oaxaca: Blinder-Oaxaca Decomposition in R

Marek Hlavac
Harvard University

Abstract

This article introduces the R package **oaxaca** to perform the Blinder-Oaxaca decomposition, a statistical method that decomposes the gap in mean outcomes across two groups into a portion that is due to differences in group characteristics and a portion that cannot be explained by such differences. Although this method has been most widely used to study gender- and race-based discrimination in the labor market, Blinder-Oaxaca decompositions can be applied to explain differences in any continuous outcome across any two groups. The **oaxaca** package implements all the most commonly used variants of the Blinder-Oaxaca decomposition for linear regression models, calculates bootstrapped standard errors for its estimates, and allows users to visualize the decomposition results.

Keywords: Blinder-Oaxaca decomposition, linear regression models, R.

1. Introduction

In this article, I introduce the R package **oaxaca** to estimate Blinder-Oaxaca decompositions for linear regression models. The Blinder-Oaxaca decomposition is a statistical method that decomposes differences in mean outcomes across two groups into a part that is due to group differences in the levels of explanatory variables and a part that is due to differential magnitudes of regression coefficients.

The Blinder-Oaxaca decomposition originated and has been widely used in the study of labor market discrimination (Blinder 1973; Oaxaca 1973). Economists and sociologists have, for instance, used it to decompose wage and earnings differences based on gender (e.g., Stanley and Jarrell 1998; Weichselbaumer and Winter-Ebmer 2005) and race (e.g., Darity, Guilkey, and Winfrey 1996; Kim 2010). Although Blinder-Oaxaca decompositions have been a mainstay of empirical research on discrimination, they can be, in principle, applied to explain differences in any continuous outcome across any two groups. Researchers have, for instance, used it to examine the assimilation of immigrants (LaLonde and Topel 1992), school enrolment rates (Borooah and Iyer 2006), health insurance coverage (Bustamante, Fang, Rizzo, and Ortega 2009), the prevalence of smoking (Bauer, Göhlmann, and Sinning 2007), or even local hunting lease rates (Munn and Hussain 2010).

Several software implementations of the Blinder-Oaxaca decomposition are already available. These include modules **oaxaca** (Jann 2008), **decomp** (Watson 2010) for Stata (StataCorp 2013) that estimate the decomposition for linear regression models, as well as Stata modules **fairlie** (Jann 2006) and **nldecompose** (Sinning, Hahn, and Bauer 2008) that implement the decomposition for a large variety of non-linear models using methods proposed in Fairlie (2005), Bauer and Sinning (2008) and Bauer and Sinning (2010). A SAS (SAS Institute 2014)

implementation of the Blinder-Oaxaca decomposition for non-linear models is also available (Fairlie 2013).

The **oaxaca** package is the first Blinder-Oaxaca decomposition package for the R statistical programming language (R Core Team 2014). It implements several types of the decomposition for linear regression models, and obtains point estimates of all decomposition components using the same estimation procedures as the Stata module **oaxaca** (Jann 2008). Standard errors are calculated using a non-parametric bootstrapping approach (Efron 1979). Unlike any other existing software implementation of the Blinder-Oaxaca decomposition, **oaxaca** enables users to generate elegant bar graph visualizations of all decomposition results.

The package is available free of charge, and can be installed from the Comprehensive R Archive Network (CRAN) (2014) in the usual way:

```
R> install.packages("oaxaca")
```

In the next section, I give a brief description of the Blinder-Oaxaca decomposition method. I then provide an overview of the **oaxaca** package's features in Section 3. In Section 4, I showcase them on an empirical example that examines the wage gap between native and foreign-born Hispanic workers in metropolitan Chicago. Section 5 concludes.

2. Blinder-Oaxaca decomposition

This section provides an overview of the Blinder-Oaxaca decomposition. It is by no means intended to be exhaustive, and primarily aims to give readers an understanding of the estimation procedures that the **oaxaca** package implements. Readers who are interested in a more comprehensive and rigorous treatment of the statistical method can refer to the excellent overview in Jann (2008), whose notation I follow with only a few minor adjustments.

The aim of the Blinder-Oaxaca decomposition is to explain how much of the difference in mean outcomes across two groups is due to group differences in the levels of explanatory variables, and how much is due to differences in the magnitude of regression coefficients (Oaxaca 1973; Blinder 1973). I will label the two groups as Group A and Group B. The mean outcome difference to be explained ($\Delta\bar{Y}$) is simply the difference of the mean outcomes for observations in Group A and Group B, denoted as $\bar{Y}_A$ and $\bar{Y}_B$, respectively:

$$\Delta\bar{Y} = \bar{Y}_A - \bar{Y}_B \tag{1}$$

2.1. Threefold decomposition

In the context of a linear regression, the mean outcome for Group $G \in \{A, B\}$ can be expressed as $\bar{Y}_G = \bar{\boldsymbol{X}}'_G\hat{\boldsymbol{\beta}}_G$, where $\bar{\boldsymbol{X}}_G$ contains the mean values of explanatory variables and $\hat{\boldsymbol{\beta}}_G$ are the estimated regression coefficients. Hence, $\Delta\bar{Y}$ can be rewritten as:

$$\Delta\bar{Y} = \bar{\boldsymbol{X}}'_A\hat{\boldsymbol{\beta}}_A - \bar{\boldsymbol{X}}'_B\hat{\boldsymbol{\beta}}_B \tag{2}$$

This expression can, in turn, be written as the sum of the following three terms:

$$\Delta\bar{Y} = \underbrace{(\bar{\boldsymbol{X}}_A - \bar{\boldsymbol{X}}_B)'\hat{\boldsymbol{\beta}}_B}_{\text{endowments}} + \underbrace{\bar{\boldsymbol{X}}'_B(\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}_B)}_{\text{coefficients}} + \underbrace{(\bar{\boldsymbol{X}}_A - \bar{\boldsymbol{X}}_B)'(\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}_B)}_{\text{interaction}} \tag{3}$$

Equation 3 is the threefold Blinder-Oaxaca decomposition of the mean outcome difference. The endowments term represents the contribution of differences in explanatory variables across groups, and the coefficients term is the part that is due to group differences in the coefficients. Finally, the interaction term accounts for the fact that cross-group differences in explanatory variables and coefficients can occur at the same time.

The threefold decomposition can also be estimated separately for each explanatory variable:

$$\underbrace{(\bar{\boldsymbol{X}}_A - \bar{\boldsymbol{X}}_B)'\hat{\boldsymbol{\beta}}_B}_{\text{endowments}} = \underbrace{(\bar{X}_{1A} - \bar{X}_{1B})\hat{\beta}_{1B}}_{\text{variable 1}} + \underbrace{(\bar{X}_{2A} - \bar{X}_{2B})\hat{\beta}_{2B}}_{\text{variable 2}} + \dots \tag{4}$$

$$\underbrace{\bar{\boldsymbol{X}}'_B(\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}_B)}_{\text{coefficients}} = \underbrace{\bar{X}_{1B}(\hat{\beta}_{1A} - \hat{\beta}_{1B})}_{\text{variable 1}} + \underbrace{\bar{X}_{2B}(\hat{\beta}_{2A} - \hat{\beta}_{2B})}_{\text{variable 2}} + \dots \tag{5}$$

$$\underbrace{(\bar{\boldsymbol{X}}_A - \bar{\boldsymbol{X}}_B)'(\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}_B)}_{\text{interaction}} = \underbrace{(\bar{X}_{1A} - \bar{X}_{1B})(\hat{\beta}_{1A} - \hat{\beta}_{1B})}_{\text{variable 1}} + \underbrace{(\bar{X}_{2A} - \bar{X}_{2B})(\hat{\beta}_{2A} - \hat{\beta}_{2B})}_{\text{variable 2}} + \dots \tag{6}$$

2.2. Twofold decomposition

Alternatively, one can estimate a twofold Blinder-Oaxaca decomposition. The twofold approach decomposes the mean outcome difference with respect to a vector of reference coefficients $\hat{\boldsymbol{\beta}}_R$. In the research literature on labor market discrimination, the reference coefficient vector has typically been interpreted to be non-discriminatory – in other words, as the set of regression coefficients that would emerge in a world of no labor market discrimination.

$$\Delta\bar{Y} = \underbrace{(\bar{\boldsymbol{X}}_A - \bar{\boldsymbol{X}}_B)'\hat{\boldsymbol{\beta}}_R}_{\text{explained}} + \underbrace{\underbrace{\bar{\boldsymbol{X}}'_A(\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}_R)}_{\text{unexplained A}} + \underbrace{\bar{\boldsymbol{X}}'_B(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_B)}_{\text{unexplained B}}}_{\text{unexplained}} \tag{7}$$

As Equation 7 shows, the twofold decomposition divides the difference in mean outcomes into a portion that is explained by cross-group differences in the explanatory variables, and a part that remains unexplained by these differences.

The unexplained portion of the mean outcome gap has often been attributed to discrimination, but may also result from the influence of unobserved variables. It can be further decomposed into two sub-components, labeled "unexplained A" and "unexplained B" above. If one interprets the reference coefficient vector to be non-discriminatory, these sub-components measure the part of the mean difference in outcomes that originates from discrimination in favor of Group A and the part that comes from discrimination against Group B, respectively.

Again, a detailed, variable-by-variable decomposition can also be estimated:

$$\underbrace{(\bar{\boldsymbol{X}}_A - \bar{\boldsymbol{X}}_B)'\hat{\boldsymbol{\beta}}_R}_{\text{explained}} = \underbrace{(\bar{X}_{1A} - \bar{X}_{1B})\hat{\beta}_{1R}}_{\text{variable 1}} + \underbrace{(\bar{X}_{2A} - \bar{X}_{2B})\hat{\beta}_{2R}}_{\text{variable 2}} + \dots \tag{8}$$

$$\underbrace{\bar{\boldsymbol{X}}'_A(\hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}_R)}_{\text{unexplained A}} = \underbrace{\bar{X}_{1A}(\hat{\beta}_{1A} - \hat{\beta}_{1R})}_{\text{variable 1}} + \underbrace{\bar{X}_{2A}(\hat{\beta}_{2A} - \hat{\beta}_{2R})}_{\text{variable 2}} + \dots \tag{9}$$

$$\underbrace{\bar{\boldsymbol{X}}'_B(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_B)}_{\text{unexplained B}} = \underbrace{\bar{X}_{2B}(\hat{\beta}_{2R} - \hat{\beta}_{2B})}_{\text{variable 1}} + \underbrace{\bar{X}_{2B}(\hat{\beta}_{2R} - \hat{\beta}_{2B})}_{\text{variable 2}} + \dots \tag{10}$$

The choice of the reference coefficients is generally up to the researcher. In the literature on labor market discrimination, it is often assumed that only one of the two groups faces discrimination – for instance, that only women or members of ethnic minorities are discriminated against. In such cases, the reference coefficients will simply be the coefficients from a regression on observations from one of the groups: either $\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}_A$ or $\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}_B$.

Some researchers have instead used a weighted average of $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_B$ as the set of reference coefficients. Reimers (1983), for example, proposes giving equal weight to coefficients from regressions on Group A and Group B observations:

$$\hat{\boldsymbol{\beta}}_R = 0.5\hat{\boldsymbol{\beta}}_A + 0.5\hat{\boldsymbol{\beta}}_B \tag{11}$$

Cotton (1988) suggests weighting the coefficients by the proportion of observations in the corresponding group:

$$\hat{\boldsymbol{\beta}}_R = \frac{n_A}{n_A + n_B}\hat{\boldsymbol{\beta}}_A + \frac{n_B}{n_A + n_B}\hat{\boldsymbol{\beta}}_B \tag{12}$$

Others still have advocated the use of coefficient estimates from a regression that pools observations from both Groups A and B, and includes (Jann 2008) or does not include (Neumark 1988) the group indicator variable as an additional regressor. The **oaxaca** package estimates results for all of the aforementioned choices of $\hat{\boldsymbol{\beta}}_R$, and also enables users to specify their own custom weights for $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_B$ to construct a weighted average-based set of reference coefficients.

2.3. Sensitivity to the choice of omitted baseline category

The results of Blinder-Oaxaca decompositions have been found to be sensitive to the researcher's choice of the omitted baseline category when categorical variables are included as covariates (Oaxaca and Ransom 1999). Typically, categorical explanatory variables are introduced as a set of indicator ("dummy") variables on the right hand side. To avoid perfect multicollinearity, one of the dummy variables is usually omitted, and represents the baseline category. The coefficients on the remaining dummy variables are then interpreted as deviations from this omitted baseline. A linear regression model that contains a categorical explanatory variable may thus have the following general form:

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \dots + \beta_{k-1} D_{k-1} + \boldsymbol{X}'\boldsymbol{\gamma} + \epsilon \tag{13}$$

where D_i, such that $i = 1, \ldots, k-1$, are indicator variables that represent individual levels of a categorical variable. Category k is the omitted baseline.

To ensure that the Blinder-Oaxaca decomposition results are invariant to the user's choice of the omitted baseline category, **oaxaca** implements a procedure proposed by Gardeazabal and Ugidos (2004). More specifically, the package transforms the above regression model into:

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 D_1 + \tilde{\beta}_2 D_2 + \tilde{\beta}_3 D_3 + \cdots + \tilde{\beta}_{k-1} D_{k-1} + \tilde{\beta}_k D_k + \boldsymbol{X}'\boldsymbol{\gamma} + \epsilon \tag{14}$$

where the new regression coefficients on the indicator variables are calculated by adding or subtracting an adjustment amount a to/from the original coefficients. The adjustment amount a is simply the sum of the original dummy coefficients $\boldsymbol{\beta}$ divided by k, the total number of categories:

$$a = \frac{\sum_{j=1}^{k-1} \beta_j}{k} \tag{15}$$

The adjustment amount is then added to the original intercept β_0:

$$\tilde{\beta}_0 = \beta_0 + a \tag{16}$$

and subtracted from each of the other regression coefficients:

$$\tilde{\beta}_i = \beta_i - a \tag{17}$$

for $i = 1, \ldots, k$. The adjusted coefficients ($\tilde{\boldsymbol{\beta}}$), as well as the results of detailed variable-by-variable Blinder-Oaxaca decompositions, will remain the same regardless of the researcher's choice of the omitted category k.

2.4. Estimation uncertainty

The **oaxaca** package provides a measure of the estimation uncertainty that accompanies each of its decomposition estimates. In particular, it reports bootstrapped standard errors based on a user-specified number (R) of sampling replicates (Efron 1979). The package uses the following procedures to calculate standard errors:

1. R resamples are randomly sampled with replacement from the relevant set of observations.
2. Decomposition estimates are calculated for each of the R resamples from Step 1.
3. The bootstrapped standard error is the standard deviation of the R decomposition estimates from Step 2.

3. Overview of the oaxaca package

The **oaxaca** package consists of the main function `oaxaca()`, which performs the Blinder-Oaxaca decompositions, as well as of a related `plot()` method that produces a bar graph visualization of the decomposition results. In this section, I offer a brief overview of these functions' capabilities. A more detailed description of the arguments and output of both functions can be obtained by typing `?oaxaca` or `?plot.oaxaca` into the R console.

3.1. Decomposition estimation: Main function `oaxaca()`

The main function `oaxaca()` performs both the threefold and the twofold variants of the Blinder-Oaxaca decomposition using observations from the data frame provided in the `data` argument. The linear regression model for the Blinder-Oaxaca decomposition is specified through the `formula` argument. Users can pass on a multiple-part formula that specifies the dependent variable (`y`), the explanatory variables (`x1`, `x2`, `x3`, etc.), as well as an indicator variable (`z`) that indicates whether an observation belongs to Group A (when `z` equals `FALSE` or `0`) or Group B (when it equals `TRUE` or `1`). These variables, along with the functional form of the model, are passed on to the `formula` argument in an object of class `"Formula"` from the **Formula** package (Zeileis and Croissant 2010).

Typically, the model formula takes the following form:

```
y ~ x1 + x2 + x3 + ...  | z
```

If the regression model contains dummies that represent a categorical variable (`d1`, `d2`, `d3`, etc.), these can be specified by adding another part to the formula:

```
y ~ x1 + x2 + x3 + ...  | z | d1 + d2 + d3 + ...
```

When categorical variable dummies are specified, the `oaxaca()` function will automatically adjust estimates to be invariant with respect to the user's choice of the omitted baseline category.

If the user does not include any other arguments, `oaxaca()` will estimate the Blinder-Oaxaca decompositions – both threefold and twofold – based on Ordinary Least Squares regressions (estimated via the standard `lm()` function), and will calculate standard errors based on 100 bootstrapping replicates. By default, `oaxaca()` estimates the twofold decomposition with Group A coefficients, Group B coefficients, their equally weighted average (Reimers 1983), a weighted average that reflects the number of observations in Groups A and B (Cotton 1988), as well with pooled coefficients – both including and excluding the group indicator variable (Neumark 1988; Jann 2008) – as the set of reference coefficients.

These defaults can, however, easily be changed. Users can use the argument `weights` to specify additional relative weights of Group A and Group B coefficients in the estimation of the twofold decomposition. They can also choose, via the `R` argument, how many bootstrapping resamples should be drawn to calculate the standard errors. Last but not least, users can use a difference regression function (argument `reg.fun`) to estimate the regression coefficients used in the decompositions. Note that, if a non-linear function such as `glm()` is chosen, the decomposition will be based on the linear systematic component – usually associated with the estimation of the corresponding latent variable – of the regression method.

The function `oaxaca()` returns an object of class `"oaxaca"`, which can then be passed on to the `plot()` method to obtain a bar graph visualization of the Blinder-Oaxaca decomposition results. The object contains lists named `threefold` and `twofold` which contain the results of the threefold and twofold decompositions, respectively. In addition, the object stores the regression coefficients used in the decomposition (component `beta`), the number of observations in each group that were used in the analysis (`n`), the number of bootstrapping replicates (`R`), the regression objects generated during the analysis (`reg`), as well as the mean values of both the dependent variable (`y`) and the explanatory variables (`x`).

3.2. Visualization: Method `plot()`

The **oaxaca** package can produce easily customizable bar charts that visually summarize the results of its Blinder-Oaxaca decompositions. All bar charts are generated using the **ggplot2** package (Wickham 2009). To visualize the decomposition results, the user simply passes an `"oaxaca"`-class object created by the main function `oaxaca()` to the `plot()` method.

Users can choose which of the estimated decompositions to visualize. The `decomposition` argument determines whether a threefold or a twofold Blinder-Oaxaca decomposition will be shown, while the `type` argument specifies whether the bar graph will contain an overall decomposition or a detailed, variable-by-variable one. If the detailed decomposition type is selected, `component.left` determines whether decomposition components or variable names will be aligned along the left side of the graph. The argument `weight` allows the user to select which of the twofold decompositions should be shown, and the `unexplained.split` argument determines whether the unexplained components ought to be split into the two discrimination subcomponents ("unexplained A" and "unexplained B").

Users can, furthermore, choose which of the variables and decomposition components will be shown (arguments `variables` and `components`), as well as their labels (`variable.labels` and `component.labels`). Standard error bars that indicate confidence intervals can be toggled by the `ci` argument, and the confidence level adjusted by `ci.level`. Several formatting options are available. The bar graph's title can be set using the `title` argument, and axes can be labeled by `xlab` and `ylab`. Finally, users can change the colors of the bars by specifying the `bar.color` argument.

4. Example: Wages of native and foreign-born workers

In this section, I use an empirical example to demonstrate the capabilities of the **oaxaca** package. In particular, I use the Blinder-Oaxaca decomposition to explain the wage gap between native and foreign-born Hispanic workers in metropolitan Chicago. I analyze data from the `chicago` data frame, included in the **oaxaca** package:

```
R> data("chicago")
```

The `chicago` data frame contains information about the demographic characteristics and labor market outcomes of 712 employed Hispanic workers in the Chicago metropolitan area. It is a subset of the 2013 Current Population Survey (CPS) Outgoing Rotation Groups (ORG) data set (Center for Economic and Policy Research 2014). These data have been used extensively in labor economics research (e.g., Holzer and Hlavac 2014).

I am interested in decomposing the wage gap between native and foreign-born workers. The wage gap could be due to group differences in the level of wage determinants such as age, gender or education. Alternatively, the gap could arise from a differential effect of these determinants on native and immigrant workers' wages. I call the `oaxaca()` function to estimate the relative magnitudes of these channels' influence:

```
R> results <- oaxaca(formula = real.wage ~ age + female + LTHS +
+    some.college + college + advanced.degree | foreign.born | LTHS +
+    some.college + college + advanced.degree, data = chicago, R = 1000)
```

As the `formula` argument indicates, the outcome variable in this decomposition is `real.wage`, the worker's real wage denominated in 2013 U.S. dollars. The values of the dependent variable had been obtained by exponentiating the natural logarithm of the workers' real wages (contained in the provided `ln.real.wage` variable):

```
R> chicago$real.wage <- exp(chicago$ln.real.wage)
```

The linear regression model includes covariates that account for the workers' age, gender and education. `LTHS` ("less than high school"), `some.college`, `college` and `advanced.degree` are indicator variables that denote the highest level of education an individual has achieved. A high school education is the omitted baseline category. The variable `foreign.born` indicates whether a worker was born outside of the United States. Group A consists of native workers, and Group B of foreign-born ones. To make sure that the choice of the omitted baseline does not affect the decomposition estimates, the `formula` argument also specifies that the categorical variables denoting the education level ought to be adjusted. Bootstrapped standard errors are calculated based on 1,000 replicates.

```
R> results$n

$n.A
[1] 287

$n.B
[1] 379

$n.pooled
[1] 666
```

The `n` component of the resulting `"oaxaca"`-class object indicates that there are $n_A = 287$ native and $n_B = 379$ foreign-born workers in the analyzed sample. The pooled analysis contains $n_A + n_B = 666$ observations.

```
R> results$y

$y.A
[1] 17.58282

$y.B
[1] 14.56725

$y.diff
[1] 3.015574
```

The y component of the resulting "oaxaca"-class object indicates that the mean real wage is $17.58 for the natives (Group A) and $14.57 for foreign-born workers, leaving the difference of approximately $3.02 to be explained by the Blinder-Oaxaca decomposition.

4.1. Threefold decomposition

First, I look at the results of the threefold Blinder-Oaxaca decomposition:

```
R> results$threefold$overall

coef(endowments)     se(endowments) coef(coefficients)   se(coefficients)
       1.6165339           0.6565025           2.8333261           0.8936198

coef(interaction)    se(interaction)
      -1.4342857           0.7953771
```

The results of the threefold decomposition suggest that, of the $3.02 difference, approximately $1.62 can be attributed to group differences in endowments (i.e., age, gender, education), $2.83 to differences in coefficients, and the remaining -$1.43 is accounted for by the interaction of the two. Next, I examine the endowments and coefficients components of the threefold decomposition variable by variable. This is most easily done by using the plot() method:

```
R> plot(results, components = c("endowments","coefficients"))
```

Figure 1 shows the estimation results for each variable, along with error bars that indicate 95% confidence intervals. In the endowments component, most variables appear to have a statistically insignificant (or only marginally significant) influence, with the sole exception of LTHS. It seems that a significant portion of the native-immigrant wage gap is driven by group differences in the proportion of individuals with less than a high school education.

```
R> summary(results$reg$reg.pooled.2)$coefficients["LTHS",]

   Estimate  Std. Error     t value    Pr(>|t|)
-2.86539843  0.89467794 -3.20271499  0.00142703
```

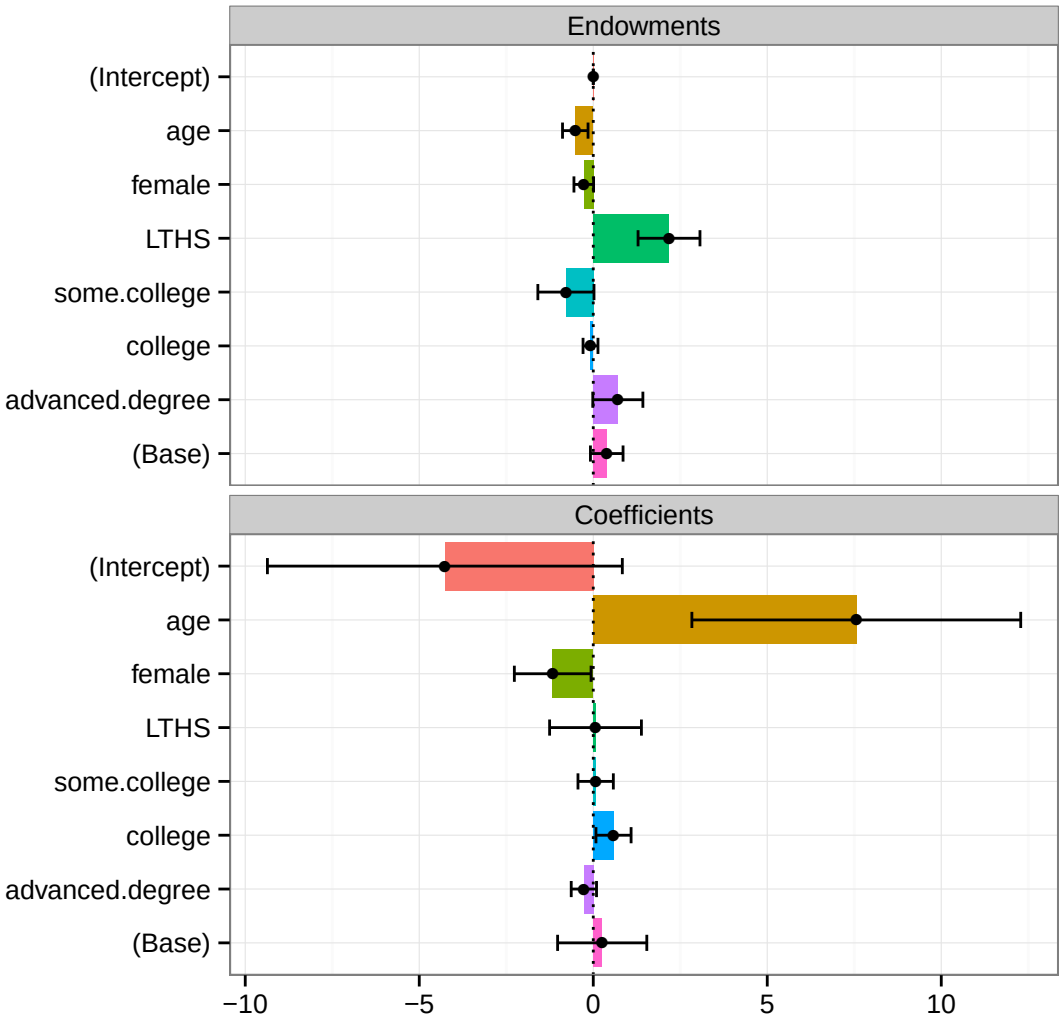

Figure 1: The endowments and coefficients components of a threefold Blinder-Oaxaca decomposition of the native vs. immigrant wage gap.

```
R> results$x$x.mean.diff["LTHS"]

      LTHS
-0.2693959
```

Individuals with less human capital tend to earn less, as can be seen from the pooled regression coefficient on `LTHS` reported above. Furthermore, the value of `x.mean.diff` shows that a greater proportion of foreign-born Hispanic workers have not attained a high school education. The difference in the educational composition of native and immigrant worker groups thus accounts for some portion of the natives' higher wages.

Similarly, most variables are either insignificant or exhibit only marginal statistical significance in the coefficients component. The only variable which achieves clear statistical significance is `age`.

```
R> results$beta$beta.diff["age"]

      age
0.1860063
```

As the difference in the `age` coefficients between natives and immigrants shows, the wage payoff of an additional year of age is greater for U.S.-born Hispanic workers by almost 19 cents. As Figure 1 makes clear, differences in the regression coefficients on `age` account for the decisive portion of the wage gap.

4.2. Twofold decomposition

Next, I look at the results of the twofold Blinder-Oaxaca decomposition. In the output below, the `weight` column indicates the relative weights of coefficients from a regression on observations from Groups A and B, respectively, in the reference coefficient vector $\hat{\boldsymbol{\beta}}_R$. The two negative weights indicate that the reference coefficients come from pooled regressions either without (`-1`) or with (`-2`) the group indicator variable included as a covariate.

```
R> results$twofold$overall

         weight coef(explained) se(explained) coef(unexplained) se(unexplained)
[1,]  0.0000000       1.6165339     0.6565025          1.399040       0.9415643
[2,]  1.0000000       0.1822482     0.7126499          2.833326       0.8936198
[3,]  0.5000000       0.8993911     0.5579216          2.116183       0.8272809
[4,]  0.5800562       0.7845676     0.5669967          2.231007       0.8254721
[5,] -1.0000000       1.3557222     0.5059496          1.659852       0.6589794
[6,] -2.0000000       0.9525717     0.5220180          2.063003       0.8269841
     coef(unexplained A) se(unexplained A) coef(unexplained B) se(unexplained B)
[1,]        1.399040e+00      9.415643e-01           0.0000000         0.0000000
[2,]        0.000000e+00      0.000000e+00           2.8333261         0.8936198
[3,]        6.995202e-01      4.707821e-01           1.4166630         0.4468099
[4,]        5.875183e-01      3.954041e-01           1.6434883         0.5183497
[5,]        9.445705e-01      3.763768e-01           0.7152816         0.2858236
[6,]        4.490852e-14      3.801248e-14           2.0630026         0.8269841
```

For presentational ease, I focus my discussion on the Neumark (1988) decomposition, which uses pooled regression coefficients (from a regression that does not include the group indicator variable `foreign.born`) as the reference coefficient set. The Neumark decomposition is denoted by `-1` in the `weights` column. The results of the overall twofold decomposition indicate that the \$3.02 wage gap between native and foreign-born Hispanic workers can be decomposed into \$1.36 that can be explained by group differences in the explanatory variables and \$1.66 that is unexplained.

Let us assume that the unexplained component of the wage gap occurs due to labor market discrimination, and that the pooled regression coefficients are non-discriminatory. The Blinder-Oaxaca decomposition would then also indicate that \$0.94 of the unexplained part originates from discrimination in favor of native Hispanic workers (component `"unexplained A"`), while \$0.72 comes from discrimination against those who are born outside of the United States (component `"unexplained B"`). The standard errors provide a sense of the uncertainty that accompanies all of the point estimates.

```
R> plot(results, decomposition = "twofold", weight = -1)
```

Figure 2 provides a variable-by-variable twofold decomposition. The results are consistent with the threefold composition. It appears that the wage gap is driven largely by the lower proportion of workers with less than a high school education among the natives (in the explained component) and by the native workers' greater returns to age.

I can explore the unexplained component even further. In Figure 3, I examine three variables from the decomposition – `age`, `female` and `college` – and visualize how much of the unexplained portion of the wage gap can be attributed to discrimination in favor of the natives, and how much is due to discrimination against the immigrants.

```
R> plot(results, decomposition = "twofold", weight = -1,
+    unexplained.split = TRUE, components = c("unexplained A",
+    "unexplained B"), component.labels = c("unexplained A" =
+    "In Favor of Natives", "unexplained B" = "Against the Foreign-Born"),
+    variables = c("age", "female", "college"), variable.labels = c("age" =
+    "Years of Age", "female" = "Female", "college" = "College Education"))
```

I use a variety of `plot()` method arguments to customize the formatting of the resulting bar graph. Through the `components` and `component.labels` arguments, I choose to display only the two subparts – `"unexplained A"` (i.e., discrimination in favor of Group A) and `"unexplained B"` (discrimination against Group B) – of the unexplained decomposition component, and attach appropriate labels to them. Similarly, I use the `variables` and `variable.labels` arguments to select and label the variables I examine.

It appears that only the discrimination components for the `age` variable (labeled `"Years of Age"` in the bar graph) achieve non-marginal statistical significance. The relative size of the bars suggests that – if we assume that the pooled regression coefficients reflect a state of non-discrimination – almost twice as much of the wage gap is explained by discrimination against foreign-born workers as it is by discrimination in favor of native ones.

The comparison would be a little easier to make if the discrimination components bar charts were presented side-by-side for each variable separately. This can be achieved by switching on

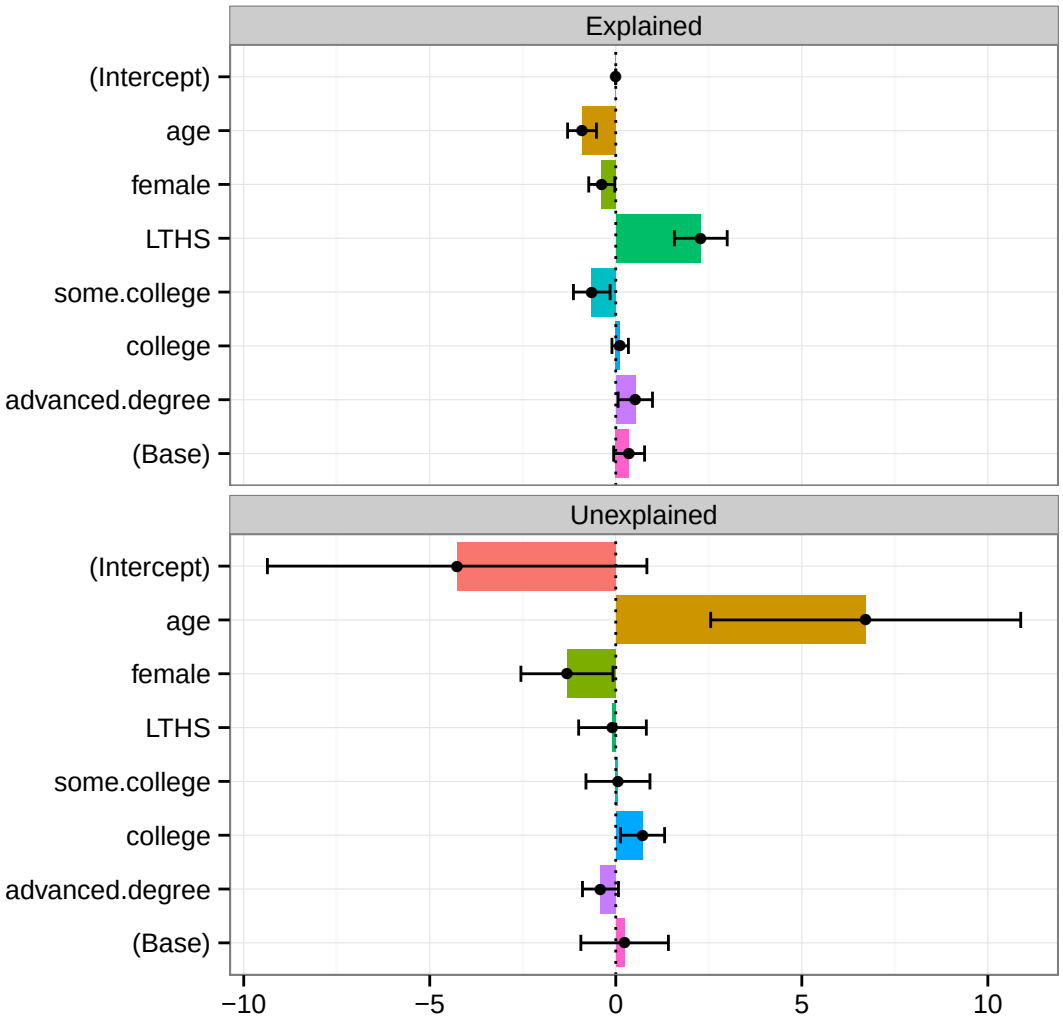

Figure 2: The explained and unexplained components of a twofold Blinder-Oaxaca decomposition of the native vs. immigrant wage gap.

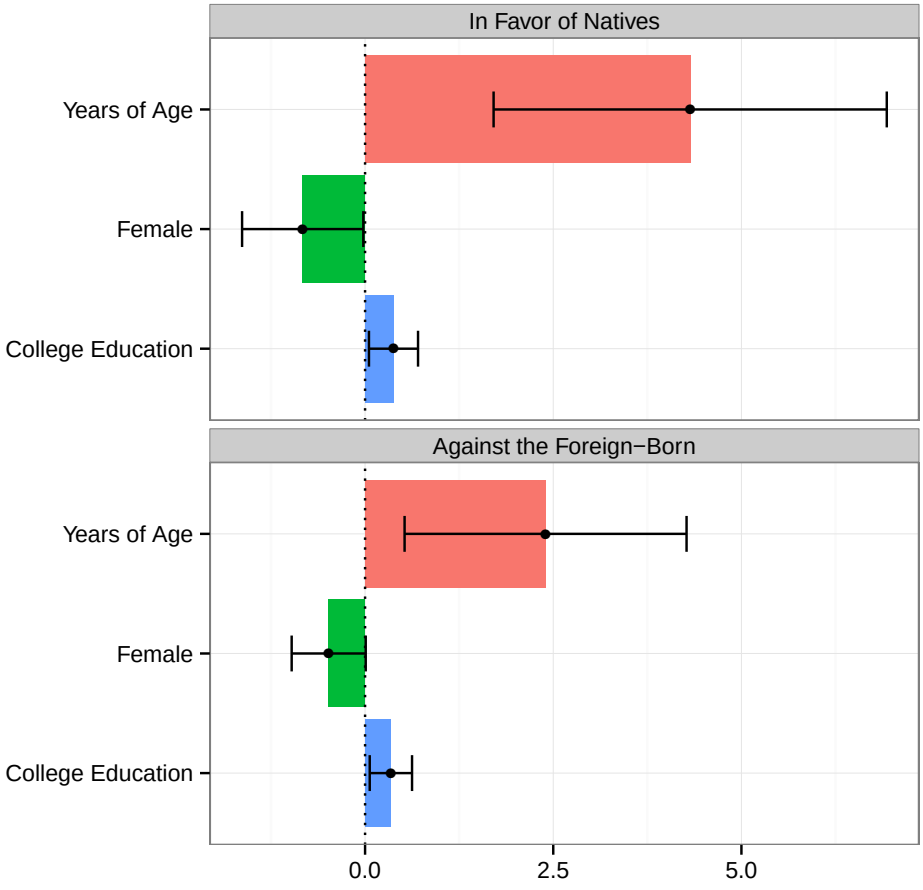

Figure 3: The unexplained portion's discrimination sub-components in a twofold Blinder-Oaxaca decomposition of the native vs. immigrant wage gap.

the `component.left` argument in the `plot()` method. The resulting bar graph is presented in Figure 4.

```
R> plot(results, decomposition = "twofold", weight = -1,
+    unexplained.split = TRUE, components = c("unexplained A",
+    "unexplained B"), component.labels = c("unexplained A" =
+    "In Favor of Natives", "unexplained B" = "Against the Foreign-Born"),
+    component.left = TRUE, variables = c("age","female","college"),
+    variable.labels = c("age" = "Years of Age", "female" = "Female",
+    "college" = "College Education"))
```

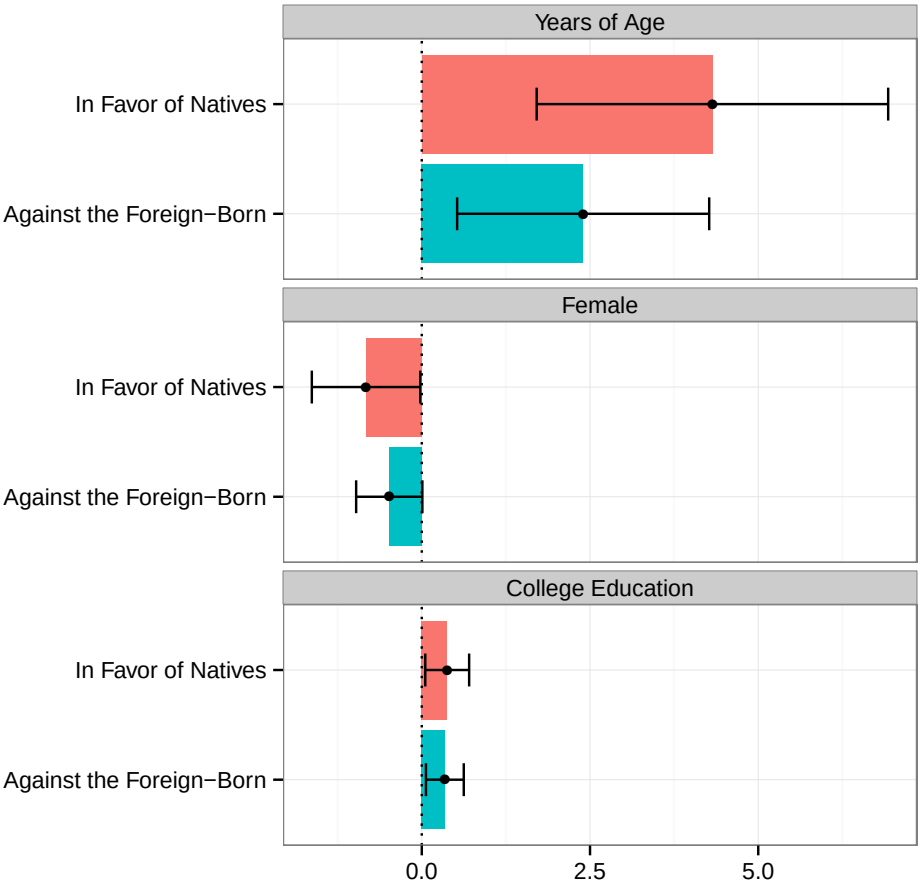

Figure 4: The unexplained portion's discrimination sub-components in a twofold Blinder-Oaxaca decomposition of the native vs. immigrant wage gap. An alternative presentation.

Specific numerical values of the point estimates of the unexplained discrimination components can, of course, be obtained directly from the "oaxaca"-class object:

```
R> variables <- c("age", "female", "college")
R> columns <- c("weight", "coef(unexplained A)", "coef(unexplained B)")
R> results$twofold$variables[[5]][variables, columns]
```

```
        weight coef(unexplained A) coef(unexplained B)
age         -1           4.3191008           2.3980443
female      -1          -0.8285489          -0.4832824
college     -1           0.3777076           0.3428246
```

To summarize, I have used the Blinder-Oaxaca decomposition to examine the wage gap between native and foreign-born Hispanic workers in the Chicago metropolitan area. The results of my analysis suggest that much of the gap can be explained by two facts:

- There are more workers with less than a high school education in the foreign-born group. Workers with a lower stock of human capital tend to command lower wages in the labor market. As a result, the relatively less-educated group of foreign-born Hispanic workers will, on average, earn lower wages than their native counterparts.

- The returns to age are greater for native workers than for the immigrants. In other words, even if the foreign-born workers had the same average age as the natives, the native group would, on average, earn higher wages than immigrants. This result makes some intuitive sense if we interpret age as potentially picking up the effect of labor market experience. The higher returns to age among the natives may, for instance, reflect the differential availability of more lucrative jobs with greater opportunities for career growth.

5. Concluding remarks

In this article, I have introduced the **oaxaca** package for the R statistical programming language. It is the first R package that allows researchers to estimate Blinder-Oaxaca decompositions, a statistical method that decomposes differences in mean outcomes across two groups into a part that is due to group differences in the levels of explanatory variables and a part that is due to differential magnitudes of regression coefficients.

oaxaca estimates threefold and twofold Blinder-Oaxaca decompositions for linear models, and also provides estimates for a detailed, variable-by-variable decomposition. Each point estimate is presented with a bootstrapped standard error that measures the corresponding estimation uncertainty.

I have demonstrated the package's capabilities through an empirical example that examines the wage gap between native and foreign-born Hispanic workers in the Chicago metropolitan area. In doing so, I have also showcased the **oaxaca** package's unique visualization features that allow users to graphically summarize the results of their decompositions.

Acknowledgments

I would like to thank Kai Gehring and Becca Goldstein for helpful comments and suggestions.

References

Bauer T, Göhlmann S, Sinning M (2007). "Gender Differences in Smoking Behavior." *Health Economics*, **16**(9), 895–909.

Bauer TK, Sinning M (2008). "An Extension of the Blinder-Oaxaca Decomposition to Nonlinear Models." *Advances in Statistical Analysis*, **92**(2), 197–206.

Bauer TK, Sinning M (2010). "Blinder-Oaxaca Decomposition for Tobit Models." *Applied Economics*, **42**(12), 1569–1575.

Blinder AS (1973). "Wage Discrimination: Reduced Form and Structural Estimates." *Journal of Human Resources*, **8**(4), 436–455.

Borooah VK, Iyer S (2006). "The Decomposition of Inter-Group Differences in a Logit Model: Extending the Oaxaca-Blinder Approach with an Application to School Enrolment in India." *Journal of Economic and Social Measurement*, **30**(4), 279–293.

Bustamante AV, Fang H, Rizzo JA, Ortega AN (2009). "Heterogeneity in Health Insurance Coverage Among US Latino Adults." *Journal of General Internal Medicine*, **24**(3), 561–566.

Center for Economic and Policy Research (2014). "CPS ORG Uniform Extracts, Version 1.9." URL http://ceprdata.org/cps-uniform-data-extracts/cps-outgoing-rotation-group/.

Comprehensive R Archive Network (CRAN) (2014). "Comprehensive R Archive Network (CRAN)." URL http://cran.us.r-project.org/.

Cotton J (1988). "On the Decomposition of Wage Differentials." *Review of Economics and Statistics*, **70**(2), 236–243.

Darity W, Guilkey DK, Winfrey W (1996). "Explaining Differences in Economic Performance Among Racial and Ethnic Groups in the USA." *American Journal of Economics and Sociology*, **55**(4), 411–425.

Efron B (1979). "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics*, **7**(1), 1–26.

Fairlie RW (2005). "An Extension of the Blinder-Oaxaca Decomposition Technique to Logit and Probit Models." *Journal of Economic and Social Measurement*, **30**(4), 305–316.

Fairlie RW (2013). *Example of Non-Linear Decomposition Technique for Logit Model*. SAS program, URL http://people.ucsc.edu/~rfairlie/decomposition/decompexample_v6.sas.

Gardeazabal J, Ugidos A (2004). "More on Identification in Detailed Wage Decompositions." *Review of Economics and Statistics*, **86**(4), 1034–1036.

Holzer HJ, Hlavac M (2014). *Diversity and Disparities: America Enters a New Century*, chapter A Very Uneven Road: U.S. Labor Markets in the Past Thirty Years. Russell Sage Foundation, New York, NY, USA.

Jann B (2006). ***fairlie**: Stata Module to Generate Nonlinear Decomposition of Binary Outcome Differentials*. Stata module, URL http://econpapers.repec.org/software/bocbocode/s456727.htm.

Jann B (2008). "The Blinder-Oaxaca Decomposition for Linear Regression Models." *Stata Journal*, **8**(4), 453–479.

Kim C (2010). "Decomposing the Change in the Wage Gap Between White and Black Men Over Time, 1980-2005: An Extension of the Blinder-Oaxaca Decomposition Method." *Sociological Methods Research*, **38**(4), 619–651.

LaLonde RJ, Topel RH (1992). *Immigration and the Work Force*, chapter The Assimilation of Immigrants in the U.S. Labor Market. The University of Chicago Press, Chicago, IL, USA.

Munn IA, Hussain A (2010). "Factors Determining Differences in Local Hunting Lease Rates: Insights from Blinder-Oaxaca Decomposition." *Land Economics*, **86**(1), 66–78.

Neumark D (1988). "Employers' Discriminatory Behavior and the Estimation of Wage Discrimination." *Journal of Human Resources*, **23**(3), 279–295.

Oaxaca RL (1973). "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review*, **14**(3), 693–709.

Oaxaca RL, Ransom MR (1999). "Identification in Detailed Wage Decompositions." *Review of Economics and Statistics*, **81**(1), 154–157.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Reimers CW (1983). "Labor Market Discrimination Against Hispanic and Black Men." *Review of Economics and Statistics*, **65**(4), 570–579.

SAS Institute (2014). *SAS/STAT Software*. SAS Institute Inc., Cary, NC, USA. URL http://www.sas.com/en_us/software/analytics/stat.html.

Sinning M, Hahn M, Bauer TK (2008). "The Blinder-Oaxaca Decomposition for Nonlinear Regression Models." *Stata Journal*, **8**(4), 480–492.

Stanley T, Jarrell SB (1998). "Gender Wage Discrimination Bias? A Meta-Regression Analysis." *Journal of Human Resources*, **33**(4), 947–973.

StataCorp (2013). *Stata Statistical Software: Release 13*. StataCorp LP, College Station, TX, USA. URL http://www.stata.com.

Watson I (2010). ***decomp**: Stata Module to Compute Decompositions of Earnings Gap*. Stata module, URL http://fmwww.bc.edu/repec/bocode/d/decomp.ado.

Weichselbaumer D, Winter-Ebmer R (2005). "A Meta-Analysis of the International Gender Wage Gap." *Journal of Economic Surveys*, **19**(3), 479–511.

Wickham H (2009). ***ggplot2**: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, NY, USA.

Zeileis A, Croissant Y (2010). "Extended Model Formulas in R: Multiple Parts and Multiple Responses." *Journal of Statistical Software*, **34**(1), 1–13. URL http://www.jstatsoft.org/v34/i01/.

Affiliation:

Marek Hlavac
Harvard University
John F. Kennedy School of Government
79 John F. Kennedy Street
Cambridge, MA 02138
United States of America
E-mail: hlavac@fas.harvard.edu