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)

2. The 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 $X_1, \ldots, X_K$ 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

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

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

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

- 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 $E(X|G=1)\beta^0$, we obtain the decomposition

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

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 $\widehat{\beta}^g$ be the estimate of β^g and $\bar{X}^g = \widehat{\mathrm{E}}(X|G = g)$ be the estimate of $\mathrm{E}(X|G = g)$. The decomposition estimate then is

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

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

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

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

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

    (see the proof in Jann 2005).
  - The last term, trace{}, is asymptotically vanishing and can be ignored.
  - To estimate $V(\bar{X}\widehat{\beta})$, plug in estimates for $V(\widehat{\beta})$ (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:

$$V(\widehat{\Delta}_X^\mu) = V((\bar{X}^0 - \bar{X}^1)\widehat{\beta}^0) \approx (\bar{X}^0 - \bar{X}^1)V(\widehat{\beta}^0)(\bar{X}^0 - \bar{X}^1)' + \widehat{\beta}^{0\prime}\left[V(\bar{X}^0) + V(\bar{X}^1)\right]\widehat{\beta}^0$$

$$V(\widehat{\Delta}_S^\mu) = V(\bar{X}^1(\widehat{\beta}^0 - \widehat{\beta}^1)) \approx \bar{X}^1\left[V(\widehat{\beta}^0) + V(\widehat{\beta}^1)\right]\bar{X}^{1\prime} + (\widehat{\beta}^0 - \widehat{\beta}^1)'V(\bar{X}^1)(\widehat{\beta}^0 - \widehat{\beta}^1)$$

- 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

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

- For the "unexplained" part we have

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

Detailed decomposition

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

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

$$\widehat{\Delta}^{\mu}_{S} = (\widehat{\beta}^{0}_{0} - \widehat{\beta}^{1}_{0}) + \sum_{k=1}^{a}(\widehat{\beta}^{0}_{k} - \widehat{\beta}^{1}_{k})\bar{X}^{1}_{k} + \sum_{k=a+1}^{b}(\widehat{\beta}^{0}_{k} - \widehat{\beta}^{1}_{k})\bar{X}^{1}_{k} + \dots$$

Example analysis

- Data: `gsoep-extract.dta`; extract from German Socio-Economic Panel (GSOEP), waves 1995, 2005, 2015, 2020
- 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 gsoep-extract, clear
(Example data based on the German Socio-Economic Panel)
. // selection
. keep if wave==2015
(29,970 observations deleted)
. keep if inrange(age, 25, 55)
(5,671 observations deleted)
. // variables
. generate lnwage = ln(wage)
(1,709 missing values generated)
. generate expft2 = expft^2
(35 missing values generated)
. summarize wage lnwage yeduc expft expft2
```

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

Summarize wages by gender

```
. bysort sex: summarize wage if wage>0 & yeduc<. & expft<., detail

------------------------------------------------------------------------------------------

-> sex = male

                    gross hourly wage
-------------------------------------------------------------
      Percentiles      Smallest
 1%         5.04           3.05
 5%         7.77           3.05
10%         9.23           3.08       Obs               2,642
25%        12.46           3.45       Sum of wgt.       2,642

50%        17.33                      Mean           19.81089
                        Largest       Std. dev.      10.89243
75%        24.58         101.33
90%        33.24         103.02       Variance       118.6451
95%        38.84         105.62       Skewness       2.237586
99%        53.79         121.42       Kurtosis       13.73294

------------------------------------------------------------------------------------------

-> sex = female

                    gross hourly wage
-------------------------------------------------------------
      Percentiles      Smallest
 1%         4.25           3.03
 5%         6.38           3.05
10%        7.685            3.1       Obs               2,820
25%        9.895           3.26       Sum of wgt.       2,820

50%       14.015                      Mean           15.53262
                        Largest       Std. dev.      8.307052
75%       19.035          69.56
90%        25.28          72.16       Variance       69.00712
95%        30.03         117.53       Skewness       2.827928
99%        41.72          119.3       Kurtosis       23.94746
```

The gender wage gap

```
. mean wage if wage>0 & yeduc<. & expft<., over(sex)
Mean estimation                          Number of obs = 5,462

--------------------------------------------------------------
             |       Mean   Std. err.     [95% conf. interval]
-------------+------------------------------------------------
  c.wage@sex |
       male  |   19.81089   .2119134      19.39545    20.22632
     female  |   15.53262   .1564308      15.22595    15.83928
--------------------------------------------------------------

. lincom c.wage@1.sex-c.wage@2.sex
 ( 1)  c.wage@1bn.sex - c.wage@2.sex = 0

------------------------------------------------------------------------------
        Mean | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         (1) |   4.278272   .2633969    16.24   0.000      3.76191    4.794635
------------------------------------------------------------------------------

. nlcom _b[c.wage@1.sex]/_b[c.wage@2.sex]
       _nl_1: _b[c.wage@1.sex]/_b[c.wage@2.sex]

------------------------------------------------------------------------------
        Mean | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _nl_1 |   1.275438   .0187385    68.07   0.000     1.238711    1.312165
------------------------------------------------------------------------------

. nlcom (_b[c.wage@1.sex]/_b[c.wage@2.sex]-1)*100
       _nl_1: (_b[c.wage@1.sex]/_b[c.wage@2.sex]-1)*100

------------------------------------------------------------------------------
        Mean | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _nl_1 |    27.5438   1.873849    14.70   0.000     23.87112    31.21647
------------------------------------------------------------------------------
```

The gender wage gap

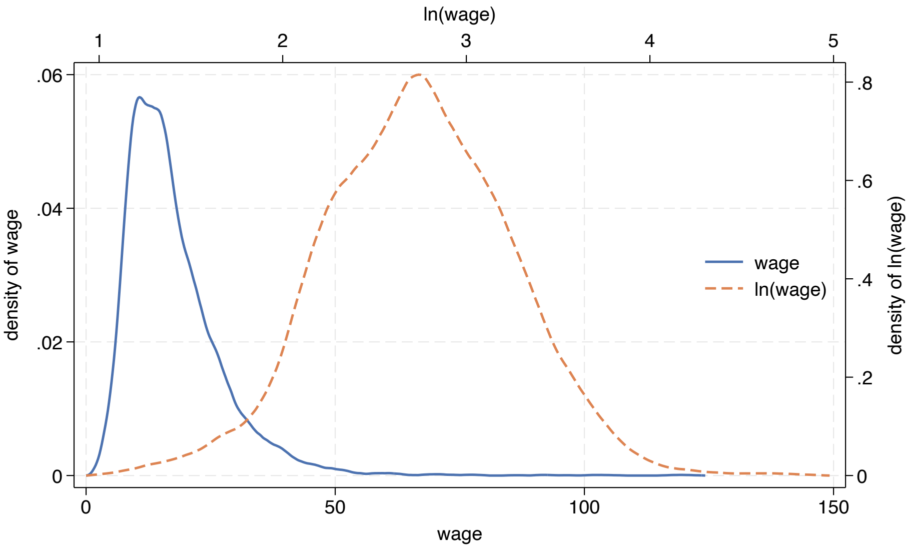
- Typically, the *logarithm* of wages is analyzed, because
  - wages can only be positive; $Y \in (0, \infty)$
  - wages have a (left) skewed distribution; taking the logarithm makes the distribution look more like a normal distribution (see next slide)
  - economic theory (Mincer 1974, Willis 1986)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

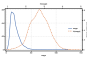
$$\ln(x/y) = \ln(x) - \ln(y) \quad \text{hence: } \exp(\ln(x) - \ln(y)) = x/y$$

- 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)
1
2
3
4
5
.06
.04
.02
0
density of wage
.8
.6
.4
.2
0
density of ln(wage)
wage
ln(wage)
0
50
100
150
wage



```
twoway (kdens wage if wage>0, 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))
```

The gender wage gap

```
. mean lnwage if yeduc<. & expft<., over(sex)
Mean estimation                         Number of obs = 5,462
```

	Mean	Std. err.	[95% conf.	interval]
c.lnwage@sex				
male	2.858357	.0098305	2.839085	2.877629
female	2.62663	.009016	2.608955	2.644305

```
. lincom c.lnwage@1.sex-c.lnwage@2.sex
 ( 1)  c.lnwage@1bn.sex - c.lnwage@2.sex = 0
```

Mean	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
(1)	.2317274	.0133389	17.37	0.000	.2055777	.257877

```
. nlcom exp(_b[c.lnwage@1.sex])/exp(_b[c.lnwage@2.sex])
       _nl_1: exp(_b[c.lnwage@1.sex])/exp(_b[c.lnwage@2.sex])
```

Mean	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
_nl_1	1.260776	.0168174	74.97	0.000	1.227814	1.293737

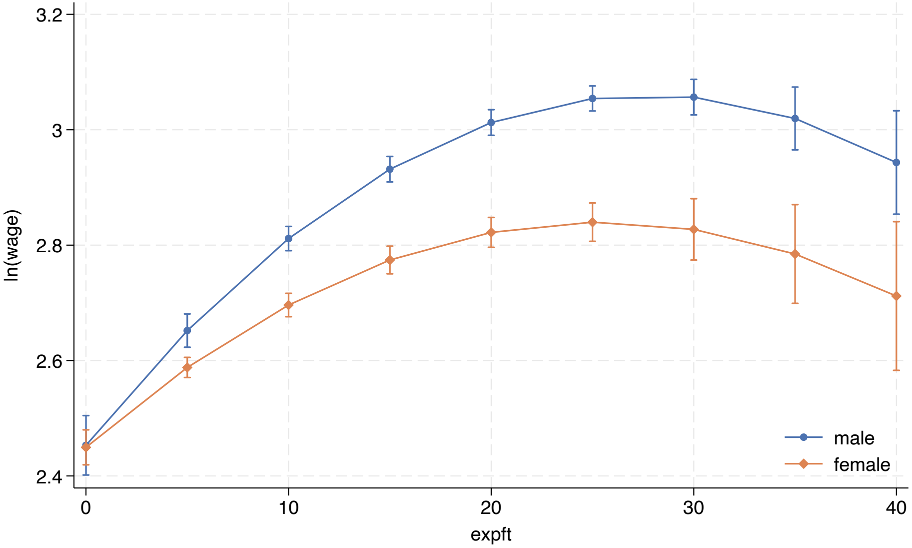
```
. nlcom (exp(_b[c.lnwage@1.sex]-_b[c.lnwage@2.sex])-1)*100
       _nl_1: (exp(_b[c.lnwage@1.sex]-_b[c.lnwage@2.sex])-1)*100
```

Mean	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
_nl_1	26.07759	1.681741	15.51	0.000	22.78144	29.37375

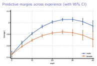
Separate wage regressions by gender

```
. bysort sex: regress lnwage yeduc expft expft2

--------------------------------------------------------------------------------------------

-> sex = male

      Source |       SS           df       MS      Number of obs   =     2,642
-------------+----------------------------------   F(3, 2638)      =    428.35
       Model |  220.878193         3  73.6260643   Prob > F        =    0.0000
    Residual |  453.426469     2,638  .171882664   R-squared       =    0.3276
-------------+----------------------------------   Adj R-squared   =    0.3268
       Total |  674.304662     2,641   .25532172   Root MSE        =    .41459

------------------------------------------------------------------------------
      lnwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       yeduc |   .0909759    .002884    31.54   0.000     .0853207    .0966311
       expft |   .0436904   .0033515    13.04   0.000     .0371187    .0502622
      expft2 |  -.0007859   .0000935    -8.41   0.000    -.0009692   -.0006026
       _cons |   1.270429   .0456615    27.82   0.000     1.180893    1.359965
------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------

-> sex = female

      Source |       SS           df       MS      Number of obs   =     2,820
-------------+----------------------------------   F(3, 2816)      =    346.11
       Model |  174.081563         3  58.0271878   Prob > F        =    0.0000
    Residual |  472.121431     2,816  .167656758   R-squared       =    0.2694
-------------+----------------------------------   Adj R-squared   =    0.2686
       Total |  646.202995     2,819  .229231286   Root MSE        =    .40946

------------------------------------------------------------------------------
      lnwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       yeduc |   .0818828   .0028763    28.47   0.000      .076243    .0875226
       expft |   .0306978   .0028001    10.96   0.000     .0252075    .0361882
      expft2 |  -.0006035   .0000964    -6.26   0.000    -.0007927   -.0004144
       _cons |    1.38513   .0400425    34.59   0.000     1.306614    1.463646
------------------------------------------------------------------------------
```

Predictive margins across experience (with 95% CI)



Predictive margins across experience (with 95% CI)




```
regress lnwage yeduc c.expft##c.expft if sex==1
margins, at(yeduc=13 expft=(0(5)40)) post
estimates store male
regress lnwage yeduc c.expft##c.expft if sex==2
margins, at(yeduc=13 expft=(0(5)40)) post
estimates store female
coefplot male female, at recast(connect) ciopts(recast(rcap)) ///
    xtitle(expft) yti(ln(wage))
```

Means of the X variables by gender

```
. mean yeduc expft expft2 if lnwage<., over(sex)
Mean estimation                          Number of obs = 5,462
```

	Mean	Std. err.	[95% conf.	interval]
c.yeduc@sex				
male	12.4788	.05532	12.37035	12.58725
female	12.73936	.0506089	12.64015	12.83858
c.expft@sex				
male	17.31501	.1812995	16.95959	17.67043
female	9.616578	.1552872	9.312153	9.921003
c.expft2@sex				
male	386.6178	6.509539	373.8565	399.3791
female	160.4562	4.509838	151.6151	169.2973

Aggregate Oaxaca-Blinder decomposition: by hand

- Explained part

```
. display %9.0g (  12.4788 -  12.73936) *  .0909759 ///
>             + ( 17.31501 -  9.616578) *  .0436904 ///
>             + ( 386.6178 -  160.4562) * -.0007859
 .1349025
```

- Unexplained part

```
. display %9.0g            ( 1.270429 -   1.38513) ///
>            +  12.73936 * ( .0909759 -  .0818828) ///
>            +  9.616578 * ( .0436904 -  .0306978) ///
>            +  160.4562 * (-.0007859 - -.0006035)
 .0968164
```

Aggregate Oaxaca-Blinder decomposition: oaxaca

```
. oaxaca lnwage yeduc expft expft2, by(sex) weight(1) nodetail
Blinder-Oaxaca decomposition                    Number of obs     =      5,462
                                                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)
```

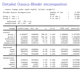
lnwage	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
overall						
group_1	2.858357	.0098343	290.65	0.000	2.839082	2.877632
group_2	2.62663	.0090195	291.22	0.000	2.608952	2.644308
difference	.2317274	.0133441	17.37	0.000	.2055734	.2578813
explained	.1349029	.0111087	12.14	0.000	.1131302	.1566756
unexplained	.0968245	.0136114	7.11	0.000	.0701467	.1235023

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

Detailed Oaxaca-Blinder decomposition

```
. oaxaca lnwage yeduc expft expft2, by(sex) weight(1)
Blinder-Oaxaca decomposition                    Number of obs     =      5,462
                                                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)

------------------------------------------------------------------------------
      lnwage | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.858357   .0098343   290.65   0.000     2.839082    2.877632
     group_2 |    2.62663   .0090195   291.22   0.000     2.608952    2.644308
  difference |   .2317274   .0133441    17.37   0.000     .2055734    .2578813
   explained |   .1349029   .0111087    12.14   0.000     .1131302    .1566756
 unexplained |   .0968245   .0136114     7.11   0.000     .0701467    .1235023
-------------+----------------------------------------------------------------
explained    |
       yeduc |  -.0237045   .0068624    -3.45   0.001    -.0371545   -.0102545
       expft |   .3363478   .0278291    12.09   0.000     .2818037    .3908919
      expft2 |  -.1777404   .0220436    -8.06   0.000    -.2209452   -.1345357
-------------+----------------------------------------------------------------
unexplained  |
       yeduc |   .1158413   .0518914     2.23   0.026      .014136    .2175466
       expft |   .1249445   .0420461     2.97   0.003     .0425357    .2073533
      expft2 |    -.02926     .02157    -1.36   0.175    -.0715366    .0130165
       _cons |  -.1147013    .060732    -1.89   0.059    -.2337338    .0043313
------------------------------------------------------------------------------
```

Detailed Oaxaca-Blinder decomposition



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 yeduc (experience: expft expft2), by(sex) weight(1)
Blinder-Oaxaca decomposition                  Number of obs     =      5,462
                                              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)
```

lnwage	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
overall						
group_1	2.858357	.0098343	290.65	0.000	2.839082	2.877632
group_2	2.62663	.0090195	291.22	0.000	2.608952	2.644308
difference	.2317274	.0133441	17.37	0.000	.2055734	.2578813
explained	.1349029	.0111087	12.14	0.000	.1131302	.1566756
unexplained	.0968245	.0136114	7.11	0.000	.0701467	.1235023
explained						
yeduc	-.0237045	.0068624	-3.45	0.001	-.0371545	-.0102545
experience	.1586074	.0091482	17.34	0.000	.1406772	.1765375
unexplained						
yeduc	.1158413	.0518914	2.23	0.026	.014136	.2175466
experience	.0956845	.0226133	4.23	0.000	.0513632	.1400057
_cons	-.1147013	.060732	-1.89	0.059	-.2337338	.0043313

```
experience: expft expft2

. estimates store unconditional
```

Bootstrap standard errors

```
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) weight(1) vce(bootstrap, reps(100))
(running oaxaca on estimation sample)
Bootstrap replications (100): .........10.........20.........30.........40.........50.........6
> 0.........70.........80.........90.........100 done
Blinder-Oaxaca decomposition                          Number of obs     =      5,462
                                                      Replications      =        100
                                                      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)
```

lnwage	Observed coefficient	Bootstrap std. err.	z	P>|z|	Normal-based [95% conf.	interval]
overall						
group_1	2.858357	.0101971	280.31	0.000	2.838371	2.878343
group_2	2.62663	.0080962	324.43	0.000	2.610761	2.642498
difference	.2317274	.0130774	17.72	0.000	.2060961	.2573586
explained	.1349029	.0116026	11.63	0.000	.1121621	.1576436
unexplained	.0968245	.0150469	6.43	0.000	.067333	.1263159
explained						
yeduc	-.0237045	.0066624	-3.56	0.000	-.0367625	-.0106464
experience	.1586074	.0104849	15.13	0.000	.1380573	.1791574
unexplained						
yeduc	.1158413	.0461041	2.51	0.012	.0254789	.2062037
experience	.0956845	.0226534	4.22	0.000	.0512847	.1400843
_cons	-.1147013	.0567826	-2.02	0.043	-.2259931	-.0034095

```
experience: expft expft2
. estimates store bootstrap
```

Analytic vs. bootstrap standard errors

```
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) 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	2.858	(0.00807)	2.858	(0.00983)	2.858	(0.0102)
group_2	2.627	(0.00771)	2.627	(0.00902)	2.627	(0.00810)
difference	0.232	(0.0112)	0.232	(0.0133)	0.232	(0.0131)
explained	0.135	(0.00768)	0.135	(0.0111)	0.135	(0.0116)
unexplained	0.0968	(0.0135)	0.0968	(0.0136)	0.0968	(0.0150)
explained						
yeduc	-0.0237	(0.000751)	-0.0237	(0.00686)	-0.0237	(0.00666)
experience	0.159	(0.00773)	0.159	(0.00915)	0.159	(0.0105)
unexplained						
yeduc	0.116	(0.0519)	0.116	(0.0519)	0.116	(0.0461)
experience	0.0957	(0.0226)	0.0957	(0.0226)	0.0957	(0.0227)
_cons	-0.115	(0.0607)	-0.115	(0.0607)	-0.115	(0.0568)
N	5462		5462		5462	

Standard errors in parentheses

Exercise 1

Post-estimation commands

- Similar to other estimation commands in Stata, `oaxaca` leaves results behind in `e(b)` and `e(V)` so that they can be processed by post-estimation commands.
- Examples are:
  - Command `test` and `testnl` to perform hypothesis tests.
  - Commands `lincom` and `nlcom` to compute linear and non-linear combinations (and the corresponding standard errors).
  - Commands such as `esttab` (Jann 2007) and `coefplot` (Jann 2014) to make tables and graphs from results.
- For many of these commands it is important to know how the elements in `e(b)` are named. Type

  ```
  . ereturn display, coeflegend
  ```

  after running `oaxaca` to display the names.

Hypothesis tests

- In its standard output, `oaxaca` displays tests of the individual components against zero.
- Depending on context, tests against other values might be required and you might also want to perform joint tests of multiple hypotheses.
- A general command to perform so-called Wald tests of simple and composite linear hypotheses, is `test`. A command for nonlinear hypotheses is `testnl`.

Hypothesis tests

```
. oaxaca lnwage yeduc expft expft2, by(sex) weight(1)
  (output omitted)
. ereturn display, coeflegend
```

lnwage	Coefficient	Legend
overall		
group_1	2.858357	_b[overall:group_1]
group_2	2.62663	_b[overall:group_2]
difference	.2317274	_b[overall:difference]
explained	.1349029	_b[overall:explained]
unexplained	.0968245	_b[overall:unexplained]
explained		
yeduc	-.0237045	_b[explained:yeduc]
expft	.3363478	_b[explained:expft]
expft2	-.1777404	_b[explained:expft2]
unexplained		
yeduc	.1158413	_b[unexplained:yeduc]
expft	.1249445	_b[unexplained:expft]
expft2	-.02926	_b[unexplained:expft2]
_cons	-.1147013	_b[unexplained:_cons]

Examples

- Test that the explained part is different from the unexplained part:

```
. test _b[overall:explained] = _b[overall:unexplained]
 ( 1)  [overall]explained - [overall]unexplained = 0
           chi2(  1) =    3.30
         Prob > chi2 =    0.0692
```

- Joint test of the contributions of `expft` and `expft2` to the explained part against zero:

```
. test _b[explained:expft] = 0
```

  (*output omitted*)

```
. test _b[explained:expft2] = 0, accum
 ( 1)  [explained]expft = 0
 ( 2)  [explained]expft2 = 0
           chi2(  2) =  301.33
         Prob > chi2 =    0.0000
```

- This is a different test than testing their joint contribution:

```
. test _b[explained:expft] + _b[explained:expft2] = 0
 ( 1)  [explained]expft + [explained]expft2 = 0
           chi2(  1) =  300.59
         Prob > chi2 =    0.0000
```

Linear and nonlinear combinations

- Close cousins of `test` and `testnl` are commands `lincom` and `nlcom`.
- Command `nlcom` is extremely useful because it can generate arbitrary combinations and transformations of results. Standard errors (and covariances between multiple results) are computed by the so-called "delta method" (linearization; first order Taylor series approximation; see, e.g., Feiveson 1999, Oehlert 1992).
- `lincom` is similar, but can only be used for linear combinations (and only computes one result at the time).

- Express the explained part and the unexplained part as percentage of the overall gap.

```
. nlcom (Percent_explained:   _b[overall:explained]  /_b[overall:difference]*100) ///
>       (Percent_unexplained: _b[overall:unexplained]/_b[overall:difference]*100)

Percent_ex~d: _b[overall:explained]  /_b[overall:difference]*100
Percent_un~d: _b[overall:unexplained]/_b[overall:difference]*100
```

lnwage	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
Percent_explained	58.2162	4.649956	12.52	0.000	49.10246	67.32995
Percent_unexplained	41.7838	4.649956	8.99	0.000	32.67005	50.89754

- Compute the percentage of the overall gap that is explained by schooling (years of education), and the percentage that is explained by work experience.

```
. nlcom (schooling:  _b[explained:yeduc] / _b[overall:difference]*100) ///
>       (experience: (_b[explained:expft] + _b[explained:expft2]) / /*
>      */_b[overall:difference]*100)

   schooling: _b[explained:yeduc] / _b[overall:difference]*100
  experience: (_b[explained:expft] + _b[explained:expft2]) / _b[overall:difference]*100
```

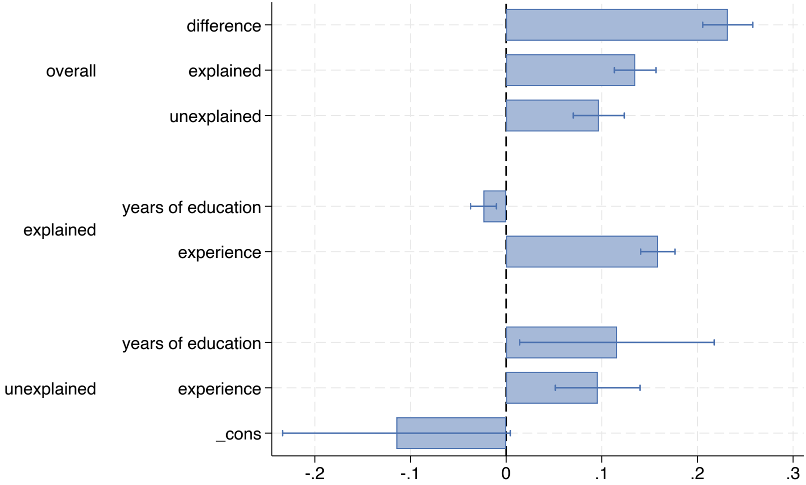
lnwage	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
schooling	-10.22947	3.272162	-3.13	0.002	-16.64279	-3.816153
experience	68.44568	5.191238	13.18	0.000	58.27104	78.62032

Tables and graphs

- To tabulate results from `oaxaca` (and export the table to LaTeX or Word etc.) you can use, for example, command `esttab` (Jann 2007). There are also various other user commands that could be employed.
  - Since Stata 17, there is also official `collect` (and `etable`, which is based on `collect`).
- For graphs, try `coefplot` (Jann 2014).
- The commands support combining results from multiple calls to `oaxaca` or `nlcom` that have been stored using `estimates store`.
- For `nlcom`, you need to specify the `post` option before tabulation and graphing is possible.

Example: graph

```
. oaxaca lnwage yeduc (experience: expft expft2), by(sex) weight(1)
  (output omitted)
. coefplot, drop(overall:group*) xline(0) ///
>     recast(bar) barwidth(.7) base(0) citop ciopts(recast(rcap))
```



Example: display means and coefficients

- Note that `oaxaca` returns the coefficients and means that are used for the decomposition in `e(b0)` and `e(V0)`. Use option `xb` to display these auxiliary statistics.

```
. oaxaca, xb

Blinder-Oaxaca decomposition                    Number of obs     =      5,462
                                                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)

------------------------------------------------------------------------------
      lnwage | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.858357   .0098343   290.65   0.000     2.839082    2.877632
     group_2 |    2.62663   .0090195   291.22   0.000     2.608952    2.644308
  difference |   .2317274   .0133441    17.37   0.000     .2055734    .2578813
   explained |   .1349029   .0111087    12.14   0.000     .1131302    .1566756
 unexplained |   .0968245   .0136114     7.11   0.000     .0701467    .1235023
-------------+----------------------------------------------------------------
explained    |
       yeduc |  -.0237045   .0068624    -3.45   0.001    -.0371545   -.0102545
  experience |   .1586074   .0091482    17.34   0.000     .1406772    .1765375
```

Example: display means and coefficients

unexplained						
yeduc	.1158413	.0518914	2.23	0.026	.014136	.2175466
experience	.0956845	.0226133	4.23	0.000	.0513632	.1400057
_cons	-.1147013	.060732	-1.89	0.059	-.2337338	.0043313

experience: expft expft2

Coefficients (b) and means (x)

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
b1						
yeduc	.0909759	.002884	31.54	0.000	.0853233	.0966285
expft	.0436904	.0033515	13.04	0.000	.0371217	.0502592
expft2	-.0007859	.0000935	-8.41	0.000	-.0009692	-.0006026
_cons	1.270429	.0456615	27.82	0.000	1.180934	1.359924
b2						
yeduc	.0818828	.0028763	28.47	0.000	.0762454	.0875201
expft	.0306978	.0028001	10.96	0.000	.0252098	.0361858
expft2	-.0006035	.0000964	-6.26	0.000	-.0007926	-.0004145
_cons	1.38513	.0400425	34.59	0.000	1.306648	1.463612
b_ref						
yeduc	.0909759	.002884	31.54	0.000	.0853233	.0966285
expft	.0436904	.0033515	13.04	0.000	.0371217	.0502592
expft2	-.0007859	.0000935	-8.41	0.000	-.0009692	-.0006026

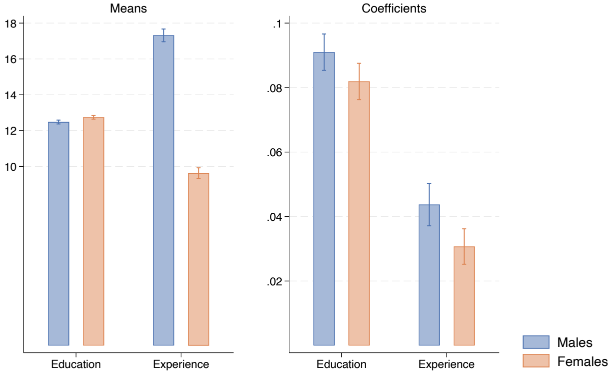
Example: display means and coefficients

_cons	1.270429	.0456615	27.82	0.000	1.180934	1.359924
x1						
yeduc	12.4788	.05532	225.57	0.000	12.37038	12.58723
expft	17.31501	.1812995	95.51	0.000	16.95967	17.67035
expft2	386.6178	6.509539	59.39	0.000	373.8594	399.3763
_cons	1	.	.	.	.	.
x2						
yeduc	12.73936	.0506089	251.72	0.000	12.64017	12.83855
expft	9.616578	.1552872	61.93	0.000	9.312221	9.920935
expft2	160.4562	4.509838	35.58	0.000	151.6171	169.2953
_cons	1	.	.	.	.	.

Example: display means and coefficients

- You can use `coefplot` to draw a graph:

```
. coefplot (. , keep(x1:) drop(_cons expft2)) ///
>          (. , keep(x2:) drop(_cons expft2)), bylabel(Means) ///
>     ||   (. , keep(b1:) drop(_cons expft2)) ///
>          (. , keep(b2:) drop(_cons expft2)), bylabel(Coefficients) ///
>     || , b(b0) v(V0) byopts(yrescale) plotlabels(Males Females) ///
>          coeflabels(yeduc = "Education" expft = "Experience") ///
>          recast(bar) barwidth(.2) base(0) citop ciopts(recast(rcap)) vertical
```



Example: graphing results from `nlcom`

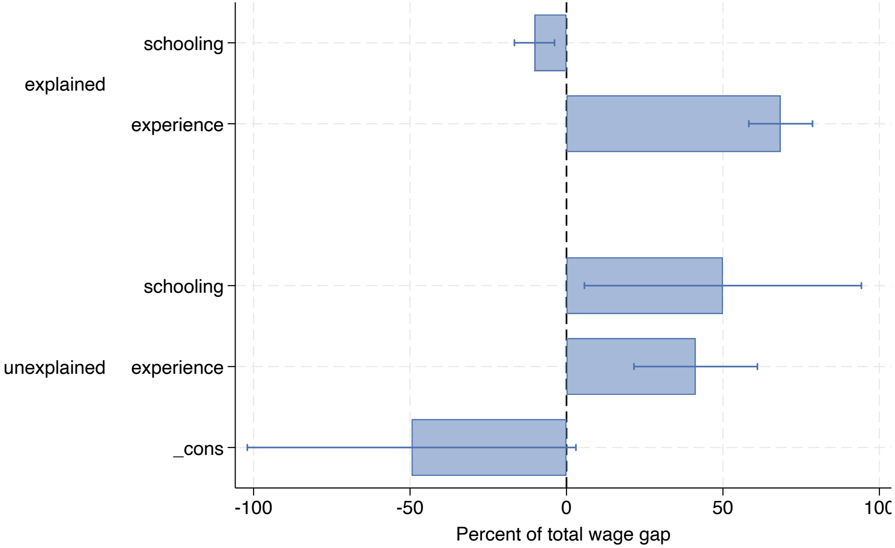
- Use the `post` option in `nlcom` to move the results to `e()` so that they can be tabulated (but be aware that this will delete original results unless they have been saved using `estimates store`).
- In the following example the detailed decomposition results are displayed as percentages of the overall gap.

```
. nlcom (e_schooling:  _b[explained:yeduc]         /_b[overall:difference]*100) ///
>       (e_experience: _b[explained:experience]    /_b[overall:difference]*100) ///
>       (u_schooling:  _b[unexplained:yeduc]       /_b[overall:difference]*100) ///
>       (u_experience: _b[unexplained:experience]  /_b[overall:difference]*100) ///
>       (u__cons:      _b[unexplained:_cons]       /_b[overall:difference]*100) ///
>       , post
```

(*output omitted*)

```
. coefplot (., keep(e_*) asequation(explained)   rename(e_* = "")  ///
>          \ ., keep(u_*) asequation(unexplained) rename(u_* = "")) ///
>     , xline(0) recast(bar) barwidth(.7) base(0) citop ciopts(recast(rcap)) ///
>       xtitle("Percent of total wage gap")
```

Example: graphing results from `nlcom`

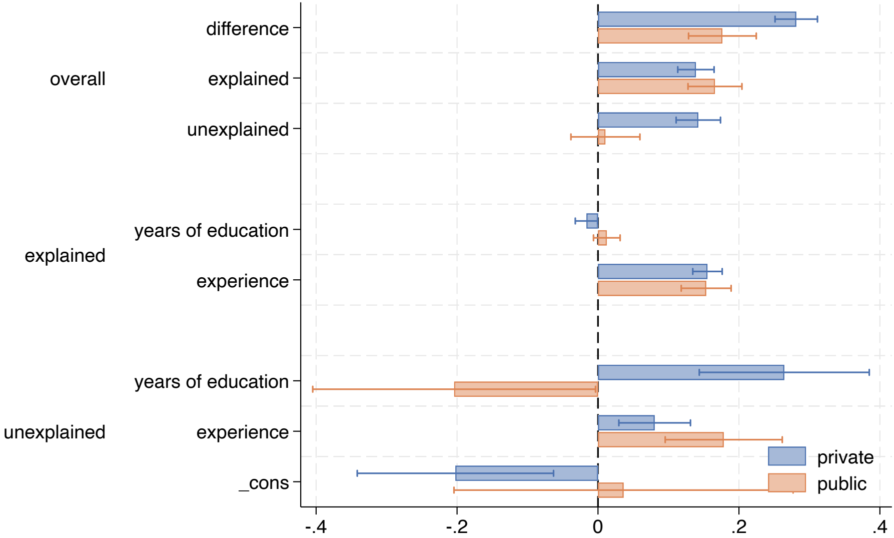


Example: graphing results from multiple decompositions

- Use `estimates store` to hold on to results from a decomposition for later processing.
- Example: wage gap in private sector vs. in public sector.

```
. oaxaca lnwage yeduc (experience: expft expft2) if public==0, by(sex) weight(1)
  (output omitted)
. estimate store private
. oaxaca lnwage yeduc (experience: expft expft2) if public==1 , by(sex) weight(1)
  (output omitted)
. estimate store public
. coefplot private public, drop(overall:group*) xline(0) ///
>     recast(bar) barwidth(.3) base(0) citop ciopts(recast(rcap))
```

Example: graphing results from multiple decompositions



Example: table

```
. oaxaca lnwage yeduc expft expft2, by(sex) weight(1) nodetail
  (output omitted)
. estimates store raw
. nlcom (explained:   _b[overall:explained]  /_b[overall:difference]*100) ///
>       (unexplained: _b[overall:unexplained]/_b[overall:difference]*100), post
  (output omitted)
. estimates store pct
. esttab raw pct using mytable.tex, replace ///
>     keep(difference explained unexplained) nostar ci wide ///
>     noobs nonumber mtitle("Decomposition" "In percent") eqlab(none)
(output written to mytable.tex)
```

- The table looks like this:

	Decomposition		In percent	
difference	0.232	[0.206,0.258]		
explained	0.135	[0.113,0.157]	58.22	[49.10,67.33]
unexplained	0.0968	[0.0701,0.124]	41.78	[32.67,50.90]

95% confidence intervals in brackets

Exercise 2

Program for tomorrow

- The index problem and the transformation problem
- Exercise 3
- Decomposition methods for nonlinear models
- Exercise 4

References

Blinder, Alan S. (1973). Wage Discrimination: Reduced Form and Structural Estimates. The Journal of Human Resources 8(4):436–455.

Feiveson, Alan H. (1999). FAQ: What is the delta method and how is it used to estimate the standard error of a transformed parameter? http://www.stata.com/support/faqs/stat/deltam.html

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 (2007). Making regression tables simplified. The Stata Journal 7(2):227–244.

Jann, Ben (2008). The Blinder-Oaxaca decomposition for linear regression models. The Stata Journal 8(4):453–479.

Jann, Ben (2014). Plotting regression coefficients and other estimates. The Stata Journal 14(4):708–737.

Mincer, Jacob (1974). Schooling, Experience and Earnings. New York and London: Columbia University Press.

References

Oaxaca, Ronald (1973). Male-Female Wage Differentials in Urban Labor Markets. International Economic Review 14(3):693–709.

Oaxaca, Ronald L., Michael Ransom (1998). Calculation of approximate variances for wage decomposition differentials. Journal of Economic and Social Measurement 24:55–61.

Oehlert, Gary W. (1992). A Note on the Delta Method. The American Statistician 46(1):27–29.

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.