RWI : Discussion Papers

Thomas K. Bauer and Mathias Sinning

An Extension of the Blinder-Oaxaca Decomposition to Non-Linear Models

No. 49

RWI
ESSEN

RWI : Discussion Papers
No. 49

Published by Rheinisch-Westfälisches Institut für Wirtschaftsforschung, Hohenzollernstrasse 1/3, D-45128 Essen, Phone +49 (0) 201/81 49-0

Editor: Prof. Dr. Christoph M. Schmidt, Ph.D.
ISSN 1612-3565 – ISBN 3-936454-76-0
ISBN-13 978-3-936454-76-5

RWI : Discussion Papers
No. 49

Thomas K. Bauer and Mathias Sinning

An Extension of the Blinder-Oaxaca Decomposition to Non-Linear Models

RWI
ESSEN

Bibliografische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

ISSN 1612-3565
ISBN 3-936454-76-0
ISBN-13 978-3-936454-76-5

Thomas K. Bauer and Mathias Sinning*

An Extension of the Blinder-Oaxaca Decomposition to Non-Linear Models

Abstract
In this paper, a general Blinder-Oaxaca decomposition is derived that can also be applied to non-linear models, which allows the differences in a non-linear outcome variable between two groups to be decomposed into a part that is explained by differences in observed characteristics and a part attributable to differences in the estimated coeffcients. Departing from this general model, we show how it can be applied to different models with discrete and limited dependent variables.

JEL Classification: C13, C20

Keywords: Blinder-Oaxaca decomposition, non-linear models

October 2006

*Thomas K. Bauer, IZA-Bonn and CEPR London; Mathias Sinning, RWI Essen. – All correspondence to Mathias Sinning, Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI Essen), Hohenzollernstr. 1–3, 45128 Essen, Germany, Tel: +49 201 / 81 49-214, Fax: +49 201 81 49-200, Email: sinning@rwi-essen.de.

1 Introduction

The decomposition method developed by Blinder (1973) and Oaxaca (1973) and generalized by Juhn, Murphy, and Pierce (1991), Neumark (1988) and Oaxaca and Ransom (1988, 1994) is a very popular descriptive tool in empirical economics, since it 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 returns to these characteristics. So far, these decomposition methods have mainly been applied in the context of linear regression models. In many cases, however, the outcome variable is non-linear, requiring the estimation of non-linear models because OLS yields inconsistent parameter estimates and in turn misleading decomposition results. In particular, since the parameter estimates of non-linear models typically differ from the marginal effects of the latent outcome variable, they cannot be used to perform a standard Blinder-Oaxaca decomposition.

A decomposition method for models with binary dependent variables has been developed by Fairlie (1999, 2003). In this paper, we generalize the Blinder-Oaxaca decomposition method to other non-linear models. Based on this generalized decomposition, we than demonstrate how the Blinder-Oaxaca decomposition can be applied to models with discrete and limited dependent variables.

2 An Extension of the Blinder-Oaxaca Decomposition to Non-linear Models

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

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

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

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

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 equation (1) displays the difference in the outcome variable between

the two groups due to differences in observable characteristics, whereas the second term shows the differential that is due to differences in coefficient estimates.

A decomposition of the outcome variable similar to equation (1) is not appropriate in the non-linear (NL) case, because the conditional expectations $E(Y_{ig}|\mathbf{X}_{ig})$ may differ from $\overline{\mathbf{X}}_g\widehat{\beta}_g$. For that reason, the decomposition of the mean difference of Y_i between the two groups has to be considered:

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

where $E_{\beta_g}(Y_{ig}|\mathbf{X}_{ig})$ refers to the conditional expectation of Y_{ig} and $E_{\beta_g}(Y_{ih}|\mathbf{X}_{ih})$ 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

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

In both equations, the first term on the right hand side again 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 the part of the differential in Y_{ig} that is due to differences in coefficients. To apply this decomposition to different non-linear models, one just has to derive the respective sample counterparts $S(\hat{\beta}_g, \mathbf{X}_{ig})$ and $S(\hat{\beta}_h, \mathbf{X}_{ig})$ of the conditional expectations $E_{\beta_g}(Y_{ig}|\mathbf{X}_{ig})$ and $E_{\beta_h}(Y_{ig}|\mathbf{X}_{ig})$ for $g, h = (A, B)$ and $g \neq h$. The following section illustrates the application of equation (2) for different models with discrete and limited dependent variables. An estimation of the corresponding components of equation (3) is straightforward. Note that this decomposition 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.

3 Discrete Dependent Variable Models

3.1 Logit and Probit Models

Discrete dependent variable models comprise binary and ordered Logit and Probit models as well as models for count data. Because binary Logit and Probit models

may be considered as a special case of ordered Logit and Probit models, the decomposition method for binary dependent variables proposed by Fairlie (1999, 2003) represents a special case of the Blinder-Oaxaca decomposition for ordered choice models. Ordered Logit and Probit models (O) are frequently used as a framework for analyzing outcomes of opinion surveys. These models are based on a latent regression of the form

$$Y_{ig}^* = \mathbf{X}_{ig}\beta_g^O + \varepsilon_{ig}^O,$$

where Y_{ig}^* is unobserved. Instead of Y_{ig}^*, only the following realizations are observed:

$$\begin{aligned} Y_{ig} &= 0 \text{ if } Y_{ig}^* \leq 0, \\ &= 1 \text{ if } 0 < Y_{ig}^* \leq \mu_1, \\ &= 2 \text{ if } \mu_1 < Y_{ig}^* \leq \mu_2, \\ &\dots \\ &= J \text{ if } \mu_{J-1} \leq Y_{ig}^*, \end{aligned}$$

where the μ's are unknown parameters to be estimated together with the coefficients β_g^O. The conditional expectation of Y_{ig} evaluated at the parameter vector β_g^O can be written as

$$\begin{aligned} E_{\beta_g^O}(Y_{ig}|\mathbf{X}_{ig}) &= F(\mu_1 - \mathbf{X}_{ig}\beta_g^O) - F(-\mathbf{X}_{ig}\beta_g^O) \\ &+ 2[F(\mu_2 - \mathbf{X}_{ig}\beta_g^O) - F(\mu_1 - \mathbf{X}_{ig}\beta_g^O)] \\ &+ ... \\ &+ J[1 - F(\mu_{J-1} - \mathbf{X}_{ig}\beta_g^O)]. \end{aligned}$$

Assuming that the error term ε_{ig}^O is normally distributed across observations leads to the ordered Probit model, where $F(\cdot)$ is defined as the cumulative standard normal distribution $\Phi(\cdot)$. The Logit model is obtained when the error term ε_{ig}^O is assumed to follow a logistic distribution, i.e. when $F(\cdot)$ represents a cumulative logistic distribution $\Lambda(\cdot)$.

Given the estimates of the parameter vector β_g^O, the sample counterparts of the

single components of the decomposition equation can be calculated by

$$\begin{aligned} S(\hat{\beta}_g^O, \mathbf{X}_{ig}) &= N^{-1} \sum_{i=1}^{N} \Big\{ [F(\hat{\mu}_1 - \mathbf{X}_{ig}\hat{\beta}_g^O) - F(-\mathbf{X}_{ig}\hat{\beta}_g^O)] \\ &+ 2[F(\hat{\mu}_2 - \mathbf{X}_{ig}\hat{\beta}_g^O) - F(\hat{\mu}_1 - \mathbf{X}_{ig}\hat{\beta}_g^O)] \\ &+ \dots \\ &+ J[1 - F(\hat{\mu}_{J-1} - \mathbf{X}_{ig}\hat{\beta}_g^O)] \Big\}. \end{aligned}$$

The sample counterpart of $E_{\beta_h^O}(Y_{ig}|\mathbf{X}_{ig})$, $S(\hat{\beta}_h^O, \mathbf{X}_{ig})$, is obtained by just replacing $\hat{\beta}_g^O$ with $\hat{\beta}_h^O$ in the above equation.[1] These sample counterparts can then be used to calculate the single parts of the decomposition equation (2) as

$$\begin{aligned} \hat{\Delta}^O &= \Big[S(\hat{\beta}_A^O, \mathbf{X}_{iA}) - S(\hat{\beta}_A^O, \mathbf{X}_{iB}) \Big] \\ &\quad + \Big[S(\hat{\beta}_A^O, \mathbf{X}_{iB}) - S(\hat{\beta}_B^O, \mathbf{X}_{iB}) \Big]. \end{aligned}$$

The Blinder-Oaxaca decomposition for ordered choice models reduces to the decomposition method for binary choice models if $J = 1$.

3.2 Count Data Models

The Poisson regression model (P), which has been widely used to study count data, assumes that the dependent variable Y_{ig} conditional on the covariates $\mathbf{X}_{ig}$ is Poisson distributed with density

$$f(Y_{ig}|\mathbf{X}_{ig}) = \frac{\exp(-\mu_{ig})\mu_{ig}^{Y_{ig}}}{Y_{ig}!}, \quad Y_{ig} = 0, 1, 2, ...$$

and conditional expectation

$$E(Y_{ig}|\mathbf{X}_{ig}) = \mu_{ig} = \exp(\mathbf{X}_{ig}\beta_g^P).$$

The sample counterpart of $E_{\beta_g^P}(Y_{ig}|\mathbf{X}_{ig})$ which is necessary to estimate the decomposition equation is given by

$$S(\hat{\beta}_g^P, \mathbf{X}_{ig}) = \overline{Y}_{g,\hat{\beta}_g^P} = \frac{1}{N_g} \sum_{i=1}^{N_g} \exp(\mathbf{X}_{ig}\hat{\beta}_g^P).$$

[1] Because the calculation of $S(\hat{\beta}_h, \mathbf{X}_{ig})$ is straightforward, we will present just the calculation of $S(\hat{\beta}_g, \mathbf{X}_{ig})$ in the remainder of the paper.

A well-known problem of the Poisson-model is the assumption that the dependent variable has the same mean and variance $\mu_{ig} = \exp(\mathbf{X}_{ig}\beta_g^P)$. If this assumption is violated, an alternative conditional distribution of the dependent variable needs to be specified that permits a more flexible specification of the variance of the dependent variable. The negative binomial (Negbin) regression model (NB) represents such an alternative. The Negbin regression model relaxes the assumption of equality of the conditional mean and the variance of the dependent variable while assuming the same form of the conditional mean as the Poisson-model. Hence, the sample counterpart of the conditional mean of the Negbin regression model is

$$S(\hat{\beta}_g^{NB}, \mathbf{X}_{ig}) = \overline{Y}_{g,\hat{\beta}_g^{NB}} = \frac{1}{N_g}\sum_{i=1}^{N_g}\exp(\mathbf{X}_{ig}\hat{\beta}_g^{NB}).$$

Different to the Poisson model, the Negbin model assumes a quadratic relationship between the variance and the mean, i.e.

$$V(Y_{ig}|\mathbf{X}_{ig}) = \mu_{ig} + \alpha\mu_{ig}^2.$$

where α is a scalar parameter to be estimated together with β_g^{NB}.

In addition to the Poisson and Negbin regression models, zero-inflated models are frequently used when analyzing count data. These models take into account that real-life data may contain excess zeros, causing a higher probability of zero values than is consistent with the Poisson and negative binomial distribution. In this case it could be assumed that zeros and positive values do not come from the same data generating process (Winkelmann 2000).

In order to investigate the probability of excess zeros, Lambert (1992) proposed a zero-inflated Poisson model, that allows for two different data generating regimes: the outcome of regime 1 (R1) is always zero, whereas the outcome of regime 2 (R2) is generated by a poisson process. In this model, the unconditional expectation of the dependent variable consists of the conditional probability of observing regime 2 and the conditional expectation of the zero-truncated density:

$$E(Y_{ig}|\mathbf{X}_{ig}) = (1 - Pr(R1|\mathbf{X}_{ig}))E(Y_{ig}|R2, \mathbf{X}_{ig}). \tag{4}$$

Lambert (1992) specifies the conditional probability of regime 1, that always leads

to a zero outcome, as a Logit model:

$$Pr(R1|\mathbf{X}_{ig}) = \frac{\exp(\mathbf{Z}_{ig}\gamma_g)}{1+\exp(\mathbf{Z}_{ig}\gamma_g)},$$

where $\mathbf{Z}_{ig}$ contains the covariates of the conditional probability of excess zeros and γ_g is the parameter vector to be estimated. The unconditional mean of the dependent variable specified by equation (4) can then be estimated for the zero-inflated Poisson and the zero-inflated Negbin model by

$$S(\hat{\beta}_g^j, \mathbf{X}_{ig}) = \frac{1}{N_g}\sum_{i=1}^{N_g}(1-(\widehat{Pr}(R1)|\mathbf{X}_{ig}))\hat{\mu}_{ig} = \frac{1}{N_g}\sum_{i=1}^{N_g}\frac{\exp(\mathbf{X}_{ig}\hat{\beta}_g^j)}{1+\exp(\mathbf{Z}_{ig}\hat{\gamma}_g^j)},$$

for $j = ZIP, ZINB$.

Hurdle models represent another modification of count data models. The hurdle model can be interpreted as a two-part model, where the first part is a binary outcome model, and the second part a truncated count data model. The unconditional mean of the dependent variable in these models is given by:

$$E(Y_{ig}|\mathbf{X}_{ig}) = Pr(Y_{ig}>0|\mathbf{X}_{ig})E(Y_{ig}|Y_{ig}>0,\mathbf{X}_{ig}).$$

According to Cameron and Trivedi (1998) the conditional expected values of Y_{ig} of the hurdle Poisson (HP) and the hurdle Negbin (HNB) model are given by

$$E(Y_{ig}|Y_{ig}>0,\mathbf{X}_{ig}) = \frac{\exp(\mathbf{X}_{ig}\beta_g^{HP})}{1-\exp(-\exp(\mathbf{X}_{ig}\beta_g^{HP}))}$$

and

$$E(Y_{ig}|Y_{ig}>0,\mathbf{X}_{ig}) = \frac{\exp(\mathbf{X}_{ig}\beta_g^{HNB})}{1-(1+\alpha\exp(\mathbf{X}_{ig}\beta_g^{HNB}))^{-\frac{1}{\alpha}}},$$

respectively. Assuming a logistic distribution for the underlying zero generating process, the unconditional expected values can be estimated by

$$S(\hat{\beta}_g^{HP}, \mathbf{X}_{ig}) = \frac{1}{N_g}\sum_{i=1}^{N_g}\frac{\exp(\mathbf{X}_{ig}\hat{\beta}_g^{HP})}{(1-\exp(-\exp(\mathbf{X}_{ig}\hat{\beta}_g^{HP})))(1+\exp(\mathbf{Z}_{ig}\hat{\gamma}_g^{HP}))}$$

and

$$S(\hat{\beta}_g^{HNB}, \mathbf{X}_{ig}) = \frac{1}{N_g}\sum_{i=1}^{N_g}\frac{\exp(\mathbf{X}_{ig}\hat{\beta}_g^{HNB})}{(1-(1+\alpha\exp(\mathbf{X}_{ig}\hat{\beta}_g^{HNB}))^{-\frac{1}{\alpha}})(1+\exp(\mathbf{Z}_{ig}\hat{\gamma}_g^{HNB}))}.$$

4 Limited Dependent Variable Models

4.1 Tobit Models

Limited dependent variable models comprise truncated regression models and models for censored and corner solution outcome variables. Technically, censored and corner solution outcome variables may be described appropriately by a Tobit model (Wooldridge 2002). While censored outcome variables are not observable for a part of the population (such as top-coded wage information or preferred labor supply), corner solution outcome variables take on the value zero with positive probability but represent a continuous random variable over strictly positive values (such as actual labor supply). In the Tobit model (TB), the dependent variable takes on the values a_1 and a_2 with positive probability and represents a continuous random variable over values between a_1 and a_2, i.e.

$$
\begin{aligned}
Y_{ig}^{*} &= \mathbf{X}_{ig}\beta_g^{TB} + \varepsilon_{ig}^{TB}, \\
Y_{ig} &= a_1 \quad \text{if} \quad Y_{ig}^{*} \leq a_1 \\
Y_{ig} &= a_2 \quad \text{if} \quad Y_{ig}^{*} \geq a_2 \\
Y_{ig} &= Y_{ig}^{*} = \mathbf{X}_{ig}\beta_g^{TB} + \varepsilon_{ig}^{TB} \quad \text{if} \quad a_1 < Y_{ig}^{*} < a_2, \\
\varepsilon_{ig} &\sim N(0, (\sigma_g^{TB})^2).
\end{aligned}
$$

If one is interested in the marginal effects of a latent censored outcome variable, the strategy would be to use the Tobit estimator in the standard Blinder-Oaxaca decomposition depicted in equation (1). However, the conventional decomposition method leads to erroneous predictions of the components of the decomposition equation if we aim at analyzing the observable corner solution outcome variable Y_{ig}. In this case, an alternative decomposition method must be applied.

Assuming homoscedastic and normal distributed error terms ε_{ig}^{TB}, the unconditional expectation of Y_{ig} given $\mathbf{X}_{ig}$ consists of the conditional expectations of Y_{ig}, weighted by the respective probabilities of observing a_1, a_2, or a value between a_1

and a_2, i.e.

$$\begin{aligned} E(Y_{ig}|\mathbf{X}_{ig}) &= a_1\Phi_1(\beta_g^{TB}, \mathbf{X}_g, \sigma_g^{TB}) + a_2\Phi_2(\beta_g^{TB}, \mathbf{X}_g, \sigma_g^{TB}) \\ &+ \Lambda(\beta_g^{TB}, \mathbf{X}_g, \sigma_g^{TB}) \left[\mathbf{X}_{ig}\beta_g^{TB} + \sigma_g^{TB}\frac{\lambda(\beta_g^{TB}, \mathbf{X}_g, \sigma_g^{TB})}{\Lambda(\beta_g^{TB}, \mathbf{X}_g, \sigma_g^{TB})}\right]. \end{aligned} \tag{5}$$

where

$$\Lambda(\beta_g^{TB}, \mathbf{X}_g, \sigma_g^{TB}) = 1 - \Phi[(\sigma_g^{TB})^{-1}(a_1 - \mathbf{X}_{ig}\beta_g^{TB})] - \Phi[(\sigma_g^{TB})^{-1}(a_2 - \mathbf{X}_{ig}\beta_g^{TB})]$$

and

$$\lambda(\beta_g^{TB}, \mathbf{X}_g, \sigma_g^{TB}) = \phi[(\sigma_g^{TB})^{-1}(a_1 - \mathbf{X}_{ig}\beta_g^{TB})] - \phi[(\sigma_g^{TB})^{-1}(a_2 - \mathbf{X}_{ig}\beta_g^{TB})].$$

$\phi(\cdot)$ represents the standard normal density function.

Equation (5) shows that a decomposition of the outcome variable similar to equation (2) is not appropriate for censored outcome variables, because the conditional expectations $E(Y_{ig}|X_{ig})$ in the Tobit model depend on the variance of the error term σ_g^{TB}. Even though the ancillary parameter σ_g^{TB} does not affect the sign of the marginal effects, it affects their magnitudes and therefore becomes important for the decomposition. Depending on which σ_g^{TB} is used in the counterfactual parts of the decomposition equation, several possibilities of decomposing the mean difference of Y_i between the two groups can be derived. Two possibilities are

$$\begin{aligned} \Delta_{AB}^{TB} &= \left[E_{\beta_A^{TB}, \sigma_A^{TB}}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta_A^{TB}, \sigma_B^{TB}}(Y_{iB}|\mathbf{X}_{iB})\right] \\ &+ \left[E_{\beta_A^{TB}, \sigma_B^{TB}}(Y_{iB}|\mathbf{X}_{iB}) - E_{\beta_B^{TB}, \sigma_B^{TB}}(Y_{iB}|\mathbf{X}_{iB})\right], \end{aligned} \tag{6}$$

and

$$\begin{aligned} \Delta_{AA}^{TB} &= \left[E_{\beta_A^{TB}, \sigma_A^{TB}}(Y_{iA}|\mathbf{X}_{iA}) - E_{\beta_A^{TB}, \sigma_A^{TB}}(Y_{iB}|\mathbf{X}_{iB})\right] \\ &+ \left[E_{\beta_A^{TB}, \sigma_A^{TB}}(Y_{iB}|\mathbf{X}_{iB}) - E_{\beta_B^{TB}, \sigma_B^{TB}}(Y_{iB}|\mathbf{X}_{iB})\right], \end{aligned} \tag{7}$$

where $E_{\beta_g^{TB}, \sigma_g^{TB}}(Y_{ig}|\mathbf{X}_{ig})$ now refers to the unconditional expectation of Y_{ig} evaluated at the parameter vector β_g^{TB} and the error variance σ_g^{TB}. In both equations, 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 the part of the differential in Y_{ig} that is due to differences in coefficients.

The two versions of the decomposition equation differ from each other as soon as large differences in the variance of the error term between the two groups exist. Note however, that the decomposition using σ_B^{TB} to calculate the counterfactual parts, as in equation (6), is more comparable to the OLS decomposition described in equation (1), since the counterfactual parts differ from $E_{\beta_B^{TB},\sigma_B^{TB}}(Y_{iB}|\mathbf{X}_{iB})$ only by using the parameter vector for group A, β_A^{TB}, rather than by using the parameter vector *and* the error variance for group A in the alternative decomposition described in equation (7).

Using the sample counterpart of equation (5),

$$\begin{aligned} S(\hat{\beta}_g^{TB}, \mathbf{X}_{ig}, \hat{\sigma}_g^{TB}) &= N^{-1} \sum_{i=1}^{N} a_1 \Phi_1(\hat{\beta}_g^{TB}, \mathbf{X}_g, \hat{\sigma}_g^{TB}) + a_2 \Phi_2(\hat{\beta}_g^{TB}, \mathbf{X}_g, \hat{\sigma}_g^{TB}) \\ &+ \Lambda(\hat{\beta}_g^{TB}, \mathbf{X}_{ig}, \hat{\sigma}_g^{TB}) \left[\mathbf{X}_{ig} \hat{\beta}_g^{TB} + \hat{\sigma}_g^{TB} \frac{\lambda(\hat{\beta}_g^{TB}, \mathbf{X}_{ig}, \hat{\sigma}_g^{TB})}{\Lambda(\hat{\beta}_g^{TB}, \mathbf{X}_{ig}, \hat{\sigma}_g^{TB})} \right], \end{aligned}$$

equation (6) can be estimated by

$$\begin{aligned} \hat{\Delta}_{AB}^{TB} &= \left[S(\hat{\beta}_A^{TB}, \mathbf{X}_{iA}, \hat{\sigma}_A^{TB}) - S(\hat{\beta}_A^{TB}, \mathbf{X}_{iB}, \hat{\sigma}_B^{TB}) \right] \\ &+ \left[S(\hat{\beta}_A^{TB}, \mathbf{X}_{iB}, \hat{\sigma}_B^{TB}) - S(\hat{\beta}_B^{TB}, \mathbf{X}_{iB}, \hat{\sigma}_B^{TB}) \right]. \end{aligned} \tag{8}$$

Similarly, equation (7) can be estimated by

$$\begin{aligned} \hat{\Delta}_{AA}^{TB} &= \left[S(\hat{\beta}_A^{TB}, \mathbf{X}_{iA}, \hat{\sigma}_A^{TB}) - S(\hat{\beta}_A^{TB}, \mathbf{X}_{iB}, \hat{\sigma}_A^{TB}) \right] \\ &+ \left[S(\hat{\beta}_A^{TB}, \mathbf{X}_{iB}, \hat{\sigma}_A^{TB}) - S(\hat{\beta}_B^{TB}, \mathbf{X}_{iB}, \hat{\sigma}_B^{TB}) \right] \end{aligned} \tag{9}$$

If the dependent variable is not truncated, i.e. if $a_1 \to -\infty$ and $a_2 \to \infty$, equations (6) and (7) reduce to the original Blinder-Oaxaca decomposition described in equation (1).

4.2 Truncated Regression Models

The results derived for the Tobit model can be easily transferred to a truncated regression model of the form

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

where the dependent variable is truncated at a lower limit a_1 and a higher limit a_2. The error terms ε_{ig}^{TR} are assumed to be homoscedastic and distributed normally with mean zero and variance $(\sigma_g^{TR})^2$. Consequently,

$$Y_{ig}|\mathbf{X}_{ig} \sim N(\mathbf{X}_{ig}\beta_g^{TR}, (\sigma_g^{TR})^2).$$

In this model, the unconditional expectation of Y_{ig} given $\mathbf{X}_{ig}$ consists of the conditional expectation of Y_{ig} weighted with the probability of observing a value between a_1 and a_2:

$$E_{\beta_g^{TR},\sigma_g^{TR}}(Y_{ig}|\mathbf{X}_{ig}) = \Lambda(\beta_g^{TR}, \mathbf{X}_g, \sigma_g^{TR}) \left[\mathbf{X}_{ig}\beta_g^{TR} + \sigma_g^{TR} \frac{\lambda(\beta_g^{TR}, \mathbf{X}_g, \sigma_g^{TR})}{\Lambda(\beta_g^{TR}, \mathbf{X}_g, \sigma_g^{TR})} \right].$$

Consequently, similar to equations (8) and (9), the components of the decomposition equation of the truncated regression model can be estimated by using the sample counterpart of the unconditional expectation:

$$S(\hat{\beta}_g^{TR}, \mathbf{X}_{ig}, \hat{\sigma}_g^{TR}) = N^{-1} \sum_{i=1}^{N} \Lambda(\hat{\beta}_g^{TR}, \mathbf{X}_{ig}, \hat{\sigma}_g^{TR}) \times \left[\mathbf{X}_{ig}\hat{\beta}_g^{TR} + \hat{\sigma}_g^{TR} \frac{\lambda(\hat{\beta}_g^{TR}, \mathbf{X}_{ig}, \hat{\sigma}_g^{TR})}{\Lambda(\hat{\beta}_g^{TR}, \mathbf{X}_{ig}, \hat{\sigma}_g^{TR})} \right].$$

5 Conclusion

In this paper, the decomposition method proposed by Blinder (1973) and Oaxaca (1973) is extended to non-linear models. The extension of the conventional decomposition method permits a decomposition of differences in a non-linear outcome variable between two groups into a part that may be explained by differences in observed characteristics and a part that is attributable to differences in the estimated coefficients.

The paper illustrates how the Blinder-Oaxaca decomposition can be applied to models with discrete and limited dependent variables. In particular, a Blinder-Oaxaca decomposition method for ordered Logit and Probit models is derived which represents a generalization of the decomposition method for binary Logit and Probit models proposed by Fairlie (1999, 2003). Moreover, the Blinder-Oaxaca decomposition is applied to count data models, including Poisson and Negative Binomial

models, zero-inflated Poisson and Negative Binomial models as well as Hurdle Poisson and Negative Binomial Models. An empirical application of the decomposition method for count data models is provided by Bauer, Göhlmann, and Sinning (2006). Finally, the Blinder-Oaxaca decomposition is extended to truncated regression and Tobit models, where the latter has been used by Bauer and Sinning (2005) to analyze differences in the savings behavior between natives and immigrants in Germany.

References

Bauer, T. K., S. Göhlmann, and M. Sinning (2006): "Gender Differences in Smoking Behavior," *RWI Discussion Papers No. 44*, pp. 1–19.

Bauer, T. K., and M. Sinning (2005): "Blinder-Oaxaca Decomposition for Tobit Models," *RWI Discussion Papers No. 32*, pp. 1–10.

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

Cameron, A. C., and P. K. Trivedi (1998): "Regression Analysis of Count Data," Cambridge University Press, Cambridge.

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.

——— (2003): "An Extension of the Blinder-Oaxaca Decomposition Technique to Logit and Probit Models," *Yale University Economic Growth Center Discussion Paper No. 873*, pp. 1–11.

Juhn, C., K. M. Murphy, and B. Pierce (1991): "Accounting for the Slowdown in Black-White Wage Convergence," in *Workers and Their Wages: Changing Patterns in the United States*, ed. by M. H. Kosters. American Enterprise Institute, Washington.

Lambert, D. (1992): "Zero-inflated Poisson Regression with an Application to Defects in Manufacturing," *Technometrics*, 34, 1–14.

Neumark, D. (1988): "Employers' Discriminatory Behavior and the Estimation of Wage Discrimination," *Journal of Human Resources*, 23, 279–295.

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

Oaxaca, R. L., and M. 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.

Winkelmann, R. (2000): "Econometric Analysis of Count Data," Springer Verlag, Berlin, Heidelberg.

Wooldridge, J. M. (2002): "Econometric Analysis of Cross Section and Panel Data," MIT Press, Cambridge, Massachusetts.