The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–23

GiniInc: A Stata package for measuring inequality from incomplete income and survival data

Long Hong
University of Wisconsin
Madison, WI, USA
long.hong@wisc.edu

Guido Alfani
Bocconi University
Milan, Italy
guido.alfani@unibocconi.it

Chiara Gigliarano
University of Insubria
Varese, Italy
chiara.gigliarano@uninsubria.it

Marco Bonetti
Bocconi University
Milan, Italy
marco.bonetti@unibocconi.it

Abstract. Quite often, observed income and survival data are incomplete due to left- or right- censoring or truncation. Measuring inequality, for instance by the Gini index of concentration, from incomplete data like this will produce biased results. We describe the Stata package ***GiniInc***, which contains three independent functions to estimate the Gini concentration index under different conditions. First, ***survgini*** computes a test statistic for the comparison of two (survival) distributions based on the non-parametric estimation of the restricted Gini index for right-censored data, using both asymptotic and permutation inference. Second, ***survbound*** computes non-parametric bounds for the unrestricted Gini index from censored data. Finally, ***survlsl*** implements maximum likelihood estimation for three commonly used parametric models to estimate the unrestricted Gini Index, both from censored and truncated data. We briefly discuss the methods, describe the package, and illustrate its use through simulated data and examples from an oncology and a historical income study.

Keywords: st0001, Gini index, income distribution, inequality, survival analysis, censored data, truncated data, survgini, survbound, survlsl

1 Introduction

The Gini Index

The Gini index is the most common statistical index used in the social sciences for measuring inequality or concentration in the distribution of a positive random variable such as income (Gini 1912, 1914). It is usually defined based on the Lorenz Curve as shown in Figure 1, in which the y-axis represents the cumulative proportion of the total income owned by the poorest percentage of the population (on the x-axis). The 45-degree line represents the case of no inequality because the total income is equally distributed in the population, and the Lorenz Curve measures how different the income distribution is from the equality distribution. Indeed, the Gini concentration index is equal to twice the area between the 45-degree line and the Lorenz Curve. So it is bounded between 0 and 1 - a Gini index of zero means no inequality, while a value of 1 represents maximal inequality.

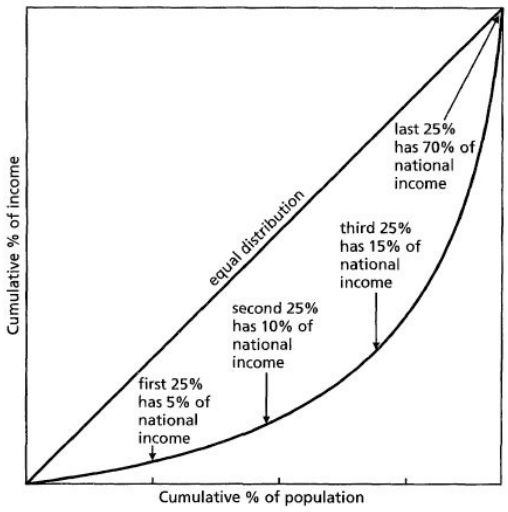

Figure 1: Lorenz Curve

The Gini index can also be expressed in the following mathematical representations. Consider a non-negative random variable X with cumulative distribution function $F(x)$. The Gini index is also written as:

$$G = \frac{\int_{\mathbb{R}^+} \int_{\mathbb{R}^+} |x_1 - x_2| dF(x_1) dF(x_2)}{2\mu},$$

where μ is the expected value of X (Gini 1912, 1914).

Although the Gini index is mainly used in economics as a measure of income or wealth inequality, it has recently been used to quantify the inequality in mortality.

Moreover, an important expression has also been developed that links the Gini index to the survival analysis:

$$G = 1 - \frac{\int_{\mathbb{R}^+} S^2(x)dx}{\int_{\mathbb{R}^+} S(x)dx}, \tag{1}$$

where $S(x) = P(X \geq x)$ (Michetti and Dall'Aglio 1957; Hanada 1983).

The literature has mostly focused on complete data, while less attention has been paid to censored or truncated data. The Stata package ***GiniInc*** focuses on the estimation of the Gini index from incomplete income or survival data.

This paper describes two specific situations. First, when the data are right censored, we briefly review existing methods to compare the distributions of two samples based on their non-parametric restricted Gini index. Such methods are implemented in the Stata function ***survgini*** (Section 2). Second, when the data are left censored or truncated, we have obtained non-parametric bounds for the (unrestricted) Gini index implemented in the Stata function ***survbound*** (Section 3). If one has an educated guess on the parametric form of the distribution, we can estimate the unrestricted Gini and its corresponding large-sample confidence interval from likelihood maximization. This is implemented in the Stata function ***survlsl*** (Section 4).

Censoring vs. Truncation

This section provides a brief review of censoring and truncation. Consider a lifetime (or income) random variable $X \geq 0$, with distribution $f_\theta(x)$. Also let $C \geq 0$ be another random variable, independent of X. Let us focus on censoring first.

> Observation of X is left-censored if one observes the *largest* between X and C, and knows which one it is. One example of such a setting is a study in which an animal is followed over time to obtain the age (X) at which it learns to perform a task. Left censoring occurs if the animal already knows how to perform the task at age (C), when observation begins. The corresponding observed data are called left-censored data.
>
> The observation of X is right-censored if one observes the *smallest* between X and C, and knows which one it is. This is a very common situation in survival studies, when one observes the time (X) from entry into the study until some event (e.g., death) occurs, and when the observation of the event is only possible until the end of the study (C time units from entry). Note that the maximum observation time for each individual here depends on the time of entry, and as such it typically differs across subjects.
>
> Right (or left) censoring can also occur with a constant C, for example if one considers the demand for tickets for a sporting event, which is to be held in a stadium that has (obviously) finite capacity. The number of tickets sold can only be as high as the stadium's capacity, which therefore makes the observation of

the demand for tickets right-censored by that fixed number (see example 19.4 in Greene (2012)).

Now, let us consider truncation.

> The variable X is left-truncated if one observes X only when $X \geq C$, and it is right-truncated if one observes X only when $X \leq C$. Truncation will often occur with respect to a constant variable $C = k$. This occurs, for example, when there exists a lower (or upper) detection threshold for an instrument, so that measurements below (or above) the threshold do not occur.

Summarizing, with censoring one does observe the individual, but the value that is collected might not correspond to the true underlying value that one is interested in (e.g. in Greene's example, the fact that all tickets were sold does not mean that the demand for tickets was exactly equal to capacity). In contrast, for truncated variables one does not even observe the individual/observation with values of the variable that are outside some range.

Let us analyze the difference between (left) censoring and truncation in a bit more detail. We focus on the case of constant C, which will be relevant in what follows. Consider a city that has an income tax system that requires its citizens to pay tax only if their incomes are above a certain fixed threshold, and suppose for simplicity that one only cares about that single variable - income. One can easily recover the income data from the taxes that have been paid, but the income data will not be available for any of those who do not pay taxes. If we do not know how many citizens are below the threshold, then the observed data in the tax records are left-truncated, since exempt individuals are simply not in the dataset. If, on the other hand, one also knows the total number of citizens in the town (including the number of tax-exempt citizens), then the income data can be augmented by adding these individuals, who are known to have an income below the threshold. The resulting more-informative dataset would now be larger and made of left-censored data.

We can describe the difference between the two cases in terms of the likelihood constructed from the two observed data distributions, when a parametric model describes the population.

Observed left-censored data can be written as the pairs $(y_1, \delta_1), ..., (y_n, \delta_n)$, where δ_i is the event indicator, which is equal to 0 if the income is left-censored by the fixed and known value k, and 1 otherwise. Denote x_i as the true underlying income of individual i, and

$$y_i = \max(x_i, k)$$
$$\delta_i = I(k \leq x_i).$$

Assume that the true incomes $x_1, ...x_n$ are an iid sample from the model density $f_\theta(x)$.

Then the likelihood function for the observed data is

$$L(\theta) \propto \prod_{i=1}^{n} \{f_\theta(y_i)\}^{\delta_i} \{F_\theta(X_i \leq k)\}^{1-\delta_i}$$

(note that $y_i = k$ whenever $\delta_i = 0$). By using maximum likelihood estimation in Stata, both the estimate $\widehat{\theta}$ and its standard error $se(\widehat{\theta})$ can be easily obtained.

In the case of left truncation, we observe only the cases such that $x_j \geq k$, but we do not know how many of the observations are missing, i.e. the cases with $x_j < k$. Denote by m the number of available observations $y_j = x_j, j = 1, ..., m$. The left-truncated income are therefore an iid sample from $f_\theta(x|X \geq k)$, and the observed data likelihood is

$$L(\theta) \propto \prod_{j=1}^{m} \frac{f_\theta(y_j)}{F_\theta(X \geq k)}.$$

Again, using maximum likelihood estimation in Stata, both the estimate $\widetilde{\theta}$ of θ and its standard error $se(\widetilde{\theta})$ can be obtained.

To illustrate the implications of the two sampling models empirically, we have generated some data from a log-normal distribution with parameters $\mu = 3$ and $\sigma = 2$. We simulated with different combinations of the number of observations n. The threshold k was set to 10, so that roughly 30% of the original simulated observations were below k and not observed. We repeated the simulation 1,000 times and report the results in Table 1, which shows that when both n and m are large, $\widehat{\mu}$ (and $\widehat{\sigma}$) and $\widetilde{\theta}$ (and $\widetilde{\sigma}$) naturally converge to the true μ (and σ). Note that the estimation under left truncation is less precise, i.e. $se(\widehat{\theta}) < se(\widetilde{\theta})$, which is somewhat expected. A thorough study of estimation under the two sampling schemes is beyond the scope of this paper.

Table 1: Empirical estimation with left censoring $(\widehat{\mu}, \widehat{\sigma})$ vs. left truncation $(\widetilde{\mu}, \widetilde{\sigma})$

n	$\widehat{\mu}$	$\widehat{\sigma}$	$\widetilde{\mu}$	$\widetilde{\sigma}$
50	2.97 (.318)	1.98 (.284)	2.71 (1.63)	1.99 (.615)
100	2.99 (.217)	1.98 (.191)	2.83 (1.17)	2.00 (.472)
300	3.00 (.124)	1.99 (.112)	2.96 (.509)	1.99 (.247)
500	3.00 (.099)	2.00 (.086)	2.97 (.391)	2.00 (.190)
1000	3.00 (.068)	2.00 (.061)	2.98 (.283)	2.00 (.138)

Standard errors in parentheses

Recall the tax example we mentioned earlier. Suppose the mayor knows that the income data of the city actually follow a log-normal distribution. The analysis above implies that he can always use the maximum likelihood estimation to recover the true parameters of the log-normal distribution when the income data are either left-censored or left-truncated.

Section 2 below discusses right-censored data with general censoring variable C. Sections 3 and 4 will focus on censored and truncated data when the variable C is constant. Specifically, Section 3 will discuss non-parametric bounds for the Gini index when the observed data are left-censored, while Section 4 will discuss the general parametric estimation of the Gini index.

2 Non-parametric restricted Gini for right-censored survival data

In Bonetti et al. (2009) a nonparametric test has been proposed based on the Gini index for testing the equality of two survival distributions from the point of view of concentration based on two independent right-censored samples from the two populations.

Let $X_1, \ldots, X_n$ be an i.i.d. sample from X observed only partially, in particular after random right censoring (independent of X). The Gini index can be modified for application to lifetime data, in which individuals have finite follow-up time for survival by defining the *restricted* Gini index

$$G_t = 1 - \frac{\int_0^t S^2(u)du}{\int_0^t S(u)du}, \tag{2}$$

rather than the traditional unrestricted Gini index G (whose integrals in (2) would run from zero to infinity, see (1)). The time t indicates the longest follow-up time in the data. The estimator proposed for the restricted Gini index for right censored data is

$$\widehat{G}_t = 1 - \frac{\int_0^t \widehat{S}^2(u)du}{\int_0^t \widehat{S}(u)du},$$

where $\widehat{S}(u)$ is the Kaplan-Meier estimator of $S(u)$; see Kaplan and Meier (1958).

The authors have shown that, under some regularity conditions, the scaled estimator $\widehat{G}_t$ follows an asymptotically normal distribution, with an explicit expression for the asymptotic variance. A test has been proposed for comparing two survival functions estimated from the independent samples of sizes n_1 and n_2. The Gini test statistic is

$$T_t := \frac{\left(\widehat{G}_{1,t} - \widehat{G}_{2,t}\right)^2}{\widehat{Var}(\widehat{G}_{1,t}) + \widehat{Var}(\widehat{G}_{2,t})}, \tag{3}$$

where $\widehat{G}_{j,t}$ is the estimator of the restricted Gini index for censored data for group j and $\widehat{Var}(\widehat{G}_{j,t})$ is the estimator of the sampling variance of $\widehat{G}_{j,t}$ for group j, $j = 1, 2$. In Bonetti et al. (2009) a simulation analysis is described, in which the Gini test is compared to other tests for the difference between two survival distributions, such as the log-rank, Wilcoxon, and Gray-Tsiatis tests.

Further, Gigliarano and Bonetti (2013) compared the asymptotic inference with a permutation approach, suggesting that the permutation test should be preferred to

the asymptotic test, especially in the case of unbalanced and small groups. The Gini permutation test is a permutation test procedure applied to the test statistic $(\widehat{G}_{1,t} - \widehat{G}_{2,t})^2$, constructed as follows:

(i) Compute the test statistic (note that this is the numerator of T_t) for the original data.

(ii) Repeat the following M times (with index $m = 1, ..., M$):

- sample a permutation π^m from all permutations of the (n_1+n_2) group labels;
- compute the test statistic value $g_t^{(m)}$ from the original data, but with the permuted group labels π^m.

(iii) Estimate the permutation distribution of the test statistic with the empirical cumulative distribution function obtained from the permutated samples.

(iv) Obtain the permutation p-value p_0 corresponding to the value of the test statistic observed on the original data from the empirical distribution estimated in (iii). If $p_0 \leq \alpha$ for the given significance level α, we reject the null hypothesis of equality of the two survival distributions.

Our Stata function ***survgini*** not only implements the asymptotic and the permutation restricted Gini tests, but also compares the log-rank and Wilcoxon tests, to allow for an immediate comparison of the results. Note that this function replicates what was available in the R package *survgini* (Gigliarano and Bonetti 2011).

2.1 Syntax of the survgini command

The syntax of ***survgini*** is:

`survgini` *time failure treatment* [*if*] [*in*] [, *options*]

where the sequence of the variable list must be *fixed*.

Table 2 demonstrates a typical data structure of the three variables in survival analysis. *time* is the time-to-event variable. *failure* is a dummy variable, which is equal to zero if the survival time is right-censored. *treatment* is a categorical variable with one standing for the first group and two standing for the second group.

The syntax also contains a set of *options* for various purposes:

- *nolastevent*: Integrate the restricted Gini statistic until the last censored or non-censored observation
- *nolinearrank*: Inactivate the production of the two linear rank tests (log-rank test and Wilcoxon test)

Table 2: An illustration of the data structure for *survgini*

time	*failure*	*treatment*
2.69678	1	1
6.38193	0	2
5.61533	0	1
⋮	⋮	⋮

- *noasymptotic*: Inactivate the asymptotic Gini test
- *nopermutation*: Inactivate the permutation Gini test
- *m(integer)*: Number of replications of permutation sampling; default = 500.

2.2 Example

We will use the survival data from the Eastern Cooperative Oncology Group (ECOG) phase III clinic trial E1690[1], which accrued patients from 1991 to 1995 and was unblinded in 1998. This trial was aimed at comparing the effect of Interferon alpha-2b chemotherapy (IFN) to observation only in patients affected by high-risk melanoma. Trial E1690 was a randomized three-arm clinical trial that compared high dose IFN, low dose IFN, and control. To illustrate ***survgini***, we use relapse-free survival (RFS)[2] data from the treatment group with high dose IFN and data from the control group (215 and 212 patients, respectively).

Figure (2) shows the Kaplan-Meier estimates for RFS for both groups. The solid line and dashed line represent the survival time of the treatment group and the observation group, respectively. The analysis of trial E1690 has showed that the high dose IFN has a significant impact on RFS (Kirkwood et al. 2000). We will now examine whether the two survival distributions are significantly different with respect to their concentrations.

The implementation of ***survgini*** is straightforward by simply inputting the three variables of interest in the *orders* as shown below. Note that when the sample size is relatively large, we shall inactivate the permutation test, which is mainly for the cases where the sample size is small. The function reports the corresponding p-values of different tests. pGiniAs, pLR, and pW stand for the p-value of the Gini asymptotic test, the log-rank test, and the Wilcoxon test, respectively. The results show that the difference in concentration between the two groups is marginally significant at 5% confidence level.

```
. survgini failtime failcens trt, noperm
Comparison among GiniAs Log-rank and Wilcoxon tests
```

1. The data are available from: http://merlot.stat.uconn.edu/∼mhchen/survbook/
2. This is the time from randomization until relapse (progression of the disease).

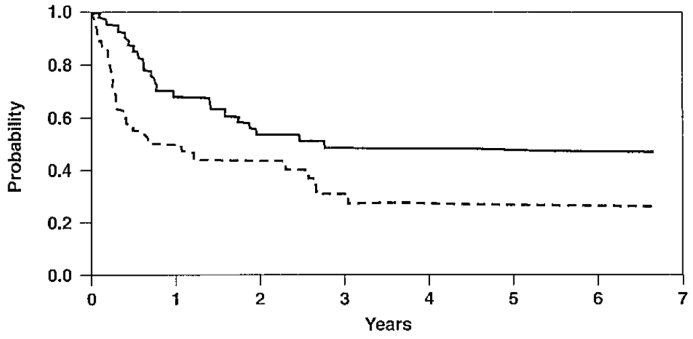

Figure 2: Kaplan-Meier Estimate of Relapse-free survival (Kirkwood et al. 2000)

```
                 |    pGiniAs        pLR         pW
-----------------+---------------------------------
        pval     |     .0526      .05391      .03506
        stat     |    3.7565      3.7154      4.4421
```

```
. return list
scalars:
            r(pGiniAs) =  .0526027215785181
                r(pLR) =  .0539137282127673
                 r(pW) =  .0350623367664427
          r(statGiniAs) =  3.756495446427073
              r(statLR) =  3.71536833497514
               r(statW) =  4.442136830000731
        r(statGiniPerm) =  .003742444738037
```

If the sample size does not seem large enough to produce reliable asymptotic results from asymptotic Gini, log-rank, and Wilcoxon tests, the permutation test can be used by inactivating the asymptotic tests as shown below.

```
. set seed 20171121

. survgini failtime failcens trt, nolin noas
Gini Permutation Test
                 |  pGiniPerm
-----------------+------------
        pval     |        .05
        stat     |      .0037

. return list
scalars:
          r(pGiniPerm) =  .05
        r(statGiniPerm) =  .003742444738037
```

However, attention must be paid to the following three direct consequences of the fact that the permutation test involves replications of permutation sampling. First, it will significantly slow down the speed of programming, especially when *m(integer)* is set to be large. Second, the result can be change slightly every time it is executed. We strongly recommend use of the ***seed*** function before using the permutation test for replication purposes. Finally, the test statistic is the numerator of Equation 3, which is not comparable to the rest of the tests whose test statistic follows an asymptotic Chi-square distribution.

3 Non-parametric Gini index estimation for left-censored data

This section explores the upper and lower bounds of the Gini index for *left-censored* income data. There are two important reasons why we consider only left-censored income data. First, if the data are right-censored, the largest unobserved value(s) could theoretically be as large as $+\infty$, thus corresponding to a true Gini index as large as 1 (which would be the upper bound). Second, if the data are left-truncated, then the proportion below the threshold is unknown, which would push the lower bound to zero. Therefore, in this section, we explore only the *left-censoring* case with a fixed censoring value k. We will discuss the truncated data case further in the next section.

3.1 Gini bounds for left-censored data

In the left-censoring setting, the percentages of incomes (survival times) below and above the threshold are known. For ease of description, we refer to incomes below (left censored survival times are more rare). Also, we observe all the incomes above the threshold, but not the ones below the threshold. While it is impossible to compute the exact Gini index for the distribution, we search for the possible upper and lower bounds of the index by using the Gini decomposition method (Yitzhaki and Schechtman 2013). The Gini index can be decomposed as follows:

$$G = s_1\pi_1 G_1 + s_2\pi_2 G_2 + \frac{\pi_1\pi_2(\mu_2 - \mu_1)}{\mu} \tag{4}$$

where indices 1 and 2 refer to below and above a fixed threshold k, respectively. The values s_j, G_j, π_j, and μ_j indicate the share of total income, the Gini index, the share of total number of observations, and the total mean, respectively for group j. G and μ refer to the entire population. In our case, π_1 and π_2 are known. G_2 and μ_2 can be easily computed from the (known) incomes above k. However, the rest of the parameters are

unknown, and they all depend on the unknown μ_1:

$$\begin{aligned}\mu(\mu_1) &= \mu_1\pi_1 + \mu_2\pi_2 \\ s_1(\mu_1) &= \frac{\mu_1\pi_1}{\mu_1\pi_1 + \mu_2\pi_2} \\ s_2(\mu_1) &= \frac{\mu_2\pi_2}{\mu_1\pi_1 + \mu_2\pi_2}.\end{aligned}$$

The value of G_1 is also unknown and it also depends on the unknown shape of the distribution below the threshold k. Thus, we can rewrite G as a function of G_1, G_2 μ_1, μ_2, π_1, and π_1, where only G_1 and μ_1 are the unknown parameters:

$$G = \underbrace{\frac{\mu_1\pi_1^2 G_1}{\mu_1\pi_1 + \mu_2\pi_2}}_{G_{<k}(\mu_1, G_1)} + \underbrace{\frac{\mu_2\pi_2^2 G_2}{\mu_1\pi_1 + \mu_2\pi_2}}_{G_{>k}(\mu_1)} + \underbrace{\frac{\pi_1\pi_2(\mu_2 - \mu_1)}{\mu_1\pi_1 + \mu_2\pi_2}}_{GB(\mu_1)}$$

and where $G_{<k}$, $G_{>k}$, and GB are the three components of Equation (4). We now examine the behavior of each component by taking a derivative of μ_1.

1. For $G_{<k}(\mu_1, G_1)$:

From the expression of $G_{<k}(\mu_1, G_1)$, we cannot tell too much about its behavior as a function of the unknown μ_1, G_1. However, since we know that $s_1 \in [0, \frac{\pi_1 k}{\pi_1 k + \pi_2\mu_2}]$ and $G_1 \in [0, 1]$, we can compute the following bounds:

$$G_{<k}(\mu_1) \in [0, \frac{\pi_1^2 k}{k\pi_1 + \mu_2\pi_2}).$$

The lower bound can be reached either by setting $G_1 = 0$ or $\mu_1 = 0$. It seems that the upper bound may be reached by setting $\mu_1 = k$ and $G_1 = 1$; however, this is not a realistic case. Indeed, if $\mu_1 = k$, then the incomes below k are equally distributed, i.e. everyone below the threshold has the same income k, which means the $G_1 = 0$. If $G_1 = 1$, then all the income is owned by one person, which means that with a relatively large sample, $\mu_1 \approx 0$. Therefore, the two conditions cannot be reached simultaneously.

2. For $G_{>k}(\mu_1)$:

$$\frac{\partial}{\partial\mu_1} G_{>k}(\mu_1) = -\frac{\mu_2\pi_1\pi_2^2 G_2}{(\mu_1\pi_1 + \mu_2\pi_2)^2} < 0,$$

so that $G_{>k}(\mu_1)$ is, in non-trivial cases, decreasing with respect to μ_1. Since $\mu_1 \in [0, k]$, it follows that:

$$G_{>k}(\mu_1) \in [\frac{\mu_2\pi_2^2 G_2}{k\pi_1 + \mu_2\pi_2}, \pi_2 G_2].$$

3. For $GB(\mu_1)$**:**

$$\frac{\partial}{\partial \mu_1} GB(\mu_1) = -\frac{\pi_1 \pi_2 \mu_2}{(\mu_1 \pi_1 + \mu_2 \pi_2)^2} < 0,$$

so that $GB(\mu_1)$ is also decreasing with respect to μ_1. Since $\mu_1 \in [0, k]$, it also follows that:

$$GB(\mu_1) \in [\frac{\pi_1 \pi_2 (\mu_2 - k)}{k\pi_1 + \mu_2 \pi_2}, \pi_1].$$

Combining all the elements from above, we are now able to construct the bounds of the overall Gini index G as follows:

$$G \geq 0 + \frac{\mu_2 \pi_2^2 G_2}{k\pi_1 + \mu_2 \pi_2} + \frac{\pi_1 \pi_2 (\mu_2 - k)}{k\pi_1 + \mu_2 \pi_2} = \frac{\mu_2 \pi_2^2 G_2 + \pi_1 \pi_2 (\mu_2 - k)}{k\pi_1 + \mu_2 \pi_2}.$$

where the equal sign can be reached by setting $\mu_1 = k$ and $G_1 = 0$;

$$G \leq \min\{1, \frac{\pi_1^2 k}{k\pi_1 + \mu_2 \pi_2} + \pi_2 G_2 + \pi_1\}$$

where the equal sign for the second term in the curly brackets cannot be reached because the first component requires $\mu_1 = k$ while the last two components require $\mu_1 = 0$.

Our Stata function ***ginibound*** computes the bounds[3], and it tries to improve on them by also implementing a numerical method to obtain the approximate upper bound using a grid search method (see below).

3.2 Syntax of the survbound command

The syntax of ***survbound*** is

`survbound` *income*`, thres`**`hold(`***real***`)`** `censorpct(`*real*`)` [`grid(`*integer*`)`]

where *income* could be any non-negative variable. In order to have a complete syntax, one has to provide the threshold for the observed left censoring, as well as the percentage of the data that are left-censored. In the example below, the threshold will be 10 (shillings) and the percentage 0.3 (30%).

There is a non-compulsory option "grid(n)", where n is an integer, that allows grid-search by taking $(n-1)^2$ simple combinations of (μ_1, G_1) to improve the upper bound. Take $n = 10$ for instance. In our data, $\mu_1 \in [0, k]$ and since the Gini index below the threshold is by definition between 0 and 1, "grid(10)" generates 9 values of μ_1: $\{\frac{k}{10}, \frac{2k}{10}, ..., \frac{9k}{10}\}$ and 9 values of $G_1 : \{.1, .2, ..., .9\}$. It then uses all 81 combinations of (μ_1, G_1) to estimate the largest possible value for the overall Gini index (G).

3. We have used the user-written function *fastgini* (Sajaia 2007) to compute G_2, since data are completely observed beyond k.

3.3 A small simulation exercise

In order to see how the bounds behave in comparison with the true Gini index in different situations, i.e. with different percentages of population below the threshold, we present a small empirical analysis using simulated data (sample size $= 10,000$) from a log-normal distribution with location parameter $\mu = 2$ and shape parameters $\sigma = 0.5$. The true Gini index for this distribution is $G = 0.28$. Different thresholds are set at the $10\%, 20\%, 30\%, 40\%, 50\%$, and 60% percentiles (π_1), respectively. We repeat the simulation $1,000$ times and report the average lower and upper bounds, and their lengths.

Table 3: Empirical average lower bound, lengths, and upper bounds for the Gini index

π_1	Lower bound	Upper (A)	Length (A)	Upper (G)	Length (G)
0.10	.2547	.3131	.0584	.3134	.0503
0.20	.2512	.3930	.1418	.3729	.1134
0.30	.2403	.4910	.2507	.4347	.1865
0.40	.2227	.6097	.3870	.4985	.2684
0.50	.1989	.7522	.5532	.5638	.3582
0.60	.1414	.9258	.7844	.6187	.4725

Upper (A) = the average of the analytic (A) upper bound.
Length (A) = the average of the difference between lower and analytic upper bounds.
Upper (G) = the average of the numeric upper bound calculated by "grid(10)".
Length (G) = the average of the difference between lower and numeric upper bounds.

As expected, Table 3 shows that the true value of the Gini index (0.28) always lies between the lower bound and upper bounds. When the percentage (π_1) of population below the threshold is low, the gap between the bounds is very tight. However, as π_1 becomes larger, then the length of the bounds could become quite big and not very useful. For example, when the $\pi_1 = 0.40$, i.e. 40% of the observations are not observed, then the non-parametric estimation of the Gini index ranges from 0.22 to 0.61, which could be of little help for describing the true Gini index.

Also, the numeric upper bound obtained by the "grid-search" method is very close to the analytic one when π_1 is small. However, as π_1 becomes larger, the gap between the numeric upper bound and the analytic upper becomes non-trivial, which sheds light on the fact that the analytic upper bound behaves poorly when π_1 gets large. In particular, when $\pi_1 = 60\%$, the upper bound is quite close to 1, which does not give as much information for the true G since $G \in [0, 1]$. The simulation exercise shows the importance of also computing the "grid-search" numeric upper bound when π_1 is large.

3.4 Example

The data are provided by the project *EINITE*[4] - Economic inequality across Italy and Europe, 1300 - 1800. The data cover historical household wealth based on the English "lay subsidies", which were levied across the country according to a uniform regulation. The lay subsidies provide a unique opportunity to study economic (wealth) inequality in late-medieval England. In the following we will use the terminology "income" throughout. Here, we use the historical data (sample size = $5,694$) from the lay subsidy of the country of Warwichshire, England, levied in 1332. The tax-paying threshold was set at 10 shillings, which means that a relatively large share of the overall population (those with wealth below 10 shillings) was exempted from paying the lay subsidy and consequently does not appear in the records. From other historical sources, the missing households can be estimated at approximately 30% of the total (Alfani and García-Montero 2018).

The implementation of ***survbound*** is straightforward by inputting the variable of interest (*income*), the value of the threshold (here, 10), and the percentage of censoring (here, 0.3). As shown below, the command computes the lower and upper bounds, and saves as the two scalars (r(lower_a) and r(upper_a)). The result shows that the overall Gini index is between 0.43 and 0.58.

```
. survbound income, thres(10) censorpct(0.30)

Non-Parametric Gini Numeric Boundaries:
```

	Lower(A)	Upper(A)
Non-Parametric Gini	.42755	.57873

```
Lower(A): Analytic lower bound
Upper(A): Analytic upper bound

. return list
scalars:
            r(lower_a) =  .4275491624079315
            r(upper_a) =  .5787303443070373
```

The grid-searching method can be easily implemented by adding the "*grid*" option. The result below shows that the upper bound found by grid-searching is slightly smaller than the analytic upper bound. The value of the upper bound computed by the "grid" option is saved as the scalar (r(upper_g)).

```
. survbound income, thres(10) censorpct(0.30) grid(10)

Non-Parametric Gini Numeric Boundaries:
```

	Lower(A)	Upper(A)	Upper(G)

4. Please refer to www.dondena.unibocconi.it/EINITE

```
------------------------+-----------------------------------
 Non-Parametric Gini    |     .42755      .57873      .53898
------------------------+-----------------------------------
Lower(A): Analytic lower bound
Upper(A): Analytic upper bound
Upper(G): Upper bound approximation by Grid-search
. return list
scalars:
            r(lower_a) =  .4275491624079315
            r(upper_a) =  .5787303443070373
            r(upper_g) =  .5389826627857929
```

4 Parametric log-scale-location models for incomplete data

If the distribution of the data is (assumed) known, one can calculate the Gini index for both censored or truncated data through maximum likelihood estimation. Because left-censoring or truncation is analogous to right-censoring or left-truncation in our settings, we only focus on left- censoring and truncation below.

4.1 Interval estimation of the Gini index for parametric models

Again, suppose that the income distribution can be represented by a probability density function $f(x)$ with the corresponding cumulative distribution function F(x). Defining μ as the mean of the distribution, the Gini coefficient can also be written as:

$$G = 1 - \frac{1}{\mu}\int_0^\infty (1 - F(y))^2 dy.$$

For three commonly used parametric log-scale-location models, the explicit analytic expressions of the Gini index are available as shown in Table 4. Note that since the Gini index is scale-invariant, it does not depend on the scale parameter.

If we can reasonably assume that the left- censored or truncated sample follows some parametric model, then we can perform maximum likelihood estimation using the likelihood functions as defined in the previous section, and obtain the estimated parameters. From these, we can thus calculate the Gini index. We can construct a large-sample confidence interval for the estimated Gini index using either a direct approach or the Delta method. Below we focus on the log-normal distribution for a detailed illustration of the two approaches.

The direct approach (C.I. 1)

Since the Gini index of the log-normal distribution is a function of σ only, we should recall the relevant properties of σ. By the asymptotic property of MLEs, $\widehat{\sigma}$ asymptotically normal:

$$\widehat{\sigma} \approx \mathcal{N}(\sigma, \frac{1}{nI(\mu, \sigma)}) \quad \text{for large } n,$$

Table 4: A selection of log-scale-location parametric models

Model	Density Function	Gini Index
Log-normal $(\mu,\ \sigma)$	$\frac{1}{x\sigma\sqrt{2\pi}}\exp[-\frac{(\ln x-\mu)^2}{2\sigma^2}]$	$2\Phi(\frac{\sigma}{\sqrt{2}})-1$
Weibull $(\alpha,\ \beta)$	$\frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}\exp[-(\frac{x}{\alpha})^{\beta}]$	$1-2^{-\frac{1}{\beta}}$
Log-logistic $(\alpha,\ \beta)$	$\frac{(\beta/\alpha)(x/\alpha)^{\beta-1}}{(1+(x/\alpha)^{\beta})^2}$	$1/\beta$

where $I(\mu,\sigma)$ is the Fisher Information. Using Slusky's theorem, the 95%-level large-sample confidence interval for σ is

$$\sigma \in \left(\widehat{\sigma} \pm z_{0.025} * \widehat{\delta}\right),$$

where $\widehat{\delta} = \sqrt{\frac{1}{nI(\widehat{\mu},\widehat{\sigma})}}$ is the standard error of $\widehat{\sigma}$, and $z_{0.025}$ the 0.975th percentile of the $N(0,1)$ distribution. Stata will produce $\widehat{\delta}$ automatically when estimating σ from the last iteration of the Newton-Raphson algorithm[5].

Once σ is estimated, one can then easily calculate the estimated Gini index using the formula in Table 4:

$$\widehat{G} = G(\widehat{\sigma}) = 2\Phi(\frac{\widehat{\sigma}}{\sqrt{2}}) - 1.$$

Note that the Gini index of the log-normal distribution is an *increasing* function of σ, as

$$\frac{\partial}{\partial\sigma}G(\sigma) = \sqrt{2} * \phi(\frac{\sigma}{\sqrt{2}}) > 0, \tag{5}$$

where ϕ is a density function of a normal distribution. Therefore, a 95% confidence interval is readily constructed as:

$$G(\sigma) \in \left(G(\widehat{\sigma} - z_{0.025} * \widehat{\delta}), G(\widehat{\sigma} + z_{0.025} * \widehat{\delta})\right). \tag{C.I. 1}$$

The delta method approach (C.I. 2)

Using the first-order derivative with respect to σ of the Gini index from Equation (5), by the delta method the following holds:

5. For more information, we refer to the Stata command ***ml***.

$$\widehat{G}(\sigma) \overset{a}{\sim} \mathcal{N}\Big(G(\sigma), [G'(\sigma)]^2 \frac{1}{nI(\mu,\sigma)}\Big) = \mathcal{N}\Big(G(\sigma), 2[\phi(\frac{\sigma}{\sqrt{2}})]^2 \frac{1}{nI(\mu,\sigma)}\Big),$$

and by Slutsky's theorem we can replace μ and σ by $\widehat{\mu}$ and $\widehat{\sigma}$, respectively. Therefore, an alternative 95%-level large-sample confidence interval for $G(\sigma)$ is

$$G(\sigma) \in \Big(G(\widehat{\sigma}) \pm z_{0.025} * \sqrt{2}\phi(\frac{\widehat{\sigma}}{\sqrt{2}}) * \widehat{\delta}\Big). \qquad \text{(C.I. 2)}$$

Note that the Gini index is a monotone function of the corresponding parameter for all three parametric models in Table 4. Therefore, the confidence intervals for the other parametric models can be obtained similarly. Table 5 summarizes the results.

Table 5: Confidence intervals for the estimated Gini index

Model	C.I. 1	C.I. 2
Log-normal $(\mu,\ \sigma)$	$G(\widehat{\sigma} \pm z_{0.025} * \widehat{\delta})$	$G(\widehat{\sigma}) \pm z_{0.025} * \sqrt{2}\phi(\frac{\widehat{\sigma}}{\sqrt{2}}) * \widehat{\delta}$
Weibull $(\alpha,\ \beta)$	$G(\widehat{\beta} \pm z_{0.025} * \widehat{\delta})$	$G(\widehat{\beta}) \pm z_{0.025} * (\widehat{\beta}^{-2} 2^{-1/\widehat{\beta}} \ln 2) * \widehat{\delta}$
Log-logistic $(\alpha,\ \beta)$	$G(\widehat{\beta} \pm z_{0.025} * \widehat{\delta})$	$G(\widehat{\beta}) \pm z_{0.025} * \widehat{\beta}^{-2} * \widehat{\delta}$

To empirically compare the two confidence intervals, we can use simulated data from three log-normal distributions with $\mu = 2$, but different shape parameters: $\sigma = 0.5, 1,$ and 1.5, respectively. The threshold k is set to $6, 5,$ and 3.5, respectively, so that roughly 30% of the original simulated observations are below k and are not observed. We repeat the simulation $1,000$ times and report, in Table 6, the probability that the true Gini index lies in the two confidence intervals and the average lengths of the intervals.

Table 6 shows that as the sample size grows, as expected, approximately 95% of the simulations produced confidence intervals that cover the true Gini index. When the sample size is small, the coverage is typically lower than 95%. The coverage results from C.I. 1 and C.I. 2 are similar, as are the average length of the two intervals.

4.2 Syntax of the survlsl command

The parametric methods mentioned above are implemented in the command ***survlsl***. The syntax of the command is as follows:

Table 6: Example - empirical comparisons of C.I.1 and C.I.2

		C.I.1		C.I.2	
True G	# Obs	Coverage	Avg length	Coverage	Avg length
	20	.91	.2166	.905	.2158
0.28	50	.94	.1366	.938	.1364
	100	.941	.0973	.94	.0972
	300	.944	.0562	.944	.0562
	20	.9	.3588	.892	.3566
0.52	50	.921	.2289	.916	.2283
	100	.948	.1628	.947	.1625
	300	.956	.0941	.954	.0941
	20	.918	.3755	.912	.3791
0.71	50	.926	.2413	.927	.2419
	100	.933	.1711	.922	.1713
	300	.951	.0996	.952	.0996

*Avg length = the average size of all the simulated confidence intervals.

`survbound` *income*`, thres`**`hold(`***real***`)`** `censorpct(`*real*`)` `model(`*string*`)`

which is very similar to ***survbound***. The main difference is that ***survlsl*** has one more compulsory option, "model(*string*)", to indicate the parametric model. The currently available models are log-normal distribution (lognormal), log-logistic distribution (loglogistic), and Weibull distribution (weibull)[6]. If one types "censorpct(0)", then ***survlsl*** treats the data as truncated, otherwise as censored.

4.3 Example

To illustrate the use of ***survlsl***, we will make use of the same historical income data mentioned previously. The only difference is that, for now, the percentage of censoring is uncertain. It might be 30% (as previously estimated) and in this case the data are left-*censored*; or the percentage can be completely unknown, and in this case the data are left-*truncated*. The key assumption is that the data follow a *log-normal* distribution. Figure 3 shows a histogram of the log income distribution, with a fitted normal distribution. The vertical line represents the log threshold, log(10). The figure suggests that the data fit a log-normal distribution well.

The syntax of ***survlsl*** is similar to that of ***suvbound***. The command computes the

6. The *exact* name should be specified in order to correctly indicate the model. For example, *survlsl* cannot recognize model(*lognorm*).

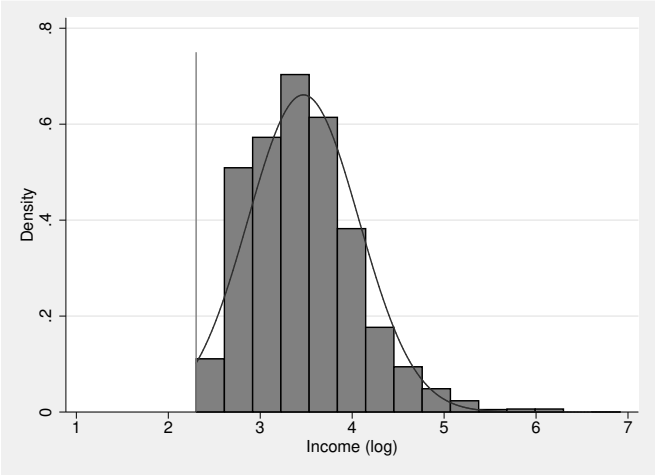

Figure 3: Distribution of the observed income data

maximum likelihood estimates of the location and scale parameters of the log-normal distribution, as well as the corresponding Gini index. One may notice that the Gini index (0.517) calculated using the parametric method is within the non-parametric bounds (0.427, 0.539) mentioned in the previous section (note that this need not be the case). All the estimates are saved in scalars and matrices. Moreover, ***survlsl*** reports the confidence intervals from both the direct approach and the delta method. Note that the confidence intervals are very tight ($n = 5694$). The results are stored in the 2×2 matrix r(conf_interval).

```
. survlsl income, thres(10) censorpct(0.30) model(lognormal)

(... MLE output omitted ...)

Left Censored Model

Estimated Parameters:
 MLE location  = 2.93999
 MLE scale     = .99122
Parametric Gini = .51663

Parametric Gini 95% Confidence Interval:
C.I. 1 is derived from the delta method;
C.I. 2 is derived from a direct approach.
```

	Lower	Upper
Conf Interval 1	.50794	.52532
Conf Interval 2	.5079	.52528

```
. return list
```

```
scalars:
               r(gini) =  .5166333406145751
              r(alpha) =  2.939988458339696
               r(beta) =  .9912194875700111
matrices:
          r(estimates) :  1 x 2
          r(variances) :  1 x 2
      r(conf_interval) :  2 x 2
```

If one does not actually know the percentage of the observations below the threshold, i.e. if the data are left truncated, then the estimation relies even more heavily on the distributional assumption. We use "censorpct(*0*)" to flag this case.

```
. survlsl income, thres(10) censorpct(0) model(lognormal)

(... MLE output omitted ...)

Left Truncated Model

Estimated Parameters:
 MLE location  = 3.39364
 MLE scale     = .67306
Parametric Gini = .36587

Parametric Gini 95% Confidence Interval:
C.I. 1 is derived from the delta method;
C.I. 2 is derived from a direct approach.
```

	Lower	Upper
Conf Interval 1	.35736	.37439
Conf Interval 2	.35733	.37436

Note that the estimates in the two scenarios are quite different. This raises a question about the assumptions: (i) the censoring percentage is 30%; and (ii) the data follow a log-normal distribution. Because if both assumptions are correct, then the two estimates should be similar. Indeed, from the truncated data likelihood we can easily estimate the proportion of below-threshold observations as 5.25%.

If we are very confident only about the log-normal assumption, then we shall trust the results produced from the later estimate since it only relies on the distributional assumption. We can also conduct some sensitivity analysis using different percentages, as shown in Table 7. On the other hand, if we are very certain about the percentage of the unobserved below the threshold, then we shall doubt the log-normal assumption.

Note that in Table 7, as one would expect, if the censoring percentage is 5.25% we recover the estimated values of .366 from the truncated analysis. The confidence intervals are also very similar, with the truncated analysis producing slightly wider intervals (as expected from Table 1). If one applied the two estimation procedures to a subset of the data (i.e. to the first 500 observations only), then the confidence interval

Table 7: A robustness check on the censoring assumption

Censoring %	30%	20%	10%	5.25%
Estimated Gini	.517	.458	.396	.366
C.I. 1	(.508, .525)	(.450, .468)	(.390, .403)	(.360, .372)
C.I. 2	(.508, .525)	(.450, .466)	(.390, .403)	(.359, .372)

(Stata outputs omitted)

under truncation would be shown to be quite a bit wider than that under censoring (data not shown).

Moreover, neither of the assumptions may be valid. Using the estimated parameters, we can draw the two log-normal distributions and compare them with the observed income data as shown in Figure 4. Even from such a simple graphic comparison, the estimated distribution under truncation fits the observed data very well, and is much better than the one under censoring. Therefore, we shall trust more the results from the truncation case. Additional work on the goodness of fit aspects may be pursued, but that is beyond the scope of this paper.

Note that one possible explanation of the discrepancy may be the presence of a mixture distribution with a large fraction of (very) small incomes (e.g. due to bundling below the threshold).

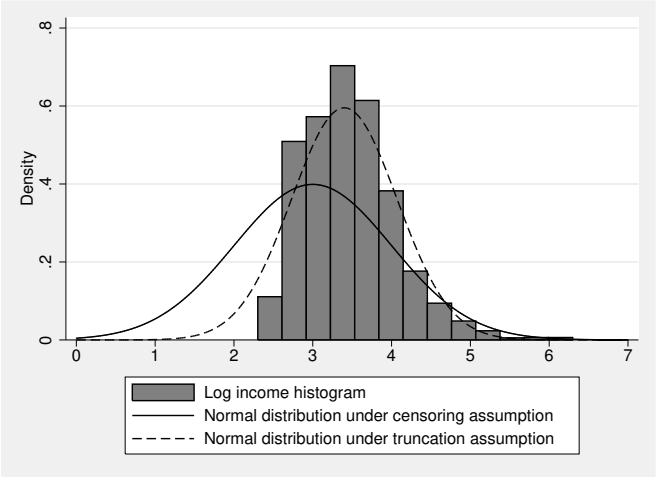

Figure 4: Distribution of the observed income data (log income)

5 Conclusions

We have developed the new Stata package ***GiniInc*** to measure the Gini index from incomplete income and survival data. First, we showed how the function **survgini** helps to compare two survival distributions with respect to their concentration. Second, we introduced two functions that work with left censored or truncated data with a fixed threshold: **survbound** calculates non-parametric Gini bounds when the data are left censored, and **survlsl** estimates the Gini index and its corresponding large-sample confidence intervals under a parametric model when the data are assumed to be known.

Plans for future developments include the addition of confidence intervals for the lower and upper bounds in ***survbound***. We would also like to extend ***survlsl*** to allow for regression models (Gigliarano et al. 2016). Lastly, we also plan to include the (more rarely encountered) case of right censoring in the parametric estimation of the Gini index.

6 References

Alfani, G., and H. García-Montero. 2018. Wealth Inequality in Preindustrial England: A Long-Term View (Thirteenth to Seventeenth Centuries). Unpublished manuscript.

Bonetti, M., C. Gigliarano, and P. Muliere. 2009. The Gini Concentration Test for Survival Data. *Lifetime Data Analysis* 453(15): 493–518.

Gigliarano, C., U. Basellini, and M. Bonetti. 2016. Longevity and Concentration in Survival Time: The log-scale-location Family of Failure Time Models. *Lifetime Data Analysis* 10(2): 254–274.

Gigliarano, C., and M. Bonetti. 2011. *Survgini - The Gini concentration test for survival data*. https://cran.r-project.org/web/packages/Survgini/index.html.

———. 2013. The Gini Test for Survival Data in Presence of Small and Unbalanced Groups. *Epidemiology, Biostatistics and Public Health* 10(2).

Gini, C. 1912. Variabilità e mutabilià. Contributo allo studio delle distribuzioni e relazioni statistiche. *Studi Economico-Giuridici dellUniversità di Cagliari III* .

———. 1914. Sulla misura della concentrazione e della Variabilità dei caratteri. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti LXXIII (part2)* 12031248.

Greene, W. 2012. *Econometric Analysis*. London, United Kingdom: Pearson.

Hanada, K. 1983. A formula of Ginis concentration ratio and its applications to life tables. *Journal of the Japan Statistical Society* 19: 293325.

Kaplan, E., and P. Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53: 457481.

Kirkwood, J., J. Ibrahim, V. Sondak, J. Richards, L. Flaherty, M. Ernstoff, T. Smith, U. Rao, M. Steele, and R. Blum. 2000. High- and Low-dose Interferon Alfa-2b in High-risk Melanoma: First Analysis of Intergroup Trial E1690/S9111/C9190. *Journal of Clinical Oncology* 18(12): 2444–58.

Michetti, B., and Dall'Aglio. 1957. La differenza semplice media. *Statistica* 7(2): 159255.

Sajaia, Z. 2007. FASTGINI: Stata module to calculate Gini coefficient with jackknife standard errors. https://ideas.repec.org/c/boc/bocode/s456814.html.

Yitzhaki, S., and E. Schechtman. 2013. *The Gini Methodology*. New York, NY: Springer.

About the authors

Long Hong completed his MSc in economics at Bocconi University in 2016, and is currently a doctoral student in economics at the University of Wisconsin - Madison. He is interested in applied econometrics and the development of statistical methods for empirical use.

Guido Alfani is an associate professor of economic history at Bocconi University, and an honorary research fellow of the University of Glasgow. His research interests are mainly in economic and social inequality, social mobility, the distribution and concentration of wealth in the early modern age, and economic trends in pre-industrial Italy.

Chiara Gigliarano is an associate professor of statistics at the University of Insubria. She is inerested in income distributions, inequality, poverty and polarization, multidimensional well-being, vulnerability, and survival analysis.

Marco Bonetti is a professor of statistics at Bocconi University, and the director of the Dondena Centre for Research on Social Dynamics and Public Policy. His research interests are in the analysis of experimental data, survival analysis, treatment effect heterogeneity, clinical trials in oncology, and methods for missing data problems.