Økonometri A
Efterår 2022

Mandatory Assignment 2
Integration af Flygtninge på Arbejdsmarkedet

This assignment is the second of three mandatory assignments in Econometrics A. All three assignments must be passed in order to go to the final exam.

The assignment may be answered in groups of max 3 students.

This assignment is similar to the final exam in Econometrics A (DatØk) in the Fall 2020.

Practical instructions to the assignment
Read the entire assignment before you begin to respond, and answer all the questions.

The answer to the assignment must be presented in a comprehensive report with relevant tables and figures. The front page of the report must be based on the template that is available on Absalon. Fill in the names and study ID of all group members on the front page.

Prepare one Python Jupyter Notebook (.ipynb) file that generates all tables and figures that appear in your report. The program should produce tables and figures in the same order as they appear in the report. Comments should clearly indicate which table or figure appearing in the report is being produced. Make sure that the do-file can be executed without any errors. The do-file must include the names and study ID of all group members.

To pass a mandatory assignment, it is required that:

- An adequate response is given to all questions in the assignment.
- More than half of the questions are answered correctly.
- Answers are written in a precise and easy to understand language.

The report must not exceed 8 A4 pages with font size 12, line space 1.5 and page margins of 2.5 cm. This includes the main text, tables and figures in the report, but not the front page.

Uploading your report
Each group must hand-in only one report in total. You must upload two files:

1. The report itself must be uploaded as a PDF file.
2. The Jupyter Notebook must be uploaded as a .ipynb file.

If needed, a free PDF converter can be found here: www.pdf995.com

If group members are assigned to different class teachers, upload the report to the class teacher of your choice. Do not upload the same report to more than ones.

Access to data
The data file Integration.dta contains all relevant information for this assignment. The file can be downloaded from the course website on Absalon.

Documentation of the data (in Danish)

Data indeholder information om 2400 flygtninge, der fik ophold i Danmark i 2014, 2015 og 2016. Vi observerer dem i 2017 og 2018, dvs. der er i alt 4800 observationer i data. Hovedparten er mænd. Variablene i data fremgår af Tabel 1 nedenfor. Alle variable er observeret for alle individer i begge perioder.

Table 1: Variabeloversigt

Variabelnavn	Indhold
id_i	Individspecifikt personnummer
$mand_i$	Dummy variabel for om personen er en mand
$ankomstaar_i$	Året hvor personen fik flygtningestatus i Danmark ("indvandringsår")
$alder_{it}$	Personens alder målt i periode t
$aarslon_{it}$	Personens årsindkomst fra lønnet beskæftigelse målt i periode t
aar_t	Tidsperiode (året for observationen).

Note: Datasættet indeholder konstruerede data og kan derfor ikke bruges til andet end at besvare denne opgave.

Introduction to the assignment (in Danish)

I denne opgave vil vi undersøge, hvordan flygtninges lønninger udvikler sig efter ankomst til Danmark. Vi tager udgangspunkt i de flygtninge, der fik ophold i Danmark i årene 2014, 2015 og 2016, hvor antallet af asylansøgere i Europa var på sit højeste.

Hvordan indvandrere klarer sig i modtagerlandene har været genstand for grundige empiriske analyser i årtier. I de fleste sammenhænge ser man, at beskæftigelsen og lønniveauet for nytilkomne indvandrere ligger væsentligt under det gennemsnitlige niveau for modtagerlandet, men over tid forbedres indvandrenes beskæftigelse og løn - et fænomen økonomer kalder "økonomisk assimilation". Mange studier fokuserer på at estimere "den økonomiske assimilationsrate", som er den årlige konvergens mod niveauet for indfødte lønmodtagere. En central variabel i disse analyser er antallet af år siden indvandring eller "years since migration" (YSM). Det vil også den centrale forklarende variabel i denne opgave. (Vi vil dog kun se på flygtninge og ikke estimere deres udfald relativt til indfødte, så vi kan ikke direkte tale om konvergens, men vi kan se flygtningenes løn forbedres over tid.)

I den første toneangivende artikel om indvandreres integration på arbejdsmarkedet, publiceret i 1978 i det anerkendte tidskrift Journal of Political Economy, estimerede Barry Chiswick en lønregression med YSM, som den centrale variabel. Barrys analyse var baseret på det amerikanske Census fra 1970 og var ret optimistisk omkring indvandrenes økonomiske assimilation. I 1985 publicerede George Borjas en artikel i Journal of Labor Economics baseret på to tværsnit af det amerikanske Census fra 1970 og 1980. Artiklen korrigerer Chiswicks hovedresultat. Ved hjælp af to tidsperioder kunne George Borjas se at "assimilationsraten" var lavere end hidtil antaget. I denne opgave vil vi også estimere en

tværsnitsmodel og en model med to perioder.

I de efterfølgende årtier er modellerne, der bruges til at studere økonomisk assimilation blevet mere rafinerede, og forskere er blevet optaget af mere politikrelevante spørgsmål, såsom hvordan diverse indsatser og reformer på integrationsområdet har på virket integrationen. De mere avancerede modeller for lønassimilation og politikevalueringerne er ofte baseret på metoder, der undervises i de senere økonometrifag. Den grundlæggende struktur er dog ikke langt fra det i ser i denne opgave.

References

Chiswick, Barry (1978). "The Effect of Americanization on the Earnings of Foreign Born Men". Journal of Political Economy, 86(5): 897-921.

Borjas, George (1985). "Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants". Journal of Labour Economics, 3(4): 463-89.

Problem 1 (20%)

a Udfør en deskriptiv analyse af datasættet: Lav en dummyvariabel for hvert ankomstår og beregn gennemsnittet af årslønnen, alder og ankomstår dummyerne separat for mænd og kvinder og for årene 2017 og 2018. Rapporter disse i en tabel og kommenter kort.

Vi ønsker at undersøge, hvordan flygtningenes lønninger udvikler sig efter ankomst til Danmark. Den første model, model (1) nedenfor, er en tværsnitsmodel, hvor i angiver personen. Den afhængige variabel er den naturlige logaritme til årslønnen, $ln(aarslon)_i$, og som forklarende variable benyttes år siden indvandring ($YSM_i = aar_i - ankomstaar_i$) og personens alder, $alder_i$.

$$ln(aarslon)_i = \tilde{\beta}_0 + \tilde{\beta}_1 alder_i + \tilde{\beta}_2 YSM_i + u_i \tag{1}$$

b Konstruer de nødvendige variable og estimer model (1) på tværsnittet for 2017. Rapporter parameterestimater og de relevante standardfejl. Fortolk på samtlige parameterestimater. Har hældningskoefficienterne det forventede fortegn?

c Beregn den estimerede samlede årlige ændring i årslønnen for flygtninge i Danmark på baggrund af estimationen, dvs. $\tilde{\beta}_1 + \tilde{\beta}_2$. Angiv estimatet og standardfejlen af $\tilde{\beta}_1 + \tilde{\beta}_2$. Hint: omskriv model (1) så I kan undersge dette ved direkte at estimere $\tilde{\beta}_1 + \tilde{\beta}_2$.

Problem 2 (20%)

Vi udvider nu model (1) ved at inkludere en variabel for ankomståret samtidig med, at vi bruger nu data fra både 2017 og 2018. Fodtegnet t angiver året, dvs. $t = 2017, 2018$. Vi får dermed model (2).

$$ln(aarslon)_{it} = \beta_0 + \beta_1 alder_{it} + \beta_2 YSM_{it} + \beta_3 ankomstaar_{it} + \varepsilon_{it} \tag{2}$$

a Estimer model (2) med OLS og rapporter parameterestimater samt de relevante standardfejl. Kommenter på resultaterne. Kan vi tillægge parametrene en kausal fortolkning?

b Det er kun i model (2), at vi medtager ankomståret som forklarende variabel. Er det også muligt at kontrollere for $ankomstaar_i$ i model (1), der estimeres for 2017 tværsnittet? Begrund dit svar.

c Antag at model (2) opfylder MLR.1-4, men at vi estimerer model (1). Under hvilken antagelse er OLS estimatoren for $\tilde{\beta}_2$ i model (1) middelret? Hvad er den forventede retning på bias på estimatet for $\tilde{\beta}_2$? Du kan antage i dit svar, at corr($alder$,YSM)=0. Hint: Benyt evt. udregningerne i appendix 3A.4.

Problem 3 (20%)

Vi udvider nu model (2) med alder kvadreret, $alder^2$, og antal år siden indvandring kvadreret, YSM_{it}^2, foruden at vi nu bruger dummyer for indvandringsår 2015 og 2016, $kohorte15_{it}$ og $kohorte16_{it}$ (2014 er det udeladte ankomstår) istedet for $ankomstaar_{it}$.

$$\begin{aligned} ln(aarslon)_{it} &= \gamma_0 + \gamma_1 alder_{it} + \gamma_2 alder_{it}^2 + \gamma_3 YSM_{it} + \gamma_4 YSM_{it}^2 \\ &\quad + \gamma_5 kohorte15_{it} + \gamma_6 kohorte16_{it} + \varepsilon_{it} \end{aligned} \tag{3}$$

a Estimer model (3) med OLS. Rapporter parameterestimater samt de relevante standardfejl. Undersøg den partielle sammenhæng mellem $ln(aarslon)_{it}$ og YSM_{it} - hvornår (målt i år siden ankomst) er lønindkomst i Danmark maksimeret?

b Vi undersøger nu specifikationen af model (3) nærmere. (i) Test simultant om $alder$, YSM og forskellene på tværs af ankomstkohorter udvikler sig lineært. Opstil nulhypotesen og alternativhypotesen baseret på parametrene i model (3), rapporter den relevante teststørrelse og konkluder. (ii) Undersøg om mænd og kvinder kan antages at følge samme model. Beskriv hvordan du gør, din nulhypotese og alternativhypotese, rapporter teststørrelsen og konkluder. Hvis du mener mænd og kvinder ikke følger samme model, kan du fortsætte kun med mænd.

c Undersøg om fejlledet i din foretrukne version af model (3) er heteroskedastisk. Hvilken variabel / hvilke variable driver eventuel heteroskedasticitet? Rapporter relevant(e) test og grafer.

Problem 4 (20%)

Antag at

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \tag{4}$$

og

$$Var(\varepsilon_i | x_i) = \frac{1}{4} \cdot x_{1i}^4 \cdot \sigma^2,$$

hvor $\sigma > 0$ (og ukendt) og $x_{1i} > 0$ for alle i.

a Forslå en WLS estimator, dvs. en transformation af model (4), så fejlledet bliver homoskedastisk. Vis at fejlledet i den transformerede model er homoskedastisk.

b Vis (fremfør de nødvendige og tilstrækkelige argumenter for) at din WLS estimator er middelret og konsistent. Lav de nødvendige antagelser for model (4) og kommenter kort på dem.

c Udregn din WLS estimators varians og argumenter for din WLS er BLUE (brug matrixnotation).

Problem 5 (20%)

I denne opgave vil vi bruge en Monte Carlo simulation til at sammenligne OLS med almindelige standardfejl, OLS med robuste standardfejl og weighted-least-squares (WLS).

Tag udgangspunkt i model (4) og antag at $\beta_0 = 2$, $\beta_1 = 3$, $\beta_2 = -1$ $\beta_3 = -1$, $x_{ji} \sim N(1, 4)$ for $j = 1, 2, 3$ og $\varepsilon_i \sim N(0,\ \frac{1}{4} \cdot x_{1i}^4)$. Det antages at x_1, x_2, x_3 og ε er indbyrdes uafhængige.

a Foretag en simulation med 1000 observationer og 1000 replikationer. Hint: Tag udgangspunkt i Jupyter Notebook'en fra Problem Set 4 eller fra forelæsningerne og ret koden til. For hvert simuleret datasæt estimeres en multipel lineær regressionsmodel som angivet i ligning (4) med OLS og med henholdsvis almindelige standardfejl og heteroskedasticitetsrobuste standardfejl. Rapporter, for begge estimationer, middelværdi og standardafvigelse for estimatet for β_1, standardfejlen på β_1 og andel gange $H_0 : \beta_1 = 3$ afvises i en tabel. Hvor ofte forventer vi at afvise en sand nulhypotese med et 5% signifikansniveau?

   Hvilken betydning har de benyttede standardfejl for:

   i. Estimatet for β_1?
   ii. Standardfejlen for β_1?
   iii. Andelen af simulationerne vi afviser den sande nulhypotese $H_0 : \beta_1 = 3$?

   [Hint 1: Husk, når du bruger `numpy`-funktionen `np.random.normal()` til at simulere normalfordelte variable, så skal du give funktionen middelværdien som det første argument, og standardafvigelsen (ikke variansen!) som det andet argument.]

   [Hint 2: Når du har kørt en regression i statsmodels, kan du bruge `.params` egenskaben i dit resultat-objekt til at tilgå parameterestimaterne. Tilsvarende kan du tilgå de ikke-robuste standardfejl med `.bse` og de robuste standardfejl med `.HC1_se`]

b Foretag en simulation med 50 observationer og 1000 replikationer og undersøg om dine konklusioner vedrørende (i)-(iii) i det foregående spørgsmål ændrer sig. Rapporter igen dine resultater i en tabel.

c Simuler data på ny både med 50 og 1000 observationer for at undersøge fordelen ved at benytte WLS i forhold til at benytte OLS med robuste standardfejl. Rapporter også disse resultater og kommenter på, hvad du finder.