An Empirical Assessment of Parametric Survival Functions for Modeling Longevity

Dennis Feehan
dfeehan@princeton.edu

DRAFT: March 1, 2011

Abstract

In this study, I present an analysis of the best available data on mortality at advanced ages with the aim of updating and expanding our understanding of which of the functional forms commonly used to parameterize survival fit high-quality data on death rates over the age of 80 the best. This has been discussed at length in the literature, and is of importance to researchers and policymakers in a host of fields. I expand on the results of previous analyses by adding new data, considering more functional forms, and improving the strategy for assessing goodness of fit. I find that, for death rates at ages 88-99, the logistic functional form commonly recommended in the literature produces moderately good fits, but that the quadratic model enjoys the most robust empirical support.

Draft: results for poster at PAA, March 31 2011.
Please do not cite

Introduction

In the developed world, life expectancy at birth continues to rise, fertility levels are low, and populations are aging at remarkable rates. (Kinsella, Phillips and Bureau, 2005; Martin and Preston, 1994). This has dramatic implications for the needs societies will face in the future. In order to fully contend with these implications, scholars and policymakers require a quantitative understanding of death rates above age 80; that is, they need a mathematical function that summarizes the age pattern of death rates at advanced ages[1] with a few parameters. This is important for several reasons. First, the continuing debate about the limits, or lack of limits, to human longevity requires a solid understanding of the empirical evidence on death rates at advanced ages to formulate and evaluate theories that describe the process of aging (Gavrilov and Gavrilova, 2009; Horiuchi and Wilmoth, 1998; Thatcher, Kannisto and Vaupel, 1998). Many of the existing theories imply that some functional forms should reproduce observed death rates at advanced ages better than others. Second, many policymakers and planners need accurate, mathematical summaries of death rates at advanced ages in order to produce forecasts and projections (Bongaarts, 2005; Tabeau, van den Berg Jeths and Heathcote, 2001). Here, a low-dimensional (few parameter) summary of death rates at advanced ages is useful because it permits projections or forecasts to focus on fewer parameters. Finally, an understanding of old-age survival patterns is important for a number of other research questions of great relevance to sociology, economics, and public policy. One example is the relationship between improvements in old-age survival and changes in savings and investment behavior; understanding how savings behavior changes as a function of improvements in survival past retirement age has dramatic implications for public policy (Sheshinski, 2009). In order to build models of behavioral responses to improvements in survival at older ages, a mathematical function that captures the essential dynamics of those changes is required.

Many functional forms have been proposed to capture the shape of mortality at advanced ages (Tabeau, van den Berg Jeths and Heathcote, 2001; Thatcher, Kannisto and Vaupel, 1998). In this article, I present an analysis of the best available data on mortality above age 80, with the aim of updating and expanding our understanding of which of the functional forms commonly used to parameterize survival fit high-quality data on death rates over the age of 80 the best. Over a decade ago, Thatcher, Kannisto and Vaupel (1998) studied

[1] In this paper, I use the term 'advanced ages' to refer to ages over 80.

the fit of several functional forms to a unique and carefully constructed database of deaths at advanced ages. My analysis is similar in spirit to that exercise, with a few important differences: first, I make full use of the considerable volume of data that have become available since 1990. I also benefit from a recent, thorough review of the data's quality (Jdanov et al., 2008), which allows me to focus on the country-years where quality is least likely to be a problem. The quantity of high-quality data available to today is therefore considerably greater than was available a decade ago. I also expand the focus of Thatcher, Kannisto and Vaupel (1998) to several new functional forms that have appeared in the literature since their study.

Finally, I make some modifications to the methods applied in Thatcher, Kannisto and Vaupel (1998); in particular, I propose the use of more principled criteria for assessing goodness-of-fit across the functional forms considered, which include an appropriate penalty for the increased complexity of functional forms with more parameters. This is very important, as it provides us with some guard against over-fitting idiosyncracies in the data that might result from small sample sizes at advanced ages.

Methods

Data

I use the Kannisto-Thatcher (K-T) database on Old Age Mortality, which is arguably the best collection of data on mortality at advanced ages available (Kannisto and Thatcher, 2009). The K-T database contains data on deaths and exposure above age 80 for 35 countries, with some of the Scandinavian data going back to the mid-18th century. However, data quality for mortality at advanced ages is a serious concern; see Jdanov et al. (2008) for a thorough summary of the problems involved in analyzing them. In order to ensure that poor-quality data do not mislead us in selecting useful functional forms, I only study country-years of data that were found to be of acceptable quality by Jdanov et al. (2008). Specifically, I retained the data from 1970-2000 from all countries where more than half of the years were of the highest quality, and the remaining years were of the second-highest quality in the assessment the authors provided[2]. This leaves us with data from Belgium, the

[2]For Poland, I start at 1971.

Czech Republic, Denmark, France, West Germany, Italy, Japan, the Netherlands, Poland, Scotland, Sweden, and Switzerland, making a total of 371 country-years of data for each sex and single year of age. For the analysis of cohort death rates, I study the birth cohorts from as early as possible to 1900 for the same countries, for a total of 381 country-cohort years of data for each sex and single year. Table 1 describes the years and cohorts in our dataset for each country.

	period start	period end	cohort start	cohort end
Belgium	1970	2000	1864	1900
Czech Republic	1970	2000	1864	1900
Denmark	1970	2000	1864	1900
France	1970	2000	1865	1900
W. Germany	1970	2000	1875	1900
Italy	1970	2000	1874	1900
Japan	1970	2000	1869	1900
Netherlands	1970	2000	1869	1900
Poland	1971	2000	1890	1900
Scotland	1970	2000	1869	1900
Sweden	1970	2000	1864	1900
Switzerland	1970	2000	1864	1900

Table 1: Country-years and country-cohorts from the Kannisto-Thatcher Database on Old Age Mortality (Kannisto and Thatcher, 2009) used in this analysis. I chose all country-years that Jdanov et al. (2008) found to be acceptable in their systematic review; that is, I choose all country-years between 1970 and 2000 for countries where more than half of the years were of highest quality, and the remaining years were of the second-highest quality in the assessment the authors provided. Cohort data are from all possible years between 1864 and 1900 for the same countries. For both cohort and period analyses, I model ages 80-99.

Functional Forms

We usually obtain data on mortality in the form of central death rates,

$$m(z) = \frac{D_z}{N_z},$$

where D_z is the number of deaths between ages z and $z+1$ in the time period being considered, and N_z is the number of person-years of exposure in the time period being considered.

name	parameters	function	sample reference
constant hazard	α	$\mu(z) = \alpha$	
Gompertz	a, b	$\mu(z) = a\exp(bz)$	
Kannisto	a, b	$\mu(z) = \frac{a\exp(bz)}{1+\exp(bz)}$	
Weibull	a, b	$\mu(z) = az^b$	
ETA	α, T	$\mu(z) = \frac{\alpha}{1-\exp(\alpha(z-T))}$	
Makeham	a, b, c	$\mu(z) = c + a\exp(bz)$	
logistic	a, b, α	$\mu(z) = c + \frac{a\exp(bz)}{1+\alpha\exp(bz)}$	
quadratic	a, b, c	$\mu(z) = a + bz + cz^2$	
extended Weibull	α, γ, b	$\mu(z) = \alpha(z-\gamma)^b$	
Lynch and Brown	α, β, γ, δ	$\mu(z) = \alpha + \beta\text{arctan}\{\gamma(z-\delta)\}$	Lynch and Brown (2001)

Table 2: The functional forms for the hazard of death at advanced ages considered in this analysis. In the functions listed, z is age, $\mu(z)$ is the force of mortality at age z, and $q(z)$ is the probability of death between ages z and $z+1$. In compiling this list, I am particularly indebted to the discussion of various functional forms, and their origins in the literature, contained in Gavrilov and Gavrilova (1991) and Thatcher, Kannisto and Vaupel (1998).

The central death rate we observe is a function of a continuous, underlying hazard of death function which I will call $\mu(z)$. The hazard function is the instantaneous probability of death at age z, conditional on surviving to z; it is also the negative logarithmic derivative of the survival function, $F(z)$; that is,

$$\mu(z) = -\frac{dF(z)}{dz}\frac{1}{F(z)} \iff F(z) = \exp\left(-\int_0^z \mu(x)dx\right).$$

It will convenient to model hazards rather than survival, though the two functions contain the same information and can be derived from one another. Although the data we are able to collect usually permit us to compute central death rates, we are generally interested in the continuous, underlying hazard that led to those death rates. It is these hazards that I will model in this exercise.

To begin with, I consider several functional forms that have been employed to model mortality in adult and advanced ages; these are listed in Table 2. Each of these forms has two or more parameters; once the parameters are fixed, the shape of the hazard funtion is completely determined.

Draft: results for poster at PAA, March 31 2011.
Please do not cite

Estimation Strategy

I fit each hazard function to each country-year and country-cohort of data using maximum likelihood. I assume that deaths in an age-sex group are distributed binomially[3] with probability given by the hazard function; that is, for a given sex,

$$D_z \sim \text{Binomial}(\pi(z, \theta), N_z)$$

where D_z is the observed number of deaths in age group z, N_z is the observed population at risk of death in the age group, θ is a vector of parameters for the hazard function being used, and $\pi(z)$ is the probability of surviving to age $z + 1$, conditional on surviving to age z, and is given by

$$\pi(z, \theta) = \exp\left(-\int_z^{z+1} \mu(z, \theta) dz\right).$$

I approximate $\pi(z, \theta)$ with $\exp(-\mu(z + 0.5))$, which is commonly done in applications like this one (Thatcher, Kannisto and Vaupel, 1998, Appendix A).

The likelihood for an observed sequence of deaths $D_1, D_2, \ldots$ and exposures, $N_1, N_2, \ldots$ is then

$$Pr(\mathbf{D}|\pi(\mathbf{z}, \theta)) = \prod_z \binom{N_z}{D_z} \pi(z, \theta)^{D_z} (1 - \pi(z, \theta))^{(N_z - D_z)}.$$

Taking logs and dropping terms that don't vary with $\pi(z, \theta)$, I have

$$ll(\mathbf{D}|\pi(\mathbf{z}, \theta)) \propto \sum_z \left[D(z) \log(\pi(z, \theta)) + (N(z) - D(z)) \log(1 - \pi(z, \theta))\right]. \tag{1}$$

This is the likelihood I maximize, as a function of θ, in order to fit each hazard function to the data.

[3] I also performed all of the analysis presented here with a Poisson specification, and the results were extremely similar.

Goodness-of-Fit

I measure the goodness of fit of the various functional forms in two ways: first, I compute the root mean squared error in the estimated number of deaths for each country-sex-year; that is, I compute

$$\text{RMSE} = \frac{1}{k}\sqrt{\sum_z \left(\hat{D}_z - D_z\right)^2},$$

where k is the number of ages being fit. This quantity gives us a measurement of how close each model's predictions come to the observed numbers of deaths, in absolute terms; its units are number of deaths.

Second, I compute Akaike's Information Criterion (AIC) for each model and use it to rank them within each country-sex-year; that is, I compute the AIC values within the country-sex-year and rank them from best to worst. The AIC is essentially a penalized log-likelihood, where the penalty is a function of the number of parameters being estimated in the model:

$$\text{AIC} = -2\mathfrak{L} + 2k, \tag{2}$$

where $\mathfrak{L}$ is the value of the maximized likelihood from Equation 1, and k is the number of parameters being estimated in the model. Although the AIC has a simple form, it can be justified as selecting the model minimizes the estimated Kullback-Leibler distance between the distribution of data implied by the model and the one seen in the data (Burnham and Anderson, 2004; Claeskens and Hjort, 2008). Absolute AIC values are not interpretable, but the ordering of models given by their AIC values is. The ranking of models by their AIC values, which I compute for each country-sex-year, is thus an indication of how well the various functional forms trade off the number of parameters estimated with the accuracy of their fit to the data. A tradeoff is necessary here because a functional form with too many parameters is at risk of overfitting – that is, if I allow too many parameters to be estimated, I risk capturing esoteric features of the dataset I happen to have observed. I would not want to evaluate theories or produce forecasts based on a model that would reproduce esoteric features of our data that would not be found in other settings.

Draft: results for poster at PAA, March 31 2011.
Please do not cite

Results

As a sample of the results, Figure 1 shows the observed and fitted death rates from the ten functional forms for one country-sex-year in the period dataset: Belgium, females, 1989. The red circles show the observed death rates, while the blue curves show the maximum likelihood fit of each functional form. I can see that the functional forms vary widely in their ability to fit the data for this country-year. For example, the four-parameter logistic form does a decent job of capturing the shape of the observed hazard, which increases and then appears to plateau, a pattern which has been commented on extensively in the literature (Gavrilov and Gavrilova, 2009; Horiuchi and Wilmoth, 1998; Rau et al., 2009). The gompertz function overestimates mortality at the oldest ages, while the extra parameter of the makeham function addresses but is not able to solve that problem entirely. The eta and constant-hazard parameterizations both have considerable trouble in fitting the data well, since they assume that hazard is constant or increasing with age, which is not what I see in these observations. The weibull also has difficulty adequately capturing the empirical death rates.

Fits like the ones in Figure 1 were produced for each country-sex-year in the period dataset, and each country-sex-cohort in the cohort dataset. In order to make conclusions about how well the functional forms fit the data in general, I will consider aggregate measures of their performance across all of the country-sex-years and country-sex-cohorts in our data.

Goodness-of-Fit

In order to examine how well each functional form fits the data, I compute two quantities. The first, root mean squared error, is a measure of how close the predictions for each functional form come to matching the actual observed numbers of death at each age. This is a measure of absolute error which does not account for model complexity. The second quantity is the Akaike Information Criterion (AIC), which includes a term that trades off between the quality of a model's absolute fit, and the number of parameters it uses.

Table 3 gives an example of the goodness of fit measures for our running example, Belgian females in the 1989 period data. The functional forms are ordered by the rank they attain using the AIC, where rank 1 is the model that fits the data the best and rank 10 is the model that fits the data the worst. The table also shows the root mean squared error in the estimated number of deaths at each age (RMSE), and the difference between each model's AIC and the minimum AIC (ΔAIC). A few things are remarkable about the results. First, comparing the AIC ranks with the RMSE shows that the AIC does not uniformly prefer models whose predictions come closest to the observed deaths in absolute terms; it trades off between a model's fit and its complexity. For example, the results from the Kannisto model are slightly less accurate at replicating the observed data than the results from the quadratic or logistic forms. However, the quadratic and logistic forms both use one more parameter than the Kannisto does. Here, the AIC suggests that the extra parameters are not justified. Note, however, that as a rule of thumb, Burnham and Anderson (2004) suggest that models with $\Delta\text{AIC} \leq 2$ have some support from the data. In this case, the rule of thumb suggests that the data also provide some support for the quadratic model.

Summaries of the distribution of RMSE between the observed and fitted death counts for the period and cohort datasets are presented in Tables 4 and 5. I see that in terms of mean RMSE, the best-performing functional form in the period dataset is the Lynch-Brown, followed by the logistic, quadratic, Makeham, Kannisto, extended Weibull, Gompertz, Weibull, ETA, and constant hazard specifications. For the cohort data, the best performing functional form, as measured by RMSE, is again Lynch-Brown, followed by Makeham, quadratic, logistic, extended Weibull, Kannisto, Gompertz, Weibull, ETA, and constant hazard. In absolute terms, the ETA and constant hazard functions, which are widely used in theoretical models in economics (Bloom et al., 2007; Sheshinski, 2009, see, for example,), have on average several times the RMSE of the other functions; on the other hand, the mean RMSE for the the top few functional forms is very similar. The RMSE results for the cohort dataset are quite similar.

The results obtained from ranking the models with the AIC are presented in Tables 6 and 7. Table 6 shows summaries of the distributions of AIC rank for the fits to the period dataset, while Table 7 shows the distributions of the AIC rank of the fits to the cohort dataset. Since the AIC is a relative measure of model fit, its absolute value is not interpretable; rather, I examine the distribution of the rank of each functional form across all country-sex-years or country-sex-cohorts. The model that performs the best in a given country-sex-year or cohort is ranked 1, while the model with the poorest performance is ranked 10[4]. For example, in Table 6, I see that across all country-sex-years, the average

[4]Recall that the AIC includes a penalty for the complexity (ie, number of parameters) in a model, so that it is possible for model A to outperform model B, even if model A is nested within model B.

	RMSE	AIC	AIC rank	ΔAIC
kannisto	40	174195.11	1	0.00
quadratic	39	174195.37	2	0.26
logistic	39	174197.43	3	2.32
lynch-brown	40	174198.36	4	3.25
makeham	41	174198.53	5	3.42
gompertz	61	174232.21	6	37.10
extended-weibull	80	174264.11	7	69.00
weibull	191	174715.46	8	520.35
eta	400	176451.93	9	2256.82
consthaz	683	180742.69	10	6547.59

Table 3: Root mean squared error (RMSE), Akaike Information Criterion (AIC), AIC ranks, and differences between each AIC and the minimum AIC (ΔAIC) for the maximum likelihood fits of the functional forms to the period data for females in Belgium, 1989. In this country-sex-year, the AIC ranking suggests that the Kannisto functional form attained the best tradeoff between number of parameters and fit. Note that some functional forms with slightly lower RMSE than the Kannisto attained were still ranked worse because the difference in RMSE did not compensate for the additional parameters they use. For example, the quadratic and logistic functional forms both predicted the absolute number of deaths more accurately than the Kannisto model, but according to the AIC, the difference in absolute fit did not make up for the extra parameter they use. The differences between each AIC and the minimum AIC, in the column labeled ΔAIC, suggest that the quadratic model is also substantially supported by the data, using Burnham and Anderson (2004)'s rule of thumb that models with $\Delta\text{AIC} \leq 2$ are plausible.

rank for the Lynch-Brown model was 3.28, which is quite good. For the period data, the best performer in terms of average AIC rank is the quadratic functional form, followed by the Kannisto, Lynch-Brown, Makeham, logistic, extended Weibull, Gompertz, Weibull, ETA, and constant hazard specifications. In the cohort data, the results are very similar: the best average AIC rank is attained by the quadratic functional form, followed by the Makeham, Lynch-Brown, Kannisto, Gompertz, logistic, extended Weibull, Weibull, ETA, and constant hazard specifications. Taken together, tables 6 and 7 show that although several of the functional forms perform well in some country-sex-years, on average the quadratic performs the best.

Figures 2 and 3 present the entire distribution of ranks for each model in the period and cohort data. Figure 2 displays all of the distributions of ranks simulatenously, while Figure 3 expands the same distributions across multiple plots for more precise comparison. In both plots, each color represents a rank, with red being the worst, blue intermediate, and green being the best. In Figure 2, the order of the functional forms is determined by their average AIC rank. We see that in the period data, on the left, the Kannisto model is ranked number 1 (best) more often than any other, but that the quadratic achieves the best average AIC rank over all of the country-sex-years. Qualitatively, the quadratic, Kannisto, Lynch-Brown, and Makeham specifications all perform quite well. The logistic specification, which Thatcher, Kannisto and Vaupel (1998) concludes is the best for fitting data at advanced ages, is decidedly in the middle of the pack. And the functional forms used in economic theory, the ETA and constant hazard, consistently perform the worst. For the cohort data, on the right, the results are similar but not identical. Again, the Kannisto model is most often ranked best, but it is only fourth in terms of average AIC rank. The best average rank is again attained by the quadratic model, with the Makeham, Lynch-Brown, and Kannisto specifications all performing well. Again, the logistic model is in the middle of the pack, and the models used in economic theory are the poorest performers.

Conclusion

Many studies of mortality at advanced ages, and related substantive topics, require that the analyst choose a functional form to summarize the hazard of death by age. In these studies, the analyst should employ principled criteria for selecting which functional form to use. In particular, the criteria used to select a model should explicitly address the tension between fitting the observed data as well as possible, on the one hand, and minimizing model complexity and, therefore, the risk of over-fitting, on the other.

In this excercise, I selected ten functional forms used in a wide spectrum of studies, from forecasting to theoretical modeling, and demonstrated the use of one powerful, principled tool – Akaike's Information Criterion – to select the most attractive model. Of the ten functional forms I considered, the quadratic model had the strongest support from a cross-national, high-quality dataset of deaths at advanced ages. Several other functional forms also performed well, but the one most commonly recommended for use in the literature, the logistic, generally produced only moderately good fits.

In many cases, there will be considerations beyond statistical ones which will limit the set of candidate models that the analyst will consider. For example, a study that seeks to evaluate the level of empirical support for theories of aging may only wish to consider functional forms suggested by one of the theories being investigated. The procedure outlined above, or a similar one, should still be employed on the reduced set of candidate models to decide which ones enjoy the strongest empirical support.

References

Bloom, D.E., D. Canning, R.K. Mansfield and M. Moore. 2007. "Demographic change, social security systems, and savings." *Journal of Monetary Economics* 54(1):92–114.

Bongaarts, J. 2005. "Long-range trends in adult mortality: Models and projection methods." *Demography* 42(1):23–49.

Burnham, K.P. and D.R. Anderson. 2004. "Multimodel inference: understanding AIC and BIC in model selection." *Sociological Methods & Research* 33(2):261.

Claeskens, G. and N.L. Hjort. 2008. *Model selection and model averaging.* Cambridge University Press.

Gavrilov, L.A. and N.S. Gavrilova. 1991. *The biology of life span: a quantitative approach.* Harwood Academic Publishers New York.

Gavrilov, Leonid A. and Natalia Gavrilova. 2009. "Moratlity Measurement at Advanced Ages." Paper presented at Population Association of America.

Horiuchi, S. and J.R. Wilmoth. 1998. "Deceleration in the age pattern of mortality at older ages." *Demography* 35(4):391–412.

Jdanov, D.A., D. Jasilionis, EL Soroko, R. Rau and J.W. Vaupel. 2008. "Beyond the Kannisto–Thatcher Database on Old-Age Mortality: An assessment of data quality at advanced ages." *MPIDR Working Papers* .

Kannisto and Thatcher. 2009.
URL: *http://www.demogr.mpg.de/?http://www.demogr.mpg.de/databases/ktdb/*

Kinsella, K.G., D.R. Phillips and Population Reference Bureau. 2005. *Global aging: The challenge of success.* Population Reference Bureau.

Lynch, S.M. and J.S. Brown. 2001. "Reconsidering mortality compression and deceleration: An alternative model of mortality rates." *Demography* 38(1):79–95.

Martin, L.G. and S.H. Preston. 1994. *Demography of aging.* National Academies Press.

Rau, Roland, Magdalena M. Muszynska, Annette Baudisch and James W. Vaupel. 2009. "At What Age Does Mortality Start To Decelerate?" Paper presented at Population Association of America.

Sheshinski, E. 2009. "Uncertain Longevity and Investment in Education." *CESifo Working Paper Series* .

Tabeau, E., A. van den Berg Jeths and C.R. Heathcote. 2001. *Forecasting mortality in developed countries: insights from a statistical, demographic, and epidemiological perspective.* Springer.

Thatcher, A.R., V. Kannisto and J.W. Vaupel. 1998. *The force of mortality at ages 80 to 120.* Odense University Press.

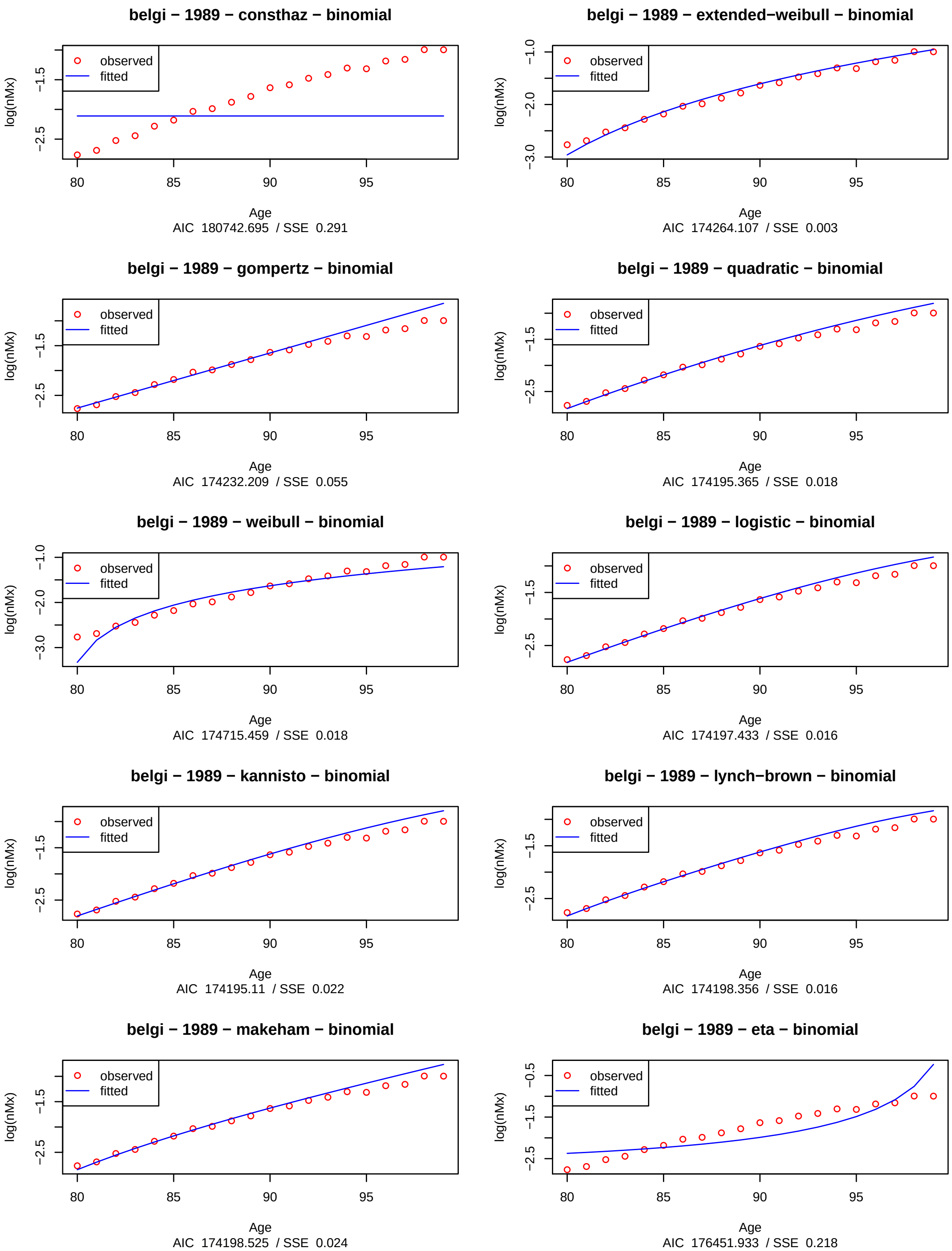

Figure 1: The functional forms fit to the period data for females in Belgium, 1989. The red circles show the observed death rates, and the blue curves show the maximum-likelihood fit of each functional form. For this country-sex-year, some functional forms fit the observed death rates much better than others. The AIC rankings and root mean squared error (RMSE) values for this country-sex-year are shown in Table 3.

	consthaz	gompertz	weibull	kannisto	makeham	extended-weibull	quadratic	logistic	lynch-brown	eta
Min	63.05	8.55	25.55	7.68	7.54	7.91	7.54	7.57	7.60	47.89
1st Qu.	243.26	37.05	85.80	31.51	30.40	33.64	29.82	29.41	28.52	157.61
Median	477.66	74.09	145.83	63.10	61.51	65.56	59.91	59.39	58.10	297.96
Mean	1071.61	159.05	321.43	132.81	128.52	133.07	125.28	124.33	123.40	648.92
3rd Qu.	1495.40	200.02	471.30	118.99	120.00	125.48	110.78	107.97	107.01	996.51
Max	6497.21	1095.57	1888.22	924.65	861.88	837.55	838.24	834.07	833.38	3223.00

Table 4: Distribution of the root mean squared error in fitted counts of deaths, for each functional form, across all country-years and both sexes in the period data. On average, the Lynch-Brown model produces the most accurate predictions of the number of deaths in each country-sex-year in absolute terms, while the constant hazard specification does the worst.

	consthaz	gompertz	weibull	kannisto	makeham	extended-weibull	quadratic	logistic	lynch-brown	eta
Min	1.24	0.01	0.18	0.01	0.03	0.10	0.01	0.01	0.00	1.24
1st Qu.	139.44	27.80	50.25	26.99	25.51	27.34	25.59	25.64	24.95	63.18
Median	250.83	45.93	86.20	44.57	41.54	44.35	41.89	42.21	41.15	101.44
Mean	584.94	95.86	191.21	91.57	85.47	89.79	85.61	86.09	83.44	237.13
3rd Qu.	865.96	149.09	284.80	143.79	129.97	145.38	131.99	131.88	127.23	376.92
Max	3943.26	473.14	1338.20	443.83	846.97	455.70	444.83	440.51	431.98	1247.55

Table 5: Distribution of the root mean squared error in fitted counts of deaths, for each functional form, across all country-cohorts and both sexes in the cohort dataset. On average, the Lynch-Brown specification produces the most accurate predictions of the number of deaths in each country-sex-cohort in absolute terms, while the constant hazard specification does the worst.

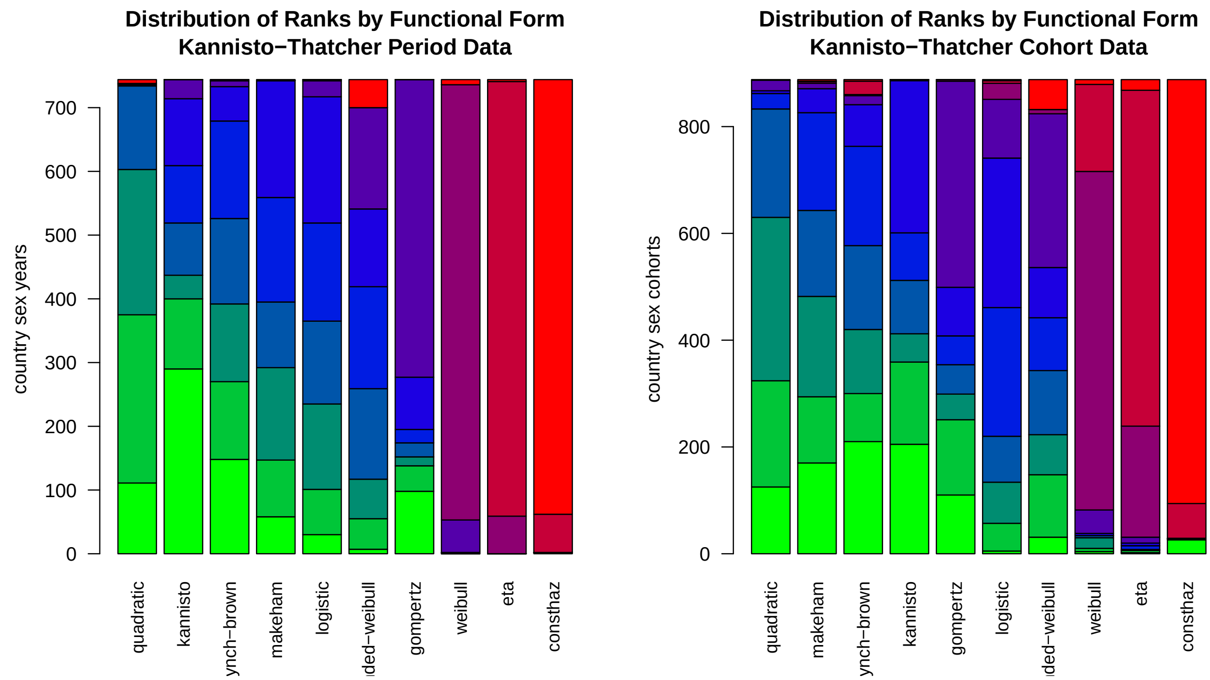

Figure 2: The distribution of model ranks, according to AIC, for the period data (left) and the cohort data (right). Red is the worst rank, blue is intermediate, and green is the best. The plots are ordered by their mean rank in the period data; an expanded visualization of these distributions is shown in Figure 3. We see that although the Kannisto specification is ranked first more than any other, the quadratic outperforms it in terms of average AIC rank in both the period and the cohort datasets. In general, it is clear that the Weibull, ETA, and constant hazard specifications perform poorly. On the other hand, the quadratic, Kannisto, Lynch-Brown, and Makeham specifications all perform quite well.

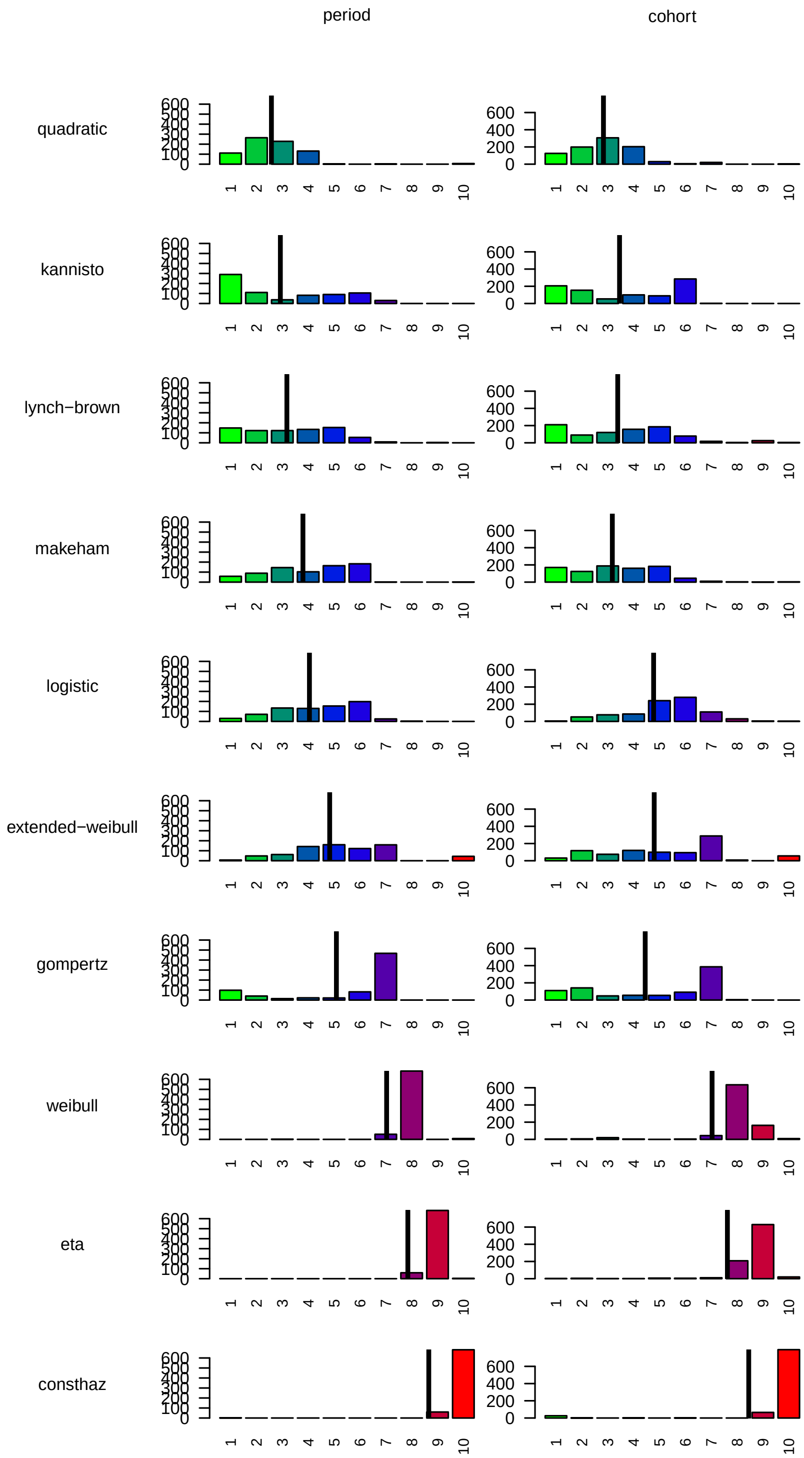

Figure 3: An expanded version of Figure 2, showing the distribution of AIC rank in the period and cohort data for each functional form on a separate plot. The color scheme is the same as in Figure 2: red colors are the worst ranks, blues are intermediate ranks, and greens are the best ranks. The solid vertical line shows the mean rank on each plot. Qualitatively, it is clear that the Weibull, ETA, and constant hazard specifications are the poorest.

	consthaz	gompertz	weibull	kannisto	makeham	extended-weibull	quadratic	logistic	lynch-brown	eta
Min	1.00	1.00	3.00	1.00	1.00	1.00	1.00	1.00	1.00	8.00
1st Qu.	10.00	6.00	8.00	1.00	3.00	4.00	2.00	3.00	2.00	9.00
Median	10.00	7.00	8.00	3.00	4.00	5.00	2.00	4.00	3.00	9.00
Mean	9.93	5.86	7.95	3.43	4.11	4.78	2.40	4.30	3.28	8.96
3rd Qu.	10.00	7.00	8.00	6.00	5.00	6.00	3.00	6.00	5.00	9.00
Max	10.00	7.00	10.00	7.00	10.00	10.00	10.00	8.00	9.00	10.00

Table 6: Distribution of rank according to AIC, for each functional form, across all country-years and both sexes. The functional form with the best fit in a country-year-sex group has rank 1, while the model with the worst fit has rank 5. We see that the logistic, gompertz, and makeham specifications trade off for the best performance, while eta and consthaz are generally in the last two spots.

	consthaz	gompertz	weibull	kannisto	makeham	extended-weibull	quadratic	logistic	lynch-brown	eta
Min	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1st Qu.	10.00	2.00	8.00	2.00	2.00	3.00	2.00	4.00	2.00	8.00
Median	10.00	6.00	8.00	4.00	3.00	5.00	3.00	5.00	4.00	9.00
Mean	9.66	4.98	7.95	3.73	3.40	4.95	2.86	5.23	3.59	8.66
3rd Qu.	10.00	7.00	8.00	6.00	5.00	7.00	4.00	6.00	5.00	9.00
Max	10.00	8.00	10.00	7.00	10.00	10.00	10.00	10.00	10.00	10.00

Table 7: Distribution of rank according to AIC, for each functional form, across all country-cohorts and both sexes. The functional form with the best fit in a country-year-sex group has rank 1, while the model with the worst fit has rank 5. We see that the logistic, gompertz, and makeham specifications trade off for the best performance, while eta and consthaz are generally in the last two spots. The cohort data are best fit by the Makeham hazard.

	consthaz	gompertz	weibull	kannisto	makeham	extended-weibull	quadratic	logistic	lynch-brown	eta
Min	337.26	0.00	23.40	0.00	0.00	0.00	0.00	0.00	0.00	127.11
1st Qu.	1692.46	4.69	154.61	0.00	1.50	2.05	0.11	2.07	0.53	627.67
Median	3719.32	18.72	316.73	2.05	2.78	4.22	1.48	3.37	1.98	1353.08
Mean	9027.20	57.87	724.95	11.30	12.18	10.47	2.39	3.54	2.08	2936.56
3rd Qu.	10263.87	60.68	848.86	9.84	7.73	8.74	2.31	3.97	3.03	4037.70
Max	76115.03	731.28	7011.81	258.47	4114.44	515.16	43.45	42.45	56.51	20041.80

Table 8: Distribution of ΔAIC, for each functional form, across all country-years and both sexes. The functional form with the best fit in a country-year-sex group has ΔAIC$=0$.

	consthaz	gompertz	weibull	kannisto	makeham	extended-weibull	quadratic	logistic	lynch-brown	eta
Min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1st Qu.	815.66	1.42	70.25	0.13	0.35	1.16	0.49	2.83	0.17	127.20
Median	1690.70	6.05	154.65	3.01	1.67	3.75	1.65	3.93	1.89	229.09
Mean	3943.32	20.49	357.66	11.47	9.85	9.66	3.25	5.76	1.77	537.97
3rd Qu.	5045.41	20.49	446.25	11.08	2.95	9.72	2.37	5.36	2.61	759.34
Max	33388.53	279.43	3765.25	193.36	5489.43	333.81	70.75	76.18	15.05	3137.12

Table 9: Distribution of ΔAIC, for each functional form, across all country-cohorts and both sexes. The functional form with the best fit in a country-year-sex group has $\Delta\text{AIC}=0$.