Decomposition methods in the social sciences

Bamberg Graduate School of Social Sciences, June 7–8, 2018

Ben Jann

University of Bern, Institut of Sociology

Oaxaca-Blinder decomposition

Contents

Introduction

- Studies by Oaxaca (1973) und Blinder (1973) analyzed the wage gap between men and women and between whites and blacks in the USA.
- For example, the gender wage gap (measured as the difference in average wages between males and females) was about 45 percent at that time (data of 1967).
- Question: How large is the part of the gender wage gap that can be attributed to gender differences in characteristics that are relevant for wages (such as education or work experience)? That is, how large is Δ_X^ν?
- The remaining part of the gap, Δ_S^ν, is due to differences in the wage structure $m()$, that is, to differences in how the characteristics are rewarded in the labor market for men and women. In the context of the gender wage gap this part is often interpreted as "discrimination".

The Oaxaca-Blinder decomposition

- The classic OB decomposition focuses on group differences in $\mu(F_Y)$, the mean of Y.
- Presumed is the following structural function:

$$Y_i^g = m^g(X_i, \epsilon_i) = \beta_0^g + \beta_1^g X_{1i} + \cdots + \beta_K^g X_{Ki} + \epsilon_i, \quad \text{for } g = 0, 1$$

- For example, Y^0 are (log) wages according to the wage structure of men, Y^1 are (log) wages according to the wage structure of women.
- Assumptions:
  - Additive linearity: $m(X, \epsilon) = X\beta + \epsilon$, that is, effects of observed and unobserved characteristics are additively separable in $m()$
  - Zero conditional mean/conditional (mean) independence: $\mathrm{E}(\epsilon|X, G) = 0$

Remark on notation: in expressions such as $X\beta$, X is a data matrix or a single row vector of values for X_1, ..., X_K and β is a corresponding column vector of coefficients. X includes a constant unless noted otherwise, i.e. $X = [1, X_1, \ldots, X_K]$.

The Oaxaca-Blinder decomposition

- In this case, Δ^μ can be written as

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

- To perform the decomposition, we now need a suitable counterfactual.
- Proposal: use $F_{Y^0|G=1}$, that is, use the counterfactual mean

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

- If $G=0$ are men and $G=1$ are women, this is the average of (log) wages we would expect for women, if they were paid like men.

The Oaxaca-Blinder decomposition

- Adding and subtracting $\mathrm{E}(X|G=1)\beta^0$, we obtain the decomposition

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

where

Δ^μ_X "explained" part, endowment effect, composition effect, quantity effect

Δ^μ_S "unexplained" part, discrimination, price effect

Estimation

- All components of the above decomposition can readily be estimated.
  - β^g can be estimated by applying linear regression to the $G = g$ subsample.
  - A suitable estimate of $\mathrm{E}(X|G = g)$ is simply the vector of means of X in the $G = g$ subsample.
  - That is, run regressions among men and women, and compute the means of X for men and women.
- Let $\hat{\beta}^g$ be the estimate of β^g and $\bar{X}^g = \hat{\mathrm{E}}(X|G = g)$ be the estimate of $\mathrm{E}(X|G = g)$. The decomposition estimate then is

$$\hat{\Delta}^\mu = \hat{\Delta}^\mu_X + \hat{\Delta}^\mu_S = (\bar{X}^0 - \bar{X}^1)\hat{\beta}^0 + \bar{X}^1(\hat{\beta}^0 - \hat{\beta}^1)$$

Standard errors

- For a long time, results from OB decompositions were reported without information on statistical inference (standard errors, confidence intervals).
- Meaningful interpretation of results, however, is difficult without information on estimation precision.
- A first suggestion on how to compute standard errors for decomposition results has been made by Oaxaca und Ransom (1998; also see Greene 2003:53–54).
- These authors, however, assume “fixed” covariates (like factors in an experimental design) and hence ignore an important source of statistical uncertainty.
- That the stochastic nature of covariates has no consequences for the estimation of (conditional) coefficients in regression models is an important insight of econometrics. However, this does not hold for (unconditional) OB decompositions.

Standard errors

- Think of a term such as $\bar{X}\hat{\beta}$, where $\bar{X}$ is a row vector of sample means and $\hat{\beta}$ is a column vector of regression coefficients (the result is a scalar). How can its sampling variance, $V(\bar{X}\hat{\beta})$, be estimated?
  - If the covariates are fixed, then $\bar{X}$ has no sampling variance. Hence:

    $$V(\bar{X}\hat{\beta}) = \bar{X}V(\hat{\beta})\bar{X}'$$

  - However, if covariates are stochastic, the sampling variance is

    $$V(\bar{X}\hat{\beta}) = \bar{X}V(\hat{\beta})\bar{X}' + \hat{\beta}'V(\bar{X})\hat{\beta} + \text{trace}\Big\{V(\bar{X})V(\hat{\beta})\Big\}$$

    (see the proof in Jann 2005).
  - The last term, trace{}, is asymptotically vanishing and can be ignored.
  - To estimate $V(\bar{X}\hat{\beta})$, plug in estimates for $V(\hat{\beta})$ (the variance-covariance matrix of the regression coefficients) and $V(\bar{X})$ (the variance-covariance matrix of the means), which are readily available.

Standard errors

- Using this result, equations for the sampling variances of the components of an OB decomposition can easily be derived.
- For example, assuming that the two groups are independent, we get:

$$
\begin{aligned}
V(\hat{\Delta}_X^{\mu}) = V((\bar{X}^0 - \bar{X}^1)\hat{\beta}^0) &\approx (\bar{X}^0 - \bar{X}^1)V(\hat{\beta}^0)(\bar{X}^0 - \bar{X}^1)' \\
&\quad + \hat{\beta}^{0\prime}\left[V(\bar{X}^0) + V(\bar{X}^1)\right]\hat{\beta}^0 \\
V(\hat{\Delta}_S^{\mu}) = V(\bar{X}^1(\hat{\beta}^0 - \hat{\beta}^1)) &\approx \bar{X}^1\left[V(\hat{\beta}^0) + V(\hat{\beta}^1)\right]\bar{X}^{1\prime} \\
&\quad + (\hat{\beta}^0 - \hat{\beta}^1)'V(\bar{X}^1)(\hat{\beta}^0 - \hat{\beta}^1)
\end{aligned}
$$

- Equations for other variants of the decomposition, for elements of the detailed decomposition (see below), and for the covariances among components can be derived similarly. Incorporation of complex survey designs (in which, e.g., the two groups are not independent) is also possible.
- An alternative is to use replication techniques such as the bootstrap or jackknife.

Detailed decomposition

- Often one is not only interested in the aggregate decomposition into an "explained" and an "unexplained" part, but one wants to further decompose the components into contributions of single covariates.
- Given the assumption of additive linearity, such detailed decompositions are easy to compute.
- For the "explained" part we have

$$\begin{aligned}\hat{\Delta}^{\mu}_{X} = (\bar{X}^0 - \bar{X}^1)\hat{\beta}^0 &= \sum_{k=1}^{K}\hat{\beta}^0_k(\bar{X}^0_k - \bar{X}^1_k) \\ &= \hat{\beta}^0_1(\bar{X}^0_1 - \bar{X}^1_1) + \cdots + \hat{\beta}^0_K(\bar{X}^0_K - \bar{X}^1_K)\end{aligned}$$

- For the "unexplained" part we have

$$\begin{aligned}\hat{\Delta}^{\mu}_{S} = \bar{X}^1(\hat{\beta}^0 - \hat{\beta}^1) &= (\hat{\beta}^0_0 - \hat{\beta}^1_0) + \sum_{k=1}^{K}(\hat{\beta}^0_k - \hat{\beta}^1_k)\bar{X}^1_k \\ &= (\hat{\beta}^0_0 - \hat{\beta}^1_0) + (\hat{\beta}^0_1 - \hat{\beta}^1_1)\bar{X}^1_1 + \cdots + (\hat{\beta}^0_K - \hat{\beta}^1_K)\bar{X}^1_K\end{aligned}$$

Detailed decomposition

- Furthermore, it is easy to subsume the detailed decomposition by sets of covariates:

$$\hat{\Delta}_X^{\mu} = \sum_{k=1}^{a} \hat{\beta}_k^0 (\bar{X}_k^0 - \bar{X}_k^1) + \sum_{k=a+1}^{b} \hat{\beta}_k^0 (\bar{X}_k^0 - \bar{X}_k^1) + \dots$$

$$\hat{\Delta}_S^{\mu} = (\hat{\beta}_0^0 - \hat{\beta}_0^1) + \sum_{k=1}^{a} (\hat{\beta}_k^0 - \hat{\beta}_k^1) \bar{X}_k^1 + \sum_{k=a+1}^{b} (\hat{\beta}_k^0 - \hat{\beta}_k^1) \bar{X}_k^1 + \dots$$

Example analysis

- Data: `gsoep.dta`; extract from GSOEP29 (2012)
- Outcome variable (Y): logarithm of gross hourly wages
- Groups (G): males vs. females
- Predictors (X): years of schooling, years of full-time work experience
- Sample selection: respondents between 25 and 55 years old
- The example requires the `oaxaca` package (Jann 2008). To install the package and view the help file, type:

```
. ssc install oaxaca, replace
. help oaxaca
```

Data preparation

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

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

Summarize wages by gender

```
. bysort bcsex: summarize wage if schooling<. & ft_experience<., detail
------------------------------------------------------------------------------
-> bcsex = [1] Maennlich
                            wage
-------------------------------------------------------------
      Percentiles      Smallest
 1%     2.583979       .3624283
 5%      6.20155       .3875969
10%     8.050941       .6395349       Obs               3,877
25%     11.57623        .744186       Sum of Wgt.       3,877

50%     16.27907                      Mean           18.28089
                        Largest       Std. Dev.       12.2374
75%     22.14839       145.3488
90%     29.71576       162.7907       Variance       149.7539
95%     36.10771       186.0465       Skewness       5.931026
99%     60.62196       287.2267       Kurtosis       88.37888
------------------------------------------------------------------------------
-> bcsex = [2] Weiblich
                            wage
-------------------------------------------------------------
      Percentiles      Smallest
 1%     2.034884       .4186046
 5%     4.651163       .5285412
10%      6.20155       .6644518       Obs               3,983
25%      8.75513       .6976744       Sum of Wgt.       3,983

50%     12.72727                      Mean           14.50449
                        Largest       Std. Dev.      17.70616
75%     17.44186       197.8295
90%     22.96512       220.5814       Variance       313.5081
95%     28.16222       227.1498       Skewness       34.66708
99%     43.77565       914.7287       Kurtosis       1694.197
```

The gender wage gap

```
. mean wage if schooling<. & ft_experience<., over(bcsex)
Mean estimation                   Number of obs    =      7,860
    _subpop_1: bcsex = [1] Maennlich
    _subpop_2: bcsex = [2] Weiblich
```

Over	Mean	Std. Err.	[95% Conf.	Interval]
wage				
_subpop_1	18.28089	.1965356	17.89563	18.66615
_subpop_2	14.50449	.2805558	13.95453	15.05445

```
. lincom _subpop_1-_subpop_2
 ( 1)  [wage]_subpop_1 - [wage]_subpop_2 = 0
```

Mean	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	3.776401	.342546	11.02	0.000	3.104919	4.447882

```
. nlcom _b[_subpop_1]/_b[_subpop_2]
       _nl_1:  _b[_subpop_1]/_b[_subpop_2]
```

Mean	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	1.260361	.0278913	45.19	0.000	1.205695	1.315027

```
. nlcom (_b[_subpop_1]/_b[_subpop_2]-1)*100
       _nl_1:  (_b[_subpop_1]/_b[_subpop_2]-1)*100
```

Mean	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	26.03608	2.789132	9.33	0.000	20.56948	31.50268

The gender wage gap

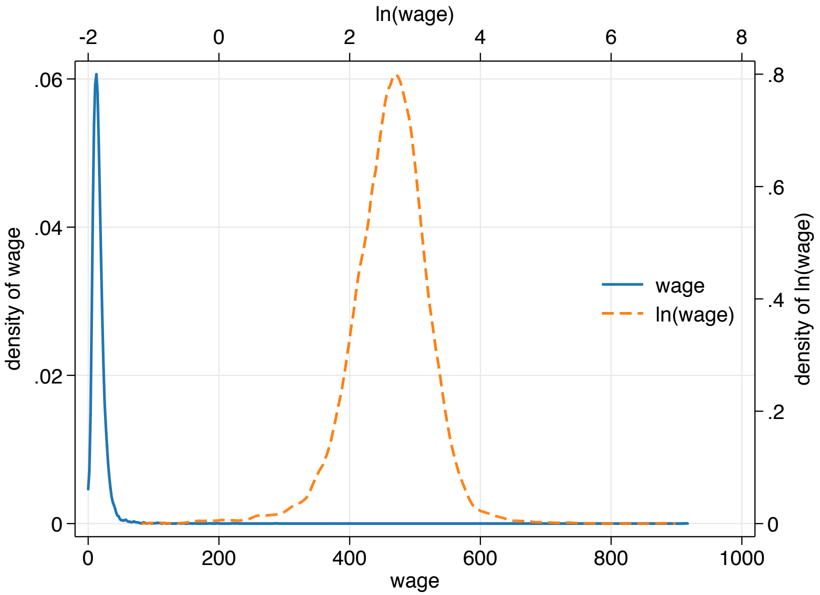
- Typically, the *logarithm* of wages is analyzed, because
  - wages can only be positive; $Y \in (0, \infty)$
  - wages have a (left) skewed distribution; taking the logarithm makes the distribution look more like a normal distribution (see next slide)
  - economic theory (Mince 1974, Willis 1986) suggests that effects on wages are relative, not absolute; differences in logs correspond to ratios on the original scale:

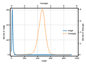
$$\ln(x/y) = \ln(x) - \ln(y) \quad \text{hence: } \exp(\ln(x) - \ln(y)) = x/y$$

- The mean difference in log wages can approximately be interpreted as the percentage difference in average wages.
  - More precisely: the mean difference in log wages corresponds to the ratio of geometric means of wages

$$\exp\left(\overline{\ln x} - \overline{\ln y}\right) = \frac{\tilde{x}}{\tilde{y}}$$

where $\tilde{x} = \sqrt[n]{x_1 x_2 \cdots x_n}$ is the geometric mean of x.


ln(wage)
-2
0
2
4
6
8
.06
.04
.02
0
density of wage
.8
.6
.4
.2
0
density of ln(wage)
wage
ln(wage)
0
200
400
600
800
1000
wage



```
. twoway (kdens wage, ll(0) ) (kdens lnwage,  yaxis(2) xaxis(2)), ///
>     xti(wage) xti(ln(wage), axis(2)) ///
>     yti(density of wage) yti(density of ln(wage), axis(2)) ///
>     legend(order(1 "wage" 2 "ln(wage)") pos(3))
(bandwidth = 2.3878868)
(bandwidth = .16802291)
```

The gender wage gap

```
. mean lnwage if schooling<. & ft_experience<., over(bcsex)
Mean estimation                   Number of obs    =      7,860
    _subpop_1: bcsex = [1] Maennlich
    _subpop_2: bcsex = [2] Weiblich
```

Over	Mean	Std. Err.	[95% Conf. Interval]	
lnwage				
_subpop_1	2.749054	.0092334	2.730954	2.767153
_subpop_2	2.498484	.0091986	2.480452	2.516516

```
. lincom _subpop_1-_subpop_2
 ( 1)  [lnwage]_subpop_1 - [lnwage]_subpop_2 = 0
```

Mean	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
(1)	.2505696	.0130334	19.23	0.000	.2250207	.2761185

```
. nlcom exp(_b[_subpop_1])/exp(_b[_subpop_2])
       _nl_1:  exp(_b[_subpop_1])/exp(_b[_subpop_2])
```

Mean	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
_nl_1	1.284757	.0167447	76.73	0.000	1.251938	1.317576

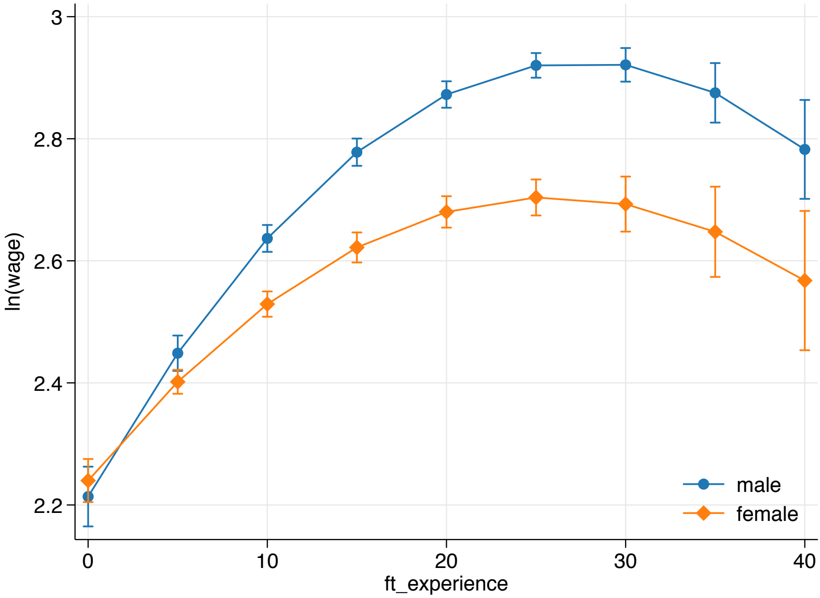
```
. nlcom (exp(_b[_subpop_1]-_b[_subpop_2])-1)*100
       _nl_1:  (exp(_b[_subpop_1]-_b[_subpop_2])-1)*100
```

Mean	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
_nl_1	28.4757	1.674472	17.01	0.000	25.1938	31.7576

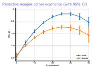
Separate wage regressions by gender

```
. bysort bcsex: regress lnwage schooling ft_experience ft_experience2

------------------------------------------------------------------------------------------------------
-> bcsex = [1] Maennlich

      Source |       SS           df       MS      Number of obs   =     3,877
-------------+----------------------------------   F(3, 3873)      =    443.01
       Model |  327.313727         3  109.104576   Prob > F        =    0.0000
    Residual |  953.834709     3,873  .246278004   R-squared       =    0.2555
-------------+----------------------------------   Adj R-squared   =    0.2549
       Total |  1281.14844     3,876  .330533652   Root MSE        =    .49626

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   schooling |   .0933227   .0029897    31.21   0.000     .0874611    .0991844
ft_experience|   .0516494      .0031    16.66   0.000     .0455717    .0577272
ft_experience2|  -.0009358   .0000859   -10.89   0.000    -.0011042   -.0007673
       _cons |   1.000596   .0487866    20.51   0.000     .9049461    1.096246
------------------------------------------------------------------------------

------------------------------------------------------------------------------------------------------
-> bcsex = [2] Weiblich

      Source |       SS           df       MS      Number of obs   =     3,983
-------------+----------------------------------   F(3, 3979)      =    352.47
       Model |  281.757765         3   93.919255   Prob > F        =    0.0000
    Residual |  1060.24333     3,979  .266459746   R-squared       =    0.2100
-------------+----------------------------------   Adj R-squared   =    0.2094
       Total |   1342.0011     3,982   .33701685   Root MSE        =     .5162

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   schooling |    .086751    .003054    28.41   0.000     .0807635    .0927385
ft_experience|   .0358245   .0029841    12.01   0.000     .0299741     .041675
ft_experience2|  -.0006908   .0000953    -7.25   0.000    -.0008777   -.0005039
       _cons |   1.112193   .0442856    25.11   0.000     1.025369    1.199018
------------------------------------------------------------------------------
```

Predictive margins across experience (with 95% CI)



Predictive margins across experience (with 95% CI)




```
regress lnwage schooling c.ft_experience##c.ft_experience if bcsex==1
margins, at(schooling=13 ft_experience=(0(5)40)) post
est sto male
regress lnwage schooling c.ft_experience##c.ft_experience if bcsex==2
margins, at(schooling=13 ft_experience=(0(5)40)) post
est sto female
coefplot male female, at recast(connect) ciopts(recast(rcap)) ///
    xtitle(ft_experience) yti(ln(wage))
```

Means of the X variables by gender

```
. mean schooling ft_experience ft_experience2 if lnwage<., over(bcsex)
Mean estimation                    Number of obs   =      7,860

    _subpop_1: bcsex = [1] Maennlich
    _subpop_2: bcsex = [2] Weiblich
```

Over	Mean	Std. Err.	[95% Conf.	Interval]
schooling				
_subpop_1	12.88664	.0445749	12.79926	12.97402
_subpop_2	12.97452	.0426577	12.8909	13.05814
ft_experience				
_subpop_1	18.38458	.1552555	18.08023	18.68892
_subpop_2	11.27442	.1418485	10.99636	11.55248
ft_experience2				
_subpop_1	431.4208	5.604688	420.4341	442.4075
_subpop_2	207.2343	4.439645	198.5314	215.9372

Aggregate Oaxaca-Blinder decomposition: by hand

- Explained part

```
. display  .0933227 * (12.88664 - 12.97452)   ///
>        +  .0516494 * (18.38458 - 11.27442)   ///
>        + -.0009358 * (431.4208 - 207.2343)
.14924057
```

- Unexplained part

```
. display (1.000596  - 1.112193 )             ///
>        + ( .0933227 -  .086751 ) * 12.97452  ///
>        + ( .0516494 -  .0358245) * 11.27442  ///
>        + (-.0009358 - -.0006908) * 207.2343
.10131182
```

Aggregate Oaxaca-Blinder decomposition: oaxaca

```
. oaxaca lnwage schooling ft_experience ft_experience2, by(bcsex) weight(1) nodetail
Blinder-Oaxaca decomposition                      Number of obs     =      7,860
                                                    Model             =     linear
Group 1: bcsex = 1                                  N of obs 1        =       3877
Group 2: bcsex = 2                                  N of obs 2        =       3983
```

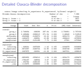
lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overall						
group_1	2.749054	.009236	297.64	0.000	2.730951	2.767156
group_2	2.498484	.0092013	271.54	0.000	2.48045	2.516518
difference	.2505696	.0130372	19.22	0.000	.2250172	.276122
explained	.1492473	.009391	15.89	0.000	.1308412	.1676533
unexplained	.1013223	.0131188	7.72	0.000	.07561	.1270346

Option `weight(1)` requests using a counterfactual as defined above; option `nodetail` suppresses the detailed decomposition.

Detailed Oaxaca-Blinder decomposition

```
. oaxaca lnwage schooling ft_experience ft_experience2, by(bcsex) weight(1)
Blinder-Oaxaca decomposition                       Number of obs     =      7,860
                                                     Model             =     linear
Group 1: bcsex = 1                                   N of obs 1        =       3877
Group 2: bcsex = 2                                   N of obs 2        =       3983
```

lnwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
overall						
group_1	2.749054	.009236	297.64	0.000	2.730951	2.767156
group_2	2.498484	.0092013	271.54	0.000	2.48045	2.516518
difference	.2505696	.0130372	19.22	0.000	.2250172	.276122
explained	.1492473	.009391	15.89	0.000	.1308412	.1676533
unexplained	.1013223	.0131188	7.72	0.000	.07561	.1270346
explained						
schooling	-.008201	.0057638	-1.42	0.155	-.0194978	.0030958
ft_experience	.3672357	.0245724	14.95	0.000	.3190748	.4153967
ft_experience2	-.2097875	.020391	-10.29	0.000	-.2497531	-.1698218
unexplained						
schooling	.0852652	.0554512	1.54	0.124	-.0234172	.1939476
ft_experience	.1784167	.048564	3.67	0.000	.0832329	.2736004
ft_experience2	-.050762	.0266193	-1.91	0.057	-.1029349	.0014109
_cons	-.1115975	.0658889	-1.69	0.090	-.2407374	.0175423



FAQ:

Huh, the contribution of schooling to the explained part is negative.

How can that be? What's going wrong?

Answer:

Negative contributions are perfectly fine. This simply means that the overall difference would even be larger if average schooling of men and women would be the same. In the example, the explanation is that schooling has a positive effect on wages and that women have, on average, slightly more schooling than men. If we eliminate this schooling advantage of women, they would be even worse off and, hence, the wage gap would increase.

Subsuming the contribution of experience

```
. oaxaca lnwage schooling (experience: ft_experience ft_experience2), by(bcsex) weight(1)
Blinder-Oaxaca decomposition                    Number of obs     =      7,860
                                                  Model           =     linear
Group 1: bcsex = 1                                N of obs 1      =       3877
Group 2: bcsex = 2                                N of obs 2      =       3983

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.749054    .009236   297.64   0.000     2.730951    2.767156
     group_2 |   2.498484   .0092013   271.54   0.000      2.48045    2.516518
  difference |   .2505696   .0130372    19.22   0.000     .2250172     .276122
   explained |   .1492473    .009391    15.89   0.000     .1308412    .1676533
 unexplained |   .1013223   .0131188     7.72   0.000       .07561    .1270346
-------------+----------------------------------------------------------------
explained    |
   schooling |   -.008201   .0057638    -1.42   0.155    -.0194978    .0030958
  experience |   .1574483   .0080355    19.59   0.000     .1416989    .1731976
-------------+----------------------------------------------------------------
unexplained  |
   schooling |   .0852652   .0554512     1.54   0.124    -.0234172    .1939476
  experience |   .1276546   .0245238     5.21   0.000     .0795889    .1757204
       _cons |  -.1115975   .0658889    -1.69   0.090    -.2407374    .0175423
------------------------------------------------------------------------------
experience: ft_experience ft_experience2

. estimates store unconditional
```

Bootstrap standard errors

```
. oaxaca lnwage schooling (experience: ft_experience ft_experience2), ///
>     by(bcsex) weight(1) vce(bootstrap, reps(100))
(running oaxaca on estimation sample)

Bootstrap replications (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100

Blinder-Oaxaca decomposition                   Number of obs     =      7,860
                                               Replications      =        100
                                               Model              =     linear
Group 1: bcsex = 1                             N of obs 1         =       3877
Group 2: bcsex = 2                             N of obs 2         =       3983

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
      lnwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.749054   .0092526   297.11   0.000     2.730919    2.767188
     group_2 |   2.498484   .0080223   311.44   0.000     2.482761    2.514207
  difference |   .2505696   .0115967    21.61   0.000     .2278404    .2732988
   explained |   .1492473   .0081171    18.39   0.000      .133338    .1651566
 unexplained |   .1013223   .0135516     7.48   0.000     .0747616     .127883
-------------+----------------------------------------------------------------
explained    |
   schooling |   -.008201   .0058454    -1.40   0.161    -.0196578    .0032558
  experience |   .1574483   .0084314    18.67   0.000      .140923    .1739735
-------------+----------------------------------------------------------------
unexplained  |
   schooling |   .0852652   .0571485     1.49   0.136    -.0267439    .1972743
  experience |   .1276546   .0251543     5.07   0.000     .0783531    .1769561
       _cons |  -.1115975    .068842    -1.62   0.105    -.2465254    .0233303
------------------------------------------------------------------------------
experience: ft_experience ft_experience2

. estimates store bootstrap
```

Analytic vs. bootstrap standard errors

```
. oaxaca lnwage schooling (experience: ft_experience ft_experience2), ///
>     by(bcsex) weight(1) fixed
```

(*output omitted*)

```
. estimates store conditional
. esttab conditional unconditional bootstrap, nogap wide se mtitle nostar nonumber
```

	conditional		unconditio~l		bootstrap	
overall						
group_1	2.749	(0.00797)	2.749	(0.00924)	2.749	(0.00925)
group_2	2.498	(0.00818)	2.498	(0.00920)	2.498	(0.00802)
difference	0.251	(0.0114)	0.251	(0.0130)	0.251	(0.0116)
explained	0.149	(0.00633)	0.149	(0.00939)	0.149	(0.00812)
unexplained	0.101	(0.0131)	0.101	(0.0131)	0.101	(0.0136)
explained						
schooling	-0.00820	(0.000263)	-0.00820	(0.00576)	-0.00820	(0.00585)
experience	0.157	(0.00639)	0.157	(0.00804)	0.157	(0.00843)
unexplained						
schooling	0.0853	(0.0555)	0.0853	(0.0555)	0.0853	(0.0571)
experience	0.128	(0.0245)	0.128	(0.0245)	0.128	(0.0252)
_cons	-0.112	(0.0659)	-0.112	(0.0659)	-0.112	(0.0688)
N	7860		7860		7860	

Standard errors in parentheses

Exercise 1

- Extend the X variables of the model by tenure ("Dauer der Betriebszugehörigkeit") and the "ISEI". Also take account of the survey design (clustering by households, sampling weights `bcphrf`).
- Compute the aggregate and detailed Oaxaca-Blinder decomposition. How did the results change compared to the specification used in the example analysis?
- Confirm the results returned by `oaxaca` by computing the aggregate Blinder-Oaxaca decomposition "by hand" (that is, estimate the means of the variables and the regression coefficients and then compute the decomposition from these outputs, and not by using `oaxaca`). Also compute the contribution of schooling in $\hat{\Delta}^{\mu}_{X}$ and $\hat{\Delta}^{\mu}_{S}$ by hand.

References

- Blinder, Alan S. (1973). Wage Discrimination: Reduced Form and Structural Estimates. The Journal of Human Resources 8(4):436–455.
- Greene, William H. (2003). Econometric Analysis. 5. Upper Saddle River, NJ: Pearson Education.
- Jann, Ben (2005). Standard errors for the Blinder-Oaxaca decomposition. 2005 German Stata Users Group meeting. https://ideas.repec.org/p/boc/dsug05/03.html.
- Jann, Ben (2008). The Blinder-Oaxaca decomposition for linear regression models. The Stata Journal 8(4):453–479.
- Mincer, Jacob (1974). Schooling, Experience and Earnings. New York and London: Columbia University Press.
- Oaxaca, Ronald (1973). Male-Female Wage Differentials in Urban Labor Markets. International Economic Review 14(3):693–709.

References

- Oaxaca, Ronald L., Michael Ransom (1998). Calculation of approximate variances for wage decomposition differentials. Journal of Economic and Social Measurement 24:55–61.
- Willis, Robert J. (1986). Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions. In Orley Ashenfelter and Richard Layard (Eds.), Handbook of Labor Economics (pp. 525-602). Amsterdam: North-Holland.