The Stata Journal (2017)
17, Number 2, pp. 490–502

Rate decomposition for aggregate data using Das Gupta's method

Jinjing Li
National Centre for Social and Economic Modelling
Institute for Governance and Policy Analysis
University of Canberra
Bruce, Australia
jinjing.li@canberra.edu.au

Abstract. Social, behavioral, and health scientists frequently decompose changes or differences in outcome variables into components of change and assess their relative importance. Many Stata commands facilitate this exercise using unit-level data, notably by applying the Blinder–Oaxaca approach. However, none of the comparable user-written commands decompose changes or differences in aggregate data despite their availability and the widespread use of corresponding decomposition techniques. In this article, I present the user-written command `rdecompose`, which decomposes aggregate or cross-classified data based on Das Gupta's (1993, *Standardization and Decomposition of Rates: A User's Manual, Volume 1*) approach, and demonstrate its application in multiple settings. This command extends the original method by allowing multiple factors and flexible functional specifications.

Keywords: st0483, rdecompose, decomposition, cross-classified, Das Gupta method

1 Introduction

Numeric values such as rates, means, percentages, and proportions are instrumental in measuring social, economic, health, and demographic outcomes. Researchers often study measures such as birth rates, prevalences of diseases, or income inequality and analyze differences between populations or changes over time; such factors reflect differences in relevant population characteristics that may directly or indirectly influence outcomes. Demographers, economists, and public health scientists traditionally apply standardization and decomposition techniques to distinguish real "rate" differences from the effects of compositional factors on measured outcomes. These techniques are used throughout the social sciences to determine why rates differ among populations (Guo et al. 2012; Wang et al. 2000; Yamaguchi 2011).

The literature uses decomposition techniques that fall broadly into two categories based on data requirements. The first category uses unit record data along with a multivariate regression-based technique known as the Blinder–Oaxaca approach (Blinder 1973). This approach has multiple Stata implementations, including the linear version from Jann (2008) and nonlinear extensions from Sinning, Hahn, and Bauer (2008) and Powers, Yoshioka, and Yun (2011). This approach relies on the availability of individual-level data. The second approach is designed for cross-classified data or contingency tables and often uses algebraic relationships rather than econometric estimations (Chevan and Sutherland 2009; Das Gupta 1993).

Although some Stata commands apply unit record-based Blinder−Oaxaca decomposition, no comparable user-written commands implement existing decomposition techniques for aggregate data,[1] despite wide availability of aggregate data and use of corresponding aggregate data-based decomposition methods in the literature. In this article, I introduce a new user-written command, `rdecompose`, for a variant of such methods known as Das Gupta's reformulation and demonstrate potential applications from a range of settings, including demography, public health, and economics.

Das Gupta's method for cross-classified data is based on incremental methodological developments in standardization and decomposition techniques that have occurred over several decades. Wolfbein and Jaffe (1946) were among the first to demonstrate the importance of decomposition by applying a double standardization technique, but Kitagawa (1955) developed a formal procedure for decomposing cross-classified data. Her work led to the simultaneous identification of separate but additive composition and rate components that summed to rate differences. In a series of articles, Das Gupta took a symmetric approach to the interaction component and developed functional relationships that allow deterministic allocation of the interaction components among cross-classified variables. This approach led to integration of the interaction between component effects into the additive main effects and facilitated interpretation of results. Das Gupta's approach also imposed few constraints on the nature of the variable and its distribution or on the specification of relationships for the outcome of interest. Thus the method can be applied flexibly to most cross-classified aggregate data.

2 Method

2.1 Standard rate decomposition with multiplicative factors

Demographers, health researchers, and social scientists must sometimes interpret differences between two crude rates of the same or comparable population. For instance, researchers may want to understand what drives differences in rates of death today versus 20 years ago. Differences between death rates may be decomposed into multiple confounding factors such as differences in age-specific death rates and age structures. Kitagawa's (1955) approach initially decomposed differences in job mobility rates between two cities into migrant status and time spent in the labor force. Das Gupta (1978, 1991, 1993) generalized the method, and Chevan and Sutherland (2009) made improvements so that the approach can be applied to any type and number of factors.[2] Unlike other decomposition methods that allow nonlinear specification of rates, Das Gupta's method yields stable results independent of the order in which factors are introduced and needs no special treatment for interaction terms.

[1] Researchers may be able to decompose crude rate data via existing individual-record based decomposition commands such as `mvdcmp` (Powers, Yoshioka, and Yun 2011) in some cases. Doing so may involve data transformation and programming.

[2] For a review of rate standardization and comparison methods, see Keiding and Clayton (2014).

Suppose rate r can be expressed by k multiplicative factors $x_1 \dots x_k$. Rate r can be a death rate, fertility rate, or any aggregate measures of interest. $x_1 \dots x_k$ are k factors (for example, age structure in the case of death rates) driving changes in the rate.

$$r(x_1 \dots x_k) = \sum_{i=1}^{k} x_i$$

If superscript a is used to denote the first population and superscript b the second population, the (unstandardized) contribution of two factors $C(x_1)$ and $C(x_2)$ to the difference of r_a and r_b in the case of two factors can be expressed mathematically as below, following Das Gupta (1991, 1993):

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

Intuitively, the contribution of a factor lies in its conditional effect on the mean values of other factors. The relative contribution of x_1 is therefore $C(x_1)/\{C(x_1) + C(x_2)\}$, and the relative contribution of x_2 is $C(x_2)/\{C(x_1) + C(x_2)\}$. This approach is generally straightforward when there are few factors. However, calculations become cumbersome as k increases because of the need to compute all possible counterfactuals (2^k) and aggregate the result. Mathematically, the contribution of the ith factor to the rate is

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

where $R(j, i)$ is the sum of all possible values of the product of $k-1$ factors (excluding x_i), from which j factors are from population a and all other factors are from population b. The number of possible values can be large when there are many factors because the number of permutations increases faster than k.

One advantage of this type of decomposition is the consideration of all possible replacements of the elementary rates of the first population with the corresponding rates of the second, thus avoiding path dependency. Das Gupta's method essentially assigns a weight to each possible path where the importance of the specific factor gradually fades when further changes to other factors are introduced in the counterfactual. This assumption differs from Shapley's decomposition approach, where all paths are treated equally (Sastre and Trannoy 2002).

The result of Das Gupta's decomposition can be presented intuitively, where observed rate differences are decomposed into different component effects with the relative contribution summing to 100%. This method can be applied to a wide range of research, such as differences in sociodemographic attributes, income inequality, and disparities in health outcome.

2.2 Generalized rate decomposition

In some cases, the crude rate cannot be represented by a series of multiplicative factors. Instead, more complex computations might be involved. In the case of death rates, for instance, researchers might need to use only summative and multiplicative operators to control for deaths attributable to different potential mechanisms and age structure. In other cases, rate functions can be more complicated. Consider, for example, $r = x_1 e^{x_2} \ln(x_3 + x_4)$. In such cases, the calculation of r can be presented in a more generic form,

$$r(x_1 \dots x_k) = f(x_1, \dots, x_k)$$

where $f(\cdot)$ is the rate function instead of a simple multiplicative equation as before. Although the principle remains the same, the calculation of the rate must be replaced by a more generic function. This necessity often increases technical complexity in practical implementation, especially when k is large. The `rdecompose` command assists researchers with such issues.

3 The rdecompose command

The `rdecompose` command implements Das Gupta's style decomposition where the aggregate rate r is calculated based on k factors and aggregated over s in the following manner:

$$r = \sum_s f(x_1 \dots x_k)$$

3.1 Syntax

The syntax of `rdecompose` is

`rdecompose` *varlist* [*if*] `, group(`*varname*`)` [`sum(`*varname*`) detail reverse`
`function(`*string*`) transform(`*varname*`) multi baseline(`*#*`)`]

`rdecompose` should be immediately followed by the names of the variables (factors) that contribute to the rates. The population group indicator also must be included in the `group()` option. An `if` condition can be used in combination with `rdecompose` if required.

3.2 Options

`group(`*varname*`)` specifies the group indicator of the two populations that will be compared. *varname* can be in numeric or string format. `group()` is required.

`sum(`*varlist*`)` indicates the population rate is an aggregated value summation over each distinct value of this variable or variables (for example, age or location).

`detail` gives more detailed output when the `sum()` option is specified.

`reverse` reverses the order of the two compared groups.

`function(`*string*`)` specifies the function form of the rate. For example, the user may specify `ln(`*factor1*`+`*factor2*`)*exp(`*factor3*`)` to be used as a function. Most Stata-supported functions can be used here. `rdecompose` assumes multiplicative operations if a function is not specified. An error message will appear if the specified function is invalid or cannot be evaluated.

`transform(`*varname*`)` converts absolute numbers into proportions within the population.

`multi` indicates there are more than two populations in the `group()` option. Specifying this option results in multiple comparisons against the baseline population, which can be specified in the `baseline()` option.

`baseline(`*#*`)` specifies the value of the group variable for the baseline population.

3.3 Output and stored results

The typical output of `rdecompose` resembles what is presented in output 1. The output reports the names of variables and rates corresponding to the two compared populations, the functional form assumed in the computation, and a table listing factors and their contributions. `rdecompose` also standardizes the total contribution into 100% for the convenience of interpretation.

Output 1: An example of the rdecompose command

```
. rdecompose size rate, sum(agegroup) group(pop)
Decomposition between pop == 1 (9800.09)
                  and pop == 2 (55800.13)
Func Form = sum(agegroup)size*rate
-------------------------------------------------------------------------
                Component         Absolute Difference    Proportion (%)
-------------------------------------------------------------------------
                     size                       15839        34.43
                     rate                       30161        65.57
-------------------------------------------------------------------------
                  Overall                       46000       100.00
-------------------------------------------------------------------------
```

Besides table output, `rdecompose` returns some estimation results as scalars, macros, and matrices. The most notable include

Scalars
`e(rate1)` contains the rate calculated for the first group
`e(rate2)` contains the rate calculated for the second group
`e(diff)` shows total differences between two groups

Macros
`e(basegroup_value)` stores the value of the group variable for the baseline population

Matrices
`e(b)` contains total contributions for each factor

4 Examples

This section demonstrates the use of `rdecompose` for a range of decomposable factors and data types. The first two examples are mostly for validation because they replicate known examples from Bongaarts (1978) and Clogg and Eliason (1988). These examples draw data from the discipline of demography and decompose changes in parity progression and total fertility rates (TFRs). Both examples are cited and discussed by Das Gupta (1993). The third example uses health expenditure data from China from 1993 to 2012 and attempts to decompose growth in health expenditures into five major factors. The fourth example shows `rdecompose`'s use in economics through an exercise of income inequality decomposition. The final example demonstrates how standard errors of decomposition results can be derived with bootstrapping.

4.1 Example 1: Explaining changes in fertility rates over time

Table 1 presents the TFR in South Korea from 1960–1970. It also shows data on proportions of married women (C_m), women not using contraception (C_c), prevalence of abortion (C_a), lactational infecundability (C_i), and total fecundity (TF) rate. The last is the level of natural fertility that the population would have attained if all women had married at an early age, practiced neither contraception nor abortion, and did not have long gestation periods during lactation. In demographic literature, these variables are collectively known as proximate determinants of fertility. TFR is expressed as a product of these five factors (that is, $\text{TFR} = \text{TF} \times C_m \times C_c \times C_a \times C_i$).

Table 1. TFRs and proximate determinants of fertility

	South Korea (1960–1970)	
Fertility measures	1960 (population 1)	1970 (population 2)
TF rate	16.158	16.573
Index of proportion married (C_m)	0.72	0.58
Index of noncontraception (C_c)	0.97	0.76
Index of induced abortion (C_a)	0.97	0.84
Index of lactational infecundability (C_i)	0.56	0.66
TFR	6.13	4.05

Source: Bongaarts (1978)

As table 1 shows, between 1960 and 1970, the TFR in South Korea declined 2.08 points from 6.13 in 1960 to 4.05 in 1970. `rdecompose` can be applied as follows to determine relative contributions of each proximate determinant factor to changes in TFR observed:

Output 2: Stata code and command output of example 1

```
. use example1-bongaarts

. rdecompose marriage noncontracept abortion lactation fecundity, group(year)

Decomposition between year == 1960 (6.13)
                  and year == 1970 (4.05)
Func Form = marriage*noncontracept*abortion*lactation*fecundity
------------------------------------------------------------------------

                Component        Absolute Difference    Proportion (%)
------------------------------------------------------------------------

                 marriage                      -1.09         52.46
            noncontracept                      -1.23         59.13
                 abortion                      -.728         35.00
                lactation                        .84        -40.38
                fecundity                       .129         -6.20
------------------------------------------------------------------------

                  Overall                      -2.08        100.00
------------------------------------------------------------------------
```

The first column in output 2 shows the factor names, and the second column reports absolute contributions of each factor to the decline in recorded fertility rates. Regarding the reduction of 2.08 children per woman between 1960 and 1970, contraception and marriage contributed to a decline of about one child each. Abortion contributed about 0.73 to the total reduction of 2.08. The last column shows relative contributions of each factor as a percent of the total difference. Here 59.1% of the decline in TFRs during the decade can be attributed to increased use of contraception. During the same time, however, the duration of lactation declined and contributed to an increase in fertility. Results derived from `rdecompose` match the outcomes reported in the original article.

4.2 Example 2: Explaining differences in desire for more children

The second example is adapted from Clogg and Eliason (1988), who examine the desire to bear more children. It illustrates decomposition using cross-classified data. Table 2 compares desire for more children in two groups of women: those with 4 or more children (represented by parity 4+) and those with 1 child (represented as parity 1). Given that age is an important determinant of fertility and that most parity 1 women are likely younger than women with 4 or more children, the question is how to isolate the effect of age composition differences in the two parity groups. `rdecompose` can be applied to isolate the effect of age composition from actual differences in rates between the two groups of women.

Table 2. Population size and percent desiring more children (rate) by age

	Parity 4+ (population 1)		Parity 1 (population 2)	
Age groups	Size (Ni)	Rate (Ti)	Size (Ni)	Rate (Ti)
20 to 24	27	37.037	363	90.083
25 to 29	152	19.079	208	76.923
30 to 34	224	15.179	96	56.25
35 to 39	239	5.021	59	20.339
40 to 44	211	6.161	48	10.417
All ages	853	11.489	774	72.093

Source: Clogg and Eliason (1988)

As shown, this example uses the `transform()` option to convert the absolute number for size into proportions within the population group. The result suggests that the rate effect contributes about 62% of the difference, whereas the size effect (size of each age group) contributes about 38% of the differences.

Output 3: Stata code and command output of example 2

```
. use example2-clogg
. rdecompose Size Rate, group(Parity) transform(Size) sum(age_group)
Decomposition between Parity == 1 (11.49)
                  and Parity == 4 (72.09)
Func Form = \sum(age_group){Size*Rate}
────────────────────────────────────────────────────────────────────────
                Component         Absolute Difference    Proportion (%)
────────────────────────────────────────────────────────────────────────

                  Size(*)                        23.1         38.07
                     Rate                        37.5         61.93
────────────────────────────────────────────────────────────────────────

                  Overall                        60.6        100.00
────────────────────────────────────────────────────────────────────────
(*) indicates transformed variables
```

Source: Zhai, Goss, and Li (2017)

4.3 Example 3: Explaining drivers of health expenditures in China

The third example is from Zhai, Goss, and Li (2017), which examines the factors driving rising health expenditures in China. Using published National Health Accounts reports and disease prevalence data from the Global Burden of Disease 2013 Study from the Institute for Health Metrics and Evaluation (2015), this example decomposes the growth of health expenditure between 1993 and 2012 in China into five factors: population increase, changes in disease prevalence rates, shifts in age structure of the population, excessive health price inflation, and changes in treatment cost per case. This example showcases a more complex decomposition with multiple factors and the `detail` option.

rdecompose allows users to aggregate results from multiple subgroups—disease and age groups in this example. The detail option allows the program to display more detailed decomposition results normally hidden by the sum process. As seen in output 4, the detail option shows the contribution to health expenditures by disease and age group. Given page limits, only partial results are presented.

The decomposition suggests that real expenditure per case (exp_percase), excess health price inflation (ehpi), and aging (aging) drove increased health expenditures in China between 1993 and 2012. Population growth (population) was a secondary factor. Reductions in disease prevalence rates (prevalence_rate) only slightly slowed the growth in expenditures. Moreover, the result suggests that more than 70% of the difference in health expenditures on neoplasms and circulatory, respiratory, endocrine, nutritional, metabolic, and digestive diseases over the period was caused by changes in expenditures per case and the excess health price inflation. Examining results in the "detail" section reveals that aging of the population contributes more to growth of expenditures on neoplasms and circulatory, endocrine, nutritional, metabolic, and digestive diseases versus other diseases.

Output 4: Stata code and command output of example 3

```
. use example3-zhai

. rdecompose prevalence_rate population ageing exp_percase ehpi, group(year)
> sum(disease age_group) detail

Decomposition between year == 1993 (124535.24)
                  and year == 2012 (1000586.64)
Func Form = \sum(disease)\sum(age_group){prevalence_rate*population*
> ageing*exp_percase*ehpi}
--------------------------------------------------------------------------
                Component          Absolute Difference    Proportion (%)
--------------------------------------------------------------------------

          prevalence_rate                       -18982         -2.17
               population                        59535          6.80
                   ageing                        98405         11.23
              exp_percase                       635946         72.59
                     ehpi                       101148         11.55
--------------------------------------------------------------------------
    Value of disease and Components           Detailed Contributions
--------------------------------------------------------------------------
            Blood    prevalence_rate              -507         -0.06
                          population               500          0.06
                              ageing               426          0.05
                         exp_percase              4183          0.48
                                ehpi               851          0.10
      Circulatory    prevalence_rate             -6919         -0.79
  (output omitted)
--------------------------------------------------------------------------
                  Overall                       876051        100.00
--------------------------------------------------------------------------
```

4.4 Example 4: Income equality decomposition

The rate decomposition method also applies to other fields such as economics. A series of studies (for example, Bargain and Callan [2010]) decomposed changes in inequalities between countries using Shorrocks–Shapley decomposition (Shorrocks 2013). Such studies often use indexes derived from counterfactual simulations to attribute the relative importance of each component. In this case, decomposing the contributing factors to the income inequality resembles a rate decomposition.

The contribution of a factor to income inequality sometimes is determined based on Shapley values, which are computed by averaging the effects of all possible permutations before and after the factor of interest is substituted. This approach avoids path dependency (Devicienti 2010; Okamoto 2011) but treats the first-order effect with the same weight as mixed effects where multiple input factors have been substituted. Alternatively, the estimation of a factor's contribution can follow Das Gupta's approach, where weights are normalized, giving greater weight to direct effects. `rdecompose` can be used for such analyses.

Because the outcome value cannot be described as a simple function, specific values can be passed via the `function()` option of the command. For instance, to assess the contribution to Gini from two factors (for example, the population structure and the tax system as in table 3), one can use `rdecompose` as demonstrated in output 5.

Table 3. Estimated Gini coefficient with two factors

Gini		Factor 1 (for example, population)	
		1	2
Factor 2 (for example, tax system)	1	0.31	0.39
	2	0.48	0.52

Output 5: Decompose contributions to Gini based on two factors

```
. rdecompose factor1 factor2, group(group)
> function(cond(factor1==1, cond(factor2==1,0.31,0.48),
> cond(factor2==1,0.39,0.52)))

Decomposition between group == 1 (0.31)
                  and group == 2 (0.52)
Func Form = cond(factor1==1, cond(factor2==1,0.31,0.48),
> cond(factor2==1,0.39,0.52))
──────────────────────────────────────────────────────────────────────
                Component         Absolute Difference    Proportion (%)
──────────────────────────────────────────────────────────────────────
                  factor1                        .06          28.57
                  factor2                        .15          71.43
──────────────────────────────────────────────────────────────────────
                  Overall                        .21         100.00
──────────────────────────────────────────────────────────────────────
```

4.5 Example 5: Bootstrap with rdecompose

rdecompose does not have the native support of standard-error estimations in its current version because sources of sampling and nonsampling errors could vary substantially for each case. Standard errors in some cases might be meaningless if population-level data are applied or possibly inaccurate because of data confidentialization processes. However, should uncertainties of input data be mathematically described, it may be possible to derive the standard errors of the estimators. One way to do this is to use the bootstrap technique (Wang et al. 2000), which can be programmed in combination with rdecompose.

For example, to consider the sampling errors in example 2, one may bootstrap the underlying sample, which can be presented as a unit record dataset. A short customized program can be written in Stata to extract the rdecompose output for each iteration of the bootstrapping process.

Output 6: A customized rdecompose program for bootstrapping

```
program mydecompose, eclass
  preserve
  collapse (count) Size= d (mean) Rate = d, by(age_group Parity)
  quietly rdecompose Size Rate, group(Parity) transform(Size) sum(age_group)
  matrix b = e(b) * 100
  ereturn post b
  restore
end
```

The dataset from example 2 needs to be transformed for the bootstrap command as shown in output 7, which includes both the commands and the results of this example.

Output 7: Results from bootstrapping

```
. use example2-clogg

. expand Size
(1,617 observations created)

. by age_group Parity, sort: generate d = _n<=round(Rate*Size/100)

. bootstrap, nowarn nodots reps(1000): mydecompose

Bootstrap results                               Number of obs     =      1,627
                                                Replications      =      1,000

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Size(*) |   23.07168   2.459112     9.38   0.000     18.25191    27.89145
        Rate |   37.53248   3.136146    11.97   0.000     31.38575    43.67922
------------------------------------------------------------------------------
```

As shown, this short program returns the standard errors of estimated rate contributions via bootstrapping. With minor changes to the program, one can estimate the standard error of the percentage values if needed.

5 Concluding remarks

This article presents a user-written command, `rdecompose`, that replicates and extends the popular rate decomposition method presented by Das Gupta (1993). This command provides researchers a user-friendly tool to decompose changes in populations, which is a task common to research in demography, health, and economics. This tool reduces the tediousness of programming large numbers of factors and relaxes the functional requirement of the original method. `rdecompose` normally assumes that the underlying relations of the rate calculation are known; however, some or all permutations can be overridden via the `function()` option. The command currently has no native support for standard-error estimation because each case may contain vastly different sources of sampling and nonsampling errors. It may be possible, however, to use bootstrap techniques to derive standard errors of the estimates.

6 Acknowledgments

The author would like to thank Yohannes Kinfu, the reviewers, and the editors for their comments and suggestions.

7 References

Bargain, O., and T. Callan. 2010. Analysing the effects of tax-benefit reforms on income distribution: A decomposition approach. *Journal of Economic Inequality* 8: 1–21.

Blinder, A. S. 1973. Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources* 8: 436–455.

Bongaarts, J. 1978. A framework for analyzing the proximate determinants of fertility. *Population and Development Review* 4: 105–132.

Chevan, A., and M. Sutherland. 2009. Revisiting Das Gupta: Refinement and extension of standardization and decomposition. *Demography* 46: 429–449.

Clogg, C. C., and S. R. Eliason. 1988. A flexible procedure for adjusting rates and proportions, including statistical methods for group comparisons. *American Sociological Review* 53: 267–283.

Das Gupta, P. 1978. A general method of decomposing a difference between two rates into several components. *Demography* 15: 99–112.

———. 1991. Decomposition of the difference between two rates and its consistency when more than two populations are involved. *Mathematical Population Studies* 3: 105–125.

———. 1993. *Standardization and Decomposition of Rates: A User's Manual, Volume 1.* Washington, DC: U.S. Department of Commerce, Economics and Statistics Administration, Bureau of the Census.

Devicienti, F. 2010. Shapley-value decompositions of changes in wage distributions: A note. *Journal of Economic Inequality* 8: 35–45.

Guo, Z., Z. Wu, C. M. Schimmele, and S. Li. 2012. The effect of urbanization on China's fertility. *Population Research and Policy Review* 31: 417–434.

Institute for Health Metrics and Evaluation. 2015. GBD compare.
http://vizhub.healthdata.org/gbd-compare.

Jann, B. 2008. The Blinder–Oaxaca decomposition for linear regression models. *Stata Journal* 8: 453–479.

Keiding, N., and D. Clayton. 2014. Standardization and control for confounding in observational studies: A historical perspective. *Statistical Science* 29: 529–558.

Kitagawa, E. M. 1955. Components of a difference between two rates. *Journal of the American Statistical Association* 50: 1168–1194.

Okamoto, M. 2011. Source decomposition of changes in income inequality: The integral-based approach and its approximation by the chained Shapley-value approach. *Journal of Economic Inequality* 9: 145–181.

Powers, D. A., H. Yoshioka, and M.-S. Yun. 2011. mvdcmp: Multivariate decomposition for nonlinear response models. *Stata Journal* 11: 556–576.

Sastre, M., and A. Trannoy. 2002. Shapley inequality decomposition by factor components: Some methodological issues. *Journal of Economics* 77: 51–89.

Shorrocks, A. 2013. Decomposition procedures for distributional analysis: A unified framework based on the Shapley value. *Journal of Economic Inequality* 11: 99–126.

Sinning, M., M. Hahn, and T. K. Bauer. 2008. The Blinder–Oaxaca decomposition for nonlinear regression models. *Stata Journal* 8: 480–492.

Wang, J., A. Rahman, H. A. Siegal, and J. H. Fisher. 2000. Standardization and decomposition of rates: Useful analytic techniques for behavior and health studies. *Behavior Research Methods, Instruments, and Computers* 32: 357–366.

Wolfbein, S. L., and A. J. Jaffe. 1946. Demographic factors in labor force growth. *American Sociological Review* 11: 392–396.

Yamaguchi, K. 2011. Decomposition of inequality among groups by counterfactual modeling: An analysis of the gender wage gap in Japan. *Sociological Methodology* 41: 223–255.

Zhai, T., J. Goss, and J. Li. 2017. Main drivers of health expenditure growth in China: A decomposition analysis. *BMC Health Services Research* 17: 185–193.

About the author

Jinjing Li is an associate professor at the National Centre for Social and Economic Modelling of the Institute for Governance and Policy Analysis, University of Canberra, Australia.