# Migration and the Labour Market: Outline of Research Question and Method

Prof. Dr. Herbert Brücker

Otto-Friedrich-University of Bamberg

Projectseminar, Meeting May 24, 2012

# Contents of today's meeting

- Repetition of last meeting
- Analysis of Borjas (QJE 2003) model
- A primer on OLS estimaton
- Description of data sets

## Repetition of last meeting

- Spatial correlation approach
  - Endogeneity problem
  - Natural experiments (Card, ILRR 1990)
  - Regional regressions (Borjas/Freeman/Katz, AER 1997)
- National level studies
  - Using the variance of foreigner share across education-experience cells
  - Simple regression analysis with dummy variables as controls
  - Estimation of parameters of structural model
  - Borjas (2003); Aydemir/Borjas (2006); Ottaviano/Peri (2006, 2008)
  - The Borjas (2003)/ Ottaviano/Peri (2012) controversy: Are natives and foreigners perfect substitutes

# Repetition of Borjas' model

- Definition of labour supply shock:

$$m_{ijt} = M_{ijt}/(M_{ijt} + N_{ijt}) \tag{1}$$

- Wage equation:

$$w_{ijt} = \theta m_{ijt} + \delta_i + x_j + \pi_t + (\delta_i \times x_j) + (\delta_i \times \pi_t) + (x_j \times \pi_t) + e_{ijt} \tag{2}$$

- Unemployment equation:

$$u_{ijt} = \theta m_{ijt} + \delta_i + x_j + \pi_t + (\delta_i \times x_j) + (\delta_i \times \pi_t) + (x_j \times \pi_t) + e_{ijt} \tag{3}$$

# Definition of variables (I/II)

- $w_{ijt}$ = wage of education group $i$ and experience group $j$ at time $t$ (in natural logarithms)
- $u_{ijt}$ = unemployment rate of education group $i$ and experience group $j$ at time $t$
- $m_{ijt}$ = migration rate in education cell $i$ and experience cell $j$ at time $t$
- $\delta_i$ = vector of fixed effects (= dummy variables) for each education group;
- $x_j$ = vector of fixed effects (= dummy variables) for each experience group; $\pi_t$ = vector of fixed time effects;
- $e_{ijt}$ = error term;
- $i = 1...4$ = index for education group;
- $j = 1...8$ index of experience group.

# Definition of variables (II/II)

- $\theta =$ is the slope parameter which measures the impact of immigration. That is what we are interested in.

- $e_{ijt} =$ is the error term (disturbance term). If the model is properly specified, the error term should be 'white noise', i.e. it should be normally distributed, have an expected value of zero and constant (finite) variance.

- Miscalleneous: The migration rate $m_{ijt}$ is defined as $M_ijt/(N_{ijt} + M_{ijt})$, where $M_{ijt}$ is the migrant labor force and $N_{ijt}$ the native labor force in education cell $i$ and experience group $j$ at time $t$.

# What is a fixed effect / dummy variable?

- Consider a a vector of dummy variables $\delta_j$ for an education group ($j = 1, 2, 3$)
- $\delta_1$ has a value of 1 if an individual/cell has education level 1, and zero if otherwise
- $\delta_2$ has a value of 1 if an individual/cell has education level 2, and zero if otherwise
- $\delta_3$ has a value of 1 if an individual/cell has education level 3, and zero if otherwise
- Considering education fixed effects /dummy variables in our model allows to control for the impact of education on the wage or the unemployment rate
- The same is true for experience and time fixed effects /dummy variables

## What is an interaction dummy?

- The Borjas' (2003) model considers also education-experience ($\delta_i \times x_j$) and education-time ($\delta_i \times \pi_t$) and experience-time ($\delta_i \times x_t$) dummies. These dummy variables are called interaction dummies.
- $\delta_i \times x_j$: This interaction dummy has a value of 1 if an individual/cell has education $i$ and experience $j$ and of zero if otherwise.
- Interaction dummies control for the interaction of two effects, e.g. for the interaction of education with time in case of skilled-biased technological progress.

## What are the effects of dummy variables?

- Considering dummy variables for education, experience and time is a save way to control for all influences which emerge from different education and experience levels of individuals and all macro effects which vary over time, e.g. the business cycle.

- Considering interaction dummies enables furthermore to control for all effects which result from the interaction between these effects.

- Altogether, this enables us to identify the pure effects of immigration without any disturbing effects.

# Can we think of alternative specifications? (I/III)

- Wage equation for natives:

$$w_{Nijt} = \theta_N m_{ijt} + \delta_i + x_j + \pi_t + (\delta_i \times x_j) + (\delta_i \times \pi_t) + (x_j \times \pi_t) + e_{ijt} \tag{4}$$

- Wage equation for migrants:

$$w_{Fijt} = \theta_F m_{ijt} + \delta_i + x_j + \pi_t + (\delta_i \times x_j) + (\delta_i \times \pi_t) + (x_j \times \pi_t) + e_{ijt}, \tag{5}$$

- where $N$ is an index for natives and $F$ for immigrants.
- Borjas (2003)' hypothesis: The slope parameter is in both regressions identical. The Ottaviano/Peri (2012) hypothesis: the slope parameter differs. (Why?)

# Can we think of alternative specifications? (II/III)

- Alternatively we can estimate:

$$
\begin{aligned}
w_{kijt} &= \theta m_{kijt} + \theta_F m_{kijt} + \delta_i + x_j + \eta_k + \pi_t \quad (6) \\
&+ (\delta_i \times x_j) + (\delta_i \times \eta_k) + (\delta_i \times \pi_t) \\
&+ (x_j \times \eta_k) + (x_j \times \pi_t) + (\eta_k \times \pi_t) + e_{ijt}
\end{aligned}
$$

- where $\eta_k$ is a dummy for natives and foreigners, respectively, and $k = 1, 2$ an index for national origin.
- The Borjas' hypothesis is here that $\theta_F = 0$. If $\theta_F \neq 0$ and this is statistically significant, than natives and foreigners are imperfect substitutes.
- Note that the $\eta_k$ and the interaction dummies where national origin is involved have also an interpretation: The nationality should be insignificant if natives and foreigners are perfect substitutes.

# Can we think of alternative specifications? (III/III)

- We can also test whether the slope parameters differ by education or experience groups, e.g.

$$
\begin{aligned}
w_{ijt} &= \theta m_{ijt} + \theta_i m_{ijt} + \delta_i + x_j + \pi_t \\
&+ (\delta_i \times x_j) + (\delta_i \times \pi_t) + (x_j \times \pi_t) + e_{ijt}
\end{aligned}
\tag{7}
$$

- where $\theta_i$ is the slope coefficient by education group $i$.

## Further issues

- The Borjas' (2003) model is based on the assumption that immigration in each education-experience cell is exogenous.

- This is only true if immigration is not affected by wage differences across education-experience groups in a country.

- This contradicts the Roy-model, which would predict that better qualified would move into a country with a high wage premium for skills and vice versa.

- Instrumental variable estimation can control for this potential endogeneity.

## What is instrumental variables (IV) estimation?

- The problem: Theoretically we assume that all right-hand-side (explanatory) variables are exogenous. This means that the migration rate should affect the wage rate, but the wage rate should not affect the migration rate. This is however often not the case.

- Definition of an instrument: An instrumental variable is a variable which is correlated with the explanatory (right-hand-side) variables, but not with the explanandum (the left-hand-side variable), i.e. in our case: a variable which is correlated with the migration rate, but not the wage rate.

# Instrumental variables (IV) estimation (I/III)

- Intuition of IV estimation: You first regress the explanatory variable (e.g. the migration rate) against the other explanatory variables plus the instrumental variables. Since these variables are exogenous, the 'estimated' migration rate is exogenous as well. Then you use the 'estimated migration rate' in a second step to estimate the 'true' impact of migration in wages.
- This is called Two-Step-Least-Squares"(2SLS) estimation.

# Instrumental variables (IV) estimation (II/III)

- Criteria for instruments:
    - The instrument should be not correlated with the LHS variable. We cannot directly test for this, but indirect tests like Hanson's $J$ test for overidentification provide indirect evidence for this. At the end you have to argue that your instrument is valid. This is very often controversial.
    - Your instrument should by strong, i.e. it should be strongly correlated with the exogenous explanatory variable (e.g. the migration rate). This can be tested (e.g. the Kleibergen-Pap Test statistics).
    - The model should not be underidentified, again you can test for this.
    - Fortunately, your Software provides the test results asap.

# Instrumental variables (IV) estimation (III/III)

- What is the key challenge?
- You have to find good instruments, i.e. instruments which are valid and strong.
- In our case the migration rate in each education-experience cell has to be instrumented.
- Sometimes lagged migration rates are used, but it is highly questionable whether lagged values are really valid instruments.
- Potentially exogenous are sending-country variables, e.g. income and unemployment rates (ideally by education and experience).
- We can perhaps estimate a potential migration variable by education-expierence fixed effects and average wage and unemployment rates of main sending countries and then create a potential migration variable which we use an IV.

## Shall we do IV estimation?

- This is not required, the entire literature does not use IV estimation in this case.
- Nevertheless, it might be an interesting step forward.
- It might be an interesting exercise for the class.
- The main issue is the data collection for the IVs.
- We discuss later in the seminar whether we have time to do so.

# The next steps ahead of us

- Literature review (done?)
- Specification of empirical model (by and large done: Read once more Borjas, 2003)
- Collection and provision of data
- Getting familiar with STATA
- Loading data into STATA
- Generation of Dummy-Variables in STATA ("gen" command in STATA)
- Description of data set and descriptive statistics
    - Descriptive statistics of variables of model ("sum" command in STATA)
    - Producing graphs and tables ("graph" command in STATA)
- Carrying out regressions ("regress" and "xtregregresscommand in STATA)
- Analysis of regression output
- Perhaps: Collecting data for IV estimation
- Carrying out IV estimation

## The data sets

- Denmark: IDA dataset
    - Administrative data derived from social security and tax records
    - Covers entire labor force. Sample period: 1990-2006.
- Germany: IEB (Ïntegrierte Erwerbsbigrafie") dataset
    - Administrative data derived from social security records and pension data
    - 5% sample. Sample period: 1992-2009.
- UK Labor force survey.
    - Quarterly survey data
    - 60,000 households. Sample period: 1993-2009.

# The data sets: problems and definitions

- Identification of foreigners
    - Danish and UK data define foreigners by foreign-born concept
    - German data define foreigners by nationality
        - Control for naturalizations by treating individuals as foreigners if they are reported as foreign nationals in first spell
        - Ethnic Germans (SSpätaussiedler") identified by programme participation
        - 2nd and 3rd generation immigrants remain in data if they possess foreign citizenship when joining the labor market
- Other issues
    - Males and females considered in all three data sets
    - Only full-time employed considered since German data do not provide hourly wage information
    - Censored wage information imputed in DE
    - Missing education information imputed in DE

# Description of Variables (I/III)

- $L$ = Employed Persons
- $U$ = Unemployed Persons
- $N$ = Labour force (Ërwerbspersonen") ($L+U=N$)
- $u$ = Unemployment rate
- $w$ = wage
- wagesum = Lohnsumme
- year = Jahr
- trend = linear deterministic time trend

# Description of variables (II/III)

- $H$ = Natives
- $F$ = Foreigners
- $q$ = education group
- $k$ = experience group
- $t$ = time index

# Description of Variables (III/III)

- ed = education group (ßtring"(text) variable)
- group = index for education group (1,2,3,4)
- ex = index for experience group (1, 2 ... 8)
- index = index for education-experience group (1,2 ... 32)
- Useful for generation of dummy variables

# Next Meeting

- , Thursday, Nay 31, 2012, 15:00 hours
- **Room RZ/00/006** (Wirtschaftsinformatik).