The Stata Journal (2008)
8, Number 4, pp. 480–492

The Blinder–Oaxaca decomposition for nonlinear regression models

Mathias Sinning
RSSS at the Australian National University, and IZA
Canberra, Australia
mathias.sinning@anu.edu.au

Markus Hahn
Melbourne Institute of Applied Economic and Social Research
The University of Melbourne
Melbourne, Australia
mhahn@unimelb.edu.au

Thomas K. Bauer
RWI Essen, Ruhr-Universität Bochum, and IZA
Bochum, Germany
thomas.bauer@ruhr-uni-bochum.de

Abstract. In this article, a general Blinder–Oaxaca decomposition for nonlinear models is derived, which allows the difference in an outcome variable between two groups to be decomposed into several components. We show how, using `nldecompose`, this general decomposition can be applied to different models with discrete and limited dependent variables. We further demonstrate how the standard errors of the estimated components can be calculated by using Stata's `bootstrap` command as a prefix.

Keywords: st0152, nldecompose, Blinder–Oaxaca decomposition, nonlinear models

1 Introduction

The decomposition method developed by Blinder (1973) and Oaxaca (1973), and generalized by Neumark (1988) and Oaxaca and Ransom (1988, 1994), allows the decomposition of outcome variables between two groups into a part that is explained by differences in observed characteristics and a part attributable to differences in the estimated coefficients. So far, these decomposition methods have mainly been applied in the context of linear regression models. Often the estimation of nonlinear models is required because ordinary least squares (OLS) yields inconsistent parameter estimates and, in turn, misleading decomposition results. Several studies have developed and applied Blinder–Oaxaca decompositions for models with binary dependent variables (Gomulka and Stern 1990; Even and Macpherson 1990; Yun 2004; Fairlie 1999, 2005). An extension of the Blinder–Oaxaca decomposition to nonlinear regression models was

developed by Bauer and Sinning (2008). Following their approach, this article describes how the Blinder–Oaxaca decomposition method can be applied to models with discrete and limited dependent variables by using `nldecompose`.

The `nldecompose` command performs a Blinder–Oaxaca decomposition of the mean outcome differential of linear and nonlinear regression models. `regress`, `logit`, `probit`, `ologit`, `oprobit`, `tobit`, `intreg`, `truncreg`, `poisson`, `nbreg`, `zip`, `zinb`, `ztp`, and `ztnb` are supported. `nldecompose` calculates different variants of the decomposition equation. However, the command does not separate the contributions of single variables. If desired, the `svy:` prefix can be used. Finally, standard errors of the components of the decomposition equation can be estimated by using the `bootstrap` option of `nldecompose`. Other available decomposition packages are `decompose`, `fairlie`, and `oaxaca`, by Ben Jann; `decomp`, by Ian Watson; and `gdecomp`, by Tamás Bartus.

The following section sets out the theoretical framework of the Blinder–Oaxaca decomposition for linear and nonlinear models, taking into account extensions of the original decomposition method that have been applied in the literature. The syntax of `nldecompose` is described in section 3. Section 4 illustrates the application of the decomposition method to different models with discrete and limited dependent variables. Section 5 summarizes and concludes the article.

2 Framework

Consider the following linear regression model, which is fitted separately for the groups $g = (A, B)$:

$$Y_{ig} = \mathbf{X}_{ig}\beta_g + \varepsilon_{ig}$$

where $i = 1, \ldots, N_g$ and $\sum_g N_g = N$. For these models, Blinder (1973) and Oaxaca (1973) propose the decomposition

$$\overline{Y}_A - \overline{Y}_B = \Delta^{\text{OLS}} = (\overline{\mathbf{X}}_A - \overline{\mathbf{X}}_B)\widehat{\beta}_A + \overline{\mathbf{X}}_B(\widehat{\beta}_A - \widehat{\beta}_B) \tag{1}$$

where $\overline{Y}_g = N_g^{-1}\sum_{i=1}^{N_g} Y_{ig}$ and $\overline{\mathbf{X}}_g = N_g^{-1}\sum_{i=1}^{N_g} \mathbf{X}_{ig}$. The first term on the right-hand side of (1) displays the difference in the outcome variable between the two groups that is due to differences in observable characteristics, whereas the second term shows the differential that is due to differences in coefficient estimates. Ben Jann provides a detailed discussion of the estimation of the Blinder–Oaxaca decomposition for linear regression models in this issue of the *Stata Journal*.

A decomposition of the outcome variable similar to (1) is not appropriate in the nonlinear (NL) case, because the conditional expectations, $E(Y_{ig} \,|\, \mathbf{X}_{ig})$, may differ from $\overline{\mathbf{X}}_g\widehat{\beta}_g$. Therefore, we rewrite (1) in terms of conditional expectations to obtain a general version of the Blinder–Oaxaca decomposition:

$$\Delta_A^{\text{NL}} = \{E_{\beta_A}(Y_{iA} \,|\, \mathbf{X}_{iA}) - E_{\beta_A}(Y_{iB} \,|\, \mathbf{X}_{iB})\} + \{E_{\beta_A}(Y_{iB} \,|\, \mathbf{X}_{iB}) - E_{\beta_B}(Y_{iB} \,|\, \mathbf{X}_{iB})\} \tag{2}$$

where $E_{\beta_g}(Y_{ig} \,|\, \mathbf{X}_{ig})$ refers to the conditional expectation of Y_{ig}, and $E_{\beta_g}(Y_{ih} \,|\, \mathbf{X}_{ih})$ refers to the conditional expectation of Y_{ih} evaluated at the parameter vector β_g, with

$g, h = (A, B)$ and $g \neq h$. Changing the reference group, an alternative expression for the decomposition is

$$\Delta_B^{\text{NL}} = \{E_{\beta_B}(Y_{iA} \,|\, \mathbf{X}_{iA}) - E_{\beta_B}(Y_{iB} \,|\, \mathbf{X}_{iB})\} + \{E_{\beta_A}(Y_{iA} \,|\, \mathbf{X}_{iA}) - E_{\beta_B}(Y_{iA} \,|\, \mathbf{X}_{iA})\} \quad (3)$$

Again the first term on the right-hand side displays the part of the differential in the outcome variable between the two groups that is due to differences in the covariates $\mathbf{X}_{ig}$, and the second term displays the part of the differential in Y_{ig} that is due to differences in coefficients.

Oaxaca and Ransom (1994) give an overview of the application of the following generalized linear decomposition:

$$\overline{Y}_A - \overline{Y}_B = (\overline{\mathbf{X}}_A - \overline{\mathbf{X}}_B)\beta^* + \overline{\mathbf{X}}_A(\beta_A - \beta^*) + \overline{\mathbf{X}}_B(\beta^* - \beta_B) \quad (4)$$

In (4), β^* is defined as a weighted average of the coefficient vectors, β_A and β_B:

$$\beta^* = \mathbf{\Omega}\beta_A + (\mathbf{I} - \mathbf{\Omega})\beta_B$$

where $\mathbf{\Omega}$ is a weighting matrix and $\mathbf{I}$ is an identity matrix. Decompositions (2) and (3), which were proposed by Blinder (1973) and Oaxaca (1973), represent special cases of the generalized equation in which $\mathbf{\Omega}$ is a null matrix or is equal to $\mathbf{I}$, respectively. The generalized version of (4) can be written as

$$\begin{aligned} \overline{Y}_A - \overline{Y}_B &= \{E_{\beta^*}(Y_{iA} \,|\, \mathbf{X}_{iA}) - E_{\beta^*}(Y_{iB} \,|\, \mathbf{X}_{iB})\} \\ &+ \{E_{\beta_A}(Y_{iA} \,|\, \mathbf{X}_{iA}) - E_{\beta^*}(Y_{iA} \,|\, \mathbf{X}_{iA})\} \\ &+ \{E_{\beta^*}(Y_{iB} \,|\, \mathbf{X}_{iB}) - E_{\beta_B}(Y_{iB} \,|\, \mathbf{X}_{iB})\} \end{aligned}$$

Different assumptions about the form of $\mathbf{\Omega}$ can be considered. Reimers (1983) and Cotton (1988) treat $\mathbf{\Omega}$ as a scalar matrix. Reimers (1983) proposes the weighting matrix $\mathbf{\Omega} = (0.5)\mathbf{I}$, while Cotton (1988) chooses the weighting matrix $\mathbf{\Omega} = s\mathbf{I}$, where s denotes the relative sample size of the majority group. In the context of a linear regression model, Neumark (1988) and Oaxaca and Ransom (1994) propose to estimate a pooled model to derive the counterfactual coefficient vector, β^*.

To apply (4) to different nonlinear models, one needs to derive the sample counterparts—$S(\widehat{\beta}_g, \mathbf{X}_{ig})$, $S(\widehat{\beta}_h, \mathbf{X}_{ig})$, and $S(\widehat{\beta}^*, \mathbf{X}_{ig})$—of the conditional expectations—$E_{\beta_g}(Y_{ig} \,|\, \mathbf{X}_{ig})$, $E_{\beta_h}(Y_{ig} \,|\, \mathbf{X}_{ig})$, and $E_{\beta^*}(Y_{ig} \,|\, \mathbf{X}_{ig})$—for $g, h = (A, B)$ and $g \neq h$.[1] The decomposition of nonlinear models shares all problems of the original Blinder–Oaxaca decomposition, such as, e.g., a potential sensitivity of the results with respect to the choice of the reference group and the specification of the regression model.

Finally, Daymont and Andrisani (1984) proposed the following extension of the Blinder–Oaxaca decomposition:

1. For example, the sample counterpart $S(\widehat{\beta}_g, \mathbf{X}_{ig})$ in a probit model is given by $N_g^{-1} \sum_{i=1}^{N_g} \Phi(\widehat{\beta}_g, \mathbf{X}_{ig})$, where $\Phi(\cdot)$ is the cumulative normal density function. Bauer and Sinning (2008) provide an overview of sample counterparts in nonlinear models.

$$\overline{Y}_A - \overline{Y}_B = (\overline{\mathbf{X}}_A - \overline{\mathbf{X}}_B)\beta_B + \overline{\mathbf{X}}_B(\beta_A - \beta_B) + (\overline{\mathbf{X}}_A - \overline{\mathbf{X}}_B)(\beta_A - \beta_B) = E + C + CE \quad (5)$$

where E is the part of the raw differential that is due to differences in endowments, C reflects the part attributable to differences in coefficients, and CE represents the part that can be explained by the interaction between C and E. In the general version of the decomposition, these components are given by the following:

$$E = \{E_{\beta_B}(Y_{iA} \mid \mathbf{X}_{iA}) - E_{\beta_B}(Y_{iB} \mid \mathbf{X}_{iB})\}$$
$$C = \{E_{\beta_A}(Y_{iB} \mid \mathbf{X}_{iB}) - E_{\beta_B}(Y_{iB} \mid \mathbf{X}_{iB})\}$$
$$CE = \{E_{\beta_A}(Y_{iA} \mid \mathbf{X}_{iA}) - E_{\beta_B}(Y_{iA} \mid \mathbf{X}_{iA})\} + \{E_{\beta_A}(Y_{iB} \mid \mathbf{X}_{iB}) - E_{\beta_B}(Y_{iB} \mid \mathbf{X}_{iB})\}$$

Similarly to (4), the single components of (5) can be estimated by using the sample counterparts $S(\widehat{\beta}_g \mid \mathbf{X}_{ig})$ and $S(\widehat{\beta}_h \mid \mathbf{X}_{ig})$ of the conditional expectations $E_{\beta_g}(Y_{ig} \mid \mathbf{X}_{ig})$ and $E_{\beta_h}(Y_{ig} \mid \mathbf{X}_{ig})$ for $g, h = (A, B)$ and $g \neq h$.

Table 1 presents examples for the sample counterpart $S(\widehat{\beta}_g, \mathbf{X}_{ig})$ of a group $g = (A, B)$. Bauer and Sinning (2008) provide a more detailed description of these sample counterparts and also discuss the Blinder–Oaxaca decomposition and the corresponding sample counterparts of count-data models and limited dependent variable models.

(*Continued on next page*)

Table 1. Examples for the sample counterpart

Command	Sample counterpart
`regress`	$\frac{1}{N_g}\sum_{i=1}^{N}\mathbf{X}_{ig}\widehat{\beta}_g$
`logit`	$\frac{1}{N_g}\sum_{i=1}^{N}\Lambda(\mathbf{X}_{ig}\widehat{\beta}_g)$, where Λ is the cumulative logistic density function
`probit`	$\frac{1}{N_g}\sum_{i=1}^{N}\Phi(\mathbf{X}_{ig}\widehat{\beta}_g)$, where Φ is the cumulative normal density function
`ologit`	$\frac{1}{N_g}\sum_{i=1}^{N}[\{\Lambda(\widehat{\mu}_1-\mathbf{X}_{ig}\widehat{\beta}_g)-\Lambda(-\mathbf{X}_{ig}\widehat{\beta}_g)\}$ $+2\{\Lambda(\widehat{\mu}_2-\mathbf{X}_{ig}\widehat{\beta}_g)-\Lambda(\widehat{\mu}_1-\mathbf{X}_{ig}\widehat{\beta}_g)\}+\cdots+$ $J\{1-\Lambda(\widehat{\mu}_{J-1}-\mathbf{X}_{ig}\widehat{\beta}_g)\}]$, where J is the number of possible outcomes and $\widehat{\mu}_1,\ldots,\widehat{\mu}_{J-1}$ are the estimated threshold values of `ologit`
`oprobit`	$\frac{1}{N_g}\sum_{i=1}^{N}[\{\Phi(\widehat{\theta}_1-\mathbf{X}_{ig}\widehat{\beta}_g)-\Phi(-\mathbf{X}_{ig}\widehat{\beta}_g)\}$ $+2\{\Phi(\widehat{\theta}_2-\mathbf{X}_{ig}\widehat{\beta}_g)-\Phi(\widehat{\theta}_1-\mathbf{X}_{ig}\widehat{\beta}_g)\}+\cdots+$ $J\{1-\Phi(\widehat{\theta}_{J-1}-\mathbf{X}_{ig}\widehat{\beta}_g)\}]$, where J is the number of possible outcomes, and $\widehat{\theta}_1,\ldots,\widehat{\theta}_{J-1}$ are the estimated threshold values of `oprobit`

3 The syntax of nldecompose

`nldecompose` is a prefix command, which means that it stands in front of the relevant regression command. `nldecompose` requires Stata 10 or higher. A simplified syntax reads as follows:

`nldecompose, by(`*varname*`)` [*options*]`:` *regcmd*

where `by(`*varname*`)` specifies the groups for which the difference in the outcome variable should be analyzed. *varname* should be defined as an indicator variable that takes on a value of 1 for the group with the higher outcome (group A) and a value of 0 for the group with the lower outcome (group B). `by(`*varname*`)` is required. *regcmd* is the command of the regression model to be decomposed. If desired, the `svy:` prefix can be used. `nldecompose` supports the following Stata commands: `regress`, `logit`, `probit`, `ologit`, `oprobit`, `tobit`, `intreg`, `truncreg`, `poisson`, `nbreg`, `zip`, `zinb`, `ztp`, and `ztnb`.

3.1 Syntax

`nldecompose, by(`*varname*`)` [`threefold omega(`*#* [`, `*#*, ...`]` | *omega_option*`)`
`regoutput bootstrap bs bsoptions(`*bootstrap_options*`) ll(`*#* | *varname*`)`
`ul(`*#* | *varname*`)`]`:` *regcmd*

3.2 Options

`by(`*varname*`)` specifies the sample of the high group (A) and the low group (B). *varname* should be defined as a dummy variable that takes on a value of 1 for the group with the higher outcome and a value of 0 for the group with the lower outcome. `by()` is required.

`threefold` performs a decomposition into three components (as described in Daymont and Andrisani [1984]).

`omega(`*#* [`, `*#*, ...`]` | *omega_option*`)` represents the general weighting matrix as specified by Oaxaca and Ransom (1994). `omega()` can either contain a scalar weight or a vector including the weights $w_1, \ldots, w_k$ on the diagonal of the weighting matrix, where k corresponds to the number of coefficients in the model. *omega_option* can be `reimers`, `cotton`, or `neumark`, referring to the corresponding weighting schemes proposed by Reimers (1983), Cotton (1988), and Neumark (1988).

`regoutput` displays the respective outputs of the two regressions.

`bootstrap` calculates bootstrap standard errors and confidence intervals.

`bs` is an alias for `bootstrap`.

`bsoptions(`*bootstrap_options*`)` specifies options of the internal `bootstrap` routine. `bsoptions()` shares all the options of Stata's `bootstrap` command (see `help bootstrap`). It can be used only in combination with the internal `bootstrap` option of `nldecompose`.

`ll(`*#* | *varname*`)` specifies the lower limit of the outcome variable. This option can include either a scalar or a variable. `ll()` can be used only with `intreg`. `ll()` or `ul()` (or both) is required if `intreg` is used.

`ul(`*#* | *varname*`)` specifies the upper limit of the outcome variable. This option can include either a scalar or a variable. `ul()` can be used only with `intreg`. `ul()` or `ll()` (or both) is required if `intreg` is used.

(*Continued on next page*)

3.3 Saved results

`nldecompose` saves the results into `r()`. Some results are only available in combination with certain options or regression commands. "AB" indicates a result based on the coefficients of the high group (A) and the characteristics of the low group (B); "BA" indicates a result based on the coefficients of the low group (B) and the characteristics of the high group (A).

Scalars

`r(raw)`	raw differential
`r(charAB)`	part of raw differential attributable to differences in characteristics (AB)
`r(coefAB)`	part of raw differential attributable to differences in coefficients (AB)
`r(intAB)`	part of raw differential attributable to interaction between `charAB` and `coefAB` (if `threefold` is specified)
`r(pcharAB)`	`charAB/raw`
`r(pcoefAB)`	`coefAB/raw`
`r(pintAB)`	`intAB/raw` (if `threefold` is specified)
`r(charBA)`	part of raw differential attributable to differences in characteristics (BA)
`r(coefBA)`	part of raw differential attributable to differences in coefficients (BA)
`r(intBA)`	part of raw differential attributable to interaction between `charBA` and `coefBA` (if `threefold` is specified)
`r(pcharBA)`	`charBA/raw`
`r(pcoefBA)`	`coefBA/raw`
`r(pintBA)`	`intBA/raw` (if `threefold` is specified)
`r(prod)`	part of raw differential attributable to productivity (if `omega()` is specified)
`r(adv)`	part of raw differential attributable to advantage of the high group (if `omega()` is specified)
`r(disadv)`	part of raw differential attributable to disadvantage of the low group (if `omega()` is specified)
`r(pprod)`	`prod/raw` (if `omega()` is specified)
`r(padv)`	`adv/raw` (if `omega()` is specified)
`r(pdisadv)`	`disadv/raw` (if `omega()` is specified)
`r(obsA)`	number of observations for group A
`r(obsB)`	number of observations for group B
`r(N_reps)`	number of bootstrap replications (if `bootstrap` is specified)
`r(level)`	confidence level in percent (if `bootstrap` is specified)
`r(wgt)`	singular weight (if `omega()` is specified as a number or scalar)

Macros

`r(regcmd)`	regression command

Matrices

`r(bootstrap)`	matrix with coefficients, standard errors, z-values, p-values, and confidence intervals (if `bootstrap` is specified)
`r(wgt)`	weighting (omega) matrix (if `omega()` is specified as a vector)

4 Examples

In this section, we use a test dataset called `nldecompose.dta` to illustrate the use of the `nldecompose` command. The data file is available together with `nldecompose.ado`. If the test dataset is located in the current directory, it can be opened with

```
. use nldecompose
(Test data for nldecompose.ado)
```

4.1 Basic syntax

Suppose that we want to decompose the outcome differential between two groups by using a linear regression model. `nldecompose` requires a variable (group indicator) indicating membership to the group with the higher outcome (group indicator = 1) or the group with the lower outcome (group indicator = 0). The group indicator in `nldecompose.dta` is named *group*. `nldecompose` further requires the specification of a regression model to be decomposed. The test data include several outcome variables as well as a set of covariates to be used for the respective regression models. To obtain estimates of the Blinder–Oaxaca decomposition in the linear case, we type

```
. nldecompose, by(group): regress y_regress x1 x2 x3
                                                Number of obs (A) =     773
                                                Number of obs (B) =     971
```

Results	Coef.	Percentage
Omega = 1		
Char	.0118193	1.282994%
Coef	.909411	98.71701%
Omega = 0		
Char	.0385519	4.184829%
Coef	.8826784	95.81517%
Raw	.9212303	100%

The decomposition results given in the above output table denote the absolute values of the components of the decomposition equation (i.e., differences in characteristics, Char; differences in coefficients, Coef; and the raw differential, Raw) and the corresponding percentages of the raw differential. `nldecompose` reports the Blinder–Oaxaca decomposition estimates for Omega = 1 and Omega = 0, which refer to different specifications of the omega matrix (see section 2). These two weighting schemes are given by default in every output table. An additional scheme can be specified by using the `omega()` option. This option can either include a numeric value (which is transformed into a scalar matrix), a matrix, or an option name. For instance, to apply the weighting scheme proposed by Neumark (1988), we type

(*Continued on next page*)

```
. nldecompose, by(group) omega(neumark): logit y_logit x1 x2 x3
                                                Number of obs (A) =     773
                                                Number of obs (B) =     971
```

Results	Coef.	Percentage
Omega = 1		
Char	-.0130632	-5.584561%
Coef	.2469787	105.5846%
Omega = 0		
Char	.0482973	20.64731%
Coef	.1856182	79.35269%
Omega = wgt		
Prod	.0084567	3.615265%
Adv	.1255278	53.66375%
Disadv	.099931	42.72099%
Raw	.2339155	100%

A third part is now added to the output table, including the components of the decomposition proposed by Oaxaca and Ransom (1994): differences due to a different productivity, Prod; advantage of the high group, Adv; and disadvantage of the low group, Disadv. The part of the differential attributable to different characteristics is negative for Omega = 1 in the above example, indicating that observable characteristics contribute to reducing the gap between the two groups. This is typically the case if the group with the lower outcome possesses a relative advantage in one or more observable characteristics. Finally, the `threefold` option of `nldecompose` allows a decomposition of the mean differential of the outcome variables into three components (Daymont and Andrisani 1984). The third component reported in the following output table denotes the interaction term, Int, presented in (5):

```
. nldecompose, by(group) threefold: tobit y_tobit x1 x2 x3, ll(0)
                                                Number of obs (A) =     773
                                                Number of obs (B) =     971
```

Results	Coef.	Percentage
Omega = 1		
Char	.469702	92.31832%
Coef	.0705736	13.871%
Int	-.0314904	-6.189324%
Omega = 0		
Char	.4382117	86.129%
Coef	.0390832	7.681677%
Int	.0314904	6.189324%
Raw	.5087853	100%

The lower and upper limits, specified as an option of the `tobit` command, are retrieved automatically by `nldecompose`. However, because the `intreg` command also permits the use of variables that include lower limits, upper limits, or both for each observation (see `help intreg`), these limits must be specified as additional options if `intreg` is used:

```
. nldecompose_old, by(group) ll(lowerlimit) ul(1000): intreg y_intreg1
> y_intreg 2 x1 x2 x3
                                                Number of obs (A) =     773
                                                Number of obs (B) =     971
```

Results	Coef.	Percentage
Omega = 1		
Char	.0459277	4.990594%
Coef	.8743578	95.00941%
Omega = 0		
Char	.1678028	18.23377%
Coef	.7524828	81.76623%
Raw	.9202855	100%

4.2 Bootstrap

Although `nldecompose` does not provide analytic standard errors, the bootstrap method can be applied to derive standard errors of the components of the decomposition equation. In addition to Stata's `bootstrap` command, the `bootstrap` option of `nldecompose` can be used. The following table includes the estimates of the Blinder–Oaxaca decomposition for a Poisson model and their standard errors, z-values, p-values, and confidence intervals.

```
. nldecompose, by(group) bootstrap bsoptions(reps(50) seed(553721)): poisson
> y_poisson x1 x2 x3
                                                Number of obs (A) =     773
                                                Number of obs (B) =     971
                                                BS Replications   =      50
```

Results	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Omega = 1						
Char	-.0019096	.0140652	-0.14	0.892	-.0294769	.0256577
Coef	.0130316	.1915354	0.07	0.946	-.3623708	.3884341
Omega = 0						
Char	.0489932	.0547485	0.89	0.371	-.058312	.1562983
Coef	-.0378711	.2042228	-0.19	0.853	-.4381404	.3623981
Raw	.011122	.1886442	0.06	0.953	-.3586138	.3808579

Although `nldecompose` has the built-in support of `bootstrap`, it is possible to bootstrap the results of `nldecompose` by using Stata's `bootstrap` command as a prefix. Here an alternative syntax is required:

`bootstrap` [*options*]`: nldecompose, by(`*varname*`)` [*options*] `regcmd(`*regcmd*`)`

This syntax can be used to obtain the same estimates as above:

```
. bootstrap charAB=r(charAB) coefAB=r(coefAB) charBA=r(charBA) coefBA=r(coefBA)
> raw=r(raw), reps(50) seed(553721) nowarn: nldecompose, by(group)
> regcmd(poisson y_poisson x1 x2 x3)
(running nldecompose on estimation sample)

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

Bootstrap results                               Number of obs      =      1744
                                                Replications       =        50

      command:  nldecompose, by(group) regcmd(poisson y_poisson x1 x2 x3)
       charAB:  r(charAB)
       coefAB:  r(coefAB)
       charBA:  r(charBA)
       coefBA:  r(coefBA)
          raw:  r(raw)
```

	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
charAB	-.0019096	.0140652	-0.14	0.892	-.0294769	.0256577
coefAB	.0130316	.1915354	0.07	0.946	-.3623708	.3884341
charBA	.0489932	.0547485	0.89	0.371	-.058312	.1562983
coefBA	-.0378711	.2042228	-0.19	0.853	-.4381404	.3623981
raw	.011122	.1886442	0.06	0.953	-.3586138	.3808579

5 Summary

In this article, a general Blinder–Oaxaca decomposition for linear and nonlinear models was derived, which allows the differences in an outcome variable between two groups to be decomposed into several components. We showed how this general decomposition can be applied to different models with discrete and limited dependent variables by using `nldecompose`. We further demonstrated how the standard errors of the estimated components can be calculated by using the `bootstrap` command. Because the article focuses on overall decompositions without separating the contributions of single variables, future work will concentrate on detailed decompositions.

The Blinder–Oaxaca decomposition for the tobit model has been applied empirically by Bauer and Sinning (Forthcoming-a, -b). Bauer, Göhlmann, and Sinning (2007) provide an application of the Blinder–Oaxaca decomposition to count-data models.

6 Acknowledgments

The authors are grateful to two anonymous reviewers for their helpful comments.

7 References

Bauer, T., S. Göhlmann, and M. Sinning. 2007. Gender differences in smoking behavior. *Health Economics* 16: 895–909.

Bauer, T. K., and M. Sinning. 2008. An extension of the Blinder–Oaxaca decomposition to nonlinear models. *Advances in Statistical Analysis* 92: 197–206.

———. Forthcoming-a. Blinder–Oaxaca decomposition for tobit models. *Applied Economics*.

———. Forthcoming-b. The savings behavior of temporary and permanent migrants in Germany. *Journal of Population Economics*.

Blinder, A. S. 1973. Wage discimination: Reduced form and structural estimates. *Journal of Human Resources* 8: 436–455.

Cotton, J. 1988. On the decomposition of wage differentials. *Review of Economics and Statistics* 70: 236–243.

Daymont, T. N., and P. J. Andrisani. 1984. Job preferences, college major, and the gender gap in earnings. *Journal of Human Resources* 19: 408–428.

Even, W. E., and D. A. Macpherson. 1990. Plant size and the decline of unionism. *Economics Letters* 32: 393–398.

Fairlie, R. W. 1999. The absence of the African-American owned business: An analysis of the dynamics of self-employment. *Journal of Labor Economics* 17: 80–108.

———. 2005. An extension of the Blinder–Oaxaca decomposition technique to logit and probit models. *Journal of Economic and Social Measurement* 30: 305–316.

Gomulka, J., and N. Stern. 1990. The employment of married women in the United Kingdom, 1970–1983. *Econometrica* 57: 171–199.

Jann, B. 2008. The Blinder–Oaxaca decomposition for linear regression models. *Stata Journal* 8: 453–479.

Neumark, D. 1988. Employers' discriminatory behavior and the estimation of wage discrimination. *Journal of Human Resources* 23: 279–295.

Oaxaca, R. 1973. Male–female wage differentials in urban labor markets. *International Economic Review* 14: 693–709.

Oaxaca, R. L., and M. R. Ransom. 1988. Searching for the effect of unionism on the wages of union and nonunion workers. *Journal of Labor Research* 9: 139–148.

———. 1994. On discrimination and the decomposition of wage differentials. *Journal of Econometrics* 61: 5–21.

Reimers, C. W. 1983. Labor market discrimination against Hispanic and black men. *Review of Economics and Statistics* 65: 570–579.

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

About the authors

Mathias Sinning is a research fellow at the Social Policy Evaluation, Analysis, and Research Centre of the Australian National University.

Markus Hahn is a research officer at the Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Thomas K. Bauer is a member of the executive board of RWI Essen and a full professor of economics at the Ruhr-Universität Bochum.