DISCUSSION PAPER SERIES

IZA DP No. 3971

Multivariate Decomposition for Hazard Rate Models

Daniel A. Powers
Myeong-Su Yun

January 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Multivariate Decomposition for Hazard Rate Models

Daniel A. Powers
University of Texas at Austin

Myeong-Su Yun
Tulane University and IZA

Discussion Paper No. 3971
January 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 3971
January 2009

ABSTRACT

Multivariate Decomposition for Hazard Rate Models

We develop a regression decomposition technique for hazard rate models, where the difference in observed rates is decomposed into components attributable to group differences in characteristics and group differences in effects. The baseline hazard is specified using a piecewise constant exponential model, which leads to convenient estimation based on a Poisson regression model fit to person-period, or split-episode data. This specification allows for a flexible representation of the baseline hazard and provides a straightforward way to introduce time-varying covariates and time-varying effects. We provide computational details underlying the method and apply the technique to the decomposition of the black-white difference in first premarital birth rates into components reflecting characteristics and effect contributions of several predictors, as well as the effect contribution attributable to race differences in the baseline hazard.

JEL Classification: C20, C41, J13

Keywords: decomposition, hazard rates, piecewise constant exponential model, Poisson regression

Corresponding author:

Daniel A. Powers
Department of Sociology
1 University Station (A1700)
University of Texas at Austin
Austin TX 78712
USA
E-mail: dpowers@mail.la.utexas.edu

Multivariate Decomposition for Hazard Rate Models

Introduction

Hazard rate models have been used by social researchers to study fertility, mortality, job mobility, and other processes involving transitions from one state to another over time. Interest generally focuses on how rates respond to changes in individual and structural characteristics or how these factors shape differences in rates across social groups. Understanding the sources of group differences in rates can inform policy makers and scholars alike about the impact of compositional differences across groups and the effects of group differences in returns-to-risk associated with certain individual-level and structural characteristics. Multivariate decomposition analysis is an appropriate tool for this purpose.

Multivariate decomposition is widely used in social research to quantify the contributions to group differences in average predictions from multivariate models. The technique utilizes the output from regression models to parcel out components of a group difference in a statistic, such as a mean or proportion, which can be attributed to compositional differences between groups (i.e., differences in characteristics or endowments) and to differences in the effects of characteristics (i.e., differences in the returns, coefficients, or behavioral responses). These techniques are equally applicable for partitioning change over time into components attributable to changing effects and changing composition.

Decomposition techniques for linear regression models have been used for many decades in sociological research. This heterogeneous collection of techniques is more

generally referred to as regression standardization (Althauser and Wigler 1972, Duncan 1969, Duncan, Featherman and Duncan 1968, Coleman and Sorenson 1970, Coleman and Blum 1971, Coleman, Berry, and Blum 1971, Winsborough and Dickinson 1971). Demographic standardization and decomposition techniques—generally referred to as component analysis—have a much longer history, and were formally developed by Kitagawa (1955) and generalized by Das Gupta (1993). This technique is also known as "shift-share" analysis, and has been used to decompose differences in rates and inequality measures (see, e.g., Shorrocks 1980; 1982, Williams 1991). Unlike regression-based approaches that rely on individual-level observational data, component and shift-share analysis utilize aggregate data, often in the form of published tables. Oaxaca (1973) and Blinder (1973) are usually credited with introducing regression decomposition in the econometric literature in the early 1970's. Although their methods are formally identical to those developed by sociological methodologists and demographers, the technique has become more commonly known as Oaxaca-Blinder, Oaxaca, or Blinder-Oaxaca decomposition.

Regression decomposition has been extended to nonlinear models including: probit (Gomulka and Stern 1990), logit (Even and Macpherson 1993, Fairlie 2005, Nielson 1998, Yun 2005a), and count models (see e.g., Bauer et al. 2007; Heitmueller 2004; Park and Lohr 2008). For linear regression, logit, and count models, the observed difference in group means, proportions, or counts (i.e., a difference in the "first moment") is additively decomposed into a characteristics (or endowments) component and a coefficient (or effects) component. It should be noted that in any given application a

researcher may be interested in one or the other of these components, such as in the portion of the total differential that could be attributed to compositional differences between groups, or to the change in characteristics over time for a single group (see e.g., Evan and Macpherson 1993 and Nielsen 1998).[1]

The rationale for extending multivariate decomposition to rate models is motivated by considering parallels with traditional approaches along with the conveniences achieved by adopting the more widely used regression-based decomposition techniques. The traditional demographic approach of component analysis is a form of decomposition that seeks to partition a difference in rates into components due to compositional differences between groups and to group differences in rates (Kitagawa 1955). Aggregate data are required for traditional component analysis, which has an advantage insofar as analysis can be carried out based on published data tables (see, e.g., Smith, Morgan, and Koropeckyj-Cox 1996). However, the increased complexity of method when extended to more than a few variables is a disadvantage. Given the limitations of the traditional approach and the advantages of carrying out analysis using individual-level observational data, we develop a convenient regression-based method for decomposing differences in rates utilizing results from multivariate models. This approach provides a link between the traditional demographic approach of multiple component analysis for differences in rates and recent regression-based decomposition approaches.

[1] It is also possible to apply a difference in differences approach by combining decompositions across groups over time into a single decomposition.

Given the widespread use of hazard rate models in applied research in a variety of disciplines, as well as longstanding interests in understanding the sources of group disparities and changes over time, extending the regression decomposition to hazard rate models is warranted. This paper develops a multivariate decomposition technique for proportional hazard rate models that are specified with a piecewise constant baseline hazard. This approach is flexible in that it can accommodate arbitrary forms of time dependence in the baseline hazard as well as nonproportional covariate effects. The decomposition is based on a generalized linear model of the same form as the logit, probit, and loglinear models for which software has been developed and extensions may be easily implemented. However, complexities are introduced that are not present in other regression decomposition methods.

In this paper we discuss refinements to the Oaxaca-Blinder decomposition method that lead to a practical approach for multivariate decomposition of a difference in rates. Section 1 reviews the standard Oaxaca-Blinder decomposition. Section 2 discusses the specification of the hazard rate model and the set up for the multivariate decomposition of rates. Section 3 discusses the detailed (covariate by covariate) decomposition, and Section 4 discusses sampling variability of the estimates. Section 5 provides an illustrative example, and Section 6 provides a discussion of extensions and limitations of the technique.

1. Oaxaca-Blinder Decomposition

The Oaxaca-Blinder technique is the most familiar and widely used decomposition technique for linear models. The approach has been applied in research on wage differentials with the goal of understanding the relative roles played by group differences in levels of certain characteristics and group differences in the effects of those characteristics on wage differentials. For example, it is often argued that the portion of the wage differential that cannot be accounted for by group differences in characteristics is the result of labor market discrimination or differences in the returns to human capital factors such as education or job experience, and differences in unmeasured factors.

Oaxaca-Blinder regression decomposition begins with a linear model estimated separately for two groups, or for one group at two time points, indexed by *j*,

$$\hat{y}_{ij} = \mathbf{x}'_{ij}\mathbf{b}_j \qquad j = 1, 2, \tag{1}$$

where $\hat{y}_{ij}$ denotes the fitted value of *y* for the *i*th individual in the *j*th group, $\mathbf{x}_{ij}$ is a collection of measured characteristics for that individual (a $K \times 1$ vector)—including a constant term—and $\mathbf{b}_j$ is the set of estimated regression coefficients (a $K \times 1$ vector). The difference in average predictions can be partitioned into the sum of two components as[2],

$$\bar{\hat{y}}_1 - \bar{\hat{y}}_2 = \bar{y}_1 - \bar{y}_2 = \mathbf{b}'_1(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) + \bar{\mathbf{x}}'_2(\mathbf{b}_1 - \mathbf{b}_2) \tag{2}$$

The first component reflects the contribution to the total differential due to group differences in the mean values of $\mathbf{x}$, holding the effects constant at group 1 levels. This

[2] The difference is often decomposed into the sum of 3 components as $E = \mathbf{b}'_1(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, $C = \bar{\mathbf{x}}'_1(\mathbf{b}_1 - \mathbf{b}_2)$, and the interaction $I = (\mathbf{b}_2 - \mathbf{b}_1)'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$.

component is called the explained component, or endowment or characteristics effect, which is generally denoted by E. The second component reflects the portion of the differential due to group differences in $\mathbf{b}$, holding the mean value of characteristics constant at group 2 levels, which is generally denoted by C. This component is called the unexplained component or coefficients effect. An equivalent decomposition, albeit with a change of sign, results from switching the roles of the comparison (group 1) and the reference group (group 2). In practice, both sets of results are reported or the results from the two separate decompositions are averaged. It is also possible to base the decomposition on results from various forms of pooled regression models (see, e.g., Jann 2008 for a review).

In addition to a decomposition of the overall difference, we are often interested in the unique contribution of each covariate to the overall difference, or the *detailed* decomposition. For example, if groups differ on levels of education and returns to education, it would be desirable to isolate the distinct contributions to the total differential attributable to differences in levels of, and returns to, education along with the unique contributions of the other predictors in the model.

The Oaxaca technique is applicable when group differences in sample means or changes in sample means over time are the focus of inference. However, many socio-demographic outcomes involve differences in predicted rates or proportions estimated from nonlinear response models. It is well known that the usual Oaxaca method of mean/coefficient substitution is not strictly applicable to nonlinear response models, hence the recent interest in extending the method to this class of models. Moreover, for

nonlinear response models, the results from the detailed decomposition are sensitive to the order in which variables enter the decomposition. Various methods have been proposed to overcome this dependency, including: averaging over all possible orders of covariate replacement (Fairlie 2005) and by determining the relative contribution of each variable to each component using a set of appropriately constructed weights (Even and Macpherson 1993, Nielson 1998, Yun 2005a).

This paper builds on previous research by Even and Macpherson (1993) and Nielsen (1998), who extend the Oaxaca-Blinder approach to binary response models. These methods, as well as several innovative extensions, have been developed in a more systematic way by Yun (2004), who addressed several weaknesses in past approaches to multivariate decomposition of nonlinear response models (see e.g., Fairlie 2005). Yun's estimator is simple to calculate and its sampling distribution can be obtained using asymptotic theory (see e.g., Yun 2005a).

2. Regression Decomposition of a Difference in Rates

We follow the logic used in previous research on multivariate decomposition of binary response models by introducing modifications for rate models. As an illustrative example, we decompose the observed difference in premarital birth rates for non-Hispanic blacks and whites using data from the 1979 cohort of the National Longitudinal Survey of Youth (NLSY). We define the empirical rate in the conventional way as the number of events divided by the total amount of exposure to risk. Let d_i be a binary

variable coded 1 if an event occurs for individual *i* at age t_i, and 0 otherwise (i.e., t_i is right censored). The observed rate can be expressed as $r = \sum d_i / \sum t_i$. The black-white difference in rates is expressed as

$$r_B - r_W = \overline{F(\mathbf{x}'_{iB}\mathbf{b}_B)} - \overline{F(\mathbf{x}'_{iW}\mathbf{b}_W)}, \tag{3}$$

where the indices *B* and *W* denote the higher-risk (non-Hispanic black) and lower-risk (non-Hispanic white) group, respectively, and $\overline{F(\mathbf{x}'_{ij}\mathbf{b}_j)}$, is computed as

$$r_j = \overline{F(\mathbf{x}'_{ij}\mathbf{b}_j)} = \frac{\sum_{i=1}^{N_j}\sum_{l=1}^{n_{ij}} F(\mathbf{x}'_{ilj}\mathbf{b}_j)}{\sum_{i=1}^{N_j}\sum_{l=1}^{n_{ij}} \Delta t_{ilj}}, \tag{4}$$

where

$$F(\mathbf{x}'_{ilj}\mathbf{b}_j) = \Lambda_{ilj} = \Lambda_{0lj}\exp(\mathbf{x}'_{ilj}\mathbf{b}_j) \quad j \in \{W, B\}, \quad (l = 1, \ldots, n_{ij}) \tag{5}$$

is the estimated cumulative, or integrated, hazard associated with the *i*th individual in the *l*th time interval from a piecewise constant exponential hazard rate model. We can view Λ_{ilj} as the expected number of events experienced by the *i*th individual in the *l*th interval of exposure to risk, assuming a time-homogeneous Poisson process with rate λ_{ilj} that is observed until either a *first* event occurs or the sub-interval of time has elapsed without the event occurring (see, e.g., Aitken and Clayton 1980, Barlow and Proschan 1975, Holford 1980). For the (piecewise constant) exponential model, the total number of events in group *j* equals the sum of the estimated integrated hazards. That is,

$$\sum_{i=1}^{N_j} d_{ij} = \sum_{i=1}^{N_j}\sum_{l=1}^{n_{ij}} \Lambda_{0lj}\exp(\mathbf{x}'_{ilj}\mathbf{b}_j), \quad j \in \{B, W\} \tag{6}$$

Each individual contributes Δt_{il} units of exposure to the *l*th time interval. It follows that an individual's total exposure equals the individual's event or censoring age, $\sum_{l=1}^{n_{ij}} \Delta t_{ilj} = t_{ij}$. The number and the widths of the age intervals are chosen exogenously and must be the same for each group. Specifically, we define a set of cut-points, $[\tau_0, \tau_1), [\tau_1, \tau_2), \ldots, [\tau_l, \infty)$, or pieces, for the piecewise constant model that are common to both groups. Then, given an individual's event or censoring time t_i , we determine an individual's exposure in the *l*th interval as,

$$\Delta t_{il} = \begin{cases} 0 & \text{if } t_i < \tau_{l-1}, \\ t_i - \tau_{l-1} & \text{if } \tau_{l-1} < t_i \le \tau_l, \\ \tau_l - \tau_{l-1} & \text{if } t_i > \tau_l. \end{cases} \tag{7}$$

This results in n_i sub-episodes of risk for individual *i*. Note that the sum of the subinterval exposures over all individuals necessarily equals the total exposure in the sample, i.e., $\sum_{i=1}^{N_j} \sum_{l=1}^{n_{ij}} \Delta t_{ilj} = \sum_{i=1}^{N_j} t_{ij}$. Combining this with Eq. (6), we can show the equivalence between the *exposure-averaged* predicted event counts and the observed rates,

$$r_j = \frac{\sum_{i=1}^{N_j} d_{ij}}{\sum_{i=1}^{N_j} t_{ij}} = \frac{\sum_{i=1}^{N_j} \sum_{l=1}^{n_{ij}} \Lambda_{0lj} \exp(\mathbf{x}'_{ilj} \mathbf{b}_j)}{\sum_{i=1}^{N_j} \sum_{l=1}^{n_{ij}} \Delta t_{ilj}}, \quad j \in \{B, W\}. \tag{8}$$

For our illustrative example, we adopt a proportional hazards model with piecewise constant hazards over 6 age intervals, [12,16), [16,18), [18,20), [20,22),

[22,24), [24+), which allows time dependence in the baseline hazard by age and is similar to the partitioning used by Wu and Martinson (1993), Powers (2001), and others. It is convenient to parameterize the baseline log hazard separately from the structural part of the model by excluding the conventional intercept and including a set of dummy variables for the 6 age intervals, $D_{i1j},\ldots,D_{i6j}$ and a corresponding set of parameters for the log baseline hazard, $a_{1j},\ldots,a_{6j}$, which results in the following model specification for the rate:

$$\lambda_{ilj} = \exp(a_{1j}D_{i1j} + a_{2j}D_{i2j} + \ldots + a_{6j}D_{i6j} + \mathbf{z}'_{ilj}\boldsymbol{\gamma}_j) = \exp(a_{lj} + \mathbf{z}'_{ilj}\boldsymbol{\gamma}_j), \qquad (9)$$

where $\mathbf{z}$ denotes the vector of predictors and $\boldsymbol{\gamma}$ denotes the corresponding vector of coefficients.

This is a proportional hazards model that is semi-parametric in the sense of a Cox (1972) proportional hazards model as the number of time intervals increases.[3] Assuming a constant exponential hazard for each piece, the integrated hazard in Eq. (5) can be written as $\Lambda_{ilj} = \Delta t_{ilj}\lambda_{ilj}$. We exploit the similarity between the loglinear model for counts and the exponential model for rates by including the logged exposure to risk in the *l*th interval ($\log \Delta t_{il}$) as an "offset" term in a Poisson regression model. It is well known that this approach yields a piecewise constant exponential hazard rate model (see, e.g., Holford 1976;1980, Laird and Oliver 1981). Eq. (5) can now be written as

$$F(\mathbf{x}'_{ilj}\mathbf{b}_j) = \exp(a_{lj} + \mathbf{z}'_{ilj}\boldsymbol{\gamma}_j + \log \Delta t_{ilj}) \qquad (10)$$

[3] In the extreme case, the number of time intervals would equal the number of unique event times. Applying this model to a data set that has been split at the unique failure times would give results identical to a Cox proportional hazard model estimated with Breslow's correction for ties.

It should be noted that while modeling is based on Eq. (10) using Poisson regression, we actually decompose the difference in the rates given in Eq. (9). That is, $\mathbf{x} = (D_1, \ldots, D_6, \mathbf{z})$ and $\mathbf{b} = (a_1, \ldots a_6, \gamma)$. There are advantages to this formulation apart from the fact that standard programs can be used to estimate the models (i.e., Stata **`glm`**, R **`glm`**, and SAS **`proc genmod`**). Nonproportional covariate effects can be introduced by replacing γ_j in Eq. (10) with γ_{lj} (i.e., by including interactions of covariates ($\mathbf{z}$) and the dummy variables (*D*) for the age intervals). Similarly, time-varying covariates can be included in the model, with possibly different values of $\mathbf{z}$ for each interval. The calculations above are facilitated by arranging the input data in the form of a split-episode data structure, with n_i periods of risk (i.e., person-periods or stacked data) allocated to individual *i*. In this case the double summations in Eq. (4) are replaced by single summations over the person-period data (see e.g., Allison 1982).

We would like to decompose the overall difference in Eq. (3) into components that reflect compositional differences between groups and differences in the effects of those characteristics between groups similar to what was done in Eq. (2). We can rewrite Eq. (3) as[4]

$$r_B - r_W = \underbrace{\{\overline{F(\mathbf{x}'_B\mathbf{b}_B)} - \overline{F(\mathbf{x}'_W\mathbf{b}_B)}\}}_{E} + \underbrace{\{\overline{F(\mathbf{x}'_W\mathbf{b}_B)} - \overline{F(\mathbf{x}'_W\mathbf{b}_W)}\}}_{C} \qquad (11)$$

The *E* component appearing in Eq. (11) is the portion of the differential attributed to compositional differences or differences in "endowments," which is the predicted

[4] We drop the individual subscript *i* on $\mathbf{x}_i$ for notational clarity.

premarital birth rate for blacks minus the predicted rate if whites experienced the same returns to risk, or behavioral responses, to characteristics as blacks. This component reflects the contribution to the difference that would have occurred if the two groups differed with respect to characteristics alone. The *C* component in Eq. (11) is the portion of the black-white gap attributable to differences in the coefficients, and reflects the contribution to the difference that would prevail if only the covariate effects differed across groups. Both groups' characteristics are held fixed at white levels to assess this component.

In the expressions above, the coefficients for the black sample are used as weights in the composition (*E*) component and the white covariate values are used as weights in the coefficient (*C*) component, making blacks the comparison group and whites the reference group in this case. The same differential (with a change in sign) can be obtained from an alternative decomposition that switches the roles of the reference and comparison groups. This is referred to as the "indexing" problem (Neumark 1988, Oaxaca and Ransom 1988; 1994).

By fixing the coefficients in the composition component to black levels, we assess the contribution to the black-white gap that would have occurred if the returns to risk associated with the covariates in the model were fixed to the values in the black sample. By fixing characteristics to white levels in the coefficient component, we assess the contribution to the differential that is due to the black-white difference in effects. An equivalent decomposition would reverse this procedure. That is, we could perform a different decomposition by weighting the composition component by the white

coefficient values while using the observed characteristics of blacks as weights in the coefficient component. Sometimes the average of the results of the two specifications is reported.

3. Detailed Decomposition

The decomposition thus far has been described at the aggregate level. To understand the unique contribution of each predictor to each component of the difference requires a detailed decomposition. That is, we wish to partition *E* and *C* into portions, E_k and C_k $(k = 1, \ldots, K)$ that represent the unique contribution of the *k*th covariate to *E* and *C*, respectively. Unlike the decomposition for a linear model, a nonlinear decomposition is sensitive to the order in which the independent variables are entered into the decomposition. This problem is referred to as "path dependence" (see e.g., Yun 2004). The two approaches to detailed decomposition outlined below provide remedies to this problem.

Fairlie (2005) adopts a multi-step procedure for a decomposition based on a logit model, focusing on the characteristics component, E. The procedure requires that we perform a one-to-one matching of comparison-group and reference-group observations based on the ranking of their respective within-group predicted response probabilities. The independent contribution of a variable to *E* is determined by evaluating a decomposition in which one covariate value from the reference group (e.g., z_{1W}) is swapped with one from the comparison group (e.g., z_{1B}). Thus, the contribution of each variable to E is equal to the difference in the average prediction when the reference

group's distribution on a variable is replaced with the comparison group's distribution on that variable while holding the distributions of the other variable constant.

This method is straightforward when the sample sizes are equal. Since this is seldom the case, modifications to the matching procedure are required. The steps suggested by Fairlie are to: (1) draw a random sample from the larger group equal in size to that of the smaller group, (2) rank each group by their respective predicted response probabilities, (3) match observations from the two samples according to their respective rankings on the predicted responses, and (4) evaluate the average group difference in the response probabilities using the sequential covariate swapping approach outlined earlier. This approach does not solve the path dependence problem unless it is accompanied by randomizing the variable swapping order in step (4). In practice, it is necessary to carry out these steps on a large number of random samples from the larger group. The results are then averaged over all the random samples.

Even and Macpherson (1993), Nielsen (1998), and Yun (2004) have suggested simpler methods for detailed decomposition using weights derived from a linearization of the decomposition equation. The detailed decompositions obtained in this way are invariant to the order that variables enter the decomposition, thus providing a solution to path dependency. It should be noted that Even and Macpherson (1993) focus on the endowment component only, whereas Nielsen (1998) focuses only on the coefficient component.

In order to derive the weights that determine the contribution of each covariate to the characteristics and coefficients effects, we consider a two-step approximation of

decomposition equation (Eq. (11)). We first approximate $\overline{F(\mathbf{x}_j'\mathbf{b}_j)}$ by evaluating $F(\mathbf{x}_j'\mathbf{b}_j)$ at the means of the covariates, i.e., $\overline{F(\mathbf{x}_j'\mathbf{b}_j)} \approx F(\overline{\mathbf{x}}_j'\mathbf{b}_j)$. For example, let us denote the characteristics and coefficients components evaluated at the covariate means as $E_M = F(\overline{\mathbf{x}}_B'\mathbf{b}_B) - F(\overline{\mathbf{x}}_W'\mathbf{b}_B)$ and $C_M = F(\overline{\mathbf{x}}_W'\mathbf{b}_B) - F(\overline{\mathbf{x}}_W'\mathbf{b}_W)$, respectively. Eq. (3) can then be expressed as

$$r_B - r_W = E_M + C_M + R_M, \tag{12}$$

where

$$R_M = (E - E_M) + (C - C_M).$$

In the 2nd step, we approximate E_M and C_M in Eq. (12) by a first-order Taylor expansion about $\overline{\mathbf{x}}_B'\boldsymbol{\beta}_B$ and $\overline{\mathbf{x}}_W'\boldsymbol{\beta}_W$.[5] The final decomposition equation after the Taylor expansion is

$$\begin{aligned} r_B - r_W &= (\overline{\mathbf{x}}_B - \overline{\mathbf{x}}_W)'\mathbf{b}_B f(\overline{\mathbf{x}}_B'\mathbf{b}_B) + \overline{\mathbf{x}}_W'(\mathbf{b}_B - \mathbf{b}_W) f(\overline{\mathbf{x}}_W'\mathbf{b}_W) + R_M + R_T \\ &= E_T + C_T + R_M + R_T \end{aligned}, \tag{13}$$

where

$$\begin{aligned} R_T &= [E_M - (\overline{\mathbf{x}}_B - \overline{\mathbf{x}}_A)'\mathbf{b}_B f(\overline{\mathbf{x}}_B'\mathbf{b}_B)] + [C_M - \overline{\mathbf{x}}_W'(\mathbf{b}_B - \mathbf{b}_W) f(\overline{\mathbf{x}}_W'\mathbf{b}_W)] \\ &= (E_M - E_T) + (C_M - C_T) \end{aligned},$$

and $f(\overline{\mathbf{x}}_j'\mathbf{b}_j) = \dfrac{dF(\overline{\mathbf{x}}_j'\mathbf{b}_j)}{d(\overline{\mathbf{x}}_j'\mathbf{b}_j)}$ is the first derivative of $F(\overline{\mathbf{x}}_j'\mathbf{b}_j)$. The quantities R_M, R_T, and $f(\overline{\mathbf{x}}_j'\mathbf{b}_j)$ are all scalars. R_M and R_T are approximation errors resulting from

[5] As noted by Yun (2004), it is also possible to derive the weights using a single approximation. For expository purposes, we consider two approximations.

evaluating $F(\cdot)$ at the mean values and by using the first order Taylor expansion, respectively. Based on Eq. (13), the *k*th weight component for E is obtained as,

$$W_{\Delta x_k} = \frac{E_{Tk}}{E_T} = \frac{b_{Bk}(\bar{x}_{Bk} - \bar{x}_{Wk}) f(\bar{\mathbf{x}}_B' \mathbf{b}_B)}{\sum_{k=1}^{K} b_{Bk}(\bar{x}_{Bk} - \bar{x}_{Wk}) f(\bar{\mathbf{x}}_B' \mathbf{b}_B)} = \frac{b_{Bk}(\bar{x}_{Bk} - \bar{x}_{Wk})}{\sum_{k=1}^{K} b_{Bk}(\bar{x}_{Bk} - \bar{x}_{Wk})}. \quad (14)$$

Similarly, the *k*th weight component for C is given by,

$$W_{\Delta b_k} = \frac{C_{Tk}}{C_T} = \frac{\bar{x}_{Bk}(b_{Bk} - b_{Wk}) f(\bar{\mathbf{x}}_W' \mathbf{b}_W)}{\sum_{k=1}^{K} \bar{x}_{Bk}(b_{Bk} - b_{Wk}) f(\bar{\mathbf{x}}_W' \mathbf{b}_W)} = \frac{\bar{x}_{Bk}(b_{Bk} - b_{Wk})}{\sum_{k=1}^{K} \bar{x}_{Bk}(b_{Bk} - b_{Wk})}, \quad (15)$$

where $\sum_k W_{\Delta x_k} = \sum_k W_{\Delta b_k} = 1.0$.

Thus, the composition weights $W_{\Delta x_k}$ reflect the contribution of the *k*th covariate to the Taylor approximation of E (E_T) as determined by the magnitude of the group difference in means weighted by the reference group's effect. Similarly, the coefficient weights $W_{\Delta b_k}$ reflect covariate *k*'s contribution to C_T as determined by the magnitude of the group difference in the effects weighted by the comparison group's mean. The weights are invariant to change in the scale of the covariates.

The raw difference can now be expressed in terms of the overall components as a sum of weighted sums of the unique contributions.

$$r_B - r_W = E + C = \sum_{k=1}^{K} W_{\Delta x_k} E + \sum_{k=1}^{K} W_{\Delta b_k} C = \sum_{k=1}^{K} E_k + \sum_{k=1}^{K} C_k \,. \quad (16)$$

This weighting method gives results that are nearly identical to the sampling and randomization procedures outlined earlier as long as enough samples are drawn.

4. Variability in Decomposition Estimates

Many applications ignore the sampling variability of the decomposition components (see, e.g., Borooah and Iyer 2005, Sweeney and Phillips 2004, Van Hook, Brown and Kwenda 2004). The characteristics and effects components do not provide information about the precision of the contributions to group differences *per se*. For this reason, it is important to gauge the sampling variability of *E* and *C* in substantive applications. Because the components used in the decomposition are functions of maximum likelihood estimates, the delta method described by Rao (1973, Pp. 321-323) can be used to derive asymptotic standard errors of the detailed contributions. Interval estimation and significance testing can be done in the usual way (see, e.g., Yun 2005a). This approach utilizes expressions for the first derivatives (i.e., gradients) of the detailed components with respect to the estimates, in addition to the variance covariance matrix of the estimates from each group, as we show next.

E and *C*, along with the detailed contributions, E_k and C_k, are nonlinear functions of the maximum likelihood estimates $\mathbf{b}$. The derivatives of E_k and C_k with respect to $\mathbf{b}$, together with the variance/covariance matrix of $\mathbf{b}$, are used to obtain the asymptotic variance-covariance matrix of the detailed components. We begin by expressing the endowment component as a weighted sum of the individual contributions, E_k,

$$E = \sum_{k=1}^{K} E_k = \sum_{k=1}^{K} W_{\Delta x_k} \{\overline{F(\mathbf{x}_B'\mathbf{b}_B)} - \overline{F(\mathbf{x}_W'\mathbf{b}_B)}\}. \qquad (17)$$

The *k*th element of the gradient vector is given by

$$\frac{\partial E_k}{\partial b_{B_k}} = W_{\Delta x_k}\left\{\overline{\frac{\partial F(\mathbf{x}'_B\mathbf{b}_B)}{\partial b_{B_k}}} - \overline{\frac{\partial F(\mathbf{x}'_W\mathbf{b}_B)}{\partial b_{B_k}}}\right\} + w_{x_k}\{\overline{F(\mathbf{x}'_B\mathbf{b}_B)} - \overline{F(\mathbf{x}'_W\mathbf{b}_B)}\}, \tag{18}$$

where

$$w_{x_k} = \frac{\partial W_{\Delta x_k}}{\partial b_{B_k}} = \frac{\overline{x}_{B_k} - \overline{x}_{W_k}}{\sum_k b_{B_k}(\overline{x}_{B_k} - \overline{x}_{W_k})} - \frac{b_{B_k}(\overline{x}_{B_k} - \overline{x}_{W_k})^2}{\left\{\sum_k b_{B_k}(\overline{x}_{B_k} - \overline{x}_{W_k})\right\}^2},$$

For nonlinear models in general, $\frac{\partial F(\mathbf{x}'_{ij}\mathbf{b}_j)}{\partial b_k} = f(\mathbf{x}'_{ij}\mathbf{b}_j)x_{ij_k}$, $j \in \{B,W\}$. For the models considered here

$$\overline{f(\mathbf{x}'_j\mathbf{b}_j)x_{j_k}} = \frac{\sum_{i=1}^{N_j}\sum_{l=1}^{n_{ij}} \partial F(\mathbf{x}'_{ilj}\mathbf{b}_j)/\partial b_{j_k}}{\sum_{i=1}^{N_j}\sum_{l=1}^{n_{ij}} \Delta t_{ilj}} = \frac{\sum_{i=1}^{N_j}\sum_{l=1}^{n_{ij}} f(\mathbf{x}'_{ilj}\mathbf{b}_j)x_{ilj_k}}{\sum_{i=1}^{N_j}\sum_{l=1}^{n_{ij}} \Delta t_{ilj}}, \tag{19}$$

which has a convenient form owing to the assumption of Poisson sampling.[6]

Letting $\text{var}(\mathbf{b}_B)$ denote the variance/covariance matrix of $\mathbf{b}_B$ and $\mathbf{E}$ denote the $K \times K$ matrix with $E_1,\ldots,E_K$ on the main diagonal and zeros elsewhere, the asymptotic (co)variances matrix of the detailed characteristics component is

$$\left(\frac{\partial \mathbf{E}}{\partial \mathbf{b}_B}\right)' \text{var}(\mathbf{b}_B)\left(\frac{\partial \mathbf{E}}{\partial \mathbf{b}_B}\right). \tag{20}$$

Following the same logic, the coefficient component can be written as the sum of individual contributions as,

$$C = \sum_{k=1}^{K} C_k = \sum_{k=1}^{K} W_{\Delta b_k}\{\overline{F(\mathbf{x}'_W\mathbf{b}_B)} - \overline{F(\mathbf{x}'_W\mathbf{b}_W)}\}. \tag{21}$$

[6] In this case $F(.) = f(.)$.

Each covariate's contribution to the overall coefficient component depends on the parameter vectors, $\mathbf{b}_B$ and $\mathbf{b}_W$. The *k*th elements of the respective gradients are

$$\frac{\partial C_k}{\partial b_B} = W_{\Delta b_k}\,\overline{f(\mathbf{x}'_W\mathbf{b}_B)x_{W_k}} + w_{b_k}\,\overline{F(\mathbf{x}'_W\mathbf{b}_B)} \tag{22}$$

and

$$\frac{\partial C_k}{\partial b_{W_k}} = w_{b_k}\,\overline{F(\mathbf{x}'_W\mathbf{b}_W)} - W_{\Delta b_k}\,\overline{f(\mathbf{x}'_W\mathbf{b}_W)x_{W_k}} \tag{23}$$

where

$$w_{b_k} = \frac{\partial W_{\Delta b_k}}{\partial b_{jk}} = \frac{\overline{x}_{Wk}}{\sum_k \overline{x}_{Wk}(b_{Bk} - b_{Wk})} - \frac{\overline{x}^2_{Wk}(b_{Bk} - b_{Wk})}{\left\{\sum_k \overline{x}_{Wk}(b_{Bk} - b_{Wk})\right\}^2}. \tag{24}$$

when $j = W$ this quantity has the opposite sign. Letting $\text{var}(\mathbf{b}_B)$ and $\text{var}(\mathbf{b}_W)$ denote the covariance matrix of the estimates from black and white models, respectively, and let $\mathbf{C}$ be a $K \times K$ matrix with $C_1, \ldots, C_K$ on the main diagonal and zeros elsewhere, the large sample (co)variance matrix of the detailed coefficient components is

$$\text{var}(\mathbf{C}) = \left(\frac{\partial \mathbf{C}}{\partial \mathbf{b}_B}\right)' \text{var}(\mathbf{b}_B)\left(\frac{\partial \mathbf{C}}{\partial \mathbf{b}_B}\right) + \left(\frac{\partial \mathbf{C}}{\partial \mathbf{b}_W}\right)' \text{var}(\mathbf{b}_W)\left(\frac{\partial \mathbf{C}}{\partial \mathbf{b}_W}\right) \tag{25}$$

Significance tests on individual components, blocks of components, or for the overall decomposition as a whole, can be carried out using Wald tests by redefining **E** and **C** to include a subset of the original set of terms along with the corresponding sub-matrices of $\text{var}(\mathbf{b}_B)$ and $\text{var}(\mathbf{b}_W)$. The variance estimates derived above assume that the

independent variables are fixed and that groups are independent. They will underestimate the true variances if this is not the case.

It would also be possible to obtain a bootstrapped distribution of the components by applying a repeated modeling approach. Alternatively, a Bayesian approach can be used to obtain the posterior distribution of each component using a Markov Chain Monte Carlo (MCMC) method as outlined by Radchenko and Yun (2003). An alternative to bootstrapping or a full Bayesian approach is to simulate the distributions of each component by drawing *M* parameter vectors for each group, carrying out the decomposition on the simulated parameter vectors, and obtaining means and variances of the resulting distributions of the decomposition components. Specifically, let

$$\mathbf{b}_j^m \sim MVN(\mathbf{b}_j, \boldsymbol{\Sigma}_{\mathbf{b}_j}) \tag{26}$$

denote the *m*th simulated parameter vector from the *j*th group, which is assumed to follow a multivariate normal distribution centered around the MLE's, with variance/covariance $\boldsymbol{\Sigma}_{\mathbf{b}}$. With no loss of generality, $\boldsymbol{\Sigma}_{\mathbf{b}}$ could be drawn from an inverse-Wishart distribution to allow for sampling variation in the covariances. Under this approach, the decomposition is carried out *M* times, resulting in a posterior predictive distribution for each quantity in the decomposition (see e.g., Lynch and Western 2004). Statistical inference can be carried on the quantities from the resulting distributions.

5. Example

Race/ethnic differences in the risk of out-of-wedlock childbearing are routinely examined using group-specific hazard rate models or models in which race/ethnicity is

included as a risk factor. Although this approach yields insight into the relative importance of key predictors of nonmarital fertility for different race/ethnic groups, it cannot answer questions about the relative contributions of race differences in characteristics and effects to the absolute race/ethnic differences in rates. In particular, to what extent is the racial difference in rates attributable to compositional differences in predictors—such as what might be reflected by group differences in socioeconomic resources and family structure—and to differences in the effects of these predictors (i.e., the group differences in behavioral responses to these characteristics)?

We decompose the observed black-white difference in premarital birth rates into compositional and return-to-risk components. The decomposition is carried out at the aggregate and detailed levels, thus allowing an assessment of the contribution of each model predictor to the racial gap. For research on first nonmarital fertility transitions, this type of analysis provides a way to assess the contributions of socioeconomic background and family structure, whose effects and distributions differ by race.

Data from the 1979 National Longitudinal Survey of Youth (NLSY79 Center for Human Resource Research 1979) are used to model first non-marital fertility transitions (i.e., first premarital birth) for blacks and whites using proportional hazards models. We adopt a parsimonious model specification using covariates that have been widely used in past research including: (1) family background characteristics: (mother's education, adjusted family income[7] and number of older siblings) and (2) family structure characteristics: (mother's age at respondent's birth, proportion of years living in single

[7] Adjusted Income $= \text{family income} / (10{,}000 \times \sqrt{\text{family size}})$.

mother family, and number of family changes up to the time of the event or before age 18, whichever occurs first). The latter two variables are computed using the 18-year living arrangement histories in the NLSY (see, e.g., Wu and Thomson 2000, Powers 2005).

The estimated black-white difference in the crude rates of first premarital birth is 0.02048 ($r_B - r_W = 0.02530 - 0.00482 = 0.02048$). To facilitate the presentation of results, we express this difference as 20.48 births (per year of age) per 1,000 women. Table 1 presents covariate means and model estimates for each group as well as the crude rates per 1,000 and race differences in rates. Table 2 provides the detailed decomposition obtained by averaging the results of separate decompositions with interchanged reference and comparison groups.[8] The contributions have been multiplied by 1,000 to reflect increases or decreases in the gap in terms of numbers of births per year of age per 1,000 women. Under the current model, compositional differences between blacks and whites (i.e., differences in levels of resources and family structure) contribute 5.16 births per 1,000 (25.2%) to the overall gap, whereas black-white differences in covariate effects (i.e., the returns-to-risk of these characteristics) contribute 15.32 births per 1,000 (74.8%) to the estimated difference.

[**Tables 1 and 2 about here**]

We first discuss the contributions of the substantive predictors to the overall premarital birth rate gap. We shall discuss the baseline hazard contribution later. Table 2 shows the detailed decomposition for the family background and family structure variables. A positive characteristic effect, E_k, indicates the amount that the black-white

[8] The results reported above were estimated using a computer routine written in R (R Development Core Team 2005) available upon request.

gap would decrease if the group difference in variable *k* would disappear. Based on the results from the proportional hazard models (Table 1), each change in family structure (or family transition) increases the risk of a first premarital birth by 16% for blacks and 34% for whites. However, whites experience fewer of these transitions on average than blacks, with means of 0.49 and 0.62 transitions, respectively. The results in Table 2 show that with respect to the (white) reference group, this compositional disadvantage for black women contributes 0.46, or about 2.2%, to the overall difference. Turning to the income effect, we see from Table 1 that a $10,000 increase in adjusted family income is associated with a 34% and 43% decrease in the risk of premarital birth for whites and blacks, respectively. Despite similar returns to income, average income in black families in the NLSY is 55 percent that of white families. From Table 2 we see that the difference in family income by race accounts for 4.37 births per 1,000 women, which comprises over 21% of the overall racial difference in rates. Among the compositional factors considered here, making family incomes and number of older siblings in the comparison population (blacks) equal to that of the reference population (whites) would produce the largest reductions in the racial gap in the premarital birth rate.

A similar interpretation applies to the effects component, C_k. A *negative* coefficient indicates the expected *increase* in the black-white gap if blacks experienced the same returns-to-risk as whites. For example, if we consider the "number of older siblings" effects reported in Table 1, each additional elder sibling is expected to increase a woman's risk of a premarital birth by 18.2% and 5.3% for white and black women, respectively. From Table 2, we find that the overall black-white gap would be expected to

increase by 2.82 births per 1,000 (7.7%) if black women were penalized by the number older siblings to the same extent as white women. Similarly, *a positive C*-coefficient reflects the expected *decrease* in the black-white gap due to equalizing an effect to the white level. For example, whites and blacks experience different returns to maternal education. Based on the results from the proportional hazards model in Table 1, each additional year of mother's schooling reduces the risk of premarital birth by 12.6% for whites and 6.3% for blacks. The decomposition results in Table 2 show that if blacks benefitted from higher levels of maternal education to the same degree as whites, then we would expect the black-white gap in the premarital birth rate to decrease by 8.11 births per 1,000, or 39.6% of the overall gap. Differences in returns to maternal education, as well as differences in the effects of maternal age at respondent's birth are the largest contributors to the overall gap.

If we were to consider a hypothetical policy designed to reduce the black-white gap in the premarital birth rate, then equalizing socioeconomic resources across groups would lead to a larger decrease in the compositional portion of the gap than would making groups more similar in terms of family structure (number of family transitions, proportion of years spent in a single mother family, and number of older siblings). However, a greater share of the total differential can be attributed to differences in the effects of maternal education and mother's age at respondent's birth, so equalizing these effects across groups would yield the greatest reduction in the black-white premarital birth rate gap. It is probably safe to say that changing behavioral responses presents a

more challenging task from a policy perspective than equalizing socioeconomic resources across groups.

Baseline Hazard Components

Compositional components involving the dummy variables for the age intervals that define the baseline hazard play the same role as the constant term in a standard multivariate decomposition. The piecewise constant hazard model effectively partitions the constant term into several pieces, with individuals differing on the number of pieces they contribute. In standard models, the mean value of the constant is always 1 and the difference in means across groups is always 0. For the decomposition of the piecewise constant hazard model, the characteristics effects associated with the pieces of the baseline hazard reflect race differences in the distribution of exposure, which in-turn is a function of race differences in the age distribution of events and censoring. The fact that the characteristics effects of the baseline hazard reported in the first panel of Table 2 are at first negative and then positive, reflects that the age distribution of events is centered at a younger age for black women and at an older age for white women.

The coefficient effects for the baseline hazard are informative about the contribution of racial differences in the age-specific baseline hazard rates to racial gap in premarital birth rates. Taken together, group differences in the logged baseline hazards (i.e., the coefficients pertaining to the age-interval of the event from the model) account for about 7 births per 1,000, or 37%, of the racial gap (Table 2). This is the expected reduction in the gap if blacks were to experience the same age-specific baseline rates as

whites. We find that the largest contributors to the differential are attributed to differences in the baseline hazards pertaining to the first 3 age intervals, which reflect the race differences in the underlying rates for teenagers.

Nonproportional Hazards

As mentioned earlier, it is possible to incorporate nonproportional effects via interactions with the age-interval dummy variables. For example, we could introduce a $D_l \times z$ interaction into Eq. (9). Adding these types of interactions presents no additional difficulty in the decomposition procedure *per se*. However, the characteristics effects for an interaction term will reflect differences in exposure in age interval *l* in addition to differences in characteristics for those at risk in age interval *l*. Thus, the characteristic effects associated with age-interval interactions are somewhat ambiguous. However, the coefficient effects have a straightforward interpretation.

In these data we find evidence of an age-varying effect of family income in the sample of non-Hispanic white women, with different effects on the risk for those aged 24 and younger and for those at older than 24. We fit a nonproportional effects model to both groups that includes one family income effect on the risk in 12-24 year old age-interval and one income effect on the risk beyond age 24. We refer to this as Model 2. Table 3 provides the relevant income effects for both groups as well as the decomposition results from Model 2 and from the original model (Model 1). We find that for white women, a $10,000 change in family income yields a 62% reduction in the risk of a premarital birth in the 12-24 age interval, whereas the same change in family income

increases the risk of a premarital birth by 63% at older ages. For black women, increased income has no significant effect on the risk of premarital births at older ages.

[**Table 3 about here**]

The decomposition results in Table 3 show that compositional differences in income levels for those at risk of events in the 12-24 age interval account for 23% of the total racial gap according to Model 2. This is similar to the contribution of 21% in Model 1, which pertains to all ages. Differences in the effects of income at younger ages account for 12% of the total gap in Model 2. Therefore, equalizing the returns to income for younger women would be expected to reduce the black-white gap by 2.5 births per 1,000 as a result of the larger race difference in age 12-24 income effects in Model 2. This is in sharp contrast to Model 1 where the effects of income are similar by race and where the racial difference in income effects comprises a negligible portion of the black-white difference in the premarital birth rate. Race differences in effects of income at older ages account for a small portion of the total gap, and equalizing income effects for older women would be expected to increase the gap in the premarital birth rate by less than 1 birth per 1,000 women.

Under Model 2, differences in characteristics account for 27.4% of the racial gap in premarital birth rates while differences in effects account for 72.6%. These results are similar to those from Model 1 (25.2% and 74.8%, respectively). While the overall contributions are similar, it should be noted that the income $\times$ age interaction in Model 2 necessarily impacts the baseline hazard, thus blurring the distinction between the baseline hazard and the structural part of the model to some extent. We find that race differences

in the baseline hazard account for 13% of the overall gap in Model 2 compared to 37% of the gap in Model 1.

6. Discussion

Multivariate decomposition provides a way to partition an *observed group difference* in statistics into portions that can be attributed to group differences in characteristics, or endowments, and to group differences in effects, or coefficients from a multivariate model. The statistics of interest could be means, proportions, counts, and (as shown here) rates. Multivariate decomposition of a difference in means involves a straightforward substitution technique based on results from OLS regression models, and has been widely used over the past several decades. Techniques for multivariate decomposition for nonlinear models have been developed more recently, but have not thus far been extended to hazard rate models.

Decomposition Using Cox Regression Models and Discrete-Time Hazard Models

Our proposed method involves a simple parametric model with a flexible functional form for the baseline hazard. Due to its popularity, a multivariate decomposition technique based on the Cox regression model would seem particularly attractive. However, applying our decomposition approach to the Cox model poses several problems. Because the baseline hazard is unspecified in Cox regression, it is not directly available as a decomposable part of the model. Moreover, although we could retrieve the baseline hazard using numerical methods, there is no guarantee that a

difference in the hazard (at age *t*) across groups would exist due to group differences in the ages that events occur. For example, in the NLSY, the youngest and oldest event ages are 11.67 and 33.83 in the black subsample and 13.83 and 35.33 in the white subsample. Therefore, applying this approach to a Cox regression model would require that the number of unique event times be the same and that the event times are equal for both groups. Despite these difficulties, there may be alternative decomposition techniques that are better suited to the Cox model. Here we offer a practical alternative to approximate the baseline hazard of the Cox regression model using a step function defined along a common set of cut points.

Discrete-time models are widely used in social science research. These models are generally estimated with logit models fit to person-period data. The decomposition of differences in predictions from discrete time logit hazard models is straightforward. We assume that data are structured in the person-period format as previously discussed. As with the piecewise constant rate models, the number of events in group *j* equals the sum of the predicted probabilities over the person-periods of exposure

$$\sum_{i=1}^{N_j}\sum_{l=1}^{n_{ij}} d_{ilj} = \sum_{i=1}^{N_j}\sum_{l=1}^{n_{ij}} F(\mathbf{x}'_{ilj}\mathbf{b}_j), \tag{27}$$

where $F(\mathbf{x}_{ilj}\mathbf{b}_j) = \dfrac{\exp(a_{lj} + \mathbf{z}'_{ilj}\boldsymbol{\gamma})}{1+\exp(a_{lj} + \mathbf{z}'_{ilj}\boldsymbol{\gamma})}$, and a_l, $\mathbf{z}$, and $\boldsymbol{\gamma}$ are defined in the same way as before. The empirical probabilities to be differenced take the form:

$$p_j = \overline{F(\mathbf{x}'_{ij}\mathbf{b}_j)} = \frac{\sum_{i=1}^{N_j}\sum_{l=1}^{n_{ij}} F(\mathbf{x}'_{ilj}\mathbf{b}_j)}{N_j}. \tag{28}$$

Some minor changes to the formulas for the standard errors of the decomposition quantities are needed. Specifically, the analogous expression for Eq. (19) is

$$\overline{f(\mathbf{x}_j'\mathbf{b}_j)x_{j_k}} = \frac{\sum_{i=1}^{N_j}\sum_{l=1}^{n_{ij}} \partial F(\mathbf{x}_{ilj}'\mathbf{b}_j)/\partial b_{j_k}}{N_j} = \frac{\sum_{i=1}^{N_j}\sum_{l=1}^{n_{ij}} f(\mathbf{x}_{ilj}'\mathbf{b}_j)x_{ilj_k}}{N_j}, \tag{29}$$

where $f(\mathbf{x}_{ilj}'\mathbf{b}_j) = \dfrac{\exp(\mathbf{x}_{ilj}'\mathbf{b}_j)}{[1+\exp(\mathbf{x}_{ilj}'\mathbf{b}_j)]^2}$.

Effect Normalization

Although often overlooked, it is well known that estimates corresponding to dummy variables in the coefficient component of the decomposition are not invariant to the choice of the reference categories for the dummy variables appearing in the model (Oaxaca and Ransom 1999). Adopting a particular reference category necessarily affects the estimates of the effects corresponding to other factor levels as well as the constant term. For models such as ours, where the baseline hazard is absorbed into the intercept, this implies that the decomposition of the baseline hazard is sensitive to the normalization of the dummy variables. This parameter invariance can be remedied by carrying out a separate decomposition for each of the possible normalizations of the dummy variables and averaging the results. Alternatively, it is possible to augment the coefficient and design matrices that are passed to the decomposition routine as suggested by Yun (2005b). Either approach yields estimates of C_k and E_k for all levels of the factors.

Table 4 shows the coefficients pertaining to the log baseline hazard for whites and blacks under alternative specifications. The coefficients appearing in the row labeled

normalization 1 are those obtained from fitting our original model (6 dummy variables for the age intervals without the conventional intercept). The baseline rates, adjusted for covariates, are given by the exponentiated values in the 2nd row. The coefficients appearing in normalization 2 are those from a model in which the baseline hazard is parameterized by including a constant term plus 5 dummy variables for the 2nd through 6th age categories. The row labeled normalization 2n contains the grand-mean centered effects, with exponentiated values in parentheses. This is an augmented parameter vector constructed from the coefficients in normalization 1. The effects corresponding to the age intervals are deviations from the grand mean, and thus sum to 0. The coefficients under normalization 2n can be obtained from those in normalization 1 by letting a_l denote a coefficient from normalization 1. The grand mean is obtained by averaging over the 6 pieces of the log baseline hazard $\bar{a} = \frac{1}{6}\sum_{l=1}^{6} a_l$. The normalized effects are $a_l^* = a_l - \bar{a}$.

Table 5 provides results from the decomposition using the grand mean centered coefficients (the a_l^*'s in normalization 2n) and the corresponding augmented design matrix. We show only the coefficients component corresponding to the baseline hazard.[9]

[**Tables 4 and 5 about here**]

We now obtain a coefficient effect for the "constant," which reflects the black-white difference in the *mean* baseline rate after adjusting for covariates. The coefficients for the age intervals show how this difference is adjusted by age. Race differences in the mean baseline rates contribute 5.64 births per 1,000 (27%) to the total differential. This

[9] An augmented covariance matrix is required to obtain standard errors under this normalization.

reflects the higher rate for black women at the mean baseline (0.21 vs. 0.12 from the exponentiated values in rows 8 and 4 of Table 4). The positive signs of the coefficient effects associated with the 12-18 age intervals reflect higher rates of early childbearing for blacks. Similarly, the negative coefficient associated with the 24+ age interval reflects the higher rate for white women in this interval.

Limitations

Multivariate decomposition is facilitated by the availability of several computer routines in standard statistical packages. For example, multivariate decomposition can be carried out in Stata using the packages: **`oaxaca`** (Jann 2008), **`gdecomp`** (Bartus 2006), and **`fairlie`** (Jann 2006), and **`nldecompose`** (Sinning, Hahn, and Bauer 2008). SAS macros also exist for Fairlie's method. Stata's **`fairlie`** package (and the SAS macro) decomposes a difference in proportions based on logit or probit models into the characteristics portion only, whereas **`gdecomp`** provides both components and extends to models for count data. The **`nldecompose`** handles a variety of nonlinear models, but does not carry out a detailed decomposition. The **`oaxaca`** package handles differences in means using results from the classical linear model.

Currently no add-on routines for commercial packages exist for carrying out a multivariate decomposition of a difference in rates. However, it would be feasible to modify existing routines such as **`gdecomp`** (Bartus 2006) to include an offset term. Another limitation might be the narrow focus on differences in first moments. The methods outlined above adhere to the standard logic of multivariate decomposition by

following in the spirit of previous research in which observed differences are the quantities to be decomposed. It is also possible to decompose a difference in logits, log-odds ratios, or relative risks. For example, we could ignore the baseline hazard in a Cox regression and decompose the difference in the log of the risk scores, where the estimated risk score is defined as $\exp(\mathbf{x}_i'\mathbf{b})$, and where $\theta_j = \overline{\mathbf{x}}_j'\mathbf{b}_j$

$$\theta_1 - \theta_2 = \mathbf{b}_1'(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) + \overline{\mathbf{x}}_2'(\mathbf{b}_1 - \mathbf{b}_2). \quad (30)$$

This leads to a much simpler decomposition involving only means and effects. However, there is no sample analog for this difference and it is model dependent, whereas the differences in sample statistics are fixed for the samples involved.

The decomposability of the baseline hazard is an important issue for the models considered here. In these data, the two groups are distributed differently across the segments of the baseline hazard, but the hazard varies across segments. That is, different weight is given to different pieces of the baseline hazard in the computation of the group-specific rates. Investigation into the sources of differences in distributions across pieces of the baseline hazard may be a fruitful area for further research. A potential source of difference is early censoring, such as what might occur when one group, on average, marries at a younger age resulting in reduced exposure to risk at later ages. In this case, decomposition might be carried out to identify an additional component due to differences in age at marriage and other sources of censoring. Thus, additional information can be used to obtain alternative decomposition estimates. However, in the absence of additional information, the method that we propose obtains estimates in a straightforward manner using only model estimates and covariates.

Another possible source of difference in the baseline hazard stems from the dynamic nature of the hazard rate model insofar as the amount of exposure at later ages depends on the conditional rates at earlier ages. Alternative approaches might be developed to take account of this when building the counterfactuals for the decomposition. For example, if the coefficients for blacks are used to compute the counterfactual rate for whites, the counterfactual could be adjusted for the fact that with the substituted coefficients, the expected exposure distribution for whites would change. This implies that part of the differential attributed to independent variables encompasses the indirect contribution through effects on the exposure distribution. For a decomposition constructed in this way, the contribution of the baseline hazard would then only reflect the differences due to group-specific differences in censoring (e.g., because of group differences in the average age at marriage).[10] However, this approach may require additional assumptions to generate the expected exposure sets. Compared to this approach, our proposed Oaxaca-type decomposition provides a straightforward approach to obtain counterfactuals.

A further limitation is that all decomposition methods are sensitive to model specification. Although the decomposition of differences in rates outlined above guarantees a partitioning of components that necessarily sum to the observed difference, the detailed results are sensitive to model specification insofar as adding or removing covariates will affect the allocation of overall difference to the constituent parts. Therefore, the context in which this method is used is important as it depends on a well-

[10] We thank a reviewer for suggesting this extension.

specified model—such as the final model among a set of competing models—a as well as a strong substantive motivation for that model.

References

Aitkin, M. and D. Clayton 1980. "The Fitting of Exponential Weibull, and Extreme Value Distributions to Complex Survival Data using GLIM." *Applied Statistics*, 29: 156-163.

Allison, P.D. 1982. "Discrete-Time Methods for the Analysis of Event Histories." Pp. 61-98. In S. Leinhardt (Ed.), *Sociological Methodology*, San Francisco: Jossey-Bass.

Althauser, R. P., and M. Wigler. 1972. "Standardization and Component Analysis." *Sociological Methods and Research* 1: 97–135.

Barlow, R. E., and F. Proschan .1975. *Statistical Theory of Reliability and Life Testing*. New York: Holt, Rinehart and Winston.

Bartus, T. 2006. "Marginal Effects and Extending the Blinder-Oaxaca Decomposition for Nonlinear Models," UK Stata User's Group Meetings 2006 05, *Stata Users Group*.

Bauer, T, S. Göhlmann, and M. Sinning,. 2007. "Gender Differences in Smoking Behavior." *Health Economics*, 16: 895-909 .

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

Borooah, V.K., and S. Iyer. 2005. "The Decomposition of Inter-Group Differences in a Logit Model: Extending the Oaxaca-Blinder Approach with an Application to School Enrollment in India." *Journal of Economic and Social Measurement*, 30: 279-293.

Center for Human Resource Research. 1979. The National Longitudinal Survey of Youth Handbook, Columbus, Ohio: The Ohio State University.

Coleman, J.S., and Z.D. Blum .1971. "Note on the Decomposition of Differences between Two Groups," Johns Hopkins University, (unpublished).

Coleman, J.S., C.C. Berry, and Z.D. Blum. 1971. "White and Black Careers during the First Ten Years of Work Experience: A Simultaneous Consideration of Occupational Status and Income Changes." Johns Hopkins University Center for Social Organization of Schools Report 76.

Cox, D.R. 1972. "Regression Models and Life Tables (with discussion)." *Journal of the Royal Statistical Society, Series B*, 28: 150-163.

Das Gupta, P. 1993. *Standardization and Decomposition of Rates: A User's Manual.* Washington, D.C.: U.S. Government Printing Office.

Duncan, O.D. 1968. "Patterns of Occupational Mobility Among Negro Men." *Demography*, 5: 11-22.

Duncan, O.D. 1969. "Inheritance of Poverty or Inheritance of Race?", pp 85-110 in D.P. Moynihan (Ed.) *On Understanding Poverty*, New York: Basic Books.

Duncan, O.D., D.L. Featherman, and B. Duncan. 1968. "Socioeconomic Background and Occupational Achievement: Extensions of a Basic Model." Washington D.C., U.S. Department of Health, Education and Welfare.

Even, W.E., and D.A. Macpherson. 1993. "The Decline of Private-Sector Unionization and the Gender Wage Gap." *Journal of Human Resources*, 28: 279-296

Fairlie, R.W. 2005. "An Extension of the Blinder-Oaxaca Decomposition Technique to Logit and Probit Models." *Journal of Economic and Social Measurement*, 30: 295-304.

Gomulka, J., and N. Stern .1990. "The Employment of Married Women in the United Kingdom 1970-83. *Econometrica*, 57: 171-199.

Heitmueller, A. 2004. "Job Mobility in Britain : Are the Scots Different? Evidence from the BHPS." *Scottish Journal of Political Economy*, 51: 329-358.

Holford, T. 1976. "Life Tables with Concomitant Information." *Biometrics*, 32: 587-597.

Holford, T. 1980. "The Analysis of Rates and of Survivorship Using Log-Linear Models." *Biometrics*, 36: 299-305.

Jann, B. 2006. "fairlie: Stata Module to Generate Nonlinear Decomposition of Binary Outcome Differentials." Available from http://ideas.repec.org/c/boc/bocode/s456727.html.

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

Kitagawa, E.M. 1955. "Components of a Difference between Two Rates." *Journal of the American Statistical Association*, 50: 1168-1194.

Laird, N., and D. Oliver.1981. "Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques." *Journal of the American Statistical Association*, 76:231-240.

Lynch, S. M., and B. Western .2004. "Bayesian Posterior Predictive Checks for Complex Models." *Sociological Methods and Research*, 32:301-335.

Nielsen, H.S. 1998. “Discrimination and Detailed Decomposition in a Logit Model.” *Economics Letters*, 61: 115-120.

Neumark, D. 1988. “Employers' Discriminatory Behavior and the Estimation of Wage Discrimination,” *Journal of Human Resources*, 23:279-295.

Oaxaca, R.L. 1973. “Male-Female Wage Differentials in Urban Labor Markets.” *International Economic Review*, 14: 693-709.

Oaxaca, R. L. and M. R. Ransom .1988. “Searching for the Effect of Unionism on the Wages of Union and Nonunion Workers,” *Journal of Labor Research*, 9: 139-148.

Oaxaca, R. L. and M. R. Ransom.1994. “On Discrimination and the Decomposition of Wage Differentials,” *Journal of Econometrics*, 61:5-21.

Oaxaca, R. L., and M. R. Ransom. 1999. “Identification in Detailed Wage Decompositions.” *Review of Economics and Statistics*, 81: 154-57.

Park, T.A. and L. Lohr. 2008. “A Oaxaca-Blinder Decomposition for Count Data Models.” *Applied Economics Letters*, (forthcoming).

Powers, D. A. 2001. “Unobserved Family Effects on the Risk of a First Premarital Birth.” *Social Science Research*, 30: 1-24.

Powers, D.A. 2005. “Effects of Family Structure on the Risk of First Premarital Birth in the Presence of Correlated Unmeasured Family Effects.” *Social Science Research*, 34: 511-537.

R Development Core Team .2005. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Radchenko, S. I., and M.-S. Yun. 2003. "A Bayesian Approach to Decomposing Wage Differentials." *Economics Letters*, 78: 431-36.

Rao, C.R. 1973. *Linear Statistical Inference and Its Applications*. New York: Wiley.

Sinning, M., M. Hahn, and T. K. Bauer .2008. "The Blinder-Oaxaca Decomposition for Nonlinear Regression Models." *The Stata Journal*. 8: 480-492.

Shorrocks, A.F. 1980. "The Class of Additively Decomposable Inequality Measures," *Econometrica*. 48: 613-625.

Shorrocks, A.F. 1982. "Inequality Decomposition by Factor Components," *Econometrica*. 50: 193-212.

Smith, H.L., S.P. Morgan, and T. Koropeckyj-Cox. 1996. "A Decomposition of Trends in Nonmarital Fertility Ratios of Blacks and Whites in the United States, 1960-1992." *Demography*, 33: 141-151.

Sweeney, M.M., and J.A. Phillips. 2004. "Understanding Racial Differences in Marital Disruption: Recent Trends and Explanations." *Journal of Marriage and Family* 66:239-250.

Van Hook, J., Brown, S.L., Kwenda, M.N. 2004. "A Decomposition of Trends in Poverty Among Children of Immigrants." *Demography*, 41: 649-670.

Williams, D.F. 1991. "Structural Change and the Aggregate Poverty Rate." *Demography*, 28: 323-332.

Winsborough, H.H., P. Dickinson. 1971. "Components of Negro-White Income Differences." Proceedings of the American Statistical Association, Social Statistics Section: 6-8.

Wu, L.L., and B. Martinson. 1993. “Family Structure and the Risk of a Premarital Birth.” *American Sociological Review*, 58: 210-232.

Wu, L.L., and E. Thomson. 2000. “Race Differences in Family Experience and Early Sexual Initiation: Dynamic Models of Family Structure and Family Change.” *Journal of Marriage and the Family*, 63: 682-696.

Yun, M-S. 2004. “Decomposing Differences in the First Moment." *Economics Letters*, 82: 275-280.

Yun, M-S. 2005a. “Hypothesis Tests when Decomposing Differences in the First Moment.” *Journal of Economic and Social Measurement*, 30: 305-319.

Yun, M-S. 2005b. “A Simple Solution to the Identification Problem in Detailed Wage Decompositions.” *Economic Inquiry*, 43: 766-772.

Table 1: Means, Effects (hazard ratios and baseline hazards), and Event Percentages.

	Blacks			Whites		
Independent Variables	**Means**	e^b	$Z^\dagger$	**Means**	e^b	$Z^\dagger$
pct. years in single mother family	0.24	1.117	1.02	0.06	2.166	2.41
number of family changes	0.62	1.162	3.82	0.49	1.336	5.87
mother's schooling	10.76	0.937	-4.38	11.87	0.874	-5.69
Adjusted family income X 10,000	0.55	0.570	-5.15	1.00	0.658	-3.09
number of older siblings	2.72	1.053	3.01	1.90	1.182	4.61
mother's age at R's birth	24.91	0.971	-4.30	25.48	0.949	-4.09
Baseline Hazard Age Intervals	**% Events**	e^b	$Z^\dagger$	**% Events**	e^b	$Z^\dagger$
[12, 16)	5.60	0.014	-16.99	0.74	0.006	-11.39
[16, 18)	15.03	0.367	-4.24	2.45	0.176	-4.30
[18, 20)	14.00	0.455	-3.30	2.58	0.244	-3.44
[20, 22)	9.87	0.473	-3.06	2.06	0.276	-3.07
[22, 24)	5.16	0.348	-4.06	0.87	0.165	-3.97
[24+)	6.26	0.194	-6.42	2.49	0.245	-3.39
Event Percentage	55.93			11.19		
Crude Rates $\times 1000$	25.30			4.82		
Black-White Difference in Rates = 20.48						
N	1,357			2,287		

$^\dagger Z = b / \mathrm{se}(b)$.

Table 2: Decomposition into Characteristics (*E*) and Coefficients (*C*) Components.

		95% CI				95% CI		
Independent Variables	***E*** (×1000)	**lower**	**upper**	**% of total**	***C*** (×1000)	**lower**	**upper**	**% of total**
pct. years in single mother family	0.89	-0.77	2.54	4.33	-1.11	-1.82	-0.39	-5.42
number of family changes	0.46	0.39	0.52	2.23	-0.81	-1.24	-0.37	-3.95
mother's schooling	1.68	-0.71	4.07	8.19	8.11	5.18	11.11	39.60
adjusted family income X 10,000	4.37	2.26	6.48	21.32	-1.09	-2.81	0.63	-5.33
number of older siblings	1.19	0.28	2.11	5.83	-2.82	-4.17	-1.46	-13.74
mother's age at R's birth	0.36	-1.16	1.89	1.77	6.00	3.60	8.39	29.28
Baseline Hazard Age Intervals								
[12, 16)	-5.02	-6.25	-3.79	-24.51	2.75	1.36	4.14	13.43
[16, 18)	-0.44	-0.83	-0.04	-2.13	1.97	0.79	3.14	9.61
[18, 20)	0.38	-0.30	1.05	1.83	1.19	0.26	2.11	5.80
[20, 22)	0.36	-0.34	1.06	1.76	0.66	0.02	1.30	3.22
[22, 24)	0.49	-0.02	1.00	2.39	0.60	0.14	1.06	2.93
[24+)	0.45	0.09	0.81	2.18	-0.13	-0.44	0.18	-0.64
	$\sum E_k = 5.16$				$\sum C_k = 15.32$			
Overall Contributions	**95%CI** [3.43 – 6.89]			25.21	**95%CI** [12.88 – 17.77]			74.79

Note: % of total is the percentage share of the differential in crude rates of 20.48 between blacks (25.30 per 1,000) and whites (4.82 per 1,000). Results are the average of two decompositions.

Table 3: Proportional and Nonproportional Effects of Family Income and Decomposition Components

	Blacks e^b	Whites e^b	*E*	% of Total	*C*	% of Total
Model 1: Proportional Effect of Family Income						
family income	0.570 *	0.658 *	4.37 *	21.32	-1.09	-5.33
Model 2: Nonproportional Effect of Family Income						
family income X age [12,24)	0.553 *	0.385 *	4.78 *	21.25	2.51 *	12.23
family income X age [24+)	0.699	1.629 *	0.12	0.45	-0.40 *	-1.95

* $p < 0.05$.

Table 4. Alternative Normalizations of the Log Baseline Hazard

	Normalization	Constant or Grand Mean	Age Intervals [12,16)	[16,18)	[18,20)	[20,22)	[22,24)	[24+)
Whites	**1**	---	-5.14	-1.74	-1.41	-1.29	-1.80	-1.41
	(exp) 1		(0.005)	(0.18)	(0.24)	(0.28)	(0.17)	(0.24)
	2	-5.14	---	3.40	3.73	3.85	3.34	3.73
	2n	-2.13	-3.00	0.39	0.72	0.84	0.33	0.72
	(exp) 2n	(0.12)	(0.05)	(1.48)	(2.05)	(2.32)	(1.39)	(2.05)
Blacks	**1**	---	-4.26	-1.00	-0.79	-0.75	-1.06	-1.64
	(exp) 1		(0.014)	(0.37)	(0.45)	(0.47)	(0.35)	(0.19)
	2	-4.26	---	3.26	3.47	3.51	3.20	2.62
	2n	-1.58	-2.68	0.58	0.79	0.83	0.52	-0.60
	(exp) 2n	(0.21)	(0.07)	(1.79)	(2.20)	(2.29)	(1.68)	(0.55)

Note: See text for details on normalizations 1, 2, and 2n.

Table 5: Normalized Coefficients Component for Baseline Hazard

		95% CI		
Variable	C $(\times 1000)$	**lower**	**upper**	**% of Total**
Constant	5.64	2.29	8.99	27.52
Baseline Hazard Age Intervals				
[12,16)	1.03	0.33	1.73	5.03
[16,18)	0.52	0.12	0.91	2.51
[18,20)	0.14	-0.16	0.44	0.69
[20,22)	-0.01	-0.23	0.21	-0.06
[22,24)	0.16	-0.03	0.35	0.78
[24+)	-0.43	-0.55	-0.32	-2.12