Recentered Influence Functions (RIF) in Stata

RIF-regression and RIF-decomposition

Fernando Rios-Avila [1]

[1]Levy Economics Institute
Bard College

Stata Conference-Chicago 2019

Table of Contents

Table of Contents

Prologue

- Interested in the commands. Download it from ssc: ssc install rif
- Latest Files: https://bit.ly/2NFM3cH

Table of Contents

Introductions: Why ?

- Once upon a time (2011), I was young(er), and came across a paper: Firpo, Fortin and Lemieux (2009): Unconditional Quantile Regressions (UQR).
- The premise was simple: A regression framework analysis to explore factors behind changes across the unconditional distributions (quantiles).
- Similar (Conditional) Quantile regression, but not quite the same.
- As many people. Sat down, read the paper and its companions many times. After understanding what it did, and apply it for my dissertation. (-rifreg-)

Introduction: Why?

- Few years later(2017), couple of papers with the method, decided to teach it in my econometrics class. There was a problem.
- Implementations of UQR in Stata were limited: -rifreg-, -xtrifreg-, -rifireg-. There was no "easy" applications for decompositions.
- I had programs that were too crude and clunky. Hard to share with students.
- So what to do: if the solution does not exist yet. Solve it yourself!

Table of Contents

How to compare distributional statistics?

- When comparing distributional statistics, one requires one of the following items:
  - Collection of data: $Y = [y_1, y_2, y_3, ..., y_N]$
  - The Cumulative distribution function $F(Y)$ or F_Y
  - The probability density function $f(Y)$ or f_Y
- Once any one of these three pieces is obtained, any distributional statistic ($v()$) can be easily estimated. And differences across two groups can be obtained straight forward.

$$\Delta v = v(G_Y) - v(F_y)$$

- Where Δv is the change in v when $F_y \rightarrow G_y$

Table of Contents

RIFs, IFs and Gateaux Derivative

- Influence Functions (IF) can be thought as a generalization of the above experiment.
- It represents the re-scaled effect that a change in the distribution from $F_y \to G_y$ has on statistic v, when the change is infinitesimally small:

$$G_Y^{y_i} = (1-\varepsilon)F_Y + \varepsilon 1_{y_i}$$

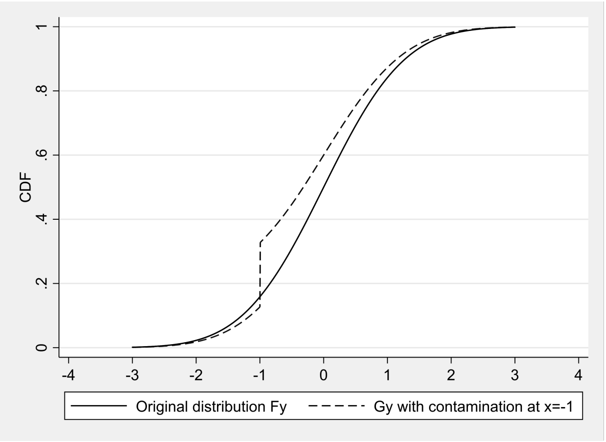
$$IF(y_i, v(F_Y)) = \lim_{\varepsilon \to 0} \frac{v(G_Y^{y_i}) - v(F_Y)}{\varepsilon}$$

- And, as introduced by FFL(2009)

$$RIF(y_i, v(F_Y)) = v(F_Y) + IF(y_i, v(F_Y))$$

- The contribution of y_i to the statistic $v()$

Visual Example of the change in F



RIF's Properties

- RIF has the following characteristics:

$$RIF(y_i, v(F_Y)) = v(F_Y) + IF(y_i, v(F_Y))$$

$$E(RIF(y_i, v(F_Y))) = v(F_Y)$$

$$E(IF(y_i, v(F_Y))) = 0$$

$$Var(v(F_Y)) = E(IF(y_i, v(F_Y))^2)$$

Why are they useful?

- Visual tool to inspect data, analyze statistics robustness to outliers (Cowel and Flatchaire, 2007)
- Simple estimation of standard errors of distributional statistic (Deville, 1999)
- Analysis of unconditional partial effects on distributional statistics based on regression and decomposition analysis (FFL, 2009,2018)

Table of Contents

How are RIF's Estimated?

- The estimation of RIFs varies in complexity depending on the statistic of interest:
  Mean:
  $$RIF(y_i, \mu_Y) = y_i$$
  Variance:
  $$RIF(y_i, \sigma_Y^2) = (y_i - \mu_Y)^2$$
  Quantile:
  $$RIF(y_i, q_Y(p)) = q_Y(p) + \frac{p - 1(y \leq q_Y(p))}{f_Y(q_Y(p))}$$
- But complexity increases for other statistics.
- In Rios-Avila (2019) I provide a collection of RIFs for a large set of distribution statistics. They include the statistics from FFL(2018), Firpo and Pinto (2016), Chung and Vankerm (2018), Cowell and Flachaire (2007), Essama-Nssah and Lambert (2012) and Heckley et al (2016).

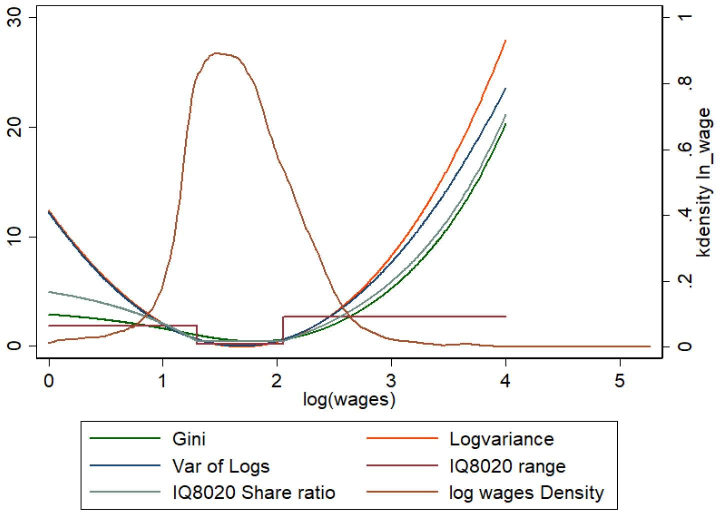
Using _grifvar()

- _grifvar() is an addon for egen(), that can be used to estimate all RIF's detailed in Rios-Avila(2019). It can be installed using (ssc install rif)
- The syntax is:
  `egen new=rifvar(oldvar) [if/in], [by() weight() rifoptions]`
  `rifoptions: Mean, variance, Coefficient of variation, standard deviation, quantile, Interquantile range, interquantile ratio, Gini, etc`
  For further detail -help rifvar-

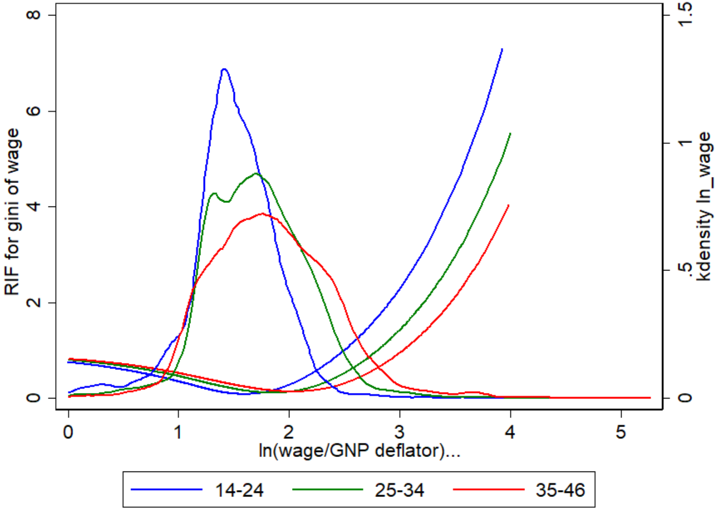
Example using _grifvar

```
webuse nlswork, clear
gen wage=exp(ln_wage)
egen rif_gini=rifvar(wage), gini
egen rif_log=rifvar(wage), logvar
egen rif_varlog=rifvar(ln_wage), var
egen rif_iqr=rifvar(ln_wage), iqr(20 80)
egen rif_iqsr=rifvar(wage), iqsr(20 80)
recode age (14/24=1 "14-24") (25/34=2 "25-34") (35/46=3 "35-46"),
gen(age_g)
egen rif_gini_age=rifvar(wage), gini by(age_g)
```

Example using _grifvar



Example using _grifvar



Example using _grifvar

Bootstrap with INEQDECO vs Mean RIF

	Observed Coef.	Bootstrap Std. Err.	Mean	Std. Err.
e_1	.1332862	.0018928	.1332862	.0019675
e_0	.1248232	.0021064	.1248232	.0021383
a05	.0635771	.0013159	.0635771	.0013272
a2	.2104676	.0023613	.2104676	.0024529
p9010	3.033107	.0167076	3.033602	.017874
p7525	1.826839	.0101527	1.825998	.0077394
gini	.2731835	.0020243	.2731836	.0020572

Table of Contents

RIF Regression: rifhdreg

- FFL(2009) Introduced the a new type of quantile regression that they call unconditional quantile regression. This was a special case of RIF regressions.
- The core of the idea was:
  - In a linear regression $y = b_0 + b_1 * x_1 + b_2 * x_2 + e$ we are modeling how changes in x's may cause a change in y.
  - $RIF(y_i, v(F_Y))$ is the contribution of an observation y_i has on the construction of statistic v.
  - then, if we model $RIF(y_i, v(F_Y)) = a_0 + a_1 * x_1 + a_2 * x_2 + e$, we are modeling how changes in X's relate to the contributions of observation i to the statistic of interest.
- FFL(2009) proposed using the RIF instead of IF. (No impact on regressions)

RIF Regression: rifhdreg

- So now that we are modeling RIF's as functions of X's. The interpretation requires some care. why?

$$RIF(y_i, v(F_Y)) = a_0 + a_1 * x_1 + a_2 * x_2 + e$$

- The simple partial effect tell us...nothing, except for few exceptions (for example Mean, FGT and Watts poverty indices).

$$\frac{\partial RIF(.)}{\partial x_1} = a_1$$

$$\frac{\partial E(RIF(.)|x_1, x_2)}{\partial x_1} = a_1$$

why? if x_1 changes for person i, that persons influence on the outcome will change in a_1. But, in a population of millions, one person won't make a difference on v.

RIF Regression: rifhdreg

- Alternatively, if we take unconditional expectations:

$$E\Big(RIF(y_i, v(F_Y))\Big) = E(a_0 + a_1 * x_1 + a_2 * x_2 + e)$$

$$v(F_Y) = a_0 + a_1 * E(x_1) + a_2 * E(x_2)$$

- So we can derive correct partial effect

$$\frac{\partial v(F_Y)}{\partial E(x_1)} = a_1$$

- a_1 is the effect that a unit change in the average value of $E(x_1)$ will have on statistic v, assuming everything else constant.
- For most v(), one needs to assumes everyone's x change in 1 unit.
- For Dummy variables, one needs to assume the change is in the Proportion of people in a particular group.

RIF Regression: rifhdreg

- Up until now, 3 other options were available for the estimation of RIF regressions: -rifreg- (FFL2009); -xtrifreg- (Borgen2016);-rifireg- Heckley et al (2016).
- the command -rifhdreg- does everything this other commands do with additional capabilities.
  - Can estimate all RIFs using _grifvar()
  - It is a wrapper around regress and reghdfe (Correira 2017). So most of their capabilities are used.
  - Different weight options, robust standard errors, fixed effects and allows for factor variables.
- It has a simple syntax:

```
rifhdreg depvar indepvar [aw pw iw] [if in],
rif(rifoptions) regress_options reghdfe_options
```

rifhdreg: Example

RIF regression with rescaled RIF:

```
rifhdreg wage age grade union tenure hours wks_work,
rif(gini) scale(1000) robust
```

```
rifhdreg wage age grade union tenure hours wks_work,
rif(ucs(80)) scale(100) robust
```

RIF regression with rescaled RIF and Fixed effects:

```
rifhdreg wage age grade union tenure hours wks_work,
rif(gini) scale(1000) vce(robust) abs(idcode) keepsingleton
rifhdreg wage age grade union tenure hours wks_work,
```

```
rif(ucs(80)) scale(100) vce(robust) abs(idcode)
keepsingleton
```

RIF Regression: rifhdreg

```
. rifhdreg wage age grade union tenure hours wks_work, rif(lor(20)) scale(100) robust

Linear regression                               Number of obs     =     18,601
                                                F(6, 18594)       =     107.15
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0402
                                                Root MSE          =      6.539
```

RIF(wage)	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.1328956	.0109562	-12.13	0.000	-.1543707	-.1114205
grade	-.2921282	.025608	-11.41	0.000	-.3423223	-.2419341
union	-.1795715	.1245977	-1.44	0.150	-.4237943	.0646514
tenure	-.08538	.0124141	-6.88	0.000	-.1097128	-.0610472
hours	.0643307	.0085744	7.50	0.000	.047524	.0811373
wks_work	.0084812	.0022489	3.77	0.000	.0040731	.0128892
_cons	15.27706	.4965183	30.77	0.000	14.30384	16.25028

```
Distributional Statistic: lor(20)
Sample Mean    RIF lor(20) :  9.8941
```

RIF Regression: rifhdreg

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	gini	ucs80	lor20	mean X	FE gini	FE ucs80	FE lor20
age	4.905	0.350	-0.133	31.39	5.424	0.398	-0.134
	(0.548)	(0.0470)	(0.0110)	(0.0454)	(1.079)	(0.0936)	(0.0183)
grade	4.604	0.0708	-0.292				
	(1.240)	(0.108)	(0.0256)				
union	0.746	-0.501	-0.180	0.236	-13.91	-1.101	0.474
	(7.176)	(0.630)	(0.125)	(0.00311)	(7.065)	(0.615)	(0.153)
tenure	-0.102	-0.129	-0.0854	4.003	0.644	-0.0584	-0.105
	(0.588)	(0.0525)	(0.0124)	(0.0303)	(1.006)	(0.0887)	(0.0189)
hours	-3.142	-0.251	0.0643	36.82	-3.037	-0.263	0.0478
	(0.327)	(0.0274)	(0.00857)	(0.0698)	(0.552)	(0.0473)	(0.0115)
wks_work	-0.288	-0.0168	0.00848	63.29	-0.204	-0.0128	0.00510
	(0.103)	(0.00885)	(0.00225)	(0.208)	(0.105)	(0.00913)	(0.00218)
_cons	185.2	35.34	15.28		218.9	34.80	12.31
	(18.44)	(1.567)	(0.497)		(40.58)	(3.511)	(0.714)
N	18601	18601	18601	18601	18601	18601	18601
rifmean	263.8	36.31	9.894		263.8	36.31	9.894

Standard errors in parentheses

Table of Contents

RIF Decomposition: oaxaca_rif

- -rifhdreg- provides a simple framework for analyzing the impact of changes in the distribution of X's on distributional statistics, at the margin. (RIF's are Local linear approximations)
- Large changes in distributions require other methods, for example decomposition methods: Oaxaca Blinder
- The premise, allow for all distributions of X's to change between two groups.
- As long as the conditional independence assumptions holds, we can apply OB to decompose differences in statistics as functions of differences in characteristics, and differences in returns to those characteristics.

RIF Decomposition: oaxaca_rif

- The simple OB framework:

$$\Delta v = v(H_y) - v(F_y)$$

$$\Delta v = \beta_h \overline{X}_h - \beta_f \overline{X}_f$$

$$\Delta v = \beta_h(\overline{X}_h - \overline{X}_f) - (\beta_h - \beta_f)\overline{X}_f$$

  This assumes a linear counterfactual $v(CF_y) = \beta_h \overline{X}_f$

- A better counterfactual can be obtained using IPW ($\omega(x)$).

$$\Delta v = v\Big(\int (H_{Y|X} * dH_X)\Big) - v\Big(\int (F_{Y|X} * dF_X)\Big)$$

$$v(CF_y) = v\Big(\int (H_{Y|X} * dF_X)\Big) = v\Big(\int (H_{Y|X} * \omega(x) * dH_X)\Big) = \beta_c \overline{X}_c$$

RIF Decomposition: oaxaca_rif

- -oaxaca_rif- is a wrapper around -oaxaca- (Jann 2008) that implements these two types of decompositions.
- It basically estimates the appropriate RIFs, uses them as dependent variables, and re-arranges the results.
- the syntax
  `oaxaca_rif depvar indepvar [aw pw iw] [if in], by(var) rif(rifoptions) IPW_options oaxaca_options`
- Many features of -oaxaca- are kept.

Oaxaca_rif example

```
bootstrap:oaxaca_rif wage age grade tenure hours wks_work,
rif(mean) by(union) swap w(1)
```

```
bootstrap: oaxaca_rif wage age grade tenure hours wks_work,
rif(mean) by(union) rwlogit(age grade tenure hours
wks_work) swap w(1)
```

```
bootstrap:  oaxaca_rif wage age grade tenure hours wks_work,
rif(gini) by(union) rwlogit(age grade tenure hours
wks_work) scale(1000) swap w(1)
```

```
bootstrap:  oaxaca_rif wage age grade tenure hours wks_work,
rif(ucs(80)) by(union) rwlogit(age grade tenure hours
wks_work) scale(100) swap w(1)
```

Oaxaca_rif example

```
Model     : Blinder-Oaxaca RIF-decomposition
Type      : Standard
RIF       : mean
Scale     : 1
Group 1: union = 1                              N of obs 1      = 4382
Group c: x2*b1                                  N of obs C      =          .
Group 2: union = 0                              N of obs 2      = 14219
```

wage	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
overall						
group_1	7.611619	.0794412	95.81	0.000	7.455917	7.767321
group_2	6.174817	.0243625	253.46	0.000	6.127067	6.222566
difference	1.436802	.0808799	17.76	0.000	1.278281	1.595324
explained	.2467854	.0639477	3.86	0.000	.1214503	.3721205
unexplained	1.190017	.1289997	9.22	0.000	.9371822	1.442852

Oaxaca_rif example

```
Group 1: union = 1                                   N of obs 1      = 4382
Group c: X1~>rw~>X2                                  N of obs C      = 4382
Group 2: union = 0                                   N of obs 2      = 14219
```

wage	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
Overall						
Group_1	7.611619	.0684663	111.17	0.000	7.477427	7.74581
Group_c	7.296463	.1015878	71.82	0.000	7.097355	7.495572
Group_2	6.174817	.0243648	253.43	0.000	6.127062	6.222571
Tdifference	1.436802	.0766688	18.74	0.000	1.286534	1.587071
ToT_Explained	.3151556	.0540289	5.83	0.000	.2092609	.4210503
ToT_Unexplained	1.121647	.1086084	10.33	0.000	.9087782	1.334515
Explained						
Total	.3151556	.0540289	5.83	0.000	.2092609	.4210503
Pure_explained	.285553	.0510947	5.59	0.000	.1854092	.3856968
Specif_err	.0296026	.0075979	3.90	0.000	.014711	.0444942
Unexplained						
Total	1.121647	.1086084	10.33	0.000	.9087782	1.334515
Reweight_err	-.0235812	.0124732	-1.89	0.059	-.0480282	.0008659
Pure_Unexplained	1.145228	.1133489	10.10	0.000	.9230683	1.367388

Oaxaca_rif example

	(1) gini	(2) ucs80	(3) lor20	(4) q10	(5) q50	(6) q90
Overall						
Group_1	246.1	34.72	10.35	1.394	1.919	2.446
	(5.946)	(0.536)	(0.116)	(0.00860)	(0.00922)	(0.00749)
Group_c	261.5	36.17	10.19	1.337	1.848	2.407
	(9.477)	(0.920)	(0.168)	(0.0107)	(0.00987)	(0.0107)
Group_2	262.8	36.44	10.03	1.197	1.683	2.309
	(2.029)	(0.167)	(0.0575)	(0.00545)	(0.00359)	(0.00480)
Explained						
Total	-15.43	-1.446	0.161	0.0565	0.0710	0.0389
	(4.287)	(0.448)	(0.0758)	(0.00642)	(0.00825)	(0.00756)
Pure_explained	-16.27	-1.451	0.227	0.0587	0.0634	0.0296
	(3.761)	(0.395)	(0.0708)	(0.00474)	(0.00621)	(0.00603)
Specif_err	0.840	0.00470	-0.0659	-0.00220	0.00765	0.00924
	(0.686)	(0.0655)	(0.0174)	(0.00659)	(0.00413)	(0.00516)
Unexplained						
Total	-1.318	-0.273	0.161	0.140	0.166	0.0978
	(9.336)	(0.937)	(0.169)	(0.0122)	(0.00957)	(0.0114)
Reweight_err	-2.119	-0.174	0.0451	0.00106	-0.000364	-0.00454
	(0.705)	(0.0647)	(0.0139)	(0.00119)	(0.00157)	(0.00160)
Pure_Unexplained	0.801	-0.0997	0.115	0.139	0.166	0.102
	(9.394)	(0.931)	(0.168)	(0.0123)	(0.00959)	(0.0119)

Standard errors in parentheses

Table of Contents

Other Extensions

- Since its inception, early this year, a few other expansions have been added to this program. Some very recent.
- -rifhdreg- It allows for SVY. Specially useful for the estimation of standard errors of distributional statistics.
- -rifhdreg- adds "over". This may be used as a partial conditional RIF. Useful for standard errors across multiple groups.
- -rifhdreg- can now estimate effects similar IPWRA treatment effects, using rwlogit or rwprobit. This is similar to Firpo and Pinto (2016). Allows for att, ate and atu. Useful for analyzing Inequality treatment effects

Other Extensions

- -rifsureg-. This would the the equivalent to sqreg, but unconditional quantile regressions.
- Handy for making plots across quantiles.
- -rifsureg2- is similar to rifsureg, but allows to simultanously estimate RIF regressions for non colinear models.
- -uqreg- Stand alone command to estimate Unconditional Partial effects for UQR with alternative model specifications (logit/probit/other)

rifsureg: Example

```
rifsureg ln_wage age grade union tenure hours wks_work,
qnts(10(10)90)

margins, dydx(grade) nose

marginsplot
```

rifsureg: Example

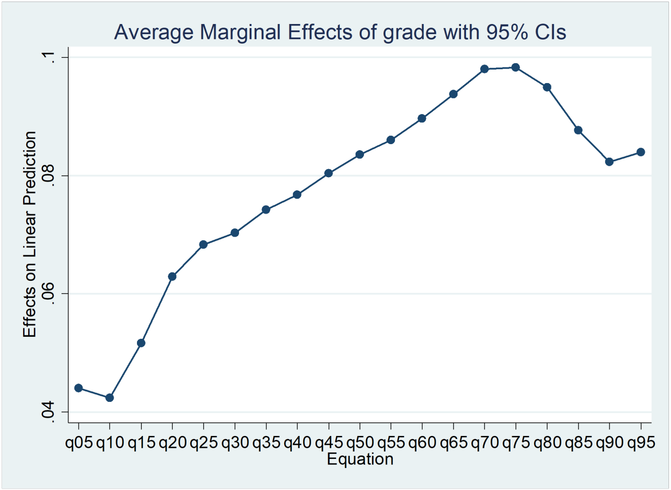


Table of Contents

Conclusions

- RIF and IF are powerful tools for analyzing and visualizing distributional statistics.
- The three main commands presented today (_grifvar, rifhdreg, oaxaca_rif) aim to facilitate the use of RIF's for this type of analysis
- Questions, comments and suggestions are welcome.
- Thank you
- Latest Files: https://bit.ly/2NFM3cH

References

Borgen, Nicolai T. 2016. "Fixed effects in Unconditional Quantile Regression." The Stata Journal no. 16 (2):403-415.

Chung, Choe, and Philippe Van Kerm. 2018. "Foreign Workers and the Wage Distribution: What Does the Influence Function Reveal." Econometrics no. 6 (2):28.

Cowell, F. and E. Flachaire (2007). 'Income Distribution and Inequality Measurement: The problem of Extreme Values.' Journal of Econometrics, 141(2):1044-1072.

Correira, Sergio. 2017. Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator. In Working Paper. Deville, J.-C. 1999, 'Variance estimation for complex statistics and estimators: linearization and residual techniques', Survey Methodology 25, 193–204

References

Essama-Nssah, B. and Lambert, P. J. (2012), Influence functions for policy impact analysis, in J. A. Bishop and R. Salas, eds, 'Inequality, Mobility and Segregation: Essays in Honor of Jacques Silber', Vol. 20 of Research on Economic Inequality, Emerald Group Publishing, chapter 6, pp. 135–159

Firpo, S., Fortin, N. M. and Lemieux, T. (2009), 'Unconditional quantile regressions', Econometrica 77(3), 953–973.

Firpo, S., Fortin, N. M. and Lemieux, T. (2018), 'Decomposing Wage Distributions using Recentered Influence Functions Regressions', Econometrics 6(41).

Firpo, S. P., and C. Pinto. (2016). "Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures." Journal of Applied Econometrics no. 31 (3):457-486

Jann, Ben. 2008. "The Blinder-Oaxaca decomposition for linear regression models." Stata Journal no. 8 (4):453-479.

References

Heckley, G., U.-G. Gerdtham, and G. Kjellsson. 2016. "A General Method for Decomposing the Causes of Socioeconomic Inequality in Health." Journal of Health Economics. 48:89-106.

Rios-Avila, F. 2019. "Recentered Influence Functions in Stata: Methods for Analyzing the Determinants of Poverty and Inequality" Working Paper 927. Levy Economics Institute.