NTNU
Det skapende universitet

Institutt for samfunnsøkonomi

Eksamensoppgave i SØK3001 – Økonometri I / Econometrics I

Faglig kontakt under eksamen: Jørn Rattsø

Tlf.: 73 59 19 34

Eksamensdato:	29. mai 2013
Eksamenstid:	5 timer
Sensurdato:	19. juni 2013
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, HP 30S eller SR-270X College
Annen informasjon:	-

Målform/språk:	Bokmål, nynorsk og engelsk
Antall sider:	10 (inkl. forside)
Antall sider vedlegg:	2

Merk! Studenter finner sensur i Studentweb. Har du spørsmål om din sensur må du kontakte instituttet ditt. Eksamenskontoret vil ikke kunne svare på slike spørsmål.

Bokmål

Alle spørsmål skal besvares.
To tabeller er vedlagt.

Oppgave 1

Den nedenstående regresjonsmodell er estimert med data for individer fra folketellingen i USA i år 2000 (805 observasjoner)

$$lWage_i = \beta_0 + \beta_1\, Exp_i + \beta_2\, Exp_i^2 + \beta_3 Edyrs_i + \beta_4\, Part_i + \beta_5\, Female_i + \beta_6\, Black_i + u_i$$

hvo lWage er log timelønn, Exp er arbeidsmarkedserfaring i år, Exp^2 er erfaring opphøyd i andre, Edyrs er antall år utdanning, Part er en dummy variable lik 1 hvis individet arbeider deltid, Female er en dummy variabel lik 1 hvis individet er kvinne, Black er en dummy variabel som er lik 1 hvis individet er sort; alle variable er målt for individ j.

Regresjonsestimater og noe statistikk ('R i andre', Korrigert 'R i andre', og summen av kvadrerte avvik SSR) er vist i tabellen under:

lWage	Koeffisient	St. avvik
Exp	0.037	0.0063
Exp^2	-0.00049	0.00015
Edyrs	0.091	0.0095
Part	0.0030	0.0581
Female	-0.225	0.0427
Black	-0.170	0.0590
Const	0.923	
'R i andre'	0.2308	
Korr. 'R i andre'	0.2251	
SSR	282.584	

a) Tolk koeffisientene, kommenter deres fortegn, størrelse og signifikans. Hva er de viktige faktorene som forklarer lønnsforskjeller gitt denne analysen?
b) Etter hvor mange år i arbeidsmarkedet begynner individets lønn å falle?
c) Vis hvordan hypotesen om at koeffisienten for Female er lik -0.1 kan testes og gjør testen.
d) Tillegg av både kubisk og kvartisk ledd (tredje ordens og fjerde ordens) i Exp fører til en reduksjon i summen av kvadrerte avvik til 281.54. Vis hvordan hypotesen om at koeffisientene til disse tilleggsvariable samtidig er lik null kan testes og gjør testen.
e) Forklar hva som måles med 'R i andre' og forskjellen mellom 'R i andre' og korrigert 'R i andre'.
f) En regresjon med sum av kvadrerte avvik fra modellen over med de samme forklaringsvariablene gir en 'R i andre' på 0.0014. Diskuter hvordan du kan bruke denne informasjonen til å teste for heteroskedastisitet.

Oppgave 2

Britiske forskere prøver å estimere effekten av migrasjon på arbeidsmarkedet i Storbritannia. De har bare gode regionale data for 1999 og estimerer den følgende modellen med minste kvadraters metode:

$$lWage_i = \beta_0 + \beta_1 \, MigPop_i + \beta_2 \, X_i + u_{it}$$

Her er $lWage_i$ den gjennomsnittlige log av reallønnen til arbeidere født i Storbritannia som lever i region i, $MigPop_i$ er prosentandelen av innflyttere i befolkningen i region i, X_i representerer et sett av kontrollvariable i region i. De finner et estimat for koeffisienten for MigPop på 0.88 med standardavvik på 0.21.

a) Gi en tolkning av koeffisienten. Tror du dette estimat med bruk av minste kvadraters metode gir et forventningsrett estimat for virkningen av migrasjon på lønninger?
b) Forskerne argumenterer at nye innflyttere delvis baserer beslutningen om hvor de skal bo i Storbritannia på eksisterende beholdning av tilsvarende innflyttere. De bruker beholdningen av innflyttere for 10 år siden som instrument for MigPop. Hvilke forutsetninger må dette instrumentet tilfredsstille? Kan du tenke deg noen grunner for at det foreslåtte instrumentet ikke tilfredsstiller disse forutsetningene.
c) Forklar hvordan forskerne kan gå fram for å finne et instrumentvariabel (IV) estimat for effekten av MigPop.
d) Med bruk av standard IV estimeringsmetode finner de en koeffisient for MigPop på 0.94 med standardavvik på 0.58. Kommenter deres funn.
e) Diskuter hvordan forskerne kunne gå fram for å teste bruken av instrumentet.

Oppgave 3

Finansdepartementet I Storbritannia har vært bekymret for at økningen i merverdiavgift (MVA) fra 17,5% til 20% den 1. Januar 2011 vil redusere samlede salg. Økonomer klarte å overbevise departementet om at økningen burde innføres i noen regioner og ikke andre. Ti regioner er unntatt for økningen i en periode på 3 måneder. Du samler inn data for de 10 regionene som er unntatt og ytterligere 10 tilsvarende regioner som har hatt økning i merverdiavgift. Den følgende tabellen viser gjennomsnittlig detaljsalg per butikk (i millioner pund) for 3 måneder før og etter økningen i MVA:

	Økning før MVA-økning	Økning etter MVA-økning
Regioner med økning	4.77	4.15
Regioner med unntak	4.34	4.21

a) Kalkuler 'difference in difference' estimat av effekten av MVA økning på detaljsalget.
b) Diskuter nødvendige forutsetninger for å tolke 'difference in difference' estimatet som en årsakseffekt av MVA på økningen i salg.

Du samler nå inn data fra alle detaljbutikker i de 10 behandlingsregionene (med MVA økning) og de 10 kontrollregionene (med unntak) om salget de tre siste månedene i 2009 og de 3 første månedene i 2010 (alle før økningen i MVA).

Du estimerer følgende regresjon basert på individ-data for de to tidsperiodene:

$$Sales_{it} = \beta_0 + \beta_1\, Treat_i + \beta_2\, Time_t + \alpha_{DiD}(Treat_i x\, Time_t) + u_{it}$$

Her er Sales detaljsalg fra butikk i tidsperiode t, Treat er en dummy variabel lik 1 hvis butikken er i en region med MVA økning og null hvis den er i en region med unntak, Time er en dummy variabel som er lik null i de siste 3 månedene i 2009 og 1 for de første 3 månedene i 2010.

c) Du estimerer en α_{DiD} lik -0.1 med standardavvik 0.02. Hva betyr dette for din 'difference in difference' estimering. Tror du estimatene i a) over- eller under-estimerer effekten av MVA på salg?

Ved bruk av data i dine behandlingsregioner (med MVA økning), estimerer du den følgende regresjon for endring i salget i butikk i over perioden med MVA økning:

$$\Delta Sales_i = \beta_0 + \beta_1 Distance_i + u_i$$

Her måler Distance avstanden fra butikk i til den nærmeste kontrollregion (region som har unntak for MVA økning). Du finner en estimert koeffisient for Distance på 0.011 med standardavvik på 0.0012.

d) Hva betyr dette for din 'difference in difference' estimering. Tror du estimatene i a) over- eller under-estimerer effekten av MVA på salg?

Nynorsk

Alle spørsmål skal besvarast.
To tabellar er vedlagt.

Oppgåve 1

Regresjonsmodellen nedanfor er estimert med data for individ fra folketellinga i USA i år 2000 (805 observasjonar)

$$lWage_i = \beta_0 + \beta_1\ Exp_i + \beta_2\ Exp_i^2 + \beta_3 Edyrs_i + \beta_4\ Part_i + \beta_5\ Female_i + \beta_6\ Black_i + u_i$$

kor lWage er log timelønn, Exp er arbeidsmarknadserfaring i år, Exp^2 er erfaring opphøgd i andre, Edyrs er antall år utdanning, Part er en dummy variable lik 1 hvis individet arbeider deltid, Female er en dummy variabel lik 1 hvis individet er kvinne, Black er en dummy variabel som er lik 1 hvis individet er sort; alle variable er målt for individ j.

Regresjonsestimater og noe statistikk ('R i andre', Korrigert 'R i andre', og summen av kvadrerte avvik SSR) er vist i tabellen under:

lWage	Koeffisient	St. avvik
Exp	0.037	0.0063
Exp^2	-0.00049	0.00015
Edyrs	0.091	0.0095
Part	0.0030	0.0581
Female	-0.225	0.0427
Black	-0.170	0.0590
Const	0.923	
'R i andre'	0.2308	
Korr. 'R i andre'	0.2251	
SSR	282.584	

a) Tolk koeffisientane, kommenter deira fortegn, størrelse og signifikans. Kva er dei viktige faktorane som forklarer lønnsforskjellar gitt denne analysen?
b) Etter kor mange år i arbeidsmarknaden begynner individets lønn å falle?
c) Vis korleis hypotesa om at koeffisienten for Female er lik -0.1 kan testes og gjør testen.
d) Tillegg av både kubisk og kvartisk ledd (tredje ordens og fjerde ordens) i Exp fører til en reduksjon i summen av kvadrerte avvik til 281.54. Vis korleis hypotesa om at koeffisientane til disse tilleggsvariable samtidig er lik null kan testes og gjør testen.
e) Forklar kva som måles med 'R i andre' og forskjellen mellom 'R i andre' og korrigert 'R i andre'.
f) Ein regresjon med sum av kvadrerte avvik fra modellen over med dei same forklaringsvariablane gir en 'R i andre ' på 0.0014. Diskuter korleis du kan bruke denne informasjonen til å teste for heteroskedastisitet.

Oppgåve 2

Britiske forskarar prøver å estimere effekten av migrasjon på arbeidsmarknaden i Storbritannia. Dei har bare gode regionale data for 1999 og estimerer den fylgjande modellen med minste kvadraters metode:

$$lWage_i = \beta_0 + \beta_1\, MigPop_i + \beta_2\, X_i + u_{it}$$

Her er $lWage_i$ den gjennomsnittlige log av reallønnen til arbeidere født i Storbritannia som lever i region i, $MigPop_i$ er prosentandelen av innflyttere i befolkningen i region i, X_i representerer et sett av kontrollvariable i region i. De finner et estimat for koeffisienten for MigPop på 0.88 med standardavvik på 0.21.

a) Gje ei tolking av koeffisienten. Tror du dette estimat med bruk av minste kvadraters metode gir et forventningsrett estimat for verknaden av migrasjon på lønninger?
b) Forskerane argumenterer at nye innflyttarar delvis baserer beslutningen om kor dei skal bo i Storbritannia på eksisterande beholdning av tilsvarende innflyttarar. Dei bruker beholdningen av innflyttarar for 10 år sidan som instrument for MigPop. Hvilke forutsetningar må dette instrumentet tilfredstille? Kan du tenke deg noen grunner for at det foreslåtte instrumentet ikke tilfredstillar desse forutsetningane.
c) Forklar korleis forskerane kan gå fram for å finne eit instrumentvariabel (IV) estimat for effekten av MigPop.
d) Med bruk av standard IV estimeringsmetode finner dei ein koeffisient for MigPop på 0.94 med standardavvik på 0.58. Kommenter funnet deira.
e) Diskuter korleis forskerane kunne gå fram for å teste bruken av instrumentet.

Oppgåve 3

Finansdepartementet i Storbritannia har vært bekymra for at auken i merverdiavgift (MVA) fra 17,5% til 20% den 1. Januar 2011 vil redusere samlede salg. Økonomer klarte å overbevise departementet om at auken burde innføres i noen regioner og ikke andre. Ti regioner er unntatt for auken i en periode på 3 månader. Du samler inn data for de 10 regionene som er unntatt og ytterligere 10 tilsvarande regioner som har hatt auke i merverdiavgift. Den fylgjande tabellen viser gjennomsnittlig detaljsalg per butikk (i millioner pund) for 3 måneder før og etter auken i MVA:

	Auking før MVA-auke	Auking etter MVA-auke
Regioner med auking	4.77	4.15
Regioner med unntak	4.34	4.21

a) Kalkuler 'difference in difference' estimat av effekten av MVA auking på detaljsalget.
b) Diskuter nødvendige forutsetningar for å tolke 'difference in difference' estimatet som en årsakseffekt av MVA på auken i salg.

Du samler nå inn data fra alle detaljbutikker i de 10 behandlingsregionane (med MVA auke) og de 10 kontrollregionene (med unntak) om salget de tre siste månadene i 2009 og de 3 første månadene i 2010 (alle før auken i MVA).

Du estimerer fylgjande regresjon basert på individ-data for de to tidsperiodane:

$$Sales_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 Time_t + \alpha_{DiD}(Treat_i x\, Time_t) + u_{it}$$

Her er Sales detaljsalg fra butikk i tidsperiode t, Treat er en dummy variabel lik 1 hvis butikken er i en region med MVA auke og null hvis den er i en region med unntak, Time er en dummy variabel som er lik null i dei siste 3 månadene i 2009 og 1 for dei første 3 månedene i 2010.

c) Du estimerer en α_{DiD} lik -0.1 med standardavvik 0.02. Kva betyr dette for din 'difference in difference' estimering. Tror du estimatane i a) over- eller under-estimerer effekten av MVA på salg?

Ved bruk av data i dine behandlingsregioner (med MVA auke), estimerer du den fylgjande regresjon for endring i salget i butikk i over perioden med MVA auke:

$$\Delta Sales_i = \beta_0 + \beta_1 Distance_i + u_i$$

Her måler Distance avstanden fra butikk i til den nærmeste kontrollregion (region som har unntak for MVA auke). Du finner en estimert koeffisient for Distance på 0.011 med standardavvik på 0.0012.

d) Kva betyr dette for din 'difference in difference' estimering. Tror du estimatane i a) over- eller under-estimerer effekten av MVA på salg?

English

All questions to be answered.
Two tables enclosed.

Question 1

The following regression is estimated on individuals from the US Census in the year 2000 (805 observations)

$$lWage_i = \beta_0 + \beta_1 \, Exp_i + \beta_2 \, Exp_i^2 + \beta_3 Edyrs_i + \beta_4 \, Part_i + \beta_5 \, Female_i + \beta_6 \, Black_i + u_i$$

where lWage is the log of hourly wage, Exp is the labour market experience in years, Exp^2 is experience squared, Edyrs is the number of years in education, Part is a dummy variable equal to 1 if the individual works part time, Female is a dummy variable equal to 1 if the individual is female, Black is a dummy variable which is equal to 1 if individual is black; all variables measured for individual i.

The regression estimates and some statistics (R-squared, Adjusted R-squared and sum of squared residuals SSR) are shown in the table below:

lWage	Coefficient	St.error
Exp	0.037	0.0063
Exp^2	-0.00049	0.00015
Edyrs	0.091	0.0095
Part	0.0030	0.0581
Female	-0.225	0.0427
Black	-0.170	0.0590
Const	0.923	
R-squared	0.2308	
Adj. R-squared	0.2251	
SSR	282.584	

a) Interpret the coefficients, comment their sign, magnitude and significance. What are the major factors explaining wage differences based on this evidence?
b) After how many years in the labour market does an individual's wage begin to decline?
c) Show how to test the hypothesis that the coefficient on Female is equal to -0.1 and do the test.
d) Adding both a cubic and a quartic term (third order and fourth order) in Exp leads to a reduction in the sum of squared residuals to 281.54. Show how to test the hypothesis that the coefficients of these additional variables are jointly equal to zero and do the test.
e) Explain what is measured by R-squared and the difference between R-squared and Adjusted R-squared.
f) A regression of the squared errors from the regression above on the same variables yields an R-squared value of 0.0014. Discuss how you can use this information to test for heteroskedasticity.

Question 2

British researchers attempt at estimating the impact of migration on the UK labour market. They only have good regional data for 1999 and estimate the following model using OLS:

$$lWage_i = \beta_0 + \beta_1\, MigPop_i + \beta_2\, X_i + u_{it}$$

Here $lWage_i$ is the mean log real wage of british born workers who live in region i, $MigPop_i$ is the percent of migrants in the population in region i, X_i represents a set of control variable in region i. They find an estimate of the coefficient on MigPop of 0.88 with a standard error of 0.21.

a) Provide an interpretation of this coefficient. Do you think that this OLS estimate gives an un-biased estimate of the impact of migration on wages?
b) The researchers argue that new migrants partly base their decision on where to reside in the UK on the current stock of similar migrants. They use the stock of migrants ten years ago in the region as an instrument for MigPop. What conditions do this instrument need to satisfy? Can you think of any reasons why the proposed instrument does not satisfy these conditions?
c) Explain how the researchers may proceed to obtain an instrument variable (IV) estimate of the effect of MigPop.
d) Using standard IV estimation methods they estimate the coefficient of MigPop to 0.94 with a standard error of 0.58. Comment on their findings.
e) Discuss how the researchers could proceed to test the use of the instrument.

Question 3

The Ministry of Finance in the UK has been worried that the increase in the value added tax (VAT) from 17,5% to 20% January 1 2011 will reduce aggregate sales. Economists managed to convince the Ministry to introduce the rise in some regions and not others. Ten counties of the UK are exempt from the increase for a 3 month period. You collect data on sales for the exempt 10 counties and for another 10 similar counties who are subject to the VAT rise. The following table shows the average retail sales per outlet (in million pounds) for the 3 months before and after the VAT increase:

	Pre-VAT increase	Post-VAT increase
VAT rise counties	4.77	4.15
Exempt counties	4.34	4.21

a) Calculate the difference in difference estimate of the impact of the VAT increase on retail sales.
b) Discuss the assumptions needed to interpret this difference in difference estimate as the causal effect of the VAT rise on sales.

You now collect data from all individual retail outlets in the 10 treatment (VAT rise) and 10 control (exempt) counties on sales for the last 3 months of 2009 and the first 3 months of 2010 (all before the VAT increase).

You estimate the following regression based on the individual data for the two time periods:

$$Sales_{it} = \beta_0 + \beta_1\, Treat_i + \beta_2\, Time_t + \alpha_{DiD}(Treat_i x\, Time_t) + u_{it}$$

Here Sales are retail sales from outlet i in time period t, Treat is a dummy variable equal to 1 if the outlet is in a VAT rise county and zero if it is in a VAT exempt country, Time is a dummy variable that is equal to zero for the last 3 months of 2009 and 1 for the first 3 months of 2010.

c) You estimate an α_{DiD} equal to -0.1 with a standard error of 0.02. What does this mean for your difference in difference estimation. Do you think the estimates in a) over- or under-estimates the effect of VAT on sales?

Using data in your treated (VAT rise) counties, you estimate the following regression on the change of retail sales in outlet i over the period of the VAT rise:

$$\Delta Sales_i = \beta_0 + \beta_1 Distance_i + u_i$$

Here Distance measures the distance from the outlet i from the nearest control county (county that is exempt from the VAT increase). You find an estimated coefficient of Distance of 0.011 with a standard error of 0.0012.

d) What does this mean for your difference in difference estimation? Do you think that the estimates in part a) over- or under-estimate the impact of VAT on sales?

Søk 3001, kandidat 10083, vår 2013

Slik får man A, svarer på spørsmålene, svarene er rett, ryddig og klar framstilling, gode tolkinger.

Oppgave 1 var 'planken' og skulle vise beherskelse av grunnleggende forståelse av estimater og tester. Etter litt lang innledning kommer presise og riktige tolkinger av estimatene i a (mange hadde problemer med prosent-effekter), b er rett tolking av annengradsledd for erfaring, c er rett t-test mot bestemt verdi, d er rett F-test av flere restriksjoner, e gir rett forståelse av R2adj, og f forkaster ikke homoskedastisitet med rett F-test. Ikke mye å klage på.

Oppgave 2 om flytting var en plage for mange. I oppgave a kunne oppgaveteksten vært tydeligere, og vi har godtatt både 0.88% og 88% effekt av 1% høyere flytting. 0.88% høres mest realistisk ut. Kunne kanskje gitt mer diskusjon av faktorer som gir mulig skjevhet: Kausaliteten går motsatt vei, utelatte variable og målefeil. b er grei om eksogenitet og relevans og c er standard 2SLS. d er rett, kunne diskutert hva som kan gi skjevhet nedover. e har test om endogenitet, kunne vært mer utfyllende om test av overidentifikasjon.

Oppgave 3 gir rett diff-in-diff i a, b har gode momenter om forventning og timing, kunne også diskutert felles trend og muligheten for spillovers, c gir grei t-test om felles trend, lavere vekst i behandlingsgruppen før økningen i VAT ventes å overestimere (negativ) VAT effekt ja, d handler om spillover, grei t-test, og lavere salg nært til kontrollgruppen indikerer spillover, data 'overdriver' salget i kontrollgruppen og overestimerer da (negativ) effekt.

Fornøyd Jørn Rattsø, faglærer.

NTNU

Kandidat nr./*Candidate no.* 10083

Dato/*Date:* 29.05.13 Side/*Page:* 1

Emnekode/*Subject* SØK3001 Antall ark/*Number of pages:* 30

Denne kolonnen er forbeholdt sensor

This column is for external examiner

Oppgave 1

Har her en multippel regresjonsmodell estimert med data for individer fra folketellingen i USA i år 2000.

Antall observasjoner, $n = 805$

Antar at modellen har blitt estimert vha minste kvadraters metode, MKM, og har fått følgende resultater:

$$\widehat{lwage}_i = 0{,}923 + \underset{(0{,}0063)}{0{,}037}Exp_i - 0{,}00049Exp_i^2$$

$$+ \underset{(0{,}0095)}{0{,}091}Edyrs_i + \underset{(0{,}0581)}{0{,}0030}Part_i - \underset{(0{,}0427)}{0{,}225}Female_i$$

$$- \underset{(0{,}0590)}{0{,}170}Black_i$$

$R^2 = 0{,}2308$, $\bar{R}^2 = 0{,}2251$

$SSR = 282{,}584$

Legger merke til at vi her har tverrsnittdata, dvs observasjoner for et tverrsnitt / fra et utvalg av befolkningen på samme tidspunkt eller periode, her år 2000.

For at estimatorene vi har funnet skal være BLUE (de beste lineære, forventningsrette og effisiente estimatorene) må følgende forutsetninger være oppfylt:

NTNU

Kandidat nr./*Candidate no.* 10083

Dato/*Date:* 29.05.13 Side/*Page:* 2

Emnekode/*Subject* SØK3001

Antall ark/*Number of pages:* 30

Denne kolonnen er forbeholdt sensor

This column is for external examiner

1. Populasjonsmodellen må være lineær i sine parametre, dvs at

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

2. Tilfeldig utvalg på størrelse n. Dette sikrer at $cov(u_i, u_j | x_{i1}, \dots x_{ik}) = 0$

3. Den betingede forventningen til restleddet gitt forklaringsvariablene må være lik 0.
$\Rightarrow E(u_i | x_{i1}, \dots x_{ik}) = 0$
Dette gir at $E(u_i) = 0$ og $cov(u_i, X) = 0$

4. Homoskedastisitet, dvs at variansen til restleddet er lik og konstant for alle observasjoner:
$$var(u_i | x_{i1}, \dots x_{ik}) = \sigma^2$$

5. Ingen perfekt kollinearitet, dvs at ingen av forklaringsvariablene kan skrives som en perfekt lineær kombinasjon av de andre forklaringsvariablene.

Under 1.-5. er, i følge Gauss Markov-teoremet, estimatorene vi finner vha MKM BLUE.

Kandidat nr./*Candidate no.* 10083

Dato/*Date:* 29.05.13 Side/*Page:* 3

Emnekode/*Subject* SØK3001 Antall ark/*Number of pages:* 30

Denne kolonnen er forbeholdt sensor

This column is for external examiner

a) I denne oppgaven skal jeg tolke koeffisientene, kommentere fortegn, størrelse og signifikans. Jeg vil teste om de er signifikante vha t-test og velger derfor å begynne med å forklare hva en t-test er.

Dersom vi i tillegg til forutsetningene 1.-5. antar

6. Normalfordelte restledd $u_i \sim N(0, \sigma^2)$

kan vi begynne med hypotesetesting. Når utvalget er stort (sånn som her) trenger vi egentlig ikke denne forutsetningen, da vi kan lene oss på sentralgrenseteoremet som sier at en variabel Z_n vil være standard asymptotisk normalfordelt dersom

$$Z_n = \frac{Y_n - \mu}{\sigma/\sqrt{n}}$$

, der $Y_1, \ldots Y_n$ er en stokastisk uavh. variabel med forv. μ og varians σ^2. Velger likevel å bygge videre på 6., for å være sikret en god fordeling.

Under 1.-6. vil $\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j))$

og $$\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N(0,1).$$

der $$var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}$$

der $SST_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$ og R_j^2 er R^2 fra regresjonen x_j på alle andre forklaringsvariabler.

Den gule kopien beholder du/*Please keep the yellow page*

Fagtrykk Trondheim AS 11.2012

NTNU

Kandidat nr./*Candidate no.* 10083

Dato/*Date:* 29.05.13 Side/*Page:* 4

Emnekode/*Subject* SØK3001 Antall ark/*Number of pages:* 30

Denne kolonnen er forbeholdt sensor

This column is for external examiner

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

Men siden σ^2 er ukjent, må denne estimeres

$\hat{\sigma}^2 = \dfrac{SSR}{n-k-1}$. Bruker dette til å finne

estimert varians og standardavvik:

$$\widehat{Var}(\hat{\beta}_j \mid x_{i1} \ldots x_{ik}) = \frac{\hat{\sigma}^2}{SST_j(1-R_j^2)}$$

og $se(\hat{\beta}_j) = \sqrt{\dfrac{\hat{\sigma}^2}{SST_j(1-R_j^2)}}$.

Erstatter da $sd(\hat{\beta}_j)$ med $se(\hat{\beta}_j)$ og får:

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

dvs t-fordelt med $n-k-1$ frihetsgrader, der n - antall observasjoner og k - antall forkl. variabler.

Ønsker så å teste signifikans, har da følgende hypotese:

$H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

↳ tosidig test

Bruker følgende testobservator:

$$t^{OBS} = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$ siden $\beta_j = 0$ under H_0.

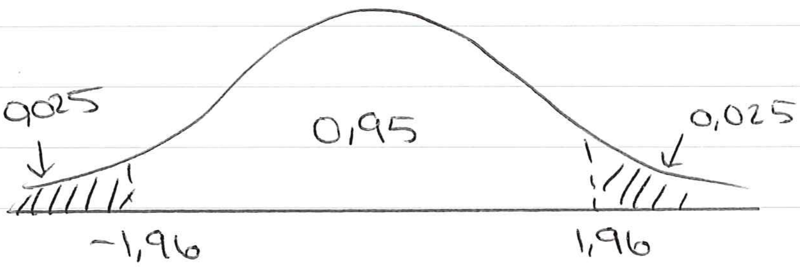
Den gule kopien beholder du/*Please keep the yellow page*

Fagtrykk Trondheim AS 11.2012

Denne kolonnen er forbeholdt sensor

This column is for external examiner

Velger så et 5% signifikansnivå og forkaster H_0 dersom $|t^{OBS}| > |t^c|$, der t^c er kritisk verdi, her lik 1,96. Forkaster altså H_0 dersom t^{OBS} befinner seg i ett av de skraverte områdene:



Ser at det er to ting som bidrar til forkastning: høy estimert verdi av β_j (da er det mindre sannsynlig at denne er lik 0) eller et lavt standardavvik fordi et lavt standardavvik indikerer et presist resultat.

Kan få to typer feil:

signi.nivå

Type I-feil: $P(\text{Forkaste } H_0 | H_0 \text{ er sann}) = \alpha$

Type II-feil: $P(\text{ikke forkaste } H_0 | H_1 \text{ er sann})$.

Bruker så dette til å teste om β_j, $j=2,\dots 6$ er signifikante. De vil altså være signifikante dersom $|t^{OBS}| > |t^c| = 1,96$.

i) $\hat{\beta}_1 = 0{,}037$

Dette sir at $t^{OBS} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{0{,}037}{0{,}0063} = 5{,}873$.

Ser at $|t^{OBS}| > t^c = 1{,}96$. Kan da forkaste H_0 om at $\beta_1 = 0$, og β_1 er altså statistisk signifikant forskjellig fra 0.

Har her log-level-form, dvs at en enhets-økning i arbeidsmarkedserfaring, dvs 1 år ekstra med erfaring gir en økning i timelønn på $\beta_1 \cdot 100\%$ dvs $0{,}037 \cdot 100 = 3{,}7\%$.

Har et positivt forhold mellom erfaring og timelønn, hvilket virker logisk.

ii) $\hat{\beta}_2 = -0{,}00049$

$\Rightarrow t^{OBS} = \frac{-0{,}00049}{0{,}00015} = -3{,}26$

Ser at denne også er signifikant. β_1 gir ikke den hele effekten av 1 års ekstra i erfaring, β_2 vil også ha noe å si!!!

Har også log-level-form her, men et negativt forhold mellom lnwage og Exp_i^2.

Har da at effekt av 1 års ekstra i arbeid blir:

$$\frac{\partial lnwage}{\partial Exp} = \beta_1 + 2\beta_2 Exp_i$$
$$= 0{,}037 - 2 \cdot 0{,}00049 Exp_i$$

Ser at det avh. av nivået på Exp_i!

Denne kolonnen er forbeholdt sensor

This column is for external examiner

iii) $\hat{\beta}_3 = 0{,}091$

$$t^{OBS} = \frac{0{,}091}{0{,}0095} = 9{,}579$$

Har at $|t^{OBS}| > |t^q = 1{,}96| \Rightarrow \beta_3$ er også statistisk signifikant, dvs at antall års utd. har en signifikant positiv effekt på timelønn.

Har her log-level-forhold mellom timelønn og lengde på utdanning $\Rightarrow$

$$\frac{1}{wage} dwage = \beta_3 \cdot dEdyrs$$

$$\frac{\frac{dwage}{wage}}{dEdyrs} = \beta_3 \quad | \cdot 100$$

$$\frac{\frac{dwage}{wage} \cdot 100}{dEdyrs} = \beta_3 \cdot 100$$

Har da en semi-elastisitet!

ser at et år til med utdanning øker timelønn med $\beta_3 \cdot 100\,\% \Rightarrow 9{,}1\,\%$. Har et positivt forhold, dette virker også logisk.

iv) $\hat{\beta}_4 = 0{,}0030$

$$t^{OBS} = \frac{0{,}0030}{0{,}0581} = 0{,}0516$$

Ikke signifikant. Det ser altså ikke ut til å

Denne kolonnen er forbeholdt sensor

This column is for external examiner

være noen forskjell timelønn om du jobber deltid eller ikke, siden β_4 forteller oss forskjellen på timelønn mellom individer som jobber deltid (PartT = 1) eller ikke (PartT = 0).

v) $\hat{\beta}_5 = -0,225$

$$t^{OBS} = \frac{-0,225}{0,0427} = 5,269$$

β_5 står i likhet med β_4 foran en dummy, og vil dermed fortelle oss forskjellen på timelønn mellom kvinner og menn. Har her at β_5 er signifikant forskjellig fra 0 siden $|t^{OBS}| > |t^c|$ ⇒ det er en signifikant forskjell på timelønn mellom kvinner og menn.

Kvinner vil ha en predikert timelønn $-0,225 \cdot 100 = -22,5\%$ lavere enn menn.

√

vi) $\hat{\beta}_6 = -0,170$

$$t^{OBS} = \frac{-0,170}{0,0590} = -2,88$$

β_6 er signifikant forskjellig fra 0. Står også foran dummy ⇒ sier at timelønn vil være 17% lavere for individer som er svarte.

√

ser fra i) - vi) at alle faktorer er viktige til å forklare lønnsforskjeller bortsett fra om du jobber deltid eller ikke, dvs variabelen Part$_i$. Dette siden β_4 ikke kan sies å være signifikant forskjellig fra 0.

b) som sagt i a) vil effekt av erfaring på timelønn være

$$\frac{\partial \ln wage}{\partial exp} = \beta_1 + 2\beta_2 Exp_i$$

Denne vil nå sitt maksimum når $\frac{\partial \ln wage}{\partial exp} = 0$. Vet at dette er et maxpunkt siden $\beta_2 < 0$ (da er andre-ordensbetingelsen oppfylt)

Finner maxpunktet:

$$\frac{\partial \ln wage}{\partial exp} = \beta_1 + 2\beta_2 Exp = 0$$

$$0{,}037 - 2 \cdot 0{,}00049 Exp = 0$$

$$2 \cdot 0{,}00049 Exp = 0{,}037$$

$$Exp = \frac{0{,}037}{2 \cdot 0{,}00049}$$

$$Exp = \underline{\underline{37{,}75}}$$

NTNU

Kandidat nr./*Candidate no.* 10083

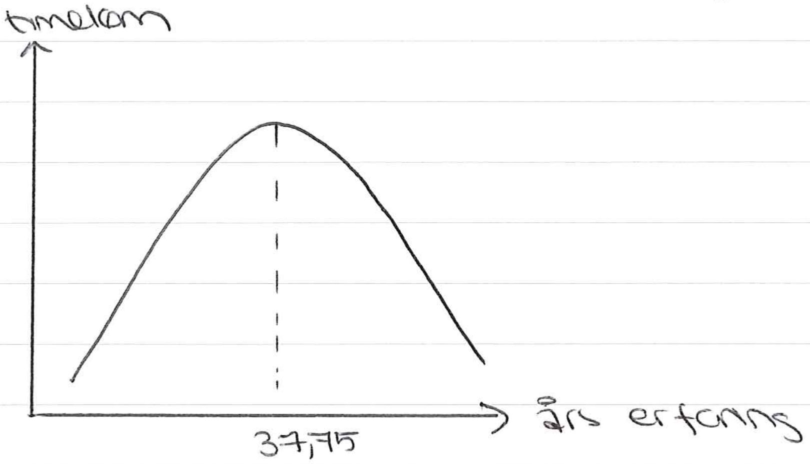
Dato/*Date:* 29.05.13 Side/*Page:* 10

Emnekode/*Subject* SØK3001 Antall ark/*Number of pages:* 30

Denne kolonnen er forbeholdt sensor

This column is for external examiner

Dvs at individets kan begynne å falle etter 37,75 år ute i arbeidsmarkedet. Frem til da opplever individet økt timelønn. Kan illustrere effekten av erfaring på timelønn vha følgende graf:



c) Kan teste om koeffisienten for Female, dvs β_5, er lik $-0,1$ vha følgende test:

$$H_0: \beta_5 = -0,1$$
$$H_1: \beta_5 \neq -0,1$$

Velger et 5% signifikansnivå, får da at $|t^{OBS}| > |t^c| = 1,96$ for at H_0 skal forkastes.

Beregner t^{OBS}:

$$t^{OBS} = \frac{\hat{\beta}_5 - (-0,1)}{se(\hat{\beta}_5)} = \frac{-0,225 + 0,1}{0,0427} = \underline{\underline{-2,9274}}$$

Ser at $|t^{OBS}| = 2,9274 > |t^c| = 1,96$

$\Rightarrow H_0$ kan forkastes.

Kan altså forkaste at $\beta_5 = -0,1$.

Den gule kopien beholder du/*Please keep the yellow page*

Fagtrykk Trondheim AS 11.2012

Denne kolonnen er forbeholdt sensor
This column is for external examiner

d) Tillegg av Exp^3 og Exp^4 gir en SSR = 281,54

Kan teste om β_7 (parameteren foran Exp^3) og β_8 (parameteren foran Exp^4) er lik 0 vha en F-test. Sammenligner da SSR fra begrenset og ubegrenset modell. SSR vil alltid øke når restriksjoner pålegges, dette siden SSR er et mål på uforklart variasjon. Ser da på relativ endring i SSR korrigert for antall restriksjoner og frihetsgrader.

Har følgende F-observator:

$$F^{OBS} = \frac{SSR_R - SSR_{UR} / q}{SSR_{UR} / n-k-1} \sim F_{q,n-k-1}$$

der SSR_R – SSR fra begrenset modell, Her vil det være vår første: $SSR_R = 282,584$

SSR_{UR} – SSR fra ubegrenset modell, dette er den som inneholder Exp^3 og Exp^4. $SSR_{UR} = 281,54$.

q – antall restriksjoner

n – antall observasjoner

k – antall forklaringsvariabler i ubegrenset modell.

Har da følgende nullhypotese:

$H_0: \beta_7 = \beta_8 = 0$

H_1: simultant $\neq 0$

Denne kolonnen er forbeholdt sensor

This column is for external examiner

Viktig her at vi tester om vi kan si at β_7 og β_8 er simultant lik 0 eller ikke.

Beregner vår F^{OBS}:

$$F^{OBS} = \frac{282{,}584 - 281{,}54 \,/\, 2}{\frac{281{,}54}{805-9}}$$

$$= 1{,}48$$

Velger 5 % signifikantnivå. $F^{C}_{2,805-8-1} = 3{,}00$ (leser av tabellen)

På samme måte som med t-test, forkastes H_0 dersom $F^{OBS} > F^{C}$. Ser at dette ikke er tilfellet her $\Rightarrow$ konkluderer med at vi ikke kan forkaste at $\beta_7 = \beta_8 = 0$, dvs at Exp^3 og Exp^4 ikke har noen effekt. Det ser altså ut til å være slik.

e) R^2 måler "goodness-of-fit", dvs forklaringskraften til en modell. $R^2 \cdot 100$ vil si hvor mye som forklares i prosent. R^2 er definert som:

$$R^2 = \frac{SSE}{SST}$$, der SSE er et mål på forklart variasjon:

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2$$

og SST er et mål på total variasjon:

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

Siden SST = SSE + SSR, der SSR er et mål på uforklart variasjon ($SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$) kan vi skrive at:

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

Når vi legger til forklaringsvariabler vil alltid R^2 øke, og vi må derfor straffe for frihetsgrader. "R^2 justert", $\bar{R}^2$, tar høyde for dette og korrigerer for nettopp dette:

√

$$\bar{R}^2 = 1 - \frac{SSR/n-k-1}{SST/n-1}.$$

"R^2 justert" er dermed bedre å bruke til å sammenligne modeller med ulikt antall forklaringsvariabler.

f) Får nå oppgitt at en regresjon med summen av kvadrerte avvik fra modellen over med samme forkl. variabler gir en $R^2 = 0{,}0014$. Skal forklare hvordan dette kan brukes til å teste for heteroskedastisitet, men ønsker først å gjøre rede for hva heteroskedastisitet er.

Antok i sted homoskedastisitet, som er at

Denne kolonnen er forbeholdt sensor
This column is for external examiner

restleddet har lik og konstant varians uavhengig av observasjonsenhetes. Ved heteroskedastisitet vil dette ikke være tilfellet for minst én av enhetene.
Har da at

$$var(u_i | x_{i1} \dots x_{ik}) = \sigma_i^2$$

Dersom $E(u_i | x_{i1} \dots x_{ik}) = 0$ fortsatt vil vi få forv. rette estimatorer, men variansen vil påvirkes. Vil da få følgende varians:

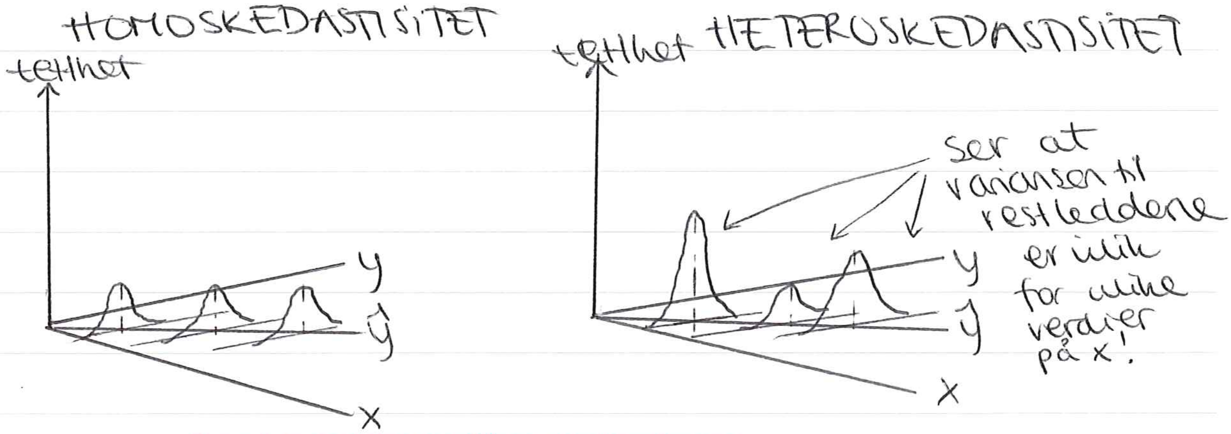
$$var(\hat{\beta}_j | x_{i1} \dots x_{ik}) = \frac{\sum (x_{ij} - \bar{x}_j)^2 \sigma_i^2}{SST_x^2}$$

Kan korrigere for dette vha Whites metode der σ_i^2 erstattes med $(\hat{u}_i)^2$. Får da robust varians/robuste standardavvik:

$$Var(\hat{\beta}_j) = \frac{\sum (x_i - \bar{x})^2 (\hat{u}_i)^2}{SST_x^2}$$

$$st(\hat{\beta}_j) = \sqrt{\frac{\sum (x_i - \bar{x})^2 (\hat{u}_i)^2}{SST_x^2}} \leftarrow$$ ROBUSTE STANDARDAVVIK

Kan illustrere forskjellen mellom homo- og heteroskedastisitet:



Kan så teste for heteroskedastisitet vha Breusch-Pagan-test. Finner da først residualene $\hat{u}_i$, fra opprinnelig regresjon. Bruker så dette til å estimere følgende:

$$\hat{u}_i^2 = \delta_0 + \delta_1 x_{i1} + \delta_2 x_{i2} + \delta_3 x_{i3} + \dots + \delta_6 x_{i6} + v_i \quad (1)$$

(Har i vår regresjonsmodell 6 forklaringsvariabler, tar med alle de).
Estimerer denne og tester så

$H_0: \delta_1 = \delta_2 = \dots = \delta_6 = 0$
$H_1:$ simultant $\neq 0$.

Dersom H0 forkastes har vi heteroskedastisitet siden parameterne foran forklaringsvariablene ikke kan ses å være simultant lik 0.

Kan teste dette vha F-test, og det er her R^2 kommer inn i bildet. Skriver om F-observatoren til å inneholde R^2 isf SSR.

$$\Rightarrow F^{OBS} = \frac{R^2_{UR} - R^2_R / q}{(1 - R^2_{UR})/n-k-1}$$

Ubegrenset modell vil være (1), mens begrenset modell vil være:

$$\hat{u}_i^2 = \delta_0 + v_i$$

Denne vil ha $R^2_R = 0$

NTNU

Kandidat nr./*Candidate no.* 10083

Dato/*Date:* 29.05.13 Side/*Page:* 16

Emnekode/*Subject* SØK3001

Antall ark/*Number of pages:* 30

Denne kolonnen er forbeholdt sensor

This column is for external examiner

Dette gir da:

antall restriksjoner

$$F^{OBS} = \frac{0{,}0014 - 0 / 6}{1 - 0{,}0014 / 805 - 6 - 1}$$

antall observasjoner

antall forkl. variable

$$= \underline{\underline{0{,}1865}}$$

Ved 5% signifikansnivå vil $F^{C}_{6,798} = 2{,}10$

Må altså ha at

$F^{OBS} > F^{C}$ for at H_0 skal forkastes.

Kan med andre ord ikke forkaste H_0 siden $F^{OBS} = 0{,}1865 < F^{C} = 2{,}10 \Rightarrow$ Har altså dessverre heteroskedastisitet.

NTNU

Kandidat nr./Candidate no. 10083

Dato/Date: 29.05.13 Side/Page: 17

Emnekode/Subject SØK3001

Antall ark/Number of pages: 30

Denne kolonnen er forbeholdt sensor

This column is for external examiner

Oppgave 2

Skal nå estimere effekten av migrasjon på arbeidsmarkedet i Storbritannia. Estimerer følgende modell via MKM:

$$lwage_i = \beta_0 + \beta_1 MigPop_i + \beta_2 X_i + u_i$$

Finner at $\hat{\beta}_1 = 0,88$, $se(\hat{\beta}_1) = 0,21$

a) Har her et log-level-forhold mellom reallønn i region i og %-andel innflyttere i region i. Dette gir at β_1 vil være en semielastisitet og gi effekt på timelønn tilsvarende $\beta_1 \cdot 100\%$ som følge av en enhets økning (her: et prosentpoengs høyere andel) i MigPop.

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

Har altså at en enhets økning i MigPop fører til en 88 % økning i reallønnen.

Det er nærliggende å tro at dette estimatet ikke er forventningsrett, da vi trolig vil ha et endogenitetsproblem. Trolig vil både andel innflyttere påvirke reallønn og reallønn påvirke andel innflyttere. Jo høyere reallønnen i region i er, desto mer attraktivt er det å flytte til nettopp denne regionen. Dersom dette er tilfellet, at vi har et endogenitetsproblem, vil vi ha en sammenheng mellom mispop og u, hvilket vi vet fra oppgave 1 at gir brudd på forutsetning 3. Har at $E(u_i \mid mispop) \neq 0$. Dette gir at estimatorene vi finner vha MKM ikke er forventningsrette eller konsistente.

Pga tidsbegrensningen denne eksamenen har, rekker jeg ikke å utlede MKM. Men MKM tar utgangspunkt i å minimere summen av kvadrerte avvik mellom observert verdi og predikert verdi, dvs residualene. Har altså følgende minimerings-problem:

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

Her:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} (lwage_i - \widehat{lwage}_i)^2$$

Dette gir følgende estimator for β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (lwage - \overline{lwage})(mispop - \overline{mispop})}{\sum_{i=1}^{n} (mispop - \overline{mispop})^2}$$

Skriver fra nå av $\sum_{i=1}^{n} = \sum$ og $lwage - w$ og $mispop - m$.

$$\Rightarrow \hat{\beta}_1 = \frac{\sum (w - \bar{w})(m - \bar{m})}{\sum (m - \bar{m})^2}$$

Dersom estimatoren er forventningsrett, vil vi ha at $E(\hat{\beta}_1)=\beta_1$. Skriver om uttrykket for $\hat{\beta}_1$ og ser at dette ikke vil være tilfellet dersom $E(u|m)\neq 0$.

$$\hat{\beta}_1 = \beta_1 + \frac{\sum u_i(m-\bar{m})}{\sum(m-\bar{m})^2} \qquad (*)$$

Tar så forv.verdien av $(*)$

$$E(\hat{\beta}_1) = \beta_1 + \frac{E(u_1|m)\sum(m-\bar{m})}{\sum(m-\bar{m})^2}$$

Dersom $E(u_1|m)\neq 0$ vil $E(\hat{\beta}_1)\neq\beta_1$ og vi vil få forventningsskjevhet. Siden det er nærliggende å tro at vi har et endogenitetsproblem her, vil $\hat{\beta}_1$ ikke være forv.rett.

Kan vise at den heller ikke vil være konsistent. Vil ha konsistens dersom sannsynlighets-grensen til $\hat{\beta}_1$ er lik β_1. Dvs at $\hat{\beta}_1$ konvergerer mot β_1 (sann verdi) når $n\to\infty$.

Tar plim av $(*)$:

$$\text{plim}(\hat{\beta}_1) = \beta_1 + \frac{\text{cov}(u,m)}{\text{var}(m)}$$

Ser at det samme gjelder for dette uttrykket. Vil mest sannsynlig ikke ha at $\text{plim}(\hat{\beta}_1)=\beta_1$ da $\text{cov}(u,m)\neq 0$ pga endogenitet.

NTNU

Kandidat nr./Candidate no. 10083

Dato/Date: 29.05.13 Side/Page: 20

Emnekode/Subject SØK3001 Antall ark/Number of pages: 30

Denne kolonnen er forbeholdt sensor
This column is for external examiner

b) For at beholdningen av innflyttere for 10 år siden skal være et gyldig instrument for mispop, må denne variabelen (kaller den z) oppfylle følgende krav:

i) Den må være eksogen, dvs at $cov(z,u)=0$. I motsetning til mispop må altså denne være ukorrelert med restleddet.

√

ii) Den må være relevant, dvs at $cov(z,m)\neq 0$. Instrumentvariabelen må altså være korrelert med mispop. Siden variansen til estimatoren for instrumentvariabelen avhenger av korrelasjonskoeff. i annen er det viktig at denne korrelasjonen er så høy som mulig for å sikre lavest mulig varians og mest mulig presise resultater.

Har at $Var(\hat{\beta}_1^{IV})=\dfrac{\sigma^2}{SST_m R^2_{z,m}}$

mens vi før hadde at

$$Var(\beta_1)=\frac{\sigma^2}{SST_m}.$$

Ser at $R^2_{z,m}$ nærmest mulig 1 (perfekt korrelasjon) vil gi lavest mulig varians.

Den gule kopien beholder du/Please keep the yellow page

Fagtrykk Trondheim AS 11.2012

Kan være at dette ikke er et godt instrument fordi det også for 10 år siden var et endogenitetsproblem, eller så kan det hende at forholdene har forandret seg såpass mye at korrelasjonen er svak.

c) Kan finne et instrumentvariabel estimat vha 2-stegs MKM.

1. steg: studerer sammenhengen mellom forklaringsvariabelen (her: Mispop) og IV:

$Mispop_i = \pi_0 + \pi_1 Z_i + \pi_2 X_i + u_i$

viktig her å ha med kontrollvariablene! Inkluder all informasjon du har.

Estimer så denne såkalte redusert-form-likningen vha vanlig OLS

$\Rightarrow \widehat{Mispop}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_1 + \hat{\pi}_2 X_i$

Bruk så $\widehat{Mispop}$ videre i 2. steg:

2. steg: Erstatt da Mispop med $\widehat{Mispop}$ funnet i 1. steg i den opprinnelige relasjonen.

$\Rightarrow wage_i = \beta_0 + \beta_1 \widehat{Mispop} + \beta_2 X_i + u_i$

Estimer så denne vha OLS. Dersom Z_1 oppfyller krav i) og ii) fra b) vil vi få et konsistent estimat for effekten av Mispop.

Kan vise dette:

(ser nå bort i fra $\beta_2 x_i$ for enkelthetens skyld, men så lenge denne er eksogen vil vi få samme resultat)

Har at $w_i = \beta_0 + \beta_1 m_i + u_i$

Trekker i fra forv. verdien til variablene:

$$w_i - E(w) = \beta_1(m_i - E(m)) + \underbrace{u_i - E(u_i)}_{=0}$$

Ganger med $(z_i - E(z))$ og tar forv. verdien av dette:

$$E(w_i - E(w))(z_i - E(z)) = \beta_1 E(m_i - E(m_i))(z_i - E(z_i))$$

$$+ \underbrace{E(u_i(z_i - E(z_i)))}_{=0 \text{ per definisjon}}$$

$$\Rightarrow \quad \beta_1^{IV} = \frac{E(w_i - E(w))(z_i - E(z))}{E(m_i - E(m_i))(z_i - E(z))}$$

$$= \frac{cov(w,z)}{cov(m,z)}$$

Erstatter så de teoretiske momentene med de empiriske.

$$\Rightarrow \quad \hat{\beta}_1^{IV} = \frac{\cancel{1/n}\sum(w_i - \bar{w})(z_i - \bar{z})}{\cancel{1/n}\sum(m_i - \bar{m})(z_i - \bar{z})}$$

$$= \frac{\sum(w_i - \bar{w})(z_i - \bar{z})}{\sum(m_i - \bar{m})(z_i - \bar{z})}$$

Denne er konsistent fordi $plim(\hat{\beta}_1^{IV}) = \beta_1^{IV}$

$$\Rightarrow plim(\hat{\beta}_1^{IV}) = \frac{cov(w,z)}{cov(m,z)} = \beta_1^{IV}.$$

Som sagt i b) vil $Var(\hat{\beta}_1^{IV}) = \frac{\sigma^2}{SST_m R^2_{m,z}}$.

Har da funnet et IV-estimat for effekten av m, Mispop, her kalt β_1^{IV}.

d) Finner nå vha av IV at
$\hat{\beta}_1^{IV} = 0,94$ og $se(\hat{\beta}_1^{IV}) = 0,58$.

Kan da finne skjevheten som oppstod da vi estimerte uten IV:
$Bias(\hat{\beta}_1^{OLS}) = \hat{\beta}_1^{OLS} - \hat{\beta}_1^{IV} = 0,88 - 0,94 = \underline{\underline{-0,06}}$

V

Hadde da en negativ skjevhet og β_1 ble underestimert. Ser nå også at $\hat{\beta}_1^{IV}$ ikke er signifikant forskjellig fra 0, da $t_{IV}^{OBS} = \frac{0,94}{0,58} = 1,62$. Har at $|t^{OBS}| < t^*$ på et 5% signifikansnivå. Før var den signifikant:

$$t_{OLS}^{OBS} = \frac{0,88}{0,21} = 4,19$$

Dette kan skyldes at den virket mer signifikant fordi den fanget opp noe av endogeniteten.

Vel

e) Det er to tester som er viktige å utføre i forbindelse med IV:

1. Er instrumentvariabelen relevant?
   Det er viktig at den er det, dette er en av forutsetningene for at vi kan bruke IV og få konsistente estimatorer. Dette sjekker vi ved å teste parameteren foran IV-variabelen

Denne kolonnen er forbeholdt sensor

This column is for external examiner

i redusert form likningen fra 1 stegi 2-stegs MKM. Har da at

$Mispop_i = \Pi_0 + \Pi_1 z_1 + \Pi_2 X_1 + u_1$

Tester da $H_0: \Pi_1 = 0$

$H_1: \Pi_1 \neq 0$ ved vanlig

t-test. Hvis Ho kan forkastes vil vi ha en relevant IV. Viktig at den forkastes med <u>god</u> margin, da vi vet at en svak sammenheng mellom IV og forklaringsvariabel vil gi høy varians og upresise resultater.

Vel

2. Er Mispop i det hele tatt endogen? Kan teste dette via en såkalt Hausman-test. Viktig å teste dette, da vi helst ikke vil bruke IV med mindre vi <u>må</u> (pga høyere varians).

Har at

$lwage = \beta_0 + \beta_1 Mispop + \beta_2 X_i + u_i \quad (1)$

Antar så at vi i tillegg har z_2 som er en eksogen variabel. Ser så på

$Mispop_i = \gamma_0 + \gamma_1 X_i + \gamma_2 z_2 + v_2 \quad (2)$

siden både X_i og z_2 er eksogene, vil vi kun ha at Mispop er endogen dersom v_2 er korrelert med u_1.

Antar at vi kan skrive $u_1 = \delta_0 v_2 + støy$

Det letteste da er å bare inkludere v_2 i (1), men siden v_2 er ubservenar må vi estimere denne; Gjør det ved å estimere (2) vha vanlig OLS og finne residualene: $\hat{v}_2$. Bruker så dette til å inkludere $\delta_0 \hat{v}_2$ i (1), får da:

$$lwage = \beta_0 + \beta_1 Mispop + \beta_2 X_i + \delta_0 \hat{v}_2 + støy$$

Estimerer så denne vha vanlig MKM og tester om

$H_0: \delta_0 = 0$
$H_1: \delta_0 \neq 0$

Hvis H0 ikke kan forkastes vil det ikke være noen sammenheng mellom v_2 og u, og mispop vil ikke være endogen.

Dersom H0 kan forkastes, er δ_0 signifikant forskjellig fra 0 og mispop er endogen.

Kan også teste for overidentifikasjon dersom vi har flere IV'er for samme forklaringsvariabel og lurer på om alle er uavhengige av restleddet. Har dermot ikke flere IV'er her, kun én, og da er ikke denne testen relevant i denne sammenhengen.

Denne kolonnen er forbeholdt sensor
This column is for external examiner

Oppgave 3

Legger her merke til at vi har sammenkoblede tverrsnittdata, og et såkalt naturlig eksperiment. Har 10 regioner som er unntatt økningen i MVA, og dette vil være kontrollgruppen. Har også 10 regioner som har hatt en økning i MVA; disse vil være behandlingsgruppen.

a) "difference-in-difference" estimatet av effekten vil være forskjellen mellom forskjellen mellom behandlingsgruppen og kontrollgruppen. Dvs at vi til å begynne med hadde en forskjell mellom behandlingsgruppen og kontrollgruppen, og skal så se på hvordan denne forskjellen mellom de har endret seg som følge av at behandlingsgruppen har fått en økning i MVA.

Finner først endringen i behandlingsgruppen:

$$4{,}15 - 4{,}77 = -0{,}62$$

Ser at salget har gått ned.

Finner så endringen i kontrollgruppen:

$$4{,}21 - 4{,}34 = -0{,}13$$

Finner så differansen mellom disse:

$$-0{,}62 - (-0{,}13) = \underline{\underline{-0{,}49}}$$

Dette vil være difference-in-difference estimatet av effekten, altså kausaleffekten

Denne kolonnen er forbeholdt sensor

This column is for external examiner

av at det ble innført en MVA-økning i behandlingsgruppen. Behandlingsgruppen har altså fått 0,49 lavere gjennomsnittlig detaljsalg ift kontrollgruppen.

b) Når vi tolker difference-in-difference estimatet er det viktig å være oppmerksom på at dette er en gjennomsnittseffekt. Må forøvrig vite at tallene vi har samlet inn for perioden før og etter, virkelig er tall for periodene før og etter. F.eks: kan MVA-økningen allerede ha vært fortsett 3 mnd før den ble innført? Hvis ja, har man kanskje allerede en effekt da. Kan også spørre seg om 3 mnd etter innført økning i MVA er lenge nok etterpå til å drøfte om denne økningen har hatt noen effekt.

Det er også viktig å stille seg spørsmålet; har det skjedd noe annet? Dvs, har vi et utelatt variabel problem? Isåfall vil vi få utelatt variabel skjevhet i estimatene våre, og difference-in-difference-estimatet vårt vil fange opp effekten av noe annet og dermed gi oss feilaktige resultater. F.eks kan det hende at det var en grunn til at nettopp de 10 regionene ikke fikk økning i MVA, kan altså stille spørsmålstegn ved om dette var tilfeldig.

NTNU

Kandidat nr./Candidate no. 10083

Dato/Date: 29.05.13 Side/Page: 28

Emnekode/Subject SØK3001

Antall ark/Number of pages: 30

Denne kolonnen er forbeholdt sensor

This column is for external examiner

c) Har nå estimert

$$sales_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 Time_t + \alpha_{DiD} \cdot Treat_i Time_i + u_{it}$$

der $Treat = \begin{cases} 1 & \text{hvis butikken er i region med MVA-økning (behandlingsgruppe)} \\ 0 & \text{hvis kontrollgruppe} \end{cases}$

$Time = \begin{cases} 1 & \text{de 3 første mnd i 2010} \\ 0 & \text{de siste 3 mnd i 2009} \end{cases}$

Estimerer så at $\hat{\alpha}_{DiD} = -0{,}1$, $se(\hat{\alpha}_{DiD}) = 0{,}02$
ser at denne er signifikant forskjellig fra 0:

$$t^{obs} = \frac{-0{,}1}{0{,}02} = -5.$$

Har at $|t^{obs}| = 5 > 1{,}96 = t^c$ på et 5% sign.nivå.
Har altså en svært signifikant kausaleffekt
(parameteren α_{DiD} vil måle difference-in difference-estimatet av effekten av en økning i MVA i behandlingsgruppen).
Siden begge periodene her er før økningen i MVA, er det nærliggende å tro at vi har utelatt variabelskjevhet. Som sagt i b) skaper dette problemer siden kausaleffekten fra a) da vil fange opp effekten av noe annet. Det betyr at min difference-in difference-estimering i a) fant feilaktig kausaleffekt. Jeg tror da at estimatet i a) ble mer negativt enn det burde ha blitt dvs at effekten ble overestimert fordi den også fanget opp denne negative effekten.

Vel

Fagtrykk Trondheim AS 11.2012

Denne kolonnen er forbeholdt sensor

This column is for external examiner

d) Estimerer nå endring i salget i butikk i over perioden med MVA-økning i behandlingsregionene:

$$\Delta sales_i = \beta_0 + \beta_1 Distance_i + u_i$$

der $\hat{\beta_1} = 0{,}011$, $se(\hat{\beta_1}) = 0{,}0012$

$$\Rightarrow t^{OBS} = \frac{\hat{\beta_1}}{se(\hat{\beta_1})} = 9{,}167$$

ser at $|t^{OBS}| > 1{,}96 = t^c \Rightarrow \hat{\beta_1}$ er statistisk signifikant forskjellig fra 0. Ser her at effekten av avstand fra butikk i til nærmeste kontrollregion har en signifikant effekt på endringen i salget i butikk i. Dess avstanden øker vil endringen bli større, hvilket vil si at beliggenheten til butikken ift kontrollregionen har noe å si for hvor mye økningen i MVA påvirker salg. Kunne vurdert å ha kontrollert for denne avstanden i analysen, da dette kunne gitt riktigere estimater.

Har altså at jo lenger vekk fra kontrollregionene, jo større endring i salg. Tror da at estimatet i a) overestimerer effekten i butikkene som ligger nær kontrollregionene, mens det underestimerer effekten av økt MVA i butikkene som ligger fjernt fra kontrollregionene. Ser fra dette at det er en viktig

Denne kolonnen er forbeholdt sensor

This column is for external examiner

at kausaleffekten kun er en gjennomsnitts
effekt, fordi den ikke tar hensyn til
forskjeller som dette.

Fagtrykk Trondheim AS 11.2012