Imputation of Annual Income in Household Panel Surveys with Partially Non-Responding Households

Joachim R. Frick, Markus M. Grabka, Olaf Groh-Samberg (SOEP, DIW Berlin)

ESRA 2009 Conference
Warsaw, 29/06 - 3/07 2009
Motivation

• Typical applications using household panel survey data:
  A) Cross-sectional perspective:
    • Welfare economics: “pooling and sharing” of economic resources across all HH members
    • Labor economics: interaction between HH members, e.g. labor supply decisions of couples
  B) Longitudinal perspective:
    • Tracing individuals and their respective household context over time (mobility)

→ Collecting complete data on all HH respondents is crucial
Concept of HH Survey

HH Questionnaire given by Adult A: Info on HH and

Child a

Child b

Personal Questionnaire: Adult A

Personal Questionnaire: Adult B

Personal Questionnaire: Adult C

completely realised HH
Concept of HH Survey

- HH Questionnaire given by Adult A: Info on HH and
  - Child a
  - Child b
- Personal Questionnaire: Adult A
- Personal Questionnaire: Adult B
- Personal Questionnaire: Adult C
Concept of HH Survey

HH Questionnaire given by Adult A: Info on HH and

Personal Questionnaire: Adult A

Personal Questionnaire: Adult B

Personal Questionnaire: Adult C

Child a

Child b

partially realised HH –
partial unit non-response
Motivation

• **Partial Unit-Non-Response (PUNR)**
  • Specific type of non-response occurring only in household (panel) surveys *prohibiting proxy-interviews for adults*
  • PUNR=Incomplete coverage of multi-respondent households *(no problem for single HH)*

**Income Analysis: PUNR = underestimation of household income aggregates !!!**

• **Open questions**
  • How can we deal with missing income information due to PUNR, esp. in panel surveys ?
  • Does the choice of the treatment of PUNR affect cross-sectional and longitudinal analyses ?
Dealing with PUNR

- **Version 1:** ignoring PUNR for income measure, only
  \[ Y(\text{PUNR})=0 \quad \text{and} \quad \text{Needs}(\text{PUNR})>0 \quad \text{assumption:} \quad Y(\text{PUNR})=0 \]

- **Version 2:** deleting non-responding individuals from PUNR-households by adjusting equivalent scale
  \[ Y(\text{PUNR})=0 \quad \text{and} \quad \text{Needs}(\text{PUNR})=0 \quad \text{assumption:} \quad Y(\text{PUNR}) \sim Y(\text{noPUNR}) \text{ within PUNR-HHs} \]

- **Version 3:** deleting PUNR-households altogether and re-scaling of weights for households at risk of PUNR
  \[ \text{PUNR-HH}=. \quad \text{assumption:} \quad Y(\text{PUNR-HH}) \sim Y(\text{noPUNR-HH} \mid \#\text{adults}>1) \]

- **Version 4:** imputation of annual incomes of PUNR
  \[ Y(\text{PUNR})>0 \quad \text{and} \quad \text{Needs}(\text{PUNR})>0 \quad \text{assumption:} \quad Y(\text{PUNR})=f(X)+e \]
Outline

1. Incidence and Selectivity of PUNR in the German SOEP
2. The Imputation Process
3. The Impact of PUNR-Treatment on inequality & poverty
4. Conclusion
1. Incidence and Selectivity of PUNR in the German SOEP
The German SOEP

• started in 1984 incl. oversampling of migrants
  – additional samples started in 1990 (East), 1994/95 (new immigrants), 1998/2000/2006 (refreshment and innovation samples), 2002 (high income sample)

• annual interviews
  – household
  – all individual household members (17+)
  – mixed mode: PAPI, CAPI, self-administered

• more information: http://www.diw.de/soep
Incidence of PUNR (%)
### Selectivity of PUNR – Results from a pooled probit regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex female</td>
<td>-0.443***</td>
</tr>
<tr>
<td>Age groups (ref. 56-65 yrs of age)</td>
<td></td>
</tr>
<tr>
<td>17-24</td>
<td>-0.003</td>
</tr>
<tr>
<td>25-40</td>
<td>-0.053*</td>
</tr>
<tr>
<td>41-55</td>
<td>-0.210***</td>
</tr>
<tr>
<td>66 and over</td>
<td>-0.304***</td>
</tr>
<tr>
<td>Relation to household head (ref: household head)</td>
<td></td>
</tr>
<tr>
<td>Partner</td>
<td>0.977***</td>
</tr>
<tr>
<td>Child</td>
<td>0.984***</td>
</tr>
<tr>
<td>Other</td>
<td>1.700***</td>
</tr>
<tr>
<td>Migration background (ref: autochthonous)</td>
<td></td>
</tr>
<tr>
<td>Native born foreigners</td>
<td>0.204***</td>
</tr>
<tr>
<td>Foreign born Germans</td>
<td>-0.264**</td>
</tr>
<tr>
<td>Foreign born foreigners</td>
<td>-0.347***</td>
</tr>
<tr>
<td>East Germany (ref: West Germany)</td>
<td></td>
</tr>
<tr>
<td>East Germany</td>
<td>-0.245***</td>
</tr>
<tr>
<td>Change in household composition (ref: no change)</td>
<td></td>
</tr>
<tr>
<td>Change</td>
<td>0.076***</td>
</tr>
<tr>
<td>First wave</td>
<td>0.528***</td>
</tr>
<tr>
<td>No. of adults in household (ref: 2)</td>
<td></td>
</tr>
<tr>
<td>3 adults</td>
<td>0.052</td>
</tr>
<tr>
<td>4+ adults</td>
<td>0.125***</td>
</tr>
<tr>
<td>No. of children in household (ref: none)</td>
<td></td>
</tr>
<tr>
<td>1 child</td>
<td>-0.078***</td>
</tr>
<tr>
<td>2 children</td>
<td>-0.163***</td>
</tr>
<tr>
<td>3+ children</td>
<td>-0.170***</td>
</tr>
<tr>
<td>Home owner (ref: tenant)</td>
<td></td>
</tr>
<tr>
<td>Home owner</td>
<td>-0.051**</td>
</tr>
<tr>
<td>Highest education of hh head/partner (ref: intermediate)</td>
<td></td>
</tr>
<tr>
<td>Lower secondary</td>
<td>0.039</td>
</tr>
<tr>
<td>Higher secondary</td>
<td>-0.102***</td>
</tr>
<tr>
<td>Tertiary</td>
<td>0.124**</td>
</tr>
<tr>
<td>Item non-response monthly household income (ref: no)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.396</td>
</tr>
<tr>
<td>Control dummies for survey year</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-1.779</td>
</tr>
<tr>
<td>Pseudo-R2</td>
<td>0.1568</td>
</tr>
<tr>
<td>Observations / Individuals</td>
<td>325414 / 45038</td>
</tr>
</tbody>
</table>

Robust standard errors. * significant at 10%; ** significant at 5%; *** significant at 1%


→ Indication for underreporting of economically active HH members
2. Imputation of PUNR
Imputation strategy

Imputation of 6 individual gross income components
(comparable to detailed information as collected in standard questionnaire):
1. labour income (all),
2. pensions (all),
3. unemployment benefits (all),
4. student grants,
5. maternity leave transfers,
6. private transfers

1. Step: Filter imputation (Y>0) using available longitudinal information
2. Step: Imputation of metric value | Filter = 1
   a) longitudinal imputation (Little & Su 1989)
   b) otherwise cross-sectional, regression based imputation
3. Step: Integration of (imputed) gross income components into
   the standard tax and SSC simulation module
   → PUNR-adjusted Post-Government Household Income
1) Filter Imputation ($Y_k > 0$)

- successive conditional regressions
  1. pension
  2. unemployment
  3.-5. maternity benefits, student grants, private transfers
  6. labour income
- binary regression model (probit)
- Dep. Variable: $\text{Filter} = 1$, if $P(Y_k > 0) > \text{Cut-off}$
  - random cut-off is normally distributed with mean 0.5, stdev. 0.2

- RHS-Variables:
  - *Longitudinal information* on receipt of $Y_k$, if at all available
  - *demographic variables*: age-groups, type of hh, no. of children, person in need of care, relation to head of hh, region, community size, SOEP-sample, change in hh-composition
  - *social structure*: home owner, (highest) education of head of hh or spouse/partner, education of children, migration background
  - *income variables*: screener income (net household incomes at month of interview), public transfers (housing subsidies, social assistance, etc.), aggregated observed individual incomes as a share of household screener income
  - *filter(s)* considering receipt of other income components (successive inclusion)
2a) Longitudinal Imputation | Filter=1

- “Row-and-column” imputation (Little & Su, 1989):
  1. row effect (longitudinal: within-person avg. Y position) $r_i = \frac{1}{k} \sum_{j=1}^{k} Y_{ij}$
  2. column effect (x-sectional: avg. yearly income) $c_t = \frac{1}{k} \sum_{i=1}^{k} Y_i$
    -> expected income position of individual $i$ at time $t$: $E_{it} = r_i * c_t$
  3. stochastic term (=deviation from observed value of nearest neighbour $n$)
    -> imputed income $Z_{it} = E_{it} * \left[ \frac{Y_{nt}}{E_{nt}} \right]$

- PUNRs with valid interview(s) within 7 year-window
  - three years before and after non-response
  - *weighted by distance* to the year of missing information
  - applies to about roughly 1/3 of all PUNRs
  - L&S is also the standard imputation procedure for imputation of INR
2b) Cross-sectional Imputation | Filter=1

Imputation of annual income $Y_k \mid Y_k>0$

- OLS-regression, conditional on full set of filters
- Dep. Variable: $\log(Y_k)$
- RHS-variables (see Filter imputation)
- stochastic term: randomly drawn residuals from observed population
- separate estimations for men and women (esp. labour)
Imputation Results: Gross Annual Labour Income

Population Share (in %)
Imputation Results: Gross Annual Labour Income

Avg. Value (in €), if Filter == 1
3. The impact of PUNR-treatment

• Relevant Income Variable: Equivalent post-government income
  – after simulation of direct taxes and SSC
  – including imputed rent
  – applying modified OECD scale (1; 0.5; 0.3)
  – in prices of year 2000
  – Relative income poverty (PL = 60 % of national median income)

• Comparison of four PUNR-treatments
  – 1: ignoring PUNR in income measure
  – 2: adjusting equivalent scale
  – 3: deleting PUNR-households and re-scaling weights for households at risk of PUNR
  – 4: imputation of annual incomes of PUNR
Hypothetical Effects of PUNR on Relative Poverty and Inequality

• any accounting for PUNR yields higher average incomes
  → increase in relative poverty threshold
  → c.p. increasing poverty (risk) among non-PUNR-HH

• versions 2 (adjust equiv. scale) and 4 (imputation) yield higher equivalent income among PUNR-HH
  → c.p. decreasing poverty risk rates for PUNR-HH

→ overall effect on poverty at a given point in time unclear

• however, due to increase in the incidence of PUNR over time, poverty trends will be affected as well

• effects on the socio-demographic structure of poor HH:
  → increase in poverty for all HH not at risk of PUNR (singles, lone parent families) due to increase in Pov-line
Poverty by PUNR-Treatment

Poverty Measures

Survey Year

V1: ignoring
V2: adj. needs
V3: deleting&re-scaling
V4: imputation

FGT(0)

FGT(1)

FGT(2)
Poverty Dynamics by PUNR-treatment

Year-to-Year Poverty mobility by PUNR-treatment (avg. 1985-2007)

V1: ignoring  V2: adj. needs  V3: deleting&re-scaling  V4: imputation

Poverty Mobility (% moving in / out of poverty)
4. Conclusion

• Facts
  – PUNR is of increasing relevance in household panels
  – PUNR is selective
  – *Consequence*: Bias in aggregated household income measures

• PUNR-treatments
  – Ignoring is “far from perfect”
  – Equivalence scale adjustments do not sufficiently correct for this bias
  – Deleting PUNR and re-scaling weights as well as (complex) imputation of missing income components yield similar results
  – However, imputation allows to maintain the entire survey population → most relevant for longitudinal (mobility) research
How to intervene?

• Increase incentives
  – Individual monetary incentive:
    • Effects on third household members: „I also want to be paid (more)“ → expensive in the short run
    • Expectation to be paid more in future waves as well → expensive solution in the long run!
  – Household level incentive for completeness
    eg. see HILDA – add AUS$50
• Proxy-Interviews for individuals with strong limitations (severe sickness, dementia)
• Drop-off questionnaire
• HOWEVER: Proxy-Interviews and drop-off questionnaires must not motivate other household members to also quit the standard interview !!!
Comments welcome

Olaf Groh-Samberg  
ogrohsamberg@diw.de

SOEP-homepage  
www.diw.de/gsoep

SOEP-hotline  
soepmail@diw.de