Package 'oaxaca'

April 17, 2022

Type Package

Title Blinder-Oaxaca Decomposition

Version 0.1.5

Date 2022-04-17

Author Marek Hlavac <mhlavac@alumni.princeton.edu>

Maintainer Marek Hlavac <mhlavac@alumni.princeton.edu>

Description An implementation of the Blinder-Oaxaca decomposition for linear regression models.

License GPL (>= 2)

Imports Formula, ggplot2, reshape2, methods, stats

LazyData yes

Collate 'oaxaca-internal.R' 'oaxaca.R'

NeedsCompilation no

Repository CRAN

Date/Publication 2022-04-17 19:10:02 UTC

R topics documented:

`chicago` *Labor market and demographic data for employed Hispanic workers in metropolitan Chicago*

Description

Data from a 2013 sample of employed Hispanic workers in metropolitan Chicago. It is a subset of the 2013 Current Population Survey (CPS) Outgoing Rotation Groups (ORG) data set provided by the Center for Economic and Policy Research in Washington, DC (CEPR, 2014).

Usage

```
data("chicago")
```

Format

A data frame containing 712 observations on 9 variables. The 9 variables contain labor market and demographic information on a sample of employed Hispanic workers in the Chicago metropolitan area.

[, 1]	age	the worker's age, expressed in years
[, 2]	female	an indicator for female gender
[, 3]	foreign.born	an indicator for foreign-born status
[, 4]	LTHS	an indicator for having completed less than a high school (LTHS) education
[, 5]	high.school	an indicator for having completed a high school education
[, 6]	some.college	an indicator for having completed some college education
[, 7]	college	an indicator for having completed a college education
[, 8]	advanced.degree	an indicator for having completed an advanced degree
[, 9]	ln.real.wage	the natural logarithm of the worker's real wage (in 2013 U.S. dollars)

Source

Center for Economic and Policy Research (CEPR). 2014. CPS ORG Uniform Extracts, Version 1.9 . Washington, DC.

Examples

```
data("chicago")
summary(chicago)
```

---

oaxaca *Blinder-Oaxaca Decomposition*

Description

`oaxaca` performs a Blinder-Oaxaca decomposition for linear regression models (Blinder, 1973; Oaxaca, 1973). This statistical method decomposes the difference in the means of outcome variables across two groups into a part that is due to cross-group differences in explanatory variables and a part that is due to differences in group-specific coefficients. Economists have used Blinder-Oaxaca decompositions extensively to study labor market discrimination. In principle, however, the method is appropriate for the exploration of cross-group differences in any outcome variable.

The `oaxaca` function allows users to estimate both a threefold and a twofold variant of the decomposition, as described and implemented by Jann (2008). It supports a variety of reference coefficient

weights, as well as pooled model estimation. It can also adjust coefficients on indicator variables to be invariant to the choice of the omitted reference category. Bootstrapped standard errors are calculated (e.g., Efron, 1979). The function returns an object of class "oaxaca" that can be visualized using the plot.oaxaca method.

Usage

```
oaxaca(formula, data, group.weights = NULL, R = 100, reg.fun = lm, ...)
```

Arguments

formula	a formula that specifies the model that the function will run. Typically, the formula is of the following form:
	
	`y ~ x1 + x2 + x3 + ... | z`
	
	where y is the dependent variable, x1 + x2 + x3 + ... are explanatory variables and z is an indicator variable that is TRUE (or equal to 1) when an observation belongs to Group B, and FALSE (or equal to 0) when it belongs to Group A.
	
	The formula can also take on an alternative form:
	
	`y ~ x1 + x2 + x3 + ... | z | d1 + d2 + d3 + ...`
	
	Here, d1 + d2 + d3 + ... are indicator ("dummy") variables that will be adjusted so that the decomposition results do not change depending on the user's choice of the reference category (Gardeazabal and Ugidos, 2004).

data	a data frame containing the data to be used in the Blinder-Oaxaca decomposition.
group.weights	a vector of numeric values between 0 and 1. These values specify the weight given to Group A relative to Group B in determining the reference set of coefficients (Oaxaca and Ransom, 1994). By default, the following weights are included in each estimation: • 0: Group A coefficients used as reference. • 1: Group B coefficients used as reference. • 0.5: Equally weighted average (each 0.5) of Group A and B coefficients used as reference, as in Reimers (1983). • an average of Group A and B coefficients weighted by the number of observations in Group A and B, following Cotton (1988). • -1: Coefficients from a pooled regression (that does not include the group indicator variable) used as reference, as suggested by Neumark (1988). • -2: Coefficients from a pooled regression (that includes the group indicator) used as reference. See Jann (2008).
R	number of bootstrapping replicates for the calculation of standard errors. No bootstrapping is performed when the value of R is set to NULL.
reg.fun	a function that estimates the desired regression model. The function must accept arguments formula and data, and be treated by functions model.frame and

model.matrix, in the same way that the standard functions lm and glm do. Additional arguments can be passed on via the ... argument. By default, an Ordinary Least Squares (OLS) regression is performed via the lm function.

... additional arguments that will be passed on to the regression function specified by reg.fun.

Value

oaxaca returns an object of class "oaxaca". The corresponding summary function (i.e., summary.oaxaca) returns the same object.

An object of class "oaxaca" is a list containing the following components:

beta a list that contains information about the regression coefficients used in estimating the decomposition. If dummy variables d1 + d2 + d3 + ... are specified in the formula argument, this list contains coefficients that have been adjusted to make estimation results invariant to the choice of the omitted baseline category (Gardeazabal and Ugidos, 2004). The beta list contains the following components:

- beta.A: coefficients from a regression on observations in Group A
- beta.B: coefficients from a regression on observations in Group B
- beta.diff: equal to beta.A - beta.B
- beta.B: a matrix that contains the reference coefficients for each of the estimated twofold decompositions

call the matched call.

n a list that contains information about the number of observations used in the analysis. It contains the following components:

- n.A: the number of observations in Group A
- n.B: the number of observations in Group B
- n.pooled: the number of observations in the pooled model that includes both Group A and Group B

R a numeric vector that contains the number of bootstrapping replicates.

reg a list that contains estimated regression objects:

- reg.A: a regression on observations in Group A
- reg.B: a regression on observations in Group B
- reg.pooled.1: a pooled regression that does not include the group indicator variable (Neumark, 1988)
- reg.pooled.2: a pooled regression that does includes the group indicator variable (Jann, 2008)

threefold a list that contains the result of the threefold Blinder-Oaxaca decomposition. It decomposes the difference in mean outcomes into three parts:

- endowments: the contribution of differences in explanatory variables across groups.
- coefficients: part that is due to group differences in the coefficients (or "effect size"). Includes differences in the model intercept.

- `interaction`: part that accounts for the fact that cross-group differences in explanatory variables and coefficients occur at the same time.

The list `threefold` contains two sub-components: `overall` and `variables`. The former is a numeric vector that stores results - coefficients (`coef`) and standard errors (`se`) - for the overall decomposition of the difference in outcomes into the three parts described above. The latter is a numeric matrix that contains the results of a variable-by-variable threefold Blinder-Oaxaca decomposition.

`twofold` a list that contains the result of the twofold Blinder-Oaxaca decomposition. It decomposes the difference in mean outcomes into two parts:

- `explained`: the portion that is explained by cross-group differences in the explanatory variables.
- `unexplained`: the remaining part that is not explained by differences in the explanatory variables. Often attributed to discrimination, but may also result from the influence of unobserved variables.

The `unexplained` part can be further decomposed into two sub-parts, `unexplained A` and `unexplained B`, that represent discrimination in favor of Group A and against Group B, respectively. See Jann (2008) for details on these sub-parts' interpretation.

The list `twofold` contains two sub-components: `overall` and `variables`. The former is a numeric matrix that stores results - coefficients (`coef`) and standard errors (`se`) - for the overall decomposition of the difference in outcomes into the two parts described above. The latter is a list of numeric matrices that contains the results of a variable-by-variable twofold Blinder-Oaxaca decomposition. In all matrices, the `weight` column indicates the weight given to Group A relative to Group B in determining the reference coefficients.

`x` a list that contains:

- `x.mean.A`: the mean values of explanatory variables for Group A
- `x.mean.B`: the mean values of explanatory variables for Group B
- `x.mean.diff`: equal to `x.mean.A - x.mean.B`

`y` a list that contains the mean values of the dependent variable (i.e., the outcome variable). It contains the following components:

- `y.A`: the mean outcome value for observations in Group A
- `y.B`: the mean outcome value for observations in Group B
- `y.diff`: the difference between the mean outcomes values in Groups A and B. Equal to `y.A - y.B`.

Please cite as:

Hlavac, Marek (2022). oaxaca: Blinder-Oaxaca Decomposition in R.
R package version 0.1.5. https://CRAN.R-project.org/package=oaxaca

Author(s)

Dr. Marek Hlavac < mhlavac at alumni.princeton.edu >
Social Policy Institute, Bratislava, Slovakia

References

Blinder, Alan S. (1973). Wage Discrimination: Reduced Form and Structural Estimates. *Journal of Human Resources*, 8(4), 436-455.

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

Efron, Bradley. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, 7(1), 1-26.

Gardeazabal, Javier and Arantza Ugidos. (2004). More on Identification in Detailed Wage Decompositions. *Review of Economics and Statistics*, 86(4), 1034-1036.

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

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

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

Oaxaca, Ronald L. and Michael R. Ransom. (1994). On Discrimination and the Decomposition of Wage Differentials. *Journal of Econometrics*, 61(1), 5-21.

Reimers, Cordelia W. (1983). Labor Market Discrimination Against Hispanic and Black Men. *Review of Economics and Statistics*, 65(4), 570-579.

See Also

plot.oaxaca

Examples

```
# set random seed
set.seed(03104)

# load data set of Hispanic workers in Chicago
data("chicago")

# perform Blinder-Oaxaca Decomposition:
# explain differences in log real wages across native and foreign-born groups
oaxaca.results.1 <- oaxaca(ln.real.wage ~ age + female + LTHS + some.college +
                                          college + advanced.degree | foreign.born,
                           data = chicago, R = 30)

# print the results
print(oaxaca.results.1)

# Next:
# - adjust gender and education dummy variable coefficients to make results
#   invariant to the choice of omitted baseline (reference category)
# - include additional weights for the twofold decomposition that give
#   weights of 0.2 and 0.4 to Group A relative to Group B in the choice
#   of reference coefficients
```

```
oaxaca.results.2 <- oaxaca(ln.real.wage ~ age + female + LTHS + some.college +
                                          college + advanced.degree | foreign.born |
                                        LTHS + some.college + college + advanced.degree,
                           data = chicago, group.weights = c(0.2, 0.4), R = 30)

# plot the results
plot(oaxaca.results.2)
```

`plot.oaxaca` *Coefficient Bar Plots for the Blinder-Oaxaca Decomposition*

Description

`plot.oaxaca` is used to generate a set of coefficient bar plots that present the results of a Blinder-Oaxaca decomposition graphically.

Usage

```
## S3 method for class 'oaxaca'
plot(x, decomposition = "threefold", type = "variables", group.weight = NULL,
     unexplained.split = FALSE, variables = NULL, components = NULL,
     component.left = FALSE, component.labels = NULL, variable.labels = NULL,
     ci = TRUE, ci.level = 0.95,
     title = "", xlab = "", ylab = "", bar.color = NULL, ...)
```

Arguments

`x`	an object of class `"oaxaca"`, typically generated by the `oaxaca` function.
`decomposition`	specifies which type of decomposition will be presented. Can be either `"threefold"` (default) or `"twofold"`.
`type`	specifies whether the results of an overall decomposition or a variable-by-variable decomposition will be presented. Can be either `"variables"` (default) or `"overall"`.
`group.weight`	a numeric value that specifies the group weight for which the twofold decomposition results will be presented. Only relevant when argument `decomposition` is set to `"twofold"`.
`unexplained.split`	a logical value that toggles whether, in the twofold decomposition, the presentation of the `unexplained` component will be split into the `unexplained A` and `unexplained B` parts. See `oaxaca` for details.
`variables`	a character vector that specifies the variables for which coefficient bar plots are requested. If NULL, plots for all variables will be produced. Only relevant when argument `type` is set to `"variables"`. This argument can also be used to determine the order in which variables are presented.

components a character vector that specifies which decomposition components will be presented. For threefold decomposition, must be a subset of "endowments", "coefficients" and "interaction". For twofold decomposition, must be a subset of "explained", "unexplained" and - if argument unexplained.split is set to TRUE - "unexplained A" and "unexplained B". This argument can also be used to determine the order in which decomposition components are presented.

component.left a logical value that specifies whether the decomposition components will be presented along the left side of the coefficient bar plot. By default, the argument is set to FALSE, and the left side of the plot lists individual variables. Only relevant when argument type is set to "variables".

component.labels
a named character vector that specifies custom labels for individual decomposition components. The character vector elements contain the new labels, while the elements' names must correspond to the appropriate decomposition components: "endowments", "coefficients", "interaction", "explained", "unexplained", "unexplained A", "unexplained B".

variable.labels
a named character vector that specifies custom labels for the presented variables. The character vector elements contain the new labels, while the elements' names must contain the appropriate variable name. Only relevant when argument type is set to "variables".

ci a logical value that toggles the presentation of confidence intervals (standard error bars) in the coefficient bar plots.

ci.level a numeric value between 0 and 1 that specifies the confidence level for the presented confidence intervals.

title a character string that contains the title of the coefficient bar plot.

xlab a character string that specifies the horizontal axis label.

ylab a character string that specifies the vertical axis label.

bar.color a named vector that specifies the color of bars in the plot. The vector elements' names contain a string that identifies either the variable or the decomposition component that the color will be applied to. If no names are specifies, the bars will be colored from top to bottom.

... additional arguments passed on to the aesthetic mapping (aes) of the ggplot function (Wickham, 2009) used to generate the coefficient bar plots inside the method.

Please cite as:

Hlavac, Marek (2022). oaxaca: Blinder-Oaxaca Decomposition in R.
R package version 0.1.5. https://CRAN.R-project.org/package=oaxaca

Author(s)

Dr. Marek Hlavac < mhlavac at alumni.princeton.edu >
Social Policy Institute, Bratislava, Slovakia

References

Wickham, Hadley. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer Science & Business Media.

See Also

oaxaca

Examples

```
# set random seed
set.seed(08544)

# load data set of Hispanic workers in Chicago
data("chicago")

# perform Blinder-Oaxaca Decomposition:
# explain differences in log real wages across native and foreign-born groups
oaxaca.results <- oaxaca(ln.real.wage ~ age + female + LTHS + some.college +
                                        college + advanced.degree | foreign.born,
                         data = chicago, R = 50)

# plot results of the threefold decomposition, variable-by-variable
# only include educational variables
# decomposition components along the left side of the plot
plot(oaxaca.results, component.left = TRUE,
     variables = c("LTHS", "some.college", "college", "advanced.degree"),
     variable.labels = c("LTHS" = "less than high school",
                         "some.college" = "some college",
                         "advanced.degree" = "advanced degree"))

# plot results of the twofold decomposition (overall results)
# equal weight for Group A and B in reference coefficient determinantion (weight = 0.5)
# unexplained portion split into A and B
plot(oaxaca.results, decomposition = "twofold", type = "overall",
     group.weight = 0.5, unexplained.split = TRUE,
     bar.color = c("limegreen", "hotpink", "steelblue"))
```

Index