

Økonometri I

Forelæsning 4

Mette Ejrnæs

18. februar 2021 kl. 10-12

Program for forelæsning:

Lektion 4: W3.1–W3.2, Math refresher: Appendix D1-D3, E1 .

- Det sidste fra kapitel 2
- Den multiple regressionsmodel
- Definition
- Egenskaber.

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.

Variansen af OLS:

Overvej hvad der påvirker standardfejlen:

Svar på følgende spørgsmål:

- ➊ Stiger standardfejlen, når der er mere variation i x ?
- ➋ Stiger standardfejlen, når der er mere variation i u ?
- ➌ Stiger standardfejlen, når vi har flere observationer?

Den multiple regressionsmodel:

Den multiple regressionsmodel

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$$

Notation

- y den afhængige variabel
- $x_1, \dots, x_k$ de k forklarende variable
- β_0 konstantled
- $\beta_1, \dots, \beta_k$ k hældningskoefficienter
- u fejleddet
- k antallet af forklarende variable
- $k + 1$ parameter i alt

Motivation:

- Ofte er der flere faktorer som påvirker y
 - Den multiple regressionsmodel kan eksplicit kontrollere for flere faktorer
 - Disse faktorer er ikke længere inkluderet i u
 - Ofte lettere at få opfyldt antagelsen om $E(u|x) = 0$
- Det gør det lettere at lave ceteris paribus fortolkninger
 - Vi kan nu eksplicit holde faktorerne $x_2, \dots, x_k$ faste, når vi bestemmer effekten af at ændre x_1 med en enhed.
- Muligt at have modeller med mere fleksible funktionelle former

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$$

..dog stadig lineær i parametrene

Multiple lineære regressionsmodel på matrix form:

Se appendix E1:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & & \ddots & \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

- y og u er $n \times 1$ matricer (vektor)
- X er en $n \times (k + 1)$ matrix
- β er en $(k + 1) \times 1$ matrix (vektor)

Multiple lineære regressionsmodel:

Modellen kan nu skrives op på følgende måde

$$\underset{n \times 1}{y} = \underset{n \times (k+1)}{X} \underset{(k+1) \times 1}{\beta} + \underset{n \times 1}{u}.$$

- OLS estimatoren kan udledes som en momentestimator (MM)
- Momenterne vi bruger er

$$E(u|X) = 0 \Rightarrow$$

$$E(u) = 0$$

$$E(ux_j) = 0, \quad j = 1, \dots, k$$

- Vi har $k + 1$ momentbetingelser (ligeså mange som antallet af parametre)

Udledning af OLS:

Det teoretiske moment ($k + 1$ momenter)

$$E(X'u) = 0$$
$$E \left[\begin{pmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{n1} \\ x_{1k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \right] = \begin{pmatrix} E(\sum_{i=1}^n u_i) \\ E(\sum_{i=1}^n x_{i1} u_i) \\ E(\sum_{i=1}^n x_{ik} u_i) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Vi finder OLS ved at erstattet dette med det empiriske moment

$$X'\hat{u} = 0$$
$$\begin{pmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{n1} \\ x_{1k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \hat{u}_i \\ \sum_{i=1}^n x_{i1} \hat{u}_i \\ \vdots \\ \sum_{i=1}^n x_{ik} \hat{u}_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

hvor $\hat{u} = y - X\hat{\beta}$. Dette er $k + 1$ ligninger.

Udledning af OLS:

Udledning

$$\begin{aligned}X'\hat{u} &= 0 \Leftrightarrow \\X'(y - X\hat{\beta}) &= 0 \Leftrightarrow \\X'y - X'X\hat{\beta} &= 0 \Leftrightarrow \\X'y &= X'X\hat{\beta}\end{aligned}$$

Hvis X har fuld rang, så eksisterer $(X'X)^{-1}$

$$\hat{\beta} = (X'X)^{-1}X'y$$

Fuld rang?:

Hvornår har en matrix fuld rang?

Søjlerne i X skal være lineært uafhængige

Et eksempel på matricer som ikke har fuld rang

- Hvis der er en variabel, som er konstant (og vi har et konstantled i modellen - første søjle er 1-taller)
- Hvis der findes en lineær relation mellem en eller flere variable
- En model hvor der indgår konstantled, alder (per 31. december i et givent år f.eks. 2017) og fødselsår
- Alder kan fås som en lineær relation mellem konstantled og fødselsår

$$alder = 2017 - fødselsår$$

- Alder og fødselsår er perfekt korreleret.
- Vi kan ikke både estimere effekten af alder og fødselsår.

Fortolkning af parametre:

Antag følgende model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$$

Prædikterede værdier af y

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$$

Ændringer i den prædikterede værdi

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \dots + \hat{\beta}_k \Delta x_k$$

Fortolkning af parametre:

Hvis vi holder alle øvrige forklarende variable konstante undtagen x_j

$$\Delta \hat{y} = \hat{\beta}_j \Delta x_j$$

Den multiple regressionsmodel gør det muligt at lave ceteris paribus fortolkninger, selvom vi i data ikke kan holde alle faktorer konstante.

Vi siger, at vi "kontrollerer for de øvrige effekter"

Lønregression:

I lønregression tilføjer vi nu en ekstra variabel

$$\log(\text{timeløn}) = \beta_0 + \beta_1 \text{uddannelse} + \beta_2 \text{erfaring} + u$$

Fortolkningen af β_1

Lønregression:

	SLR	MLR
Afhængig. var.	$\log(\text{timeløn})$	$\log(\text{timeløn})$
<i>Const</i>	4.350	4.105
<i>uddannelse</i>	0.045	0.048
<i>erfaring</i>	-	0.014
<i>N</i>	1,046	1,046
<i>R</i> ²	0.140	0.275

OLS estimatoren i en model med $k=2$:

OLS estimatoren for en multipel regressionsmodel kan bestemmes ved at lave to simple regressionsestimationer.

Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Man kan bestemme OLS-estimatoren for β_1 ved at udføre to simple lineære regressionsmodeller.

Procedure:

- 1 Regresser x_1 på x_2 og en konstant og gem residualerne som $\hat{r}_1$
- 2 Regresser y på $\hat{r}_1$

OLS estimatoren i en model med $k=2$:

Trin 1

Regresser x_1 på x_2 og gem residualerne som $\hat{r}_1$

$$x_1 = \alpha_0 + \alpha_1 x_2 + r_1$$

Vi kan nu beregne residualerne $\hat{r}_1 = x_1 - \hat{\alpha}_0 - \hat{\alpha}_1 x_2$.

Trin 2

Regresser y på $\hat{r}_1$

$$y = \delta_0 + \beta_1 \hat{r}_1 + e$$

OLS estimatoren kan fås som

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i \hat{r}_{1i}}{\sum_{i=1}^n \hat{r}_{1i}^2}$$

I skal vise dette i den første obligatoriske opgave.

Dette kaldes for **Frisch-Waugh Teoremet**

Mekaniske egenskaber ved OLS:

- For en OLS estimation med konstantled gælder der altid, at
 - $\sum_{i=1}^n \hat{u}_i = 0$.
 - $\sum_{i=1}^n \hat{u}_i x_{ji} = 0$ for $j = 1, \dots, k$.
- Punktet $(\bar{y}, \bar{x}_1, \bar{x}_2, \dots, \bar{x}_k)$ ligger på regressionslinjen dvs.
$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \dots + \hat{\beta}_k \bar{x}_k$$
- Gennemsnittet af $\hat{y}$ er lig $\bar{y}$

Prædikterede værdier og residualer:

Dekomponering af y :

- Den prædikterede værdi $\hat{y}$.
- Residualen $\hat{u}_i = y_i - \hat{y}_i$

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

$$SST = SSE + SSR.$$

Goodness of fit:

Ud fra 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 :

- $R^2 \in [0, 1]$.
- Jo højere R^2 er, jo mere af variationen er forklaret.
- Flere variable vil altid få R^2 til at stige (eller være uændret).
- Man skal ikke alene bruge R^2 til at afgøre, hvor mange variable der skal med i modellen.
- R^2 kan ikke sammenlignes, hvis den afhængige variabel er forskellig

Goodness of fit:

Antag vi har en multipel lineær regressionsmodel:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$$

Vi laver nu en omskalering af variabelen x_1 : $\tilde{x}_1 = 2x_1$. $\tilde{x}_1$ indsættes istedet for x_1 og modellen estimeres med OLS.

Hvilket af følgende udsagn er sandt:

- R^2 falder
- SSR stiger
- SSR er uændret

Regressionsmodel uden konstantled:

Det kan nogle gange være relevant at have en regressionsmodel **uden** konstantled.

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$$

Så skal man være opmærksom på, at nogle af egenskaberne ikke holder længere:

- Summen af residualerne $\sum_{i=1}^n \hat{u}_i$ er ikke nødvendigvis 0.
- R^2 kan blive negativ.
- Hvis der faktisk er et konstantled i den "sande model", vil OLS estimationen af en model uden konstantled være en **ikke-middelret estimator**.

Nøglebegreber:

- Multipel regressionsmodel - en generalisering af den simple lineære regressionsmodel.
- Matrixnotation for regressionsmodellen.
- Udledning af momentestimator

Næste gang på tirsdag d. 23.2:

Lektion 5: W3.3–W3.4+ Appendix E.2