Combining Forward and Backward Estimates of Mortality

Dan A. Black, University of Chicago, NORC, and IZA
Yu-Chieh Hsu, University of Chicago and NORC
Seth G. Sanders, Duke University and NORC
Lowell J. Taylor, Carnegie Mellon University, NORC, and IZA

January 21, 2014

Abstract

Demographers often form estimates by combining information from two data sources—a challenging problem when one or both data sources are incomplete. A classic example entails the construction of mortality rates, which requires death counts for sub-populations under study and corresponding base population estimates. Approaches typically entail "back projection," as in Wrigley and Schofield's (1981) seminal analysis of historical English data, or "inverse" or "forward projection" as used by Lee (1985) in his important reanalysis of that work. Our paper generalizes and shows how forward and backward approaches can be optimally combined, using a generalized method of moments (GMM) framework. We apply the method to the estimation of mortality rates for relatively small sub-populations within the U.S. (in particular, birth state by birth cohort by race by gender cells), combining data from Vital Statistics records and Census samples.

Keywords: Mortality, forward estimation, backward estimation

We gratefully acknowledge support from Eunice Kennedy Shriver National Institute of Child Health and Human Development (RO1 HD062747). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development or the National Institutes of Health. We also wish to thank the Population Center at the University of Chicago for support. We appreciate comments from seminar participants at the University of Chicago, the Population Association of America, Princeton University, and the University of California at Berkeley. We particularly thank Donald Bogue, Ronald Lee, and Shiro Horiuchi for insightful comments on an earlier draft.

1 Introduction

Inference in demography often entails piecing together data from multiple, often incomplete, sources. A classic example is the estimation of mortality rates, which requires an estimate of deaths for the population under study *and* a corresponding estimate of the population at risk. Wrigley and Schofield's (1981) classic work, for instance, draws on entries of baptisms, marriages, and burials in Anglican parish registers to construct "back projection" estimates of mortality rates for the English population in the 16th through 19th century. Using those same data, Lee (1985) alternatively constructs mortality estimates using "inverse" or "forward projection." Variants of these basic approaches have been used in many studies. Often the required data are available from statistical agencies, as in the empirical work below, in which we use death records from the Vital Statistics registry, and estimate the population at risk using Census samples.

Our innovation is to show how researchers can improve inferences about mortality by optimally combining forward and backward methods using Vital Statistics and Census data. We illustrate the use of our method by providing estimates of relatively small sub-populations of the U.S.—cells constructed at the cohort × birth state × race × gender level.

To set the key idea, suppose we are interested in estimating mortality for a particular group i—for instance, black men born in Georgia in 1932—over a particular time period, e.g., 1990 through 2000. There are three common demographic methods to do this, one using Census data only and two which combine Census and Vital Statistics data.

In setting up the Census-only approach, let N_i^{90} be a count of individuals in group i in 1990 and N_i^{00} the corresponding count in 2000. Then, assuming there is no net migration,

$$ {}_{10}q_i^{\mathrm{C}} = \frac{N_i^{90} - N_i^{00}}{N_i^{90}} \tag{1} $$

gives the 10-year mortality rate. This method works well if complete counts of group i are available. For example, the 100% Summary Tape Files from the U.S. Census for 1990 and 2000 provide counts of the number of men and women by single year of age by race for the U.S. as a whole. Birth state (an additional element required for our analysis), however, was collected only on the "long form," giving a 1-in-6 sample, and researchers will typically have access only through Public Use Microsamples (PUMS) to 1-in-20 or 1-in-100 samples, depending on the Census year. When this is the case, researchers must use *estimates* of N_i^{90} and N_i^{00} in (1), leading to the estimator

$$ {}_{10}\hat{q}_i^{\mathrm{C}} = \frac{\hat{N}_i^{90} - \hat{N}_i^{00}}{\hat{N}_i^{90}}, \tag{2} $$

where S_i^{90} is the sample counts of group i in 1990, $\hat{N}_i^{90} = \omega^{90} S_i^{90}$ has $\omega^{90} = 20$ given the sampling rate of 1-in-20 in the 1990 PUMS, and, similarly, $\hat{N}_i^{00} = \omega^{00} S_i^{00}$ with $\omega^{00} = 20$.[1]

[1] In fact, inflation factors differ in both the 1990 and 2000 Census samples, so our estimator of $\hat{N}_i^{90}$ is slightly more complicated. For the remainder of the paper we restrict attention to a constant sampling rate in a Census year, which greatly reduces notation with little loss in insight. When we turn to our empirical example that has individual-level inflation factors (weights), we use these in forming estimates.

The Census-only approach is valuable in developing countries where Censuses are largely complete but Vital Statistics registries are not (or in similar historical circumstances). Lleras-Muney's (2005) innovative work uses this estimator in the U.S. to calculate cohort-specific mortality by state of birth for the purpose of inferring the impact of state education policies on late-life mortality.

This Census-only method has two clear disadvantages. First, samples used to estimate $\hat{N}_i^{90}$ and $\hat{N}_i^{00}$ are often quite small, which can lead to imprecise estimation of ${}_{10}\hat{q}_i^{\text{C}}$.[2] Second, analysts often require annual mortality rates, rather than mortality over a 10-year period. This method cannot produce such estimates.

The second commonly used method combines Census and Vital Statistics data using "forward projection." To estimate the 10-year mortality rate using this method, the researcher counts the number of deaths of group i between 1990 and 2000 and divides by the corresponding 1990 estimate of the group i population. Thus the forward estimator is

$${}_{10}\hat{q}_i^{\text{F}} = \frac{{}_{10}D_i^{90}}{\hat{N}_i^{90}}, \tag{3}$$

where ${}_{10}D_i^{90} = \sum_{t=90}^{99} D_i^t$.[3]

The final method, "back projection," counts the number of deaths of group i between 1990 and 2000 but divides by an alternative estimate of 1990 group i population—the estimated 2000 population plus the number of deaths occurring between 1990 and 2000. Thus the backward estimator is

$${}_{10}\hat{q}_i^{\text{B}} = \frac{{}_{10}D_i^{90}}{(\hat{N}_i^{00} + {}_{10}D_i^{90})}. \tag{4}$$

Either the forward or backward projection method can be used also to estimate *annual* mortality rates. For example, the mortality rate from the forward estimator for group i in 1991 is

$$D_i^{91}/(\hat{N}_i^{90} - D_i^{90})$$

while the backward estimator is

$$D_i^{91}/(\hat{N}_i^{00} + {}_{9}D_i^{91}).$$

An important use of the *difference* between the forward and backward projection estimates is to assess the quality of Census or Vital Statistics data. For example, Palloni and Kominski (1984) use the difference in the two estimators to assess the incompleteness of Vital Statistics data in several Latin American countries. An under-appreciated possibility is that even when data quality is high, ${}_{10}\hat{q}_i^{\text{F}}$ and ${}_{10}\hat{q}_i^{\text{B}}$ will differ because of sampling variation. Such differences become larger the smaller is the size of subpopulation i—clearly a concern in our

[2] Indeed, using this method Lleras-Muney (2005) reports a non-negligible proportion of mortality estimates for birth state × cohort cells imply *negative* mortality.

[3] As a practical matter, given the timing of the Census, D_i^t is the count of number of deaths of group i between April 1, year t and March 31, year $t+1$. ${}_{10}D_i^{90}$ is thus the number of deaths of group i between April 1, 1990 and March 31, 2000.

example, as we can expect to encounter small sub-populations when we estimate mortality for such groups as black men born in Georgia in 1932.

Without a formal statistical theory it is not clear what researchers should do with two consistent but differing estimates of group i mortality. A naive approach would be to take a simple average of the two. Below we provide theoretical arguments that show why this approach can work well for some applications (and Hsu (2012) provides empirical examples in which this approach does indeed work well). More importantly, we show how this naive estimator is related to an efficient generalized method of moments (GMM) estimator. We do this to clarify why equal weighting is sometimes effective, and also to highlight circumstances under which a researcher would want to weigh one of the two estimators more heavily. Our GMM estimator produces optimal weights, and these weights turn out to have intuitively appealing properties.

In Section 2 of our paper we show that the naive estimator is actually the solution to a minimum distance (MD) estimator. The MD estimator then generalizes to a two-step estimator—a generalized method of moments (GMM) estimator first proposed in the seminal work of Hansen (1982). For interest sake, we also formulate a constrained maximum likelihood (ML) estimator for the problem at hand and demonstrate a close relationship between the GMM and ML approaches in the Appendix.

In Section 3 of our paper we turn to an illustration: racial differences of the mortality of men born in the Great Depression by state of birth in later life. We show that conditioning on the state of birth greatly improves the accuracy of the regression model regardless of the estimator used. We find strong evidence that black-white mortality differences increased, results driven by the improvement in Southern white mortality. Indeed, we cannot reject the hypothesis that for blacks and whites born in the North, there was no increase in black-white mortality gaps.

In Section 4 we provide concluding remarks.

2 Estimating Mortality Using Two Data Sources

Our problem is conceptually quite simple. Suppose that in an initial period, designated period 0, we have a Census dataset that randomly samples a population of N^0 individuals using a known sampling rate, say 1 in ω^0.[4] In period 1 we similarly have a Census that samples at a rate of 1 in ω^1. Estimates of the population sizes for subset i in period 0 and 1 are just

$$\hat{N}_i^0 = \omega^0 S_i^0 \quad \text{and} \quad \hat{N}_i^1 = \omega^1 S_i^1. \tag{5}$$

[4] We denote the first Census as 0 and the second as 1 to make the discussion general and to save on notation. As noted above, period 0 is 1990 and period 1 is 2000. Again, for ease of exposition, we assume that sampling weights do not vary within period, though we do use Census sampling weights in our estimation below.

We have already defined the Census-only estimate of mortality rate for group i,

$$ {}_T\hat{q}_i^{\mathrm{C}} = \frac{(\hat{N}_i^0 - \hat{N}_i^1)}{\hat{N}_i^0}, \tag{6} $$

where T is the length of time between period 0 and period 1. As we suggest in the introduction, estimator (6) is likely to be quite noisy when samples are small. We can do much better if we replace the numerator in (6) with the proper death count from Vital Statistics over the time T, which we denote ${}_TD_i$. As for the denominator of (6), we could use $\hat{N}_i^0$ estimated with period 0 Census data, in which case we have a forward estimator. Alternatively, we can exploit a direct relationship, $N_i^0 = N_i^1 + {}_TD_i$,[5] and form a backward estimator with the denominator $\hat{N}_i^1 + {}_TD_i$ instead of $\hat{N}_i^0$. Either approach to estimating the denominator, N_i^0, is likely to be noisy; intuitively, one would like to use both pieces of information in forming inferences.

Our problem, then, is to combine the data to find the *best* estimate of the number of individuals of type i in time 0 for use in the denominator of our estimator, i.e., the consistent estimator that minimizes asymptotic variance. We start with a simple, intuitively sensible *minimum distance estimator.*

2.1 A Minimum Distance Estimator

In constructing our estimate of the size of the group i population, N_i^0, we use the relationships

$$ \begin{aligned} \mathrm{E}\left\{\omega^0 S_i^0 - N_i^0\right\} &= 0, \\ \mathrm{E}\{\omega^1 S_i^1 + {}_TD_i - N_i^0\} &= 0. \end{aligned} \tag{7} $$

The expressions in (7), which involve expectations, are often called *moment restrictions.*[6] Given that our goal is to find estimators that fit equations (7) "well," an intuitively attractive idea is to find value $\hat{N}_i^{0,\mathrm{MD}}$ that minimizes the expression,

$$ \begin{bmatrix} N_i^0 - \omega^0 S_i^0 & N_i^0 - \omega^1 S_i^1 - {}_TD_i \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} N_i^0 - \omega^0 S_i^0 \\ N_i^0 - \omega^1 S_i^1 - {}_TD_i \end{bmatrix}. \tag{8} $$

To find this *minimum distance estimator,* we simply solve the problem,

$$ \min_{N_i^0} V = \left(N_i^0 - \omega^0 S_i^0\right)^2 + \left(N_i^0 - \omega^1 S_i^1 - {}_TD_i\right)^2. \tag{9} $$

[5] In general, $N_i^0 = N_i^1 + {}_TD_i + {}_TE_i$ where when i is defined by state of birth the net emigration of the group ${}_TE_i \approx 0$; if a black born in Georgia migrates to New York he appears in New York but his state of birth is still recorded as Georgia. It is only net emigration out the U.S. entirely that creates an issue for our estimation below, which focuses on birth state. The problem is considerably more difficult if groups were defined by state of *residence.* In this case it would be necessary to estimate state-to-state net migration of group i to calculate our base estimates.

[6] Our restrictions assume that ${}_TD_i$, the death counts for individuals in group i, have been accurately recorded in Vital Statistics records. If this number is thought to be recorded with error, and the error process can be modeled, we would instead have three moment restrictions.

V is a strictly convex function of N_i^0 that has a first-order condition,

$$\frac{dV}{dN_i^0} = 2\left(\hat{N}_i^{0,\mathrm{MD}} - \omega^0 S_i^0\right) + 2\left(\hat{N}_i^{0,\mathrm{MD}} - \omega^1 S_i^1 - {}_T D_i\right) = 0, \tag{10}$$

which leads to the resulting estimator,

$$\hat{N}_i^{0,\mathrm{MD}} = \frac{1}{2}\left(\omega^0 S_i^0\right) + \frac{1}{2}\left(\omega^1 S_i^1 + {}_T D_i\right). \tag{11}$$

The minimum distance estimator is simply the average of the two potential Census estimators proposed above. Because the samples are approximately independent (only about 0.0025 of the population will appear in two consecutive 1-in-20 PUMS), we stand to gain a great deal of efficiency by using *two* samples to construct our estimate of N_i^0.

With our estimator of N_i^0 in place we can easily construct our mortality rate estimate. Let ${}_T q_i$ be the mortality *rate* for group i between time 0 and time 1. Our estimator of this object, based on the minimum distance (MD) approach, is simply

$${}_T q_i^{\mathrm{MD}} = \frac{{}_T D_i}{\hat{N}_i^{0,\mathrm{MD}}}, \tag{12}$$

i.e., the ratio of the observed deaths to our minimum distance estimate of the number of people in group i who were alive at time 0.

This clearly is a consistent estimator, and it has the advantage of using all available data in a simple and coherent way. The estimator is easy to implement, e.g., with simple commands in any statistical package or spreadsheet program. An important paper by Hansen (1982), though, establishes a generalization of the MD estimator that has optimal properties, in terms of minimizing the estimator's asymptotic variance. We turn to that estimator next.

2.2 A GMM Estimator

The idea of Hansen's *generalized method of moments* (GMM) estimator is to undertake a minimization exercise, such as the one given in (8), but in which the matrix in the interior of (8) is *not* the identity matrix, but rather a 2 × 2 symmetric matrix, W^{-1}, the inverse of the covariance matrix from the vector of "moment restrictions," which in our case is

$$\begin{aligned} W &= \mathrm{E}\left\{\begin{bmatrix} N_i^0 - \omega^0 S_i^0 \\ N_i^0 - \omega^1 S_i^1 - {}_T D_i \end{bmatrix}\begin{bmatrix} N_i^0 - \omega^0 S_i^0 & N_i^0 - \omega^1 S_i^1 - {}_T D_i \end{bmatrix}\right\} \\ &= \begin{bmatrix} \left(\omega^0\right)^2 S^0 p_i^0(1-p_i^0) & 0 \\ 0 & \left(\omega^1\right)^2 S^1 p_i^1(1-p_i^1) \end{bmatrix}, \end{aligned} \tag{13}$$

where p_i^0 and p_i^1 are, respectively, the probability in period 0 that an observation from the complete sample S^0 is a member of group i, and the analogous probability in period 1.[7] The

[7] Put another way, $p_i^0 = \frac{N_i^0}{N^0}$ and $p_i^1 = \frac{N_i^1}{N^1}$. We of course do not directly observe p_i^0 or p_i^1, since N_i^0 and N_i^1 are unknown.

terms in W are easy to find as our particular problem entails draws from two independent binomial processes.[8] Hansen proves that the use of W^{-1} is optimal in terms of minimizing the asymptotic variance of the estimator.

As we do not know the values of $[p_i^0,\ p_i^1]$ in advance, we cannot simply substitute W^{-1} for the 2×2 identity matrix in equation (8), and proceed with the minimization problem. Instead, Hansen (1982) suggests a two-step estimator. The first step is the simple minimum distance estimation given in Section 2.1 above. The idea is to use the estimator (11) to consistently estimate $[p_i^0,\ p_i^1]$, and to use *those* to estimate the covariance matrix. Thus we form $\hat{W}^{-1}$ using equation (13), but replacing each p_i^0 and p_i^1 with our estimates, $\hat{p}_i^0$ and $\hat{p}_i^1$. The second step then entails finding the value $\hat{N}_i^{0,\text{GMM}}$ that minimizes

$$\left[\begin{array}{cc} \hat{N}_i^{0,\text{GMM}} - \omega^0 S_i^0 & \hat{N}_i^{0,\text{GMM}} - \omega^1 S_i^1 - {}_T D_i \end{array}\right] \left[\begin{array}{cc} (\omega^0)^2 S^0 \hat{p}_i^0 (1-\hat{p}_i^0) & 0 \\ 0 & (\omega^1)^2 S^1 \hat{p}_i^1 (1-\hat{p}_i^1) \end{array}\right]^{-1} \times \left[\begin{array}{c} \hat{N}_i^{0,\text{GMM}} - \omega^0 S_i^0 \\ \hat{N}_i^{0,\text{GMM}} - \omega^1 S_i^1 - {}_T D_i \end{array}\right], \tag{14}$$

which yields the necessary condition,

$$\frac{\hat{N}_i^{0,\text{GMM}} - \omega^0 S_i^0}{(\omega^0)^2 S^0 \hat{p}_i^0 (1-\hat{p}_i^0)} + \frac{\hat{N}_i^{0,\text{GMM}} - \omega^1 S_i^1 - {}_T D_i}{(\omega^1)^2 S^1 \hat{p}_i^1 (1-\hat{p}_i^1)} = 0. \tag{15}$$

Following a series of algebraic steps we can show that the resulting estimator is

$$\begin{aligned} \hat{N}_i^{0,\text{GMM}} = & \left[\frac{\left((\omega^0)^2 S^0 \hat{p}_i^0 (1-\hat{p}_i^0)\right)^{-1}}{\left((\omega^0)^2 S^0 \hat{p}_i^0 (1-\hat{p}_i^0)\right)^{-1} + \left((\omega^1)^2 S^1 \hat{p}_i^1 (1-\hat{p}_i^1)\right)^{-1}}\right] \omega^0 S_i^0 \\ & + \left[\frac{\left((\omega^1)^2 S^1 \hat{p}_i^1 (1-\hat{p}_i^1)\right)^{-1}}{\left((\omega^0)^2 S^0 \hat{p}_i^0 (1-\hat{p}_i^0)\right)^{-1} + \left((\omega^1)^2 S^1 \hat{p}_i^1 (1-\hat{p}_i^1)\right)^{-1}}\right] \left(\omega^1 S_i^1 + {}_T D_i\right). \end{aligned} \tag{16}$$

Notice that as in (11), we are using a weighted sum of two consistent estimators of N_i^0 for our estimator, but in the GMM case we use asymptotically *optimal* weights, which include objects that are estimated in the first stage of the estimation procedure.

Finally, having found the GMM estimate of N_i^0, our estimate of the mortality rate for group i from time 0 to time 1, based on the GMM approach, is

$${}_T q_i^{\text{GMM}} = \frac{{}_T D_i}{\hat{N}_i^{0,\text{GMM}}}. \tag{17}$$

To build intuition for this estimator, consider the case in which the inflation weights are

[8] Conceptually, the Census finds the entire population, and samples a fraction of these individuals for public use releases. Then, for example, in period 0 each of these individuals has a p_i^0 probability of belonging to group i and a $1-p_i^0$ probability of being in some other group. Estimates of the first moment have variance $S^0 p_i^0 (1-p_i^0)$.

the same in the two consecutive Census periods, $\omega^0 = \omega^1 = \omega$. Then (16) reduces to

$$\begin{aligned}\hat{N}_i^{0,\mathrm{GMM}} &= \left[\frac{\left(\hat{p}_i^0(1-\hat{p}_i^0)\right)^{-1}}{\left(\hat{p}_i^0(1-\hat{p}_i^0)\right)^{-1} + \left(\hat{p}_i^1(1-\hat{p}_i^1)\right)^{-1}}\right] \omega S_i^0 \\ &+ \left[\frac{\left(\hat{p}_i^1(1-\hat{p}_i^1)\right)^{-1}}{\left(\hat{p}_i^0(1-\hat{p}_i^0)\right)^{-1} + \left(\hat{p}_i^1(1-\hat{p}_i^1)\right)^{-1}}\right] \left(\omega S_i^1 + {}_TD_i\right). \end{aligned} \tag{18}$$

Given this expression, consider two cases:

First suppose $\hat{p}_i^0 \approx \hat{p}_i^1$. This approximation applies when mortality is very low for group i. In this case, the weights (in brackets) are approximately $\frac{1}{2}$; the two estimates of N_i^0 are given roughly equal weight. That is

$$\hat{N}_i^{0,\mathrm{GMM}} \approx \frac{1}{2}\left[\omega S_i^0 + \left(\omega S_i^1 + {}_TD_i\right)\right]. \tag{19}$$

It is easy to show how this relates the forward and backward mortality estimators. If (19) were to hold exactly,

$$\begin{aligned}{}_Tq_i^{\mathrm{GMM}} &= \frac{{}_TD_i}{\hat{N}_i^{0,\mathrm{GMM}}} = \frac{{}_TD_i}{\frac{1}{2}\left(\omega S_i^0 + \omega S_i^1 + {}_TD_i\right)} \\ &= \left[\frac{\frac{1}{2}\omega S_i^0}{\frac{1}{2}\left(\omega S_i^0 + \omega S_i^1 + {}_TD_i\right)}\right]\frac{{}_TD_i}{\omega S_i^0} + \left[\frac{\frac{1}{2}\left(\omega S_i^1 + {}_TD_i\right)}{\frac{1}{2}\left(\omega S_i^0 + \omega S_i^1 + {}_TD_i\right)}\right]\frac{{}_TD_i}{\omega S_i^1 + {}_TD_i} \\ &= \left[\frac{\omega S_i^0}{\omega S_i^0 + \omega S_i^1 + {}_TD_i}\right]{}_T\hat{q}_i^{\mathrm{F}} + \left[\frac{\omega S_i^1 + {}_TD_i}{\omega S_i^0 + \omega S_i^1 + {}_TD_i}\right]{}_T\hat{q}_i^{\mathrm{B}}.\end{aligned}$$

As mortality is low in our example, ${}_TD_i$ is small relative to N_i^0, and in turn $\omega S_i^0 \approx \omega S_i^1$. So in this case our GMM estimate is very close to

$${}_Tq_i^{\mathrm{GMM}} \approx \frac{\left({}_Tq_i^{\mathrm{F}} + {}_Tq_i^{\mathrm{B}}\right)}{2},$$

i.e., the simple average of the forward and backward estimators.

Next, consider the opposite case, in which $\hat{p}_i^1$ has declined nearly to 0. This happens when the cohort has nearly become extinct, which would be most common at very old ages. In this case, careful inspection of (18) shows that the GMM estimator places a weight slightly less than 1 on the second term, and a weight slightly greater than 0 on the first term.

In short, we have a somewhat counterintuitive result: When the Census uses the same sampling scheme in two consecutive periods, and estimators are formed using complete death counts (from Vital Statistics), the weight given the first Census sample (period 0) will never be greater than $\frac{1}{2}$, even though the Census count for the sub-population in period 0 is larger than in period 1. Moreover, when the sub-population is substantially smaller in period 1 than in period 0, as is typical at older ages, the weight given to the period 1 Census *increases.*

The proper intuition for the result comes from focusing on the extreme case, in which *all* individuals in a cohort have died. In this important case our estimator converges to "extinct generation estimation"—a methodology used in many important papers, e.g., Elo and Preston (1994).[9] Given our assumption that death counts in the Vital Statistics are accurate, for a cohort that is *extinct* in period 1 we can form a perfect estimate the number of people who were alive in period 0 simply by counting recorded deaths between the two periods. In this case the imperfect information from the period 0 Census can be entirely disregarded. With this intuition in place, now notice that as a cohort near extinction, S_i^1 approaches 0 (and will typically be much smaller than ${}_TD_i$ for typical sampling rates), so the GMM procedure effectively places progressively higher weight on the death counts, relative to Census samples, as a means of determining the base with which to estimate mortality (in (17)).

As we have noted, use of the Vital Statistics data allows researchers to estimate annual mortality rates. There is, of course, no Census estimate of the population in intercensal years, but we can construct an estimate using this simple difference equation:

$$\widehat{N}_i^{t+1} = \widehat{N}_i^t - D_i^t \tag{20}$$

where $\widehat{N}_i^t$ is a population estimate obtained from the GMM, backward, or forward estimate of the population. This difference equation does imply that errors in measurement will be correlated across birth state by cohort cells so standard errors should be clustered to account for such correlation.

3 Application

Our application entails the estimation and analysis of men's mortality in mid-life—ages approximately 51 to 70—by sex, race, and birth state, for people born during the 1930s. To put this work in context, we mention two important strands of literature.

First, a vast literature focuses on black-white disparities in health outcomes—including mortality—in the twentieth century. Measured in terms of life expectancy, racial disparity has decreased over the century, but remains high. According to recent life tables produced at the Division of Vital Statistics (Arias, 2010), the gap in life expectancy at birth between whites and blacks born in the U.S. declined from 10.4 years for cohorts born 1919-1921 (with life expectancies of 57.4 for whites and 47.0 for blacks) to a historic low of 5.0 for the cohort born in 2006 (78.2 for whites and 73.2 for blacks).[10]

There are many proximate medical causes for the mortality gap, including black-white disadvantages in mortality due to diseases of the heart, cancer, cerebrovascular disease, diabetes mellitus, and pneumonia and influenza (e.g., Levine, et al., 2001). Importantly, for our purposes, the incidence of life-threatening disease (and other threats, such as violence) varies substantially across local areas in the U.S. For example, in a seminal paper, McCord and Freeman (1990) estimated the rate of survival beyond the age of 40 for black men in

[9] Elo and Preston provide reference to Vincent's (1951) seminal use of this method.

[10] These estimates are from period life tables, which calculate life expectancy for a hypothetical cohort that experiences current rates of age-specific mortality throughout its lifetime.

Harlem, circa 1960-1980, to be lower than for men in Bangladesh. Geronimus, Bound, and Colen (2011) provide more recent location-specific statistics, by race, for a geographically diverse set of locations, and similarly demonstrate high variation in mortality rates, and in black-white differences in mortality rates, across locations.[11]

A second important literature focuses on the "long reach" of health threats in early childhood and *in utero* (Barker, 1990 and 1995), particularly conditions of nutritional deficiencies during these crucial periods of human physical development. This idea plays an important role, for example, in Fogel's (2004) analysis of the long-run decline in mortality, and is analyzed in a great many important studies. More generally, deprivation in childhood can lead to poor health outcomes later in life via a number of potential behavioral mechanisms related to the intergenerational transmission of socio-economic wellbeing.

Some of the research on the role of early-childhood circumstances on later-life mortality focuses specifically on the African American population. For instance, even using a relatively small sample of 582 older African Americans, Preston, Hill, and Drevenstedt (1998) were able to show that children who were exposed to the most unhealthy childhood environments were less likely to reach age 85 than those living in more favorable environments. In their study, mortality risks at young ages and mortality risks at older ages are shown to be positively correlated for this population, suggesting that assaults on health early in life adversely affect mortality at all subsequent ages for the population. Similarly, Hayward and Gorman (2004) study associations between childhood socioeconomic conditions and men's mortality, and Warner and Hayward (2006) assess the extent to which childhood and adulthood conditions account for the racial gap in men's mortality.[12]

Against this backdrop, there is clear value in being able to evaluate variation in later-life health outcomes conditional on one's location of birth. There is a small literature on this topic. Fang, *et al.* (1996), for example, explore the high rate of mortality from cardiovascular causes among blacks in New York City, finding that there is substantial variation among blacks based on their place of birth. In particular, Southern-born blacks had higher rates of mortality from cardiovascular disease than those of their Northeastern-born counterparts. Greenberg and Schneider (1992), as another example, examine black mortality by place of birth and residence. That paper suggests that blacks who migrated from the South had higher mortality rates than blacks born in other regions in the United States.

In short, there are good reasons to believe that mortality in adulthood might vary by state of birth in interesting and important ways. As we have mentioned, our focus is on black and white individuals born during the 1930s. We then assess annual mortality rates for these individuals for 1990 through 2000. Thus we are looking at mortality in the mid-life (at ages 51 through 70) for men born in 15 states: the nine Southern states with the largest African American populations—Alabama, Arkansas, Georgia, Louisiana, Mississippi,

[11] A major challenge in this literature is its difficulty in sorting out the extent to which bad environmental factors within high-mortality locations cause poor health, or conversely, people who have better resources and better health avoid such neighborhoods.

[12] See also work by Costa, *et al.* (2007), showing that black men in the early twentieth century have higher incidence of infectious disease, leading them to have higher prevalence rates of chronic conditions, such as arteriosclerosis, at older ages.

North Carolina, South Carolina, Tennessee, and Virginia—and the Northern industrial states of Illinois, Indiana, New Jersey, New York, Ohio, and Pennsylvania.[13] Because of legacy of slavery, the Southern states have relatively large African Americans born within their borders, but even in the 1930s the number of Northern-born African Americans can be modest. Hence, we focus on six large industrial states, which did have reasonably large numbers of black births.

To see how the three estimators compare, in Panel A of Table 1 we compare the log difference in black and white mortality using the GMM, forward, and backward estimates of mortality. Despite using the same numerator (Vital Statistic's counts of deaths by state-of-birth and age), the backward and forward estimators are correlated only at 0.72. In comparing these two estimators to the GMM estimator, we see that, as the theory predicts, the backward estimator is more highly correlated with the GMM estimator than the forward estimator.

In Panel B of Table 1, we provide summary statistics for black mortality rate, white mortality rates, and differences in black-white mortality rates by age for men born in our 15 states. The black mortality rate exceeds the white mortality rate substantially at each age, and the differences are remarkably stable, starting out at about 0.8 percent at early ages and peaking at about 1.4 percent at age 67. As we shall see, however, these aggregated numbers hide a great deal of heterogeneity.

3.1 Mortality Estimates by Birth State for Men Born 1930–1939: A Comparison of Three Estimators

Our goal in this section of empirical results is to compare estimates of mortality using the GMM, forward and backward estimators mentioned in our methodological section for a substantive problem: Age-specific black-white mortality gaps.

Given that we are estimating mortality in state-of-birth×race×age×birth-cohort cells, in many cases we are estimating mortality on the basis of relatively small samples. We wish to compare black-white differences in mortality and see whether that differences increased or decreased for the cohorts born in the 1930s. We begin by comparing the log differences in mortality rates for black men ($\ln(q_{bas})$) and white men ($\ln(q_{was})$), and for the following regression:

$$\ln(q_{bas}) - \ln(q_{was}) = \alpha_a + \beta_s + \tau(\text{Cohort}) + \tau_n(\text{Cohort} \times \text{North}) + \epsilon_{cs}, \qquad (21)$$

where a indicates age ($a = 51, \ldots, 70$), s indicates state, "Cohort" is an index of birth cohorts (0 indicates 1930, 1 indicates 1931, *et cetera*), "North" is an indicator for being one of the six Northern states, and c indicates birth cohort ($c = 1930, \ldots, 1939$). Thus, we compare

[13] Of course our methodology could be applied to the study of mortality at younger ages and at older ages as well—both of which are interesting. One reason we do not study mortality at younger ages for our cohorts of study is a lack of consistently reported data on state of birth in available death records prior to 1978. Researchers who look at black-white mortality at older ages would do well to consider age reporting issues raised by Preston, *et al.* (1999).

trends in black-white mortality differences in 15 states—the 9 Southern states where the most African Americans were born in the 1930s, as well as 6 large comparison states from the North.[14] Because we expect errors to be correlated within a state and cohort, we cluster our standard errors within the state and cohort, resulting in 150 clusters (10 cohorts by 15 states).

In Panel A of Table 2, we estimate equation (21) while constraining the state-of-birth coefficients to be zero ($\beta_s = 0$) for equations in which we use, respectively, the GMM estimator for the dependent variable, the forward estimator for the dependent variable, and the backward estimator for the dependent variable. Notice that across regressions only the measure of the dependent variable differs. In Panel B, we provide corresponding regressions that allow for state-of-birth fixed effects. Several features of the results warrant mention:

First, both the forward estimator and the backward estimator produce substantially worse fits, as measured by the R^2, than the comparable GMM estimator. Also, not surprisingly, the standard errors of the coefficient estimates when using the forward and backward estimators are substantially higher than when using the GMM estimator. For instance, in Panel B the standard error on the *trend coefficient* when using the forward estimator is approximately 1.5 times the size as the corresponding standard error when using GMM. Similarly, the standard error on the *interaction coefficient* when using the forward estimator is 1.4 times that of the corresponding GMM standard error.

Second, we have a potentially important substantive observation: the inclusion of state-of-birth fixed effects substantially increases the R^2 in all models. The F-test for inclusion of state fixed effects rejects the hypothesis that $\beta_s = 0$ at a significance level in excess of 0.0001.

Finally, inclusion of state of birth affects inferences about the black-white mortality gap. Estimates from Panel A, with no state of birth indicators, suggest that black-white male mortality gaps are widening over the period we study, but this inference is based on a *trend coefficient* that is significant only at the 0.10 level in the regression using GMM mortality estimates, and is not statistically significant at conventional levels when using either the forward or backward mortality estimates. In contrast, estimates from Panel B, which included state of birth indicators, provide convincing evidence that black-white mortality is widening, and suggest that this phenomenon is driven completely by men born in the South. We emphasize that this later inference is very robust when we use the efficient GMM estimates of mortality (with statistical significance on the *trend* of 0.01) but is less convincing when we use either the forward or backward mortality estimates (with statistical significance on the *trend* of 0.10).

The GMM estimator should, according to statistical theory, reduce the inherent sampling variation in the Census data's measures of cohort by state of birth. Because the GMM combines information from the Censuses and the count of deaths from Vital Statistics, it produces a more precise estimate of mortality rates by reducing the noise in our populations

[14] In terms of the notation in the previous section, demographic "group i" is now a single cell given by age, state-of-birth, and birth cohort (e.g., black 60 year old men born in 1932 in Georgia). We have $n = 1,650$ groups: 10 cohorts × 15 states × 11 years.

estimates. This application, in our view, perfectly illustrates this improved precision. The model better fits the data and the resulting standard errors of the parameter estimates are reduced when using the less noisy measure of the dependent variable. Of course, in our application we focused on states with relatively large African American birth cohorts; we would expect more improvement yet from states with smaller birth cohorts.

4 Conclusion

This paper establishes a simple GMM estimator for the purposes of drawing statistical inference when demographers combine data from two sources. To our knowledge, this is the first application of GMM statistical procedures for the purpose of demographic estimation.

We develop an example that demonstrates the estimator. Our application is a potentially valuable one. We are able to estimate, quite accurately, differences in the mortality rates of black and white men by birth cohort and birth states for cohorts born during the Great Depression. We find that state-of-birth effects are an important correlate with black-white mortality differences. Moreover, for Southern-born men, we find strong evidence that racial differences in mortality rates increased for this generation. For Northern-born men, however, we cannot reject the hypothesis that there was no change in the black-white mortality gap.

As we have mentioned, natural future use of GMM estimation might include the examination of mortality by race, gender, and birth state over more states, more cohorts, and more ages. Also, these methods would be useful for analyses that look at death rates by cause of death.

More generally, GMM procedures are potentially useful for estimating other objects of interest in demography—fertility rates, marriage rates, migration, etc.—or for conducting data validation when more than one data source is available to estimate a population parameter.

Table 1: Black-White Mortality Rates by Age, Cohort Born Between 1930 and 1939 in Selected States

Panel A: Correlation of the Three Measure of the Dependent Variable, the Logarithm of the Annual Death Rate

	GMM Estimates	**Forward Estimates**	**Backward Estimates**
GMM Estimates	1.000	—	—
Forward Estimates	0.904	1.000	—
Backward Estimates	0.947	0.720	1.000

Panel B: GMM Estimates of Black and White Mortality Rates by Age

Age	**Black Male Mortality Rate**	**White Male Mortality Rate**	**Differences**	**N**
51	0.015	0.007	0.008	15
52	0.015	0.008	0.008	30
53	0.016	0.008	0.008	45
54	0.017	0.009	0.008	60
55	0.018	0.010	0.008	75
56	0.020	0.011	0.009	90
57	0.021	0.012	0.009	105
58	0.023	0.013	0.010	120
59	0.024	0.014	0.010	135
60	0.026	0.016	0.010	150
61	0.028	0.017	0.011	150
62	0.030	0.019	0.011	135
63	0.032	0.020	0.012	120
64	0.035	0.022	0.012	105
65	0.036	0.023	0.013	90
66	0.039	0.025	0.013	75
67	0.041	0.027	0.014	60
68	0.042	0.030	0.012	45
69	0.045	0.032	0.013	30
70	0.048	0.035	0.013	15

Notes: Sample is men born between 1930 and 1939 in the states of Alabama, Arkansas, Georgia, Illinois, Indiana, Louisiana, Mississippi, North Carolina, New Jersey, New York, Ohio, Pennsylvania, South Carolina, Tennessee, and Virginia.

Table 2: Black-White Male Mortality Differences, 1930–1939 Birth Cohorts for Selected States

Panel A. No State of Birth Indicators

	GMM Estimator	Forward Estimator	Backward Estimator
Age indicators	Yes	Yes	Yes
Trend	0.0052*	0.0061*	0.0041
	(0.00283)	(0.00339)	(0.00365)
North × Trend	0.0012	-0.0049	0.0066
	(0.00330)	(0.00415)	(0.00398)
Ratios of standard error to GMM standard error	1.000, 1.000	1.198, 1.258	1.290, 1.206
R^2	0.247	0.194	0.229
N	1,650	1,650	1,650

Panel B. State of Birth Indicators Included

Age indicators	Yes	Yes	Yes
Trend	0.0069***	0.0062*	0.0069*
	(0.00236)	(0.00349)	(0.00324)
North × Trend	-0.0074*	-0.0052	-0.0084
	(0.00439)	(0.00616)	(0.00564)
Ratios of standard error to GMM standard error	1.000, 1.000	1.478, 1.403	1.373, 1.285
R^2	0.347	0.311	0.310
N	1,650	1,650	1,650

Notes: States include Alabama, Arkansas, Georgia, Illinois, Indiana, Louisiana, Mississippi, North Carolina, New Jersey, New York, Ohio, Pennsylvania, South Carolina, Tennessee, and Virginia. Standard errors are clustered at the state-of-birth by birth cohort. Dependent variable is log difference of black and white mortality differences. Ages range from 51 to 70.

*Significant at 10 percent level.
**Significant at 5 percent level.
***Significant at 1 percent level.

References

Arias, E., 2010. "United States Life Tables, 2006," *National Vital Statistics Reports,* 58(21), 1-40.

Barker, D. J. P., 1990. "The Fetal and Infant Origins of Adult Disease," *British Medical Journal,* 301, 1111.

Barker, D. J. P., 1995. "Fetal Origins of Coronary Heart Disease," *British Medical Journal,* 311, 171-74.

Costa, Dora L., Lorens Helmchen, and Sven Wilson, 2007. "Race, Infection, and Arteriosclerosis in the Past," *Proceedings of the National Academy of Science,* 104(33), 13219-24.

Elo, Irma T., and Samuel H. Preston, 1994. "Estimating African-American Mortality from Inaccurate Data," *Demography,* 31(3), 427-458.

Fang, Jing, Shantha Madhavan, and Michael Alderman, 1996. "The Association Between Birthplace and Mortality from Cariovascular Causes among Black and White Residents of New York City," *New England Journal of Medicine,* 335(21), 1545-51.

Fogel, Robert, 2004. *The Escape from Hunger and Premature Death, 1700-2100.* Cambridge University Press.

Geronimus, Arline T., John Bound, and Cynthia G. Colen, 2011. "Excess Black Mortality in the United States and in Selected Black and White High-Poverty Areas, 1980-2000," *American Journal of Public Health,* 101(4), 720-729.

Greenberg, Michael, and Dona Schneider, 1992. "Region of Birth and Mortality of Blacks in the United States," *International Journal of Epidemiology,* 21(2), 324-328.

Hansen, Lars Peter, 1982. "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica,* 50(4), 1029-1054.

Hayashi, Fumio, 2000. *Econometrics,* Princeton University Press.

Hayward, Mark D., and Bridget K. Gorman, 2004. "The Long Arm of Childhood: The Influence of Early-Life Social Conditions on Men's Mortality," *Demography*, 41(1), 87-107.

Hsu, Yu-Chieh, 2012. *Three Essays on Measurement and Evaluation of Mortality,* Ph.D. Dissertation, Carnegie Mellon University.

Lee, R. D., 1985. "Inverse Projection and Back Projection: A Critical Appraisal, and Comparative Results for England, 1539 to 1871," *Population Studies,* 39(2), 233-248.

Levine, Robert, James Foster, Robert Fullilove, Mindy Fullilove, Nathaniel Briggs, Pamela Hull, Baqar Husaini, and Charles Hennekens, 2001. "Black-White Inequalities in Mortality and Life Expectancy, 1933-1999: Implications for Healthy People 2010," *Public Health Reports,* 116, 474-83.

Lleras-Muney, Adriana, 2005. "The Relationship between Education and Adult Mortality in the United States," *Review of Economic Studies,* 72(1), 189-221.

McCord, C., and H. P. Freeman, 1990. "Excess Mortality in Harlem," *New England Journal of Medicine*, 322(3), 173-77.

Palloni A., and R. Kominski, 1984. "Estimation of Adult Mortality Using Forward and Backward Projections," *Population Studies,* 38(3), 479-493.

Preston, Samuel H., Irma T. Elo, and Quincy Stewart, 1999. "Effects of Age Misreporting on Mortality Estimates at Older Ages," *Population Studies,* 53(2), 165-177.

Preston, Samuel H., Mark Hill, and Greg Drevenstedt, 1998. "Childhood Conditions that Predict Survival to Advanced Ages Among African Americans," *Social Science Medicine,* 47(9), 1231-46.

Vincent, P., 1951. "La Mortalité des Vieillards," *Population,* 6, 181-204.

Warner, David F., and Mark D. Hayward, 2006. "Early-Life Origins of the Race Gap in Men's Mortality," *Journal of Health and Social Behavior,* 47, 209-26.

Wrigley, E. A., and R. S. Schofield, 1981. *The Population History of England 1541-1871: A Reconstruction,* Harvard University Press.

Appendix. Comparison to the Maximum Likelihood Estimator

Hansen (1982) establishes the optimal properties of the GMM estimator among the class of method of moments estimators. However, GMM estimation is not familiar in demographic research, so many readers might find it helpful to compare the GMM approach to the more familiar idea of maximum likelihood (ML).

As we have seen, our problem boils down to estimating the fraction of the population that is in group i in time 0, which we designate p_i^0. To set the stage, recall that if one wanted to estimate that parameter using ML based solely on Census data in time 0, the goal would be to choose the estimator that maximizes the log of

$$\mathcal{L} = \frac{S^0!}{S_i^0!(S^0 - S_i^0)!} {p_i^0}^{S_i^0} (1 - p_i^0)^{(S^0 - S_i^0)}. \tag{22}$$

The ML estimator is easily found here by taking the derivative of the log likelihood with respect to p_i^0 and setting to 0. The resulting estimator is the mean, $\tilde{p}_i^0 = S_i^0 / S^0$.

Our ML problem, incorporating data from both the Census and from Vital Statistics, is a bit harder. In this case we want to maximize the joint log likelihood of p_i^0 and p_i^1, given by

$$\ln\left[{p_i^0}^{S_i^0} (1 - p_i^0)^{(S^0 - S_i^0)}\right] + \ln\left[{p_i^1}^{S_i^1} (1 - p_i^1)^{(S^1 - S_i^1)}\right] + C \tag{23}$$

(where C is a constant that is independent of the parameters), subject to the constraint

$$\omega^0 p_i^0 S^0 - \omega^1 p_i^1 S^1 - {}_T D_i = 0. \tag{24}$$

Carrying out the constrained maximization problem, and following an extensive series of algebraic steps, we can show that the ML estimates are the values, $\tilde{p}_i^0$ and $\tilde{p}_i^1$, that solve

$$\omega^0 \tilde{p}_i^0 S^0 - \omega^1 \tilde{p}_i^1 S^1 - {}_T D_i = 0, \tag{25}$$

$$\frac{1}{\omega^0 S^0 \tilde{p}_i^0 (1 - \tilde{p}_i^0)} \left[\tilde{p}_i^0 S^0 - S_i^0\right] + \frac{1}{\omega^1 S^1 \tilde{p}_i^1 (1 - \tilde{p}_i^1)} \left[\tilde{p}_i^1 S^1 - S_i^1\right] = 0. \tag{26}$$

Below we use numerical methods to solve (25) and (26) to form ML estimates. Then with the estimates of p_i^0, we proceed to estimate mortality using

$${}_T q_i^{ML} = \frac{{}_T D_i}{\tilde{p_i}^0 N^0}, \tag{27}$$

where $\tilde{p_i}^0 N^0$ serves to estimate N_i^0 (the denominator for the estimator) for each demographic group.

Our interest here is the comparison of the ML estimator to the GMM procedure outlined above. Recall that the GMM estimator of N_i^0 is a two step estimator in which one first gets the minimum distance estimators, $\hat{N}_i^{0,MD}$ and $\hat{N}_i^{1,MD}$, and uses those to find $\hat{p}_i^0$ and $\hat{p}_i^1$. These values are then used in a second stage, using equation (16), to find the second-stage estimator of N_i^0. In principle one could similarly find a second-stage estimate of N_i^1, and then use *those* second-stage estimators to get updated estimates of p_i^0 and p_i^1. These new estimates could again be used in (16) to get third-stage estimators. The process could be repeated in a fourth stage, and so on, until the exercise converges to fixed points, say $\check{p}_i^0$ and $\check{p}_i^1$. Suppose such fixed points satisfy (16), but now with $\check{N}_i^0 = \check{p}_i^0 N^0$ on the left-hand side, and with $\check{p}_i^0$ replacing $\hat{p}_i^0$ and $\check{p}_i^1$ replacing $\hat{p}_i^1$ on the right-hand side. Following many algebraic manipulations, we find that these "iterated GMM" estimates must then also solve

$$\omega^0 \check{p}_i^0 S^0 - \omega^1 \check{p}_i^1 S^1 - {}_T D_i = 0, \tag{28}$$

$$\frac{1}{\omega^0 S^0 \check{p}_i^0 (1-\check{p}_i^0)} \left[\check{p}_i^0 S^0 - S_i^0\right] + \frac{1}{\omega^1 S^1 \check{p}_i^1 (1-\check{p}_i^1)} \left[\check{p}_i^1 S^1 - S_i^1\right] = 0. \tag{29}$$

Notice that the solution for iterated GMM, (28) and (29), takes precisely the same form as the equations that solve ML, (25) and (26); if we were to take the two-step GMM procedure and iterate as an n-step procedure we would converge to the ML estimates. In short, GMM can be thought of here as the first two steps in an iterative process that solves ML.

As Hayashi (2000) notes (see pages 481-482), in general GMM is less efficient than ML. The exception is in such cases as ours—when one can exploit knowledge of the parametric form of the density function in forming the weighting matrix W^{-1}. While MLE is a sensible method to use for our problem, both ML and GMM are asymptotically efficient, and the GMM approach is considerably easier to implement.