Proxy variable, målefejl og manglede data

Økonometri A

Bertel Schjerning

Program

Proxy Variable (W9.2)

Målefejl (W9.4)

- Målefejl i den afhængige variabel
- Målefejl i forklarende variable

Dataproblemer (W9.5)

- Manglende observationer
- Dataudvælgelse
- Eksterme observationer

Motivation

Nogen gange har vi ikke præcis de variable, vi kunne ønske os.

Fx hvis vi

- Skal estimere effekten af skat på arbejdsudbuddet, skal vi kende folks marginalskat.
- Ønsker at kontrollere for “evne” i regressionen af løn på uddannelse.
- Bruger survey data, hvor folk selv har skulle udfylde formue mv.

Motivation

Overordnet skelner vi mellem:

- **Proxy variable**: Proxy for en uobserveret variabel, som **ikke har en præcis kvantitativ fortolkning**.
  - Fx helbred eller evner, som kan være svære at kvantificere.
  - Vi er ofte ikke interesserede i effekten af variablen, men ønsker at kontrollere for den.

- **Målefejl**: Variable opgjort med målefejl, som har en **præcis kvantitativ fortolkning**
  - Fx udgifter til forbrug (har en præcis betydning), men når vi måler det, kan der være fejl i selvrapporterede udgifter til forbrug
  - Folk husker forkert eller har glemt indkøb.
  - Vi ofte er interesserede i parameteren til variablen.

Proxy Variable

Proxy variable

Antag at den sande model er givet ved

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^* + u \tag{1}$$

hvor vi er interesserede i β_1. x_2^* er således kun med i modellen for at den opfylder MLR.1-MLR.4. Det antager vi her.

x_2^* er dog uobserveret.

I stedet observerer vi en proxy x_2 som er korreleret med x_2^*:

$$x_2^* = \delta_0 + \delta_1 x_2 + v, \text{ med } \delta_1 \neq 0 \tag{2}$$

Spørgmålet er nu om vi kan få et middelret estimat ved at inkludere x_2 i stedet for x_2^*

Proxy variable

Hvis vi indsætter ligning (2) i ligning (1) får vi

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_2 + v) + u \\ &= \beta_0 + \beta_2 \cdot \delta_0 + \beta_1 x_1 + \beta_2 \delta_1 x_2 + u + \beta_2 v \\ &= \tilde{\beta}_0 + \beta_1 x_1 + \tilde{\beta}_2 x_2 + e \end{aligned}$$

Dvs. vi kan stadig få et middelret estimat af β_1, hvis MLR.1-MLR.4 er opfyldt i ovenstående model.

MLR.4 er opfyldt hvis:

$$\begin{aligned} E(e|x_1, x_2) &= E(\beta_2 v|x_1, x_2) + E(u|x_1, x_2) \\ &= \beta_2 E(v|x_1, x_2) + 0 = 0 \end{aligned}$$

Dvs. hvis $E(v|x_1, x_2) = 0$

Proxy variable: Eksempel

Antag at MLR.1-MLR.4 er opfyldt for den følgende model

$$\log(løn) = \beta_0 + \beta_1 uddannelse + \beta_2 evner + u \tag{3}$$

Vi observerer ikke folks sande evner, men vi kan observere

$$evner = \delta_0 + \delta_1 IQ + v \tag{4}$$

OLS estimation af (3) med IQ som proxy for evner er middelret, hvis

$$E(v|uddannelse, IQ) = 0 \tag{5}$$

Er det en rimelig antagelse?

Proxy variable

Mere generelt kan vi sammenligne bias ved at helt at undlade at kontrollere for x_2^* eller ved at anvende en proxy.

Bias ved undladelse

$$\text{plim } \hat{\beta}_1^u - \beta_1 = \beta_2 \frac{\text{cov}(x_1, x_2^*)}{\text{var}(x_1)} = \beta_2 \frac{\delta_1 \text{cov}(x_1, x_2) + \text{cov}(x_1, v)}{\text{var}(x_1)}$$

Bias med proxy

$$\text{plim } \hat{\beta}_1^m - \beta_1 = \beta_2 \frac{\text{cov}(\hat{r}_1, v)}{\text{var}(\hat{r}_1)} = \beta_2 \frac{\text{cov}(x_1, v)}{\text{var}(\hat{r}_1)}$$

Hvor $\hat{r}_1$ er residualerne fra en regression af x_1 på x_2:
$\hat{r}_1 = x_1 - \text{cov}(x_1, x_2)/\text{var}(x_2)x_2$

Bias med en proxy er generelt - men ikke altid - mindre end uden.

Målefejl

Målefejl i den afhængige variabel

Antag at vi har følgende model

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u,$$

hvor MLR.1-MLR.4 er opfyldt.

- y^* er **uobserveret**.
- I stedet observerer vi $y = y^* + e$
- e er en målefejl: $e = y - y^*$.

Hvad er konsekvensen ved at anvende y i stedet for y^*?

Estimationsmodel

$$y = y^* + e = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u + e$$

Målefejl i den afhængige variabel

Vil OLS estimatoren være en middelret og konsistent estimator?

Det kræver, at der for det nye fejlled $u + e$ gælder:

$$\begin{aligned} E(u+e|x) &= E(u|x) + E(e|x) \\ &= 0 + E(e|x) = 0 \end{aligned}$$

- Antag at $E(e) = 0$ (ikke kritisk antagelse).
- Vi behøver også, at $E(e|x) = 0$. Det vil gælde, hvis e er uafhængig af $x_1, x_2, .., x_k$.
- Så vil **OLS estimatoren** i en model med målefejl i den afhængige variabel stadig være **middelret og konsistent**.

Målefejl i den afhængige variabel

Er $E(e|x) = 0$ en realistisk antagelse?

- Ofte ja.
- Men ikke altid.
- Hvis rige underrapporterer deres indkomst/formue.

Målefejl i den afhængige variabel

Hvad med variansen på OLS estimatoren?

- Hvis variansen af fejlleddet (σ_u^2) og målefejlen (σ_e^2) er konstant
- og e og u er uafhængige, får vi

$$\text{var}(u + e) = \sigma_u^2 + \sigma_e^2 > \sigma_u^2$$

Dvs. større varians af fejlleddet, hvis der er målefejl i den afhængige variabel.

$\rightarrow$ Større varians på parameterestimatoren $var(\hat{\beta}_j|x)$.

Multiplikative målefejl

Nogle gange giver det mere mening at antage, at målefejlen er multiplikativ.

- Fx hvis størrelsen af målefejlene er proportionale med y variablen.
- Afhængig variabel:

$$y = y^* \cdot a, \quad a > 0$$

Løsning: en model med $\log(y)$

$$\begin{aligned} \log(y) &= \log(y^*) + \log(a) \\ &= \log(y^*) + e, \end{aligned}$$

hvor $e = \log(a)$.

Dvs. vi er tilbage i setuppet fra før.

Målefejl i forklarende variable

Vi betragter nu følgende model:

$$y = \beta_0 + \beta_1 x^* + u,$$

hvor MLR.1-MLR.4 er opfyldt.

- x^* er ikke observeret.
- I stedet observerer vi $x = x^* + e$

To tilfælde:

1. $E(e) = 0, Cov(e, x) = 0$
2. $E(e) = 0, Cov(e, x^*) = 0$ - Klassisk målefejl.

Tilfælde 1: $Cov(e, x) = 0$

Indsæt $x^* = x - e$ i modellen

$$\begin{aligned} y &= \beta_0 + \beta_1 x^* && +u \\ &= \beta_0 + \beta_1(x - e) && +u \\ &= \beta_0 + \beta_1 x && +u - \beta_1 e \end{aligned}$$

OLS er konsistent, da det nye fejlled er $u - \beta_1 e$ og der gælder:

$$\text{plim } \hat{\beta}_1 - \beta_1 = \frac{cov(u - \beta_1 e, x)}{var(x)} = \frac{cov(u, x)}{var(x)} - \beta_1 \frac{cov(e, x)}{var(x)}$$

Denne form for målefejl svarer til proxy variable (og normalt ikke det vi opfatter som målefejl).

Tilfælde 2: $Cov(e, x^*) = 0$

Indsæt igen $x^* = x - e$ i modellen

$$y = \beta_0 + \beta_1 x + u - \beta_1 e.$$

For det nye fejled gælder nu:

$\text{cov}(u - \beta_1 e, x) =$

Tilfælde 2: $Cov(e, x^*) = 0$

Den asymptotiske bias af $\hat{\beta}_1$ er nu

$$
\begin{aligned}
p\lim \hat{\beta}_1 - \beta_1 &= \frac{cov(u - \beta_1 e, x)}{var(x)} \\
&= -\beta_1 \frac{\sigma_e^2}{var(x^* + e)} \\
&= -\beta_1 \frac{\sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2}
\end{aligned}
$$

Det betyder at den forventede værdi af $\hat{\beta}_1$ er

$p\lim \hat{\beta}_1 =$

Tilfælde 2: $Cov(e, x^*) = 0$

Konklusion: OLS estimatoren er biased imod 0, da

$$0 < \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2} < 1$$

- Hvis $\beta_1 > 0$ så er $p\lim \hat{\beta}_1 < \beta_1$
- Hvis $\beta_1 < 0$ så er $p\lim \hat{\beta}_1 > \beta_1$
- Jo større målefejlsvariansen er, jo større er den asymptotisk bias.

Dette kalder vi typisk for **attenuation bias**.

For flere forklarende variable bliver det mere kompliceret.

- Generelt vil alle estimaterne være biased, hvis der er målefejl i en af de forklarende variable.

Tilfælde 2: $Cov(e, x^*) = 0$

Quiz

Sand model

$$y = 1 + 2x^* + u$$
$$x = x^* + e$$

hvor

- $cov(e, x^*) = 0$ (tilfælde 2: Klassisk målefejl)
- $x^* \sim N(2, 1), u \sim N(0, 1), e \sim N(0, 1)$

Estimationsmodel

$$y = \beta_0 + \beta_1 x + v$$

Hvad er den asymptotisk forventede værdi af OLS $\hat{\beta}_1$?

Monte Carlo Simulation

Den sande model

$$y = \beta_0 + \beta_1 x^* + u.$$

Setup og antagelser:

- Antallet af replikationer 1000
- Antallet af observationer $n = 100$
- $x = x^* + e$
- $\text{Cov}(e, x^*) = 0$ (Case 2).
- Parametre: $\beta_0 = 1, \beta_1 = 2.$
- Fordelingen: $x^* \sim N(2,1), u \sim N(0,1), e \sim N(0,1)$

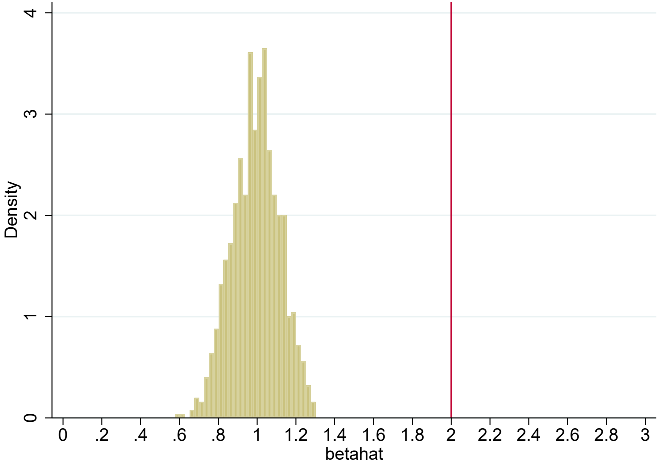
Regressionsmodel

$$y = \beta_0 + \beta_1 x + v$$

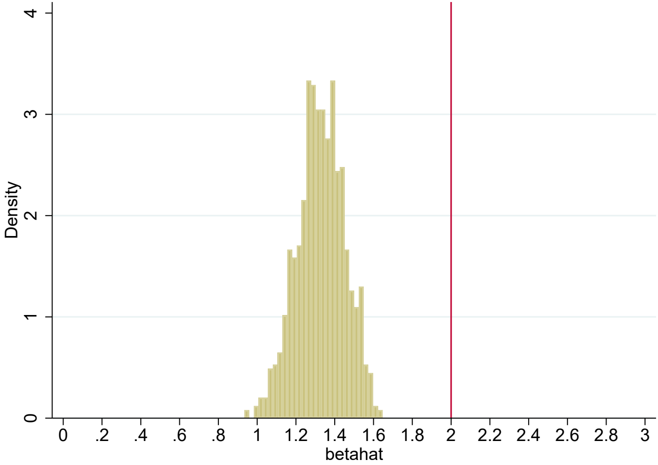
Monte Carlo Simulation: Stata eksempel

```
*DEFINE PROGRAM THAT SPECIFIES THE DGP
program olsdata, rclass
drop _all

*SET NUMBER OF CURRENT OBSERVATIONS
set obs 100

*DATA GENERATING PROCESS
* true process
generate xstar = 2+1*rnormal()
generate e1 =rnormal(0,$var_e)
generate u = rnormal()
generate y = 1+2*xstar+u
* observed x with measurement error
generate x = xstar+e1

*CALCULATE OLS ESTIMATES AND OLS SLOPE IN B1
regress y x
return scalar b1=_b[x]
end

*SIMULATE PROGRAM 10000 TIMES
simulate betahat=r(b1), seed(117) reps(1000) nodots:olsdata
```

Monte Carlo Simulation: Stata eksempel $var(e) = 1$



Monte Carlo Simulation: Stata eksempel $var(e) = 0.5$



Dataproblemer

Dataproblemer

Indtil nu har vi antaget, at **MLR.2** er opfyldt:

- MLR.2 : En tilfældig/repræsentativ stikprøve: uafhængige fejlled trukket tilfældigt fra populationen.

Det er ikke altid tilfældet:

- Manglende observationer: tilfældigt eller ikke tilfældigt
- Dataudvælgelse: endogen eller eksogen udvælgelse
- Korrelerede fejlled (kun delvist dækket i Wooldridge)

Derudover skal vi snakke kort om problemer med

- Ekstreme observationer

Manglende observationer

Manglende observationer for en eller flere af variablene.

Er det et problem?

- Ud over at manglende observationer reducerer n.

Det er vigtigt at vide, hvorfor observationerne mangler.

- Hvis sandsynligheden for at en observation mangler er korreleret med fejlleddet vil OLS være biased.
- Fx hvis
  - Det kun er studerende, som er ekstraordinært glade/sure over et fag, svarer på evaluaeringerne.
  - Det kun er "særlige kloge" lavt uddannede som tager en IQ test.

Hvis observationerne mangler **tilfældigt**, vil OLS stadig være **middelret og konsistent**.

Opgave om manglende observationer

Vi skal se på, hvordan manglende observationer kan påvirke estimationen vha et simulationseksperiment.

Til simulationseksperimentet bruger vi, at

$$x \sim N(2,1), v \sim U(0,1), e \sim N(0,1),$$

Ud fra disse variable danne u, y og m

$$\begin{aligned} u &= e + 2(v - 0.5), \\ m &= 1(v > 0.2) \\ y &= 1 + x + u \quad \text{hvis } m = 1 \end{aligned}$$

Dvs. den sande model er $y = 1 + x + u$, men desværre observerer vi kun y når variablen $m = 1$

Opgave om manglende observationer

Vi benytter simulationsprogrammet på Absalon *11 MC Manglede observationer.do* til at svare på følgende spørgsmål:

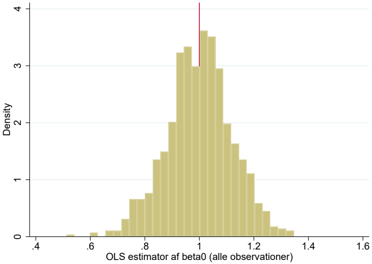
- Estimer modellen, hvor alle observationer bruges. Er OLS estimatoren for β_0 og β_1 middelret og konsistent?
- Estimer modellen, hvor der kun anvendes observationer hvor $m = 1$. Er OLS estimatoren for β_0 og β_1 middelret og konsistent?
- Overvej hvilke aspekter af modellen som er kritiske for dine resultater?

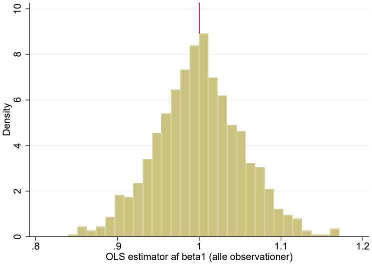
Opgave om manglende observationer: Stata

```
*DEFINE PROGRAM THAT SPECIFIES THE DGP
program olsdata, rclass
drop _all

*SET NUMBER OF CURRENT OBSERVATIONS
set obs 500

*DATA GENERATING PROCESS
generate v = (runiform())
generate u = rnormal()+ 2*(v-0.5)
generate x = 2+1*rnormal()
generate m =(v>0.2)
generate y = 1+1*x+u
*replace y=. if m==0

*CALCULATE OLS ESTIMATES AND OLS SLOPE IN B1
regress y x if m==1
return scalar b1=_b[x]
return scalar b0=_b[_cons]
end

*SIMULATE PROGRAM 10000 TIMES
simulate beta1hat=r(b1) beta0hat=r(b0), seed(117) reps(1000) nodots:olsdata
```

Opgave om manglende observationer: Stata

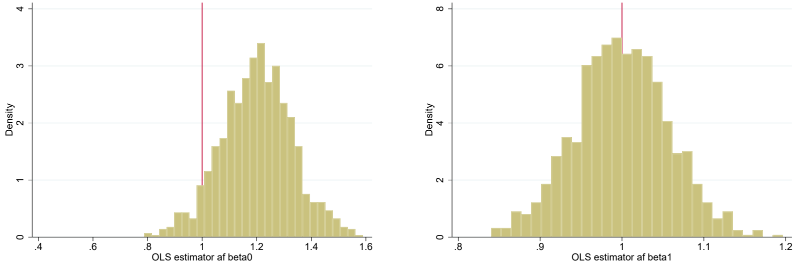
Fordeling af estimater for alle observationer





Opgave om manglende observationer: Stata

Fordeling af estimater for observationer med $v > 0.2$.



Dataudvælgelse

Nogen gange er det os som økonometrikere, som udvælger data baseret på forskellige karakteristika (fx alder).

Her skelner vi mellem

- **Eksogen dataudvælgelse:** baseret på forklarende variable.
- **Endogen dataudvælgelse:** baseret på den afhængige variabel.
- **Stratificeret data:** Oversampling af observation med givne forklarende variable.

Nogle gange kommer vi til at udvælge data uden eksplicit at tænke over det

Som ved manglende observation, medfører ikke alle typer dataudvælgelse bias.

Dataudvælgelse: Eksogen

Eksogen dataudvælgelse er baseret på en eller flere af de forklarende variable.

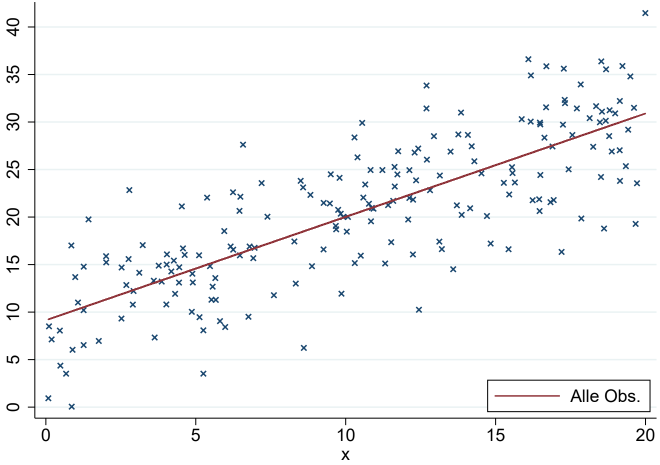
- Det er i udgangspunktet ikke et problem.
- Hvis vi antager at $E(y|x)$ er den samme funktion for alle dele af populationen (MLR.1), vil vi få det samme estimate uanset hvilken x-interval vi kigger.
- I så fald er OLS estimatoren stadig middelret og konsistent.

Hvis vi ikke bruger den rigtige funktionelle form, vil estimaterne ændre sig ved forskellige stikprøver.

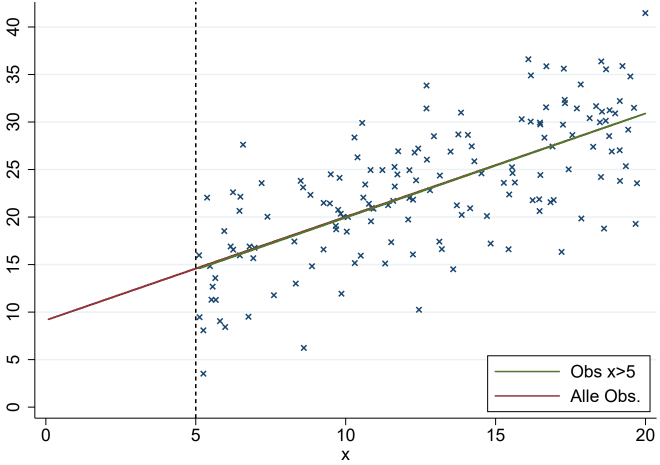
- Det er dog ikke problem. Tværtimod er det nyttig information.

Stratifisering er et eksempel på eksogen dataudvælgelse.

Dataudvælgelse: Stata eksempel



Dataudvælgelse: Stata eksempel



Datavdvælgelse: Endogen

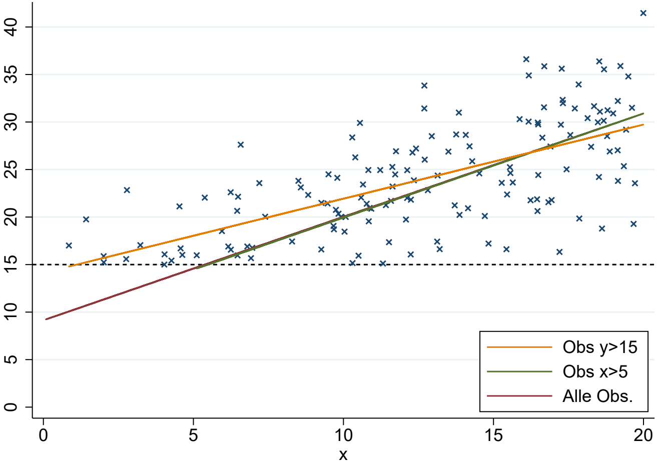
Endogen dataudvælgelse er baseret på den afhængige variabel - enten eksplicit eller implicit.

Eksempler

- **Effekten af uddannelse på lønnen blandt lavtlønnede**
  Uddannelse reducerer sandsynligheden for at være lavt lønnet.
  Så vi ender et sample af alle lavt uddannede + de "dårligste" af de højt uddannede.
- **Effekten af uddannelse på timelønnen**
  Vi observerer kun timelønnen for personer i arbejde.
  Der kan være systematiske forskelle på personer i arbejde og personer, som ikke arbejder.

OLS estimatoren vil generelt **ikke** være **middelret og konsistent**.

Dataudvælgelse: Stata eksempel



Eksterme observationer

Ekstreme observationer er observationer, som skiller sig ud relativt til resten af stikprøven.

Ektreme observationer kan have stor betydning for OLS estimaterne.

- OLS minimerer de kvadrerede residualer. Ektreme observationer kommer derfor til at betyde meget.
- Medfører store standardfejl.

Hvorfor optræder der ekstreme observationer?

- Datafejl fx indtastningsfejl.
- Data behandling. Fx beregning af timeløn for personer med meget få timer.
- Nogle observationer er bare ektreme, fx Bill Gates, Google mv.

Eksterme observationer

Hvad skal vi gøre med ekstreme observationer?

- Hvis det er en fejl, skal de fjernes (det er bare ikke altid, man ved det med sikkerhed).
- Estimer med og uden de ekstreme observationer og se hvor stor forskel det gør.
- Omdan variablene. Log i stedet for niveau, ranks, dummierne for at variablen er over et givent niveau.
- Der findes estimatorer, som er mindre følsomme overfor ekstreme observationer.

Opsummering

Opsummering

- Proxy variable kan reducere bias og i visse tilfælde genskabe middelrethedden af OLS.
- Målefejl i den afhængige variable medfører typisk kun større varians på OLS estimaterne.
- (Klassisk) målefejl i forklarende variable medfører bias mod 0.
- Manglende observation kan skabe bias, hvis det er korreleret med fejlleddet (og x'erne).
- Dataudvælgelse på x, inkl. stratifisering, er typisk OK.
- Udvælgelse på y medfører typisk bias.