ECONnect
NTNU

Faktor

- en eksamensavis utgitt av ECONnect

Eksamensbesvarelse:
SØK3001 – Økonometri I

Eksamen: Våren 2009
Antall sider: 22

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
Elise Caspersen (Fagdagansvarlig)	elise@econnect-ntnu.no
Pål Christian Vågbø (Bedriftansvarlig)	paal@econnect-ntnu.no
Tormod Hagerup (Økonomi/Samfunnsøk.)	tormod@econnect-ntnu.no
Tiril Toftedahl	tiril@econnect-ntnu.no
Louis Dieffenthaler	louis@econnect-ntnu.no
Tone Hedvig	tone@econnect-ntnu.no
Ole Christian Grytten	ole@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: Eksamensbesvarelsene har i varierende grad feil og mangler, både oppsett og innhold. De vil også kun vise en av flere mulige fremgangsmåter. ECONnect står ikke ansvarlig for selve faginnholdet.

Kommentar fra sensor:

Besvarelsen framstår som svært ryddig, veldisponert og poengtert. Besvarelsen indikerer at kandidaten behersker sentrale økonometriske metoder svært godt og han/hun er i stand til å anvende metodekunnskapene på de problemstillinger som reises i oppgaveteksten. Besvarelsen er klart best på oppgave 1 og 2, mens oppgave 3 inneholder noen uklarheter som bør kommenteres. På oppgave 3b bes for eksempel kandidaten om å konstruere et 95% konfidensintervall for effekten av ølskatt på 0,50$ per kasse. Kandidaten velger da å lage konfidensintervallet rundt den estimerte parameteren for ølskatt, 0.45. Her ville imidlertid det mest naturlige være å konstruere konfidensintervallet rundt effekten av 0,50$. Det gir et konfidensintervall for effekten på 0.45 • 0.5 ± 1.96 • 0.22• 0.5. Altså et konfidensintervall på (-0.44, -0.01) som inkluderer verdier av effekten svært nær null. På spm c) spørres det om effekten av at inntekten i delstaten øker med 1%. Kandidaten kommenterer at koeffisientestimatet betyr at antall dødsfall per 10000 innbyggere øker med 0.0181 når inntekten øker med 1%. Dette er effekten det spørres om. Tallet 14.66 som han/hun kommer fram til like nedenfor er ukommentert og bidrar lite til forståelse. Dessuten konstrueres det igjen et konfidensintervall for parameteren og ikke for effekten av 1% økning i inntekten.

Norges teknisk-naturvitenskapelige universitet
Institutt for samfunnsøkonomi

EKSAMENSOPPGAVE I SØK3001
ØKONOMETRI I

Faglig kontakt under eksamen: Jørn Rattsø

Tlf.: 9 19 34

Eksamensdato: Tirsdag 2. juni 2009

Eksamenssted: Dragvoll

Eksamenstid: 5 timer

Studiepoeng: 15

Tillatte hjelpemidler: Flg formelsamling: Knut Sydsæter, Arne Strøm og Peter Berck (2006): Matematisk formelsamling for økonomer, 4utg. Gyldendal akademiske. Knut Sydsæter, Arne Strøm, og Peter Berck (2005): Economists' mathematical manual, Berlin.
Enkel kalkulator Citizen SR-270x el. HP 30S.

Sensur: 23. juni 2009

Antall sider bokmål: 2
Antall sider nynorsk: 2
Antall sider engelsk: 2

Merk! Studentene må primært gjøre seg kjent med sensur på stud.web eller ved å oppsøke sensuroppslagene. Evt telefoner om sensur må rettes til instituttet. Eksamenskontoret vil ikke kunne svare på slike telefoner.

Bokmål

Eksamen består av tre oppgaver som alle skal besvares

Oppgave 1

Forkklar hva slags skjevhet vi kan få i estimatene hvis vi utelater en relevant variabel, også kalt underspesifikasjon av modellen.

Anta en modell med en avhengig variabel og en forklaringsvariabel og hvor en relevant variabel er utlatt. Hvis vi antar at den utelatte variabelen har positiv effekt på den avhengige variabelen, hva kan vi da si om skjevheten i estimatet for forklaringsvariabelen? Forklar hvilke faktorer som bestemmer størrelsen på skjevheten.

Diskuter hvordan vi bør gå fram i analysen hvis den utelatte variabelen ikke er observerbar.

Oppgave 2

En analyse av kjønnsforskjeller blant toppledere i USA er basert på data for 46 670 bedrifter. Data omfatter variablene WAGE (årslønn, målt i dollar), MARKET VALUE (markedsverdi av bedriften, målt i millioner dollar), RETURN (aksjeavkastning, målt i prosent). Kjønn er representert med dummyvariabelen FEMALE som har verdi 1 når topplederen er kvinne og verdi 0 når det er en mann. Fotskrift i angir bedrift i hvor i= 1,..,n og n = 46 670.

En regresjonsanalyse av sammenhengen mellom lønn og kjønn ser slik ut, hvor standardavvik er oppgitt i parenteser under de estimerte parametrene:

$$\ln(WAGE)_i = \underset{(0.01)}{6.48} - \underset{(0.05)}{0.44}\, FEMALE_i$$

a) Forklar hvordan koeffisienten med verdi -0.44 skal tolkes.

b) Tilsier dette resultatet at kvinnelige toppledere diskrimineres?

I en utvidelse av analysen estimeres følgende relasjon:

$$\ln(WAGE)_i = \underset{(0.03)}{3.86} - \underset{(0.04)}{0.28}\, FEMALE_i - \underset{(0.004)}{0.37}\ln(MARKETVALUE)_i + \underset{(0.003)}{0.004}\, RETURN_i$$

c) Forklar hvorfor estimert koeffisient for dummyvariabelen som representerer kjønnsforskjellen nå er redusert til -0.28.

d) Gi en tolkning av effekten av markedsverdi på toppledernes lønnsnivå.

e) Er det større sannsynlighet for at store bedrifter har kvinnelige toppledere?

Oppgave 3

Det er gjort en panelanalyse med data for 48 stater i USA i perioden 1982 til 1988 for å analysere faktorer som påvirker trafikkdødsfall. Data omfatter 336 observasjoner, trafikkdødsfall er representert ved variabelen DEATH og målt i dødsfall per 10.000 innbyggere. Estimatene for følgende variable rapporteres: BEERTAX (ølskatt målt i dollar per kasse øl), UNEMP (arbeidsledighetsrate målt i prosent) og REALINC (realinntekt per innbygger). Gjennomsnittlig trafikkdødsfall per 10.000 innbyggere i denne perioden er 2 og gjennomsnittlig ølskatt er 0.50 $ per kasse. Fotskrift i angir stat og t angir år.

En enkel analyse av sammenhengen mellom trafikkdødsfall og ølskatt gir følgende relasjon (standardavvik er oppgitt i parantes under estimert parameter, konstantledd er utelatt):

$$DEATH_{it} = \underset{(0.05)}{0.36} BEERTAX_{it}$$

a) Gi en tolkning av hva dette estimeringsresultat sier som sammenhengen mellom ølskatt og trafikkdødsfall.

I en utvidelse av analysen brukes fixed effects konstanter for hver stat og tidsdummier for år og følgende estimeringsresultat oppnås, hvor konstantene er utelatt:

$$DEATH_{it} = -\underset{(0.22)}{0.45} BEERTAX_{it} - \underset{(0.012)}{0.063} UNEMP_{it} + \underset{(0.47)}{1.81} \ln(REALINC)_{it}$$

b) Gitt dette estimeringsresultat, hvor stor virkning vil en økning av ølskatten med 0.50 $ per kasse ha på trafikkdødsfall i gjennomsnitt? Hva er konfidensintervallet for effekten?

c) Staten New Jersey har 8.1 millioner innbyggere. Anta at New Jersey har en ølskatt på 1$ per kasse i 1988. Hvor mange liv kan vi si at ølskatten redder i løpet av et år? Konstruer et 95-% konfidensintervall for effekten.

d) De økonomiske variablene synes å ha betydning for trafikkdødsfallene. Gi en tolking av effekten på trafikkdødsfall av arbeidsledighet og inntektsnivå.

e) Anta at realinntekten i New Jersey øker med 1% neste år. Gi en prediksjon av endringen i trafikkdødsfall i staten neste år. Konstruer et 90-% konfidensintervall for prediksjonen.

f) Det er reist spørsmål om arbeidsledigheten har større effekt på trafikkdødsfall i de vestlige statene i USA. Hvordan kan du teste en slik hypotese? Vis hvordan du vil spesifisere regresjonsmodellen for å gjøre slik test og hvilken test du vil bruke.

g) En forsker vil undersøke hvordan snøforholdene påvirker trafikkdødsfall. Hun samler inn data om snøforhold og legger en variabel til den utvidede modellen ovenfor. Sammenlign følgende framgangsmåter for å analysere snøforholdene: 1) Forskeren samler inn data om gjennomsnittlig snøfall i hver stat og bruker dette som variabel (AVERAGESNOW). 2) Forskeren samler inn data om snøfall i hver stat for hvert år i datamaterialet og bruker denne som variabel (SNOW).

NTNU

Kandidatnr.

Dato: 02.06.09 Side: 1

Fagnr/emnekode SØK 3001 Antall ark: 22

Denne kolonnen er forbeholdt sensor

Oppg 1

Antar enkel lineær populasjonsmodell;

$$y = \beta_0 + \beta_1 x_1 + u$$ (u: restledd)

Antar tilfeldig utvalg, dvs $Cov(u_i, u_j) = 0$ for $i \neq j$
Får da empirisk modell

(1) $y_i = \beta_0 + \beta_1 x_{i1} + u_i \qquad i = 1, 2, \ldots, n$

Antar variasjon i forklaringsvariabel, x.

Ønsker å estimere modellen vha minste kvadraters metode, MKM. Estimatene vi finner for β_0 og β_1, $\hat{\beta}_0$ og $\hat{\beta}_1$ gir så predikert verdi på $\hat{y}$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}$$

og vi finner residualer (det observerbare motstykket til restleddene) som

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

MKM går ut på å velge $\hat{\beta}_0$ og $\hat{\beta}_1$ slik at den kvadrerte summen av residualer minimeres, slik at vi får (en slags) gjennomsnittlig sammenheng mellom x_1 og y.

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1})^2 \equiv SSR$$

(I det følgende skriver jeg $\sum_{i=1}^{n} = \sum$)

FOB

1) $\frac{\partial SSR}{\partial \hat{\beta}_0} = 0 \rightarrow 2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1}) \cdot (-1) = 0$

$$\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1}) = 0$$

Ganger med $\frac{1}{n}$ på begge sider

$$\frac{1}{n} \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1}) = 0$$

$$\frac{1}{n} \sum y_i - \frac{1}{n} n \hat{\beta}_0 - \hat{\beta}_1 \frac{1}{n} \sum x_{i1} = 0$$

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}_1 = 0 \qquad \text{der } \bar{y} = \frac{1}{n} \sum y_i, \ \bar{x}_1 = \frac{1}{n} \sum x_{i1}$$

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

NTNU

Kandidatnr.

Dato: 02.06.09 Side: 2

Fagnr/emnekode SØK 3001 Antall ark: 22

Denne kolonnen er forbeholdt sensor

ii) $\frac{\partial SSR}{\partial \beta_1} = 0 \rightarrow 2\sum(y_i - \hat\beta_0 - \hat\beta_1 x_i)(-x_i) = 0$

$$\sum(y_i - \hat\beta_0 - \hat\beta_1 x_i)\, x_i = 0$$

Setter inn for $\hat\beta_0$

$$\sum_i (y_i - \bar y + \hat\beta_1 \bar x - \hat\beta_1 x_i)\, x_i = 0$$

$$\sum_i (y_i - \bar y) x_i = \hat\beta_1 \sum_i (x_i - \bar x_i) x_i$$

$$\hat\beta_1 = \frac{\sum (y_i - \bar y) x_i}{\sum (x_i - \bar x) x_i}$$

Som er det samme som

$$\boxed{\hat\beta_1 = \frac{\sum (y_i - \bar y)(x_i - \bar x_i)}{\sum (x_i - \bar x_i)^2}} \quad \#$$

Omskriver videre...

Har at $\bar y = \beta_0 + \beta_1 \bar x_1 + \bar u$ (snittet av (1))

Trekker dette fra (1)

$$y_i - \bar y = \beta_1 (x_i - \bar x_1) + (u_i - \bar u)$$

Inn i uttrykket for $\hat\beta_1$:

$$\hat\beta_1 = \frac{\sum (\beta_1 (x_i - \bar x_1) + (u_i - \bar u))(x_i - \bar x_1)}{\sum (x_i - \bar x_1)^2}$$

$$= \frac{\beta_1 \sum (x_i - \bar x_1)^2 + \sum (x_i - \bar x)(u_i - \bar u)}{\sum (x_i - \bar x_1)^2}$$

$$= \beta_1 + \frac{\sum (x_i - \bar x_1) u_i}{\sum (x_i - \bar x_1)^2}$$ siden $\sum (x_i - \bar x_1)(u_i - \bar u)$

$$= \sum (x_i - \bar x_1) u_i - \bar u \underbrace{\sum (x_i - \bar x_1)}_{n\bar x_1 - n\bar x_1 = 0}$$

Hvis vi nå antar

(*) $E(u_i \mid x_i) = 0$

dvs $E(u_i) = 0$

$Cov(x_i, u_i) = 0$,

Vil

$$E(\hat\beta_1 \mid X_1) = \beta_1 + \frac{\sum (x_i - \bar x_1) \overbrace{E(u_i \mid x_i)}^{0}}{\sum (x_i - \bar x_1)^2} = \beta_1$$

Den gule kopien beholder du

Så dersom (*) gjelder, dvs at (1) er sann modell er $\hat{\beta}_1$ forventningsrett!

Men ved underspesifikasjon av modellen, kan vi f.eks anta at sann modell er gitt ved

(2) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$, der $E(u_i | x_{i1}, x_{i2}) = 0$

Dermed er altså x_{i2} utelatt variabel i (1)

Fra sann modell finnes u

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

og

$$y_i - \bar{y} = \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} + \bar{x}_2) + (u_i - \bar{u})$$

Setter nå dette inn i uttrykket for $\hat{\beta}_1$ i (#)

$$\hat{\beta}_1 = \frac{\sum\left(\beta_1(x_{i1}-\bar{x}_1) + \beta_2(x_{i2}+\bar{x}_2) + (u_i - \bar{u})\right)(x_{i1}-\bar{x}_1)}{\sum(x_{i1}-\bar{x})^2}$$

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

$$= \beta_1 + \frac{\beta_2 \sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum(x_{i1}-\bar{x})^2} + \frac{\sum(x_{i1}-\bar{x}_1)u_i}{\sum(x_{i1}-\bar{x}_1)^2}$$

Tar nå forventningen på begge sider, betinget på x_{i1} og x_{i2}

$$E(\hat{\beta}_1 | x_{i1}, x_{i2}) = \beta_1 + \frac{\beta_2 \sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum(x_{i1}-\bar{x}_1)^2} + \frac{\sum(x_{i1}-\bar{x}_1)\overbrace{E(u_i|x_{i1}, x_{i2})}^{0}}{\sum(x_{i1}-\bar{x}_1)^2}$$

(merk at vi ikke setter forventning på x_1 og x_2 leddene ettersom vi jo forventningen gjøres betinget på disse)

√

$$E(\hat{\beta}_1 | x_{i1}, x_{i2}) = \beta_1 + \beta_2 \hat{\delta}_1$$

der $\hat{\delta}_1 = \frac{\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum(x_{i1}-\bar{x}_1)^2}$

Som vi jo ser er ~~mu~~ estimatet av δ_1 fra likningen $x_{i2} = \delta_0 + \delta_1 x_{i1} + v$ $(E[v|x_{i1}] = 0)$

NTNU

Kandidatnr.

Dato: 02.06.09 Side: 4

Fagnr/emnekode SØK 3001

Antall ark: 22

Denne kolonnen er forbeholdt sensor

Dermed ser vi at når vi utelater en relevant variabel (dvs $\beta_2 \neq 0$) vil vi få forventningsskjevhet i MKM-estimatet på β_1 dersom utelatt variabel, x_2, er korrelert med inkludert forklaringsvariabel, x_1, i ligning (1).
Dette kalles utelatt variabelskjevhet, som altså kan oppstå ved underspesifikasjon av modellen.

Som sagt er x_2 her relevant variabel (påvirker altså y, $\beta_2 \neq 0$), og vi har skjevhet dersom x_1 og x_2 er korrelerte, dvs $\hat{\delta}_1 \neq 0$.

Vi får oppgitt at utelatt variabel har positiv effekt på avhengig variabel, dvs $\beta_2 > 0$.
Størrelsen på skjevheten vil da avhenge av om x_2 er positivt eller negativt korrelert med x_1.

Hvis

* $\hat{\delta}_1 > 0$, positiv korrelasjon mellom x_1 og x_2,

er $\beta_2 \hat{\delta}_1 > 0$, slik at $E(\hat{\beta}_1 | x_1, x_2) > \beta_1$

$\Rightarrow$ utelatt variabel gir overestimert effekt av x_1 på y (positiv skjevhet)

Intuisjon: økt x_1 øker x_2 når de er positivt korrelerte. Økt x_2 øker y når $\beta_2 > 0$, og denne effekten kommer i tillegg til den isolerte effekten av x_1. Dermed ser effekten av x_1 ut til å være større enn den virkelig er.

* $\hat{\delta}_2 < 0$, negativ korrelasjon

er $\beta_2 \hat{\delta}_1 < 0$, slik at $E(\hat{\beta}_1 | x_1, x_2) < \beta_1$

$\Rightarrow$ underestimering (negativ skjevhet)

* $\hat{\delta}_2 = 0$, $\beta_2 \hat{\delta}_1 = 0$ $\rightarrow$ $E(\hat{\beta}_1 | x_1, x_2) = \beta_1$

$\Rightarrow$ ingen skjevhet

Vi ser altså at det er β_2 og $\hat{\delta}_1$ som er avgjørende for skjevheten: Jo større effekt av x_2 på y (større β_2) og jo større korrelasjon (pos eller negativ) mellom x_1 og x_2 - jo større er denne skjevheten.

Den gule kopien beholder du

Denne kolonnen er forbeholdt sensor

Fra analysen så langt ser vi at vi for å få et forventningsrett estimat på effekten av x_1 på y, bør kontrollere for x_2 (altså estimere (2))

Det kan imidlertid være at grunnen til at variabelen utelates er at den ikke observeres. F.eks. hvis vi ønsker å finne effekten av utdanning (educ) på lønn (wage), vil medfødt evne (abil) gjerne være korrelert med educ ($\delta_1 \neq 0$) og påvirke wage ($\beta_2 \neq 0$), men dette er en uobserverbar variabel.
Hva gjør man da?

1) Overser problemet og estimerer modell med utelatt variabel

$$wage_i = \beta_0 + \beta_1 educ_i + u_i$$

↳ $\beta_2 abil_i$ inngår

Vi kan her forutsi skjevheten da abil og educ trolig er positivt korrelert og abil trolig har positiv effekt på wage. Dermed kan vi gå ut fra at β_1 vil overestimere effekten av educ på wage, og ta dette med i betraktningen.

2) Bruke en proxy på den utelatte variabelen. F.eks kan vi bruke IQ-score som proxy på abil.
Dette vil være en forbedring av β_1-estimatet sammenlignet med 1), men IQ vil ikke fange opp alt av abil, slik at vi fortsatt ikke kan si oss fornøyd.
(Vil få måle feil problem her...)

3) Antar nå at vi ikke bare har tverrsnittsdata, men at vi har fulgt de samme individene, $i = 1, 2, \dots n$, over flere år, slik at vi har paneldata.
Får da følgende modell

$$(3) \quad y_{it} = \beta_0 + \beta_1 x_{it1} + u_{it} \qquad i = 1, 2, \dots n \quad t = 1, 2, \dots T$$

Antar at vi kan dele opp restleddet

$$u_{it} = \eta_i + \varepsilon_{it}$$

√ der η_i - individspesifikk komponent som fanger opp restleddsvariable som er spesifikt for individet

Denne kolonnen er forbeholdt sensor

og da ikke varierer over tid. Herunder kan vi dermed plassere åbil.

ε_{it} er den delen av restleddet som i tillegg til å variere mellom individ, også har tidsdimensjon

Antar her

i) $E(\varepsilon_{it}, \varepsilon_{js}) = \begin{cases} \sigma_\varepsilon^2 & \text{hvis } i=j,\ t=s \\ 0 & \text{ellers} \end{cases}$ (homoskedastisitet) (ingen seriekorrelasjon)

ii) $E(\eta_i, \eta_j) = \begin{cases} \sigma_\eta^2 & \text{hvis } i=j \\ 0 & \end{cases}$ (sikres ved tilfeldig utvalg)

iii) $E(\eta_i, \varepsilon_{jt}) = 0$ for alle i, j, t (uavhengighet)

iv) $E(\varepsilon_{it} \mid x_{it1}) = 0$

Men her ser at åbil inngår i η_i, slik at

$$E(\eta_i \mid x_{it1}) \neq 0$$

og dermed vil MKM anvendt på (3) gi utelatt variabel skjevhet for $\hat{\beta}_1$.

Vi skriver istedet (3) med individspesifikt konstantledd

(4) $y_{it} = \beta_{0i} + \beta_1 x_{it1} + \varepsilon_{it}$ der $\beta_{0i} = \beta_0 + \eta_i$

Denne likningen summeres over alle t, og deles på T,

(5) $\bar{y}_i = \beta_{0i} + \beta_1 \bar{x}_{i1} + \bar{\varepsilon}_i$ der $\begin{cases} \bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} \\ \bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it} \\ \bar{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{it} \end{cases}$

→ er individspesifikke gjennomsnitt.

Merk at $\frac{1}{T}\sum_{t=1}^{T} \beta_{0i} = \frac{1}{T} \cdot T\beta_{0i} = \beta_{0i}$

siden leddet ikke varierer i tidsdimensjonen

NTNU

Kandidatnr.

Dato: 02.06.09 Side: 7

Fagnr/emnekode SØK 3001 Antall ark: 22

Denne kolonnen er forbeholdt sensor

Trekker (5) fra (4) og får

(6) $y_{it} - \bar{y}_i = \beta_1(x_{it1} - \bar{x}_{i1}) + (\varepsilon_{it} - \bar{\varepsilon}_i)$

✓ Her vil da få kvittet oss med hele utelatt variabel-problemet ettersom vi her har transformert bort det individspesifikke restleddet.
⇒ fixed effect transformasjonen.

En alternativ transformasjonsmetode er den såkalte first difference.
(4) på første differens gir

(7) $\Delta y_{it} = \beta_{0i} + \beta \Delta x_{it} + \Delta \varepsilon_{it}$ hvor $\Delta y_{it} = y_{it} - y_{it-1}$

$\Delta x_{it} = x_{it} - x_{it-1}$

Trekker vi (7) fra (5) får vi igjen transformert bort problemleddet $\Delta \varepsilon_{it} = \varepsilon_{it} - \varepsilon_{it-1}$

(8) $y_{it} - \Delta y_{it} = \hat{\beta}_1(x_{it} - \Delta x_{it}) + \varepsilon_{it} - \Delta \varepsilon_{it}$

Både (6) og (8) kan estimeres ved bruk av MKM og vi gi forventningsrette estimat på β_1 (antatt at vi ikke har seriekorrelerte restledd, hvis vi det bør (8) benyttes ettersom restleddet her er mer "gunstig").

Både FE og FD er god løsning.
Problemet kan være at vi har lite (eller ingen) variasjon i forklaringsvariabel dersom denne varierer lite over tid - husk at vi kun utnytter variasjon i tidsdimensjonen her, vi har lukket ut individvariasjon. Dette vil isåfall gi upresist estimat på β_1 (stor varians).
Det er dessuten ikke sikkert vi har paneldata tilgjengelig.

4) Instrumentvariabelmetoden (IV)

Går nå tilbake til modellen

(1) $y_i = \beta_0 + \beta_1 x_{i1} + u_i$

der problemet er at $\mathrm{Cov}(x_{i1}, u_i) \neq 0$; et endogenitets-problem pga utelatt variabel.

Anta ~~variabel~~ at det finnes en variabel, z_1, som har følgende egenskaper:

i) $Cov(x_1, z_1) \neq 0$
Variabelen z er altså korrelert med forklarings-variabel x.
→ relevans.
Dette kan vi teste ved å estimere modellen
$$x_{1i} = \pi_0 + \pi_1 z_{1i} + \omega_i \qquad E(\omega_i | z_{i1}) = 0$$
og kjøre t-test på $\hat{\pi}_1$. Vi får da utføre nullhypotesen H_0: $\pi_1 = 0$, og dersom vi får forkastet denne (se mer om t-test i oppg 2) er z_1 relevant instrument for x_1.

√

ii) $Cov(z_1, u) = 0$
Variabelen z skal altså ikke være korrelert med restleddet
– eksogenitet
Dette vil si at z_1 ikke skal påvirke y, når vi har kontrollert for effekten via x_1

Med disse egenskapene kan z_1 brukes som instrument for x_1. Intuisjonen er at x_1 består av en eksogen og en endogen del, hvor den eksogene delen kan representeres av z_1.

Skal vise at vi ved å bruke z_1 som IV får konsistent $\hat{\beta}_1$, dvs at $\hat{\beta}_1$ går mot β_1 i sannsynlighetsgrense når $n \to \infty$.

Når vi fortsatt antar $E(u_i) = 0$, er

$$E(y) = \beta_0 + \beta_1 E(x_1)$$

Trekker dette fra (1)

$$y_i - E(y) = \beta_1 (x_{i1} - E(x_1)) + u_i$$

Ganger med $(z_{i1} - E(z_1))$ og tar forventning på begge sider

$$E[(y_i - E(y))(z_{i1} - E(z_1))] = \beta_1 E[(x_{i1} - E(x_1))(z_{i1} - E[z_1])]$$
$$+ E[u_i(z_{i1} - E(z_1))]$$

$$Cov(y_i, z_1) = \beta_1 + Cov(x_1, z_1) + Cov(u, z_1)$$

Gitt ii) får vi

$$\beta_1 = \frac{Cov(y, z_1)}{Cov(x_1, z_1)}$$

Erstatter vi nå populasjonsmomentene med sine empiriske momenter finner vi IV-estimator ved momentmetoden som

$$\hat{\beta}_1^{IV} = \frac{\frac{1}{n}\sum(y_i - \bar{y})(z_{i1} - \bar{z}_1)}{\frac{1}{n}\sum(x_{i1} - \bar{x}_1)(z_{i1} - \bar{z}_1)} = \frac{\sum(y_i - \bar{y})(z_{i1} - \bar{z}_1)}{\sum(x_{i1} - \bar{x}_1)(z_{i1} - \bar{z}_1)}$$

Tar nå sannsynlighetsgrensen på hver side

$$\text{plim}\, \hat{\beta}_1^{IV} = \frac{\text{plim}\, \frac{1}{n}\sum(y_i - \bar{y})(z_{i1} - \bar{z}_1)}{\text{plim}\, \frac{1}{n}\sum(x_{i1} - \bar{x}_1)(z_{i1} - \bar{z}_1)}$$

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

der den første likheten følger av at empiriske momenter vil konvergere mot sine respektive sanne populasjonsverdier når $n \to \infty$, og den andre likheten følger om ii) er sann.

Ser at også i) er nødvendig i argumentet, ettersom vi ikke kan ha null i teller!

→ $\hat{\beta}_1^{IV}$ er konsistent estimator, og vi har derfor her en god løsning på endogenitetsproblemet utelatt variabel påfører oss.

Merk midlertid at variansen til $\hat{\beta}_1^{IV}$ er større enn $\hat{\beta}_1^{MKM}$ (får ved å estimere (1)), ettersom variasjonen i z er mindre enn variasjonen i x - siden z jo bare fanger opp den eksogene delen av x - og dette gir altså mindre presist β_1-estimat. Jo mindre korrelasjon mellom z_1 og x_1, jo større er variansen i $\hat{\beta}_1^{IV}$.

Merk også at $\hat{\beta}_1^{IV}$ alternativt kan finnes ved 2stegs MKM.
Første steg vil være å estimere redusert form ligningen, $x_{i1} = \pi_0 + \pi_1 z_{i1} + w_i$, ved MKM,

NTNU

Kandidatnr.

Dato: 02.06.09 Side: 10

Fagnr/emnekode SØK 3001

Antall ark: 22

Denne kolonnen er forbeholdt sensor

hvorfra vi finner predikerte verdier for X_{i1}

$$\hat{X}_{i1} = \hat{\pi}_0 + \hat{\pi}_1 Z_{i1}$$

Andre steg blir så å estimere ligning (1), men hvor X_{i1} erstattes med predikert verdi

$$Y_i = \beta_0 + \beta_1 \hat{X}_{i1} + \text{støy}_i$$

Ved MKM også her finner vi samme $\hat{\beta}_1$ som ved instrumentvariabelmetoden

2stegs MKM er mer praktisk dersom vi har flere variable som kan brukes som instrument for X_1. Prosedyren er da tilsvarende som over bare at vi får flere Z-er i redusert form ligningen som så brukes til å finne predikert X.

Vi har nå sett på 4 metoder for hvordan vi kan gå frem dersom utelatt variabel ikke er observerbar

Merk at utfordringen i løsning 4) er å finne et instrument Z.

Den gule kopien beholder du

Tapir Uttrykk 04/09

NTNU

Kandidatnr.

Dato: 02.06.09 Side: 11

Fagnr/emnekode SØK 3001

Antall ark: 22

Denne kolonnen er forbeholdt sensor

Oppg 2 Analyse av kjønnsforskjeller mellom toppledere i USA

n = 46 670 (bedrifter)

Wage - årslønn (dollar)
Marketvalue - markedsverdi av bedriften (millioner dollar)
return - (aksjeavkastning, målt i %)

Female = 1 når toppleder er kvinne
0 når mann

Estimerer modellen

(1) $\ln(wage)_i = \beta_0 + \beta_1 Female_i + u_i$

og får

$$\widehat{\ln(wage)_i} = \underset{(0,01)}{6,48} - \underset{(0,05)}{0,44}\,Female_i$$

a) $\hat{\beta}_1 = -0,44$
hvordan tolkes β_1?

Differensierer ligning (1)

$$\frac{1}{wage}\,dwage = \beta_1\, dFemale$$

$$\beta_1 = \frac{\frac{dwage}{wage}}{dFemale}$$

$$100\cdot\beta_1 = \frac{\frac{dwage}{wage}\cdot 100}{dFemale}$$

Altså gir $100\cdot\beta_1$ %-vis endring i wage når topplederen er kvinne og ikke mann

V Estimert modell sier altså at kvinner har 44% lavere lønn enn menn!

V NB! Dette er en tilnærming! Eksakt %-vis forskjell er gitt ved $100(e^{\hat{\beta}_1}-1) = -35,6$

Den gule kopien beholder du

Tapir Uttrykk 04/09

NTNU

Kandidatnr.

Dato: 02.06.09 Side: 12

Fagnr/emnekode SOK 3001 Antall ark: 22

Denne kolonnen er forbeholdt sensor

b) La meg først få forklare t-test generelt, da denne vil gå igjen i oppgaven

Forutsetninger

MLR1 : multippel lineær populasjonsmodell

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

Har altså k forklaringsvariable

MLR2 : tilfeldig utvalg, dvs $Cov(u_i, u_j) = 0$ for $i \neq j$

Får da empirisk modell

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + u_i$$

MLR3: Variasjon i alle forklaringsvariable, og ikke eksakt multikollinaritet (dvs at ingen av forklaringsvariablene må være en lineær kombinasjon av de andre)

MLR4 : $E(u \mid x_1, x_2, \dots x_k) = 0$

Forventet restledd betinget på forkl variablene skal være null

$E(\hat{\beta}_j \mid x_1, x_2 \dots x_k) = \beta_j$

Dersom MLR1 - MLR4 holder er MKM-estimatorene forventningsrette (jf oppg 1)

MLR5 : Homoskedastisitet, dvs konstant restleddsvarians

$$Var(u \mid x_1, x_2, \dots, x_k) = \sigma^2$$

kan også skrives : $Var(y \mid x_1, x_2, \dots, x_k) = \sigma^2$

Dersom MLR1 - MLR5 holder er MKM-estimatorene BLUE (best linear unbiased estimator), dvs at de er estimatorene med minst varians når vi også krever forventningsretthet

Vi får her at varians til estimator:

$$Var(\hat{\beta}_j \mid x_1, x_2, \dots, x_k) = \frac{\sigma^2}{SST_j(1-R_j^2)} \qquad j = 1, 2 \dots k$$

der $SST_j = \sum (x_{ij} - \bar{x}_j)^2$

og R_j^2 kommer fra regresjonen

$$x_{ij} = \alpha_0 + \alpha_1 x_{i1} + \dots + \alpha_{j-1} x_{ij-1} + \alpha_{j+1} x_{ij+1} + \dots + \alpha_k x_{ik} + \text{støy}_i$$

Standardavvik: $sd(\hat{\beta}_j) = \sqrt{\dfrac{\sigma^2}{SST_j(1-R_j^2)}}$

Restleddsvariansen, σ^2, vil være ukjent parameter, og vi har forventningsrett estimator ved

$$\hat{\sigma}^2 = \frac{SSR}{n-k-1} \quad , \quad E(\hat{\sigma}^2) = \sigma^2$$

Slik at estimert varians på β_1-estimatet ved MKM er

$$\widehat{Var}(\hat{\beta}_1 | x_1, \ldots x_k) = \frac{\hat{\sigma}^2}{SST_1(1-R_1^2)}$$

Og får estimert standardavvik (=standardfeil)

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{SST_1(1-R_1^2)}}$$

MLR6: normalfordelt restledd, $u \sim N(0, \sigma^2)$

Under MLR1 - MLR6 er

$$\hat{\beta}_i \sim N(\beta_i, Var(\hat{\beta}_i)) \quad \text{normalfordelt}$$

$$\frac{\hat{\beta}_i - \beta_i}{sd(\hat{\beta}_i)} \sim N(0,1) \quad \text{standard normalfordelt}$$

og da med estimert standardavvik

$$\frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} \sim t_{n-k-1}$$ t-fordelt med $n-k-1$ frihetsgrader (df)

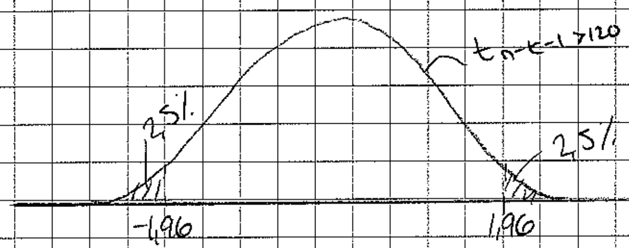
Har nullhypotesen: $H_0: \beta_i = 0$, mot alternativhypotesen $H_1: \beta_i \neq 0$

Hvis H_0 er sann, og vi har $n > 120$, får vi at

$\frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$ er t-fordelt med > 120 df.

Dette betyr at det er 5% sannsynlighet for at verdien på variabelen er større enn 1,96 eller mindre enn $-1{,}96$.

Grafisk



Her er kritisk verdi, $c = 1{,}96$, og signifikansnivå 5%.

Dersom $t^{obs} = \hat{\beta}_j / se(\hat{\beta}_j)$ er større enn 1,96 eller mindre enn -1,96 kan vi forkaste H_0 med 5% sannsynlighet for at vi forkaster riktig hypotese (dvs type 1 feil).
Altså forkaster vi H_0 dersom
$|t^{obs}| > c$ for gitt signifikansnivå, α,
og vi sier isåfall at β_j er signifikant forskjellig fra null på α%-nivå, da det er uten sannsynlig å observere en slik t^{obs} dersom H_0 er sann.

For $n > 120$ har vi at

$c = 2{,}56$ for $\alpha = 1\%$ (antar, siden ikke har fordeling tilgjengelig, og ikke husker helt!)

I oppgaven er $\hat{\beta}_1 = -0{,}44$, $se(\hat{\beta}_1) = 0{,}05$

Vi skal undersøke om kvinnelige toppledere diskrimineres, dvs

$H_0: \beta_1 = 0$ ingen diskriminering
$H_1: \beta_1 < 0$ diskriminering
↳ ensidig test

✓ $t^{obs} = \frac{-0{,}44}{0{,}05} = -8{,}8$

For ensidig test er $c = 1{,}96$ på 2,5%-nivå

Har at $t^{obs} < -c$, for alle meningsfulle signifikansnivå.
Gitt at modell (1) er sann modell tilsier altså resultatet at kvinner klart diskrimineres.

✓ Her er det midlertid rimelig å anta at ligning (1) utelater relevante forklaringsvariable som er korrelerte med female, og som vi bør kontrollere

NTNU

Kandidatnr.

Dato: 02.06.09 Side: 15

Fagnr/emnekode SØK 3001 Antall ark: 22

Denne kolonnen er forbeholdt sensor

for år å unngå utelatt variabelskjevhet i β_1-estimatet (jf oppg 1). Altså er det sannsynlig at $E(u|female) \neq 0$

c)

(2) $\ln(wage)_i = \beta_0 + \beta_1 female_i + \beta_2 \ln(marketvalue)_i + \beta_3 return_i + v_i$

Estimert

$\ln(wage)_i = \underset{(0{,}03)}{3{,}86} - \underset{(0{,}04)}{0{,}28} female_i - \underset{(0{,}004)}{0{,}37} \ln(mark\ value)_i + \underset{(0{,}003)}{0{,}04} return_i$

Ser nå at $\hat{\beta}_1$ faller til $-0{,}28$.
Dette skyldes nok at vi nå har inkludert variable
✓ som var korrelert med female, og dermed
har rettet opp (iallefall noe av) skjevheten i b).
Ser spesielt at

$$t^{obs}_{\hat{\beta}_2} = \frac{-0{,}37}{0{,}004} = -9{,}25$$

som viser at marketvalue er relevant forklaringsvariabel for wage (β_2 er klart statistisk signifikant forskjellig fra null; $|t^{obs}_{\beta_2}| > c$ for alle meningsfulle α).

(Siden $t^{obs}_{\hat{\beta}_3} = \frac{0{,}04}{0{,}03} = 1{,}33$, er ikke β_3 statistisk signifikant forskjellig fra null for noe tilfredsstillende signifikansnivå, og jeg velger å ikke diskutere denne nærmere - ser ikke ut til å være relevant)

Siden estimatet på β_1 faller ved inkludering av den relevante variabelen "marketvalue", og marketvalue har en signifikant negativ effekt på wage, er det utfra diskusjonen i oppg 1 rimelig å anta at female og marketvalue er positivt korrelert, slik at $\hat{\beta}_1$ underestimeres når marketvalue utelates.

d) Skal tolke β_2

Differensierer (2) mhp wage og marketvalue (mv)

$$\frac{1}{wage}\, dwage = \beta_2 \frac{1}{mv}\, dmv$$

$$\beta_2 = \frac{\frac{dwage}{wage}}{\frac{dmv}{mv}} = \frac{\frac{dwage}{wage} \cdot 100}{\frac{dmv}{mv} \cdot 100}$$

Den gule kopien beholder du

Altså ser vi at β_2 gir %-vis endring i toppledernes kvinneandel for 1% økning i markedsverdi, med andre ord elastisiteten.

Med $\hat{\beta}_2 = -0,37$ har vi at 1% økning i markedsverdien, reduserer topplederkvinna med 0.37%

✓

e) Dersom store bedrifter er ment i betydningen høy markedsverdi, tyder resultatene her på at store bedrifter har større sannsynlighet for å ha kvinnelige toppledere. Dette ut fra diskusjonen i c) der vi fant at korrelasjonen mellom female og markedsverdi er positiv.

Oppgave 3

Panelanalyse
48 stater i USA, perioden 82-88 (7 år)
Data: 336 observasjoner, $n \cdot T = 48 \cdot 7$

DEATH - trafikkdødsfall per 10 000 innbyggere

BEERTAX - ølskatt målt i dollar pr kasse øl

UNEMP - arbeidsledighetsrate i %

REALINC - realinntekt pr innbygger

$\overline{DEATH} = 2$

$\overline{BEERTAX} = 0,50$

a)
$$DEATH_{it} = \underset{(0,05)}{0,36}\, BEERTAX_{it} \qquad \begin{matrix} i = 1,2,\dots,n \\ t = 1,2,\dots,T \end{matrix}$$

Altså har vi her et estimat, $\hat{\beta}_1 = 0,36$, som sier at dersom ølskatt pr kasse øl øker med 1 dollar, vil antall dødsfall per 10 000 innbyggere øke med 0,36.
Sammenhengen er sterk statistisk signifikant ($t^{obs}_{\hat{\beta}_1} = 7,2$)

Dette er kontraintuitivt. En skulle tro at økt ølskatt skulle gi lavere trafikkdødsfall (gjennom lavere ølkonsum → mindre fyllekjøring → færre ulykker).

Igjen lukter det utelatt variabelskjevhet.
Dette kan være både statspesifikke variable (varierer mellom stater, men ikke over tid), og tidsspesifikke variable (varierer over tid, men ikke mellom stater, gjerne makrovariable for USA som helhet).

Vi kan kontrollere for disse i en slengvei ved å inkludere tidsdummyer og statspesifikke dummyer - eller ekvivalent: fixed effects konstanter for hver stat.

NTNU

Kandidatnr.

Dato: 02.06.09 Side: 18

Fagnr/emnekode Sok 3001 Antall ark: 22

Denne kolonnen er forbeholdt sensor

Dette er gjort, og når vi utelater konstantene får vi (og vi har inkludert også unemp og realinc til kontroll)

$$DEATH_{it} = -0,45\,BEERTAX_{it} - 0,063\,UNEMP_{it} + 1,81\ln(REALINC)_{it}$$
$$\quad(0,22) \qquad\qquad (0,012) \qquad\qquad (0,47)$$

b) Nå er effekten av BEERTAX : $\hat{\beta}_1 = -0,45$

Resultatet sier altså at en økning av ølskatten per ølkasse med 1 dollar vil redusere antall dødsfall per 10 000 innbygger med 0,45. Dermed må vi ha at en økning av ølskatten med 0,50 $ per kasse må redusert antall trafikkdødsfall pr 10 000 innbyggere med $0,5 \cdot 0,45 = 0,225$. (Merk: alt annet likt, i gjennomsnitt)

Har at

$$t^{obs}_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-k-1}$$

($n-k-1 > 120$ her!)

Da har vi (i oppg 2b) sett at

$$P\left(-1,96 \le \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \le 1,96\right) = 95\%$$

omformet

$$P\left[\hat{\beta}_1 - 1,96\cdot se(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + 1,96\,se(\hat{\beta}_1)\right] = 0,95$$

Slik at et 95%-konfidensintervall for effekten av ølskatt på trafikkdødsfall er

$$[-0,45 - 1,96\ 0,22 \quad , \quad -0,45 + 1,96\cdot 0,22]$$

$$= \underline{\underline{[-0,8812\ ,\ -0,0188]}}$$

Merk at dette ikke betyr at det er 95% ssh. for at β_1 ligger i dette intervallet, men at i 19 av 20 slike analyser vil sann effekt, β_1, ligge i konfidensintervallet

NTNU

Kandidatnr.

Dato: 02.06.09 Side: 19

Fagnr/emnekode SØK 3001

Antall ark: 22

Denne kolonnen er forbeholdt sensor

c) New Jersey har 8,1 millioner innbyggere
Har ølskatt på 18 per kasse i 1988

Ølskatten vil redde 0,45 pr 10 000 innbygger

Dermed vil ølskatten redde

$$\frac{8\,100\,000}{10\,000} \cdot 0{,}45 = 364{,}5$$

mennesker i staten i 1988

For å finne 95% konfidensintervall her, bruker vi resultatene fra b)

this $\beta_1 = -0{,}8812$ reddes.

$$\frac{8\,100\,000}{10\,000} \cdot 0{,}8812 \approx 714 \quad \text{mennesker}$$

this $\beta_1 = -0{,}0188$ reddes

$$\frac{8\,100\,000}{10\,000} \cdot 0{,}0188 = 15{,}2 \quad \text{mennesker}$$

95% konfidensintervallet her blir altså

$$[15{,}2 \ , \ 714]$$

d) $t^{obs}_{\hat{\beta}_2} = \frac{-0{,}063}{0{,}012} = -5{,}25$ $|t^{obs}| > 1{,}96$

Arbeidsledighetsraten har en statistisk signifikant (negativ) effekt på trafikkdødsfall.
Nærmere bestemt viser resultatene at 1%-poengs økning i arbeidsledighetsraten reduserer antall dødsfall pr 10 000 innbyggere med 0,063.
(alt annet likt)

$t^{obs}_{\hat{\beta}_3} = \frac{1{,}81}{0{,}47} = 3{,}85 > 1{,}96$

→ også ln Realinc har en statistisk signifikant effekt på trafikkdødsfall, er positiv.

Modellen differensiert mhp DEATH og REALINC gir

$$d\,DEATH = \beta_3 \frac{1}{REALINC} \cdot dREALINC$$

Denne kolonnen er forbeholdt sensor

$$\beta_3 = \frac{dDEATH}{\frac{dREALINC}{REALINC}}$$

Slik at

$$\frac{\beta_3}{100} = \frac{dDEATH}{\frac{dREALINC}{REALINC} \cdot 100}$$

Dermed viser $\beta_3/100$ endringen i DEATH for 1% økning i realinc.
Når realinntekten pr innbygger øker med 1% øker dødsfall pr 10 000 innbyggere med 0,0181.
Dette kan f.eks ha å gjøre med at alkoholkonsum og trafikk øker med økt inntekt

alt annet likt

e) Realinntekten i New Jersey øker med 1% neste år. Skal anne prediksjon av endring i trafikkdødsfall i staten neste år

1% økning i realinntekt vet vi øker dødsfall pr 10 000 innbyggere med 0,0181

Da økes antall trafikkdødsfall i NJ:

$$\frac{8\ 100\ 000}{10\ 000} \cdot 0{,}0181 = \underline{\underline{14{,}66}}$$

Finner først 90% -konfidensintervall for β_3.
Antar (har ikke tabell) at $c = 1{,}645$ for $\alpha = 10\%$
Altså har vi

$$P\left(-1{,}645 < \frac{\hat{\beta}_3 - \beta_3}{Se(\hat{\beta}_3)} < 1{,}645\right) = 90\%$$

Og vi får på samme måte som tidligere 90% konfidensintervall som

$$\left[\hat{\beta}_3 - 1{,}645 \cdot Se(\hat{\beta}_3)\ ,\ \hat{\beta}_3 + 1{,}645 \cdot Se(\hat{\beta}_3)\right]$$

$$= \left[1{,}03685\ ,\quad 2{,}58315\right]$$

Denne kolonnen er forbeholdt sensor

Hvis $\beta_3 = 1,03685$, øker trafikkdødsfall med

$$\frac{8\,100\,000}{10\,000} \cdot \frac{1,03685}{100} = 8,398$$

Hvis $\beta_3 = 2,58315$, øker trafikkdødsfall med

$$\frac{8\,100\,000}{10\,000} \cdot \frac{2,58315}{100} = 20,924$$

Altså får vi 90% konfidensintervall for prediksjonen:

$$\underline{[8,398 \ , \ 20,924]}$$

f) Hypotese: arbeidsledighet har større effekt på trafikkdødsfall i de vestlige statene i USA.

Definerer (statspesifikk) dummyvariabel, V

$$V_i = \begin{cases} 1 & \text{hvis staten er av de vestlige statene} \\ 0 & \text{hvis østlig} \end{cases}$$

Jeg ville så inkludert et interaksjonsledd mellom denne variabelen og arbeidsledigheten i modellen, altså:
(for ikke med konstantledd var vel...)

$$DEATH_{it} = \beta_1 BEERTAX_{it} + \beta_2 UNEMP_{it} + \beta_3 \ln(REALINC_{it}) + \delta_1 V_i UNEMP_{it} + støy_i$$

Effekten av UNEMP på DEATH er nå

$$\beta_2 + \delta_1 V_i$$

- som altså avhenger om staten ligger på vest eller østsida.

Test hvorvidt effekten av arbeidsledighet er større for vestlige stater blir da å teste nullhypotesen

$H_0: \delta_1 = 0$ mot alternativ: $H_1: \delta_1 < 0$
(husk at $\beta_2 < 0$)

Denne kolonnen er forbeholdt sensor

Testen blir altså å estimere modellen vha vanlig MKM. Deretter testes H_0 ved vanlig t-test. Dersom $t_{\frac{\alpha}{2}}^{obs} < -c$ for valgt α forkastes H_0, og vi har bevist at i vestlige stater er det større effekt av ølavgiften på trafikkdødsfall

g) Snøforholds påvirkning på trafikkdødsfall.

1) $\widehat{DEATH}_{it} = \hat{\beta}_1 BEERTAX_{it} + \hat{\beta}_2 UNEMP_{it} + \hat{\beta}_3 \ln(REALINC)_{it}$

$+ \hat{\beta}_4 AVERAGESNOW_i$

hvor AVERAGESNOW er gjennomsnittlig (over tid) snøfall i hver stat $i = 1, \ldots, 48$.
Dette er altså en statspesifikk variabel.
Men vi må huske at i den utvidede modellen har vi inkludert fixed effects konstanter for hver stat. AVERAGESNOW er en variabel som inkluderes i konstanten (jf individspesifikt konstantledd, jf oppg 1). Når vi har inkludert fixed effect konstanter kan vi ikke ha AVERAGESNOW som forklaringsvariabel, for da vil vi få perfekt multikollinaritet mellom forklaringsvariable. Husker MLR3 forbød dette.

Forskeren kan altså ikke kjøre regresjon med AVERAGESNOW som forklaringsvariabel i denne modellen, og har derfor ingen nytte av å samle inn slike data

2) $\widehat{DEATH}_{it} = \hat{\beta}_1 BEERTAX_{it} + \hat{\beta}_2 UNEMP_{it} + \hat{\beta}_3 \ln(REALINC)_{it}$

$+ \hat{\beta}_5 SNOW_{it}$

Her har variabelen både variasjon både mellom stater og over tid. Sistnevnte gjør at vi ikke får perfekt multikollinaritet som over, og MKM på denne vil fungere!

Forskeren bør dermed velge denne fremgangsmåten for å analysere snøforholdene.