HISTORICAL COVERAGE PATTERNS OF BLACKS IN THE UNITED STATES CENSUS: AN APPLICATION OF AGE-PERIOD-COHORT ANALYSIS

Katherine M. Condon, Ph.D.
Population Division, U.S. Census Bureau

This abstract is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on methodological issues are those of the author and not necessarily those of the U.S. Census Bureau.

INTRODUCTION

This paper builds on work that examined coverage patterns in a series of censuses from 1880 to 1950 for the White Native-born population by fitting a generalized linear model to predict population by 5-year age groups, census, and cohort (Condon, et al., 2008). In this research, the analysis will be extended to examine coverage patterns of Blacks by 5-year age groups, census and cohort for the period 1900 to 1950, using the same methodology. The population data will come from published census data for 1900 to 1950. Average relative rates of coverage and both the mean absolute percent error (MAPE) and mean algebraic percent error (MALPE) will be estimated for the Black population. In addition, there will be a comparison of the results with those found with this method for the White Native-born population (1900-1950).

METHODOLOGY

This is a method for estimating enumeration completeness for United States censuses from 1900 to 1950 for the Black population using an age-period-cohort accounting framework. This is a useful methodology for studying enumeration completeness for specific sub-populations (e.g., White native-born population by Condon, et al., 2008), because it is a relatively "closed" population, i.e., population change between two points in time is due to births, deaths, or aging (Glenn, 2005). Using age-period-cohort analysis, it is proposed to estimate the pattern of undercount for each census and each age group. The results from this new approach will offer a useful check on earlier demographic estimates of census completeness. In addition, this approach addresses the "substantive issues relating to aging and social and cultural change …" (Glenn, 2005: 1).

Like the demographic analysis method for estimating enumeration completeness, the objective is to estimate the pattern of coverage for each census and each age group in the United States. However, unlike demographic analysis, the model can evaluate a longer time series when one limits the sub-group under evaluation. In this case, I limit the analysis to the Black population in the United States. In addition, there is good information from another source on survivorship between the censuses. Another benefit of this alternative methodology is that while demographic analysis is "… oriented towards either period or cohort analysis," this alternative methodology will allow for a simultaneous analysis of both period and cohort dynamics, since the model "… expresses the number of events observed (in this case enumerated population) … in terms of effects of age, cohort and period" (Willekens and Baydar, 1984).

There are two major limitations with using this method. The first is that the overall undercount level cannot be directly assessed and the estimates produced are all initially relative to that unknown quantity. Compared to the demographic analysis method, age-period-cohort modeling suggests relative completeness of several censuses. The second limitation lies in the parameterization step, i.e., how sensitive are the estimates to the constraints chosen. This issue will be discussed in more detail in the paper.

Basic Problem

In general we can state the problem as follows, the number of people of a particular age enumerated in a specific census, N(a, t), depends upon six factors:

1) the extent of migration into and out of the country;
2) the accuracy of age-reporting, or the extent of age heaping or depletion for age a;
3) cohort survivorship to age a, [p(a, t-a)];
4) the number of births t-a years earlier [B(t-a)];
5) the overall completeness of the census taken at time t, which can be viewed as a "period effect"; and
6) the relative completeness with which people age a are enumerated. If this factor is constant across censuses, there is an age-period interaction.

Some of these factors may be ignored if we limit the population analyzed to Blacks. The first factor, net migration, can be ignored because migration across national boundaries was minimal during this period for Blacks in the United States. While data on immigrants by race are not available, immigrants from Africa and the Caribbean (including Haiti and the Dominican Republic) during 1900 to 1950 made up less than three percent of all immigrants who obtained legal permanent residence status.[1] Age-misreporting could be corrected, using the age-adjustment methods initially developed by Zelnik (1959). However, it was found in Condon, et al. (2008), that the results were not sensitive to age-heaping adjustment. The third factor, survivorship, is obtainable from life table data. The other factors can be estimated from the age-period-cohort (APC) analysis of census data.

Basic Form of the APC Model

In the basic form of the APC model presented here, age groups and period intervals are not equal but constant in length (i.e., 5-year age groups).[2] This will allow for more confidence in the age constraints chosen. However, compared to an APC model with equal age groups and period intervals, an APC model with age groups and period intervals that are unequal but constant in length will require one more identification constraint as discussed in Fienberg and Mason (1978), as well as some changes in the mathematics of the model.

For the basic model, let i refer to the 5-year age groups for Blacks, for convenience labeled 1, 2, … I. Let j refer to the decennial censuses, for convenience labeled 1, 2, … J (with the applied version of the model using U.S. censuses 1900 to 1950). Let T_{ij} be the "true" number of people in age i and census j, and let N_{ij} be the number people actually counted. Then the reporting rate

[1] Immigration Statistics Yearbook, Table 2 http://www.dhs.gov/xlibrary/assets/statistics/yearbook/2007/table02.xls.
[2] In Condon and Pullum (1988), the preliminary research using age-period-cohort modeling of underenumeration used 10-year age groups and 10-year period intervals. Fienberg and Mason (1978) refer to this as a case of equally and evenly spaced age groups and period intervals. However, using five-year age groups will require another identification constraint as discussed in Fienberg and Mason (1978), and some changes in the mathematics of the model to accommodate this added identification constraint.

for (i, j) is $R_{ij} = N_{ij}/T_{ij}$. The survivorship ratios are of the form $S_{ij} = T_{ij}/T_{i-1,j-1}$. The arrays N_{ij} (enumerated Black population) and S_{ij} (survivorship, taken from life tables) are known. The problem is to estimate the arrays of T_{ij} and R_{ij}.

In trying to specify a form for the array, R_{ij}, three things must be noted. First, there is an overall undercount/overcount (i.e., coverage) that is impossible to estimate within the model and will be referred to as M. The model produces estimates of the deviation from this overall rate, by ages (rows) and censuses (columns). Second, one can not assume that the model includes cohort (diagonal) effects, because these would be confounded with the cohort sizes, which are to be estimated. Substantively, there is no obvious reason to assume that cohort membership influences the probability of enumeration independently of age, census, and birth cohort size. Third, the form of the multiplicative model is somewhat arbitrary, since one could say that R_{ij} (reporting rate) or $R_{ij} - 1$ (undercount rate) is a product of effects. I have chosen to factor R_{ij}, which means that the various factors would take the value 1.00 in the absence of any undercount.

If we knew the true count for one cell in each diagonal (or cohort), then by using the survivorship factors one could fill in the entire table. For a "basic set" of cells when examining 5-year age groups, we take all those which are in rows 1 and 2 or column 1 (i.e., age group 0-4 years and 5-9 years and the 1900 census).[3]

I use the survivorship array S_{ij} to generate another array called F_{ij}. S_{ij} is obtained as the ratio of two consecutive L values for the period between two censuses (i.e., the entry i=3, for ages 10-14, will be ${}_5L_{10}/{}_5L_0$ from the life table for the 10-year interval between j and j+1. F_{ij} is the proportion of people in the "basic set" of cells who can be expected to survive to cell (i,j). All of the entries in the first two rows or first column of this array will take the value 1.0 when analyzing for 5-year age groups. Otherwise, F_{ij} will be the product of all the survivorship ratios S_{ij} itself. One can therefore use the array F_{ij} and the basic set (T_{ij} with i=1 or j=1) to get the full array T_{ij}:

$$T_{ij} = \begin{cases} T_{i-j+1,1}\, F_{ij} & \text{if } j<1 \\ T_{i,j-i+1}\, F_{ij} & \text{if } j>1 \end{cases}$$

Decomposing the factor R_{ij} results in the following structure of R_{ij} being equal to $M*A_{i.}*B_{.j}$ (where $A_{i.}$ and $B_{.j}$ are factors for row i [age effect] and column j [census or period effect], respectively). M is the overall undercount/overcount (i.e., coverage). Putting aside the constraints that are necessary for estimation, we have the following objective: to estimate the basic set of T_{ij} (I+J-1 numbers), the main effect M (1 number), the row reporting factors $A_{i.}$ (I numbers) and the column reporting factors $B_{.j}$ (J numbers) to fit the observed array N_{ij} (I x J numbers). Note that M is not actually an overall reporting factor, because it is impossible to estimate that quantity with the data at hand. Therefore, from the previous relationships, we have a multiplicative model:

$$N_{ij} = \begin{cases} T_{i-j+1,1} * F_{ij} * M * A_{i.} * B_{.j} & \text{if } j<1 \\ T_{i,j-i+1} * F_{ij} * M * A_{i.} * B_{.j} & \text{if } j>1 \end{cases}$$

[3] In the original research (Condon and Pullum, 1988) which looked at 10-year age groups and census years 1880-1980, the "basic set" of cells took all those which were in row 1 or column 1 (i.e., row 1 was age group 0-9 years and column 1 was census year 1880).

For estimation, we take logs of both sides, (using lower case letters for logs and reordering the terms):

$$n_{ij} = \begin{cases} m + a_{i.} + b_{.j} + t_{i-j+1,1} + f_{ij} & \text{if } j<1 \\ m + a_{i.} + b_{.j} + t_{i,j-i+1} + f_{ij} & \text{if } j>1 \end{cases}$$

or

$$n_{ij} = m + a_{i.} + b_{.j} + c_k + f_{ij}$$

where $k = j-i+1$ and $c_k = t_{i-j+1,1}$ if $j<1$ or $c_k = t_{i,j-i+1}$ if $j>1$. That is, c_k is interpreted as the log of the "basic set" of true frequencies in the first column, moving up the column and then moving to the right along the top row (for 10-year age groups) or top two rows (for 5-year age groups) of the table. This form of the equation has a main effect, row effects, column effects, diagonal effects, all to be estimated, and a known term f_{ij}. The model can be estimated using SAS PROC GENMOD with a Poisson error distribution and offset by f_{ij}. By defaults, those cells with higher population counts exert more weight in the estimations. Weighting each cell according to the inverse of its observed count, in order to counteract the greater emphasis of the maximum likelihood estimation procedures upon the larger cells, could solve this problem. In this first approximation, however, weighting will not be used.

Identification Constraints

To make the set of constraints on the estimated parameters be interpretable, the next step in the analysis is to reparameterize the model. In Condon and Pullum (1988) using 10-year age groups, the initial set of specifications was that the main effect was zero (on the log scale), and the sum of the row effects and sum of the column effects were zero (on the log scale). Specifying that the main effect is zero, the sum of the age effects is zero and the sum of the period effects, allows us to conceptualize our results as deviations from some overall average completeness of enumeration.

After these conditions have been specified, it is still necessary to set two additional constraints for the model to be identified (Feinberg and Mason, 1978). The final identifications specify that two consecutive pairs of age group effects are equal. The choice of the final identification constraints is significant.[4] Forcing two age groups to have equal age effects – if the effects are not in fact equal – may result in the signs of the coefficients being reversed. On the other hand, the model is somewhat robust in that the relative size of coefficients does not fluctuate wildly when the constraints are varied. Thus, I have chosen the following constraints for the planned analysis.

In choosing these last two identification constraints, one would expect – from demographic estimates of absolute undercounts – that the effect of being in a particular age group is different for males and females. Thus, the particular constraints will be chosen more on empirical than substantive groups. However, the robustness of the model will be tested given different identification constraints.

If enumeration were complete, the number in a cohort at age B, relative to the number in the cohort at an earlier age A (F'_{ij}) would equal the probability of cohort survival from age A to age

[4] Alternatively, with some modification to the model, the constraint could have been to set the effect of two periods to be equal.

B in an accurate cohort life table (F_{ij}). Following this logic the observed number in a cohort in a particular census, relative to the observed number in the cohort at its initial appearance in the matrix (F''_{ij}), should be similar to the survivorship probabilities from life tables (F_{ij}). Assuming that the survivorship probabilities from life tables are accurate, deviations of F''_{ij} and F_{ij} point to a particular pattern of error in enumeration. Two age groups with similar patterns of deviation of F''_{ij} from F_{ij} should then show similar patterns of error in enumeration.

REFERENCES

Condon, Katherine M.; Judson, Dean; and Robinson, J. Gregory. 2008. "Historical Coverage Patterns in the United States Census: An Application of Age-Period-Cohort Analysis." Presented at the Southern Demographic Association annual conference Oct 30-Nov 2, 2008, Greenville, SC.

Condon, Katherine M. and Pullum, Thomas. 1988. "Underenumeration in the United States Census 1880-1980: An application of Age-Period-Cohort Analysis." Presented at the annual meeting of the Population Association of America, New Orleans, LA, April 21-23, 1988.

Fienberg, S. E. and Mason, W. M. 1978. "Identification and estimation of age-period-cohort models in the analysis of discrete archival data." In *Sociological Methodology*, K. F. Schuessler (editor). San Francisco, CA: Jossey-Bass.

Glenn, Norval D. 2005. *Cohort Analysis*. 2nd edition. Thousand Oaks, CA: Sage Publications.

Immigration Statistics Yearbook. http://www.dhs.gov/xlibrary/assets/statistics/yearbook/2007/table02.xls

Willekens, Frans and Baydar, Nazli. 1984. "Age-period-cohort models for forecasting fertility." Netherlands Interuniversity Demographic Institute (N.I.D.I.) Working Paper no. 45. Voorburg, the Netherlands.

Zelnik, M. 1959. *Estimates of Annual Births and Birth Rates for the White Population of the United States from 1855 to 1934*. Ph.D. Dissertation. Princeton University, Princeton, NJ.