Decomposition Methods in the Social Sciences

GESIS Training Course
January 29 – February 1, 2024, Cologne

Johannes Giesecke (Humboldt University Berlin)
Ben Jann (University of Bern)

7. RIF decomposition

Beyond the mean: recap

- 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).
  - Estimating $\nu(F_{Y^g|G\neq g})$ via recentered influence function regression (Firpo et al. 2007, 2009)
  - 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 2005, Melly 2005, 2006)
  - Estimating $F_{Y^g|G\neq g}$ by distribution regression (Chernozhukov et al. 2013)
- We have already looked at reweighting. Now, we will look at recentered influence function (RIF) regression.

Approach based on RIF regression

- A simple approach that was porposed by Firpo et al. (2007, 2009) 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.
- Decompositions based on RIF regression has several advantages:
  - It is computationally quite easy.
  - It offers an easy way to obtain detailed decomposition of composition effect.
  - It offers an easy way to obtain consistent standard errors.

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 $\text{IF}(y; \nu, F_Y)$ 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 statistic (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

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

- 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

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

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

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

- 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:

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

- 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://sites.google.com/view/nicole-m-fortin/data-and-programs.

  [scroll down to „RIF Regression (*.ado files)from Firpo, Fortin and Lemieux (2009)“]

  - The RIF variables stored by `rifreg` can then be used in `oaxaca`.
- There is also a relatively new package called `rif` (Rios-Avila 2020) that streamlines 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`
- Highly accurate influence functions for a very large number of statistics can also be computed by command `dstat` (Jann 2020; type `ssc install dstat`).
  - The procedure is to call `dstat` with option `rif()` to save the RIF, then apply `oaxaca` to the RIF.

Example analysis: private–public gap in wage inequality

```
. use gsoep-extract, clear
(Example data based on the German Socio-Economic Panel)
. keep if wave==2015
(29,970 observations deleted)
. keep if inrange(age, 25, 55)
(5,671 observations deleted)
. generate lnwage = ln(wage)
(1,709 missing values generated)
. generate expft2 = expft^2
(35 missing values generated)
. svyset psu [pw=weight], strata(strata)
Sampling weights: weight
             VCE: linearized
     Single unit: missing
        Strata 1: strata
 Sampling unit 1: psu
           FPC 1: <zero>
. summarize wage lnwage yeduc expft expft2 public
```

Variable	Obs	Mean	Std. dev.	Min	Max
wage	5,600	17.57278	9.858855	3.03	121.42
lnwage	5,600	2.736721	.5062968	1.108563	4.799255
yeduc	7,121	12.28823	2.783974	7	18
expft	7,274	11.63359	9.556508	0	39.5
expft2	7,274	226.6548	293.3739	0	1560.25
public	5,770	.2353553	.4242574	0	1

```
. drop if missing(lnwage, yeduc, expft, public) // remove unused observation
(1,851 observations deleted)
```

We look at the variance of log wages

`rifreg` computes the RIF and then applies `regress`

```
. rifreg lnwage yeduc expft expft2 ///
>     [aw=weight] /// rifreg does not allow pweights
>     if public==0, variance retain(RIF)
(1,274 missing values generated)
```

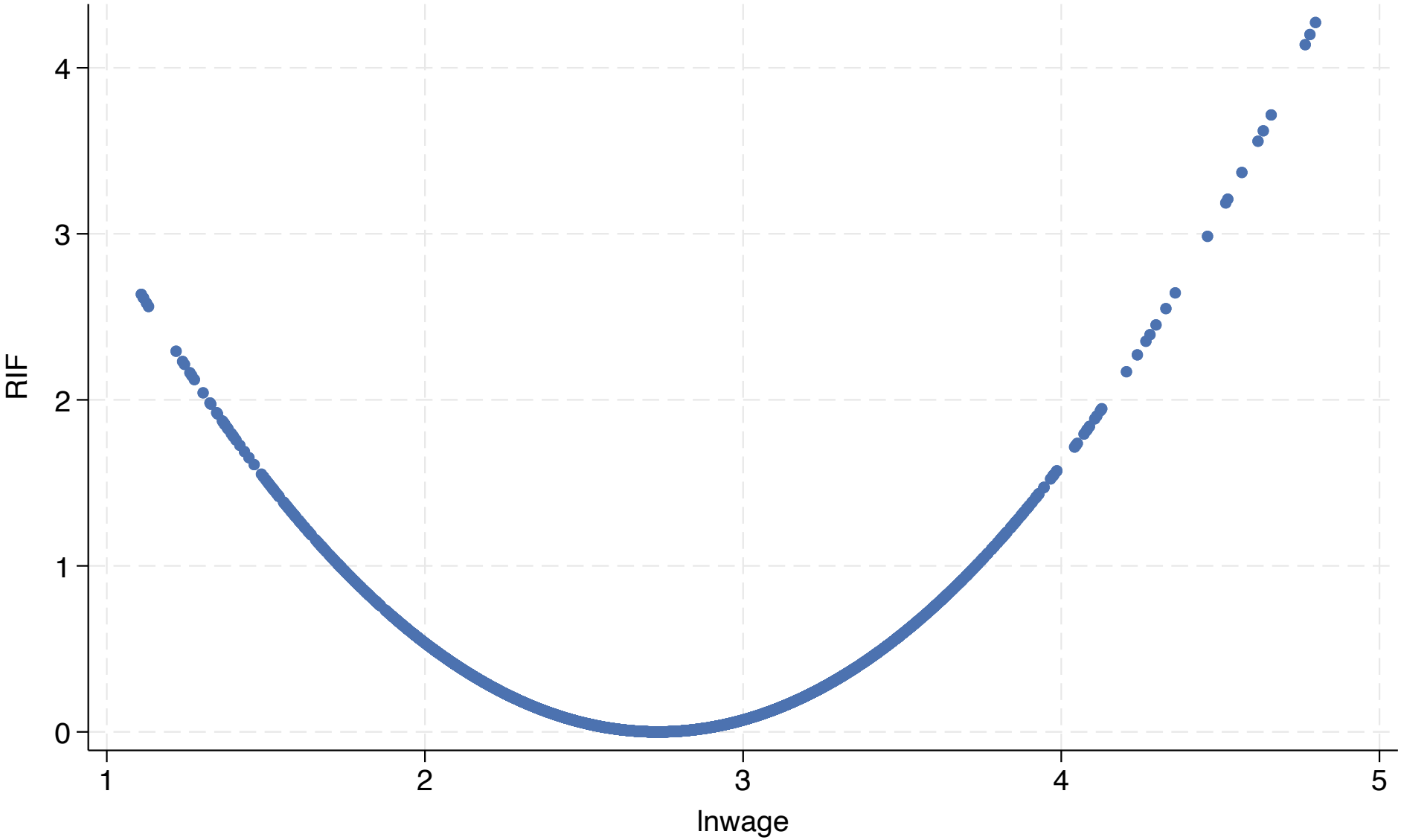
Source	SS	df	MS	Number of obs =	4184
				F(3, 4180) =	20.18
Model	8.4861949	3	2.82873163	Prob > F =	0.0000
Residual	585.814051	4180	.140146902	R-squared =	0.0143
				Adj R-squared =	0.0136
Total	594.300246	4183	.142075125	Root MSE =	.37436

RIF	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
yeduc	.0110638	.0022301	4.96	0.000	.0066916	.0154361
expft	-.0079049	.0021117	-3.74	0.000	-.0120449	-.0037648
expft2	.0001593	.0000626	2.54	0.011	.0000365	.000282
_cons	.1776732	.031722	5.60	0.000	.1154813	.2398651

```
. regress RIF yeduc expft expft2 [pw=weight], noheader
(sum of wgt is 9,231,938.5954959)
```

RIF	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
yeduc	.0110638	.0048886	2.26	0.024	.0014797	.020648
expft	-.0079049	.0042057	-1.88	0.060	-.0161503	.0003406
expft2	.0001593	.0001063	1.50	0.134	-.0000492	.0003677
_cons	.1776732	.0838222	2.12	0.034	.013337	.3420093

How does the RIF of the variance look like?



RIF decomposition (using `rifreg` and `oaxaca`)

```
. quietly rifreg lnwage [aw=weight] if public==0, variance retain(RIFprivate)
. quietly rifreg lnwage [aw=weight] if public==1, variance retain(RIFpublic)
. generate double RIF = cond(public==1, RIFpublic, RIFprivate)
. oaxaca RIF yeduc (experience: expft expft2), by(public) weight(1) svy
Blinder-Oaxaca decomposition
Number of strata =    15                        Number of obs     =      5,458
Number of PSUs   = 2,036                        Population size   = 12,146,771
                                                Design df         =      2,021
                                                Model             =     linear
Group 1: public = 0                             N of obs 1        =      4,184
Group 2: public = 1                             N of obs 2        =      1,274
    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)
```

RIF	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
overall						
group_1	.250799	.0098871	25.37	0.000	.2314091	.2701889
group_2	.1968692	.0178071	11.06	0.000	.1619471	.2317913
difference	.0539298	.0203778	2.65	0.008	.0139661	.0938935
explained	-.0207047	.0079636	-2.60	0.009	-.0363224	-.005087
unexplained	.0746345	.0205978	3.62	0.000	.0342395	.1150296
explained						
yeduc	-.0181894	.0082889	-2.19	0.028	-.0344451	-.0019337
experience	-.0025153	.0020614	-1.22	0.223	-.006558	.0015273
unexplained						
yeduc	.1189914	.1087014	1.09	0.274	-.0941871	.3321699
experience	.0791311	.0569362	1.39	0.165	-.0325288	.1907909
_cons	-.1234879	.1439615	-0.86	0.391	-.4058165	.1588406

```
experience: expft expft2
. drop RIF*
```

RIF decomposition (using `rifvar` and `oaxaca`)

```
. egen double RIF = rifvar(lnwage), var by(public) weight(weight)
. oaxaca RIF yeduc (experience: expft expft2), by(public) weight(1) svy
Blinder-Oaxaca decomposition
Number of strata =    15                        Number of obs     =      5,458
Number of PSUs   = 2,036                        Population size   = 12,146,771
                                                Design df         =      2,021
                                                Model             =     linear
Group 1: public = 0                             N of obs 1        =      4,184
Group 2: public = 1                             N of obs 2        =      1,274
    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)
```

RIF	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
overall						
group_1	.250799	.0098871	25.37	0.000	.2314091	.2701889
group_2	.1968692	.0178071	11.06	0.000	.1619471	.2317913
difference	.0539298	.0203778	2.65	0.008	.0139661	.0938935
explained	-.0207047	.0079636	-2.60	0.009	-.0363224	-.005087
unexplained	.0746345	.0205978	3.62	0.000	.0342395	.1150296
explained						
yeduc	-.0181894	.0082889	-2.19	0.028	-.0344451	-.0019337
experience	-.0025153	.0020614	-1.22	0.223	-.006558	.0015273
unexplained						
yeduc	.1189914	.1087014	1.09	0.274	-.0941871	.3321699
experience	.0791311	.0569362	1.39	0.165	-.0325288	.1907909
_cons	-.1234879	.1439615	-0.86	0.391	-.4058164	.1588406

```
experience: expft expft2
. drop RIF
```

RIF decomposition (using oaxaca_rif)

```
. oaxaca_rif lnwage yeduc (experience: expft expft2) ///
>     [pw=weight], by(public) wgt(1) rif(var) cluster(psu)
No Reweighted Strategy Choosen
Estimating Standard RIF-OAXACA using RIF:var
Model  : Blinder-Oaxaca RIF-decomposition
Type   : Standard
RIF    : var
Scale  : 1
Group 1: public = 0 x1*b1                          N of obs 1      = 4184
Group c: x2*b1                                     N of obs C      =     .
Group 2: public = 1 x2*b2                          N of obs 2      = 1274
                                (Std. err. adjusted for 2,036 clusters in psu)
```

lnwage	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
overall						
group_1	.250799	.0098877	25.36	0.000	.2314195	.2701784
group_2	.1968692	.0177784	11.07	0.000	.1620242	.2317142
difference	.0539298	.0203614	2.65	0.008	.0140222	.0938374
explained	-.0207047	.0079782	-2.60	0.009	-.0363417	-.0050677
unexplained	.0746345	.0206196	3.62	0.000	.0342209	.1150482
explained						
yeduc	-.0181894	.0083108	-2.19	0.029	-.0344784	-.0019005
experience	-.0025153	.0020632	-1.22	0.223	-.0065592	.0015285
unexplained						
yeduc	.1189914	.1088647	1.09	0.274	-.0943795	.3323624
experience	.0791311	.0568256	1.39	0.164	-.032245	.1905071
_cons	-.1234879	.1439222	-0.86	0.391	-.4055703	.1585944

```
experience: expft expft2
```

RIF decomposition (using dstat and oaxaca)

```
. quietly dstat (variance(0)) lnwage [pw=weight], over(public) rif(RIF, compact)
. oaxaca RIF yeduc (experience: expft expft2), by(public) weight(1) svy
Blinder-Oaxaca decomposition
Number of strata =    15                        Number of obs     =      5,458
Number of PSUs   = 2,036                        Population size   = 12,146,771
                                                Design df         =      2,021
                                                Model             =     linear
Group 1: public = 0                             N of obs 1        =      4,184
Group 2: public = 1                             N of obs 2        =      1,274
    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)
```

RIF	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
overall						
group_1	.250799	.0098871	25.37	0.000	.2314091	.2701889
group_2	.1968692	.0178071	11.06	0.000	.1619471	.2317913
difference	.0539298	.0203778	2.65	0.008	.0139661	.0938935
explained	-.0207047	.0079636	-2.60	0.009	-.0363224	-.005087
unexplained	.0746345	.0205978	3.62	0.000	.0342395	.1150296
explained						
yeduc	-.0181894	.0082889	-2.19	0.028	-.0344451	-.0019337
experience	-.0025153	.0020614	-1.22	0.223	-.006558	.0015273
unexplained						
yeduc	.1189914	.1087014	1.09	0.274	-.0941871	.3321699
experience	.0791311	.0569362	1.39	0.165	-.0325288	.1907909
_cons	-.1234879	.1439615	-0.86	0.391	-.4058164	.1588406

```
experience: expft expft2
. drop RIF*
```

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 we will try to construct the reweighted RIF decomposition manually.

Reweighted RIF decomposition (using oaxaca_rif)

```
. oaxaca_rif lnwage yeduc (experience: expft expft2) ///
>     [pw=weight], by(public) cluster(psu) wgt(1) rif(var) ///
>     rwlogit(c.yeduc##c.expft##c.expft)
Estimating Reweighted RIF-OAXACA using RIF:var
Model  : Blinder-Oaxaca RIF-decomposition
Type   : Reweighted
RIF    : var
Scale  : 1
Group 1: public = 0 x1*b1                          N of obs 1      = 4184
Group c: X1~>rw~>X2 or x2*b1                       N of obs C      = 4184
Group 2: public = 1 x2*b2                          N of obs 2      = 1274
                                (Std. err. adjusted for 2,036 clusters in psu)
```

lnwage	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
Overall						
group_1	.250799	.0098048	25.58	0.000	.231582	.270016
group_c	.2678071	.0052595	50.92	0.000	.2574986	.2781155
group_2	.1968692	.0182572	10.78	0.000	.1610858	.2326526
tdifference	.0539298	.0207419	2.60	0.009	.0132764	.0945832
t_explained	-.0170081	.0110085	-1.54	0.122	-.0385843	.0045681
t_unexplained	.0709379	.0377126	1.88	0.060	-.0029776	.1448533
explained						
total	-.0170081	.0110085	-1.54	0.122	-.0385843	.0045681
p_explained	-.0209147	.0116102	-1.80	0.072	-.0436702	.0018408
specif_err	.0039066	.0132571	0.29	0.768	-.0220768	.0298901

Reweighted RIF decomposition (using `oaxaca_rif`)

p_explained						
yeduc	-.0185053	.012529	-1.48	0.140	-.0430617	.006051
experience	-.0024094	.0016408	-1.47	0.142	-.0056253	.0008065
specif_err						
yeduc	.1059227	.080579	1.31	0.189	-.0520092	.2638546
experience	-.0274095	.0339557	-0.81	0.420	-.0939614	.0391424
_cons	-.0746065	.0992839	-0.75	0.452	-.2691994	.1199864
unexplained						
total	.0709379	.0377126	1.88	0.060	-.0029776	.1448533
rwg_error	.0000329	.0015476	0.02	0.983	-.0030004	.0030662
p_unexplained	.070905	.0376173	1.88	0.059	-.0028235	.1446335
p_unexplained						
yeduc	.0132832	.1802001	0.07	0.941	-.3399024	.3664689
experience	.1065032	.1037578	1.03	0.305	-.0968585	.3098648
_cons	-.0488814	.2429951	-0.20	0.841	-.5251432	.4273803
rwg_error						
yeduc	.0001014	.0006713	0.15	0.880	-.0012143	.0014171
experience	-.0000685	.0011994	-0.06	0.954	-.0024193	.0022823

experience:
expft
expft2

Exercise 7

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.

Jann, Ben (2020). dstat: Stata module to compute summary statistics and distribution functions including standard errors and optional covariate balancing. Available from http://ideas.repec.org/c/boc/bocode/s458874.html.

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.

References

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.

Rios-Avila, Fernando (2020). Recentered influence functions (RIFs) in Stata: RIF regression and RIF decomposition. The Stata Journal 20(1):51–94