Rahimi and Hashemi Nazari
Emerg Themes Epidemiol (2021) 18:12
https://doi.org/10.1186/s12982-021-00100-9

Emerging Themes in Epidemiology

ANALYTIC PERSPECTIVE Open Access

A detailed explanation and graphical representation of the Blinder-Oaxaca decomposition method with its application in health inequalities

Ebrahim Rahimi[1] and Seyed Saeed Hashemi Nazari[2*]

Abstract

This paper introduces the Blinder-Oaxaca decomposition method to be applied in explaining inequality in health outcome across any two groups. In order to understand every aspect of the inequality, multiple regression model can be used in a way to decompose the inequality into contributing factors. The method can therefore be indicated to what extent of the difference in mean predicted outcome between two groups is due to differences in the levels of observable characteristics (acceptable and fair). Assuming the identical characteristics in the two groups, the remaining inequality can be due to differential effects of the characteristics, maybe discrimination, and unobserved factors that not included in the model. Thus, using the decomposition methods can identify the contribution of each particular factor in moderating the current inequality. Accordingly, more detailed information can be provided for policy-makers, especially concerning modifiable factors. The method is theoretically described in detail and schematically presented. In the following, some criticisms of the model are reviewed, and several statistical commands are represented for performing the method, as well. Furthermore, the application of it in the health inequality with an applied example is presented.

Keywords: Health Inequality, Decomposition methods, Contributing factors

Introduction

Inequality is one of the most obvious facts perceived in human life, referring to differences affecting the individual way of life. In spite of this simple meaning, inequality involves complexities hindering a consensus on its definition. Thus, inequality has been subject to numerous research projects. This concept is generally defined according to different needs and conditions of individuals. It is therefore related to the conditions and characteristics of recipients rather than providers of special services [1].

In some cases, inequality is used interchangeably with inequity. Any judgment to what extent inequality should be considered inequity will depend on the unfairness [2].

Discrimination, alongside the sense of inequity, refers to a situation in which belonging to a particular group sets the ground for preference or non-preference of that group [3]. In spite of identical capabilities and characteristics, the individuals in each group receive different benefits and services, related to the specific position of that group compared to others.

In the realm of public healthcare, inequality is a term referring to differences, dissimilarities and disparities visible in the health status of individuals or groups, which

*Correspondence: saeedh_1999@yahoo.com
[2] Prevention of Cardiovascular Disease Research Center, Department of Epidemiology, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Velenjak St., Chamran Highway, Tehran, Iran
Full list of author information is available at the end of the article

is indirectly applied as a tool to measure health inequity. In other words, one instance of inequity is the systematic and avoidable inequalities in groups with identical characteristics [4].

Despite the overall improvement in global health and hygiene, inequality has escalated over recent centuries. This issue can be mitigated inevitably through identification of the contributing factors. However, achievement of equity in the health sector is known as a major challenge facing the relevant authorities (2, 5). Although a great deal of investigation has been conducted on disparities in the health sector and medical sciences, very little research has focused on how to reduce it [5]. The first step in the formulation, design and implementation of effective interventions to reduce health inequalities is an investigation into the contributing factors and causes [6].

To that end, and in order to understand every aspect of inequality, multiple regression model can be used in a way to decompose inequality into its components. In 1973, Blinder [7] and Oaxaca [8] proposed a new method to examine the factors associated with racial/gender wage inequality and discrimination in the labor market. This method can be applied to explain inequalities in health outcome across any two groups.

The aim of the method is to explain how much of the difference in mean outcome between two groups is due to group differences in the levels of observed characteristics (acceptable and fair) and how much is due to discrimination, but may also result from the differential effect of the observed characteristics (group difference in the magnitude of regression coefficients) as well as other unknown associated factors. In fact, existence of inequality despite identical individual characteristics can be rooted in unknown factors that affect the outcomes. Thus, the difference in the outcome can be adjusted by mitigating the difference in the level of associated factors across comparison groups. The rest will concern unmeasurable, unobserved factors [9, 10]. Therefore, this method can be employed to identify the contribution of each factor into inequality.

Main text

Blinder-Oaxaca (B-O) decomposition method

Sometimes it is essential to decompose the mean difference in a specific continuous outcome between 2 groups (Group 1 and Group 2) to determine the factors contributing to that difference. For this purpose, multiple regression model can be employed. In terms of statistical measures, this particular decomposition method can be considered a combination of t-test and multiple regression models. Assuming that the outcome value (Y) is explained by K variables ($x_1, \ldots. x_k$) in the linear regression model, the mean predicted outcome for group g (1 and 2) can be expressed as follows:

$$\overline{Y}^g = \beta_0^g + \sum_{j=1}^{k} \beta_j^g \overline{x}_j^g$$

where $\overline{x}_j$ is the mean value of each predictor and β is the estimated regression coefficient.

Thus, the mean difference in outcome between the 2 groups (1 and 2) is as follows:

$$\Delta\overline{Y} = \left(\beta_0^1 - \beta_0^2\right) + \sum_{j=1}^{k} (\beta_j^1 \overline{x}_j^1 - \beta_j^2 \overline{x}_j^2) \quad (1)$$

The mean difference of outcome is the sum of the effects of different components, including: (1) Average difference between the level of each observable variable (x_j); (2) differential effects (β_j) of these variables in the 2 comparison groups, and (3) basic difference which includes the effect of unknown variables that are not included in the model. One question worth asking is “How large is the contribution of each of model components to this difference?".

To answer this question, the levels of explanatory variables and regression coefficients in the two groups are alternately assumed identical to achieve the net effect of each component. In fact, a counterfactual approach is adopted to replace the coefficients and the variables levels of the equation for one group to corresponding values for the other group (reference). Accordingly, the expected change in a group mean outcome is obtained when this group gets the predictor values and regression coefficients of the reference. In this procedure, the contribution of each component can be estimated [9, 11].

If Group 1 (or its outcome) is selected as the reference, the expected change in predictors level and regression coefficients of Group 2 and subsequently the change in its outcome will be considered.

The equation exclusive to Group 1 can be reformulated from the perspective of Group 2 as follows:

$$\begin{aligned}
\overline{Y}^1 &= \beta_0^1 + \sum_{j=1}^{k} \beta_j^1 \overline{x}_j^1 \\
&= \beta_0^1 + \sum_{j=1}^{k} \left[\beta_j^2 + (\beta_j^1 - \beta_j^2\right)] \\
&\sum_{j=1}^{k} \left[\overline{x}_j^2 + \left(\overline{x}_j^1 - \overline{x}_j^2\right)\right] \\
&= \beta_0^1 + \sum_{j=1}^{k} \beta_j^2 \overline{x}_j^2 + \sum_{j=1}^{k} \beta_j^2 \left(\overline{x}_j^1 - \overline{x}_j^2\right) \\
&+ \sum_{j=1}^{k} \overline{x}_j^2 (\beta_j^1 - \beta_j^2) + \sum_{j=1}^{k} \left(\overline{x}_j^1 - \overline{x}_j^2\right)\left(\beta_j^1 - \beta_j^2\right)
\end{aligned}$$

The above equation involves $\beta_j^1 = \beta_j^2 + (\beta_j^1 - \beta_j^2)$ and $\overline{x}_j^1 = \overline{x}_j^2 + \left(\overline{x}_j^1 - \overline{x}_j^2\right)$, which can be replaced in Eq. 1 to decompose the mean difference in outcome into 4 components as follows:

$$\Delta\overline{Y} = \underbrace{(\beta_0^1 - \beta_0^2)}_{B} + \underbrace{\sum_{j=1}^{k} \beta_j^2(\bar{x}_j^1 - \bar{x}_j^2)}_{E} + \underbrace{\sum_{j=1}^{k} \bar{x}_j^2\left(\beta_j^1 - \beta_j^2\right)}_{C} + \underbrace{\sum_{j=1}^{k} (\bar{x}_j^1 - \bar{x}_j^2)\left(\beta_j^1 - \beta_j^2\right)}_{I} \tag{2}$$

The decomposition shown in this equation is formulated from the perspective of Group 2, when Group 1 is selected as the reference.

Accordingly, the predicted difference (D) can be decomposed into 4 components (B, E, C and I); in other words, the contribution of each component in the difference can be estimated:

1. The first component (B) is attributed to basic differences. It includes the effects of unobservable variables not taken into account (i.e. not included in the model).
2. The second component (E) indicates change in group 2's mean predicted outcome when it meets the group 1 (Reference)'s covariates level:

$$\sum_{j=1}^{k} \beta_j^2(\overline{x}_j^1 - \overline{x}_j^2) = \sum_{j=1}^{K} \left(\beta_j^2\overline{x}_j^1 - \beta_j^2\overline{x}_j^2\right)$$

   In other words, a portion of the difference (D) that explained by group differences in the level of observable explanatory variables (explained component). This portion is known as "endowments effect".
3. The third component(C) is a part of the difference represents a change in group 2's mean predicted outcome when that group meets the regression coefficients of the other group:

$$\sum_{j=1}^{k} \overline{x}_j^2\left(\beta_j^1 - \beta_j^2\right) = \sum_{j=1}^{K} \left(\beta_j^1\overline{x}_j^2 - \beta_j^2\overline{x}_j^2\right)$$

   It involves a portion of the difference (D) caused by the differential effect of the observable variables on outcome across the 2 comparison groups. It cannot be explained by the level of observable explanatory variables (unexplained component). This portion of the difference is known as "coefficients effect".
4. The fourth component (I) involves an interaction due to simultaneous effect of differences in endowments and coefficients [11, 12].

Figure 1 schematically displays the decomposition of group difference in mean predicted outcome from the perspective of Group 2, when Group 1 has been considered a reference (Eq. 2)

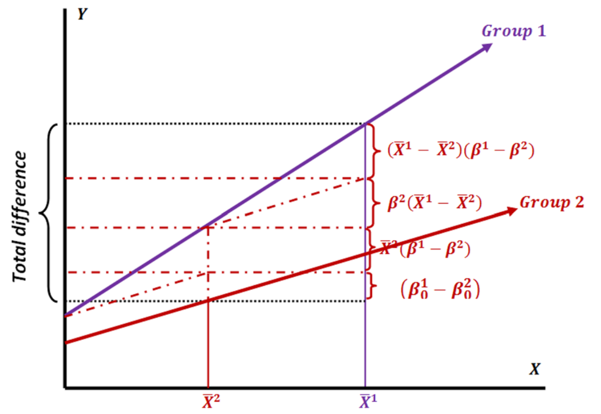

Y
Group 1
$(\overline{X}^1 - \overline{X}^2)(\beta^1 - \beta^2)$
$\beta^2(\overline{X}^1 - \overline{X}^2)$
Group 2
$\overline{X}^2(\beta^1 - \beta^2)$
$(\beta_0^1 - \beta_0^2)$
Total difference
X
$\overline{X}^2$
$\overline{X}^1$

Fig. 1 Decomposition of the group difference in mean predicted outcome (the interaction model) by selecting Group 1 as the reference (from the perspective of Group 2)

Similarly, when Group 2 is selected as the reference, the expected change in mean predicted outcome can be expressed from the perspective of Group 1 as follows:

$$\Delta\bar{Y} = (\beta_0^1 - \beta_0^2) + \sum_{j=1}^{k} \beta_j^1(\bar{x}_j^1 - \bar{x}_j^2) + \sum_{j=1}^{k} \bar{x}_j^1\,(\beta_j^1 - \beta_j^2) - \sum_{j=1}^{k}(\bar{x}_j^1 - \bar{x}_j^2)(\beta_j^1 - \beta_j^2) \quad (3)$$

D B E C I

To achieve the above Eq. 3, the covariates level and regression coefficients of Group 2 are reformulated from the perspective of Group 1 as follows

$$\beta_j^2 = \beta_j^1 - \left(\beta_j^1 - \beta_j^2\right)$$

$$\bar{x}_j^2 = \bar{x}_j^1 - \left(\bar{x}_j^1 - \bar{x}_j^2\right)$$

And then replace their corresponding values in the Eq. 1. Thus, the difference (D) in Eq. 3 can be decomposed into 4 partitions (B, E, C and I): The first component (B) and fourth component (I) are related to the same factors expressed in the following Eq. 2. The second component (E), however, measures expected change in group 1's mean predicted outcome if this group has the group 2(Reference)'s covariates level:

$$\sum_{j=1}^{k} \beta_j^1(\bar{x}_j^1 - \bar{x}_j^2) = \sum_{j=1}^{K}\left(\beta_j^1\bar{x}_j^1 - \beta_j^1\bar{x}_j^2\right)$$

The third component (C) is similarly a part of difference measures the expected change in group 1's mean predicted outcome when this group meets the regression coefficients of Group 2:

$$\sum_{j=1}^{k} \bar{x}_j^1\left(\beta_j^1 - \beta_j^2\right) = \sum_{j=1}^{K}\left(\beta_j^1\bar{x}_j^1 - \beta_j^2\bar{x}_j^1\right)$$

Figure 2 schematically depicts the decomposition conditions where Group 2 has been selected as a reference.

In Eqs. 2 and 3 (Figs. 1 and 2) the first component (B), $\left(\beta_0^1 - \beta_0^2\right)$, denotes to the differences between two groups that cannot be explained by the observed covariates (X). In fact, this difference is due to unobserved variables. On the other hand, the coefficient component (part C) is also unexplained by those differences. Then, we can combine these two components (B and C) into unexplained part (U), yielding the three-fold decomposition,

$$\Delta\overline{Y} = \sum_{j=1}^{k} \beta_j^2(\bar{x}_j^1 - \bar{x}_j^2) + \sum_{j=1}^{k} \bar{x}_j^2\left(\beta_j^1 - \beta_j^2\right) + \sum_{j=1}^{k}(\bar{x}_j^1 - \bar{x}_j^2)\left(\beta_j^1 - \beta_j^2\right) \quad (4)$$

$$\Delta\bar{Y} = \sum_{j=1}^{k} \beta_j^1(\bar{x}_j^1 - \bar{x}_j^2) + \sum_{j=1}^{k} \bar{x}_j^1\,(\beta_j^1 - \beta_j^2) - \sum_{j=1}^{k}(\bar{x}_j^1 - \bar{x}_j^2)(\beta_j^1 - \beta_j^2) \quad (5)$$

D = E U I

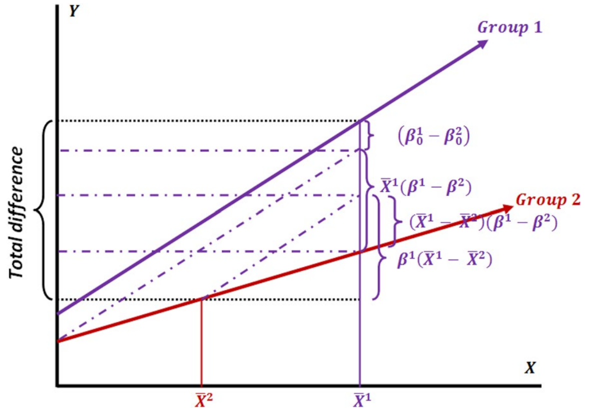

Y
Group 1
$(\beta_0^1 - \beta_0^2)$
$\bar{X}^1(\beta^1 - \beta^2)$
Group 2
$(\bar{X}^1 - \bar{X}^2)(\beta^1 - \beta^2)$
$\beta^1(\bar{X}^1 - \bar{X}^2)$
Total difference
X
$\bar{X}^2$
$\bar{X}^1$

Fig. 2 Decomposition of difference (the interaction model) by selecting Group 2 as the reference (from the perspective of Group 1)

In other words, if we assume that there are no relevant unobservable explanatory variables, the total unexplained part (U) in the Eqs. 4 and 5 will be equal to the C component in the Eqs. 2 and 3.

In this approach, the difference in mean predicted outcome (D) contains three components (E, U and I): The first component (E) is explained by the difference in the level of the covariates, the second component (U) arises from the differential effect of all those covariates (the unexplained part), and the third component (I) involves an interaction caused by the simultaneous group differences in the covariates level and their coefficients.

Up to now we had postulated that one of the groups 1 or 2 has the best achievable outcome and the other group should reach to this outcome. Another approach is that we suppose that there is a nondiscriminatory condition (marked by a nondiscriminatory vector of coefficients) that both groups should reach to this condition. Therefore, this approach requires the definition of nondiscriminatory conditions or reference coefficients. Sometimes even this nondiscriminatory condition can be the situation of one of the comparison groups (which we can call it reference coefficient).

Suppose β^* is the nondiscriminatory condition or reference coefficient, the overall equation for decomposition of $\Delta\overline{Y}$ will be:.

$$\Delta\bar{Y} = \sum_{j=1}^{k} \beta_j^*(\bar{x}_j^1 - \bar{x}_j^2) + [\sum_{j=1}^{k} \bar{x}_j^1 (\beta_j^1 - \beta_j^*) + \sum_{j=1}^{k} \bar{x}_j^2 (\beta_j^* - \beta_j^2)] \quad (6)$$

Endowments effect   Discrimination effect

In this way the outcome difference has been decomposed into two components (two-fold decomposition). The first component is the part of the group difference that is explained by the differences in the levels of observed characteristics. This is also called “endowments effect”. The second component refers to the part of the gap that is due to differences of β s with the non-discriminating β*. It also catches differences in level of unobservable variables and also their differential (discriminating) effects. This component determines the unexplained portion of the disparity. If all the unobserved covariates were in the model and measured, it comprised just the difference of β *s* with the non-discriminating β*. This portion is sometimes considered as “discrimination effects”.

β* is always between β^1 and β^2 , or equal to both or one of them. Then, we have $\beta^1 \geq \beta^* \geq \beta^2$ or $\beta^1 \leq \beta^* \leq \beta^2$.

If $\beta^1 > \beta^* > \beta^2$, we have positive discrimination "in favor of" group 1 and negative discrimination “against" group 2 and if $\beta^1 < \beta^* < \beta^2$, then we have positive discrimination in "in favor of" group 2 and negative discrimination "against" Group 1.

There is also a case that one of the two groups experiences discrimination and the non-discriminating β* will simply be the coefficients from the other group. In such case, $\beta^1 = \beta^* > \beta^2$ or $\beta^1 > \beta^* = \beta^2$. If we replace β* with β^1 in the Eq. 6, we reach to the Eq. 7 and if we replace β* with β^2 we reach to the Eq. 8

$$\Delta\overline{Y} = \sum_{j=1}^{k} \beta_j^1(\bar{x}_j^1 - \bar{x}_j^2) + \sum_{j=1}^{k} \bar{x}_j^2\left(\beta_j^1 - \beta_j^2\right) \quad (7)$$

$$\Delta\bar{Y} = \sum_{j=1}^{k} \beta_j^2(\bar{x}_j^1 - \bar{x}_j^2) + \sum_{j=1}^{k} \bar{x}_j^1 (\beta_j^1 - \beta_j^2) \quad (8)$$

D = Ec + Uc

.

Therefore, we have a twofold decomposition of the difference in mean predicted outcome (D):

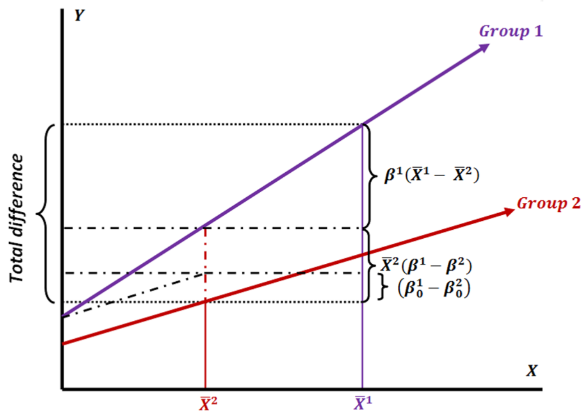


Fig. 3 Decomposition of outcome difference using the group 1 coefficients as the reference

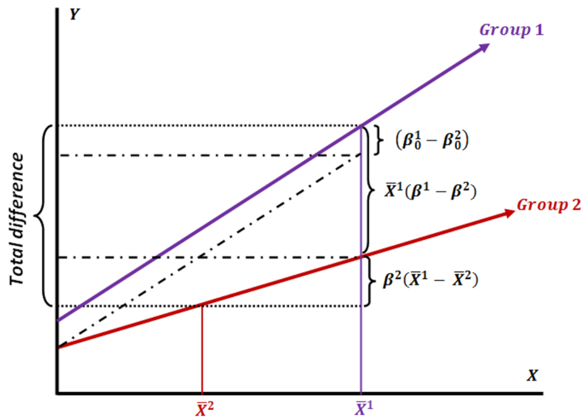


Fig. 4 Decomposition of outcome difference using the group 2 coefficients as the reference

1. The Unexplained component (Uc): This is exactly similar to the U part of the three-fold decomposition (Eq. 4 and 5). It arises from the differential effect of observable variables and also differential effect (β) and level of unobservable variables. This determines the unexplained portion of the gap.
2. The Explained component (Ec): This part is the combination of E and I parts of the three-fold decomposition (Eqs. 4 and 5). Although this component is called the explained component in two-fold decomposition in many texts but some part of it (the interaction part) is in fact the simultaneous difference of coefficients and covariates level in both groups.

   Hence, if somebody wants the crude explained component, three folds' decomposition can provide this crude explained part [11, 13].

Therefore, Eqs. 7 and 8 can be considered a specific form of Eqs. 4 and 5, where components *E* and *I* have been integrated. Thus:

$$\sum_{j=1}^{k}\beta_j^2(\overline{x}_j^1-\overline{x}_j^2)+\sum_{j=1}^{k}(\overline{x}_j^1-\overline{x}_j^2)\left(\beta_j^1-\beta_j^2\right)=\sum_{j=1}^{k}\beta_j^1(\overline{x}_j^1-\overline{x}_j^2)$$

and

$$\sum_{j=1}^{k}\beta_j^1(\overline{x}_j^1-\overline{x}_j^2)-\sum_{j=1}^{k}(\overline{x}_j^1-\overline{x}_j^2)\left(\beta_j^1-\beta_j^2\right)=\sum_{j=1}^{k}\beta_j^2(\overline{x}_j^1-\overline{x}_j^2)$$

Figures 3 and 4 schematically demonstrate the decomposition conditions where the group 1 and group 2 coefficients has been selected as a reference, respectively.

It is not clear which regression coefficient should be selected as a reference (Eqs. 7and 8). This is known as "index problem" [9, 14–18]. Reimers [19] suggests using the average regression coefficients over both groups $(\frac{\beta_i^1+\beta_i^2}{2})$, while Cotton [15] expresses the sum of coefficients weighted by each group size $(\frac{n^1\beta_\mathrm{i}^1+n^2\beta_\mathrm{i}^2}{\mathrm{N}})$ for β_j^*. In this regard, Neumark proposes the use of regression coefficients from a pooled model over both groups as an estimate for nondiscriminatory conditions [16, 17].

Nonlinear extension of B-O decomposition method

Although the primary application of the proposed B-O decomposition is based on the linear regression model, several researchers, including Yun and Fairlie, have proposed a nonlinear version of decomposition [14, 20–22], which has been widely used in the decomposition of inequalities in the health sector [23–27].

As mentioned, the original B-O decomposition of the 2-group disparity in the average value of the response variable, Y, can be expressed as:

$$\Delta\overline{Y}=[\beta^1\left(\overline{X}^1-\overline{X}^2\right)]+[\overline{X}^2\left(\beta^1-\beta^2\right)]$$

where $\overline{X}$ is a row vector of average values of the explanatory variables and β is a vector of coefficient estimates for each group 1 and 2. In this case, the coefficient estimates of group 1, β^1, have been assumed to be as the reference. According to Fairlie[28], the decomposition for a nonlinear equation, $Y=F(X\beta)$ can be expressed as follows:

$$\Delta\bar{Y} = \underset{\substack{\blacktriangle \\ \textit{Explained Part}}}{\left[\frac{1}{N^1}\sum_{i=1}^{N^1} F(\beta^1 X_i^1) - \frac{1}{N^2}\sum_{i=1}^{N^2} F(\beta^1 X_i^2)\right]} + \underset{\substack{\blacktriangle \\ \textit{Unexplained Part}}}{\left[\frac{1}{N^2}\sum_{i=1}^{N^2} F(\beta^1 X_i^2) - \frac{1}{N^2}\sum_{i=1}^{N^2} F(\beta^2 X_i^2)\right]} \tag{9}$$

where N^g is the sample size for group g (1 or 2), and $\Delta\overline{\mathrm{Y}}$ represents the difference in "mean predicted probability of outcome" between two groups with N^1 and N^2 individuals. This alternative expression for the decomposition is used because in non-linear transformations of Y, $\overline{Y}$ does not necessarily equal $F\left(\overline{X}\beta\right)$. The original B-O decomposition is a special case of Eq. 9 in which $F(X_i\beta) = X_i\beta$ [28].

Similarly, another expression for the decomposition is:

$$\Delta\overline{Y} = \left[\frac{1}{N^1}\sum_{i=1}^{N^1} F\left(\beta^2 X_i^1\right) - \frac{1}{N^2}\sum_{i=1}^{N^2} F\left(\beta^2 X_i^2\right)\right] + \left[\frac{1}{N^1}\sum_{i=1}^{N^1} F\left(\beta^1 X_i^1\right) - \frac{1}{N^1}\sum_{i=1}^{N^1} F\left(\beta^2 X_i^1\right)\right] \tag{10}$$

where the vector of coefficient estimates for group 2 is used as the reference.

Detailed decomposition

In the detailed decomposition one can determine the relative contribution of each factor (X variables) to each one of explained and unexplained components. This can be achieved by sequentially substituting variables levels/coefficients of one group with those of another group while keeping other variables in the model constant.

Using linear regression based decomposition; the detailed decomposition is not a complicated task because each component is obtained simply by summing over the contribution of each predictor to the each component. In nonlinear method, however, performing the detailed decomposition is not as straightforward. In other words, the application of the original (linear) method to nonlinear decomposition models has some conceptual problems that affect the results [28–31]. The first problem is known as "identification problem", that is, for nominal (categorical) variables, as the predictors, the decomposition estimates depend on the choice of the base (omitted) category. One solution, proposed by Yun [30], is computing normalized effects. It is equivalent to averaging the coefficients effects of a set of dummy variables while changing the reference groups.

Another problem is "path dependency". Unlike linear models, nonlinear decomposition is sensitive also to the order of variables being included into the decomposition process (path dependency) [22, 28, 29, 32]. One solution to this issue has been suggested by Fairlie, which involves randomly ordering the variables across replications of the decomposition. This procedure requires one-to-one matching of individuals from the 2 comparing groups, thus there should be equal number of individuals in both groups $\left(\mathrm{N}^1 = \mathrm{N}^2\right)$. Otherwise (which is usually the case), a random subsample of the majority group 1 (which is usually equal to the sample size of minority group 2) will be selected and then matched according to the predicted probability for the response variable of each person. In fact, the individual observations in each group will be separately arranged based on predicted probability and then matched according to ranking. This procedure will match the individual characteristics in both groups. Thus, the matched observations (one-to-one) will determine the contribution of each factor to the outcome difference. Thus, the multiple sub-samples (e.g. 100 or 1000 times) are selected and the mean estimate is considered as the final estimate [14, 24, 28, 33]. Using logit coefficient estimates (β^*) from pooled sample or from the appropriate reference group, the independent contribution of $\mathrm{X_J}$ to the gap can be expressed as:

$$\frac{1}{N^2}\sum_{i=1}^{N^2} F\left(X_{ji}^1\beta_j^* + \sum_{J\neq j} X_{ji}^1\beta_j^*\right) - F\left(X_{ji}^2\beta_j^* + \sum_{J\neq j} X_{ji}^1\beta_j^*\right)$$

or

$$\frac{1}{N^2}\sum_{i=1}^{N^2} F\left(X_{ji}^2\beta_j^* + \sum_{J\neq j} X_{Ji}^1\beta_J^*\right) - F\left(X_{ji}^2\beta_j^* + \sum_{J\neq j} X_{ji}^2\beta_j^*\right)$$

One simpler strategy to overcome this issue involves using weights as well [22, 34, 35]. According to Yun [22], detailed decomposition using weights can be expressed as follows:

$$\overline{Y}^1 - \overline{Y}^2 = \sum_{k=1}^{k} W^{\Delta}_{X_k}\left[\overline{F(x^1\beta^1)} - \overline{F(x^2\beta^1)}\right] + \sum_{k=1}^{k} W^{\Delta}_{\beta_k}\left[\overline{F(x^2\beta^1)} - \overline{F(x^2\beta^2)}\right] \tag{11}$$

$$\text{If } \sum_{k=1}^{k} W^{\Delta}_{X_k} = \sum_{k=1}^{k} W^{\Delta}_{X_k} = 1$$

where $W^{\Delta}_{X_k}$ and $W^{\Delta}_{\beta_k}$ represent the weight of *K*th variable in the linearization of the explained and unexplained components of inequality, respectively [22, 32]:

$$W^{\Delta}_{X_k} = \frac{\beta^1_k\left(\overline{X}^1_k - \overline{X}^2_k\right)}{\sum_{k=1}^{k} \beta^1_k\left(\overline{X}^1_k - \overline{X}^2_k\right)}$$

$$W^{\Delta}_{\beta_k} = \frac{\overline{X}^1_k\left(\beta^1_k - \beta^2_k\right)}{\sum_{k=1}^{k} \overline{X}^1_k\left(\beta^1_k - \beta^2_k\right)}$$

The Fairlie method mainly focuses on the explained portion of inequality without calculating the contribution of the differential effect from each factor to the unexplained part [14]. Nonetheless, that can be achieved through the practical technique proposed by Power et al. [32].

Implementation of the decomposition in related Software

The Oaxaca command package is available for Stata [9], R [36] and SAS Macro% BO_decomp [37] to perform the blinder-oaxaca decomposition. In addition, Stata provides several packages developed for implementation of various forms of Blinder-Oaxaca decomposition into non-linear models, including fairlie [38], gdecomp [39], mvdcmp [32], and nldecompose [40] (Table 1).

Conclusions

Drawbacks of the B-O decomposition models

The B-O decomposition methods have been subject to some criticisms that mostly focus on the model specification and the selection of the explanatory variables for the model [11, 30, 41].

The decomposition does not consider the different distribution of outcome among the individuals of each group [3]. It provides only information about the difference in mean predicted outcome between the 2 groups which is different from crude difference to the extent that distribution of other covariates between two groups are different.

The decomposition estimates also vary depending on the choice of reference group. There is often no compelling reason to choose the best group.

The 2 groups are not comparable due to unknown factors, putting the estimates subject to the effect of selection bias. In addition, the measurement errors that are systematically different for the groups can distort the results. Thus, there will be inefficiencies in the estimation of coefficients and consequently the "unexplained” component [3]. Additionally, if there is some misspecification in the outcome model and the average of residuals of the outcome model is not zero, there will be also another unexplained part which is the difference between among difference of mean observed outcomes and difference of mean predicted outcomes.

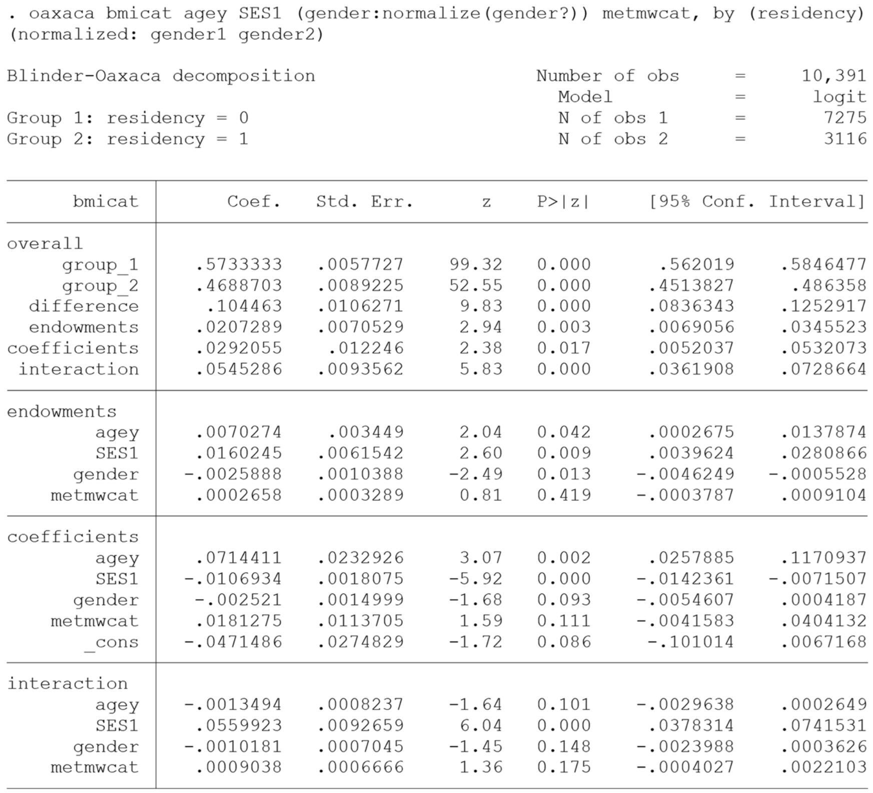
As we mentioned before, the unexplained part is sometimes referred to as the discrimination part. But for this part to be the exact discriminatory part, all the determinant of the outcome should exist in the model, otherwise this discriminatory part may be over or under estimated. Omitted variables or information bias can be some caused of this over/under estimation [15].

Application of B-O method in health issues

This B-O decomposition method can be applied to explain inequalities in health outcome across any two groups, which defined based on race, gender, socioeconomic status, and so on. Using the method, the inequality can be decomposed into two general components;

Table 1 Different Stata command packages for decomposition of outcome differences between the two groups

Command	Description (estimation_command)
Oaxaca	Linear decomposition; the default (linear), logit decomposition (logit), probit decomposition (probit)
Fairlie	Logit model; the default (logit), probit model (probit)
Gdecomp	Logit model; the default (logit), probit model (probit), Poisson regression (poisson), negative binomial regression (nbreg)
Mvdcmp	Probit model (probit), Poisson regression (poisson), negative binomial regression (nbreg), complementary log–log model (cloglog)
Nldecompose	Linear regression (regress), logit model (logit), probit model (probit), Ordinal logit model (ologit), Ordinal probit model (probit), Tobit model (tobit), Interval regression (intreg), Truncated Gaussian regression (truncreg), Poisson regression (poisson), negative binomial regression (nbreg), *zero-inflated Poisson* (zip), zero inflated negative binomial (zinb), *Zero-truncated* poisson (ztp), *Zero-truncated* negative binomial (ztnb)



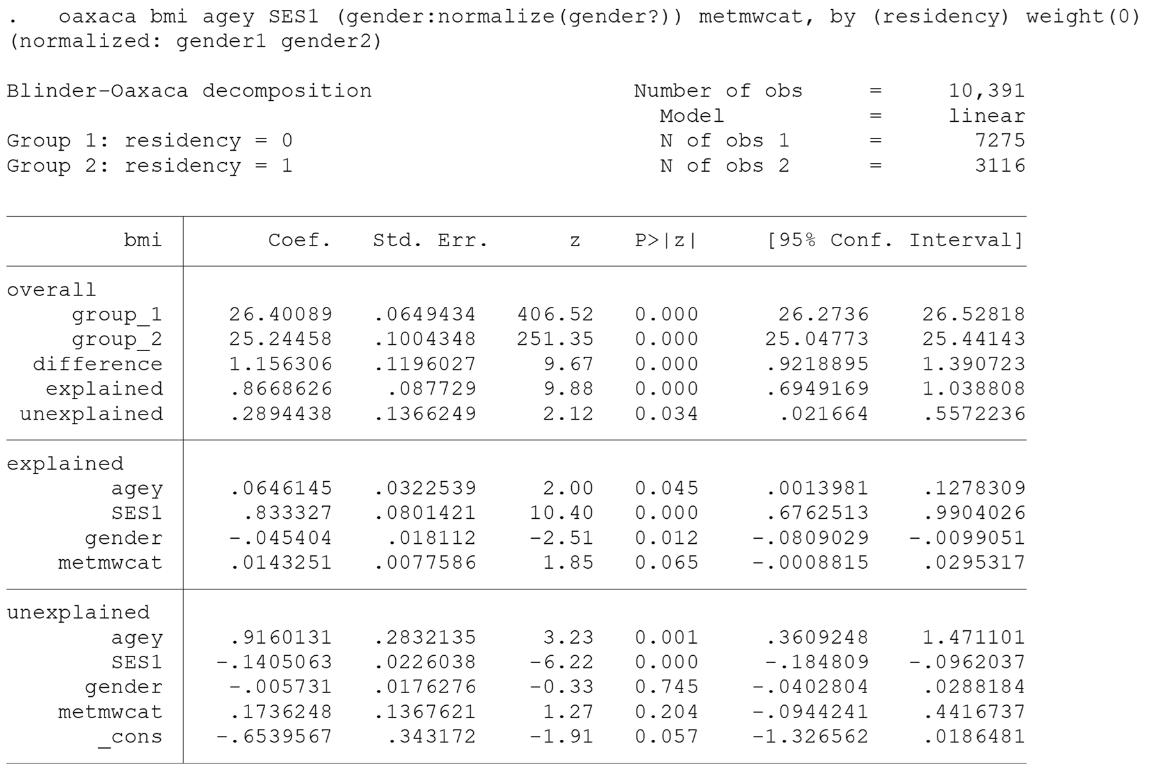
```
. oaxaca bmicat agey SES1 (gender:normalize(gender?)) metmwcat, by (residency) logit threefold(reverse)
(normalized: gender1 gender2)

Blinder-Oaxaca decomposition                       Number of obs     =     10,391
                                                     Model           =      logit
Group 1: residency = 0                               N of obs 1      =       7275
Group 2: residency = 1                               N of obs 2      =       3116

------------------------------------------------------------------------------
      bmicat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   .5733333   .0057727    99.32   0.000      .562019    .5846477
     group_2 |   .4688703   .0089225    52.55   0.000     .4513827     .486358
  difference |    .104463   .0106271     9.83   0.000     .0836343    .1252917
  endowments |   .0207289   .0070529     2.94   0.003     .0069056    .0345523
coefficients |   .0292055    .012246     2.38   0.017     .0052037    .0532073
 interaction |   .0545286   .0093562     5.83   0.000     .0361908    .0728664
-------------+----------------------------------------------------------------
endowments   |
        agey |   .0070274    .003449     2.04   0.042     .0002675    .0137874
        SES1 |   .0160245   .0061542     2.60   0.009     .0039624    .0280866
      gender |  -.0025888   .0010388    -2.49   0.013    -.0046249   -.0005528
    metmwcat |   .0002658   .0003289     0.81   0.419    -.0003787    .0009104
-------------+----------------------------------------------------------------
coefficients |
        agey |   .0714411   .0232926     3.07   0.002     .0257885    .1170937
        SES1 |  -.0106934   .0018075    -5.92   0.000    -.0142361   -.0071507
      gender |   -.002521   .0014999    -1.68   0.093    -.0054607    .0004187
    metmwcat |   .0181275   .0113705     1.59   0.111    -.0041583    .0404132
       _cons |  -.0471486   .0274829    -1.72   0.086     -.101014    .0067168
-------------+----------------------------------------------------------------
interaction  |
        agey |  -.0013494   .0008237    -1.64   0.101    -.0029638    .0002649
        SES1 |   .0559923   .0092659     6.04   0.000     .0378314    .0741531
      gender |  -.0010181   .0007045    -1.45   0.148    -.0023988    .0003626
    metmwcat |   .0009038   .0006666     1.36   0.175    -.0004027    .0022103
------------------------------------------------------------------------------
```

Output 1 Threefold (interaction) Blinder-Oaxaca decomposition for non-linear models using rural adults (group 2) as the reference (from perspective of Group 1)

The first component is explained by the differences in the level (distribution or average) of observed related factors or characteristics between 2 comparison groups. The second represents the rest of inequality that not explained by such differences. In fact, existence of inequality despite identical individual characteristics can be rooted in unknown or un-measurable factors that affect the health outcomes. It may also result from "differential effect" of the observed characteristics (group difference in the magnitude of regression coefficients) across comparison groups. Each of these components can be decomposed into smaller components depending on the number of characteristics (variables) with the potential to create inequality (detailed decomposition).

Statistically, the "differential effect" in the decomposition model arises from the interaction effect of the related factors with the group's indicator and can be interpreted in two ways; one depends on the nature of the variable itself, referring to the "behavioral response" of the explanatory variables in the two comparison groups such as different health behaviors and/or different individual tendency toward that behavior. One instance is the likelihood of smoking or deciding to start smoking in different communities or different social groups. Such difference can be due to cultural, environmental or attitudinal disparities in those communities [42]. The other is affected from the outside due to discrimination between the two groups in terms of associated factors (characteristics) such as unequal accessibility to health care services, and different quality of education, which in turn leads to different outcomes in the two groups. For example, if we assume that education level is the only known factor contributing to health behavior, the group difference in health behavior, despite the same level of education, can be due to varying qualities of education. This can be attributed to "differential effect" of education on health behavior in the absence of information on the quality of education.

Therefore, in addition to individual and social factors contributing to health, there are unequal or even unfair macro policies, social and economic programs playing a crucial role [5]. Assuming that individual and social characteristics remain the same in different subgroups of the population, inequality is expected to persist because of different government policies and programs. This implies that groups with specific characteristics receive different health programs and even at different quality. Hence, the conditions for inequality are set by physical, cultural,



```
.   oaxaca bmi agey SES1 (gender:normalize(gender?)) metmwcat, by (residency) weight(0)
(normalized: gender1 gender2)

Blinder-Oaxaca decomposition                          Number of obs     =     10,391
                                                        Model           =     linear
Group 1: residency = 0                                  N of obs 1      =       7275
Group 2: residency = 1                                  N of obs 2      =       3116

------------------------------------------------------------------------------
         bmi |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   26.40089   .0649434   406.52   0.000       26.2736    26.52818
     group_2 |   25.24458   .1004348   251.35   0.000      25.04773    25.44143
  difference |   1.156306   .1196027     9.67   0.000      .9218895    1.390723
   explained |   .8668626    .087729     9.88   0.000      .6949169    1.038808
 unexplained |   .2894438   .1366249     2.12   0.034       .021664    .5572236
-------------+----------------------------------------------------------------
explained    |
        agey |   .0646145   .0322539     2.00   0.045      .0013981    .1278309
        SES1 |    .833327   .0801421    10.40   0.000      .6762513    .9904026
      gender |   -.045404    .018112    -2.51   0.012     -.0809029   -.0099051
    metmwcat |   .0143251   .0077586     1.85   0.065     -.0008815    .0295317
-------------+----------------------------------------------------------------
unexplained  |
        agey |   .9160131   .2832135     3.23   0.001      .3609248    1.471101
        SES1 |  -.1405063   .0226038    -6.22   0.000      -.184809   -.0962037
      gender |   -.005731   .0176276    -0.33   0.745     -.0402804    .0288184
    metmwcat |   .1736248   .1367621     1.27   0.204     -.0944241    .4416737
       _cons |  -.6539567    .343172    -1.91   0.057     -1.326562    .0186481
------------------------------------------------------------------------------
```

Output 2 Twofold (Discrimination) Blinder-Oaxaca decomposition for linear models using the coefficients from rural adults' model as the reference coefficients

social and economic status of a community, giving rise to different, and in some cases, unfair opportunities.

In summary, the above decomposition methods can be applied in the health sciences in an effort to identify the contribution of each unevenly distributed factor as well as their different effects to the gap. It can therefore be indicated to what extent the average outcome varies according to changes in each factor while assuming the other factors are constant. Moreover, it will determine the overall share of unknown factors in creating inequality. In fact, the residual difference will be estimated while assuming that the distribution of observed factors remains identical [43].

It concludes that the application of the decomposition methods in the health inequality can identify the relative contribution of each particular factor in moderating the current inequality. Therefore, more detailed information can be provided for government planners and policymakers, especially concerning modifiable factors [23, 44].

Applied example

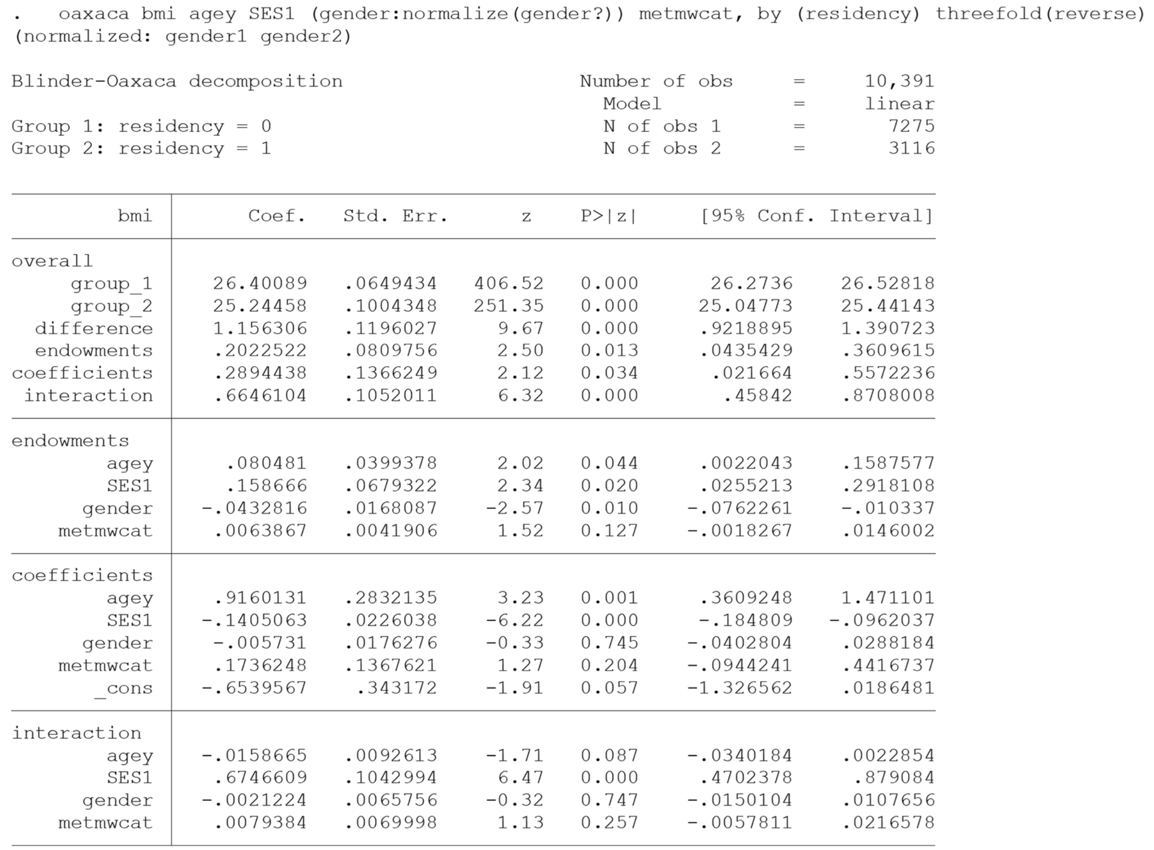
We illustrate the Blinder- Oaxaca Decomposition model using the available data from the 2011 STEPS Non-communicable Disease Risk Factors Survey of Iran. The survey was a population-based study according to STEPwise approach to the WHO non-communicable disease risk factor surveillance [45, 46].

The primary outcome was risk of obesity and overweight between urban and rural adults (residency: 0 = Urban, 1 = Rural). A set of variables including age (agey), gender (1 male, 2 female), socioeconomic stutus (SES) and physical activity (metmwcat: at least 600 MET-minutes per week) were considered as the predictors. The decomposition analysis was conducted in the Stata statistical software (v.14) using an updated Oaxaca package described by Yun [22]. The package included methods to handle the path dependency and identification problems [28–30]. The analysis has been performed for adults aged 15–69 years.

The threefold (interaction) decomposition type

We decompose predicted rural–urban difference in Body Mass Index (bmi) using the Blinder-Oaxaca decomposition for linear models. The overall and detailed results are presented in output 3.

"oaxaca" command in Stata computes the threefold decomposition from the perspective of Group 2 (Eq. 4), unless "threefold(reverse)", "weight()" or "pooled" is specified. In this example, "threefold(reverse)" option expresses the threefold decomposition from the



```
.   oaxaca bmi agey SES1 (gender:normalize(gender?)) metmwcat, by (residency) threefold(reverse)
(normalized: gender1 gender2)

Blinder-Oaxaca decomposition                    Number of obs      =     10,391
                                                  Model            =     linear
Group 1: residency = 0                            N of obs 1       =       7275
Group 2: residency = 1                            N of obs 2       =       3116

------------------------------------------------------------------------------
         bmi |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   26.40089   .0649434   406.52   0.000      26.2736    26.52818
     group_2 |   25.24458   .1004348   251.35   0.000     25.04773    25.44143
  difference |   1.156306   .1196027     9.67   0.000     .9218895    1.390723
  endowments |   .2022522   .0809756     2.50   0.013     .0435429    .3609615
coefficients |   .2894438   .1366249     2.12   0.034      .021664    .5572236
 interaction |   .6646104   .1052011     6.32   0.000       .45842    .8708008
-------------+----------------------------------------------------------------
endowments   |
        agey |    .080481   .0399378     2.02   0.044     .0022043    .1587577
        SES1 |    .158666   .0679322     2.34   0.020     .0255213    .2918108
      gender |  -.0432816   .0168087    -2.57   0.010    -.0762261    -.010337
    metmwcat |   .0063867   .0041906     1.52   0.127    -.0018267    .0146002
-------------+----------------------------------------------------------------
coefficients |
        agey |   .9160131   .2832135     3.23   0.001     .3609248    1.471101
        SES1 |  -.1405063   .0226038    -6.22   0.000     -.184809   -.0962037
      gender |   -.005731   .0176276    -0.33   0.745    -.0402804    .0288184
    metmwcat |   .1736248   .1367621     1.27   0.204    -.0944241    .4416737
       _cons |  -.6539567    .343172    -1.91   0.057    -1.326562    .0186481
-------------+----------------------------------------------------------------
interaction  |
        agey |  -.0158665   .0092613    -1.71   0.087    -.0340184    .0022854
        SES1 |   .6746609   .1042994     6.47   0.000     .4702378     .879084
      gender |  -.0021224   .0065756    -0.32   0.747    -.0150104    .0107656
    metmwcat |   .0079384   .0069998     1.13   0.257    -.0057811    .0216578
------------------------------------------------------------------------------
```

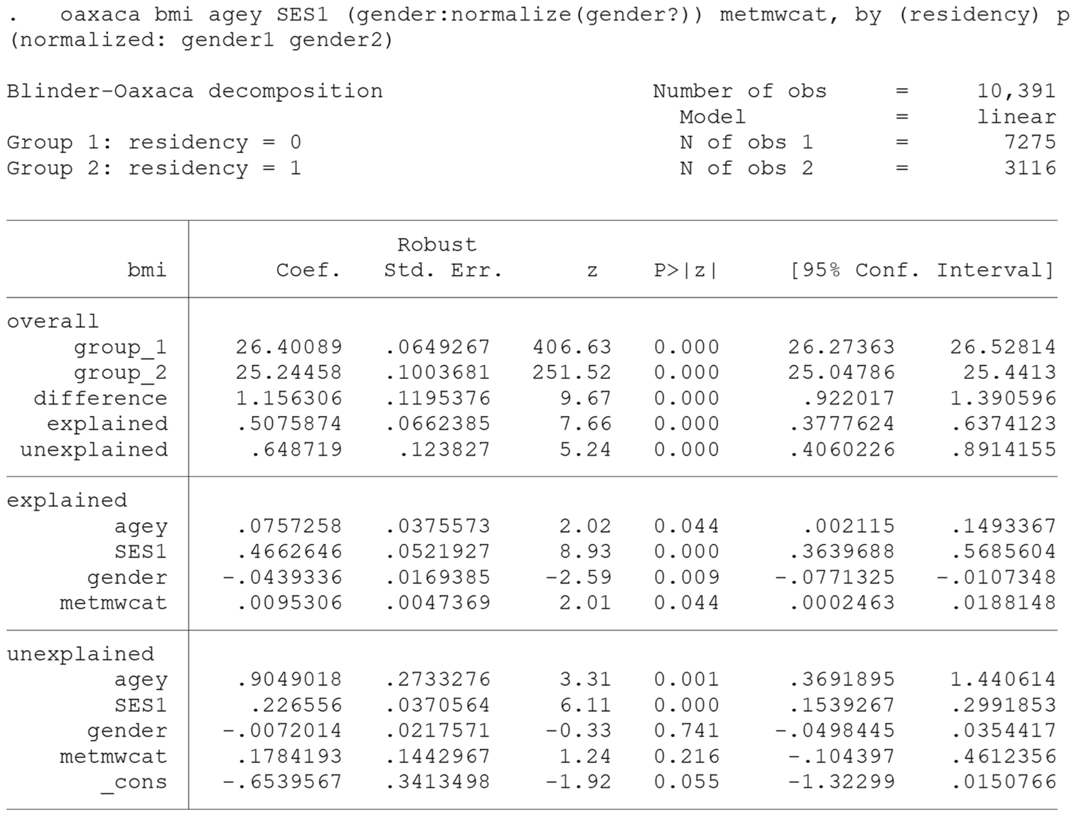
Output 3 Threefold (interaction) Blinder-Oaxaca decomposition for linear models using rural adults (group 2) as the reference (from perspective of Group 1)

perspective of Group 1 (Eq. 5). That means group 2 (i.e., rural adults with a low average BMI) are selected as the reference for analysis. As discussed in the text, for nominal predictors such as gender, the detailed decomposition estimates depend on the choice of the base (omitted) category (Identification Problem). A solution is to perform the decomposition based on "normalized" effects (gender:normalize(gender?)) which recognizes sets of dummy variables representing nominal predictor and converts the coefficients so that the results are constant to the choice of the baseline [30].

In our sample, the mean predicted BMI is 26.4 for urban adults and 25.24 for rural adults, yielding a BMI disparity of 1.156. In general, only about 17.5% (0.202/1.156) of the disparity was due to the different distribution of the predictors (endowments). Among them, SES contributed the most (0.158/1.156 = 13.72%). In other words, reducing the difference of SES between the rural and urban adults will lead to a reduction of approximately 14% in the disparity. Furthermore, about 25% (0.43/1.24) of the disparity was attributed to the differential effect of the covariate entered in the model (coefficients effect) including general effect of unknown factors (_cons). This component specifies the unexplained portion of the disparity. The differential effect of age (0.916/1.156 = 79.2%) had the greatest contribution to this part of the disparity, followed by physical activity (metmwcat) and SES. The negative contribution of SES implies that removing the rural/urban difference in SES widens the disparity. Moreover, the 'interaction part' refers to the gap that is explained by the interaction between the endowment and coefficient effects.

Similarly, the predicted rural–urban difference in prevalence of overweight and obesity (bmicat) has been decomposed using B-O decomposition for non-linear models. The results are presented in Output 1. As shown, the prevalence remained significantly higher among urban (57.34%) compared to urban adults (46.89%) controlling for age, SES, gender and physical activity. The decomposition results indicate that different level of SES and differential effect of age had the greatest contribution to the difference as well.

The "oaxaca" command in Stata also supports the nonlinear decomposition for binary outcome. "logit" causes



```
.   oaxaca bmi agey SES1 (gender:normalize(gender?)) metmwcat, by (residency) pooled
(normalized: gender1 gender2)

Blinder-Oaxaca decomposition                        Number of obs      =     10,391
                                                      Model            =     linear
Group 1: residency = 0                                N of obs 1       =       7275
Group 2: residency = 1                                N of obs 2       =       3116

------------------------------------------------------------------------------
             |               Robust
         bmi |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   26.40089   .0649267   406.63   0.000     26.27363    26.52814
     group_2 |   25.24458   .1003681   251.52   0.000     25.04786     25.4413
  difference |   1.156306   .1195376     9.67   0.000      .922017    1.390596
   explained |   .5075874   .0662385     7.66   0.000     .3777624    .6374123
 unexplained |    .648719    .123827     5.24   0.000     .4060226    .8914155
-------------+----------------------------------------------------------------
explained    |
        agey |   .0757258   .0375573     2.02   0.044      .002115    .1493367
        SES1 |   .4662646   .0521927     8.93   0.000     .3639688    .5685604
      gender |  -.0439336   .0169385    -2.59   0.009    -.0771325   -.0107348
    metmwcat |   .0095306   .0047369     2.01   0.044     .0002463    .0188148
-------------+----------------------------------------------------------------
unexplained  |
        agey |   .9049018   .2733276     3.31   0.001     .3691895    1.440614
        SES1 |    .226556   .0370564     6.11   0.000     .1539267    .2991853
      gender |  -.0072014   .0217571    -0.33   0.741    -.0498445    .0354417
    metmwcat |   .1784193   .1442967     1.24   0.216     -.104397    .4612356
       _cons |  -.6539567   .3413498    -1.92   0.055     -1.32299    .0150766
------------------------------------------------------------------------------
```

Output 4 Twofold (Discrimination) Blinder-Oaxaca decomposition for linear models using the coefficients from pooled model over both groups as the reference coefficients

the non-linear decomposition for a binary outcome to be computed using the weighting method described by Yun [22].

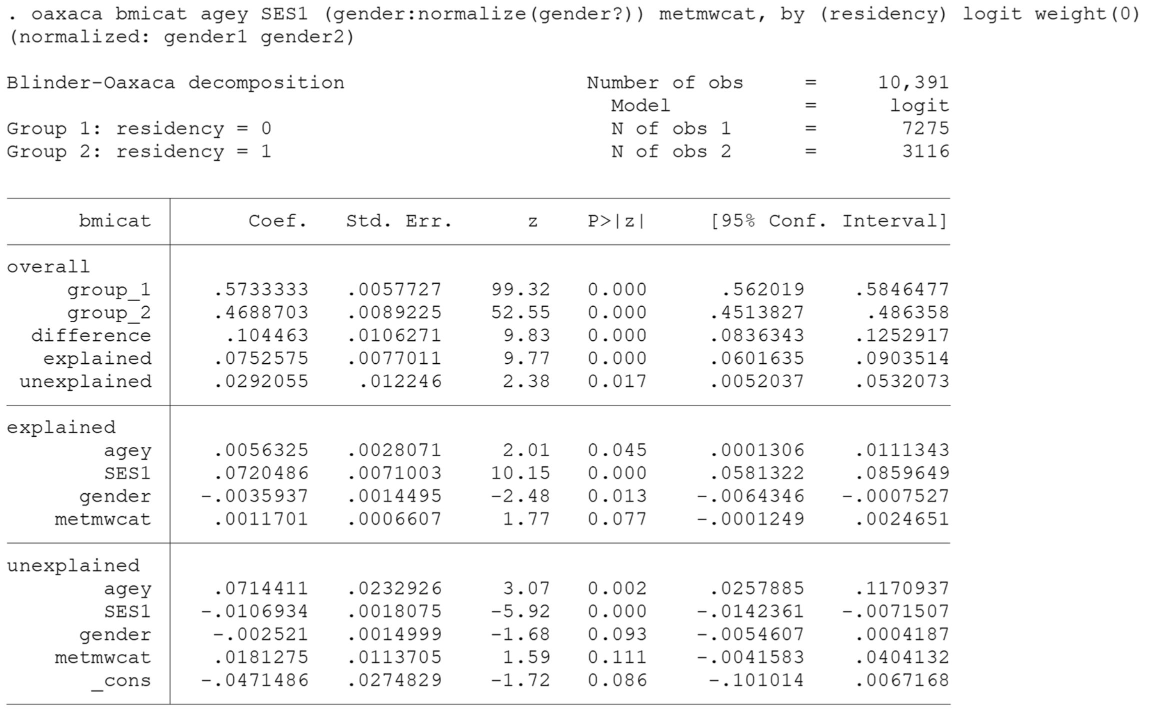
The threefold decomposition results indicate that the mean predicted BMI is generally higher in urban than in rural adults and that the prevalence of obesity and overweight is consequently higher in urban adults. This is attributed to an obesogenic environment that promotes obesity-related behaviours such as unhealthy diet and insufficient physical activity. Consistent with this findings, in many developing countries urbanization and its related lifestyle changes, are considered significant risk factors for obesity and overweight. However, it was not the case in the study of Trivedi et al. [47].

Persistent of the disparity after adjusting for some obesity-related factors call for further investigation in this issue, which suggests that much of the difference between urban and rural residents is also driven by other unknown factors. Moreover, the findings suggest that the effect of age on obesity and overweight risk is different across the both rural and urban adults. To be more precise, the urban adults experienced higher obesity risk with increasing age than rural ones. This is in line with WHO report discussing that in developing countries, rural adults still maintaining a classic lifestyle gained little weight with age [48]. Accordingly, effective programs are needed to help elderly urban adults reduce high risks for obesity and unhealthy lifestyles.

As a general conclusion, obesity risk reduction policies need to consider not only rural/urban adults but also how it interacts with associated factors that make some subgroups more vulnerable than others. Generally, the application of the decomposition methods in the health inequality can identify the relative contribution of each particular factor in moderating the inequality. Therefore, more detailed information can be provided for government planners and policy-makers, especially concerning modifiable factors [23, 44].

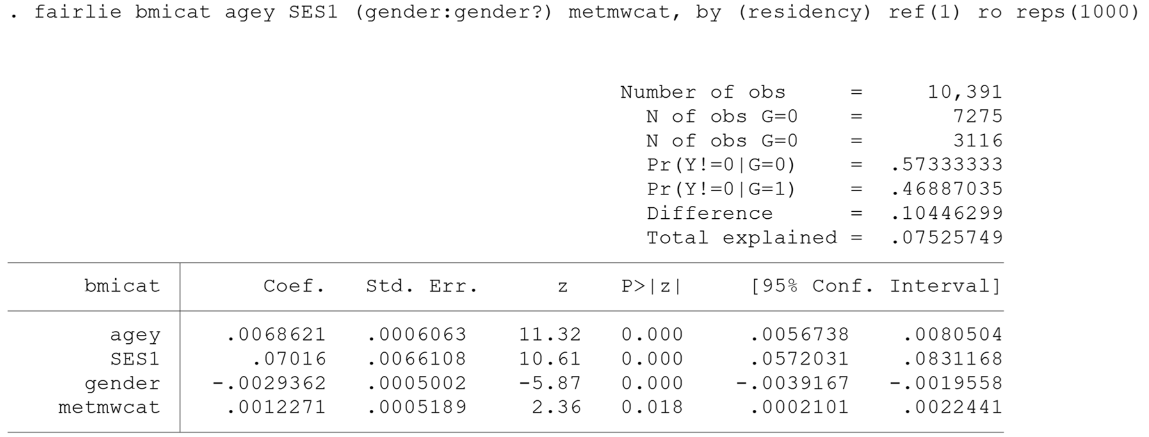
The twofold (discrimination) decomposition type

An alternative decomposition commonly used in the discrimination literature is the twofold decompositin (Eqs. 6, 7 and 8). In “Oaxaca” command in Stata, this decomposition can be performed, where “weight()” or “pooled” specifies the choice of the reference coefficients.



```
. oaxaca bmicat agey SES1 (gender:normalize(gender?)) metmwcat, by (residency) logit weight(0)
(normalized: gender1 gender2)

Blinder-Oaxaca decomposition                      Number of obs     =     10,391
                                                    Model           =      logit
Group 1: residency = 0                              N of obs 1      =       7275
Group 2: residency = 1                              N of obs 2      =       3116

------------------------------------------------------------------------------
      bmicat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   .5733333   .0057727    99.32   0.000      .562019    .5846477
     group_2 |   .4688703   .0089225    52.55   0.000     .4513827     .486358
  difference |    .104463   .0106271     9.83   0.000     .0836343    .1252917
   explained |   .0752575   .0077011     9.77   0.000     .0601635    .0903514
 unexplained |   .0292055    .012246     2.38   0.017     .0052037    .0532073
-------------+----------------------------------------------------------------
explained    |
        agey |   .0056325   .0028071     2.01   0.045     .0001306    .0111343
        SES1 |   .0720486   .0071003    10.15   0.000     .0581322    .0859649
      gender |  -.0035937   .0014495    -2.48   0.013    -.0064346   -.0007527
    metmwcat |   .0011701   .0006607     1.77   0.077    -.0001249    .0024651
-------------+----------------------------------------------------------------
unexplained  |
        agey |   .0714411   .0232926     3.07   0.002     .0257885    .1170937
        SES1 |  -.0106934   .0018075    -5.92   0.000    -.0142361   -.0071507
      gender |   -.002521   .0014999    -1.68   0.093    -.0054607    .0004187
    metmwcat |   .0181275   .0113705     1.59   0.111    -.0041583    .0404132
       _cons |  -.0471486   .0274829    -1.72   0.086     -.101014    .0067168
------------------------------------------------------------------------------
```

Output 5 Twofold (Discrimination) Blinder-Oaxaca decomposition for non-linear models using the coefficients from rural adults' model as the reference coefficients



```
. fairlie bmicat agey SES1 (gender:gender?) metmwcat, by (residency) ref(1) ro reps(1000)


                                                      Number of obs     =     10,391
                                                        N of obs G=0    =       7275
                                                        N of obs G=0    =       3116
                                                        Pr(Y!=0|G=0)    =  .57333333
                                                        Pr(Y!=0|G=1)    =  .46887035
                                                        Difference      =  .10446299
                                                        Total explained =  .07525749
------------------------------------------------------------------------------
      bmicat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        agey |   .0068621   .0006063    11.32   0.000     .0056738    .0080504
        SES1 |     .07016   .0066108    10.61   0.000     .0572031    .0831168
      gender |  -.0029362   .0005002    -5.87   0.000    -.0039167   -.0019558
    metmwcat |   .0012271   .0005189     2.36   0.018     .0002101    .0022441
------------------------------------------------------------------------------
```

Output 6 Fairlie decomposition model using the coefficients from rural adults' model as the reference coefficients

The results after using the "weight(0)" option are presented in output 2. It indicates that the coefficients from the group 2's (rural adults') model are used as the reference (non-discriminating). On the contrary, "weight(1)" specifies group 1 coefficients as the standard. In our case, the rural adults (Group 2) with a lower average of BMI was preferred as the reference.

As is evident from the output 2, the "unexplained" component is exactly similar to the "coefficients" component of the three-fold decomposition (Output 3). This component is often used as a measure for discrimination, but it also subsumes the effects of group differences in unobserved predictors. For this part to be the exact discriminatory part, all the determinant of the outcome should exist in the model, otherwise this part may be over or under estimated.

As shown, the differences in the level of observed covariates (the explained component) accounted for about 75% (0.867/1.156) of the total disparity. This component is the combination of "endowments" and "interaction" parts of the three-fold decomposition (Output 3). Although this component is called the explained component in two-fold decomposition in many texts, but some part of it (the interaction part) is in fact the simultaneous difference of coefficients and covariates level in both groups. Hence, if somebody wants the crude explained component, three folds' decomposition can provide this crude explained part.

In the Output 4, "pooled" specifies that the coefficients from the pooled model over all cases be used for the decomposition. The results also indicate that different level of SES and differential effect of age had the greatest contribution to the difference. However, it is clearly shown that the decomposition estimates vary depending on the choice of reference group (index problem). There is often no compelling reason to choose the best group.

Alternatively, the twofold decomposition can be requested for non-linear models. In the Output 5, the predicted rural–urban difference in the prevalence of overweight and obesity (bmicat) has been decomposed using Blinder-Oaxaca decomposition for non-linear models.

An alternative non-linear decomposition command for binary outcome is available as "fairlie" [14].

The primary outcome is difference in the proportion of obesity and overweight (bmicat) between the rural and urban adults (residency: must be coded as 0 and 1). Accordingly, the technique decompose the rural/urban difference in "mean predicted probability of outcome". However, it mainly focuses on the explained portion of inequality without calculating the contribution of the differential effect from each factor to the unexplained part [14].

The main concern with the non-linear model is sensitive to the order of variables being included into the decomposition process (path dependency). The fairlie technique solving the problem by randomly ordering the variables across replications of the decomposition [28].

"ro" option causes the ordering of variables to be randomized in the analysis. reps(#) defines the number of decomposition replications. Thus, the multiple random sub-samples (e.g. 100 or 1000 times) of the majority group (equal to the sample size of minority group) are selected and the mean estimate is considered as the final estimate [14, 24, 28, 33].

ref(#) specifies the reference coefficients to be used with the decomposition. "ref(1)" indicates that the coefficients from the group = = 1 model (rural adults') are used. It is equivalent to the "weight (0)" in twofold Blinder-Oaxaca decomposition models.

Outputs 5 and 6 reports estimates from two decomposition methods, the non-linear Blinder-Oaxaca technique and the fairlie technique, for the rural/urban disparity in the prevalence of overweight and obesity. The prevalence also remained significantly higher among urban compared to urban adults controlling for age, SES, gender and physical activity. This is consistent with all the above reports. Approximately, 72% (0.0753/0.104) of disparity explained by rural/urban differences in these predictors. Among them, SES contributed the most.

Abbreviations

B-O: Blinder-Oaxaca decomposition method; Uc: Unexplained Component; Ec: Explained Component; BMI: Body Mass Index; SES: Socioeconomic Status.

Acknowledgements

The authors greatly appreciate the center for NCDs control of Iran's Ministry of Health and Medical Education for providing the data for applied example.

Authors' contributions

The first author (RE) contributed substantially to the conception of the original idea, the writing of the manuscript and to performing the analysis. The second author (SSHN) conceived the original idea, provided critical revision of the article and supervised the project. Both authors read and approved the final manuscript.

Funding

The authors received no funding for this work.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

We used an available secondary data from the 2011 STEPS Non-communicable Disease Risk Factors Survey of Iran. The authors confirm that the data has no identifying information.

Consent for publication

Both authors of the paper consent to publication.

Competing interests

The authors have no conflicts of interest.

Author details

[1] Department of Public Health, Mamasani Higher Education Complex for Health, Shiraz University of Medical Sciences, Shiraz, Iran. [2] Prevention of Cardiovascular Disease Research Center, Department of Epidemiology, School of Public Health and Safety, Shahid Beheshti University of Medical Sciences, Velenjak St., Chamran Highway, Tehran, Iran.

Received: 18 October 2020 Accepted: 15 July 2021
Published online: 06 August 2021

References

1. Morris S, Sutton M, Gravelle H. Inequity and inequality in the use of health care in England: an empirical investigation. Soc Sci Med. 2005;60(6):1251–66.

2. Braveman P, Gruskin S. Defining equity in health. J Epidemiol Community Health. 2003;57(4):254–8.
3. Ospino CG, Vasquez PR, Narvaez NB. Oaxaca-Blinder wage decomposition: methods, critiques and applications. A literature review. Revista de Economía del Caribe. 2010;5:237–74.
4. Kawachi I, Subramanian SV, Almeida-Filho N. A glossary for health inequalities. J Epidemiol Community Health. 2002;56(9):647–52.
5. Graham, H. and M.P. Kelly, Health inequalities: concepts, frameworks and policy. 2004: Health Development Agency London.
6. World Health Organization, Handbook on health inequality monitoring with a special focus on low- and middle-income countries. 2013. Available from: https://www.who.int/docs/default-source/gho-documents/health-equity/handbook-on-health-inequality-monitoring/handbook-on-health-inequality-monitoring.pdf?sfvrsn=d27f8211_2.
7. Blinder AS. Wage discrimination: reduced form and structural estimates. J Hum Resources. 1973;8:436–55.
8. Oaxaca R. Male-female wage differentials in urban labor markets. Int Economic Rev. 1973;14:693–709.
9. Jann B. The Blinder-Oaxaca decomposition for linear regression models. Stand Genomic Sci. 2008;8(4):453–79.
10. O'Donnell OA, Wagstaff A. Analyzing health equity using household survey data: a guide to techniques and their implementation. World Bank Publications, 2008.
11. Jones FL, Kelley J. Decomposing differences between groups: a cautionary note on measuring discrimination. Sociological Methods Res. 1984;12(3):323–43.
12. Daymont TN, Andrisani PJ. Job preferences, college major, and the gender gap in earnings. J Hum Resources. 1984;19:408–28.
13. "Blinder-Oaxaca Decomposition Technique ." International Encyclopedia of the Social Sciences. Retrieved June 17, 2021 from Encyclopedia.com: https://www.encyclopedia.com/social-sciences/applied-and-social-sciences-magazines/blinder-oaxaca-decomposition-technique. Accessed 23 July 2021.
14. Fairlie RW. An extension of the Blinder-Oaxaca decomposition technique to logit and probit models. J Econ Soc Meas. 2005;30(4):305–16.
15. Cotton J. On the decomposition of wage differentials. Rev Economics Statistics. 1988;70:236–43.
16. Neumark D. Employers' discriminatory behavior and the estimation of wage discrimination. J Hum Resources. 1988;23:279–95.
17. Oaxaca RL, Ransom MR. On discrimination and the decomposition of wage differentials. J Econometrics. 1994;61(1):5–21.
18. Farhat JB, Mijid N. Do women lag behind men? A matched-sample analysis of the dynamics of gender gaps. a matched-sample analysis of the dynamics of gender gaps (May 21, 2016), 2016.
19. Reimers CW. Labor market discrimination against Hispanic and black men. Rev Economics Statistics. 1983;65:570–9.
20. Fairlie RW. The absence of the African-American owned business: an analysis of the dynamics of self-employment. J Law Econ. 1999;17(1):80–108.
21. Bauer TK, Sinning M. An extension of the Blinder-Oaxaca decomposition to nonlinear models. AStA. 2008;92(2):197–206.
22. Yun M-S. Decomposing differences in the first moment. Econ Lett. 2004;82(2):275–80.
23. Averett SL, Stacey N, Wang Y. Decomposing race and gender differences in underweight and obesity in South Africa. Econ Hum Biol. 2014;15:23–40.
24. Mehta HB, et al. Application of the nonlinear Blinder-Oaxaca decomposition to study racial/ethnic disparities in antiobesity medication use in the United States. Res Social Adm Pharm. 2013;9(1):13–26.
25. Liu H, Fang H, Zhao Z. Urban–rural disparities of child health and nutritional status in China from 1989 to 2006. Econ Hum Biol. 2013;11(3):294–309.
26. Brick A, Nolan A. Maternal country of birth differences in breastfeeding at hospital discharge in Ireland. Econ Social Rev. 2014;45(4):455–84.
27. Lhila A, Long S. What is driving the black–white difference in low birthweight in the US? Health Econ. 2012;21(3):301–15.
28. Fairlie RW. Addressing path dependence and incorporating sample weights in the nonlinear blinder-oaxaca decomposition technique for logit, probit and other nonlinear models. University of California, Santa Cruz Working Paper, 2016. Retrieved from http://people.ucsc.edu/~rfairlie/decomposition/decomprevisted_v9.docx.
29. Rahimi E, et al. Decomposing gender disparity in total physical activity among Iranian adults. Epidemiol Health. 2017;39:e2017044.
30. Yun MS. A simple solution to the identification problem in detailed wage decompositions. Econ Inq. 2005;43(4):766–72.
31. Yun M-S. Identification problem and detailed Oaxaca decomposition: a general solution and inference. J Econ Soc Meas. 2008;33(1):27–38.
32. Powers DA, Yoshioka H, Yun M-S. mvdcmp: multivariate decomposition for nonlinear response models. Stata J. 2011;11(4):556–76.
33. Nie P, Sousa-Poza A. Food insecurity among older Europeans: Evidence from the Survey of Health, Ageing, and Retirement in Europe. 2016, University of Hohenheim, Faculty of Business, Economics and Social Sciences.
34. Nielsen HS. Discrimination and detailed decomposition in a logit model. Econ Lett. 1998;61(1):115–20.
35. Even WE, Macpherson DA. The decline of private-sector unionism and the gender wage gap. J Hum Resources. 1993;28:279–96.
36. Hlavac M. Oaxaca: blinder-oaxaca decomposition in r. browser download this paper, 2014.
37. Lewis T, Ezoua S. A Simple SAS® Macro to Perform Blinder-Oaxaca Decomposition. 2016; http://analytics.ncsu.edu/sesug/2016/SD-162_Final_PDF.pdf.
38. Jann B. Fairlie: stata module to generate nonlinear decomposition of binary outcome differentials. 2006.
39. Bartus T. Marginal effects and extending the Blinder–Oaxaca decomposition for nonlinear models. in United Kingdom Stata Users' Group Meetings 2006. 2006. Stata Users Group.
40. Sinning M, Hahn M, Bauer TK. The Blinder-Oaxaca decomposition for nonlinear regression models. Stand Genomic Sci. 2008;8(4):480–92.
41. Oaxaca RL, Ransom MR. Identification in detailed wage decompositions. Rev Econ Stat. 1999;81(1):154–7.
42. Amin V, Lhila A. Decomposing racial differences in adolescent smoking in the US. Econ Hum Biol. 2016;22:161–76.
43. Dubowitz T, et al. Racial/ethnic differences in US health behaviors: a decomposition analysis. Am J Health Behav. 2011;35(3):290–304.
44. Fortin N, Lemieux T, Firpo S. Decomposition methods in economics. Handbook Labor Economics. 2011;4:1–102.
45. World Health Organization (WHO). STEPwise Approach to NCD Risk Factor Surveillance (STEPS): Country Data & Reports (cited on 2021). https://www.who.int/teams/noncommunicable-diseases/surveillance/systems-tools/steps. Accessed May 2020.
46. WHO STEPS Surveillance Manual, The WHO STEPwise approach to noncommunicable disease risk factor surveillance. Last Updated: 26 January 2017. https://www.who.int/ncds/surveillance/steps/STEPS_Manual.pdf?ua=1. Accessed May 2021.
47. Trivedi T, et al. Obesity and obesity-related behaviors among rural and urban adults in the USA. 2015.
48. Organization, W.H., Obesity: preventing and managing the global epidemic. 2000.

Publisher's Note