Oaxaca-Blinder decomposition of changes in means and inequality: A simultaneous approach

Arthur Charpentier, Emmanuel Flachaire

Economics Bulletin

Volume 44, Issue 1

Oaxaca-Blinder decomposition of changes in means and inequality: A simultaneous approach

Arthur Charpentier
UQAM

Emmanuel Flachaire
Aix-Marseille University, AMSE

Abstract

In this paper, we show that a decomposition of changes in inequality, with the mean log deviation index, can be obtained directly from the Oaxaca-Blinder decompositions of changes in means of incomes and log-incomes. It allows practitioners to conduct simultaneously empirical analyses to explain which factors account for changes in means and in inequality indices between two distributions with strictly positive values.

Paper #4 in Special issue "In memory of Pr. Michel Terraza" This work was supported by the Natural Sciences and Engineering Research Council of Canada Grant NSERC-2019-07077, and by the French government under the "France 2030" investment plan managed by the French National Research Agency (reference: ANR-17-EURE-0020, ANR-17-CE41-0007, ANR-19-FRAL-0006) and from Excellence Initiative of Aix-Marseille University - A*MIDEX.
Citation: Arthur Charpentier and Emmanuel Flachaire, (2024) "Oaxaca-Blinder decomposition of changes in means and inequality: A simultaneous approach", *Economics Bulletin*, Volume 44, Issue 1, pages 308-320
Contact: Arthur Charpentier - charpentier.arthur@uqam.ca, Emmanuel Flachaire - emmanuel.flachaire@univ-amu.fr.
Submitted: May 24, 2023. **Published:** March 30, 2024.

Economics Bulletin

Picture credits: Virginie Terraza

Special issue “In memory of Professor Michel Terraza”

1 Introduction

The Oaxaca-Blinder decomposition is widely used in empirical studies to analyze the sources of changes in income distributions (Oaxaca 1973, Blinder 1973). It separates changes due to differences in characteristics between two groups, from other effects. Initially developped to explain the male-female wage difference in means (gender gap), this method has been extended to various distributional measures (DiNardo et al. 1996, Fortin et al. 2011).

This paper focuses on the Oaxaca-Blinder decomposition of changes in inequality indices. When comparing inequality between two years or two countries, difference in inequality may be mainly driven by differences in characteristics between two groups (age, education, etc.), while there would be no difference if individuals from the two groups share, on average, the same characteristics. With an Oaxaca-Blinder decomposition, we are then concerned by measuring the share of the difference in inequality which is not due to differences in characteristics between individuals in the two groups. Such decomposition is usually done with regression based on the Recentered Influence Function (RIF), when applied to inequality indices (Ferreira et al. 2017, Firpo et al. 2018, Rios-Avila 2020, Gradín 2020).

In this paper, we show that the use of RIF regression is not required to obtain a decomposition of changes in inequality indices. To decompose changes in inequality, we consider the mean log deviation (MLD) measure, also known as the second Theil index or Generalized Entropy index with parameter zero, which is defined for strictly positive values. The MLD index has an attractive property compared to other inequality indices. It is the only relative inequality measure which respects both the principle of transfer (a transfer from an individual to a poorer individual cannot increase inequality) and the principle of monotonicity in distance (when a rich get richer inequality cannot decrease), see Cowell and Flachaire (2018). Here, we show another attractive feature of the MLD index. An Oaxaca-Blinder decomposition of changes in MLD inequality indices can be obtained directly from the results of a standard decomposition of changes in means of incomes and log-incomes. In other words, standard Oaxaca-Blinder decomposition analysis can be used to simultaneously explain which factors account for changes in means and in inequality indices between two distributions.

In section 2, we present the standard Oaxaca-Blinder decomposition of changes in means. In section 3, we show that an Oaxaca-Blinder decomposition of changes in MLD indices can be derived from the results obtained on the decomposition of changes in means. Finally, an application is presented in section 4.

2 Decomposing difference in means

The Oaxaca-Blinder decomposition is mainly used to study changes in means between two groups A and B of individuals. Let us consider the following linear regression models:

$$\log y_A = X_A\beta_A + \varepsilon_A \tag{1}$$
$$\log y_B = X_B\beta_B + \varepsilon_B \tag{2}$$

where $\mathbb{E}(\varepsilon_A|X_A) = \mathbb{E}(\varepsilon_B|X_B) = 0$. From an OLS estimation, we have

$$\overline{\log y}_A = \bar{X}_A\hat{\beta}_A \tag{3}$$
$$\overline{\log y}_B = \bar{X}_B\hat{\beta}_B \tag{4}$$

where $\overline{\log y}_A$, $\overline{\log y}_B$, $\bar{X}_A$ and $\bar{X}_B$ are the means of the variables for observations in group A and group B, including a constant vector. By substracting these two equations, the difference of log-incomes means between group A and B is then defined as follows:

$$\overline{\log y}_A - \overline{\log y}_B = \bar{X}_A\hat{\beta}_A - \bar{X}_B\hat{\beta}_B \tag{5}$$

By adding and substracting $\bar{X}_A\hat{\beta}_B$, we obtain:

$$\Delta\mu = \overline{\log y}_A - \overline{\log y}_B = \underbrace{\bar{X}_A(\hat{\beta}_A - \hat{\beta}_B)}_{\Delta_\mu^S \text{ (unexplained)}} + \underbrace{(\bar{X}_A - \bar{X}_B)\hat{\beta}_B}_{\Delta_\mu^X \text{ (explained)}} \tag{6}$$

It is a standard Oaxaca-Blinder aggregated decomposition, into explained and unexplained effects:

- The explained effect reflects the differences in (average) covariates between A and B. For instance, it is different from zero when the average level of education is not the same in A and B.
- The unexplained effect reflects the differences of the "effects" of the covariates between A and B. For instance, education may explain income differences among individuals in A more strongly than in B.

The additive property of linear regression models allows us to decompose the explained and unexplained effects in greater details. Indeed, the contributions of the kth covariate to the explained and unexplained effects in (6) are, respectively, equal to:

$$\Delta_{\mu,k}^X = (\bar{X}_{A,k} - \bar{X}_{B,k})\hat{\beta}_{B,k} \tag{7}$$
$$\Delta_{\mu,k}^S = \bar{X}_{A,k}(\hat{\beta}_{A,k} - \hat{\beta}_{B,k}) \tag{8}$$

However, the detailed decomposition analysis of the unexplained effect produces arbitrary results. It is known to have an identification problem: the coefficients effect of sets of dummy variables and the intercept are sensitive to the choice of reference groups (Oaxaca and Ransom 1999, Yun 2005). One solution is to adjust the coefficients (Gardeazabal and Ugidos 2004).

The shares of the explained and unexplained effects are sensitive to the decomposition method. Indeed, another decomposition can be obtained by adding and substracting $\bar{X}_B\hat{\beta}_A$ in (5):

$$\Delta\mu = \overline{\log y}_A - \overline{\log y}_B = \underbrace{\bar{X}_B(\hat{\beta}_A - \hat{\beta}_B)}_{\Delta_\mu^S \text{ (unexplained)}} + \underbrace{(\bar{X}_A - \bar{X}_B)\hat{\beta}_A}_{\Delta_\mu^X \text{ (explained)}} \tag{9}$$

The difference comes from the reference group used in the decomposition:

- decomposition in (6) is based on $\bar{X}_A\hat{\beta}_B$, that is, what would be the log-income mean of individuals in A if they would be living in B,
- decomposition in (9) is based on $\bar{X}_B\hat{\beta}_A$, that is, what would be the log-income mean of individuals in B if they would be living in A.

The choice of the reference group is arbitrary and the decomposition can be viewed as arbitrary as well. There is no general guidance to this choice (Fortin et al. 2011). Nevertheless, the reference group has some economic meaning in the decomposition of changes in inequality. Indeed, the unexplained component in (6) can be interpreted as an average treatment effect on individuals in group A:

$$\Delta_\mu^S = \bar{X}_A(\hat{\beta}_A - \hat{\beta}_B) = \overline{\log y}_A - \bar{X}_A\hat{\beta}_B \tag{10}$$

The term $\bar{X}_A\hat{\beta}_B$ is a counterfactual, it represents the mean of log-incomes induced by individuals in A if they would be living in B. Similarly, the unexplained component in (9) can be interpreted as an average treatment effect on individuals in group B.

3 Decomposing difference in inequality indices

In this section, we show that the Oaxaca-Blinder decomposition of the difference of MLD inequality indices can be obtained from the decomposition of the difference of log-incomes means from the previous section.

From a sample of incomes $\{y_1, y_2, \ldots, y_n\}$, the MLD index is computed as follows:

$$\nu = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{y_i}{\bar{y}}\right) = \log\bar{y} - \overline{\log y} \tag{11}$$

It is the difference between the log of incomes mean and the mean of log-incomes. The difference of the MLD indices between individuals from groups A and B is then equal to:

$$\nu_A - \nu_B = \log(\bar{y}_A/\bar{y}_B) + \overline{\log y}_B - \overline{\log y}_A \tag{12}$$

In other words, the difference of the MLD inequality indices between two groups can be rewritten as a function of the difference of the mean of log-incomes defined in (6) and (9):

$$\Delta\nu = \log(\bar{y}_A/\bar{y}_B) - \Delta\mu \tag{13}$$

We need to further decompose $\log(\bar{y}_A/\bar{y}_B)$ into explained and unexplained components. Let us consider the following linear regression models:

$$y_A = X_A\delta_A + \eta_A \tag{14}$$
$$y_B = X_B\delta_B + \eta_B \tag{15}$$

where $\mathbb{E}(\eta_A|X_A) = \mathbb{E}(\eta_B|X_B) = 0$. With OLS estimators $\hat{\delta}_A$ and $\hat{\delta}_B$, and by adding and substracting $\bar{X}_A\hat{\delta}_B$, an Oaxaca-Blinder decomposition is given by

$$\Delta\xi = \overline{y}_A - \overline{y}_B = \underbrace{\bar{X}_A(\hat{\delta}_A - \hat{\delta}_B)}_{\Delta_\xi^S \text{ (unexplained)}} + \underbrace{(\bar{X}_A - \bar{X}_B)\hat{\delta}_B}_{\Delta_\xi^X \text{ (explained)}} \tag{16}$$

By adding and substracting $\log(\bar{X}_A\hat{\delta}_B)$ to $\log(\bar{y}_A/\bar{y}_B)$, we have

$$\log(\bar{y}_A/\bar{y}_B) = \log\bar{y}_A - \log(\bar{X}_A\hat{\delta}_B) + \log(\bar{X}_A\hat{\delta}_B) - \log\bar{y}_B \tag{17}$$
$$= \underbrace{\log(\bar{X}_A\hat{\delta}_A) - \log(\bar{X}_A\hat{\delta}_B)}_{\text{unexplained}} + \underbrace{\log(\bar{X}_A\hat{\delta}_B) - \log(\bar{X}_B\hat{\delta}_B)}_{\text{explained}} \tag{18}$$

From (6), (13) and (18), the aggregate decomposition of changes in MLD indices - into explained and unexplained effects - is equal to:

$$\Delta_\nu^X = \log(\bar{X}_A\hat{\delta}_B) - \log(\bar{X}_B\hat{\delta}_B) - \Delta_\mu^X \qquad \text{(explained)} \tag{19}$$
$$\Delta_\nu^S = \log(\bar{X}_A\hat{\delta}_A) - \log(\bar{X}_A\hat{\delta}_B) - \Delta_\mu^S \qquad \text{(unexplained)} \tag{20}$$

A detailed decomposition of the unexplained effect can be obtained by using the following approximation, which holds for small log differences:

$$\log(\bar{X}_A\hat{\delta}_A) - \log(\bar{X}_A\hat{\delta}_B) \approx \frac{\bar{X}_A(\hat{\delta}_A - \hat{\delta}_B)}{\bar{X}_A\hat{\delta}_B} = \frac{\Delta_\xi^S}{\bar{X}_A\hat{\delta}_B} \tag{21}$$

Hence, from (8), (20) and (21) we have

$$\Delta^S_{\nu,k} \approx \frac{\bar{X}_{A,k}(\hat{\delta}_{A,k} - \hat{\delta}_{B,k})}{\bar{X}_A\hat{\delta}_B} - \bar{X}_{A,k}(\hat{\beta}_{A,k} - \hat{\beta}_{B,k}) = \frac{\Delta^S_{\xi,k}}{\bar{X}_A\hat{\delta}_B} - \Delta^S_{\mu,k} \tag{22}$$

where $\Delta^S_{\nu,k}$, $\Delta^S_{\xi,k}$ and $\Delta^S_{\mu,k}$ are the contributions of the variable k to the unexplained effect of changes in, respectively, inequality, income means and log-income means. With a similar development and approximation, from (7) and (19), the contribution of the variable k to the explained effect of changes in inequality is given by

$$\Delta^X_{\nu,k} \approx \frac{(\bar{X}_{A,k} - \bar{X}_{B,k})\hat{\delta}_{B,k}}{\bar{X}_B\hat{\delta}_B} - (\bar{X}_{A,k} - \bar{X}_{B,k})\hat{\beta}_{B,k} = \frac{\Delta^X_{\xi,k}}{\bar{X}_B\hat{\delta}_B} - \Delta^X_{\mu,k} \tag{23}$$

The unexplained component (20) can be interpreted as an average treatment effect on individuals in group A:

$$\begin{aligned}
\Delta^S_\nu &= \log(\bar{X}_A\hat{\delta}_A) - \log(\bar{X}_A\hat{\delta}_B) - \bar{X}_A(\hat{\beta}_A - \hat{\beta}_B) && (24)\\
&= \log\bar{y}_A - \overline{\log y_A} - \log(\bar{X}_A\hat{\delta}_B) + \bar{X}_A\hat{\beta}_B && (25)\\
&= \nu_A - \log(\bar{X}_A\hat{\delta}_B) + \bar{X}_A\hat{\beta}_B && (26)
\end{aligned}$$

The term $\log(\bar{X}_A\hat{\delta}_B) - \bar{X}_A\hat{\beta}_B$ is a counterfactual, it represents the income inequality induced by individuals in A if they would be living in B.

Standard errors and confidence intervals can be computed by bootstraping, using the following procedure: [1] resample with replacement in the OLS residuals of (1) to form a vector of the same size and generate new values of log-incomes by adding it to (3) ; [2] resample with replacement in the OLS residuals of (2) to form a vector of the same size and generate new values of log-incomes by adding them to (4) ; [3] Perform a decomposition with the new values of log-incomes ; [4] Repeat the previous steps R times ; [5] The bootstrap standard error is the standard deviation of the R decomposition estimates, and the bootstrap confidence interval is obtained from the 0.025 and 0.975 sample quantiles of the R decomposition estimates.

4 Application

The data are obtained from the application in Hlavac (2018), on labor wages and demographic characteristics of 712 employed hispanic workers in the Chicago metropolitan area.

	group A	group B	difference	explained		unexplained	
			Δ	Δ^X	s.e.	Δ^S	s.e
μ ($\overline{\log y}$)	2.6967	2.5534	0.1434	0.0769	(0.0213)	0.0665	(0.0395)
ξ ($\overline{y}$)	17.5828	14.5672	3.0156	1.6165	(0.4076)	1.3991	(0.8021)
ν (MLD)	0.1702	0.1254	0.0448	0.0283	(0.0114)	0.0165	(0.0256)

Table 1: Aggregated decomposition of changes in means of log-wages, wages, and in MLD indices between native (A) and foreign-born (B) workers

4.1 Difference in means

Table 1 (first line) shows the aggregated decomposition of changes in means of log-wages between native (Group A) and foreign-born workers (Group B), with native workers taken as reference group. The difference in means between natives and foreign-born workers is equal to 0.1434, of which 0.0769 can be attributed to differences in characteristics (i.e. age, gender, education) and the remaining 0.0665 is unexplained.

Figure 1 shows the detailed decomposition of changes in means of log-wages for each variable, computed from (7) and (8), with 95% bootstrap confidence intervals (R = 10,000). In the explained components, all variables are statistically significant. It suggests that the explained part of the difference in means is driven by differences in characteristics between the two groups. In the unexplained components, only the variable age is clearly statistically significant. It suggests that the payoff of an additional year of age is greater for native workers.

Overall, this empirical analysis shows that the log-wage mean of native hispanic workers is greater than that of foreign-born hispanic workers. In addition, native workers have, on average, different characteristics and they enjoy greater returns to age. Similar conclusions can be obtained from the decomposition of changes in means of wages, rather than log-wages (middle line in Table 1 and Figure 2)

4.2 Difference in inequality indices

Table 1 (last line) shows the aggregated decomposition of changes in MLD inequality indices, obtained from (13)–(20). The results show that there is more inequality among native than foreign-born workers (0.1702 vs. 0.1254). The difference in inequality between the two groups is equal to 0.0448. It is decomposed into 0.0283, which can be attributed to differences in characteristics (i.e. age, gender, education), and the remaining 0.0165 which is

unexplained. These results suggest that if average characteristics of foreign-born and native workers were the same, the overall inequality difference between the two groups would be smaller (0.0165 vs. 0.0448).

Figure 3 shows the approximated detailed decomposition of changes in MLD indices for each variable, computed from (22) and (23), with 95% bootstrap confidence intervals. In the explained components, the variable *advanced degree* is statistically significant. It suggests that the explained part of the difference in inequality is partly driven by the proportion of highly educated individuals, which is larger in native hispanic workers. In the unexplained components, no variable is statistically significant.

Overall, this empirical analysis shows that inequality difference between native and foreign-born hispanic workers is mainly driven by more highly educated native hispanic workers than foreign-born workers.

5 Conclusion

In this paper, it is shown that a decomposition of difference in inequality indices can be obtained directly from standard Oaxaca-Blinder decomposition of difference in log-income and income means. An application illustrates how to investigate decomposition of differences in means and in inequality simultaneously.

References

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

Cowell, F. A. and E. Flachaire (2018). Inequality measures and the median: Why inequality increased more than we thought. PEP discussion paper, STICERD, LSE.

DiNardo, J., N. M. Fortin, and T. Lemieux (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, **64**, 1001–1044.

Ferreira, F. G., S. P. Firpo, and J. Messina (2017). Ageing poorly? Accounting for the decline in earnings inequality in Brazil, 1995-2012. World Bank Policy Research Working Paper 8018.

Firpo, S., N. M. Fortin, and T. Lemieux (2018). Decomposing wage distributions using recentered influence function regressions. *Econometrics*, **6**(2), 28.

Fortin, N., T. Lemieux, and S. Firpo (2011). Decomposition methods in economics. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume A, pp. 1–102. North-Holland: Elsevier.

Gardeazabal, J. and A. Ugidos (2004). More on identification in detailed wage decompositions. *Review of Economics and Statistics*, **4**(1034-1036), 86. F

Gradín, C. (2020). Quantifying the contribution of a subpopulation to inequality an application to Mozambique. *Journal of Economic Inequality*, **18**, 391419.

Hlavac, M. (2018). oaxaca: Blinder-Oaxaca decomposition in R. R package version 0.1.4. https://CRAN.R-project.org/package=oaxaca.

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

Oaxaca, R. L. and M. R. Ransom (1999). Identification in detailed wage decompositions. *Review of Economics and Statistics*, **81**, 154–157.

Rios-Avila, F. (2020). Recentered influence functions (RIFs) in Stata: RIF regression and RIF decomposition. *Stata Journal*, **20**, 5194.

Yun, M.-S. (2005). A simple solution to the identification problem in detailed wage decompositions. *Economic Inquiry*, **43**, 766–772.

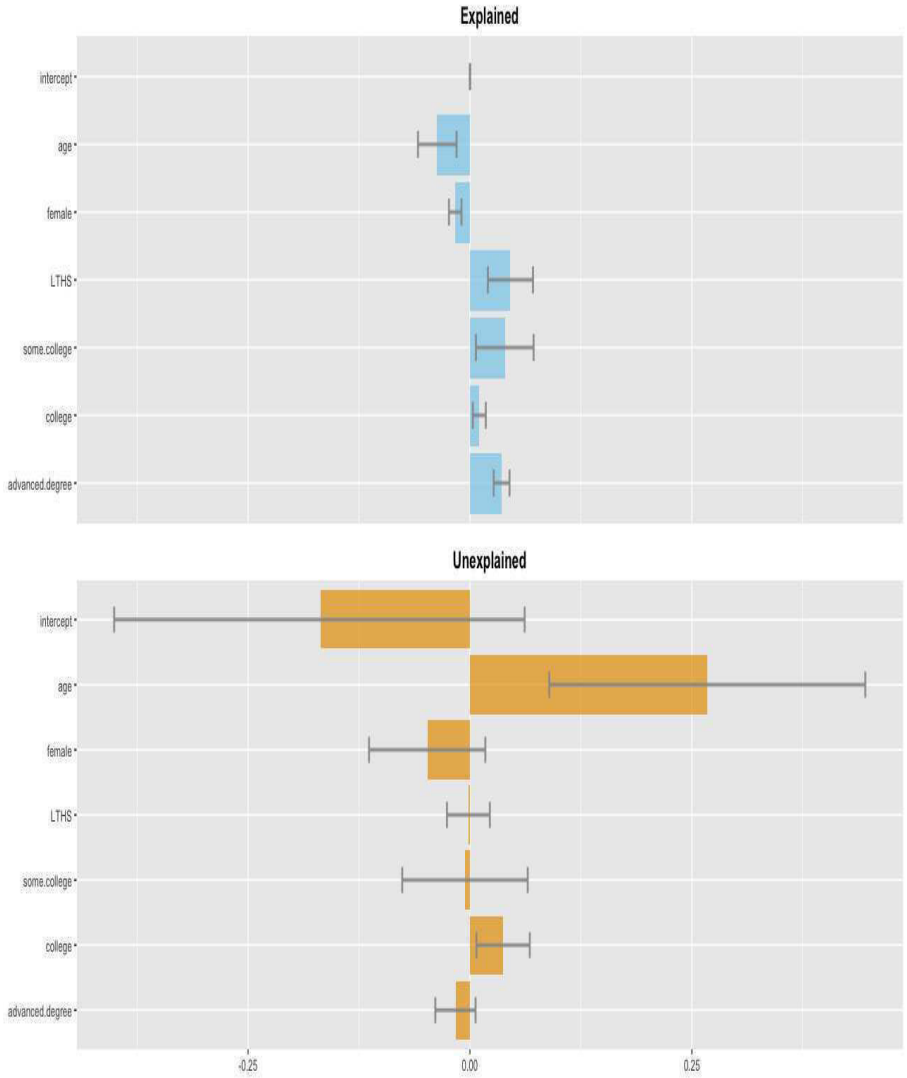

Figure 1: **Log-wage gap** - The detailed explained and unexplained components of a Oaxaca-Blinder decomposition on the native vs. foreign-born immigrant log-wage gap in means.

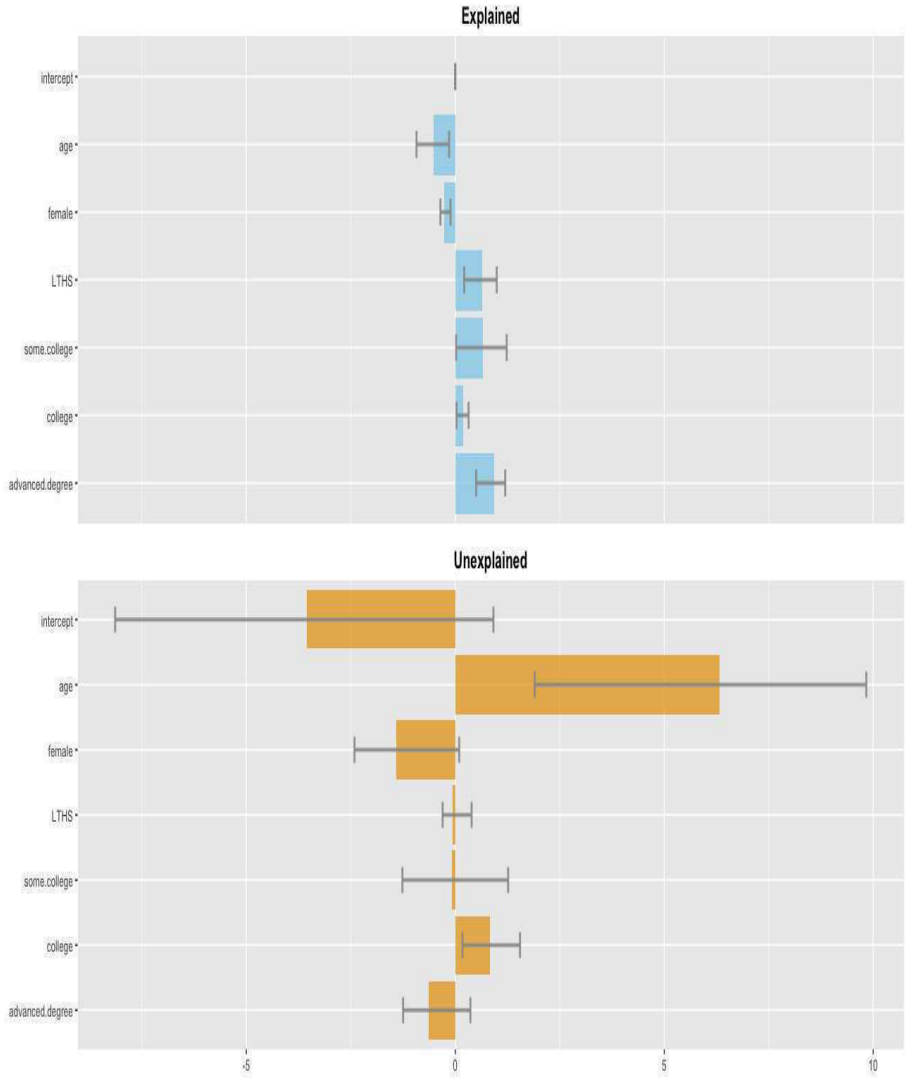

Figure 2: **Wage gap** - The detailed explained and unexplained components of a Oaxaca-Blinder decomposition on the native vs. foreign-born immigrant wage gap in means

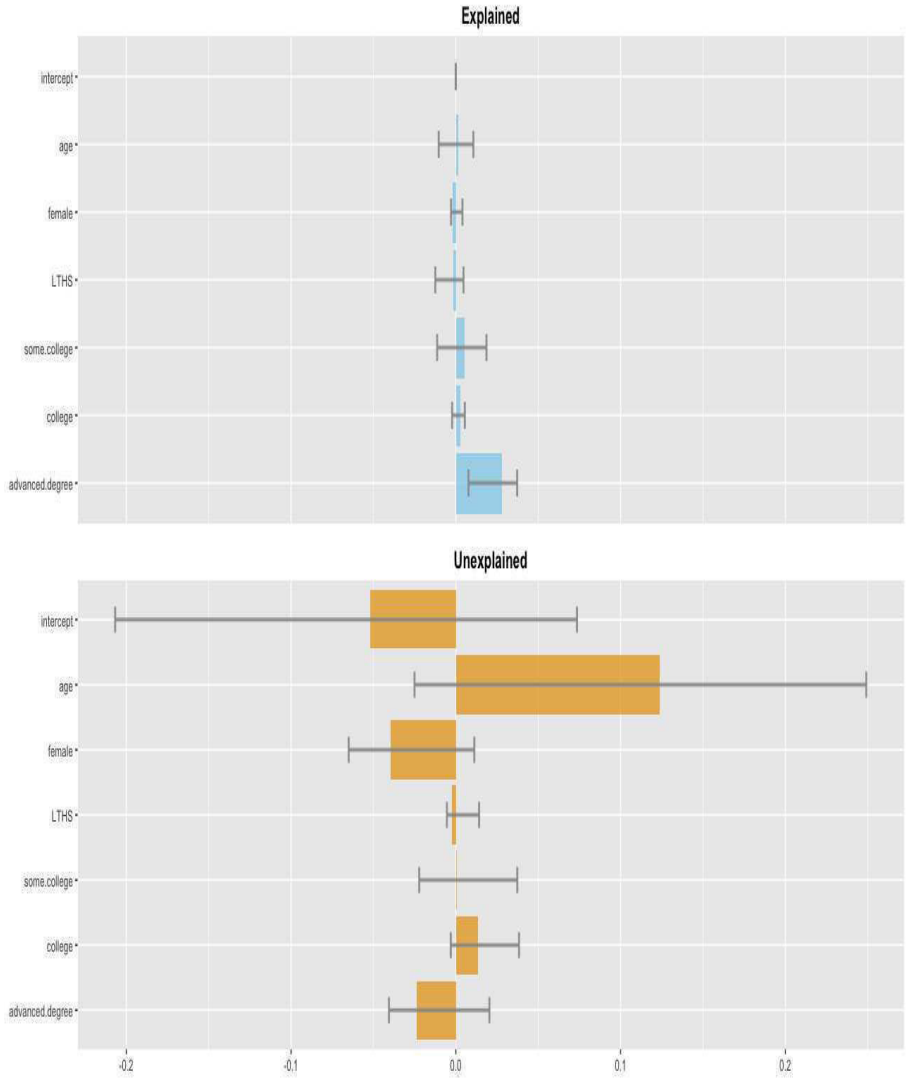

Figure 3: **MLD inequality gap** - The detailed explained and unexplained components of a Oaxaca-Blinder decomposition on the native vs. foreign-born immigrant gap in mean-logarithmic deviation inequality indices.