Late Cohort Analysis with CenSoc Data: Potential Problems and Considerations for Researchers

Maria Osborne[†]
December 16, 2023

Summary

In this brief report, we explore the consequences of conducting mortality analysis using CenSoc data from later birth cohorts (approximately 1930 onwards). For late cohorts, deaths below age 65 are commonly observable within the CenSoc mortality coverage windows of 1975-2005 (CenSoc-DMF dataset) and 1988-2005 (BUNMD and CenSoc-Numident datasets). We show that coverage of these deaths at ages below 65 is deficient in Social Security Administration mortality data. Consequently, the population distribution of deaths at younger ages is not well represented and can lead to problems with mortality analyses. We recommend researchers use cohorts with ample mortality coverage at ages 65+, limit analyses to ages 65+, and create weights for decedents of younger ages if they are included.

1. Overview

CenSoc datasets link records in the 1940 decennial Census to Social Security Administration (SSA) morality records. The CenSoc-Numident links the census to the Berkeley Unified Numident Mortality Database (BUNMD) and covers deaths from 1988-2005, while the CenSoc-DMF links the census to the Death Master File and covers the period 1975-2005. As shown in Figure 1 and Figure 2, these datasets include persons born from the late 19th century through March of 1940. Peak birth cohort coverage is reached at about 1915-1920 for the CenSoc-Numident and 1910-1920 for the CenSoc-DMF. Coverage rapidly increases for the CenSoc-Numident from around 1910-1915, as one the variables used for matching records (place of birth) becomes much more common in SSA records among these cohorts. CenSoc-DMF coverage is better at earlier cohorts than the CenSoc-Numident because the earlier start date of the window of deaths (1975 vs. 1988) increases our ability to capture those born earlier. Additionally, the CenSoc-DMF does not use place of birth to match census records with mortality records and so match rates are not dependent on that variable's availability.

[†] Department of Demography, University of California, Berkeley. Please contact censoc@berkeley.edu for questions.

In general, we recommend that CenSoc data users include cohorts in the range of 1900-1925 in their analyses, depending on the dataset and research question. These cohorts tend to maximize sample size, have the fullest mortality coverage at older ages, and allow for observation of adult characteristics such as income and completed education in the 1940 Census. However, it is possible to use cohorts up to and including 1940. In this brief report, we explore some of the potential analytical issues that can arise from using "later" cohorts (defined as cohorts after about 1930) that researchers should consider.

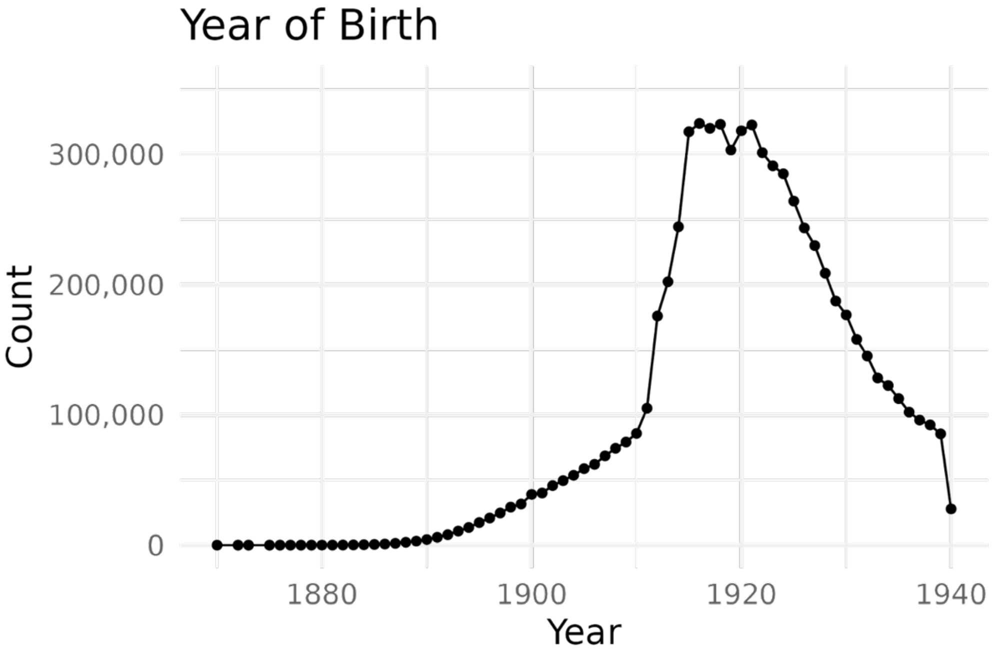

Figure 1: Number of records per birth year in the CenSoc-Numident.

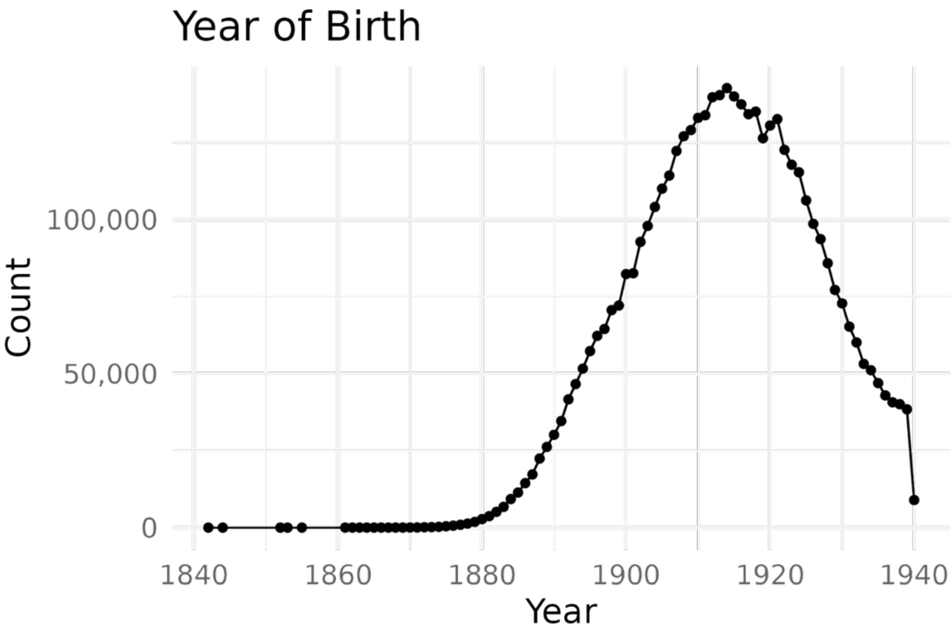

Figure 2: Count of records per birth year in the CenSoc-DMF.

2. Death Age Coverage in Social Security Records and Implications for Late Cohorts

Because deaths are observed within a set window of time, each cohort is observed at a different range of ages. The later the cohort, the earlier we can observe them in Social Security mortality records. The CenSoc-Numident, for example, includes deaths from 1988-2005. For the cohort of 1910, therefore, the lowest observable death age is 77 years. For the cohort of 1935, however, the lowest age is only 52. This is important because public SSA mortality records are less likely to capture deaths at younger ages.

Specifically, death coverage is best at age 65 and above in our data. We can see in Figure 3 that compared to gold-standard Human Mortality Database[‡] data, BUNMD and Social Security DMF death coverage is nearly complete for 65–100-year-olds in the high-coverage windows for each dataset. (The BUNMD and Social Security DMF are the mortality data used to create the CenSoc-Numident and CenSoc-DMF linked datasets, respectively.) Mortality coverage worsens for younger age groups in all years. While there are many deaths below age 65 included in public SSA data, the chance of a such deaths being captured is lower than deaths at older ages.

[‡] Available from: https://www.mortality.org

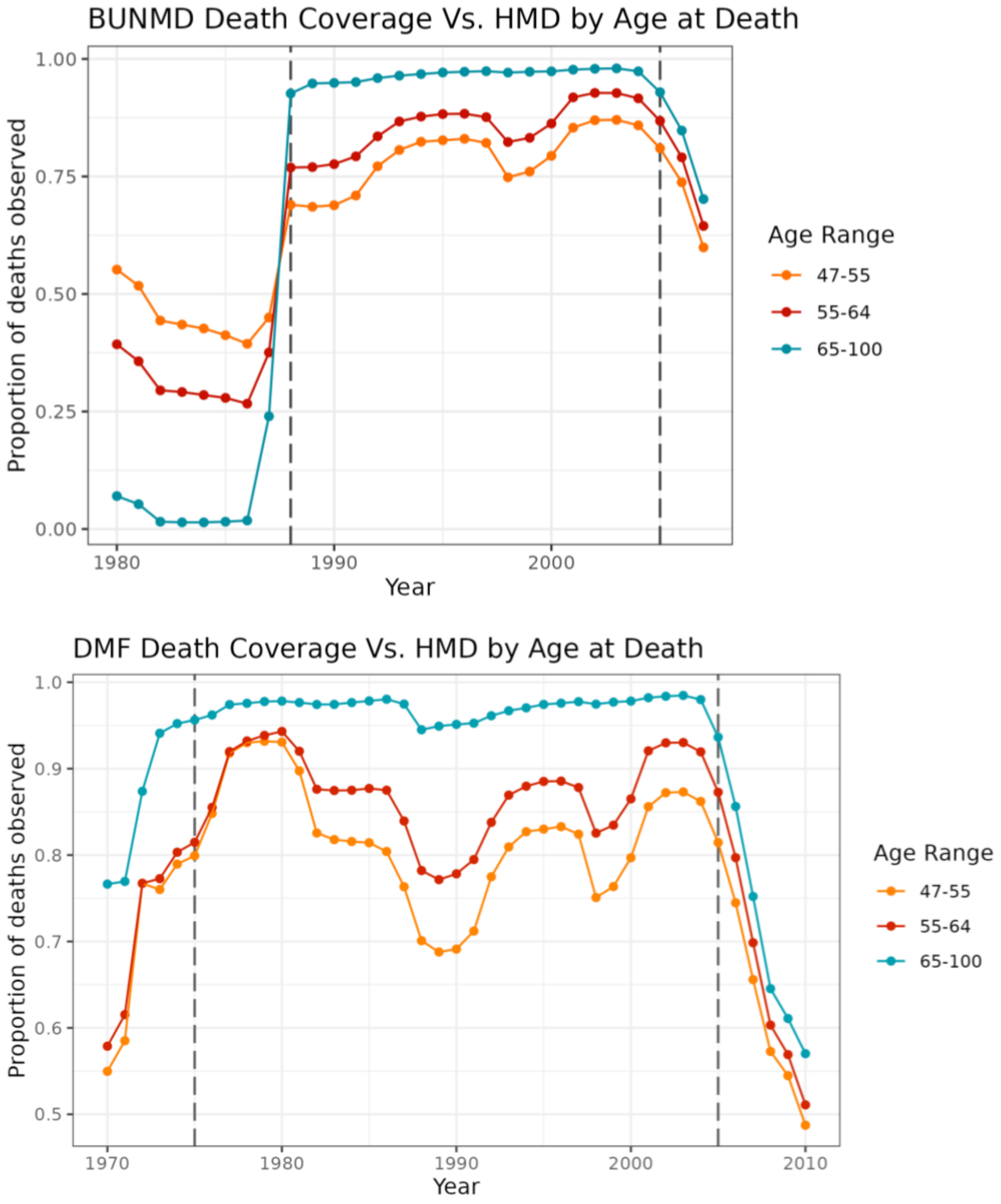

Figure 3: BUNMD vs. Human Mortality Database (HMD) coverage (top) and Social Security DMF vs. HMD coverage (bottom). The high coverage windows for each dataset are marked by dashed lines. Within the high coverage time window, the ratio of 65–100-year-old deaths BUNMD and DMF compared to the HMD are close to 1, indicating that the BUNMD and Social Security DMF capture most deaths occurring in the United States at those ages. The proportion of HMD deaths observed worsens with younger age groups.

The full explanation for incomplete mortality coverage below 65 is not known, other than the possibility that the SSA may be less likely to interact with for younger persons. 65 is the age at which most people born before 1940 became eligible for full Social Security retirement benefits. Thus, Americans have a vested interest in applying for Social Security before this age, and the SSA has a vested interest in recording deaths above age 65 in order to prevent fraud. However, the full explanation for age-specific differences in coverage in publicly released data is largely opaque. This has multiple consequences for analyzing cohorts observable at ages below 65. First, CenSoc-observable deaths at younger ages may not be representative of the population dying at these ages in unknown and unobservable ways. Additionally, differential coverage at different ages means that the distribution of ages at death for younger cohorts could be inaccurately represented in CenSoc data, which is particularly problematic in the context of some parametric modeling strategies.

3. Linear Regression Example

In Figure 4 below, we run the same linear model for individual cohorts 1915-1939 in the CenSoc-Numident. For each cohort, we estimate the effect on longevity in years of being male relative to being female. We choose this example because the results are simple to present, and general sex differences in mortality are substantial and well-documented. Only ages 65 and above are included, so this effect is contingent on survival to either age 65 or the lowest observable age for each cohort. Our results show that, for example, males born in 1919 live about one year less on average than females. For this cohort, this difference in age at death is contingent to survival to age 69, the lowest observable age of death within the window of 1988-2005. In actuality, the difference in life expectancy at age 69 for men and women of this cohort may be greater, as results from linear regression are likely attenuated due to truncation. See Goldstein et al. (2023) for a more thorough discussion of this phenomenon.

This truncation becomes increasingly severe for later cohorts. For cohorts after 1933, the effect shrinks to zero or nearly zero. Very few ages at death are observed at this point. For the cohort of 1939, which did not turn 65 years old until 2004, the only death ages observed at 65+ in the window of 1988-2005 are 65 and 66. Thus, very little variation in age at death is observable for that cohort, and any extant differences in older-age longevity might be difficult to capture.

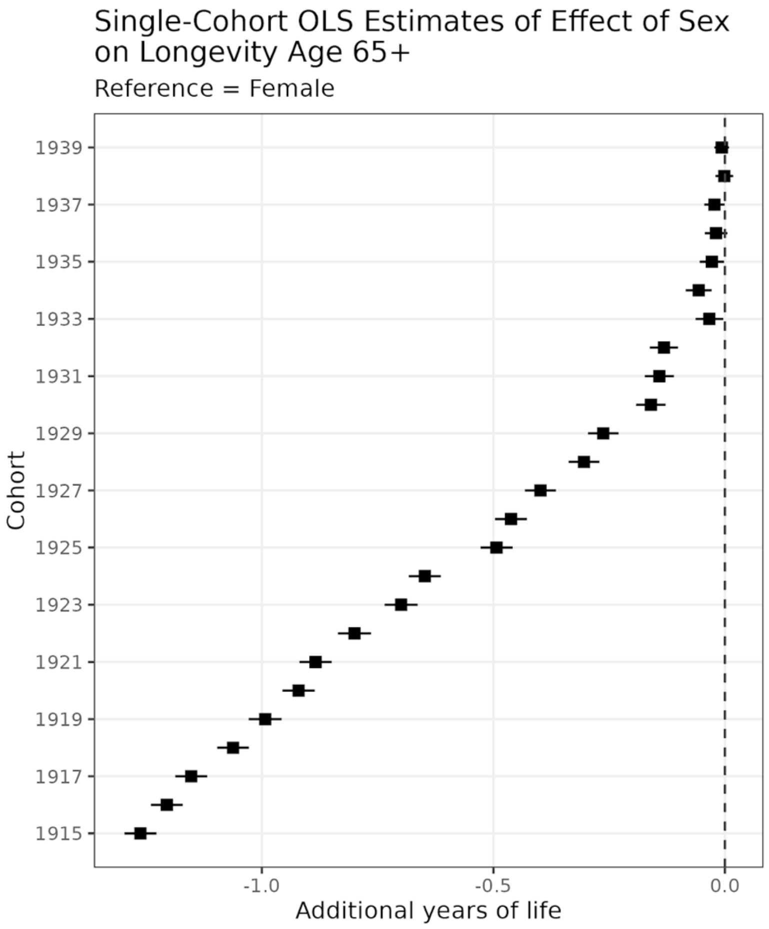

Figure 4: Single-cohort OLS models of the effect of maleness on longevity, CenSoc-Numident data. Only ages 65+ are included in these models, and the observed affect is dependent upon survival to either 65 or the lower observable age for each cohort. For later cohorts, effect sizes attenuate towards zero.

4. Gompertz Parametric Example

The lack of deaths at age 65 and above for later cohorts is also problematic in a parametric framework. In Figure 5, we fit the age-specific death counts for select cohorts to a Gompertz curve using maximum likelihood estimation. For cohorts up to about 1930, the parameterization works well. However, lack of information for later cohorts makes the model fit clearly less reliable after this point. For the cohort of 1935, for example, the Gompertz model estimates a slope parameter β of 0.18, an extremely high value for human populations, and shows age-specific death counts dropping to near zero past age 80[§].

[§] A reasonable value for β would be around 0.1. See page 1041-1042 of Missov et al. (2015) for estimated values of β (denoted by b in this paper) in human populations using Human Mortality Database data. For male populations, β usually falls between 0.07 and 0.011.

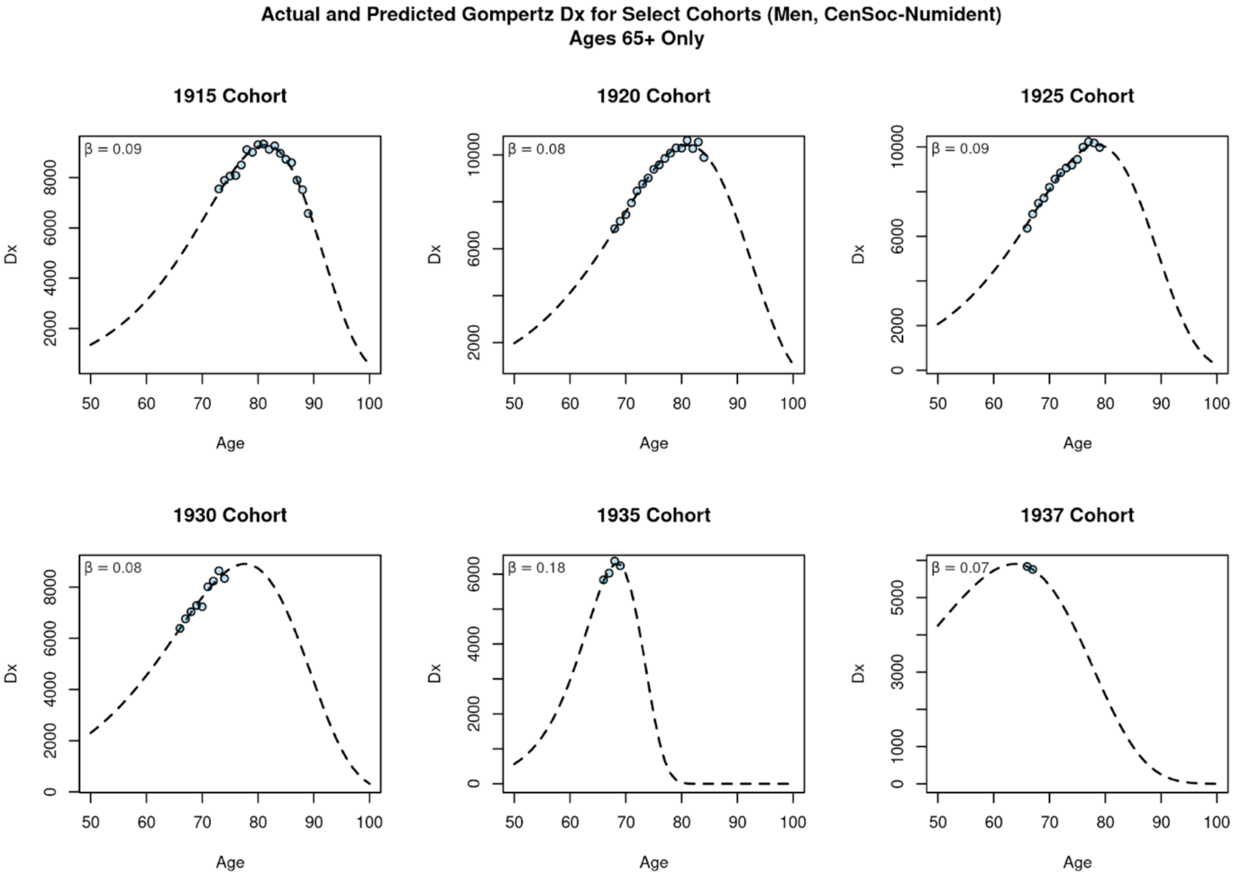

Figure 5: Observed death (blue circles) vs. modeled (dashed lines) for men in individual cohorts. Modeled death counts were calculated using Gompertzian maximum likelihood estimation. The Gompertz "slope" parameter β is shown in the upper left corner of each plot. For cohorts later than about 1930, the model fit clearly becomes less resemblant of a human Gompertz distribution as the data truncation becomes severe.

4. Including Deaths Below Age 65

In the previous examples, we showed that estimation for later cohorts is difficult when limited to ages 65+. However, deaths at younger ages are published in CenSoc datasets and available to be used by researchers. Generally, we do not advise using these younger ages, for the reasons explored in Section 2. Here, we will consider the impact of including these younger deaths in analyses.

In Figure 6 below, we again fit Gompertz distributions to individual cohorts, but this time include all observable death ages. We can see parameterizations for the cohorts of c. 1930 and later, which include several ages of death below 65, are again not ideal. For the cohorts of 1930, 1935, and 1940, β is estimated to be very high, at 0.17-0.19. Estimated death counts fall off extremely sharply after peaking, leading to almost no deaths by the mid 80's. Again, these are unlikely features for a human population. This parameterization could result from a few factors: first, death coverage in our data increases as age of death increases from 47-65, leading to an artificially steep gradient in death counts among younger ages. Second, estimation of the Gompertz curve could be less reliable when data are far away from the empirical modal age of death. Comparing the 1930 cohort in Figure 5 and Figure 6, we note how inclusion of deaths below 65 dramatically affects the model fit, even in this cohort where most observable deaths are at age 65+.

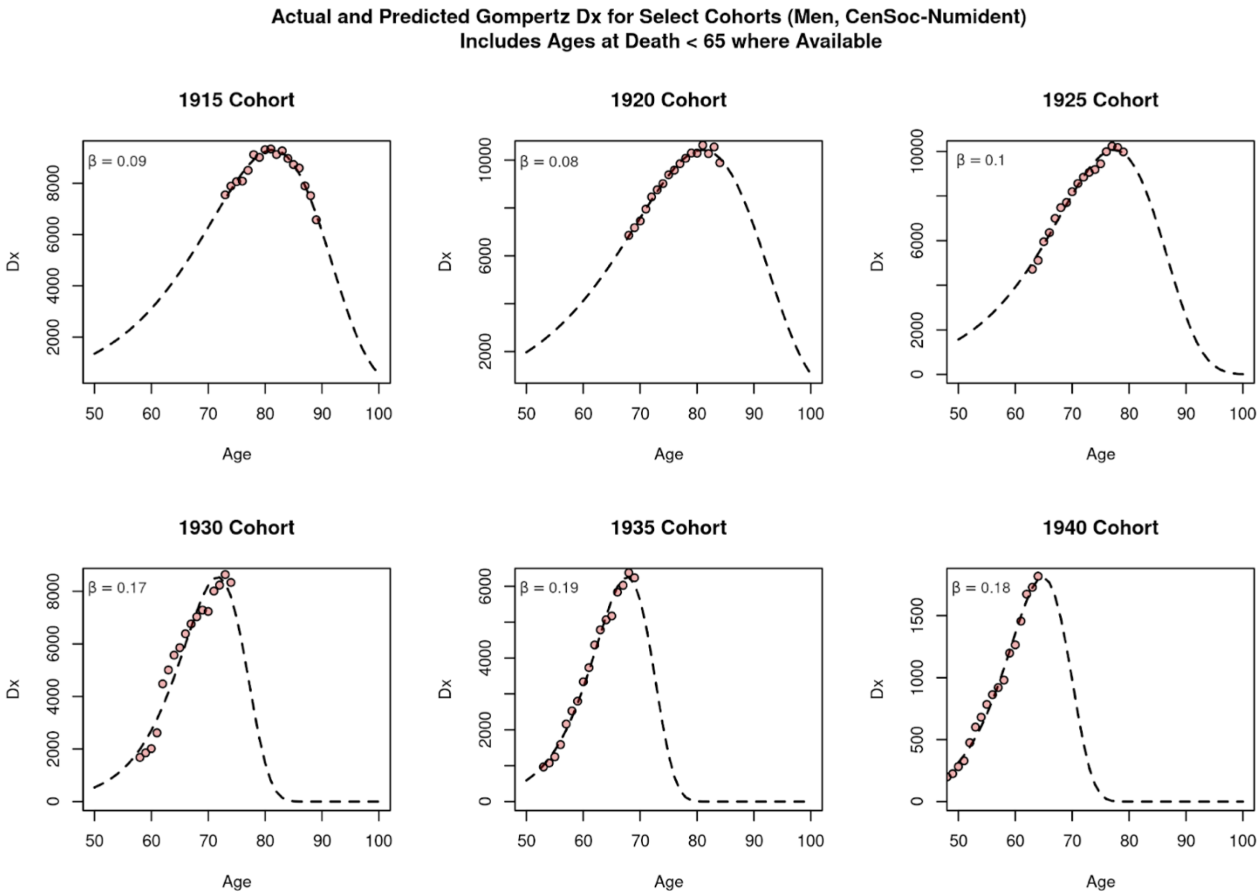

Figure 6: Observed deaths (red circles) vs. modeled (dashed lines) for men in individual cohorts when all observable ages at death are included. Modeled death counts were calculated using Gompertzian maximum likelihood estimation. The Gompertz "slope" parameter β is shown in the upper left corner of each plot. For later cohorts, inclusion of younger deaths causes the model to estimate very high β values and low modal ages at death. Inclusion of deaths below age 65 affects the model fit even in cohorts like 1930, where most observable deaths occur at 65 or older. This may be because deaths under 65 are relatively undercounted, leading to an unrealistically steep gradient in death counts at younger ages.

The fact that representation of deaths below age 65 is comparably poor is most consequential in this parametric context, which depends on an accurate representation of the population's age distribution of deaths. The implications of including younger ages at death in other frameworks, such as linear regression, are less clear. While we know that deaths below age 65 are less likely to be included in Social Security mortality data, we do not know why certain deaths are not included. Social Security deaths at younger ages could be representative of the population dying, or they could differ from the population dying in important ways. Ultimately, the data generation process is not clear.

In Figure 7, we compare estimates of the effect of sex on longevity for individual cohorts using ages 65+, to estimates using all observable ages. (Before 1923, only ages 65+ are observable in the window 1988-2005 so unweighted estimates are equivalent.) Estimates are contingent on survival until the lowest observable age at death, which differs for each cohort.

In this example, estimates for later cohorts are no longer attenuated towards zero. It is not known to what extent the effects obtained for later cohorts are influenced by possible non-representiveness at younger ages. Within a single-cohort framework, one should also consider the highest observable age in our interpretation of results. The cohort of 1939, for example, is observed between ages 48 and 66. Our estimate for this particular cohort, therefore, is conditional on dying in this relatively young age range.

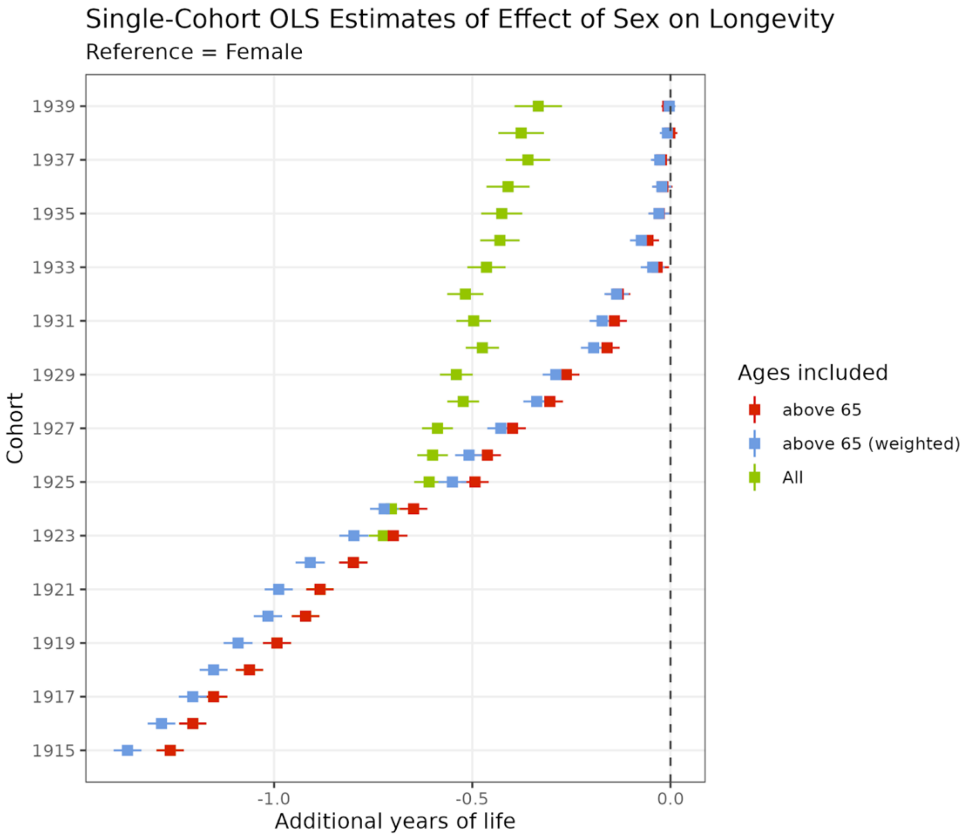

Figure 7: OLS estimates of the effect of sex on longevity for single cohorts. The red points are the same estimates shown in Figure 4 and only include deaths at ages 65 and older. The blue points show weighted estimates for deaths age 65+, and are very similar to the unweighted estimates. In green we show the (unweighted) estimates when all observable ages at death are included in each model. Unlike models that only included ages 65+, these models do not attenuate to zero among younger cohorts.

4. Conclusions and Recommendations

CenSoc mortality coverage is best at age 65+, as deaths at these ages are more likely to be recorded by the Social Security Administration and/or included in publicly released versions of SSA mortality data. The reasons for this are not fully known. Researchers should be aware that cohorts younger than 1922 (CenSoc-Numident) or 1909 (CenSoc-DMF) contain ages below 65, and that death data at these ages may or may not be representative of the population dying. Further, undercounts of deaths at younger ages compared to older ages may misrepresent the true distribution of deaths in the population. We recommend the following:

1. Choose cohorts with significant mortality data at ages 65+ available, and limit analyses to ages 65+. Recommend cohort ranges include approximately 1910-1925 for the CenSoc-

Numident/BUNMD and 1900-1925 for the CenSoc-DMF. Use the person-weights published in CenSoc datasets, which are available for deaths aged 65-100.

2. If deliberately using late cohorts or younger ages at death, consider creating weights for these data. This is particularly important if using a parametric methodology that relies on estimating an accurate distribution of death ages, such as the truncated Gompertz MLE method utilized here. Using version 2 files of the BUNMD, CenSoc-Numident, and CenSoc-DMF is recommended if computing weights for ages outside the 65-100 range. These datasets include weights by age/cohort/sex using Human Mortality Database data, which are freely available to the public. Researchers may download HMD data from https://www.mortality.org and use the code on the CenSoc development GitHub to construct weights for any ages and cohorts of their choosing.
3. Regardless of cohorts used, researchers should be aware of how **truncation** (the fact that CenSoc deaths fall within a set time window, which creates cohort-specific age windows) shape their results. Whether the sample is a single cohort or a group of cohorts, researchers should find the minimal and maximal age at death represented. If using a method like linear regression on age at death, results are contingent upon survival to a lowest observable age and/or death before a highest observable age.
4. In general, results from linear regression will be more highly attenuated as the data become more highly truncated. (For a fuller discussion of this phenomenon, see Goldstein et al. (2023).) In the context of late cohort data, this is particularly relevant because such cohorts are severely truncated for ages 65+. For any statistically significant coefficient measurable with late cohort data, the "actual" effect (the effect we would measure in absence of truncation) could be many times greater in magnitude. If that statistically significant coefficient is a false positive, however, researchers could mistakenly report large but specious effects.