Occupational choices for the risk of death in different gender groups

Introduction

Survival analysis, a statistical technique, is employed to scrutinize the duration until the occurrence of a specific event of interest. In this report, we delve into the survival journey of a selected cohort within our study population and dissect the various factors that influence the time it takes for the event of interest to transpire. Notably, factors such as gender, occupation, and regional disparities may play pivotal roles in shaping an individual's risk of death as they age.

One of the central objectives of this report is to elucidate the intricate relationship between occupation and death, specifically when dissected by gender. By exploring this multifaceted connection, we aim to gain a more profound understanding of the nuanced ways in which occupational choices may impact the risk of death among individuals of different genders. This investigation is poised to provide valuable insights into the intersections of occupation and gender in the context of death.

Data

The data for this analysis consist of 2.2 million individuals' human time data including birth date, death date, lifespan, occupation, region, gender, etc, whose birth date ranges from 1500 to 1900 . Since only a subset of the subjects will die during our study, we censored 2022 to fill the missing lifespan with the difference of 2022 and birth date. In addition, we created the died variable to record the status of each individual during our study. Our left truncation in terms of year should be 1500, so we create the time by subtracting the birth date with 1500 and square root it to mitigate the skew.

Methodology

Semi-parametric proportional hazards (PH) model

$$h(t|X = x) = \exp(\beta' x)h_0(t).$$

In the equation, t represents time, $h(t|X = x)$ is the hazard function at time period t, which is also called instantaneous event rate, telling how the risk of the event happening changes over time t. $h_0(t)$ is the baseline hazard, which corresponds to the value of the hazard if all the independent variables are equal to 0. In our analysis, the baseline hazard has been adjusted to 1900 instead of 1500, because we are interested in the situation of recent years. With two kinds of gender and four categories of occupation, we constructed 8 baseline hazards for different combinations of gender and occupation. X contains a set of independent variables. β' are the regression coefficients for the independent variables, like time and region.

Specifically, we empirically conducted the 4 degree of freedom regression spline on time variables concerning the potential issues of polynomial regression. First, the polynomial regression might be badly scaled. If the power of polynomials is lower, the limited parameter could not capture the features of the dataset. If the power of polynomials is higher, it might predict with unexpected high values and low values which is not ideal. Secondly, the leverage of polynomial regression might cause the predicted response only weekly depending on its corresponding predictors, thus reducing the accuracy.

Result

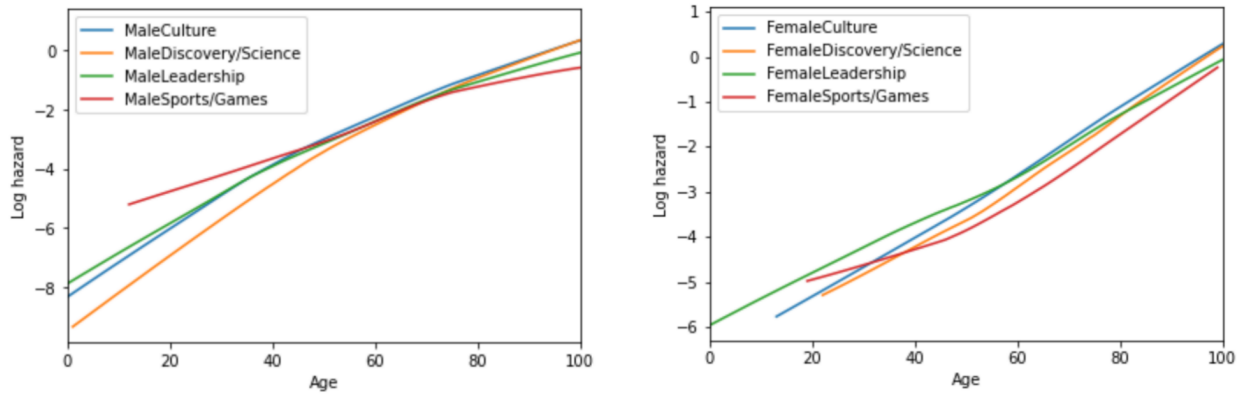
In our dataset, while the number of records for males outweighs that of females, the proportional distribution of various occupations across genders remains similar.

Table 1 Summary of Hazard Regression

	log HR	log HR SE	HR	t	P>|t|	[0.025	0.975]
un_region[T.America]	-0.1744	0.0521	0.8399	-3.3498	0.0008	0.7584	0.9302
un_region[T.Asia]	-0.1439	0.0541	0.8660	-2.6573	0.0079	0.7788	0.9630
un_region[T.Europe]	-0.1750	0.0516	0.8394	-3.3928	0.0007	0.7587	0.9287
un_region[T.Oceania]	-0.2170	0.0906	0.8049	-2.3954	0.0166	0.6740	0.9613
bs(time, 4)[0]	-3.1821	0.1523	0.0415	-20.8947	0.0000	0.0308	0.0559
bs(time, 4)[1]	1.8495	0.0778	6.3563	23.7849	0.0000	5.4578	7.4028
bs(time, 4)[2]	-2.9113	0.0836	0.0544	-34.8043	0.0000	0.0462	0.0641
bs(time, 4)[3]	-5.1085	0.3418	0.0060	-14.9450	0.0000	0.0031	0.0118

If we pay attention to the p-value associated with the HR in Table 1, a low p-value (typically below 0.05) indicates that the HR is statistically significant, suggesting that the predictor variable has a significant impact on the hazard rate and our regression with the predictors including region and 4 degree of freedom spline of time is valid.

Graph 1 Log Hazard over age for Male/Female occupations



Observing the data presented in Graph 1, where the x-axis represents age and the y-axis illustrates the logarithm of hazard, we discern trends within the realm of cultural occupations. In this context, we observe that the male group exhibits a steeper initial slope in hazard as age advances, which subsequently transitions into a relatively flatter trajectory. In contrast, the female group demonstrates a hazard rate that appears relatively consistent across various age groups. Further exploration of the male group's log-hazard curve reveals a noteworthy transition point occurring at approximately 60-80 years of age. Despite conducting a regression analysis on log-hazard, our findings still suggest that males working in cultural occupations tend to experience a reduction in the rate of death risk increase, in an exponential context, beyond the age range of 60-80 years.

For Discovery/Science, Leadership, Sports occupations, an intriguing pattern emerges when comparing the male and female groups. Initially, the male group displays a sharper upward

slope at the onset of age, which gradually levels off, while the female group exhibits an inversely flatter trajectory at the outset of age, which subsequently becomes more pronounced. Additionally, the transition points reveal distinctive age ranges for each gender group. For males, this pivotal transition occurs around the age range of 60-80, marking a notable shift in their hazard rates. Conversely, for females, the change point falls within the 40-60 age range. This implies that, beyond the age of 40-60, females experience an increasing rate of death risk, whereas males tend to undergo a reduction in the rate of death risk increase, especially within an exponential context. These findings underscore the unique dynamics at play within Discovery/Science, Leadership, Sports occupations and the different age-related patterns of death risk between genders.

Particularly noteworthy are the observations within the realm of sports occupations. Initially, among males, sports occupations exhibit the highest risk of mortality at the onset of age. In contrast, the female group, while also engaging in sports occupations, does not consistently exhibit the same elevated risk. Surprisingly, within the female group, Discovery/Science occupations emerge as the category with the highest death risk before the age of 30.