Project 4

Analysis of Age and Maximum Heart Rate: A Survival Analysis Approach

Weronika Dyszkiewicz, Katarzyna Mamla

Faculty of Mathematics and Information Science
Warsaw University of Technology

Warsaw, 27th January 2025

Table of contents

1. Project Overview
    - Description of dataset
    - Survival Analysis approach
2. Descriptive Analysis and Relationships
3. Kaplan-Meier method
    - Comparison of survival curves
4. Weibull Regression
    - Cox-Snell residuals
5. Cox Proportional Hazards Model
    - Can the Cox Proportional Hazards Model be used to predict age?
6. Conclusions

Background

- The dataset used for our project is based on the study of Sydó et al. (2014).
- This study aimed at knowing the relationship between age and the maximum heart rate during a 21-minute treadmill exercise known as Bruce protocol-standardized diagnostic test used in the evaluation of cardiac function and physical fitness, developed by American cardiologist Robert A. Bruce.

Figure: Treadmill for functional diagnostics for competitive athletes (1980) (source: wikipedia.org/Bruce_protocol).

Variables in the Data Set

- `Age`: Age of individuals (in years).
- `Female`: Gender indicator (0 = male, 1 = female).
- `res_hr`: Resting heart rate [bpm]
- `peak_hr`: Maximum heart rate during the exercise [bpm].
- `Finished.test`: Whether the participant completed the exercise (0 = No, 1 = Yes).

Objective of the Project

- Investigate the relationship between **age** and **maximum heart rate**.
- Use **descriptive analysis** to summarize key data features and visualize the patterns between age and maximum heart rate.
- Apply **survival analysis** techniques to treat peak heart rate as a time-to-event variable, exploring how age and other factors impact the time to reach maximum heart rate.
- Explore gender differences in peak heart rate, considering the role of resting heart rate and exercise completion.
- Apply **Weibull regression** to understand the effect of age and sex on the time to reach peak heart rate and assess the significance of these factors.
- Estimate **a Cox proportional hazards model** to examine the influence of age and sex on the time to reach peak heart rate, and evaluate their statistical significance.

Methodology Comparison

Approach in Sydó et al. (2014):

- Traditional statistical methods (e.g. regression analysis) were used to model the relationship between age, gender, and peak heart rate.
- Linear regression to derive sex-specific formulas for predicting peak heart rate (e.g. HR=220 - 0.95 · age for men, adjusted formula for women).
- Focused on differences in heart rate responses by sex and age.

Proposed Survival Analysis Approach:

- Survival analysis interprets `peak_hr` as the time to an event (e.g. reaching maximum exertion).
- Kaplan-Meier curves estimate survival probabilities (likelihood of reaching peak heart rate) for males and females.
- Uses Weibull regression and Cox proportional hazards models to assess the influence of variables (e.g. age, gender) on peak heart rate.

Justification for Survival Analysis Approach

Survival analysis is a strong choice due to the following:

- **Censoring in the Data:** The `Finished.test` acts as a right-censoring indicator, handling incomplete data for participants who did not complete the treadmill test. Survival analysis methods, like Kaplan-Meier and Cox regression, can account for this and provide robust estimates.
- **Time-to-Event Nature of `peak_hr`:** By interpreting `peak_hr` as the time to an event, survival analysis aligns naturally with the concept of endurance under stress. This approach models how factors like age and gender influence the ability to sustain peak exertion during exercise.

Data Interpretation in Survival Analysis Context

- **Event**: The event of interest is reaching the maximum heart rate (`peak_hr`).
- **Censoring:** Participants who did not complete the test ('dropped out') are treated as censored data.
  - **Right-censored data**: We only know the test duration until the dropout, not the exact time to the event.
  - Example: A participant stops due to fatigue or health issues before reaching `peak_hr`.
- **Finished.test as the right censoring indicator:**
  - `Finished.test` = **0**: The event (reaching maximum heart rate) was fully observed.
  - `Finished.test` = **1**: The participant did not reach peak heart rate, so the observation is censored.
- Survival analysis accounts for these cases, preserving their contribution to the model without introducing bias.

Summary Statistics

Variable	Min	1st Qu.	Median	Mean	Max
Age	40.00	47.00	53.00	54.23	87.00
res_HR	37.00	67.00	74.00	75.20	148.00
peak_HR	72.00	160.00	170.00	168.40	255.00

Table: Summary statistics for variables in dataset.

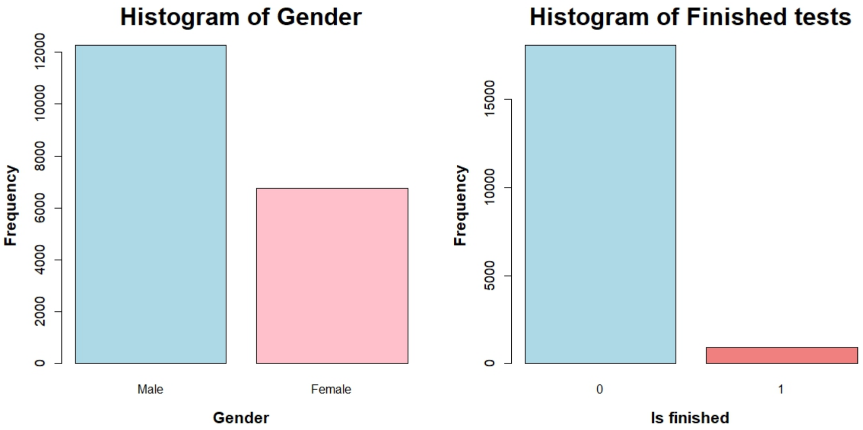

Figure: Distribution of gender and finished test categories. The dataset includes 12,264 observations for males (`Female` = 0) and 6,748 observations for females (`Female` = 1). The dataset contains 18,072 observations (`Finished.test` = 0) and 940 censored observations (`Finished.test` = 1).

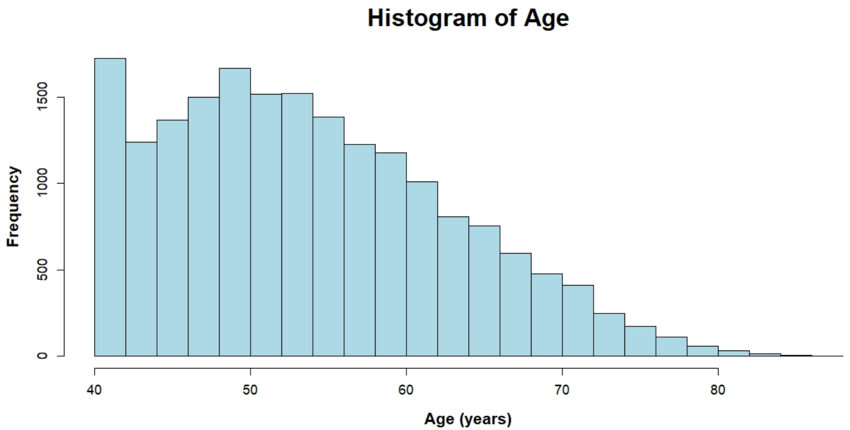

Figure: Distribution of participant ages, ranging from 40 to 87 years, with a higher frequency of participants in the 40–60 years range.

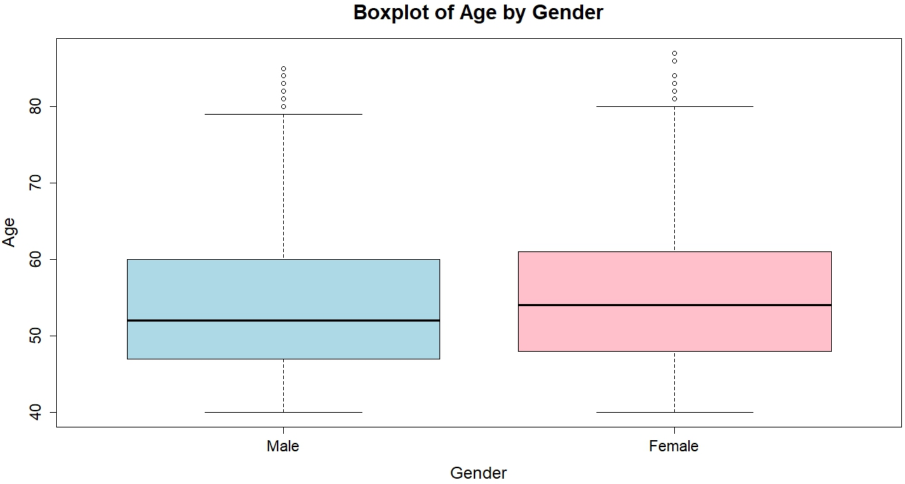

Figure: Comparison of the age distribution between males and females, showing similar medians and ranges for both groups.

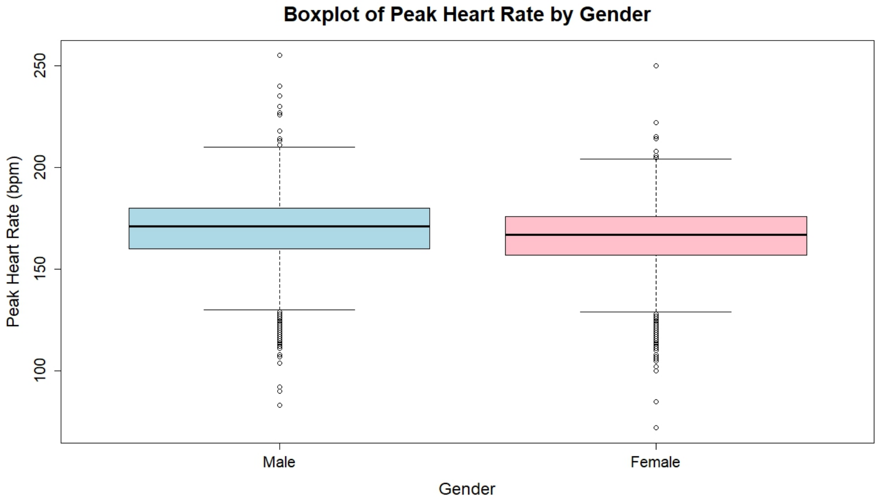

Figure: The boxplots display the peak heart rate distributions for males and females. The median peak heart rate for each group differs significantly, what was reflected in the results of the Mann-Whitney U Test (the peak heart rate distributions for both males and females were found to be non-normally distributed by Kolmogorov-Smirnov tests).

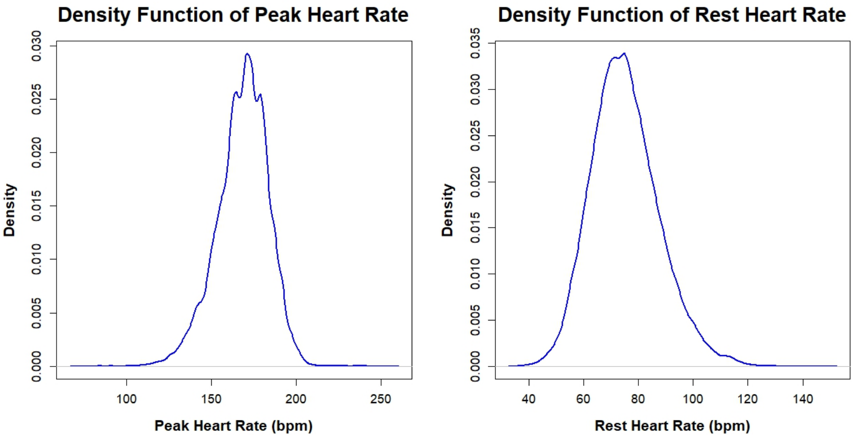

Figure: The most individuals in the dataset achieved peak heart rates in the typical range for exercise, but some individuals experienced lower or higher than average peak rates. Most values of resting heart rate lie in the range of approximately 60 to 80 bpm, indicating a normal rate for the majority.

Relationship between Age and Peak Heart Rate

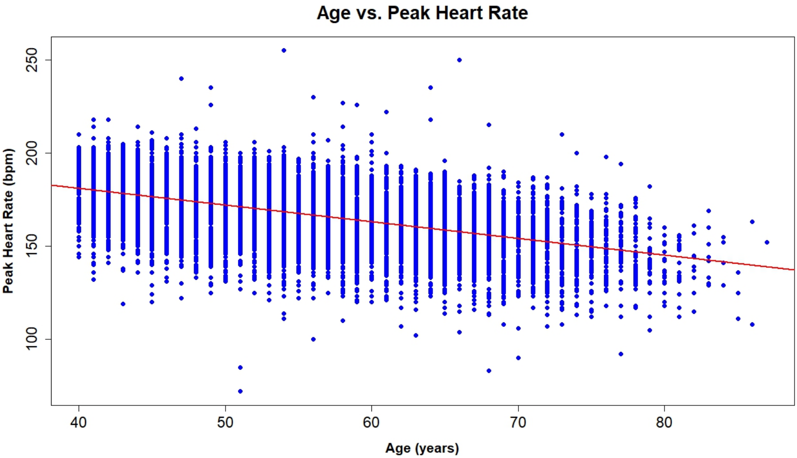

Figure: There is a clear negative trend between Age and Peak Heart Rate: as age increases, peak heart rate decreases. The significant scatter around the regression lines indicates variability in peak heart rates for individuals of the same age.

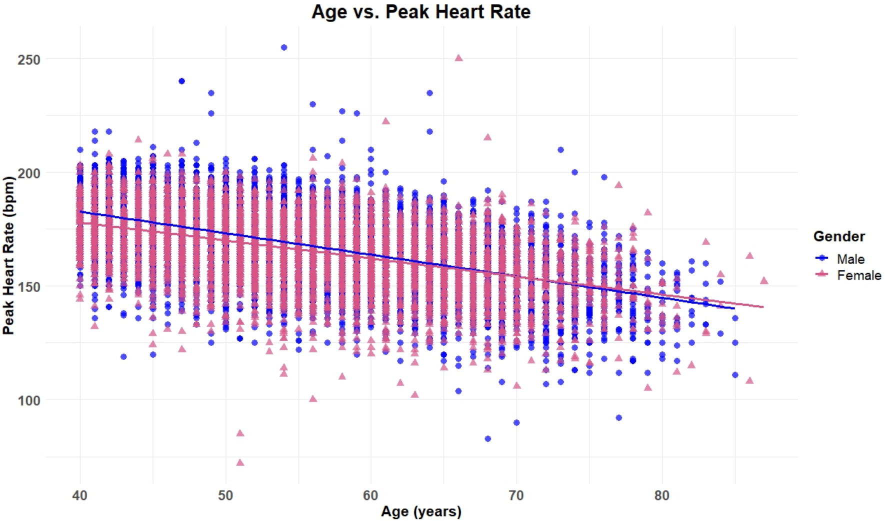

Figure: The slopes of the regression lines for males and females are similar, indicating that the decline in peak heart rate with age occurs at a similar rate for both genders. However, the lines suggest slightly higher peak heart rates for males than females at most ages.

Measures of Association: Correlation

1. **Pearson correlation between Age and Peak Heart Rate**:

$$r_{xy} = -0.54$$

   There is a moderate negative relationship between the two variables, meaning that as one variable increases, the other tends to decrease, but not strongly. In our case as age increases, the maximum heart rate decreases.

2. **Linear regression model to predict Peak Heart Rate based on Age:** Regression equations for peak HR were

$$217.32 - 0.9 \cdot \text{Age}$$

   The slope represents the rate of change in peak heart rate for each unit increase in age.

   - This indicates that for each additional year of age, the maximum heart rate (peak heart rate) decreases by 0.9 beats per minute (bpm) on average.
   - The relationship is considered statistically significant (p-value: $< 2.2e - 16$).

Kaplan-Meier Estimator: Overview

Kaplan-Meier Estimator:

- Non-parametric method used to estimate the survival function from time-to-event data.
- Incorporate right-censored data (cases where the event of interest is not observed during the study period).

Formula:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

In our case:

- $\hat{S}(t)$: Estimated survival probability that an individual's heart rate has not yet reached its peak (event hasn't occurred) by a given t (specific heart rate value).
- t_i: A specific peak heart rate value where an event (reaching peak heart rate) occurs.
- d_i: Number of individuals who reach their peak heart rate at t_i.
- n_i: Number of individuals still being tested and at risk of reaching their peak heart rate just before t_i.

Kaplan-Meier Survival Curves

Using the Kaplan-Meier estimator, we estimate the survival curve for `peak_hr` for all individuals without covariates

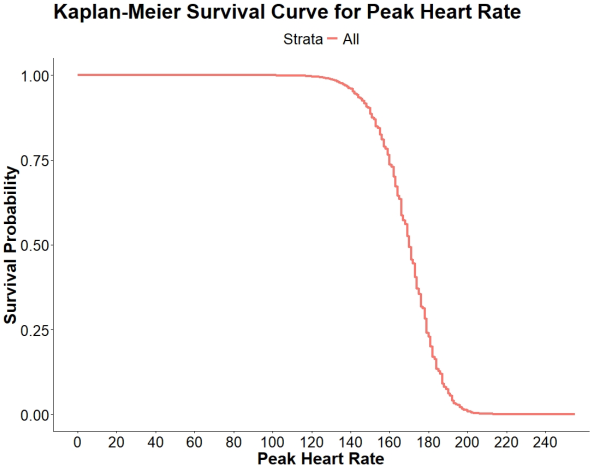

Figure: Kaplan-Meier survival curve estimating the probability of participants (overall, male, female) not reaching their peak heart rate during the exercise test.

Comparison of Peak Heart Rate: Male vs Female

Next we constructed Kaplan-Meier curves for two groups.

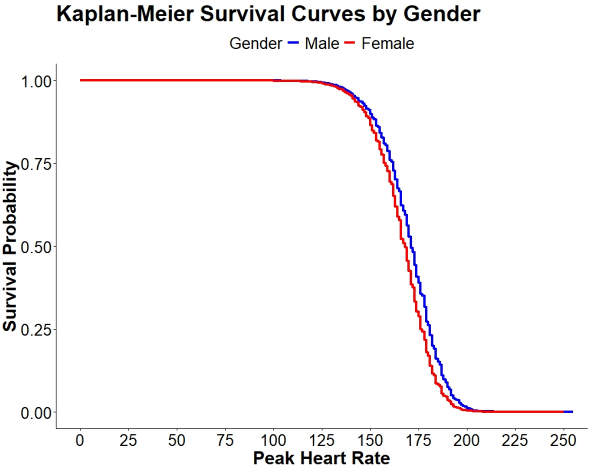

Figure: Kaplan-Meier survival curves for males (blue) and females (red). The higher survival probabilities for males indicate that they tend to sustain lower heart rates for longer during the exercise, while females reach their peak heart rates more quickly.

We perform two statistical tests to compare `peak_hr` between males and females:

- Suitable tests: Log-rank test or Peto-Peto.
- We hypothesize that there is a significant difference in peak heart rate between males and females.

Log-rank: tests the overall difference in survival distributions between groups.
Peto-Peto: focuses more on the early event times (i.e., differences between groups in the initial stages of the event).

The null hypothesis is the same for log-rank and Peto-Peto test:

$$H_0 : S_1(t) = S_2(t) \quad \text{for all } t,$$

where $S_1(t)$ and $S_2(t)$ are the survival functions for groups 1 and 2.

Example: In both tests p-value is $< 2e - 16$ and this indicates a significant difference between the survival curves for males and females. It means that the null hypothesis of no difference between the groups should be rejected.

Weibull Regression Analysis

We perform a Weibull regression analysis to model `peak_hr` as the outcome variable and the remaining variables in the data set as covariates. It assumes the outcome follows a Weibull distribution.

The fit of a Weibull distribution to data can be visually assessed using a Weibull plot. The Weibull plot is a plot of the empirical cumulative distribution function $\widehat{F}(t_i)$ of data on special axes in a type of Q–Q plot. The axes are

$$\ln(x) \ \text{versus} \ \ln(-\ln(1-\widehat{F}(t_i))).$$

The reason for this change of variables is that the cumulative distribution function can be linearized.

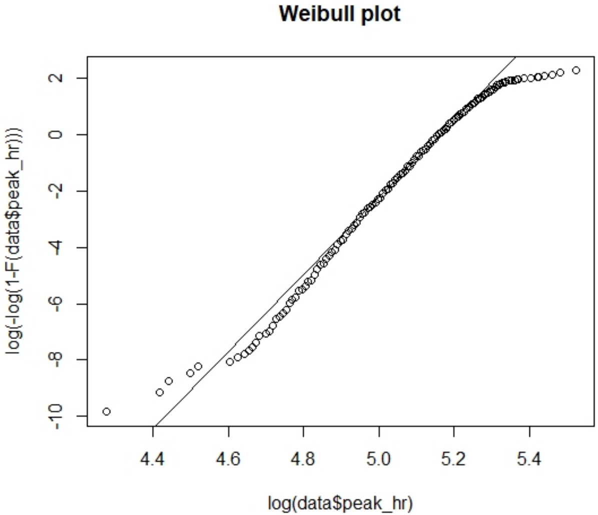

Figure: The alignment of most data points with the line suggests that the Weibull distribution is a reasonable model for our data, except for possible deviations at the extremes.

Importance of Age and Sex based on Weibull model

We use two approaches to assess the significance of Age and Sex as covariates in explaining the outcome variable.

1. Firstly, we use a forward stepwise selection method. We start with a null model that has no covariates, and then progressively add covariates such as Age, Sex and Resting Heart Rate to see if they significantly improve the model.

Model	**AIC**
`Surv_obj ~ 1`	145973.4
`Surv_obj ~ data$age`	140690.0
`Surv_obj ~ data$age + data$rest_hr`	139130.0
`Surv_obj ~ data$age + data$rest_hr + data$female`	138739.8

2. Then, we use ANOVA test to compare two models to check whether the inclusion of additional variables improves the fit of the model significantly. Both Sex and Age are significant predictors of the outcome variable, with extremely low p-values indicating their strong influence. The substantial decrease in -2 Log-Likelihood and the large deviance when adding each predictor further support their importance in improving the model's fit.

Residual Analysis: Weibull Model

After fitting the Weibull model, we perform a residual analysis:

- Use Cox-Snell residuals to check the fit of the model.
- Perform a statistical test (e.g., Kolmogorov-Smirnov test) to validate the distribution of the residuals.

A model fits the data well if the Cox-Snell residuals follow an exponential distribution of parameter 1. Komologorov-Smirnov test is used to assess whether this is the case.

Our dataset: p-value is extremely small, indicating that the null hypothesis that the data follows the exponential distribution is rejected. To confirm it, we visualised the comparison to the expected distribution.

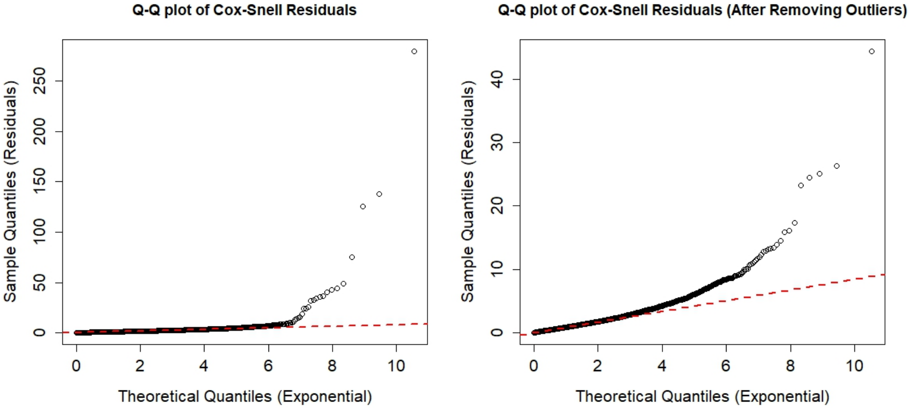

Figure: Both sets of residuals deviate from the expected exponential distribution, but removing outliers leads to a small reduction in this deviation. However, neither dataset appears to fit well to the assumed exponential distribution based on these plots.

Cox Proportional Hazards Model - Key Results

We estimate a Cox proportional hazards model for peak_hr.
The covariates included in the model are Age, Female, and Rest_hr.
The baseline category for female is 0 (male). The model equation is given by:

$$h(t) = h_0(t) \cdot \exp\left(0.06835 \cdot \text{age} + 0.30228 \cdot \text{female1} - 0.02583 \cdot \text{rest\_hr}\right)$$

Variable	β (coef)	e^{β} (Hazard rate)	$e^{-\beta}$	**p-value**
data$age	0.0683	1.071	0.934	<2e-16 ***
data$female1	0.3023	1.353	0.739	<2e-16 ***
data$res_HR	-0.0258	0.975	1.026	<2e-16 ***

Table: Summary of Cox Proportional Hazards Model

Interpretation:

- For data$age, each additional year increases the hazard rate by 7% ($e^{0.0683} \approx 1.071$).
- For data$female1, being female increases the hazard rate by 35% compared to males ($e^{0.3023} \approx 1.353$).
- For data$res_HR, each unit increase in resting heart rate decreases the hazard rate by 2.5% ($e^{-0.0258} \approx 0.975$).

Significance of Age and Sex in the Cox Model

We tested the significance of age and sex using two models:

- **Age**: A reduced model excluding `female` was compared with the full model using the Likelihood Ratio Test (LRT). The result $(p < 2.2 \times 10^{-16})$ indicates that age significantly influences the outcome variable.
- **Sex**: Similarly, a reduced model excluding age was compared to the full model. The result $(p < 2.2 \times 10^{-16})$ shows that `sex` (female vs male) also significantly influences the outcome variable.

Stepwise Model: A stepwise AIC model confirmed that all variables (age, `female`, and `rest_hr`) are significant predictors of the outcome.

Prediction of Age Based on the Estimated Model

The Cox proportional hazards model is typically used to assess the impact of covariates (such as age, gender, resting heart rate) on the time to an event (e.g. reaching peak heart rate). While the model does not directly predict age, it provides valuable information about how age influences the likelihood of reaching the event.

To predict the age of an individual, we would need to rearrange the model. If we have an individual's hazard of reaching peak heart rate, along with other known covariates (e.g. resting heart rate, gender), we could use the estimated coefficients from the model to solve for their age. This is done by rearranging the hazard function to isolate age.

$$\text{Age} = \frac{\ln\left(\frac{h(t)}{h_0(t)}\right) - \beta \cdot \text{Other Covariates}}{\beta_{Age}}.$$

While this approach allows for an estimate of age, it's important to remember that the Cox model is designed to quantify relative hazard rather than to directly predict a specific individual's age.

Thus, while the model can provide insights into age's role in the timing of an event, predicting age from the hazard rate is not its intended use.

Summary

- Unlike the method used in the article, our approach takes advantage of the time-to-event nature of the data, using survival analysis methods.
- Survival analysis accounts for participants who didn't reach their peak heart rate (censored data), ensuring unbiased estimates by including their incomplete outcomes in the model.
- The visualization of the relationship suggest variability in peak heart rates for individuals of the same age, potentially influenced by factors like fitness level, health status, or measurement variability.
- Log-rank and Peto-Peto tests confirmed significant gender differences in peak heart rate.
- Removing outliers does not improve the fit of the Weibull regression, as indicated by the Cox-Snell residuals.
- Each additional year increases the hazard of reaching peak heart rate by 7% (Cox model). Females reach peak heart rate faster, with a hazard 35% higher than males.
- The Weibull model and Cox model validate age, gender, and resting heart rate as significant predictors of peak heart rate.

Bibliography

[1] Nóra Sydó, Sahar S. Abdelmoneim, Sharon L. Mulvagh, Béla Merkely, Martha Gulati, Thomas G. Allison *Relationship Between Exercise Heart Rate and Age in Men vs Women*, Mayo Clinic Proceedings, December 2014, Volume 89, Issue 12, Pages 1664-1672. `http://dx.doi.org/10.1016/j.mayocp.2014.08.018`

[2] `https://en.wikipedia.org/wiki/Weibull_distribution`

[3] Lectures of the Biostatistics course 2024/2025, MiNI, Politechnika Warszawska

Thank you for your attention!