Estimating Mortality by State of Birth and Race Using Census and Vital Statistics Data

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

January 31, 2012

Abstract

Demographers often estimate population parameters by combining data from separate sources. We propose a generalized method of moments (GMM) estimator for such cases, and to illustrate this statistical approach we estimate mortality rates using data from the U.S. Census and from the Vital Statistics. Specifically, we estimate mortality rates by race, gender, birth cohort, and State of birth for cohorts born in the 1930s. On a substantive level, resulting estimates are interesting: In mid-life (ages 40 through 60), State-of-birth effects are quite large for men, but not for women. Among both black and white men, mortality is generally higher for individuals born in relatively low-income States—typically Southern States.

Keywords: mortality, racial disparity in mortality, generalized method of moments (GMM) estimation

We gratefully acknowledge support for this research from the National Institute for Child Health and Development and the Population Center at the University of Chicago. This work is preliminary and incomplete. Comments welcome.

1 Introduction

The two most important sources for inference in demography are data from Census records, and official Vital Statistics, which record key individual-level events such as births and deaths. An important problem in empirical demography is how to optimally combine data from such separate sources for the purpose of estimation.

In this paper we work through one example. The empirical issue we have in mind is a typical problem in demography—the estimation of mortality rates for subsets of the population. In our case we are interested in studying differences in mortality that emerge in mid-adulthood—at ages 40 through 60—by race, gender, and State of birth. Our focus on the State of birth is potentially important if State-level variation in conditions affecting prenatal and early childhood health has a meaningful impact on mid-life mortality.

One obvious, and useful, way to evaluate mortality rates by State of birth is to estimate ten-year mortality rates using U.S. Census data, which include birth State as a data element. For example, suppose we want to estimate middle-age mortality for black men born in Georgia in 1932 (denoted here as group i) over the period 1980 through 1990. Let N_i^0 be the count of individuals in this cohort in 1980, a Census year, and N_i^1 be the comparable count from the following decennial Census, taken in 1990. Then $(N_i^0 - N_i^1)/N_i^0$ is the ten-year mortality rate (from, roughly, age 48 to 58). We do not have the detail needed to estimate mortality with "short form" data (which covers the population); estimates must be formed by using "long form" data, which merely sample the population. These samples are generally 1-in-5 or 1-in-6 for restricted use data, and 1-in-20 or 1-in-100 for the Public Use Microsamples (PUMS). Thus, to continue our example, if we wish to estimate middle-age mortality for black men born in Georgia in 1932, we need to rely on samples: Letting S_i^0 and S_i^1 denote samples from the PUMS, a natural estimator is $(S_i^0 - S_i^1)/S_i^0$. Indeed, this estimator has been put to good use in such work as Lleras-Muney (2005).[1]

A clear problem with the approach outlined in the previous paragraph is that estimates based on small samples, as with the cohort of black men born in Georgia in 1932, are likely to be quite noisy. Indeed, when we turn to estimation below, due to sampling variation it is quite common for such estimates to actually be negative.

We can easily improve on the accuracy of inferences by incorporating an additional data source—data on deaths from the detailed mortality files. These data provide a presumably highly accurate counting of deaths, and include sex, race, age, and State of birth for recorded deaths. Thus, we could use these data to make an accurate assessment of the *number* of deaths to black men born in Georgia in 1932 between 1980 and 1990. On their own, the data cannot be used to estimate death *rates,* of course, because there is no base. Still, intuition suggests that this information is useful for the purpose of inference in our context.

[1] Lleras-Muney uses the estimator for calculating cohort-specific mortality by birth State for white individuals, for the purpose of inferring the impact of State education policies on later-life mortality. Notice that the estimator requires an additional assumption that there is no international out-migration or return-migration for the group—an assumption that is probably fine for the example cohort (and, in any event, can be checked with available data). Also, as discussed below, the estimator needs some modification if the sample provided is weighted.

Our problem, then, is how to optimally combine data from the three sources—the two Census samples and Vital Statistics records—to estimate mortality rates by race, gender, birth cohort, and State of birth.

In Section 2 of our paper we set out an intuitive minimum distance (MD) estimator, and then generalize the solution to an easily-implemented two-step estimator—an optimal 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 Section 3 of our paper we turn to our substantive application. We estimate the effect of birth State on middle-age mortality rates for black and white individuals born in 15 States during the 1930s. We demonstrate the viability of the GMM approach, and show how this approach substantially improves our ability to draw inferences, compared to estimation using Census data only. Our estimates document substantial variation in State-of-birth effects for men, especially black men, though not for women. We make an additional empirical contribution to the literature on the social forces that impact mortality by showing a negative relationship, at the level of State of birth, between mid-life mortality and household income in childhood.

In Section 4 we provide concluding remarks.

2 Estimating Mortality Using Two Data Sources

Our problem is conceptually quite simple. Suppose that in 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, resulting in a sample of $S^0 = N^0/\omega^0$ individuals. An example is the 1-in-20 public use sample of the U.S. Census. In period 1 (10 years later in the case of the U.S. Decennial Census) we similarly have a Census that samples at a rate of 1 in ω^1, resulting in a sample of $S^1 = N^1/\omega^1$ individuals. Our interest is the mortality rates of a subset i of the population, for example selected on the basis of birth cohort, State of birth, gender, race, etc. Let S_i^0 be the count of such individuals in year 0 and let S_i^1 be the corresponding count in year 1.

As we mention in the introduction, Census data alone are sufficient to estimate 10-year mortality rates of interest here. The actual mortality rate for group i,

$$\frac{(N_i^0 - N_i^1)}{N_i^0},$$

can be estimated by

$$\frac{(\omega^0 S_i^0 - \omega^1 S_i^1)}{\omega^0 S_i^0}, \tag{1}$$

if all individuals in the population are sampled at the same rates in each of the two periods, i.e., if ω^0 and ω^1 are the appropriate "inflation factors" in the respective periods. Of course, if inflation factors differ across individuals in the sample, our estimator is slightly more complicated. Now index all individuals with j and let $j \in i$ indicate individuals in group i. Then mortality is estimated by

$$\frac{\left(\sum_{j\in i} \omega_j^0 - \sum_{j\in i} \omega_j^1\right)}{\sum_{j\in i} \omega_j^0}.$$

For the remainder of this section we restrict attention to the case given in (1). We present this case because there is no additional insight that follows from allowing for differing inflation weights, but there is considerable notational clutter.[2]

As we mention in the introduction, estimator (1) is likely to be quite noisy when samples are small. We can do much better if we have a count from Vital Statistics data of the number of individuals in group i who have died between time 0 and time 1. We let D_i be such a count, and suppose that it has been accurately recorded. Given that we have the number of deaths in group i we can estimate the death rate using $D_i/(\omega^0 S_i^0)$, because $\omega^0 S_i^0$ is a consistent estimate of the number of people in group i alive in time 0 (given the 1-in-ω^0 sample). This is likely to be a substantial improvement over the estimator in (1) because here at least we have an accurate *numerator.*

We can potentially do better yet by exploiting a direct relationship between N_i^0 (e.g., the population of black men born in Georgia in 1932 observed in 1980) and N_i^1 (that same population in 1990):

$$N_i^0 = N_i^1 + D_i + E_i, \tag{2}$$

where E_i is the net emigration of the group. For the United States during the last half of the 20th century, $E_i \approx 0$ for middle-aged individuals born in the U.S., so we could estimate N_i^0 with either $\omega^0 S_i^0$, as in the previous paragraph, or with

$$\omega^1 S_i^1 + D_i. \tag{3}$$

Either approach to estimating N_i^0 is likely to be noisy; intuitively, one would like to use both pieces of information in forming inferences.[3]

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

$$\begin{aligned} \mathrm{E}\left\{\omega^0 S_i^0 - N_i^0\right\} &= 0, \\ \mathrm{E}\{\omega^1 S_i^1 + D_i - N_i^0\} &= 0. \end{aligned} \tag{4}$$

[2] When we turn to estimation, though, we have samples in which inflation factors differ across individuals, so we use appropriate weights in forming estimates.

[3] Because D_i comes from Vital Statistics records, it is precisely measured. The problem comes with samples S_i^0 and S_i^1, as they are small. (In practice, samples often become smaller as a cohort ages, of course.)

The expressions in (4), which involve expectations, are often called *moment restrictions.*[4] Given that our goal is to find estimators that fit equations (4) "well," an intuitively attractive idea is to find value $\hat{N_i^0}$ that minimizes the expression,

$$\begin{bmatrix} N_i^0 - \omega^0 S_i^0 & N_i^0 - \omega^1 S_i^1 - D_i \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} N_i^0 - \omega^0 S_i^0 \\ N_i^0 - \omega^1 S_i^1 - D_i \end{bmatrix}. \tag{5}$$

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

$$\min_{N_i^0} V = \left(N_i^0 - \omega^0 S_i^0\right)^2 + \left(N_i^0 - \omega^1 S_i^1 - D_i\right)^2. \tag{6}$$

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

$$\frac{dV}{dN_i^0} = 2\left(\hat{N}_i^0 - \omega^0 S_i^0\right) + 2\left(\hat{N}_i^0 - \omega^1 S_i^1 - D_i\right) = 0, \tag{7}$$

which leads to the resulting estimator,

$$\hat{N_i^0} = \frac{1}{2}\left(\omega^0 S_i^0\right) + \frac{1}{2}\left(\omega^1 S_i^1 + D_i\right). \tag{8}$$

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 parameter estimate of interest. Let d_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

$$d_i^{\text{MD}} = \frac{D_i}{\hat{N_i^0}}, \tag{9}$$

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. Furthermore, as we show below, the estimator works well in our application. An important paper by Hansen (1982), though, establishes a generalization of the minimum distance estimator that has optimal properties, in terms of minimizing the estimator's asymptotic variance. We turn to that estimator next.

[4] Our restrictions assume that D_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.

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 (5), but in which the matrix in the interior of (5) 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

$$\begin{aligned} W &= \mathrm{E}\left\{ \left[\begin{array}{c} N_i^0 - \omega^0 S_i^0 \\ N_i^0 - \omega^1 S_i^1 - D_i \end{array} \right] \left[\begin{array}{cc} N_i^0 - \omega^0 S_i^0 & N_i^0 - \omega^1 S_i^1 - D_i \end{array} \right] \right\} \\ &= \left[\begin{array}{cc} \left(\omega^0\right)^2 S^0 p_i^0 (1 - p_i^0) & 0 \\ 0 & \left(\omega^1\right)^2 S^1 p_i^1 (1 - p_i^1) \end{array} \right], \end{aligned} \tag{10}$$

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.[5] The terms in W are easy to find as our particular problem entails draws from two independent binomial processes.[6] Hansen proves that the use of W^{-1} is optimal in terms of minimizing the asymptotic variance of the estimator.

Since we don't 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 (5), 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 2.1. The idea is to use the estimator in (8) 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 (10), 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 $\hat{\hat{N}}_i^0$ that minimizes

$$\begin{aligned} &\left[\begin{array}{cc} \hat{\hat{N}}_i^0 - \omega^0 S_i^0 & \hat{\hat{N}}_i^0 - \omega^1 S_i^1 - D_i \end{array} \right] \left[\begin{array}{cc} (\omega^0)^2 S^0 \hat{p}_i^0 (1 - \hat{p}_i^0) & 0 \\ 0 & (\omega^1)^2 S^1 \hat{p}_i^1 (1 - \hat{p}_i^1) \end{array} \right]^{-1} \times \\ &\qquad \left[\begin{array}{c} \hat{\hat{N}}_i^0 - \omega^0 S_i^0 \\ \hat{\hat{N}}_i^0 - \omega^1 S_i^1 - D_i \end{array} \right], \end{aligned} \tag{11}$$

which yields the necessary condition,

$$\left((\omega^0)^2 S^0 \hat{p}_i^0 (1 - \hat{p}_i^0)\right)^{-1} \left(\hat{\hat{N}}_i^0 - \omega^0 S_i^0 \right) + \left((\omega^1)^2 S^1 \hat{p}_i^1 (1 - \hat{p}_i^1)\right)^{-1} \left(\hat{\hat{N}}_i^0 - \omega^1 S_i^1 - D_i \right) = 0. \tag{12}$$

[5] 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 don't directly observe p_i^0 or p_i^1, since N_i^0 and N_i^1 are unknown.

[6] 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 $S^0 p_i^0 (1 - p_i^0)$.

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

$$\begin{aligned}\hat{\hat{N}}_i^0 = &\left[\frac{\left((\omega^0)^2 S^0 \hat{p}_i^0(1-\hat{p}_i^0)\right)^{-1}}{\left((\omega^0)^2 S^0 \hat{p}_i^0(1-\hat{p}_i^0)\right)^{-1} + \left((\omega^1)^2 S^1 \hat{p}_i^1(1-\hat{p}_i^1)\right)^{-1}}\right] \omega^0 S_i^0 \\ &+ \left[\frac{\left((\omega^1)^2 S^1 \hat{p}_i^1(1-\hat{p}_i^1)\right)^{-1}}{\left((\omega^0)^2 S^0 \hat{p}_i^0(1-\hat{p}_i^0)\right)^{-1} + \left((\omega^1)^2 S^1 \hat{p}_i^1(1-\hat{p}_i^1)\right)^{-1}}\right] \left(\omega^1 S_i^1 + D_i\right). \end{aligned} \tag{13}$$

As in (8), 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

$$d_i^{\text{GMM}} = \frac{D_i}{\hat{\hat{N}}_i^0}. \tag{14}$$

To build intuition for this estimator, consider the case in which the inflation weights are the same in the two samples and the population size is the same in periods 0 and 1. Then (13) reduces to

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

Now consider two cases. First suppose $\hat{p}_i^0 \approx \hat{p}_i^1$. This will be true when the mortality rate is very low. In this case, the weights (in brackets) are approximately $\frac{1}{2}$; the two estimates of N_i^0 are given roughly equal weight. 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, 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.

The second case shows that the GMM estimate, in the extreme case, converges to "extinct generation estimation"—a methodology used in many important papers, e.g., Elo and Preston (1994).[7] The intuition is straightforward. If death counts from Vital Statistics are accurate, for an *extinct cohort* we can estimate the number of people in the cohort who were alive in a prior period simply by counting recorded deaths. When a cohort near extinction, S_i^1 approaches 0 (and will be much smaller than D_i), and 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 (14)).

As an alternative to GMM, researchers could calculate mortality by estimating the base from the Census in period 0 (ignoring information from period 1), while estimating the

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

numerator using Vital Statistics data. Our analysis shows that such a weighting scheme is sub-optimal, and deviates from optimality dramatically when mortality rates are high.[8]

2.3 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

$$\mathcal{L} = \frac{S^0!}{S_i^0!(S^0 - S_i^0)!} {p_i^0}^{S_i^0} (1 - p_i^0)^{(S^0 - S_i^0)}. \tag{15}$$

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

$$\ln \left[{p_i^0}^{S_i^0} (1 - p_i^0)^{(S^0 - S_i^0)} \right] + \ln \left[{p_i^1}^{S_i^1} (1 - p_i^1)^{(S^1 - S_i^1)} \right] + C \tag{16}$$

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

$$\omega^0 p_i^0 S^0 - \omega^1 p_i^1 S^1 - D_i = 0. \tag{17}$$

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

$$\omega^0 \tilde{p}_i^0 S^0 - \omega^1 \tilde{p}_i^1 S^1 - D_i = 0, \tag{18}$$

$$\frac{1}{\omega^0 S^0 \tilde{p}_i^0 (1 - \tilde{p}_i^0)} \left[\tilde{p}_i^0 S^0 - S_i^0 \right] + \frac{1}{\omega^1 S^1 \tilde{p}_i^1 (1 - \tilde{p}_i^1)} \left[\tilde{p}_i^1 S^1 - S_i^1 \right] = 0. \tag{19}$$

Below we use numerical methods to solve (18) and (19) to form ML estimates. Then with the estimates of p_i^0, we proceed to estimate mortality using

$$d_i^{ML} = \frac{D_i}{\tilde{p_i}^0 N^0}, \tag{20}$$

[8] As Shiro Horiuchi pointed out to us, in terms of the previous literature, we are essentially combining "forward" and "backward" methods for estimating mortality (see, e.g., Bennett and Horiuchi, 1981, for a discussion of classical methods in indirect estimation, with a focus on problems that arise when deaths are under-registered).

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, $\hat{N}_i^0$ and $\hat{N}_i^1$, 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 (13), 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 (13) 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 (13), 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

$$\omega^0 \check{p}_i^0 S^0 - \omega^1 \check{p}_i^1 S^1 - D_i = 0, \tag{21}$$

$$\frac{1}{\omega^0 S^0 \check{p}_i^0 (1-\check{p}_i^0)} \left[\check{p}_i^0 S^0 - S_i^0\right] + \frac{1}{\omega^1 S^1 \check{p}_i^1 (1-\check{p}_i^1)} \left[\check{p}_i^1 S^1 - S_i^1\right] = 0. \tag{22}$$

Notice that the solution for iterated GMM, (21) and (22), takes precisely the same form as the equations that solve ML, (18) and (19); 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.

As for the minimum distance (MD) estimator, it is consistent but not efficient. It is, however, simpler even than GMM, and in our example below it works extremely well.[9]

3 Application: The Role of Birth State in Shaping Adult Mortality

3.1 The Research Question and Basic Empirical Strategy

Our application entails the estimation and analysis of mortality in mid-life—ages approximately 40 to 60—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.

[9] Altongi and Segal (1996) show that MD can have better small sample properties than GMM, though their concerns are not applicable in our context. In our application (which does entail relatively small samples in some instances) GMM and ML estimates are nearly indistinguishable, and in turn the MD estimates are very similar to those estimates.

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 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) who assess the extent to which childhood and adulthood conditions account for the race gap in men's mortality.[12]

[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.

[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.

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.

Given this literature, it seems likely 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 10-year mortality rates for these individuals for 1980 through 1990. Thus we are looking at mortality in the mid-life (at ages approximately 40 through 60).[13] By focusing on cohorts born in the 1930s, we are studying individuals who were born during the Great Depression—an era of great deprivation in many U.S. States. With credible estimates of mid-life mortality rates, generated using our GMM approach, we can begin to ask such basic questions as, "Is mid-life mortality higher among those born in States that had relatively high rates of childhood poverty?"

As for the States we study, we note that as of the 1930s, most African Americans were born in the Southern States, so we include the nine largest States in terms of births of black individuals. For comparison, we include six large Northern States. As we will see, there are substantial challenges to estimating mortality among black men and women in these latter States because of the relatively smaller sample sizes in the Census.

In Section 3.2 we focus on methodology—comparing estimates using the various methods we have described above—and then in Sections 3.3 and 3.4 we give substantive results about the role of State of birth in shaping mortality in mid-life among cohorts of black and white individuals born in the 1930s.

3.2 Mortality Estimates by Birth State for Individuals Born 1930–1939: A Comparison of Four Estimators

Our goal in this first section of empirical results is to compare estimates of mortality using the four estimators mentioned in our methodological section: (1) the estimator based on Census data only, e.g., as used by Lleras-Muney (2005), and the additional estimators that use both Census and Vital Statistics data, as discussed above, i.e. (2) the MD estimator, (3) the GMM estimator, and (4) the ML estimator.

Given that we are estimating mortality in State of birth × sex × race × birth cohort cells, in many cases we are estimating mortality on the basis of relatively small samples. With

[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).

this in mind, consider estimates for New York, presented in graphical form in Figure 1. The panels provide four sets estimates of the 10-year mortality rate, from 1980 through 1990, for the birth cohorts 1930–1939.

Estimates based on Census data along are extremely noisy—so noisy that it would be difficult to draw inferences about the extent to which mortality varies by birth cohort, gender, or race. The variability in estimates is especially great for black men and women. This makes sense, since these estimates are based on smaller samples than for white men and women. By way of comparison, the ML, MD, and GMM approaches provide extremely similar estimates, and these estimates are sensible. Mortality rates are lower for the more recent cohorts than for earlier cohorts, they are lower for women than for men, and they are lower for whites than for blacks.

Our primary purpose in producing mortality estimates is to make comparisons across State of birth, gender, and race. For this purpose, MD, GMM, and ML estimates—based on Census and Vital Statistics data—prove to be much more useful than the noisier counterparts drawn from the Census data alone. To establish this point, we begin with one demographic group, black men, and for that group estimate the following regression:

$$\ln(d_{cs}) = \alpha_c + \beta_s + \epsilon_{cs}, \tag{23}$$

where c indicates birth cohort ($c = 1930, \ldots, 1939$), and s indicates State. As noted above, we use 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]

In short, we are estimating the impact of State of birth on log ten-year mortality rates in a regression model that has a full set of birth cohort effects for black men born 1930–1939. As for the State-of-birth effects, for ease of interpretation we normalize them to average 0.[15]

Column (1) of Table 1 shows estimated State effects for our regression if we use estimates of mortality with Census data only, taken from the 1980 and 1990 public use samples. Standard errors are very large; the evidence does not allow us to draw firm conclusions about the importance of State of birth on later-life mortality for our population. This result is not surprising. As illustrated in Figures 1, our dependent variable here is being estimated with a great deal of noise. As it turns out, our mortality measures are so noisy that we simply cannot draw useful inferences about the role of birth State in shaping later-life mortality for African American men.

Columns (2) through (4) of Table 1 then shows estimated State effects from our regression when we use estimates of mortality using Census and Vital Records data, using, respectively, MD, GMM, and ML procedures. In contrast to results reported in Column (1) of Table 1, in these columns we find very interesting results. We notice that estimated State-of-birth effects are quite precisely estimated. Also, the MD, GMM, and ML approaches give very

[14] In terms of the notation in the previous section, demographic "group i" is now a single cell given by c and s (e.g., black men born in 1932 in Georgia). We have $n = 150$ groups: 10 cohorts × 15 States.

[15] Following Haisken-DeNew and Schmidt (1997) and subsequent authors, such as Kovak (2011), we norm the State effects to average 0. Specifically, the estimated State coefficients are re-normed from a first-step standard regression, expressing the results as deviation from a weighted average (with equal weight across States).

similar answers. Indeed the correlation between estimated coefficients using MD and GMM approaches exceeds 0.999, and the correlation between estimated coefficients using ML and GMM approaches is greater than 0.998.

To conclude our comparison of estimators, we recall that there is yet another way of combining data from the Census and Vital Statistics to calculate 10-year mortality rates: we could simply estimate the base from the 1980 Census data alone, and the deaths from Vital Statistics records. Notice that we are thereby giving *zero* weight to information available from the 1990 Census data, and as noted above, this is sub-optimal. Still, this is a simple and well-used estimator, and to get a sense of how such estimates compare to our GMM estimates, we calculated 10-year mortality for *all* 50 States of birth plus DC by sex × race × birth cohort cells for the 1930-1939 cohorts under study. If we restrict attention to the 1685 cells that have an initial sample size greater than 100, the two estimators give reasonably similar answers: the correlation between the estimates is 0.972. However, for the 235 cells with 100 or fewer observations, the correlation is only 0.748.[16] Our theory shows why we are better off with the GMM approach, and this investigation shows, not surprisingly, that the advantages to GMM are most important when we have smaller samples.

For the remainder of our empirical work we use the GMM approach, which is attractive for both its simplicity and optimal asymptotic properties.

3.3 Birth-State Effects for Mid-Life Mortality

We proceed with our analysis of mortality focusing first on men. Panel A of Table 2 shows that for the 1930-1939 cohorts, ten-year mortality, 1980 to 1990, was approximately twice as high for black men as white men; mortality was 140 per 1000 for blacks and 69 per 1000 for whites. In first set of columns in Panel B, we again report results using GMM estimates for State effects given in column (3) of Table 1, but now order States from lowest to highest mortality according to these GMM estimates. Our results indicate substantial variation across States in the ten-year mortality rates for black men. For example, the ten-year mortality rate is 18 log points higher than the 15-State average for men born in South Carolina, while it is more than 19 log points lower than the average for those born in Ohio.[17] Every statistically significant positive State effect is for a Southern State, while every statistically significant negative State effect is for a Northern State.

A major interest is the comparison of mortality patterns for the black and white populations. So in right-hand columns of Table 2, we repeat the analysis for the white population. The variation across States is lower for white men than for black men, but the general pattern is similar. Every Southern State has a positive estimated coefficient, while every Northern State has a negative estimated coefficient.

Table 3 repeats our exercise, but for women. Mortality rates are much lower for women than for men, but as with men, mid-life mortality rates are substantially higher among blacks

[16] Overall the unweighted correlation is 0.889. (In all calculations we exclude cases for which initial cell sizes were 0, which leaves 1920 cells overall.)

[17] Log points are quite close to percentages, so those born in South Carolina have mortality rates that are approximately 18 percent above the average.

than whites. As for State effects, patterns are quite different for women than for men. For black women, mortality is approximately 12 percent higher than average for those born in Georgia and is approximately 9 percent higher for those born in South Carolina. Otherwise, estimated State effects do not significantly differ from 0. For white women, State effects are quite precisely estimated and several estimates are significantly different than 0. However, the variation in estimated State effects is quite small, and we do not observe the distinctive South-North pattern in estimated coefficients that observed for white men (in Table 2).

3.4 The Relationship Between Mid-Life Mortality and Birth-State Childhood Conditions

The primary purpose of our paper is to introduce GMM estimation for the purpose of estimating mortality from multiple sources, and to demonstrate its use in one application. Having calculating mid-life mortality for the 1930-1939 cohorts by State of birth, though, we proceed in this section to provide an example of how statistics of the sort we generate can be used for further analysis. The idea here is to see if mid-life mortality measured by State of birth is correlated with State-level average household conditions for these individuals when they were children.

As a first step we use data from the 1940 U.S. Census to examine State-level characteristics of households that had children aged 1-10, i.e., children born 1930-1939. We undertake this exercise for black households and white households. Results are given in Table 4.

Column (1) gives household income. To adjust for inflation to 2011 buying power in the usual way—using the consumer price index (CPI) adjustment from the Bureau of Labor Statistics—one can multiply by 16.12. Undertaking this exercise underscores that at the end of the Great Depression young children typically lived in households with very low income. For example, the inflation-adjusted annual household income for black children born in Mississippi (the poorest households) is only \$2,628. The highest-income households are whites in New Jersey, with an inflation-adjusted average income of \$26,635. Notice that there was extremely high variation across States and races in household income.

Column (2) gives household income per household member. This figure is especially low for black children born in Mississippi, Georgia, and South Carolina; in these States the inflation-adjusted income per household member was on the order of \$1,000. At the other extreme, white children in New York and New Jersey lived in households with inflation-adjusted income per household member equal to approximately \$6,000.

Finally, column (3) gives the average years of schooling for the household head for the households we examine. These education measures are very low for blacks in Southern States, especially such States as Louisiana, Georgia, and South Carolina. Average education was generally highest for whites in Northern States.

Table 5 presents coefficient estimates from State-level regressions that specify log of the 10-year death rate to be a function of cohort fixed effects and household characteristics measured in 1940:

$$\ln(d_{cs}) = \gamma_c + \beta x_s + \epsilon_{cs}, \tag{24}$$

for cohorts $c = 1930, 1931, \ldots, 1939$, where x_s is the relevant State-level measure of household characteristics. We estimate this regression separately for men and women, and also then separately by race. We enter the State-level characteristics as (1) the natural logarithm of household income, (2) the log of household income per person in the household, and (3) schooling of the household head, and we do so in three separate regressions. In all cases we estimate standard errors by clustering at the State level.

Results are not statistically significant for women. Given estimates presented in Table 3, this is not surprising; variation in mid-life mortality across birth State is not large for women. For men, on the other hand, mortality is seen to be negatively correlated with the household income and education measures. This is true for both black and white men. For the household income measures we have a log-log specification, so the estimate coefficients are elasticities. Thus, we see that a 10 percent increase in childhood household income per household member is associate with a 1.38 percent decrease in mid-life mortality for black men and a 2.31 percent decrease in mid-life mortality for white men.

Additional steps in analyzing these results might take any number of directions. For example, it would be important to examine cause of death as a means of establishing potential causal pathways.[18] Particularly important in this endeavor would be an effort to understand why State-level childhood characteristics seem to be important for mid-life mortality rates of men but not women. A second line of inquiry might include efforts to examine the role of the Great Migration for black men and women, and to examine patterns of rural-to-urban migration for the population generally. Yet another use for these data would be to look at how State-level health and education policies might have affected subsequent mortality.

Of course, future analyses could examine annual mortality rates (instead of 10-year mortality rates), and do so for more years than we have thus far studied. Also, it would be valuable to look at earlier and later cohorts.[19]

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, and we compare inferences with better-known maximum likelihood (ML) methodology. Asymptotic properties are the same for the GMM and ML estimators, and in our application results are nearly identical, but GMM estimation is much easier to implement.

Our application is a potentially valuable one. We are able to estimate, quite accurately, mid-life mortality rates for blacks and whites by gender, birth cohort, and birth States for cohorts born during the Great Depression. We find that among men there is interesting and important variation: men born in the South have generally higher mortality than men born

[18] Note that our GMM estimation procedure would potentially be very valuable for this purpose.

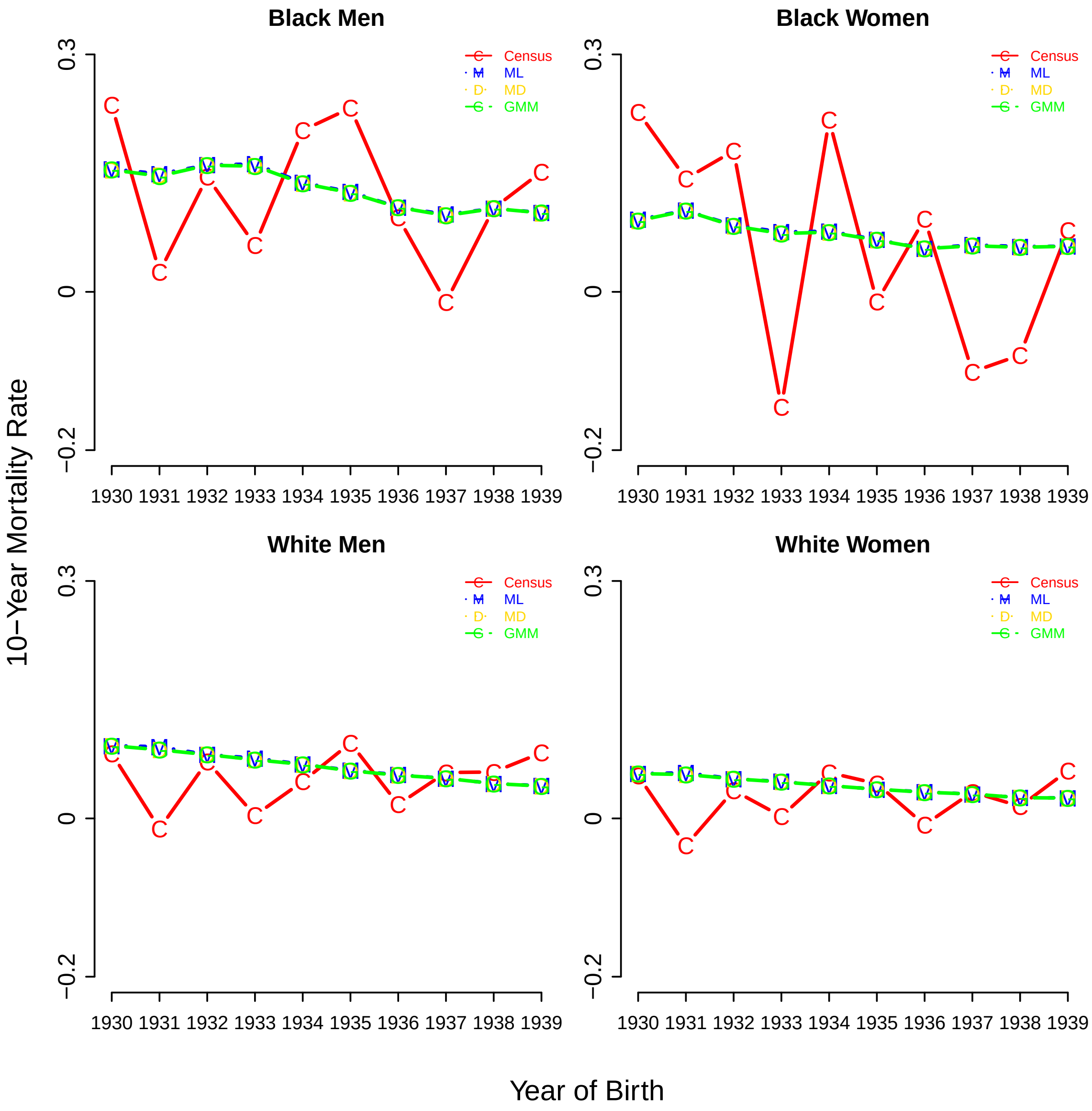
[19] One challenge for researchers who want to study black-white mortality differences at old ages using U.S. Census data is properly accounting for age misreporting. See, e.g., Preston, Elo, and Stewart (1999).

in the North, and birth-State variation is especially for African American men. For men, mid-life mortality is negatively correlated with the socio-economic status of households in the individuals' birth States. In contrast, there are only modest birth-State differences in mid-life mortality rates for women.

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.

Figure 1: Ten-Year Mortality Rates, 1980 to 1990, for New York, by Cohort (1930–1939)



Source: Authors' calculations, data from 1980 and 1990 Census and Vital Statistics.

Table 1: Ten-Year Mortality (1980-1990) by Birth State, Black Men Born 1930-1939: Comparison of State Effects Using Four Estimates

State (S indicates South, and N North)	(1) Estimates Using Census Data	(2) Estimates Using MD	(3) Estimates Using GMM	(4) Estimates Using ML
Alabama (S)	-0.1528	-0.0039	-0.0043	-0.0076
	(0.2919)	(0.0158)	(0.0161)	(0.0157)
Arkansas (S)	-0.0865	0.0223	0.0208	0.0217
	(0.1363)	(0.0167)	(0.0164)	(0.0168)
Georgia (S)	-0.1353	0.1357	0.1359	0.1326
	(0.2420)	(0.0152)	(0.0147)	(0.0147)
Illinois (N)	0.0808	-0.0294	-0.0285	-0.0260
	(0.3617)	(0.0228)	(0.0231)	(0.0217)
Indiana (N)	0.9419	-0.1624	-0.1611	-0.1398
	(0.2482)	(0.0550)	(0.0550)	(0.0550)
Louisiana (S)	0.1526	-0.0161	-0.0144	-0.0197
	(0.1617)	(0.0190)	(0.0195)	(0.0193)
Mississippi (S)	-0.1092	0.0496	0.0493	0.0500
	(0.1975)	(0.0172)	(0.0173)	(0.0156)
New Jersey (N)	0.2734	0.0074	0.0077	0.0073
	(0.2219)	(0.0223)	(0.0235)	(0.0206)
New York (N)	-0.0149	-0.0478	-0.0468	-0.0508
	(0.2055)	(0.0244)	(0.0243)	(0.0244)
North Carolina (S)	0.1357	0.0995	0.0999	0.0950
	(0.1611)	(0.0196)	(0.0197)	(0.0196)
Ohio (N)	-0.1618	-0.1978	-0.1974	-0.2007
	(0.2235)	(0.0405)	(0.0405)	(0.0403)
Pennsylvania (N)	0.0676	-0.0720	-0.0718	-0.0749
	(0.1797)	(0.0214)	(0.0216)	(0.0212)
South Carolina (S)	-0.1854	0.1836	0.1820	0.1810
	(0.2677)	(0.0106)	(0.0110)	(0.0095)
Tennessee (S)	-0.2442	0.0366	0.0355	0.0363
	(0.2130)	(0.0278)	(0.0280)	(0.0274)
Virginia (S)	-0.5620	-0.0053	-0.0068	-0.0044
	(0.4344)	(0.0174)	(0.0173)	(0.0185)

Note: Authors' calculations using U.S. Census data, 1980 and 1990, and Vital Statistics data. Mortality in each cohort is first calculated as described in the text. Log of mortality is the dependent variable in a regression that includes cohort indicator variables, and State indicator variables. State effects, normed to average 0, and are reported here. Standard errors are in parentheses. $n = 150$ (15 States and 10 cohorts).

Table 2: Ten-Year Mortality (1980-1990) by Birth State, Men Born 1930-1939

A. Deaths per 1000

Blacks	Whites
140	69

B. State Effects

	Blacks		Whites
South Carolina (S)	0.1820	South Carolina (S)	0.1123
	(0.0110)		(0.0121)
Georgia (S)	0.1359	Mississippi (S)	0.1044
	(0.0147)		(0.0136)
North Carolina (S)	0.0999	Georgia (S)	0.0976
	(0.0197)		(0.0128)
Mississippi (S)	0.0493	Alabama (S)	0.0937
	(0.0173)		(0.0120)
Tennessee (S)	0.0355	Tennessee (S)	0.0862
	(0.0280)		(0.0066)
Arkansas (S)	0.0208	North Carolina (S)	0.0594
	(0.0164)		(0.0117)
New Jersey (N)	0.0077	Virginia (S)	0.0506
	(0.0235)		(0.0120)
Alabama (S)	-0.0043	Arkansas (S)	0.0406
	(0.0161)		(0.0102)
Virginia (S)	-0.0068	Louisiana (S)	0.0029
	(0.0173)		(0.0135)
Louisiana (S)	-0.0144	New York (N)	-0.0973
	(0.0195)		(0.0098)
Illinois (N)	-0.0285	Pennsylvania (N)	-0.1034
	(0.0231)		(0.0112)
New York (N)	-0.0468	New Jersey (N)	-0.1037
	(0.0243)		(0.0084)
Pennsylvania (N)	-0.0718	Indiana (N)	-0.1112
	(0.0216)		(0.0080)
Indiana (N)	-0.1611	Ohio (N)	-0.1113
	(0.0550)		(0.0096)
Ohio (N)	-0.1974	Illinois (N)	-0.1206
	(0.0405)		(0.0062)

Note: Authors' calculations, U.S. Census data from 1980 and 1990, and Vital Statistics from 1980–1990. Mortality in each cohort is first calculated using GMM (as described in the text). Log of mortality is the dependent variable in a regression that includes cohort indicator variables, and State indicator variables. State effects, normed to average 0, and are reported here. Standard errors are in parentheses. $n = 150$ in each regression.

Table 3: Ten-Year Mortality (1980-1990) by Birth State, Women Born 1930-1939

A. Deaths per 1000

Blacks	Whites
74	39

B. State Effects

	Blacks		Whites
Georgia (S)	0.1229	Georgia (S)	0.0528
	(0.0222)		(0.0116)
South Carolina (S)	0.0897	Alabama (S)	0.0456
	(0.0179)		(0.0133)
Pennsylvania (N)	0.0266	New York (N)	0.0382
	(0.0213)		(0.0077)
Tennessee (S)	0.0208	Arkansas (S)	0.0217
	(0.0181)		(0.0160)
Ohio (N)	0.0052	South Carolina (S)	0.0172
	(0.0358)		(0.0188)
Mississippi (S)	0.0022	Virginia (S)	0.0149
	(0.0143)		(0.0097)
Arkansas (S)	0.0010	Tennessee (S)	0.0096
	(0.0261)		(0.0113)
Louisiana (S)	-0.0114	Illinois (N)	0.0012
	(0.0234)		(0.0069)
North Carolina (S)	-0.0235	Mississippi (S)	-0.0139
	(0.0181)		(0.0152)
New York (N)	-0.0255	Louisiana (S)	-0.0162
	(0.0313)		(0.0158)
Virginia (S)	-0.0255	Ohio (N)	-0.0257
	(0.0172)		(0.0166)
New Jersey (N)	-0.0389	North Carolina (S)	-0.0262
	(0.0457)		(0.0160)
Illinois (N)	-0.0469	New Jersey (N)	-0.0332
	(0.0440)		(0.0128)
Alabama (S)	-0.0476	Pennsylvania (N)	-0.0412
	(0.0205)		(0.0084)
Indiana (N)	-0.0492	Indiana (N)	-0.0448
	(0.0500)		(0.0114)

Note: Authors' calculations, U.S. Census data from 1980 and 1990, and Vital Statistics from 1980–1990. Mortality in each cohort is first calculated using GMM (as described in the text). Log of mortality is the dependent variable in a regression that includes cohort indicator variables, and State indicator variables. State effects, normed to average 0, and are reported here. Standard errors are in parentheses. $n = 150$ in each regression.

Table 4: Household Characteristics in 1940 for Children Born 1930–1939

A. Blacks

State	(1) Household Income	(2) Household Income per Capita	(3) Years of Schooling of Household Head
Alabama (S)	299	72.9	3.88
Arkansas (S)	216	74.9	4.84
Georgia (S)	316	65.8	3.59
Illinois (N)	771	181.3	7.21
Indiana (N)	752	136.3	7.13
Louisiana (S)	282	74.6	3.34
Mississippi (S)	163	59.1	4.41
New Jersey (N)	912	151.0	6.44
New York (N)	914	191.3	7.54
North Carolina (S)	365	84.0	4.32
Ohio (N)	773	154.1	7.22
Pennsylvania (N)	764	162.0	6.68
South Carolina (S)	257	61.3	3.82
Tennessee (S)	379	99.7	5.20
Virginia (S)	501	103.0	4.35

B. Whites

State	(1) Household Income	(2) Household Income per Capita	(3) Years of Schooling of Household Head
Alabama (S)	574	164.2	6.98
Arkansas (S)	404	126.7	7.19
Georgia (S)	678	188.6	7.06
Illinois (N)	1293	326.3	8.82
Indiana (N)	1052	272.3	8.99
Louisiana (S)	804	232.3	6.74
Mississippi (S)	503	150.8	8.09
New Jersey (N)	1652	371.5	8.46
New York (N)	1455	364.5	8.62
North Carolina (S)	735	185.1	6.91
Ohio (N)	1220	298.5	9.07
Pennsylvania (N)	1179	273.3	8.19
South Carolina (S)	772	204.6	7.30
Tennessee (S)	624	168.0	6.80
Virginia (S)	940	231.1	6.80

Note: Authors' calculations. Data from public use files of the 1940 U.S. Census.

Table 5: Relationship Between Mid-Life Mortality and 1940 Household Characteristics, Cohorts Born 1930–1939

A. Women

Variable	All			Blacks			Whites		
Log(HH Inc.)	-0.019			-0.025			-0.017		
	(0.0199)			(0.0186)			(0.0273)		
Log(HHI/Person)		-0.016			-0.047			-0.006	
		(0.0268)			(0.0308)			(0.0333)	
Education			-0.011			-0.015			-0.009
			(0.0078)			(0.0102)			(0.0094)
N	300	300	300	150	150	150	150	150	150

B. Men

Variable	All			Blacks			Whites		
Log(HH Inc.)	-0.171***			-0.083**			-0.201***		
	(0.0286)			(0.0338)			(0.0337)		
Log(HHI/Person)		-0.212***			-0.138***			-0.231***	
		(0.0317)			(0.0403)			(0.0368)	
Education			-0.076***			-0.043**			-0.088***
			(0.0088)			(0.0149)			(0.0084)
N	300	300	300	150	150	150	150	150	150

Note: Author's calculations. The dependent variable is the log of GMM mortality estimates, as described in the text. Explanatory variables are calculated using 1940 Census data for households with children born 1930–1939. Log(HH Inc.) is the natural logarithm of household income, Log(HHI/Person) is the log of household income per household member, and Education is years of schooling of household head. Each coefficient reflects a separate regression with indicator variables for year of birth included as covariates. Standard errors, given in parentheses, are clustered at the State of birth. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

References

Altonji, Joseph G., and Lewis M. Segal, 1996. "Small-Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business & Economic Statistics*, 14(3), 353-366.

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.

Bennett, Neil G., and Shiro Horiuchi, 1981. "Estimating the Completeness of Death Registration in a Closed Population," *Population Index,* 472(2), 207-221.

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, 1992. "Effects of Early-Life Conditions on Adult Mortality: A Review," *Population Index,* 58(2), 186-212.

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.

Haisken-DeNew, John P., and Christoph M. Schmidt, 1997. "Interindustry and Interregion Differentials: Mechanics and Interpretation," *The Review of Economics and Statistics,* 79(3), 516-521.

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

Hansen, Lars Peter, John Heaton, and Amir Yaron, 1996. "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business & Economic Statistics,* 14(3), 262-280.

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.

Imbens, Guido W., 2002. "Generalized Method of Moments and Empirical Likelihood," *Journal of Business & Economic Statistics (JBES Twentieth Anniversary Issue on the Generalized Method of Moments),* 20(4), 493-506.

Kovak, Brian, 2011. “Regional Labor Market Effects of Trade Policy: Evidence from Brazilian Liberalization,” Working Paper, Carnegie Mellon University.

Levine, Robert, James Foster, Robert Fullilove, Mindy Fullilove, Nathaniel Briggs, Pamela Hull, Baqar Husaini, 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.

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.

Preston, Samuel H., Irma T. Elo, Ira Rosenwaike, and Mark Hill, 1996. “African-American Mortality at Older Ages: Results of a Matching Study,” *Demography,* 33(2), 193-209.

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.