

Økonometri I

Forelæsning 3

Mette Ejrnæs

16. februar 2021 kl. 8-10

Program for forelæsning:

Lektion 3: W2.3–W2.6 .

- Funktionel form.
- Antagelser for den simple lineære regressionsmodel.
- Middelrethed af OLS estimator.
- Variansen af OLS estimator.

Prædikterede værdier og residualer:

Variationen i y kan også dekomponeres:

- Total variation (SST) $SST = \sum_{i=1}^n (y_i - \bar{y})^2$ (variationen i y).
- Forklaret variation (SSE) $SSE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$.
- Residual variationen (SSR) $SSR = \sum_{i=1}^n \hat{u}_i^2$.
- Den totale variation

$$SST = SSE + SSR.$$

- Ekstra opgave: Vis ovenstående. Hint

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \dots$$

Goodness of fit:

Udfra dekomponeringen af den totale variation kan man konstruere et mål for, hvor meget af variation regressionsligningen forklarer

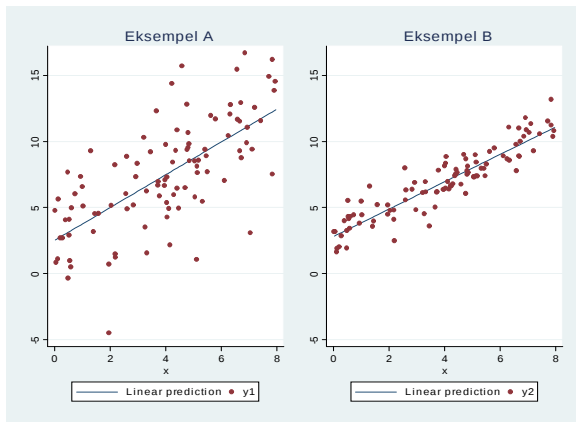
$$R^2 = \frac{SSE}{SST}.$$

Der gælder om R^2 :

- R^2 ligger mellem 0 og 1.
- Jo højere R^2 er jo mere af variationen er forklaret.

Eksempler:

Besvar spørgsmålet: Hvilken figur har det største R^2 ?



Eksempel lønligningen:

- Estimation af simpel model for timeløn og uddannelse:

$$timeløn = \beta_0 + \beta_1 uddannelse + u.$$

- Vi benytter data fra Danmarks Statistik.
- Data indeholder 1.078 observationer for personer som er tilfældigt udvalgt i 1994.
- For hvert individ har vi information om timelønnen, køn, antal års uddannelse, single, alder og erhvervserfaring.
- Datasættet er afgrænset således:
 - 20-68 årige.
 - Lønmodtagere.
 - Timeløn på mere end 40 kr.
 - Vi tilføjer nu: $Uddannelse > 0$
- Data kan findes på Absalon.

Eksempel løn ligningen:

Estimation med STATA.

- STATA program

```
-----  
-----  
clear all  
cd "c:\undervisning\data"  
use wage  
regress wage edu if edu>0  
-----
```

- Resultater:

$$\widehat{timeløn} = 60.23 + 6.69 \cdot uddannelse_i, \quad i = 1, \dots, 1,046$$

- Fortolkning?

Enhederne og funktionel form:

- Kan (simpel) regressionsanalyse anvendes på mere generelle modeller? Ja.
- Det er muligt at lave lineære transformationer af modellen, men parameterestimerne ændrer sig
- Man behøver ikke at have en lineær relation mellem y og x (så længe modellen er lineær i parametrene)

$$g(y) = \beta_0 + \beta_1 f(x) + u.$$

- Fortolkningen af parametrene bliver anderledes.

Enhederne og funktionel form:

Regresionsmodel:

$$timeløn_i = \beta_0 + \beta_1 uddannelse_i + u_i$$

Hvad sker der, når man benytter andre enheder for den afhængige variabel?

F.eks. måler timelønnen i 2010-DKK:

$$timeløn_i^{2010} = timeløn_i^{1994} \times 1.37$$

Hvad sker der, hvis man benytter andre enheder for den forklarende variabel?

F.eks. måler uddannelse i måneder i stedet for år?

$$uddannelse_i^{måned} = 12 \times uddannelse_i^{år}$$

Hvad hvis man bruger $\log(timeløn_i)$ i stedet for $timeløn_i$ som den afhængige variabel?

Lønregression:

:

$$timeløn_i = \beta_0 + \beta_1 uddannelse_i + u_i$$

	(1)	(2)	(3)	(4)
Afhængig. var.	$timeløn^{94}$	$timeløn^{10}$	$timeløn^{94}$	$\log(timeløn^{94})$
<i>Const</i>	60.23	82.53	60.23	4.35
<i>uddannelse</i> ^{år}	6.69	9.17	-	0.05
<i>uddannelse</i> ^{måned}	-	-	0.558	-
<i>N</i>	1,046	1,046	1,046	1,046
<i>R</i> ²	0.142	0.142	0.142	0.140

Bemærk: 32 observationer er UDELADT, fordi observationerne havde uddannelse=0.

Lønregression:

I arbejdsmarkedsøkonomi estimerer man ofte lønligningen med følgende regressionsmodel (dog er der som regel flere forklarende variable):

$$\log(\text{timeløn}_i) = \beta_0 + \beta_1 \text{uddannelse}_i + u_i$$

Fortolkningen af parameteren, β_1 :

- Den relative ændring af timelønnen ved at tage et års ekstra uddannelse (ceteris paribus)
- $100\beta_1$ er ca. det procentvise afkast af et års ekstra uddannelse

Bemærk: vi antager samme procentvise afkast uanset niveauet af uddannelse.

Funktionel form:

Almindelige former

Model	Afh var.	Forkl. var.	Fortolkning af β_1	Elasticitet
Level-level	y	x	$\frac{\partial y}{\partial x}$	$\beta_1 \frac{x}{y}$
Log-level	$\log(y)$	x	$\frac{1}{y} \frac{\partial y}{\partial x}$	$\beta_1 x$
Level-log	y	$\log(x)$	$x \frac{\partial y}{\partial x}$	$\frac{\beta_1}{y}$
Log-log	$\log(y)$	$\log(x)$	$\frac{x}{y} \frac{\partial y}{\partial x}$	β_1

Funktionel form:

Den lineære regressionsmodel skal være lineær i parametrene, og ikke nødvendigvis i variablene.

Er de følgende modeller en lineær regressionsmodel?

Svar på Absalon.

- ❶ $y_i = \beta_0 + \beta_1 \sqrt{x_i} + u_i$
- ❷ $\exp(y_i) = \beta_0 + \beta_1 \sqrt{x_i + 3} + u_i$
- ❸ $Y_i = AL_i^{\beta_1} u_i$
- ❹ $\log y_i = \beta_0 x_i^{\beta_1} + u_i$
- ❺ Hvilken model kan laves om så den bliver lineære?

OLS estimatoren er middelret

Definition (see appendix C2)

Antag vi har en estimator **b** (stokastisk variabel)

En estimator **b** for β er **middelret** hvis:

$$E(\mathbf{b}) = \beta$$

for alle værdier af β

- Middelrethed er en statistisk egenskab.
- Hvorfor er det vigtigt, at en estimator er middelret?

OLS estimatoren er middelret

Antagelser for den simple lineære regressionsmodel

SLR.1 Lineær i parametrene:

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

SLR.2 Tilfældig udvælgelse:

Vi har tilfældigt udvalgte observationer $\{(x_i, y_i) : i = 1, \dots, n\}$ fra en population. Det betyder, at observationerne er uafhængige.

SLR.3 Variation i x :

I datasættet må x ikke antage den samme værdi for alle observationer

SLR.4 Den betingede middelværdi af fejleddet skal være 0: $E(u|x) = 0$

Middelret OLS estimator:

Theorem 2.1: Middelrethed af OLS estimatoren

Under antagelse af SLR.1–SLR.4, er OLS estimatoren **middelret**:

$$E(\hat{\beta}_0) = \beta_0$$

$$E(\hat{\beta}_1) = \beta_1$$

Bevis for $E(\hat{\beta}_1) = \beta_1$

Bevis for OLS er middelret:

Ingredienser til beviset:

SLR.1: Model: $y_i = \beta_0 + \beta_1 x_i + u_i$

SLR.3: $SST_x = \sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$

SLR.4: $E(u|x) = 0$

OLS estimatoren: $\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n x_i (x_i - \bar{x})}$,

Regneregler:

- $\sum_{i=1}^n (x_i - \bar{x}) = 0$
- $SST_x = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i (x_i - \bar{x})$
- $E(a(x)y + b(x)|x) = a(x)E(y|x) + b(x)$ (se Appendix Math refresher B-4f)

Variansen af OLS estimatoren:

- Selvom OLS estimatoren er middelret, er det ikke det samme som, at OLS estimatoren er lig den sande parameter ($\hat{\beta}_1$ kan være forskellig fra β_1).
- OLS estimatoren er en stokastisk variabel, og derfor har den en varians (og en middelværdi, som vi lige har vist er lig de sande parameter).

Vi vil nu finde variansen af OLS estimatoren, men så behøver vi en ekstra antagelse:

SLR.5 Variansen af fejleddet er konstant

$$\text{Var}(u|x) = \sigma^2.$$

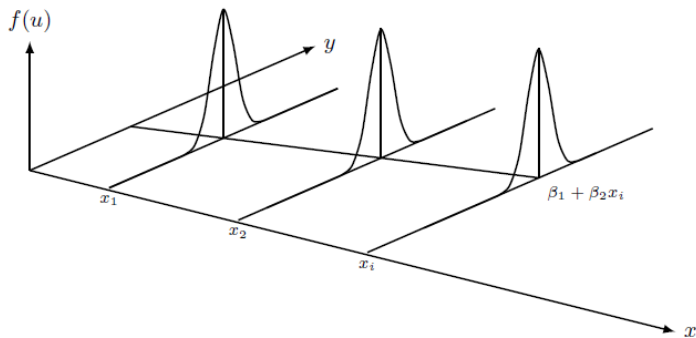
Variansen af OLS estimatoren:

SLR.5 Variansen af fejleddet er konstant

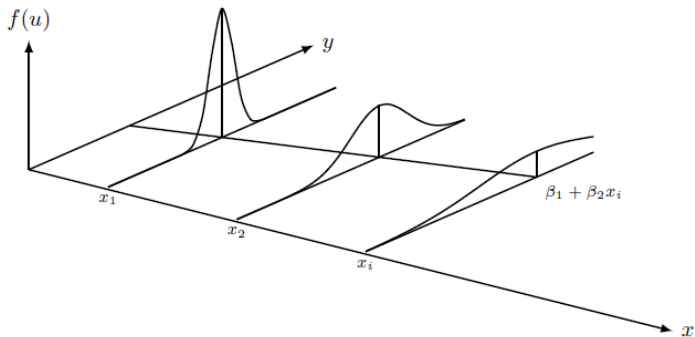
$$\text{Var}(u|x) = \sigma^2.$$

- Vi behøver antagelserne SLR.1-SLR.4 for at vise, at OLS er middelret, men vi behøver **ikke** SLR.5.
- Antagelserne SLR1-SLR.5 kaldes **Gauss-Markov** antagelserne
- Antagelsen SLR.5 gør det lettere at udlede variansen af OLS estimatoren, men vi kan også udlede variansen af OLS estimatoren under andre antagelser (det kommer i kapitel 8).

Homoskedasticitet:



Heteroskedasticitet:



Homoskedasticitet:

Hvad betyder homoskedasticitet?

- Det betyder at variansen er konstant (ikke afhænger af x).
- Vi kan vise, at når SLR.5 er opfyldt, så gælder.

$$\begin{aligned} \text{Var}(u|x) &= E([u - E(u|x)]^2|x) \\ &= E([u]^2|x) \\ &= E(u^2|x) = \sigma^2. \end{aligned}$$

Der gælder også, at:

$$\sigma^2 = E(u^2|x) = E(u^2) = \text{Var}(u)$$

Altså den ubetingede varians af fejlleddet er også σ^2 .

Variansen af OLS:

Teorem 2.2

Under antagelse af SLR.1.-SLR.5 gælder der, at variansen af OLS estimatoren er:

$$\begin{aligned} \text{Var}(\hat{\beta}_1|x) &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \text{Var}(\hat{\beta}_0|x) &= \frac{\sigma^2 \frac{1}{n} \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Bevis for variansen af OLS:

Ingredienser til beviset:

OLS estimatoren: $\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$

Antagelser:

- SLR.1: Model: $y_i = \beta_0 + \beta_1 x_i + u_i$
- SLR.2: observationerne er uafhængige.
- SLR.3: $SST_x = \sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$
- SLR.4: $E(u|x) = 0$
- SLR.5: $Var(u) = \sigma^2$

Regneregler:

- $Var(a(x)y + b(x)|x) = [a(x)]^2 Var(y|x)$
- Hvis y_1 og y_2 er uafhængige: $Var(y_1 + y_2) = Var(y_1) + Var(y_2)$

Variansen af OLS:

Variansen af fejleddet σ^2 er ukendt.

Vi kan bruge følgende estimator for variansen af fejleddet

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{n-2}$$

Teorem 2.3

Under antagelse af SLR.1-SLR.5 er estimatoren for variansen af fejleddet middelfret:

$$E(\hat{\sigma}^2) = \sigma^2$$

Vi benytter i beviset at

$$E((\hat{\beta}_1 - \beta_1)^2 | x) = \text{Var}(\hat{\beta}_1 | x)$$

Variansen af OLS (Bevis):

Bevis:

Vi antager at SLR.1-SLR.5 er opfyldt. Vi ved at

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)x_i \quad (1)$$

Da gennemsnittet af residualerne er lig 0 har vi

$$\begin{aligned} 0 &= \frac{1}{n} \sum \hat{u}_i = \bar{u} - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)\bar{x} \Leftrightarrow \\ -(\hat{\beta}_0 - \beta_0) &= -\bar{u} + (\hat{\beta}_1 - \beta_1)\bar{x} \end{aligned} \quad (2)$$

Ved at indsætte (2) i (1) får vi

$$\begin{aligned} \hat{u}_i &= u_i - \bar{u} - (\hat{\beta}_1 - \beta_1)(x_i - \bar{x}) \Rightarrow \\ (\hat{u}_i)^2 &= (u_i - \bar{u})^2 + (\beta_1 - \hat{\beta}_1)^2(x_i - \bar{x})^2 - \\ &\quad 2(u_i - \bar{u})(\hat{\beta}_1 - \beta_1)(x_i - \bar{x}) \end{aligned}$$

Variansen af OLS (Bevis):

Vi regner nu på

$$\begin{aligned}\sum_{i=1}^n (\hat{u}_i)^2 &= \sum_{i=1}^n (u_i - \bar{u})^2 + \sum_{i=1}^n (\beta_1 - \hat{\beta}_1)^2 (x_i - \bar{x})^2 \\ &\quad - 2 \sum_{i=1}^n (u_i - \bar{u})(\hat{\beta}_1 - \beta_1)(x_i - \bar{x}) \\ &= \sum_{i=1}^n (u_i - \bar{u})^2 + (\beta_1 - \hat{\beta}_1)^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &\quad - 2(\hat{\beta}_1 - \beta_1) \sum_{i=1}^n (u_i - \bar{u})(x_i - \bar{x}) \\ &= \sum_{i=1}^n (u_i - \bar{u})^2 + (\beta_1 - \hat{\beta}_1)^2 SST_x - 2(\hat{\beta}_1 - \beta_1) \sum_{i=1}^n u_i(x_i - \bar{x})\end{aligned}$$

Variansen af OLS (Bevis):

Første led:

$$E\left(\sum_{i=1}^n (u_i - \bar{u})^2 | x\right) = (n-1)\sigma^2$$

Andet led:

$$\begin{aligned} E((\beta_1 - \hat{\beta}_1)^2 SST_x | x) &= SST_x \text{Var}(\hat{\beta}_1 | x) \\ &= SST_x \frac{\sigma^2}{SST_x} = \sigma^2 \end{aligned}$$

Variansen af OLS (Bevis):

Tredje led:

Her benytter vi at

$$(\hat{\beta}_1 - \beta_1) = \frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{SST_x}.$$

Tredje led er derfor

$$\begin{aligned} 2(\hat{\beta}_1 - \beta_1) \sum_{i=1}^n u_i(x_i - \bar{x}) &= 2 \frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{SST_x} \sum_{i=1}^n u_i(x_i - \bar{x}) \\ &= 2 \frac{(\sum_{i=1}^n u_i(x_i - \bar{x}))^2}{SST_x} \end{aligned}$$

Variansen af OLS (Bevis):

Vi har nu at

$$\begin{aligned} E\left(\left(\sum_{i=1}^n u_i(x_i - \bar{x})\right)^2 \mid x\right) &= \text{Var}\left(\sum_{i=1}^n u_i(x_i - \bar{x})\right) \\ &= \left(\sum_{i=1}^n \text{Var}(u_i(x_i - \bar{x}) \mid x)\right) \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 \text{Var}(u_i \mid x) \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2 \\ &= \sigma^2 SST_x \end{aligned}$$

Første lighedstegn fås ved at $E((\sum_{i=1}^n u_i(x_i - \bar{x})) \mid x) = 0$ (SLR.4). Andet lighedstegn fås ved at benytte, at observationerne er uafhængige (SLR.2). Den betingede varians af en sum er da summen af de betingede varianser. Fjerde lighedstegn fås ved at bruge SLR.5.

Variansen af OLS (Bevis):

$$\begin{aligned} E\left(\sum_{i=1}^n (\hat{u}_i)^2\right) &= (n-1)\sigma^2 + \sigma^2 - 2\frac{SST_x}{SST_x}\sigma^2 \\ &= (n-2)\sigma^2 \end{aligned}$$

Variansen af OLS:

Vi kan nu udregne **standardfejlen** for estimatoren (ikke standardafvigelsen).

Under antagelserne SLR.1-SLR.5 er standardfejlen for OLS estimatoren

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Standardfejllen giver et mål for variationen i OLS estimatoren (som skyldes den tilfældige stikprøve).

Vi skal bruge standardfejlen når vi skal til at lave hypotesetest.

Nøglebegreber fra forelæsningsen:

- Antagelserne for OLS: SLR1-SLR.5
- OLS estimatoren er middelret når SLR.1-SLR.4 er opfyldt
- Variansen af OLS estimatoren er (når SLR.1-SLR.5)

$$\text{Var}(\hat{\beta}_1 | x) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

hvor variansen kan estimeres som

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{n-2}$$

Næste gang torsdag d. 18.2:

Lektion 4: W3.1–W3.2

Forberedelse: repeter regneregler for middelværdier og varians (tag quizzen)