

IZA DP No. 1795

## Blinder-Oaxaca Decomposition for Tobit Models

Thomas K. Bauer  
Mathias Sinning

October 2005

# Blinder-Oaxaca Decomposition for Tobit Models

**Thomas K. Bauer**

*RWI Essen, University of Bochum,  
CEPR London and IZA Bonn*

**Mathias Sinning**

*RWI Essen*

Discussion Paper No. 1795  
October 2005

IZA

P.O. Box 7240  
53072 Bonn  
Germany

Phone: +49-228-3894-0  
Fax: +49-228-3894-180  
Email: [iza@iza.org](mailto:iza@iza.org)

Any opinions expressed here are those of the author(s) and not those of the institute. Research disseminated by IZA may include views on policy, but the institute itself takes no institutional policy positions.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit company supported by Deutsche Post World Net. The center is associated with the University of Bonn and offers a stimulating research environment through its research networks, research support, and visitors and doctoral programs. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

## ABSTRACT

### **Blinder-Oaxaca Decomposition for Tobit Models<sup>\*</sup>**

In this paper, a decomposition method for Tobit-models is derived, which allows the differences in a censored 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 coefficients. The method is applied to a decomposition of the gender wage gap using German data.

JEL Classification: C24, J31

Keywords: Blinder-Oaxaca decomposition, Tobit model, wage gap

Corresponding author:

Thomas K. Bauer  
RWI Essen  
Hohenzollernstr. 1-3  
D-45128 Essen  
Germany  
E-mail: [bauer@rwi-essen.de](mailto:bauer@rwi-essen.de)

---

<sup>\*</sup> The authors are grateful to Michael Fertig and Harald Tauchmann for helpful comments on an earlier draft of this paper.

# 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, since it permits the decomposition of the difference in an outcome variable between two groups into a part that is explained by differences in the observed characteristics of these groups and a part that is due to differences in the estimated coefficients. Among other applications, the Blinder-Oaxaca decomposition has been used in numerous studies of wage-differentials between males and females or between different ethnic groups (Altonji and Black 1999). In these studies, the unexplained part of the decomposition is interpreted as discrimination.

So far, the Blinder-Oaxaca-decomposition and its various generalizations have mainly be used in linear regression models. A decomposition method for models with binary dependent variables has been developed by Fairlie (1999, 2003). In many cases, however, the censoring of outcome variables requires the estimation of limited dependent variable models. In such situations, OLS might yield in inconsistent parameter estimates and in turn misleading decomposition results. This paper aims at providing a solution to this problem by deriving a decomposition method for Tobit-models. To illustrate this method, we apply it to the gender wage gap using German data.

## 2 Blinder-Oaxaca Decomposition for Tobit Models

Consider the following linear regression model, which is estimated separately for the groups  $g = m, f$

$$Y_{ig} = \mathbf{X}_{ig}\beta_g + \varepsilon_{ig}, \tag{1}$$

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

$$\begin{aligned}\bar{Y}_m - \bar{Y}_f &= \Delta^{OLS} = [E_{\beta_m}(Y_{im}|\mathbf{X}_{im}) - E_{\beta_m}(Y_{if}|\mathbf{X}_{if})] \\ &\quad + [E_{\beta_m}(Y_{if}|\mathbf{X}_{if}) - E_{\beta_f}(Y_{if}|\mathbf{X}_{if})] \\ &= (\bar{X}_m - \bar{X}_f)\hat{\beta}_m + \bar{X}_f(\hat{\beta}_m - \hat{\beta}_f),\end{aligned}\tag{2}$$

where  $\bar{Y}_g = N_g^{-1} \sum_{i=1}^{N_g} Y_{ig}$  and  $\bar{\mathbf{X}}_g = N_g^{-1} \sum_{i=1}^{N_g} \mathbf{X}_{ig}$ .  $E_{\beta_g}(Y_{ig}|\mathbf{X}_{ig})$  refers to the conditional expectation of  $Y_{ig}$  evaluated at the parameter vector  $\beta_g$ . The first term on the right hand side of equation (2) 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.

Given  $\mathbf{X}_{ig}$ , the linear model is a good approximation of the expected value of the outcome variable  $E(Y_{ig}|\mathbf{X}_{ig})$  for values of  $\mathbf{X}_{ig}$  close to the mean. If the outcome variable  $Y_{ig}$  is censored, however, the use of OLS may lead to biased estimates of the parameter vector and hence misleading results of the decomposition.

To illustrate the Blinder-Oaxaca decomposition for censored regression models, we consider a Tobit model, where the distribution of the dependent variable is censored from above at the point  $a_1$  and from below at the point  $a_2$ , i.e.

$$\begin{aligned}Y_{ig}^* &= \mathbf{X}_{ig}\beta_g + \varepsilon_{ig}, \\ Y_{ig} &= a_1 \quad \text{if } Y_{ig}^* \leq a_1 \\ Y_{ig} &= a_2 \quad \text{if } Y_{ig}^* \geq a_2 \\ Y_{ig} &= Y_{ig}^* = \mathbf{X}_{ig}\beta_g + \varepsilon_{ig} \quad \text{if } a_1 < Y_{ig}^* < a_2, \\ \varepsilon_{ig} &\sim N(0, \sigma_g^2).\end{aligned}\tag{3}$$

The unconditional expectation of  $Y_{ig}$  given  $\mathbf{X}_{ig}$  consists of the conditional expectations of  $Y_{ig}$  weighted with the respective probabilities of being censored (from above or below) or uncensored:

$$\begin{aligned}E(Y_{ig}|\mathbf{X}_{ig}) &= a_1\Phi_1(\beta_g, \mathbf{X}_g, \sigma_g) + a_2\Phi_2(\beta_g, \mathbf{X}_g, \sigma_g) \\ &\quad + \Lambda(\beta_g, \mathbf{X}_g, \sigma_g) \left[ \mathbf{X}_{ig}\beta_g + \sigma \frac{\lambda(\beta_g, \mathbf{X}_g, \sigma_g)}{\Lambda(\beta_g, \mathbf{X}_g, \sigma_g)} \right],\end{aligned}\tag{4}$$

where  $\Phi_1(\beta_g, \mathbf{X}_g, \sigma_g) = \Phi[\sigma_g^{-1}(a_1 - \mathbf{X}_{ig}\beta_g)]$ ,  $\Phi_2(\beta_g, \mathbf{X}_g, \sigma_g) = \Phi[\sigma_g^{-1}(a_2 - \mathbf{X}_{ig}\beta_g)]$ ,  $\Lambda(\cdot) = \Phi_2(\cdot) - \Phi_1(\cdot)$  and  $\lambda(\beta_g, \mathbf{X}_g, \sigma_g) = \phi[\sigma_g^{-1}(a_1 - \mathbf{X}_{ig}\beta_g)] - \phi[\sigma_g^{-1}(a_2 - \mathbf{X}_{ig}\beta_g)]$  for  $g = m, f$ .  $\phi(\cdot)$  represents the standard normal density function and  $\Phi(\cdot)$  is the cumulative standard normal density function.

Equation (4) shows that a decomposition of the outcome variable similar to equation (2) is not appropriate if the dependent variable is censored. Particularly, in contrast to the linear regression model, the conditional expectations  $E(Y_{ig}|X_{ig})$  in the Tobit model depend on the variance of the error term  $\sigma_g$ . Consequently, there are several possibilities to decompose the mean difference of  $Y_i$  between the two groups depending on which  $\sigma_g$  is used in the counterfactual parts of the decomposition equation. Two possible decompositions are

$$\begin{aligned} \Delta_f^{Tobit} &= [E_{\beta_m, \sigma_m}(Y_{im}|\mathbf{X}_{im}) - E_{\beta_m, \sigma_f}(Y_{if}|\mathbf{X}_{if})] \\ &\quad + [E_{\beta_m, \sigma_f}(Y_{if}|\mathbf{X}_{if}) - E_{\beta_f, \sigma_f}(Y_{if}|\mathbf{X}_{if})], \end{aligned} \quad (5)$$

and

$$\begin{aligned} \Delta_m^{Tobit} &= [E_{\beta_m, \sigma_m}(Y_{im}|\mathbf{X}_{im}) - E_{\beta_m, \sigma_m}(Y_{if}|\mathbf{X}_{if})] \\ &\quad + [E_{\beta_m, \sigma_m}(Y_{if}|\mathbf{X}_{if}) - E_{\beta_f, \sigma_f}(Y_{if}|\mathbf{X}_{if})], \end{aligned} \quad (6)$$

where  $E_{\beta_g, \sigma_g}(Y_{ig}|\mathbf{X}_{ig})$  now refers to the conditional expectation of  $Y_{ig}$  evaluated at the parameter vector  $\beta_g$  and the error variance  $\sigma_g$  for  $g = f, m$ . 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 may differ from each other, if large differences in the variance of the error term between the two groups exist. Note however, that the decomposition using  $\sigma_f$  to calculate the counterfactual parts, as in equation (5), is more comparable to the OLS decomposition described in equation (2), since the counterfactual parts differ from  $E_{\beta_f, \sigma_f}(Y_{if}|\mathbf{X}_{if})$  only by using the parameter vector for group  $m$ ,  $\beta_m$ , rather than by using the parameter vector *and* the error variance for group  $m$  in the alternative decomposition described in equation (6).

Using the *sample counterpart* of equation (4),

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

equation (5) can be estimated by

$$\hat{\Delta}_f^{Tobit} = \left[ S(\hat{\beta}_m, \mathbf{X}_{im}, \hat{\sigma}_m) - S(\hat{\beta}_m, \mathbf{X}_{if}, \hat{\sigma}_f) \right] + \left[ S(\hat{\beta}_m, \mathbf{X}_{if}, \hat{\sigma}_f) - S(\hat{\beta}_f, \mathbf{X}_{if}, \hat{\sigma}_f) \right] \quad (7)$$

Similarly, equation (6) can be estimated by

$$\hat{\Delta}_f^{Tobit} = \left[ S(\hat{\beta}_m, \mathbf{X}_{im}, \hat{\sigma}_m) - S(\hat{\beta}_m, \mathbf{X}_{if}, \hat{\sigma}_m) \right] + \left[ S(\hat{\beta}_m, \mathbf{X}_{if}, \hat{\sigma}_m) - S(\hat{\beta}_f, \mathbf{X}_{if}, \hat{\sigma}_f) \right] \quad (8)$$

If the dependent variable is not censored, i.e. if  $a_1 \rightarrow -\infty$  and  $a_2 \rightarrow \infty$ , both equations reduce to the original Blinder-Oaxaca decomposition described in equation (2).

### 3 Empirical Illustration: Gender Wage Gap in Germany

To illustrate how the Blinder-Oaxaca decomposition for Tobit models works, we analyze the gender wage gap using data from the German Socioeconomic Panel (GSOEP) for the year 2004. We estimate the following wage equation separately for males ( $m$ ) and females ( $f$ ):

$$\ln(w_{ig}) = \mathbf{X}_{ig} \beta_g + \varepsilon_{ig}, \quad (9)$$

for  $g = m, f$ , where  $w_{ig}$  refers to the gross hourly wage rate of individual  $i$  in group  $g$ . The explanatory variables  $\mathbf{X}_i$  include the years of completed schooling, potential labor market experienced (calculated as *Age - Years of Schooling - 6*) and potential labor market experience squared, the number of children, and dummy

variables for married individuals, part-time workers, immigrants, and persons residing in East-Germany. We restrict our sample to working individuals aged 16 to 65. We eliminated all observations with missing values for at least one of the variables used in the analysis, which yields a sample of 3,610 observations for men and 2,465 observations for women.

Since the wage information in the GSOEP is not censored, we apply in a first step the original Blinder-Oaxaca decomposition described in equation (2) using the results of OLS-estimates of the regression model (9). In a second step, we censor the distribution of gross hourly wages at the lower and upper 10th percentile and estimate equation (9) by OLS using the transformed wage information as dependent variable to show the potential bias in the estimation results and wage decomposition when ignoring that the dependent variable is censored. In a final step, we use the transformed wage variable and estimate equation (9) using a Tobit model and apply the Tobit-Blinder-Oaxaca decompositions described in equations (7) and (8)<sup>1</sup>. To be able to test the different decomposition results against each other, we obtained standard errors for the decomposition parts by bootstrapping with 1000 replications.

Table 1 reports the results from this analysis. The estimated coefficients of the OLS and Tobit-models reported in Parts A and B of Table 1, respectively, have the expected signs and are statistically significant at conventional levels. When using the artificial censored dependent variable, the Tobit estimates perform slightly better than the OLS-estimates, i.e. are closer to the respective estimation results when using the original uncensored wage information.

Based on the uncensored wage information, the results of the decomposition analysis reported in column 1 of Table 2 (which does not differ between the OLS and the various Tobit-decomposition methods) shows that more than 67% of the wage differential between men and women is attributable to differences in observable

---

<sup>1</sup>We censored the dependent variable artificially in our example because we want to compare censored and uncensored estimates. Artificial censoring, however, does also permit an alternative decomposition strategy, which is based on the unconditional expectation of the latent dependent variable,  $E(Y_{ig}^*|\mathbf{X}_{ig}) = E(Y_{ig}^*)$ , instead of the unconditional expectation  $E(Y_{ig}^*|\mathbf{X}_{ig})$ . In such a case, it is sufficient to estimate the parameters of the Tobit model and to calculate the components of equation (2).



characteristics.

When using the original Blinder-Oaxaca decomposition ( $\Delta^{OLS}$ ), censoring the dependent variable from below or from both sides of the wage distribution increases the unexplained part of the wage differential, while the decomposition results do not change very much when wages are censored just from above. Furthermore, for left-censoring and censoring from both sides of the wage distribution the Tobit decomposition methods perform better than the original Blinder-Oaxaca decomposition. However, in our example t-tests demonstrate that the differences in the decomposition results between the uncensored and the three censored estimations are not statistically significant in all cases.

## 4 Conclusion

In this paper, a decomposition method for Tobit-models is derived. This method allows the decomposition of the difference in a censored outcome variable between two groups into a part that is explained by differences in the observed characteristics and a part attributable to differences in the estimated coefficients of these characteristics. Using data of the GSOEP, we find that the major part of the wage differential between men and women is attributable to differences in observable characteristics. In our application, applying the Tobit decomposition method produces better results than the original Blinder-Oaxaca decomposition when wages are censored from below and from both sides of the wage distribution. However, in our example the differences between the various decomposition methods are not statistically significant.

## References

- ALTONJI, J., AND R. BLACK (1999): "Race and Gender in the Labor Market," in *Handbook of Labor Economics, Vol. 3C*, ed. by O. Ashenfelter, and D. Card. Elsevier Science, Amsterdam.
- BLINDER, A. S. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 8, 436–455.
- 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. Koster. American Enterprise Institute, Washington.
- 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.

**Table 1: Estimation Results**

	uncensored		left-censored		right-censored		left/right-censored	
	Men	Women	Men	Women	Men	Women	Men	Women
<b>A: OLS estimates</b>								
Education (Yrs.)	0.085 (0.003)***	0.077 (0.004)***	0.082 (0.003)***	0.074 (0.003)***	0.065 (0.002)***	0.070 (0.004)***	0.064 (0.002)***	0.068 (0.003)***
Experience	0.027 (0.003)***	0.035 (0.004)***	0.024 (0.003)***	0.029 (0.003)***	0.024 (0.003)***	0.033 (0.003)***	0.021 (0.003)***	0.027 (0.003)***
Experience <sup>2</sup> × 10 <sup>-2</sup>	-0.031 (0.007)***	-0.062 (0.008)***	-0.027 (0.006)***	-0.048 (0.006)***	-0.030 (0.006)***	-0.059 (0.007)***	-0.026 (0.005)***	-0.046 (0.006)***
Constant	1.223 (0.054)***	1.136 (0.064)***	1.306 (0.048)***	1.247 (0.052)***	1.482 (0.046)***	1.238 (0.060)***	1.540 (0.040)***	1.339 (0.048)***
R <sup>2</sup>	0.38	0.26	0.40	0.30	0.36	0.25	0.41	0.31
<b>B: Tobit estimates</b>								
Education (Yrs.)	0.085 (0.003)***	0.077 (0.004)***	0.086 (0.003)***	0.083 (0.004)***	0.085 (0.003)***	0.075 (0.004)***	0.085 (0.003)***	0.080 (0.004)***
Experience	0.027 (0.003)***	0.035 (0.004)***	0.026 (0.003)***	0.035 (0.004)***	0.027 (0.003)***	0.035 (0.004)***	0.025 (0.003)***	0.034 (0.003)***
Experience <sup>2</sup> × 10 <sup>-2</sup>	-0.031 (0.007)***	-0.062 (0.008)***	-0.029 (0.007)***	-0.059 (0.007)***	-0.033 (0.007)***	-0.061 (0.008)***	-0.030 (0.006)***	-0.058 (0.007)***
Constant	1.223 (0.054)***	1.136 (0.064)***	1.230 (0.052)***	1.060 (0.061)***	1.215 (0.054)***	1.166 (0.062)***	1.242 (0.051)***	1.108 (0.059)***
McFadden Pseudo R <sup>2</sup>	0.27	0.19	0.29	0.22	0.27	0.19	0.29	0.23

*Notes:* 3,610 observations for men and 2,465 observations for women. Standard errors in parentheses. \* significant at 10%; \*\* significant at 5%; \*\*\* significant at 1%. Additional variables used: number of children and dummy-variables for marital status, part-time employment, immigrants and East-Germany.

**Table 2: Decomposition Results**

	uncensored	left-censored	right-censored	left/right-censored
$\widehat{\Delta}^{OLS}$	0.326*** (0.014)	0.301*** (0.012)	0.288*** (0.013)	0.268*** (0.011)
Explained Part	0.220*** (0.025)	0.173*** (0.019)	0.198*** (0.023)	0.153*** (0.017)
in % of $\widehat{\Delta}^{OLS}$	67.6*** (7.7)	57.3*** (6.1)	68.8*** (8.3)	57.1*** (6.2)
Unexplained Part	0.105*** (0.026)	0.128*** (0.019)	0.089*** (0.025)	0.115*** (0.017)
in % of $\widehat{\Delta}^{OLS}$	32.3*** (7.7)	42.6*** (6.1)	31.1*** (8.3)	42.8*** (6.2)
$\widehat{\Delta}_f^{Tobit}$	0.326*** (0.014)	0.301*** (0.013)	0.293*** (0.013)	0.270*** (0.011)
Explained Part	0.220*** (0.025)	0.189*** (0.019)	0.194*** (0.024)	0.164*** (0.017)
in % of $\widehat{\Delta}^{Tobit}$	67.6*** (7.7)	62.7*** (6.1)	66.3*** (8.2)	60.6*** (6.4)
Unexplained Part	0.105*** (0.026)	0.112*** (0.019)	0.098*** (0.025)	0.106*** (0.018)
in % of $\widehat{\Delta}^{Tobit}$	32.3*** (7.7)	37.2*** (6.1)	33.6*** (8.2)	39.3*** (6.4)
$\sigma_f$	0.453	0.420	0.439	0.399
$\widehat{\Delta}_m^{Tobit}$	0.326*** (0.014)	0.301*** (0.013)	0.293*** (0.013)	0.270*** (0.011)
Explained Part	0.220*** (0.025)	0.187*** (0.018)	0.195*** (0.024)	0.163*** (0.017)
in % of $\widehat{\Delta}^{Tobit}$	67.6*** (7.7)	62.1*** (6.0)	66.5*** (8.3)	60.3*** (6.3)
Unexplained Part	0.105*** (0.026)	0.114*** (0.019)	0.098*** (0.025)	0.107*** (0.018)
in % of $\widehat{\Delta}^{Tobit}$	32.3*** (7.7)	37.8*** (6.0)	33.4*** (8.3)	39.6*** (6.3)
$\sigma_m$	0.455	0.431	0.444	0.412

*Notes:* Decomposition results based on the regression results in Table 1. Bootstrapped (1,000 replications) standard errors in parentheses. \* significant at 10%; \*\* significant at 5%; \*\*\* significant at 1%.

# Appendix

TABLE: Descriptive Statistics

	Wages		Education		Experience	
	(1)	(2)	(3)	(4)	(5)	(6)
	Men	Women	Men	Women	Men	Women
Uncensored	15.879 (0.239)	11.654 (0.234)	12.398 (0.070)	12.263 (0.082)	23.707 (0.268)	22.664 (0.379)
Left-censored	15.977 (0.236)	11.836 (0.230)	-	-	-	-
Right-censored	14.847 (0.164)	11.360 (0.164)	-	-	-	-
Left/Right-censored	14.945 (0.160)	11.542 (0.159)	-	-	-	-

Notes: 3,610 observations for men and 2,465 observations for women. Standard deviations in parentheses.