Institute for Employment Research

The Research Institute of the Federal Employment Agency

Beat the Heap

An Imputation Strategy for Valid Inferences from Rounded Income Data

Jörg Drechsler
(Institute for Employment Research)

&

Hans Kiesl
(University of Applied Sciences Regensburg)

Q2014, Vienna

Income questions in surveys

- Information on income is highly relevant:
  - to measure inequality, discrimination, poverty, etc
  - for political decisions (laws, labor market programs etc)
- Exact information on income is hard to obtain:
  - considered sensitive information (high nonresponse rates)
  - most respondents approximate their income (high probability of rounding)
- Agencies often address nonresponse problem
- Rounding problem is left to the user

Rounding for income questions

- presumably respondents often do not report their exact income
- Czajka and Denmead (2008) find that 28-30% of earners report amounts divisible by $5,000, and 16-17% report amounts divisible by $10,000 in the CPS and the ACS for their income in 2002.
- Rounding problem not limited to yearly income
- example for the monthly income from the panel study "Labor Market and Social Security"

Income divisible by	1,000	500	100	50	10	5
Relative frequency (%)	13.75	23.83	59.79	67.29	78.70	81.60

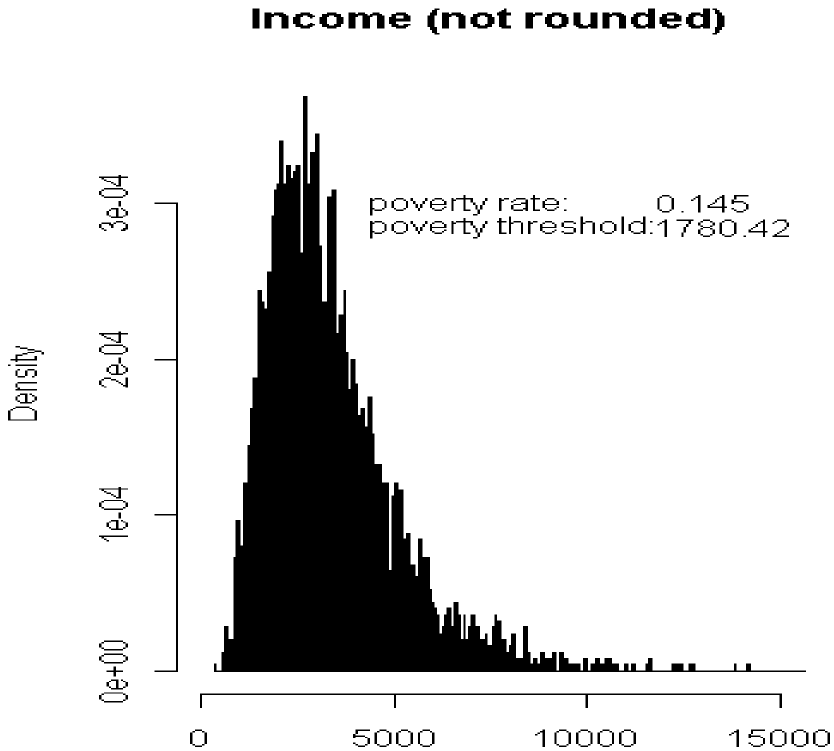
- analyst needs to address the problem to obtain valid inferences

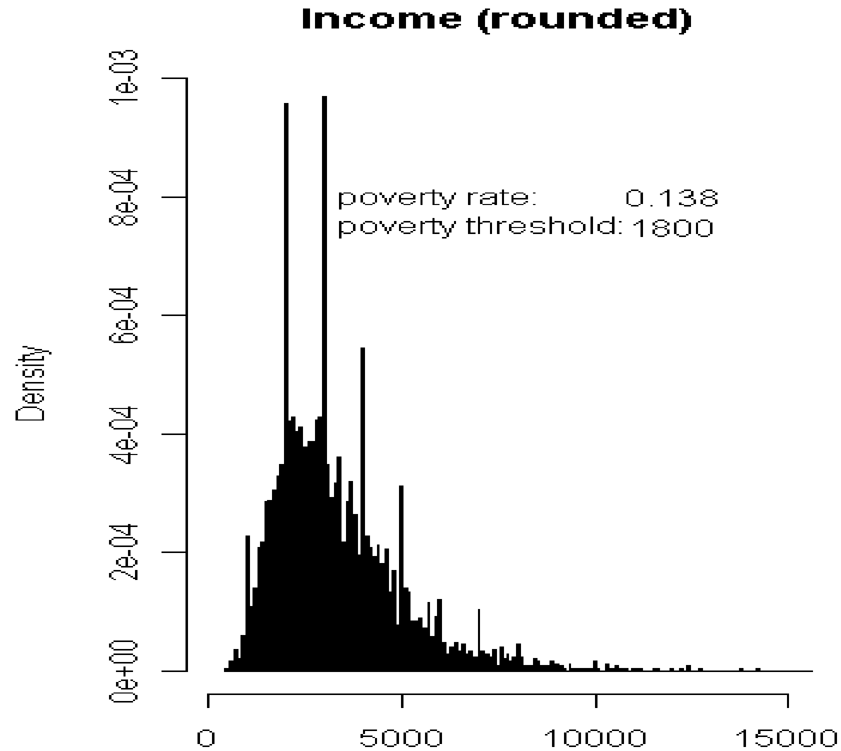
Is rounding problematic?

- affects the marginal distribution
- variance estimate is biased
- affects the quantiles of the distribution

- Illustration
- let $f(\text{family income}) \sim \log N(8, 0.47)$
- let $p_{round}(0, 10, 100, 1000) = (0.1, 0.4, 0.4, 0.1)$
- quantity of interest: Poverty rate (percentage of households with an income < 60% of the median income)

Simulation results





Adjusting for rounding error

- rounding error could be corrected at the analysis stage
- we suggest to address the problem at the data processing stage
- advantages
  - data producer has more information available
  - data user lacks the capacity to deal with the problem adequately
  - data user has own problems to worry about so data deficiencies should be kept at a minimum
  - different data users will get consistent results
- disadvantages:
  - more work for the data providing agency

Rounding error correction through MI

- imputation easy if rounding intervals were known for each record
  - simply impute by drawing from a truncated distribution (Schenker et al. (2006))
- rounding interval is unknown
- rounding interval needs to be estimated
- define the joint distribution for income and the tendency to round
- imputation approach is related to Heitjan and Rubin (1991)
- standard model for income:

$$\ln(inc) \mid X \sim N(\beta_0 + X'\beta_1, \sigma^2)$$

- Probit model for the rounding

$$r \mid \log(inc), Z \sim N(\alpha_1 \log(inc) + Z'\alpha_2, \tau^2)$$

Imputation

- Obtain ML estimates from joint model for income and rounding
- Draw a value from the approximate posterior distribution of the parameters

$$\hat{\Phi} \sim MVN(\Phi_{ML}, \mathrm{I}(\Phi_{ML}))$$

- For given parameter values, impute by rejection sampling:
  1) Draw values for (log(*inc*), *r*) from a truncated bivariate normal with truncation points defined by the maximum rounding interval given the observed data.
  2) Accept drawn values if imputed income is consistent with observed income given the imputed rounding parameter *r.*
  3) Otherwise draw again.

- Repeat everything *m* times

Simulation study

- generate a population based on variables from the panel study “Labor Market and Social Security (PASS)”
- true income distribution in the population needs to be known

$$\log(income) = \alpha + \beta_1 \cdot hhsize + \beta_2 \cdot unemp\_benefits + \beta_3 \cdot age + \beta_4 \cdot livspace + \varepsilon$$

- model rounding behavior
  - assume rounding tendency only depends on income

$$r = \gamma \cdot \log(income) + \varepsilon$$

  - rounding bases $(1, 5, 10, 50, 100, 500, 1000)$
  - rounding behavior can be modeled as a 7 category probit model
  - use $\hat{\gamma}$ and estimated thresholds from the PASS survey to round income in the population

Simulation study

- repeatedly draw simple random samples with $n = 1{,}000$
- impute true income using two different models
  - always assume widest possible rounding interval (naïve approach)
  - estimate rounding probabilities from the data (improved imputation approach)
- generate $m = 5$ imputed datasets for both approaches
- quantity of interest: poverty rate
- repeat whole process of sampling, rounding, imputation and analysis $1{,}000$ times

Simulation results

- poverty rate in the population: 18.46 %

	mean($\hat{pr}$)	Var($\hat{pr}$)	mean($\widehat{Var}(\hat{pr})$)	Variance ratio	95% Coverage rate
True income	18.44	$2.49*10^{-5}$	$2.62*10^{-5}$	1.05	95.3
Rounded income	19.20	$3.27*10^{-5}$	$2.63*10^{-5}$	0.80	67.4
Naive imputation	18.02	$2.20*10^{-5}$	$3.19*10^{-5}$	1.45	92.5
Improved imputation	18.52	$2.34*10^{-5}$	$3.02*10^{-5}$	1.29	97.6

Application to the panel study “Labor Market and Social Security (PASS)”

- household survey that aims at measuring the social effects of labor market reforms
- conducted yearly since 2006
- dual frame survey (approximately 6,000 households in each frame)
  - sample from the Federal Employment Agency’s register data containing all persons receiving unemployment benefits
  - address based sample of the general population
- contains a large number of socio-demographic, employment-related, and benefit related characteristics
- head of household is asked to estimate the total monthly household income

Imputation models

- linear regression model for log(income)
- Explanatory variables:

household size	5 categories
deprivation index	range: 0-21
living space	range: 7-903 square meters
type of household	8 categories
amount of debt	7 categories
income from savings	yes/no (not available for wave 1)
amount of savings	8 categories (not available for wave 1)
age of respondent	range: 15-99
unemployment benefits	yes/no
weight	range: 24.95-186,000

- categories that contain less than 5% of the records are collapsed

Imputation models

- probit model for rounding variable
- Explanatory variable: log(income)
- posterior predictive simulations to evaluate the quality of the models
- only complete cases are included
- starting values for the maximum likelihood estimation from regressions based on the original data
- number of imputations: $m = 25$

Poverty rate before and after correction (95% confidence interval in brackets)

wave	original data	corrected data
wave 1	17.31 (15.79;18.83)	16.35 (15.14;17.55)
wave 2	16.91 (15.76;18.05)	16.98 (15.69;18.27)
wave 3	14.27 (12.22;16.33)	15.40 (13.91,16.90)
wave 4	14.89 (13.64;16.15)	14.61 (13.40;15.81)
wave 5	16.34 (14.80;17.88)	15.75 (14.41;17.10)
wave 6	15.95 (14.42;17.48)	16.27 (14.81;17.72)

Conclusions

- rounding can lead to biased estimates
- addressing this potential bias at the data processing stage can be beneficial
- multiple imputation can be a tool to address the bias problem
- probability for rounding also needs to be estimated
- future work
    - address nonresponse in the variables
    - investigate rounding effects when family income is derived from various components

Institute for Employment Research

The Research Institute of the Federal Employment Agency

Thank you for your attention

joerg.drechsler@iab.de

Setting up the likelihood

- Parameter vector $\Phi = (\beta_0, \beta_1, \sigma, \ \alpha_1, \alpha_2, \tau, k_0, k_1, k_2, k_3, k_4, k_5)$
- Likelihood:

$$L(\Phi \mid \mathbf{z}, inc_{obs}) = \prod_i f(\mathbf{z}_i, inc_{obs,i}) \mid \Phi)$$

$$= \prod_i \iint f_{\ln(inc_{true}),r}(z_i, inc_{obs,i}, j_i, r_i \mid \Phi) dj_i, dr_i$$

$$= \prod_i \iint_{A(obs-inc_i)} f_{\ln(inc_{true}),r}(j_i, r_i, z_i \mid \Phi) dj_i \ dr_i$$

because

$$f(inc_{obs,i} \mid r_i, \mathbf{z}_i, inc_{true,i}) = \delta(r_i, inc_{true,i}, inc_{obs,i})$$

where $A(obs - inc_i)$ is the set of possible values for $(\ln(inc), r)$, determined by the observed income obs-inc_i

Example

- observed income = 850
- possibly rounded to the closest 1,5,10,50 Euros

$$g(\mathbf{z}_i, inc_{obs,i}, \Phi) = \int_{\ln(849.5)}^{\ln(850.5)} \int_{-\infty}^{k_0} f_{\ln(inc),r}(i, r \mid z_i, \Phi)\, dr\, di + \int_{\ln(847.5)}^{\ln(852.5)} \int_{k_0}^{k_1} f_{\ln(inc),r}(i, r \mid z_i, \Phi)\, dr\, di$$

$$+ \int_{\ln(845)}^{\ln(855)} \int_{k_1}^{k_2} f_{\ln(inc),r}(i, r \mid z_i, \Phi)\, dr\, di + \int_{\ln(825)}^{\ln(875)} \int_{k_2}^{k_3} f_{\ln(inc),r}(i, r \mid z_i, \Phi)\, dr\, di$$

Joint model

- Joint model for income and the rounding indicator *r*

$$r, \log(inc) \mid Z \sim N(\mu, \Sigma)$$

with

$$\boldsymbol{\mu} = \begin{pmatrix} \beta_0 + Z'\beta_1 \\ \alpha_1\beta_0 + Z'(\alpha_2 + \alpha_1\beta_1) \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & \alpha_1 \cdot \sigma^2 \\ \alpha_1 \cdot \sigma^2 & \tau^2 + \alpha_1^2 \cdot \sigma^2 \end{pmatrix}$$

Setting up the likelihood

- Parameter vector $\Phi = (\beta_0, \beta_1, \sigma, \ \alpha_1, \alpha_2, \tau, k_0, k_1, k_2, k_3, k_4, k_5)$

- Likelihood:

$$L(\Phi \mid \mathbf{z}, inc_{obs}) = \prod_i f(\mathbf{z}_i, inc_{obs,i} \mid \Phi)$$
$$= \prod_i g(\mathbf{z}_i, inc_{obs,i}, \Phi)$$

with

$$g(\mathbf{z}_i, inc_{obs,i}, \Phi) = \iint_{A(inc_{obs,i})} f_{\ln(inc_i), r_i}(i, r, z_i \mid \Phi) dr\, di$$

where $A(inc_{obs,i})$ is the set of possible values for $(\ln(inc), r)$, determined by the observed income $inc_{obs,i}$

Estimating the poverty rate from the PASS data

- estimated household income is translated into available income as defined by the OECD
- But: income subject to rounding
- Goal: get unbiased results by accounting for the rounding
- Impute “unrounded” data

posterior simulations for the income model

- use parameters from ML estimation
- generate m=1,000 income imputations based on model parameters
- check whether posterior distribution of the imputations for each record cover the reported income value for those records for which the reported income was known not to be rounded
- if imputation model is correct, true (observed) income should be covered in the region [α/2% quantile; 1-α/2% quantile] of the imputed values with a probability of 1-α.
- Compute percentage of records for which this is true and compare with expected percentage

Expected	Empirical Coverage (in %)					
Cov. (in %)	wave 1	wave 2	wave 3	wave 4	wave 5	wave 6
99.00	98.69	94.87	98.03	98.21	96.28	97.94
95.00	95.86	92.96	94.15	94.43	93.75	95.14
90.00	93.11	90.27	90.66	90.06	89.95	90.78

posterior simulations for the rounding behavior model

- re-round imputed data based on estimated rounding probabilities
- generate m=100 imputations of unrounded income
- round each income value k=100 times according to the predicted rounding probabilities
- compare occurrence of “round” values in the original data and the re-rounded data

Income divisible by	5	10	50	100	500	1,000
Observed income (%)	3.51	12.73	8.04	37.34	10.11	13.37
Unrounded income (%)	10.03	8.28	1.15	1.06	0.13	0.27
Re-rounded income (%)	2.64	13.33	9.85	46.64	8.62	9.59