Arbeid med mikrodata (registerdata), ØASØK4700

Jørgen Modalsli, OsloMet, februar 2025

Referanser til lærebok: Borjas, Labor Economics, Ninth edition

Våren 2025: Vi har jobbet med **Innledning**, **Utdanning og arbeid** og **Data explorer kapittel 3** på forelesning.

Før forelesning:

Logg på microdata.no med BankID, så du vet du har tilgang

Innledning

Vi begynner med følgende øvingsoppgaver:

https://www.microdata.no/wp-content/uploads/2022/04/Oppgavesett-innforende-microdata-kurs-basic.pdf (side 1-15)

Gå gjennom trinnene i øvingsoppgavene.

Se også annen informasjon på kurssiden: https://www.microdata.no/kurs/

Utdanning og arbeid

Eksempel på studie av inntekt og utdanningsnivå (ref. kapittel 6).

Kommandoer som skal brukes på microdata.no er skrevet med **fet skrift.** Om du har data i minnet fra før kan det hende du må starte med å skrive **clear**

- Koble til databasen
    - 37 er siste versjon når dette skrives. Det går også an å bruke en tidligere versjon men da har man ikke de nyeste dataene
    - "db" er navnet vi gir databasen. Man kan også bruke et annet navn, for eksempel "kapittel6"
      require no.ssb.fdb:37 as db
- Lag et nytt datasett. Dette kan vi også kalle hva vi vil, vi velger "utdanning"
  create-dataset utdanning

- Importer en del sentrale variable

  import db/BEFOLKNING_KJOENN as kjønn

  import db/BEFOLKNING_FOEDSELS_AAR_MND as faarmnd

  import db/BEFOLKNING_STATUSKODE 2023-01-01 as regstat
- Bare behold de som er bosatt i Norge per januar 2023 (altså ikke de som er utvandret, døde, mm.)

  keep if regstat=='1'
- Lag en variabel for alder. Først lager vi en for fødselsår, så regner vi om til alder pr. 2023

  generate fodselsaar = int(faarmnd/100)

  generate alder = 2022 - fodselsaar
- Sjekk at aldersfordelingen ser hensiktsmessig ut

  tabulate alder
- Legg til informasjon om inntekt og alder

  import db/INNTEKT_LONN 2022-12-31 as inntekt

  import db/NUDB_BU 2022-12-31 as utdanningskode
- Legg til informasjon om foreldres utdanning

  import db/NUDB_SOSBAK as familiebakgrunn
- Lag en variabel for utdanningsnivå (se https://www.ssb.no/klass/klassifikasjoner/36)

  generate utdanningsnivå = substr(utdanningskode,1,1)
- Konverter variabelen for utdanning fra "string" til "numeric"

  destring utdanningsnivå

  tabulate utdanningsnivå
- Lag en variabel for "antall år utdanning" (her litt omtrentlig regnet)

  generate utdanning_aar = .

  replace utdanning_aar = 7 if utdanningsnivå==1

  replace utdanning_aar = 10 if utdanningsnivå ==2

  replace utdanning_aar = 12 if utdanningsnivå ==3

  replace utdanning_aar = 13 if utdanningsnivå ==4

  replace utdanning_aar = 14 if utdanningsnivå ==5

  replace utdanning_aar = 16 if utdanningsnivå ==6

  replace utdanning_aar = 18 if utdanningsnivå ==7

  replace utdanning_aar = 21 if utdanningsnivå ==8

Nå kan vi kjøre noen analyser. Før vi gjør det, ser vi på variablene at de er hensiktsmessige.

Vi begrenser populasjonen vår til personer mellom 30 og 50 år, i denne omgang.

keep if alder>=30

keep if alder<=50

Vi ser på log inntekt (da antar vi at utdanning betyr mer for lave enn for høye inntekter)

generate log_inntekt = ln(inntekt)

summarize inntekt log_inntekt utdanning_aar

regress log_inntekt utdanning_aar

regress log_inntekt utdanning_aar i.kjønn

Hva med "ability bias"? La oss prøve å kontrollere for karakterer på grunnskolen. Vi bruker grunnskolen fordi det er endogent hvem som tar videregående.

import db/NUDB_KURS_GRPOENG as karakterpoeng

Vi leser i dokumentasjonen for karakterpoeng-variabelen https://microdata.no/discovery/variable/no.ssb.fdb/37/NUDB_KURS_GRPOENG?searchString=karakter at denne er endret ca 2007. De som var 16 år i 2007 er ca 32 år i 2023. I tillegg er vi nysgjerrige på hvor langt tilbake karaktervariabelen finnes, det vil si hvor gamle personer vi har karakterdata for.

Hvor mange observasjoner (utropstegn! betyr motsatt, "sysmiss" betyr at variabelen mangler; !sysmiss er altså at det finnes informasjon):

tabulate alder if !sysmiss(karakterpoeng)

Det ser ut som vi bør basere analysen på personer som er 37 år og yngre.

Gjennomsnittlige poeng:

tabulate alder, summarize(karakterpoeng)

Det ser ut som variabelen gjør et hopp mellom de som er 31 og 32 år. I tillegg mellom de som er 37 og 38, men vi vet at det er fordi det er mye færre observasjoner for 38-åringer.

Vi bør altså videre basere oss på de som er fra og med 32 til og med 37 år.

Regresjon for inntekt og utdanning for denne gruppen, med stadig flere kontroller:

regress log_inntekt utdanning_aar i.kjønn if alder>=32 & alder<=37

regress log_inntekt utdanning_aar i.kjønn i.alder if alder>=32 & alder<=37

regress log_inntekt utdanning_aar i.kjønn alder if alder>=32 & alder<=37

regress log_inntekt utdanning_aar i.kjønn i.alder karakterpoeng if alder>=32 & alder<=37

Til diskusjon:

- Hva er effekten av ett år mer utdanning på inntekt?
- Hvilken av regresjonene over er "best" til å svare på spørsmålet? Hva er fordeler og ulemper med de ulike oppsettene?

Data explorer kapittel 6: Utdanningsfelt

Videre: Se for eksempel på oppgaven "Data explorer" i Borjas kap. 6 (s. 250). Dette bygger videre på oppgaven over.

Forsøk å svare på denne for et gitt år.

Du kan bruke mye av koden fra forrige avsnitt.

Det er en del forskjeller på IPUMS-datasettet Borjas viser til og det du har tilgang til av norske data. Det er derfor ikke sikkert du får gjort alle de samme begrensningene av datasettene / valgene som er satt opp i oppgaven.

Du trenger ikke bruke "sampling weights" når du bruker registerdata.

Du kan finne fagfelt ved å bruke kommandoen:

generate fagfelt = substr(utdanningskode,2,1)

Kodene finner du i notat 2016/30 her: https://www.ssb.no/utdanning/norsk-standard-for-utdanningsgruppering , side 10-15. Der ser du også hvordan du evt. kan bruke mer detaljerte utdanningsfelt.

Data explorer kapittel 3: Hvem har lav inntekt?

Her kan du begynne med blanke ark igjen (skriv "clear", men lagre gjerne kommandoene som skript først, ref. instruksjonsnotat fra microdata.no)

Kulepunktene under viser hvilke variable / konsepter fra microdata som er aktuelt å bruke i stedet for eksemplene i læreboka. Du må selv velge hensiktsmessige definisjoner av gruppering etter utdanningsnivå, osv.

Grunnlagsvariable ("variables to download"):

- År: Settes ved uttrekk av variable (f. eks pr 31-12-2022)
- Alder, kjønn: Se tidligere oppgave
- Race, hispanic: Ikke relevant i norske data.
    - Men kan se på innvandringsbakgrunn:
    - BEFOLKNING_INVKAT for en kategorisering i innvandrere og ikke-innvandrere
    - BEFOLKNING_FODELAND for en kategorisering etter fødeland
- Paid by the hour: Ikke relevant
- Det er vanskelig å bruke timelønn og antall timer, fordi de er definert på jobb-nivå og ikke på person-nivå. Vi bruker derfor årslønn og stillingsprosent i stedet.
- Årslønn: INNTEKT_LONN
- Stillingsprosent: ARBLONN_PERS_SUM_STILLINGSPST
- Utdanningsnivå: Se tidligere oppgave
- Sampling weight: Ikke relevant

Sample to use: Forsøk å definere denne selv.

Created variables:

- Årslønn observeres i data. Kan justeres for stillingsprosent hvis du ønsker.
- Sett et rimelig nivå på lavlønn. Se for eksempel: https://www.ssb.no/arbeid-og-lonn/lonn-og-arbeidskraftkostnader/artikler/hva-er-vanlig-lonn-i-norge

Gjør "Statistical analysis" og "Questions for discussion".

Eventuelt, om man vil jobbe med timelønn i stedet for årslønn, kan man lese mer om hvordan man setter opp datasettet her:

Koble sammen data om arbeidsforhold på persondatasett - microdata.no

Data explorer kapittel 4: Hvem ble rammet av COVID-19?

Her begynner du med blanke ark igjen (skriv "clear")

- Sett år for første utvalg av data til rett før covid (f. eks januar 2020)
- Bruk månedlig lønn (se dokumentasjonen for hva som er korrekt vairabel, og hent ut per 2020-01-31, 2020-02-28, 2020-03-31, osv), ikke ukelønn som det står i oppgaven
- Trenger heller ikke her bruke "sampling weights"

Gjør "Statistical analysis" og "Questions for discussion"