Rios-Avila, Fernando

**Working Paper**

# Recentered influence functions in Stata: Methods for analyzing the determinants of poverty and inequality

Working Paper, No. 927

**Provided in Cooperation with:**
Levy Economics Institute of Bard College

*Suggested Citation:* Rios-Avila, Fernando (2019) : Recentered influence functions in Stata: Methods for analyzing the determinants of poverty and inequality, Working Paper, No. 927, Levy Economics Institute of Bard College, Annandale-on-Hudson, NY

This Version is available at:
https://hdl.handle.net/10419/209170


Working Paper No. 927

# Recentered Influence Functions in Stata: Methods for Analyzing the Determinants of Poverty and Inequality

by

**Fernando Rios-Avila**
Levy Economics Institute of Bard College

**April 2019**

The Levy Economics Institute Working Paper Collection presents research in progress by Levy Institute scholars and conference participants. The purpose of the series is to disseminate ideas to and elicit comments from academics and professionals.

Levy Economics Institute of Bard College, founded in 1986, is a nonprofit, nonpartisan, independently funded research organization devoted to public service. Through scholarship and economic research it generates viable, effective public policy responses to important economic problems that profoundly affect the quality of life in the United States and abroad.

Levy Economics Institute
P.O. Box 5000
Annandale-on-Hudson, NY 12504-5000
http://www.levyinstitute.org



ISSN 1547-366X

**ABSTRACT**

Recentered influence functions (RIFs) are statistical tools popularized by Firpo, Fortin, and Lemieux (2009) for analyzing unconditional partial effects on quantiles in a regression analysis framework (unconditional quantile regressions). The flexibility and simplicity of these tools has opened the possibility of extending the analysis to other distributional statistics using linear regressions or decomposition approaches. In this paper, I introduce three Stata commands to facilitate the use of RIFs in the analysis of outcome distributions: `rifvar()` is an `egen` extension used to create RIFs for a large set of distributional statistics; `rifhdreg` facilitates the estimation of RIF regressions, enabling the use of high-dimensional fixed effects; and `oaxaca_rif` to implement Oaxaca-Blinder type decomposition analysis (RIF decompositions).

**KEYWORDS:** Recentered Influence Functions; Unconditional Partial Effects; Unconditional Quantile Regression; RIF Regressions; Distributional Statistics; Oaxaca-Blinder; RIF Decomposition

**JEL CLASSIFICATIONS:** C13; C18; I14; I30; J31

## 1. INTRODUCTION

Influence functions (IF) are statistical tools that have been used for analyzing the robustness of distributional statistics, or functionals, to small disturbances in data (Cowell and Flachaire 2007) or for a simplified strategy to estimate asymptotic variances of complex statistics (Cowell and Flachaire 2015; Deville 1999). More recently, Firpo, Fortin, and Lemieux (2009) suggested the use of IFs, specifically recentered influence functions (RIFs), as a tool for analyzing the impact that changes in the distribution of explanatory variables X have on the unconditional distribution of Y.

The method introduced by Firpo, Fortin, and Lemieux (2009) focused on the estimation of unconditional quantile regression (UQR), which allows the researcher to obtain partial effects of explanatory variables on any unconditional quantile of the dependent variable. The simplest version of this methodology, referred to as recentered influence functions–ordinary least squares (RIF-OLS), is easily implemented making use of the user-written command `rifreg` by the same authors.[1] As part of their conclusions, the authors highlight the potential extensions of this strategy for analyzing other distributional statistics, as well as the potential usefulness for generalizing the traditional Oaxaca-Blinder (OB) decomposition for analyzing differences of outcome distributions across groups.[2]

After its introduction, UQR became a popular method for analyzing and identifying the distributional effects on outcomes in terms of changes in observed characteristics in areas such as labor economics, income and inequality, health economics, and public policy. The potential simplicity and flexibility the methodology offers for the analysis of any distributional statistics also motivated subsequent research to expand the use of RIFs in the framework of regression analysis.

[1] The user-written program `rifreg` implements the estimation of what the authors call RIF-OLS for UQRs.
[2] The use of RIF regressions within the OB decomposition approach has been discussed in the review of decomposition methods in Fortin, Lemieux, and Firpo (2011) and, more recently, in Firpo, Fortin, and Lemieux (2018).

In a recently published paper, Firpo, Fortin, and Lemieux (2018) discuss the application of RIF regressions for the variance and Gini coefficients, with emphasis on the generalization of the OB decomposition.[3] Borgen (2016), building on the work of Firpo, Fortin, and Lemieux (2009), provides the command `xtrifreg` for the efficient estimation of UQRs in the presence of a single high-dimensional fixed effect, but limited to quantile regressions.

Cowell and Flachaire (2007), for their analysis on the sensitivity of inequality measures to the presence of extreme values, provide IFs for the most commonly used inequality indices including the Atkinson index, the generalized entropy index, and the logarithmic variance index. Essama-Nssah and Lambert (2012), who discuss the use of RIF regressions and OB decompositions for the analysis of distributional changes, provide a large set of IFs and RIFs for distributional statistics relevant for policy analysis, including Lorenz and generalized Lorenz ordinates, Foster-Greer-Thorbecke (FGT) poverty indices, and Watts and Sen poverty indices. Most recently, Heckley, Gerdtham, and Kjellsson (2016) examine the use of RIFs for measures of health inequality, with emphasis on bivariate rank-dependent concentration indices. While there is no available command associated to this implementation, the authors provide a Stata do-file used for their analysis.

While RIF regressions and RIF decompositions have become important tools of analysis in the empirical literature, to the best of my knowledge, there are only limited attempts to provide a simplified framework to allow the use of RIFs as a standard analytical tool. Within the statistical software Stata, only the user-written command `xtrifreg` is readily available from the ssc archives. The command that started it all, `rifreg`, is limited in the estimation of RIF statistics and is not available in the ssc archives, although it can be accessed manually from the author's website.[4] Furthermore, while RIFs and reweighted regressions are broadly used for the generalization of the OB decomposition for statistics beyond the mean, no commands exist to facilitate such analysis.

[3] While recently published, the working paper version of the paper dates from 2007 and was cited in their 2009 paper as part of the extensions on the use of RIFs in regression analysis. The program `rifreg` estimates all the RIFs proposed in the paper.
[4] The command can be accessed at: https://faculty.arts.ubc.ca/nfortin/rifreg.zip

This paper introduces three Stata commands that aim to facilitate the use of RIF regressions and RIF decompositions. The first command, `rifvar()`, is a *byable* plug-in extension that works with `egen` and can be used to estimate RIFs for a large set of distributional statistics, such as those described in Firpo, Fortin, and Lemieux (2018), Essama-Nssah and Lambert (2012), Cawley, Cowell, and Flachaire (2007), and Heckley, Gerdtham, and Kjellsson (2016). The second, `rifhdfe,` is a wrapper program for `regress` (StataCorp) and `reghdfe` (Correira 2017), that in combination with `rifvar()` is used to estimate RIF regressions in the presence of high-dimensional fixed effects. Finally, the third command, `oaxaca_rif`, is a wrapper around `oaxaca` (Jann 2008) that can be used to implement standard and reweighted OB decompositions (see Fortin, Lemieux, and Firpo [2011] and Firpo, Fortin, and Lemieux [2018]).

The rest of this paper is structured as follows. Section 2 provides an overview for understanding what IFs are and how they are estimated. Section 3 introduces and explains the use of `rifvar` to estimate RIF variables. Section 4 describes the use of `rifhdreg` for the estimation of RIF regressions. Section 5 describes the use of `oaxaca_rif` for the estimation of standard and reweighted decompositions using RIF decompositions. Section 6 provides an illustration of the commands, and section 7 concludes.

## 2. RECENTERED INFLUENCE FUNCTIONS AND DISTRIBUTIONAL STATISTICS

### 2.1. Distributional Statistics: Basics

When analyzing social welfare, inequality, poverty, or any other measure that describes the distributional characteristics of an outcome of interest, it is necessary to have access to one of the following pieces of information. The most common scenario is to have access to the full set of values corresponding to each observation in the population or sample. If the sample/population is finite of size *n*, it can be referred to as $Y = [y_1, y_2, \dots, y_n]$, where each $y_i$ is the outcome of interest (i.e., income) of the *ith* person.

The second scenario is one where one may or may not know the income level of each individual in the sample, but one knows the relative position of all individuals compared to the rest of the population (cumulative distribution function or cdf) or how frequent or common it is to observe an individual with any given level of income (probability density function or pdf). Using the function $F_Y()$ to refer to the cdf and $f_Y()$ to the pdf, the vector of information required for analyzing distributions can be more briefly written as a set of ordered pairs, $F_Y = [(y, F_Y(y))|y \in \mathbb{R}]$ or $f_Y = [(y, f_Y(y))|y \in \mathbb{R}]$, where y represents any real number (generally positive when referring to income).[5] This simply means that if one has access to any of these vectors of information $(Y, F_Y, f_Y)$, any distributional statistic can be derived.

Let us call the $v(.)$ a functional that uses all the information contained in $Y$, $F_Y$, or $f_Y$ to estimate a distributional statistic of Y. This functional can be used to estimate statistics relevant to policy analysis like the mean, qth quantile, poverty indices, or inequality indices. To measure the impact a change in the distribution of income will have on the distributional statistic, one can simply compare the indices swapping the cdf from the observed distribution $F_Y$ to the ex post distribution $G_Y$.[6] This can be written as follows:

$$\Delta v = v(G_Y) - v(F_Y) \tag{1}$$

Thus, $\Delta v$ is the change in the distributional statistic generated by a change in distribution from $F_y \rightarrow G_y$. This change can be as large as implying that everyone in the population receives a fixed transfer (shifting the function $F_y()$ to the right),[7] or as simple as having a new person (with random income) added to the sample, changing the rankings of everyone in the sample. The first scenario is a simplified example of what DiNardo, Fortin, and Lemieux (1996) used for analyzing changes in the distribution of wages. The second scenario is an exercise that can be used for understanding the definition of IFs and RIFs.

[5] It is useful to remember that the functions $F_Y()$ and $f_Y()$ are not arbitrary functions and obey a strict set of properties and relations among them to represent well-defined distribution functions: $f_Y(y) \geq 0 \; \forall \; h \in \mathbb{R}$, $\int_{-\infty}^{y} f_Y(x)dx = F_Y(y)$, $dF_Y(y) = f_Y(y)$, $F_Y(-\infty) = 0$, $F_Y(\infty) = 1$, and $F_Y(y_1) \leq F_Y(y_2) \Leftrightarrow y_1 \leq y_2$.

[6] $G_Y()$ and its counterpart $g_Y()$ have the same properties as $F_Y()$ and $f_Y()$.

[7] This thought experiment can also be thought of as if suddenly people with lower income disappear while people with higher income become more numerous.

### 2.2. Estimations of IFs and RIFs: Gâteaux Derivative

The thought experiment of adding of a new person to a sample can also be considered as a case of data contamination in the original distribution, and equation (1) can be used to estimate the influence of this thought experiment on the statistic $v$. The problem with this example is the magnitude of the change $\Delta v$ will depend on the magnitude of the change from $F_Y() \rightarrow G_Y()$. For a population of size N, changes to the distribution caused by one additional person will be larger compared to the same experiment with a population of size 2N. A solution is to standardize the change in the statistic $\Delta v$ with respect to some measure that quantifies the change of the distribution $\left(\Delta(G_Y - F_Y)\right)$:

$$\Delta^s v = \frac{\Delta v}{\Delta(G_Y - F_Y)} = \frac{v(G_Y) - v(F_Y)}{\Delta(G_Y - F_Y)} \tag{2}$$

This process can be extended to measure $\Delta^s v$ for an infinitesimally small change in the distribution function from $F_Y \rightarrow G_Y$. This idea is what lies behind the Gâteaux derivative, a generalization of the directional derivative for functional analysis. The derivative is used to construct IFs, which can be used as measures of robustness of functionals to data outliers (Hampel 1974) and facilitate the visualization of the structure of the distributional statistic as a function of the available data. Before proceeding to the formal definition of the IF, it is useful to revisit and formalize the thought experiment just described.

Assume the observation to be introduced in the sample has an income equal to $y_c$. Since this is the only element of that distribution, its cdf can be characterized as follows:

$$H_{Y_c}(y) = 0 \; \forall \; y < y_c \; \& \; H_{Y_c}(y) = 1 \; \forall \; y \geq y_c \tag{3}$$

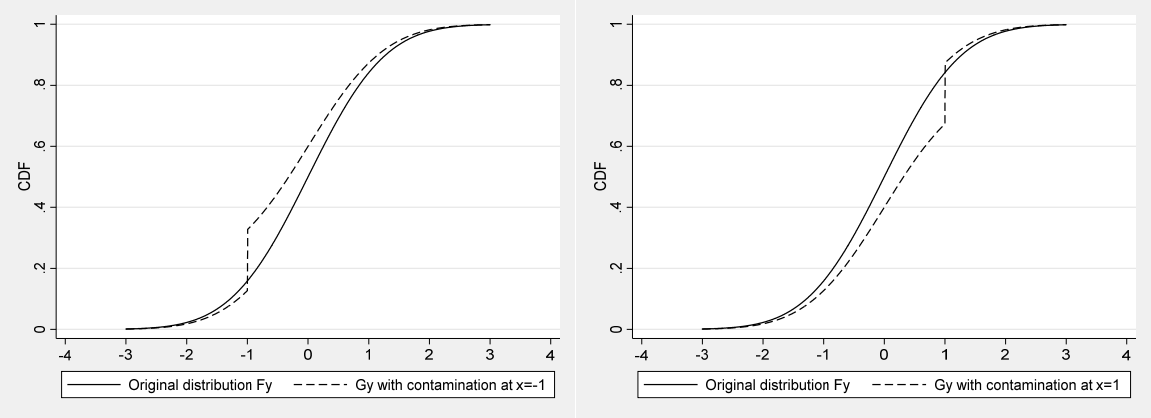
This indicates that the distribution $H_{Y_c}$ only puts mass at the value $y_c$.[8] With this definition, we can construct the distribution that would be observed in $G_Y$, combining the observed distribution from $F_Y$ and $H_{Y_c}$:

[8] Conversely, this means that $dH_{y_c}(y) = 0 \; \forall y \neq y_c$ and $dH_{y_c}(h) = \infty \; if \; y = y_c$ .

$$G_Y = (1 - \varepsilon)F_Y + \varepsilon H_{Y_c} \quad (4)$$

In other words, $G_Y$ is the resulting distribution when the original distribution $F_Y$ is transformed in the direction of $H_{Y_c}$. This expression also quantifies the change in the distribution when moving from $F_Y \rightarrow G_Y$ as $\varepsilon$. Figure 1 provides a graphical example of the changes observed as a result of this change in the distribution (contamination) of the distribution function.

**Figure 1. Comparison between Original and Contaminated Distributions**




With this last concept in place, we can finally provide the formal definition of the IF:

$$IF\left(y_c; v(F_Y)\right) = \lim_{\varepsilon \to 0} \frac{v\left((1-\varepsilon)F_Y + \varepsilon H_{Y_c}\right) - v(F_Y)}{\varepsilon} = \frac{\partial v(F_Y \rightarrow H_{Y_c})}{\partial \varepsilon} \quad (4)$$

The IF is a directional derivative that shows how the distributional statistic $v$ would change when there is a small change in the distribution $F_y$ that gives more weight to observations with values $y_c$.

It is important to notice that the IF will be different for each point of reference $y_c$ (contamination point) used for its estimation. Also, as discussed in Firpo, Fortin, and Lemieux (2009), Cowell and Flachaire (2015), and Essama-Nssah and Lambert (2012), the IF has the following properties:

$$\int IF\big(y; v(F_Y)\big)\, dF_Y = 0 \tag{5}$$

$$v\big(F_y\big) \sim N\left(v(F_Y), \frac{\sigma_{IF}^2}{N}\right) \tag{6}$$

$$\sigma_{IF}^2 = \int IF\big(y; v(F_Y)\big)^2\, dF_Y \tag{7}$$

Instead of using the IF directly, Firpo, Fortin, and Lemieux (2009) propose the use of the recentered version of the statistics, referred to as the RIF, which is equivalent to the first two terms of the von Mises (1947) linear approximation of the corresponding distributional statistic $v$:

$$RIF\big(y_i; v(F_Y)\big) = v(F_Y) + IF\big(y_i; v(F_Y)\big) \tag{8}$$

This expression maintains the same properties of IFs and can be used directly for the estimation of standard errors of any statistic for which an RIF exists. While this change has no impact for the estimation of RIF regressions—other than changes in the intercept in the case of RIF-OLS—using the RIF is crucial for the implementation of RIF decompositions.

Many authors have examined the properties of distributional statistics and have derived the corresponding RIFs for a myriad of statistics. Appendix 1 provides a list of all distributional statistics, the corresponding RIFs, and the sources where they have been obtained and are currently available in the program that is described in next section.

## 3. ESTIMATING RIFS: `egen newvar=rifvar(oldvar), [options]`

The estimation of RIFs is a task of variable complexity. Some statistics have a simple mathematical expression that requires few lines of code to define the corresponding RIF. The easiest example is the RIF for the mean, since the RIF mean for any value $y_i$ is simply itself. Other statistics, however, may require many intermediate steps to correctly define their corresponding RIF.

The user-written command `rifreg` can be used to estimate RIF regressions and create the corresponding RIF, but it is limited to the analysis of the variance, quantile, and Gini coefficient. If one is interested in the analysis of other distributional statistics, I suggest the use of a new command called `rifvar()`.[9] This is a plug-in program that adds new functions to the command `egen`, facilitating the estimation of RIFs for a large set of distributional statistics.[10]

The syntax of the command is as follows:

```
egen [type] newvar = rifvar(varname) [if exp] [in range] [,
by(varname) weight(varname) rifoptions ]
```

where `varname` is the variable being analyzed and `newvar` is the new variable name where the RIF will be stored, given the restrictions set by `[if]` and/or `[in]`. All statistics allow for the use of the options `by(varname)`, used to indicate the variables over which the RIF will be estimated (i.e., sex or race group), and `weight(varname)`, used to indicate weights for the estimation of the RIFs.

`rifoptions` allows the user to specify which distributional statistics are used to obtain the RIF statistic. Table 1 provides a detailed list of the statistics that are currently available for estimation. The column *options* indicate the options, required or otherwise, needed to estimate the RIF associated to the statistic named in the *description* column. In appendix 2, a summary of performance simulation is provided to show how well the RIF standard errors approximate to the simulated standard errors for all these statistics.

[9] Internally, the program is stored in a file named `grifvar.ado`. `rifvar()` builds on the `rifreg` command and the do-file provided in the appendix in Heckley, Gerdtham, and Kjellsson (2016). All codes were adapted to allow for the estimations by groups.
[10] See appendix 1 for the full set of statistics, formulas, and sources.

**Table 1. `rifvar rifoptions`**

| Options | Description |
|---|---|
| `mean` | Sample mean |
| `var` | Variance |
| `q(#p) [kernel(kernel) bw(#)]` | pth Quantile, where $0 < p < 100$. The options `kernel`(.) and `bw`(.) are not required. The default is to use a gaussian kernel. All kernel functions available for the command `kdensity` are also allowed.[1] Unless otherwise specified, the Silverman's plug-in optimal bandwidth is used. |
| `iqr(#p1 #p2) [kernel(kernel) bw(#)]` | Interquantile range: q(p2) - q(p1), where $0 < p_1 \leq p_2 < 100$. Options for bandwidth (`bw`) and kernel function (`kernel`) are the same as for the quantile case. |
| `gini` | Gini inequality index. |
| `cvar` | Coefficient of variation |
| `iqratio(#p1 #p2) [kernel(kernel) bw(#)]` | Interquantile ratio: q(p2)/q(p1), where $0 < p_1 \leq p_2 < 100$. Options for bandwidth (`bw`) and kernel function (`kernel`) are the same as for the quantile case. |
| `entropy(#a)` | Generalized entropy index with sensitivity parameter #a. |
| `atkin(#e)` | Atkinson inequality index with inequality aversion #e>0 |
| `logvar` | Logarithmic variance |
| `glor(#p)` | Generalized Lorenz ordinate at #p where $0 < p < 100$ |
| `lor(#p)` | Lorenz ordinate at #p where $0 < p < 100$ |
| `ucs(#p)` | Share of income held by richest 1-p%. 1-lor(#p) |
| `iqsr(#p1 #p2)` | Interquantile share ratio: (1-lor(#p2))/lor(#p1), where $0 < p_1 \leq p_2 < 100$. |
| `mcs(#p1 #p2)` | Share of income held by people between #p1 and #p2: lor(#p2)-lor(#p1). where $0 < p_1 < p_2 < 100$. |
| `pov(#a) pline(#\|varname)` | FGT poverty index given sensitivity parameter #a. For a=0 one obtains the poverty head count, a=1 poverty gap and a=2 poverty severity. FGT are defined based on the poverty line `pline(.)`, which can be a scalar (fixed poverty line) or a variable (variable poverty line) |
| `watts(#povline)` | Watts poverty index. It requires a number or variable to define the poverty line. |
| `sen(#povline)` | Sen poverty index. It requires a number to define the poverty line. |
| `tip(#p) pline(#)` | Three I's of poverty (TIP) curve ordinate at $p$ for poverty line defined by `pline`(#). where $0 < p < 100$. |
| `agini` | Absolute Gini |
| `acindex(varname)` | Absolute concentration index using `varname` as the rank variable. |
| `cindex(varname)` | Concentration index using `varname` as the rank variable. |
| `eindex(varname) lb(#)ub(#)` | Erreygers index using `varname` as the rank variable, with lower bound #lb and upper bound #ub, and $\#lb < \#ub$ |
| `arcindex(varname) lb(#)` | Attainment relative concentration index using `varname` as the rank variable, with lower bound #lb |
| `srindex(varname) ub(#)` | Shortfall relative concentration index using `varname` as the rank variable, with upper bound #ub |
| `windex(varname) lb(#)ub(#)` | Wagstaff concentration index using `varname` as the rank variable, with lower bound #lb and upper bound #ub, and $\#lb < \#ub$ |

**Note:** [1]To use other kernel functions, the full kernel name should be used and not their abbreviations (i.e., `biweight` instead of `bi`), with one exception: to request using the epanechnikov kernel, one should use `epan`.

## 4. RIF-REGRESSION: `rifhdreg`

As previously indicated, IFs and RIFs have been used in statistics as a tool for analyzing the robustness of statistics to outliers and as a method to draw statistical inferences from complex statistics (Cowell and Flachaire 2015; Deville 1999; Efron 1982; Hampel 1974). A recently popularized use by Firpo, Fortin, and Lemieux (2009), Heckley, Gerdtham, and Kjellsson (2016), and Essama-Nssah and Lambert (2012) is the estimation of RIF regressions.

Firpo, Fortin, and Lemieux (2009) use this strategy to estimate unconditional partial effects (UPE) of small changes in the distribution of the dependent variable characteristics X on the distributional statistic $v$. The authors use this strategy for the estimation of UQRs using a linear model as the easiest method for approximating this partial effect.[11] Paraphrasing the original paper, the intuition behind RIF regressions can be described as follows.

Assume there is a joint distribution function $f_{Y,X}(y,x)$ that determines the linear and nonlinear relationships between the dependent variable $Y$ and all independent/exogenous variables $X$. Similar to the standard linear model, under the exogeneity assumption of $X$, the interest does not fall on estimating the full joint density function, but only the conditional distribution function $f_{Y|X}(Y|X=x) = \frac{f_{Y,X}(y,x)}{f_X(x)}$, so that $F_{Y|X}(y|X=x) = \int_{-\infty}^{y} f_{Y|X}(y|X=x)dz$. By definition, following the law of iterated expectations, the following is true:

$$F_Y(y) = \int F_{Y|X}(Y|X=x)\, dF_X(x) \tag{9}$$

which simply states the unconditional cumulative distribution of y can be obtained by integrating (averaging) the marginal cumulative $F_{Y|X}(Y|X=x)$ over all possible realizations of X.

[11] For the special case of quantiles, the RIF is defined as $q_Y(p) + \frac{p-1(y\leq q_Y(p))}{f(q_Y(p))}$. Since the only element of this expression that varies across observations is $1\big(y \leq q_Y(p)\big)$, Firpo, Fortin, and Lemieux (2009) propose to model this element using three methods: a linear probability model (RIF-OLS), a probit model (RIF-probit), or a nonparametric binomial model (RIF-NP).

Next, assume that one is interested in analyzing the distributional statistic $v(F_Y)$. Based on the concepts described previously and the properties of IFs and RIFs (see equations [5] to [7]), the statistic $v$ can be rewritten as:

$$v\big(F_Y(y)\big) = \int RIF\big(y; v(F_Y)\big)\, dF_Y(y) \tag{10a}$$

$$v\big(F_Y(y)\big) = v\big(F_Y(y)\big) + \int IF\big(y; v(F_Y)\big)\, dF_Y(y) \tag{10b}$$

Again, using the law of iterated expectations, equation (9) can be used to rewrite equation (10b) as a function of the distribution of the explanatory variables $F_X(x)$:

$$v\big(F_Y(y)\big) = v\big(F_Y(y)\big) + \int\int IF\big(y; v(F_Y)\big)\, dF_{y|X}(y|X = x)\, dF_X(x) \tag{11}$$

An intuitive interpretation of this expression is: if there is a small change in the distribution of the exogenous characteristics $\Delta Fx$, assuming that the conditional distribution $F_{y|X}$ is constant, it will generate a change in the unconditional distribution $\Delta F_Y$ which will translate into a change in the statistic $\Delta v\big(F_Y(y)\big)$. This is measured by averaging the IFs through the changes in the distribution $\Delta F_Y$.

An alternative way of expressing this equation using the law of iterated expectations is the following:

$$v\big(F_Y(y)\big) = \int E\big(RIF\big(y; v(F_Y)\big)|X = x\big)\, dF_X(x) \tag{12}$$

This last equation is used by Firpo, Fortin, and Lemieux (2009) to validate the use of RIFs related to regression analysis. Assuming a linear approximation of the relationship between Ys and Xs, then OLS can be used to estimate a linear model to capture how changes in $\Delta F_X$ relate to changes in $\Delta v_Y$. The difference with the standard OLS model is that RIF-OLS uses the estimated $RIF\big(y_i; v(F_Y)\big)$ for each observation $y_i$ in the data as the dependent variable and regresses it against all the variables of interest[12]:

[12] This is a two-step process that is done internally within the user-written commands `rifreg` and `xtrifreg`.

$$RIF\big(y_i; v(F_Y)\big) = X_i'\beta + \varepsilon_i, E(\varepsilon_i) = 0 \quad (13)$$

While the use of OLS directly relates the RIF regression to standard regression analysis, some differences in the interpretation exist. In the standard OLS, the typical interpretation of the coefficients is that for the average person (in the sample), a one-unit increase in X will increase $y$ in $\beta$ units (for that average person), everything else held constant. The interpretation from the RIF-OLS is slightly different. To obtain the UPE on the statistic $v$, one first needs to obtain unconditional expectations on both sides of equation (13):

$$v(F_Y) = E\left(RIF\big(y_i; v(F_Y)\big)\right) = E(X_i'\beta) + E(\varepsilon_i) = \bar{X}'\beta \quad (14)$$

From here the UPE is given by:

$$\frac{\partial v(F_Y)}{\partial \bar{X}_k} = \beta \quad (15)$$

Based on equation (15), the correct interpretation of the UPE is if the distribution of $x_k$ changes such that the unconditional average increases by one unit ($\Delta\bar{X}_k$=1), the expected change in the distributional statistic is equal to $\beta$. This interpretation identifies one weakness of using a linear regression that includes the explanatory variables linearly: it only captures one aspect of the distribution of X, the unconditional mean. This can be easily mended by including higher-order polynomials and interactions that would better capture some of the nonlinear relationships across the explanatory variables. For example, assuming only one exogenous variable, the following specification could be applied:

$$RIF\big(y_i; v(F_Y)\big) = \beta_0 + \beta_1 X_i + \beta_2 (X_i - \bar{X})^2 + \varepsilon_i \quad (16)$$

Considering the unconditional expectations, one obtains the following:

$$v(F_Y) = \beta_0 + \beta_1 \bar{X} + \beta_2 E((X_i - \bar{X})^2) = \beta_0 + \beta_1 \bar{X} + \beta_2 Var(X) \quad (17)$$

The UPE can now obtained as a function of changes in two moments of the unconditional distribution of X: the mean and the variance.[13]

The estimation of RIF regressions in Stata, under the linearity assumption, is easily implemented using the user-written command `rifreg` (for Gini, variance, and quantiles) or `xtrifreg` (for quantiles with one high-dimensional fixed effect). However, there are no commands for the estimation of RIF regressions for other distributional statistics, nor for when two or more high-dimensional fixed effects need to be estimated.

For the estimation of both RIF regressions under both scenarios, I introduce the command `rifhdreg`. This command is a wrapper that uses the same two-step procedure used in `rifreg` and `xtrifreg`. The first step estimates the corresponding RIF for each observation in the sample of interest for a specific distributional statistic using the previously introduced `rifvar()` command. The second step uses the RIF as the dependent variable and estimates a linear model—using the official command `regress` (official Stata command) when no fixed effects are used and `reghdfe` (Correira 2017) when fixed effects are specified—to estimate the RIF-OLS models. The syntax of the command is as follows:

```
rifhdreg depvar [indepvars] [if] [in] [weight], rif(rifvar)
[retain(str) replace abs(varlist) scale(real) regress_options
reghdfe_options ]
```

The main difference with the `regress` and `reghdfe` commands is that `rifhdreg` requires specifying the distributional statistic of interest with the option `rif(rifvar)`.

`rif(rifvar)` specifies the statistic of interest, internally estimating the corresponding RIF. It uses the same syntax presented in table 1. For example, to estimate the RIF regression for the interquantile share ratio, one can type `rifhdreg y x1 x2 x3, rif(iqsr(10 90))`

[13] While not yet explored, it is possible to use RIFs of other distributional statistics of X as explanatory variables to better capture how changes in the distribution of X ($\Delta F_X$) affect the distribution statistic of $y$ ($\Delta v(F_Y)$).

`retain(str)` as an option that saves the internally constructed RIF for the restricted sample used in the regression under a newly named variable.

`replace` allows for saving the internally constructed RIF if the variable name specified in `retain(str)` already exists.

`abs(varlist)` identifies the fixed effects to be absorbed. Each variable listed here represents one set of fixed effects.

`scale(real)` is used to provide a value and rescale the dependent variable, and is useful for statistics like the Lorenz ordinate or quantile shares, which are measures that fall between 0 and 1. The default option is 1 (norescaling).

When `abs(varlist)` is used, `rifhdreg` calls for `reghdfe` to estimate the RIF-OLS model, and all options in `reghdfe` are available. Otherwise it uses `regress` to estimate the model, allowing for all `regress` options.[14]

The command reports OLS asymptotic standard errors by default, but one can request other standard errors allowed in `regress` or `reghdfe` commands. Based on the recommendation provided in Firpo, Fortin, and Lemieux (2009) and the simulations presented in appendix 1, bootstrapped standard errors should be used when the statistic of interest is the unconditional quantile or related statistics, as well as for the Atkinson inequality index. For the correct estimation of bootstrap standard errors, one should use the `bootstrap` prefix before the `rifhdreg` command.

[14] It should be noticed that while all `regress` and `reghdfe` options are permitted, not all of them may be appropriate for the estimation of RIF regressions. I recommend using them with caution.

## 5. RIF DECOMPOSITION: `oaxaca_rif`

As previously described, one of the main advantages of RIF regressions is they can easily be used to analyze factors affecting inequality due to small changes in distribution characteristics. Often, however, there is interest in analyzing the impact of a large change in the distribution of the characteristics: specifically, comparing and decomposing the gaps in the distributional statistics between two groups, taking into account any linear and nonlinear differences in the joint distribution across those groups. RIF regressions will not be appropriate for analyzing the impact of such change, and a different strategy is required.

The OB decomposition is one of the most extensively used methodologies in labor economics and aims to analyze outcome differences between two groups (Blinder 1973; Oaxaca 1973). These differences are characterized as functions of differences in characteristics (composition effect) and differences in coefficients associated with those characteristics (*wage* structure effect). The OB decomposition is done by estimating separate regressions for each one of the groups of interest, effectively accounting for all possible interactions between the grouping variable and the relationships between outcome and exogenous variables.

While the original methodology was created to analyze outcome differences at the mean, several papers that followed provided extensions and refinements to extend the analysis to other distributional statistics (see Fortin, Lemieux, and Firpo [2011] for a review). Additionally, under the assumptions of ignorability (conditional independence) and overlapping support, the aggregate structure effect can be identified and interpreted as a treatment effect.

Firpo, Fortin, and Lemieux (2018) describe the use of RIF regressions in combination with a reweighted strategy (DiNardo, Fortin, and Lemieux 1996) as a feasible methodology for decomposing differences in distributional statistics beyond the mean. This is referred to as RIF decomposition. This methodology has three advantages compared to other strategies in the literature: the simplicity of its implementation, the possibility of obtaining detailed contributions

of individual covariates on the aggregate decomposition,[15] and the possibility of expanding the analysis to any statistic for which an RIF can be defined. This strategy can be described as follows.

Assume there is a joint distribution function that describes all relationships between the dependent variable Y, the exogenous characteristics X, and the categorical variable T: $(f_{Y,X,T}(y_i, x_i, T_i))$. Since there are only two groups based on T, the joint probability distribution function and cumulative distribution of Y conditional on T can be written as:

$$f^k_{Y,X}(y,x) = f^k_{Y|X}(Y|X) f^k_X(X) \tag{18a}$$

$$F^k_Y(y) = \int F^k_{Y|X}(Y|X) dF^k_X(X) \tag{18b}$$

where the subscript *k* indicates that the density is conditional on $T = k$ with $k \in [0,1]$. To analyze the differences between groups 0 and 1 for a given distributional statistic $v$, the cumulative conditional distribution of Y can be used to calculate the gap:

$$\Delta v = v_1 - v_0 = v(F^1_Y) - v(F^0_Y) \tag{19a}$$

$$\Delta v = v\left(\int F^1_{Y|X}(Y|X) dF^1_X(X)\right) - v\left(\int F^0_{Y|X}(Y|X) dF^0_X(X)\right) \tag{19b}$$

From equation (19b) it is easy to see that differences in the statistics of interest $\Delta v$ will arise because of differences in the distribution of Xs $\left(dF^1_X(X) \neq dF^0_X(X)\right)$ or because of differences in the relationships between Y and X $\left(F^1_{Y|X}(Y|X) \neq F^0_{Y|X}(Y|X)\right)$. In the context of the standard OB decomposition, this is equivalent to comparing differences in average characteristics and differences in coefficients.

To identify how important differences in characteristics (composition effect) and differences in coefficients (wage structure effect or coefficients effect) are for explaining the overall gap in the

[15] Akin to the standard regression analysis, the identification of the detailed contribution of covariates also requires the zero conditional mean assumption. In other words, any other variable not accounted for in the model has a distribution that is independent from the measured characteristics X.

distributional statistic $\Delta v$, it is necessary to create a counterfactual scenario. Define the counterfactual statistic $v_c$ as follows:

$$v_c = v(F_Y^c) = v\left(\int F_{Y|X}^0(Y|X)dF_X^1(X)\right) \tag{20}$$

Using this counterfactual, the gap in the distribution statistic $v$ can be disaggregated into two components:

$$\Delta v = \underbrace{v_1 - v_c}_{\Delta v_S} + \underbrace{v_c - v_0}_{\Delta v_X} \tag{21}$$

where $\Delta v_X$ reflects the gap attributed to differences in characteristics, while $\Delta v_S$ would reflect the differences attributed to the relationships between Y and X.[16] The difficulty is the identification of the counterfactual statistic $v_c$, because the combination of characteristics and outcomes is not observed in the data. Based on the review in Fortin, Lemieux, and Firpo (2011), two broad strategies have been suggested for the identification of the counterfactual statistic $v_c$. The first strategy follows the standard OB decomposition, using linear regressions and their approximations to identify $v_c$. Specifically, following equation (14), separate RIF regressions can be estimated for each group, so the counterfactual statistic can be identified as follows:

$$v_1 = E\left(RIF\left(y_i; v(F_Y^1)\right)\right) = \bar{X}^{1\prime}\hat{\beta}^1 \tag{22a}$$

$$v_0 = E\left(RIF\left(y_i; v(F_Y^1)\right)\right) = \bar{X}^{0\prime}\hat{\beta}^0 \tag{22b}$$

$$v_c = \bar{X}^{1\prime}\hat{\beta}^0 \tag{22c}$$

This alternative mirrors the standard OB decomposition where $\Delta v_X = (\bar{X}^1 - \bar{X}^0)\beta^0$ and $\Delta v_S = \bar{X}^1\left(\hat{\beta}^1 - \hat{\beta}^0\right)$. The main disadvantage of this strategy, discussed in Barsky et al. (2002) in the context of conditional means, is the counterfactual statistic $v_c$ may be incorrectly identified if

[16] In the labor economics literature, this is often referred to as the wage structure effect.

the model is misspecified,[17] or if the local approximation obtained using RIF cannot be extended beyond the local extrapolations. The alternative is to use a semiparametric reweighting approximation, as discussed in Barsky et al. (2002) and DiNardo, Fortin, and Lemieux (1996), to identify the counterfactual distribution $F^0_{Y|X}(Y|X)dF^1_X(X)$ based on the observed data. This procedure can be described as follows.

The problem of identifying the counterfactual scenario is that we do not directly observe the distribution of outcomes and characteristics that the counterfactual distribution $F^c_{Y|X}$ implies (see equation [20]). However, from an abstract point of view, it is possible to obtain an approximation for the counterfactual distribution by multiplying the observed distribution of characteristics $dF^0_X(X)$ with a factor $\omega(X)$, so it resembles the distribution $dF^1_X(X)$:

$$F^c_Y = \int F^0_{Y|X}(Y|X)dF^1_X(X) \cong \int F^0_{Y|X}(Y|X)dF^0_X(X)\omega(X) \tag{23}$$

Using Bayes rule, the reweighting factor $\omega(X)$ can be identified as follows:

$$\omega(X) = \frac{dF^1_X(X)}{dF^0_X(X)} = \frac{dF_{X|T}(X|T=1)}{dF_{X|T}(X|T=0)} = \frac{dF_{T|X}(T=1|X)}{dF_T(T=1)}\frac{dF_T(T=0)}{dF_{T|X}(T=0|X)} = \frac{1-P}{P}\frac{P(T=1|X)}{1-P(T=1|X)} \tag{24}$$

where $p$ is the proportion of people in group T=1 and $P(T=1|X)$ is the conditional probability of someone with characteristics X being part of group 1. In other words, to identify the counterfactual distribution $F^c_{Y|X}$, one can estimate the reweighting factor $\omega(X)$ using parametric or nonparametric methods to estimate the conditional probability $P(T=1|X)$. As described in Firpo, Fortin, and Lemieux (2018), in practice, a probit or logit model can be used to estimate this conditional probability.

Once these reweighting factors are obtained, equation (22c) is estimated as:

$$v_c = E\left(RIF\left(y_i; v(F^c_Y)\right)\right) = \bar{X}^{c\prime}\hat{\beta}^c \tag{25}$$

[17] Notice that the concept of misspecification here also includes the idea of accounting for changes in the whole distribution of X, not only the mean.

And the decomposition components are now defined as:

$$\Delta v = \underbrace{\bar{X}^{1\prime}\left(\hat{\beta}_1 - \hat{\beta}_c\right)}_{\Delta v_s^p} + \underbrace{(\bar{X}^1 - \bar{X}^c)'\hat{\beta}_c}_{\Delta v_s^e} + \underbrace{(\bar{X}^c - \bar{X}^0)'\hat{\beta}_0}_{\Delta v_X^p} + \underbrace{\bar{X}^{c\prime}\left(\hat{\beta}_C - \hat{\beta}_0\right)}_{\Delta v_X^e} \quad (26)$$

The components $\Delta v_s^p + \Delta v_s^e$ correspond to the OB aggregate wage structure effect, whereas $\Delta v_X^p + \Delta v_X^e$ correspond to the aggregate composition effect. These two components are further decomposed into a pure wage structure ($\Delta v_s^p$) and composition effect ($\Delta v_X^p$), plus two components that can be used to assess the overall fitness of the model. $\Delta v_s^e$ is the reweighting error that is used evaluate the quality of the reweighting strategy and is expected to go to zero in large samples. $\Delta v_X^e$ is the specification error and is used to assess the importance of departures from linearity in the model specification or the RIF approximation.

The implementation of OB decomposition in Stata is simple. The most popular command used for this type of analysis is the user-written command `oaxaca` (Jann 2008), which can be used for many of the extensions that have been developed for the analysis of average differences across groups. Extending the OB decomposition analysis to statistics other than the mean can easily be done by carefully calculating RIFs for the conditional distributions and using them as the dependent variable using the `oaxaca` command. However, no formal implementation of the estimation of RIF decompositions and the hybrid reweighted RIF decomposition is currently available.

For the estimation of these two types of decompositions, I present the command `oaxaca_rif`. This program is a wrapper around `oaxaca` that uses processes suggested in Firpo, Fortin, and Lemieux (2018) for the estimation of the standard RIF decomposition (following equation [22c]) or the hybrid reweighted decomposition (equation [26]).

The syntax of the command is as follows:

```
oaxaca_rif depvar [indepvars] [if] [in] [weight] , by(groupvar)
rif(rifvar) [swp  wgt(#) cluster(varname) scale(real) retain(str)
replace  rwlogit(varlist) rwprobit(varlist)]
```

The internal syntax of `oaxaca_rif` allows the use of most of the options available in `oaxaca`. Parallel to the `rifhdreg` command, `oaxaca_rif` requires the option `rif(rifvar)` to define the distributional statistic to be used for the decomposition analysis. Internally, it calls on `rifvar()` to estimate the RIF for each group defined by `by()`, and follows equation (22c) to identify the counterfactual and implement the decomposition.

The default option is to estimate the distributional statistic gap between observations with the lowest value in the grouping variable minus the observations in the group with the highest value, (group1-group2).

`swp` can be used to request the gap to be estimated in the opposite order, (group2-group1).

`wgt(#)` is used to define the counterfactual distribution. The default is the value 0, which identifies the decomposition according to equation (22c). Using `wgt(1)` instead uses $v_c = \bar{X}^{0\prime}\hat{\beta}^1$ as the counterfactual.

`scale(real)` is used to provide a value and rescale the dependent variable. It is useful for statistics like the Lorenz ordinate or quantile shares, which are measures that fall between 0 and 1. The default value is 1 (no-rescaling).

`retain(str)` and `replace` are used to store the internally generated RIF in a new variable or replace the existing variable if `replace` is used. For the reweighted decomposition, this option does not generate the RIF for the counterfactual option.

`rwlogit(varlist)` and `rwprobit(varlist)` are options used to specify the estimation of the reweighting factors using a logit or a probit model. When used, the command estimates the reweighted RIF decomposition.

For the reweighted standard decomposition, the command first estimates the probability model, then estimates the reweighting factor $\omega(X)$, and obtains the RIFs for the three scenarios. The decomposition output is obtained from two separate decompositions to identify the four components detailed in equation (26).

The variables included in the options `rwlogit(varlist)` or `rwprobit(varlist)` may or may not be the same as the ones used in the specification of the main model. Adding higher-order polynomials and interactions for the estimation of the conditional probability will improve the quality of the reweighting, but may create problems of overfitting and violation of the overlapping assumption. Similarly, interactions and polynomials may reduce problems of error due to model misspecification, but they can make the model more difficult to interpret.

For standard errors, the default option is to report robust standard errors, equivalent to using the robust option in the `oaxaca` command.[18] When the reweighted decomposition is requested, the command reports robust standard errors clustered at the individual level. This is done because for one of the internally estimated decompositions (see equations [22b] and [25]) the same sample is used for both groups. The option `cluster(varname)` supersedes the individual cluster option. Weights are allowed when robust or clustered standard errors are used.

As shown for the case of RIF regressions, asymptotic and robust standard errors may not be appropriate for the decomposition of statistics related to quantiles or the Atkinson index. Furthermore, since the reweighting factors $\omega(X)$ are estimated variables, standard errors of the decomposition components need to be adjusted. Given the complexity of estimating asymptotic standard errors in the framework of RIF decompositions, the suggested alternative is to use

[18] Robust standard errors for the `oaxaca` command are not available in all version of the command. For `oaxaca_rif` to work properly, be sure to have the latest version, which at the time of writing this paper is version 4.0.5.

bootstrapped standard errors throughout. Bootstrapped standard errors can be obtained using the `bootstrap` prefix. Bootstrapped standard errors cannot be used in combination with weights.

## 6. ILLUSTRATION: CHANGES IN INEQUALITY IN THE UNITED STATES, 1998 VERSUS 2018

To illustrate the use of RIF regressions and decompositions for the analysis of poverty and inequality, in this section the above-described commands are used to analyze the determinants of wage inequality in the US, exploring the changes between 1998 and 2018. For this example, data from the March Current Population Survey (CPS) is gathered for both years, concentrating on families, excluding members from other families within the household or any individual who is not a relative of the head of the household. Real family income is used as the dependent variable for the poverty analysis,[19] while real family income per capita is used for the inequality analysis.

Only variables that capture household characteristics are used as explanatory variables. These include: an indicator if the household is single headed, average age of the head and spouse (if present), number of children 0–17 years old in the household, number of people 25–64 years old (excluding head and spouse), and number of people 65 or above living in the household. I also control for noncitizenship status, Hispanic status, and nonwhite-couple status of the household. To account for socioeconomic characteristics, I control for the highest educational attainment of husband or wife, whether the household rents a house, and if either or both the husband and wife are currently employed. Controls for regions are also included.

Table 2 provides the results of the RIF regressions for the years 1998 and 2018, using the Gini coefficient, the income share held by the upper 10 percent and lower 40 percent, and poverty severity as the distributional statistics of interest. Robust standard errors using survey weights are reported. This table also includes the weighted average characteristics for both years.

[19] In the US, poverty lines are defined at the household level, not individual level.

**Table 2. Determinants of Inequality in the US: RIF Regression Approach**

| | Gini Index | | Share of Income Bottom 40% | | Share of Income Top 10% | | Poverty Severity | | Avg. Characteristics | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1998 | 2018 | 1998 | 2018 | 1998 | 2018 | 1998 | 2018 | 1998 | 2018 |
| Singe-headed household | 118.8* | 110.4* | -6.409* | -5.378* | 7.329* | 7.664* | 4.865* | 4.127* | 30.3 | 34.6 |
| | (4.420) | (4.528) | (0.183) | (0.162) | (0.479) | (0.521) | (0.187) | (0.158) | | |
| Average householder age | | | | | | | | | | |
| 24–44 | -58.99* | -35.92* | 3.611* | 1.507* | -3.446* | -3.228+ | -4.014* | -2.413* | 52.7 | 40.9 |
| | (6.992) | (11.880) | (0.380) | (0.432) | (0.616) | (1.356) | (0.531) | (0.507) | | |
| 44–64 | -24.76* | -28.87+ | 1.808* | 0.951+ | -1.166 | -2.867+ | -4.065* | -2.825* | 29.8 | 37.7 |
| | (7.596) | (11.770) | (0.388) | (0.425) | (0.713) | (1.346) | (0.512) | (0.481) | | |
| 65+ | -126.8* | -80.86* | 9.041* | 4.409* | -4.776* | -5.129* | -9.499* | -6.724* | 13.1 | 17.9 |
| | (9.108) | (12.720) | (0.439) | (0.458) | (0.897) | (1.455) | (0.544) | (0.514) | | |
| #children | 33.97* | 36.61* | -1.774* | -1.912* | 2.418* | 2.288* | 1.405* | 1.176* | 1.169 | 1.002 |
| | (1.633) | (1.606) | (0.087) | (0.073) | (0.155) | (0.168) | (0.105) | (0.097) | | |
| # HH members ages 25–64 | -50.26* | -44.12* | 2.193* | 2.253* | -4.104* | -2.898* | -1.564* | -1.327* | 0.111 | 0.156 |
| | (4.315) | (3.429) | (0.207) | (0.159) | (0.434) | (0.350) | (0.148) | (0.128) | | |
| # Elderly (65+) | -56.05* | -37.77* | 3.000* | 2.457* | -3.668* | -1.649^ | -1.926* | -1.579* | 0.024 | 0.036 |
| | (9.217) | (8.813) | (0.453) | (0.378) | (0.935) | (0.938) | (0.239) | (0.302) | | |
| =1 Husband or wife | 37.73* | 17.04* | -2.245* | -0.658* | 1.940* | 1.378* | 0.883* | -0.106 | 12.5 | 18.4 |
| is Hispanic | (5.666) | (4.776) | (0.261) | (0.204) | (0.579) | (0.513) | (0.245) | (0.195) | | |
| =1 Husband or wife | 22.90* | 25.03* | -1.066* | -1.110* | 1.941* | 1.847* | -0.122 | 0.214 | 14.8 | 22 |
| is noncitizen | (6.208) | (5.150) | (0.266) | (0.203) | (0.660) | (0.573) | (0.223) | (0.181) | | |
| =1 is not a white couple | 7.428 | 7.267 | -1.004* | -0.511* | -0.778^ | 0.236 | 1.724* | 0.685* | 17.5 | 23.4 |
| | (4.596) | (4.618) | (0.227) | (0.180) | (0.448) | (0.517) | (0.224) | (0.167) | | |
| Highest educational attainment | | | | | | | | | | |
| Less than high school | 70.84* | 57.23* | -4.405* | -3.711* | 3.470* | 2.417* | 2.946* | 1.749* | 12.7 | 7.69 |
| | (4.441) | (4.703) | (0.260) | (0.272) | (0.363) | (0.382) | (0.331) | (0.385) | | |
| Some college | -25.23* | -30.60* | 1.193* | 1.606* | -1.966* | -2.004* | -1.062* | -1.148* | 27.8 | 27.5 |
| | (3.807) | (4.312) | (0.192) | (0.184) | (0.369) | (0.463) | (0.164) | (0.180) | | |
| College | -0.458 | -42.65* | -0.357 | 1.474* | -0.783 | -4.058* | -1.261* | -1.215* | 19.1 | 24.2 |
| | (5.689) | (5.166) | (0.235) | (0.202) | (0.616) | (0.576) | (0.142) | (0.180) | | |
| Grad school | 136.0* | 28.65* | -5.805* | -1.842* | 11.43* | 1.329 | -1.214* | -1.139* | 11.6 | 19 |
| | (10.510) | (7.559) | (0.364) | (0.257) | (1.201) | (0.880) | (0.147) | (0.178) | | |
| Share of employed couples | -112.7* | -99.53* | 6.654* | 5.243* | -6.461* | -5.660* | -6.989* | -7.372* | 68.7 | 66.6 |
| | (5.617) | (5.820) | (0.254) | (0.220) | (0.579) | (0.654) | (0.288) | (0.263) | | |
| Census regions | | | | | | | | | | |
| Midwest | -21.98* | -28.45* | 1.094* | 1.252* | -1.430+ | -2.460* | -0.405+ | 0.126 | 23.7 | 21.4 |
| | (5.735) | (7.034) | (0.234) | (0.247) | (0.623) | (0.815) | (0.162) | (0.195) | | |
| South | 7.979 | -6.794 | -0.339 | 0.412^ | 0.744 | -0.467 | 0.248 | 0.131 | 35.3 | 37.8 |
| | (5.541) | (6.654) | (0.225) | (0.228) | (0.603) | (0.774) | (0.175) | (0.177) | | |
| West | 6.694 | -4.119 | -0.348 | 0.186 | 0.314 | -0.483 | -0.474* | -0.306^ | 21.9 | 23.4 |
| | (6.240) | (7.142) | (0.250) | (0.245) | (0.683) | (0.831) | (0.170) | (0.186) | | |
| Constant | 491.9* | 517.6* | 10.37* | 10.30* | 34.88* | 38.00* | 9.187* | 8.939* | | |
| | (9.297) | (13.570) | (0.467) | (0.505) | (0.888) | (1.537) | (0.564) | (0.569) | | |
| Observations | 49816 | 66618 | 49816 | 66618 | 49816 | 66618 | 49816 | 66618 | | |
| E(RIF) | 454.3 | 469.9 | 13.35 | 12.36 | 33.61 | 34.15 | 2.69 | 2.623 | | |

**Note:** Robust standard errors in parenthesis. ^ p<0.1, + p<0.05, * p<0.01.

Inequality in the US has risen steadily for the last few decades. Based on data from the US Census Bureau, the Gini index has increased from 0.459 in 1997 to 0.482 in 2017. Despite this increase in inequality, average household income increased from $75,915 to $86,220, and the poverty rate decreased from 13.3 percent to 12.3 percent over the same period of time. The estimations for the constrained data used here are very similar. Based on the estimates from table 2, inequality based on the Gini coefficient increased from 0.454 in 1997 to 0.470 in 2017, average real household income increased from $73,759 to $83,213, and poverty declined from 11.63 percent to 10.15 percent. While there are many theories that have been provided to explain the reasons behind the rising inequality in the US, this exercise will concentrate on aspects that are more related to the socioeconomic characteristics of households.

Overall, all models hint at the same direction of the estimated effects, with some differences in the magnitude of the effect the characteristics have on specific income inequality measures. In general, an increase in the share of single-headed households, an increase in the number of children living in the household, and an increase in the presence of minority households (mixed-raced households, noncitizen households, or Hispanic households) are related to increases in inequality as measured by the Gini coefficient, a reduction in the share of income held by the bottom 40 percent of the population, and an increase in the share of income held by the top 10 percent of the population. These characteristics are also related to increases in poverty severity. One exception is that the presence of nonwhite households seems to have no impact on the Gini coefficient, but reduces the share of income held by the top and the bottom of the population. This may imply that nonwhite households are more likely to be middle-income households, thus they are more likely to increase the share of income held by the middle class.

One interesting result of the regressions is that the aging of the population may have a negative effect on inequality in the long run. If the share of older households increases, it would reduce the Gini, increase the share of income held by the bottom 40 percent of the population, and reduce income held by the top 10 percentile.

If the share of households where the average age is above 24 increases, it would reduce the Gini, increase the share of income held by the bottom 40 percent of the population, and reduce income held by the top 10 percentile.

The effect, however, may depend on which segment of the population is aging faster. A faster increase in the share of households in the 44–64 age bracket would increase rather than decrease inequality.

In terms of socioeconomic characteristics, changes in educational attainment have an ambiguous effect on inequality and poverty. If the share of households with less than a high school education declines, it will have a large impact on inequality. The estimates indicate that a 5 percentage point decline (about a 33 percent reduction from the current average) can reduce inequality by 3 Gini points, assuming that the 5 percentage point decline translates into a 5 percentage point increase in high school–educated households. If the increase is observed among households with some college and college education, the inequality-decreasing effect of better education may be larger yet (in 2017, but not 1997). However, if the change translates into an increase in households with a graduate education, inequality may increase. In all scenarios, an improvement in education has the potential to reduce poverty severity.

The only variable in the model that may capture the health of the economy, as well as the labor force participation of the households, is the share of employment among the couples. The estimations suggest that increasing the share of employment in the population—most likely by generating more jobs in the economy—has a strong effect in reducing inequality. The models suggest that a 5 percentage point increase in the share of employed householders could reduce the Gini index by about 5 points, potentially reducing the poverty severity index by 0.33, or almost 10 percent of the observed levels in 1998 and 2018. Finally, the regional dummies indicate that even after controlling for other demographic characteristics, the Midwest region experienced the lowest levels of inequality and the second-lowest poverty severity in 1998. By 2017, inequality in the region declined further. This decline was accompanied by a change in the poverty severity index, bringing it closer to levels seen elsewhere in the US.

Since both composition factors and structural factors have changed in 20 years, the next step could be to implement a decomposition analysis to see how those changes explain the increase in income inequality in the US. For the implementation, the same model specification is used for the estimation of the reweighting factors, using a logit model for the estimation. The results are shown in table 3. For simplicity, robust standard errors are provided in parenthesis.[20]

[20] Robust standard errors may be incorrect since we are not correcting for the errors introduced in the first stage of the estimation of the reweighting factors.

**Table 3. Determinants of the Changes in Inequality in the US: RIF Decomposition**

| | Gini | | Share of Income: Bottom 40% | | Share of Income: Top 10% | | Poverty Severity | |
|---|---|---|---|---|---|---|---|---|
| Average RIF 2018 | 462.8* | (1.956) | 12.28* | (0.071) | 32.97* | (0.222) | 2.623* | (0.058) |
| Counterfactual 1998 B's 2008 X's | 450.7* | (2.947) | 13.01* | (0.112) | 32.62* | (0.334) | 2.259* | (0.070) |
| 1998 | 439.4* | (2.398) | 13.46* | (0.095) | 31.66* | (0.266) | 2.690* | (0.072) |
| Total difference | 23.37* | (3.095) | -1.180* | (0.119) | 1.308* | (0.347) | -0.0675 | (0.093) |
| | **Gini** | | **Share of Income: Bottom 40%** | | **Share of Income: Top 10%** | | **Poverty Severity** | |
| | **Composition effect** | **Coefficient effect** | **Composition effect** | **Coefficient effect** | **Composition effect** | **Coefficient effect** | **Composition effect** | **Coefficient effect** |
| Aggregate decomposition | 11.26* | 12.12* | -0.455* | -0.725* | 0.961* | 0.347 | -0.432* | 0.364* |
| | (1.591) | (3.677) | (0.064) | (0.138) | (0.177) | (0.418) | (0.038) | (0.092) |
| Pure composition effect | 18.65* | | -0.761* | | 1.588* | | -0.402* | |
| | (1.437) | | (0.058) | | (0.157) | | (0.039) | |
| Pure coefficient effect | | 13.42* | | -0.788* | | 0.448 | | 0.370* |
| | | (3.481) | | (0.129) | | (0.402) | | (0.086) |
| Specification error | -7.390* | | 0.306* | | -0.628* | | -0.0299^ | |
| | (0.665) | | (0.028) | | (0.076) | | (0.017) | |
| Reweighting error | | -1.303 | | 0.0633 | | -0.101 | | -0.0053 |
| | | (0.869) | | (0.045) | | (0.062) | | (0.034) |
| **Detailed Decomposition** | **Pure composition effect** | **Pure coefficient effect** | **Pure composition effect** | **Pure coefficient effect** | **Pure composition effect** | **Pure coefficient effect** | **Pure composition effect** | **Pure coefficient effect** |
| Single-headed household | 6.711* | 6.443* | -0.387* | -0.0879 | 0.366* | 0.689* | 0.239* | 0.107 |
| | (0.294) | (2.074) | (0.016) | (0.093) | (0.020) | (0.217) | (0.012) | (0.084) |
| Age composition | 1.711+ | 35.70* | -0.0457^ | -2.376* | 0.148^ | 1.257 | -0.282* | 1.372^ |
| | (0.720) | (9.864) | (0.027) | (0.543) | (0.082) | (0.863) | (0.023) | (0.727) |
| HH composition | -6.138* | 2.409 | 0.337* | -0.0685 | -0.393* | 0.103 | -0.362* | -0.213 |
| | (0.706) | (4.531) | (0.035) | (0.171) | (0.064) | (0.517) | (0.025) | (0.163) |
| HH demographics | 2.628* | 0.902 | -0.135* | 0.153 | 0.146+ | 0.467 | 0.141* | -0.217* |
| | (0.541) | (3.220) | (0.023) | (0.124) | (0.059) | (0.370) | (0.019) | (0.078) |
| HH education | 9.835* | -11.50* | -0.322* | 0.09 | 1.063* | -1.728* | -0.308* | 0.177 |
| | (1.080) | (4.402) | (0.037) | (0.203) | (0.126) | (0.457) | (0.017) | (0.192) |
| Region | 0.425* | -5.25 | -0.0130+ | -0.0484 | 0.0469+ | -1.169 | 0.00905+ | 0.199 |
| | (0.161) | (7.997) | (0.006) | (0.286) | (0.018) | (0.931) | (0.005) | (0.184) |
| Share of employed couples | 3.476* | 5.154 | -0.195* | -0.357 | 0.211* | 0.543 | 0.161* | -0.740* |
| | (0.280) | (6.672) | (0.015) | (0.269) | (0.020) | (0.745) | (0.013) | (0.269) |
| Constant | | -20.44 | | 1.907* | | 0.286 | | -0.315 |
| | | (15.79) | | (0.738) | | (1.626) | | (0.847) |

**Note:** ^ p<0.1, + p<0.05, * p<0.01, Age composition: includes all age group categories; HH composition: includes the number of children, number of working-age adults, and number of elderly in the household; HH demographics: aggregates the citizenship status, Hispanic status, and race status of the household; Region: aggregates the effect of all regions.

The top panel of table 3 summarizes the observed changes in inequality in the US. As described before, inequality in the US has increased by 23 Gini points over the period of study. This was reflected by an increase in the concentration of resources at the top of the income distribution by 1.3 percentage points, as well as a decline in the share of income held by the bottom 40 percent of the population (1.2 percentage point decline), and a very small but statistically nonsignificant decline in poverty severity.

Considering the counterfactual income distributions, changes in the distribution of observed characteristics explain between 40 percent to 75 percent of the observed changes in income inequality. The pure composition effect is larger, with a statistically significant specification error in all models, suggesting that a more flexible model should be used to better capture the composition effect. Interestingly, the decomposition regarding poverty suggests that changes in characteristics contributed to a reduced poverty severity index from 2.690 to 2.259, but the changes in the coefficients effect were large enough to counteract that effect. Except for the poverty severity analysis, the changes in the returns structure in the economy (coefficient effect) further increased inequality in the US. Nevertheless, while still large in magnitude, no statistically significant evidence is observed in regards to the impact of the coefficient effect on the share of income held by the top 10 percent of the income distribution.

The detailed decomposition provides us with more interesting details regarding the observed changes in inequality in the US. The increase in the share of single-headed households contributed to an increase in inequality and poverty severity, an effect that was further deepened by the coefficients effect. The rapid decline of the share of households in the 24–44 year-old group counteracted other inequality-reducing age structure changes in the population. Although the results are less precise, changes in the age composition had a small impact on increasing inequality, but contributed to a reduction of poverty severity. Differences in the coefficients effect, however, had a much larger impact, pushing toward higher inequality. For poverty severity, while changes in household age structure help in reducing poverty severity, the coefficients effect had a larger effect in increasing poverty severity.

The decline in the number of children in the household and increase in number of working-age adults has contributed toward a reduction of inequality and poverty severity, with almost no effect from changes in the coefficients effect. However, the observed increases in the share of minority households in the population (Hispanic households, noncitizen households, and nonwhite households) seem to be related to increases in inequality. Interestingly, while a similar effect is found in terms of changes in characteristics and increases in poverty severity, the coefficients effect suggests poverty severity decreased in terms of income returns structure.

Between 1998 and 2018, the share of households with a householder that has at least a college education increased by 12 percentage points. This change had one of the largest effects of increasing inequality but reducing poverty severity. Possibly due to this increase in the supply of highly educated workers, however, the excess returns to education declined, which is reflected in a decreasing coefficients effect on the Gini coefficient and the share of income held at the top of the population. Last but not least, the decline in the share of employment among householders has increased inequality and poverty severity. However, the weaker relationship between poverty severity and lower employment rates observed in 2018 made the coefficients effect of the share of employed couples more than compensate for the decline in employment

## 7. CONCLUSIONS

Influence functions (IFs) and recentered influence functions (RIFs) are important statistical tools that can be used to analyze the robustness of statistics to outliers, and obtain asymptotic standard errors of otherwise complex distributional statistics (Cowell and Flachaire 2007; Deville 1999). Firpo, Fortin, and Lemieux (2009) expand on this literature by proposing the use of RIFs in the context of regression and decomposition analysis. This is a simple strategy for analyzing UPEs on any distributional statistics for which an RIF can obtained.

This paper revises the intuition behind the IF and RIF, and briefly discusses the setup under which they can be used for regressions and decomposition analysis. To facilitate the implementation of these strategies and make RIFs an easy-to-use tool for the applied

econometrician, I introduced three new Stata commands: `rifvar()`, `rifhdreg`, and `oaxaca_rif`. As an illustration, a simple analysis of the determinants of inequality in the US between 1998 and 2018 was presented. It provides interesting results in regards to the demographic changes the US experienced between these years, and how they relate to the experienced increase in inequality and poverty.

**REFERENCES**


Barsky, Robert, John Bound, Kerwin Kofi Charles, and Joseph P. Lupton. 2002. "Accounting for the Black-White Wealth Gap: A Nonparametric Approach." *Journal of the American Statistical Association* 97(459): 663–73.

Blinder, Alan S. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *The Journal of Human Resources* 8(4): 436–55.

Borgen, Nicolai T. 2016. "Fixed effects in Unconditional Quantile Regression." *The Stata Journal* 16(2): 403–15.

Chung, Choe, and Philippe Van Kerm. 2018. "Foreign Workers and the Wage Distribution: What Does the Influence Function Reveal." *Econometrics* 6(2): 28.

Correira, Sergio. 2017. "Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator." Unpublished mansucript. Available at: http:// scorreia.com/research/hdfe.pdf

Cowell, Frank A., and Emmanuel Flachaire. 2007. "Income Distribution and Inequality Measurement: The Problem of Extreme Values." *Journal of Econometrics* 141(2): 1044–72.

———. 2015. "Statistical Methods for Distributional Analysis." In Anthony B. Atkinson and François Bourguignon (eds.), *Handbook of Income Distribution*. Amsterdam: Elsevier.

Davies, James B., Nicole M. Fortin, and Thomas Lemieux. 2017. "Wealth inequality: Theory, measurement and decomposition." *Canadian Journal of Economics/Revue canadienne d'économique* 50(5): 1224–61.

Deville, Jean-Claude. 1999. "Variance estimation for complex statistics and estimators: Linearization and residual techniques." *Survey Methodology* 25(2): 193–203.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach." *Econometrica* 64(5): 1001–44.

Efron, B. 1982. *The Jacknife, the Bootstrap and Other Resampling Plans*, Vol. 38. Philadelphia: Society for Industrial and Applied Mathematics.

Essama-Nssah, Boniface, and Peter J. Lambert. 2012. "Influence Functions for Policy Impact Analysis." In John A. Bishop and Rafael Salas (eds.), *Inequality, Mobility and Segregation: Essays in Honor of Jacques Silber.* Bigley, UK: Emerald Group Publishing Limited.

Firpo, Sergio P., Nicole M. Fortin, and Thomas Lemieux. 2009. “Unconditional Quantile Regressions.” *Econometrica* 77(3): 953–73.

———. 2018. “Decomposing Wage Distributions Using Recentered Influence Function Regressions.” *Econometrics* 6(3): 41.

Firpo, Sergio P., and Cristine Pinto. 2016. “Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures.” *Journal of Applied Econometrics* 31(3): 457–86.

Fortin, Nicole M., Thomas Lemieux, and Sergio P. Firpo. 2011. “Decomposition Methods in Economics.” In Orley Ashenfelter and David Card (eds.), *Handbook of Labor Economics.* Amsterdam: Elsevier.

Hampel, Frank R. 1974. “The Influence Curve and its Role in Robust Estimation.” *Journal of the American Statistical Association* 69(346): 383–93.

Heckley, Gawain, Ulf- G. Gerdtham, and Gustav Kjellsson. 2016. “A General Method for Decomposing the Causes of Socioeconomic Inequality in Health.” *Journal of Health Economics* 48: 89–106.

Jann, Ben. 2008. “The Blinder-Oaxaca decomposition for linear regression models.” *The Stata Journal* 8(4):453–79.

Oaxaca, Ronald. 1973. “Male-Female Wage Differentials in Urban Labor Markets.” *International Economic Review* 14(3): 693–709.

von Mises, Richard. 1947. “On the Asymptotic Distribution of Differentiable Statistical Functions.” *The Annals of Mathematical Statistics* 18(3): 309–48.

**APPENDIX 1. FUNCTIONAL STATISTICS AND RIFS**

| Statistic | Definition | RIF | Source |
|---|---|---|---|
| Mean | $\mu_Y = \int y dF_Y(y)$ | $RIF(y, \mu_Y) = y$ | Firpo, Fortin, and Lemieux (2018) |
| Variance | $\sigma_Y^2 = \int (y - \mu_Y)^2 dF_Y(y)$ | $RIF(y, \sigma_Y^2) = (y - \mu_Y)^2$ | Firpo, Fortin, and Lemieux (2018) |
| pth quantile | $q_Y(p) = F_Y^{-1}(p)$ | $RIF\big(y, q_Y(p)\big) = q_Y(p) + \frac{p - 1\big(y \le q_Y(p)\big)}{f\big(q_Y(p)\big)}$ | Firpo, Fortin, and Lemieux (2018) |
| Interquantile range | $iqr_Y(p_1, p_2) = q_Y(p_2) - q_Y(p_1)$ | $RIF\big(y, iqr_Y(p_1, p_2)\big) = RIF(y, q_Y(p_1), F_Y) - RIF(y, q_Y(p_2), F_Y)$ | Firpo, Fortin, and Lemieux (2018) |
| Gini | $Gini_Y = 1 - \frac{2}{\mu_Y} R_Y$<br>$R_Y = \int_0^1 GL_Y(p) dp$<br>$GL_y(p) = \int_{-\infty}^{q_Y(p)} y\, dF_Y(y)$ | $RIF(y, Gini_Y) = 1 + \frac{2}{\mu_Y^2} R_Y - \frac{2}{\mu_Y}\big[y\big(1 - F_Y(y)\big)\big]$ | Firpo, Fortin, and Lemieux (2018) |
| Coefficient of variation | $cv_Y = \frac{\sigma_Y}{\mu_Y}$ | $RIF(y, cv_Y) = cv_Y + \frac{1}{2}\frac{(y - \mu_Y)^2 - (v - \mu_Y^2)}{\mu_Y * \sqrt{v - \mu_Y^2}} - \frac{\sqrt{v - \mu_Y^2}}{v}(y - \mu_Y)$<br>$v = \int y^2 dF_Y(y)$ | Firpo and Pinto (2016) |
| Interquantile ratio | $iqratio_Y(p_1, p_2) = \frac{q_Y(p_2)}{q_Y(p_1)}$ | $RIF\big(y, iqratio_Y(p_1, p_2)\big) = iqratio_Y(p_1, p_2) + \frac{1}{q_Y(p_1)}\left(\frac{p_2 - 1\big(y \le q_Y(p_2)\big)}{f\big(q_Y(p_2)\big)} - \frac{q_Y(p_2)}{q_Y(p_1)}\frac{p_1 - 1\big(y \le q_Y(p_1)\big)}{f\big(q_Y(p_1)\big)}\right)$ | Chung and Van Kerm (2018) |
| Generalized entropy index, $\alpha \notin \{0,1\}$ | $I_{Y_E}^{\alpha} = \frac{1}{\alpha(\alpha - 1)}\left(\frac{v}{{\mu_Y}^{\alpha}} - 1\right)$<br>$v = \int y^{\alpha} dF_Y(y)$ | $RIF\big(y, I_{Y_E}^{\alpha}\big) = I_E^{\alpha} + \frac{y^{\alpha} - v}{\alpha(\alpha - 1)\mu_Y^{\alpha}} - \frac{v}{(\alpha - 1)\mu_Y^{\alpha + 1}}(y - \mu_Y)$ | Cowell and Flachaire (2007) |

| | | | |
|---|---|---|---|
| Generalized entropy index, $\alpha = 1$ | $I_{Y_E}^{1} = \frac{v}{\mu_Y} - \log \mu_Y$<br>$v = \int y \ln y \, dF_Y(y)$ | $RIF\left(y, I_{Y_E}^{1}\right) = I_{Y_E}^{1} + \frac{1}{\mu_Y}(y \ln y - v) - \frac{v + \mu_Y}{{\mu_Y}^2}(y - \mu_Y)$ | Cowell and Flachaire (2007) |
| Generalized entropy index, $\alpha = 0$ | $I_{Y_E}^{0} = \log \mu_Y - v$<br>$v = \int \ln y \, dF_Y(y)$ | $RIF\left(y, I_{Y_E}^{0}\right) = I_{Y_E}^{0} - (\ln y - v) + \frac{1}{\mu_Y}(y - \mu_Y)$ | Cowell and Flachaire (2007) |
| Atkinson index $\epsilon > 0 \; and \; \epsilon \neq 1$ | $I_A^{\epsilon} = 1 - \frac{v^{\frac{1}{1-\epsilon}}}{\mu_Y}$<br>$v = \int y^{1-\epsilon} dF_Y(y)$ | $RIF\left(y, I_{Y_A}^{\epsilon}\right) = I_A^{\epsilon} + \frac{v^{\frac{\epsilon}{1-\epsilon}}}{(\epsilon - 1)\mu_Y}(y^{1-\epsilon} - v) + \frac{v^{\frac{1}{1-\epsilon}}}{\mu_y^2}(y - \mu_Y)$ | Cowell and Flachaire (2007) |
| Atkinson index $\epsilon = 1$ | $I_A^{1} = 1 - \frac{e^{v}}{\mu_Y}$<br>$v = \int \ln y \, dF_Y(y)$ | $RIF\left(y, I_{Y_A}^{1}\right) = I_A^{1} - \frac{e^{v}}{\mu_Y}(\ln y - v) + \frac{e^{v}}{\mu_Y^2}(y - \mu_Y)$ | Cowell and Flachaire (2007) |
| Logarithmic variance | $LV_Y = \int \left(\ln \frac{y}{\mu_Y}\right)^2 dF_Y(y)$ | $RIF(y, LV_Y) = LV_Y + [(\log y)^2 - v_1] - 2 \log \mu_Y (\ln y - v_2) - \frac{2}{\mu_Y}(v_2 - \ln \mu_Y)(y - \mu_Y)$<br>$v_1 = \int (\ln y)^2 dF_Y(y) \, ; v_2 = \int \ln y dF_Y(y)$ | Cowell and Flachaire (2007) |
| Generalized Lorenz ordinate | $GL_y(p) = \int_{-\infty}^{q_Y(p)} y \, dF_Y(y)$ | $RIF\left(y, GL_y(p)\right) = pq_Y(p) + \left(y - q_Y(p)\right)\left(y < q_Y(p)\right)$ | Essama-Nssah and Lambert (2012) |
| Lorenz ordinate | $L_Y(p) = \frac{GL_y(p)}{\mu_Y}$ | $RIF\left(y, L_Y(p)\right) = L_Y(p)\left(1 - \frac{y}{\mu_Y}\right) + \frac{pq_Y(p)}{\mu_Y} + \left(\frac{y - q_Y(p)}{\mu_Y}\right)\left(y < q_Y(p)\right)$<br>$IF\left(y, L_Y(p)\right) = -\frac{y}{\mu_Y}L_Y(p) + \frac{pq_Y(p)}{\mu_Y} + \left(\frac{y - q_Y(p)}{\mu_Y}\right)\left(y < q_Y(p)\right)$ | Essama-Nssah and Lambert (2012) |
| Upper class share | $ucs_Y(p) = 1 - L_Y(p)$ | $RIF\left(y, ucs_Y(p)\right) = ucs_Y(p) - IF\left(y, L_Y(p)\right)$ | No source |
| Interquantile share ratio | $Iqrs_Y(p_1, p_2) = \frac{1 - L_Y(p_2)}{L_Y(p_1)}$ | $RIF\left(y, Iqrs_Y(p_1, p_2)\right) = Iqrs_Y + \frac{1}{L_Y(p_1)}\left(-IF\left(y, L_Y(p_2)\right) - \frac{1 - L_Y(p_2)}{L_Y(p_1)} IF\left(y, L_Y(p_1)\right)\right)$ | No source |
| Middle-class share | $mcs_Y(p_1, p_2) = L_Y(p_2) - L_Y(p_1)$ | $RIF\left(y, mcs_Y(p_1, p_2)\right) = RIF\left(y, L_Y(p_2)\right) - RIF\left(y, L_Y(p_1)\right)$ | Davies, Fortin, and Lemieux (2017) |

| | | | |
|---|---|---|---|
| Foster-Greer-Thorbecke poverty indices | $FGT_Y(\alpha, Z) = \int_{-\infty}^{Z} \left(\frac{Z-y}{Z}\right)^{\alpha} dF_Y(y)$ $\alpha \geq 0, Z = poverty\ line$ | $RIF\left(y, FGT_Y(\alpha, Z)\right) = \left(\frac{Z-y}{Z}\right)^{\alpha} (y \leq Z)$ | Essama-Nssah and Lambert (2012) |
| Watts index | $W_Y(Z) = \int_0^Z \ln\frac{Z}{y} dF_Y(y)$ | $RIF\left(y, W_Y(Z)\right) = \ln\frac{Z}{y} (y < Z)$ | Essama-Nssah and Lambert (2012) |
| Sen index | $S_Y(Z) = \frac{2}{ZF_Y(Z)} \int_0^Z (z - y)\left(F_Y(Z) - F_Y(y)\right) dF_Y(y)$ | $RIF(y,) = RIF\left(y, S_Y(Z)\right) = -\frac{1}{F_Y(Z)} S_Y(Z) - \frac{2}{ZF_Y(Z)} \int_0^y \left(F_Y(Z) - F_Y(x)\right) dx + 2$ | Essama-Nssah and Lambert (2012) |
| TIP curve ordinate | $TIP_Y(Z,p) = \int_0^x (z - y) dF_Y(y)$ $x = \min\left(Z, q_Y(p)\right)$ | $RIF\left(y, TIP_Y(Z,p)\right) = \begin{bmatrix} (Z-y) * (Z > y)\ if\ Z < q_Y(p) \\ p\left(Z - q_Y(p)\right) + \left(q_Y(p) - y\right)(q_Y(p) > y)\ if\ Z \geq q_Y(p) \end{bmatrix}$ | Essama-Nssah and Lambert (2012) |
| Absolute Gini | $Agini_Y = 2\int_{-\infty}^{\infty} (y - \mu_Y)(F_Y(y) - 0.5) dF_Y(y) = 2Cov\left(y, F_Y(y)\right)$ | $RIF(y. Agini_Y) = -Agini_Y + \left(\mu_y - y\right) + 2\left(yF_y(y) - GL_Y\left(F_y(y)\right)\right)$ | Essama-Nssah and Lambert (2012) |
| For the following indices, one assumes that the data used can be written as: $(H, Y) = [(h_1, y_1), (h_2, y_2), \ldots (h_n, y_n)]$. The joint probability function for $H$ and $F_Y$ are $f_{H,F_Y}$ and $F_{H,F_Y}$, and the data contamination is $\delta_{h,y}(h_c, y_c) = 1$ if $h_i \geq h_c$ & $F(y_i) \geq F(y_c)$ | | | |
| Absolute concentration index | $ACI\left(h, F_{H,F_Y}\right) = 2Cov\left(h, F_Y(y)\right)$ where h is the variable interest and y the ranking variable. | $RIF\left(h, ACI\left(h, F_{H,F_Y}\right)\right) = ACI\left(h, F_{H,F_Y}\right) + IF\left(h, ACI\left(h, F_{H,F_Y}\right)\right)$ $IF\left(h, ACI\left(h, F_{H,F_Y}\right)\right) = -2ACI\left(h, F_{H,F_Y}\right) + (\mu_H - h) + 2\left(hF_{H,F_Y} - \int^y \int^{\infty} h\, f_{H,F_Y} dh\, dF_y(x)\right)$ | Heckley, Gerdtham, and Kjellsson (2016) |
| Concentration index | $CI(h, F_Y) = \frac{ACI\left(h, F_{H,F_Y}\right)}{\mu_H}$ | $RIF\left(h, CI\left(h, F_{H,F_Y}\right)\right) = CI\left(h, F_{H,F_Y}\right) + \frac{\mu_H - h}{\mu_H^2} ACI\left(h, F_{H,F_Y}\right) + \frac{1}{\mu_H} IF\left(h, ACI\left(h, F_{H,F_Y}\right)\right)$ | Heckley, Gerdtham, and Kjellsson (2016) |
| Erreygers index | $EI\left(h, F_{H,F_Y}, ub, lb\right) = \frac{4ACI\left(h, F_{H,F_Y}\right)}{ub - lb}$ | $RIF\left(h, EI\left(h, F_{H,F_Y}, ub, lb\right)\right) = EI\left(h, F_{H,F_Y}, ub, lb\right) + \frac{4}{ub - lb} IF\left(h, ACI\left(h, F_{H,F_Y}\right)\right)$ | Heckley, Gerdtham, and Kjellsson |

| | | | |
|---|---|---|---|
| | | | (2016) |
| Attainment relative concentration index | $ARI\left(h, F_{H,F_Y}, lb\right) = \frac{ACI\left(h, F_{H,F_Y}\right)}{\mu_H - lb}$ | $RIF\left(h, ARI\left(h, F_{H,F_Y}, lb\right)\right) = ARI\left(h, F_{H,F_Y}, lb\right) + \frac{\mu_H - h}{(\mu_H - lb)^2} ACI\left(h, F_{H,F_Y}\right) + \frac{1}{\mu_H - lb} IF\left(h, ACI\left(h, F_{H,F_Y}\right)\right)$ | Heckley, Gerdtham, and Kjellsson (2016) |
| Shortfall relative concentration index | $SRI\left(h, F_{H,F_Y}, ub\right) = \frac{ACI\left(h, F_{H,F_Y}\right)}{ub - \mu_H}$ | $RIF\left(h, SRI\left(h, F_{H,F_Y}, ub\right)\right) = SRI\left(h, F_{H,F_Y}, ub\right) + \frac{h - \mu_H}{(ub - \mu_H)^2} ACI\left(h, F_{H,F_Y}\right) + \frac{1}{ub - \mu_H} IF\left(h, ACI\left(h, F_{H,F_Y}\right)\right)$ | Heckley, Gerdtham, and Kjellsson (2016) |
| Wagstaff index | $WI\left(h, F_{H,F_Y}, ub, lb\right) = \frac{(ub - lb)ACI\left(h, F_{H,F_Y}\right)}{(ub - \mu_H)(\mu_H - lb)}$ | $RIF\left(h, WI\left(h, F_{H,F_Y}, ub, lb\right)\right) = WI\left(h, F_{H,F_Y}, ub, lb\right) + \frac{(h - \mu_H)(ub - lb)(ub + lb - 2\mu_H)}{\left((ub - \mu_H)(\mu_H - lb)\right)^2} ACI\left(h, F_{H,F_Y}\right) + \frac{ub - lb}{(ub - \mu_H)(\mu_H - lb)} IF\left(h, ACI\left(h, F_{H,F_Y}\right)\right)$ | Heckley, Gerdtham, and Kjellsson (2016) |

## APPENDIX 2. RIF STATISTICS AND STATISTICAL INFERENCE

This appendix provides the Monte Carlo simulation results used to evaluate the use of RIFs for statistical inference regarding distributional statistics. This particular use of RIFs has been discussed in Cowell and Flachaire (2015), Deville (1999), and Efron (1982).

For the statistics provided below, I use a sample of 2,500 observations of two variables, $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$, drawn from a jointly standard normal distribution, with variance 1 and correlation=0.5. To simulate data that resemble income distributions, the random draws for $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are used to create draws from a chi2 distribution with 5 degrees of freedom, using an inverse transformation approach:

$$ix_{i,k} = F^{-1}_{\chi^2(5)}\left(\Phi\left(x_{i,k}\right)\right) \text{ for } k = 1 \,\&\, 2$$

where $F^{-1}_{\chi^2(5)}(.)$ is the inverse cumulative function corresponding to a chi2 distribution with 5 degrees of freedom, and $\Phi(.)$ is the normal cdf. Based on this data structure, 10,000 repetitions are drawn and the RIFs detailed in table A1 are obtained for estimating the RIFs' standard errors. For the bivariate distributional concentration, indices are estimated for variable $ix_{i,1}$ based on the ranking from $ix_{i,2}$.

A simple look at the results, in particular the ratio between the simulated standard error and the average standard error obtained from using the RIFs, shows that for most statistics the results are robust, with two exceptions.

The largest biases seem to be associated with the estimation of standard errors for sample quantile, interquantile range, and interquantile ratio statistics, in particular when using quantiles at the lower end of the distribution (10th). On average, the RIFs' standard errors overstate the simulation-based standard errors by almost 10 percent compared to the simulated standard errors. This has also been described in Firpo, Fortin, and Lemieux (2009), who indicate that the

estimation of the sample density increases the complexity for the estimation of the asymptotic standard errors for UQRs, suggesting instead the use of bootstrapped standard errors.

One also observes that the variance associated to the Atkinson statistics understates the standard errors by almost 9 percent for an inequality aversion =2. Additional simulations (not shown here) suggest that the size of the bias increases with the degree of inequality aversion, but is negligible for low levels of inequality aversion. This also suggests the use of bootstraped standard errors when one is interested in drawing conclusions in regards to this inequality index.

The use of RIFs for the estimation of standard errors seems to be robust for all other statistics, with an average bias of less than 1 percent, based on the current simulation.

**Table A1. Simulation Results: Evaluating Asymptotic Performance of RIFs for the Estimation of Statistics Sample Errors**

| Statistic | Distributional statistic $v$ | Simulation-based standard error | Avg. RIF standard error | Ratio[1] |
|---|---|---|---|---|
| Mean | 5.0003 | 0.0630 | 0.0632 | 1.0031 |
| Variance | 9.9946 | 0.4170 | 0.4166 | 0.9991 |
| 10th quantile | 1.6111 | 0.0490 | 0.0533 | 1.0865 |
| 50th quantile | 4.3525 | 0.0727 | 0.0739 | 1.0160 |
| 90th quantile | 9.2384 | 0.1623 | 0.1596 | 0.9831 |
| 50/10 interquantile range | 2.7414 | 0.0733 | 0.0753 | 1.0276 |
| 90/50 interquantile range | 4.8859 | 0.1538 | 0.1520 | 0.9877 |
| Gini index | 0.3394 | 0.0044 | 0.0044 | 1.0030 |
| Coefficient of variation | 0.6321 | 0.0106 | 0.0105 | 0.9945 |
| 50/10 interquantile Ratio | 2.7037 | 0.0801 | 0.0860 | 1.0740 |
| 90/50 interquantile Ratio | 2.1229 | 0.0420 | 0.0420 | 1.0019 |
| Entropy index e=0 | 0.2130 | 0.0060 | 0.0060 | 1.0006 |
| Entropy index e=1 | 0.1868 | 0.0050 | 0.0050 | 1.0006 |
| Entropy index e=2 | 0.1998 | 0.0067 | 0.0066 | 0.9944 |
| Atkinson index a=1 | 0.1919 | 0.0048 | 0.0049 | 1.0007 |
| Atkinson index a=1.5 | 0.2930 | 0.0079 | 0.0079 | 0.9933 |
| Atkinson index a=2 | 0.3995 | 0.0144 | 0.0132 | 0.9140 |
| Logarithmic variance | 0.5355 | 0.0192 | 0.0192 | 0.9972 |
| Generalized lorenz ordinate at p=20 | 0.3079 | 0.0080 | 0.0080 | 1.0065 |
| Generalized lorenz ordinate at p=40 | 0.9080 | 0.0174 | 0.0175 | 1.0060 |
| Generalized lorenz ordinate at p=60 | 1.7812 | 0.0285 | 0.0286 | 1.0052 |
| Generalized lorenz ordinate at p=80 | 3.0037 | 0.0423 | 0.0423 | 1.0011 |
| Lorenz ordinate at p=20 | 0.0616 | 0.0014 | 0.0014 | 1.0042 |
| Lorenz ordinate at p=50 | 0.2616 | 0.0031 | 0.0031 | 1.0051 |
| Lorenz ordinate at p=80 | 0.6007 | 0.0037 | 0.0037 | 1.0035 |
| Upper class share at p=20 | 0.9384 | 0.0014 | 0.0014 | 1.0042 |
| Upper class share at p=50 | 0.7384 | 0.0031 | 0.0031 | 1.0051 |
| Upper class share at p=80 | 0.3993 | 0.0037 | 0.0037 | 1.0035 |
| Interquantile share ratio 90/10 | 10.8464 | 0.4258 | 0.4252 | 0.9987 |
| Interquantile share ratio 80/20 | 6.4894 | 0.1851 | 0.1856 | 1.0028 |
| Interquantile share ratio 60/40 | 3.5463 | 0.0692 | 0.0694 | 1.0033 |
| Middle-class share 10/90 | 0.7422 | 0.0031 | 0.0031 | 1.0049 |
| Middle-class share 20/80 | 0.5391 | 0.0033 | 0.0033 | 1.0053 |
| Middle-class share 40/60 | 0.1746 | 0.0016 | 0.0016 | 1.0041 |
| FGT(a=0,Z=2.5) headcount | 0.2235 | 0.0084 | 0.0083 | 0.9949 |
| FGT(a=1,Z=2.5) poverty gap | 0.0777 | 0.0036 | 0.0036 | 1.0058 |
| FGT(a=2,Z=2.5) poverty severity | 0.0389 | 0.0023 | 0.0023 | 1.0082 |
| Watts poverty index Z=2.5 | 0.1154 | 0.0062 | 0.0062 | 1.0061 |

| | | | | |
|---|---|---|---|---|
| Sen poverty index Z=2.5 | 0.1072 | 0.0047 | 0.0048 | 1.0046 |
| TIP ordinate at p=10 z=2.5 | 0.1411 | 0.0039 | 0.0039 | 1.0034 |
| TIP ordinate at p=25 z=2.5 | 0.1942 | 0.0090 | 0.0091 | 1.0058 |
| TIP ordinate at p=50 z=2.5 | 0.1942 | 0.0090 | 0.0091 | 1.0058 |
| Absolute Gini | 1.6963 | 0.0307 | 0.0308 | 1.0028 |
| Absolute concentration index | 0.8521 | 0.0356 | 0.0354 | 0.9948 |
| Concentration index | 0.1705 | 0.0066 | 0.0066 | 0.9941 |
| Erreygers index lb(1) ub(9) | 0.4261 | 0.0178 | 0.0177 | 0.9948 |
| Attainment relative concentration index lb(1) | 0.2130 | 0.0082 | 0.0082 | 0.9947 |
| Shortfall relative concentration index ub(9) | 0.2132 | 0.0106 | 0.0106 | 0.9965 |
| Wagstaff index lb(1) ub(9) | 0.4262 | 0.0178 | 0.0177 | 0.9948 |

**Note:** Monte Carlo simulation using 10,000 repetitions. The ratio is defined as the ratio between the average RIF standard errors and the simulation-based standard errors.