SIEPR Discussion Paper No. 17-013

Addressing Path Dependence and Incorporating Sample Weights in the Nonlinear Blinder-Oaxaca Decomposition Technique for Logit, Probit and Other Nonlinear Models

By

Robert W. Fairlie

Stanford Institute for Economic Policy Research
Stanford University
Stanford, CA 94305
(650) 725-1874

The Stanford Institute for Economic Policy Research at Stanford University supports research bearing on economic and public policy issues. The SIEPR Discussion Paper Series reports on research and policy analysis conducted by researchers affiliated with the Institute. Working papers in this series reflect the views of the authors and not necessarily those of the Stanford Institute for Economic Policy Research or Stanford University

Addressing Path Dependence and Incorporating Sample Weights in the Nonlinear Blinder-Oaxaca Decomposition Technique for Logit, Probit and Other Nonlinear Models

Robert W. Fairlie
Professor of Economics, University of California, Santa Cruz
Visiting Scholar, Stanford University
Research Associate, NBER
rfairlie@ucsc.edu

April 2017

Abstract

The Blinder-Oaxaca decomposition technique is widely used to identify and quantify the separate contributions of differences in measurable characteristics to group differences in an outcome of interest. The use of a linear probability model and the standard Blinder-Oaxaca decomposition, however, can provide misleading estimates when the dependent variable is binary, especially when group differences are very large for an influential explanatory variable. A simulation method of performing a nonlinear decomposition that uses estimates from a logit, probit or other nonlinear model was first developed in a *Journal of Labor Economics* article (Fairlie 1999). This nonlinear decomposition technique has been used in nearly a thousand subsequent studies published in a wide range of fields and disciplines. In this paper, I address concerns over path dependence in using the nonlinear decomposition technique. I also present a straightforward method of incorporating sample weights in the technique.

I thank Eric Aldrich and Ben Jann for comments and suggestions, and Brandon Heck for research assistance.

1. Introduction

The Blinder-Oaxaca decomposition technique has been used extensively to examine the potential causes of inter-group differences in outcome variables. A problem arises, however, if the outcome is qualitative and the coefficients are from a logit, probit, multinomial logit, or other nonlinear model. These coefficient estimates cannot be used directly in the standard Blinder-Oaxaca decomposition equations. Additionally, the use of a linear probability model and the standard Blinder-Oaxaca decomposition can provide misleading estimates especially when group differences are very large for an influential explanatory variable. A solution to this problem is a simulation algorithm first developed and published in the *Journal of Labor Economics* (Fairlie 1999) and revised slightly later in the same journal (Fairlie and Robb 2007). The technique uses the original nonlinear equation, such as a logit or probit, for both estimation and decomposition. Software code for the technique has been written for Stata, SAS and R making it relatively easy to implement in practice.[1]

The nonlinear decomposition technique addresses the concern with the Blinder-Oaxaca technique when group differences are large for an independent variable. The concern is related to the problem with the possibility of predicted probabilities lying outside of the (0,1) interval using the linear probability model, but is potentially more problematic. The decomposition expression essentially involves calculating the difference between the predicted probability for one group using the other group's regression coefficients and the predicted probability for that group using its own regression coefficients. Even at the means, the predictions involving one group with

[1] See http://people.ucsc.edu/~rfairlie/decomposition/.

another group's coefficients could be much lower than 0 or much larger than 1 resulting in misleading contribution estimates in the decomposition. Procedures that partially linearize the decomposition can suffer from a similar concern because they might load all of the weight on the one explanatory variable that has the extreme difference between groups even if the total difference is constrained.

The nonlinear decomposition technique due to (Fairlie 1999) has been used extensively in the literature to examine group differences across a wide range of outcomes, choices of groups, fields, and disciplines. The technique has been used to explore the potential causes of racial and gender differences in many different economic outcomes similar to the original applications of the Blinder-Oaxaca technique (Blinder 1973 and Oaxaca 1973).[2] The causes of differences in other individual characteristics have also been examined using the Fairlie (1999) nonlinear decomposition technique. For example, the technique has been used to examine differences between low-IQ and high-IQ individuals in stock market participation rates (Grinblatt, Keloharju, and Linnainmaa: *Journal of Finance* 2011) and religion on child survival in India (Bhalotra, Valente, and van Soest: *Journal of Health Economics* 2010). The technique is not limited to exploring the potential causes of differences in race, gender or other individual characteristics, however, and has also be used to study differences over time, geographies and school types. For example, the technique has recently been used to analyze the causes of changes over time in mortality rates (Finks, Osborne and Birkmeyer: *New England Journal of Medicine* 2013) and childlessness (Hayford: *Demography* 2013), differences between

[2] Racial and gender differences in other fields and disciplines have been examined with the technique. For example, it has been used to study the causes of gender differences in college major choice (Zafar: *Journal of Human Resources* 2013), low cholesterol (Sambamoorthi et al.: *Women's Health Issues* 2012), and racial differences in appendicitis (Livingston and Fairlie: *JAMA: Surgery* 2012).

contiguous and noncontiguous countries in conflicts (Reed and Chiba: *American Journal of Political Science* 2010), and differences between school types in teacher turnover rates (Stuit and Smith: *Economics of Education Review* 2012).

The main concern with the nonlinear decomposition technique is over path dependence due to the arbitrarily selected ordering of variables. Although not problematic in many applications, there is the concern that the decomposition estimates could be sensitive to the ordering of variables because of the nonlinearity of the prediction equations. This paper presents a simple and straightforward method of addressing the concern. Specifically, path dependence is addressed by randomly ordering the variables across replications of the decomposition. Randomly ordering variables preserves the summing up property in each replication and the correlation across characteristics within individuals, but removes the arbitrariness of the order in which variables are chosen when switching group distributions. With enough replications the procedure will converge to the decomposition in which the contribution from each variable is calculated from the average of all possible orderings of variables.

This paper also discusses how to incorporate sample weights in the decomposition. Similar to the unweighted decomposition in which a white subsample is randomly chosen, the weighted decomposition also involves drawing a white subsample, but in this case the probabilities of being randomly chosen are proportional to the sample weights. A black sample of equal size is also drawn randomly with weights proportional to sample weights.

Finally, the paper provides an empirical example that demonstrates the problem with the linear Blinder-Oaxaca technique when there are large differences between groups in an independent variable. Furthermore, I demonstrate how partially linearized techniques for performing the decomposition can also provide misleading contribution estimates in some situations.

2. Nonlinear Decomposition Technique

For a linear regression, the standard Blinder-Oaxaca decomposition of the white/black gap (male/female, North//South, U.S./Country X, etc...) in the average value of the dependent variable, Y, can be expressed as:

(2.1) $$\overline{Y}^W - \overline{Y}^B = \left\lfloor (\overline{X}^W - \overline{X}^B)\hat{\beta}^W \right\rfloor + \left\lfloor \overline{X}^B(\hat{\beta}^W - \hat{\beta}^B) \right\rfloor,$$

where $\overline{X}^j$ is a row vector of average values of the independent variables and $\hat{\beta}^j$ is a vector of coefficient estimates for race *j*. Following Fairlie (1999), the decomposition for a nonlinear equation, $Y = F(X\hat{\beta})$, can be written as:

(2.2) $$\overline{Y}^W - \overline{Y}^B = \left[\sum_{i=1}^{N^W} \frac{F(X_i^W \hat{\beta}^W)}{N^W} - \sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^W)}{N^B} \right] + \left[\sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^W)}{N^B} - \sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^B)}{N^B} \right],$$

where N^j is the sample size for race *j*. This alternative expression for the decomposition is used because $\overline{Y}$ does not necessarily equal $F(\overline{X}\hat{\beta})$.[3] In both (2.1) and (2.2), the first term in brackets represents the part of the racial gap that is due to group differences in distributions of *X*, and the second term represents the part due to differences in the group processes determining levels of *Y*. The second term also captures the portion of the racial gap due to group differences in unmeasurable or unobserved endowments. Similar to

[3] Note that the Blinder-Oaxaca decomposition is a special case of (2.2) in which $F(X_i\beta) = X_i\beta$.

most previous studies applying the decomposition technique, I do not focus on this "unexplained" portion of the gap because of the difficulty in interpreting results (see Jones 1983 and Cain 1986 for more discussion).

To calculate the decomposition, define $\overline{Y}^{j}$ as the average probability of the binary outcome of interest for race *j* and *F* as the cumulative distribution function from the logistic distribution. Equation (2.2) will hold exactly for the logit model that includes a constant term because the average value of the dependent variable must equal the average value of the predicted probabilities in the sample.[4] The equality does not hold exactly for the probit model, in which F is defined as the cumulative distribution function from the standard normal distribution, but holds very closely as an empirical regularity.

An equally valid expression for the decomposition is:

$$(2.3)\quad \overline{Y}^{W} - \overline{Y}^{B} = \left[\sum_{i=1}^{N^W} \frac{F(X_i^W \hat{\beta}^B)}{N^W} - \sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^B)}{N^B}\right] + \left[\sum_{i=1}^{N^W} \frac{F(X_i^W \hat{\beta}^W)}{N^W} - \sum_{i=1}^{N^W} \frac{F(X_i^W \hat{\beta}^B)}{N^W}\right],$$

In this case, the black coefficient estimates, $\hat{\beta}^{B}$ are used as weights for the first term in the decomposition, and the white distributions of the independent variables, $\overline{X}^{W}$ are used as weights for the second term. This alternative method of calculating the decomposition often provides different estimates, which is the familiar index problem with the Blinder-Oaxaca decomposition technique. Alternatively, the first term of the decomposition expression could be weighted using coefficient estimates from a pooled sample of the

[4] In contrast, the predicted probability evaluated at the means of the independent variables is not necessarily equal to the proportion of ones, and in the sample used below it is larger because the logit function is concave for values greater than 0.5.

two groups as suggested in Oaxaca and Ransom (1994) or all racial and ethnic groups as suggested in Fairlie and Robb (2007). I return to this issue below.

The first terms in (2.2) and (2.3) provide an estimate of the contribution of racial differences in the entire set of independent variables to the racial gap in the dependent variable. Estimation of the total contribution is relatively simple as one only needs to calculate two sets of predicted probabilities and take the difference between the average values of the two. Identifying the contribution of group differences in specific variables to the racial gap, however, is not as straightforward. To simplify, first assume that $N_B = N_W$ and that there exists a natural one-to-one matching of black and white observations. Using coefficient estimates from a logit regression for a pooled sample, $\hat{\beta}^*$, the independent contribution of X_1 to the racial gap can then be expressed as:

$$(2.4)\quad \frac{1}{N^B}\sum_{i=1}^{N^B} F(\hat{\alpha}^* + X_{1i}^W\hat{\beta}_1^* + X_{2i}^W\hat{\beta}_2^*) - F(\hat{\alpha}^* + X_{1i}^B\hat{\beta}_1^* + X_{2i}^W\hat{\beta}_2^*).$$

Similarly, the contribution of X_2 can be expressed as:

$$(2.5)\quad \frac{1}{N^B}\sum_{i=1}^{N^B} F(\hat{\alpha}^* + X_{1i}^B\hat{\beta}_1^* + X_{2i}^W\hat{\beta}_2^*) - F(\hat{\alpha}^* + X_{1i}^B\hat{\beta}_1^* + X_{2i}^B\hat{\beta}_2^*).$$

The contribution of each variable to the gap is thus equal to the change in the average predicted probability from replacing the black distribution with the white distribution of that variable while holding the distributions of the other variable constant. A useful property of this technique is that the sum of the contributions from individual variables will be equal to the total contribution from all of the variables evaluated with the full sample.

One problem, however, is that unlike in the linear case, the independent contributions of X_1 and X_2 depend on the value of the other variable. This implies that the choice of a variable as X_1 or X_2 (or the order of switching the distributions) is potentially important in calculating its contribution to the racial gap. I discuss a straightforward solution to address this problem of path dependence in the next section.

Standard errors can also be calculated for these estimates. Following Oaxaca and Ransom (1998), I use the delta method to approximate standard errors. To simplify notation, rewrite (2.4) as:

$$(2.6)\ \hat{D}_1 = \frac{1}{N^B}\sum_{i=1}^{N^B} F(X_i^{WW}\hat{\beta}^*) - F(X_i^{BW}\hat{\beta}^*).$$

The variance of $\hat{D}_1$ can be approximated as:

$$(2.7)\ Var(\hat{D}_1) = \left(\frac{\delta\hat{D}_1}{\delta\hat{\beta}^*}\right)' Var(\hat{\beta}^*)\left(\frac{\delta\hat{D}_1}{\delta\hat{\beta}^*}\right).$$

where $\frac{\delta\hat{D}_1}{\delta\hat{\beta}^*} = \frac{1}{N^B}\sum_{i=1}^{N^B} f(X_i^{WW}\hat{\beta}^*)X_i^{WW} - f(X_i^{BW}\hat{\beta}^*)X_i^{BW}$ and f is the logistic probability density function.

In practice, the sample sizes of the two groups are rarely the same and a one-to-one matching of observations from the two samples is needed to calculate (2.4), (2.5), and (2.7). In this example, it is likely that the black sample size is substantially smaller than the white sample size. A convenient method to address this problem is to draw a random subsample of whites with or without replacement of equal size to the full black sample (N^B).[5] Each observation in the randomly drawn white subsample is matched

[5] The choice over drawing the white subsample with or without replacement is unimportant if enough replications of the procedure are performed. As more replications are performed the decomposition

randomly to an observation in the full black sample. Originally, the technique involved matching the white subsample and the full black sample by their respective rankings in predicted probabilities (Fairlie 1999).[6] More recently, however, the technique has been revised to randomly match the white subsample and full black sample (Fairlie and Robb 2007), which is more in line with the goal of hypothetically matching all white observations to all black observations.[7]

The decomposition estimates obtained from this procedure depend on the randomly chosen subsample of whites. Ideally, the results from the decomposition should approximate those from matching the entire white sample to the entire black sample. A simple method of approximating this hypothetical decomposition is to draw a large number of random subsamples of whites, randomly match each of these randomly drawn subsamples of whites to the full black sample, and calculate separate decomposition estimates. The mean value of estimates from the separate decompositions is calculated and used to approximate the results for the entire white sample.[8]

To ensure that the full white distribution is approximated a large number of replications should be performed. Depending on computational speed and complexity of the model, I recommend drawing 1,000 subsamples if feasible.[9] Increasing the number of

estimates will converge to the same value. As discussed further below, however, sampling with replacement is preferred when incorporating sample weights because sampling weights could differ widely across observations.

[6] To match by predicted probabilities, one set of coefficient estimates (white, black or pooled) are first used to calculate predicted probabilities for each black and white observation in the sample (Fairlie 1999). Then each observation in the randomly chosen white subsample and full black sample is separately ranked by the predicted probabilities and matched by their respective rankings.

[7] Fairlie (2003, 2005) finds estimates that are not overly sensitive to this choice.

[8] An example of the code used in Stata is provided in Appendix A1.

[9] See Appendix A2 for setting the number of replications to 1,000 in Stata. In Fairlie (2005) I find that estimates for the main specification are identical to the 4th decimal place using 10,000 simulations for all contributions except two groups of variables (which were both less than 0.0001 different). In fact, using only 100 simulations provided contribution estimates that are identical to the 4th decimal place except for

replications might be important in some applications, and when randomizing the order of variables or using sample weights as discussed below. When randomizing the order of variables it is especially important to increase the number of replications (1,000 should be considered a minimum number for most applications).

Pooled Estimates

As discussed above, an alternative to weighting the first term of the decomposition expression using coefficient estimates from a white sample as presented in (2.2) or the black sample as presented in (2.3) is to use coefficient estimates from a pooled sample of the two groups as suggested in Oaxaca and Ransom (1994).[10] Both groups contribute to the estimation of the parameters instead of only one group. It is essential to include a dummy variable for black race in the regression to remove any influence on the coefficients from racial differences that are correlated with any of the explanatory variables.

An alternative approach, which is becoming increasingly popular when studying racial differences, however, is to use the full sample of all races to estimate the coefficients (Fairlie and Robb 2007). This version of the pooled sample is advantageous in that it incorporates the full market response and does not exclude rapidly growing groups of the population (i.e. Hispanics and Asians).[11] Again, it is important to include the full set of racial and ethnic dummy variables in the regression specification. In the

only two groups of variables (which are both less than 0.0002 different). But, this insensitivity might not hold for more complicated models.

[10] Appendix A1 provides an example of Stata code for using the pooled estimates. Appendix A3 provides an example where instead the white sample is used to estimate the coefficients (i.e. equation 2.2) and Appendix A4 provides an example where the black sample is used (i.e. equation 2.3).

[11] An example of the Stata code is provided in Appendix A5.

end, the choice across these alternative methods of calculating the first term of the decomposition is difficult and depends on the application with many studies reporting results for more than one specification.

3. Path Dependence

A potential concern in using the nonlinear decomposition technique is the effect of the ordering of variables on the results. As noted above, because of the nonlinearity of the decomposition equation the results may be path dependent. Often researchers will experiment with reversing the order of switching distributions of variables, and in many cases the results will not be overly sensitive to that ordering. One important feature of the decomposition technique is that the total contribution, however, remains unchanged because the sum of the individual contributions, regardless of their order, must equal the total contribution defined in (2.2) or (2.3).

The sensitivity of estimates to the reordering of variables, however, depends on the application. The initial location in the logistic distribution and the total movement along the distribution from switching distributions of other variables both contribute to how sensitive the results are to the ordering of variables. If the results vary across different orderings of the variables then randomizing the ordering of variables can solve the problem. In fact, the ordering of switching distributions can be conveniently randomized at the same time as drawing the random subsample of whites. By using a large number of replications the procedure will approximate the average decomposition across all possible orderings of variables while preserving the summing up property.

Table 1 presents decomposition estimates for an application for explaining racial differences in computer ownership originally discussed in Fairlie (2003, 2005). The nonlinear decomposition results from the original ordering of variables are presented along with the random ordering of variables and the reverse order ordering of variables.[12] There are some differences in contribution estimates for a few variables between the original ordering of variables and the reverse ordering of variables. Reassuringly, the random ordering of variables results in decomposition estimates that lie between estimates from the two orderings of variables. Although the differences in contribution estimates resulting from different orderings of variables are relatively small, with the availability of faster computing randomizing the order of variables is straightforward. Faster computers have made this increasingly feasible even with very large datasets and complicated underlying regression models.

4. Comparison to Linear and Partially Linearized Methods

Concerns over path dependence have been the main criticism of the Fairlie (2009) nonlinear decomposition technique, and has been used to justify use of the linear Blinder-Oaxaca technique or the alternative "partially linearized" technique (e.g. Yun 2004 and Even and Macpherson 1990) even with a logit or probit model. Although these techniques do not suffer from path dependence they are vulnerable to the possibility of inflated estimates on contributions from independent variables in which group differences

[12] The procedure of randomizing the order of switching variables with each replication is easily implemented in Stata. See Appendix A6 for an example of the Stata code. It is recommended, however, to increase the number of replications and check to make sure that the decomposition estimates are similar when performing the procedure with a few different random draws. SAS code for randomizing the order of variables is also available (see http://people.ucsc.edu/~rfairlie/decomposition/decompexamplerandom_v7.sas).

are extremely large. The contribution estimate from independent variable, X_1, in the Blinder-Oaxaca linear decomposition is the following from (2.1):

$$(4.1)\quad C_1 = (\overline{X}_1^W - \overline{X}_1^B)\hat{\beta}_1^W .$$

The linear expression of (2.1) allows for separability of contributions for X_1 and X_2, and thus removes concerns over path dependence. One major weakness in the linear setting is that there is no restriction on how large this contribution estimate can be even when the dependent variable is constrained to equal 0 or 1. This concern is similar to the concern over predictions when using OLS to estimate an equation in which the dependent variable lies between 0 and 1. Even at the means, the predictions involving one group with another group's coefficients could be much lower than 0 or much larger than 1 resulting in misleading contribution estimates in the decomposition.

It is not difficult to find an empirical example where this potentially presents a problem. One example is that GNP per capita in the United States is \$48,000 compared with \$1,400 in India (World Bank 2015). The contribution estimate, (4.1), from this cross-country difference in per-capital income is thus 46,600 * $\hat{\beta}_1$. One would only need a coefficient of at least 0.00002 (which implies a small effect of 2 percentage points for every \$1,000 in per-capita income) to generate a contribution estimate of 1.

Another "partially linearized" method used to perform the decomposition is to combine this linear contribution estimate for each independent variable with the total contribution from all variables using the nonlinear function (e.g. probit or logit) as displayed in the first half of (2.2):

$$(4.2)\quad \gamma_T = \sum_{i=1}^{N^W} \frac{F(X_i^W \hat{\beta}^W)}{N^W} - \sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^W)}{N^B} ,$$

The partially linearized contribution for independent variable, X_1, is:

$$(4.2)\ L_1 = \frac{C_1}{\sum_{k=1}^{K} C_k} \gamma_T ,$$

where k=1,…,K for each independent variable (or group of variables) in the decomposition, and $\hat{\beta}_1^W$ is from the nonlinear estimation technique (e.g. probit or logit) for calculating C_1 in (4.1). The contribution estimate is essentially a linear partition of the total nonlinear contribution from all included variables. It limits the total potential influence of group differences in all variables combined, but does not prevent one variable with a very large group difference from capturing all or most of the total contribution from the nonlinear difference.

Illustrative Example with Large Group Difference in an Independent Variable

To illustrate these concerns with an empirical example, I use data from the 2013 American Community Survey (ACS). The dependent variable is computer ownership which is equal to 0 or 1. The independent variables included in this illustrative example are housing value, education level, and age. To maximize differences between groups for illustrative purposes I choose the highest housing value state for whites (California) and the lowest housing value state for Latinos (Oklahoma). Average house prices are $355,000 for whites and $49,000 for Latinos. The difference in percentage terms is similar to the national gap in net worth between whites and Latinos (U.S. Census Bureau 2015), but the ACS does not include information on net worth, only house values.

Table 2 reports estimates from all three decomposition methods: the nonlinear technique, Blinder-Oaxaca technique, and the partially linearized technique for the gaps in computer ownership. The nonlinear technique indicates that ethnic/racial differences in house values and education levels explain large portions of the gaps in computer ownership. Group differences in house values explain 13.7 percentage points (or 49 percent) of the 28 percentage point gap in computer ownership rates. Group differences in education levels, which are also large, explain another 6.4 percentage points (or 23 percent) of the computer gap. The education disparity is large between whites and Latinos (14.5 and 10.9 years of schooling, respectively).

Using the Blinder-Oaxaca decomposition I find a much larger contribution estimate for white/Latino differences in house values. I find that group differences in house values explain 39 percentage points of the gap (which is more than 100% of the gap). Although there is always the possibility that one factor can explain more than 100% of the gap, the key point here is that this estimate is three times larger than the contribution estimate from the nonlinear decomposition technique which through the use of a logit or probit model forces a limit on predicted probabilities being less than 1 no matter how large house values are. The Blinder-Oaxaca also provides larger contribution estimates from group differences in education and age, but these are more similar to the contribution estimates from the nonlinear technique.

By definition the partially linearized technique provides the same total contribution estimate from group differences in all three independent variables combined as the nonlinear technique. For both the partially linearized technique and the nonlinear technique group differences in house values, education and age explain a total of 20.7

percentage points (or 74 percent) of the 28 percentage point gap in computer ownership rates. What is of most concern, however, is that the technique places almost all of this weight on group differences in house values and very little weight on the contribution from group differences in education (which as noted above are also very large). The partially linearized technique provides a contribution estimate from house value differences of 18.9 percentage points of the total 20.7 percentage points explained by all of the variables. The partially linearized technique provides a contribution estimate of only 1.6 percentage points from education differences.

Although admittedly, the example is chosen to maximize differences in house values to illustrate potential problems with the Blinder-Oaxaca and partially linearized techniques, large group differences in independent variables are not uncommon.[13] In fact, the magnitude and group difference in net worth for the country are not that different from these house value differences. The mean level of net worth among non-Latino whites is $435,000 and the mean level of net worth among Latinos is $86,000 for the United States (U.S. Census Bureau 2015). Decompositions involving cross country differences often involve much larger group differences. For example, as noted above GNP per capita in the United States is $48,000 compared with $1,400 in India (World Bank 2015).

In the end, the Blinder-Oaxaca and partially linearized techniques may face problems when group differences in a key independent variable are large. In this case,

[13] The patterns hold for other large states with large Latino populations and low house prices, such as Texas. The Blinder-Oaxaca technique provides a very large house value contribution and the partially linearized technique provides a very small education contribution.

and in general, it is safer to use the nonlinear decomposition technique to avoid overly large contribution estimates or loading up too much weight on one variable.

4. Sample Weights

This section turns to a discussion of including sample weights in the nonlinear decomposition. To simplify the presentation, the decomposition equations presented above do not include sample weights. Without sample weights, a random white (or group 1) subsample is drawn of the same sample size as the minority (or group 2) full sample for convenience for matching distributions. An easy method of incorporating sample weights, first suggested by Ben Jann (Jann 2006), is to draw the white subsample with replacement where the sampling probabilities are proportional to the sample weights. A minority subsample should also be drawn with replacement where the sampling probabilities are proportional to the sample weights. The decomposition technique presented above is nearly identical to this procedure for incorporating sample weights when the sample weights are the same within the white and minority samples and the minority sample size is used. The only difference is that the full minority sample differs from the minority sample drawn with replacement in each iteration. As the number of replications of the procedure increases estimates of the mean value across all replications will converge to each other.

A few possible choices for the sample size to match the white and black (minority) subsamples include the full black sample size (N^B), the full white sample (N^W), or the average of the two. The decision over the number of observations drawn from the white and black samples is arbitrary, however, because the convergence in the precision of results depends on the total number of white and black observations matched (which is a function of the matching sample size and the number of replications). Choosing the smaller black sample size for each iteration, for example, could be offset by increasing the number of replications. For any chosen sample size for matching, the expected value of results is equivalent to the original decomposition if the weights are equal within the white and black samples or are independent from the variables. See Appendix A7 for an example of the Stata code to perform the decomposition with sample weights. SAS code is also now available (see Appendix and http://people.ucsc.edu/~rfairlie/decomposition/).

5. Summary

The nonlinear decomposition technique developed in Fairlie (1999) has been used to identify the underlying causes of group differences in outcomes in nearly one thousand studies in several different fields, across many outcomes, and for a wide range of groups. Because the technique uses the original nonlinear equation, such as a logit or probit, for both estimation and decomposition it does not suffer from the potential problem of generating predictions outside of the (0,1) interval or misleading estimates from the linear Blinder-Oaxaca decomposition (or partially linearized techniques) when group differences are very large for an influential explanatory variable. Concerns over path dependence due to the ordering of variables in the nonlinear decomposition technique are addressed by randomly ordering the variables and increasing the number of replications of the procedure. Sample weights are also easily included in the decomposition by randomly drawing a minority subsample in addition to a white subsample and randomly drawing each subsample in proportion to the original sample weights. The random ordering and sample weights are easy to implement with existing Stata and SAS code.

References

Bhalotra, Sonia, Christine Valente, and Arthur van Soest. 2010. "The puzzle of Muslim advantage in child survival in India," *Journal of Health Economics*, 29: 191–204

Blinder, Alan S. 1973. "Wage Discrimination: Reduced Form and Structural Variables." Journal of Human Resources, 8, 436-455.

Cain, Glen G. 1986. "The Economic Analysis of Labor Market Discrimination: A Survey," Handbook of Labor Economics, Vol. 1, eds. O. Ashenfelter and R. Laynard, Elsevier Science Publishers BV.

Even, William E., and David A. Macpherson. 1990. "Plant size and the decline of unionism." *Economics Letters* 32(4): 393-398.

Fairlie, Robert W. 1999. "The Absence of the African-American Owned Business: An Analysis of the Dynamics of Self-Employment," Journal of Labor Economics, 17(1): 80-108.

Fairlie, Robert W. 2003. "An Extension of the Blinder-Oaxaca Decomposition Technique to Logit and Probit Models," Yale University, Economic Growth Center Discussion Paper No. 873.

Fairlie, Robert W. 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, Robert W. and Alicia M. Robb. 2007. "Why are Black-Owned Businesses Less Successful than White-Owned Businesses: The Role of Families, Inheritances, and Business Human Capital," *Journal of Labor Economics*, 25(2): 289-323.

Finks, Jonathan F., Nicholas H. Osborne, and John D. Birkmeyer. 2011. "Trends in Hospital Volume and Operative Mortality for High-Risk Surgery," *New England Journal of Medicine*, 364:2128-2137.

Grinblatt, Mark, Matti Keloharju, and Juhani Linnainmaa. 2011. "IQ and Stock Market Participation," *The Journal of Finance*, 66(6): 2121-2164.

Hayford, Sarah R. 2013. "Marriage (Still) Matters: The Contribution of Demographic Change to Trends in Childlessness in the United States," *Demography*, 50(5): 1641-1661.

Livingston, Edward H., and Robert W. Fairlie, 2012. "Little Effect of Insurance Status or Socioeconomic Condition on Disparities in Minority Appendicitis Perforation Rates," *Journal of the American Medical Association (JAMA): Surgery (Archives of Surgery)*, 147(1): 11-17.

Jann, Ben. 2006. fairlie: Stata module to generate nonlinear decomposition of binary outcome differentials. Available from http://ideas.repec.org/c/boc/bocode/s456727.html.

Jones, F.L. 1983. "On Decomposing the Wage Gap: A Critical Comment on Blinder's Method," Journal of Human Resources, 18(1): 126-130.

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

Oaxaca, Ronald, and Michael Ransom. 1994. "On Discrimination and the Decomposition of Wage Differentials," Journal of Econometrics, 61, 5-21.

Oaxaca, Ronald, and Michael Ransom. 1998. "Calculation of Approximate Variances for Wage Decomposition Differentials," *Journal of Economic and Social Measurement*, 24, 55-61.

Reed, William and Daina Chiba. 2010. "Decomposing the Relationship Between Contiguity and Militarized Conflict," *American Journal of Political Science,* 54(1): 61-73.

Sambamoorthi, Usha, Sophie Mitra, Patricia A. Findley, and Leonard M. Pogach. 2012. "Decomposing Gender Differences in Low-Density Lipoprotein Cholesterol among Veterans with or at Risk for Cardiovascular Illness," *Women's Health Issues*, 22(2): e201–e208.

Stuit, DA, and TM Smith. 2012. "Explaining the gap in charter and traditional public school teacher turnover rates," *Economics of Education Review*.

Yun, Myeong-Su. 2004. "Decomposing differences in the first moment." *Economics letters* 82(2): 275-280.

Zafar, Basit. 2013. "College major choice and the gender gap." *Journal of Human Resources* 48(3): 545-595.

Appendix
SAS and Stata Code

SAS

SAS Programs (both programs can incorporate sample weights if needed)

SAS Program for Specified Ordering of Variables:
http://people.ucsc.edu/~rfairlie/decomposition/decompexample_v7.sas

SAS Program for Randomized Ordering of Variables:
http://people.ucsc.edu/~rfairlie/decomposition/decompexamplerandom_v7.sas

Example Dataset to Use with Programs:
http://people.ucsc.edu/~rfairlie/decomposition/finaldecomp00.sas7bdat

Stata

- to install procedure	ssc install fairlie
- or to update version	ssc install fairlie, replace
- to obtain help on procedure	help fairlie

Examples:

To load dataset for examples:

```
use http://people.ucsc.edu/~rfairlie/decomposition/finaldecomp00.dta
```

A1: Nonlinear decomposition using pooled (white and black) coefficient estimates

```
fairlie homecomp female age (educ:hsgrad somecol college) (marstat:married prevmar) if
  white==1|black==1, by(black) pooled(black)
```

A2: Adding more replications to A1

```
fairlie homecomp female age (educ:hsgrad somecol college) (marstat:married prevmar) if
  white==1|black==1, by(black) pooled(black) reps(1000)
```

A3: Using white coefficient estimates

```
fairlie homecomp female age (educ:hsgrad somecol college) (marstat:married prevmar) if
  white==1|black==1, by(black)
```

A4: Using black coefficient estimates

```
fairlie homecomp female age (educ:hsgrad somecol college) (marstat:married prevmar) if
  white==1|black==1, by(black) reference(1)
```

A5: Using pooled (all races) coefficient estimates

```
generate black2 = black==1 if white==1|black==1
fairlie homecomp female age (educ:hsgrad somecol college) (marstat:married prevmar),
  by(black2) pooled(black latino asian natamer)
```

A6: Randomly ordering variables

```
fairlie homecomp female age (educ:hsgrad somecol college) (marstat:married prevmar),
  by(black2) pooled(black latino asian natamer) ro reps(1000)
```

A7: Including sample weights

```
fairlie homecomp female age (educ:hsgrad somecol college) (marstat:married prevmar)
  [pw=wgt], by(black2) pooled(black latino asian natamer) reps(1000)
```

Table 1

Non-Linear Decompositions of Black/White Gaps in Home Computer Rates

Orginal Ordering, Reverse Ordering and Random Ordering of Variable Groups

	Specification			
	(1)	(2)	(3)	(4)
Modification to decomposition	Orginal Order	Reverse Order	Random Order	Random Order
White computer ownership rate	0.7278	0.7278	0.7278	0.7278
Black computer ownership rate	0.4175	0.4175	0.4175	0.4175
Black/White gap	0.3103	0.3103	0.3103	0.3103
Contributions from racial differences in:				
Sex and age	-0.0001	-0.0002	-0.0002	-0.0002
	(0.0002)	(0.0005)	(0.0004)	(0.0004)
	0.0%	-0.1%	0.0%	0.0%
Marital status and children	0.0154	0.0237	0.0206	0.0207
	(0.0011)	(0.0016)	(0.0014)	(0.0014)
	5.0%	7.6%	6.6%	6.7%
Education	0.0329	0.0510	0.0407	0.0409
	(0.0010)	(0.0011)	(0.0010)	(0.0010)
	10.6%	16.4%	13.1%	13.2%
Income	0.1005	0.0768	0.0886	0.0883
	(0.0019)	(0.0020)	(0.0019)	(0.0019)
	32.4%	24.8%	28.6%	28.5%
Region	0.0062	0.0047	0.0057	0.0057
	(0.0012)	(0.0010)	(0.0011)	(0.0011)
	2.0%	1.5%	1.8%	1.8%
Central city status	0.0003	-0.0009	-0.0002	-0.0002
	(0.0014)	(0.0012)	(0.0014)	(0.0014)
	0.1%	-0.3%	-0.1%	-0.1%
All included variables	0.1552	0.1552	0.1552	0.1552
	50.0%	50.0%	50.0%	50.0%
Number of replications	1,000	1,000	1,000	5,000

Table 2
Non-Linear, Linear and Partially Linearized Decompositions of Latino/White Gaps in Home Computer Rates

		Specification	
	(1)	(2)	(3)
Modification to decomposition	Nonlinear	Linear (Bl.-Oaxaca)	Partially Linearized
White (CA) computer ownership rate	0.8979	0.8979	0.8979
Latino (OK) computer ownership rate	0.6181	0.6181	0.6181
Latino/White gap	0.2798	0.2798	0.2798
Contributions from racial differences in:			
House values	0.1368	0.3915	0.1885
	(0.0071)	(0.0413)	(0.0086)
	48.9%	139.9%	67.4%
Education	0.0641	0.0777	0.0163
	(0.0087)	(0.0111)	(0.0031)
	22.9%	27.8%	5.8%
Age	0.0058	0.0130	0.0018
	(0.0054)	(0.0073)	(0.0016)
	2.1%	4.6%	0.6%
All included variables	0.2067	0.4822	0.2067
	73.9%	172.3%	73.9%
Number of replications	1,000	N/A	N/A