2AMS11 - Survival Analysis for Data Scientists

Group Assignment - 2024/2025

INSTRUCTIONS:

- You should work in groups of **three** or **four** students. Exceptions are possible with an adequate justification, but you must contact me at least two weeks before the due-date with such a request.
- Your report should be in **English** and be submitted via CANVAS, as a **single pdf file**. The report is due on **October 20th, 2024 at 23:59**.
- Ideally, you should use the statistical software `R` for this project (but this is not mandatory and you can use a different computing language). Please include the code you used (properly commented) in a separate document. You will not be assessed on the code. That being said, **your report should contain all the information needed to understand and be able to replicate your analysis and results**, without having to look at the code.
- Use of Artificial Intelligence (AI) tools and/or Large Language Models (LLMs) is NOT ALLOWED, particularly to prepare the report and code. If the usage of AI and LLM tools is suspected in breach of these directives, the situation will be reported to the examination committee.

Preamble

The main goal of this project is to gain experience in conducting sound data analysis and inference in scenarios involving time-to-event data. This will require you to perform exploratory data analysis, critically develop statistical models and think carefully about the assumptions being made, and use the tools of survival analysis and reliability theory to draw meaningful inference on the processes underlying the data you have available.

It is crucial to document clearly your work. Particularly important when using statistical tools, it is essential to identify clearly what assumptions are being made, what models are being considered. These choices must be supported with qualitative and quantitative arguments. In some parts of this report you must tackle estimation and hypothesis testing problems. You must clearly describe these in the context of your model (e.g., indicate what is the null and alternative hypothesis being considered).

When preparing your report you should have in mind a stakeholder that is a technically minded person at the company. Your report should be well structured and feature an introduction, explaining the general context and goals of the project. This should be followed by modeling considerations, where all the assumptions made are clearly described and motivated (based on qualitative and possibly quantitative arguments). It is important to be very critical about any choices you make. In addition, you should carefully describe the methods used. In particular, when using statistical tests, clearly indicate the test you are using, what are the hypothesis being tested, and what are the main assumptions of the testing procedure. Your report should end with a short conclusion summarizing your findings.

Although you may deviate from this structure, a suitable structure might be:

- **Abstract** (a very brief description of the challenge, methods used and main findings)
- **Introduction**
- **Methodology and Modeling Assumptions**
- **Results**
- **Conclusions and recommendations**

Introduction

This assignment pertains warranty data from a model of a fictitious robot vacuum cleaner, the *DirtSlurper3000*, which is manufactured by *IButler*, a company focussed on robotic solutions for everyday life. This robot is highly advanced, and uses ultra-deep learning technology for ultra-deep cleaning ability. Furthermore, the manufacturers of the DirtSlurper3000 are keen in offering a 5 year limited warranty, meaning that, if the device failed within a 5 year period counting from the registration date, the owner can send it for repair without costs (within the normal limitations of wear and tear).

Robot vacuum cleaners are complex machines, relying on a number of sensors, motion actuators, a high-performance battery and so on. Of particular interest for this assignment are three major components of the robot, namely, the **Battery** (that powers all the systems in the whole device), the **IR sensor**, which is used to create "images" of the environment, and get rough distance estimates, and the **Impact sensor**, which allows the robot to physically "touch" and "bump" into objects in the environment. The three components are essential for the working of the robot. DirtSlurpers are proud members of the Internet-of-Things, and therefore share data with the manufacturer to monitor and improve their products. After purchasing a robot, the owner needs to go through a registration process, where they provide some information about the environment the robot will operate in. Among other questions, they are asked if there are cats or dogs in the house. While in operation, and after a number of cleaning runs, the robot also updates a score indicating the perceived amount of carpeting in the environment. This *carpet score* (taking values between 1 and 9, where a large number indicates a highly carpeted environment) is used to adjust the operation of the robot. Finally, data about the usage patterns is also collected, and in particular the total cleaning time from registration is also logged.

The goal of this project is to characterize the lifetime of the three components of the robot listed above, namely, the Battery, IR sensor, and Impact sensor. The manufacturer of the DirtSlurper3000 has collected data from registered devices from January 2015 onwards, until the the end of December 2019. Some of the registered devices were sent for repair in that period. If that was the case, the date in which request for repair was issued is noted, as well as the total cleaning time up to that point. In addition, also noted are the part or parts that have failed (note that the robot has other components besides the three we are interested in). If a device was not issued a repair request before 31 December 2019 then the total cleaning time up to that date is noted instead. For all the devices in the database, the declaration if (hairy) pets are in the environment being cleaned, as well as the carpet score at the time the data was collected is also registered. The data is available in CANVAS in the file `DirtSlurper3000.csv`.

The managers of IButler feel the number of DirtSlurper3000 sent for repair is higher than what they anticipated, and are concerned some of the components used in the robot do not conform to the required specifications. However, to pursue any action (legal or not) they need to find strong evidence towards such claims, and therefore it is crucial to accurately characterize the lifetime of each of the components.

1 Exploratory data analysis

Import the data into `R` (or the software of your choice). The commands `read.table` or `read.csv` might come in handy. Get acquainted with the dataset, and identify possible issues in the data that might require some pre-processing.

2 Modeling

The main goal of the project is to characterize the lifetime of three main components of the vacuum cleaner, namely, the Battery, the IR sensor, and the Impact sensor. Note that the lifetime of the components might depend on various factors, such as the properties of the environment the robot is operating in, as well as the intensity of usage. Think carefully about the modeling assumptions you can make, including the possible censoring and truncation mechanisms present in the way data was collected. Carefully justify all the choices and assumptions made.

3 Inference

Based on the model choices above characterize the lifetime of the three components of interest. Note that the lifetime of each component might depend on the way it is used. In other words, environmental factors might influence the reliability of each part. For each of the three components identify which factors play a role in the characterization of the lifetime. Contrast your findings with some of the information provided by the manufacturers of each component:

- The IR sensor manufacturer guarantees that $L_{10} \geq 2000$ days. **Note:** L_{10} is a widely used quantity in reliability engineering. In statistical terms this is the 10% quantile of the distribution. In other words, 90% of the parts will survive at least 2000 days.
- The battery manufacturer guarantees that the battery will retain at least 80% capacity after 2400 hours of continuous use. Excluding extreme use cases (i.e., 2400 hours or more of continuous use) the manufacturer guarantees $L_{10} \geq 1000$ days. For repair purposes, a battery with less than 80% capacity needs to be replaced.
- The Impact sensor is built by IButler and an integral part of the design of the DirtSlurper3000. No specifications concerning reliability have been provided to you.