ECONnect NTNU

Faktor

- en eksamensavis utgitt av ECONnect

Pensumsammendrag: SØK3001 – Økonometri I

Forfatter: Drago Bergholt
E-post: bergholt@stud.ntnu.no

Skrevet: Våren 2010
Antall sider: 71

Om ECONnect:

ECONnect er en frivillig studentorganisasjon for studentene på samfunnsøkonomi- og finansøkonomistudiet ved NTNU. Vi arbeider for økt faglig kompetanse blant våre studenter samt tettere kontakt med næringslivet. Det gjør vi ved å arrangere fagdager, gjesteforelesninger, bedriftspresentasjoner m.m. I dag går det ca. 200 studenter på bachelornivå (1.-3. klasse) og ca. 70 studenter på masternivå (4.-5. klasse). Studentene på masternivå er fordelt på de to linjene samfunnsøkonomi (ca. 50 stk) og finansiell økonomi (ca. 20 stk). Mer om ECONnect og aktuelle arrangementer på www.econnect-ntnu.no.

ECONnect består av følgende personer ved utgivelsestidspunkt:

Bjørn Bergholt (Leder)	bjorn@econnect-ntnu.no
Sophie S. Strømman (Bedriftsansvarlig)	sophie@econnect-ntnu.no
Maiken Weidle (Fagdagsansvarlig)	maiken@econnect-ntnu.no
Joakim Bjørkhaug (Økonomi- og IT-ansvarlig)	joakim@econnect-ntnu.no
Elise Caspersen	elise@econnect-ntnu.no
Tiril Toftedahl	tiril@econnect-ntnu.no
Louis Dieffenthaler	louis@econnect-ntnu.no
Andreas H. Jung	andreas@econnect-ntnu.no
Mari Benedikte Ellingsen	mari@econnect-ntnu.no
Herman Westrum Thorsen	herman@econnect-ntnu.no

Post- og besøksadresse:	*Organisasjonsnummer:*	*Hjemmeside:*
ECONnect, NTNU Dragvoll Institutt for samfunnsøkonomi Bygg 7, Nivå 5 7491 Trondheim	NO 994 625 314	www.econnect-ntnu.no

Merk: Alle pensumsammendrag og tekster som utgis av Faktor er skrevet av og for studenter. ECONnect står ikke ansvarlig for selve faginnholdet. Spørsmål om teksten kan rettes til tekstforfatteren.

Drago Berghølt

Pensumsammendrag: SØK3001 – ØKONOMETRI I

Enkel regresjon:

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad (1)$$

Modell: $y = a + bx$ der $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$ (2)

Forutsetninger:

SLR1: $y = \beta_0 + \beta_1 x + u$ (lineær i parametrene)

SLR2: Tilfeldig trekking fra populasjonen

SLR3: Variasjon i x $\rightarrow$ $Var(x) > 0$

SLR4: $E(u_i | x_i) = 0$ $\Rightarrow$ $E(u_i) = 0$ og $Cov(x_i, u_i) = 0$

SLR5: $Var(u_i | x_i) = \sigma^2$

OLS-metoden: (MKM)

Sann modell: $y_i = \beta_0 + \beta_1 x_i + u_i$

Estimert modell: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\Big\} \; y_i - \hat{y}_i = \hat{u}_i$

Vil minimere $\sum_{i=1}^{n} (\hat{u}_i)^2$:

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

$$\Rightarrow \frac{\partial MK}{\partial \beta_0} = \sum_{i=1}^{n} 2 (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot (-1) = 0 \quad (3)$$

$$\Rightarrow \frac{\partial MK}{\partial \beta_1} = \sum_{i=1}^{n} 2 (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot (-x_i) = 0 \quad (4)$$

(1) & (2) kalles normalligningene

Av normalligning (1):

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

$$\Rightarrow \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n} n\cdot\hat{\beta}_0 - \frac{1}{n}\hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \quad \Leftrightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (5)$$

Setter (5) inn i minimeringsproblemet:

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \sum_{i=1}^{n}\left(y_i - \left(\bar{y} - \hat{\beta}_1\bar{x}\right) - \hat{\beta}_1 x_i\right)^2$$

$$= \sum^{n}\left[\left(y_i - \bar{y}\right) - \hat{\beta}_1\left(x_i - \bar{x}\right)\right]^2$$

$\Rightarrow$ Minimerer mhp. $\hat{\beta}_1$:

$$\sum_{i=1}^{n} 2\left(y_i - \bar{y} - \hat{\beta}_1\left(x_i - \bar{x}\right)\right)\cdot(-1)\left(x_i - \bar{x}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) - \hat{\beta}_1 \sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right) = 0$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \qquad (6)$$

(5) & (6) kalles OLS-estimatorene:

(5) $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

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

Forventningsrett & konsistent estimator for β_1:

- Forventningsrett:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (6)$$

Har at siden

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad (7)$$

må også

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u} \quad (8)$$

Dette gir

$$(y_i-\bar{y}) = \beta_1(x_i-\bar{x}) + (u_i-\bar{u}) \quad (9)$$

Setter (9) inn i (6):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left[\beta_1(x_i-\bar{x})+(u_i-\bar{u})\right](x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}\beta_1(x_i-\bar{x})^2 + \sum_{i=1}^{n}(u_i-\bar{u})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \beta_1 \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \frac{\sum_{i=1}^{n}(u_i-\bar{u})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \beta_1 + \frac{\sum_{i=1}^{n}(u_i-\bar{u})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^{n}u_i(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (10)$$

Tar forventningsverdien på begge sider og får:

$$E(\hat{\beta}_1) = \beta_1 + \frac{\sum_{i=1}^{n}E(u_i|x_i)(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \beta_1 \quad (11)$$

- Konsistens:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(u_i-\bar{u})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (10)$$

$$= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(u_i-\bar{u})(x_i-\bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (12)$$

Tar sannsynlighetsgrensen på begge sider og får:

$$\operatorname*{plim}_{n\to\infty}(\hat{\beta}_1) = \operatorname*{plim}_{n\to\infty}\left[\beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(u_i-\bar{u})(x_i-\bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$$

$$= \beta_1 + \frac{\operatorname{plim}\left[\frac{1}{n}\sum_{i=1}^{n}(u_i-\bar{u})(x_i-\bar{x})\right]}{\operatorname{plim}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]}$$

$$= \beta_1 + \frac{Cov(u,x)}{Var(x)} = \beta_1 \quad (13)$$

der vi har brukt at $Cov(u,x)=0$. Generelt antar vi at de empiriske momentene konvergerer i sannsynlighet mot sine respektive teoretiske momenter.

Alternativt kan vi ta utgangspunkt i (6):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (6)$$

Sannsynlighetsgrensen gir:

$$\operatorname*{plim}_{n\to\infty}(\hat{\beta}_1) = \frac{\operatorname{plim}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})\right]}{\operatorname{plim}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]} \quad (14)$$

$$= \frac{Cov(y,x)}{Var(x)} = \beta_1 \quad (15)$$

Dette alternativet viser til momentmetoden

Momentmetoden:

- Bruker at $E(u_i|x_i)=0$ og $Cov(u_i|x_i)=0$.
- Tar forventningsverdien på begge sider av (1):

$$E(y_i) = \beta_0 + \beta_1 E(x_i) + E(u_i) \qquad (16)$$

- Trekker (14) fra (1) og får:

$$y_i - E(y_i) = \beta_1 (x_i - E(x_i)) + (u_i - E(u_i))$$

- Multipliserer gjennom med $x_i - E(x_i)$ og tar forventningen på begge sider:

$$(x_i - E(x_i))(y_i - E(y_i)) = \beta_1 (x_i - E(x_i))^2 + (u_i - E(u_i))(x_i - E(x_i))$$

$$\Rightarrow E[(x_i - E(x_i))(y_i - E(y_i))] = \beta_1 E[(x_i - E(x_i))^2] + E[(u_i - E(u_i))(x_i - E(x_i))]$$

$$\Rightarrow Cov(y,x) = \beta_1 Var(x) + Cov(u,x)$$

$$= \beta_1 Var(x)$$

$$\Rightarrow \beta_1 = \frac{Cov(y,x)}{Var(x)} \qquad (15)$$

- Forholdet mellom teoretiske & empiriske momenter:

$\hat{\beta}_0$: (16) gir (5):

$$E(y_i) = \beta_0 + \beta_1 E(x_i) + E(u_i) \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\hat{\beta}_1$: (15) gir (14):

$$\beta_1 = \frac{Cov(y,x)}{Var(x)} \Rightarrow \hat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Dekomponering av variasjonen i y – forklaringsmål:

- Normalligningene (3) & (4) gir at:

$$\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)=\sum_{i=1}^{n}\left(\hat{u}_i\right)=0 \tag{17}$$

$$\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)x_i=\sum_{i=1}^{n}\left(\hat{u}_i x_i\right)=0 \tag{18}$$

- Definerer sum of squares:

$$SST \equiv \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 \tag{19}$$

$$SSE \equiv \sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2 \tag{20}$$

$$SSR \equiv \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{n}\left(\hat{u}_i\right)^2 \tag{21}$$

- $SST = SSE + SSR$:

$$\begin{aligned}
SST &= \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 \\
&= \sum_{i=1}^{n}\left[\left(y_i-\hat{y}_i\right)+\left(\hat{y}_i-\bar{y}\right)\right]^2 \\
&= \sum_{i=1}^{n}\left[\left(y_i-\hat{y}_i\right)^2+2\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\bar{y}\right)+\left(\hat{y}-\bar{y}\right)^2\right] \\
&= \sum_{i=1}^{n}\left(y_i-\hat{y}\right)^2+2\sum_{i=1}^{n}\hat{u}_i\left(\hat{y}_i-\bar{y}\right)+\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2 \\
&= SSR+2\left(\sum_{i=1}^{n}\hat{u}_i\hat{y}_i-\bar{y}\sum_{i=1}^{n}\hat{u}_i\right)+SSE \\
&= SSE+SSR+2\left(\sum_{i=1}^{n}\hat{u}_i\hat{\beta}_0+\sum_{i=1}^{n}\hat{u}_i\hat{\beta}_1 x_i-\bar{y}\sum_{i=1}^{n}\hat{u}_i\right) \\
&= SSE+SSR+2\left(\hat{\beta}_0\sum_{i=1}^{n}\hat{u}_i+\hat{\beta}_1\sum_{i=1}^{n}\hat{u}_i x_i-\bar{y}\sum_{i=1}^{n}\hat{u}_i\right)
\end{aligned}$$

(17) & (18) gir at det siste leddet er null:

$$SST = SSE + SSR \tag{22}$$

Den multipple determinasjonskoeffisienten R^2:

- Definisjon:

$$R^2 \equiv \frac{SSE}{SST} \tag{23}$$

$$= \frac{SST - SSR}{SST}$$

$$= 1 - \frac{SSR}{SST} \tag{24}$$

der $0 \leq R^2 \leq 1$

- Justert R^2: $\bar{R}^2$

$$\bar{R}^2 = 1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n-1}} \Rightarrow \text{„Straffer" parameterrike likninger}$$

Variansen til $\hat{\beta}_1$:

- Tar utgangspunkt i (10):

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} u_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^{n} u_i d_i}{SST_x} \tag{25}$$

der $d_i \equiv (x_i - \bar{x})$ & $SST_x \equiv \sum_{i=1}^{n} (x_i - \bar{x})^2$

Utnytter at $Var(ax) = a^2 Var(x)$ og får:

$$Var(\hat{\beta}_1 | x) = 0 + \frac{\sum_{i=1}^{n} Var(u_i | x_i) \cdot d_i^2}{SST_x^2} = \frac{\sum_{i=1}^{n} \sigma_i^2 d_i^2}{SST_x^2} \tag{26}$$

Hvis homoskedastisitet, dvs hvis $\sigma_i = \sigma_j = \sigma$:

$$Var(\hat{\beta}_1 | x) = \frac{\sigma^2 \sum_{i=1}^{n} d_i^2}{SST_x^2} = \frac{\sigma^2 \cdot SST_x}{SST_x^2} = \frac{\sigma^2}{SST_x} \tag{27}$$

Estimering av varians & standardavvik

- Variansen til u_i: $\text{Var}(u_i|x_i) = \sigma^2$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{u}_i)^2 = \frac{SSR}{n-2} \quad (28)$$

- Standardavviket til u_i:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{SSR}{n-2}} = \frac{\sqrt{SSR}}{\sqrt{n-2}} \quad (29)$$

- Variansen til $\hat{\beta}_1$:

Erstatter den teoretiske variansen til u i (27) med den empiriske:

$$\widehat{\text{Var}}(\hat{\beta}_1|x) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (30)$$

- Standardavviket til $\hat{\beta}_1$:

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \quad (31)$$

Multippel regresjon:

- Modell: $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$ (1)

  der $\bar{y} = \beta_0 + \beta_1 \bar{x}_1 + \ldots + \beta_k \bar{x}_k + \bar{u}$ (2)

- Forutsetninger:

  MLR 1: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ innebærer at $\text{cov}(u_i, u_j | x) = 0$

  MLR 2: Tilfeldig trekking fra populasjonen

  MLR 3: Variasjon i X & ikke perfekt kollinearitet mellom navn. variabler

  MLR 4: $E(u_i | x_{i1}, \ldots, x_{ik}) = 0$ → $E(u_i) = 0$, $\text{Cov}(u_i, x_{ij}) = 0$

  MLR 5: $\text{Var}(u_i | x_{i1}, \ldots, x_{ik}) = \sigma^2$

- forventning:

  MLR 4 innebærer at $E(\hat{\beta}_j | X) = \beta_j$

- Multippel determinasjonskoeffisient:

  Har tidligere vist at

  $$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

  Justert R^2: $\bar{R}^2$

  $$\bar{R}^2 = 1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n+1}} \quad (3)$$

  (3) "straffer" parameterrike likninger

- Variansen til $\hat{\beta}_j$:

$$\text{Var}(\hat{\beta}_j | X) = \frac{\sigma^2}{SST_j (1 - R_j^2)} \quad (4)$$

der $SST_j \equiv \sum_{i=1}^{n} (x_{ij} - \overline{x_j})^2$

og R_j^2 er R^2 fra

$$x_{ij} = \alpha_0 + \alpha_1 x_{i1} + \ldots + \alpha_{j-1} x_{j-1} + \alpha_{j+1} x_{j+1} + \ldots + \alpha_{ik} x_k + u_i^j$$

- Estimering av varians & standardavvik:

Erstatter teoretiske moment med empiriske:

$$\widehat{\text{Var}}(\hat{\beta}_1 | X) = \frac{\hat{\sigma}^2}{SST(1 - R_j^2)}$$

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

Standardavvik:

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST(1 - R_j^2)}}$$

Hypotesetesting / inferens:

- Generell testing:

  H_0: Påstand A er sann

  H_1: Påstand A er ikke sann / H_0 er ikke sann

  Hypotesene H_0 & H_1 er gjensidig utelukkende og fyller tilsammen hele utfallsrommet.

- t-testen: Testing av enkelt parameter

  t-testens testobservator $t_{\hat{\beta}_1}$ er gitt ved:

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1^0}{\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} = \frac{\hat{\beta}_1 - \beta_1^0}{se(\hat{\beta}_1)} \sim t_{n-k-1} \qquad (1)$$

  Behold H_0 hvis $|t_{\hat{\beta}_1}| < t_c$,

  forkast H_0 hvis $|t_{\hat{\beta}_1}| \geq t_c$

- F-testen: Testing av flere parametre:

  SSR vil alltid øke når en modell pålegges restriksjoner. Ønsker imidlertid å teste om økningen i SSR er så stor at vi kan avvise hypotesen om at restriksjonen er sann. Altså; er SSR_r tilstrekkelig større enn SSR_{ur}?

Testobservatoren F_{obs} er gitt ved:

$$F_{obs} \equiv \frac{\frac{SSR_r - SSR_{ur}}{q}}{\frac{SSR_{ur}}{n-k-1}} \sim F_{q,\,n-k-1} \qquad (2)$$

Behold H_0 hvis $F_{obs} < F_c$,

forkast H_0 hvis $F_{obs} \geq F_c$

Alternativt:

$$R^2 = 1 - \frac{SSR}{SST}$$

$$\Rightarrow SSR = SST(1 - R^2)$$

$$\Rightarrow F_{obs} = \frac{\frac{SSR_r - SSR_{ur}}{q}}{\frac{SSR_{ur}}{n-k-1}}$$

$$= \frac{\frac{SST(1-R_r^2) - SST(1-R_{ur}^2)}{q}}{\frac{SST(1-R_{ur}^2)}{n-k-1}}$$

$$= \frac{\frac{1 - R_r^2 - 1 + R_{ur}^2}{q}}{\frac{1-R_{ur}^2}{n-k-1}}$$

$$= \frac{\frac{R_{ur}^2 - R_r^2}{q}}{\frac{1-R_{ur}^2}{n-k-1}} \qquad (3)$$

- Test av helhetlig signifikans:

H_0: Ingen variabler bidrar til forklaring.

$$\Rightarrow R_r^2 = 0 \quad \text{og} \quad q = k$$

$$\Rightarrow F_{obs} = \frac{\frac{R^2}{k}}{\frac{1-R^2}{}} \qquad (4)$$

- Spesialtilfelle: Bruk av t-test

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

$$H_0: \beta_1 = \beta_2 \rightarrow \beta_1 - \beta_2 = 0$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_2 (x_{i1} - x_{i1}) + u_i$$

$$= \beta_0 + (\beta_1 - \beta_2) x_{i1} + \beta_2 (x_{i1} + x_{i2}) + u_i$$

$$= \beta_0 + \theta x_{i1} + \beta_2 z_i + u_i$$

der $\theta \equiv \beta_1 - \beta_2$ & $z_i \equiv x_{i1} + x_{i2}$

$$H_0: \theta = 0$$

- Konfidensintervall:

1. Konfidensintervall for β_1 når σ er kjent:

Siden $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{n})$ vil den normaliserte variabelen være fordelt slik at

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Har da at et 95% konfidensintervall gis ved:

$$P\left(-1{,}96 \leq \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{n}}} \leq 1{,}96\right) = 0{,}95$$

$$\Rightarrow \hat{\beta}_1 \pm 1{,}96 \cdot \frac{\sigma}{\sqrt{n}}$$

2. Konfidensintervall for $\hat{\beta}_1$, når σ er ukjent:

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

Siden $\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

er (når vi setter inn det empiriske mom.):

$\frac{\hat{\beta}_1 - \beta_1}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t_{n-1}$

$\Rightarrow$ 95% konfidensintervall gitt ved:

$\hat{\beta}_1 \pm 1{,}96 \frac{\hat{\sigma}}{\sqrt{n}}$

- Konfidensintervall i regresjonslikningen:

Siden $\frac{\hat{\beta}_j - \beta_j}{Se(\hat{\beta}_j)} \sim t_{n-k-1}$

er konfidensintervallet til den ukjente β_j gitt ved:

$$\hat{\beta}_j \pm t_c \cdot Se(\hat{\beta}_j) \qquad (J)$$

der t_c er den $(1 - \frac{1}{2}p)$ persentilen i en t_{n-k-1}-fordeling (p er p-verdi). f.eks. $p = 0{,}05$ gir $(1 - \frac{1}{2}p) = 1 - \frac{0{,}05}{2} = 0{,}975$ persentil.

Eksempel: $\hat{y} = -4{,}38 + \underset{(0{,}47)}{1{,}084}\, x_1 + \underset{(0{,}060)}{0{,}0217}\, x_2$ $\quad n = 32,\ R^2 = 0{,}918$

95%

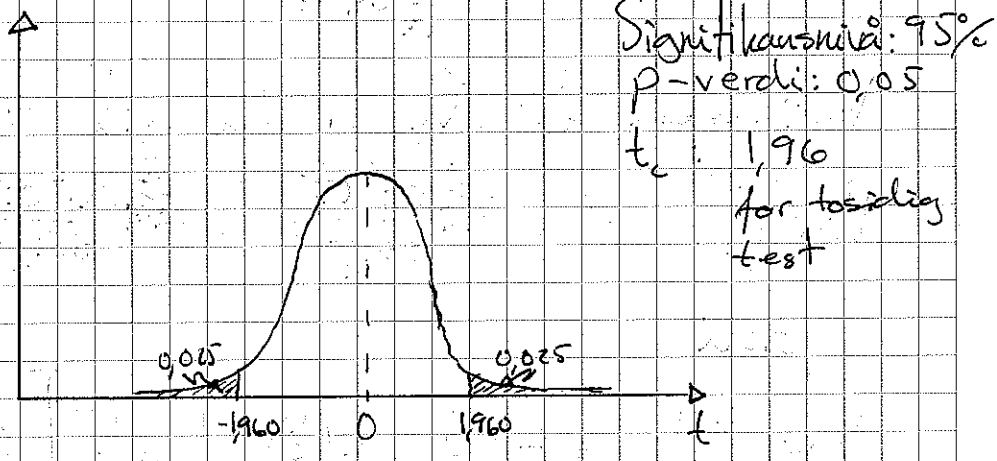
Konf.int.v.: $1{,}084 \pm 2{,}045 \cdot 0{,}060 \Rightarrow [0{,}961,\ 1{,}21]$

- p-verdi: Det minste signifikansnivået som en nullhypotese kan forkastes ved, gitt observert t-verdi.

Alternativ formulering:

Sannsynligheten for å observere t_{obs} gitt at nullhypotesen er sann.

- t-fordelingen:



forkaster H_0 hvis $|t_{obs}| \geq 1,96$

Heteroskedastisitet:

Hartidligere antatt at $Var(u_i|X) = \sigma^2$, altså homoskedastisitet.

Generelt innebærer heteroskedastisitet at $Var(u_i|X) = \sigma_i^2$, altså ikke konstant varians i restledd.

Tar utgangspunkt i:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(u_i-\bar{u})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \beta_1 + \frac{\sum_{i=1}^{n} u_i d_i}{SST_x} \quad \text{der } d_i = x_i - \bar{x} \text{ og } SST_x = \sum_{i=1}^{n}(x_i-\bar{x})^2$$

Variansen til $\hat{\beta}_1$ er da gitt ved:

$$Var(\hat{\beta}_1) = 0 + \frac{\sum_{i=1}^{n} d_i^2 \sigma_i^2}{SST_x^2} = \frac{\sum_{i=1}^{n} d_i^2 \sigma_i^2}{SST_x^2}$$

Fordi

$$Var(\hat{\beta}_1|X) = E\left[\left(\hat{\beta}_1 - E(\hat{\beta}_1)\right)^2\right] = E\left[\left(\beta_1 + \frac{\sum_{i=1}^{n} u_i d_i}{SST_x} - \beta_1\right)^2\right]$$

$$= E\left[\left(\frac{\sum_{i=1}^{n} u_i d_i}{SST_x}\right)^2\right] = E\left[\frac{\left(\sum_{i=1}^{n} u_i d_i\right)^2}{SST_x^2}\right]$$

$$= \frac{\sum_{i=1}^{n} Var(u_i|X) \cdot d_i^2}{SST_x^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\sigma_i^2}{SST_x^2}$$

Så lenge ikke $\sigma_i^2 = \sigma_j^2 = \sigma^2$ har vi heteroskedastisitet.

- Konsekvenser av heteroskedastisitet:
  Ved heteroskedastisitet gjelder ikke de vanlige MKM-formlene for varians & standardavvik. Ei heller t-testen eller F-testen eller LM-testen. Alle disse baseres nemlig på konstant varians i restleddet.
  Heteroskedastisitet gir imidlertid ikke skjevhet eller inkonsistens i OLS-estimatene, ei heller i goodness-of-fit-målene R^2 & $\bar{R}^2$.
- Kan løse problemer knyttet til heteroskedastisitet vha korrigering eller hensyn til heteroskedastisiteten.

- Estimering av heteroskedastisitetskorrigerte eller robuste standardavvik:

Benytter metoden til Kenneth White:

$$\widehat{Var}(\hat{\beta}_1 | X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma_i^2}{SST_x^2}$$

⇒ Setter inn for $\hat{u}_i^2$ istedet for σ_i^2:

$$\widehat{Var}(\hat{\beta}_1 | X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \hat{u}_i^2}{SST_x^2}$$ (i bivariate analyser)

⇒ Finner deretter standardavviket:

$$\widehat{se}(\hat{\beta}_1 | X) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \hat{u}_i^2}{SST_x^2}}$$

$$= \frac{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \hat{u}_i^2}}{SST_x}$$

Fremgangsmåten er altså å først estimere parametrene vha OLS, & deretter bruke residualene til å estimere heteroskedastisitetskorrigerte standardavvik. Dermed kan man ta i bruk inferens.

Ved multippel regresjon:

$$\widehat{Var}(\hat{\beta}_j | X) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}$$

$$\Rightarrow \widehat{se}(\hat{\beta}_1 | X) = \frac{\sqrt{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}}{SSR_j^2}$$

der $\hat{r}_{ij}$ er i'ende residual i regresjonen med x_j som avh. variabel og alle andre x som uavhengige, og SSR_j er sum of squared residuals fra denne regresjonen.

- Testing for heteroskedastisitet:

Kan ta utgangspunkt i:

$$y_i = \beta_0 + \sum_{j=1}^{k_1} \beta_j x_{ij} + u_i$$

H_0: $\mathrm{Var}(u|X) = \sigma^2$ eller $E(u^2|X) = \sigma^2$

H_1: H_0 er ikke sann.

H_1 kan spesifiseres på mange måter, f.ks slik: $u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_1 x_2 + \text{støy}$

Nullhypotesen $\delta_1 = \delta_2 = \delta_3 = 0$ kan testes vha. F-testen eller LM-testen.

Slik testing for heteroskedastisitet kalles Breusch-Pagan-testing for heteroskedastisitet (test av hvilketlig sign. der $F_{obs} \equiv \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}}$)

Altså:

Estimerer $\hat{u}_i^2 = \hat{\delta}_0 + \hat{\delta}_1 x_{i1} + \hat{\delta}_2 x_{i2} + \hat{\delta}_3 x_{i1} x_{i2}$

og tester H_0: $\delta_1 = \delta_2 = \delta_3 = 0$

White-Test: $\hat{u}^2 = \delta_0 + \delta_1 y + \delta_2 y^2$

Testing for heteroskedastisitet kan gjøres ved å teste forskjellige funksjonsformer der u^2 er på venstresiden & x_j er på høyresiden.

- Vektet OLS (WLS):

Er en transformasjon slik at restleddet får konstant varians.

Antar at $y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + u_i$ der

Variansen er gitt ved: $\text{Var}(u_i|X) = \sigma^2 \cdot h_i(X)$

Dividerer nå gjennom strukturligningen med $\frac{1}{\sqrt{h_i}}$ dette gir:

$$\frac{y_i}{\sqrt{h_i}} = \frac{\beta_0}{\sqrt{h_i}} + \sum_{j=1}^{k} \frac{\beta_j x_{ij}}{\sqrt{h_i}} + \frac{u_i}{\sqrt{h_i}}$$

Variansen blir da

$$\text{Var}\left(\frac{u_i}{\sqrt{h_i}}\middle|X\right) = \left(\frac{1}{\sqrt{h_i}}\right)^2 \cdot \text{Var}(u_i|X) = \frac{1}{h_i} \cdot \sigma^2 \cdot h_i = \sigma^2$$

Her har altså restleddet konstant varians.

Ved transformasjon av strukturligningen får vi konstant varians i restleddet, og vi kan da benytte oss av vanlige inferensmetoder.

Metoden der vi estimerer h_i ($\hat{h}_i$) kalles Feasable GLS.

Funksjonsform:

- logaritmer:

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

$\Rightarrow \frac{1}{y} \cdot dy = \beta_1 \cdot dx$

$\Rightarrow$ Marginal endring i x gir relativ endring i y (målt i prosent).

Ved betydelig effekt på dy av dx er følgende et mer presist estimat:

$\%\Delta y = 100 \cdot (e^{\hat{\beta}_1 dx_1} - 1)$

Eksempel:

$\widehat{\ln y} = 9,23 + 0,306 x$

$\Rightarrow \%\Delta y = 100 \cdot (e^{0,306 \cdot 1} - 1) = 35,8\%$

$\Rightarrow$ hadde bare vært 30,6% hvis estimert direkte.

$y = \beta_0 + \ln x$

$\Rightarrow dy = \beta_1 \frac{1}{x} \cdot dx$

$\Rightarrow$ Prosentvis endring i x gir endring i y.

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

$\rightarrow \frac{1}{y} \cdot dy = \beta_1 \frac{1}{x} dx$

$\Rightarrow$ Prosentvis endring i x gir prosentvis endring i y. Mål på elastisitetsform.

- Dummy:

$y = \beta_0 + \beta_1 x_1$ der $x_1 = 1$ hvis event inntreffer

og $x_1 = 0$ hvis event ikke inntreffer.

Forskjellige skjæringspunkt & stigningstall:

$y = \beta_0 + \beta_1 D_{MALE} + \beta_2 X + \beta_3 D_{MALE} \cdot X$

Skjæringspunkt:

Menn: $\beta_0 + \beta_1$

Kvinner: β_0

Stigningstall:

Menn: $\beta_2 + \beta_3$

Kvinner: β_2

Dummyer kan også brukes til å skille/kategorisere variabler på nominalnivå, f.kx kommuner.

- Sammenligning av to grupper: Chows test

Generell modell: $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$

Ønsker å teste om fks. kvinner og menn har forskjellige skjæringspunkt & stigningstall.

Problem: Kan være vanskelig å oppnå signifikans når mange variabler er involvert. Vha. dummyer

Løsning:

Beregne modellen kun for kvinner, finn SSR_1.

Beregne modellen kun for menn, finn SSR_2.

Beregne modellen for hele populasjonen, finn SSR_p.

Chows F-test:

$$F = \frac{SSR_p - (SSR_1 + SSR_2)}{\frac{k+1}{SSR_1 + SSR_2}} \Big/ \frac{}{n-2(k+1)}$$

H_0: Det er ingen forskjell i parametrene mellom kvinner & menn.

- Dummy i venstresidevariabelen: Linear probability model (LPM). Merk: Kan ta verdier >1 & <0.

- Annengradsledd:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i$$

$$\Rightarrow dy = \beta_1 dx + 2\beta_2 dx \cdot x$$

$$\Rightarrow \frac{dy}{dx} = \beta_1 + 2\beta_2 x_i$$

- Samspillsledd:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + u_i$$

$$\Rightarrow dy = \beta_1 dx_1 + \beta_3 x_2 \cdot dx_1$$

$$\Rightarrow \frac{dy}{dx_1} = \beta_1 + \beta_3 x_2$$

Valg mellom ikke-nestede modeller:

- Valg mellom nestede modeller, dvs. der én modell fremkommer som en annen modell med null restriksjoner, kan gjøres vha. en F-test eller LM-test.
- ikke-nestede modeller: Kan ikke skrives som spesialtilfeller av hverandre.

Eksempel: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$ (1)

$y_i = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + v_i$ (2)

- Merk: Hvis begge modellene har samme venstreside variabel, vil vi alltannet likt velge den med størst forklaringskraft ($\bar{R}^2$). Imidlertid mye annet enn $\bar{R}^2$ som spiller inn.
- Alternativ testprosedyre 1: Mizon & Richard

Slå sammen (1) & (2):

$$y_i = \alpha_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_1 z_{i1} + \gamma_2 z_{i2} + w_i$$

Tester deretter om: A: $\beta_1 = \beta_2 = 0$

B: $\gamma_1 = \gamma_2 = 0$

- Hvis kun én av hypotesene forkastes kan vi velge modell.
- Hvis begge forkastes er ingen gyldige forenklinger av en mer generell modell.
- Hvis ingen forkastes er testen inkonklusiv.

- Alternativ testprosedyre? Davidson & Mackinnon intuisjon: Hvis (1) er sann er de estimerte verdiene fra (2) insignifikante når de puttes inn i (1).

Fremgangsmåte:

1. Estimér (2) for å finne $\hat{y}_i$.
2. Estimer variansen $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \delta_1 \hat{y}_i + u_i$ av (1).
3. Gjennomfør en t-test av $\hat{\delta}_1$. Forkast (1) dersom H_0 forkastes.
4. Gjenta samme prosedyre med utg.p. i (2).

Merk: Også her kan testen være inkonklusiv.

Endogen forklaringsvariabel:

Kilder til endogenitet:

1. Utelatte variable
2. Målefeil
3. Simultanitet

Krav til eksogenitet i forklaringsvariabelen:

$Cov(x, u) = 0$

1. Utelatt variabelskjevhet:

Sann modell: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$

Estimert modell: $y_i = \beta_0 + \beta_1 x_{i1} + v_i$ der $v_i = \beta_2 x_{i2} + u_i$

MKM-estimatoren for β_1:

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_{i1} - \bar{x}_1)}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} \quad \text{der} \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

og $\bar{y} = \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \bar{u}$

$$= \frac{\sum_{i=1}^{n}\left[\beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + (u_i - \bar{u})\right](x_{i1} - \bar{x}_1)}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$

$$= \frac{\sum_{i=1}^{n}\beta_1(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n}\beta_2(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) + \sum_{i=1}^{n}(u_i - \bar{u})(x_{i1} - \bar{x}_1)}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$

$$= \beta_1 \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} + \beta_2 \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} + \frac{\sum_{i=1}^{n}(u_i - \bar{u})(x_{i1} - \bar{x}_1)}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$

$$= \beta_1 + \beta_2 \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} + \frac{\frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})(x_{i1} - \bar{x}_1)}{\frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$

$$\operatorname{plim}_{n\to\infty}(\tilde{\beta}_1) = \beta_1 + \beta_2 \cdot \frac{\operatorname{Cov}(x_1, x_2)}{\operatorname{Var}(x_1)} = \beta_1 + \beta_2 \delta_1$$

der $\delta_1 = \frac{\operatorname{cov}(x_1, x_2)}{\operatorname{Var}(x_1)}$

Positiv variabelskjevhet: $\tilde{\beta}_1 > \beta_1$

1.1. $\beta_2 > 0$ & $\delta_1 > 0$

Utelatt variabel påvirker avhengig variabel positivt og korrelerer samtidig positivt med inkludert variabel.

1.2. $\beta_2 < 0$ & $\delta_2 < 0$

Utelatt variabel påvirker avhengig variabel negativt & korrelerer samtidig negativt med inkludert variabel.

Negativ variabelskjevhet: $\tilde{\beta}_1 < \beta_1$

1.3. $\beta_2 < 0$ & $\delta_1 > 0$

Utelatt variabel påvirker avhengig variabel negativt & korrelerer samtidig positivt med inkludert variabel.

1.4. $\beta_2 > 0$ & $\delta_1 < 0$

Utelatt variabel påvirker avhengig variabel positivt & korrelerer samtidig negativt med inkludert variabel.

Ingen variabelskjevhet: $\tilde{\beta}_1 = \beta_1$

1.5 $\beta_2 = 0$

Utelatt variabel har ingen innvirkning på avhengig variabel.

1.6 $\delta_1 = 0$

Utelatt variabel korrelerer ikke med inkludert variabel.

Variansen til $\tilde{\beta}_1$:

Har at $\quad Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_1(1-R_1^2)}$

$\quad Var(\tilde{\beta}_1) = \frac{\sigma^2}{SST_1}$

Ser at så lenge $Cov(x_1, x_2) \neq 0$, er $Var(\tilde{\beta}_1) < Var(\hat{\beta}_1)$

Dette innebærer at dersom $\beta_2 = 0$, er det best å utelate x_2 fra modellen!

Intuisjon: Å inkludere irrelevante variable fører bare til høyere varians i $\hat{\beta}_1$ gitt multikollinearitetsproblemet.

2. Målefeil:

Antar at den riktige modellen er gitt ved:

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

Målefeil når de sanne y^* & x^* ikke observeres, men hhv:

$y = y^* + \varepsilon$

$x = x^* + e$

Presisering av forutsetninger:

$E(u) = E(\varepsilon) = E(e) = 0$

$Var(u) = \sigma_u^2$, $Var(\varepsilon) = \sigma_\varepsilon^2$, $Var(e) = \sigma_e^2$

$Cov(u,x) = Cov(u,\varepsilon) = Cov(u,e) = Cov(\varepsilon,e) = 0$

Tilfelle 1: Kun målefeil i venstresidevariabelen.

Estimerer altså:

$y_i = \beta_0 + \beta_1 x_i + (u_i + \varepsilon_i) = \beta_0 + \beta_1 x + v$ der $v = u + \varepsilon$

Hvorvidt $\hat{\beta}_1$ er forventningsrett eller ikke avhenger av $Cov(v,x)$, siden $\hat{\beta}_1 = \beta_1 + \frac{Cov(v,x)}{Var(x)}$

Har allerede antatt at $Cov(u,x)=0$. Så lenge $Cov(\varepsilon,x)=0$ er da også $Cov(v,x)=0$. Dette er normalt en rimelig antagelse.

Imidlertid er variansen til restleddet større:

$$Var(v) = Var(u+\varepsilon) = \sigma_u^2 + \sigma_\varepsilon^2 > \sigma_u^2$$

Dette innebærer høyere varians & standardavvik for OLS-estimatorene.

Tilfelle 2: Kun målefeil i forklaringsvariabelen, målefeilen er ukorrelert med observert verdi ($Cov(e,x)=0$)

Estimerer altså:

$y = \beta_0 + \beta_1(x-e) + u = \beta_0 + \beta_1 x + v$ der $v = u - \beta_1 e$

Siden vi har antatt at $Cov(u,x)=0$, & samtidig at $Cov(e,x)=0$, må så også $Cov(v,x)=0$. Dermed er $\tilde{\beta}_1 = \beta_1 + \frac{Cov(v,x)}{Var(x)} = \beta_1$

Variansen til restleddet er imidlertid større:

$$Var(v) = Var(u-\beta_1 e) = \sigma_u^2 + \beta_1^2 \sigma_e^2 > \sigma_u^2$$

Dermed høyere varians & st. avvik for OLS-estimatene.

Tilfelle 3: Kun målefeil i forklaringsvariabelen, målefeilen er ukorrelert med uobserverbar variabel ($Cov(e, x^*)=0$)

Estimerer altså:

$y = \beta_0 + \beta_1 (x - e) + u = \beta_0 + \beta_1 x + v$ der $v = u - \beta_1 e$

Siden $Cov(e, x^*) = 0$ er

$Cov(e, x) = Cov(e, x^* + e) = \sigma_e^2$

Dermed er

$Cov(v, x) = Cov(u - \beta_1 e, x^* + e) = -\beta_1 \sigma_e^2$

Her er mao. restleddet i den estimerte ligningen korrelert med den inkluderte forklaringsvariabelen.

Videre er:

$Var(x) = Var(x^* + e) = \sigma_{x^*}^2 + \sigma_e^2$

Dette gir:

$\tilde{\beta}_1 = \beta_1 + \frac{Cov(v, x)}{Var(x)} = \beta_1 + \frac{-\beta_1 \sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2} = \beta_1 \left(1 - \frac{\sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2}\right)$

Ser at: $\beta_1 > 0$ gir $\tilde{\beta}_1 < \beta_1$

$\beta_1 < 0$ gir $\tilde{\beta}_1 > \beta_1$

Altså systematiske underestimering av absoluttverdien til effekten av X.

Videre er:

$$\hat{\beta}_1 = \beta_1 \left(1 - \frac{\sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2}\right)$$

$$= \beta_1 \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2}\right) \quad \leftrightarrow \quad \beta_1 \frac{Var(x^*)}{Var(x)}$$

$$= \beta_1 \left(\frac{1}{1 + \frac{\sigma_e^2}{\sigma_{x^*}^2}}\right)$$

Ser at desto større variansen til målefeilen er relativt til variansen til den sanne forklaringsvariabelen, desto mer alvorlig er målefeilen.

Noise to signal ratio: $\frac{\sigma_e^2}{\sigma_{x^*}^2}$

3. Simultanitet:

Grunnleggende illustrasjon av simultanitetsproblemet:

$$y = \beta_1 x + \alpha_1 z_1 + u_1 \quad (1)$$

men også

$$x = \beta_2 y + \alpha_2 z_2 + u_2 \quad (2)$$

Altså:

$x \Rightarrow y$

men også

$y \Rightarrow x$

Mao. determineres y & x simultant av z_1 & z_2.

Estimering på redusert form ved å putte (2) inn i (1):

$$y = \beta_1(\beta_2 y + \alpha_2 z_2 + u_2) + \alpha_1 z_1 + u_1$$

$$\Rightarrow (1 - \beta_1\beta_2) y = \beta_1\alpha_2 z_2 + \alpha_1 z_1 + (u_1 + \beta_1 u_2)$$

$$\Rightarrow y = \frac{\alpha_1}{1-\beta_1\beta_2} z_1 + \frac{\beta_1\alpha_2}{1-\beta_1\beta_2} z_2 + \frac{u_1 + \beta_1 u_2}{1-\beta_1\beta_2}$$

$$= \Pi_1 z_1 + \Pi_2 z_2 + V \quad (3)$$

der $\Pi_1 = \frac{\alpha_1}{1-\beta_1\beta_2}$, $\Pi_2 = \frac{\beta_1\alpha_2}{1-\beta_1\beta_2}$ & $v = \frac{u_1+\beta_1 u_2}{1-\beta_1\beta_2}$

Estimering av (3) vha OLS vil i prinsippet gi forventningsrette & konsistente estimatorer for Π_1 & Π_2.

Problemet er imidlertid å identifisere β_1!

Generelt vil OLS-estimering av (1) gi skjeve estimater av både β_1 & α_1.

Hvis man ønsker å finne kausaleffekten av x på y:

1. Estimerer redusert form-regresjonen

$$x = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 \qquad (4)$$

der z_2 betraktes som mulig instrument.

2. Gitt at γ_2 er signifikant estimeres (1) der x byttes ut med $\hat{x}$ fra (4).

Kausaleffekten kan altså estimeres vha IV-2SLS så lenge (2) inneholder minst én eksogen variabel som er ekskludert i (1).

Mulige løsninger ved endogenitet i forklaringsvariabelen:

1. Overse problemet.
2. Proxy & OLS
3. Paneldata & fixed effects eller first difference
4. Instrumentvariabelmetoden

Tidsseriedata:

Modell: $y_t = \beta_0 + \beta_1 x_t + u_t$

Forutsetninger:

TS 1: $y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t$

TS 2: Variasjon i X & ikke perfekt kollinearitet mellom uavhengige variabler.

TS 3: $E(u_t | X) = 0$

TS 4: $\text{Var}(u_t | X) = \sigma^2$

TS 5: $\text{Corr}(u_t, u_s | X) = 0$ for alle $t \neq s$

Altså ingen serielkorrelasjon

TS 6: Restleddene er normalfordelt;

$N \sim (0, \sigma^2)$.

Under antagelsene TS1 - TS3 er OLS-estimatorene forventningsrette:

$E(\hat{\beta}_j) = \beta_j$

Finite distributed lag (FDL) model av orden q:

$$y_t = \alpha_0 + \beta_0 x_t + \beta_1 x_{t-1} + \dots + \beta_q x_{t-q} + u_t$$

Kortidseffekten (umiddelbar effekt): β_0

Langtidseffekten: $\beta_0 + \beta_1 + \dots + \beta_q$

(Merk at det her er snakk om varig endring i x).

Varians & standardavvik i $\hat{\beta}_j$:

Gitt TS1 - TS5 er:

$$\mathrm{Var}(\hat{\beta}_j | X) = \frac{\sigma^2}{SST_j(1-R_j^2)}$$

der $SST_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$

og R_j^2 er R^2 fra

$x_{ij} = \alpha_0 + \alpha_1 x_{t1} + \dots$

Estimatet er:

$$\widehat{\mathrm{Var}}(\hat{\beta}_j | X) = \frac{\hat{\sigma}^2}{SST_j(1-R_j^2)}$$

der $\hat{\sigma}^2 = \frac{SSR}{n-k-1}$

Standardavviket estimeres følgelig som:

$$\mathrm{Se}(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SST_j(1-R_j^2)}}$$

Autoregressiv modell: Koyck-modellen

Utgangspunkt i en infinite distributed lag (IDL) modell:

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \ldots + u_t \quad (1)$$

Koyck: Antar geometrisk avtagende vekter på parametrene:

$\beta_j = \gamma \cdot \rho^j$ slik at $\beta_0 = \gamma$, $\beta_1 = \gamma \cdot \rho$, osv. (2)

der $0 < \rho < 1$

Poeng: Får nå bare to parametre å estimere:

$$y_t = \alpha + \gamma x_t + \gamma\rho x_{t-1} + \gamma\rho^2 x_{t-2} + \ldots + u_t \quad (3)$$

Tilbakedaterer alle variabler i (3) med 1:

$$y_{t-1} = \alpha + \gamma x_{t-1} + \gamma\rho x_{t-2} + \gamma\rho^2 x_{t-3} + \ldots + u_{t-1} \quad (4)$$

Multipliserer (4) med ρ:

$$\rho y_{t-1} = \rho\alpha + \gamma\rho x_{t-1} + \gamma\rho^2 x_{t-2} + \gamma\rho^3 x_{t-3} + \ldots + \rho u_{t-1} \quad (5)$$

Trekker nå (5) fra (3):

$$y_t - \rho y_{t-1} = (1-\rho)\alpha + \gamma x_t + (u_t - \rho u_{t-1})$$

$$\Rightarrow y_t = \alpha_0 + \rho y_{t-1} + \gamma x_t + v_t \quad (7)$$

der $\alpha_0 = (1-\rho)\alpha$ & $v = u_t - \rho u_{t-1}$

(7) kalles en autoregressiv modell.
Denne kan i prinsippet estimeres vha OLS, bortsett fra at v_t bryter med prinsippet om ingen seriekorrelasjon. Skal foreløpig se bort fra dette.

Kortidseffekt & langtidseffekt i en autoregressiv modell:

Modellen: $y_t = \beta_0 + \beta_1 x_t + \rho y_{t-1} + u_t$

Kortidseffekt: $dy_t = \beta_1 \cdot dx_t \Rightarrow \frac{dy_t}{dx_t} = \beta_1$ (8)

Langtidseffekt: $d\bar{y} = \frac{\beta_1}{1-\rho} d\bar{x} \Rightarrow \frac{d\bar{y}}{d\bar{x}} = \frac{\beta_1}{1-\rho}$ (9)

siden $y_t = y_{t-1} = \bar{y}$ & $x_t = \bar{x}$

Merk: Når $0 < \rho < 1$ er prosessen stabil og langtidseffekten av en varig økning i x er $\frac{\beta_1}{1-\rho}$

Desto nærmere ρ er 1, desto tregere skjer tilpasningen.

Autoregressiv Distributed Lag (ADL) modell:

Utvider den autoregressive modellen med tilhørende daterte verdier på forklaringsvariablene:

$$y_t = \alpha + \rho y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t \quad (10)$$

Kortidseffekten: β_0

Langtidseffekten: $\frac{\beta_0 + \beta_1}{1-\rho}$

Fordi:

$$y_t = y_{t-1} = \bar{y} \quad \& \quad x_t = x_{t-1} = \bar{x}$$

$$\Rightarrow \bar{y} = \alpha + \rho\bar{y} + (\beta_0 + \beta_1)\bar{x}$$

$$\Rightarrow \bar{y} = \frac{\alpha}{1-\rho} + \frac{\beta_0 + \beta_1}{1-\rho}\bar{x} = \mu + \beta\bar{x} \quad (11)$$

$$\Rightarrow \frac{d\bar{y}}{d\bar{x}} = \frac{\beta_0 + \beta_1}{1-\rho} = \beta \quad (12)$$

Trend:

Trendproblem hvis følgende ligninger er signif.:

$$y_t = \beta_0 + \beta_1 t + u_t$$

$$x_t = \alpha_0 + \alpha_1 t + u_t$$

Får da spurious sammenheng i $y_t = \gamma_0 + \gamma_1 x_t + u_t$

Kan forstås som et utelatt variabelproblem.

Kan løses ved å inkludere tidstrenden:

$$y_t = \gamma_0 + \gamma_1 x_t + \gamma_2 t + u_t \quad (13)$$

Alternative formuleringer av trend:

$$\ln y_t = \gamma_0 + \gamma_1 x_t + u_t$$

$$y_t = \gamma_0 + \gamma_1 x_t + \delta_0 D_t + \delta_0 D_{t-1} \ldots + u_t$$

Error correction model (likevekts- eller feiljusteringsmodell):

Antar følgende modell: (10):

$$y_t = \alpha + \rho y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + u_t \quad (10)$$

$$\Rightarrow y_t - y_{t-1} = (\rho - 1) y_{t-1} + \alpha + \beta_0 x_t + \beta_1 x_{t-1} + u_t$$

$$= (\rho - 1) y_{t-1} + \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \beta_0 (x_{t-1} - x_{t-1})$$

$$= (\rho - 1) y_{t-1} + \alpha + \beta_0 (x_t - x_{t-1}) + (\beta_0 + \beta_1) x_{t-1} + u_t$$

$$= (\rho - 1) \left[y_{t-1} - \frac{\alpha}{1-\rho} - \frac{\beta_0 + \beta_1}{1-\rho} x_{t-1} \right] + \beta_0 (x_t - x_{t-1}) + u_t$$

$$\Rightarrow \Delta y_t = -\lambda (y_{t-1} - \mu - \beta x_{t-1}) + \beta_0 \Delta x_t + u_t \quad (14)$$

der $\lambda = 1 - \rho$, $\Delta y_t = y_t - y_{t-1}$,

$\mu = \frac{\alpha}{1-\rho}$, $\Delta x_t = x_t - x_{t-1}$

$\beta = \frac{\beta_0 + \beta_1}{1-\rho}$,

i (14) kan $\mu + \beta x_{t-1}$ tolkes som likevektsverdien på y på tidspunkt t-1 eller gitt x_{t-1}. Denne er også gitt av (11). Avviket i $(y_{t-1} - \mu - \beta x_{t-1})$ kan slik sett tolkes som avvik fra likevekten på tidspunkt t-1.

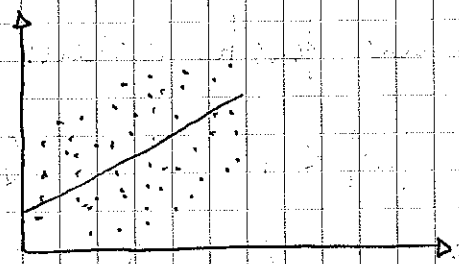
Dersom positivt avvik bidrar $0 < \lambda < 1$ til at $\Delta y_t < 0$.

Dersom negativt avvik bidrar $-1 < \lambda < 0$ til at $\Delta y_t > 0$.

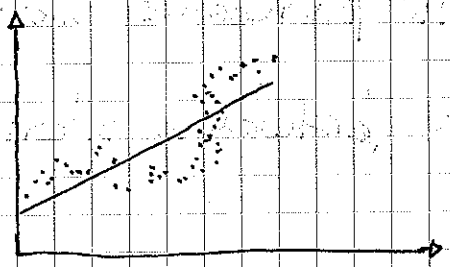
Generelt uttrykker $-\lambda (y_{t-1} - \mu - \beta_0 x_{t-1})$ en mekanisme i modellen som medfører at y utvikler seg i retning av den langsiktige likevektsbanen over tid.

Seriekorrelasjon:

Fravær av seriekorrelasjon:



Tilstedeværelse av seriekorrelasjon:



Seriekorrelasjon dersom restleddene er avhengige av hverandre, dvs dersom $Corr(u_t, u_s) \neq 0$

- Dersom strikt eksogenitet ($E(u_t|X)=0$) vil OLS gi forventningsrette estimatorer.
- Dersom kontemporær eksogenitet ($E(u_t|x_t)=0$) vil OLS gi konsistente, men ikke nødvendigvis forventningsrette estimatorer.
- Seriekorrelasjon innebærer at de vanlige inferensmetodene ikke lenger er gyldige.

Hvordan variansen til OLS-estimatene påvirkes av seriekorrelasjon:

Anta autoregressiv prosess av første orden:

$u_t = \rho_1 u_{t-1} + e_t$ der $|\rho| < 1$ (1)

$E(e_t|X) = 0$, $Var(e_t) = \sigma_e^2$, $Cov(e_t, e_s) = 0$

Dersom $\rho > 0$ har vi positiv førsteordens s.korrelasjon.

Dersom $\rho < 0$ har vi negativ førsteordens s.korrelasjon.

Har at:

$$E(u_t u_{t-1}) = E\left((\rho_1 u_{t-1} + e_t) u_{t-1}\right) = \rho_1 E(u_{t-1}^2) + E(e_t u_{t-1})$$

$$= \rho_1 \sigma_u^2 \qquad (2)$$

(1) innebærer at:

$$u_{t-1} = \rho_1 u_{t-2} + e_{t-1} \qquad (3)$$

Setter (3) inn i (1):

$$u_t = \rho_1 (\rho_1 u_{t-2} + e_{t-1}) + e_t$$

$$= \rho_1^2 u_{t-2} + \rho_1 e_{t-1} + e_t$$

Har da at:

$$E(u_t u_{t-2}) = E\left[(\rho_1^2 u_{t-2} + \rho_1 e_{t-1} + e_t) u_{t-2}\right]$$

$$= \rho_1^2 \sigma_u^2 + \rho_1 E(e_{t-1} u_{t-1}) + E(e_t u_{t-2}) = \rho_1^2 \sigma_u^2$$

Generelt:

$$E(u_t u_{t-s}) = \rho_1^s \sigma_u^2 \qquad (4)$$

Korrelasjonskoeffisienten er da generelt gitt ved:

$$\text{Corr}(u_t, u_{t-s}) = \frac{E(u_t u_{t-s})}{\sqrt{\text{Var}(u_t)\,\text{Var}(u_{t-s})}}$$

$$= \frac{\rho_1^s \sigma_u^2}{\sigma_u^2}$$

$$= \rho_1^s \qquad (5)$$

Hvis positiv seriekorrelasjon vil korrelasjonskoeffisienten avta i tid mellom restleddene.

Hvis negativ serieKorrelasjon vil også korr. koeffisienten avta i tid i absoluttverdi, men forteget vil skifte med verdien på s.

Har tidligere funnet at

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{t=1}^{t} (u_t - \bar{u})(x_t - \bar{x})}{\sum_{t=1}^{t} (x_t - \bar{x})^2}$$

Variansen er da:

$$\text{Var}(\hat{\beta}_1 | X) = \frac{\text{Var}\left(\sum_{t=1}^{t} (u_t - \bar{u})(x_t - \bar{x})\right)}{\left(\sum_{t=1}^{t} (x_t - \bar{x})^2\right)^2}$$

der $\text{Var}(u_t x_t | X)$ generelt blir ganske komplisert, og gjør vanlige inferensmetoder ugyldige.

Testing for serie korrelasjon:

1. ordens seriekorrelasjon:

1. Estimer modellen

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t$$

for å finne alle $\hat{u}_t$.

2. Estimer modellen

$$u_t = \rho_1 u_{t-1} + e_t \quad \text{(med eller uten konstantledd)}$$

3. Test hypotesen $H_0: \rho_1 = 0$

q. ordens seriekorrelasjon:

Tilsvarende, men estimer $u_t = \rho_1 u_{t-1} + \ldots + \rho_q u_{t-q} + e_t$

i 2. Alternativt kan man også inkludere forkl.variablene.

$H_0: \rho_1 = \rho_2 = \ldots = \rho_q = 0$

Durbin-Watson-testen for seriekorrelasjon:

$$DW \equiv \frac{\sum_{t=2}^{T}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T}(\hat{u}_t)^2} = \frac{\sum_{t=2}^{T}(\hat{u}_t^2 - 2\hat{u}_t\hat{u}_{t-1} + \hat{u}_{t-1}^2)}{\sum_{t=1}^{T}\hat{u}_t^2}$$

$$\approx \frac{\sum_{t=2}^{T}(2\hat{u}_t^2 - 2\hat{u}_t\hat{u}_{t-1})}{\sum_{t=1}^{T}\hat{u}_t^2}$$

$$= 2 \cdot \frac{\sum_{t=2}^{T}\hat{u}_t^2 - \sum_{t=2}^{T}\hat{u}_t\hat{u}_{t-1}}{\sum_{t=1}^{T}\hat{u}_t^2}$$

$$= 2\,(1 - \hat{\rho}_1) \qquad (6)$$

Test hvorvidt DW er signifikant forskjellig fra 2.
Positiv s.korr. dersom DW<2, negativ s.korr. dersom DW>2.

Korrigering for seriekorrelasjon av 1. orden:

Utgangspunkt i:

$$y_t = \beta_0 + \beta_1 x_t + u_t \quad (7)$$

$$u_t = \rho u_{t-1} + e_t \quad (8)$$

Ønsker nå å transformere (7) slik at seriekorrelasjonen forsvinner.

Dette kan gjøres ved først å tidsforsyve (7) med 1 og deretter multiplisere med ρ:

$$\rho y_{t-1} = \rho\beta_0 + \rho\beta_1 x_{t-1} + \rho u_{t-1} \quad (9)$$

Trekker deretter (9) fra (8):

$$y_t - \rho y_{t-1} = (1-\rho)\beta_0 + \beta_1 (x_t - \rho x_{t-1}) + (u_t - \rho u_{t-1})$$

$$= (1-\rho)\beta_0 + \beta_1 (x_t - \rho x_{t-1}) + (\rho u_{t-1} + e_t - \rho u_{t-1})$$

$$= (1-\rho)\beta_0 + \beta_1 (x_t - \rho x_{t-1}) + e_t \quad (10)$$

$$\Rightarrow \quad \tilde{y}_t = \gamma_0 + \beta_1 \tilde{x}_t + e_t \quad (11)$$

der $\tilde{y}_t = y_t - \rho y_{t-1}$, $\gamma_0 = (1-\rho)\beta_0$ & $\tilde{x}_t = x_t - \rho x_{t-1}$

Gitt antagelser om e_t er (11) ikke seriekorrelert, og kan derfor estimeres vha. OLS.

Fremgangsmåte for korrigering av seriekorrelasjon i praksis:

1. Estimer (7) $y_t = \beta_0 + \beta_1 x_t + u_t$ & finn residualene $\hat{u}_t$.
2. Estimer ρ ved å bruke OLS på hjelperegresjonen
   (8) $\hat{u}_t = \rho \hat{u}_{t-1} + \text{støy}$.
3. Bruk $\hat{\rho}$ til å finne de transformerte
   $\tilde{y}_t = y_t - \hat{\rho} y_{t-1}$ & $\tilde{x}_t = x_t - \hat{\rho} x_{t-1}$
4. Estimer (11) $\tilde{y}_t = \gamma_0 + \beta_1 \tilde{x}_t + e_t^*$
   der e_t^* symboliserer støy pga $\hat{\rho}$.

Spesialtilfelle: $\rho = 1$

$\Rightarrow \tilde{y}_t = y_t - y_{t-1} = \Delta y_t$

$\Rightarrow \tilde{x}_t = x_t - x_{t-1} = \Delta x_t$

$\Rightarrow \gamma_0 = (1-1)\beta_0 = 0$

Da reduseres (11) til:

$$\Delta y_t = \beta_1 \Delta x_t + e_t \qquad (12)$$

Alternativ strategi:

Antar at sann modell er gitt ved:

$$y_t = \beta_0 + \alpha_1 y_{t-1} + \gamma_0 x_t + \gamma_1 x_{t-1} + u_t \quad (13)$$

Anta at vi underestimerer modellen slik:

$$y_t = \beta_0 + \gamma_0 x_t + v_t \quad (14)$$

$$\text{der } v_t = \alpha_1 y_{t-1} + \gamma_1 x_{t-1} + u_t \quad (15)$$

Her er (15) korrelert med

$$v_{t-1} = \alpha_1 y_{t-2} + \gamma_1 x_{t-2} + u_{t-1} \quad (16)$$

siden y_{t-1} ifl. (13) avhenger av y_{t-2} & x_{t-2}.

Dette kan åpenbart korrigeres ved å inkludere utelatte variable y_{t-1} & x_{t-1} i (14) slik at vi får (13).

Dermed blir strategien å starte med en svært generell dynamisk modell, og stegvis utelukke variablene med lavest t-verdi.

Denne prosessen kan fortsette så lenge man ikke får seriekorrelerte restledd.

Sammenkoblede tverrsnittsdata (pooled cross sections) & paneldata:

Sammenkoblede tverrsnittsdata:

Forskjellige tverrsnittsutvalg fra forskjellige tidsperioder kobles sammen (pooled).

Paneldata:

Skiller seg fra sammenkoblede tverrsnittsdata ved at hvert enkelt tverrsnitt inneholder informasjon om de samme enhetene.

Chow-test for strukturelle endringer over tid:

Alternativ 1:

Anta at vi ønsker å teste hvorvidt

H_0: $\delta_0 = \delta_1 = \delta_2 = 0$ i modellen

$$y_i = \beta_0 + \delta_0 D_i + \beta_1 x_{i1} + \delta_1 D_i x_{i1} + \beta_2 x_{i2} + \delta_2 D_i x_{i2} + u_i \quad (1)$$

der $D_i = 1$ hvis periode 2, og $= 0$ hvis periode 1.

Test H_0 vha. $F_{3,\, n_1+n_2-6}$

Alternativ 2:

1. Estimer

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i \quad (2)$$

kun basert på enhetene fra periode 1. Finn SSR_1.

2. Estimer (2) kun basert på enhetene fra periode 2. Finn SSR_2.

3. Estimer (2) med alle observasjoner. Finn SSR_P.

4. Beregn

$$F = \frac{\frac{SSR_P - SSR_1 - SSR_2}{3}}{\frac{SSR_1 + SSR_2}{n_1 + n_2 - 6}} \sim F_{3,\, n_1+n_2-6}$$

Alternativ 1 & 2 av Chow-testen er ekvivalente og gir samme resultat.

Politikkanalyse med sammenkoblede tverrsnittsdata:

Difference-in-difference-estimatoren

Estimat før behandling:

$\hat{y}_{t-1} = \hat{\beta}_0 + \hat{\beta}_1 D$ der $D=1$ for de som skal behandles på tidspunkt t.

Estimat etter behandling:

$y_t = \hat{\gamma}_0 + \hat{\gamma}_1 D$

Forskjell i forskjell før & etter behandling:

$\hat{\delta} = \hat{\beta}_1 - \hat{\gamma}_1$

Komplett modell generelt:

$y_i = \beta_0 + \gamma_0 D_2 + \beta_1 D_T + \gamma_1 D_2 \cdot D_T + \text{kontrollvariable} + u$

der $D_2 = 1$ hvis periode 2 (etter behandling)

og $D_T = 1$ hvis behandlet gruppe

β_0 - ikke-behandlede før behandling

γ_0 - Endring for alle over tid

β_1 - Tidsuavhengig forskjell mellom behandlede & ikke-behandlede

γ_1 - Effekt av behandling

Illustrasjon av diff-in-diff-estimatoren:

	For	Etter	Etter ÷ For
Kontroll	β_0	$\beta_0 + \gamma_0$	γ_0
Behandling	$\beta_0 + \beta_1$	$\beta_0 + \gamma_0 + \beta_1 + \gamma_1$	$\gamma_0 + \gamma_1$
Beh. ÷ kontr.	β_1	$\beta_1 + \gamma_1$	γ_1

Paneldata:

- Balanserte paneldata: Har opplysninger om alle enheter på alle tidspunkter.
- Ubalanserte paneldata har i prinsipp de samme egenskapene som balanserte paneldata.
- Antall observasjoner i balansert datasett: $N \cdot T$
- Paneldatamodell:

$$y_{it} = \beta_0 + \beta_1 x_{it} + u_{it} \qquad (t=1,2) \qquad (1)$$

Et problem med (1) er at denne antagelig vil feilestimeres pga. utelatt variabel problem. Kan skille mellom variabler som er konstante og variabler som varierer over tid:

$$y_{it} = \beta_0 + \delta_0 D_2 + \beta_1 x_{it} + \alpha_i + u_{it} \qquad (2)$$

Ved estimering av (2) får vi:

$$\hat{y}_{it} = \hat{\beta}_0 + \hat{\delta} D_2 + \hat{\beta}_1 x_{it} + \hat{v}_{it} \text{ der } \hat{v}_{it} = \hat{\alpha}_i + \hat{u}_{it} \qquad (3)$$

Denne vil gi skjevhet så lenge α_i korrelerer med x_{it}.

First Difference (FD):

Kan transformere bort individspesifikke & tidsuavhengige faktorer α_i ved å differensiere (3):

År 1: $y_{i1} = \beta_0 + \beta_1 x_{i1} + \alpha_i + u_{i1}$ (4)

År 2: $y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + \alpha_i + u_{i2}$ (5)

Trekker (5) fra (4):

$y_{i2} - y_{i1} = \delta_0 + \beta_1 (x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$

$\Rightarrow \Delta y_{it} = \delta_0 + \beta_1 \Delta x_{it} + \Delta u_{it}$ (6)

Poeng: Har transformert bort α_i i (6).

Globale gjennomsnitt:

Modell: $y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + \beta_3 q_t + u_{it}$ (7)

$$\bar{y} = \frac{1}{NT} \cdot \sum_{i=1}^{n} \sum_{t=1}^{T} y_{it} \quad (8)$$

$$\bar{x} = \frac{1}{NT} \cdot \sum_{i=1}^{n} \sum_{t=1}^{T} x_{it} \quad (9)$$

$$\bar{z} = \frac{1}{NT} \cdot \sum_{i=1}^{n} \sum_{t=1}^{T} z_i = \frac{1}{NT} \cdot T z_i = \frac{T}{NT} \sum_{i=1}^{n} z_i = \frac{1}{N} \sum_{i=1}^{n} z_i \quad (10)$$

$$\bar{q} = \frac{1}{NT} \sum_{i=1}^{n} \sum_{t=1}^{T} q_t = \frac{N}{NT} \cdot \sum_{t=1}^{T} q_t = \frac{1}{T} \sum_{t=1}^{T} q_t \quad (11)$$

Forutsetninger i paneldatamodell:

Modellen:

$$y_{it} = \alpha + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + \gamma_1 z_{i1} + u_{it} \quad (12)$$

$$u_{it} = \eta_i + \varepsilon_{it} \quad (13)$$

1. $E(\varepsilon_{it} \mid x_{it}, z_{i1}) = 0$
2. $E(\varepsilon_{it} \cdot \varepsilon_{js}) = \begin{cases} \sigma_\varepsilon^2 & \text{for } i=j \text{ og } t=s \\ 0 & \text{ellers} \end{cases}$
3. $E(\eta_i \cdot \eta_j) = \begin{cases} \sigma_\eta^2 & \text{for } i=j \\ 0 & \text{ellers} \end{cases}$
4. $E(\varepsilon_{it} \eta_j) = 0$ for alle i, j og t.

Dersom også

5. $E(\eta_i \mid x_{itk}, z_{i1}) = 0$

er u_{it} uavhengig av både x_{itk} og z_{i1}. Da kan OLS brukes direkte på (12).

Fordeler ved å anvende OLS direkte på (12):

1. Utnytter all variasjon i data; både i tverrsnitt & tid.
2. Gir mulighet til å estimere effekten av tverrsnittsspesifikke variable som ikke varierer over tid.

Ulempe ved å anvende OLS direkte på (12):

Vil gi skjeve estimatorer dersom

$$E(\eta_i | x_{it1}, \ldots, z_{i1}) \neq 0$$

Dette vil være tilfellet dersom vi har utelukket individspesifikke faktorer som er korrelert med de inkluderte.

(12) kan skrives:

$$y_{it} = \alpha + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + \gamma_1 z_{i1} + \eta_i + \varepsilon_{it}$$

$$= \alpha_i + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + \gamma_1 z_{i1} + \varepsilon_{it} \qquad (14)$$

der $\alpha_i = \alpha + \eta_i$

Estimeringsprosedyrer som kan anvendes på (14) er:

1. Estimering av OLS der vi inkluderer dummyvariable for N-1 av tverrsnittsenhetene for å identifisere alle η_i (ekvivalent med å identifisere alle α_i).
2. Transformere bort η_i vha. Fixed Effects.
3. Transformere bort η_i vha. First Difference.

Fixed Effects

1. Beregn individspesifikke gjennomsnitt:

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} \qquad \bar{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{it}$$

$$\bar{x}_{ij} = \frac{1}{T}\sum_{t=1}^{T} x_{itj}$$

$$\bar{z}_i = \frac{1}{T}\sum_{t=1}^{T} z_i = \frac{1}{T} T z_i = z_i$$

Tar utgangspunkt i (14), summerer over alle t & dividerer med T:

$$\bar{y}_i = \alpha_i + \beta_1 \bar{x}_{i1} + \ldots + \beta_k \bar{x}_{ik} + \gamma_0 z_i + \bar{\varepsilon}_i \qquad (15)$$

2. Trekker (15) fra (14):

$$y_{it} - \bar{y}_i = \alpha_i - \alpha_i + \beta_1 (x_{it1} - \bar{x}_{i1}) + \ldots + \beta_k (x_{itk} - \bar{x}_{ik}) + \gamma_1 (z_i - z_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

$$= \beta_1 (x_{it1} - \bar{x}_{i1}) + \ldots + \beta_k (x_{itk} - \bar{x}_{ik}) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

$$\Rightarrow \ddot{y}_{it} = \beta_1 \ddot{x}_{it1} + \ldots + \beta_k \ddot{x}_{itk} + \ddot{\varepsilon}_{it} \qquad (16)$$

Estimering av (16) gir de såkalte fixed effects-estimatorene (within estimator).

First Difference:

1. Tar igjen utgangspunkt i (14). Tilbakedaterer denne med én periode og får:

$$y_{it-1} = \alpha_i + \beta_1 x_{it-11} + \ldots + \beta_k x_{it-1k} + \gamma_1 z_i + \varepsilon_{it-1} \quad (17)$$

2. Trekker (17) fra (14):

$$y_t - y_{t-1} = \alpha_i - \alpha_i + \beta_1 (x_{it1} - x_{it-1}) + \ldots + \beta_k (x_{itk} - x_{it-1k})$$
$$+ \gamma_1 (z_i - z_i) + (\varepsilon_{it} - \varepsilon_{it-1})$$
$$= \beta_1 (x_{it1} - x_{it-11}) + \ldots + \beta_k (x_{itk} - x_{it-1k}) + (\varepsilon_{it} - \varepsilon_{it-1})$$

$$\Rightarrow \Delta y_{it} = \beta_1 \Delta x_{it} + \ldots + \beta_k \Delta x_{it} + \Delta \varepsilon_{it} \quad (18)$$

Estimering av (18) gir de såkalte first difference-estimatorene.

Valg mellom FE & FD:

Hvis stor T & liten N foretrekkes gjerne FD.

Ved serielkorrelasjon i Eit foretrekkes gjerne FE.

Ofte lurt å rapportere begge.

fordeler & ulemper ved FE & FD:

- Fjerner individspesifikke komponenter & eliminerer dermed et endogenitets- (utelatt variabel-) problem. Både FE & FD gir konsistente & forventningsrette estimatorer.
- Utnytter bare variasjon i tid, siden individspesifikke egenskaper er transformert bort.
- Andre begrensninger: Vi har ikke alltid paneldata, og vi kan være interessert i individspesifikke variable som ikke varierer over tid. FE & FD løser heller ikke problemer med utelatte variable som varierer i tid & er korrelert med inkluderte forklaringsvariable.

Instrumentvariabelmetoden & 2-SLS:

Den avhengige variabelen er endogen når den er korrelert med restleddet:

$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + u_{i,t}$ der $Cov(u_{i,t}, x_{i,t}) \neq 0$

Med utgangspunkt i denne sammenhengen:

$$Cov(y,z) = \beta_1 Cov(x,z) + Cov(u,z) \quad (1)$$

$$\Rightarrow \beta_1 = \frac{Cov(y,z)}{Cov(x,z)} - \frac{Cov(u,z)}{Cov(x,z)}$$

$$= \frac{Cov(y,z)}{Cov(x,z)} \quad (2)$$

gitt at:

$$Cov(u,z) = 0 \quad \text{dvs. eksogen} \quad (3)$$

$$Cov(x,z) \neq 0 \quad \text{dvs. relevant} \quad (4)$$

Ser at $\underset{n\to\infty}{plim}(\hat{\beta}_1) = \beta_1$

Bytter ut teoretiske momenter med empiriske:

$$\hat{\beta}_1^{iv} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})} \quad (5)$$

Homoskedastisitetsantagelsen:

$$E(u^2|z) = \sigma^2 = Var(u) \quad (6)$$

Kan vises at gitt (3), (4) & (6) er den asymptotiske variansen til $\hat{\beta}_1^{iv}$:

$$Var(\hat{\beta}_1^{iv}) = \frac{\sigma^2}{n\sigma_x^2 \rho_{x,z}^2} \quad (7)$$

Den empiriske motparten er:

$$\widehat{Var}(\hat{\beta}_1^{iv}) = \frac{\hat{\sigma}^2}{SST_x \cdot R_{x,y}^2} \quad (8)$$

og

$$Se(\hat{\beta}_1^{iv}) = \frac{\hat{\sigma}}{\sqrt{SST_x \cdot R_{x,z}^2}} \quad (9)$$

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

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

$R_{x,z}^2$ er R^2 fra $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 z_i + \text{støy}$

Sammenlignet med $Var(\hat{\beta}_1^{OLS})$:

$$\widehat{Var}(\hat{\beta}_1^{iv}) = \frac{\hat{\sigma}^2}{SST_x \cdot R_{x,z}^2} > \frac{\hat{\sigma}^2}{SST_x} = \widehat{Var}(\hat{\beta}_1^{OLS})$$

$\Rightarrow$ IV-estimering går alltid ut over variansen til estimatoren.

Verdt å merke seg at (4) innebærer at:

$x = \beta_0 + \beta_1 z + u$ må gi signifikant $\beta_1 \neq 0$

siden $\beta_1 = \frac{Cov(x,z)}{Var(z)}$

Merk: R^2 gir ikke mye mening i instrumenterte regresjoner; kan ta negative verdier og >1. Gjelder også SSR, slik at F-tester ikke lenger kan brukes. Tilpassede F-tester kan imidlertid settes opp.

IV med multippel regresjon:

Anta:

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + u_1 \qquad (10)$$

Kaller (10) for strukturlikningen der x_1 mistenkes å være korrelert med restleddet u_1.

Normallikningen i IV-metoden er da gitt ved:

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 z_{i1}) &= 0 \\ \sum_{i=1}^{n} z_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 z_{i1}) &= 0 \\ \sum_{i=1}^{n} z_{i2} (y_i - \beta_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 z_{i1}) &= 0 \end{aligned} \qquad (11)$$

der z_{i2} er instrumentvariabelen. Likningssettet (11) gir 3 likninger til å estimere de tre ukjente $\hat{\beta}_0$, $\hat{\beta}_1$ & $\hat{\beta}_2$.

For å teste (4) kan man sette opp følgende redusert form-likning:

$$x_{i1} = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + u_2 \qquad (12)$$

og teste $H_0: \beta_2 = 0$.

Redusert-form-likningen skal inneholde alle de eksogene variablene z.

2 Stage Least Squares (2-SLS):

Ved bruk av mer enn én instrumentvariabel for x kan vi benytte oss av 2-SLS:

Tar igjen utgangspunkt i (10), men tenker oss nå to instrumenter z_2 & z_3.

Den beste instrumentvariabelen er da estimatet $\hat{x}_1$ i redusert form-regresjonen:

$$x_1 = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 \qquad (13)$$

Kravet (4) kan da testes der $H_0: \beta_2 = \beta_3 = 0$ må forkastes for at instrumentene skal være relevante.

Så lenge IV-variabelen er gyldig kan $\hat{x}_{i1}$ erstatte z_{i2} i (11).

Dette er identisk med å beregne

$$y_1 = \beta_0 + \beta_1 \hat{x}_1 + z_1 + u_1 \qquad (14)$$

vha OLS. Derfor kalles dette 2-SLS-estimatoren for $\hat{\beta}_1^{2SLS}$.

2-SLS kan også brukes ved instrumentering av flere endogene forklaringsvariable. Metoden er ekvivalent med den beskrevet frem til nå, men man trenger minst én eksogen variabel (instrument) som ikke er inkludert i strukturlikningen per endogene forklaringsvariabel.

Testing for endogenitet:

Motivasjon: OLS-estimatorene er mer effisiente når de er eksogene, derfor ikke et mål i seg selv med instrumentvariable.

Fremgangsmåte:

1. Estimer redusert form-regresjonen.
   Finn residualene $\hat{u}_2$.
2. Estimer strukturlikningen inkludert leddet $\rho_1 \hat{u}_2$. Test $H_0: \rho_1 = 0$. Hvis H_0 forkastes kan vi konkludere at x_1 er endogen $Corr(u_1, u_2) \neq 0$

Testing for overidentifisering:

Når man bare har ett instrument for den endogene forklaringsvariabelen er denne eksakt identifisert. Da kan man bare resonnere rundt gyldigheten av (3):

$Cov(z, u) = 0$

Når man har flere instrumenter kan man teste om man har flere enn nødvendig vha overidentifiseringstester.

Gitt homoskedastisitet kan man gjøre følgende:

1. Estimer strukturligningen vha 2SLS for å finne 2SLS-residualene $\hat{u}_{i1}$.
2. Estimer $\hat{u}_{i1}$ som funksjon av alle eksogene variable z. Finn R^2 kalt R^2_1.
3. H_0: Alle instrumenter er ukorrelert med u_1. Test H_0 vha at $n \cdot R^2_1 \sim \chi^2_q$ der q er antall instrumenter minus antall endogene forklaringsvariable. Hvis H_0 forkastes er i det minste noen instrumenter eksogene.