Population Research and Policy Review (2023) 42:36
https://doi.org/10.1007/s11113-023-09785-z

ORIGINAL RESEARCH

Mortality Modeling of Partially Observed Cohorts Using Administrative Death Records

Joshua R. Goldstein[1] · Maria Osborne[1] · Serge Atherwood[2] · Casey F. Breen[1]

Received: 2 September 2022 / Accepted: 6 March 2023

Abstract

New advances in data linkage provide mortality researchers with access to administrative datasets with millions of mortality records and rich demographic covariates. Although these new datasets allow for high-resolution mortality research, administrative mortality records often have technical limitations, such as limited mortality coverage windows and incomplete observation of survivors. We describe a method for fitting truncated distributions that can be used for estimating mortality differentials in administrative data. We apply this method to the CenSoc datasets, which link the United States 1940 Census records to Social Security administrative mortality records. Our approach may be useful in other contexts where administrative data on deaths are available. As a companion to the paper, we release the R package *gompertztrunc*, which implements the methods introduced in this paper.

Keywords Mortality estimation · Truncation · Statistical methods · Gompertz

Introduction

Researchers increasingly have access to large-scale individual-level administrative mortality records with rich demographic covariates (Bailey et al., 2022; Goldstein et al., 2021; Mehta et al., 2016). Despite the promise of these new administrative

✉ Casey F. Breen
caseybreen@berkeley.edu

Joshua R. Goldstein
josh.goldstein@berkeley.edu

Maria Osborne
mariaosborne@berkeley.edu

Serge Atherwood
satherwood@berkeley.edu

[1] Department of Demography, University of California, Berkeley, Berkeley, USA

[2] Philip R. Lee Institute for Health Policy Studies, University of San Francisco, San Francisco, USA

Published online: 26 April 2023

datasets, these records are often only available for individuals who have died, without information on survivors. This situation of having "deaths without denominators" makes it impossible to calculate occurrence–exposure mortality rates and to use the conventional tools of individual-level survival analysis (Alexander, 2018).

In this research, we present statistical methods for estimating mortality rates and differentials from a limited age window of death records. The situation we address is when we observe the distribution of age of death among members of a cohort for a left and right (henceforth, "doubly") truncated range of ages. For example, the Social Security Numident records publicly released by the United States National Archives contains nearly every record of individuals who died after age 65 from 1988 to 2005. For a very old cohort—e.g., those born in 1900—extinct cohort methods[1] might allow traditional analysis and the reconstruction of the population at risk at each age. However, for most cohorts, we see only a portion of those who die, making it difficult to know how many people with certain characteristics were alive at each age. Our approach is to take advantage of the distributional information about the limited ages we observe. Using maximum likelihood methods that include information on the ages at which each cohort is truncated allows us to infer mortality rates without observing the full population at risk. This approach also allows multivariate methods in which mortality rates depend on individual risk factors.

The method we propose contrasts with the approach that is sometimes taken of comparing average ages of death within the observed age windows, both directly and using multivariate regression (Lleras-Muney et al., 2020; Saavedra, 2013). Such an approach can be biased in the presence of double truncation. The direction and magnitude of this bias are data dependent (Greene, 2003), so there are not any guaranteed empirical regularities. However, the observed direction in most empirical studies is towards 0, meaning that estimates from regression on age of death of truncated samples will generally be attenuated (biased towards 0) relative to the untruncated case.

Our immediate application of our new method is to the CenSoc datasets, which link 1940 U.S. Census records on individual characteristics with Social Security death records (Goldstein et al., 2021). However, the problem addressed by these methods may be applicable to other sources of population-level administrative mortality records. These include the Chilean mortality registry release of a deaths-only dataset capturing deaths occurring from 2016 to 2021 (Chang et al., 2022; Chilean Ministry of Health, 2022), large-scale death registration data from genealogical databases (Kaplanis et al., 2018; Koylu et al., 2021; Otterstrom & Bunker, 2013), and centenarian databases such as the International Database on Longevity[2]. Our methods may also be relevant in non-human contexts, such as the captive-cohort studies by Carey and colleagues, as well as capture–recapture approaches (Carey &

[1] For extinct cohorts (in which all members have died), it is possible to calculate mortality rates using the classical "extinct generations" methods. First, the total number of survivors at a given age can be found by summing up all the deaths occurring above that age. Then, the age-specific mortality rates can be estimated using the age-specific ratios of deaths to survivors.

[2] These data are available here: www.supercentenarians.org (Belzile et al., 2021)

Roach, 2020). Further, this approach may be useful for researchers using survey data linked to mortality records in settings where differential match rates for population sub-groups affect estimates of mortality. Specifically, occurrence–exposure methods that depend on correct linkage of death records can incorrectly infer that a difficult-to-match person is a survivor, resulting in the so-called "Methusala effect" (Black et al., 2017).

The remainder of the paper proceeds as follows: in the "Background" section, we provide an overview of statistical methods for working with truncated data. The "Motivating Examples" section gives motivating examples for our new method, and the "Methods: Maximum Likelihood Estimation of Doubly Truncated Data" section introduces our Gompertz maximum likelihood estimation method. The "Results" section presents analyses demonstrating that our approach allows us to remove the bias introduced by truncation and produce estimates comparable to external estimates using untruncated data. We conclude in the "Discussion" section with a discussion of key considerations for applying this method and directions for future research.

Background

The statistical problem of making inferences with truncated[3] data has a long history in astronomy, economics, and survival analysis (Ying et al., 2020). A classic example from the demographic literature involves making inferences about changing population height (and health) from the heights of military recruits, who typically had to meet minimum height requirements and therefore under-counted short people (Wachter & Trussell, 1982). Another canonical example from survival analysis literature is retrospective autopsy-confirmed studies of Alzheimer's disease, wherein a subject is only observed if they die during the study's doubly truncated observation window (Rennert & Xie, 2018).

In the context of survival analysis, left truncation is not usually considered a problem, with standard methods accounting for occurrences and exposure in the observed period. The more problematic double truncation, in which the only individuals observed are those experiencing an event within a left and right truncated time window, has become the focus of a growing body of research. The first non-parametric maximum likelihood estimation methods (NPMLE) for doubly truncated

[3] For background, we distinguish between truncated and censored survival data (Mandel, 2007). For censored survival data, the process generating the data is such that for some observations we only know whether they occurred before or after a given time point. An example of a censored dataset is the National Health Interview Survey (NHIS) Linked Mortality Files, which tracks deaths for participants in the NHIS between 1997 and 2002; any participant who did not die between 2000 and 2005 is alive. Thus, there is a measure of survivorship (population denominator) and researchers can use conventional survival analysis techniques such as Cox-proportional hazard models. In contrast, the data generating process for truncated data is such that we only observe deaths that occurred between the left and right truncation bounds, with no information on survivors. An example of a doubly truncated dataset is CenSoc-Numident, which only includes those who died between 1988 and 2005, with no information on those who died outside of this window.

survival data were introduced by Efron and Petrosian (1999). Recently, Rennert and Xie Rennert & Xie (2018) introduced methods for a Cox regression model with doubly truncated data.

Regression of truncated normal data is a standard part of the econometrics toolkit (Greene, 2003). Regression estimates that do not account for truncation will be be biased, generally towards zero (Greene, 2003). Maximum likelihood estimates that account for truncation can remove this type of bias. The methods we present here can be seen as a way of adapting the parametric approach used in econometrics to mortality modeling. Like the econometric approach, we rely on parametric distributional assumptions and maximum likelihood estimation. But in our case, we use the Gompertz distribution rather than the normal distribution, in line with the almost universally observed tendency of mortality rates to rise exponentially with age (Gompertz, 1825). Another important difference is that the multivariate models we consider take the form seen in survival analysis, with covariates influencing the level of risk (usually proportionally), in contrast to regression models that predict the timing of an event.

Motivating Examples

The mortality differences observed between groups will tend to be smaller when there is truncation of youngest and oldest ages. We illustrate this tendency using two examples. The first, using simulated data, allows visualization of the effect of truncation in a simple regression framework. The second, using ages of death of men and women born in 1900 in Sweden, shows the effect of truncation when comparing two group means.

A Regression Example

The regressions on age of death for person i have the form

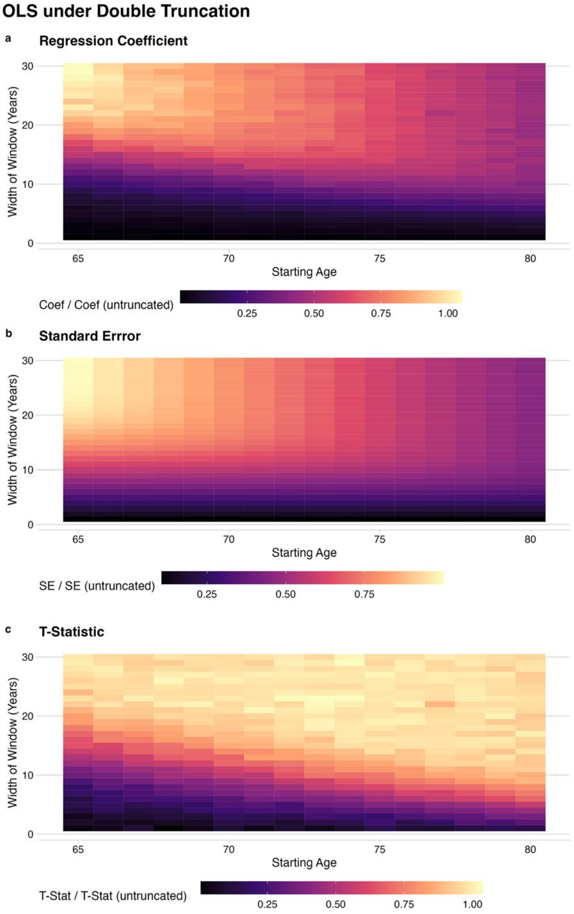
$$\text{age\_at\_ death}_i = \beta_0 + \gamma_t t_i + \boldsymbol{\beta} Z_i + \epsilon_i,$$

where β_0 is a general intercept, γ_t is the intercept for individuals born in year t, and $\boldsymbol{\beta}$ is the effect of a covariate Z_i on age of death. This regression accounts for the composition of birth cohorts, which is critical to include since people born earlier in time will be observed dying at older ages.

To illustrate the relationship between degree of truncation and estimated regression coefficients, we conduct a simulation study. Specifically, we simulate[4] a population where half of all individuals have high socioeconomic status and the other half have low socioeconomic status. We designed the simulation such that the high socioeconomic status individuals live on average 1 year longer than individuals with low

[4] We draw our simulated ages of death from a Gompertz distribution with parameters $b = 0.1$ and $a = 0.0001$.

Fig. 1 The ratio of truncated to untruncated regression statistics: standard errors (**a**), regression coefficients (**b**), and t-statistics (**c**) for different truncation windows. A value of 1 corresponds to the truncated and untruncated statistic having the same value; values less than 1 correspond to the truncated statistic being smaller than the untruncated statistic. (Color figure online)



socioeconomic status. We then simulate 480 different mortality datasets, systematically varying the width of the truncation window width (1–30 years, increments of 1 year) and the age at which the truncation window starts (ages 65–80, increments of 1 year). On each simulated dataset, we fit a regression model estimating the association between high socioeconomic status and longevity. Figure 1 plots the regression coefficients, standard errors, and t-statistics $\left(\frac{\beta_1}{\mathrm{SE}(\beta_1)}\right)$ for each dataset compared to the untruncated baseline. When the degree of truncation is higher (smallest window width), the coefficients, standard errors, and *t*-statistics are all attenuated. This simulation demonstrates that double truncation can substantially attenuate regression coefficients.

To more clearly illustrate why double truncation can substantially attenuate regression coefficients, the panels of Fig. 2 below show the effect of a covariate on age of death, using simulated data drawn from normal distributions. The left panel shows the complete sample along with the estimated "effect" of the covariate (0.45). The right panel shows what happens if observations were limited to

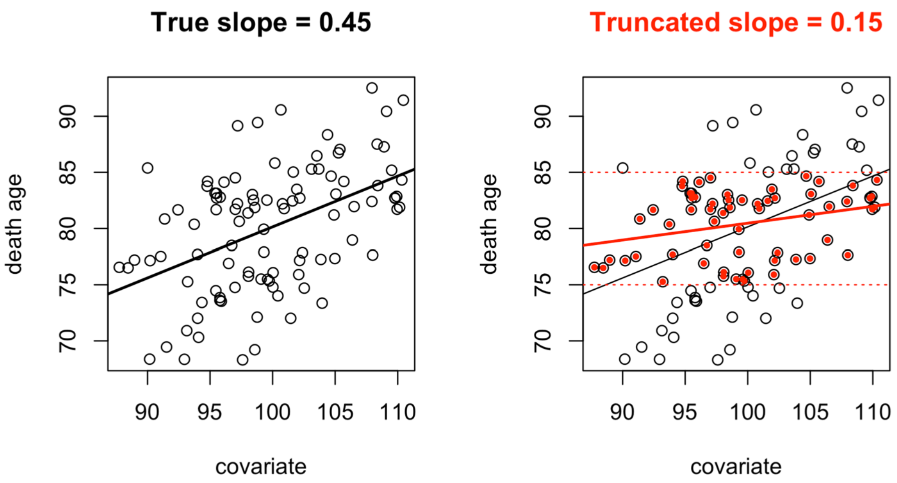


Fig. 2 Simulated example of regression on effect of covariate on death age based on full data (left) and truncated death ages only (right). The truncation reduces the estimated "effect" by about 2/3, from 0.45 to 0.15, in this example. Simulation is a sample of 100 individuals following $y = 30 + 0.5x + \epsilon$, with $x \sim N(\mu = 100, \sigma = 5)$ and $\epsilon \sim N(\mu = 0, \sigma = 5)$

ages 75 to 85. We can see that truncation removes the earliest and latest deaths, which also tend to be at the extremes of the covariate. The result is a flattening of the estimated slope to 0.15, only about one-third of its "true" value.

An Empirical Example

For our second example, we turn to real mortality data by comparing the ages of death of two groups, men and women born in 1900 in Sweden, based on cohort life tables in the Human Mortality Database (HMD). For this example, we focus on life expectancy at age 65, the average age of death conditional on survival to age 65 ($e(65) + 65$). According to HMD, the average age of death for Swedish women who died after their 65th birthday was 82.7 years, and the average age for men was 79.1 years, a difference of 3.6 years. The distribution of deaths over age 65 and the corresponding averages are shown in the top panel of the figure, along with the count of deaths from the HMD cohort life table.

Now we consider the case of what would happen if we only had access to death counts from 1980 to 1989 for these male and female cohorts. We would then have death counts only from ages 80 to 89, instead of all deaths above age 65. The middle panel highlights the counts of deaths in this narrower age window and shows the corresponding averages. Now, the longevity difference between women and men would appear to be only about 0.5 years, less than one-fifth of the difference in untruncated means. Truncation downwardly biases the observed differences in longevity.

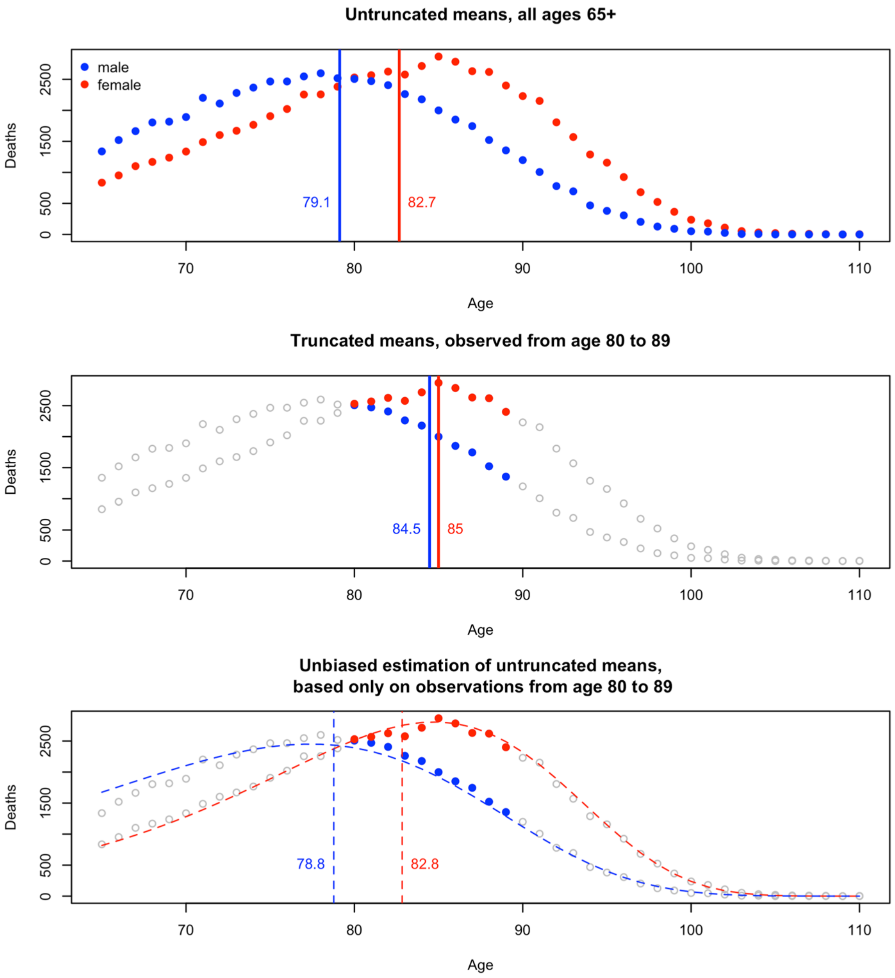


Fig. 3 Swedish deaths over 65 for male and female cohorts born in 1900. Data are from the Human Mortality Database. The top panel shows the full distribution and the mean ages of death. The middle panel shows an artificially truncated age window covering ages 80 through 89 and the mean ages within this window. The bottom panel shows the results of fitting a Gompertz curve to the truncated observations using maximum likelihood and the estimates of the implied untruncated means. The estimates are very close to the actual values, and the estimated magnitude of the sex difference in $e(65)$ is correct. (Color figure online)

Accounting for Truncation

It is possible to account for truncation explicitly using maximum likelihood estimation methods. Our approach, which we detail in the next section, allows a researcher to obtain estimates free of artificial downward bias, even when death counts are available for a limited range of ages. The results of applying this method (assuming Gompertz mortality) is shown in the lower panel of Fig. 3. We used the observed data from ages 80 to 89 to estimate the parameters of the Gompertz distribution. These estimated parameters were then used to describe the full distributions and their means. The dashed lines in the lower panel show the estimated full distribution (based only on the observations from ages 80 to 89) and the corresponding (now without bias) estimates of the means.

In this example, our estimate of the difference between life expectancy at age 65, based only on observations from ages 80 to 89, is about 4.0 years. This estimate is not exactly equal to the "true" value of 3.6 years, but is no longer subject to the strong downward bias from truncation shown in the middle panel, where the estimated difference in life expectancy is only 0.5 years.

In theory, the maximum likelihood procedure will give accurate estimates, as long as the distributional assumptions are correct. In practice, each estimate will be subject to the random effects of finite population size. Additional error is generated by departures from the underlying distribution that is assumed—in this case, the Gompertz distribution.

This example illustrates our two main points:

1. Estimation of group differences in average age of death will be downwardly biased, often greatly, by truncation.
2. Estimates that account for truncation can correct for this bias and provide a more accurate description of group differences.

These same points carry over from the comparison of the means of two groups to the more general case of multiple regression on age at death. The estimates of "effects" on age of death will generally be biased toward zero, and accounting for truncation using maximum likelihood aims to remove this bias.

Methods: Maximum Likelihood Estimation of Doubly Truncated Data

With parametric assumptions, maximum likelihood methods enable estimation of age-specific mortality with multivariate predictors, adapting the usual method of parametric survival analysis to our specific case of observing only those individuals who die. The likelihood associated with a set of observed ages of death x_i with parameters θ (e.g., the intercept and slope of the log-Gompertz curve, which may themselves be functions of covariates) is given by the product of the normalized densities, with truncation on the right at age x_i^r and on the left at age x_i^l:

$$L(\theta) = \prod_i L_i(\theta) = \prod_i \frac{f(x_i|\theta)}{F(x_i^r|\theta) - F(x_i^l|\theta)}, \tag{1}$$

where f is the density and F is the cumulative distribution.

For example, a proportional hazards model for the effect of covariates on mortality takes the following form. The hazard of an individual i aged x with covariates Z_i is given by

$$h_i(x|\beta) = h_0(x)e^{\beta Z_i}, \tag{2}$$

with a baseline hazard schedule over age x of $h_0(x)$. If the baseline is Gompertz,

$$h_i(x|\beta) = a_0 e^{b_0 x} e^{\beta Z_i}, \tag{3}$$

where a_0 and b_0 are baseline Gompertz parameters. Alternatively, we can write

$$h_i(x|\beta) = a_i e^{b_0 x}, \tag{4}$$

where a_i is the individual-level Gompertz intercept, modeled as $a_0 e^{(\beta Z_i)}$.

In this case, the observed data would contain for each person values x_i for the age of death, Z_i for covariates (e.g., years of education, place of birth), and the right and left truncation ages x_i^r and x_i^l for each cohort. The model estimates would be the parameter values $\widehat{a_0}, \widehat{b_0}$, and $\widehat{\boldsymbol{\beta}}$.

The truncated maximum likelihood approach works with a single or multiple cohorts. In the latter case, the truncation ages x_i^r and x_i^l can vary and be indexed for each individual according to the ages of observation for the cohort an individual belongs to. In the results presented here, we model common a_0 and b_0 Gompertz parameters across cohorts, with a common effects $\boldsymbol{\beta}$ also applying to all cohorts. However, it is also possible to allow the "Gompertz intercept" (a_0) to vary among cohorts, as well as the "Gompertz slope" (b_0) and even the effects ($\boldsymbol{\beta}$).

The MLE standard errors are found using conventional methods, calculating the square root of the diagonal elements of the inverse negative Hessian matrix (Greene, 2003). The standard errors of our estimated parameters generally increase with the degree of truncation, holding constant sample size. Intuitively, when we are only able to see observations from part of the underlying distribution, our estimates of the underlying distribution will be more uncertain. Compared to estimates of the truncated data from conventional methods such as OLS regression on age of death or Cox-Proportional Hazards models, our MLE estimates are generally larger in absolute value and have larger standard errors. For convenience, researchers may wish to use conventional methods in the exploratory research phase.

One of the crucial assumptions of our method is that the age distribution of deaths within the window of observation is proportional to the true underlying cohort life tables. With linked data such as CenSoc, this means that linkage rates may vary by group and by cohort, but within each group and cohort they should have no variability by age. Changes in the completeness of death registration over time can distort the age distribution of deaths observed for a cohort. Migration can also distort the age distribution of observed deaths. We do not address issues with the underlying data in this paper.[5] Instead, we focus on methods of estimation to be used with accurate, if limited, deaths by age.

[5] In the CenSoc data, we address death registration completeness by applying weights that reproduce the national number of deaths by age and year. When studying immigrants, we limit observations to those that we know were present in the United States at the beginning of our observation window. Similar approaches should be taken when working with other datasets.

As a companion to the article, we have developed a publicly available R package *gompertztrunc* that implements the methods described in this paper (Breen et al., 2022). The package can be installed from the Comprehensive R Archive Network (CRAN).

Results

To assess the accuracy of estimates from the truncated Gompertz approach, we analyze the relationship between education and old-age mortality. We use education as our substantive example because of the extensive literature investigating the relationship between education and longevity (Halpern-Manners et al., 2020; Lleras-Muney et al., 2020; Rogers et al., 2010). This analysis enables us to (1) assess the goodness-of-fit of the model to a single cohort, (2) show how our approach removes the bias in regression estimates that do not account for truncation, and (3) illustrate how the application of our method to truncated cohorts in the CenSoc data, for which only deaths are available, produces estimates that are comparable to other sources using full occurrence–exposure estimates.

A Single Cohort Example, with Graphical Diagnostics

We first apply our truncated Gompertz model to estimate the association between high school completion and longevity in the CenSoc-DMF dataset. The CenSoc-DMF file links the 1940 Census to mortality records from the Death Master File (DMF), which has nearly complete death coverage from 1975 to 2005 (see Online Appendix Sect. B for details on CenSoc datasets). We restrict our analysis to the birth cohort of 1910 because focusing on a single cohort allows for a straightforward graphical assessment of the goodness-of-fit of our model.

To assess model fit, we compare our empirical and model-based distributions of ages of death. Figure 4 shows a histogram of the number of deaths for this cohort for those who completed high school (panel a) and those who did not (panel b). The solid lines plot the corresponding Gompertz-model-based estimates of the distribution of deaths within the cohort. To construct these figures, we first plot the empirical distribution of deaths for the cohort of 1910, disaggregated by high school completion status. Next, we overlay onto each histogram our truncated Gompertz-model-based estimate of the distribution of deaths, disaggregated by high school status. We calculate our model-based estimates of the distribution of deaths separately by high school completion status. Specifically, we convert the estimated hazards (h_x) into the estimated number of deaths (d_x) using life table relationships. We then scale the number of model-estimated deaths to exactly match the number of empirical deaths within our observed truncated window.

Panel (c) shows the log hazard rate for the inferred observed (dashed lines) and the modeled (solid lines) log hazard rates. Since true hazard rates are not computable due to the lack of population denominators, we use the results of the model to estimate rates. The modeled Gompertz hazard schedule is used to estimate

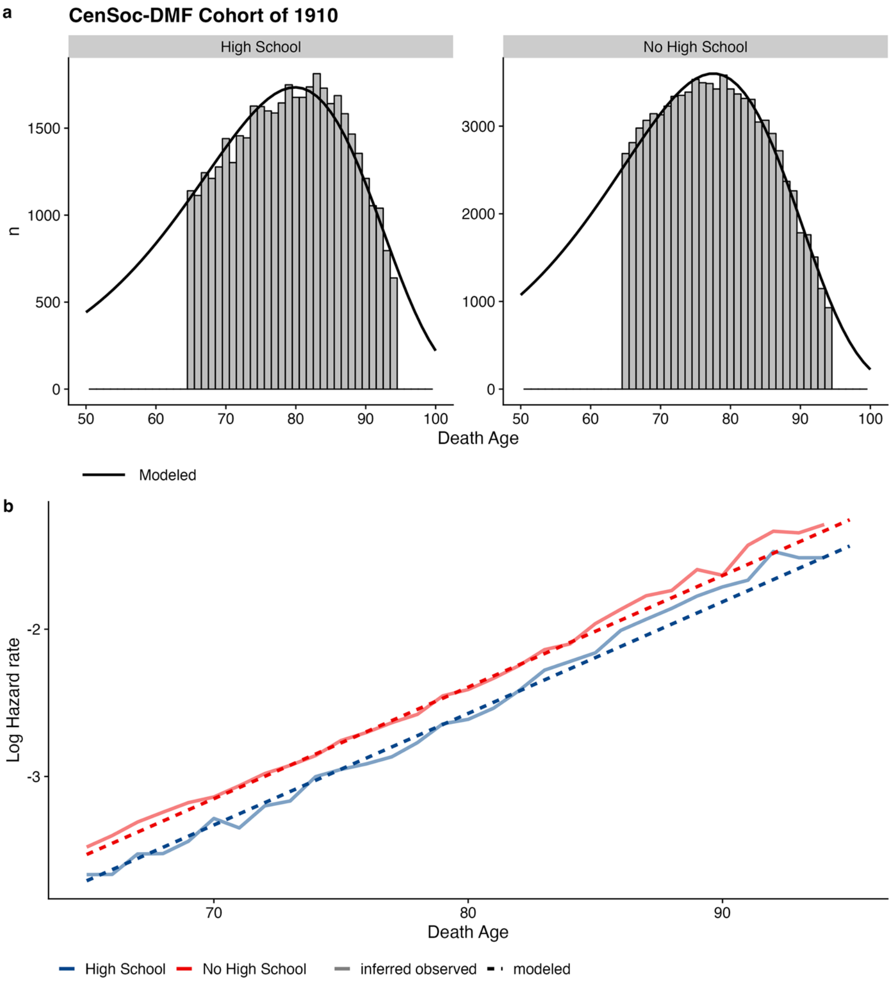


Fig. 4 For CenSoc-DMF cohort of 1910: **a** shows a histogram of age of death for those who did and did not complete high school; the black curves show our Gompertz-based model estimates of the number of deaths. **b** plots the model-based hazard ratio estimates (solid lines) vs. the inferred observed hazard ratios (dashed lines) for those who completed high school (blue) and those who did not (red). (Color figure online)

survivorship to age 66 (or the lowest fully observable age of death) for each level of a covariate. Then, real death counts are used to compute hazard rates for each subsequent observable age. In this application, we see that the inferred observed hazards closely track the modeled hazards and are approximately proportional.

These graphical diagnostic checks allow researchers to check whether two key modeling assumptions are met: that the distribution of deaths within a cohort follows a Gompertz distribution, and coefficients have proportional effects on hazards

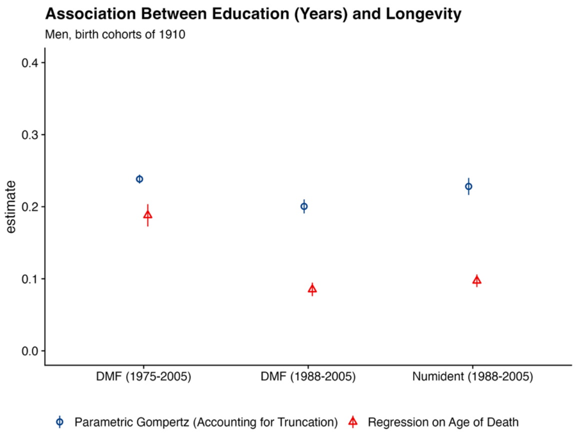


Fig. 5 Association between education (years) and longevity, using two methods. Comparison of regression on age of death (red) to the Gompertz maximum likelihood approach (blue). Results demonstrate downwardly biased coefficients for regression on age of death because truncation is not accounted for. For a more direct comparison between our Gompertz model results and estimates from linear regression on age of death, we converted our hazard ratios into estimates of additional life expectancy at age 65. See Online Appendix Sect. A.1 for details on this conversion. (Color figure online)

at all ages. The *gompertztrunc* R package includes functions to implement these diagnostic checks.

Evidence of Reduced Bias

To demonstrate that our truncated Gompertz model produces estimates with reduced bias, we fit our model to three datasets varying in their degree of truncation. Our truncated Gompertz model produces consistent estimates of the association between education and longevity, while conventional methods produce estimates downwardly biased by the truncation.

Specifically, we fit models using ordinary least squares (OLS) regression on age of death[6] and the Gompertz MLE approach described by Eqs. (1) and (2) to two different CenSoc datasets (Goldstein et al., 2021). The first is based on death records in the Death Master File (DMF), and contains nearly complete coverage of men dying at age 65 and older from 1975 to 2005. The second is based on more recently released Numident data provided through the National Archives (Breen & Goldstein, 2022). Numident data include more detailed information but span a more limited high coverage period for men over age 65: 1988 to 2005. The two datasets

[6] We use OLS regression on age of death to illustrate how conventional methods perform in the presence of double truncation, because OLS regression has been applied to analyze CenSoc data (Fletcher & Noghanibehambari, 2021; Lleras-Muney et al., 2020). We note that survival methods such as Cox proportional hazards regression models are also generally biased in the presence of double truncation (Rennert & Xie, 2018).

overlap to a large degree, since most deaths in the Numident also are captured in the Death Master File, so we would expect similar estimates from both datasets.

Figure 5 compares the potentially biased results of regression on age at death to the MLE approach that accounts for truncation. In this example, we analyze cohorts born from 1905 to 1914. This minimizes truncation in the distribution of men dying at age 65 and older captured in the DMF, which includes records from 1975 to 2005. For the birth cohort of 1910, the DMF includes records from ages 65 to 95. In contrast, the age window covered by the Numident, which includes records from 1988 to 2005, is narrower, ages 78 to 95.

The estimates shown in red are from the OLS regression on age at death. Because they do not account for truncation, the estimates based on a narrower age range are biased downwards. Whereas the estimate from the wide-coverage DMF is about 0.2 years of increased longevity at age 65, the estimate from Numident is about half this size, or about 0.1. To check if truncation differences are responsible for the difference in regression estimates, we artificially truncated the DMF data to the same year coverage as the Numident. Comparing the estimates in red in the second column (from the DMF truncated to 1988 to 2005) and third column (from Numident for the same years), we see that they now give similar estimates of about 0.1.

The blue estimates are from the Gompertz proportional hazards model. To allow comparison with the regression model, we have re-calculated the estimated proportional hazards effects in terms of the difference in remaining life expectancy at age 65. Now the estimates from our Gompertz proportional hazards models for all three sources are highly comparable, with overlapping uncertainty intervals.

Our conclusions from these results are that, for the CenSoc datasets,

1. Regression on age of death produces downwardly biased longevity effects over age 65 with truncated data.
2. The magnitude of this bias depends on the degree of truncation.
3. It is possible to obtain less biased estimates using maximum likelihood approaches that explicitly account for truncation. We obtain similar estimates for the less severely truncated data (DMF, 1975–2005), the narrower coverage data (Numident, 1988–2005), and the artificially truncated data (DMF, 1988–2005).

Comparison with External Estimates

The above analysis demonstrates that estimation that accounts for truncation can create comparable estimates of the educational gradient across CenSoc datasets of varying source and truncation degree. One can also compare these estimates to those from other studies estimating the association between education and longevity. Rogers et al. (2010) published estimates of U.S. adult mortality risk by educational degree, including post-secondary degrees. Recently, Halpern-Manners et al.

Table 1 Estimates of the association between an additional year of schooling and mortality for men from four studies

Study	Dataset	Birth cohorts	Death window	Method	Hazard ratio	e(35)
Lleras-Muney et al. (2020)	CenSoc-Numident	1906–1915	1988–2005†	OLS	–	0.099 (0.089, 0.108)
Lleras-Muney et al. (2020)	Census Tree	1906–1915	1941–2019	OLS	–	0.404 (0.355, 0.452
Halpern-Manners et al. (2020)	SSDMF-NUMIDENT Linked to 1940 Census	1910–1920	1960–2013	OLS	–	0.387 (0.379, 0.395)
Rogers et al. (2010)*	National health interview survey (NHIS) linked mortality files	1908–1928	1997–2002	Cox PH Model	0.945	0.544
This study (2022)	CenSoc-DMF	1910–1920	1975–2005†	Gompertztrunc	0.964 (0.964, 0.965)	0.434 (0.424, 0.444)
This study (2022)	CenSoc-Numident	1910–1920	1988–2005†	Gompertztrunc	0.960 (0.958, 0.961)	0.451 (0.438, 0.464)

Despite substantial differences in study design—each study uses different birth cohorts, methods, and mortality data—estimates generally align closely across studies. Notably, the one exception is the Lleras-Muney et al. (2020) estimate from OLS regression on age of death using the highly truncated CenSoc-Numident dataset. This estimate is highly attenuated (biased towards 0). The "OLS" method corresponds to Ordinary Least Squares (OLS) regression on age of death,"Cox PH Model" corresponds to the Cox proportional hazards model, and "Gompertztrunc" corresponds to the method introduced in this paper. We convert hazard ratios to estimates of $e(35)$ assuming deaths follow a Gompertz distribution with parameters $a = 3.34 \times 10^{-}5$ and $b = 0.1$. (See Online Appendix Sect. A.1 for details)

*We calculate the effect of an additional year of education from the Rogers et al. (2010) study by assuming those in the "less than 12 years of education" had on average 8 years of education. We assume those in the "BA" category had 16 years of education, 8 years more than those in the "less than 12 year of education" category. We then calculate the annualized hazard rate as: $hr_{(annualized)} = \left(hr_{(\mathrm{ba})}/hr_{(<12\ \mathrm{years})}\right)^{1/8} = (0.748/1.178)^{(1/8)} = 0.9448$. The Halpern-Manners et al. (2020) estimate includes controls for occupation category, family size, U.S.-born parents, dummies for age at 1940 census, and race. The Lleras-Muney et al. (2020) estimates are from a regression including birth year and state of birth fixed effects.

†Doubly truncated dataset

(2020) used an internal version of the Numident and DMF linked to the 1940 Census to estimate association between education and longevity. Furthermore, Lleras-Muney et al. (2020) estimated the educational gradient across states using both the untruncated Census Tree dataset[7] and the truncated CenSoc-Numident dataset, finding the double truncation of the CenSoc-Numident downwardly biased estimates of the education gradient by a factor of nearly 4.

In Table 1, we compare our estimates of the association between an additional year of education and longevity to those from the three studies. Despite substantial difference in study design—each study uses different birth cohorts, methods, and mortality data—the estimates generally align closely. However, there is one notable exception: the estimate using conventional linear regression on age of death with the highly truncated CenSoc-Numident dataset is highly attenuated, only 25% of the full estimate. This example demonstrates that using this method can explicitly account for double truncation and produce estimates comparable to those made from untruncated data.

Required Width of Truncation Window

The Gompertz proportional hazards model introduced in this paper requires a sufficiently wide window of deaths to produce reliable estimates. To illustrate, we constructed subsamples of the CenSoc-Numident cohort of 1910 with differently truncated windows of observed deaths by systematically varying the start and end ages of our observed window. For each subsample, we used the truncated Gompertz model to estimate the hazard ratio for the association between an additional year of education and longevity.

Figure 6 displays the estimated hazard ratios from this exercise. Generally, when the truncation window becomes too narrow, our model estimates become unreliable. For example, when we consider estimates from truncated windows less than five years in length, the estimated hazard ratios are close to 1, erroneously indicating that education has almost no association with longevity. Windows that do not contain the empirical modal age at death are generally less reliable. However, estimates from truncated windows 10+ years in length are highly consistent. Broadly, we recommend that caution be exercised if the mortality window is fewer than 12 years in length. Relatedly, when combining multiple cohorts for a common estimate, care should be taken to exclude cohorts with severe truncation, as this risks downwardly biasing estimates even within the MLE framework.

[7] The Census Tree data matches the 1940 Census to vital statistics from the FamilySearch database, a genealogy platform with over 12.6 million registered users.

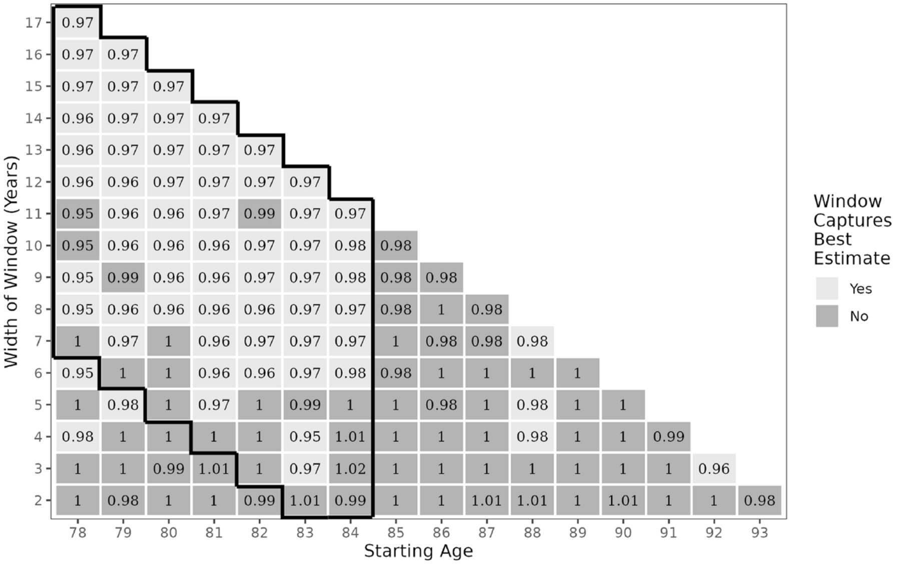


Fig. 6 Acceptably wide observation windows. Values are estimated hazard ratios of the education gradient for men born in 1910, using CenSoc-Numident data, with different combinations of truncation limits. The best estimate of the hazard ratio is 0.97, which is computed using the widest observation window possible (17 years). Light gray tiles are observation windows where the model captures the best estimate of the hazard ratio in a 95% confidence interval. Estimates inside the black outline indicate that the truncation window contains the modal age at death, which is 84 years in these data. This experiment indicates that truncation windows of approximately 10–12 years or more usually produce reliable estimates. Windows as narrow as 8–9 years may also be sufficient for estimation if the data contain the mode of the empirical mortality distribution

Discussion

In this paper, we provide a new method for estimating age-specific mortality rates from counts of deaths for partially observed (truncated) cohorts. Our principal goal was to estimate mortality rates from newly released administrative data records, such as the death records from the U.S. Social Security Administration. However, we anticipate this method will be applicable to other sources, such as oldest-old mortality studies where only deaths are counted. Our approach may also be used as an alternative to occurrence–exposure methods when there are concerns about numerator–denominator mismatch.

We propose a maximum likelihood approach that explicitly accounts for age truncation at lower and higher ages and estimates the truncation-adjusted parameters of the Gompertz mortality model. Our approach can be used for a single cohort or multiple cohorts combined. In this paper, we model the effect of covariates in a proportional hazards framework, but this model may be extended to include variation in the Gompertz slope of mortality by age. We fit these models using standard numerical optimization, but Bayesian approaches may also be used, particularly to add multi-level structure to the model.

We test our method by estimating the effect of educational attainment on male old-age mortality. First, we find that, in this setting, the proportional hazards' Gompertz model fits well, using graphical displays to show the linearity and proportionality of death rates broken down by educational attainment. Second, in the presence of double truncation, our method produces larger estimates (less attenuation bias) of the association between education and mortality than does the existing approach of regressing on age at death. Even when the window of observed ages changes, our method produces comparable results. Finally, we find that our estimates of the magnitude of the effect of education on mortality are comparable with published estimates using other data sources that do not suffer from truncation.

Together, these results give us confidence that the truncated Gompertz model can be applied to other topics that influence mortality, including a wide range of socioeconomic covariates. Topics that can be studied with the CenSoc data include the effects of wages and occupation, family structure, immigration by country of origin, geographic variation, and any other question for which census data can be of use. As a complement to the paper, we introduce a publicly available R package *gompertztrunc* implementing the methods described in this paper.

Despite the promise of this approach, there are several considerations and limitations for researchers implementing this method that warrant discussion. First of all, a number of computational issues can arise when maximizing the likelihood. These include dependency on starting values, particularly for complex models, and are especially pronounced for smaller samples or higher degrees of truncation. We note that a sufficiently wide window of deaths is required to produce reliable estimates. If the degree of truncation is high (window of death coverage is less than 10 to 12 years), we recommend researchers exercise caution in applying this method. Sensitivity to starting values and instability with smaller sample sizes are problems general to maximum likelihood estimation.[8]

In general, the computational demands of the current approach are intensive, and we run into additional computational challenges when introducing models with a large number of parameters. Large-scale fixed-effect models, for example of siblings or small geographic areas, are not possible with our current computational approach.

Our proposed method may also not be appropriate in certain contexts. We emphasize that the method is only as good as the data: if the distribution of deaths in the population is not well represented by the sample of available data, then estimates for mortality can be mistaken with patterns of undercounts. Researchers should take care to validate the mortality coverage of their data and consider using weights to account for potential disparities in coverage by age and period. High net migration occurring in the ages of observation can also distort cohort death distributions, and

[8] It is common practice to use OLS estimates to set starting values; by default, the *gompertztrunc* R packages uses estimates from linear regression converted to hazards but also allows users to customize starting values. Since regression estimates can be be biased downwards in the presence of truncation, researchers may also want to scale up starting values determined by OLS. For example, simulations imply that OLS estimates of remaining years of life for CenSoc-Numident data are usually deflated by a factor of 2–3. (However, note that OLS estimates and hazard ratios do not scale linearly (Pang and Hanley, 2021); see Online Appendix Sect. A.1 for details on converting between the two).

ideally only immigrants present at the beginning of the truncation window should be included in studies. As discussed previously, the method is less reliable if there are very few years of data available, especially if observed data are suspected to exclude the peak of the mortality distribution.

Our parametric method relies on the Gompertz model, which is well known to fit well for adult populations in their 70s and 80s. At older and younger ages, however, the Gompertz model may not necessarily be ideal, and it may become important to include a Makeham term at younger ages or a tapering “logistic” term at older ages. We also assume a common Gompertz slope parameter and proportional hazards. Further, the lack of population denominators makes it impossible to assess proportionality with common methods such as plotting Kaplan-Meier curves. Researchers can, however, use the graphical diagnostic tools in the *gompertztrunc* R package to roughly gauge proportionality from inferred observed hazard rates, as in Fig. 4.

As a final consideration, researchers studying change over time from one birth cohort to the next should be cautious in interpreting their results. Observed cohort trends could result from real changes in effects over time, age effects introduced by observation of different cohorts in different age ranges, variation in departures from model assumptions, or a combination of these. Researchers should verify that the individual cohort data are representative of the population distributions and not subject to extreme degrees of truncation. If there are sufficient data to do so, researchers may want to restrict all cohorts of interest to a common range of ages at death.

Many of these limitations represent promising extensions to our parametric Gompertz model and avenues for future research. We are investigating techniques for incorporating large numbers of fixed effects (e.g., sibling fixed effects). The functional form of the covariates (and of the baseline hazard) is flexible. We use the Gompertz form here with the proportional hazards assumption (estimating one common Gompertz slope b_0 for all cohorts and sub-groups). But it is possible to fit hierarchical models that allow for arbitrary structure in how parameters co-vary. For example, one might want to allow the Gompertz slope to vary (slowly) over time and differ between sub-groups. It is also possible to use other parametric forms besides the Gompertz, incorporating mortality deceleration at very old ages, or a Makeham constant. It is also possible, as Alexander (2018) has shown, to estimate the principal components of known age schedules (e.g., from HMD) and penalize estimates that differ from these schedules, as described by linear combinations of these components.

The truncated Gompertz MLE approach proposed here offers a method of mortality estimation for situations where population denominators are either unavailable or unreliable. While most literature on truncated data focuses on one-sided truncation, our method offers a solution for circumstances where data are doubly truncated, especially where doubly truncated survival analysis methods are impossible due to lack of information on the population at risk. Our parametric Gompertz method is straightforward and appropriate in contexts that are well described by Gompertz’s law. Our method helps ameliorate bias introduced by double truncation, allowing researchers to make estimates directly comparable to those from conventional methods applied to untruncated data. We anticipate the increasing utility of this method in the growing landscape of administrative death records that often lack a measure of the population exposed to risk.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11113-023-09785-z.

Acknowledgements For helpful feedback and background, the authors would like to thank Monica Alexander, Leslie Root, Nathan Seltzer, Jordan Weiss, members of the CenSoc Working Group, and participants in the 2023 Population Association of America session "Methodological Innovations in Mortality." The authors acknowledge financial support from NIA Grant R01AG058940. Replication code is available on the Open Science Framework: https://doi.org/10.17605/OSF.IO/D6QHF.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Alexander, M. (2018). Deaths without denominators: Using a matched dataset to study mortality patterns in the United States. Preprint SocArXiv.

Bailey, M. J., Lin, P. Z., Shariq Mohammed, A. R., Mohnen, P., Murray, J., Zhang, M., & Prettyman, A. (2022). *LIFE-M: The longitudinal, intergenerational family electronic micro-database*.

Belzile, L. R., Davison, A. C., Rootzén, H., & Zholud, D. (2021). Human mortality at extreme age. *Royal Society Open Science, 8*(9), 202097.

Black, D. A., Hsu, Y.-C., Sanders, S. G., Schofield, L. S., & Taylor, L. J. (2017). The Methuselah effect: The pernicious impact of unreported deaths on old-age mortality estimates. *Demography, 54*(6), 2001–2024.

Breen, C., & Goldstein, J. R. (2022). Berkeley unified numident mortality database: Public administrative records for individual-level mortality research. *Demographic Research, 47*, 111–142.

Breen, C., Osborne, M., & Goldstein, J. R. (2022). *Gompertztrunc: Conducting maximum likelihood estimation with truncated mortality data*.

Carey, J. R., & Roach, D. A. (2020). *Biodemography: An introduction to concepts and methods*. Berlin: Princeton University Press.

Chang, S.-W., Zitko, P., Passi-Solar, Á., Danaei, G., & Margozzini, P. (2022). Changes in cardiovascular mortality in Chile during the COVID-19 pandemic. *Heart, 108*, 1716.

Chilean Ministry of Health, Department of Statistics and Health Information. (2022). *Mortality records*. Retrieved from https://deis.minsal.cl/#datosabiertos

Efron, B., & Petrosian, V. (1999). Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association, 94*(447), 824–834.

Fletcher, J., & Noghanibehambari, H. (2021). *The effects of education on mortality: Evidence using college expansions*. Berlin: National Bureau of Economic Research.

Goldstein, J. R., Alexander, M. B., Casey, F., Miranda-González, A., Menares, F., Osborne, M., & Yildirim, U. (2021). *CenSoc mortality file: Version 20*. Berlin: University of California.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London, 115*, 513.

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Berlin: Prentice Hall.

Halpern-Manners, A., Helgertz, J., Warren, J. R., & Roberts, E. (2020). The effects of education on mortality: Evidence from linked U.S. census and administrative mortality data. *Demography, 57*(4), 1513–1541.

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, M., Bhatia, G., MacArthur, D. G., Price, A. L., & Erlich, Y. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science (New York), 360*(6385), 171–175.

Koylu, C., Guo, D., Huang, Y., Kasakoff, A., & Grieve, J. (2021). Connecting family trees to construct a population-scale and longitudinal geo-social network for the U.S. *International Journal of Geographical Information Science, 35*(12), 2380–2423.

Lleras-Muney, A., Price, J., & Yue, D. (2022). The association between educational attainment and longevity using individual level data from the 1940 census. *Journal of Health Economics, 84*, 102649.

Mandel, M. (2007). Censoring and truncation-highlighting the differences. *The American Statistician, 61*(4), 321–324.

Mehta, N. K., Elo, I. T., Engelman, M., Lauderdale, D. S., & Kestenbaum, B. M. (2016). Life expectancy among U.S.-born and foreign-born older adults in the United States: Estimates from linked social security and medicare data. *Demography, 53*(4), 1109–1134.

Otterstrom, S. M., & Bunker, B. E. (2013). Genealogy, migration, and the intertwined geographies of personal pasts. *Annals of the Association of American Geographers, 103*(3), 544–569.

Pang, M., & Hanley, J. A. (2021). Translating all-cause mortality rate ratios or hazard ratios to age-, longevity-, and probability-based measures. *American Journal of Epidemiology, 190*(12), 2664–2670.

Rennert, L., & Xie, S. X. (2018). Cox regression model with doubly truncated data: Cox regression model with doubly truncated data. *Biometrics, 74*(2), 725–733.

Rogers, R. G., Everett, B. G., Zajacova, A., & Hummer, R. A. (2010). Educational degrees and adult mortality risk in the United States. *Biodemography and Social Biology, 56*(1), 80–99.

Saavedra, M. H. (2013). *Early childhood conditions and mortality: Evidence from Japanese American internment*.

Wachter, K. W., & Trussell, J. (1982). Estimating historical heights. *Journal of the American Statistical Association, 77*(378), 279–293.

Ying, Z., Wen, Yu., Zhao, Z., & Zheng, M. (2020). Regression analysis of doubly truncated data. *Journal of the American Statistical Association, 115*(530), 810–821.