Use of symbolic regression
for fitting survival functions in actuarial science

Keywords:
actuarial science, demographics, symbolic regression, survival function, life insurance

Moscow, 2024

1 Preface

In this article, ChatGPT[1] was employed to align the text with academic style guidelines, identify and correct errors, and enhance the overall phrasing. It is important to note that ChatGPT was solely utilized to improve existing content and not to generate text from scratch.

2 Introduction to survival data analysis

In actuarial science one of the core question is: ”what is the probability that a person of age x will survive to age $x + t$?”. Without this information, it is impossible to accurately calculate insurance premiums. When both x and t are integers, the probability can be easily derived using **life tables**, such as the one shown below[2]:

Table 1: *Actuarial life table for female population. Adapted from the USA Social Security Administration [2]*

x	l_x
0	100 000
1	99 494
2	99 455
3	99 432
4	99 415
...	...
109	20
110	9
111	4
112	2
113	1

In this table, x represents age, and l_x denotes the number of survivors out of 100 000 initially born. The survival probability from age x to $x + t$ is calculated as:

$$_tp_x = \frac{l_{x+t}}{l_x}$$

However, the actual terms under which acturial firms operate are rarely constrained to integer ages. Thus, the need for continual interpolation of the life table arises. Such interpolation is often called as a **survival function**. In this article we want to propose a new approach to fitting such functions. But before it, we need to introduce existing methods and approaches.

2.1 Survival function and Mortality Intensity

A survival function $s(x)$ (also known as a mortality law) expresses the probability that an individual will survive beyond age x[3]:

$$s(x) = P(T > x)$$

The function must conform to the following properties:

- $s(0) = 1$ (All individuals are alive at birth),
- $s(+\infty) = 0$ (Everyone eventually dies),
- $s(x)$ is continuous and strictly decreasing (ensuring that death is inevitable but not instantaneous).

[1] https://chatgpt.com/

[2] We will use this data for the rest of an article

[3] T is the random variable that denotes the time of a persons death.

Sometimes it is more convenient to consider not the survival function itself, but its derivative — the mortality intensity, which defined as and roughly approximates the instantaneous probability of death at age x:

$$\mu_x = -\frac{s'(x)}{s(x)}$$

Given μ_x, we can reconstruct the survival function through integration:

$$s(x) = \exp\left(\int_0^x \mu_t \, \mathrm{d}t\right)$$

2.2 Estimating survival function

For integer ages, the survival function can be approximated by:

$$s(n) = \frac{l_n}{l_0}$$

Figure (1) presents the survival function at whole ages. The graph highlights several critical phases in the mortality curve. Most notably, mortality intensity during the first year of life is significantly higher compared to subsequent years. This is evident in the initial steep decline of the survival function. After this, the mortality rate remains relatively stable and low until around age 40. From there, a marked and consistent rise in mortality is observed, peaking at age 95, where the curve begins to flatten. Capturing these complex transitions—low mortality in early adulthood followed by a steep increase in older age—poses a significant challenge when constructing an accurate and smooth survival function.

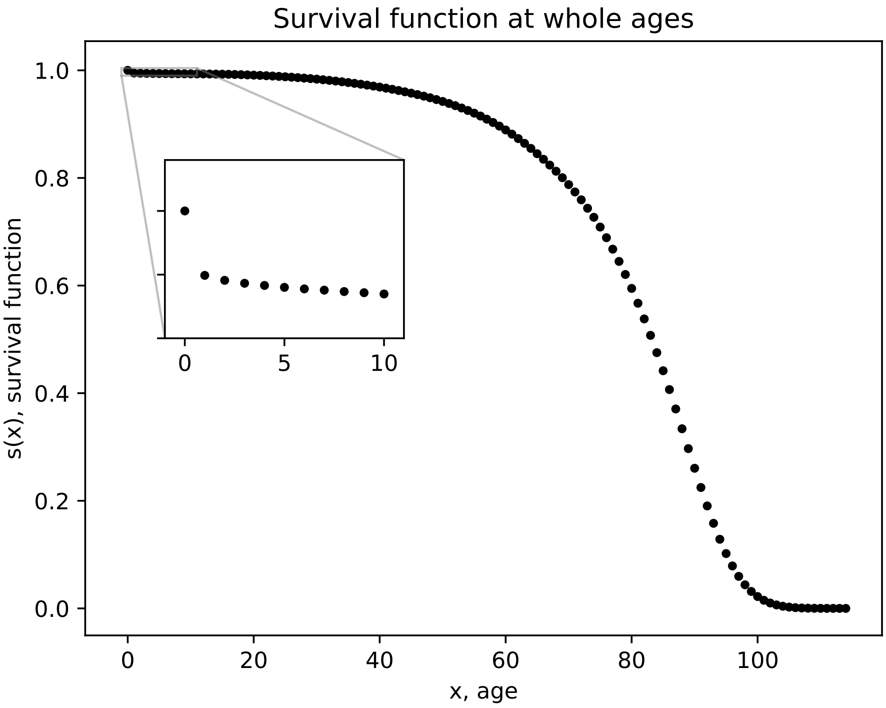


Figure 1: Survival function at integer values

However, a key challenge lies in interpolating the survival function for non-integer ages. The mortality rate varies dramatically, especially in early life (high at birth, much lower in youth) and then rises sharply after age 40. This creates a complex shape that is difficult to model precisely.
There are two common approaches to survival function interpolation:

- Interpolating between integer ages, as discussed in Section 2.2.1.
- Finding an analytical form (a ”mortality law”) that fits the data, as discussed in Section 2.2.2.

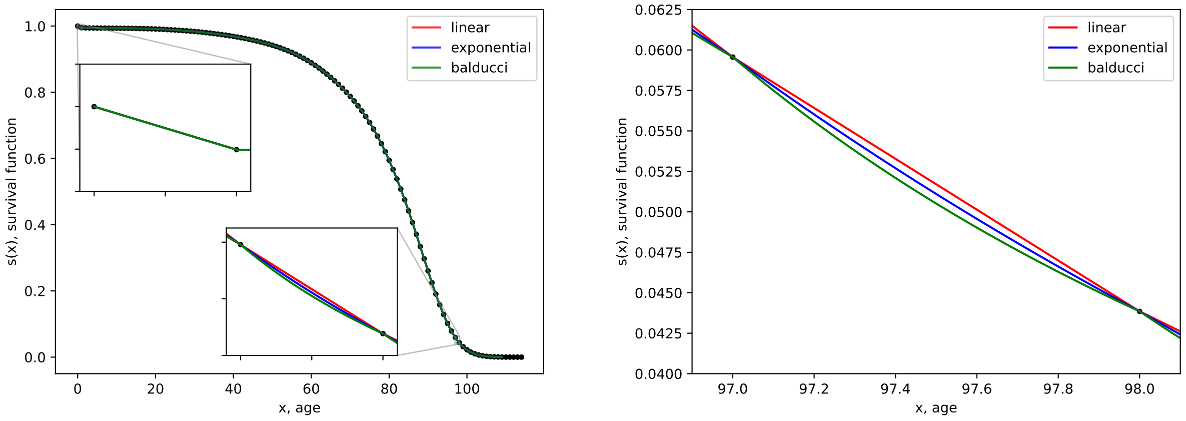
2.2.1 Interpolating Between Two Integer Ages

There are three common hypotheses for how the survival function behaves between two consecutive integer ages n and $n+1$:

- **Linear interpolation**: $s(x) = a_n + b_n x$ where $n \leqslant x < n+1$.
- **Constant force of Mortality (Exponential interpolation)**: $s(x) = a_n e^{b_n x}$ where $n \leqslant x < n+1$
- **Balducci Hypothesis (Hyperbolic or Harmonic interpolation)**: $\dfrac{1}{s(x)} = a_n + b_n x$ where $n \leqslant x < n+1$

Each of these interpolation methods is uniquely defined by two points. Given any two consecutive ages, we can establish two equations, such as $s(n) = a_n + b_n \cdot n$ and $s(n+1) = a_n + b_n \cdot (n+1)$, allowing us to solve for the coefficients a_n and b_n. This enables the interpolation of the survival function for fractional ages.

In figure (2) we observe that all three approaches – linear, exponential, and Balducci – yield nearly indistinguishable results at early ages, as shown in the first zoomed-in section. However, as we move toward older ages (second zoom or subfigure (2b)), the differences between the interpolations become more pronounced, highlighting the growing divergence in their predictions.



***(a)** Interpolation on the whole curve*

***(b)** Zoom in of the interpolation to the ages 97-98*

***Figure 2:** Interpolations of survival function*

2.2.2 Well known mortality laws

We will now provide a brief overview of the most widely known mortality laws, with this section drawing heavily from [3] and [4]. For a more detailed explanation and practical applications of these functions, we recommend referring to the original article.

One of the earliest attempts to derive an analytical expression for the survival function $s(x)$ was made by De Moivre in 1725. He proposed a simple linear form:

$$s(x) = \frac{\omega - x}{\omega} = 1 - \frac{x}{\omega}$$

While notable for its simplicity, this model significantly deviates from real-world data. In 1825, the English mathematician Benjamin Gompertz introduced a more robust formula:

$$s(x) = exp\left(-\frac{B}{\alpha}(e^{\alpha x - 1})\right)$$

Gompertz's law provides an excellent fit for middle-aged individuals, though it tends to deviate at both younger and older ages. To address this, William Makeham refined the formula in 1860 and 1869 by incorporating additional terms to account for the quadratic relationship between age and mortality:

$$s(x) = exp\left(-Ax - \frac{H}{2}x^2 - \frac{B}{\alpha}(e^{\alpha x - 1})\right)$$

In figure 3, we plot the fitted values of both Gompertz and Makeham's laws. As seen, both models exhibit significant divergence from empirical data on many intervals, illustrating the limitations of these traditional approaches.

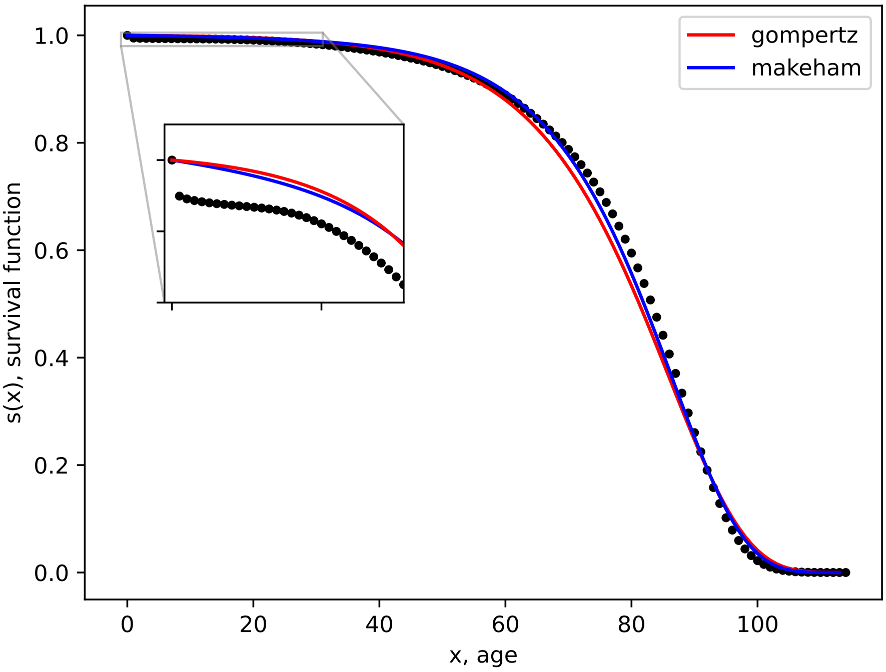


Figure 3: *Mortality laws*

There are numerous other mortality laws that have been developed over time to better capture the nuances of survival data. Ones worth mentioning are:

Weibull (power) mortality law:

$$s(x) = \exp\left(-(a/g) \times x^g\right)$$

According to [3], this model provides the best fit for male survival rates, offering a more flexible approach compared to traditional laws.

And logistic mortality function:

$$s(x) = \exp\left(-bt - (b/k)\ln\frac{a + (b-a)e^{-kx}}{b}\right)$$

This function, as noted by [3], is considered to offer the best fit for female survival rates, reflecting the different patterns of mortality across genders.

3 Symbolic Regression for Survival Function Discovery

3.1 Introduction to symbolic regression

Since a deep understanding of the mechanics behind symbolic regression is not necessary to appreciate its practical applications, we will provide a brief introduction to the concept, focusing on its benefits and limitations. For a more detailed discussion, we refer the reader to [1].

Symbolic regression is a branch of regression analysis aimed at discovering the mathematical expression that best fits a given dataset. What distinguishes symbolic regression from traditional methods is that it does not rely on a predefined model. Instead of enforcing a certain model to data, it searches for the model that best captures the relationships within the data, giving us clear analytical expression that can easily be interpreted and used for actuarial calculations.

In this paper, we employ a classic approach to symbolic regression using genetic programming. This method, inspired by Darwinian evolution, can be summarized as follows:

1. Initialize a population of mathematical expressions based on the defined operator set.
2. Repeat until the stopping criteria are met:
   - Identify the best-performing candidates based on the data.
   - Use these candidates to generate a new population through mutation and crossover.

This evolutionary algorithm enables the discovery of functions that fit the data well. However, several important caveats should be noted. First, as with all statistical modeling, there is a trade-off between complexity and accuracy. While more complex functions may fit the data better, they are often harder to interpret and may lead to overfitting. Additionally, symbolic regression can be computationally intensive, especially for large datasets. The algorithm is also non-deterministic, meaning it may produce different results with each run. Finally, enforcing specific constraints, such as continuity or strict positivity, can be challenging within this framework.
For the purposes of this study, we utilize the PySR Python module to implement symbolic regression.

3.2 Fitting continuous survival function

The instinctive approach for any researcher encountering this problem is to seek a single continuous function that adequately fits all the data. However, effectively capturing the complexities of human mortality often results in analytical forms that are considerably intricate. Additionally, since the algorithm must compute the error for all data points, the search process can be computationally demanding.

When we apply symbolic regression to our dataset, we derive the following equation[4]:

$$s(x) = \exp\left(-\frac{12x}{x - 50224.38\,(0.0084x - 1)^3 + 2.53}\right)$$

Although this equation does not conform to any established mortality laws, it provides an excellent fit for our data, as illustrated in figure 4. When comparing it to the fits suggested by the Makeham and Gompertz models, we observe that symbolic regression yields superior results. This can be attributed to the nature of the algorithm itself. While traditional mortality laws strive to establish a single analytical form applicable to multiple life tables, symbolic regression focuses on finding the best fit specifically for our dataset, thus offering a distinct advantage.

However, it is important to note that the fitted function does not adequately capture the mortality patterns in early ages. To enhance our fit, we can explore the possibility of employing a piecewise function.

3.2.1 Fitting piecewise function

Utilizing a piecewise approach may generate a set of simpler functions that can more effectively capture specific segments of the mortality curve. However, this method may introduce challenges concerning continuity between the segments.

The piecewise function discovered by the algorithm can be expressed as follows:

$$s(x) = \begin{cases} \frac{2.59\cdot 10^{-5}x^6+0.99x+0.31}{2.61\cdot 10^{-5}x^6+x+0.31}, & 0 \leqslant x \leqslant 11 \\ 0.99 - 1.93 \cdot 10^{-6}\,(0.62x-1)^3, & 10 \leqslant x \leqslant 51 \\ 0.96\left(3.65 \cdot 10^{-14}x^6 \log\left(\log(x+2.35)+0.98\right)^6 - 1\right)^2, & 50 \leqslant x \leqslant 91 \\ e^{9.54\cdot 10^{-18}\left(-x+\log\left(\left(-x+\log\left((x-0.79)^2\right)+0.79\right)^2\right)+0.79\right)^9}, & 90 \leqslant x \leqslant 120 \end{cases}$$

[4] We select the best fit, irrespective of complexity

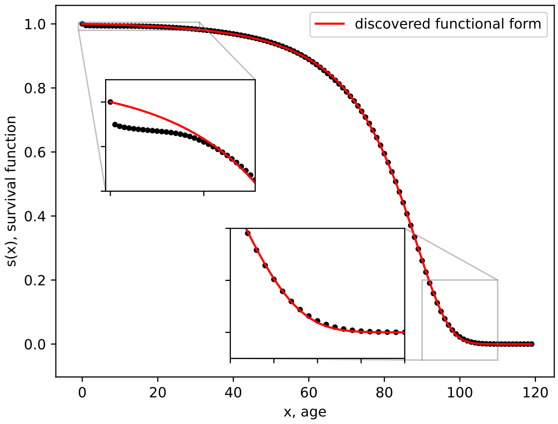


Figure 4: *Symbolic regression best fit*

As illustrated in figure 5, this piecewise function offers a significantly improved fit for early ages. However, it is crucial to note that this function lacks continuity, which may pose challenges for actuarial calculations. Furthermore, this run of the algorithm did not capture the interval from 50 to 91 effectively.

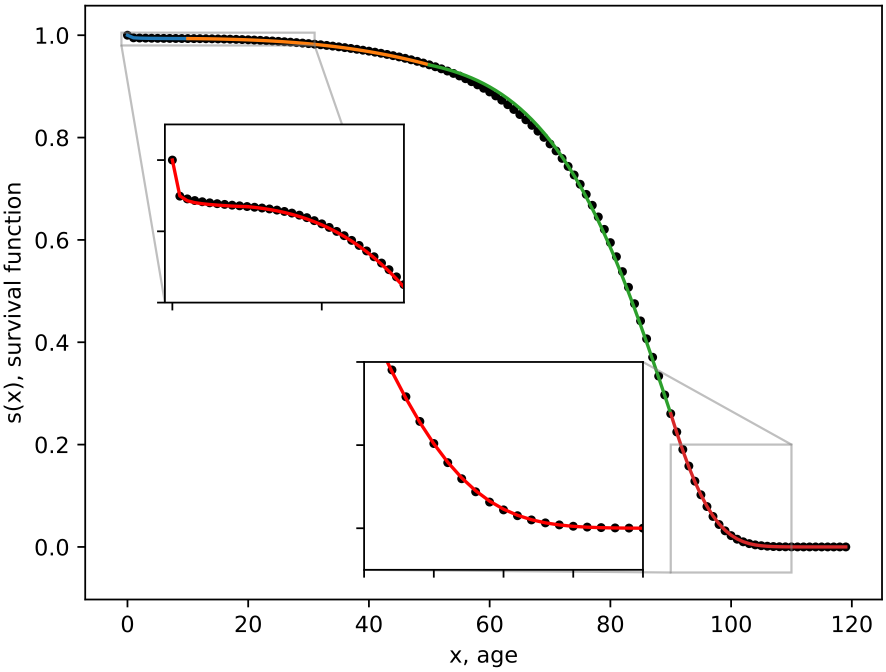


Figure 5: *Symbolic regression piecewise best fit*

In summary, symbolic regression provides a powerful means of uncovering complex relationships in survival data, enabling researchers to find tailored solutions that traditional models may overlook.

4 Conclusion

In this article, we have presented symbolic regression as an elegant solution to the long-standing challenge of fitting survival functions. We are excited to see how the proposed approach will be implemented in the industry and how it may contribute to the ongoing search for innovative and improved general solutions to this complex problem.

References

[1] Nour Makke **and** Sanjay Chawla. *Interpretable Scientific Discovery with Symbolic Regression: A Review*. 2023. arXiv: 2211.10873 [cs.LG]. URL: https://arxiv.org/abs/2211.10873.

[2] Social Security Administration of the USA. *Actuarial Life Table 2021*. Accessed: 2024-10-13. 2024. URL: https://www.ssa.gov/oact/STATS/table4c6.html.

[3] David L. Wilson. *The analysis of survival (mortality) data: Fitting Gompertz, Weibull, and logistic functions*. **volume** 74. 1. 1994, **pages** 15–33. DOI: https://doi.org/10.1016/0047-6374(94)90095-7. URL: https://www.sciencedirect.com/science/article/pii/0047637494900957.

[4] М. А. Скорик Л. В. Иванова. Ю. Н. Миронкина Н. В. Звездина. *Актуарные расчеты : учебник и практикум для вузов*. Accessed: 2024-10-13. Moscow: Издательство Юрайт, 2024. ISBN: 978-5-534-17936-1. URL: https://urait.ru/bcode/534006.