rdecompose: Decompose Aggregate Values in Stata

Jinjing Li[1] Yohannes Kinfu[2]

[1]NATSEM, Institute of Governance and Policy Analysis
[2]CeRAPH, Health Research Institute
University of Canberra

24 September 2015
Oceania Stata User Group Meeting 2015

Outline

Decomposition

- Micro-data based decomposition
  - Blinder-Oaxaca decomposition approach etc.
- Macro-data based decomposition
  - Rate Decomposition (Kitagaw and Gupta's decomposition models, various Gini decomposition etc.)

Decomposition in Stata

- Well developed commands for Blinder-Oaxaca type decomposition (e.g. oaxaca, nldecompose)
- Some highly specialised rate decomposition (e.g. Gini decomposition with descogini)
- No readily available command for general rates/aggregate data decomposition

Generalised Rate Decomposition

- Assume that the rate r can be expressed by k factors

$$r\left(x_1 \cdots x_k\right) = \prod_{i=1}^{k} x_i$$

2 Factors Decomposition

- In the case of $k = 2$

$$\begin{cases} C(x_1) = & \frac{1}{2}(x_2^a + x_2^b)(x_1^a - x_1^b) \\ C(x_2) = & \frac{1}{2}(x_1^a + x_1^b)(x_2^a - x_2^b) \end{cases}$$

- Intuitively speaking, the contribution of the factor is conditionally on the mean value of the other factors.
- We then standardised the contribution from $C(x_1)$ and $C(x_2)$

3 or more factors

Different specification

- In the case where $k \geq 3$

$$C(x_i) = \sum_{j=1}^{k-1} \frac{R(j-1,i)}{k \begin{pmatrix} k-1 \\ j-1 \end{pmatrix}} (x_i^a - x_i^b)$$

- where $R(j,i)$ is the sum of all possible values of the product of $k-1$ factors (excluding x_i), out of which j factors from population a and all other factors from population b.

3 or more factors

Generalised specification

- It is also possible that the rate function $r\left(x_1 \cdots x_k\right)$ is more than the simple product function, e.g. $newborn = \sum_{age} fertility \cdot women_{age}$
- sum over a specified group is a common operation in cross-classifed data

3 or more factors

Different specification

- The number of permutations increase much faster than k
  - for six factor decomposition, we need to calculate $r(\cdot)$ 192 times
  - No publicly available software (in any language or statistical package) to handle large k
  - Gupta published some Fortran code for small and medium size k but it requires the end user to tweak the code for each case
- In most cases, we also need to aggregate values over a group (e.g. age groups, location etc.)
- Mostly done in Excel or manual calculations, which are prone to mistakes

rdecompose

- We developed a new Stata command **rdecompose** to assist decompositions using Gupta's method
- rdecompose currently supports decomposition where the aggregate rates r is calculated based on k factors, and aggregated over s, i.e. $r = \sum_s f(x_1 \cdots x_k)$
  - ability to decompose with any arbitrary number of factors
  - ability to automatically aggregate values over a group
  - ability to specify non-standard functional form instead of product only (e.g. $x_1 e^{x_2} \ln(x_3 + x_4)$)
  - ability to interact with other commands for further processing
  - in Stata

rdecompose syntax

rdecompose *variables* [**if** exp], **group**(variable) [**sum**(varlist) **detail reverse function**(string) **transform**(variable) **multi baseline**(#)]

variables: factors that contribute to the rates

group: population identifier (string or numeric)

sum: indicates the rate is the sum of the specified variable (Default: none)

function: specifies the funciton form (Default: $f(\cdot) = \prod_{i=1}^{k} x_i$)

Example 1: Data on total fertility and proximate determinants of fertility

- Example from Gupta(1994)
- Decompose total fertility rate in Korea between 1960 and 1970
- Following Moreno(1991)

$$TFR = C_m C_c C_x \cdots C_{others}$$

Example 1: Data on total fertility and proximate determinants of fertility

- Data as in Stata

year	Marriage	Contraception	Abortion	Lactation	Fecundity
1970	.58	.76	.84	.66	16.573
1960	.72	.97	.97	.56	16.158

Table: Fertility Rate Decomposition in Korea

Example 1: Data on total fertility and proximate determinants of fertility

```
. rdecompose Marriage Contraception Abortion Lactation Fecundity , group(year)

Decomposition between year == 1960 (6.13)
                  and year == 1970 (4.05)
Func Form = Marriage*Contraception*Abortion*Lactation*Fecundity
```

Component	Absolute Difference	Proportion (%)
Marriage	−1.09	52.46
Contraception	−1.23	59.13
Abortion	−.728	35.00
Lactation	.84	−40.38
Fecundity	.129	−6.20
Overall	−2.08	100.00

Number of Obs : 10

Example 2: Data on demand for additional children in Nepal

- Example data from Clogg and Eliason(1998) on population size and percent desiring more children
- Decompose the difference between women with one child and women with 4+ children

age group	Age composition	Rate	Parity
20-24	27	37.037	One child
25-29	152	19.079	One child
	.....		
20-24	363	90.083	4+ Children
25-29	208	76.923	4+ Children

Table: National Fertility Survey, Clogg and Eliason(1988)

Example 2: Data on demand for additional children in Nepal

```
rdecompose Age_composition Rate , group( Parity ) transform( Size ) sum(
    age_group )

Decomposition between Parity == 1 (11.49)
                  and Parity == 2 (72.09)
Func Form = \sum(age_group){Size*Rate}
```

Component	Absolute Difference	Proportion (%)
Age_composition(*)	23.1	38.07
Rate	37.5	61.93
Overall	60.6	100.00

(*) indicates transformed variables Number of Obs : 20

Example 3 : Global Burden of Disease Data

- Latest IHME Data (2015)
- Decompose the mortality rate due to *ageing effect* and the change in *Communicable disease*, *Non-communicable disease* and *Injuries* between developed and developing countries.

Example 3 : Global Burden of Disease Data

- Data as in Stata

age_group	age_structure	CDM	NCD	Injuries	group
		.....			
1-4 years	0.464026	284.31	55.8	46.33	1
5-9 years	0.054843	47.17	18.6	19.13	1
		.....			
80 years and above	0.006002	1433.16	11589.4	491.64	1
		.....			
1-4 years	0.044478	4.39	12.23	9.98	2
5-9 years	0.053709	1.21	6.19	5.64	2
		.....			
80 years and above	0.045763	650.37	9489.93	329.66	2

Table: Disease Burden between Developed and Developing countries

Example 3 : Global Burden of Disease Data

```
. rdecompose age_structure CDM NCD Injuries , group(group) sum(age_group_i) func
    (age_structure*( CDM + NCD + Injuries ))

Decomposition between group == 1 (576.46)
                  and group == 2 (993.90)
Func Form = \sum(age_group_i){age_structure*( CDM + NCD + Injuries )}
```

Component	Absolute Difference	Proportion (%)
age_structure	843	201.88
CDM	−200	−47.86
NCD	−193	−46.17
Injuries	−32.8	−7.85
Overall	417	100.00

Number of Obs : 160

Example 4 : China Health Expenditure

- Ongoing project between Health Research Institute at University of Canberra and China National Health Develop Research Centre
- Decompose the increase of total health expenditure between 1993 to 2012 into
  - Prevalence by age and disease group
  - Population Size
  - Demographic Ageing
  - Expenditure per case
  - Excess Health Inflation

Example 4 : China Health Expenditure

```
Decomposition between year == 1993 (124535.22)
                  and year == 2012 (2475451.03)
Func Form = \sum(disease_group)\sum(agegroup){prevalencerate*population*ageing*
    exppercase*healthpriceinflation}
```

Component	Absolute Difference	Proportion (%)
prevalencerate	−77942	−3.32
population	139708	5.94
ageing	191016	8.13
exppercase	1536974	65.38
healthpricein~n	561159	23.87
Overall	2350916	100.00

Saved Results

- rdecompose also returns some results in scalar/macro/matrix format for further processing
  - Scalar e(N) contains the number of observations used in the estimation
  - Scalar e(rate1) contains the rate calculated for the first group
  - Scalar e(rate2) contains the rate calculated for the second group
  - Scalar e(diff) shows the total differences between two groups
  - Macro e(basegroup_value) shows the baseline group value
  - Matrix e(b) contains the total contributions for each factor

Summary

- **rdecompose** decomposes aggregated rates from two populations using Gupta's decomposition model
  - Wide range of applications in the field of demography, health, economics etc.
  - Support flexible a functional form and avoid cumbersome calculations with a large number of factors

- Limitations and Next steps
  - Other common data transformations
  - Other rate decomposition models
  - Improved support for more than two populations (currently with the option multi)