Decomposition methods in the social sciences

Fall semester 2019, Monday 14-16, Fabrikstrasse 8, B 306
(Exercises in PC Lab, B 003)

Ben Jann

University of Bern, Institut of Sociology

Further approaches

Beyond the mean

- The discussed Oaxaca-Blinder procedures and their extensions to non-linear models focus on the decomposition of differences in the expected value (mean) of an outcome variable.
- In many cases, however, one is interested in other distributional statistics, say the Gini coefficient or the D9/D1 quantile ratio, or even in whole distributions (density curves, Lorenz curves).
- The basic setup is the same; an estimate of $F_{Y^g|G\neq g}$ is needed to be able to compute a decomposition such as

$$\begin{aligned}
\Delta^\nu &= \nu\big(F_{Y|G=0}\big) - \nu\big(F_{Y|G=1}\big) \\
&= \big\{\nu\big(F_{Y|G=0}\big) - \nu\big(F_{Y^0|G=1}\big)\big\} + \big\{\nu\big(F_{Y^0|G=1}\big) - \nu\big(F_{Y|G=1}\big)\big\} \\
&= \Delta^\nu_X + \Delta^\nu_S
\end{aligned}$$

where

$$F_{Y^g|G\neq g}(y) = \int F_{Y|X,G=g}(y|x) f_{X|G\neq g}(x)\, dx$$

Beyond the mean

- Several approaches have been proposed in the literature:
  - Estimating $F_{Y^g|G\neq g}$ by reweighting (DiNardo et al. 1996).
  - Imputing values for Y^g in group $G \neq g$
    - based on regression residuals (Juhn et al. 1993)
    - based on quantile regression (Machado and Mata [illegible], Melly 2005, 2006)
  - Estimating $F_{Y^g|G\neq g}$ by distribution regression (Chernozhukov et al. 2013)
  - Estimating $\nu(F_{Y^g|G\neq g})$ via recentered influence function regression (Firpo et al. 2007, 2009)
- Last time we looked at reweighting, today we will do the rest.

Contents

JMP 1993

- The goal is to "impute" counterfactual outcomes at the individual level, i.e. to answer, for example, for each women in the sample how much she would earn if she was paid like a man.
- If such counterfactual individual-level outcomes can be generated in a "realistic" way, then we can compute decompositions for arbitrary distributional statistics, by comparing the distribution of counterfactual outcomes with distributions of observed outcomes.
- JMP propose a procedure for generating the counterfactual outcomes that makes use of residuals from regression models.

JMP 1993

- Assume that an additive linear model

$$Y_i = X_i\beta^g + \upsilon_i = X_i\beta^g + h^g(\epsilon_i)$$

can be used to describe Y in group g. Think of β^g "returns to observables" and $h^g()$ as "returns to unobservables".

- We can now construct counterfactual outcomes for group 1.
- JMP propose to do this in two steps.
  - In the first step, impute residuals based on the group 0 residual distribution:

$$Y_i^{C1} = X_i\beta^1 + \upsilon_i^C \quad \text{for each } i \text{ in group 1}$$

  - In the second step, also adjust the "returns to observables":

$$Y_i^{C2} = X_i\beta^0 + \upsilon_i^C \quad \text{for each } i \text{ in group 1}$$

JMP 1993

- We can then compute a decomposition as

$$
\begin{aligned}
\Delta^{\nu} &= \nu\left(F_{Y|G=0}\right) - \nu\left(F_{Y|G=1}\right) \\
&= \left\{\nu\left(F_{Y|G=0}\right) - \nu\left(F_{Y^{C2}|G=1}\right)\right\} \\
&\quad + \left\{\nu\left(F_{Y^{C2}|G=1}\right) - \nu\left(F_{Y^{C1}|G=1}\right)\right\} \\
&\quad + \left\{\nu\left(F_{Y^{C1}|G=1}\right) - \nu\left(F_{Y|G=1}\right)\right\} \\
&= \Delta_X^{\nu} + \Delta_\beta^{\nu} + \Delta_\upsilon^{\nu}
\end{aligned}
$$

where

Δ_X^{ν}	part due to differential composition of observables
Δ_β^{ν}	part due to differential returns of observables
Δ_υ^{ν}	part due to differential returns and composition of unobservables

JMP 1993

- The question is how to impute υ
- Let $\tau_i = F_{\upsilon|G=1}(\upsilon_i)$ be the rank of the residual of observation i in the residual distribution of group 1.
- The proposal by JMP is then to set υ_i^C to quantile τ_i from the residual distribution of group 0:

$$\upsilon_i^C = F^{-1}_{\upsilon|G=0}(\tau_i)$$

- The procedure makes a very strong assumption: the residuals are independent of X (e.g. no heteroscedasticity). A much better approach would be to use conditional ranks given X, but it is unclear how to implement this in practice.
- Stata implementation: `ssc install jmpierce`

Example

```
. use gsoep29, clear
(BCPGEN: Nov 12, 2013 17:15:52-251 DBV29)
. // selection
. generate age = 2012 - bcgeburt
. keep if inrange(age, 25, 55)
(10,780 observations deleted)
. // compute gross wages and ln(wage)
. generate wage = labgro12 / (bctatzeit * 4.3) if labgro12>0 & bctatzeit>0
(1,936 missing values generated)
. generate lnwage = ln(wage)
(1,936 missing values generated)
. // X variables
. generate schooling = bcbilzeit if bcbilzeit>0
(318 missing values generated)
. generate ft_experience = expft12 if expft12>=0
(15 missing values generated)
. generate ft_experience2 = expft12^2 if expft12>=0
(15 missing values generated)
. // summarize
. summarize wage lnwage schooling ft_experience ft_experience2 bcsex

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |      8,090    16.26903    15.21083   .3624283   914.7287
      lnwage |      8,090    2.615219    .5944705  -1.014929   6.818627
   schooling |      9,708    12.76118     2.73677          7         18
ft_experie~e |     10,011    13.41052    10.03473          0         39
ft_experie~2 |     10,011    280.5277    324.8873          0       1521
-------------+---------------------------------------------------------
       bcsex |     10,026    1.539896    .4984306          1          2
. drop if missing(lnwage,schooling,ft_experience,bcsex)
(2,166 observations deleted)
```

Example

```
. regress lnwage schooling ft_experience ft_experience2 if bcsex==1
  (output omitted)
. estimates store male
. regress lnwage schooling ft_experience ft_experience2 if bcsex==2
  (output omitted)
. estimates store female
. jmpierce male female, reference(1) statistics(mean p10 median p90)
Juhn-Murphy-Pierce decomposition (reference estimates: male)
                T           Q           P           U
  mean   .2505696   .14842295    .1013223   .00082434
   p10  .26098967   .17473984   .06000555   .02624428
median  .24613309   .15090537   .10408449  -.00885677
   p90  .25770116   .12742758   .14843249  -.01815891
T = Total difference (male-female)
Q = Contribution of differences in observable quantities
P = Contribution of differences in observable prices
U = Contribution of differences in unobservable quantities and prices
```

Approach based on conditional quantiles

- The JMP decomposition, at least if based in *unconditional* residual ranks, is not very convincing due to its simplifying assumptions.
- An approach that is much more data-driven has been suggested by Machado and Mata (2005) (MM).
- The basic idea is to impute Y^C by inverting the conditional distribution of Y from the other group:

$$Y_i^C = F^{-1}_{Y|X,G=0}(F_{Y|X,G=1}(Y_i|X_i), X_i)$$

- $F^{-1}_{Y|X,G=0}(\tau, X)$ can be estimated by quantile regression:

$$F^{-1}_{Y|X,G=0}(\tau, X) = Q^0_\tau(Y|X) = X\beta^0_\tau$$

Approach based on conditional quantiles

- Because $\tau(Y|X) = F_{Y|X}(Y|X)$ follows a uniform distribution, MM suggest a simulation procedure, where values for τ are drawn from a uniform distribution.
    1. Draw values τ_j, $j = 1, \ldots, J$, from $U(0,1)$.
    2. For each j
        - Estimate quantile regression for τ_j in group 0:
          $$F^{-1}_{Y|X,G=0}(\tau_j, X) = Q^0_{\tau_j}(Y|X) = X\beta^0_{\tau_j}$$
        - Estimate quantile regression for τ_j in group 1:
          $$F^{-1}_{Y|X,G=1}(\tau_j, X) = Q^1_{\tau_j}(Y|X) = X\beta^1_{\tau_j}$$
        - Draw a single observation j from group 1 and predict
          $$Y^C_j = X_j\beta^0_{\tau_j} \quad \text{and} \quad \hat{Y}_j = X_j\beta^1_{\tau_j}$$
    3. Compute the decomposition by comparing Y^C and $\hat{Y}$:
       $$\Delta^\nu_S = \nu(F_{Y^C}) - \nu\big(F_{\hat{Y}}\big)$$
       $$\Delta^\nu_X = \Delta^\nu - \Delta^\nu_S$$

Approach based on conditional quantiles

- As Melly (2005, 2006) shows, the simulation procedure proposed by MM is more complicated then necessary.
- An equivalent but much more efficient approach is to compute quantile regressions in group 0 over a regular grid of τ values (e.g., 99 quantile regressions from $\tau_1 = 0.01$ to $\tau_J = 0.99$), then derive the conditional distribution $F_{Y|X,G=0}$ from these quantile regressions, and then obtain the counterfactual marginal distribution of Y^C by integrating the conditional distribution over the group 1 sample (see Melly 2006 for details).
- Stata implementation of the variant proposed by Melly:
  - `net install rqdeco, from("https://sites.google.com/site/mellyblaise/")`

Example

```
. generate byte female = bcsex==2 if bcsex<.

. rqdeco lnwage schooling ft_experience ft_experience2, by(female) ///
>     quantiles(.1 .5 .9) vce(bootstrap)
Fitting base model
(bootstrapping ..................................................)

(BCPGEN: Nov 12, 2013 17:15:52-251 DBV29)

Decomposition of differences in distribution using quantile regression
         Total number of observations                          7860
           Number of observations in group 0                   3877
           Number of observations in group 1                   3983

         Number of quantile regressions estimated               100

The variance has been estimated by bootstraping the results 50 times
```

Component	Effects	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Quantile .1						
Raw difference	-.262877	.013866	-18.96	0.000	-.290053	-.235701
Characteristics	-.201693	.024503	-8.23	0.000	-.249718	-.153668
Coefficients	-.061184	.013065	-4.68	0.000	-.086792	-.035576
Quantile .5						
Raw difference	-.242743	.007749	-31.33	0.000	-.257931	-.227555
Characteristics	-.13804	.010884	-12.68	0.000	-.159372	-.116709
Coefficients	-.104702	.007858	-13.32	0.000	-.120104	-.0893
Quantile .9						
Raw difference	-.252084	.013004	-19.39	0.000	-.277571	-.226596
Characteristics	-.11231	.012991	-8.64	0.000	-.137772	-.086847
Coefficients	-.139774	.011474	-12.18	0.000	-.162263	-.117285

Approach based distribution regression

- As Chernozhukov et al. (2013) show, the conditional distribution $F_{Y|X}$ can also be estimated directly by what they call "distribution regression".
- The idea is to estimate a separate model for each value of Y (or, e.g., for a grid of Y values) in group 0:

$$F(y|X, G = 0) = \Lambda(X\beta^y)$$

where [illegible]) is a suitable link function. A simple example is to use the logistic function. In this case, β^y is estimated by running a logit model of $I(Y_i \leq y)$ on X in group 0.

Approach based distribution regression

- We can then estimate the counterfactual (marginal) distribution for group 1 by averaging over predictions from these models

$$F_{Y^C}(y) = \frac{1}{N^1} \sum_{i:G=1} \Lambda(X_i \beta^y)$$

and compute whatever statistic we are interested in to obtain the decomposition (e.g. specific quantiles by inverting F_{Y^C}), with

$$\Delta_S^\nu = \nu(F_{Y^C}) - \nu\left(F_{Y|G=1}\right)$$
$$\Delta_X^\nu = \Delta^\nu - \Delta_S^\nu$$

- Stata implementation:
  - `net install counterfactual, from("https://sites.google.com/site/mellyblaise/")`

Example

```
. generate byte female = bcsex==2 if bcsex<.
. cdeco lnwage schooling ft_experience ft_experience2, group(female) ///
>     quantiles(.1 .5 .9) method(logit)
(bootstrapping ..........................................................................
> ........................)
Conditional model                         logit
Number of regressions estimated           98
The variance has been estimated by bootstraping the results 100 times.
No. of obs. in the reference group        3877
No. of obs. in the counterfactual group   3983
```

Differences between the observable distributions (based on the conditional model)

Quantile	Quantile effect	Pointwise Std. Err.	Pointwise [95% Conf.	Interval]	Functional [95% Conf.	Interval]
.1	.26216	.025637	.211913	.312408	.197426	.326895
.5	.241162	.014198	.213335	.268989	.205312	.277012
.9	.262364	.017729	.227615	.297113	.217597	.307132

Effects of characteristics

Quantile	Quantile effect	Pointwise Std. Err.	Pointwise [95% Conf.	Interval]	Functional [95% Conf.	Interval]
.1	.156821	.035838	.086579	.227063	.049855	.263787
.5	.122898	.012913	.097589	.148207	.084357	.161439
.9	.12458	.015719	.093772	.155389	.077664	.171497

Effects of coefficients

Quantile	Quantile effect	Pointwise Std. Err.	Pointwise [95% Conf.	Interval]	Functional [95% Conf.	Interval]
.1	.105339	.042985	.021091	.189588	-.004791	.21547
.5	.118264	.016872	.085196	.151332	.075037	.161491
.9	.137784	.016174	.106084	.169483	.096346	.179222

1. Juhn-Murphy-Pierce 1993

2. Approach based on conditional quantiles

3. Approach based distribution regression

4. Approach based on RIF regression

Approach based on RIF regression

- The above procedures (conditional quantiles, distribution regression) have several drawbacks:
  - Quite complicated and computationally intensive.
  - No easy way to obtain detailed decomposition of composition effect (at least not without path dependency).
  - No easy way to obtain consistent standard errors (apart from bootstrap).
- A simple approach that solves these problems is based on so-called RIF regression (RIF = recentered influence function). RIF regression allows approximate Oaxaca-Blinder type decompositions for almost any distributional statistic of interest.

Influence functions

- An influence function is a function that quantifies how a target statistic changes in response to small changes in the data. That is, for each value y, the influence function $\mathrm{IF}(y; \nu, F_Y)$ provides an approximation of how the functional $\nu(F_Y)$ changes if a small probability mass is added at point y.
- Influence functions are used in robust statistics to describe the robustness properties of various statistics (a robust statistic has a bounded influence function).
- There is also a close connection to the sampling variance of a statistic. The asymptotic sampling variance of a statistic is equal to the sampling variance of the mean of the influence function. Therefore, influence functions provide an easy way to estimate standard errors for many statistics (e.g. inequality measures).

RIF regression

- For example, the influence function of quantile Q_p is

$$\mathrm{IF}(y; Q_p, F_Y) = \frac{p - I(y \leq Q_p)}{f_Y(Q_p)}$$

- Influence functions are centered around zero (that is, have an expected value of zero). To center an influence function around the statistic of interest, we can simply add the statistic to the influence function. This is called a recentered influence function

$$\mathrm{RIF}(y; \nu, F_Y) = \nu(F_Y) + \mathrm{IF}(y; \nu, F_Y)$$

- The idea now is to model the conditional expectation of $\mathrm{RIF}(y; \nu, F_Y)$ using regression models, e.g. using a linear model

$$\mathrm{E}(\mathrm{RIF}(Y; \nu, F_Y)|X) = X\gamma$$

- Coefficient γ thus provides an approximation of how $\nu(F_Y)$ reacts to changes in X.

RIF regression decomposition

- In practice, taking the example of a quantile, we would first compute the sample quantile $\widehat{Q}_p$ and then use kernel density estimation to get $\widehat{f}(\widehat{Q}_p)$, the density of Y at point $\widehat{Q}_p$.
- $\text{RIF}(Y_i; Q_p, F_Y)$ is then computed for each observation by plugging these estimates in to the above formula.
- Finally, we regress $\text{RIF}(Y_i; Q_p, F_Y)$ on X to get an estimate of γ.
- Using the coefficients from RIF regression in two groups, we can perform an Oaxaca-Blinder type decomposition for Q_p. For example:

$$\hat{\Delta}^{Q_p} = \hat{\Delta}_X^{Q_p} + \hat{\Delta}_S^{Q_p} = (\bar{X}^0 - \bar{X}^1)\hat{\gamma}^0 + \bar{X}^1(\hat{\gamma}^0 - \hat{\gamma}^1)$$

- A similar procedure can be followed for any other statistic $\nu(F_Y)$. All you have to know is the influence function, which is usually easy to find in the statistical literature.

Stata implementation

- Command `rifreg` provides RIF regression for quantiles, the Gini coefficient, and the variance. It can be obtained from https://economics.ubc.ca/faculty-and-staff/nicole-fortin/.
  - The RIF variables stored by `rifreg` can then be used in `oaxaca`.
- Influence functions for a variety of (robust) estimates of location, scale, skewness, and kurtosis can be obtained by command `robreg` (type `ssc install robreg`).
  - The procedure is to call `robreg` with option `generate()` to save the IF, then add the value of the estimate to the IF to obtain the RIF, the apply `oaxaca` to the RIF.
- There is also a relatively new package called `rif` that streamlined the computation of the RIF and subsequent application if `oaxaca`.
  - Type: `ssc install rif`
  - egen function to generate RIFs: `help rifvar`
  - streamlined RIF-OB decomposition: `help oaxaca_rif`

Example analysis: private–public gap in wage inequality

```
. use gsoep29, clear
(BCPGEN: Nov 12, 2013 17:15:52-251 DBV29)

. // selection

. generate age = 2012 - bcgeburt

. keep if inrange(age, 25, 55)
(10,780 observations deleted)

. // compute gross wages and ln(wage)

. generate wage = labgro12 / (bctatzeit * 4.3) if labgro12>0 & bctatzeit>0
(1,936 missing values generated)

. generate lnwage = ln(wage)
(1,936 missing values generated)

. // X variables

. generate schooling = bcbilzeit if bcbilzeit>0
(318 missing values generated)

. generate ft_experience = expft12 if expft12>=0
(15 missing values generated)

. generate ft_experience2 = expft12^2 if expft12>=0
(15 missing values generated)

. generate public = oeffd12==1 if oeffd12>0
(2,274 missing values generated)

. // summarize

. summarize wage lnwage schooling ft_experience ft_experience2 public
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	8,090	16.26903	15.21083	.3624283	914.7287
lnwage	8,090	2.615219	.5944705	-1.014929	6.818627
schooling	9,708	12.76118	2.73677	7	18
ft_experie~e	10,011	13.41052	10.03473	0	39
ft_experie~2	10,011	280.5277	324.8873	0	1521

Example analysis: private–public gap in wage inequality

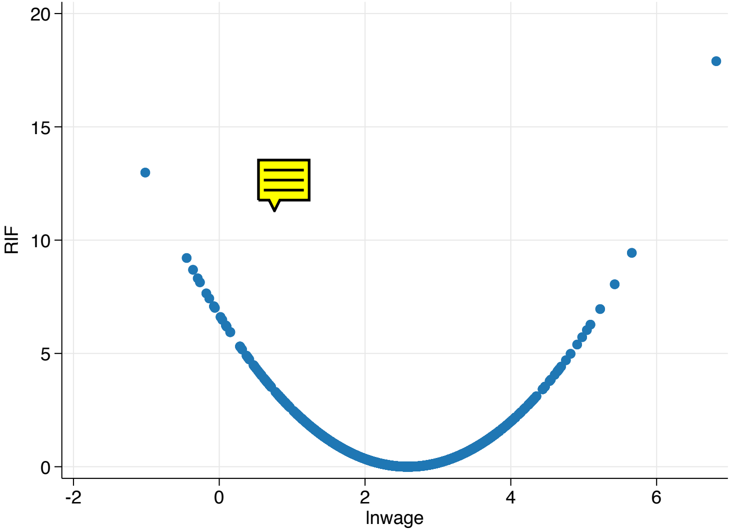
```
. rifreg lnwage schooling ft_experience ft_experience2 if public==0, variance retain(RIF)
(1,912 missing values generated)
      Source |       SS       df       MS              Number of obs =    5476
-------------+------------------------------           F(  3,  5472) =   21.01
       Model |  37.9668705     3  12.6556235           Prob > F      =  0.0000
    Residual |  3296.44132  5472  .602419832           R-squared     =  0.0114
-------------+------------------------------           Adj R-squared =  0.0108
       Total |  3334.40819  5475  .609024327           Root MSE      =  .77616

------------------------------------------------------------------------------
         RIF |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   schooling |   .0226022   .0040773     5.54   0.000      .014609    .0305954
ft_experience |  -.014324   .0038439    -3.73   0.000    -.0218596   -.0067885
ft_experience2 | .0002986   .0001136     2.63   0.009     .0000758    .0005214
       _cons |    .201826   .0589913     3.42   0.001     .0861797    .3174723
------------------------------------------------------------------------------

. regress RIF schooling ft_experience ft_experience2, noheader
------------------------------------------------------------------------------
         RIF |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   schooling |   .0226022   .0040773     5.54   0.000      .014609    .0305954
ft_experience |  -.014324   .0038439    -3.73   0.000    -.0218596   -.0067885
ft_experience2 | .0002986   .0001136     2.63   0.009     .0000758    .0005214
       _cons |    .201826   .0589913     3.42   0.001     .0861797    .3174723
------------------------------------------------------------------------------
```

Example analysis: private–public gap in wage inequality

```
. scatter RIF lnwage
. drop RIF
```



Example analysis: private–public gap in wage inequality

```
. quietly rifreg lnwage if public==0, variance retain(RIFprivate)

. quietly rifreg lnwage if public==1, variance retain(RIFpublic)

. generate double RIF = cond(public==1, RIFpublic, RIFprivate)

. oaxaca RIF schooling (experience: ft_experience ft_experience2), by(public) ///
>     weight(1) robust

Blinder-Oaxaca decomposition                      Number of obs     =      7,388
                                                    Model           =      linea
Group 1: public = 0                                 N of obs 1      =       547
Group 2: public = 1                                 N of obs 2      =       1912

------------------------------------------------------------------------------
             |               Robust
         RIF |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   .3694755   .0105488    35.03   0.000     .3488003    .3901508
     group_2 |   .2041335   .0132183    15.44   0.000     .1782262    .2300409
  difference |    .165342   .0169115     9.78   0.000      .132196     .198488
   explained |  -.0289454   .0057364    -5.05   0.000    -.0401886   -.0177023
 unexplained |   .1942874   .0175807    11.05   0.000     .1598299    .2287449
-------------+----------------------------------------------------------------
explained    |
   schooling |   -.025752   .0056895    -4.53   0.000    -.0369033   -.0146008
  experience |  -.0031934   .0017221    -1.85   0.064    -.0065687    .0001819
-------------+----------------------------------------------------------------
unexplained  |
   schooling |     .34344   .1057709     3.25   0.001     .1361328    .5507472
  experience |   .0831629   .0591501     1.41   0.160    -.0327692     .199095
       _cons |  -.2323155   .1481584    -1.57   0.117    -.5227006    .0580697
------------------------------------------------------------------------------
experience: ft_experience ft_experience2

. drop RIF*
```

Example analysis: private–public gap in wage inequality

```
. quietly robstat lnwage, over(public) generate(RIF) stat(sd)

. generate double RIF = cond(public==1, RIF1+_b[1], RIF0+_b[0])

. oaxaca RIF schooling (experience: ft_experience ft_experience2), by(public) ///
>     weight(1) robust

Blinder-Oaxaca decomposition                     Number of obs     =      7,388
                                                   Model           =     linear
Group 1: public = 0                                N of obs 1      =       5476
Group 2: public = 1                                N of obs 2      =       1912

------------------------------------------------------------------------------
             |               Robust
         RIF |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |    .607845   .0086764    70.06   0.000     .5908395    .6248505
     group_2 |   .4518114   .0146242    30.89   0.000     .4231484    .4804744
  difference |   .1560336   .0170044     9.18   0.000     .1227056    .1893615
   explained |  -.0238077   .0047182    -5.05   0.000    -.0330552   -.0145602
 unexplained |   .1798413   .0174173    10.33   0.000     .1457039    .2139786
-------------+----------------------------------------------------------------
explained    |
   schooling |  -.02[illegible]   .0046797    -4.53   0.000    -.0303531   -.0120092
  experience |  -.00[illegible]   .0014164    -1.85   0.064    -.0054027    .0001496
-------------+----------------------------------------------------------------
unexplained  |
   schooling |   .2915819   .1064475     2.74   0.006     .0829487    .5002151
  experience |   .1245912     .06056     2.06   0.040     .0058957    .2432867
       _cons |  -.2363318   .1523642    -1.55   0.121    -.5349601    .0622965
------------------------------------------------------------------------------
experience: ft_experience ft_experience2

. drop RIF*
```

Example analysis: private–public gap in wage inequality

```
. egen double RIF = rifvar(lnwage), std by(public)

. oaxaca RIF schooling (experience: ft_experience ft_experience2), by(public) ///
>     weight(1) robust

Blinder-Oaxaca decomposition                    Number of obs     =      7,388
                                                  Model           =     linear
Group 1: public = 0                               N of obs 1      =       5476
Group 2: public = 1                               N of obs 2      =       1912

------------------------------------------------------------------------------
             |               Robust
         RIF |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |    .607845   .0086772    70.05   0.000      .590838     .624852
     group_2 |   .4518114   .0146281    30.89   0.000     .4231409    .4804819
  difference |   .1560336   .0170081     9.17   0.000     .1226984    .1893688
   explained |  -.0238099   .0047186    -5.05   0.000    -.0330582   -.0145615
 unexplained |   .1798435    .017421    10.32   0.000      .145699     .213988
-------------+----------------------------------------------------------------
explained    |
   schooling |  -.0211831   .0046801    -4.53   0.000    -.0303559   -.0120103
  experience |  -.0026268   .0014166    -1.85   0.064    -.0054032    .0001496
-------------+----------------------------------------------------------------
unexplained  |
   schooling |   .2916146   .1064706     2.74   0.006      .082936    .5002932
  experience |   .1246399   .0605738     2.06   0.040     .0059175    .2433623
       _cons |   -.236411    .152399    -1.55   0.121    -.5351075    .0622855
------------------------------------------------------------------------------
experience: ft_experience ft_experience2

. drop RIF
```

Example analysis: private–public gap in wage inequality

```
. oaxaca_rif lnwage schooling (experience: ft_experience ft_experience2), by(public) ///
>     wgt(1) rif(std)
No Reweighted Strategy Choosen
Estimating Standard RIF-OAXACA using RIF:std
Model   : Blinder-Oaxaca RIF-decomposition
Type    : Standard
RIF     : std
Scale   : 1
Group 1: public = 0                                 N of obs 1      = 5476
Group c: x2*b1                                      N of obs C      =          .
Group 2: public = 1                                 N of obs 2      = 1912
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overall						
group_1	.607845	.0086772	70.05	0.000	.590838	.624852
group_2	.4518114	.0146281	30.89	0.000	.4231409	.4804819
difference	.1560336	.0170081	9.17	0.000	.1226984	.1893688
explained	-.0238099	.0047186	-5.05	0.000	-.0330582	-.0145615
unexplained	.1798435	.017421	10.32	0.000	.145699	.213988
explained						
schooling	-.0211831	.0046801	-4.53	0.000	-.0303559	-.0120103
experience	-.0026268	.0014166	-1.85	0.064	-.0054032	.0001496
unexplained						
schooling	.2916146	.1064706	2.74	0.006	.082936	.5002932
experience	.1246399	.0605738	2.06	0.040	.0059175	.2433623
_cons	-.236411	.152399	-1.55	0.121	-.5351075	.0622855

experience: ft_experience ft_experience2

Reweighted RIF decomposition

- RIF regression provides linear approximations of effects of *small* changes in the data on the statistic of interest. However, effects on statistics such as inequality measures are likely to be highly nonlinear and interaction effects are also likely.
- It might therefore be important to use a flexible specification of the RIF regression.
- Since in the decomposition we evaluate potentially *large* changes, Firpo et al. (2018) suggest to combine the RIF decomposition with reweighting (analogous to the reweighted OB decomposition). This will quantify the specification error.
- `oaxaca_rif` has a built-in option to perform such reweighted RIF decompositions (although standard errors may not be reliable). In the exercises next week we will try to construct the reweighted RIF decomposition manually.

Example: gender wage gap at different quantiles

```
. use gsoep29, clear
(BCPGEN: Nov 12, 2013 17:15:52-251 DBV29)
. // selection
. generate age = 2012 - bcgeburt
. keep if inrange(age, 25, 55)
(10,780 observations deleted)
. // compute gross wages and ln(wage)
. generate wage = labgro12 / (bctatzeit * 4.3) if labgro12>0 & bctatzeit>0
(1,936 missing values generated)
. generate lnwage = ln(wage)
(1,936 missing values generated)
. // X variables
. generate schooling = bcbilzeit if bcbilzeit>0
(318 missing values generated)
. generate ft_experience = expft12 if expft12>=0
(15 missing values generated)
. generate ft_experience2 = expft12^2 if expft12>=0
(15 missing values generated)
. // group variable
. generate byte female = bcsex==2 if bcsex<.
. // summarize
. summarize wage lnwage schooling ft_experience ft_experience2 female
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	8,090	16.26903	15.21083	.3624283	914.7287
lnwage	8,090	2.615219	.5944705	-1.014929	6.818627
schooling	9,708	12.76118	2.73677	7	18
ft_experie~e	10,011	13.41052	10.03473	0	39
ft_experie~2	10,011	280.5277	324.8873	0	1521
female	10,026	.5398963	.4984306	0	1

```
. drop if missing(lnwage,schooling,ft_experience,female)
(2,166 observations deleted)
```

Example: gender wage gap at different quantiles

```
. oaxaca_rif lnwage schooling (experience: ft_experience ft_experience2), ///
>     by(female) wgt(1) rif(q(10)) ///
>     rwlogit(c.schooling##c.ft_experience##c.ft_experience)
Estimating Reweighted RIF-OAXACA using RIF:q(10)
Model   : Blinder-Oaxaca RIF-decomposition
Type    : Reweighted
RIF     : q(10)
Scale   : 1
Group 1: female = 0                               N of obs 1      = 3877
Group c: X1~>rw~>X2                               N of obs C      = 3877
Group 2: female = 1                               N of obs 2      = 3983

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Overall      |
     Group_1 |   2.086449   .0176101   118.48   0.000     2.051934    2.120965
     Group_c |   1.811982   .0403479    44.91   0.000     1.732902    1.891063
     Group_2 |   1.837762   .0168858   108.83   0.000     1.804666    1.870857
 Tdifference |   .2486877   .0243977    10.19   0.000     .2008691    .2965063
ToT_Explained|   .2744669   .0352501     7.79   0.000     .2053779    .3435558
ToT_Unexplained| -.0257792  .0436244    -0.59   0.555    -.1112813     .059723
-------------+----------------------------------------------------------------
Explained    |
       Total |   .2744669   .0352501     7.79   0.000     .2053779    .3435558
Pure_explained|  .2310962   .0196185    11.78   0.000     .1926447    .2695477
  Specif_err |   .0433707   .0306093     1.42   0.157    -.0166224    .1033638
-------------+----------------------------------------------------------------
Pure_explained|
   schooling |  -.0050829   .0030646    -1.66   0.097    -.0110894    .0009237
  experience |   .2361791   .0192384    12.28   0.000     .1984725    .2738857
-------------+----------------------------------------------------------------
Specif_err   |
   schooling |  -.3512984   .1666451    -2.11   0.035    -.6779167   -.0246801
  experience |  -.1641361   .0771264    -2.13   0.033     -.315301   -.0129711
       _cons |   .5588051   .2425005     2.30   0.021      .083513    1.034097
-------------+----------------------------------------------------------------
Unexplained  |
       Total |  -.0257792   .0436244    -0.59   0.555    -.1112813     .059723
Reweight_err |  -.0259947   .0150057    -1.73   0.083    -.0554054     .003416
Pure_Unexplained| .0002155  .0390915     0.01   0.996    -.0764024    .0768335
-------------+----------------------------------------------------------------
Pure_Unexplained|
```

Example: gender wage gap at different quantiles

```
. oaxaca_rif lnwage schooling (experience: ft_experience ft_experience2), ///
>     by(female) wgt(1) rif(q(50)) ///
>     rwlogit(c.schooling##c.ft_experience##c.ft_experience)
Estimating Reweighted RIF-OAXACA using RIF:q(50)
Model   : Blinder-Oaxaca RIF-decomposition
Type    : Reweighted
RIF     : q(50)
Scale   : 1
Group 1: female = 0                             N of obs 1      = 3877
Group c: X1~>rw~>X2                             N of obs C      = 3877
Group 2: female = 1                             N of obs 2      = 3983

------------------------------------------------------------------------------
          lnwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
Overall          |
         Group_1 |   2.790358   .0099056   281.70   0.000     2.770943    2.809772
         Group_c |    2.64092   .0150704   175.24   0.000     2.611382    2.670457
         Group_2 |   2.543904   .0099248   256.32   0.000     2.524452    2.563357
     Tdifference |    .246453   .0140222    17.58   0.000       .21897     .273936
    ToT_Explained |   .1494378   .0123544    12.10   0.000     .1252236     .173652
  ToT_Unexplained |   .0970152   .0179589     5.40   0.000     .0618165     .132214
-----------------+------------------------------------------------------------
Explained        |
           Total |   .1494378   .0123544    12.10   0.000     .1252236     .173652
   Pure_explained |   .1393007   .0086627    16.08   0.000     .1223221    .1562793
      Specif_err |   .0101371   .0091117     1.11   0.266    -.0077214    .0279957
-----------------+------------------------------------------------------------
Pure_explained   |
       schooling |  -.0064585   .0038593    -1.67   0.094    -.0140226    .0011056
      experience |   .1457592   .0072054    20.23   0.000     .1316368    .1598816
-----------------+------------------------------------------------------------
Specif_err       |
       schooling |  -.0989773   .0406031    -2.44   0.015     -.178558   -.0193966
      experience |  -.0765073   .0175105    -4.37   0.000    -.1108273   -.0421872
           _cons |   .1856217   .0449814     4.13   0.000     .0974598    .2737835
-----------------+------------------------------------------------------------
Unexplained      |
           Total |   .0970152   .0179589     5.40   0.000     .0618165     .132214
    Reweight_err |  -.0141555   .0106512    -1.33   0.184    -.0350314    .0067204
 Pure_Unexplained |   .1111707   .0147706     7.53   0.000     .0822208    .1401206
-----------------+------------------------------------------------------------
Pure_Unexplained |
```

Example: gender wage gap at different quantiles

```
. oaxaca_rif lnwage schooling (experience: ft_experience ft_experience2), ///
>     by(female) wgt(1) rif(q(90)) ///
>     rwlogit(c.schooling##c.ft_experience##c.ft_experience)
Estimating Reweighted RIF-OAXACA using RIF:q(90)
Model   : Blinder-Oaxaca RIF-decomposition
Type    : Reweighted
RIF     : q(90)
Scale   : 1
Group 1: female = 0                            N of obs 1      = 3877
Group c: X1~>rw~>X2                            N of obs C      = 3877
Group 2: female = 1                            N of obs 2      = 3983

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Overall      |
     Group_1 |   3.391892   .0133722   253.65   0.000     3.365683    3.418101
     Group_c |   3.289035   .0157856   208.36   0.000     3.258096    3.319974
     Group_2 |   3.134464   .0131988   237.48   0.000     3.108595    3.160333
 Tdifference |   .2574281   .0187889    13.70   0.000     .2206025    .2942538
ToT_Explained|   .1028573   .0125369     8.20   0.000     .0782855     .127429
ToT_Unexplained| .1545709   .0204306     7.57   0.000     .1145276    .1946141
-------------+----------------------------------------------------------------
Explained    |
       Total |   .1028573   .0125369     8.20   0.000     .0782855     .127429
Pure_explained|  .1327299   .0109632    12.11   0.000     .1112424    .1542173
  Specif_err |  -.0298726    .009321    -3.20   0.001    -.0481414   -.0116039
-------------+----------------------------------------------------------------
Pure_explained|
   schooling |  -.0072561   .0043494    -1.67   0.095    -.0157809    .0012686
  experience |    .139986   .0097284    14.39   0.000     .1209186    .1590534
-------------+----------------------------------------------------------------
Specif_err   |
   schooling |   .0922572   .0647929     1.42   0.154    -.0347346     .219249
  experience |  -.0092558   .0183394    -0.50   0.614    -.0452004    .0266888
       _cons |   -.112874   .0758837    -1.49   0.137    -.2616033    .0358552
-------------+----------------------------------------------------------------
Unexplained  |
       Total |   .1545709   .0204306     7.57   0.000     .1145276    .1946141
Reweight_err |  -.0130203    .010007    -1.30   0.193    -.0326337    .0065932
Pure_Unexplained| .1675911  .0182325     9.19   0.000      .131856    .2033263
-------------+----------------------------------------------------------------
Pure_Unexplained|
```

References

- Chernozhukov, Victor, Iván Fernández-Val, Blaise Melly (2013). Inference on Counterfactual Distributions. Econometrica 81(6):2205–2268.
- DiNardo, John E., Nicole Fortin, Thomas Lemieux (1996). Labour Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. Econometrica 64(5):1001–1046.
- Firpo, Sergio, Nicole Fortin, Thomas Lemieux (2007). Decomposing Wage Distributions using Recentered Influence Function Regressions. Working paper.
- Firpo, Sergio, Nicole M. Fortin, Thomas Lemieux (2009). Unconditional Quantile Regressions. Econometrica 77:953–973.
- Firpo, Sergio, Nicole M. Fortin, Thomas Lemieux (2018). Decomposing Wage Distributions Using Recentered Influence Function Regressions. Econometrics 6(2): 28 (DOI:10.3390/econometrics6020028).
- Juhn, Chinhui, Kevin M. Murphy, Brooks Pierce (1993). Wage Inequality and the Rise in Returns to Skill. Journal of Political Economy 101(3):410–442.
- Machado, José A. F., José Mata (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. Journal of Applied Econometrics 20(4):445–465.
- Melly, Blaise (2005). Decomposition of differences in distribution using quantile regression. Labour Economics 12(4):577–590.
- Melly, Blaise (2006). Estimation of counterfactual distributions using quantile regression. University of St. Gallen, Discussion Paper.