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)

3. Index problem & transformation problem

Some issues with the Oaxaca-Blinder decomposition

- The OB decomposition seems useful and easy to understand, but there are several complications we need to discuss.
  - **The index problem**
  - **The transformation problem / base category problem**
  - Functional form

Contents

The index problem

- The choice of the counterfactual distribution used for the decomposition is consequential for the results.
- Up to now we used $F_{Y^0|G=1}$, that is, we asked: "How would the distribution of wages of women look like if they were paid like men?"
- This leads to decomposition

$$\begin{aligned}\Delta^\mu &= (\mathrm{E}(X|G=0) - \mathrm{E}(X|G=1))\beta^0 + \mathrm{E}(X|G=1)(\beta^0 - \beta^1) \\ &= \Delta^\mu_X + \Delta^\mu_S\end{aligned}$$

since

$$\mu\left(F_{Y^0|G=1}\right) = \mathrm{E}(X|G=1)\beta^0$$

- We might as well use another counterfactual, and this would change our results!

The index problem

- For example, we could base the decomposition on $F_{Y^1|G=0}$.
  - "How would the distribution of wages of men look like if they were paid like women?"
- Since

$$\mu(F_{Y^1|G=0}) = \mathrm{E}(X|G=0)\beta^1$$

the decomposition would then be

$$\begin{aligned}
\Delta^\mu &= \mathrm{E}(X|G=0)\beta^0 - \mathrm{E}(X|G=1)\beta^1 \\
&= \mathrm{E}(X|G=0)\beta^0 - \mathrm{E}(X|G=0)\beta^1 + \mathrm{E}(X|G=0)\beta^1 - \mathrm{E}(X|G=1)\beta^1 \\
&= \mathrm{E}(X|G=0)(\beta^0 - \beta^1) + (\mathrm{E}(X|G=0) - \mathrm{E}(X|G=1))\beta^1 \\
&= \Delta^\mu_S + \Delta^\mu_X
\end{aligned}$$

The index problem

- What is the difference between these two variants of the decomposition?
- If using $F_{Y^0|G=1}$:

$\widehat{\Delta}^{\mu}_{X} = (\bar{X}^0 - \bar{X}^1)\widehat{\beta}^0$	How much lower would average wages of men be, if they had the same endowments as women?
$\widehat{\Delta}^{\mu}_{S} = \bar{X}^1(\widehat{\beta}^0 - \widehat{\beta}^1)$	How much higher would average wages of women be, if they were paid like men?

- If using $F_{Y^1|G=0}$:

$\widehat{\Delta}^{\mu}_{X} = (\bar{X}^0 - \bar{X}^1)\widehat{\beta}^1$	How much higher would average wages of women be, if they had the same endowments as men?
$\widehat{\Delta}^{\mu}_{S} = \bar{X}^0(\widehat{\beta}^0 - \widehat{\beta}^1)$	How much lower would average wages of men be, if they were paid like women?

1. The index problem
   - The three-fold decomposition
   - Nondiscriminatory wage structure
   - Example analysis
   - Relation to treatment effects
   - Using a common characteristics distribution

2. The transformation problem
   - Transformation of covariates
   - Base level of categorical covariates
   - Normalization to solve the base level problem
   - "Industry decomposition"

The three-fold decomposition

- This difference in interpretation suggests yet another approach: the three-fold decomposition (see Winsborough and Dickinson 1971).
- From the view of women:

$$
\begin{aligned}
\widehat{\Delta}^{\mu} &= \widehat{\Delta}^{\mu}_{X} + \widehat{\Delta}^{\mu}_{S} + \widehat{\Delta}^{\mu}_{XS} \\
&= (\bar{X}^{0} - \bar{X}^{1})\widehat{\beta}^{1} + \bar{X}^{1}(\beta^{0} - \beta^{1}) + (\bar{X}^{0} - \bar{X}^{1})(\beta^{0} - \beta^{1})
\end{aligned}
$$

- From the view of men:

$$
\begin{aligned}
\widehat{\Delta}^{\mu} &= \widehat{\Delta}^{\mu}_{X} + \widehat{\Delta}^{\mu}_{S} + \widehat{\Delta}^{\mu}_{XS} \\
&= (\bar{X}^{0} - \bar{X}^{1})\widehat{\beta}^{0} + \bar{X}^{0}(\beta^{0} - \beta^{1}) + (\bar{X}^{0} - \bar{X}^{1})(\beta^{1} - \beta^{0})
\end{aligned}
$$

- The first two terms illustrate how wages of one group are affected if we change endowments or coefficients to the level of the other group.
- Such a decomposition is consistent in the sense that both terms refer to the same group, either to women or to men.

The three-fold decomposition

- The last term is an interaction term accounting for the fact that differences in endowments and coefficients exist simultaneously between the two groups.
- It captures whether there is a "double disadvantage" for women (or a "double advantage" for men) in the sense that men's coefficients are larger than women's coefficients for covariates for which women have lower levels than men, or whether differences in coefficients and in covariate levels offset each other.
- From the view of women, the interaction term will be positive in case of "double disadvantage" and negative in the offsetting scenario.
- From the view of men, the interaction term will be negative in case of "double advantage" and positive in the offsetting scenario.

Nondiscriminatory wage structure

- Yet another approach is to think of a "non-discriminatory" potential outcome Y^* defined as

$$Y^* = m^*(X, \epsilon) = X\beta^* + \epsilon$$

- The relevant counterfactuals then are $F_{Y^*|G=0}$ for men and $F_{Y^*|G=1}$ for women with

$$\mu(F_{Y^*|G=0}) = \mathrm{E}(X|G=0)\beta^* \quad \text{and} \quad \mu(F_{Y^*|G=1}) = \mathrm{E}(X|G=1)\beta^*$$

- The decomposition then is

$$\begin{aligned}
\widehat{\Delta}^\mu &= \bar{X}^0\widehat{\beta}^0 - \bar{X}^1\widehat{\beta}^1 \\
&= \bar{X}^0\widehat{\beta}^* - \bar{X}^1\widehat{\beta}^* + \bar{X}^0\widehat{\beta}^0 - \bar{X}^0\widehat{\beta}^* + \bar{X}^1\widehat{\beta}^* - \bar{X}^1\widehat{\beta}^1 \\
&= (\bar{X}^0 - \bar{X}^1)\widehat{\beta}^* + \left(\bar{X}^0(\widehat{\beta}^0 - \widehat{\beta}^*) + \bar{X}^1(\widehat{\beta}^* - \widehat{\beta}^1)\right) \\
&= \widehat{\Delta}^\mu_X + \widehat{\Delta}^\mu_S
\end{aligned}$$

Nondiscriminatory wage structure

- The unexplained part Δ_S^μ can further be subdivided into

$$\widehat{\Delta}_{S^0}^\mu = \bar{X}^0(\widehat{\beta}^0 - \widehat{\beta}^*) \quad \text{(“discrimination” in favor of men)}$$

and

$$\widehat{\Delta}_{S^1}^\mu = \bar{X}^1(\widehat{\beta}^* - \widehat{\beta}^1) \quad \text{(“discrimination” against women)}$$

- How should the “non-discriminatory” β^* be determined?
- Two special cases:
  - If $\beta^* = \beta^0$, then the wage structure of men is viewed as non-discriminatory and we end up with our first decomposition variant.
  - If $\beta^* = \beta^1$, then the wage structure of women is viewed as non-discriminatory and we end up with our second decomposition variant.

Nondiscriminatory wage structure

- Let W be a diagonal matrix of weights, such that

$$\beta^* = W\beta^0 + (I - W)\beta^1$$

- The two special cases above then correspond to $W = I$ and $W = 0$.
- Other proposals are:
  - Reimers (1983): Set $W = 0.5I$ such that

$$\widehat{\beta}^* = 0.5\widehat{\beta}^0 + 0.5\widehat{\beta}^1$$

  - Cotton (1988): Set $W = \widehat{p}^0 I$ where $p^0 = \Pr(G = 0)$ such that

$$\widehat{\beta}^* = \widehat{\Pr}(G = 0)\widehat{\beta}^0 + \widehat{\Pr}(G = 1)\widehat{\beta}^1$$

  - Neumark (1988), Oaxaca and Ransom (1994): Set $W = \Omega$ where

$$\Omega = (X^{0\prime}X^0 + X^{1\prime}X^1)^{-1}X^{0\prime}X^0$$

    which is equivalent to estimating β^* by a pooled regression over both groups (without distinguishing the groups), that is,

$$\widehat{\beta}^* = (X'X)^{-1}X'Y$$

Nondiscriminatory wage structure

- The last proposal (Neumark 1988, Oaxaca and Ransom 1994) seems attractive, but is affected by omitted variable bias with the consequence that some of the unexplained group difference is moved into the explained part (see Jann 2008):
  - Assume a simple model of log wages (Y) on experience (X) with sex-specific intercepts α^0 and α^1 due to discrimination, that is

$$Y = \begin{cases} \alpha^0 + \beta X + \epsilon & \text{if } G = 0 \\ \alpha^1 + \beta X + \epsilon & \text{if } G = 1 \end{cases}$$

  - Let $\alpha^0 = \alpha$ and $\alpha^1 = \alpha + \delta$, where δ is the discrimination parameter. The model can then be expressed as

$$Y = \alpha + \beta X + \delta G + \epsilon$$

  - However, in the $W = \Omega$ approach we estimate

$$Y = \alpha^* + \beta^* X + \epsilon^*$$

Nondiscriminatory wage structure

- Following from the theory on omitted variables, the explained part of the decomposition can then be written as

$$\begin{aligned}\Delta_X^\mu &= (E(X|G=0) - E(X|G=1))\beta^* \\ &= (E(X|G=0) - E(X|G=1))\left\{\beta + \delta\frac{\mathrm{Cov}(X,G)}{\mathrm{Var}(X)}\right\}\end{aligned}$$

- If men on average have more experience than women, then the covariance between X and G is negative and the explained part of the decomposition gets overstated (given $\beta > 0$ and $\delta < 0$).
- In essence, the difference in wages between men and women is partially explained by, well, gender.

- Hence, a final proposal is:
  - Fortin (2008), Jann (2008): estimate β^* by a pooled regression over both groups controlling group membership, that is

$$\widehat{\beta}^* = ((X,G)'(X,G))^{-1}(X,G)'Y$$

  - In this case $\widehat{\Delta}_S^\mu = -\widehat{\delta}$, where $\widehat{\delta}$ is the coefficient of G in the pooled regression. (A distinction of $\Delta_{S^0}^\mu$ and $\Delta_{S^1}^\mu$ does not make sense in this case.)

Regression models

```
. svy: regress lnwage yeduc expft expft2 if sex==1
```
(*output omitted*)
```
. estimates store male
. svy: regress lnwage yeduc expft expft2 if sex==2
```
(*output omitted*)
```
. estimates store female
. svy: regress lnwage yeduc expft expft2 if sex<.
```
(*output omitted*)
```
. estimates store omega
. svy: regress lnwage yeduc expft expft2 i.sex
```
(*output omitted*)
```
. estimates store pooled
. esttab male female omega pooled, not nogap mtitle nonumber nostar varwidth(14)
```

	male	female	omega	pooled
yeduc	0.0829	0.0789	0.0812	0.0809
expft	0.0357	0.0313	0.0352	0.0325
expft2	-0.000593	-0.000541	-0.000557	-0.000534
1.sex				0
2.sex				-0.123
_cons	1.430	1.404	1.393	1.488
N	2642	2820	5462	5462

Using the male coefficients ($W = I$)

```
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) svy weight(1)
Blinder-Oaxaca decomposition

Number of strata =     15                      Number of obs     =      5,462
Number of PSUs   =  2,037                      Population size   = 12,152,217
                                               Design df         =      2,022
                                               Model             =     linear
Group 1: sex = 1                               N of obs 1        =      2,642
Group 2: sex = 2                               N of obs 2        =      2,820

    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)

------------------------------------------------------------------------------
             |             Linearized
      lnwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.862735   .0162749   175.90   0.000     2.830818    2.894652
     group_2 |   2.657428   .0149503   177.75   0.000     2.628108    2.686747
  difference |   .2053074   .0204701    10.03   0.000     .1651628     .245452
   explained |   .0904872   .0151554     5.97   0.000     .0607654     .120209
 unexplained |   .1148202   .0211416     5.43   0.000     .0733585    .1562819
-------------+----------------------------------------------------------------
explained    |
       yeduc |  -.0191954   .0096981    -1.98   0.048    -.0382146   -.0001761
  experience |   .1096826   .0129638     8.46   0.000     .0842587    .1351065
-------------+----------------------------------------------------------------
unexplained  |
       yeduc |   .0520745   .0969211     0.54   0.591    -.1380012    .2421502
  experience |   .0370006   .0424742     0.87   0.384    -.0462972    .1202985
       _cons |   .0257451   .1178852     0.22   0.827     -.205444    .2569342
------------------------------------------------------------------------------
experience: expft expft2
```

Using the female coefficients ($W = 0$)

```
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) svy weight(0)

Blinder-Oaxaca decomposition

Number of strata =     15                     Number of obs     =      5,462
Number of PSUs   =  2,037                     Population size   = 12,152,217
                                              Design df         =      2,022
                                              Model             =     linear
Group 1: sex = 1                              N of obs 1        =      2,642
Group 2: sex = 2                              N of obs 2        =      2,820

    explained: (X1 - X2) * b2
  unexplained: X1 * (b1 - b2)
```

lnwage	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
overall						
group_1	2.862735	.0162749	175.90	0.000	2.830818	2.894652
group_2	2.657428	.0149503	177.75	0.000	2.628108	2.686747
difference	.2053074	.0204701	10.03	0.000	.1651628	.245452
explained	.0739469	.0135139	5.47	0.000	.0474443	.1004495
unexplained	.1313605	.0184168	7.13	0.000	.0952426	.1674784
explained						
yeduc	-.0182642	.0092275	-1.98	0.048	-.0363606	-.0001678
experience	.0922111	.0107177	8.60	0.000	.0711922	.11323
unexplained						
yeduc	.0511433	.0951882	0.54	0.591	-.1355338	.2378204
experience	.0544721	.0534284	1.02	0.308	-.0503084	.1592526
_cons	.0257451	.1178852	0.22	0.827	-.205444	.2569342

```
experience: expft expft2
```

Threefold decomposition from the view of females

```
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) svy threefold

Blinder-Oaxaca decomposition

Number of strata =    15                    Number of obs     =      5,462
Number of PSUs   = 2,037                    Population size   = 12,152,217
                                            Design df         =      2,022
                                            Model             =     linear
Group 1: sex = 1                            N of obs 1        =      2,642
Group 2: sex = 2                            N of obs 2        =      2,820

   endowments: (X1 - X2) * b2
 coefficients: X2 * (b1 - b2)
  interaction: (X1 - X2) * (b1 - b2)

------------------------------------------------------------------------------
             |             Linearized
      lnwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.862735   .0162749   175.90   0.000     2.830818    2.894652
     group_2 |   2.657428   .0149503   177.75   0.000     2.628108    2.686747
  difference |   .2053074   .0204701    10.03   0.000     .1651628     .245452
  endowments |   .0739469   .0135139     5.47   0.000     .0474443    .1004495
coefficients |   .1148202   .0211416     5.43   0.000     .0733585    .1562819
 interaction |   .0165403   .0131549     1.26   0.209    -.0092582    .0423388
-------------+----------------------------------------------------------------
endowments   |
       yeduc |  -.0182642   .0092275    -1.98   0.048    -.0363606   -.0001678
  experience |   .0922111   .0107177     8.60   0.000     .0711922      .11323
-------------+----------------------------------------------------------------
coefficients |
       yeduc |   .0520745   .0969211     0.54   0.591    -.1380012    .2421502
  experience |   .0370006   .0424742     0.87   0.384    -.0462972    .1202985
       _cons |   .0257451   .1178852     0.22   0.827     -.205444    .2569342
-------------+----------------------------------------------------------------
interaction  |
       yeduc |  -.0009312   .0017947    -0.52   0.604    -.0044509    .0025885
  experience |   .0174715   .0135218     1.29   0.196    -.0090467    .0439896
------------------------------------------------------------------------------
experience: expft expft2
```

Threefold decomposition from the view of males

```
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) svy threefold(reverse)

Blinder-Oaxaca decomposition

Number of strata =     15                          Number of obs     =      5,462
Number of PSUs   = 2,037                           Population size   = 12,152,217
                                                   Design df         =      2,022
                                                   Model             =     linear
Group 1: sex = 1                                   N of obs 1        =      2,642
Group 2: sex = 2                                   N of obs 2        =      2,820

   endowments: (X1 - X2) * b1
 coefficients: X1 * (b1 - b2)
  interaction: (X1 - X2) * (b2 - b1)

------------------------------------------------------------------------------
             |             Linearized
      lnwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.862735   .0162749   175.90   0.000     2.830818    2.894652
     group_2 |   2.657428   .0149503   177.75   0.000     2.628108    2.686747
  difference |   .2053074   .0204701    10.03   0.000     .1651628     .245452
  endowments |   .0904872   .0151554     5.97   0.000     .0607654     .120209
coefficients |   .1313605   .0184168     7.13   0.000     .0952426    .1674784
 interaction |  -.0165403   .0131549    -1.26   0.209    -.0423388    .0092582
-------------+----------------------------------------------------------------
endowments   |
       yeduc |  -.0191954   .0096981    -1.98   0.048    -.0382146   -.0001761
  experience |   .1096826   .0129638     8.46   0.000     .0842587    .1351065
-------------+----------------------------------------------------------------
coefficients |
       yeduc |   .0511433   .0951882     0.54   0.591    -.1355338    .2378204
  experience |   .0544721   .0534284     1.02   0.308    -.0503084    .1592526
       _cons |   .0257451   .1178852     0.22   0.827     -.205444    .2569342
-------------+----------------------------------------------------------------
interaction  |
       yeduc |   .0009312   .0017947     0.52   0.604    -.0025885    .0044509
  experience |  -.0174715   .0135218    -1.29   0.196    -.0439896    .0090467
------------------------------------------------------------------------------
experience: expft expft2
```

Using average coefficients ($W = 0.5I$)

```
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) svy weight(0.5)
Blinder-Oaxaca decomposition

Number of strata =     15                     Number of obs      =      5,462
Number of PSUs   =  2,037                     Population size    = 12,152,217
                                              Design df          =      2,022
                                              Model              =     linear
Group 1: sex = 1                              N of obs 1         =      2,642
Group 2: sex = 2                              N of obs 2         =      2,820

    explained: (X1 - X2) * b
  unexplained: X1 * (b1 - b) + X2 * (b - b2)
               with b = .5 * b1  + (1 - .5) * b2

------------------------------------------------------------------------------
             |             Linearized
      lnwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.862735   .0162749   175.90   0.000     2.830818    2.894652
     group_2 |   2.657428   .0149503   177.75   0.000     2.628108    2.686747
  difference |   .2053074   .0204701    10.03   0.000     .1651628     .245452
   explained |   .0822171   .0127629     6.44   0.000     .0571872     .107247
 unexplained |   .1230903   .0187033     6.58   0.000     .0864107      .15977
-------------+----------------------------------------------------------------
explained    |
       yeduc |  -.0187298   .0094231    -1.99   0.047    -.0372097   -.0002498
  experience |   .1009468   .0097855    10.32   0.000     .0817562    .1201375
-------------+----------------------------------------------------------------
unexplained  |
       yeduc |   .0516089   .0960543     0.54   0.591    -.1367669    .2399847
  experience |   .0457363   .0477872     0.96   0.339     -.047981    .1394537
       _cons |   .0257451   .1178852     0.22   0.827     -.205444    .2569342
------------------------------------------------------------------------------
experience: expft expft2
```

Using weighted average ($W = \widehat{p}^0 I$)

```
. summarize sex [aw=weight] if !missing(lnwage,yeduc,expft,expft2)

    Variable |     Obs      Weight        Mean   Std. dev.       Min        Max
-------------+-----------------------------------------------------------------
         sex |   5,462  12152217.3    1.480727   .4996742          1          2
. local p_m = 2 - r(mean)
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) svy weight(`p_m')
Blinder-Oaxaca decomposition
Number of strata =    15                        Number of obs     =      5,462
Number of PSUs   = 2,037                        Population size   = 12,152,217
                                                Design df         =      2,022
                                                Model             =     linear
Group 1: sex = 1                                N of obs 1        =      2,642
Group 2: sex = 2                                N of obs 2        =      2,820
    explained: (X1 - X2) * b
  unexplained: X1 * (b1 - b) + X2 * (b - b2)
               with b = .519273 * b1  + (1 - .519273) * b2
------------------------------------------------------------------------------
             |             Linearized
      lnwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.862735   .0162749   175.90   0.000     2.830818    2.894652
     group_2 |   2.657428   .0149503   177.75   0.000     2.628108    2.686747
  difference |   .2053074   .0204701    10.03   0.000     .1651628     .245452
   explained |   .0825358   .0128009     6.45   0.000     .0574314    .1076402
 unexplained |   .1227716   .0187604     6.54   0.000     .0859798    .1595633
-------------+----------------------------------------------------------------
explained    |
       yeduc |  -.0187477   .0094322    -1.99   0.047    -.0372457   -.0002498
  experience |   .1012836   .0098412    10.29   0.000     .0819837    .1205834
-------------+----------------------------------------------------------------
unexplained  |
       yeduc |   .0516269   .0960877     0.54   0.591    -.1368145    .2400682
  experience |   .0453996   .0475756     0.95   0.340    -.0479027     .138702
       _cons |   .0257451   .1178852     0.22   0.827     -.205444    .2569342
------------------------------------------------------------------------------
experience: expft expft2
```

Using pooled model without controlling group ($W = \Omega$)

```
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) svy omega
Blinder-Oaxaca decomposition

Number of strata =     15                       Number of obs     =      5,462
Number of PSUs   =  2,037                       Population size   = 12,152,217
                                                Design df         =      2,022
                                                Model             =     linear
Group 1: sex = 1                                N of obs 1        =      2,642
Group 2: sex = 2                                N of obs 2        =      2,820

    explained: (X1 - X2) * b
  unexplained: X1 * (b1 - b) + X2 * (b - b2)
               with b from pooled model (without group dummy)
```

lnwage	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
overall						
group_1	2.862735	.0162749	175.90	0.000	2.830818	2.894652
group_2	2.657428	.0149503	177.75	0.000	2.628108	2.686747
difference	.2053074	.0204701	10.03	0.000	.1651628	.245452
explained	.0950124	.0127953	7.43	0.000	.0699191	.1201056
unexplained	.1102951	.0166832	6.61	0.000	.0775771	.143013
explained						
yeduc	-.0187903	.0094532	-1.99	0.047	-.0373293	-.0002513
experience	.1138026	.0099399	11.45	0.000	.0943091	.1332961
unexplained						
yeduc	.0516694	.0960832	0.54	0.591	-.1367629	.2401017
experience	.0328806	.0491595	0.67	0.504	-.0635279	.1292891
_cons	.0257451	.1178852	0.22	0.827	-.205444	.2569342

experience: expft expft2

Using pooled model including group dummy

```
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) svy pooled
Blinder-Oaxaca decomposition

Number of strata =    15                    Number of obs     =      5,462
Number of PSUs   = 2,037                    Population size   = 12,152,217
                                            Design df         =      2,022
                                            Model             =     linear
Group 1: sex = 1                            N of obs 1        =      2,642
Group 2: sex = 2                            N of obs 2        =      2,820

    explained: (X1 - X2) * b
  unexplained: X1 * (b1 - b) + X2 * (b - b2)
               with b from pooled model (including group dummy)

------------------------------------------------------------------------------
             |             Linearized
      lnwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.862735   .0162749   175.90   0.000     2.830818    2.894652
     group_2 |   2.657428   .0149503   177.75   0.000     2.628108    2.686747
  difference |   .2053074   .0204701    10.03   0.000     .1651628     .245452
   explained |   .0824755     .01262     6.54   0.000     .0577259    .1072251
 unexplained |   .1228319   .0185895     6.61   0.000     .0863753    .1592885
-------------+----------------------------------------------------------------
explained    |
       yeduc |  -.0187109   .0094136    -1.99   0.047    -.0371722   -.0002496
  experience |   .1011864   .0096291    10.51   0.000     .0823024    .1200704
-------------+----------------------------------------------------------------
unexplained  |
       yeduc |     .05159   .0960992     0.54   0.591    -.1368738    .2400539
  experience |   .0454968   .0487542     0.93   0.351     -.050117    .1411105
       _cons |   .0257451   .1178852     0.22   0.827     -.205444    .2569342
------------------------------------------------------------------------------
experience: expft expft2
```

Relation to treatment effects

- There is a close relation between some of the above decompositions and the **regression adjustment** estimator (RA) from the treatment effects literature.

- Let males be the "control group" and females be the "treatment group". Let δ^{ATE} be the average treatment effect, δ^{ATT} be the ATE on the treated, and δ^{ATC} be the ATE in the control group. We then get the following results for the unexplained part of the decomposition.

  If using the male coefficients ($W = I$): $\widehat{\Delta}^{\mu}_{S} = -\widehat{\delta}^{ATT}$

  If using the female coefficients ($W = 0$): $\widehat{\Delta}^{\mu}_{S} = -\widehat{\delta}^{ATC}$

  If using the reverse weighted average ($W = \widehat{p}^{1} I$): $\widehat{\Delta}^{\mu}_{S} = -\widehat{\delta}^{ATE}$

Relation to treatment effects

- From this perspective, a reverse weighted average with $W = \widehat{p}^1 I$ such that
$$\widehat{\beta}^* = \widehat{\Pr}(G = 1)\widehat{\beta}^0 + \widehat{\Pr}(G = 0)\widehat{\beta}^1$$
might make sense (although we have not seen it in the literature).
- Furthermore, as noted above, if using a pooled model including a group dummy, then $\widehat{\Delta}^{\mu}_{S} = -\widehat{\delta}$ where $\widehat{\delta}$ is a regression adjustment estimate of $\delta^{ATT} = \delta^{ATC} = \delta^{ATE}$ under the assumption that there is no treatment effect heterogeneity.

Using the male coefficients ($W = I$)

```
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) svy weight(1) nodetail

Blinder-Oaxaca decomposition

Number of strata =     15                         Number of obs     =      5,462
Number of PSUs   = 2,037                          Population size   = 12,152,217
                                                  Design df         =      2,022
                                                  Model             =     linear
Group 1: sex = 1                                  N of obs 1        =      2,642
Group 2: sex = 2                                  N of obs 2        =      2,820

    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)

------------------------------------------------------------------------------
             |             Linearized
      lnwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.862735   .0162749   175.90   0.000     2.830818    2.894652
     group_2 |   2.657428   .0149503   177.75   0.000     2.628108    2.686747
  difference |   .2053074   .0204701    10.03   0.000     .1651628     .245452
   explained |   .0904872   .0151554     5.97   0.000     .0607654     .120209
 unexplained |   .1148202   .0211416     5.43   0.000     .0733585    .1562819
------------------------------------------------------------------------------

. teffects ra (lnwage yeduc expft expft2) (sex) [pw=weight], nolog atet vce(cluster psu)

Treatment-effects estimation                      Number of obs     =      5,462
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
                                     (Std. err. adjusted for 2,037 clusters in psu)
-----------------------------------------------------------------------------------
                  |               Robust
           lnwage | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
------------------+----------------------------------------------------------------
ATET              |
              sex |
(female vs male)  |  -.1148202   .0210755    -5.45   0.000    -.1561274    -.073513
------------------+----------------------------------------------------------------
POmean            |
              sex |
            male  |   2.772248   .0201007   137.92   0.000     2.732851    2.811644
-----------------------------------------------------------------------------------
```

Using the female coefficients ($W = 0$)

```
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) svy weight(0) nodetail
Blinder-Oaxaca decomposition

Number of strata =     15                       Number of obs     =      5,462
Number of PSUs   = 2,037                        Population size   = 12,152,217
                                                Design df         =      2,022
                                                Model             =     linear
Group 1: sex = 1                                N of obs 1        =      2,642
Group 2: sex = 2                                N of obs 2        =      2,820

    explained: (X1 - X2) * b2
  unexplained: X1 * (b1 - b2)

------------------------------------------------------------------------------
             |             Linearized
      lnwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.862735   .0162749   175.90   0.000     2.830818    2.894652
     group_2 |   2.657428   .0149503   177.75   0.000     2.628108    2.686747
  difference |   .2053074   .0204701    10.03   0.000     .1651628     .245452
   explained |   .0739469   .0135139     5.47   0.000     .0474443    .1004495
 unexplained |   .1313605   .0184168     7.13   0.000     .0952426    .1674784
------------------------------------------------------------------------------

. teffects ra (lnwage yeduc expft expft2) (sex) [pw=weight], nolog atet tlevel(1) vce(cluster psu)
Treatment-effects estimation                    Number of obs     =      5,462
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
                                      (Std. err. adjusted for 2,037 clusters in psu)
-----------------------------------------------------------------------------------
                  |               Robust
           lnwage | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
------------------+----------------------------------------------------------------
ATET              |
              sex |
(male vs female)  |   .1313605   .0184164     7.13   0.000      .095265     .167456
------------------+----------------------------------------------------------------
POmean            |
              sex |
          female  |   2.731374   .0167228   163.33   0.000     2.698598    2.764151
-----------------------------------------------------------------------------------
```

Using reverse weighted average ($W = \widehat{p}^1 I$)

```
. quietly summarize sex [aw=weight] if !missing(lnwage,yeduc,expft,expft2)
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) svy weight(`=r(mean)-1') nodetail
Blinder-Oaxaca decomposition
Number of strata =     15                        Number of obs      =      5,462
Number of PSUs   = 2,037                         Population size    = 12,152,217
                                                 Design df          =      2,022
                                                 Model              =     linear
Group 1: sex = 1                                 N of obs 1         =      2,642
Group 2: sex = 2                                 N of obs 2         =      2,820
    explained: (X1 - X2) * b
  unexplained: X1 * (b1 - b) + X2 * (b - b2)
               with b = .480727 * b1  + (1 - .480727) * b2
```

lnwage	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
overall						
group_1	2.862735	.0162749	175.90	0.000	2.830818	2.894652
group_2	2.657428	.0149503	177.75	0.000	2.628108	2.686747
difference	.2053074	.0204701	10.03	0.000	.1651628	.245452
explained	.0818983	.0127299	6.43	0.000	.0569332	.1068633
unexplained	.1234091	.0186494	6.62	0.000	.0868352	.1599831

```
. teffects ra (lnwage yeduc expft expft2) (sex) [pw=weight], nolog vce(cluster psu)
Treatment-effects estimation                     Number of obs      =      5,462
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
                                  (Std. err. adjusted for 2,037 clusters in psu)
```

lnwage	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
ATE						
sex (female vs male)	-.1234091	.0186048	-6.63	0.000	-.1598738	-.0869444
POmean						
sex male	2.819235	.0167258	168.56	0.000	2.786453	2.852017

1. The index problem
   - The three-fold decomposition
   - Nondiscriminatory wage structure
   - Example analysis
   - Relation to treatment effects
   - Using a common characteristics distribution

2. The transformation problem
   - Transformation of covariates
   - Base level of categorical covariates
   - Normalization to solve the base level problem
   - “Industry decomposition”

Using a common characteristics distribution

- Yet another perspective is to focus on characteristics rather than on coefficients. That is, rather than thinking of reference coefficients we might think of a reference distribution to which the two groups are adjusted.
- The solution by Reimers (1983) with $\widehat{\beta}^* = (\widehat{\beta}^0 + \widehat{\beta}^1)/2$ is an interesting case under this perspective. For the unexplained part we get

$$
\begin{aligned}
\Delta_S^\mu &= \bar{X}^0(\widehat{\beta}^0 - \widehat{\beta}^*) + \bar{X}^1(\widehat{\beta}^* - \widehat{\beta}^1) \\
&= \bar{X}^0(\widehat{\beta}^0 - \tfrac{\widehat{\beta}^0 + \widehat{\beta}^1}{2}) + \bar{X}^1(\tfrac{\widehat{\beta}^0 + \widehat{\beta}^1}{2} - \widehat{\beta}^1) \\
&= \bar{X}^0(\tfrac{\widehat{\beta}^0 - \widehat{\beta}^1}{2}) + \bar{X}^1(\tfrac{\widehat{\beta}^0 - \widehat{\beta}^1}{2}) \\
&= \bar{X}^*(\widehat{\beta}^0 - \widehat{\beta}^1)
\end{aligned}
$$

with $\bar{X}^* = \frac{\bar{X}^0 + \bar{X}^1}{2}$.

Using a common characteristics distribution

- That is, the Reimers decomposition has intuitive appeal because it handles coefficients and characteristics according to the same logic; the decomposition is based on an (unweighted) average of coefficients as well as an (unweighted) average of characteristics. Only the Reimers decomposition has this property.
- A formally equivalent approach has already been suggested by Kitagawa (1955).
- In general, a good way of thinking about the "unexplained" part of the decomposition is to ask how the difference between the groups would look like if they had the same distribution of characteristics.
- Also the gap-closing estimand by Lundberg (2022) can be seen in this way. It asks how large the gap would be if a manipulable treatment were set to the same value (or same conditional distribution) in both groups.The gap-closing estimand, however, has an interventionist focus on a specific treatment.

Detailed decomposition: some conceptual complications

- A problem with the detailed decomposition of the "unexplained" part Δ_S^μ of the OB decomposition is that it is not invariant against (uninformative) transformations of the covariates (X variables).
- Furthermore, for categorical covariates, the results of the detailed decomposition of Δ_S^μ depend on the choice of the base/reference category.
- Some authors speak of an "identification" problem in this context. As argued by Fortin et al. (2011), however, it is more a conceptual problem of interpretation.
- The detailed decomposition of the "explained" part Δ_X^μ is more robust against these problems. Here only the contributions of the single categories of a categorical variable depend on the choice of the base category, but the sum across categories is not affected. Likewise, uninformative transformations of continuous covariates do not change the results of the detailed decomposition of Δ_X^μ.

1. The index problem
   - The three-fold decomposition
   - Nondiscriminatory wage structure
   - Example analysis
   - Relation to treatment effects
   - Using a common characteristics distribution

2. The transformation problem
   - Transformation of covariates
   - Base level of categorical covariates
   - Normalization to solve the base level problem
   - "Industry decomposition"

Transformation of covariates

- Assume that a location shift (e.g. mean centering) is applied to variable X_k, that is,
$$\tilde{X}_k = X_k + \gamma$$
- Consequences of the transformation:
  - Change in the expected value of the variable:
$$\mathsf{E}(\tilde{X}_k) = \mathsf{E}(X_k + \gamma) = \mathsf{E}(X_k) + \gamma$$
  - The slope parameter β_k of the variable in a regression model is not affected, that is, $\tilde{\beta}_k = \beta_k$. Likewise, all other slope parameters are unaffected.
  - However, the intercept β_0 changes:
$$\begin{aligned}
& E(Y) = \beta_0 + \beta_k E(X_k) + \textstyle\sum_{j\neq k} \beta_j E(X_j) \\
\Rightarrow \quad & \beta_0 = \mathsf{E}(Y) - \beta_k E(X_k) - \textstyle\sum_{j\neq k} \beta_j E(X_j) \\
\Rightarrow \quad & \tilde{\beta}_0 = \mathsf{E}(Y) - \beta_k (E(X_k) + \gamma) - \textstyle\sum_{j\neq k} \beta_j E(X_j) \\
& \quad = \mathsf{E}(Y) - \beta_k E(X_k) - \textstyle\sum_{j\neq k} \beta_j E(X_j) - \beta_k \gamma = \beta_0 - \beta_k \gamma
\end{aligned}$$

Transformation of covariates

- How does this affect the detailed decomposition results?
- There is no problem for the detailed decomposition of the "explained" part (as long as the same transformation is applied in both groups):

$$
\begin{aligned}
\Delta^{\mu}_{X,\tilde{X}_k} &= \tilde{\beta}^0_k(\mathsf{E}(\tilde{X}_k|G=0) - \mathsf{E}(\tilde{X}_k|G=1)) \\
&= \beta^0_k(\mathsf{E}(X_k|G=0) + \gamma - \mathsf{E}(X_k|G=1) - \gamma) \\
&= \beta^0_k(\mathsf{E}(X_k|G=0) - \mathsf{E}(X_k|G=1)) \\
&= \Delta^{\mu}_{X,X_k}
\end{aligned}
$$

Transformation of covariates

- The detailed decomposition of the unexplained part, however, may change:

$$\begin{aligned}\Delta^{\mu}_{S,\tilde{\beta}_0} &= (\tilde{\beta}^0_0 - \tilde{\beta}^1_0) = ((\beta^0_0 - \beta^0_k\gamma) - (\beta^1_0 - \beta^1_k\gamma)) \\ &= (\beta^0_0 - \beta^1_0) - \gamma(\beta^0_k - \beta^1_k) \\ &\neq (\beta^0_0 - \beta^1_0) = \Delta^{\mu}_{S,\beta_0}\end{aligned}$$

$$\begin{aligned}\Delta^{\mu}_{S,\tilde{\beta}_k} &= (\tilde{\beta}^0_k - \tilde{\beta}^1_k)\,\mathrm{E}(\tilde{X}_k|G=1) \\ &= (\beta^0_k - \beta^1_k)(\mathrm{E}(X_k|G=1) + \gamma) \\ &= (\beta^0_k - \beta^1_k)\,\mathrm{E}(X_k|G=1) + \gamma(\beta^0_k - \beta^1_k) \\ &\neq (\beta^0_k - \beta^1_k)\,\mathrm{E}(X_k|G=1) = \Delta^{\mu}_{S,\beta_k}\end{aligned}$$

Example: Years of education centered at mean

```
. 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 yeduc [aw=weight] if !missing(lnwage, yeduc, expft, expft2)
    Variable |     Obs      Weight        Mean   Std. dev.       Min        Max
-------------+-----------------------------------------------------------------
       yeduc |   5,462  12152217.3    12.82105   2.794755          7         18
. generate c_yeduc = yeduc - r(mean)
(188 missing values generated)
```

Original result

```
. oaxaca lnwage yeduc (experience: expft*), by(sex) weight(1) svy

Blinder-Oaxaca decomposition

Number of strata =     15                    Number of obs      =      5,462
Number of PSUs   =  2,037                    Population size    = 12,152,217
                                             Design df          =      2,022
                                             Model              =     linear
Group 1: sex = 1                             N of obs 1         =      2,642
Group 2: sex = 2                             N of obs 2         =      2,820

    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)

------------------------------------------------------------------------------
             |             Linearized
      lnwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.862735   .0162749   175.90   0.000     2.830818    2.894652
     group_2 |   2.657428   .0149503   177.75   0.000     2.628108    2.686747
  difference |   .2053074   .0204701    10.03   0.000     .1651628     .245452
   explained |   .0904872   .0151554     5.97   0.000     .0607654     .120209
 unexplained |   .1148202   .0211416     5.43   0.000     .0733585    .1562819
-------------+----------------------------------------------------------------
explained    |
       yeduc |  -.0191954   .0096981    -1.98   0.048    -.0382146   -.0001761
  experience |   .1096826   .0129638     8.46   0.000     .0842587    .1351065
-------------+----------------------------------------------------------------
unexplained  |
       yeduc |   .0520745   .0969211     0.54   0.591    -.1380012    .2421502
  experience |   .0370006   .0424742     0.87   0.384    -.0462972    .1202985
       _cons |   .0257451   .1178852     0.22   0.827     -.205444    .2569342
------------------------------------------------------------------------------
experience: expft expft2
```

Result using transformed covariate

```
. oaxaca lnwage c_yeduc (experience: expft*), by(sex) weight(1) svy

Blinder-Oaxaca decomposition

Number of strata =     15                   Number of obs      =      5,462
Number of PSUs   =  2,037                   Population size    = 12,152,217
                                            Design df          =      2,022
                                            Model              =     linear
Group 1: sex = 1                            N of obs 1         =      2,642
Group 2: sex = 2                            N of obs 2         =      2,820

    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)

------------------------------------------------------------------------------
             |             Linearized
      lnwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.862735   .0162749   175.90   0.000     2.830818    2.894652
     group_2 |   2.657428   .0149503   177.75   0.000     2.628108    2.686747
  difference |   .2053074   .0204701    10.03   0.000     .1651628     .245452
   explained |   .0904872   .0151554     5.97   0.000     .0607654     .120209
 unexplained |   .1148202   .0211416     5.43   0.000     .0733585    .1562819
-------------+----------------------------------------------------------------
explained    |
     c_yeduc |  -.0191954   .0096981    -1.98   0.048    -.0382146   -.0001761
  experience |   .1096826   .0129638     8.46   0.000     .0842587    .1351065
-------------+----------------------------------------------------------------
unexplained  |
     c_yeduc |   .0004835   .0009707     0.50   0.618    -.0014202    .0023872
  experience |   .0370006   .0424742     0.87   0.384    -.0462972    .1202985
       _cons |   .0773361   .0554875     1.39   0.164    -.0314825    .1861546
------------------------------------------------------------------------------
experience: expft expft2
```

Base level of categorical covariates

- Changing the base level of a categorical covariate has consequences both for the detailed decomposition of Δ_X^μ and Δ_S^μ.
- Let d_j, $j = 1, \dots, J$, be a set of indicator variables representing a categorical variable D that has J levels ($d_j = 1$ if $D = j$ and else 0).
- The contribution of D to Δ_X^μ then is:

$$\Delta_{X,D}^\mu = \beta_{d_1}^0(\bar{d}_1^0 - \bar{d}_1^1) + \beta_{d_2}^0(\bar{d}_2^0 - \bar{d}_2^1) + \cdots + \beta_{d_J}^0(\bar{d}_J^0 - \bar{d}_J^1)$$

- To estimate the coefficients, one of the levels has to be omitted (the base level). This is equivalent to constraining its coefficient to be zero.
- That is, what we are estimating are coefficients

$$\beta_{d_j^o} = \beta_{d_j} - \beta_{d_o}$$

Base level of categorical covariates

- If we omit the first level, we have

$$\Delta^{\mu}_{X,D} = 0(\bar{d}^0_1 - \bar{d}^1_1) + \beta^0_{d^1_2}(\bar{d}^0_2 - \bar{d}^1_2) + \cdots + \beta^0_{d^1_J}(\bar{d}^0_J - \bar{d}^1_J)$$

- If we omit the second level, we have

$$\Delta^{\mu}_{X,D} = \beta^0_{d^2_1}(\bar{d}^0_1 - \bar{d}^1_1) + 0(\bar{d}^0_2 - \bar{d}^1_2) + \cdots + \beta^0_{d^2_J}(\bar{d}^0_J - \bar{d}^1_J)$$

- We clearly see that individual contributions of the single indicators variables will be different depending on the choice of the base level.

Base level of categorical covariates

- However, the sum across the contributions of all indicators will always be the same. Because $\bar{d}_o = 1 - \sum_{j \neq o} \bar{d}_j$ and $\beta_{d_j^o} = \beta_{d_j} - \beta_{d_o}$ we have, for example,

$$
\begin{aligned}
\Delta^{\mu}_{X,D} &= \beta^0_{d_2^1}(\bar{d}_2^0 - \bar{d}_2^1) + \cdots + \beta^0_{d_J^1}(\bar{d}_J^0 - \bar{d}_J^1) \\
&= (\beta^0_{d_2} - \beta^0_{d_1})(\bar{d}_2^0 - \bar{d}_2^1) + \cdots + (\beta^0_{d_J} - \beta^0_{d_1})(\bar{d}_J^0 - \bar{d}_J^1) \\
&= \beta^0_{d_2}(\bar{d}_2^0 - \bar{d}_2^1) + \cdots + \beta^0_{d_J}(\bar{d}_J^0 - \bar{d}_J^1) \\
&\qquad\qquad - \beta^0_{d_1}(\bar{d}_2^0 - \bar{d}_2^1 + \cdots + \bar{d}_J^0 - \bar{d}_J^1) \\
&= \beta^0_{d_2}(\bar{d}_2^0 - \bar{d}_2^1) + \cdots + \beta^0_{d_J}(\bar{d}_J^0 - \bar{d}_J^1) \\
&\qquad\qquad - \beta^0_{d_1}\{(1 - \bar{d}_1^0) - (1 - \bar{d}_1^2)\} \\
&= \beta^0_{d_1}(\bar{d}_1^0 - \bar{d}_1^1) + \beta^0_{d_2}(\bar{d}_2^0 - \bar{d}_2^1) + \cdots + \beta^0_{d_J}(\bar{d}_J^0 - \bar{d}_J^1)
\end{aligned}
$$

- That is, independent of the choice of the base level, we always get the same expression.

Base level of categorical covariates

- Now consider the effect on the contributions to Δ_S^μ. Omitting the first indicator, we have

$$\Delta_S^\mu = (\beta_0^0 - \beta_0^1) + (\beta_{d_2^1}^0 - \beta_{d_2^1}^1)\bar{d}_2^1 + \cdots + (\beta_{d_J^1}^0 - \beta_{d_J^1}^1)\bar{d}_J^1 + \ldots$$

- Changing the base level has consequences for the estimated coefficients of the dummy variables (see above), but it also affects the intercept β_0.
- The intercept is equal to the expectation of Y given all covariates are zero. All $(J-1)$ included indicators being zero implies that the omitted category applies. That is, the intercept reflects the conditional outcome in the base category.
- Hence, the difference in intercepts between the two groups, $\beta_0^0 - \beta_0^1$, refers to the difference in conditional outcomes in the base category. Changing the base category changes the meaning of $\beta_0^0 - \beta_0^1$.
- Naturally, also the contributions of single indicators – as well as the sum over the contributions of all included indicators – will change.

Example: Education as categorical variable

```
. recode casmin ///
>     (1 2 = 1 "low") ///
>     (4 6 = 2 "medium general") ///
>     (3 5 7 = 3 "medium vocational") ///
>     (8 9 = 4 "high") ///
>     (else = .) ///
>     , into(casmin4)
(5,724 differences between casmin and casmin4)

. tab casmin4, gen(casmin4_)
```

RECODE of casmin (level of educational (CASMIN))	Freq.	Percent	Cum.
low	763	10.65	10.65
medium general	502	7.01	17.66
medium vocational	3,989	55.67	73.33
high	1,911	26.67	100.00
Total	7,165	100.00	

```
. sum casmin4_*
```

Variable	Obs	Mean	Std. dev.	Min	Max
casmin4_1	7,165	.1064899	.3084851	0	1
casmin4_2	7,165	.0700628	.2552706	0	1
casmin4_3	7,165	.5567341	.4968055	0	1
casmin4_4	7,165	.2667132	.442272	0	1

First level as base category

```
. oaxaca lnwage casmin4_2 casmin4_3 casmin4_4 (experience: expft*), by(sex) weight(1) svy
Blinder-Oaxaca decomposition

Number of strata =    15                    Number of obs      =      5,493
Number of PSUs   = 2,047                    Population size    = 12,276,729
                                            Design df          =      2,032
                                            Model              =     linear
Group 1: sex = 1                            N of obs 1         =      2,653
Group 2: sex = 2                            N of obs 2         =      2,840

    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)

-----------------------------------------------------------------------------
             |             Linearized
      lnwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+---------------------------------------------------------------
overall      |
     group_1 |   2.862164   .0160457   178.38   0.000     2.830697    2.893632
     group_2 |   2.658932   .0147418   180.37   0.000     2.630022    2.687843
  difference |   .2032321   .0200916    10.12   0.000     .1638297    .2426344
   explained |   .1028131   .0139542     7.37   0.000     .0754471    .1301791
 unexplained |    .100419   .0206278     4.87   0.000     .0599651    .1408728
-------------+---------------------------------------------------------------
explained    |
   casmin4_2 |   .0028628   .0040742     0.70   0.482    -.0051273    .0108528
   casmin4_3 |  -.0083128   .0066494    -1.25   0.211    -.0213531    .0047276
   casmin4_4 |   .0038851   .0141355     0.27   0.783    -.0238365    .0316066
  experience |    .104378   .0117252     8.90   0.000     .0813834    .1273726
-------------+---------------------------------------------------------------
unexplained  |
   casmin4_2 |    .014117   .0063769     2.21   0.027     .0016111    .0266229
   casmin4_3 |    .036901   .0477265     0.77   0.440     -.056697     .130499
   casmin4_4 |   .0408286   .0250638     1.63   0.103    -.0083249     .089982
  experience |    .045385   .0390684     1.16   0.246    -.0312333    .1220032
       _cons |  -.0368126   .0962395    -0.38   0.702    -.2255509    .1519258
-----------------------------------------------------------------------------
experience: expft expft2
```

Third level as base category

```
. oaxaca lnwage casmin4_1 casmin4_2 casmin4_4 (experience: expft*), by(sex) weight(1) svy

Blinder-Oaxaca decomposition

Number of strata =     15                        Number of obs      =      5,493
Number of PSUs   =  2,047                        Population size    = 12,276,729
                                                 Design df          =      2,032
                                                 Model              =     linear
Group 1: sex = 1                                 N of obs 1         =      2,653
Group 2: sex = 2                                 N of obs 2         =      2,840

    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)
```

lnwage	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
overall						
group_1	2.862164	.0160457	178.38	0.000	2.830697	2.893632
group_2	2.658932	.0147418	180.37	0.000	2.630022	2.687843
difference	.2032321	.0200916	10.12	0.000	.1638297	.2426344
explained	.1028131	.0139542	7.37	0.000	.0754471	.1301791
unexplained	.100419	.0206278	4.87	0.000	.0599651	.1408728
explained						
casmin4_1	-.0042448	.0029696	-1.43	0.153	-.0100686	.001579
casmin4_2	.00041	.0007939	0.52	0.606	-.0011469	.0019668
casmin4_4	.0022699	.0082584	0.27	0.783	-.0139259	.0184657
experience	.104378	.0117252	8.90	0.000	.0813834	.1273726
unexplained						
casmin4_1	-.0031795	.0041288	-0.77	0.441	-.0112765	.0049176
casmin4_2	.01093	.0053341	2.05	0.041	.0004692	.0213908
casmin4_4	.0232233	.0122923	1.89	0.059	-.0008835	.04733
experience	.045385	.0390684	1.16	0.246	-.0312333	.1220032
_cons	.0240602	.0515252	0.47	0.641	-.0769876	.125108

```
experience: expft expft2
```

Aggregate decomposition (first=base)

```
. oaxaca lnwage (casmin: casmin4_2 casmin4_3 casmin4_4) (experience: expft*), ///
>     by(sex) weight(1) svy

Blinder-Oaxaca decomposition

Number of strata =     15                       Number of obs     =      5,493
Number of PSUs   =  2,047                       Population size   = 12,276,729
                                                Design df         =      2,032
                                                Model             =     linear
Group 1: sex = 1                                N of obs 1        =      2,653
Group 2: sex = 2                                N of obs 2        =      2,840

    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)

------------------------------------------------------------------------------
             |             Linearized
      lnwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.862164   .0160457   178.38   0.000     2.830697    2.893632
     group_2 |   2.658932   .0147418   180.37   0.000     2.630022    2.687843
  difference |   .2032321   .0200916    10.12   0.000     .1638297    .2426344
   explained |   .1028131   .0139542     7.37   0.000     .0754471    .1301791
 unexplained |    .100419   .0206278     4.87   0.000     .0599651    .1408728
-------------+----------------------------------------------------------------
explained    |
      casmin |  -.0015649   .0090668    -0.17   0.863     -.019346    .0162162
  experience |    .104378   .0117252     8.90   0.000     .0813834    .1273726
-------------+----------------------------------------------------------------
unexplained  |
      casmin |   .0918466   .0742085     1.24   0.216    -.0536861    .2373793
  experience |    .045385   .0390684     1.16   0.246    -.0312333    .1220032
       _cons |  -.0368126   .0962395    -0.38   0.702    -.2255509    .1519258
------------------------------------------------------------------------------
casmin: casmin4_2 casmin4_3 casmin4_4
experience: expft expft2
```

Aggregate decomposition (third=base)

```
. oaxaca lnwage (casmin: casmin4_1 casmin4_2 casmin4_4) (experience: expft*), ///
>     by(sex) weight(1) svy

Blinder-Oaxaca decomposition

Number of strata =     15                        Number of obs      =      5,493
Number of PSUs   =  2,047                        Population size    = 12,276,729
                                                 Design df          =      2,032
                                                 Model              =     linear
Group 1: sex = 1                                 N of obs 1         =      2,653
Group 2: sex = 2                                 N of obs 2         =      2,840

    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)
```

lnwage	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
overall						
group_1	2.862164	.0160457	178.38	0.000	2.830697	2.893632
group_2	2.658932	.0147418	180.37	0.000	2.630022	2.687843
difference	.2032321	.0200916	10.12	0.000	.1638297	.2426344
explained	.1028131	.0139542	7.37	0.000	.0754471	.1301791
unexplained	.100419	.0206278	4.87	0.000	.0599651	.1408728
explained						
casmin	-.0015649	.0090668	-0.17	0.863	-.019346	.0162162
experience	.104378	.0117252	8.90	0.000	.0813834	.1273726
unexplained						
casmin	.0309738	.0150801	2.05	0.040	.0013999	.0605478
experience	.045385	.0390684	1.16	0.246	-.0312333	.1220032
_cons	.0240602	.0515252	0.47	0.641	-.0769876	.125108

```
casmin: casmin4_1 casmin4_2 casmin4_4
experience: expft expft2
```

1. The index problem
   - The three-fold decomposition
   - Nondiscriminatory wage structure
   - Example analysis
   - Relation to treatment effects
   - Using a common characteristics distribution

2. The transformation problem
   - Transformation of covariates
   - Base level of categorical covariates
   - Normalization to solve the base level problem
   - "Industry decomposition"

Normalization

- „Normalization“ of the coefficients associated with categorical variables has been suggested as a solution to the problem that the choice of the base level changes the detailed decomposition results.
- One solution are so-called “deviation contrasts” (equivalent to “effect coding”): the coefficients of the indicators reflect deviations from the unweighted average across categories (balanced grand mean).
- The decomposition results based on coefficients that have been normalized using the deviation contrast transform are independent of the choice of the base level.
- Furthermore, the results are equal to the (unweighted) average of the results one would get from a series of decompositions in which the categories are used one after another as the base level (Yun 2008).

Normalization

- The deviation contrast normalization works as follows:
  - Again, let d_j, $j = 1, \ldots, J$, be a set of indicator variables and $\beta_{d_j^o}$ be the corresponding coefficients (with $\beta_{d_o^o} = 0$).
  - Determine
$$c = \frac{\beta_{d_1^o} + \cdots + \beta_{d_J^o}}{J}$$
  - Compute the transformed coefficients
$$\tilde{\beta}_0 = \beta_0 + c \quad \text{and} \quad \beta_{d_j} = \beta_{d_j^o} - c$$
(note that $\sum_{j=1}^{J} \beta_{d_j} = 0$).
  - Use coefficients $\tilde{\beta}_0$ and β_{d_j} to perform the decomposition instead of the original coefficients.
  - An alternative to transforming the coefficients would be to apply restricted least-squares estimation with restriction $\sum_{j=1}^{J} \beta_{d_j} = 0$.

Illustration of deviation contrasts (first=base)

```
. svy, noheader: regress lnwage casmin4_2 casmin4_3 casmin4_4 expft*
(running regress on estimation sample)
```

lnwage	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
casmin4_2	.2541662	.0622926	4.08	0.000	.1320022	.3763301
casmin4_3	.2806464	.041122	6.82	0.000	.2000006	.3612921
casmin4_4	.6933972	.0456906	15.18	0.000	.6037918	.7830026
expft	.0341931	.0037561	9.10	0.000	.026827	.0415592
expft2	-.0005688	.000107	-5.32	0.000	-.0007785	-.000359
_cons	2.072234	.0492694	42.06	0.000	1.97561	2.168857

```
. devcon, groups(casmin4_1 casmin4_2 casmin4_3 casmin4_4)
Transformed regress coefficients                    Number of obs    =      5493
```

lnwage	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
casmin4_1	-.3070524	.0325804	-9.42	0.000	-.3709468	-.243158
casmin4_2	-.0528863	.0383571	-1.38	0.168	-.1281097	.0223371
casmin4_3	-.0264061	.0185573	-1.42	0.155	-.0627995	.0099873
casmin4_4	.3863448	.0231512	16.69	0.000	.3409423	.4317473
expft	.0341931	.0037561	9.10	0.000	.026827	.0415592
expft2	-.0005688	.000107	-5.32	0.000	-.0007785	-.000359
_cons	2.379286	.0314595	75.63	0.000	2.31759	2.440982

Illustration of deviation contrasts (third=base)

```
. svy, noheader: regress lnwage casmin4_1 casmin4_2 casmin4_4 expft*
(running regress on estimation sample)
```

lnwage	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
casmin4_1	-.2806464	.041122	-6.82	0.000	-.3612921	-.2000006
casmin4_2	-.0264802	.0502395	-0.53	0.598	-.1250065	.0720461
casmin4_4	.4127509	.023991	17.20	0.000	.3657013	.4598004
expft	.0341931	.0037561	9.10	0.000	.026827	.0415592
expft2	-.0005688	.000107	-5.32	0.000	-.0007785	-.000359
_cons	2.35288	.0271146	86.78	0.000	2.299705	2.406055

```
. devcon, groups(casmin4_1 casmin4_2 casmin4_3 casmin4_4)
Transformed regress coefficients                  Number of obs     =      5493
```

lnwage	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
casmin4_1	-.3070524	.0325804	-9.42	0.000	-.3709468	-.243158
casmin4_2	-.0528863	.0383571	-1.38	0.168	-.1281097	.0223371
casmin4_3	-.0264061	.0185573	-1.42	0.155	-.0627995	.0099873
casmin4_4	.3863448	.0231512	16.69	0.000	.3409423	.4317473
expft	.0341931	.0037561	9.10	0.000	.026827	.0415592
expft2	-.0005688	.000107	-5.32	0.000	-.0007785	-.000359
_cons	2.379286	.0314595	75.63	0.000	2.31759	2.440982

Normalized results

```
. oaxaca lnwage normalize(casmin4_*) (experience: expft*), by(sex) weight(1) svy
(normalized: casmin4_1 casmin4_2 casmin4_3 casmin4_4)

Blinder-Oaxaca decomposition

Number of strata =     15                  Number of obs     =      5,493
Number of PSUs   =  2,047                  Population size   = 12,276,729
                                           Design df         =      2,032
                                           Model             =     linear
Group 1: sex = 1                           N of obs 1        =      2,653
Group 2: sex = 2                           N of obs 2        =      2,840

    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)
```

lnwage	Coefficient	Linearized std. err.	t	P>|t|	[95% conf.	interval]
overall						
group_1	2.862164	.0160457	178.38	0.000	2.830697	2.893632
group_2	2.658932	.0147418	180.37	0.000	2.630022	2.687843
difference	.2032321	.0200916	10.12	0.000	.1638297	.2426344
explained	.1028131	.0139542	7.37	0.000	.0754471	.1301791
unexplained	.100419	.0206278	4.87	0.000	.0599651	.1408728
explained						
casmin4_1	-.0048524	.0033364	-1.45	0.146	-.0113955	.0016908
casmin4_2	.0000589	.000429	0.14	0.891	-.0007824	.0009001
casmin4_3	.0011898	.0011475	1.04	0.300	-.0010606	.0034402
casmin4_4	.0020387	.0074175	0.27	0.783	-.0125079	.0165853
experience	.104378	.0117252	8.90	0.000	.0813834	.1273726
unexplained						
casmin4_1	-.0061591	.0032829	-1.88	0.061	-.0125974	.0002791
casmin4_2	.0079432	.0039587	2.01	0.045	.0001796	.0157069
casmin4_3	-.0345822	.0222469	-1.55	0.120	-.0782113	.0090469
casmin4_4	.0067243	.0123664	0.54	0.587	-.0175278	.0309764
experience	.045385	.0390684	1.16	0.246	-.0312333	.1220032
_cons	.0811078	.0593154	1.37	0.172	-.0352175	.197433

```
experience: expft expft2
```

Normalized results = average across all possible variants

```
. forv j = 1(1)4 {
  2.     local casmin casmin4_1 casmin4_2 casmin4_3 casmin4_4 _cons
  3.     local casmin: subinstr local casmin "casmin4_`j'" ""
  4.     quietly oaxaca lnwage `casmin' (experience: expft*), by(sex) weight(1) svy
  5.     estimates store m`j'
  6. }
. estout m?, keep(unexplained:casmin4_* unexplained:_cons) order(casmin4_1) collab(none)
```

	m1	m2	m3	m4
unexplained				
casmin4_1		-.0140835	-.0031795	-.0073735
casmin4_2	.014117		.01093	.006726
casmin4_3	.036901	-.1265532		-.0486765
casmin4_4	.0408286	-.0371546	.0232233	
_cons	-.0368126	.2328253	.0240602	.1043581

```
. mata: mean(editmissing(st_matrix("r(coefs)"), 0)')'
```

	1
1	-.0061591142
2	.0079432497
3	-.0345821791
4	.0067243174
5	.0811077549

Weighted normalization

- An alternative – and probably superior – variant of the normalization uses coefficients that reflect deviations from the weighted average across categories (observation-weighted grand mean), where the weights are proportional to the probabilities of the categories (Kennedy 1986, Haisken-DeNew and Schmidt 1997).
- That is, use

$$c = \Pr(D = 1)\beta_{d_1^o} + \cdots + \Pr(D = J)\beta_{d_J^o}$$

such that

$$\sum_{j=1}^{J} \Pr(D = 1)\beta_{d_j} = 0$$

- This limits the influence of sparsely populated categories and makes results more robust against recoding the categorical variable (i.e. combining several sparsely populated categories into one will not have much of an effect on the results; see Kim 2013).

"Industry decomposition"

- Yet another type of normalization is to compute

$$\Delta^{\mu}_{S,d_j} = (\beta^0_0 - \beta^1_0) + (\beta^0_{d^o_j} - \beta^1_{d^o_j}) + \sum_{k=1}^{K} (\beta^0_k - \beta^1_k)\bar{X}^1_k$$

  as suggested by Horrace and Oaxaca (2001), where d_j, $j = 1, \ldots, J$, is again a set of indicator variables and X_k, $k = 1, \ldots, K$, are all other covariates, and then normalize the contributions using

$$\%\Delta^{\mu}_{S,d_j} = \frac{\bar{d}^1_j \Delta^{\mu}_{S,d_j}}{\Delta^{\mu}_S} \quad \text{since } \Delta^{\mu}_S = \sum_{j=1}^{K} \bar{d}^1_j \Delta^{\mu}_{S,d_j}$$

  as suggested by Fortin et al. (2011)
- This makes sense, for example, if we want to know how much different industries contribute to the unexplained wage gap, controlling for differential composition of the industries with respect to the X variables and taking into account the industry size.

Comment

- There is always a certain arbitrariness to the different normalization approaches. There is no right or wrong; what makes sense may depend on context.
- Fortin et al. (2011) suggest that it may be more fruitful to chose the omitted category based on substantive reasoning and stick to the original results. This requires more thinking about how the results can be meaningfully interpreted in a specific case.

Exercise 3

References

- Cotton, Jeremiah (1988). On the Decomposition of Wage Differentials. The Review of Economics and Statistics 70(2):236–243.
- Fortin, Nicole M. (2008). The Gender Wage Gap among Young Adults in the United States. The Importance of Money versus People. Journal of Human Resources 43(4):884–918.
  - Working paper: Fortin, Nicole M. (2006). Greed, altruism, and the gender wage gap. http://faculty.arts.ubc.ca/nfortin/Fortinat8.pdf
- Fortin, Nicole, Thomas Lemieux, Sergio Firpo (2011). Decomposition Methods in Economics. Pp. 1–102 in: O. Ashenfelter and D. Card (eds.). Handbook of Labor Economics. Amsterdam: Elsevier.
- Haisken-DeNew, John P., Christoph M. Schmidt (1997). Inter-Industry and Inter-Regional Differentials: Mechanics and Interpretation. The Review of Economics and Statistics 79(3):516–521.
- Horrace, William C., Ronald L. Oaxaca (2001). Inter-Industry Wage Differentials and the Gender Wage Gap: An Identification Problem. Industrial and Labor Relations Review 54(3):611–618.

References

- Jann, Ben (2008). The Blinder-Oaxaca decomposition for linear regression models. The Stata Journal 8(4):453–479.
- Kennedy, Peter (1986). Interpreting Dummy Variables. The Review of Economics and Statistics 68(1):174–175.
- Kim, ChangHwan (2013). Detailed Wage Decompositions. Revisiting the Identification Problem. Sociological Methodology 43:346–363.
- Lundberg, Ian (2022). The Gap-Closing Estimand: A Causal Approach to Study Interventions That Close Disparities Across Social Categories. Sociological Methods & Research. DOI: 10.1177/00491241211055769
- Neumark, David (1988). Employers' Discriminatory Behavior and the Estimation of Wage Discrimination. The Journal of Human Resources 23(3):279–295.
- Oaxaca, Ronald L., Michael R. Ransom (1994). On discrimination and the decomposition of wage differentials. Journal of Econometrics 61(1):5–21.
- Reimers, Cordelia W. (1983). Labor Market Discrimination Against Hispanic and Black Men. The Review of Economics and Statistics 65(4):570–579.

References

- Winsborough, H. H., Peter Dickinson (1971). Components of Negro-White Income Differences. Pp. 6–8 in: Proceedings of the Social Statistics Section. Washington, DC: American Statistical Association.
- Yun, Myeong-Su (2008). Identification problem and detailed Oaxaca decomposition: A general solution and statistical inference. Journal of Economic and Social Measurement 33:27–38.