Estimating equivalized personal income for school students in Austria

Dominik Ernst

Johannes Gussenbauer

uRos 2024, Athens, 2024-11-29

www.statistik.at

Independent statistics for evidence-based decision making

STATISTICS AUSTRIA
The Information Manager

Setting

- Aim: Ministry of Education wants info on socio-economic background of school students
- We have:
  - Micro data for every person in the formal education system in Austria
    - Basic biographic info: age, gender, nationality, language, type of education
  - Micro data in registers on the Austrian population such as..:
    - Student → parent relationship
    - Highest attained level of education
    - Employment status
    - Migration (to/from Austria), country of birth
    - Income from tax statements (yearly net income + transfer income)
- All data sources individually linkable
  - (well, mostly..)

The problem

- Data is linked by unique personal identifiers (bPK-AS)
- Except no income data for persons with only transfer income
  - i.e. unemployment benefits, family allowance, pension, ..)
- Missingness not evenly distributed among citizens (students in our case)
- Disproportionally affects students..
  - who are immigrants and/or non-Austrian citizens
  - from lower income classes
- Possible bias!

Equivalized personal income

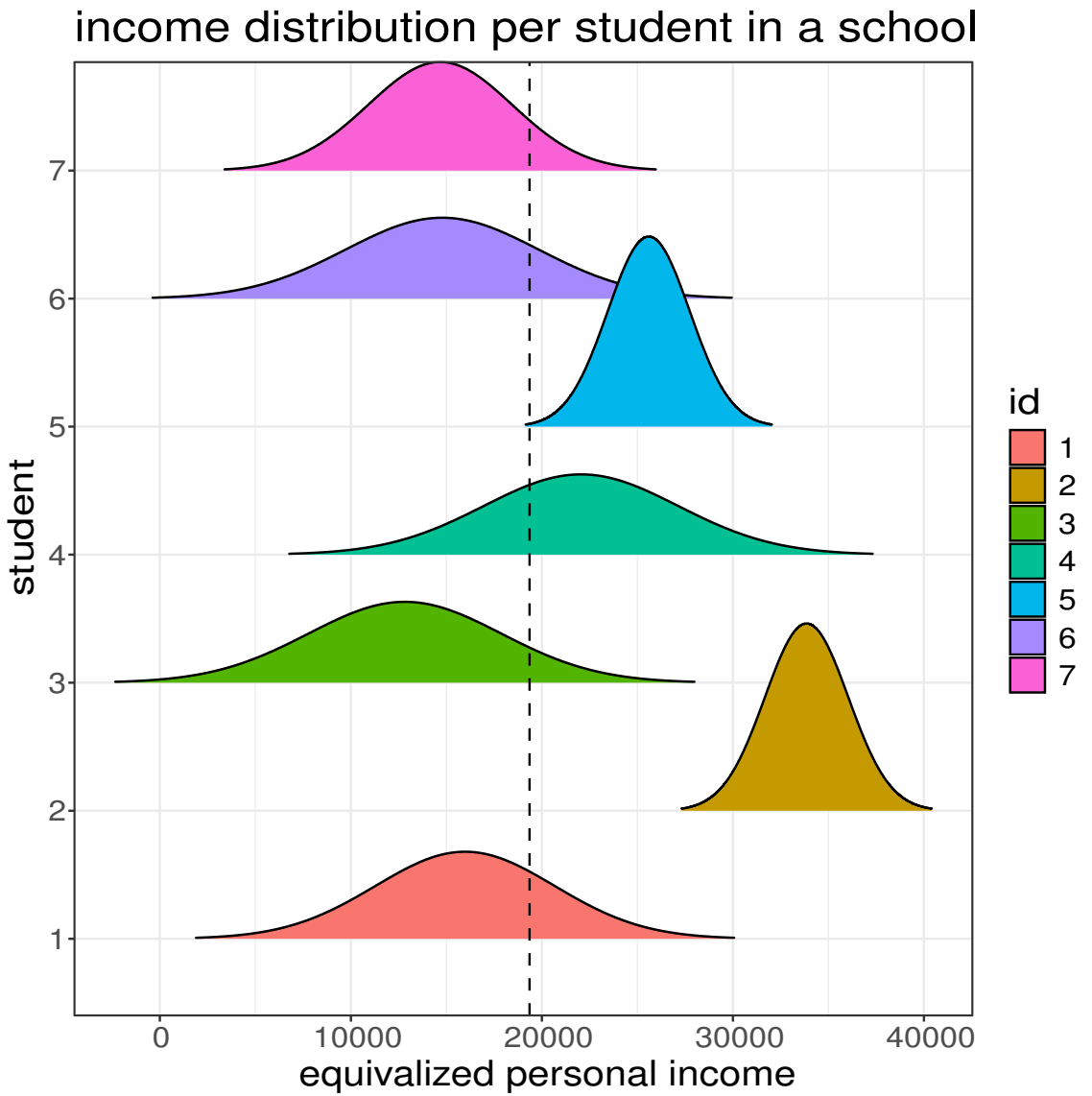
- Income that a student would have to have if they were single and adult to allow for the same equivalent standard of living as in the current household
- Pros:
  - Respects household structure (number of adult persons, children per household)
  - Respects cost of living
- Surveyed via EU-SILC
- Sample data from EU-SILC also individually linkable with registers
- (only for residents in Austria ofc)

Income estimates

- ML model on EU-SILC data + registers (xgboost)
  - Dependent: personal equivalized income
  - Students: age, gender, nationality, country of birth, language, type of school, household status (two parents, single parent)
  - Parents: age, gender, nationality, country of birth, education, employment status, income from tax statements (if available)
  - Household: size, federal state, political district, aggregated income
- Predict income on student population
  - Estimated equivalized personal income for each student
  - Better coverage than with income data alone
- Cons: estimates with limited information, might be less accurate

Bootstrap variance for equivalized income

- Bootstrap on EU-SILC replicate weights (100 samples)
  - Estimated income distribution on an individual level
  - Assumed as normal with mean and variance computed from bootstrap samples



Usage: Lower 20% income quantile

Computation of a “poverty indicator” used by the Ministry of Education

- Ratio of students with lower income than the (global) 20% quantile
- ..within a school (would be possible for different populations too)
- Soft classification using normal distributions above:
  - Probability for category “income < 20% quantile” is value of distribution function at the 20% quantile
  - Sums of probabilities yield expected number of students under 20% qu. for each school
  - Takes estimation variance into account
  - Takes into account how close students are to classification boundary

Applications

- Data dissemination for the Ministry of Education (as required by law)
  - Summary statistics per school
    i.e. gender, nationality, language, education of parents, employment status of parents, ratio of students under the 20% income quantile (see earlier)
- Study on factors benefitting graduating a secondary school in Austria
  - Personal equivalized income as control variable
  - High impact even when controlling for education and status of employment (of parents)
- National Report on Education 2024 (“Nationaler Bildungsbericht 2024”) by Ministry of Educ.
  - Ratio of students with equivalized income < 20% quantile
    Dependent on e.g. migratory status, country of origin, …
  - To be published in December 2024

Notes on implementation

- Whole project is an R package
  - No namespace conflicts, easy code reloading (one keyboard shortcut)
- (Almost) no scripts, only functions
  - Functions provide clear boundaries and interfaces between different parts of code
- Referential transparency
  - Functions have no side effects, result only depends on parameters (were applicable)
  - Easier reasoning and testing of individual functions
- Entire data pipeline is just one long pipe in R
  - From loading authentic data, processing, and saving results

Please address queries to

Dominik Ernst

dominik.ernst@statistik.gv.at

STATISTIK AUSTRIA

Guglgasse 13, 1110 Wien