Mortality Modeling of Partially Observed Cohorts Using Administrative Death Records

Joshua R. Goldstein Casey Breen Maria Osborn
Serge Atherwood Leslie Root
Nathan Seltzer Jordan Weiss

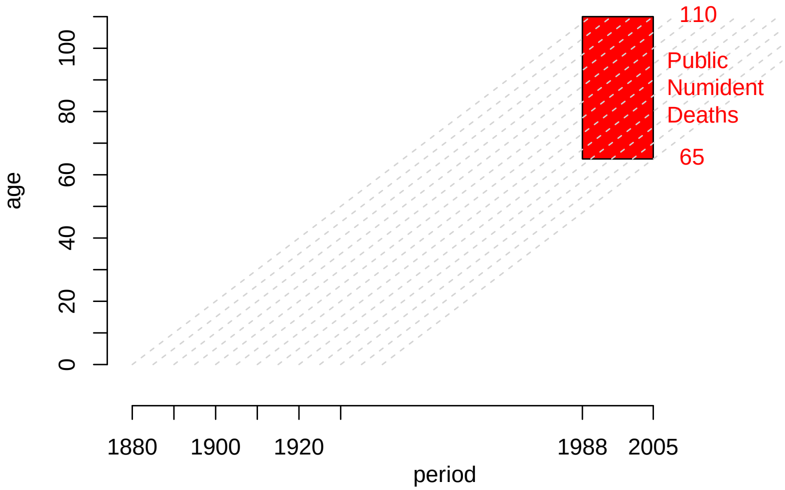
UC Berkeley, Department of Demography

Presentation for PAA 2022,
"Methodological Innovations in Mortality"
April 7, 2022

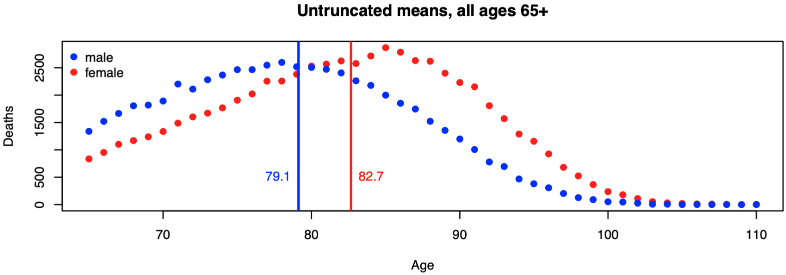
Our problem

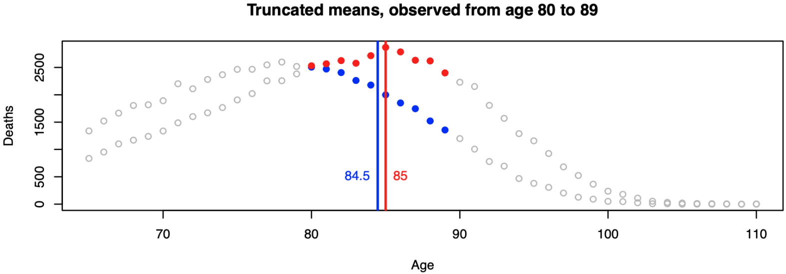
- Administrative data often complete only for limited years
- So we see only limited variation in ages at death
- This truncation biases all “effects’ (education, weather, income, neighborhood, race, ...) toward zero
- Our goal is to remove this bias and get “true” effect

Our Truncation Problem: Public-released Social Security Numident Data

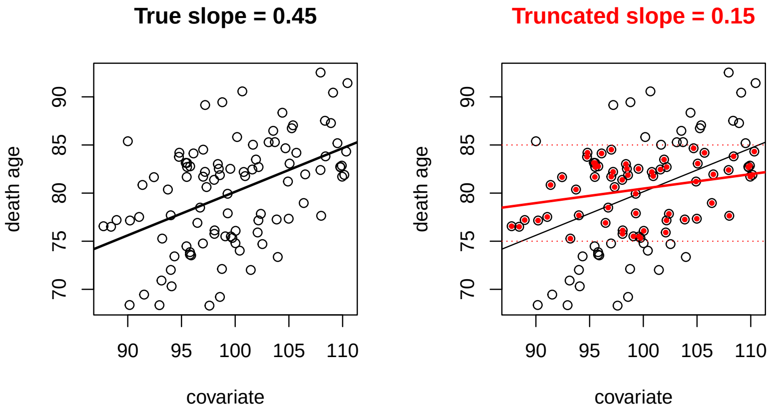


Artificial example of truncation bias (Sweden 1900 cohort – Human Mortality Database)





Regression example, simulation with covariate



Truncation removes lowest lows and highest highs, which tend to be at extremes of covariate, flattening the estimated slope.

Our Solution: Maximum Likelihood Accounting for Truncation

- Assume Gompertz mortality, with proportional effects (other choices allowed)
- Conditional probability of observed death age, given truncation

$$f(x \mid L < x < R) = \frac{f(x)}{F(R) - F(L)}$$

- Maximize the likelihood

$$\mathsf{L} = \prod_i \frac{f(x_i)}{F(R_i) - F(L_i)}$$

Note: direct output is proportional effect on hazard, but can transform to $e(65)$

Application to education, cohort of 1915

In CenSoc Numident, we have complete deaths for 1988-2005, ages 73-90. Can artificially make window narrower and see what happens

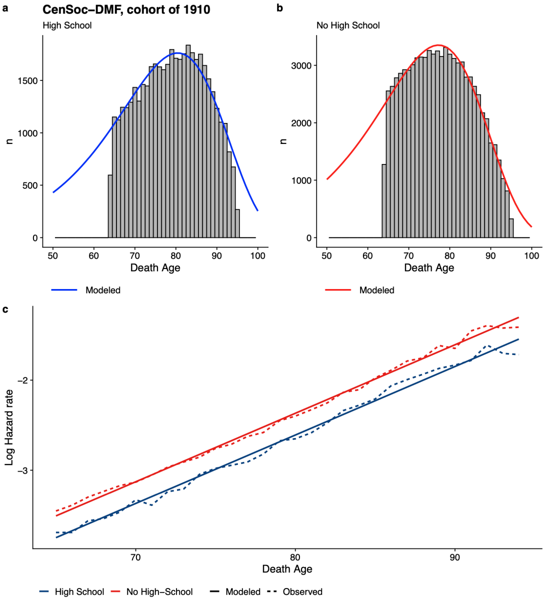
```
Marginal effect of 1-year of education on e(65)
```

	regression	Our method
1988-2005	0.17	0.63
1991-2002	0.09	0.58

Two important results:

1. Truncation has huge downward bias (estimates only 1/4 the size)
2. Differential truncation influences regression, but not our method

Graphical diagnostics



Applications

Method motivated by

- CenSoc (www.censoc.org) linked data sets to 1940 Census
- Berkeley Unified Numident Mortality Database

(These very large individual-level data sets allow analysis of mortality based on any individual or small area characteristics in 1940 census)

But other potential applications

- Matched data, false survivorship, and "Methusela Effect"
- Deaths-only data from medfly experiments and biodemography
- Other deaths-only data (genealogies, death indices etc)

Conclusions

- Partially observed cohort mortality (without denominators) can still produce good estimates
- Don't just regress on age-at-death, or you'll be biased.
- Easy to take truncation into account using maximum likelihood
- Good results (comparable to estimates in literature)
- `gompertztrunc` package for R