Modeling Survival Trends: Gender and Era Insights from the BHHT Dataset

Understanding lifespan trends over time has important implications for various sectors, including insurance, healthcare, and public policy. Identifying and modeling these trends enable stakeholders to recognize patterns and make informed decisions. This study uses the Brief History of Human Time (BHHT) dataset to model and examine survival trends from 1500 to 2000, with a specific focus on gender differences. By analyzing these trends, this study seeks to determine whether survival disparities between males and females exist across centuries and how they evolve. The findings can provide insights into gender-specific longevity patterns and their potential causes.

To prepare for survival analysis, basic data summaries were conducted to characterize the dataset. The BHHT dataset comprises observational biographical information on "notable people" from Wikipedia, including variables such as birth year, lifespan (missing if the individual is still alive), gender (male, female, and other), and region (Europe, America, Asia, Oceania, and Africa). For this analysis, only the male and female categories were used due to the limited sample size in the "other" category. Additionally, the dataset is male-dominated, with males accounting for almost six times more observations (389,684) than females (66,959). Observations are most abundant for individuals born in the 1900s, with declining counts for earlier centuries. The 2000s have the fewest observations, as fewer individuals born in this era have passed away.

Survival analysis methods were employed to investigate differences in survival across eras and genders. First, Marginal Survival Functions were used to estimate the proportion of people alive at a given time, representing the probability of survival up to a specific age. This was done using the Kaplan-Meier estimator. Additionally, Marginal Hazard Functions were calculated to estimate the instantaneous rate of mortality, capturing the probability of death (hazard) at a specific time for those still alive. The Nelson-Aalen estimator was used for Proportional Hazard Modeling to explore the relationship between survival and explanatory variables. Several modeling techniques were employed, including linear, quadratic, cubic, and cubic spline interacting for birth year. Cubic splines, with four degrees of freedom, were chosen for their flexibility in capturing non-linear trends without overfitting. Separate models were created for males, females, and the interaction between gender and birth year, with both linear and spline terms analyzed.

The Marginal Survival Functions reveal distinct survival patterns across eras and genders. Figure 1 shows that from the 1500s to the 1800s, there is a steady increase in the proportion of individuals surviving to a given age. The 1900s show the highest survival rates, with significant differences becoming evident around age 30. When categorized by gender, females consistently exhibit higher survival proportions than males as age increases, with a noticeable divergence around age 40. By age 80, approximately 60% of females are still alive compared to only 40% of males.

The Marginal Hazard Functions provide additional insights, showing a decline in the estimated percentage of individuals dying as the eras become more recent. In Figure 2, females display lower hazard rates overall, with the largest gender differences emerging around ages 40

and 50. These findings suggest that females tend to have a lower risk of mortality throughout their lifespans compared to males.

Proportional Hazard Modeling offered further insights into survival trends. A linear model indicated a statistically significant relationship between birth year and the log hazard, with a hazard ratio of 0.9953 (log hazard ratio of -0.0047) and a confidence interval of [0.9952, 0.9953] with a p-value much less than 0.05. This suggests we have significant evidence that the risk of mortality decreases as the birth year increases. Quadratic and cubic terms added similar significance, but the cubic spline model provided the most detailed interpretation. The spline revealed significant differences in hazard contributions across eras and genders. For instance, in Figure 3, males contributed less to the log hazard compared to females before 1700. After this point, males showed higher contributions. The general trend showed a steady decline in hazard contributions until around 1650, followed by a plateau until 1800 and a sharp decrease thereafter. Interestingly, hazard contributions increased slightly for both genders around 1980. When looking at the log hazard rate coefficient for females, interacting with birth year, we see a slight negative interaction (-0.0013), with a very small (less than 0.05) p-value.

The Marginal Survival Functions clearly demonstrate that survival rates have improved significantly over time, with the 1900s exhibiting the highest survival proportions. For the first time in recorded history, a notable proportion of individuals born in the 1900s survived to age 100, likely due to advancements in healthcare and living conditions. Gender differences are also apparent, with females consistently showing higher survival rates than males, particularly after age 40. These differences become more pronounced between ages 60 and 90, where the survival gap widens substantially.

Marginal Hazard Functions reveal a similar pattern, with mortality risk increasing exponentially with age but at a slower rate for females. These trends underscore the importance of considering gender when analyzing survival patterns, as they highlight distinct longevity advantages for females.

Proportional Hazard Modeling adds further nuance to these findings. The cubic spline analysis captures complex survival trends that are not evident from simpler models. The steady decline in hazard contributions until 1650 may reflect gradual improvements in societal conditions, while the plateau from 1650 to 1800 could be associated with factors such as widespread diseases and wars. The steep decline in the 1900s aligns with the introduction of modern medicine and public health initiatives. However, the slight increase in hazard contributions in the late 20th century raises questions, such as the impact of lifestyle changes or emerging health risks. Additionally, the log hazard rate (-0.0013) along with the very small p-value we see for females suggests that, compared to males, there is very strong evidence that females have decreased risk of mortality throughout the years.

Gender-specific trends in the Proportional Hazard Models also provide valuable insights. Males initially exhibited lower contributions to hazard rates compared to females, but this shifted around 1700. Possible explanations include changes in childbirth-related mortality, underrepresentation of females in historical records, or societal factors that could impact male and female life spans differently. These findings emphasize the need for more comprehensive data to fully understand the historical context of survival disparities.

This analysis provides strong evidence of gender differences in survival trends over time. Females tend to have higher survival rates than males, with these differences becoming more

pronounced in later life stages. Proportional Hazard Modeling highlights complex patterns in survival across eras, showing overall improvements in survival but raising questions about recent trends. While these findings offer valuable insights, further research with more comprehensive datasets is needed to deepen our understanding of the underlying factors driving these trends.

Figures

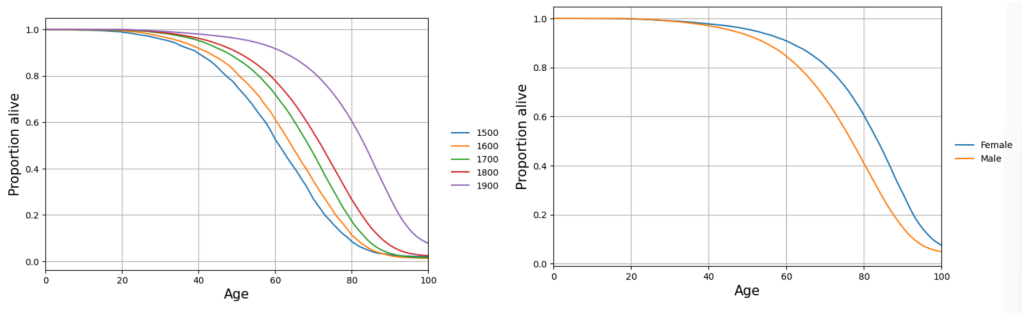

Figure 1: Marginal Survival functions categorized by era (left) and gender (right).

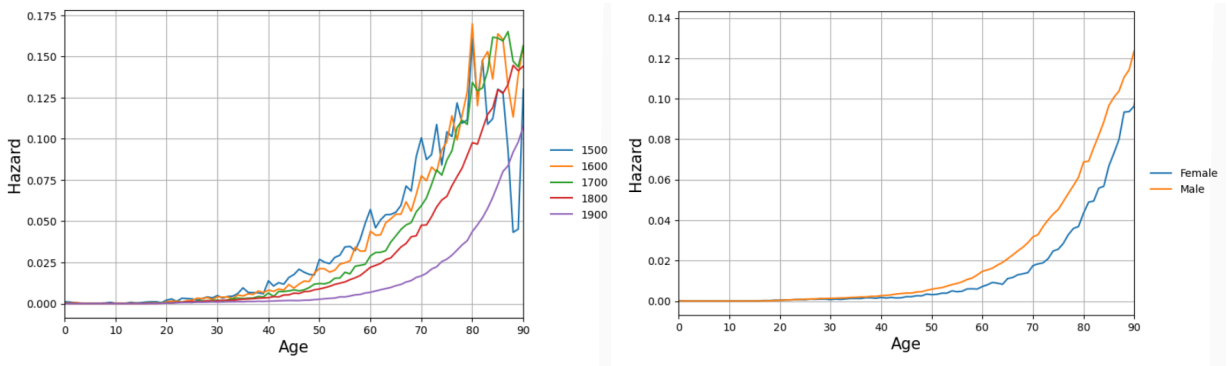

Figure 2: Marginal hazard functions categorized by era (left) and gender (right).

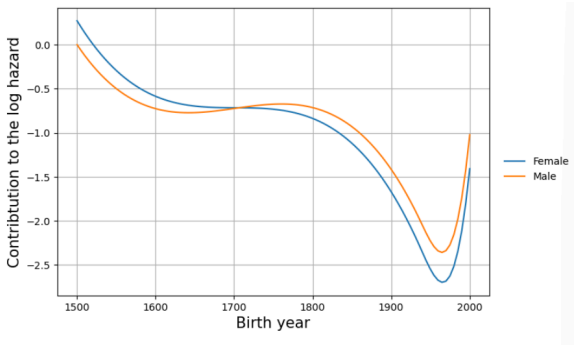

Figure 3: Using splines to model the main effect using linear terms for the interaction. This model is categorized with gender.