Harmonising socio-economic data

Claire Crawford
Our remit

• To harmonise measures of socio-economic status across studies
• Which measures?
  – Education, social class, income (of parents and children)
• Which studies?
  – The UK’s four national birth cohort studies:
    • NSHD (1946)
    • NCDS (1958)
    • BCS (1970)
    • MCS (2000/2001)
  – Plus ALSPAC and BHPS/Understanding Society
Why harmonise socio-economic data?

- To answer substantive research questions in which some measure of socio-economic status is the outcome or the main control of interest
  - **Educational inequalities**: how much more likely is someone from a high SES background to go to university than someone from a low SES background, and how has this changed across cohorts?
  - **Social mobility**:
    - Intra-generational: how likely is someone who starts working in a particular job or with a particular income level to move up or down the social spectrum as they get older?
    - Inter-generational: how predictive is SES of parents in childhood (variously defined) of SES in adulthood, and how this has changed across cohorts?
- To use as a control variable in studies comparing changes over time in other outcomes, e.g. links between childhood and adult health
What are the issues with harmonisation?

• Examples of conceptual issues:
  – The % of the population falling into each group changes over time
    • Expansion of HE means many more now go to university: is it an equally informative measure when 5% or 50% of people have degrees?
    • Structure of economy has changed dramatically: does it mean the same to have a manual job now as it did 50 years ago? What about female labour force participation?
  – Definitions change over time
    • Is having 5 GCSEs at grades A*-C the same as having 5 O-levels at grades A-C?
    • Different benefits and tax credits are introduced and taken away over time: what should be included in a total measure of family income?
What are the issues with harmonisation?

- Examples of data-related issues:
  - Ideally want measures taken at the same age using the same questions
  - But rarely happens in practice, so need to check what is available when
  - Often means harmonising lowest common denominator
  - Whether that is acceptable depends on your question of interest
Case study: harmonising income

• Want to create a continuous measure of total net family income
• What information do we have? (e.g. at age 16)
• NCDS:
  – Continuous estimate of net earnings from main job
  – Continuous estimate of partners’ net earnings
  – Continuous estimate of other income
• BCS:
  – Banded estimate of gross total income
• What can we get from this?
Imputing within band: age 16 BCS data

- We would like to know where within the band each family falls
- Use another dataset in a similar year to learn about this
- Family Expenditure Survey asks respondents for continuous income
  - Also includes a number of covariates which feature in BCS
    - e.g. parents’ age, education, work status, social class
- Split FES sample into income bands similar to BCS
- Regress income within band on different combinations of covariates for an 80% sample and use to predict income for other 20%
- Compare prediction to actual income reported to find best fit
- Apply final regression specification to BCS data
Other issues to be overcome

- Removing income of non-family members in household
  - Predict the share of income from non-family members using an alternative dataset and remove

- Imputing missing values of individual components
  - But only where confident about value (e.g. child benefit)

- Top-coding reported values
  - Judgement call; undertake sensitivity analysis to check implications

- Changing gross to net income
  - Impute tax rate using knowledge of tax code each year
An application: estimating social mobility

• Previous research has used the cohort studies to estimate whether Britain has become more or less socially mobile over time
• Economists estimate intergenerational income or earnings mobility
  – Link between parental income/earnings and own income/earnings
• A typical model looks like this:

\[ \log(\text{Income}_i) = \beta_0 + \beta_1 \log(\text{ParInc}_i) + \beta_2 \text{Age}_i + \beta_3 \text{Age}^2_i + \epsilon_i \]
  – where \( \beta_1 \) is the coefficient of interest and gives the percentage increase in income in adulthood resulting from a 1% rise in income in childhood
Estimating intergenerational income mobility

- Previous UK studies (e.g. Blanden et al., 2004) have estimated link between family income at age 16 and sons’ earnings in adulthood, and shown how estimates of $\beta_1$ changed between NCDS and BCS.

- The new harmonised measures will add to our knowledge by:
  - Enabling us to provide first UK estimates of link between income in childhood and income in adulthood (and how it has changed over time).
  - Allowing us to break down these correlations into different components of income, to investigate roles of:
    - Partnership and assortative mating
    - The tax and benefit system
  - Paving the way for us to include groups that are omitted from the majority of previous studies – e.g. women and the unemployed.
Moving towards net family income

- Start from the specification estimated by existing UK literature, i.e. gross earnings as an outcome and sample of employed males only
- Break down net family income into its constituent parts and estimate relationship between parents’ income and components of income
- Total net family income = Gross Earnings + Partners gross earnings + Other Income + Benefit Income - Taxes
  - Tells us something about partnership formation/assortative mating
  - Tells us something about redistribution
- Add individuals omitted by previous studies (women and unemployed)
RESULTS
Conclusion

• Harmonisation vital to answer some research questions
• Ideally questions would be designed with comparability in mind
  – Trade-off between comparability with past (less good?) measures and better data that could be the starting point for comparability in future
• Is the lowest common denominator sufficient for your purposes? Could this be bettered?
• Seek advice from researchers who understand underlying measures
  – Medics may be best equipped to harmonise blood pressure measures
  – Sociologists to harmonise social class measures
• Always robustness check your results
• And be upfront about the assumptions and limitations of your choices
Education

• Parents’ education
  – Age left full-time education (NSHD, NCDS, BCS, MCS)
  – Highest qualification (BCS, ALSPAC, MCS)

• Cohort members’ education
  – NSHD, NCDS, BCS, ALSPAC
  – Grade A*-C at O-level/CSE/GCSE in Maths and English
  – Number of O-levels/CSEs/GCSEs at Grades A*-C
Income

- Continuous total net family income
  - When CM was age 10-12 in BCS, ALSPAC, MCS and BHPS/US
  - When CM was age 16-18 in NCDS, BCS, ALSPAC and BHPS/US
  - When CM was age 33/34 and 42 in NCDS and BCS
Social class

• 1990 Registrar-General’s social class
  • For fathers when cohort member was age 10/11
    – NSHD, NCDS, BCS, MCS
    – (ALSPAC at age 8)
    – (BHPS/US reported by cohort members asked to think back to age 14)
• Cohort members at age 42/43
  – NSHD, NCDS, BCS