Presentation at the 12th UK Stata Users Group meeting

London, 11-12 Septermber 2006

Marginal effects and extending the Blinder-Oaxaca decomposition to nonlinear models

Tamás Bartus

Institute of Sociology and Social Policy

Corvinus University, Budapest

E-mail: tamas.bartus@uni-corvinus.hu

Outline

- Brief description of the Blinder-Oaxaca decomposition
- Motivation
- Extending the Blinder-Oaxaca decomposition to nonlinear models
    - marginal effects approach
    - estimation of standard errors
- The **gdecomp** command
- Empirical example
- Discussion

Introduction I. Blinder-Oaxaca decomposition

Substantive problem

- (1) To what extent can observed racial/gender differences be attributed to to the fact that returns to characteristics $x_1 \ldots x_K$ (*endowments*) is lower among blacks/women than among whites/men?
- (2) To what extent would the observed group difference be further reduced if blacks/women had the same endowment than whites/men do, provided there is no difference in returns to characteristics $x_1 \ldots x_K$?

Statistical solution to the problem (Blinder 1973, Oaxaca 1973)

- *Estimation stage*. Estimation of $E(Y_g) = a_g + \mathbf{b}_g \mathbf{x}_g$ for each racial/gender group g (g ={0,1})
- *Post-estimation stage* Calculation of three quantities:
  - $E = (\mathbf{x}_1 - \mathbf{x}_0)\mathbf{b}_1$ <u>e</u>ndowment effect
  - $C = (\mathbf{b}_1 - \mathbf{b}_0)\mathbf{x}_0$ <u>c</u>oefficient effect (“explained discrimination”)
  - $U = a_1 - a_0$ <u>u</u>nexplained part

Introduction II. Motivation

Main motivation

- The postestimation stage of standard decomposition is not valid if nonlinear models are used in the estimation stage; there are some decomposition results for nonlinear models (Fairlie 1999, Yun 2004)
- Several important measures of (dis)advantage are categorical or count variables, like unemployment, number of children, teenage pregnancy, marital status, imprisonment (see the concept of underclass)
- Available user-written programs (decomp, decompose and oaxaca) do not extend decomposition to nonlinear models

Other ambitions

- Graphical interpretation
- Providing detailed decomposition, that is, identifying individual contributions of variables to C and E

  Note: objection to detailed decomposition is the "identification problem" (Oaxaca-Ransom 1999, Gelbach 2002): C and U parts are sensitive to the choice of the reference category of dummies and to changes in the scaling of continuous variables

Extending the Blinder-Oaxaca decomposition to nonlinear models I. The idea

Unpacking the Blinder-Oaxaca solution

The Blinder-Oaxaca decomposition methodology can be viewed as a package of two different ideas

- *Substantive idea*: the valid mathematical representations of the effect of discrimination and the effect of differences in endowments are $(\mathbf{r}_1-\mathbf{r}_0)\mathbf{x}_0$ and $(\mathbf{x}_1-\mathbf{x}_0)\mathbf{r}_1$, where **r** is a vector summarizing returns to the vector of relevant characteristics, **x.**
- *Statistical idea*: $\mathbf{r}=\mathbf{b}$ –coefficients are rates of returns – if linear regression is applied in the estimation stage

Suggested extension to nonlinear models

If nonlinear models were used in the estimation stage, $\mathbf{r}=\mathbf{b}$ obviously does not hold. Solution:

- The substantive idea should be considered to be true, whatever statistical model is used in the estimation stage.
- Although **r#b** after nonlinear models, the substantive idea suggests that $\mathbf{r}=\mathbf{m}$ should hold, where **m** is the vector of marginal effects (or partial changes). *Proof presented on the next pages*

Extending the Blinder-Oaxaca decomposition to nonlinear models II. Proof

Claim: $E(Y_1) - E(Y_0) = (\mathbf{m}_1 - \mathbf{m}_0)\mathbf{x}_0 + (\mathbf{x}_1 - \mathbf{x}_0)\mathbf{m}_1$ **m** is the vector of marginal effects (or partial changes)

Proof

- Starting point is the decomposition presented in Fairlie (1999):

$$\overline{\mathrm{Y}}_1 - \overline{\mathrm{Y}}_0 = \left[\frac{1}{N_1} \sum_{k=1}^{N_1} F(x_1 b_1) - \sum_{k=1}^{N_0} F(x_0 b_1) \right] + \left[\frac{1}{N_0} \sum_{k=1}^{N_0} F(x_0 b_1) - \sum_{k=1}^{N_0} F(x_0 b_0) \right], \tag{1}$$

  where $F(\bullet)$ denotes the cumulative probability function.

- Taylor-series expansion around sample means transforms (1) into

$$\overline{\mathrm{Y}}_1 - \overline{\mathrm{Y}}_0 = \left[F(\bar{x}_1 b_1) - F(\bar{x}_0 b_1) \right] + \left[F(\bar{x}_0 b_1) - F(\bar{x}_0 b_0) \right] + R_1 \tag{2}$$

  where R_1 is the residual reflecting the omission of higher-order terms.

Extending the Blinder-Oaxaca decomposition to nonlinear models II. Proof (continued)

- Following Yun (2004), the two terms in brackets in (2) can be approximated using two first-order Taylor series expansions around $F(\bar{x}_1 b_1)$ and $F(\bar{x}_0 b_1)$. Then (2) can be written as

$$\overline{Y}_1 - \overline{Y}_0 = f(\bar{x}_1 b_1) b_1 (\bar{x}_1 - \bar{x}_0) + f(\bar{x}_0 b_0) \bar{x}_0 (b_1 - b_0) + (R_1 + R_2), \tag{3}$$

where $f(\bullet)$ is the probability density function and R_2 is again a residual term reflecting the omission of higher-order terms.

- Using a first-order Taylor series expansion $f(\bar{x}_0 b_0)$ can be approximated as $f(\bar{x}_1 b_1)$. Thus (3) becomes

$$\overline{Y}_1 - \overline{Y}_0 = f(\bar{x}_1 b_1) b_1 (\bar{x}_1 - \bar{x}_0) + \bar{x}_0 \left[f(\bar{x}_1 b_1) b_1 - f(\bar{x}_0 b_0) b_0 \right] + (R_1 + R_2 + R_3), \tag{4}$$

where R_3 is again a residual term reflecting the omission of higher-order terms.

- Note that the terms $f(\bar{x}_g b_g) b_g$ are marginal effects in group *g*. Equation (4) can compactly be written as

$$\overline{Y}_1 - \overline{Y}_0 \approx m_1 (\bar{x}_1 - \bar{x}_0) + \bar{x}_0 (m_1 - m_0). \tag{5}$$

Estimation of standard errors

Constructing the variance-covariance matrix

- Following Jann (2005), the separate variance-covariance matrices for endowment and coefficient effects are

$$\mathbf{V_E} = \left(\bar{\mathbf{x}}_1^{\mathrm{T}} - \bar{\mathbf{x}}_0^{\mathrm{T}}\right)\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0\right)\mathbf{V_1} \qquad \text{and} \qquad \mathbf{V_{CU}} = \bar{\mathbf{x}}_0^{\mathrm{T}}\bar{\mathbf{x}}_1\left(\mathbf{V_1} - \mathbf{V}_0\right).$$

- The above matrices are accumulated into the $\mathbf{V} = \begin{bmatrix} \mathbf{V_E} & \mathbf{0} \\ \mathbf{0} & \mathbf{V_{CU}} \end{bmatrix}$ matrix, and

- only the diagonal elements of **V** are kept (otherwise **V** is not positive definite).

Assumptions made

- $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_0$ are fixed; their sampling variance is ignored (this can easily be relaxed, see Jann 2005)

- endowment and coefficient effects are independent

The **gdecomp** command I. Syntax

Syntax

gdecomp *groupvar* [, *options*] : *estimation_command*

gdecomp graph *varname* [, *twoway_options*] *This is not documented yet*

where

groupvar specifies a binary (numeric) variable identifying the two groups
(The group with lower/higher $\overline{Y}$ is identified as group 0/1);

estimation_command should begin with a command supported by **margeff**
(Note: the *Y* and *X* variables are in the *varlist* of *estimation_command*);

varname is one of the *varlist* in *estimation_command*; and

options are

dxweight(*high* |*low***) reverse eform level(#) noheader nocoef**

dummies(*varlist_1* [\ *varlist_2* ..]**)**

The **gdecomp** command II. Options

dxweight(*high* |*low***)**

- **dxweight (***high***)** implies that E = $m_1(\bar{x}_1 - \bar{x}_0)$ this is the default
- **dxweight (***low***)** implies that E = $m_0(\bar{x}_1 - \bar{x}_0)$

reverse

- The group with higher (lower) $\overline{\mathrm{Y}}$ is identified as group 0 (1)
- Useful if large values of Y measure outcomes which are negatively valued

eform

- Means that *depvar* is the natural logarithm of the outcome under study
- Marginal effects will be changes in the exponential of linear prediction

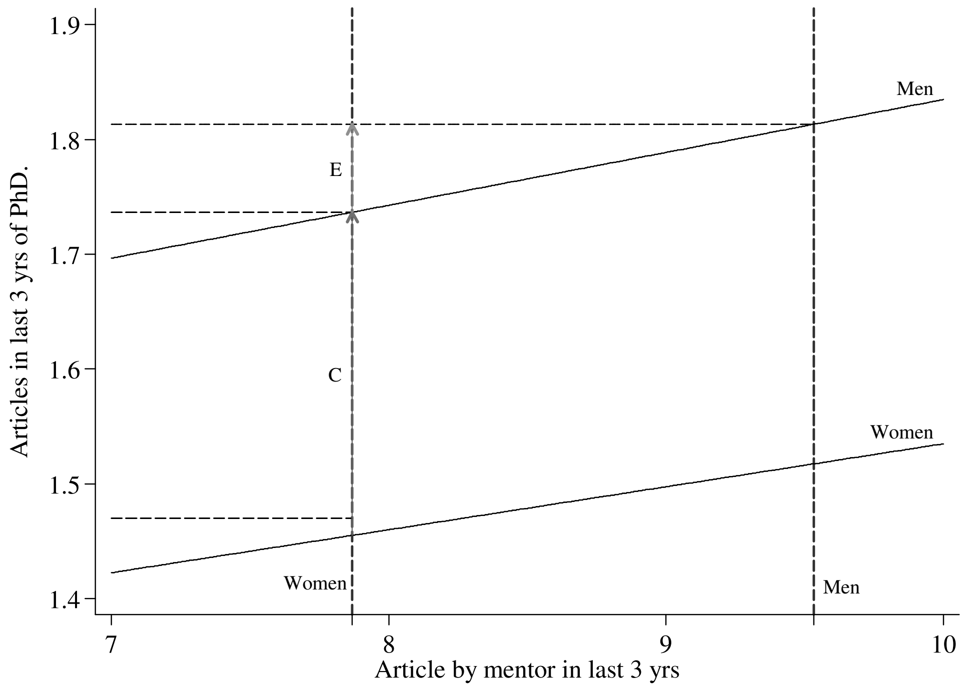
noheader / **nocoef** suppresses the display of overall / detailed decomposition results.

dummies(*varlist_1* [\ *varlist_2* ..]**)** see the help file for **margeff**

The **gdecomp** command III. The (undocumented) **graph** subcommand

This command displays the group-specific partial regression lines and visualizes the C+U and E effects:

```
. gdecomp fem : poisson art ment kidbin
. gdecomp graph ment
```



This example refers to Empirical example III.

Data and variables described on next page.

Legend

C / E = Effect of C+U / endowment effect

What you can see is that

- C here measures “total discrimination”
- Regression lines are parallel, U dominates the C+U component.
- Endowment effect is relatively small

Empirical example I. Data, variables, summary statistics

Data: Scientific Productivity of Biochemistry Phd students, used in Long (1997)

On-line availability: http://www.indiana.edu/~jslsoc/stata/socdata/couart2.dta

Definition and means of variables

Variable	Definition	Men (N=494)	Women (N=421)
fem	Sex: 1=female, 0=male.		
art	Articles in last 3 years of PhD.	1.88	1.47
lnart	Log of art + .5.	0.51	0.36
artbin	1 = 1 or more article in last 3 years of PhD, 0 = otherwise	0.72	0.67
ment	Article by mentor in last 3 years	9.53	7.87
kidbin	At least one child aged <= 5.	0.47	0.19

How to explain the gender difference in scientific productivity?

(Assume for the sake of presentation that the difference is substantial and statistically significant)

Empirical Example II. Decomposition using linear regression: results

```
. gdecomp fem : regress lnart ment kidbin

Decomposition of differences in expected value of lnart after regress
High outcome group: Men -  Low outcome group: Women

Observed difference                             .14900966
Residual difference                             2.776e-17

------------------------------------------------------------------------------
       lnart |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Model        |
           E |  -.0035613   .0220912    -0.16   0.872    -.0468593    .0397368
           C |  -.0267357   .0532107    -0.50   0.615    -.1310269    .0775554
-------------+----------------------------------------------------------------
E            |
        ment |   .0427713   .0054856     7.80   0.000     .0320196    .0535229
      kidbin |  -.0463325   .0213993    -2.17   0.030    -.0882744   -.0043907
-------------+----------------------------------------------------------------
C            |
        ment |  -.0086776   .0476413    -0.18   0.855    -.1020528    .0846975
      kidbin |  -.0180581   .0237001    -0.76   0.446    -.0645094    .0283932
-------------+----------------------------------------------------------------
U            |
       _cons |   .1793067   .0841065     2.13   0.033     .0144609    .3441524
```

Empirical Example II. Decomposition using linear regression: interpretation

Interpretation:

- Overall, neither the E nor the C part is significant.
- Detailed decomposition shows that both ment and kidbin have significant endowment effects. If women had as good mentors (as many kids) than men then women would publish more (less).
- The U part is statistically significant. But the C part is not significant, returns to observed characteristics do not depend on gender
- So, would the scientific productivity of the average woman increase if she were treated in the same way as the average man? The command

```
. lincom [U]_cons+[Model]C
```

reveals that the increase in productivity would be 0.15. This is approximately the observed difference.

Empirical Example III. Decomposition using poisson regression: results

```
. gdecomp fem : poisson art ment kidbin

Decomposition of differences in expected value of art after poisson
High outcome group: Men -  Low outcome group: Women

Observed difference                         .4122823
Residual difference                         .05424126

------------------------------------------------------------------------------
         art |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Model        |
           E |  -.0218253   .0340048    -0.64   0.521    -.0884736    .0448229
           C |    .041618   .0645936     0.64   0.519    -.0849831    .1682191
-------------+----------------------------------------------------------------
E            |
        ment |   .0765719   .0046746    16.38   0.000     .0674098    .0857341
      kidbin |  -.0983973    .033682    -2.92   0.003    -.1644127   -.0323818
-------------+----------------------------------------------------------------
C            |
        ment |   .0678438   .0544729     1.25   0.213    -.0389211    .1746088
      kidbin |  -.0262258   .0347136    -0.76   0.450    -.0942632    .0418116
-------------+----------------------------------------------------------------
U            |
       _cons |   .3382484   .1059164     3.19   0.001      .130656    .5458408
```

Empirical Example III. Decomposition using poisson regression: interpretation

Interpretation:

- About 10 per cent of observed difference is residual. Residual difference reflects the losses during linearization, the term $(R_1 + R_2 + R_3)$ in Eq. (4).
- Again, we find
  - significant endowment effects of ment and kidbin – but no significant overall endowment effect;
  - a significant U part, but a not significant C part
- So, would the scientific productivity of the average woman increase if she were treated in the same way as the average man? Here the answer is yes: the command

```
. lincom [U]_cons+[C]kidbin+[C]ment
```

reveals that the improvement is almost 0.4 articles (p<0.01), which is approximately the observed difference.

Empirical Example IV. Decomposition using logistic regression: results

```
. gdecomp fem  : logit artbin ment kidbin

Decomposition of differences in probability of artbin == 0 after logit
High outcome group: Men -  Low outcome group: Women

Observed difference                        .05486263
Residual difference                        -.00514448

------------------------------------------------------------------------------
      artbin |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Model        |
           E |   .0017739   .0122614     0.14   0.885     -.022258    .0258058
           C |  -.0745998   .0410109    -1.82   0.069    -.1549797    .0057802
-------------+----------------------------------------------------------------
E            |
        ment |   .0212166    .004637     4.58   0.000     .0121282     .030305
      kidbin |  -.0194427   .0113508    -1.71   0.087    -.0416898    .0028044
-------------+----------------------------------------------------------------
C            |
        ment |   -.065841   .0387393    -1.70   0.089    -.1417687    .0100866
      kidbin |  -.0087587   .0134597    -0.65   0.515    -.0351392    .0176218
-------------+----------------------------------------------------------------
U            |
       _cons |   .1328329   .0588945     2.26   0.024     .0174018    .2482641
```

Empirical Example IV. Decomposition using logistic regression: interpretation

Interpretation:

- Again, about 10 per cent of observed difference is residual.
- Again, we find
  - significant endowment effect of ment – but no significant overall endowment effect;
  - a significant U part, but a not significant C part
- So, would the scientific productivity of the average woman increase if she were treated in the same way as the average man? Here the linear combination

```
. lincom [U]_cons+[Model]C
```

  lacks statistical significance.

Discussion

Progress made

- Extending the decomposition methodology for some nonlinear models
- Detailed decomposition results for each variable
  - Warning: C and U parts are sensitive to the choice of the reference category of dummies and to changes in the scaling of continuous variables (this is the "identification problem")
  - But the linear combination of U and C remains "identified" (Gelbach 2002)
  - Detailed decomposition might be useful; in our example, the nonsignificant E part hides significant individual contributions

Still missing

- Variance estimation: relaxing the assumption of fixed sample means
- Graphical interpretation (work under progress)

References

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

Fairlie, R. W. (2003). An extension of the Blinder-Oaxaca decomposition technique to logit and probit models. Yale University, Economic Growth Center, Discussion paper no. 873.

Gelbach, J. (2002). Identified Heterogeneity in Detailed Wage Decomposition. Unpublished paper (On-line: http://glue.umd.edu/~gelbach/papers/comment-or/submitted-12-16-02-comment-or.pdf)

Jann, B. (2005). Standard Errors for the Blinder–Oaxaca Decomposition. Paper presented on the 3rd German Stata Users Group Meeting, Berlin. (On-line: http://repec.org/dsug2005/oaxaca_se_handout.pdf)

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

Oaxaca, R.L., Ransom, M.R. (1999). Identification in Detailed Wage Decompositions. The Review of Economics and Statistics 81: 154-157.

Yun, M-S. 2004. Decomposing differences in the first moment. Economics Letters 82: 275-280.