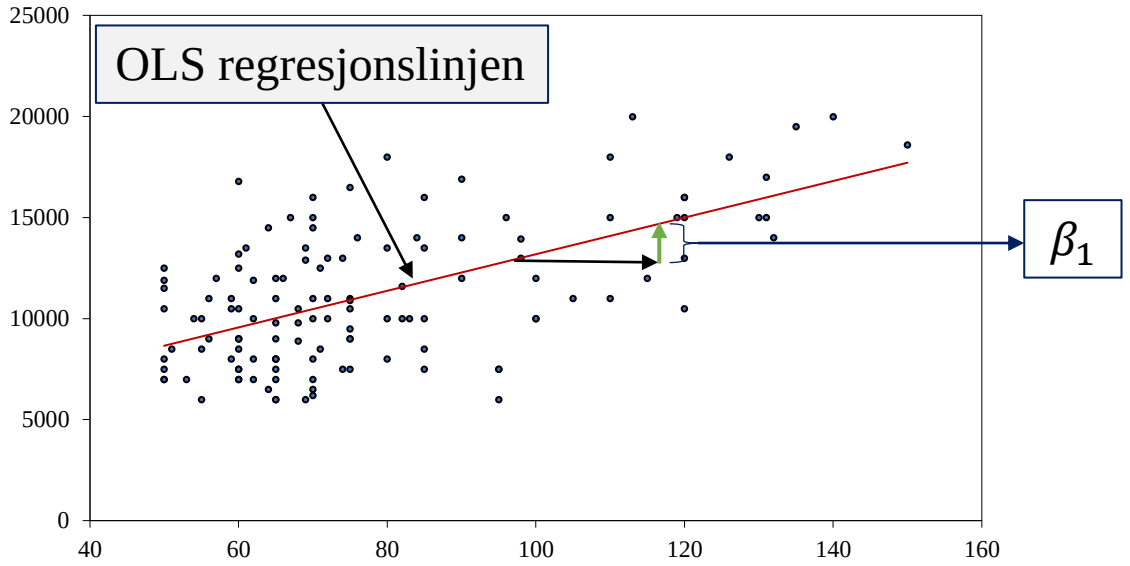
Positiv virkning på leieprisen:

→ Økning i arealet fører til økt leieprisen

Areal og leieprisen



Areal virkning på leieprisen



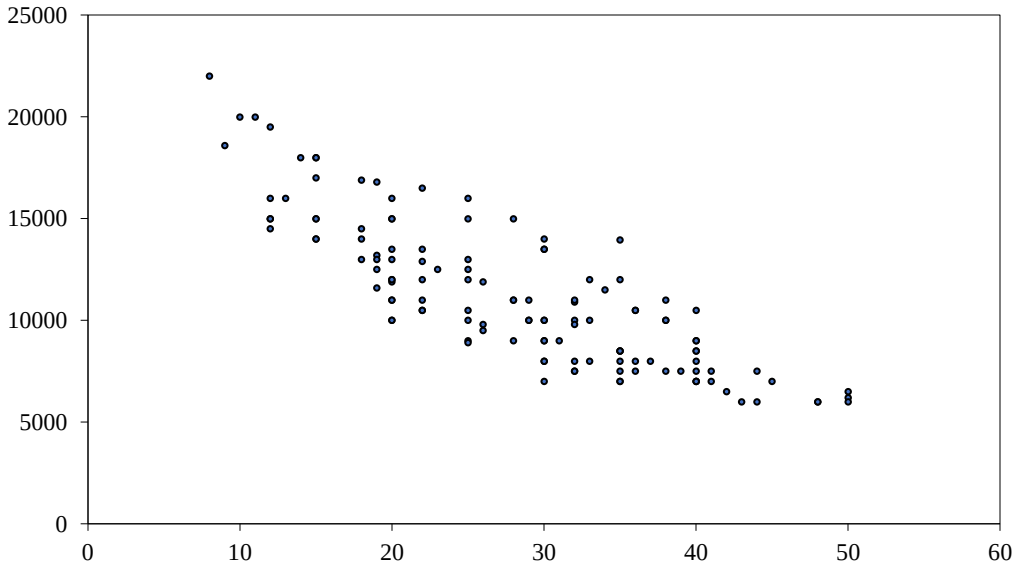
$$Pris_i = \beta_0 + \beta_1(Areal_i) + \beta_2(Alder_i) + \cdots + Feilledd_i$$

ID	Dato	Pris	Antall rom	Areal	Alder	Fylke	Kommune	Post no	Adress
1	01.01.2011	10000	2	80	15	Oslo	Oslo	320	Kirkegate 52
2	01.01.2011	15000	4	130	25	Trondelag	Trondheim	7093	Sandmovegen 7
3	01.01.2011	8000	1	55	32	Viken	Nøtterøy	3032	Teieveien 2A
4	02.01.2011	11000	2	70	22	Agder	Arendal	4203	Agdergate 11
5	02.01.2011	15000	2	100	15	Oslo	Oslo	320	Juni Plassen 1

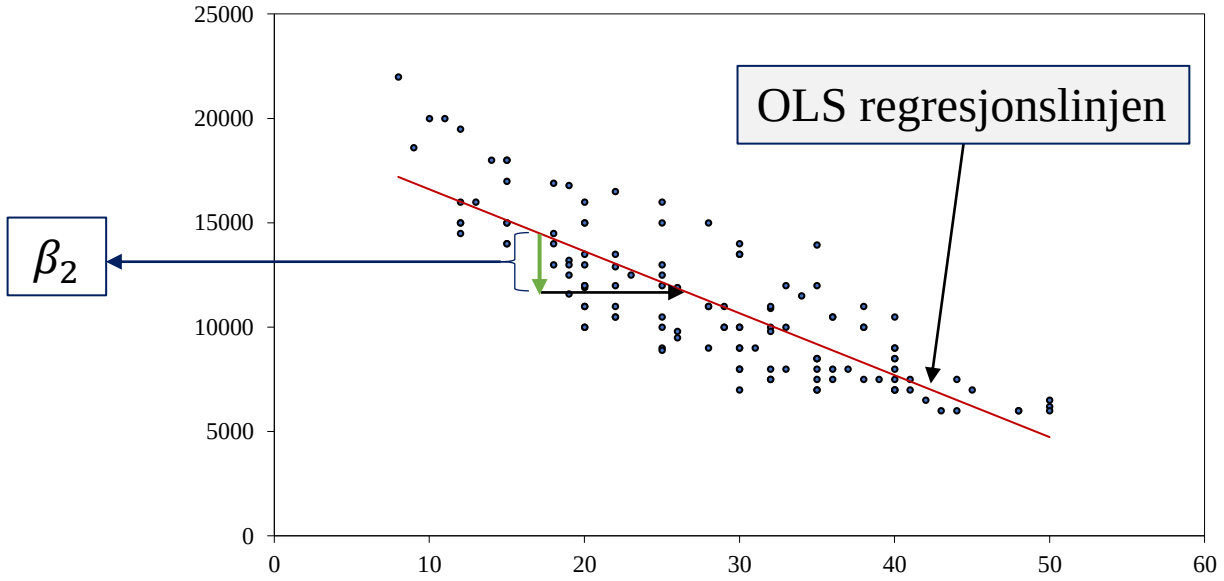
Negativ virkning på leieprisen:

→ Økning i Alder fører til redusert leieprisen

Alder og leieprisen



Alder virkning på leieprisen



Tverrsnitt (Cross sectional Data)

ID	Dato	Pris	Antall rom	Areal	Alder	Fylke	Kommune	Post no	Adress
1	01.01.2011	10000	2	80	15	Oslo	Oslo	320	Kirkegate 52
2	01.01.2011	15000	4	130	25	Trondelag	Trondheim	7093	Sandmovegen 7
3	01.01.2011	8000	1	55	32	Viken	Nøtterøy	3032	Teieveien 2A
4	02.01.2011	11000	2	70	22	Agder	Arendal	4203	Agdergate 11
5	02.01.2011	15000	2	100	15	Oslo	Oslo	320	Juni Plassen 1

$$Pris_i = \beta_0 + \beta_1(Rom_i) + \beta_2(Alder_i) + \cdots + Feilledd_i$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + e_i$$

Panel Data

ID	Dato	Pris	Antall rom	Areal	Alder	Fylke	Kommune	Post no	Adress
1	01.01.2011	10000	2	80	15	Oslo	Oslo	320	Kirkegate 52
1	01.01.2013	11000	2	80	17	Oslo	Oslo	320	Kirkegate 52
2	01.01.2011	8000	1	55	32	Viken	Nøtterøy	3032	Teieveien 2A
2	02.01.2012	8500	1	55	33	Viken	Nøtterøy	3032	Teieveien 2A
3	02.01.2011	15000	2	100	15	Oslo	Oslo	320	Juni Plassen 1
3	02.01.2017	17500	2	100	21	Oslo	Oslo	320	Juni Plassen 1

$$Pris_{it} = \beta_0 + \beta_1(Rom_{it}) + \beta_2(Alder_{it}) + \cdots + Feilledd_i t$$

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + e_{it}$$

Regresjon vs. korrelasjon

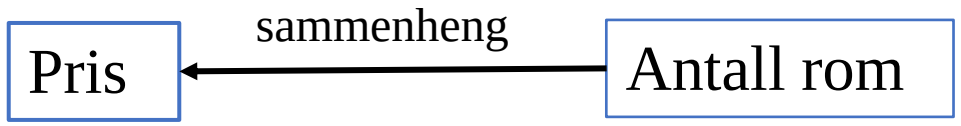
ID	Dato	Pris	Antall rom
1	01.01.2011	10000	2
2	01.01.2011	15000	4
3	01.01.2011	8000	1
4	02.01.2011	11000	2
5	02.01.2011	15000	2

Avhengig variabel: Pris (Y_i)

Uavhengig/forklaring variabel: Pris (X_i)

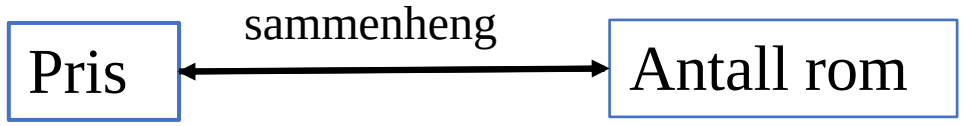
Regresjon (kausalsammenheng)

$$\text{Regresjonskoeffisient} = \frac{Cov(Y_i, X_i)}{var(X_i)}$$



Korrelasjon

$$\text{Korrelasjonskoeffisient} = \frac{cov(Y_i, X_i)}{\sigma_{Y_i}\sigma_{X_i}}$$



Forutsetninger for analyse: Klassisk lineær regresjonsmodell

1) Regresjons modellen er *lineær* i koeffisientene og er korrekt spesifisert
2) *Homoskedastisitet*: Restleddet har konstant varians (hvis ikke konstant → *heteroskedastisitet*)
3) Ingen *Multikollinearitet*: Ingen av forklaringsvariablene kan skrives som prefekt lineær funksjon av none av de andre
4) Restleddet er *normalfordelt*
5) **Alle forklaringsvariablene er ukorrelerte med restleddet (Zero conditional mean)**

Endogenitet problem

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

} (Forutsetning 5)

Forutsetning 5: Alle forklaringsvariablene er ukorrelert med restleddet (Zero conditional mean)

$$\boldsymbol{cov(X, e) = 0} \quad \textbf{eller} \quad \boldsymbol{E(e) = 0}$$

Endogen variabel: En variable som er korrelert med restleddet, som har en (eller flere) av de fire punktene nevnt ovenfor

Populasjonsmodell: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + e_i$

Konsekvensen av endogenitet

Anta at:

- ❑ Vi har en utvalg (sample) som skal benyttes i analyse
- ❑ Brukte OLS metoden til å estimere populasjonsmodellen ovenfor
- ❑ Vi fikk tallfestet (estimerte) effektene ($\widehat{\beta_1}, \widehat{\beta_2}$, ...) av forklaringsvariablene på Y_i

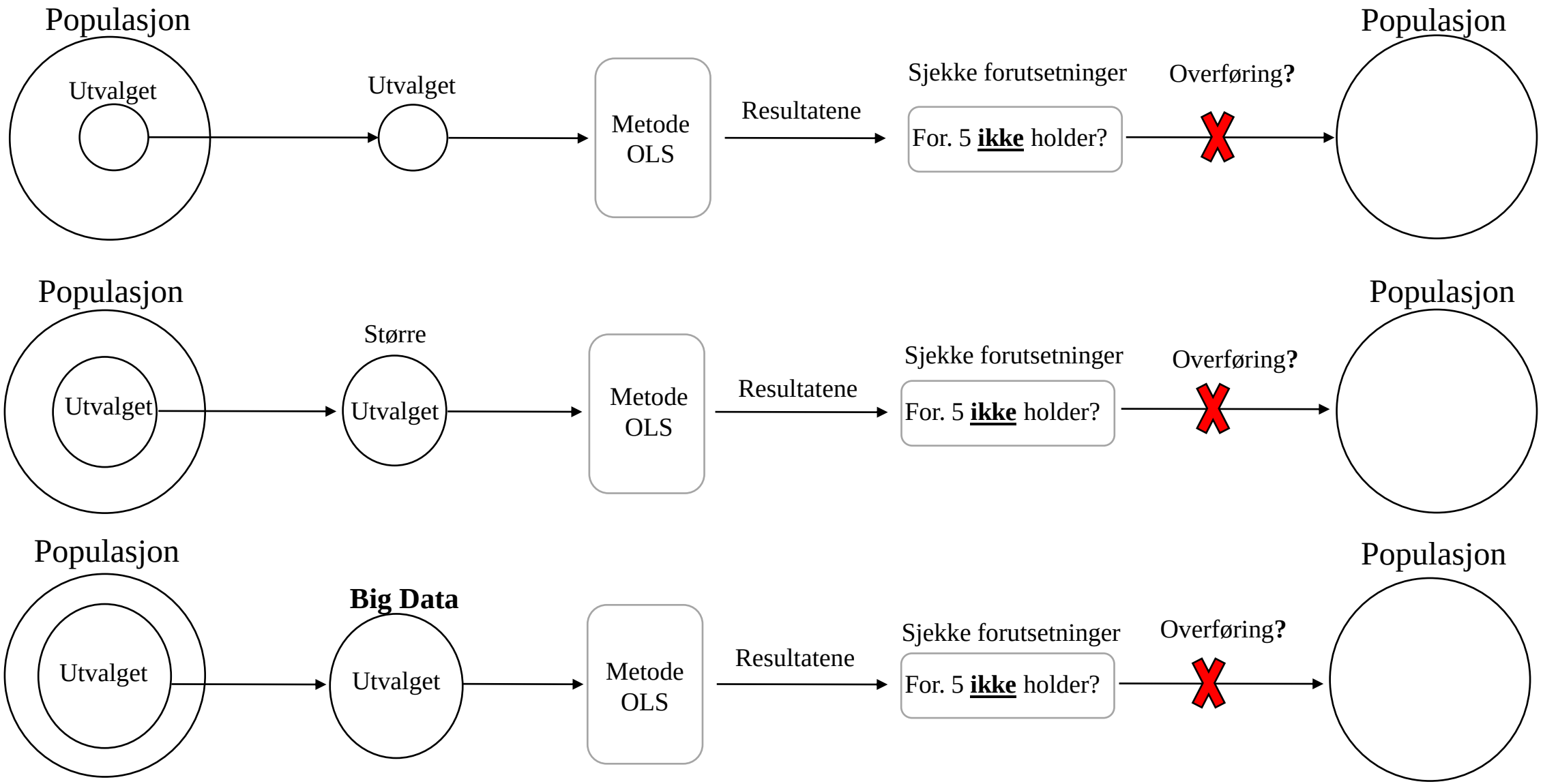
Estimertmodell: $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_{1i} + \widehat{\beta_2} X_{2i} + \cdots + \widehat{e_i}$

Eksempel: Hvis forklaringsvariabelen X_{1i} har endogenitet problem vil den estimert koeffisienten ($\widehat{\beta_1}$) være **bias**

Estimerte effekt ⟶ $\widehat{\beta_1} \neq \beta_1$ ⟵ Populasjon/ faktisk effekt

- ➢ Koeffisienten til ***endogen*** variabelen kan ***ikke*** tolkes som ***kausalsammenheng!!***
- ➢ Dermed kan vi ikke generalisere resultatene fra OLS (estimerte koeffisientene) til populasjon

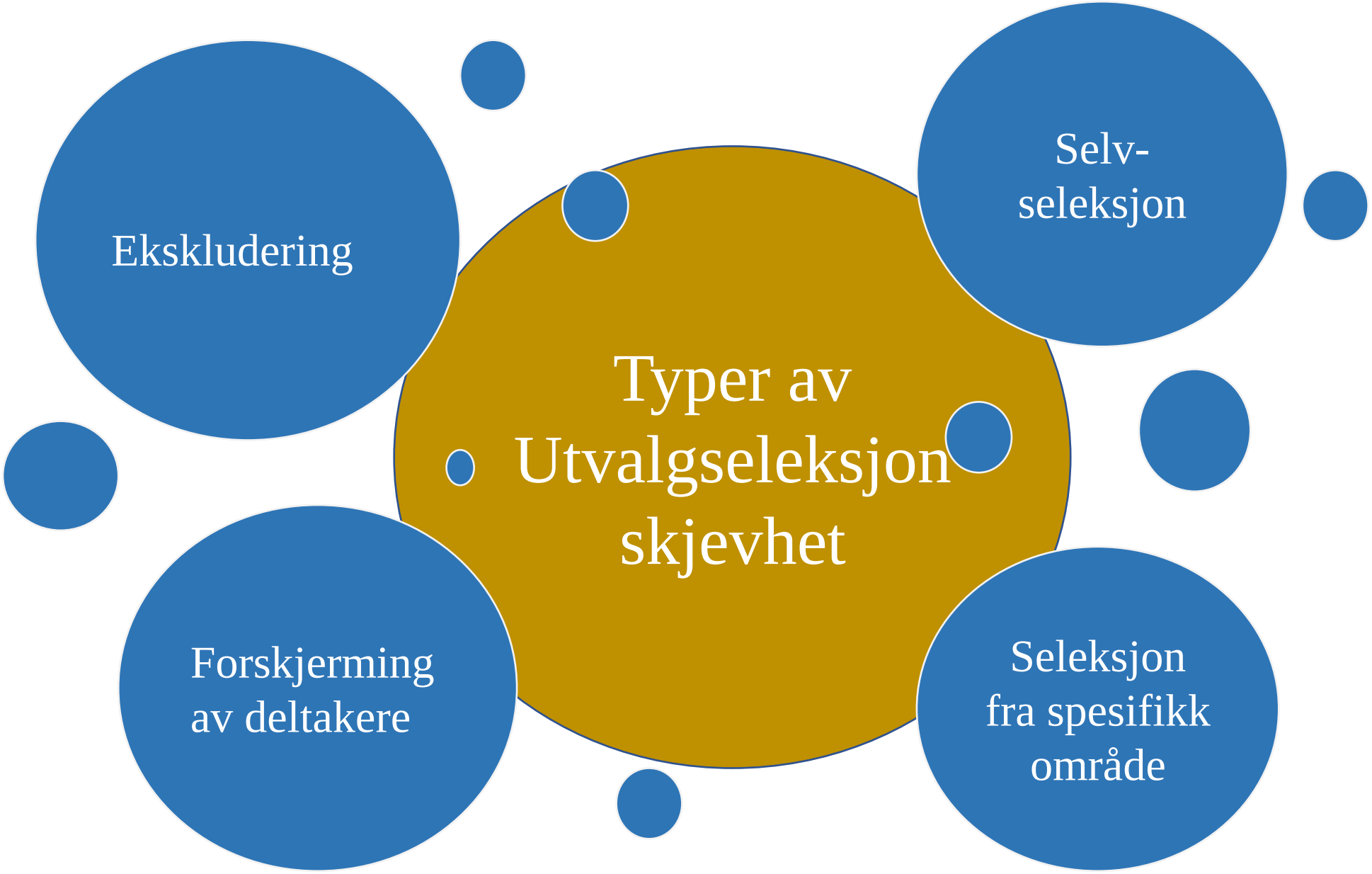
Endogenitet problem og størrelsen på utvalget



Sample Selection Bias (Non- random sample)

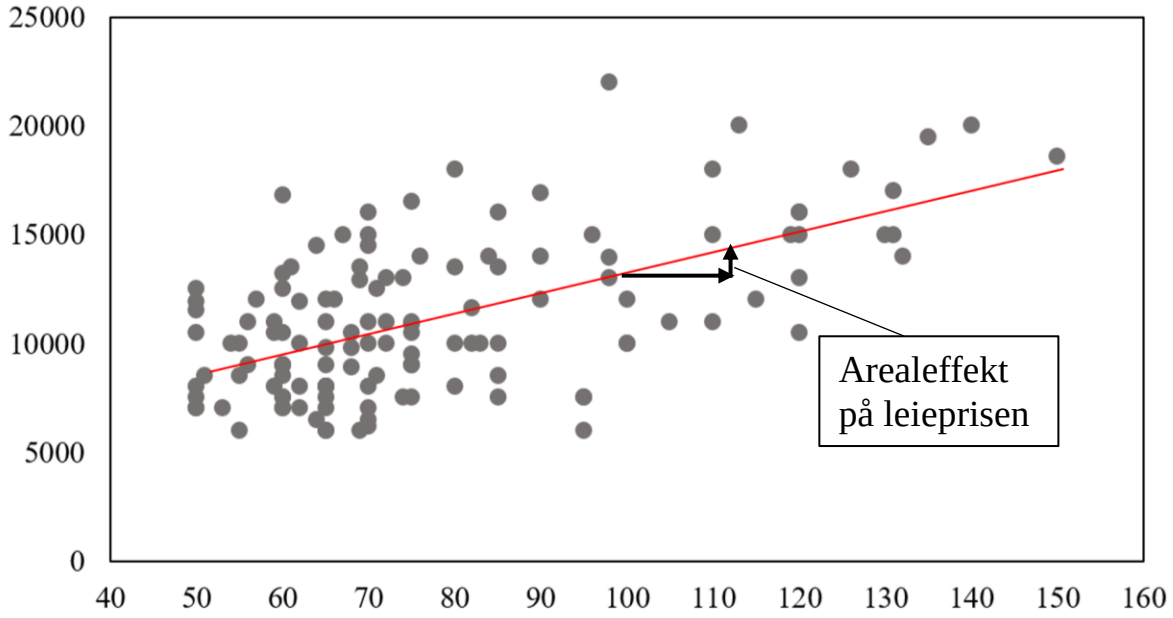
- Utvalgseleksjon skjevhet kan påvirke den statistiske analysen av et utvalg (OLS)
- Statistiske parameterne (estimerte koeffisientene fra OLS) kan være overdrevet (positive biased) eller undervurdert (negative biased)
  - Variabelen kalles endogen →sin koeffisient kan ikke tolkes som kausalsammenheng
  - Dermed resultatene er ikke representativ for hele befolkningen, dvs. resultatene av den statistiske analysen (OLS) kan ikke overføres/ generaliseres til populasjon

Sample Selection Bias

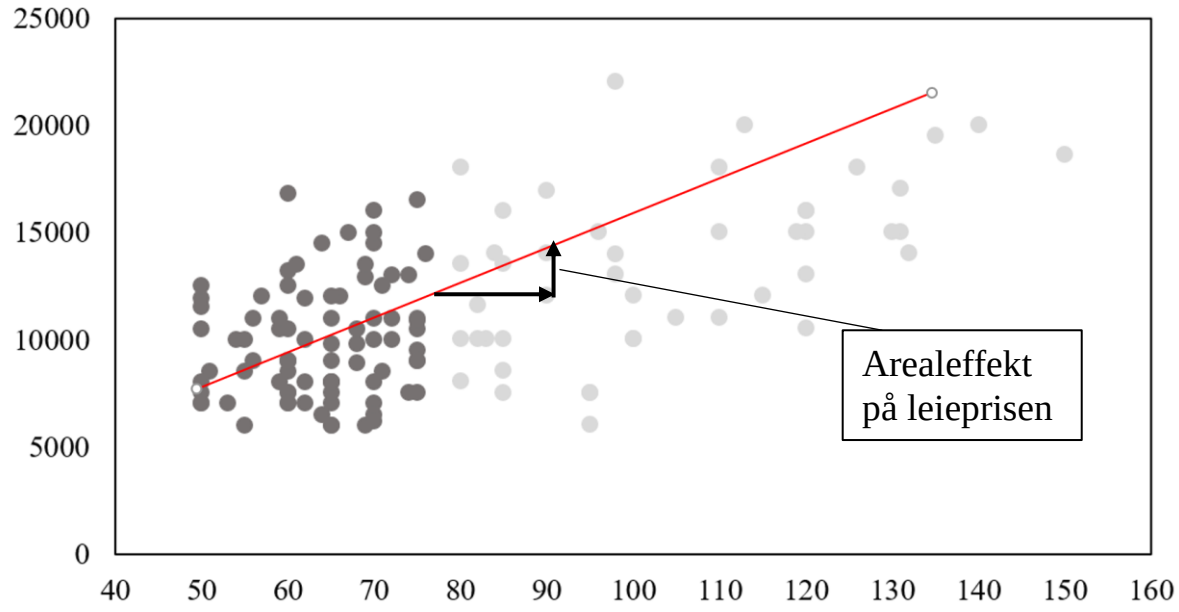


Sample Selection Bias

Areal og leieprisen (hele samplet)



Areal og leieprisen (samplet for boliger <80 kvm)



- Koeffisienten er overdrevet (positive bias)
- Arealeffekt på prisen er høyere enn faktisk

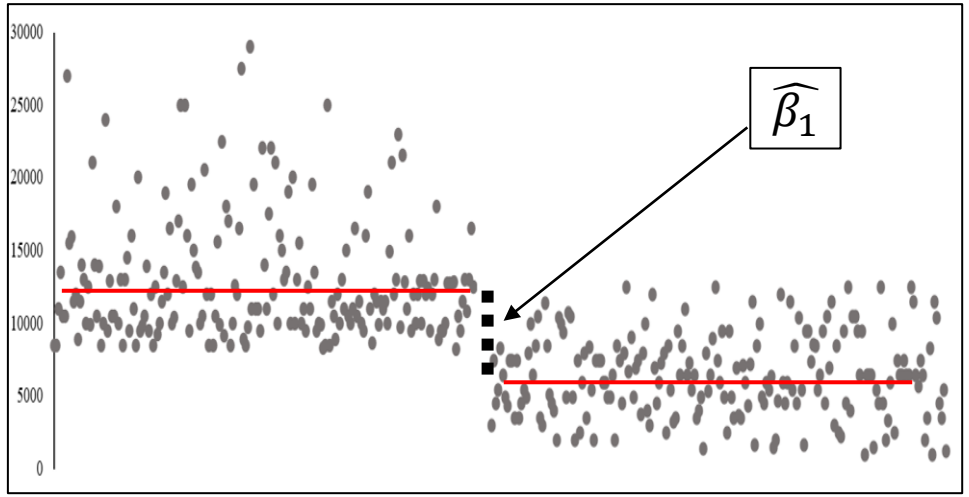
Sample Selection Bias

$$Pris_i = \beta_0 + \beta_1(Merket)_i + \epsilon_i$$

$$Merket \begin{cases} = 1 \; hvis \; bolig \; er \; merket \\ = 0 \qquad\qquad\qquad ellers \end{cases}$$

Selv-seleksjon

Leieprisen (hele samplet)

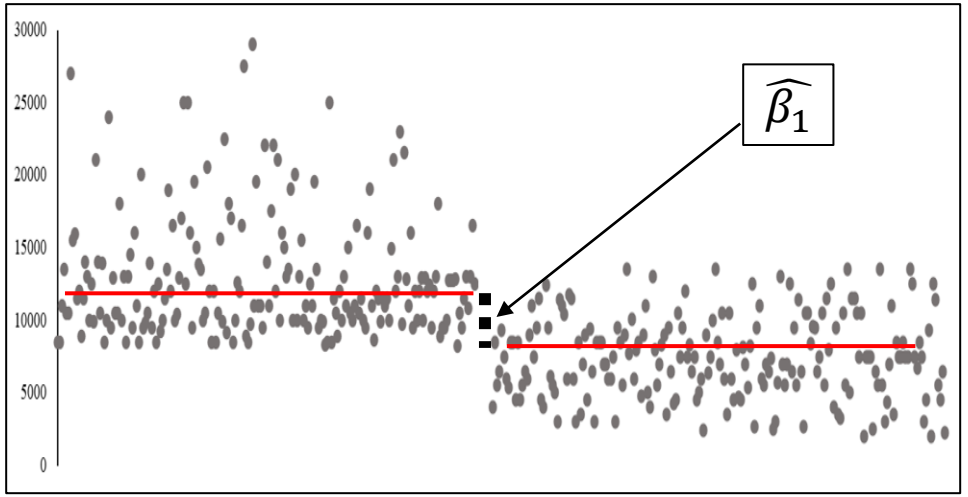


Energimerket Ikke-energimerket

- Hvis det er de dyre boliger som er merket (ikke tilfeldig)
- Seleksjon bias→ her er det positive bias ($\widehat{\beta_1} > \beta_1$)

Ingen *selv-seleksjon*

Leieprisen (hele samplet)



Energimerket Ikke-energimerket

- Etter håndtering av Seleksjon bias problemet

  → ($\widehat{\beta_1} = \beta_1$)

NB: Her antar vi at alle andre OLS forutsetninger er OK!!

Løsning?

Heckman Two-steps Correction model/method

Anta at vi vil identifisere og kvantifisere ***priseffekten*** av energimerking

Hoved modell: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 A_i + \beta_3 B_i + \beta_4 C_i + \beta_5 D_i + \beta_6 E_i + \beta_7 F_i + \beta_8 G_i + e_i$

Hvor Y_i = leieprisen

X_i = hedoniske variable av boligen som areal, antall rom, alder,….

Energimerking av bolig: A B C D E F G NON

Merket Ikke-merket

$$Merket(M_i) \begin{cases} = 1 & hvis\ bolig\ er\ merket \\ = 0 & NON\ \ (ikke - merket) \end{cases}$$

Sample Selection Bias

Hoved modell: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 A_i + \beta_3 B_i + \beta_4 C_i + \beta_5 D_i + \beta_6 E_i + \beta_7 F_i + \beta_8 G_i + e_i$

- ❑ Anta at vi mistenker at estimerte koeffisientene til energimerking ($\widehat{\beta_2} - \widehat{\beta_8}$) er utsatt for selv-seleksjon problem
- ❑ At det er , for eksempel, kun huseiere for allerede dyre boliger som velger å merke boligene sine
- ❑ Hvis dette er tilfellet, vil estimerte koeffisientene ($\widehat{\beta_2} - \widehat{\beta_8}$ være biased, dvs. at de er ikke like faktiske $\beta_1, \beta_2, \ldots$

Heckman Two-steps Correction model/method

Først trenger vi en instrument ($\boldsymbol{Z_i}$)

Step 1: Probit model
Bruk hele sampelet

$$M_i = \alpha_0 + \alpha_1 X_i + \alpha_2 \boldsymbol{Z_i} + \epsilon_i$$

$$M_i \begin{cases} = 1 & hvis\ bolig\ er\ merket \\ = 0 & NON\ \ (ikke - merket) \end{cases}$$

Instrument ($\boldsymbol{Z_i}$) skal tilfredsstille følgende to forutsetninger:

$Corr(Z_i, M_i) \neq 0$ Instrument er en variabel som forklarer variabelen M_i

$Corr\ (Z_i, Y_i) = 0$ Instrument er ikke en variabel som forklarer boligprisen direkte (Y_i)

Sample Selection Bias

Hoved modell: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 A_i + \beta_3 B_i + \beta_4 C_i + \beta_5 D_i + \beta_6 E_i + \beta_7 F_i + \beta_8 G_i + e_i$

Heckman Two-steps Correction model/method

Step 1: Probit model
Bruk hele sampelet

$$M_i = \alpha_0 + \alpha_1 X_i + \alpha_2 \boldsymbol{Z_i} + \epsilon_i$$

$$M_i \begin{cases} = 1 & hvis\ bolig\ er\ merket \\ = 0 & NON\ \ (ikke - merket) \end{cases}$$

- $\boldsymbol{Z_i}$ er total månedlige antall merket boliger på kommune nivå
- Det er mulig å bruke annen instrument som tilfredsstille de to nevnt forutsetninger

Predikerte $\widehat{M_i}$ fra **Step 1**

Prediker/beregne (IMR)

$$\boldsymbol{\widehat{\lambda_i}} = \frac{f(\widehat{\alpha_0} + \widehat{\alpha_2} X_i + \widehat{\alpha_3} Z_i)}{F(\widehat{\alpha_0} + \widehat{\alpha_1} X_i + \widehat{\alpha_3} Z_i)}$$

IMR ($\boldsymbol{\widehat{\lambda_i}}$): Inverse Mills Ratio

Step 2: OLS (hoved modell)
Bruk kun merket sampelet

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

Leg merke til at $G_i$er ikke inkludert i **Step 2** fordi vi bruker kun merket sampelet

Løsning?

$$Pris_i = \beta_0 + \beta_1 (Merket)_i + \epsilon_i$$

$$Merket \begin{cases} = 1 \; hvis \; bolig \; er \; merket \\ = 0 \qquad\qquad\qquad eller \end{cases}$$

Heckman Two-steps Correction model/method

kun merket sampelet $Y_i = \beta_0 + \beta_1 X_i + \beta_2 A_i + \beta_3 B_i + \beta_4 C_i + \beta_5 D_i + \beta_6 E_i + \beta_7 F_i + \boldsymbol{\beta_8 \widehat{\lambda_i}} + e_i$

Hvis $\boldsymbol{\beta_8} \neq 0$ → Seleksjon bias ☹

Eller hvis $corr(\epsilon_i, e_i) \neq 0$ → Seleksjon bias ☹

Konklusjon: $\beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ er biased

(positive eller negative bias?)

Hvis $\boldsymbol{\beta_8} = 0$ → Ingen Seleksjon bias ☺

Eller hvis $corr(\epsilon_i, e_i) = 0$ → Ingen Seleksjon bias ☺

Konklusjon: $\beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ er ikke biased (OK)

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 en eller flere variable har seleksjonsproblemer:
  - ➢ Koeffisienten(e) kan ikke tolkes som kausalsammenheng

Hva skal man gjøre?

- ❑ Finn en 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) ☺