

A Simple SAS® Macro to Perform Blinder-Oaxaca Decomposition

Taylor Lewis¹, U.S. Office of Personnel Management

Stanislas Ezoua¹, U.S. Office of Personnel Management

ABSTRACT

Blinder-Oaxaca decomposition is a straightforward statistical method that emerged in the econometrics literature as a way to explain differences observed between groups with respect to the mean of a continuous variable. To give a few examples, the procedure has been used to investigate potential pay differentials between males and females, or whether health disparities exist amongst individuals of varying socioeconomic statuses. In this paper, we provide background on the fundamental concepts and objectives behind Blinder-Oaxaca decomposition, and present a general-purpose macro that analysts interested in conducting the technique might find helpful.

1. BACKGROUND

This article presents a simple SAS® macro to carry out an econometric technique to decompose the overall difference in the mean of a continuous variable for two disjoint groups. The notion behind the approach is to independently model the outcome in either group as a function of observed covariates, and then rearrange terms such that the difference is partitioned into an “explained” portion and an “unexplained” portion. Because the technique was introduced in two contemporaneous works—one by Blinder (1973) and another by Oaxaca (1973)—we refer to it in this paper as *Blinder-Oaxaca decomposition*, although we acknowledge and advise the reader that there appears to be no universally accepted nomenclature in the literature. Sometimes scholars credit only one of the authors, and other times they refer to it as Oaxaca-Blinder decomposition.

Blinder-Oaxaca decomposition is a popular tool for investigating the existence of potential gender wage disparities (Weichselbaumer and Winter-Ebmer, 2005). Indeed, the technique has been employed for that purpose by both the U.S. Government Accountability Office (GAO, 2009) and the United States Office of Personnel Management (OPM, 2014) within the realm of the federal workforce. For the OPM (2014) study, an in-house SAS macro was developed to facilitate the task of conducting numerous iterations of the method amongst various covariate combinations and population domains. The macro presented in this paper is a generalized version of that in-house macro, something we hope analyst can adapt to other contexts, either in its current form or with minimal modifications.

The paper is structured as follows. In Section 2 we introduce the formal mathematical details of the method; in Section 3 we introduce our macro and demonstrate its use with a fictitious, example data set. Finally, in Section 4 we discuss enhancements to the macro we intend to make in the future.

2. METHOD

In this section we walk through the formal mathematical derivation for one of the most basic forms of Blinder-Oaxaca decomposition, the one we have programmed into the macro presented in this paper. The derivation follows what is outlined in Gosse (2002) and Busch and Holst (2009). Be advised that there are numerous variants and extensions to the general approach we present. For more background on these alternatives, we refer the reader to Jann (2008), O'Donnell et al. (2008), Fortin et al. (2011), and Hlavac (2014).

Suppose there are two mutually exclusive groups (e.g., males and females), which we can denote as A and B , and the aim is to measure the difference in the two groups' respective mean outcomes, $\bar{y}^A - \bar{y}^B$, for some continuous variable y (e.g., wages). Rather than estimating this quantity simply by subtracting the two respective marginal means, we might posit that the two population means are governed by a linear function of p parameters associated with a set of explanatory factors (e.g., demographics, measures of human capital). More formally, we would assume that the value of the outcome for unit i in group g ($g = A, B$) can be modeled as

$$y_i^g = \beta_0^g + \sum_{j=1}^p \beta_j^g x_{ji}^g + \varepsilon_i^g \quad (1)$$

where ε_i^g is a residual term assumed to be normally distributed with mean zero and constant variance.

In this setting, the difference between the two groups' means can be expressed as

¹ The opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect those of the U.S. Office of Personnel Management.

$$\bar{y}^A - \bar{y}^B = \left(\beta_0^A + \sum_{j=1}^p \beta_j^A \bar{x}_j^A \right) - \left(\beta_0^B + \sum_{j=1}^p \beta_j^B \bar{x}_j^B \right) \quad (2)$$

where $\bar{x}_j^g$ represents the mean of predictor variable j for group g . In the case of a categorical predictor variable, this could represent the mean of a 0/1 indicator variable, which can be interpreted as the proportion of cases falling within the given level of that variable.

The next step is to add and subtract the term $\sum_{j=1}^p \beta_j^A \bar{x}_j^B$ to the right side of the equation as follows:

$$\bar{y}^A - \bar{y}^B = \left(\beta_0^A + \sum_{j=1}^p \beta_j^A \bar{x}_j^A \right) - \left(\beta_0^B + \sum_{j=1}^p \beta_j^B \bar{x}_j^B \right) + \left(\sum_{j=1}^p \beta_j^A \bar{x}_j^B - \sum_{j=1}^p \beta_j^B \bar{x}_j^B \right) \quad (3)$$

We can conceive of the term $\sum_{j=1}^p \beta_j^A \bar{x}_j^B$ as a *counterfactual*, the projection of group B's covariates into group A's environment. For example, in the context of wage discrimination, if group A refers to males and group B refers to females, this term represents the wages females would have earned under the males' wage structure.

From here, we rearrange terms to arrive at the expression

$$\bar{y}^A - \bar{y}^B = \sum_{j=1}^p \beta_j^A (\bar{x}_j^A - \bar{x}_j^B) - \sum_{j=1}^p (\beta_j^B - \beta_j^A) \bar{x}_j^B + (\beta_0^A - \beta_0^B) \quad (4)$$

The terms on the right side of equation (4) can be decomposed into two components:

1. The *endowment effect*, $\sum_{j=1}^p \beta_j^A (\bar{x}_j^A - \bar{x}_j^B)$, which represents the “explained” portion of the difference, one accounted for by covariate differences.
2. The *residual effect*, $\sum_{j=1}^p (\beta_j^B - \beta_j^A) \bar{x}_j^B + (\beta_0^A - \beta_0^B)$, which represents the “unexplained” complementary portion of the difference. While this portion is often referred to as the inherent “discrimination,” it is important to bear in mind that this assumes there are no lingering, unobserved factors at play—in other words, that the two models linking the p explanatory variables to the expected outcome are correctly specified.

Note that the decomposition in equation (4) is given from the perspective of group A, which is to say group A is the *reference group*, the environment used to assess the existence of any kind of gap. This is because the endowment effect measures the expected change in group A's outcome if group A had group B's covariates. Be advised that, although our macro uses the expression in equation (4), one can derive an alternative specification with group B as the reference group (see Appendix 5 of Gosse, 2002). Changing the reference group does not have the mere effect of, say, changing the signs of terms; rather, it can change results dramatically.

As may be evident from the formulations above, the respective endowment and residual effects can be partitioned further by particular predictor variables and also by particular levels within a categorical predictor variable. Our macro performs these more granular decompositions *en route* to the overall decomposition. Also, the quantities in equation (4) are only known if there is information available for the entire population. This actually was the case with both the GAO (2009) and OPM (2014) studies, but one can substitute in sample-based estimates otherwise.

3. THE %BO_DECOMP MACRO

In this section we discuss the structure and parameters of our macro, which have named `%BO_decomp`. We then illustrate its execution using an example data set, and discuss how the generated output is can be interpreted. The syntax to actually compile the macro is located in the appendix.

Table 1 summarizes the six parameters of our macro. The first is the `dsname=`, which is used to specify the input data set, assumed to contain case-level data to be modeled—that is, the input to the macro is raw data, not model parameters. The continuous outcome measure is specified via the `depvar=` parameter. The purpose of the `groupvarref=` parameter is to identify the variable on the input data set used to define the two groups. For example, specifying `groupvar=(SEX='M')` would assign group A as males, and all other cases (i.e., females) as group B.

You can provide a list of numeric predictor variables in the `numvar=` parameter, and a list of categorical predictor variables in the `charvar=` parameters. At least one of these two parameters must be non-missing for the macro to function properly. The underlying modeling procedure exploited is PROC GLM, which defaults to assigning a categorical predictor variable's reference level to be the last in the sort-ordered listing of levels. Optionally, you can specify one or more reference levels to override this default, using syntax such as `charvarref=supervs (ref='N')`.

Parameter	Purpose
dsname	Specify input data set
depvar	Specify continuous outcome measure on input data set (e.g., <code>depvar=WAGES</code>)
groupvarref	Specify how to define reference group (e.g., <code>groupvarref=(sex='M')</code>)
numvar	List of numeric variables in the model (e.g., <code>numvar=AGE YEARS_EDUC</code>) (if applicable)
charvar	List of character variables in the model (e.g., <code>charvar=MINORITY SUPERVISOR</code>) (if applicable)
charvarref	(Optional) List of reference levels for particular character variables in the model (e.g., <code>charvarref=SUPERVISOR (REF='N')</code>)

Table 1. Parameters of the %BO_decomp Macro

In essence, the `%BO_decomp` macro uses this information to fit two linear regression models, one for each group, and then consolidates the results to output two tabular summaries.

Let us now walk through a simple example utilizing the macro. The training data set PAY consists of record-level wage information for 3,209 employees of a hypothetical organization. Variables pertinent to the present exposition include:

- `LOG_SALARY` – the natural logarithm of the employee's annual salary. (The natural logarithm transformation is frequently used in the wage discrimination literature to help induce normality in the residual terms of a model, since the distribution of untransformed wages tends to be heavily right-skewed.)
- `SEX` – employee gender, coded M/F.
- `MINORITY` – indicator of whether employee is of a minority race and/or ethnicity.
- `AGE` – numeric variable representing an employee's age.
- `WORKUNIT` – agency work unit, AG01, AG02, AG03, or AG04.

Suppose the objective is to model `LOG_SALARY` for each gender as a function of `AGE`, `MINORITY`, and `WORKUNIT`. The program below demonstrates syntax to execute our version of Blinder-Oaxaca decomposition under this assumed model. As is typically done in the wage discrimination literature, we have opted to treat males as group A, the reference group. In addition, we have instructed the macro to omit the dummy variable associated with employees who have `MINORITY='N'`, such that the non-omitted level can be interpreted as the percent change in an employee's annual salary relative to non-minority employees.

From the first summary we find the overall gap is positive at 0.068, meaning $\bar{y}^A > \bar{y}^B$, or that the average log wages of men is greater than that of females. The explained portion of the gap is 0.011, whereas the unexplained portion is 0.057. Hence, the portion of the gap that can be explained by covariates is $0.011 / 0.068 = 15.75\%$. Furthermore, we can observe from the overall summary that an employee's work unit contributes most in explaining the gap, whereas minority status and age actually detract from the gap (i.e., work in the opposite direction, in effect narrowing the gap), albeit modestly. Still, with 85% of the gap unaccounted for, it would be prudent to seek out additional variables such as education level or occupation type that might help explain a larger portion of the observed difference.

```
%BO_decomp(
  dsname=pay,
  depvar=log_salary,
  groupvarref=(sex='M'),
  numvar=age,
  charvar=minority workunit,
  charvarref=minority (ref='N')
);
```

1) Overall Summary								
Level	Gap	Exp.	Unexp.	% Exp.				
Overall	0.068	0.011	0.057	15.75%				
Intercept	-0.084	0.000	-0.084	0.00%				
Minority	-0.018	-0.000	-0.017	(0.27%)				
age	0.231	-0.002	0.233	(3.29%)				
workunit	-0.061	0.013	-0.074	19.30%				
2) Variable-Specific Summary								
Parameter	Coef A	Mean A	Coef B	Mean B	Gap	Exp.	Unexp.	% Exp.
Intercept	11.137	1.000	11.053	1.000	-0.084	0.000	-0.084	0.00%
Minority_N	0.000	0.646	0.000	0.789	0.000	0.000	0.000	0.00%
Minority_Y	-0.001	0.354	-0.084	0.211	-0.018	-0.000	-0.017	(0.27%)
age	0.006	46.194	0.011	46.555	0.231	-0.002	0.233	(3.29%)
workunit_AG01	-0.021	0.070	-0.054	0.120	-0.003	0.001	-0.004	1.58%
workunit_AG02	0.126	0.085	-0.008	0.049	-0.002	0.005	-0.007	6.85%
workunit_AG03	-0.215	0.731	-0.298	0.765	-0.057	0.007	-0.064	10.87%
workunit_AG04	0.000	0.114	0.000	0.066	0.000	0.000	0.000	0.00%

The second summary provides a more granular breakdown of the figures comprising the overall gap, and also reports the $\bar{x}_j^g$'s, or the mean values of the predictor variables by group, in the columns labeled "Mean A" and "Mean B," as well as the two sets of model parameters in the columns labeled "Coef A" and "Coef B." For dichotomous and continuous predictors (e.g., MINORITY and AGE), the two sets of gap measures are equivalent in either summary table, but the explained and unexplained portions of the gap are summed for the $K - 1$ non-omitted parameters of a categorical predictor variable with $K (> 2)$ categories to arrive at the figures given in the overall summary table. Also, note that the intercept's impact on the gap is strictly unexplained and only affects the overall gap measure.

To help edify how the pieces are combined to arrive at the overall gap figures, let us consider employees in work unit AG03. Relative to employees working in AG04, the reference category, employees working in AG03 tend to make less, which is somewhat more pronounced for females than males (-0.298 versus -0.215). In addition, a larger share of females is employed by that division than males (76.5% versus 73.1%). The end result is a contribution of $(-0.215) \times (0.731 - 0.765) = 0.007$ to the gap, or $.007 / 0.068 = 10.87\%$. Analogous computations are made for the other model parameters.

4. BRIEF DISCUSSION AND IDEAS FOR FUTURE ENHANCEMENTS

In this paper we presented a simple SAS macro to perform Blinder-Oaxaca decomposition, an econometric technique to decompose the mean difference between two groups into factor-specific explained and unexplained components. The larger the overall explained portion, the weaker the probability that some form of systematic discrimination is occurring. Of course, this is predicated on the assumption that the underlying models are correctly specified— influential factors unobserved or otherwise omitted from the model can invalidate any interpretations to be gleaned from the technique.

Considering we refer to our macro as "simple," it is perhaps of little surprise that we can enumerate several worthwhile enhancements that could be made to it. We hope to follow up in the future with an updated version of the macro incorporating options for the user to implement one or more of the following:

- Exponentiate model results to convert from natural logarithm scale to original scale (if applicable).
- Include more sophisticated extensions and variants to the basic decomposition we derived in Section 2 as discussed in Jann (2008), O'Donnell et al. (2008), Fortin et al. (2011), and Hlavac (2014).
- Allow for alternatives to treating group A's model coefficients as the reference, such as those proposed by Reimers (1983), Cotton (1988), Neumark (1988), and Oaxaca and Ransom (1994).
- As Oaxaca and Ransom (1999) note, results of the technique can be sensitive to which dummy variable is omitted for a K -category predictor variable. An alternative parameterization proposed by Gardeazabal and Ugidos (2004) makes results invariant to the choice of a categorical predictor variable's reference level.

- Currently, our macro is only applicable to continuous outcome variables, but Fairlie (2005) discusses extensions to logit and probit models for binary outcome variables.
- Mentioned earlier, the applications of Blinder-Oaxaca decomposition discussed in GAO (2009) and OPM (2014)—the latter being the impetus for this paper—covered an entire finite population, in which case there was no sampling error. A nonparametric bootstrap (Efron and Tibshirani, 1993; Lohr, 2012) could be embedded to facilitate making inferences when data is only available for a sample.
- Produce simple visualizations of the decomposition in tandem with measures of uncertainty (if applicable) such as those shown in Hlavac (2014).

REFERENCES

- Blinder, A. (1973). "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, **8**, pp. 436 – 455.
- Busch, A., and E. Holst. (2009). "Glass Ceiling Effect and Earnings – The Gender Pay Gap in Managerial Positions in Germany," Discussion Paper 905 of the Deutsches Institut für Wirtschaftsforschung. Available online at: <http://www.diw.de/documents/publikationen/73/99981/dp905.pdf>.
- Cotton, J. (1988). "On the Decomposition of Wage Differentials," *Review of Economics and Statistics*, **70**, pp. 236 – 243.
- Efron, B., and R. Tibshirani. (1993). *An Introduction to the Bootstrap*. New York, NY: Chapman and Hall.
- Fairlie, R. (2005). "An Extension of the Blinder-Oaxaca Decomposition Technique to Logit and Probit Models," *Journal of Economic and Social Measurement*, **30**, pp. 305 – 316.
- Fortin, N., Lemieux, T., and S. Firpo. (2011). "Decomposition Methods in Economics," in *Handbook of Labor Economics Volume 4A*, by Ashenfelter, O., and D. Card (eds.), pp. 1 – 102.
- Gardeazabal, J., and A. Ugidos. (2004). "More on Identification in Detailed Wage Decompositions," *Review of Economics and Statistics*, **86**, pp. 1034 – 1036.
- Gosse, M. (2002). "Working Paper: The Gender Pay Gap in the New Zealand Public Service," Working Paper Number 15 of the State Services Commission. Available online at: <https://www.ssc.govt.nz/node/7421>.
- Hlavac, M. (2014). "Oaxaca: Blinder-Oaxaca Decomposition in R," Available online at SSRN: <http://ssrn.com/abstract=2528391>.
- Jann, B. (2008). "The Blinder–Oaxaca Decomposition for Linear Regression Models," *The Stata Journal*, **8**, pp. 453 – 479.
- Lohr, S. (2012). "Using SAS® for the Design, Analysis, and Visualization of Complex Surveys," *Proceedings of the SAS Global Forum*. Available on-line at: <http://support.sas.com/resources/papers/proceedings12/343-2012.pdf>.
- Neumark, D. (1988). "Employers' Discriminatory Behavior and the Estimation of Wage Discrimination," *Journal of Human Resources*, **23**, pp. 279 – 295.
- O'Donnell, O., van Doorslaer, E., Wagstaff, A., and M. Lindelow. (2008). *Analyzing Health Equity Using Household Survey Data: A Guide to Techniques and Their Implementation*. Washington, DC: The World Bank.
- Oaxaca, R. 1973. "Male–Female Wage Differentials in Urban Labor Markets," *International Economic Review*, **14**, pp. 693 – 709.
- Oaxaca, R. and M. Ransom. (1994). "On Discrimination and the Decomposition of Wage Differentials," *Journal of Econometrics*, **61**, pp. 5 – 21.
- Oaxaca, R. and M. Ransom. (1999). "Identification in Detailed Wage Decompositions," *Review of Economics and Statistics*, **81**, pp. 154 – 157.
- Reimers, C. (1983). "Labor Market Discrimination Against Hispanic and Black Men," *Review of Economics and Statistics*, **65**, pp. 570 – 579.
- United States Government Accountability Office. (2009). "Women's Pay: Gender Pay Gap in the Federal Workforce Narrows as Differences in Occupation, Education, and Experience Diminish," Report GAO-09-279. Washington, DC: Government Printing Office.
- United States Office of Personnel Management. (2014). "Governmentwide Strategy on Advancing Pay Equality in the Federal Government" Available online at: <https://www.opm.gov/policy-data-oversight/pay-leave/reference-materials/reports/governmentwide-strategy-on-advancing-pay-equality-in-the-federal-government.pdf>.

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

APPENDIX. CODE TO COMPILE THE %BO_DECOMP MACRO

```
%macro BO_decomp
(
  dsname=,          /* input data set */
  depvar=,          /* dependent variable */
  groupvarref=,     /* reference definition for the dependent variable */
  numvar=,          /* list of continuous predictor variables (leave blank if none) */
  charvar=,         /* list of character variables (leave blank if none) */
  charvarref=       /* list reference group definitions(s) (leave blank if none)
                    for example: supervs (ref='Y') */
);
ods listing close;
%let nocharvar=%str( );
%if charvar eq %then &nocharvar=*;
* create a dichotomous A/B grouping variable where B denotes reference group;
data &dsname._bo;
  length groupvar $1;
  set &dsname;
  if &groupvarref then groupvar='A';
  else groupvar='B';
run;

* create macro variables housing associated counts of &groupvar levels;
proc sql noprint;
  select count(*) into :groupvarA_count from &dsname._bo where groupvar="A";
  select count(*) into :groupvarB_count from &dsname._bo where groupvar="B";
quit;

* fit regression models separately for each group;
proc sort data=&dsname._bo; by groupvar; run;
ods output ParameterEstimates=betas;
proc glm data=&dsname._bo;
  by groupvar;
  &nocharvar class &charvarref &charvar;
  model log_salary = &numvar /* numeric variables */
                &charvar /* character variables (w/o the (ref=' ') */
                / solution;
quit;

* re-orient group-specific parameter estimates;
proc sort data=betas; by Parameter; run;
data betas;
  length Parameter $200;
  merge betas (where=(groupvar="A") rename=(Estimate=Estimate_A))
        betas (where=(groupvar="B") rename=(Estimate=Estimate_B));
  by Parameter;
keep Dependent Parameter Estimat;;
Parameter=trim(scan(Parameter,1))||'_'||trim(scan(Parameter,2));
Parameter=compress(Parameter);
if scan(Parameter,2,"_")=' ' then Parameter=scan(Parameter,1,"_");
run;

* merge in means of the predictor variables;
ods output statistics=means0
      domain=means;
proc surveymeans data=&dsname._bo mean;
  var &numvar
      &charvar;
  domain groupvar;
run;
data means;
```

```

length Parameter $200;
merge means0
      means (where=(groupvar="A") rename=(Mean=Mean_A))
      means (where=(groupvar="B") rename=(Mean=Mean_B));
Parameter=trim(VarName)||'_'||trim(VarLevel);
Parameter=compress(Parameter);
if scan(Parameter,2,"_") in(' ' '.') then Parameter=scan(Parameter,1,"_");
keep Parameter Mea;
run;
proc sort data=means; by Parameter; run;

* create a summary table;
data all;
  merge betas
        means;
  by Parameter;
  if Mean =. then Mean=1;
  if Mean_A=. then Mean_A=1;
  if Mean_B=. then Mean_B=1;
  explained =Estimate_A*(Mean_A - Mean_B);
  unexplained=Mean_B*(Estimate_B - Estimate_A);
run;
* sum components for multi-factor effects;
data sums;
  set all;
  Parameter=scan(Parameter,1,"_");
run;
proc means data=sums noprint nway;
  class Parameter;
  var explained unexplained;
output out=sums (keep=Parameter explained unexplained) sum=;
run;
proc means data=sums noprint nway;
  var explained unexplained;
output out=sums0 (keep=explained unexplained) sum=;
run;
proc sql noprint;
  select explained into :sum_e from sums0;
  select unexplained into :sum_u from sums0;
quit;
data all;
length Parameter $200;
  set all
        sums
        sums0;
  pct_explained =explained /(&sum_e+&sum_u);
  pct_unexplained=unexplained/(&sum_e+&sum_u);
run;
proc sort data=all; by Parameter Estimate_A; run;
data all;
  set all;
  Gap=explained+unexplained;
  if Parameter=' ' then Parameter='Overall';
  if Parameter='Overall' then do;
    pct_explained=explained/(explained+unexplained);
    pct_unexplained=unexplained/(explained+unexplained);
  end;
  merge_counter=_N_;
run;

ods listing;
* overall summary report;

title '1) Overall Summary';
proc sql;

```

```

select Parameter as Level format=$15.,
       Gap format=6.3,
       explained label='Exp.' format=6.3,
       unexplained label='Unexp.' format=6.3,
       pct_explained label='% Exp.' format=percent8.2
from all
where (Parameter='Overall' or Dependent=' ');
quit;

title '2) Variable-Specific Summary';
* regression parameter level summary report;
proc sql;
  select Parameter format=$15.,
         Estimate_A label='Coef A' format=6.3, Mean_A label='Mean A' format=6.3,
         Estimate_B label='Coef B' format=6.3, Mean_B label='Mean B' format=6.3,
         Gap format=6.3, explained label='Exp.' format=6.3, unexplained
label='Unexp.' format=6.3,
         pct_explained label='% Exp.' format=percent8.2
  from all
  where Dependent ne ' ';
quit;
title;
%mend;

```

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name:	Taylor Lewis
Enterprise:	U.S. Office of Personnel Management
Address:	1900 E St. NW, Room 4332
City, State ZIP:	Washington, DC 20415
E-mail:	Taylor.Lewis@opm.gov

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.