The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–10

The REDI package: Random Empirical Distribution Imputation of continuous from categorical incomes

Molly M. King
Santa Clara University
Santa Clara, CA / USA
mollymkingphd@gmail.com

Abstract.

Researchers seek to convert categorical to continuous incomes for myriad reasons, but existing parametric and nonparametric approaches are limited. A new community-contributed command `redi` implements the method for Random Empirical Distribution Imputation described by King (forthcoming) in *Sociological Methodology*. This random cold-deck approach to imputing categorical incomes uses a real-world reference dataset to draw continuous incomes for observations of interest. The `redi` package can be used to reconcile categories between datasets or across years. The package produces continuous income observations that are nonparametric, bin consistent, and area- and variance-preserving.

Keywords: st0001, redi, imputation, categorical data, binned data, income distribution, top censored, top-coded

1 Categorical and continuous incomes

Surveys often collect income data using categories. However, when researchers seek to compare these data across years or panels, they often face the challenge of changing or overlapping categorical boundaries or even just changes in meaning due to inflation (Ligon 1989; Hout 2004). Converting categorical to continuous incomes may be valuable for further analyses, particularly when looking at interactions.

Existing methods used to estimate continuous values for observations from categorical income data each have important limitations: parametric estimators assume a particular probability distribution function to estimate continuous income values (von Hippel et al. 2017); nonparametric estimators, such as midpoint approaches, assign all observations in a bin to a single value (Ligon 1989; Hout 2004). Top incomes are notably challenging to approximate (Piketty and Saez 2014).

An alternative is to use existing empirical data about known distributions of income to estimate the value of observations within a given category. Rather than using income bins contingent on categories to maintain confidentiality in smaller surveys, for instance, the REDI method can impute from the distribution of real U.S. incomes. This minimizes assumptions about the shape of the distribution and aligns with existing empirical knowledge.

The `redi` package uses a real-world dataset known for its accuracy on income – the Current Population Survey March Annual Social and Economic Supplement (CPS ASEC) (Flood et al. 2021) – as the reference dataset to impute continuous income from categorical boundaries in a research dataset. The `redi` package implements the Random Empirical Distribution Imputation method proposed by King (forthcoming). The package takes a research dataset and imputes an income from the CPS ASEC for each observation, based on the categorical boundaries of the income range. For top incomes, `redi` uses the special freature that the CPS ASEC is not top coded, allowing for a realistic reflection of U.S.-based income inequality. I first present how the `redi` package computes these imputations. Then I discuss advantages, assumptions, and limitations for the package. Finally, I present syntax, demonstrate its use with an example, and close with a discussion of appropriate use cases.

2 Computing Empirical Imputations

The `redi` package works with two datasets: a research dataset and a reference dataset. The research dataset is the one containing the categorical data the user wishes to convert to continuous values. The reference dataset is an existing dataset with continuous values, created from empirical methods using the same sampling frame and method, from which the `redi` package draws observations.

2.1 Categories formally stated

In a research dataset with binned data, there are M income categories,

$$I = \{1, 2, ..., ..., M\},$$

in ascending order of income level. Each category has an upper income limit (U_b) and a lower income limit (L_b), except the top and bottom categories, which are open-ended. Most methods treat the lower-end of the bottom category (L_1) as 0, which seems to be generally justified. There are several approaches for approximating the top category, the most popular of which is using a Pareto distribution, which is a convenient but inexact fit to the upper tail of income limits (Hout 2004; Blanchet et al. 2018). The `redi` package introduced here does not require either assumption about the bottom or top categories.

2.2 Drawing from the reference dataset

The `redi` package uses an independent reference dataset with continuous incomes to impute continuous incomes for categorical variables in a research dataset of interest. The `redi` package begins by counting the number of observations (N_b) in each income category of the research dataset bounded by a lower (L_b) and upper bound (U_b). It then assigns a discrete value from within those bounds (L_b to U_b) to each of the N_b observations. Observations are drawn by implementing sampling with replacement from

picture(0,0)(0,0)(1,0)523.95749 (0,0)(0,-1)668.4975 (523.95749,0)(0,-1)668.4975 (0,-668.4975)(1,0)523.95749 picture

the reference dataset, using uniformly distributed random integer values on the interval $(1, N_b)$.[1] For an illustration, see Figure 1. As a result, with a large enough sample, the income values sampled from all income bins approximate a continuous distribution. If the research dataset is much larger than the reference dataset, then observations from the reference dataset may be sampled multiple times. For more details on the sampling method and a comparison of the `redi` program to other imputation approaches, see King (forthcoming).

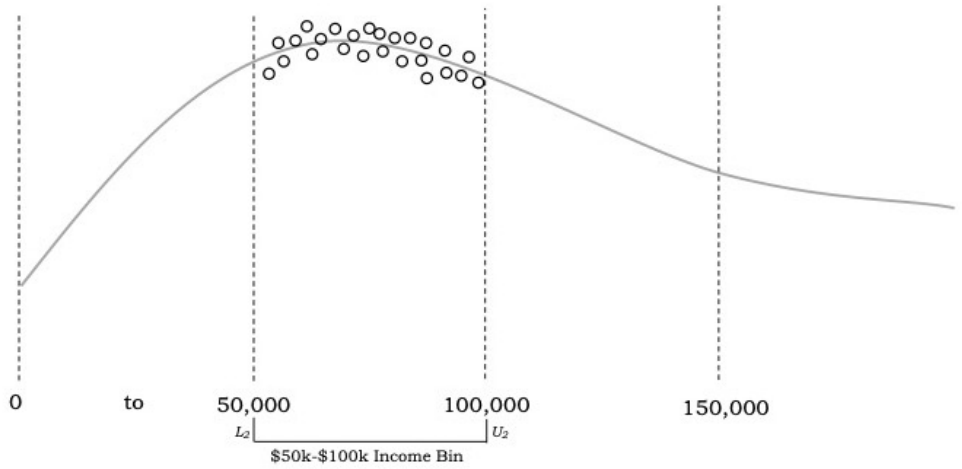

Figure 1: Illustration of how the `redi` package samples within categorical income boundaries to provide continuous income values from a reference dataset. The horizontal axis represents a continuous distribution of income with vertical dashed lines indicating upper and lower bounds for each category in the research dataset of interest. Open circles correspond to example draws of individual discrete incomes from the reference dataset. The curved line is a scaled probability density dunction representing the probability of sampling each discrete income from the reference dataset, smoothed across the entire income distribution.

2.3 Reference dataset choice: Current Population Survey Annual Social and Economic Supplement (CPS ASEC)

The reference dataset used for the `redi` package, again, is the Current Population Survey Annual Social and Economic Supplement (CPS ASEC). The CPS ASEC interviews 200,000 individuals from around 100,000 households of the non-institutionalized U.S. population per year, asking about their income in the previous calendar year, family

1. These random integer values are generated using the Stata command [FN] **runiformint**. Each value from within the income bounds (L_b to U_b) in the reference dataset has probability of being assigned to each observation in the research dataset drawn from a $(1, N_b)$-uniform variable.

and household characteristics, and program participation (U. S. Census Bureau and U. S. Bureau of Labor Statistics 2018). To avoid collinearity among members of the same household, the `redi` package selects only one member from each household. The data are publicly available to all researchers and quite reliable: self-reported earnings in the CPS ASEC are 96.2% to 99.0% of benchmarks generated using administrative data (Rothbaum 2015). The CPS ASEC is also used as the official data source for national poverty estimates in the U.S. (U. S. Census Bureau 2018).

Prior to using the `redi` package, the user must download the CPS ASEC reference dataset from the IPUMS CPS website (cps.ipums.org) for the year(s) of interest. The variables needed are: YEAR, ASECWTH, HHINCOME, and PERNUM. Place the downloaded CPS ASEC dataset in the current working directory and name the file "cps_reference.dta".[2]

3 Advantages, Assumptions, and Limitations

3.1 Advantages

As mentioned earlier, the main advantage of the `redi` package is that it can be used to compute continuous incomes from categorical ones, allowing the user to reconcile bins between datasets or across years. For example, in working with datasets spanning many years, a researcher may find that the income categories provided to survey respondents have changed over time. By employing the `redi` package on the research dataset, the user can reconcile these disparate income categories and convert all income responses into a single, inflation-adjusted dollar amount (for example, a variable measuring income in 2022 dollars across all datasets).

Other advantages include computing an income distribution that is nonparametric, bin consistent, and area- and variance-preserving. Because it employs an empirically determined distribution of U.S. income to impute continuous incomes – the Current Population Survey March Annual Social and Economic Supplement (CPS ASEC) (Flood et al. 2021) – as the reference dataset, the `redi` package minimizes assumptions about the shape of the distribution. Since `redi` draws from this reference dataset randomly with replacement, the resulting values are independent and identically distributed (IID), escaping a common failure of imputed observations. Since the CPS ASEC also has no minimum or top-coded values after 2011 (Minnesota Population Center 2018), `redi` also makes no assumptions about top or bottom categories. This is a comparative advantage relative to other methods (and packages) which fit a Pareto distribution to the top one or two income categories (e.g., von Hippel and Powers 2015; Jargowsky and Wheeler 2018).

2. Additional details on using the CPS ASEC Public Use Microdata, including technical documentation and details about analysis using survey weights, are available at https://www.census.gov/topics/population/foreign-born/guidance/cps-guidance/using-cps-asec-microdata.html.

3.2 Assumptions and Limitations

The `redi` package can be used with three key assumptions:

1. the sample frame and sampling method are substantively identical in the research and reference datasets;

2. the income question is the same in the reference and research datasets; and

3. income is not top-coded in the reference dataset.

Assumption (1) concerns matching the sampling frame and method between the research and reference datasets. The CPS ASEC is a solid reference dataset for household, family, and individual level income at the U.S. nationally representative level and also at smaller regional levels including states and census tracts. The researcher will need to select that portion of the CPS ASEC which best represents the income data from their research dataset.

Assumption (2) must be assessed by the researcher to assure that the definition of income being used in the research dataset is the same as that used in the CPS ASEC. The first important check, of course, is to use the same unit of income: household, family, or individual. Second, it is important to be sure that the research dataset includes the same income sources as the CPS ASEC in its survey question or calculation: the CPS ASEC asks about income from over 50 sources, including wage and salary, self-employment, interest and dividends, Social Security, pensions, family assistance, worker compensation, and unemployment compensation (Rothbaum 2015).

Assumption (3) is met by using the CPS ASEC dataset as the reference dataset, as long as the user is working after the year 2011. Beginning in 2011, the CPS ASEC uses a rank proximity procedure to switch top incomes with near-neighbor approximate values. Individual incomes at risk of being identified are rounded to two significant digits and swapped within particular bounds (Minnesota Population Center 2018). This has no influence on the `redi` method becuase it uses the full CPS ASEC dataset to draw the observations for the research dataset, so swapping incomes in the highest category does not influence the calculations. This is particularly true because `redi` does not rely on correlated predictors. The Census Bureau uses a similar procedure between 1996 and 2010. However, users interested in using these older data should first investigate the details of the different top-coding procedures across years.

Limitations also apply to the usefulness of the `redi` package. The performance of the method depends on the quality of the reference data (Hu and Salvucci 2001) and how well these match the data-generating process for the research data. Additionally, the `redi` package does not allow for correlated predictors of income in calculating continuous income observations. However, this also means that the package does not require them, so a high-fidelity model does not need to exist.

4 Syntax

`redi` *incvar year* `, generate(`*newvar*`)` [`cpstype(`*string*`) inflationyear(#)`]

where *incvar* requests the name of the categorical income variable in the original research dataset; *incvar* may be either a string variable (with the categories as text) or a numeric variable (with the categories storied as value labels). The command handles missing values in the *incvar* variable input by translating all missing values for the *incvar* variable to the value "98". The user will want to verify that none of the existing codes for *incvar* are meaningfully assigned "98" prior to using the `redi` command. At the conclusion of the program, "98" values for *incvar* are automatically decoded back to missing values. Since all values are converted back to ".", these may not be the exact missing values from the original research data.

The *year* argument asks the user to specify the name of the year variable in the original research dataset.

Finally, `generate(`*newvar*`)` specifies a name for the new continuous income variable calculated using the redi method. In the process of producing the continuous value, the `redi` command will also generate a lower-bound variable (`incvar_lb`) and an upper-bound variable (`incvar_ub`) for the continuous income variable drawn from the reference dataset. These can be used to verify the new continuous variable or dropped at the researcher's convenience.

4.1 Options

`cpstype(`*string*`)` specifies the type of income reference variable to use from the CPS ASEC reference dataset. Permitted options are "household", "family", or "respondent"-level income.

`inflationyear(#)` specifies the year to which the data should be inflated using the R-CPI-U-RS. The year should be specified as a 4-digit number. If no inflation adjustment is desired, the user should not specify.

4.2 Notes on the inflation option

With the inflation option, the `redi` command produces both the new continuous income variable *newvar* and *newvar_inf#*, another new variable adjusted for inflation using the specified inflation year. If the `inflationyear(#)` option is specified, the `redi` package uses the Consumer Price Index retroactive series using current methods with all items (R-CPI-U-RS) dataset to calculate inflation.[3] The Retroactive Series (R-CPI-U-RS) estimates the Consumer Price Index for Urban Consumers from 1978, using current methods that incorporate these improvements over the entire time span. Using

3. The Consumer Price Index retroactive series using current methods with all items (R-CPI-U-RS) is available from the U.S. Bureau of Labor Statistics website (http://www.bls.gov/cpi/research-series/r-cpi-u-rs-home.htm).

the inflationyear option for automatically downloads this dataset for use in inflation adjustment. The year specified indicates the year that should be used for inflation-adjusted dollars. Using this option produces a variable named *newvar_inf#*.

Without specifying an inflation year, the `redi` command produces the continuous income variable *newvar*calculated in the dollar value corresponding to the year of the original research dataset.

5 Examples

The example described here deploys a small General Social Survey (GSS) dataset that is publicly available and can be easily downloaded at https://www.ssc.wisc.edu/sscc/pubs/sfs/gss_sample.dta Social Science Computing Cooperative (2016); Smith et al. (2015). In keeping with recommendations on transparent and open social science (Freese and King 2018), code for replication of this demonstration can be found at the GitHub repository folder "redi_package" at https://github.com/mollymking/redi_code.

The demonstration using GSS data demonstrates two uses of the `redi` program. The first calculates a new continuous income variable named `finc_continuous` from categorical family income variable `income` and year variable `year`. The family income type needs to be specified as an option.

```
. redi income year, gen(finc_continuous) cpstype(family)
```

The second example demonstrates a similar use of the `redi` program, adding inflation for the resulting continuous respondent income values, using the minimum abbreviations for the options:

```
. redi rinc_continuous year, g(rinc_continuous) cps(respondent) inf(2020)
```

This example produces not only the new variable `rinc_continuous`, but also `rinc_continuous_inf2020`, a continuous income variable inflated to 2020 dollars using the R-CPI-U-RS. A snapshot of our data structure (see Figure 2) demonstrates that the command decodes the original categorical data label into an upper (`rincome_ub`) and lower (`rincome_lb`) categorical boundary before drawing a random value between the two to produce our new continuous income value (`rinc_continuous`).

6 Discussion

The `redi` (Random Empirical Distribution Imputation) program discussed here transforms categorical numerical data into continuous values using a real-world income distribution. This provides users with a resource for producing a distribution of discrete, empirically-generated values for categorical income data, even when few or no covariates are available.

Use of the `redi` package is thus appropriate when three conditions are met:

picture(0,0)(0,0)(1,0)523.95749 (0,0)(0,-1)668.4975 (523.95749,0)(0,-1)668.4975 (0,-668.4975)(1,0)523.95749 picture

	rincome	rincome_ub	rincome_lb	rinc_continuous	rinc_cont~2020
.	$4000 TO 4999	4999	4000	4,026	4,406
.	$5000 TO 5999	5999	5000	5,100	5,582
.	$5000 TO 5999	5999	5000	5,280	5,779
.	$5000 TO 5999	5999	5000	5,364	5,871
.	$5000 TO 5999	5999	5000	5,364	5,871
	$6000 TO 6999	6999	6000	6,659	7,288
.	$8000 TO 9999	9999	8000	9,348	10,231
.	$8000 TO 9999	9999	8000	8,500	9,303
.	$8000 TO 9999	9999	8000	8,904	9,745
.	$8000 TO 9999	9999	8000	9,900	10,835
.	$10000 - 14999	14999	10000	13,497	14,772
.	$10000 - 14999	14999	10000	10,320	11,295
.	$10000 - 14999	14999	10000	13,259	14,511

Figure 2: Data produced by the `redi` command in Stata applied to GSS example data Social Science Computing Cooperative (2016); Smith et al. (2015). The column `rincome` displays the variable label (or text) for the original categorical income variable. The `redi` program decoded this textual category information to produce an upper (`rincome_ub`) and lower (`rincome_lb`) bound for each observation. The program then drew randomly within these bounds from the CPS ASEC to calculate a continuous income value, stored in the variable `rinc_continuous` in the figure. Since the user specified an inflation year (2020), the program also calculated income inflated to 2020 dollars using the R-CPI-U-RS and stored this value in `rinc_continuous_inf2020`.

1. The user has categorical income data that she seeks to convert to continuous income data.

2. The user has an available dataset from which she can draw the continuous reference distribution. The `redi` program is currently implemented such that the CPS ASEC satisfies this condition, but the user must ensure that the sampling frame and method are similar to her research dataset.

3. The user has determined that hot-deck multiple imputation is not appropriate or possible because a well-specified model is not yet available or auxiliary variables have not been identified (`redi`) does not account for correlated predictors of income).

The principal advantages of the `redi` program are that it can be used to reconcile income data across multiple income categories and/or years; it preserves variance for later operations; and it can handle top incomes with the use of the CPS ASEC reference dataset. The `redi` method is also bin consistent, nonparametric and produces a continuous distribution, with the option of inflating incomes to more recent values.

7 Acknowledgments

I am grateful to Jeremy Freese for his significant help with the syntax and implementation of this program. I thank Christof Brandtner, Jeremy Freese, and David Grusky for their feedback on the methodological paper that serves as the foundation for this program.

8 References

Blanchet, T., B. Garbinti, J. Goupille-Lebret, and C. Martínez-Toledano. 2018. Applying Generalized Pareto Curves to Inequality Analysis. *AEA Papers and Proceedings* 108: 114–18. https://www.aeaweb.org/articles?id=10.1257/pandp.20181075et al Analysis.pdf?dl=0.

Flood, S., M. King, R. Rodgers, S. Ruggles, and J. R. Warren. 2021. Integrated Public Use Microdata Series, Current Population Survey: Version 8.0 [dataset]. https://doi.org/10.18128/D030.V9.0.

Freese, J., and M. M. King. 2018. Institutionalizing Transparency. *Socius: Sociological Research for a Dynamic World* 4: 1–7.

von Hippel, P. T., D. J. Hunter, and M. Drown. 2017. Better Estimates from Binned Income Data: Interpolated CDFs and Mean-Matching. *Sociological Science* 4: 641–655. https://www.sociologicalscience.com/articles-v4-26-641/.

von Hippel, P. T., and D. A. Powers. 2015. RPME: Stata Module to Compute Robust Pareto Midpoint Estimator. https://ideas.repec.org/c/boc/bocode/s457962.html.

Hout, M. 2004. Getting the Most Out of the GSS Income Measures.

Hu, M.-x., and S. Salvucci. 2001. A Study of Imputation Algorithms.

Jargowsky, P. A., and C. A. Wheeler. 2018. Estimating Income Statistics from Grouped Data: Mean-Constrained Integration over Brackets. *Sociological Methodology* 48(1): 337–374.

King, M. M. REDI for Binned Data: A Random Empirical Distribution Imputation method for estimating continuous incomes. *Sociological Methodology* Forthcoming.

Ligon, E. 1989. The Development and Use of a Consistent Income Measure for the General Social Survey. *GSS Methodological Report 64* .

Minnesota Population Center. 2018. CPS Income And Tax Variables User's Note: Missing Cases, N.I.U. Cases, Top Codes And Bottom Codes. https://cps.ipums.org/cps/inctaxcodes.shtmltopcodes.

Piketty, T., and E. Saez. 2014. Inequality in the long run. *Science* 344(6186): 838.

Rothbaum, J. L. 2015. Comparing Income Aggregates: How do the CPS and ACS Match the National Income and Product Accounts, 2007-2012.

Smith, T. W., P. Marsden, M. Hout, and J. Kim. 2015. General Social Surveys, 1972-2014 [machine-readable data file]. http://gss.norc.org/.

Social Science Computing Cooperative. 2016. Stata for Students. Madison, WI: University of Wisconsin - Madison. https://www.ssc.wisc.edu/sscc/pubs/sfs/sfs-files.htm.

U. S. Census Bureau. 2018. How the Census Bureau Measures Poverty. www.census.gov/hhes/www/poverty/about/overview/measure.html.

U. S. Census Bureau, and U. S. Bureau of Labor Statistics. 2018. Current Population Survey March Annual Social and Economic Supplement (CPS-ASEC). https://cps.ipums.org/cps/asec$_s ample_n otes.shtml$.

9 Author information

About the author

Molly M. King is an assistant professor in the Department of Sociology at Santa Clara University. Previously, she earned her PhD from Stanford University, where she was a National Science Foundation Graduate Research Fellow. She studies knowledge inequalities and the implications of these inequalities for people's lives. A computational sociologist by training, she uses mixed methods, paired with a commitment to open science, to understand how the identities of disability, gender, and class influence information acquisition. Most recently, she is exploring these questions related to climate change. Find her CV, publications, and code at www.mollymking.com.