Prædiktioner og funktionel form

Økonometri A

Bertel Schjerning

Program

Prædiktioner (W6.4)

Funktionel form (W9.1)

- Test for misspecifikation

Ekstern og intern validitet

Opsummering

Prædiktioner

Prædiktioner

Indtil nu har vi fokuseret på at finde et kausalt estimatat af effekten af en central variable x_j (fx klassestørrelse).

Alternativt kunne vi fokuserer på at lave prædiktioner af $\hat{y}$

$$\hat{y} = \hat{E}(y|x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_n x_n \tag{1}$$

Så længe vi tror på at alle $\hat{\beta}$'er kan fortolkes kausal, er denne ligning blot en generalisering af vores normale fortolkning af $\hat{\beta}$ som

$$\Delta\hat{y} = \Delta\hat{E}(y|x) = \hat{\beta}_k \Delta x_k \tag{2}$$

Fortolkning: Hvis vi ændrer x_k med Δx_k, så ændrer $\hat{y}$ sig med $\hat{\beta}_k \Delta x_k$.

Prædiktioner og usikkerhed

Prædiktioner afhænger af vores estimaterede parameter.

- $\Rightarrow$ Prædiktioner er stokastiske størrelse

Hvad er usikkerheden på vores prædiktioner?

- For en enkelte observation?
- For en gruppe med samme x'er?

Vi starter med det sidste.

Prædiktioner og usikkerhed

Betragt en simpel regressionsmodel

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

Definer prædiktionen for $x = c$ som

$$\hat{\theta}_c = \hat{E}(y|x = c) = \hat{\beta}_0 + \hat{\beta}_1 c \tag{3}$$

For at sige noget om usikkerheden på $\hat{\theta}_c$, skal vi kende $var(\hat{\theta}_c|x)$.

Problem: $\hat{\theta}_c$ afhænger af flere stokatiske variable (alle β'erne).

Prædiktioner og usikkerhed

Definer prædiktionen for $x = c$ som

$$\hat{\theta}_c = \hat{E}(y|x = c) = \hat{\beta}_0 + \hat{\beta}_1 c \tag{4}$$

Løsning: Omskriv modellen:

$$\begin{aligned}
\hat{\beta}_0 &= \hat{\theta}_c - \hat{\beta}_1 c \\
\hat{y}_0 &= \hat{\beta}_0 + \hat{\beta}_1 x = \hat{\theta}_c - \hat{\beta}_1 c + \hat{\beta}_1 x \\
\hat{y}_0 &= \hat{\theta}_c + \hat{\beta}_1 (x - c)
\end{aligned}$$

Dvs. ved at fratrække c fra x'erne svarer interceptet i modellen til prædiktionen for $x = c$

Vi kan nu let finde variansen af $\hat{\theta}_c$ (en parameter) i stedet for at skulle kombinere varianserne/kovarianserne for af flere parametre $\hat{\beta}_0$ og $\hat{\beta}_0$.

Prædiktioner og usikkerhed

Wooldridge foreslår at vi estimaterer modellen for alle relevante c og aflæser standardfejlen for interceptet.

Vi kender dog standardfejlen for interceptet (se lektion 2)

$$var(\hat{\beta}_0|x) = \frac{\sigma^2 \frac{1}{n} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5}$$

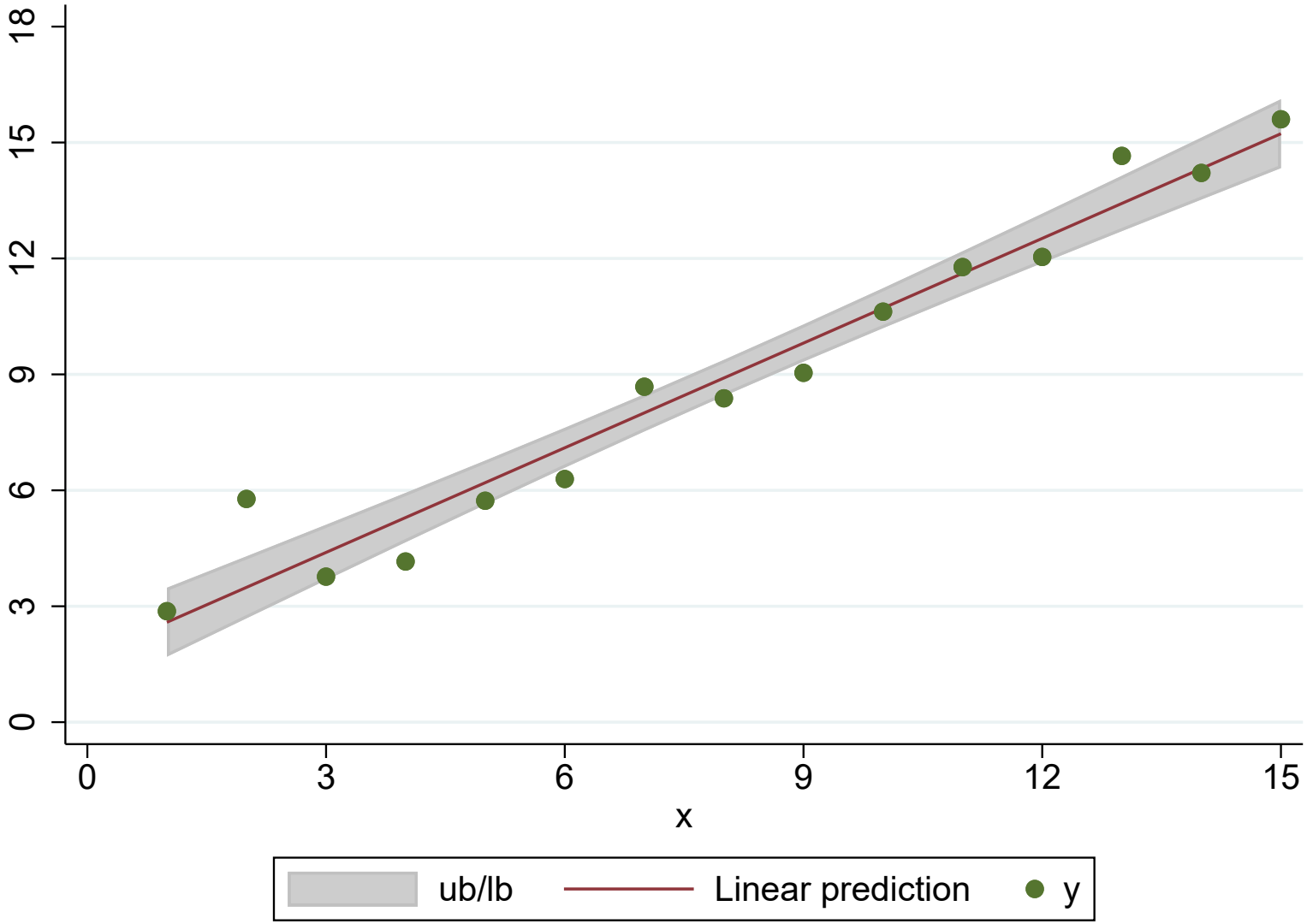
Dvs. ved at fratrække c fra x gælder

$$var(\hat{\theta}_c|x) = \frac{\sigma^2 \frac{1}{n} \sum_{i=1}^{n}(x_i - c)^2}{\sum_{i=1}^{n}(x_i - c - (\bar{x} - c))^2} = \frac{\sigma^2 \frac{1}{n} \sum_{i=1}^{n}(x_i - c)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6}$$

Bemærk: $var(\hat{\theta}_c|x)$ er mindst ved $c = \bar{x}$
$\min_c 1/n \sum_{i=1}^{n}(x_i - c)^2 \Rightarrow$ FOC: $-2/n \sum_{i=1}^{n}(x_i - c) = 0 \Rightarrow c = \bar{x}$

Prædiktioner og usikkerhed

Størst varians for x'er langt fra gennemsnittet



Prædiktioner og usikkerhed

Vi kan også være interesseret i konfidensintervallet for prædiktionen for en enkelte observation.

- Wooldridge kalder dette for prædiktionsintervallet. Stata kalder det for forecast error.

Prædiktionen for en enkelte observation i er stadig $\hat{y}_i = E(y|x = x_i)$

Prædiktioner og usikkerhed

Vi kan også være interesseret i konfidensintervallet for prædiktionen for en enkelte observation.

- Wooldridge kalder dette for prædiktionsintervallet. Stata kalder det for forecast error.

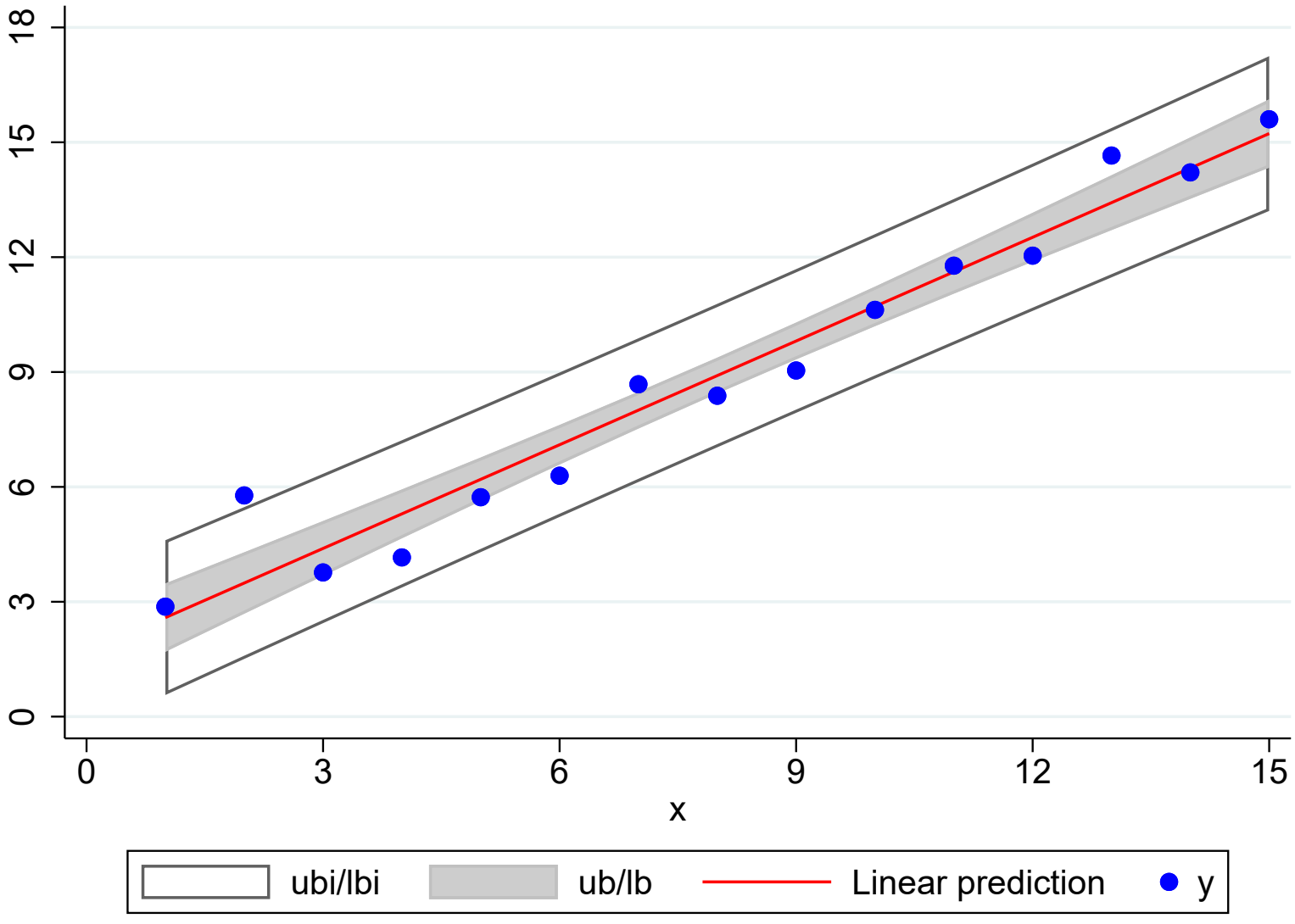
Prædiktionen for en enkelte observation i er stadig $\hat{y}_i = E(y|x = x_i)$

Vi er dog nød til at tage højde for observationerne ligger spredt omkring $E(y|x = x_i)$: $y_i = E(y|x = x_i) + u_i$

$$var(y|x_i) = var(\hat{y}|x_i) + var(u|x_i) = var(\hat{\theta}_{x_i}|x_i) + \sigma^2 \tag{7}$$

Første del er det samme som før og er en funktion af $1/n$, anden del er ny og reduceres ikke med n.

Prædiktioner og usikkerhed



Prædiktioner og usikkerhed

```
clear all
set seed 83
set obs 15

gen x = _n
gen u = rnormal()
gen y = 1 + x + u

reg y x
predict yhat, xb
predict yhat_se,  stdp
predict yhat_sei, stdf

gen lb = yhat - invnormal(0.975)*yhat_se
gen ub = yhat + invnormal(0.975)*yhat_se

gen lbi = yhat - invnormal(0.975)*yhat_sei
gen ubi = yhat + invnormal(0.975)*yhat_sei
```

Funktionel form

Funktionel form

Vi siger, at vores model har forkert funktionel form, hvis y eller x'erne indgå forkert i modellen. Fx hvis

1. Den "sande model" indeholder $\ln(y)$, mens vi anvender y
2. Den "sande model" indeholder $\ln(x)$, mens vi anvender x
3. Den "sande model" indeholder x^2, mens vi kun anvender x lineært.

Case 1) Vi kender allerede effekten fra analysen af udeladte variable.

Case 2) kan og også fortolkes som en udeladt variabel:

Sand model : $y = \beta_0 + \beta_1 \ln(x) + u$
Estimeret model: $y = \beta_0 + \beta_1 x + v$, hvor $v = \beta_1(\ln(x) - x) + u$

Vi har altså en udeladt variabel $\ln(x) - x$, som er en funktion af x.

Funktionel form

Den "sande model":

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u \tag{8}$$

Hvis vi estimerer en model uden x^2, får vi

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{cov(x, x^2)}{var(x)} \tag{9}$$

Dvs. forkert funktionel form er basalt set et brud på MLR.4.

- Men et brud vi kan fikse/teste for.
- Mere om det nedenfor.

Funktionel form

Det kan være nyttigt at skelne mellem to former for funktionel form misspecifikation

1. Misspecifikation af variable, som er nødvendig for at få et kausalt estimate af en interessant variable.
2. Misspecifikation af den variable vi er interesseret i.

Eksempel: Effekten af kompensationsgraden af dagpenge på længden af ledighed

$$y_i = \beta_0 + \beta_1(DS/w_i) + u_i \tag{10}$$

Hvor DS er dagpengesatsen, w_i er individuelle løn, og y_i er længden af ledighed.

Funktionel form

Problem 1: Hvis er DS ens for alle, så kommer al variation i kompensationsgraden fra w_i, som kan være korreleret med u.

Kan vi løse det ved at kontrollere for w_i?

- Delvist: Vi kan kontrollere for fx w_i lineært
- Men hvis vi kontrollerer for $1/w_i$ er der perfekt multikollinearitet mellem $1/w_i$ og DS/w_i.

Når vi ikke frit kan vælge variablens funktionelle form, afhænger kausaliteten af vores estimater af en **funktionel form antagelse**

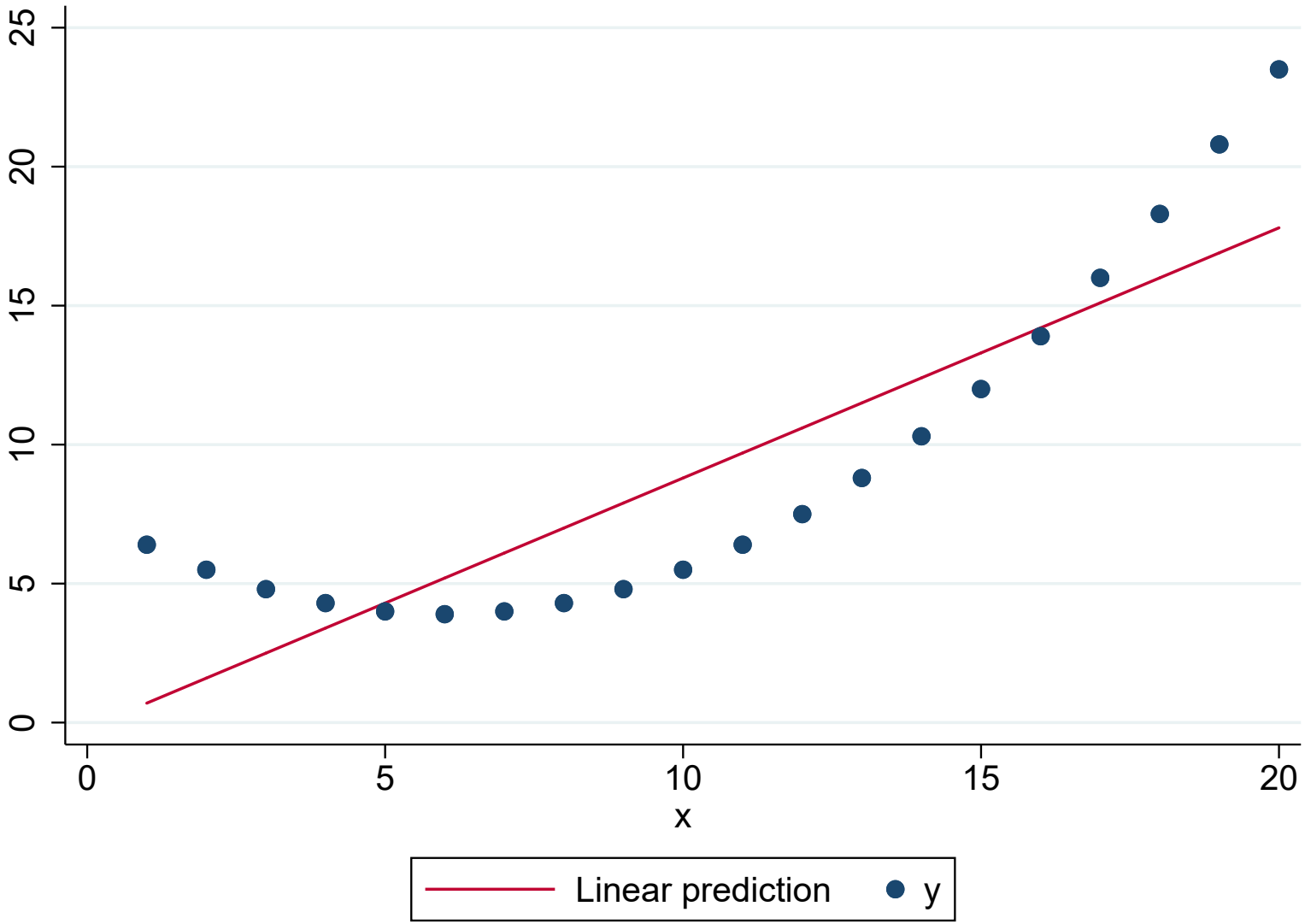
Løsning: Find variation i DS som ikke er korreleret med w_i (fx variation pga reform).

Funktionel form

Problem 2: Hvad hvis vi vælger den forkerte form på den variabel, som vi er interesseret i (her DS/w_i)?

- I udgangspunktet et mindre problem
- Vi fanger den gennemsnitlige effekt af DS/w_i i vores sample.
    - OLS er den "bedste" lineære approksimation til $E(y|x)$.
- Kan være problematisk, hvis vi bruger vores model ukritisk i "udkanten" af vores sample.

Funktionel form



Test for misspecifikation

Betragt modellen:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \tag{11}$$

Simpel test af om den funktionelle form i (11) **er rigtig**

- Tilføje yderligere led af andre funktioner af x til modellen
- Hvis MLR.1-MLR.4 er overholdt i (11), vil de yderligere led være insignifikante.
- Hvilke funktionelle former skal vi prøve?
  - Vi kan approksimere en ukendt funktionel form med polynomier, fx x^2 mv.
  - Alternativt kan vi bruge dummier.

Test for misspecifikation: RESET

REgression **S**pecification **E**rror **T**est er en generel test for misspecifikation.

Ide: Tilføj funktioner af de $\hat{y}$ til modellen.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + u. \tag{12}$$

- Sparer frihedsgrader sammenlignet med funktioner af alle individuelle forklarende variable.

Test for korrekt funktionel form er

$$H_0 : \delta_1 = \delta_2 = 0 \tag{13}$$

Testet kan udføres som et F-test $\sim F(2, n - k - 1 - 2)$.

Testet siger ikke noget om, *hvordan* modellen er misspeficieret

Test for misspecifikation: Ikke-"nestede" alternativer

Hvordan vælger vi mellem ikke-nestede alternativer?

$$\text{model 1} \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{14}$$

$$\text{model 2} \quad y = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + u \tag{15}$$

Metode 1 (Mizon and Richard):

Estimer følgende udvidet model:

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \ln(x_1) + \gamma_4 \ln(x_2) + u. \tag{16}$$

Test hypoteserne

- Model 1 er rigtig: $H_0 : \gamma_3 = \gamma_4 = 0$
- Model 2 er rigtig: $H_0 : \gamma_1 = \gamma_2 = 0$

Test for misspecifikation: Ikke-"nestede" alternativer

Metode 2 (Davidson and McKinnon):

Hjælperegressionsmodel 1:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \hat{\hat{y}} + u,$$

hvor $\hat{\hat{y}}$ er den prædikterede værdi fra model (2).

- Hypotese: $H_0 : \theta_1 = 0$

Hjælperegressionsmodel 2:

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + \theta_2 \hat{y} + u,$$

hvor $\hat{y}$ er den prædikterede værdi fra model (1).

- Hypotese: $H_0 : \theta_2 = 0$

Test for misspecifikation: Ikke-"nestede" alternativer

Konklusionen er ikke altid entydig.

- Begge modeller kan blive afvist
  - Prøv en anden funktionel form
- Ingen af modellerne bliver afvist.
  - Brug andre kriterier til at vurdere, hvilken model som er bedst
  - Selvom en model ikke bliver afvist, er det ikke nødvendigvis den rigtige model.
  - Den kan afvises i forhold til en tredje model.
- Ikke "nestede" modeller, hvor den afhængige variabel er forskellig, er mere komplicerede at behandle.

Husk: Overvej om den funktionelle form er kritisk for jeres konklusioner.

Ekstern og intern validitet

Ekstern og intern validitet

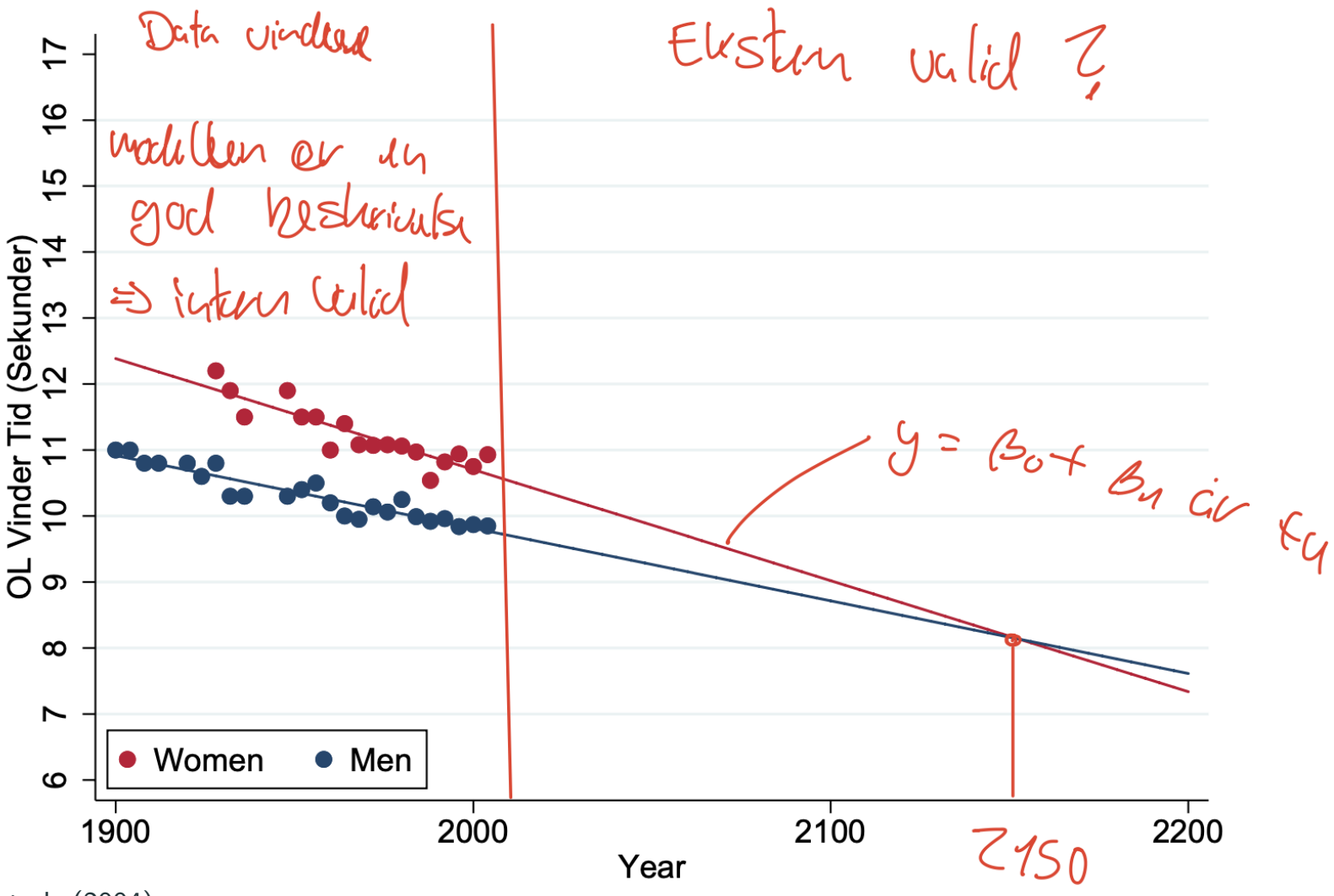
Når vi snakker om, hvorvidt vi kan stole på en models prædiktioner skelner vi ofte mellem intern og ekstern validitet.

Vi siger at hvis vores model er en god beskrivelse af sammenhængen mellem x og y...

- **i sample** så er den **intern valid**
- **uden for sample** så er den **ekstern valid**

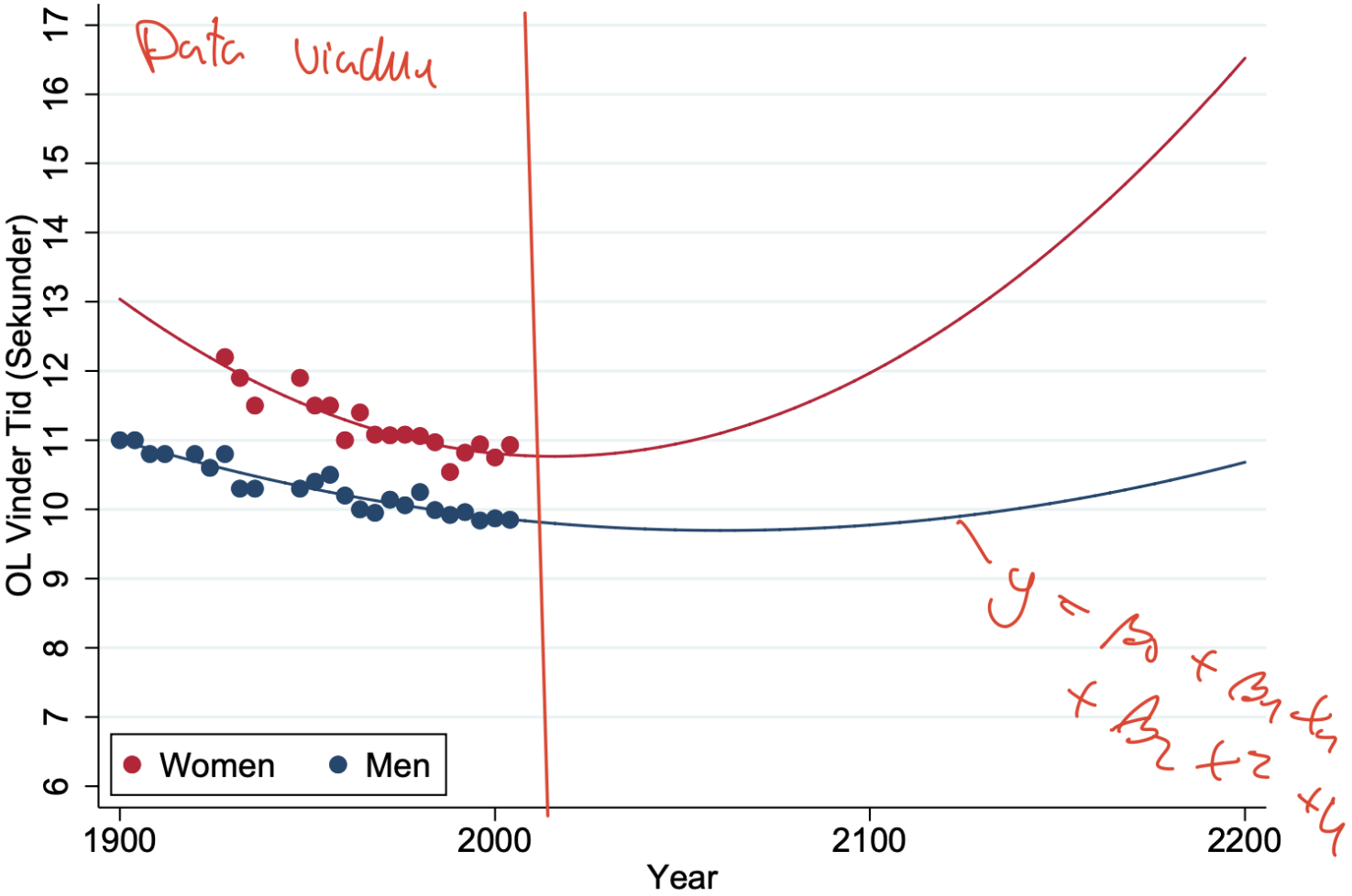
Generelt beror ekstern validitet på stærkere funktionel form antagelser end intern validitet.

Ekstern og intern validitet: 100 meter løb til OL



Tatem et al. (2004)

Ekstern og intern validitet: 100 meter løb til OL



Tatem et al. (2004)

Opsummering

Opsummering

Usikkerheden på prædiktioner er størst længst fra "midten" af sample.

Funktionel form misspefikation.

- Grundlæggende et brud på MLR.4
- Kan identifiseres og fikses.
- Ikke altid et problem. Afhænger af modellens formål.
- Vær opmærksom på identifikation via funktionelle form antagelser.

Intern vs. ekstern validitet

- Intern validitet: "Within-sample" prædiktioner.
- Ekstern validitet: "Out-of-sample" prædiktioner.