Imputing Missing Information on Census-unlinked Vital Events: A Simple Bayesian Approach

Isaac Sasson
Department of Social Policy
London School of Economics and Political Science
i.sasson@lse.ac.uk

ABSTRACT

Previous studies have shown that demographic estimates based on census-unlinked vital events are potentially biased—especially when broken down by sociodemographic groups (e.g., when assessing social disparities in mortality rates). The problem is further exacerbated in the presence of missing information on birth and death certificates, because omitting those cases from the numerator (but not the denominator) will necessarily underestimate important demographic quantities. This study shows that standard imputation methods are generally unfit to address the problem and instead develops a novel imputation method using Bayes’ theorem, taking into account all available information from both the numerator (vital events) and denominator (census data). The new method is illustrated using U.S. vital registry data from 1990, where nearly one quarter of death certificates failed to record the decedent’s level of educational attainment. The results confirm that this method outperforms standard approaches to missing data imputation where vital events are census-unlinked.
INTRODUCTION

Previous studies have shown that demographic estimates based on census-unlinked vital events are potentially biased (Shkolnikov et al. 2007), especially when broken down by sociodemographic groups (e.g., when assessing social disparities in mortality rates). The problem is further exacerbated in the presence of missing information on birth and death certificates, because omitting those cases from the numerator (but not the denominator) will necessarily underestimate important demographic quantities. Similarly, maximum likelihood estimation and Bayesian multiple imputation—now considered “state of the art” in missing data imputation (Schafer & Graham 2002)—are less than optimal when dealing with census-unlinked data, because they rely on information from a single data source rather than all available information (i.e., vital events and the at-risk population). Using Bayes’ theorem, this paper develops a simple and straightforward imputation method make optimal use of available data and overcome those difficulties. First, I introduce the method and illustrate it using U.S. vital registry data from 1990, where nearly one quarter of death certificates failed to record the decedent’s level of educational attainment. Finally, I compare the results with conventional data imputation methods.

METHODOLOGY

Estimating social disparities in mortality is often sought after and particularly by educational attainment. This requires a death count in the numerator (taken from the vital registry) and the number of person-years of exposure to the risk of mortality in the denominator (generally estimated using the mid-year population from the census). Although the vital registry system captures practically all deaths occurring in the U.S., information reported on the deceased is often incomplete. For example, in 1990, 23.6 percent of all death certificates were missing information on educational attainment. This is because seven states failed to report altogether whereas the remaining states had an average of 10.0 percent missing information. By 2000, the average level of missing information declined to 4.0 percent across the states, with three states still failing to report.
Since those vital events are unlinked to the census (at-risk) population, omitting them from the numerator when estimating mortality rates will result in significant bias. But how should one assign deaths to the most plausible educational attainment group? While conventional imputation methods tend to focus on the distribution of education in the vital registry (numerator), additional information can in fact be borrowed from the distribution of education in the at-risk population (denominator). Bayes’ rule makes this relationship clear

\[
p(\text{Education}|\text{Death}, X) = p(\text{Death}|\text{Education}, X) \cdot p(\text{Education}|X) \cdot \frac{1}{p(\text{Death}|X)}
\]  

(1)

such that the distribution of education in the vital registry, on the left-hand side of the equation, depends on three terms on the right-hand side: (a) the probability of mortality conditional on educational attainment; (b) the marginal distribution of education in the at-risk population; and (c) the marginal distribution of mortality. This relationship can be further conditioned on \(X\), a vector of covariates including age group, race, gender, and state of occurrence.

Although Equation 1 describes a mathematical identity, it can also be used as an imputation device where each component is estimated separately. We take advantage of the fact that for some quantities population data are practically complete, whereas other quantities involve missing information and have to be estimated. Since \(p(\text{Education}|X)\) depends only on the at-risk population and \(p(\text{Death}|X)\) does not depend on education, both can be derived from fully observed information. On the other hand, \(p(\text{Death}|\text{Education}, X)\) depends on missing data, as per our original problem, and has to be estimated. I estimate this quantity at the national level using states with nearly complete data—less than 10 percent missing—the convention followed by official National Center for Health Statistics publications (National Center for Health Statistics 1993).

Using Equation 1, we can now estimate \(p(\text{Education}|\text{Death}, X)\), the distribution of educational attainment among those who have died (i.e., appear in the vital registry), and impute missing values by randomly drawing from it, conditional on additional covariates (e.g., age, race,
gender, and state of occurrence). The last step is repeated ten times, averaging death counts across repetitions to yield the final estimates.

This imputation approach is preferable to other methods because it maximizes the use of available information from both the numerator and denominator. Essentially, it assumes that the relationship between education and mortality is equivalent among observed and unobserved cases, weighted by the educational composition and overall level of mortality in each age-gender-race subgroup and state of occurrence—a strategy that is particularly useful in states with high proportions missing.

DATA

To illustrate this method, I estimate age-specific mortality rates and life expectancy at age 25, $e_{25}$, for non-Hispanic white Americans by gender and educational attainment. All-cause death counts are taken from the 1990 Multiple Cause of Death (MCD) public use files (Centers for Disease Control and Prevention 2013) and stratified by age, gender, race, and educational attainment. Person-years in the denominator are estimated by the respective midyear census population, using the 5% Integrated Public Use Microdata Sample (Ruggles et al. 2010).

In both the numerator and denominator data sources, age was recoded to 5-year groups starting at 25-29 and ending with an open interval at 90+. Race was categorized as non-Hispanic white and black, excluding other race categories and persons of Hispanic origin due to small death counts or poor reporting. Educational attainment in the MCD (numerator) data is classified in single years ranging from zero to five or more years of college. Next, I recoded educational attainment in the census (denominator) data to match the MCD classification of completed years of schooling (0-11; 12; 13-15; 16+). All categories below 12 years were collapsed and recoded as 0-11. Those with more than one year of college education or an Associate’s degree were classified as 13-15, and those with a Bachelor’s degree or higher placed in the 16+ category. Finally, all those who reported
completing the 12th grade (with or without diploma), obtaining a general equivalency degree (GED), or completing “some college credit, but less than one year” were coded as 12 years.

RESULTS

Table 1 shows the number of deaths in each educational attainment group by gender, following the imputation method developed in the previous section. Table 2 shows the estimated life expectancy at age 25 for each of those groups using the new imputation method (Column A) and two benchmark methods. Column B provides a conservative estimate, whereby missing educational attainment is randomly assigned based solely on the distribution of education in the census population (denominator). In other words, it assumes that the risk of mortality is equal across educational attainment groups. Column C, on the other hand, uses multiple imputation based solely on information available in the vital registry (numerator).

The results suggest that estimates vary significantly according to the missing data imputation method used. Most importantly, estimates based on the new methodology generally conform with the well-known education-mortality gradient, whereas the other benchmark methods do not. Specifically, conservative imputation tends to overestimate life expectancy for the least educated and underestimate for the college educated relative to other imputation methods. Although multiple imputation yields results more similar to the method developed in this paper, it performs less well in states failing to report altogether (results not shown).

CONCLUSION

This paper develops a simple imputation method for census-unlinked data based on Bayes’ theorem, with the advantage of using all available information from both numerator (vital events) and denominator (census population), for which data are complete. By weighting those two data sources, it provides a viable alternative for current imputation methods—especially for subnational units with little to no information at all.
REFERENCES


Table 1: Death counts by gender and educational attainment, U.S. 1990

<table>
<thead>
<tr>
<th>Education (years)</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-11</td>
<td>338,760</td>
<td>342,673</td>
</tr>
<tr>
<td>12</td>
<td>338,761</td>
<td>313,921</td>
</tr>
<tr>
<td>13-15</td>
<td>97,574</td>
<td>98,480</td>
</tr>
<tr>
<td>16+</td>
<td>78,695</td>
<td>116,707</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>853,790</strong></td>
<td><strong>871,781</strong></td>
</tr>
</tbody>
</table>
Table 2: $e_{25}$ by educational attainment, non-Hispanic white Americans 1990

<table>
<thead>
<tr>
<th>Education (years)</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>0-11</td>
<td>46.0</td>
<td>46.7</td>
</tr>
<tr>
<td>12</td>
<td>48.7</td>
<td>47.1</td>
</tr>
<tr>
<td>13-15</td>
<td>49.7</td>
<td>52.8</td>
</tr>
<tr>
<td>16+</td>
<td>52.1</td>
<td>51.6</td>
</tr>
<tr>
<td>Total</td>
<td>49.3</td>
<td>49.3</td>
</tr>
</tbody>
</table>

Imputation methods:
A. New method (uses both numerator and denominator)
B. Conservative estimate (uses denominator only)
C. Multiple imputation (uses numerator only)