

Økonometri I

Forelæsning 2

Mette Ejrnæs

11. Februar 2021 kl. 10-12

Program for forelæsning:

Lektion 2: W2.1–W2.3 + appendix 2A, C4

- Motivation for den lineære regression model.
- Udledning af OLS.
- Goodness of fit.
- Eksempel: Lønregression.

Lektion 3: W2.3–W2.6

- Funktional form.
- Antagelser for den simple lineære regression model.
- Middelrethed af OLS estimator.
- Variansen af OLS estimator.

Motivationen for regressionsanalyser:

Vi er interesseret i at kunne forklare variationen i y med variationen x .
Eksempler:

- Hvordan påvirker gødning produktionen af sojabønner?
- Hvordan påvirker uddannelse timelønnen?
- Hvordan påvirker lønnen for den administrerende direktør profitten?
- Hvordan afhænger væksten i BNP af landenes initiale BNP?

Definition af simpel linære regressionsmodel

Definition af SLR: Simpel Lineære Regression model:

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

y : Afhængig variable

x : Forklarende variable

u : Unobserveret fejllid

β_0 : Intercept (konstantled)

β_1 : Hældningskoefficient

Motivationen for regressionsanalyser:

Når vi specificerer en lineær regressionsmodel, skal vi forholde os til følgende:

- Spørgsmål 1: Hvordan håndterer vi, at der kan være mange andre faktorer end x som påvirker y ?
- Spørgsmål 2: Hvordan er den funktionelle form, som giver en god beskrivelse af sammenhængen mellem x og y ?
 - Er det bedre at lade y afhænge af $\log(x)$?
 - Er det bedre at modellere $\log(y)$, som funktion af x ?
- Spørgsmål 3: Kan parameteren β_1 fortolkes, som som en **ceteris paribus** effekt eller kausal effekt?
 - Det er ofte nødvendigt her med yderligere antagelser.

Definition af SLR:

Når vi opstiller en lineær regressionsmodel har vi ofte implicit besvaret spørgsmålene:

Spørgsmål 1: Hvordan håndterer vi, at der kan være mange andre faktorer end x som påvirker y ?

- Alle faktorer, som påvirker y bortset fra x , er indeholdt i fejlleddet u .
- Fejlleddet indeholder:
 - Udeladte variable (omitted variable).
 - Målefejl.

Definition af SLR:

Spørgsmål 1:

- Hvad indeholder fejleddet u i eksemplet, hvor y er udbyttet af sojabønner og x er gødningsmængden?

Definition af SLR:

Spørgsmål 1:

- Hvad indeholder fejlleddet u i eksemplet, hvor y er udbyttet af sojabønner og x er gødningsmængden?
- Hvad indeholder fejlleddet u i eksemplet, hvor y er timeløn og x er uddannelse (målt i år)?

Definition af SLR:

Spørgsmål 2:

- Vi antager at y er en **lineær** funktion af x .
- Den lineære sammenhæng er restriktiv og ikke altid tilfredsstillende.
 - Eksempel: hvordan forestiller du dig sammenhængen mellem gødning og sojabønner ser ud? % højst sandsynligt aftagende på et tidspunkt / der er en optimal mængde gødning.

Definition af SLR:

Spørgsmål 3: **ceteris paribus** fortolkninger?

- Vi kan ikke lave **ceteris paribus** fortolkninger uden ekstra antagelser.
- Parameteren β_1 beskriver, hvordan y afhænger af x , hvis $\Delta u = 0$.
- Hvornår vil det være et problem:
 - Eks 1: Gødning og udbytte? Uobserverbar jordkvalitet.
 - Eks 2: Timeløn og uddannelse? Uobserverbare evner.
- Antagelser vedr. u er helt centrale i økonometri.
 - Ofte er antagelserne meget svære at validere.

Antagelser om fejleddet:

- Antagelse 2.5 i Wooldridge

$$E(u) = 0$$

- Antagelse 2.6 i Wooldridge.
- Betingede middelværdi af u er

$$E(u|x) = E(u).$$

- Det er ofte kritisk og en helt central antagelse.
- Hvad betyder antagelsen?
 - Når x og u er uafhængige gælder der $E(u|x) = E(u)$ og alle funktioner af x er ukorrelede med u . Specielt gælder at $cov(u, x) = 0$.
 - (se Math Refresher B4, CE.3-CE.5).

Antagelser om fejlleddet:

Eksempel: Timeløn og uddannelse.

- Model

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

- Fejlleddet indeholder evner, produktivitet.
- Er følgende antagelser rimelige?

$$\begin{aligned} E(evner | uddannelse = 9) &= E(evner | uddannelse = 17) \\ E(evner | folkeskolen) &= E(evner | kandidatudd.). \end{aligned}$$

Estimation:

- Vi er interesseret i at estimere β_0 og β_1 fra modellen:

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

- En måde er at benytte Maximum Likelihood estimation (som I har set i SandStat):
 - Lav en antagelse om fordelingen af u f.eks. at u er normalfordelt.
 - Opskriv likelihood funktionen og maksimer likelihood funktionen mht. β_0 og β_1 .
- En anden tilgang er **Method of Moments** (momentestimation) (MM):
 - kræver ikke, at vi antager om fordelingen af u .
 - kræver kun at

$$E(u|x) = E(u) = 0$$

Eksempel:

- Vi har en variabel y , hvor $E(y) = \mu$.
- μ er ukendt og vi ønsker at estimere μ .
- Vi har n observationer af variabelen y .
- Hvad vil være en naturlig estimator for μ ?
- Den empiriske middelværdi (gennemsnittet)

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i.$$

er middelret estimator for μ (se math refresher C2) .

- Ideen ved MM er at erstatte det teoretiske moment (populationsmomentet) (middelværdi, varians, tredjemomentet) med det empiriske moment (f.eks. at erstatte $E(y)$ med $\frac{1}{n} \sum_{i=1}^n y_i$).
- Vi benytter her "Store tals lov" ("law of large numbers")

Ideen bag momentestimation:

Varansen af y , hvor $E(y) = \mu_y$:

$$\text{Var}(y) = \sigma_y^2 = E \left[(y - \mu_y)^2 \right].$$

MM estimator for variansen:

$$\hat{\sigma}_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2, \quad i = 1, \dots, n.$$

Kovariansen mellem x og y :

$$\text{Cov}(x, y) = \sigma_{xy} = E \left[(x - \mu_x)(y - \mu_y) \right].$$

MM estimator af kovariansen:

$$\hat{\sigma}_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad i = 1, \dots, n.$$

Udledning af OLS estimatoren:

Regression model:

$$y = \beta_0 + \beta_1 x + u. \quad (1)$$

Antagelser:

$$E(u) = 0 \quad (2)$$

$$E(u|x) = 0. \quad (3)$$

- Omformuler antagelse (3): $E(u|x) = 0 \implies \text{cov}(x, u) = 0$.

Udledning af OLS estimatoren:

Regression model:

$$y = \beta_0 + \beta_1 x + u. \quad (1)$$

Antagelser:

$$E(u) = 0 \quad (2)$$

$$E(u|x) = 0. \quad (3)$$

- Omformuler antagelse (3): $E(u|x) = 0 \implies \text{cov}(x, u) = 0$.
- Så,

$$\text{cov}(x, u) \equiv E[(x - E(x))(u - E(u))] \quad (\text{def}) \quad (4)$$

$$= E[(x - E(x))u] \quad (\text{anvend antagelse 2})$$

$$= E[xu] - E[E(x)u]$$

$$= E[xu] - E(x)E(u) \quad (\text{tag forventningen})$$

$$= E(xu) = 0. \quad (5)$$

Udledning af OLS estimatoren:

Vi finder først de relevante populationsmomenter.

Udtryk fejlleddet u ved x og y ved at bruge modellen (1):

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

Indsæt udtrykket i antagelserne (2) og (3):

- $E(u) = E(y - \beta_0 - \beta_1 x) = 0$
- $E(xu) = E(x(y - \beta_0 - \beta_1 x)) = 0.$

Dette er teoretiske momenter (populationsmomenter). De tilsvarende datamomenter er:

- $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
- $\frac{1}{n} \sum_{i=1}^n (x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)) = 0.$

Man kan nu bestemme estimatorerne for β_0 og β_1 ved at løse ligningssystemet (2 ligninger, 2 ubekendte).

Udledning af OLS estimatoren:

Ingredienser til udledningen:

Regneregler for summer:

- $\frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$
- $\frac{1}{n} \sum_{i=1}^n a x_i = \frac{a}{n} \sum_{i=1}^n x_i = a \bar{x}$
- $\frac{1}{n} \sum_{i=1}^n a = \frac{a}{n} \sum_{i=1}^n 1 = a$
- $\frac{1}{n} \sum_{i=1}^n x_i (y_i + z_i) = \frac{1}{n} \sum_{i=1}^n x_i y_i + \frac{1}{n} \sum_{i=1}^n x_i z_i$

Antagelse

- $\sum_{i=1}^n (x_i - \bar{x})^2 > 0$

Udledning af OLS estimatoren:

OLS estimatoren:

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

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 en 40 kr.
- Data kan findes på Absalon.

Eksempel lønningen:

Estimation med STATA.

- STATA program

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

- Resultater:

$$\widehat{timeløn} = 90.3 + 4.3 * uddannelse_i, \quad i = 1, \dots, 1,078$$

- Fortolkning?

Prædikterede værdier og residualer:

- Population regressionsmodellen

$$E(y|x) = \beta_0 + \beta_1 x.$$

- Ved at bruge vores parameterestimer kan vi finde den prædikterede værdi af y

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x.$$

- Residualerne kan beregnes som

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x.$$

Prædikterede værdier og residualer:

- For en OLS estimation med konstantled gælder der altid, at
 - $\sum_{i=1}^n \hat{u}_i = 0$.
 - $\sum_{i=1}^n \hat{u}_i x_i = 0$.
 - Hvorfor er dette ikke så overraskende...
- OLS estimatoren kan også udledes ved at minimere summen af kvadrerede residualer (det er det man gør i ML estimation):

$$\sum_{i=1}^n (\hat{u}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

- Appendix 2A (side 65) viser, at OLS estimatoren løser dette minimeringsproblem.
- De empiriske momenter er første ordensbetingelserne til dette problem.

Prædikterede værdier og residualer:

Dekomponering af y :

- Den prædikterede værdi $\hat{\beta}_0 + \hat{\beta}_1 x$.
- Residualen $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$.
- $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$.
- Positive residualer viser, at den faktiske værdi er større end den prædikterede værdi.
- Negative residualer viser, at den faktiske værdi er mindre end den prædikterede værdi.

Quizz: prædikterede værdier og residualer:

Svar på følgende spørgsmål på Absalon:

Lad $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ og $\hat{u}_i = y_i - \hat{y}_i$

Spørgsmål: SANDT eller FALSK

$$A : \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \sum_{i=1}^n y_i$$

$$B : \frac{1}{n} \sum_{i=1}^n y_i = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n x_i$$

$$C : \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) \hat{u}_i = 0$$

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.

Nøglebegreber fra forelæsningen:

Momentestimatoren: OLS

Husk:

Forskel på

- Populationsparametre β_0 og β_1 og
- Estimatet (baseret på data og en "regneregel") : $\hat{\beta}_0$ og $\hat{\beta}_1$

Forskel på

- Statistisk antagelse: $E(u|x) = E(u) = 0$
- Mekaniske egenskaber for residualerne: $\sum_{i=1}^n \hat{u}_i = 0$ og $\sum_{i=1}^n \hat{u}_i x_i = 0$.

Næste gang:

Tirsdag 16.2

Lektion 3: W2.3–W2.6

- Husk holdundervisningen starter i næste uge
- Information om holdundervisningen se Absalon.
- Husk at have adgang til en computer med STATA og "Problem set 1".