GiniInc: A Stata Package for Measuring Inequality from Incomplete Income and Survival Data

Long Hong, Guido Alfani, Marco Bonetti
Bocconi University

Chiara Gigliarano
University of Insubria

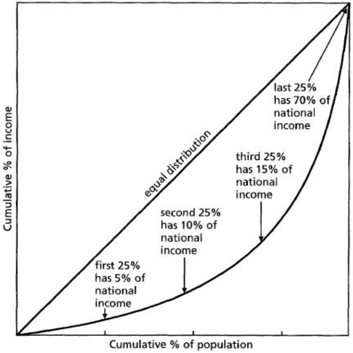
Italian Stata User Group Meeting

17 November, 2016

The Gini Index

- The Gini Index is commonly used in measuring concentration in the distribution of a positive random variable
- The Gini index G is equal to *twice the concentration area*, or the area between the 45 degree line and the Lorenz curve.

Figure 1: Lorenz curve



Introduction

- Consider random variable $X \geq 0$ with cdf F, $F(x) = F_X(x) = P(X \leq x)$
  - Survival function S, $S(x) = S_X(x) = 1 - F_X(x)$
  - finite expected value $\mu = \int_{\Re^+} (1 - F(x))\, dx$ and variance $Var(X)$
- The Gini coefficient of concentration for F (Gini 1912, 1914):

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

  - Invariant under scale changes; Bounded between 0 and 1.
- An alternative expression for G (Michetti and Dall'Aglio 1957, Hanada 1983):

$$G = 1 - \frac{\int_{\Re^+} S^2(u) du}{\int_{\Re^+} S(u) du}$$

Introduction

- The Gini Index is commonly used to
  - Measure the income or wealth inequality in Economics
  - Evaluate inequality in health and in life expectancy
- Literature has focused on complete data
  - Less attention: censored or truncated data
- **GiniInc**: Measuring Gini index using incomplete income and survival data

Outline

1. Introduction
2. Right censoring
   - Mainly survival data
   - Gini concentration tests
   - Stata Illustration
3. Left censoring + truncation
   - With fixed threshold - mainly income data
   - Parametric + non-parametric estimation
   - Stata Illustration
4. Further Developments
5. Conclusions

Part 1: Right Censoring

Survival Data

- In a clinical trial, patients may be randomized to two groups.
- Observation ends at different points for different patients (right-censoring)
- Hypothesis: censoring is independent of survival time

Questions

- Can we *calculate* Gini index for each group non-parametrically?
- Can we *compare* two survival distributions w.r.t their concentration?

Restricted Gini Index and Asymptotic Test

- For right censored data, we define Restricted Gini Index:

$$\hat{G}_t = 1 - \frac{\int_0^t \hat{S}^2(u)du}{\int_0^t \hat{S}(u)du}, \text{ where } t \text{ is the longest follow-up time in the data}$$

- Under some regularity conditions, $\hat{G}_t$ has a normal *asymptotic* distribution:

$$\sqrt{n}(\hat{G}_t - G_t) \to N(0, \tau_t), \text{ where } \tau_t \text{ is the asymptotic variance}$$

- Gini test statistics T follows χ^2 distribution with df 1 under null hypothesis

$$T := \frac{(\hat{G}_{1,t} - \hat{G}_{2,t})^2}{\hat{Var}(\hat{G}_{1,t}) - \hat{Var}(\hat{G}_{2,t})}, \text{ where } \hat{Var}(\hat{G}_t) \text{ is sample variance of } \hat{G}_t$$

(BGM, 2009)

Permutation Test

- Permutation test procedure applied to $\hat{G}_t$, especially when sample size is small
- Compute the test by M permuted samples
- Estimate the permutation distribution of $\hat{G}_t$ with the empirical cumulative distribution function

$$\hat{F}_{\hat{G}_t}(g) = \hat{P}(\hat{G}_t \leq g) = \frac{1}{M}\sum_{m=1}^{M} I(g_t^{(m)} \leq g)$$

where $g_t^{(m)}$ is the Gini statistic from permutation sample m.

(GB, 2013)

Cure Rate Models

- Patient population
  - non-cured patients $(1-\theta)$: event of interest before censoring point
  - cured patients (θ): no longer affected by disease $(X=+\infty)$
  - survival function of patient population: $S(x)=\theta+(1-\theta)S^*(x)$, where $S^*(x)$ is survival function of the non-cured.

- Other tests for difference between populations with cure rate models (but not only): a family of linear rank tests:
  - Gray-Tsiatis test
  - Log-Rank test
  - Wilcoxon test

(See BGM (2009) for results and for comparisons)

Illustration: *survgini*

Data Structure

Time	Censor	Treat
4.54483	1	1
1.28131	1	1
4.86242	0	1
2.69678	1	1
6.38193	0	2
5.61533	0	1
⋮	⋮	⋮

Variables

Time	Time-to-event variable
Censor	Censor indicator; censor=0 if right-censored, censor=1 otherwise
Treat	Treatment group; treat=1 for first group, treat=2 for second group

Illustration: *survgini*

Syntax:

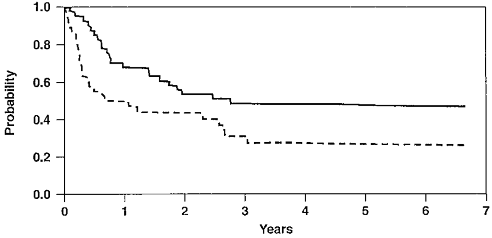
- **survgini** *time censor treat* [if] [in] [, *options*]
- *options*
  - *nolastevent*: Integrate restricted Gini statistic until the last observation
  - *nolinearrank*: Inactivate linear rank tests (log-rank test and Wilcoxon test)
  - *noasymptotic*: Inactivate asymptotic Gini test
  - *nopermutation*: Inactivate permutation Gini test
  - *m(integer)*: Number of replications of permutation sampling; default = 500.

Illustration: *survgini*

Survival Data

- Phase III melanoma clinical trial E1690 by the Eastern Cooperative Oncology Group (ECOG), available from http://merlot.stat.uconn.edu/∼mhchen/survbook/
- Patients randomized to *treatment* group with IFN high dose and *control* group (215 and 212 respectively for the two groups)

Kaplan-Meier Estimate of Relapse-free survival (RFS)



(Kirkwood et al., 2000)

Illustration: *survgini*

- Implement **survgini**:

```
.  survgini time censor treat

Comparison among GiniAs pGiniPerm Log-rank and Wilcoxon tests

             |   pGiniAs  pGiniPerm       pLR        pW
-------------+----------------------------------------
        pval |     .0526        .06     .05391     .03505


.  return list

scalars:
          r(pGiniPerm) =  .06
            r(pGiniAs) =  .0526027215785181
                 r(pW) =  .0350489502070617
                r(pLR) =  .0539137282127673
```

- Difference marginally significant at 5% confidence level

Part 2. Left Censoring + Truncation

Income Data

- Household incomes can only be obtained from tax documentations.
- However, such documentation does not exist for a certain percentage of the poor, who did not reach the income threshold of paying tax.
- The **threshold** k is usually documented, but the percentage of the poor may be estimated (*censored*) or may not be estimated (*truncated*).

Questions

- How can we estimate Gini index non-parametrically?
- If income data fit some parametric model well, can we estimate Gini index parametrically?

Non-parametric Gini bounds

- **Censoring** case only
  - π: population share; μ: income mean; G: Gini index
  - Below threshold k: π_1 known; μ_1 **unknown**; G_1 **unknown**
  - Above threshold k: π_2, G_2, μ_2 *all* known
- **Gini bounds**

$$\frac{\mu_2\pi_2^2 G_2 + \pi_1\pi_2(\mu_2 - k)}{k\pi_1 + \mu_2\pi_2} \leq G < \frac{\pi_1^2 k}{k\pi_1 + \mu_2\pi_2} + \pi_2 G_2 + \pi_1$$

  - Lower bound is reached when $\mu_1 = k$ and $G_1 = 0$
  - Upper bound cannot be reached
- Numerical method ("Grid-search"):
  - Different combinations of possible μ_1 and G_1 to search the *upper bound* numerically

Illustration: *survbound*

Example

- Historical household *income* data ($n = 5,694$) in Warwickshire, England.
- *30%* of the household's incomes are not documented
- because their incomes are below the tax-paying *threshold*, 10 shillings.

Syntax:

- **survbound** income, theshold(*real*) censorpct(*real*) [grid(*integer*)]
  - threshold: 10 (shillings)
  - censorpct: 0.3 (30%)
  - grid(n): allow grid-search by taking $(n-1)^2$ possible combinations of (μ_1, G_1)

    $$\text{Example n} = 10 \begin{cases} \mu_1 \in [0,10] \text{ available values of } \mu_1 : \{1, 2, ..., 9\} \\ G_1 \in [0,1] \text{ available values of } G_1 : \{.1, .2, ..., .9\} \end{cases}$$

Illustration: *survbound*

- Implement **survbound**:

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

Non-Parametric Gini Numeric Boundaries:

--------------------------------------------
                     |  Lower(A)    Upper(A)
---------------------+----------------------
 Non-Parametric Gini |  .4275492    .5787303
--------------------------------------------
Lower(A): Analytic lower bound
Upper(A): Analytic upper bound



. return list

scalars:
            r(lower_a) =  .4275491624079314
            r(upper_a) =  .5787303443070373
```

Illustration: *survbound*

- Allow "Grid-search" for upper bound:

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

Grid-Search Gini Numeric Boundaries:

-------------------------------------------------------
                     |  Lower(A)   Upper(A)   Upper(G)
---------------------+---------------------------------
 Non-Parametric Gini |  .4275492   .5787303   .5389827
-------------------------------------------------------
Lower(A): Analytic lower bound
Upper(A): Analytic upper bound
Upper(G): Upper bound approximation by Grid-search


. return list

scalars:
            r(lower_a) =  .4275491624079314
            r(upper_a) =  .5787303443070373
            r(upper_g) =  .5389826627857929
```

Parametric Gini index for some models

- Three commonly used **log-scale-location** models and the corresponding Gini index:

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

(GBB, 2016)

Parametric Gini index

Maximum likelihood estimation (MLE)

- Left *Censoring*, i.e. percentage below k is **known**
  - Observation: $y_i = \max(t_i, k)$; Censor Indicator: $\delta_i = I(k \leq t_i)$; where k is threshold, and t_i follows $f_\theta(t)$
  - Likelihood function: $L(\theta) \propto \prod_{i=1}^{N} \{f_\theta(y_i)\}^{\delta_i} \{F_\theta(T_i \leq k)\}^{1-\delta_i} \Rightarrow$ MLE $\hat{\theta}$
- Left *Truncation*, i.e. percentage below k is **unknown**
  - Number of obs: $N' = \sum_{i=1}^{N} \delta_i$; Observation: $y_j = \max(t_j, k)$ where k is threshold, and t_i follows $f_\theta(t | T \geq k)$
  - Likelihood function: $L(\theta) \propto \prod_{j=1}^{N'} \frac{f_\theta(y_j)}{F_\theta(T_j \geq k)} \Rightarrow$ MLE $\tilde{\theta}$
- Comparison of truncation and censoring
  - When N (and N') large, both $\hat{\theta}$ and $\tilde{\theta}$ converge to the true θ
  - But $se(\hat{\theta}) < se(\tilde{\theta})$, since $N > N'$

Illustration: *survlsl*

Example

- Historical tax threshold: 10 shillings
- Percentage of households is not sure:
  - Maybe 30%: data left *censored*
  - Maybe unknown: data left *truncated*
- **Assume**: Income follows lognormal distribution

Syntax

- **survlsl** income , theshold(*real*) censorpct(*real*) model(*string*)
  - threshold: 10
  - censorpct: 0.3 if censored; **0** if truncated
  - model: lognormal [others: weibull, loglogistic]

Illustration: *survlsl*

- Implement **survlsl** if data is censored (30%)

```
.  survlsl income, thres(10) censorpct(0.3) model(lognormal)

(... MLE iterations omitted ...)
(... MLE output omitted...)

Left Censored Model

Estimated Parameters:
 MLE location   = 2.9399885
 MLE scale      = .99121949

Parametric Gini = .51663334

. return list

scalars:
          r(gini)  =  .5166333406145753
          r(alpha) =  2.939988458339696
          r(beta)  =  .9912194875700111

matrices:
          r(estimates) :  1 x 2
          r(variances) :  2 x 2
```

Illustration: *survlsl*

- Implement **survlsl** if data is truncated

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

(... MLE iterations omitted ...)
(... MLE output omitted...)

Left Truncated Model

Estimated Parameters:
 MLE location   = 3.3936353
 MLE scale      = .67306033

Parametric Gini = .36587256

. return list

scalars:
          r(gini)  =  .3658725615414207
          r(alpha) =  3.393635300767038
          r(beta)  =  .6730603287696443

matrices:
          r(estimates) :  1 x 2
          r(variances) :  2 x 2
```

Further Developments

- **survgini**
  - Non-parametric Gini index and its corresponding confidence interval
- **survbound**
  - Confident intervals of the non-parametric Gini bounds
- **survlsl**
  - Goodness of fit: is the assumption of Log-normal distribution valid?
  - Confidence interval for parametric Gini index
  - Regression: Gini depends on covariates such as institution and culture (see, e.g. GBB (2016) for regression with one covariate)

Conclusions

- **GiniInc**: Measuring Gini index using incomplete income and survival data
- Right censoring using survival data example (**survgini**) to
  - Compare two survival distributions w.r.t their concentration
- Left censoring + truncation with fixed threshold using income data example
  - Calculate non-parametric Gini bounds if data is censored (**survbound**)
  - Compute parametric Gini index if data is truncated or censored (**survlsl**)

More are coming in the next release!

Selected References

Bonetti, M., Gigliarano, C., and Muliere, P. (2009).
The gini concentration test for survival data.
Lifetime Data Analysis, 453(15).

Gigliarano, C., Basellini, U., and Bonetti, M. (2016).
Longevity and concentration in survival times: the log-scale-location family of failure time models.
Lifetime Data Analysis, 10(2):1–21.

Gigliarano, C. and Bonetti, M. (2013).
The gini test for survival data in presence of small and unbalanced groups.
EBPH Epidemiology, Biostatistics and Public Health, 10(2).