Producing poverty estimates with missing data and measurement error in Kwazulu-Natal, South Africa

Alessandra Garbero
LSHTM
Outline

- Data
- Research aim
- Methodological issues
- Missing data treatment
- Results
- Implications & way forward
Data: ACDIS

Africa Centre Demographic Information System (ACDIS)

- It has been collecting data on more than 11,000 households and 85,000 individuals in part of the Umkhanyakude district of KwaZulu-Natal since the beginning of 2000 (Solarsh et al. 2002; Hosegood and Timaeus 2005 in press).

- Households visited every 6 months to obtain reports of births, deaths, migration and changes in household membership (rounds).

- Detailed socioeconomic data (HSE), including household expenditure data, are collected on alternate visits.

- Verbal autopsies are used to determine causes of death.

- Study population includes all non-resident and resident members.
### Data description (ACDIS)

#### Table 1: Household socio-economic waves, visit dates, rounds and number of households.

<table>
<thead>
<tr>
<th>Visit date</th>
<th>Rounds</th>
<th>HH</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSE1 Feb-Sep 2001</td>
<td>3 or 4</td>
<td>10839</td>
</tr>
<tr>
<td>HSE2 Feb 2003-Aug 2004</td>
<td>8,9,10</td>
<td>10815</td>
</tr>
<tr>
<td>HSE3 Jan-Sep 2005</td>
<td>12</td>
<td>9736</td>
</tr>
<tr>
<td>HSE4 Jan-Aug 2006</td>
<td>14</td>
<td>9338</td>
</tr>
<tr>
<td>HSE5 Jul-Dec 2007</td>
<td>17</td>
<td>10290</td>
</tr>
</tbody>
</table>

**Eligibility criteria for households:**

1. The household must have at least one resident member at a bounded structure within the DSA.

2. Must have household head who is acknowledged by the other members and who is also a member of the household.

Research aims

- Evaluation of the quality of the consumption module in the ACDIS data
  - Addressing measurement error and missing information
- Production of poverty estimates based on consumption for HSE 3
- Build poverty (or welfare) indicator of interest for each socio-economic wave
- Ultimately study changes in consumption-poverty induced by AIDS mortality in Kwazulu-Natal
Brief introduction

Methodological issues
- Attrition: typical of longitudinal data

Relevant to the calculation of a consumption-based welfare indicator:

Survey Design issues: Measurement error in the welfare proxies (consumption)
- Zeros
- Missing values

→ Using this data won’t help discriminating poor/non-poor
→ Overestimation of poverty
# HSE 3 – Consumption module

No. hh 9398

<table>
<thead>
<tr>
<th>Food expenditure items</th>
<th>No. zeros</th>
<th>No. missing</th>
<th>% zeros</th>
<th>% missing values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mealie meal (maize)</td>
<td>43</td>
<td>167</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Rice</td>
<td>460</td>
<td>671</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Beans</td>
<td>678</td>
<td>252</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Samp</td>
<td>4854</td>
<td>125</td>
<td>52</td>
<td>1</td>
</tr>
<tr>
<td>Flour</td>
<td>3287</td>
<td>150</td>
<td>35</td>
<td>2</td>
</tr>
<tr>
<td>Cooking oil</td>
<td>211</td>
<td>139</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Sugar and Salt</td>
<td>187</td>
<td>328</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Tea and Coffee</td>
<td>583</td>
<td>1487</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>Milk</td>
<td>3427</td>
<td>455</td>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>Vegetables</td>
<td>345</td>
<td>1713</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>Meat</td>
<td>225</td>
<td>2405</td>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>Bread</td>
<td>1543</td>
<td>2508</td>
<td>16</td>
<td>27</td>
</tr>
<tr>
<td>Tinned goods</td>
<td>3453</td>
<td>1346</td>
<td>37</td>
<td>14</td>
</tr>
<tr>
<td>Soap</td>
<td>575</td>
<td>749</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Non-food items</td>
<td>No. zeros</td>
<td>No. Missing</td>
<td>% zeros</td>
<td>% missing values</td>
</tr>
<tr>
<td>---------------------------------------------------------</td>
<td>-----------</td>
<td>-------------</td>
<td>---------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Education</td>
<td>1594</td>
<td>0</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>Water</td>
<td>7774</td>
<td>294</td>
<td>83</td>
<td>3</td>
</tr>
<tr>
<td>Payment for goods bought by hire-purchase or lay-buy</td>
<td>7021</td>
<td>765</td>
<td>75</td>
<td>8</td>
</tr>
<tr>
<td>Health</td>
<td>5535</td>
<td>2135</td>
<td>59</td>
<td>23</td>
</tr>
<tr>
<td>Transportation</td>
<td>1170</td>
<td>2826</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td>Religious Expenses</td>
<td>4243</td>
<td>2323</td>
<td>45</td>
<td>25</td>
</tr>
<tr>
<td>Life insurance, Burial societies, Funeral policies</td>
<td>4335</td>
<td>1729</td>
<td>46</td>
<td>18</td>
</tr>
<tr>
<td>Electricity</td>
<td>4071</td>
<td>262</td>
<td>43</td>
<td>3</td>
</tr>
<tr>
<td>Other usual expenses</td>
<td>5523</td>
<td>1567</td>
<td>59</td>
<td>17</td>
</tr>
<tr>
<td>Fuel</td>
<td>2739</td>
<td>660</td>
<td>29</td>
<td>7</td>
</tr>
<tr>
<td>Large expenses</td>
<td>6925</td>
<td>2241</td>
<td>74</td>
<td>24</td>
</tr>
</tbody>
</table>
Recommendations for constructing consumption aggregates (Deaton & Zaidi, 2002)

How to deal with missing information

- Standard practice in LSMS type surveys
- Check data for outliers; miscoding or misunderstanding of units for quantities causes errors in unit values.

Imputations

- “impute the price or quantity using market reference (i.e. looking at the average quantity or average price reported by other households with similar characteristics, such as place of residence, size, etc.)”. WB

- No prices and no quantities in ACDIS (just amount spent per month)
  - Design issues (fewer and missing key items: home production; clothing expenditures; aggregate items)

- Imputations with cluster medians: Missing prices or unit values → proxy the median of 'similar' households in the neighborhood or geographical area
Rationale for ad-hoc approach

- High % missing and zero values
- Necessary to test different imputation methodologies
- Ultimately comparison of poverty lines (and robustness of results)
  1. Absolute PL (2$)
  2. Relative PL based on Tobit regression + ICE (Royston)
  3. Relative PL based on medians of cluster (cluster defined as Isigodi area) ; crude imputation
  4. Relative poverty line (PL) based on naïve scenario (no imputation)
Two stage imputation procedure (FOOD ITEMS)

- Started with Food expenditures
- 2 stages imputation
- **1 stage**: Dealt with zero values via Tobit regression model
  - Estimated predicted probabilities based on regression model
  - Identified false zeros according to prediction model
  - Set false zeros to missing when positive predicted probabilities
  - Instead of fitted value, value attributed via imputation model (2nd stage)
Second stage Imputation (Food Items)

- **Deal with missing values in food expenditure items**
- **Method**: ICE - multiple imputation by chained equations (Royston 2004 StataJ, van Buuren *et al* Stat Med 1999)
  - Basic idea: Multiple regression, (in general)
  - Replace missing values with “plausible” substitutes (based on distribution of given data)
  - Do this several times, create $m > 1$ datasets (~10) → 5/10
  - Analyse datasets individually, but identically
  - Combine the estimates, get confidence intervals using Rubin’s rules
  - Predictors and choice of model: need careful thought
  - Major limitation: **MAR** (probability of missingness *does not* depend on unobserved information)
    → Sensitivity analysis needed → work in progress
NON-FOOD ITEMS: 1 only stage

- Only missing data problem addressed via ICE
- Settled on an average (across imputations) value of food expenditures (via Rubin’s rule)
- Used as input to imputation of non-food expenditures
- Final stage → aggregated each imputed non-food items and derived total expenditure or total consumption
Also...Crude imputation

- Performed a **crude imputation** based on median of Isigodi (for comparison purposes)
- Isigodi: areas in the DSA for which single indunas are responsible (chiefs) ➝ key predictor of missingness
- Replaced missing value with median expenditure of the area where the hh belong
Robustness: How (& how much) do contaminated data affect poverty measures?

Results: Poverty lines & FGT measures

1. Absolute PL
2. Derived relative poverty lines based on 50% of median consumption as proxied by PCE; imputed PCE (ICE) and Imputed w/median PCE (crude)

Compared three scenarios:
- Relative Poverty line (PL) based on naive scenario (no imputation) = 83.7 R
- RPL based on imputation ICE= 99.4 R
- RPL based on imputation-median = 103.2R

So does imputation methodology matter?
Results: ABSOLUTE PL & FGTs

PL 240 \( R = 2\)PPP

**ABSOLUTE PL**

Foster-Greer-Thorbecke poverty indices, FGT(\(a\))

<table>
<thead>
<tr>
<th></th>
<th>All obs</th>
<th>(a=0)</th>
<th>(a=1)</th>
<th>(a=2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Naive</strong></td>
<td></td>
<td>0.67679</td>
<td>0.31438</td>
<td>0.17918</td>
</tr>
<tr>
<td><strong>ICE</strong></td>
<td></td>
<td>0.59909</td>
<td>0.24415</td>
<td>0.12546</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td></td>
<td>0.57899</td>
<td>0.23230</td>
<td>0.11859</td>
</tr>
</tbody>
</table>

FGT(0): headcount ratio (proportion poor)

FGT(1): average normalised poverty gap

FGT(2): average squared normalised poverty gap
Relative PL (50% Median PCE)

**PL (83) 50% of Median consumption – NAIVE scenario**
Number of observations poor = 1448
Mean of xpc amongst the poor = 61.749
Mean of poverty gaps (poverty line - xpc) amongst the poor = 21.251

<table>
<thead>
<tr>
<th>Foster-Greer-Thorbecke poverty indices, FGT(a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All obs</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Naive</td>
</tr>
</tbody>
</table>

**PL (99) – 50 % of median consumption when ICE PCE**
Number of observations poor = 1268
Mean of impxpc amongst the poor = 77.051
Mean of poverty gaps (poverty line - impxpc) amongst the poor = 21.949

<table>
<thead>
<tr>
<th>All obs</th>
<th>a=0</th>
<th>a=1</th>
<th>a=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICE</td>
<td>0.13747</td>
<td>0.03048</td>
<td>0.01023</td>
</tr>
</tbody>
</table>

**PL (103) – 50 % of median consumption when Median PCE (Isigodi)**
Number of observations poor = 1297
Mean of m_impxpc amongst the poor = 78.897
Mean of poverty gaps (poverty line - m_impxpc) amongst the poor = 24.103

<table>
<thead>
<tr>
<th>All obs</th>
<th>a=0</th>
<th>a=1</th>
<th>a=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>0.14063</td>
<td>0.03291</td>
<td>0.01171</td>
</tr>
</tbody>
</table>

FGT(0): headcount ratio (proportion poor)
FGT(1): average normalised poverty gap
FGT(2): average squared normalised poverty gap
Comparison of Imputed (ICE) vs. Non-imputed PCE (log-scale)

PL 240 Rands=2$ a day PPP
Comparison of Imputed (ICE) vs. Imputed (Median-Isigodi) log-scale

PL 240 Rands=2$ a day PPP
Assets vs. Poor/Non-Poor

Number of assets by poor/non-poor (ICE-imputation)
Performance of Asset indices vs Imputed per capita expenditure

Quintiles of asset index based on PCA
Conclusion

Lack of Gold Standard: Important to understand the relative ranking of HH

Results quite robust to the choice of imputation method; however:
- Crude imputation too generous
- **ICE performs better**: statistically more accurate; Standard errors
- Additional predictors add more information
- Imputation decreases the proportion poor and shift the distribution up
- ICE takes into account variability due to imputation uncertainty
- Eventually final model of interest better estimated

Results quite robust to the choice of PL
- When using different PCE measures, relative PL derived from imputed PCE quite close
- However when building poverty profiles method matter! (choice of PL)

FGTs quite consistent
- Using an absolute PL, FGT0 and FGT1 (incidence and depth of poverty) lower in imputed measures

Additional remark: Expenditure based measures discriminate better than asset indices based on PCA
Way forward

- Sensitivity analysis based on NMAR
- Extend analysis to other waves
- Compare imputation results overtime
  - For household with minimal or no imputations what’s the trend between each wave?
- Build imputed welfare indicator of interest
- Assess impact of AIDS mortality on welfare indicator of interest