

Poverty and Inequality with Complex Survey Data

Guilherme Jacob, Anthony Damico, and Djalma Pessoa

2016-12-09

Contents

1	Introduction	5
1.1	Installation	5
1.2	Complex surveys and statistical inference	6
1.3	Linearization	6
1.4	Influence function	7
1.5	The variance estimator	8
1.6	Influence functions - Examples	8
1.7	Linearization by influence function - Examples	8
1.8	Structure of the library	9
1.9	FGT indicator	11
2	Poverty Indices	15
2.1	At Risk of Poverty Ratio and Threshold (svyarpr, svyarpt)	15
2.2	The Gender Pay Gap (svygpg)	15
2.3	Quintile Share Ratio (svyqsr)	15
2.4	Relative Median Income Ratio (svyrmir)	15
2.5	Relative Median Poverty Gap (svyrmpg)	15
2.6	Median Income Below the At Risk of Poverty Threshold (svypoormed)	15
2.7	Foster-Greer-Thorbecke class (svyfgt)	15
3	Inequality Measurement	17
3.1	Theoretical aspects of inequality	17
3.2	Lorenz Curve (svylorenz)	17
3.3	Measures derived from the Lorenz Curve	17
3.4	Entropy-based Measures	17
4	Multidimensional Indices	19
4.1	Alkire-Foster Class and Decomposition (svyafc, svyafcdec)	19
4.2	Bourguignon (1999) inequality class (svybmi)	19

Chapter 1

Introduction

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$.

For now, you have to install the development version of **bookdown** from Github:

```
devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading #.

To compile this example to PDF, you need to install XeLaTeX.

The library convey aims at estimating measures of poverty and income concentration. There are already at least two libraries covering this subject: vardpoor and Laeken. The main difference between the library convey and these two is that the convey strongly hinges on the survey library.

1.1 Installation

- the latest released version from CRAN with

```
install.packages("convey")
```

- the latest development version from github with

```
devtools::install_github("djalmapessoa/convey")
```

[This may present how to install R, RStudio and required packages. Providing brief information about survey and MonetDBLite may also be recommended.]

You can label chapter and section titles using {#label} after them, e.g., we can reference Chapter 1.1. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter 3.

Figures and tables with captions will be placed in **figure** and **table** environments, respectively.

```
par(mar = c(4, 4, .1, .1))  
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the **fig:** prefix, e.g., see Figure 1.1. Similarly, you can reference tables generated from **knitr::kable()**, e.g., see Table 1.1.

```
knitr::kable(  
  head(iris, 20), caption = 'Here is a nice table!',
```

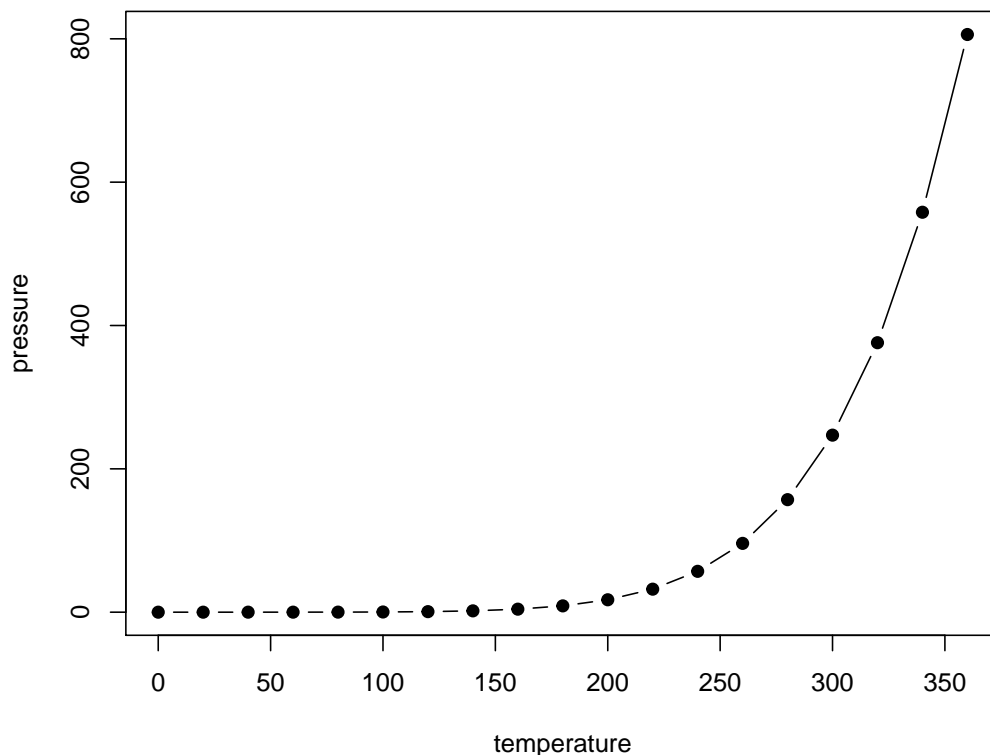


Figure 1.1: Here is a nice figure!

```
booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2016) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

1.2 Complex surveys and statistical inference

[I think we should have a discussion about what is complex survey, its importance and so on. We can use a book Djalma wrote.]

1.3 Linearization

Some measures of poverty and income concentration are defined by non-differentiable functions so that it is not possible to use Taylor linearization to estimate their variances. An alternative is to use **Influence functions** as described in (Deville, 1999) and (Osier, 2009). The library **convey** implements this methodology to work with **survey.design** objects and also with **svyrep.design** objects.

Some examples of these measures are:

- At-risk-of-poverty threshold: $arpt = .60q_{.50}$ where $q_{.50}$ is the income median;
- At-risk-of-poverty rate $arpr = \frac{\sum_U 1(y_i \leq arpt)}{N} \cdot 100$
- Quintile share ratio

Table 1.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

$$qsr = \frac{\sum_U 1(y_i > q_{.80})}{\sum_U 1(y_i \leq q_{.20})}$$

- Gini coefficient $1 + G = \frac{2 \sum_U (r_i - 1) y_i}{N \sum_U y_i}$ where r_i is the rank of y_i .

Note that it is not possible to use Taylor linearization for these measures because they depend on quantiles and the Gini is defined as a function of ranks. This could be done using the approach proposed by Deville (1999) based upon influence functions.

1.4 Influence function

Let U be a population of size N and M be a measure that allocates mass one to the set composed by one unit, that is $M(i) = M_i = 1$ if $i \in U$ and $M(i) = 0$ if $i \notin U$

Now, a population parameter θ can be expressed as a functional of M $\theta = T(M)$

Examples of such parameters are:

- Total: $Y = \sum_U y_i = \sum_U y_i M_i = \int y dM = T(M)$
- Ratio of two totals: $R = \frac{Y}{X} = \frac{\int y dM}{\int x dM} = T(M)$
- Cumulative distribution function: $F(x) = \frac{\sum_U 1(y_i \leq x)}{N} = \frac{\int 1(y \leq x) dM}{\int dM} = T(M)$

To estimate these parameters from the sample, we replace the measure M by the estimated measure \hat{M} defined by: $\hat{M}(i) = \hat{M}_i = w_i$ if $i \in s$ and $\hat{M}(i) = 0$ if $i \notin s$.

The estimators of the population parameters can then be expressed as functional of the measure \hat{M} .

- Total: $\hat{Y} = T(\hat{M}) = \int y d\hat{M} = \sum_s w_i y_i$
- Ratio of totals: $\hat{R} = T(\hat{M}) = \frac{\int y d\hat{M}}{\int x d\hat{M}} = \frac{\sum_s w_i y_i}{\sum_s w_i x_i}$
- Cumulative distribution function: $\hat{F}(x) = T(\hat{M}) = \frac{\int 1(y \leq x) d\hat{M}}{\int d\hat{M}} = \frac{\sum_s w_i 1(y_i \leq x)}{\sum_s w_i}$

1.5 The variance estimator

The variance of the estimator $T(\hat{M})$ can be approximated by:

$$Var \left[T(\hat{M}) \right] \cong var \left[\sum_s w_i z_i \right]$$

The **linearized** variable z is given by the derivative of the functional:

$$z_k = \lim_{t \rightarrow 0} \frac{T(M + t\delta_k) - T(M)}{t} = IT_k(M)$$

where, δ_k is the Dirac measure in k : $\delta_k(i) = 1$ if and only if $i = k$.

This **derivative** is called **Influence Function** and was introduced in the area of **Robust Statistics**.

1.6 Influence functions - Examples

- Total:

$$\begin{aligned} IT_k(M) &= \lim_{t \rightarrow 0} \frac{T(M + t\delta_k) - T(M)}{t} \\ &= \lim_{t \rightarrow 0} \frac{\int y d(M + t\delta_k) - \int y dM}{t} \\ &= \lim_{t \rightarrow 0} \frac{\int y d(t\delta_k)}{t} = y_k \end{aligned}$$

- Ratio of two totals:

$$\begin{aligned} IR_k(M) &= I \left(\frac{U}{V} \right)_k (M) = \frac{V(M) \times IU_k(M) - U(M) \times IV_k(M)}{V(M)^2} \\ &= \frac{Xy_k - Yx_k}{X^2} = \frac{1}{X} (y_k - Rx_k) \end{aligned}$$

1.7 Linearization by influence function - Examples

- At-risk-of-poverty threshold:

$$arpt = 0.6 \times m$$

where m is the median income.

$$z_k = -\frac{0.6}{f(m)} \times \frac{1}{N} \times [I(y_k \leq m - 0.5)]$$

- At-risk-of-poverty rate:

$$arpr = \frac{\sum_U I(y_i \leq t)}{\sum_U w_i} .100$$

$$z_k = \frac{1}{N} [I(y_k \leq t) - t] - \frac{0.6}{N} \times \frac{f(t)}{f(m)} [I(y_k \leq m) - 0.5]$$

where:

N - population size;

t - at-risk-of-poverty threshold;

y_k - income of person k ;

m - median income;

f - income density function;

1.8 Structure of the library

In the library convey, there are some basic functions that produces the linearized variables of some estimates that often enter in the definition of measures of concentration and poverty. For example the `quantile` which is linearized by the function `svyiqalpha`. Other example is the function `svyisq` that linearizes the total below a quantile of the variable.

From the linearized variables of these basic estimates it is possible by using rules of composition, valid for influence functions, to derive the influence function of more complex estimates. By definition the influence function is a Gateaux derivative and the rules rules of composition valid for Gateaux derivatives also hold for Influence Functions.

The following property of Gateaux derivatives was often used in the library convey. Let g be a differentiable function of m variables. Suppose we want to compute the influence function of the estimator $g(T_1, T_2, \dots, T_m)$, knowing the Influence function of the estimators $T_i, i = 1, \dots, m$. Then the following holds:

$$I(g(T_1, T_2, \dots, T_m)) = \sum_{i=1}^m \frac{\partial g}{\partial T_i} I(T_i)$$

In the library convey this rule is implemented by the function `contrastinf` which uses the R function `deriv` to compute the formal partial derivatives $\frac{\partial g}{\partial T_i}$.

For example, suppose we want to linearize the `Relative median poverty gap`(`rmpg`), defined as the difference between the at-risk-of-poverty threshold (`arpt`) and the median of incomes less than the `arpt` relative to the `arprt`:

$$rmpg = \frac{arpt - medpoor}{arpt}$$

where `medpoor` is the median of incomes less than `arpt`.

Suppose we know how to linearize `arpt` and `medpoor`, then by applying the function `contrastinf` with

$$g(T_1, T_2) = \frac{(T_1 - T_2)}{T_1}$$

we linearize the `rmpg`.

1.8.1 Examples of use of the library convey

In the following examples we will use the data set `eusilc` contained in the libraries `vardpoor` and `Laeken`.

```
library(vardpoor)
data(eusilc)
```

Next, we create an object of class `survey.design` using the function `svydesign` of the library `survey`:

```
library(survey)
des_eusilc <- svydesign(ids = ~rb030, strata = ~db040, weights = ~rb050, data = eusilc)
```

Right after the creation of the design object `des_eusilc`, we should use the function `convey_prep` that adds an attribute to the survey design which saves information on the design object based upon the whole sample, needed to work with subset designs.

```
library(convey)
des_eusilc <- convey_prep( des_eusilc )
```

To estimate the at-risk-of-poverty rate we use the function `svyarprt`:

```
svyarprt(~eqIncome, design=des_eusilc)
```

```
      arpr      SE
eqIncome 0.14444 0.0028
```

To estimate the at-risk-of-poverty rate for domains defined by the variable `db040` we use

```
svyby(~eqIncome, by = ~db040, design = des_eusilc, FUN = svyarprt, deff = FALSE)
```

```
      db040 eqIncome      se
Burgenland   Burgenland 0.1953984 0.017202243
Carinthia     Carinthia 0.1308627 0.010610622
Lower Austria Lower Austria 0.1384362 0.006517660
Salzburg      Salzburg 0.1378734 0.011579280
Styria        Styria 0.1437464 0.007452360
Tyrol         Tyrol 0.1530819 0.009880430
Upper Austria Upper Austria 0.1088977 0.005928336
Vienna        Vienna 0.1723468 0.007682826
Vorarlberg    Vorarlberg 0.1653731 0.013754670
```

Using the same data set, we estimate the quintile share ratio:

```
# for the whole population
svyqsr(~eqIncome, design=des_eusilc, alpha= .20)
```

```
      qsr      SE
eqIncome 3.97 0.0426
```

```
# for domains
svyby(~eqIncome, by = ~db040, design = des_eusilc,
      FUN = svyqsr, alpha= .20, deff = FALSE)
```

```
      db040 eqIncome      se
Burgenland   Burgenland 5.008486 0.32755685
Carinthia     Carinthia 3.562404 0.10909726
Lower Austria Lower Austria 3.824539 0.08783599
Salzburg      Salzburg 3.768393 0.17015086
Styria        Styria 3.464305 0.09364800
Tyrol         Tyrol 3.586046 0.13629739
```

```
Upper Austria Upper Austria 3.668289 0.09310624
Vienna Vienna 4.654743 0.13135731
Vorarlberg Vorarlberg 4.366511 0.20532075
```

These functions can be used as S3 methods for the classes `survey.design` and `svyrep.design`.

Let's create a design object of class `svyrep.design` and run the function `convey_prep` on it:

```
des_eusilc_rep <- as.svrepdesign(des_eusilc, type = "bootstrap")
des_eusilc_rep <- convey_prep(des_eusilc_rep)
```

and then use the function `svyarpr`:

```
svyarpr(~eqIncome, design=des_eusilc_rep)
```

```
      arpr      SE
eqIncome 0.14444 0.0027
```

```
svyby(~eqIncome, by = ~db040, design = des_eusilc_rep, FUN = svyarpr, deff = FALSE)
```

```
      db040 eqIncome se.eqIncome
Burgenland Burgenland 0.1953984 0.016861380
Carinthia Carinthia 0.1308627 0.013276697
Lower Austria Lower Austria 0.1384362 0.007348499
Salzburg Salzburg 0.1378734 0.009979870
Styria Styria 0.1437464 0.006918992
Tyrol Tyrol 0.1530819 0.010436504
Upper Austria Upper Austria 0.1088977 0.006480643
Vienna Vienna 0.1723468 0.006690668
Vorarlberg Vorarlberg 0.1653731 0.014722833
```

The functions of the library `convey` are called in a similar way to the functions in library `survey`.

It is also possible to deal with missing values by using the argument `na.rm`.

```
# survey.design using a variable with missings
svygini( ~ py010n , design = des_eusilc )
```

```
      gini SE
py010n NA NA
```

```
svygini( ~ py010n , design = des_eusilc , na.rm = TRUE )
```

```
      gini      SE
py010n 0.64606 0.0036
```

```
# svyrep.design using a variable with missings
# svygini( ~ py010n , design = des_eusilc_rep ) get error
svygini( ~ py010n , design = des_eusilc_rep , na.rm = TRUE )
```

```
      gini      SE
py010n 0.64606 0.0035
```

1.9 FGT indicator

Foster and all(1984) proposed a family of indicators to measure poverty.

The class of *FGT* measures, can be defined as

$$p = \frac{1}{N} \sum_{k \in U} h(y_k, \theta),$$

where

$$h(y_k, \theta) = \left[\frac{(\theta - y_k)}{\theta} \right]^\gamma \delta \{y_k \leq \theta\},$$

where: θ is the poverty threshold; δ the indicator function that assigns value 1 if the condition $\{y_k \leq \theta\}$ is satisfied and 0 otherwise, and γ is a non-negative constant.

When $\gamma = 0$, p can be interpreted as the ratio of poor people, and for $\gamma \geq 1$, the weight of poor people increases with the value γ , (Foster and all, 1984).

The poverty measure FGT is implemented in the library convey by the function `svyfgt`. The argument `thresh_type` of this function defines the type of poverty threshold adopted. There are three possible choices:

1. `abs` – fixed and given by the argument `thresh_value`
2. `relq` – a proportion of a quantile fixed by the argument `proportion` and the quantile is defined by the argument `order`.
3. `reln` – a proportion of the mean fixed the argument `proportion`

The quantile and the mean involved in the definition of the threshold are estimated for the whole population. When $\gamma = 0$ and $\theta = .6 * MED$ the measure is equal to the indicator `arpr` computed by the function `svyarpr`.

Next, we give some examples of the function `svyfgt` to estimate the values of the FGT poverty index.

Consider first the poverty threshold fixed ($\gamma = 0$) in the value 10000. The headcount ratio (FGT0) is

```
svyfgt(~eqIncome, des_eusilc, g=0, abs_thresh=10000)
```

```
      fgt0      SE
eqIncome 0.11444 0.0027
```

The poverty gap (FGT1) ($\gamma = 1$) index for the poverty threshold fixed at the same value is

```
svyfgt(~eqIncome, des_eusilc, g=1, abs_thresh=10000)
```

```
      fgt1      SE
eqIncome 0.032085 0.0011
```

To estimate the FGT0 with the poverty threshold fixed at $0.6 * MED$ we fix the argument `type_thresh="relq"` and use the default values for `percent` and `order`:

```
svyfgt(~eqIncome, des_eusilc, g=0, type_thresh= "relq")
```

```
      fgt0      SE
eqIncome 0.14444 0.0028
```

that matches the estimate obtained by

```
svyarpr(~eqIncome, design=des_eusilc, .5, .6)
```

```
      arpr      SE
eqIncome 0.14444 0.0028
```

To estimate the poverty gap (FGT1) with the poverty threshold equal to $0.6 * MEAN$ we use:

```
svyfgt(~eqIncome, des_eusilc, g=1, type_thresh= "reln")
```

	fgt1	SE
eqIncome	0.051187	0.0011

Chapter 2

Poverty Indices

[I think this is a good start. I don't think that gender pay gap, quantiles and totals are measures of poverty. Consider another chapter on other wellbeing measures.]

2.1 At Risk of Poverty Ratio and Threshold (svyarpr, svyarpt)

2.2 The Gender Pay Gap (svygpg)

2.3 Quintile Share Ratio (svyqsr)

2.4 Relative Median Income Ratio (svyrmir)

2.5 Relative Median Poverty Gap (svyrmpg)

2.6 Median Income Below the At Risk of Poverty Threshold (svy-poormed)

2.7 Foster-Greer-Thorbecke class (svyfgt)

Chapter 3

Inequality Measurement

[Present an introduction to what is inequality].

3.1 Theoretical aspects of inequality

3.1.1 Desirable properties of inequality measures

3.2 Lorenz Curve (svylorenz)

3.3 Measures derived from the Lorenz Curve

3.3.1 Gini index (svygini)

3.3.2 Amato index (svyamato)

3.3.3 Zenga Index and Curve (svyzenga, svyzengacurve)

3.4 Entropy-based Measures

3.4.1 Atkinson index (svyatk)

3.4.2 Generalized Entropy and Decomposition (svygei, svygeidec)

3.4.3 J-Divergence Entropy and Decomposition (svyjdiv, svyjdivdec)

3.4.4 Rényi Divergence (svyrenyi)

Chapter 4

Multidimensional Indices

We have finished a nice book.

4.1 Alkire-Foster Class and Decomposition (svyafc, svyafcdec)

4.2 Bourguignon (1999) inequality class (svybmi)

Bibliography

- Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25(2):193–203.
- Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality. *Journal of the European Survey Research Association*, 3(3):167–195.
- Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.
- Xie, Y. (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.3.2.