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)

8. Further distribution decompositions

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 and RIF decomposition. We will now discuss the remaining approaches.

Contents

1. Juhn-Murphy-Pierce 1993
2. Approach based on conditional quantiles
3. Approach based on distribution regression

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 the 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

$$Y_i = X_i\beta^g + \epsilon_i^g \quad \text{with } \epsilon_i^g = h^g(\upsilon_i)$$

  can be used to describe Y in group g. Think of β^g as "returns to observables", $h^g()$ as "returns to unobservables", and υ_i as the values of the unobservables.
- We can now construct counterfactual outcomes for group 1 based on counterfactual residuals ϵ_i^C based on the group 0 residual distribution (see below on how to obtain counterfactual residuals).
- JMP propose to do this in two steps.
  - In the first step, impute observations by replacing the observed residuals by the counterfactual residuals:

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

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

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

JMP 1993

- We can then compute a decomposition as

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

where

Δ_X^{ν} part due to differential composition of observables
Δ_{β}^{ν} part due to differential returns of observables
Δ_{ϵ}^{ν} part due to differential returns and composition of unobservables

JMP 1993

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

$$\epsilon_i^C = F_{\epsilon^0}^{-1}(\tau_i)$$

- 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:

```
net install jmpierce, replace ///
    from(https://raw.githubusercontent.com/benjann/jmpierce/main/)
```

(also on SSC, but the SSC version currently does not support the decomposition of the variance)

Example

```
. 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)
. 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)
```

Example

```
. regress lnwage yeduc expft expft2 [pw=weight] if public==0
  (output omitted)
. estimates store private
. regress lnwage yeduc expft expft2 [pw=weight] if public==1
  (output omitted)
. estimates store public

. jmpierce private public, reference(1) statistics(mean d9010 d9050 d5010 variance)
Juhn-Murphy-Pierce decomposition (reference estimates: private)
                   T           Q           P           U
    mean  -.13395921  -.12931139  -.00769482     .003047
   d9010   .18151069  -.05949926   .07679462   .16421533
   d9050   .19098759   .04404283   .04680848   .10013628
   d5010   -.0094769  -.10354209   .02998614   .06407905
variance    .0538351   .00009581   .01535953   .03837977
T = Total difference (private-public)
Q = Contribution of differences in observable quantities
P = Contribution of differences in observable prices
U = Contribution of differences in unobservable quantities and prices
```

1. Juhn-Murphy-Pierce 1993

2. Approach based on conditional quantiles

3. Approach based on distribution regression

Approach based on conditional quantiles

- The JMP decomposition, at least if based on *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 i from group 1 and predict

$$Y_j^C = X_i\beta^0_{\tau_j} \quad \text{and} \quad \widehat{Y}_j = X_i\beta^1_{\tau_j}$$

  3. Compute the decomposition by comparing Y^C and $\widehat{Y}$:

$$\Delta^\nu_S = \nu(F_{Y^C}) - \nu\left(F_{\widehat{Y}}\right)$$
$$\Delta^\nu_X = \Delta^\nu - \Delta^\nu_S \quad \text{where } \Delta^\nu \text{ is the observed difference}$$

Approach based on conditional quantiles

- As Melly (2005, 2006) shows, the simulation procedure proposed by MM is more complicated than necessary.
- An equivalent but much more efficient approach is as follows (also see Chernozhukov et al. 2013):
  - Estimate J quantile regression in group 0 over a regular grid of τ values between 0 and 1 (e.g., 100 quantile regressions from $\tau_1 = .005$ to $\tau_J = 0.995$ in steps of 0.01). Let $\beta^0_{\tau_j}$, $j = 1, \ldots, J$ be the coefficient vectors of these quantile regressions.
  - Use these coefficients to generate J predictions $\tilde{Y}^0_{ij} = X_i \beta^0_{\tau_j}$ for each observation i in group 1. In this way, a "realistic" counterfactual distribution of outcome values is generated for each observation, given the observation's values of X.
  - Compute any statistic of interest from these values across all observations in group 1.
    - ★ If n^1 is the number of observations in group 1, there are $n^1 \times J$ counterfactual outcome values; each of the J values from observation i is weighted by sampling weight w_i from this observation.

Approach based on conditional quantiles

- Let $\tilde{\nu}^0$ denote the statistic computed in this way.
  - It can be interpreted as the value of the statistic we would get if outcomes in group 1 would come about according to the outcome mechanism of group 0.
  - Likewise, $\tilde{\nu}^0$ can be interpreted as the covariate-adjusted value of the statistic in group 0 (i.e. the value of the statistic we would obtain if group 0 had the X distribution of group 1).
- The decomposition can then be obtained as follows:

$$\begin{aligned}\Delta^{\nu} &= \nu^0 - \nu^1 \\ &= \left\{\nu^0 - \tilde{\nu}^0\right\} + \left\{\tilde{\nu}^0 - \nu^1\right\} \\ &= \Delta_X^{\nu} + \Delta_S^{\nu}\end{aligned}$$

where ν^0 and ν^1 are the observed values of the statistic in group 0 and 1, respectively.

Approach based on conditional quantiles

- Because the procedure will have some approximation error (depending on the grid size J), an alternative suggestion is to obtain the decomposition as

$$\Delta_X^\nu = \widehat{\nu}^0 - \tilde{\nu}^0$$
$$\Delta_S^\nu = \tilde{\nu}^0 - \widehat{\nu}^1$$

  where $\widehat{\nu}^0$ is the value of the statistic obtained from fitted data in group 0 (i.e. applying the above procedure to group 0, with coefficients $\beta^0_{\tau_j}$ estimated in group 0) and $\widehat{\nu}^1$ is the fitted statistic in group 1 (i.e. applying the above procedure to group 1, with coefficients $\beta^1_{\tau_j}$ estimated in group 1).
- Approximation errors can be inspected by looking at $\nu^0 - \widehat{\nu}^0$ and $\nu^1 - \widehat{\nu}^1$, respectively.

Approach based on conditional quantiles

- Naturally, we can also obtain an alternative decomposition based on a covariate-adjusted statistic for group 1 (i.e. obtain $\tilde{\nu}^1$ by applying coefficients $\beta^1_{\tau_j}$ to group 0), that is

$$\Delta^{\nu}_{X} = \tilde{\nu}^1 - \widehat{\nu}^1$$
$$\Delta^{\nu}_{S} = \widehat{\nu}^0 - \tilde{\nu}^1$$

- Furthermore, we could obtain a decomposition in which we adjust both groups to the pooled distribution of characteristics across both groups. To do this, compute $\tilde{\nu}^0_p$ and $\tilde{\nu}^1_p$ by applying $\beta^0_{\tau_j}$ and $\beta^1_{\tau_j}$, respectively, to the pooled sample and then define the decomposition as

$$\Delta^{\nu}_{X} = (\widehat{\nu}^0 - \tilde{\nu}^0_p) + (\tilde{\nu}^1_p - \widehat{\nu}^1)$$
$$\Delta^{\nu}_{S} = \tilde{\nu}^0_p - \tilde{\nu}^1_p$$

Approach based on conditional quantiles

- Stata implementation of the procedure by Blaise Melly:
  - command `cdeco` in package `counterfactual`

```
net install counterfactual, ///
    from("https://raw.githubusercontent.com/bmelly/Stata/main/")
```

- New command `cdist` by Ben Jann:
  - available from SSC

```
ssc install cdist, replace
```

  - Option `method(qr)` requests the quantile regression method.

- Standard errors: need to bootstrap; this is unfortunate because the procedure is computationally expensive (J or even $2 \times J$ quantile regressions).

Example

Estimate counterfactuals

```
. cdist lnwage yeduc c.expft##c.expft [pw=weight], ///
>     by(public) method(qr) ///
>     statistics(mean iqr(10 90) iqr(50 90) iqr(10 50) variance)
group 0: fitting models 0%....20%....40%....60%....80%....100%
enumerating predictions ... done
group 1: fitting models 0%....20%....40%....60%....80%....100%
enumerating predictions ... done
Counterfactual distribution estimation          Number of obs      =      5,458
                                                Pooled             =         no
Group 0: public = 0                             N of obs 0         =      4,184
Group 1: public = 1                             N of obs 1         =      1,274
                                                Estimation method  =         qr
                                                Grid size          =        100
```

lnwage	Coefficient
obs0	
mean	2.732109
iqr(10,90)	1.242421
iqr(50,90)	.6530616
iqr(10,50)	.5893595
variance	.250799
fit0	
mean	2.731791
iqr(10,90)	1.259786
iqr(50,90)	.6502121
iqr(10,50)	.6095739
variance	.2504928

Example

Estimate counterfactuals

adj0	
mean	2.860572
iqr(10,90)	1.308552
iqr(50,90)	.654276
iqr(10,50)	.654276
variance	.2651937
obs1	
mean	2.866068
iqr(10,90)	1.06091
iqr(50,90)	.462074
iqr(10,50)	.5988364
variance	.1968692
fit1	
mean	2.865989
iqr(10,90)	1.02002
iqr(50,90)	.4510847
iqr(10,50)	.5689356
variance	.1914739
adj1	
mean	2.770909
iqr(10,90)	1.007829
iqr(50,90)	.4388932
iqr(10,50)	.5689356
variance	.1904105

covariates: yeduc expft c.expft#c.expft

Decomposition with private sector wage structure as reference:

```
. cdist lincom (Difference:fit0-fit1) ///
>              (Explained:fit0-adj0) ///
>              (Unexplained:adj0-fit1)
   Difference: fit0-fit1
    Explained: fit0-adj0
  Unexplained: adj0-fit1
```

lnwage	Coefficient
Difference	
mean	-.1341982
iqr(10,90)	.2397657
iqr(50,90)	.1991275
iqr(10,50)	.0406383
variance	.0590189
Explained	
mean	-.1287806
iqr(10,90)	-.0487659
iqr(50,90)	-.0040638
iqr(10,50)	-.0447021
variance	-.014701
Unexplained	
mean	-.0054176
iqr(10,90)	.2885316
iqr(50,90)	.2031913
iqr(10,50)	.0853403
variance	.0737199

```
covariates: yeduc expft c.expft#c.expft
```

Could also type: `cdist decomp`

Decomposition with public sector wage structure as reference:

```
. cdist lincom (Difference:fit0-fit1) ///
>              (Explained:adj1-fit1) ///
>              (Unexplained:fit0-adj1)
```

```
   Difference: fit0-fit1
    Explained: adj1-fit1
  Unexplained: fit0-adj1
```

lnwage	Coefficient
Difference	
mean	-.1341982
iqr(10,90)	.2397657
iqr(50,90)	.1991275
iqr(10,50)	.0406383
variance	.0590189
Explained	
mean	-.0950801
iqr(10,90)	-.0121915
iqr(50,90)	-.0121915
iqr(10,50)	4.44e-16
variance	-.0010633
Unexplained	
mean	-.0391181
iqr(10,90)	.2519572
iqr(50,90)	.2113189
iqr(10,50)	.0406383
variance	.0600823

```
covariates: yeduc expft c.expft#c.expft
```

Could also type: `cdist decomp, reverse`

Approximation error:

```
. cdist lincom (obs0-obs1) (obs0-fit0) (obs1-fit1)
```

lnwage	Coefficient
obs0-obs1	
mean	-.1339592
iqr(10,90)	.1815107
iqr(50,90)	.1909876
iqr(10,50)	-.0094769
variance	.0539298
obs0-fit0	
mean	.0003182
iqr(10,90)	-.0173649
iqr(50,90)	.0028495
iqr(10,50)	-.0202144
variance	.0003062
obs1-fit1	
mean	.0000793
iqr(10,90)	.0408902
iqr(50,90)	.0109894
iqr(10,50)	.0299008
variance	.0053953

```
covariates: yeduc expft c.expft#c.expft
```

Estimate counterfactuals based on pooled sample

```
. cdist lnwage yeduc c.expft##c.expft [pw=weight], ///
>     by(public) method(qr) pooled ///
>     statistics(mean iqr(10 90) iqr(50 90) iqr(10 50) variance)
group 0: fitting models 0%....20%....40%....60%....80%....100%
enumerating predictions ... done
group 1: fitting models 0%....20%....40%....60%....80%....100%
enumerating predictions ... done
```

```
Counterfactual distribution estimation          Number of obs      =      5,458
                                                Pooled             =        yes
Group 0: public = 0                             N of obs 0         =      4,184
Group 1: public = 1                             N of obs 1         =      1,274
                                                Estimation method  =         qr
                                                Grid size          =        100
```

lnwage	Coefficient
obs0	
mean	2.732109
iqr(10,90)	1.242421
iqr(50,90)	.6530616
iqr(10,50)	.5893595
variance	.250799
fit0	
mean	2.731791
iqr(10,90)	1.259786
iqr(50,90)	.6502121
iqr(10,50)	.6095739
variance	.2504928

Estimate counterfactuals based on pooled sample

adj0	
mean	2.762694
iqr(10,90)	1.280105
iqr(50,90)	.6583398
iqr(10,50)	.6217654
variance	.2570453
obs1	
mean	2.866068
iqr(10,90)	1.06091
iqr(50,90)	.462074
iqr(10,50)	.5988364
variance	.1968692
fit1	
mean	2.865989
iqr(10,90)	1.02002
iqr(50,90)	.4510847
iqr(10,50)	.5689356
variance	.1914739
adj1	
mean	2.793725
iqr(10,90)	1.015956
iqr(50,90)	.4470208
iqr(10,50)	.5689356
variance	.1923145

covariates: yeduc expft c.expft#c.expft

Decomposition results:

```
. cdist lincom (Difference:fit0-fit1) ///
>              (Explained:fit0-adj0+adj1-fit1) ///
>              (Unexplained:adj0-adj1)
   Difference: fit0-fit1
    Explained: fit0-adj0+adj1-fit1
  Unexplained: adj0-adj1
```

lnwage	Coefficient
Difference	
mean	-.1341982
iqr(10,90)	.2397657
iqr(50,90)	.1991275
iqr(10,50)	.0406383
variance	.0590189
Explained	
mean	-.1031671
iqr(10,90)	-.024383
iqr(50,90)	-.0121915
iqr(10,50)	-.0121915
variance	-.0057119
Unexplained	
mean	-.0310311
iqr(10,90)	.2641487
iqr(50,90)	.2113189
iqr(10,50)	.0528297
variance	.0647308

```
covariates: yeduc expft c.expft#c.expft
```

Could also type: `cdist decomp`

1. Juhn-Murphy-Pierce 1993

2. Approach based on conditional quantiles

3. Approach based on distribution regression

Approach based on 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 distinct value of Y (or, e.g., for a regular grid of Y values) in group 0:

$$F(y_j|X, G = 0) = \Lambda(X\beta_j^0)$$

where y_j, $j = 1, \ldots, J$, are the (ordered) evaluation points and where $\Lambda()$ is a suitable link function. A simple example is to use the logistic function. In this case, β_j^0 is estimated by running a logit model of $I(Y_i \leq y_j)$ on X in group 0.

Approach based on distribution regression

- We can then recover the counterfactual distribution by averaging over predictions from these models

$$\tilde{F}_j^0 = F_{\tilde{Y}^0}(y_j) = \frac{1}{n^1} \sum_{i:G=1} \Lambda(X_i \beta_j^0)$$

  and compute whatever statistic $\tilde{\nu}^0$ we are interested in.
  - Let $\tilde{w}_j^0 = \tilde{F}_j^0 - \tilde{F}_{j-1}^0$. We can then compute the statistic of interest from a dataset composed of outcome values y_j and weights $\tilde{w}_j^0$, $j = 1, \ldots, J$.
- In an analogous way we can compute $\tilde{\nu}^1$, $\widehat{\nu}^0$, and $\widehat{\nu}^1$ and obtain decompositions as described above.

Approach based on distribution regression

- Advantages of distribution regression over the quantile regression approach:
  - Less computational burden / faster.
  - Also works well with discrete outcome variables (or with outcome variables that are affected by heaping).
  - Analytic standard errors potentially easier to derive.
- For both procedures it is unclear how a detailed decomposition could be implemented (apart from the stepwise approach).
- Stata implementations:
  - `cdeco` with option `method(logit)` or `method(probit)`
  - `cdist` with option `method(logit)` (or omitting the option)

Example

Estimate counterfactuals

```
. cdist lnwage yeduc c.expft##c.expft [pw=weight], ///
>     by(public) vce(bootstrap, cluster(psu) strata(strata)) ///
>     statistics(mean iqr(10 90) iqr(50 90) iqr(10 50) variance)
(running cdist on estimation sample)
Bootstrap replications (50): .........10.........20.........30.........40.........50 done
Counterfactual distribution estimation
Number of strata = 15                             Number of obs     =      5,458
                                                  Replications      =         50
                                                  Pooled            =         no
Group 0: public = 0                               N of obs 0        =      4,184
Group 1: public = 1                               N of obs 1        =      1,274
                                                  Estimation method =      logit
                                                  Grid size         =        100
                                  (Replications based on 2,036 clusters in psu)
```

lnwage	Observed coefficient	Bootstrap std. err.	z	P>|z|	Normal-based [95% conf.	interval]
obs0						
mean	2.732109	.0150032	182.10	0.000	2.702703	2.761515
iqr(10,90)	1.242421	.0344752	36.04	0.000	1.174851	1.309991
iqr(50,90)	.6530616	.0257307	25.38	0.000	.6026304	.7034928
iqr(10,50)	.5893595	.0190542	30.93	0.000	.5520139	.6267052
variance	.250799	.0106892	23.46	0.000	.2298485	.2717495
fit0						
mean	2.746857	.0150586	182.41	0.000	2.717343	2.776371
iqr(10,90)	1.249085	.0329118	37.95	0.000	1.184579	1.313591
iqr(50,90)	.6670258	.0271929	24.53	0.000	.6137288	.7203228
iqr(10,50)	.5820589	.016753	34.74	0.000	.5492237	.6148941
variance	.2584741	.0107833	23.97	0.000	.2373393	.2796089

Example

Estimate counterfactuals

adj0						
mean	2.875583	.0210684	136.49	0.000	2.83429	2.916876
iqr(10,90)	1.327446	.039338	33.74	0.000	1.250345	1.404547
iqr(50,90)	.6934347	.0320959	21.61	0.000	.630528	.7563414
iqr(10,50)	.634011	.0234155	27.08	0.000	.5881175	.6799046
variance	.2756854	.0139956	19.70	0.000	.2482545	.3031162
obs1						
mean	2.866068	.023606	121.41	0.000	2.819801	2.912335
iqr(10,90)	1.06091	.0578079	18.35	0.000	.9476091	1.174212
iqr(50,90)	.462074	.0408853	11.30	0.000	.3819404	.5422077
iqr(10,50)	.5988364	.0474808	12.61	0.000	.5057758	.691897
variance	.1968692	.0174649	11.27	0.000	.1626386	.2310998
fit1						
mean	2.878078	.023514	122.40	0.000	2.831992	2.924165
iqr(10,90)	1.059537	.0632833	16.74	0.000	.935504	1.18357
iqr(50,90)	.4619973	.0464755	9.94	0.000	.370907	.5530876
iqr(10,50)	.5975397	.0478598	12.49	0.000	.5037363	.6913431
variance	.1938113	.0178147	10.88	0.000	.1588951	.2287274
adj1						
mean	2.789151	.0160864	173.39	0.000	2.757622	2.82068
iqr(10,90)	.9551249	.0456049	20.94	0.000	.865741	1.044509
iqr(50,90)	.4319139	.0220012	19.63	0.000	.3887923	.4750354
iqr(10,50)	.523211	.0458837	11.40	0.000	.4332807	.6131413
variance	.1752892	.0186573	9.40	0.000	.1387215	.2118568

covariates: yeduc expft c.expft#c.expft

Decomposition with private sector wage structure as reference:

```
. cdist decomp
        Delta: fit0 - fit1
        Chars: fit0 - adj0
        Coefs: adj0 - fit1
```

(Replications based on 2,036 clusters in psu)

lnwage	Observed coefficient	Bootstrap std. err.	z	P>|z|	Normal-based [95% conf.	interval]
Delta						
mean	-.1312213	.0292517	-4.49	0.000	-.1885535	-.0738892
iqr(10,90)	.1895478	.0695029	2.73	0.006	.0533247	.3257709
iqr(50,90)	.2050285	.048191	4.25	0.000	.1105759	.2994812
iqr(10,50)	-.0154808	.0510892	-0.30	0.762	-.1156138	.0846523
variance	.0646628	.0186697	3.46	0.001	.0280709	.1012547
Chars						
mean	-.128726	.0201552	-6.39	0.000	-.1682295	-.0892225
iqr(10,90)	-.078361	.0417923	-1.88	0.061	-.1602724	.0035504
iqr(50,90)	-.0264089	.0395551	-0.67	0.504	-.1039354	.0511176
iqr(10,50)	-.0519521	.017997	-2.89	0.004	-.0872256	-.0166786
variance	-.0172113	.0085588	-2.01	0.044	-.0339862	-.0004364
Coefs						
mean	-.0024954	.0216075	-0.12	0.908	-.0448452	.0398545
iqr(10,90)	.2679088	.0739714	3.62	0.000	.1229276	.41289
iqr(50,90)	.2314374	.0581914	3.98	0.000	.1173844	.3454905
iqr(10,50)	.0364714	.0526682	0.69	0.489	-.0667564	.1396991
variance	.0818741	.0204653	4.00	0.000	.0417629	.1219853

covariates: yeduc expft c.expft#c.expft

Decomposition with public sector wage structure as reference:

```
. cdist decomp, reverse
        Delta: fit0 - fit1
        Chars: adj1 - fit1
        Coefs: fit0 - adj1
                                        (Replications based on 2,036 clusters in psu)
```

lnwage	Observed coefficient	Bootstrap std. err.	z	P>|z|	Normal-based [95% conf.	interval]
Delta						
mean	-.1312213	.0292517	-4.49	0.000	-.1885535	-.0738892
iqr(10,90)	.1895478	.0695029	2.73	0.006	.0533247	.3257709
iqr(50,90)	.2050285	.048191	4.25	0.000	.1105759	.2994812
iqr(10,50)	-.0154808	.0510892	-0.30	0.762	-.1156138	.0846523
variance	.0646628	.0186697	3.46	0.001	.0280709	.1012547
Chars						
mean	-.0889272	.0189864	-4.68	0.000	-.12614	-.0517145
iqr(10,90)	-.1044121	.0490264	-2.13	0.033	-.2005021	-.008322
iqr(50,90)	-.0300834	.0418087	-0.72	0.472	-.112027	.0518602
iqr(10,50)	-.0743287	.0415146	-1.79	0.073	-.1556958	.0070384
variance	-.0185221	.0084378	-2.20	0.028	-.0350598	-.0019844
Coefs						
mean	-.0422941	.0197941	-2.14	0.033	-.0810898	-.0034983
iqr(10,90)	.2939599	.0528631	5.56	0.000	.1903501	.3975696
iqr(50,90)	.235112	.0309342	7.60	0.000	.1744821	.2957418
iqr(10,50)	.0588479	.0476024	1.24	0.216	-.034451	.1521468
variance	.0831849	.0193213	4.31	0.000	.0453158	.1210539

```
covariates: yeduc expft c.expft#c.expft
```

Approximation error:

```
. cdist lincom (obs0-obs1) (obs0-fit0) (obs1-fit1)
                                (Replications based on 2,036 clusters in psu)
```

lnwage	Observed coefficient	Bootstrap std. err.	z	P>|z|	Normal-based [95% conf.	interval]
obs0-obs1						
mean	-.1339592	.0292303	-4.58	0.000	-.1912495	-.0766689
iqr(10,90)	.1815107	.0660189	2.75	0.006	.0521159	.3109054
iqr(50,90)	.1909876	.0437096	4.37	0.000	.1053184	.2766567
iqr(10,50)	-.0094769	.0507861	-0.19	0.852	-.1090159	.0900621
variance	.0539298	.0184469	2.92	0.003	.0177746	.090085
obs0-fit0						
mean	-.0147477	.0007782	-18.95	0.000	-.0162729	-.0132225
iqr(10,90)	-.0066636	.010126	-0.66	0.510	-.0265101	.013183
iqr(50,90)	-.0139642	.0084083	-1.66	0.097	-.0304441	.0025158
iqr(10,50)	.0073006	.0058572	1.25	0.213	-.0041792	.0187804
variance	-.0076751	.0018863	-4.07	0.000	-.0113721	-.0039781
obs1-fit1						
mean	-.0120098	.0008344	-14.39	0.000	-.0136452	-.0103744
iqr(10,90)	.0013735	.0194013	0.07	0.944	-.0366524	.0393994
iqr(50,90)	.0000768	.0112647	0.01	0.995	-.0220017	.0221552
iqr(10,50)	.0012968	.0148375	0.09	0.930	-.0277841	.0303777
variance	.0030579	.0022359	1.37	0.171	-.0013244	.0074402

```
covariates: yeduc expft c.expft#c.expft
```

Example

Estimate counterfactuals based on pooled sample

```
. cdist lnwage yeduc c.expft##c.expft [pw=weight], ///
>     by(public) vce(bootstrap, cluster(psu) strata(strata)) pooled ///
>     statistics(mean iqr(10 90) iqr(50 90) iqr(10 50) variance)
(running cdist on estimation sample)
Bootstrap replications (50): .........10.........20.........30.........40.........50 done
Counterfactual distribution estimation
Number of strata = 15                             Number of obs     =      5,458
                                                  Replications      =         50
                                                  Pooled            =        yes
Group 0: public = 0                               N of obs 0        =      4,184
Group 1: public = 1                               N of obs 1        =      1,274
                                                  Estimation method =      logit
                                                  Grid size         =        100
                                  (Replications based on 2,036 clusters in psu)
```

lnwage	Observed coefficient	Bootstrap std. err.	z	P>|z|	Normal-based [95% conf.	interval]
obs0						
mean	2.732109	.0129398	211.14	0.000	2.706748	2.757471
iqr(10,90)	1.242421	.0294319	42.21	0.000	1.184736	1.300107
iqr(50,90)	.6530616	.0276282	23.64	0.000	.5989114	.7072118
iqr(10,50)	.5893595	.0149937	39.31	0.000	.5599723	.6187467
variance	.250799	.0083359	30.09	0.000	.234461	.2671369
fit0						
mean	2.746857	.0129563	212.01	0.000	2.721463	2.772251
iqr(10,90)	1.249085	.0297404	42.00	0.000	1.190795	1.307375
iqr(50,90)	.6670258	.0276863	24.09	0.000	.6127616	.7212901
iqr(10,50)	.5820589	.0139332	41.78	0.000	.5547504	.6093674
variance	.2584741	.0076506	33.78	0.000	.2434792	.2734689

Example

Estimate counterfactuals based on pooled sample

adj0						
mean	2.777747	.0121977	227.73	0.000	2.75384	2.801654
iqr(10,90)	1.258292	.033836	37.19	0.000	1.191975	1.32461
iqr(50,90)	.6729186	.0300941	22.36	0.000	.6139351	.731902
iqr(10,50)	.5853739	.0189371	30.91	0.000	.5482578	.62249
variance	.2656264	.0079471	33.42	0.000	.2500504	.2812023
obs1						
mean	2.866068	.0254851	112.46	0.000	2.816118	2.916018
iqr(10,90)	1.06091	.0651645	16.28	0.000	.9331904	1.188631
iqr(50,90)	.462074	.0418605	11.04	0.000	.3800289	.5441191
iqr(10,50)	.5988364	.0552766	10.83	0.000	.4904963	.7071765
variance	.1968692	.0201122	9.79	0.000	.1574499	.2362884
fit1						
mean	2.878078	.0254437	113.12	0.000	2.82821	2.927947
iqr(10,90)	1.059537	.0679712	15.59	0.000	.9263158	1.192758
iqr(50,90)	.4619973	.0408366	11.31	0.000	.3819591	.5420355
iqr(10,50)	.5975397	.0563505	10.60	0.000	.4870947	.7079846
variance	.1938113	.0205897	9.41	0.000	.1534562	.2341664
adj1						
mean	2.810491	.0243257	115.54	0.000	2.762813	2.858168
iqr(10,90)	.988214	.0545556	18.11	0.000	.8812871	1.095141
iqr(50,90)	.4572902	.0271629	16.84	0.000	.4040519	.5105284
iqr(10,50)	.5309238	.0518094	10.25	0.000	.4293794	.6324683
variance	.1811762	.0205492	8.82	0.000	.1409006	.2214518

covariates: yeduc expft c.expft#c.expft

Decomposition results:

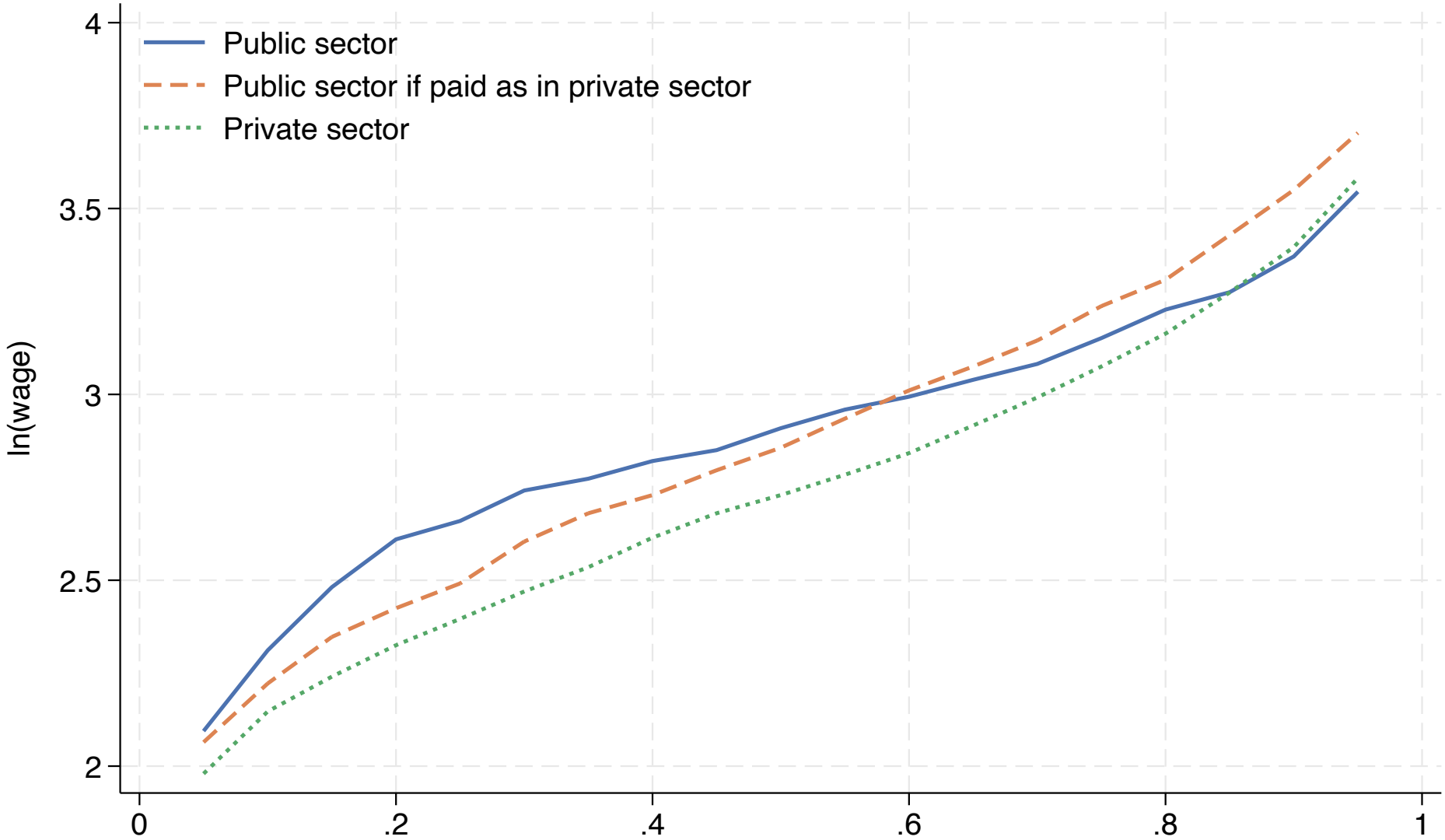
```
. cdist decomp
        Delta: fit0 - fit1
        Chars: fit0 - adj0 + adj1 - fit1
        Coefs: adj0 - adj1
                                (Replications based on 2,036 clusters in psu)
```

lnwage	Observed coefficient	Bootstrap std. err.	z	P>|z|	Normal-based [95% conf.	interval]
Delta						
mean	-.1312213	.0295784	-4.44	0.000	-.1891939	-.0732487
iqr(10,90)	.1895478	.0723926	2.62	0.009	.0476608	.3314347
iqr(50,90)	.2050285	.0486416	4.22	0.000	.1096928	.3003643
iqr(10,50)	-.0154808	.0587826	-0.26	0.792	-.1306926	.099731
variance	.0646628	.0220069	2.94	0.003	.02153	.1077955
Chars						
mean	-.0984777	.0132868	-7.41	0.000	-.1245193	-.072436
iqr(10,90)	-.0805306	.0529143	-1.52	0.128	-.1842407	.0231794
iqr(50,90)	-.0105999	.0363903	-0.29	0.771	-.0819235	.0607238
iqr(10,50)	-.0699308	.035008	-2.00	0.046	-.1385453	-.0013163
variance	-.0197874	.0082933	-2.39	0.017	-.036042	-.0035328
Coefs						
mean	-.0327437	.0256445	-1.28	0.202	-.083006	.0175186
iqr(10,90)	.2700784	.0590796	4.57	0.000	.1542845	.3858724
iqr(50,90)	.2156284	.0369729	5.83	0.000	.1431628	.288094
iqr(10,50)	.05445	.0552288	0.99	0.324	-.0537964	.1626965
variance	.0844502	.0231367	3.65	0.000	.0391031	.1297973

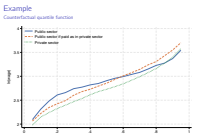
```
covariates: yeduc expft c.expft#c.expft
```

Example

Counterfactual quantile function



Example




```
. cdist lnwage yeduc c.expft##c.expft [pw=weight], by(public) quantile(#19)
```

(*output omitted*)

```
. coefplot (, keep(fit1:)) (, keep(adj0:)) (, keep(fit0:)) ///
>      , at(_coef) noci recast(line) ytitle(ln(wage)) ///
>        plotlabels("Public sector" ///
>                   "Public sector if paid as in private sector" ///
>                   "Private sector")
```

Exercise 8

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.

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.