RIF-regressions: tools for analyzing distributional statistics

Fernando Rios-Avila

Levy Economics Institute Bard College

Introduction

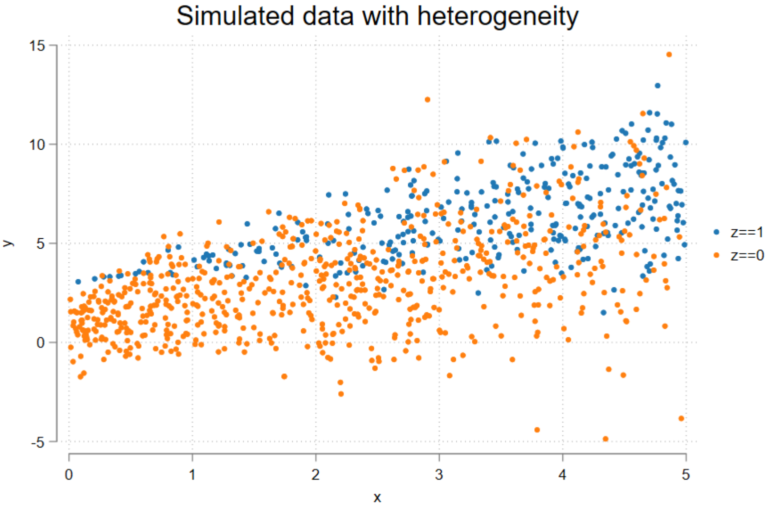
Unconditional quantile regressions (UQR) via RIF (Recentered Influence functions) was introduced by Firpo, Fortin, and Lemieux (2009) as a computationally simple strategy to estimate Unconditional partial effects on quantiles.

While Conditional QR, could be used to identify effects across conditional distributions (at the margin), UQR identifies effect on unconditional distributions.

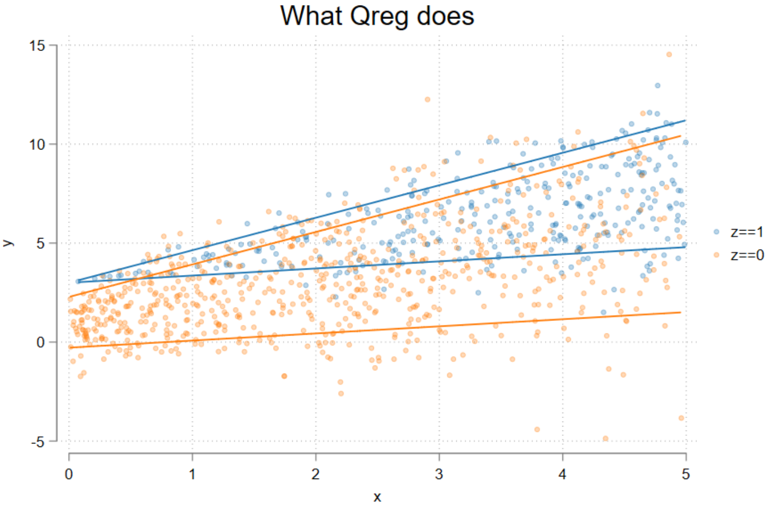
These effects, however, are different from unconditional treatment effects, which compares differences in statistics, across two (or more) distributions.

Since then, RIF-regressions have been used to analyze other statistics. See FFL(2018), Firpo and Pinto (2016), Chung and Vankerm (2018), Cowell and Flachaire (2007), Essama-Nssah and Lambert (2012) and Heckley et al (2016).

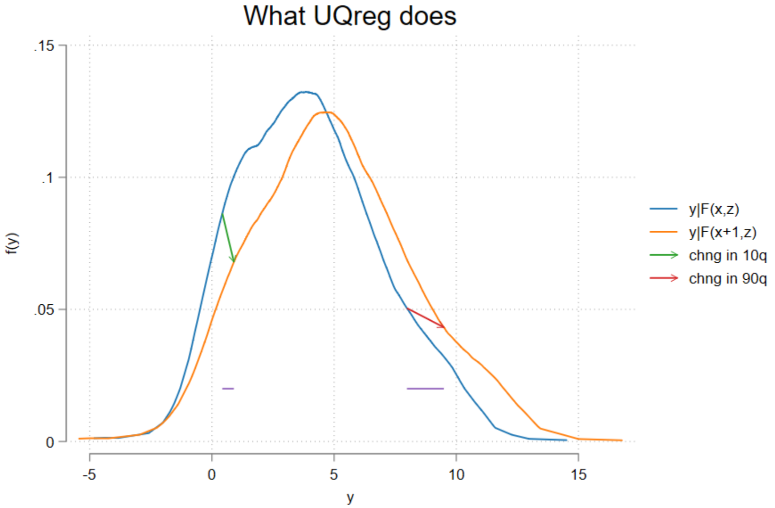
Simulated Data



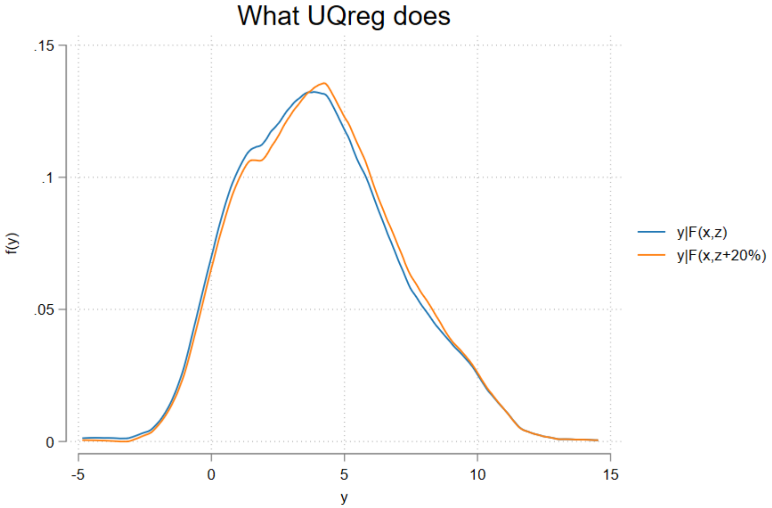
Contrasting QR methods



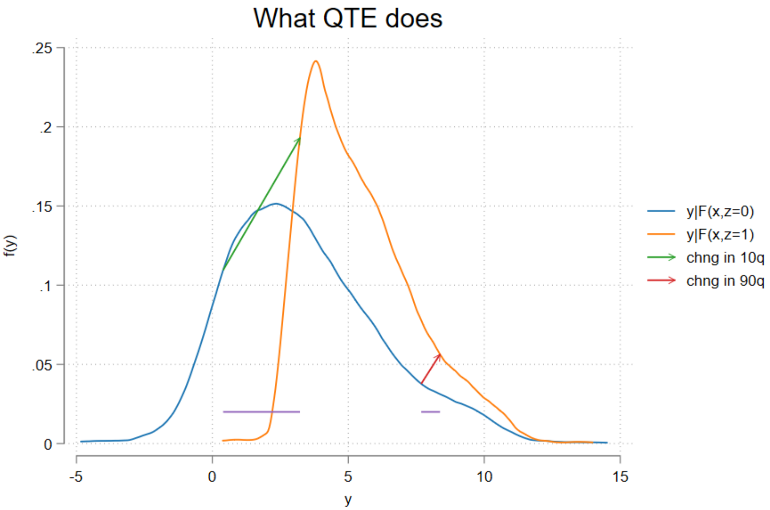
Contrasting QR methods



Contrasting QR methods



Contrasting QR methods



Building Blocks

To understand better what RIF-regressions do in general, it will be useful to set a unifying framework.

Assume that the outcome y is a function of observed x and unobserved characteristics e, such that.

$$y = g(x, e)$$

Thus, if we observe x and e, the conditional CDF is given by:

$$F(Y|X, e) = 1(Y \geq g(x, e))$$

Under the same framework, the unconditional distribution of y would be given by:

$$F(y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(Y|X, e)f(x, e)de\ dx$$

Building Blocks

It is important to notice that this unconditional distribution still "conditions", but on distributions, not specific values. ($F(Y)$ depends on $f(X, e)$).

Since we do not observe e, one alternative is to "integrate" over the unobserved factors e. However, to do so, we need to impose the exogeneity/independence assumption $f(x, e) = f(x|e)f(e) = f(x)f(e)$.

This is an stronger assumption than the zero conditional mean $E(e|X) = 0$ and Homoscedasticity assumption $Var(e|X) = c$ combined. However under this assumption:

$$F(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(Y|X, e)f(x)f(e)de\, dx$$

So we can integrate over the error

$$F(y) = \int_{-\infty}^{\infty} F(Y|X)dF(x)$$

Building Blocks

Implicit assumption: Conditional distributions of Y are fixed $F(Y|X)$. The only things that change are characteristics $F(x)$.

This has parallels with standard linear regressions:

$$F(y) = \int_{-\infty}^{\infty} F(Y|X) dF(x)$$

$$y_i = b_0 + b_1 * x_i + e_i$$

This does not mean $F(y|X)$ is constant across X, but that it is fixed to changes in X.

Side Question: How do we simulate changes in $dF(x)$?

- Full distribution simulations.
- Local simulations (swapping one observation at a time)
- Reweighting $dG(x) = w(x) * dF(x) \quad \forall x \in R$

Building Blocks

How do we analyze a the change in $F(y)$ caused by a change in $F(x)$?

As stated, the change in Y can be analyzed graphically. (how the overall distribution changes).

However, this may be impractical.

- You need separate simulations for each variable in the model.
- Or different Weighting factors.
- Or local simulations (may be simplest to implement).

The alternative is to use a single statistic that summarizes the distribution of y, and focus on how that summary statistic changes when $F(x)$ changes.

RIF-Recentered Influence Function

Call this summary (or distributional) statistic v. This is a function that depends on ALL values of y or on $f(y)$ or $F(y)$
Simplest example:

$$Mean : \mu_y = v(F(y)) = \int_{-\infty}^{\infty} yf(y)dy$$

This would be the unconditional mean, because it looks over the whole distribution. However, we can also write it as conditional with respect to the distribution of x

$$Mean : \mu_y = v(F(y)) = \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f(y|x)f(x)dxdy$$

$$\mu_y = v(F(y)) = \int_{-\infty}^{\infty} E(y|X)f(x)dx$$

RIF-Recentered Influence Function

How v changes when there is a change in $f(x)$?

Long chain of events $\Delta f(x) \rightarrow \Delta f(y) \rightarrow \Delta \mu_y$

Call F(y) original distribution and G(y) the distribution after $F(x)$ changes. The effect that this has on μ_y can be measured as

$$\Delta \mu_y = v(G(y)) - v(F(y))$$

However the "influence" of this effect should be a standardized by the change in the distribution itself.

$$\frac{\Delta \mu_y}{\Delta F(y)} = Influence(v, F(y)) = \frac{v(G(y)) - v(F(y))}{||G(y) - F(y)||}$$

And this bring us the first definition of Influence function:

$$IF(v, F(y)) = lim_{G(y) \rightarrow F(y)} \frac{v(G(y)) - v(F(y))}{||G(y) - F(y)||}$$

RIF-Recentered Influence Function

$$IF(v, F(y)) = lim_{G(y) \rightarrow F(y)} \frac{v(G(y)) - v(F(y))}{||G(y) - F(y)||}$$

This expression corresponds to a directional derivative of a functional (the distributional statistic). The Rate of changes in the statistic with respect to changes in the distribution. In other words, the IF measure the "standardized" effect on the statistic v, at the "margin". Even more technical. What is $G(y)$

$$G(y) = \epsilon * 1(y \geq y_i) + (1 - \epsilon) * F(y)$$

where $\epsilon * 1(y \geq y_i)$ is a "contamination" error on the distribution of $F(y)$. Thus

$$IF(y_i, F(y), v) = lim_{\epsilon \rightarrow 0} \frac{v(G(y)) - v(F(y))}{\epsilon}$$

RIF-Recentered Influence Function

Consider the simplest case. The mean. How would an additional observation affect the unconditional mean?

$$\text{if} \quad y_i > \mu_y \rightarrow \Delta\mu_y > 0$$

$$\text{if} \quad y_i < \mu_y \rightarrow \Delta\mu_y < 0$$

In fact, the impact on μ_y will be:

$$\Delta\mu_y = (y_i - \mu_y) * \frac{1}{N+1}$$

But the IF will be:

$$IF(y_i, \mu_y) = y_i - \mu_y$$

FFL (2009) introduce at this point the idea of the Re-centered IF

$$RIF(y_i, \mu_y) = \mu_y + IF(y_i, \mu_y) = y_i$$

RIF-Recentered Influence Function

From the technical point of view, The RIF statistic can be thought as a first order approximation or linearization of any functional v. Or the consequence of a Taylor expansion.

$$v(F(y)) \approx v(F_0(y)) + \sum \left(\frac{\partial v(.)}{\partial F(Y)} \right) * \Delta F(Y)$$

$$v(F(y)) \approx v(F_0(y)) + \frac{1}{N} \sum IF(i_{th}, v, F(y))$$

$$v(F(y)) \approx \frac{1}{N} \sum RIF(i_{th}, v, F(y))$$

From here, two properties can be derived: (Deville, 1999).

$$E(RIF(i_{th}, v, F)) = v(F(y))$$

$$Var(E(RIF(i_{th}, v, F))) = Var(v(F(y)))$$

RIF-Recentered Influence Function Regression

Reconsider the main question. We want to measure how the distribution of Y changes when the distribution of X changes:

$$F(y) = \int_{-\infty}^{\infty} F(Y|X)dF(x)$$

However, instead of focusing on the overall distribution of Y, we can concentrate on how the unconditional mean μ_y changes when $F(X)$ changes.

$$v(\mu_y, F(y)) = v\left(\mu_y, \int_{-\infty}^{\infty} F(Y|X)dF(x)\right)$$

On the left side of the equation, we can proxy the effects on the unconditional mean (or any other statistic) by using the RIF, which is the value itself.

On the right side of the equation we face the same problem. However, we can capture changes in the distribution of X also using RIFs, and first order approximations. (although higher order may also be important)

$$RIF(i_{th}, v, F(y)) = a_0 + a_1 * RIF\left(i_{th}, \mu, F(x)\right) + \varepsilon_i$$

And for the unconditional mean, and a linear approximation we have:

$$y_i = a_0 + a_1 * x_i + \varepsilon_i$$

Remarks:

- The dependent variable can be the RIF for any statistic. Including unconditional quantiles.
- For independent variables, one can use either higher order approximations (polynomial or interactions)
- Or use RIFs for other moments of the distribution, like Variance, covariances and Kurtosis. (centered polynomials)

The most common approach is, of course, simply use polynomials or interactions of explanatory variables.

RIF-regressions, Estimation

RIF-regressions can be estimated using any of the battery of methodologies we are already familiar with.

The estimation basically involves 3 steps.

- Define sample of interest.
- Estimate the RIF for the dependent variable and distributional Statistic of interest.
- Estimate a model using the RIF as dependent variable using flexible specification. OLS, logit, poisson.

The goal is to use a model that may best capture the linear or nonlinear relationship between the RIF of the distributional statistic, and RIF of explanatory variables.

FFL(2009), for example, proposes that RIF-UQR can be estimated via OLS, but that given the nature of the Quantile RIF, probit or logit models could also be used.

Other methods could be potentially used as well.

RIF-regressions, Estimation

In Stata, there are many commands that allows you to implement RIF-OLS regressions. You have `rifreg`, `rifireg`, `rifhdreg`.

My preference, is the use of `rifhdreg`, as it handles many RIF's, factor notation, weights, and even allows you to estimate QTE via RIF.

```
Syntax:
rifhdreg depvar [indepvar] [if in] [weight], [reg options] rif(rif options)
[over(overvar) rwlogit( indepvar)]
```

The most important option is `rif()`, where you indicate which statistic you are interested in analyzing.

Options `over()` and `rwlogit()` can also be used to estimate treatment effects under exogeneity. SE need correction.

RIF-regressions, Interpretation

While RIF-regressions produce an output similar to the standard LR model, the interpretation requires some considerations.

Recall the model we are trying to estimate:

$$RIF_i(v, F_y) = a_0 + a_1 * x_i + e_i$$

The standard procedure for the estimation of marginal effects requires us to obtain the conditional mean, of this equation, and derive it, with respect to the variable of interest.

$$E(RIF_i|X) = a_0 + a_1 * X \quad \rightarrow \quad \frac{\partial E(RIF_i|X)}{\partial X} = a_1$$

With the exception of the mean (and poverty), this conditional expectation is meaningless, because:

$$E(RIF_i(v, F_y)|X = x) \neq E(RIF_i(v, F_{y|X=x})) = v(y|X = x)$$

RIF-regressions, Interpretation

A better way of thinking about this is of taking unconditional expectations of the whole equation.

$$E(RIF_i) = v(y) = a_0 + a_1 * E(X) \quad \rightarrow \quad \frac{\partial E(RIF_i)}{\partial E(X)} = a_1$$

In this case, we are measuring how will the statistic v would change if there is a marginal change in $E(X)$. This is, again, a location shift in the distribution in x, or simply, if everybody in the sample experiences a 1 unit increase in x.

FFL(2009) call this the Unconditional partial effect. An effect, at the margin, of how a general increase in x would affect the distributional statistic v.

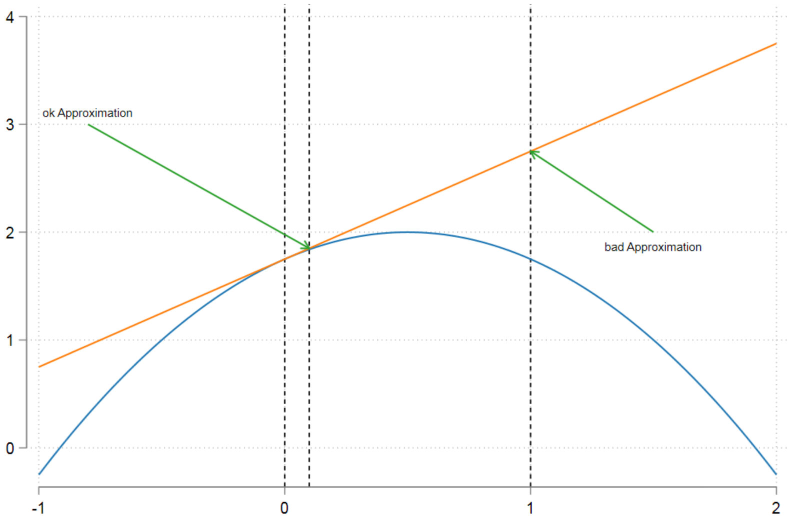
IE. If X is correlated with an increase in the RIF of an observation, then We should expect the overall statistic to increase as well.

RIF-regressions, Interpretation

When your dependent variable is binary or with a limited range (0-5), particular care is needed.

- RIF-reg are useful for obtaining effects at the margin (good approximations for small local changes).
- a 1 unit change for dummies, or for dependent variables with limited ranges may be too large.
- This means, the approximation may not be appropriate.
- Similar to making linear extrapolations in nonlinear models, or non linear explanatory variables.

RIF-regressions, Interpretation



Example:Analyzing Inequality in South Africa

```
. rifmean income_pcp , rif(q(10), q(50), q(90), gini, lor(40), ucs(90) )

Mean estimation                          Number of obs = 88,906

-----------------------------------------------------------------
                 |       Mean   Std. err.     [95% conf. interval]
-----------------+-----------------------------------------------
rif_income_pcp_1 |   3204.528    20.5559      3164.239    3244.818
rif_income_pcp_2 |   11382.28   53.54245      11277.34    11487.23
rif_income_pcp_3 |   72454.33    602.472      71273.49    73635.17
rif_income_pcp_4 |   .6537654   .0016039      .6506218    .6569091
rif_income_pcp_5 |   .0639714   .0004294      .0631298    .0648129
rif_income_pcp_6 |   .5341139   .0022211      .5297606    .5384672
-----------------------------------------------------------------
```

Example:Analyzing Inequality in South Africa

```
. rifmean expenditure_pcp , rif(q(10), q(50), q(90) , gini, lor(40), ucs(90))

Mean estimation                                  Number of obs = 88,906

----------------------------------------------------------------------
                      |       Mean   Std. err.     [95% conf. interval]
----------------------+-----------------------------------------------
rif_expenditure_pcp_1 |   3275.501    17.1785      3241.831    3309.171
rif_expenditure_pcp_2 |   9907.519   40.46346      9828.211    9986.827
rif_expenditure_pcp_3 |   51069.12   413.4056      50258.85    51879.39
rif_expenditure_pcp_4 |   .6165029   .0017325      .6131073    .6198985
rif_expenditure_pcp_5 |   .0814099   .0005105      .0804094    .0824104
rif_expenditure_pcp_6 |   .5079673   .0022486        .50356    .5123745
----------------------------------------------------------------------
```

Example:Analyzing Inequality in South Africa

Income	(1) q(10)		(2) q(90)		(3) gini		(4) m4	
Female	-0.060	(0.001)	-0.131	(0.000)	-0.274	(0.620)	0.411	(0.000)
Secondary S in~e	0.241	(0.000)	0.315	(0.000)	-5.119	(0.000)	0.318	(0.000)
Secondary S Co~e	0.429	(0.000)	1.725	(0.000)	-0.467	(0.381)	0.177	(0.000)
College+	0.262	(0.000)	4.988	(0.000)	54.652	(0.000)	0.053	(0.000)
age_hh	0.027	(0.000)	0.021	(0.000)	-0.075	(0.000)	53.255	(0.000)
Household size	-0.026	(0.000)	-0.041	(0.000)	0.020	(0.765)	6.002	(0.000)
couple=1	0.053	(0.003)	0.256	(0.000)	0.582	(0.292)	0.554	(0.000)
sh_nchild05	-0.005	(0.000)	-0.017	(0.000)	0.011	(0.422)	12.820	(0.000)
sh_nchild615	-0.005	(0.000)	-0.011	(0.000)	0.010	(0.384)	17.998	(0.000)
sh_wrk_wmen	0.005	(0.000)	0.009	(0.000)	-0.020	(0.000)	34.069	(0.000)
sh_wrk_men	0.008	(0.000)	0.007	(0.000)	-0.081	(0.000)	41.933	(0.000)
Constant	6.331	(0.000)	9.390	(0.000)	71.035	(0.000)		
Observations	68548		68548		68579		68579	
rifmean	8.079		11.192		65.377			

p-values in parentheses

Example:Analyzing Inequality in South Africa

Expenditure	(1) q(10)		(2) q(90)		(3) gini		(4) m4	
Female	-0.081	(0.000)	-0.133	(0.000)	-0.571	(0.339)	0.411	(0.000)
Secondary S in~e	0.411	(0.000)	0.334	(0.000)	-5.479	(0.000)	0.318	(0.000)
Secondary S Co~e	0.484	(0.000)	1.633	(0.000)	1.775	(0.002)	0.177	(0.000)
College+	0.409	(0.000)	4.576	(0.000)	57.754	(0.000)	0.053	(0.000)
age_hh	0.010	(0.000)	0.019	(0.000)	0.094	(0.000)	53.255	(0.000)
Household size	-0.096	(0.000)	-0.071	(0.000)	0.524	(0.000)	6.002	(0.000)
couple=1	-0.002	(0.902)	0.310	(0.000)	2.667	(0.000)	0.554	(0.000)
sh_nchild05	-0.005	(0.000)	-0.012	(0.000)	-0.080	(0.000)	12.820	(0.000)
sh_nchild615	-0.003	(0.000)	-0.010	(0.000)	-0.065	(0.000)	17.998	(0.000)
sh_wrk_wmen	0.003	(0.000)	0.006	(0.000)	0.003	(0.516)	34.069	(0.000)
sh_wrk_men	0.003	(0.000)	0.003	(0.000)	-0.016	(0.001)	41.933	(0.000)
Constant	7.824	(0.000)	9.501	(0.000)	53.635	(0.000)		
Observations	68579		68579		68579		68579	
rifmean	8.094		10.841		61.650			

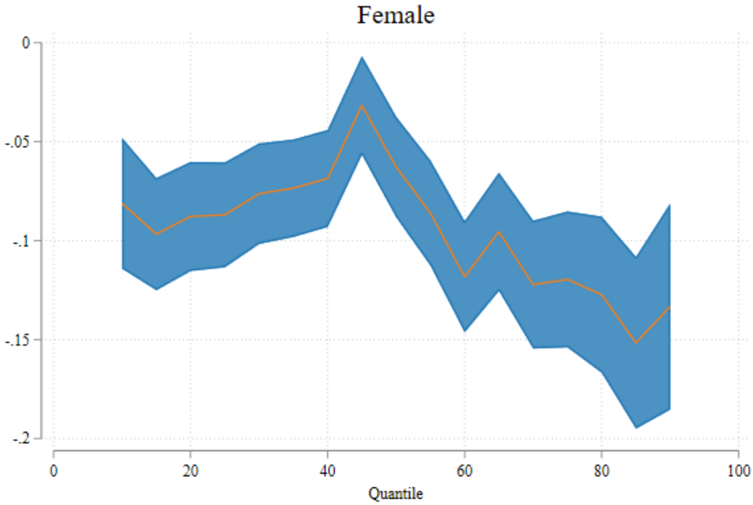
Example:Analyzing Inequality in South Africa

Treatment effects of Sex (HH) on Expenditure

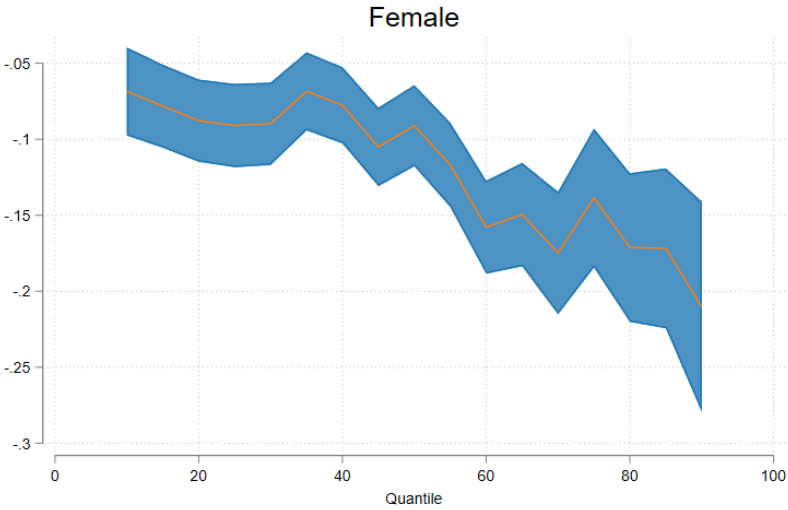
	(1) q(10)		(2) q(90)		(3) gini		(4) m4	
Female hh	-0.069	(0.000)	-0.210	(0.000)	-1.907	(0.024)	0.411	(0.000)
Secondary S in~e	0.404	(0.000)	0.257	(0.000)	-6.561	(0.000)	0.318	(0.000)
Secondary S Co~e	0.445	(0.000)	1.992	(0.000)	6.615	(0.000)	0.177	(0.000)
College+	0.435	(0.000)	3.996	(0.000)	59.089	(0.000)	0.053	(0.000)
age_hh	0.008	(0.000)	0.019	(0.000)	0.043	(0.148)	53.255	(0.000)
Household size	-0.069	(0.000)	-0.063	(0.000)	0.468	(0.000)	6.002	(0.000)
couple=1	0.031	(0.049)	0.313	(0.000)	2.824	(0.000)	0.554	(0.000)
sh_nchild05	-0.005	(0.000)	-0.017	(0.000)	-0.149	(0.000)	12.820	(0.000)
sh_nchild615	-0.003	(0.000)	-0.014	(0.000)	-0.118	(0.000)	17.998	(0.000)
sh_wrk_wmen	0.002	(0.000)	0.005	(0.000)	0.015	(0.215)	34.069	(0.000)
sh_wrk_men	0.002	(0.000)	0.004	(0.000)	-0.003	(0.717)	41.933	(0.000)
Constant	7.851	(0.000)	9.550	(0.000)	56.645	(0.000)		
Observations	68579		68579		68579		68579	
rifmean	8.125		10.756		60.452			

p-values in parentheses

Example, UPE of Female HH



Example, QTE of Female HH



Example:Analyzing Education, health and Life SF in South Africa

Population 15-40: Erreygers's and Wagstaff Indices

	Mean	Std. err.	[95% conf.	interval]
rif_educ_status_1	.207547	.0028534	.2019543	.2131398
rif_educ_status_2	.2125617	.0028988	.20688	.2182434

	Mean	Std. err.	[95% conf.	interval]
rif_health_1	.0662624	.0029254	.0605286	.0719962
rif_health_2	.071507	.003162	.0653095	.0777045

	Mean	Std. err.	[95% conf.	interval]
rif_life_satis_1	.1883218	.0031558	.1821364	.1945073
rif_life_satis_2	.2020434	.003374	.1954302	.2086566

Example:Analyzing Education, health and Life SF in South Africa

	(1) EDUC	(2) Health	(3) Life_Sa~s	(4) m4
Male	0.000	0.000	0.000	0.489*
Female	-1.276*	0.846	0.200	0.511*
Age of each ho~r	0.675*	0.204*	0.072	28.297*
Female hh	-1.783	0.309	1.045	0.411*
Secondary S in~e	-10.140*	-1.752*	0.110	0.318*
Secondary S Co~e	2.204*	4.549*	9.663*	0.177*
College+	51.266*	11.328*	43.398*	0.053*
age_hh	-0.090*	0.021	0.037	53.255*
Household size	-0.219	-0.556*	-0.083	6.002*
couple=1	0.876	1.109	4.481*	0.554*
sh_nchild05	0.101*	-0.055*	-0.014	12.820*
sh_nchild615	0.016	0.030	-0.004	17.998*
sh_wrk_wmen	0.062*	0.018*	0.042*	34.069*
sh_wrk_men	0.032*	0.009	0.029*	41.933*
Constant	4.991*	0.507	6.574*	
Observations	28311	28329	27624	68579
rifmean	20.755	6.626	18.832	

* $p<0.05$

Example:Analyzing Education, health and Life SF in South Africa

TE	(1) EDUC	(2) Health	(3) Life_Sa~s	(4) m4
Male	0.000	0.000	0.000	0.489*
Female	-0.658	0.713	0.213	0.511*
Age of each ho~r	0.726*	0.156*	0.030	28.297*
Female hh	-2.223*	-0.146	0.962	0.411*
Secondary S in~e	-9.550*	-1.674*	0.057	0.318*
Secondary S Co~e	5.219*	4.966*	8.880*	0.177*
College+	58.526*	11.731*	41.566*	0.053*
age_hh	-0.052	0.021	0.043	53.255*
Household size	-0.158	-0.575*	0.060	6.002*
couple=1	0.263	1.025	4.353*	0.554*
sh_nchild05	0.100*	-0.052	-0.004	12.820*
sh_nchild615	0.002	0.029	0.007	17.998*
sh_wrk_wmen	0.073*	0.017	0.044*	34.069*
sh_wrk_men	0.022*	0.015	0.035*	41.933*
Constant	0.956	1.737	6.416*	
Observations	28311	28329	27624	68579
rifmean	20.601	6.210	19.407	

* $p<0.05$

Conclusions

- In this presentation, I provided a general review of RIF regressions theory and estimation.
- RIF, by default estimates effects at the margin (UPE). But can be used to estimate distributional Effects
- IPW can be combined with RIF to estimate Distributional TE. But Standard errors need correction
- Their application is straight forward with the commands rifhdreg and oaxaca_rif

Thank you

Questions? friosavi@levy.org

Material: https://tinyurl.com/rifsa21

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. 2020. "Recentered Influence Functions in Stata: RIF regressions and RIF Decomposition" The Stata Journal, 20(1).