Decomposition of non-linear models using simulated residuals

François-Charles Wolff

Document de Travail
Working Paper

Lemna

EA 4272

Decomposition of non-linear models using simulated residuals

François-Charles Wolff*

2012/24

*LEMNA - Université de Nantes et INED

INSTITUT d'ECONOMIE et de MANAGEMENT de NANTES - IAE

Laboratoire d'Economie et de Management Nantes-Atlantique
Université de Nantes
Chemin de la Censive du Tertre – BP 52231
44322 Nantes cedex 3 – France
www.univ-nantes.fr/iemn-iae/recherche
Tél. +33 (0)2 40 14 17 17 – Fax +33 (0)2 40 14 17 49

UNIVERSITÉ DE NANTES

Decomposition of non-linear models using simulated residuals[#]

François-Charles Wolff[*]

March 2012

Abstract: This paper proposes to decompose non-linear models deduced from a latent regression framework using the latent dependent outcome as dependent variable and the Oaxaca-Blinder decomposition technique. Values of the unobserved latent outcome are obtained using simulated residuals.

Keywords: Blinder-Oaxaca decomposition, non-linear models, simulated residuals

JEL Classification: C20

[#] I am indebted to an anonymous reviewer for helpful comments. The usual disclaimer applies.
[*] Corresponding author. LEMNA, Université de Nantes, BP 52231 Chemin de la Censive du Tertre, 44322 Nantes Cedex 3, France and INED, Paris, France. Tel: 33 2 40141779.
E-mail: francois.wolff@univ-nantes.fr http://www.sc-eco.univ-nantes.fr/~fcwolff

To my father, who passed away on March 11, 2012

1. Introduction

A common approach used to identify the cause of gender or racial differences in labor market outcomes is the well-known Blinder-Oaxaca decomposition (Blinder, 1973, Oaxaca, 1973). The mean difference in wages between two groups, for instance men and women, is divided into one component that is explained by differences in characteristics like education or experience and one component attributable to differences in the return to these observable characteristics. This unexplained part of the gap is most often seen as a measure for discrimination.

One reason of the success of this decomposition technique lies in its large application potential. As emphasized in Jann (2008), this technique may be used to study group differences in any outcome variable. When the dependent variable is continuous, the decomposition is very easy to implement as it only requires estimates from linear regressions for the outcome of interest and sample means of the covariates used in the regressions. However, the problem becomes a little bit more complex when the outcome variable is non-linear since OLS estimates cannot be used directly in the Blinder-Oaxaca decomposition equation.

Recently, a few papers have attempted to generalize the Blinder-Oaxaca decomposition to non-linear models. Fairlie (1999) provides a decomposition method for binary dependent variables explained by either Logit or Probit models. Yun (2004) proposes a general methodology for decomposing differences in the first moment when the dependent variable may be expressed as a once differentiable function of a linear combination of covariates. Finally, Bauer and Sinning (2008) rewrite the usual decomposition equation in terms of conditional expectation to obtain a general version of the Oaxaca-Blinder decomposition that can be applied to models with either discrete or limited dependent variables (see also Sinning et al., 2008).

In this paper, I propose an alternative procedure to decompose non-linear models deduced from a latent regression framework by a non-linear mapping, which includes as special cases Probit, ordered Probit or Tobit models among others. Instead of working on the conditional expectation of the observed dependent variable, I decompose the corresponding latent outcome following the

Oaxaca-Blinder decomposition technique. Values of the unobserved latent dependent variable are obtained using simulated residuals (Gouriéroux et al., 1987)[1].

The remainder of my presentation is organized as follows. In the next Section, after a brief description of the Oaxaca-Blinder decomposition, I explain how to simulate the latent outcome corresponding to a discrete or limited dependent variable. In Section 3, I perform a Monte-Carlo analysis using a binary dependent variable. Section 4 concludes.

2. Decomposition methodology

For the presentation, I consider the Oaxaca-Blinder decomposition in the context of a linear model. In a setting with two groups $g = (A, B)$, I denote by Y_g^* a continuous dependent variable and assume that Y_g^* is explained by a vector of explanatory variables X_g following a linear regression model:

$$Y_g^* = X_g\beta_g + \varepsilon_g \tag{1}$$

with β_g the corresponding estimates and ε_g a random perturbation. The gap between the mean outcomes Y_B^* and Y_A^* may be expressed as (Blinder, 1973, Oaxaca, 1973):

$$Y_B^* - Y_A^* = (X_B - X_A)\beta_B + (\beta_B - \beta_A)X_A \tag{2}$$

From (2), the gap in Y^* is the sum of a characteristics effect and a coefficients effect. The first term $(X_B - X_A)\beta_B$ corresponds to differences attributable to observable characteristics, while the second term $(\beta_B - \beta_A)X_A$ measures differences in the return to these characteristics. As discussed in Oaxaca and Ransom (1994), there are several matrix of weights available to decompose $Y_B^* - Y_A^*$:

$$Y_B^* - Y_A^* = (X_B - X_A)(\Omega * \beta_B + (I - \Omega) * \beta_A) + (\beta_B - \beta_A)(X_B(I - \Omega) + X_A\Omega) \tag{3}$$

where I is the identity matrix and Ω is a matrix of weight. (2) is the case when $\Omega = I$, but other possibilities for the Oaxaca-Blinder decomposition are for instance $\Omega = 0$ or $\mathrm{diag}(\Omega) = 0.5$. In the case of a non-linear model, the problem with (2) or (3) is that the conditional expectation $E(Y_g^*|X_g)$ is likely to differ from $X_g\beta_g$.

[1] In the context of linear models, Yun (2007) proposes an extension of the Oaxaca-Blinder decomposition using generalized residuals, which refer to conditional expectations of residuals.

In what follows, I consider models deduced from a latent regression framework by a non-linear mapping, which includes a large set of discrete and limited dependent models (like univariate Probit, ordered Probit or Tobit regressions). In that case, the dependent variable is defined by $Y_g = f(Y_g^*)$ where Y_g^* is a continuous, latent outcome such that $Y_g^* = X_g\beta_g + \varepsilon_g$. If the latent variable Y_g^* were observed, then a simple strategy to decompose the gap $Y_B^* - Y_A^*$ would consist in an application of (2) or (3).

A solution to this problem of unobservability was suggested more than twenty years ago by Gouriéroux et al. (1987). Drawing on simulated residuals, the idea consists in simulating the values of the unobserved latent variables Y_B^* and Y_A^*. The methodology involves two steps. First, maximum likelihood estimates $\hat{\beta}_g$ of β_g are obtained from the appropriate specification given Y_g and f. Then, one simulates values $\tilde{Y}_g^*$ for the latent dependent outcome Y_g^* given Y_g, X_g and $\hat{\beta}_g$.

For the sake of illustration, consider the case of a Probit model. The dependent variable Y_g is such that $Y_g = 1$ if $Y_g^* > 0$ and $Y_g = 0$ when $Y_g^* \leq 0$, f is the normal distribution function and residuals are such that $\varepsilon_g \sim N(0;1)$. Application of simulated residuals is implemented as follows. First, estimation of two Probit models gives $\hat{\beta}_g$ for both groups A and B. Then, residuals ε_g are drawn from $N(0;1)$ and $\tilde{Y}_g^*$ is the first value satisfying $\tilde{Y}_g^* = X_g\hat{\beta}_g + \varepsilon_g > 0$ when $Y_g = 1$ and $\tilde{Y}_g^* = X_g\hat{\beta}_g + \varepsilon_g \leq 0$ when $Y_g = 0$. As a final step, I estimate $\tilde{Y}_B^* = X_B\tilde{\beta}_B + \epsilon_B$ and $\tilde{Y}_A^* = X_A\tilde{\beta}_A + \epsilon_A$ using OLS and apply a linear decomposition to the difference $\tilde{Y}_B^* - \tilde{Y}_A^* = (X_B - X_A)\tilde{\beta}_B + (\tilde{\beta}_B - \tilde{\beta}_A)X_A$ following (2)[2].

3. A Monte Carlo experiment

I evaluate the decomposition procedure by undertaking a Monte Carlo experiment, which consists of 200 replications on a sample of 10,000 observations. The explanatory variables are drawn from normal distribution with $X_A \sim N(4;1)$ and $X_B \sim N(5;1)$. Then, I calculate the values of the latent dependent variables such that $Y_A^* = 2 + 0.5 * X_A + \varepsilon_A$ and $Y_B^* = 2 + 0.6 * X_B + \varepsilon_B$, with $\varepsilon_A \sim N(0;1)$ and $\varepsilon_B \sim N(0;1)$. I set to 0.5 the proportion of individuals belonging to each group.

Since Y_A^* and Y_B^* are continuous, implementation of the Oaxaca-Blinder decomposition is straightforward. By construction, when the reference group is $\Omega = I$, then $Y_B^* - Y_A^* = (X_B - X_A)\beta_B +$

[2] The second-stage estimator $\tilde{\beta}$ is a strongly consistent estimator of β (see Gouriéroux et al., 1987, Theorem 1).

$(\beta_B - \beta_A)X_A$ and the contributions of $(X_B - X_A)\beta_B$ and $(\beta_B - \beta_A)X_A$ amounts to 60% and 40% respectively. Conversely, when $\Omega = 0$, then $Y_B^* - Y_A^* = (X_B - X_A)\beta_A + (\beta_B - \beta_A)X_B$ and the gap is explained equally by differences in characteristics (50%) and differences in coefficients (50%).

I define my binary dependent variable as follows. Denoting by $\bar{Y}^*$ the average value of the dependent variable on the whole sample (it is equal to 4.5 by construction), I define two dummy variables Y_g $(g = A, B)$ such that $Y_g = 1$ when $Y_g^* > \bar{Y}^*$ and $Y_g = 0$ otherwise. Then, I proceed in the following way. First, I calculate the Oaxaca-Blinder decomposition using the latent observed outcomes $Y_g^* - \bar{Y}^*$[3]. As shown in Table 1, I obtain the expected values for the characteristics and coefficients effects: respectively 60% and 40% when $\Omega = I$, and 50% and 50% when $\Omega = 0$. Then, I focus on the binary outcomes Y_g and implement a non-linear Probit decomposition following Sinning et al. (2008). Very similar results are obtained since 59.3% of the gap $Y_B - Y_A$ is explained by differences in characteristics when $\Omega = I$. The corresponding percentage is 51.0% when $\Omega = 0$.

Next, I suppose that the latent variables Y_g^* are unobserved, but information on the binary outcomes Y_g is available. I turn to simulated residuals to get simulated values of the unobserved latent variables $\tilde{Y}_g^*$. As shown in Table 1, the estimates $\tilde{\beta}_A$ and $\tilde{\beta}_B$ obtained from OLS regressions are very similar to the original parameters since I get $\tilde{\beta}_A = 0.501$ and $\tilde{\beta}_B = 0.604$. When decomposing $\tilde{Y}_B^* - \tilde{Y}_A^*$ following the Oaxaca-Blinder procedure, the values found for the characteristics and coefficients effects are very close to expectations: 60.4% for the characteristics effect when $\Omega = I$ and 50.0% when $\Omega = 0$.

I perform additional simulations on the binary dependent variables defined as $Y_g = 1$ when the latent outcomes Y_g^* exceed the 25th percentile of Y^* and the 75th percentile of Y^* respectively. I find that decomposing the observed binary variables and the latent outcomes leads to substantial differences. For instance, when $\Omega = I$, the weight of the characteristics effect obtained from a non-linear decomposition is 54.2% at the 25th percentile instead of 60% with the Oaxaca-Blinder technique. Of course, this is due to the different definition of the dependent variables[4]. When decomposing the simulated values of the latent outcomes, I again find estimated weights that are very close to expectations. The contribution of the characteristics effect is 60.4% at the 25th percentile and 60.1% at the 75th percentile when $\Omega = I$. Furthermore, $\tilde{\beta}_A$ and $\tilde{\beta}_B$ are almost equal to 0.5 and 0.6.

[3] By definition, $Y_B^* - Y_A^* = (Y_B^* - \bar{Y}^*) - (Y_A^* - \bar{Y}^*)$.

[4] If the purpose is for instance to explain the labor force participation of men and women, then Y_g^* is a measure of the propensity to work while Y_g is equal to one when an individual is working.

Table 1. Decomposition from Monte Carlo simulations

Decomposition		$Y_g = 1_{Y_g^* > \bar{Y}^*}$			$Y_g = 1_{Y_g^* > P25(Y^*)}$			$Y_g = 1_{Y_g^* > P75(Y^*)}$		
		Oaxaca	Probit	Simulated	Oaxaca	Probit	Simulated	Oaxaca	Probit	Simulated
$\Omega = I$										
Dif. in characteristics	proportion	0.600	0.593	0.604	0.600	0.542	0.604	0.600	0.649	0.601
	mse	0.018	0.023	0.029	0.018	0.033	0.029	0.018	0.020	0.034
Dif. in coefficients	proportion	0.400	0.407	0.396	0.400	0.458	0.396	0.400	0.351	0.399
	mse	0.018	0.023	0.029	0.018	0.033	0.029	0.018	0.020	0.034
$\Omega = 0$										
Dif. in characteristics	proportion	0.500	0.510	0.500	0.500	0.605	0.500	0.501	0.411	0.500
	mse	0.018	0.023	0.026	0.017	0.021	0.028	0.017	0.029	0.028
Dif. in coefficients	proportion	0.500	0.490	0.500	0.500	0.395	0.500	0.499	0.589	0.500
	mse	0.018	0.023	0.026	0.017	0.021	0.028	0.017	0.029	0.028
Estimated parameters										
β_A	coef	0.501	0.501	0.501	0.500	0.500	0.501	0.500	0.500	0.500
	mse	0.015	0.020	0.023	0.013	0.019	0.022	0.013	0.027	0.029
β_B	coef	0.601	0.604	0.604	0.600	0.604	0.605	0.599	0.599	0.600
	mse	0.013	0.021	0.024	0.014	0.027	0.029	0.015	0.023	0.025

Source: calculations from author.
Note: results from Monte Carlo simulations with 200 replications (N=10,000).

A difficulty with the Probit model is that the error variance is not identified (and thus normalized to one). This normalization imposes equality of error variances between the two groups in the decomposition, but the true error variances may very well be different. Unfortunately, the linear decomposition of the simulated latent variable will perform poorly in such situations. The problem is that knowing the binary outcomes Y_A and Y_B provides no information on the variances of the residuals of the latent variables. An assumption on error variances is needed to draw simulated residuals, and results from the decomposition of the simulated latent variable will clearly be conditional on the assumption made about the error terms[5].

For some non-linear models, data provides information on the variances of ε_A and ε_B. Consider for instance the case of a Tobit model. With the same latent outcome $Y_g^* - \bar{Y}^*$ as before, the dependent variable is now $Y_g = Y_g^* - \bar{Y}^*$ if $Y_g^* > \bar{Y}^*$ and $Y_g = 0$ if $Y_g^* \leq \bar{Y}^*$. Estimation of Tobit models respectively for Y_A and Y_B gives not only the coefficients $\hat{\beta}_A$ and $\hat{\beta}_B$, but also the error variances $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$. Application of simulated residuals thus requires to draw residuals such that $\varepsilon_A \sim N(0; \hat{\sigma}_A^2)$ and $\varepsilon_B \sim N(0; \hat{\sigma}_B^2)$. In additional simulations (not reported), I show that in that case the decomposition based on the latent simulated outcome performs very accurately with unequal error variances. When $\Omega = I$, I find a weight of 59.9% for the characteristics effect with $\hat{\sigma}_A^2 = 0.5$ and $\hat{\sigma}_B^2 = 1$ and of 60.1% with $\hat{\sigma}_A^2 = 1.5$ and $\hat{\sigma}_B^2 = 1$. These results are similar to those obtained from the

[5] In an appendix available upon request, I present results from Monte Carlo simulations with residuals ε_A and ε_B drawn from normal distribution with different variances. With unequal error variances, the contribution of the characteristics component when $\Omega = I$ increases with the variance of ε_A and ε_B (the reverse pattern is found when $\Omega = 0$). At the same time, it is interesting to think about the meaning of different error variances for ε_A and ε_B. If I consider for instance labor force participation by gender, an assumption is needed on the error variance of the unobserved propensity to work for men and women respectively. It seems very difficult to claim on a priori grounds that the error variance is higher either for men or women, so that the assumption of equal error variances is certainly not unreasonable.

Oaxaca-Blinder decomposition, while the characteristics effect is very different when implementing the non-linear Tobit decomposition (respectively 76% and 46.9%).

4. Conclusion

In this paper, I propose to decompose non-linear models deduced from a latent regression framework using the latent dependent outcome as dependent variable. Since the latent outcome is unobserved, I turn to simulated residuals and apply the standard Oaxaca-Blinder to the simulated latent variable. Results from a Monte Carlo experiment indicate that this procedure provides accurate estimates of the characteristics and coefficients effects. This combination of simulated residuals and linear decomposition is straightforward to implement and may be of interest to researchers interested in decomposing models with discrete or limited dependent variables.

References

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

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.

Gouriéroux, C., Monfort, A., Renault, E., Trognon, A., 1987. Simulated residuals. Journal of Econometrics 34, 201-252.

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

Oaxaca, R.L., 1973. Male-female wage differentials in urban labor markets. International Economic Review 14, 693-709.

Oaxaca, R.L., Ransom, M.R., 1994. On discrimination and the decomposition of wage differentials. Journal of Econometrics 61, 5-21.

Sinning, M., Hahn, M., Bauer, T.K., 2008. The Blinder–Oaxaca decomposition for nonlinear regression models. Stata Journal 8, 480-492.

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

Yun, M.-S., 2007. An extension of the Oaxaca decomposition using generalized residuals. Journal of Economic and Social Measurement 32, 15-22.

Appendix, not for publication

Part A. Decomposition of a Probit model, with different error variances

Setting:
$X_A \sim N(4;1)$, $X_B \sim N(5;1)$, $Y_A^* = 2 + 0.5 * X_A + \varepsilon_A$, $Y_B^* = 2 + 0.6 * X_B + \varepsilon_B$
Dependent variable: $Y_g = 1_{Y_g^* > \bar{Y}^*}$

Table A. Decomposition from Monte Carlo simulations - Probit model

Decomposition		$\varepsilon_A \sim N(0;0.5)$, $\varepsilon_B \sim N(0;1)$			$\varepsilon_A \sim N(0;1.5)$, $\varepsilon_B \sim N(0;1)$			$\varepsilon_A \sim N(0;0.5)$, $\varepsilon_B \sim N(0;1.5)$		
		Oaxaca	Probit	Simulated	Oaxaca	Probit	Simulated	Oaxaca	Probit	Simulated
$\Omega = I$										
Dif. in characteristics	proportion	0.598	0.468	0.400	0.601	0.691	0.724	0.599	0.383	0.301
	mse	0.019	0.017	0.018	0.022	0.032	0.039	0.028	0.021	0.020
Dif. in coefficients	proportion	0.402	0.532	0.600	0.399	0.309	0.276	0.401	0.617	0.699
	mse	0.019	0.017	0.018	0.022	0.032	0.039	0.028	0.021	0.020
$\Omega = 0$										
Dif. in characteristics	proportion	0.498	0.609	0.665	0.499	0.426	0.398	0.498	0.680	0.749
	mse	0.012	0.019	0.023	0.026	0.027	0.030	0.014	0.024	0.029
Dif. in coefficients	proportion	0.502	0.391	0.335	0.501	0.574	0.602	0.502	0.320	0.251
	mse	0.012	0.019	0.023	0.026	0.027	0.030	0.014	0.024	0.029
Estimated parameters										
β_A	coef	0.499	1.000	0.999	0.499	0.332	0.332	0.499	0.999	1.000
	mse	0.007	0.031	0.033	0.022	0.019	0.022	0.007	0.031	0.035
β_B	coef	0.599	0.600	0.601	0.602	0.603	0.603	0.600	0.401	0.401
	mse	0.015	0.020	0.022	0.014	0.022	0.025	0.023	0.020	0.022

Source: calculations from author
Note: results from Monte Carlo simulations with 200 replications (N=10,000).

Part B. Decomposition of a Tobit model, with different error variances

Setting:
$X_A \sim N(4;1)$, $X_B \sim N(5;1)$, $Y_A^* = 2 + 0.5 * X_A + \varepsilon_A$, $Y_B^* = 2 + 0.6 * X_B + \varepsilon_B$
Dependent variable: $Y_g = Y_g^* - \bar{Y}^*$ if $Y_g^* > \bar{Y}^*$ and $Y_g = 0$ if $Y_g^* \leq \bar{Y}^*$

Table B. Decomposition from Monte Carlo simulations - Tobit model

Decomposition		$\varepsilon_A \sim N(0;0.5)$, $\varepsilon_B \sim N(0;1)$			$\varepsilon_A \sim N(0;1.5)$, $\varepsilon_B \sim N(0;1)$			$\varepsilon_A \sim N(0;0.5)$, $\varepsilon_B \sim N(0;1.5)$		
		Oaxaca	Tobit	Simulated	Oaxaca	Tobit	Simulated	Oaxaca	Tobit	Simulated
$\Omega = I$										
Dif. in characteristics	proportion	0.600	0.750	0.599	0.602	0.469	0.601	0.599	0.800	0.601
	mse	0.018	0.011	0.026	0.022	0.036	0.030	0.027	0.011	0.037
Dif. in coefficients	proportion	0.400	0.250	0.401	0.398	0.531	0.399	0.401	0.200	0.399
	mse	0.018	0.011	0.026	0.022	0.036	0.030	0.027	0.011	0.037
$\Omega = 0$										
Dif. in characteristics	proportion	0.500	0.527	0.500	0.498	0.097	0.499	0.498	0.642	0.497
	mse	0.012	0.014	0.017	0.027	0.055	0.035	0.013	0.015	0.019
Dif. in coefficients	proportion	0.500	0.473	0.500	0.502	0.903	0.501	0.502	0.358	0.503
	mse	0.012	0.014	0.017	0.027	0.055	0.035	0.013	0.015	0.019
Estimated parameters										
β_A	coef	0.499	0.499	0.499	0.498	0.500	0.500	0.499	0.498	0.498
	mse	0.008	0.016	0.017	0.022	0.026	0.031	0.007	0.015	0.015
β_B	coef	0.599	0.599	0.598	0.602	0.602	0.602	0.600	0.600	0.602
	mse	0.016	0.017	0.021	0.014	0.016	0.020	0.022	0.025	0.030

Source: calculations from author
Note: results from Monte Carlo simulations with 200 replications (N=10,000).