



UMEÅ UNIVERSITY

Failure Probability and Lifetime Estimation for Industrial Robots

A Logistic Regression and Lifetime
Analysis Approach

Authors:

Erik Fahlbeck Carlsson

Martin Herbert

Supervisors:

Leif Nilsson

Master thesis, 30 Credits

M.SC. Industrial Engineering and Management, Industrial Statistics, 300 Credits

Spring term 2023

Copyright © 2023 Erik Fahlbeck Carlsson and Martin Herbert
All rights reserved

Failure Probability and Lifetime Estimation for Industrial Robots

Submitted in partial fulfillment of the requirements for the degree Master of Science in
Industrial Engineering and Management
Department of Mathematics and Mathematical Statistics
Umeå University
SE – 901 87 Umeå, Sweden

Supervisor:
Leif Nilsson, Umeå University

Examiner:
Håkan Lindkvist, Umeå University

Abstract

The ability to handle and process data for information extraction is getting more and more important. Using extracted data from the business to improve productivity is seen as an important part in developing the business processes. In this thesis, industrial robots and their survival times are analyzed. The work is about predicting the probability that a specific robot will fail during a specified time period. Also, survival analysis is conducted where the median lifetime and conditional median lifetime for industrial robots are estimated.

Two approaches are used, logistic regression and survival analysis. A logistic regression model is made to predict the probability for different industrial robots to break during a specified time period. The logistic model achieves an accuracy of 0.694 with even higher accuracy regarding high – and low risk robots. The survival analysis uses a Cox PH model to check validity for proportional hazards and then a parametric model with Weibull distribution is fitted. The parametrical survival model is used to estimate the median lifetime and the remaining median lifetime for the robots. The estimated probabilities and lifetimes can be used as an indication of which robots are in risk of failure.

KEYWORDS: LIFETIME ANALYSIS, LOGISTIC REGRESSION, PREDICTION, LIFETIME ESTIMATION, INDUSTRIAL ROBOTS

Sammanfattning

SANNOLIKHET FÖR FEL OCH LIVSLÄNGDSESTIMERING FÖR INDUSTRIELLA ROBOTAR: MED LOGISTISK REGRESSION OCH LIVSLÄNGDSANALYS

Förmågan att hantera och använda data för att få ut information blir mer och mer viktigt. Att använda extraerad data från verksamheten i syfte att utveckla produktiviteten är sedd som en viktig del i utvecklingen av organisatoriska processer. I den här uppsatsen analyseras industriella robotar och deras överlevnadstider där sannolikheten att en specifik robot kommer gå sönder under en specificerad tidsperiod predikteras. Dessutom genomförs en överlevnadsanalys där medianlivstiden och den villkorliga medianlivstiden för industriella robotar estimeras.

Två tillvägagångssätt används, logistisk regression och överlevnadsanalys. En logistisk regressionsmodell är gjord för att prediktera sannolikheten att olika industriella robotar går sönder under en specificerad tidsperiod. Den logistiska modellen når en precision på 0.694 med ännu högre precision för robotar med hög – och låg risk. Överlevnadsanalysen använder en Cox PH modell för att kolla giltigheten gällande proportionella hazards för att sedan inpassa en parametrisk model med Weibullfördelning. Den parametriska överlevnadsmodellen används för att estimeras medianlivstiden och den återstående medianlivstiden för robotar. De estimerade sannolikheterna och livslängderna kan användas som en indikation för vilka robotar som skulle kunna gå sönder.

NYCKELORD: LIVSLÄNGDSANALYS, LOGISTISK REGRESSION, PREDIKTION,
LIVSLÄNGDSESTIMERING, INDUSTRIELLA ROBOTAR

Acknowledgements

We would like to thank our supervisor from Umeå University, Leif Nilsson, for your domain knowledge and great experience. The discussions with you have always been constructive and helpful. Your calm approach and clear instructions have without a doubt helped us come through when things looked grim.

Secondly, thanks to the company for giving us the opportunity for writing this thesis. We would also like to send our gratitude to fellow employees at the company for showing curiosity and being helpful towards us.

Contents

List of figures	viii
List of tables	ix
List of symbols	x
1 Introduction	1
1.1 Problem statement	1
1.2 Background	1
1.3 Aim	2
1.4 Related work	2
1.5 Motivation of approach	3
1.6 Outline	3
2 Theory	4
2.1 Logistic regression	4
2.1.1 Classification logistic regression	4
2.1.2 Overfitting	4
2.1.3 Cross-validation	4
2.1.4 Leave-One-Out Cross-Validation	4
2.2 Akaike Information Criterion	4
2.3 Log-Likelihood	5
2.4 Pearson correlation coefficient	5
2.5 Step-wise Feature Selection	5
2.5.1 Forward selection	5
2.5.2 Backward elimination	5
2.5.3 Bidirectional elimination	5
2.6 Confusion Matrix	5
2.6.1 Precision, Sensitivity and specificity measure	6
2.7 Outlier	6
2.8 Moment of inertia	6
2.9 Normalization	6
2.10 Survival analysis	6
2.10.1 Survivor function	7
2.10.2 Hazard function	7
2.10.3 Kaplan-Meier curve	7
2.10.4 Cox Proportional Hazard model	7
2.10.5 Weibull distribution	8
2.10.6 Fitting a Weibull survival model with covariates	8
2.10.7 Expected lifetime and median lifetime	8
2.10.8 Conditional survival analysis	8
2.10.9 Schoenfeld's global test	9
3 Method	10
3.1 Description of the production line	10
3.1.1 Robots in the production	10
3.2 Description of the data	11
3.3 Data handling	12
3.4 Description of final data	13
3.5 Logistic Regression	13
3.6 Survival analysis	14

4	Results	16
4.1	Logistic regression classification	16
4.2	Survival analysis	19
4.2.1	Cox PH model	19
4.2.2	Parametric model	22
4.2.3	Expected robot lifetimes	24
5	Discussion	27
5.1	Data limitations	27
5.2	Logistic regression	27
5.3	Survival analysis	28
6	Conclusions	30
7	Recommendations	32
	References	33
	Appendix	34
A.1	Additional Figures	34

List of Figures

1	An example of a six axis industrial robot.	10
2	Correlation between covariates in the full logistic model.	17
3	Correlation between covariates in the final logistic model.	18
4	The Schoenfeld residual plot for the covariate Mass. The x-axis show the active usage time for the robots. The y-axis show the value of the Schonefeld residuals, the axis is denoted Beta(t) for Mass since it show if the Beta-value needs to change for different times. The line show the time component of Beta and the dotted lines show a 95% confidence interval for the estimation. If zero is within the confidence interval for all time points, this indicate that the Beta-value does not change with time.	21
5	The Kaplan-Meier survival curve (black step function) and estimated Weibull survival curve (red smooth function) when all covariates are assumed to be zero. The dashed lines are the marked 95-percent confidence interval of the Kaplan-Meier curve.	23
6	The hazard function from the data (black and not smooth function) to the estimated Weibull hazard function (red smooth function) when all covariates are assumed to be zero.	24
7	Predicted median lifetime against actual lifetime, both in active working hours, for censored data and the final survival analysis model.	25
8	Comparison between the estimated risk for a nine year period the logistic model and the predicted median lifetime of active working hours from the final survival model.	26
A.1	The Schoenfeld residual plot for covariates Mass, CoGz, CoGL and Jx0.	34
A.2	The Schoenfeld residual plot for covariates J5, J6, FrameCoGy and FrameCoGz.	35
A.3	The Schoenfeld residual plot for covariates UpperArmMass, UpperArmCoGx, UpperArmCoGy and StrainAxis1.	36
A.4	The Schoenfeld residual plot for covariates StrainAxis2, StrainAxis5 and Mass*J6.	37
A.5	The Kaplan-Meier curve and estimated Exponential survival curve.	38
A.6	The Kaplan-Meier curve and estimated Log-logistic survival curve.	38
A.7	The Kaplan-Meier curve and estimated Gamma survival curve.	39
A.8	The Kaplan-Meier curve and estimated Log-normal survival curve.	39
A.9	The corresponding hazard function to the Exponential survival curve.	40
A.10	The corresponding hazard function to the Log-logistic survival curve.	40
A.11	The corresponding hazard function to the Log-Gamma survival curve.	41
A.12	The corresponding hazard function to the Log-normal survival curve.	41

List of Tables

1	Confusion matrix for a two class classification.	5
2	Variables in the worked data set. *Explanatory variable only in the logistic data set. **Response variable together with Y_{event} only in the survival data set.	13
3	Structure of the final data set for logistic regression.	13
4	Chosen interactions.	14
5	Structure of the final data set for survival analysis.	14
6	Significant covariates in full logistic regression model.	16
7	AIC sores logistic regression models.	18
8	Confusion matrix logistic regression from LOOCV.	18
9	Performance of logistic regression model, evaluated with LOOCV.	18
10	Covariates of final logistic regression model.	19
11	Estimated probability of failure with the final logistic model during the studied time period. 10 out of 111 randomly selected robots displayed.	19
12	Significant covariates in full survival analysis model.	19
13	AIC sores survival analysis model.	19
14	Covariates of final survival analysis model.	20
15	Output of the performed Schoenfeld's global test.	22
16	AIC-scores for each distribution.	22
17	Covariates of final Weibull distribution model.	24
18	Estimated number of active hours from installation date for different robots and estimated number of remaining active hours from the end of the studied period. Example with 10 of 111 randomly selected robots displayed.	25

List of Ackronyms

CBM - Condition based maintenance

PH - Proportional hazard

LOOCV - Leave-one-out cross-validation

AIC - Akaike information criterion

TN - True negative

FP - False positive

FN - False negative

TP - True positive

CoG - Center of gravity

1 Introduction

1.1 Problem statement

When a robot failure occurs, it can bring parts of the production line to a standstill, as different sections in the production process is dependent on each other. The cost of the production not running when planned can be high. A robot failure also leads to a cost to repair the robot or to purchase a new robot. Not knowing which robots are in risk of failure makes it hard to plan which robots that could be exchanged for a more powerful robot type. It also makes it hard to know how many robots can be expected to fail during the coming years and therefore hard to plan repairs and installations of new robots. If it is expected that many robots will be exchanged during the coming years, this can be valuable information. The installations of new robots can then be planned longer in advance and don't need to come as a surprise and hence could effect the production less.

The problem with robot failure is not only the risk of a section of the production line to stand still. Having to change engines, gearboxes or buy and install new robots is expensive. It would obviously be in the interest of ██████ to increase the lifetime of robots and robot components, as this would lead to both lower costs of maintaining the robots and lower purchasing costs. To make it possible to increase the lifetime of robots, finding what factors effecting the lifetime of a robot is of interest. This gives the opportunity to evaluate robot data, analyzing if it's possible to find any significant factors behind robot malfunction through statistical models and predicting when robots might break.

1.2 Background

The biggest production facility for ██████ regarding ██████ are located in ██████. The facility is modern with robot production based manufacturing and computerized monitoring systems. The facility has a capability of producing ██████. The production process is complicated, involving many steps with a lot of different treatments and parts. Hence, the staff is required to be skilled in the production process, being able to react quickly to different circumstances regarding the manufacturing process. The production process is divided into different sections. This thesis is done within the section of production called ██████. The most common tasks for the robots in this section is welding and handling/moving different objects in the production. Many robots are working here and are vital for the production to proceed. Since every robot has a specific task to perform, one errant robot can mean that parts of the production line needs to stop. A stop in the production line can be costly as it possibly makes the output of the factory lower than planned. To reach the planned production rate, overtime is sometimes needed to meet the output demand.

A trend in the industry is that more and more data is collected. The increased amount of data is opening new opportunities. Terms like big data, data analytics and machine learning are today widely used. Big data is a term which describes a large amount of data, this data contain information that can be analysed. This is where the big data analytics comes in, which is the process of examining large data sets to uncover hidden patterns. Sometimes patterns can be found using easier methods like data visualisation or regression models. Sometimes the hidden patterns of the data is harder to uncover, in that case some machine learning model might be useful. Using data efficiently opens new opportunities like better customer service or improved operational efficiency that can come from the analytical findings [2].

Some claim that the changes that happens in the industry are so decisive that it is a fourth step of the industrial revolution, called Industry 4.0. This includes interconnected computers, smart materials and intelligent machines that are communicating with each other to make decisions with minimal human involvement. Implementing manufacturing and business processes with smarter machines and devices may offer advantages in many ways. Advantages such as manufacturing productivity, resource efficiency and waste reduction [3].

At the factory, many steps is automated and managed by robots. The robots are hard coded and performs the task without intelligence. The data that the robots provide is used to some extent, but there is a lot of possibilities to do more analysis. With use of statistical learning methods the systems can be smarter and used for supervision in decision making. Like for example deciding how robots should

be programmed to avoid movements that are unnecessary harmful for the robot.

In the production line, the robots working are industrial robots from [REDACTED]. [REDACTED] is a international technology company [REDACTED]. The company has been around for a long time and creates solutions that connects hands on tasks with software to optimize how things are manufactured, moved or operated [REDACTED].

1.3 Aim

The idea of this thesis is to use data from production robots to create models that estimate the risk of failure and the expected lifetime for production robots. The aim is that the estimates will give an overview for which robots that have a high risk of failure and for how long robots are expected to last without reparation. This information can be used as supervision when planning which robots will need to be changed. The robots with high estimated risk of failure can be kept an eye on. In the best case, problems are be found before the robot sends a warning signal, which might help to avoid a standstill somewhere along the production line. The risk and lifetime estimations can also be used when deciding which robots should be replaced with a bigger/more powerful robot model. The aim is also to gain knowledge about what factors that are effecting the risk of failure and use this information to increase the life expectancy for robots. For example if some specific tasks that the robots perform are found to decrease the life expectancy, the robot programmer could take this into account to avoid specific movements or changing the robot to a more robust or bigger robot type. This work is carried out during the spring semester 2023 and is stationed in [REDACTED] in close relation with the [REDACTED].

The questions that this thesis aim to answer can be summarized as:

1. What is the risk of robot failure during a time period and how can it be estimated?
2. What is the lifetime for new robots and how can it be estimated?
3. What is the remaining lifetime of robots already in use and how can it be estimated?
4. What factors are affecting the robot lifetime?

Some important delimitations are:

- Only robots of a certain model are analyzed, this is due to insufficient number of other robot models in the data.
- Only analyzes the risk until the first time a repair is made on a robot, risks and life expectancy's are not estimated for robots that have been repaired and are in use now.

1.4 Related work

More and more interest is shown towards using data collected from industries enabling extraction of information and insights in the business. This is done so that conclusions can be made and substantiated decisions taken towards increasing industries efficiency [2], [3].

Companies are beginning to use new technologies that arises to increase their understanding of different processes and their efficiencies. The new interest regarding efficiency and effectiveness of predictive maintenance has convinced the robotics company Comau to investigate some techniques. Not only to evaluate different components lifetime but also predicting specific faults that could arise with industrial robots and other machines. While checking related work to this master thesis, the investigated article used four different techniques to tackle this problem: Survival analysis, Convolutional Neural Networks, Random Forest and k-Nearest Neighbours. More specifically, a Cox model approach was chosen carrying out the Survival Analysis. Based on accordance on the featured data set applying PCA analysis, an output list of possible robot candidates to become broken could be generated [5].

Enabling data-driven predictions for robot health indicators, data acquisition and storage have been the key part to pave the way, this is found in another article in the exploration of related work. Robots in the production have played a big part increasing productivity in the last decades and six-axis robots plays an important role of continuous development for automated production systems. To increase the

potential even further, increasing reliability is always of interest. The article shows that predictions of unexpected faults increases the reliability with prediction of such faults based on data-driven models, giving the potential to arrange maintenance actions just before failure called predictive maintenance. To make this possible, suitable data had to be selected as well as preparing the data to be on the right format. Therefore, several supervised machine learning models for the classification and prognosis of faults in a six-axis industrial robot are compared. The comparison includes:

Performance of a:

- Support Vector Machine (SVM)
- Gaussian Process Classifier (GPC)
- Artificial Neural Net (ANN)
- Random Forest (RF)

fault classification of a:

- Support Vector Machine (SVM)
- adaBoost-Regressor (GPC)
- Artificial Neural Net (ANN)

fault based predictions based on:

- Fast Fourier Transform (FFT)
- Continuous Wavelet Analysis (CWA)

[6].

The related work above takes a more strict machine learning approach and non-parametric survival analysis when investigating lifetime estimation for industrial robots. Therefore, more emphasis was put on parametric survival analysis in this thesis.

1.5 Motivation of approach

The first question about the risk of robot failure and how it can be estimated can be seen as a classification problem, either the robot failed during the time period or not. The approach for this question is to implement a classifier using logistic regression. Logistic regression is estimating the probability of an event occurring within a specific time interval given some covariates, this can be implemented on the robot data to estimate the risk of failure for specific robots during a time period.

To answer the second and third question about the expected lifetime for robots, survival analysis is used. To get a better understanding about the risk of failure for robots it is reasonable not only to model if the robot did fail but also the time to event. Survival analysis is a collection of statistical procedures that have the time an event occurs as response variable. Both semi-parametric and parametric models are implemented. To gain information about what covariates affecting the lifetime of robots the semi-parametric models are sufficient, but to be able to predict what is happening after the studied period a parametric model is needed.

To answer the forth and final question, about the factors affecting robot lifetimes, the models made to answer the first three questions are analysed. Which covariates that are significant and the change of the model output per unit change in the covariates are analyzed. This analysis hopefully gives insights in what factors are affecting robot lifetime.

1.6 Outline

Theory regarding used models and approach can be found in section 2. Methods regarding performed analysis can be found in Section 3 with results in Section 4. Following discussions and conclusions regarding the results is to be found in Section 5 and Section 6. Recommendations for further research can be found in Section 7. Attachments can be found in the Appendix **A1**.

2 Theory

2.1 Logistic regression

Logistic regression can be used when the response variable is categorical. That can for example be if an event occurred during a time period or not, the response can in that case be represented as:

$$Y = \begin{cases} 0, & \text{No} \\ 1, & \text{Yes} \end{cases}$$

Logistic regression is used to describe probability of the response variable Y belonging to a category. This is done by taking the covariate vector $\mathbf{X} = (x_1, x_2, \dots, x_n)$ into account while doing the prediction. To the covariates, the bias β_0 and a vector β_1 with specific β -values that corresponds to a covariate is estimated. Meaning the covariates: $\mathbf{X} = (x_1, x_2, \dots, x_n)$ have an associated β -vector: $\beta_1 = (\beta_{11}, \beta_{12}, \dots, \beta_{1n})$. A β_{1i} -value is the expected per unit change in x_i . The probability $p(\mathbf{X})$ that $Y = 1$ is modelled with the logistic function:

$$p(\mathbf{X}) = \frac{e^{\beta_0 + \beta_1 \mathbf{X}}}{1 + e^{\beta_0 + \beta_1 \mathbf{X}}}.$$

The logistic function $p(\mathbf{X})$ is bounded between 0 and 1. Regardless the value of \mathbf{X} , the logistic function will always produce an S-shaped curve which gives a sensible prediction [7].

2.1.1 Classification logistic regression

Logistic regression models the probability that Y belongs to one of the two categories. A specified threshold value can be used to assign an observation to a category. For example, the model predicts that an event occurs if the estimated probability $\hat{p}(\mathbf{X}) > 0.5$, here the threshold value is 0.5.

2.1.2 Overfitting

Overfitting is when a model is made to complex for being generalized to data outside the training data. This occur when the model learns patterns in the training data that are caused by chance rather than true properties of the data. This is not desired since the fit obtained will not reflect accurate estimates of the response on new observations. If a model is not overfitted, it should have the same accuracy on the test data as it have on the training data [7].

2.1.3 Cross-validation

Cross-validation is a resampling method which means that the same statistical method are used multiple times using different subsets of the data. Cross-validation can be used to evaluate a model's performance by estimating the test error associated with a particular statistical learning method [7].

2.1.4 Leave-One-Out Cross-Validation

Leave-one-out cross-validation (LOOCV) involves splitting the set of observations into two parts. Here, the two subsets are not of comparable size. Instead, a single observation, (x_i, y_i) is the validation set and the remaining observations, (x_j, y_j) , are the training set where $j \in 1, \dots, n$ and $j \neq i$. The method is fit on the $n - 1$ training observations and a prediction, \hat{y}_i is made for the excluded observation based on its value x_1 . This is done for every observation in the data set [7].

2.2 Akaike Information Criterion

The Akaike Information Criterion (AIC) is an information criterion which is a function of a regression model's explanatory power and complexity by maximizing the log likelihood of the predictor coefficients. The explanatory power of a model (goodness of fit) increases the criterion in the desirable direction while the complexity of the model works in the reversed direction balancing the criterion in the undesirable direction. For AIC, as the criterion decreases, the model becomes more desirable [8].

2.3 Log-Likelihood

Likelihood is a peculiar concept which isn't a probability but a proportional to a probability. The likelihood of a hypothesis (H) given some data (D) is a probability of obtaining D given that H is true multiplied by a constant K :

$$L(H) = K * P(D|H).$$

For two different hypotheses, H_1 and H_2 , there is evidence for H_1 over H_2 if the probability of the data under H_1 is greater than under H_2 . The maximum of the likelihood function is found by taking the logarithm of the likelihood function and derive it. Since maximizing $\log[f(y)]$ will also maximize $f(y)$, the maximum of the likelihood function will be found. Taking the logarithm of the likelihood function makes it easier, as it changes multiplication to addition and the derivative of $\log[y]$ is $1/y$ [9].

2.4 Pearson correlation coefficient

The Pearson correlation coefficient is a statistical measure that measures how strong the linear relationship between two random variables are. It has been applied to a various different alignments in statistics like data analysis, classification, data analysis and clustering [10].

2.5 Step-wise Feature Selection

Stepwise selection algorithms make decisions in an automated way on whether to keep a variable in the model or add a variable to the model. At every iteration, the model makes the decision based on the best score of the information criterion. If the score of the information criterion was more favourable in the previous step, the algorithm terminates with the model from the previous step [8].

2.5.1 Forward selection

Forward selection typically considers adding one or more covariates to a set and starts with an empty set of covariates [11].

2.5.2 Backward elimination

Backward elimination typically considers removing one or more covariates of a set and starts with the whole set of covariates [11].

2.5.3 Bidirectional elimination

Bidirectional search starts from both sides simultaneously, from an empty set and from the whole set. This enables the algorithm to consider larger and smaller covariate subsets simultaneously [11].

2.6 Confusion Matrix

To evaluate the predictions of the classifier a confusion matrix is a good tool. A confusion matrix which is associated with a classifier of size $n*n$ shows the predicted classification and the actual classification. The number of different classes is n , and a confusion matrix for $n = 2$ can be seen in Table 1. The predicted positive is the observations that the model predicts that an event occurred for and the predicted negative is the observations that the model predicted that the event did not occur for. The actual values show if an event actually occurred in the data.

Table 1: Confusion matrix for a two class classification.

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

The matrix can be interpreted as follows:

- *TN* (True Negative): Predicted value $Y = 0$, true value $Y = 0$.
- *FP* (False Positive): Predicted value $Y = 1$, true value $Y = 0$.

- *FN* (False Negative) Predicted value $Y = 0$, true value $Y = 1$.
- *TP* (True Positive): Predicted value $Y = 1$, true value $Y = 1$.

Where the prediction accuracy and classification error from the matrix can be obtained in the following way [12]:

$$\text{Accuracy} = \frac{TN + TP}{TN + FP + FN + TP}$$
$$\text{Error} = 1 - \text{Accuracy}.$$

2.6.1 Precision, Sensitivity and specificity measure

The precision and sensitivity measures are easy to interpret and together summarizes classification performance on each class. The sensitivity measure indicates the probability for false negative errors:

$$\text{Sensitivity} = \frac{TP}{TP + FN}.$$

The precision measure indicates the probability for false positive errors:

$$\text{Precision} = \frac{TP}{TP + FP}.$$

The specificity measure indicates the probability for true negative errors:

$$\text{Specificity} = \frac{TN}{TN + FP}.$$

Furthermore, in the context of error magnitude, $\leq 10\%$ is considered an excellent result, 10-20% good, 30-40% moderate and $> 40\%$ poor [13].

2.7 Outlier

An outlier is a point for which an observation is far from the predicted value made by the model. Outliers can arise due to a number of reasons, like incorrect recording of an observation during the data collection process or just by chance or bad luck [7].

2.8 Moment of inertia

In this work, the moment of inertia is applied when a robot is lifting an object, since the robot both lifts it and rotates it at the same time. For an object with mass in a linear motion, inertia can be described as the resistance on the objects acceleration. The body's linear movement is opposed by the inertial force which is equal to the negative of the product of the objects mass and its linear acceleration. The same applies for an object in a rotational motion. The angular acceleration of the object will be exposed by a resisting inertial moment which is equal to the negative of the object's moment of inertia multiplied by the angular acceleration with SI unit $kg * m^2$ [14].

2.9 Normalization

To be able to interpret the effect of different covariates from the β -values in the models, the data is normalized. This is done by removing the mean of all covariates and divide by the standard deviation of the covariate. When this is done all covariates have mean 0 and standard deviation 1 [15].

2.10 Survival analysis

Survival analysis is a topic where the time until an event is analyzed. It is often applied in medical studies, where survival times after different treatments are analyzed, but the same applications can be used far beyond medicine. Survival analysis is useful in all settings when it is not only of interest if an event happened, but also after how long time. In a time period, where different objects are studied, hopefully some objects survive until the end of the study. Such an objects survival time is said to be right censored. It is known that the object has survived during the study but it is unknown what the true survival time is since it extends beyond the time of the study [7].

2.10.1 Survivor function

The survivor function is denoted $S(t)$ and gives the probability that an object survives longer than some specified time t . Meaning that $S(t) = P(T > t)$ gives the probability that the survival time T exceeds a specified time t . This is fundamental for survivor analysis, since obtaining survival probabilities for different values of t provides vital encapsulated information from survival data. Theoretically, t ranges from 0 to infinity. Though, the survivor function is nonincreasing:

- At time $t = 0$, $S(t) = S(0) = 1$; since at the start of the study the object hasn't been exposed yet, the probability of surviving past time 0 is 1.
- At time $t = \infty$, $S(t) = S(\infty) = 0$; since theoretically, if the study would go on forever, eventually no object would survive and hence the survivor function must eventually move to zero [16].

2.10.2 Hazard function

The hazard function, denoted by $h(t)$, gives the instantaneous possibility per unit in time for an event to occur, given that an object has survived until time t . $h(t)$ equals the limit as Δt approaches zero of a probability statement about survival. The probability statement is divided by Δt where Δt stands for a small interval in time [16],

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

2.10.3 Kaplan-Meier curve

The general formula for a Kaplan-Meier (KM) curve gives the survival probability for an object: $j = 1, 2, \dots, n$, at failure time t_j and looks like the following:

$$S(t_j) = S(t_{j-1}) * Pr(T > t_j | T \geq t_j).$$

The probability of surviving to time t_{j-1} is given by:

$$S(t_{j-1}) = \prod_{i=1}^{j-1} Pr(T > t_i | T \geq t_i),$$

where T is a random variable representing the survival time and t is a specific value for T . The estimated probability of surviving a time step is:

$$\hat{Pr}(T > t_i | T \geq t_i) = 1 - \frac{d_i}{n_i}$$

where d_i is the number of deaths and n_i is the number at risk in time period i [16].

2.10.4 Cox Proportional Hazard model

For modelling the survival data, a popular model in the form of the Cox proportional hazard (PH) model is used. The Cox PH model is a semi parametrical model, meaning that the distribution of the outcome is unknown even though the regression parameters are known. The Cox PH model gives an expression for the hazard function at time t where the hazard function contains two parts. The first part, $h_0(t)$ is called the baseline hazard and is only dependent on t . The other part, $exp(\beta\mathbf{X})$, is scaling the baseline hazard and is only dependent on the covariate vector \mathbf{X} , but not the time. The Cox model hazard function is described as:

$$h(t, \mathbf{X}) = h_0(t)exp(\beta\mathbf{X}).$$

Since the hazard is describing the probability of an event occurring "right now" it can be hard to interpret. To get something that's easier to interpret a corresponding Cox model survival function can be converted from the hazard function. This survival function formula is what determines the adjusted survival curves given each objects covariates \mathbf{X} . The formula for the survival function at time t is described as:

$$S(t, \mathbf{X}) = [S_0(t)]^{exp(\beta\mathbf{X})},$$

where $S_0(t)$ is the Kaplan-Meier estimation at time t [16].

2.10.5 Weibull distribution

For analyzing reliability and lifetime data, the Weibull distributions is one of the most widely used distributions. It is used in many different fields like engineering, material science and economics to mention a few. The survival function for the two-parameter Weibull distribution looks like the following:

$$S(t) = \exp(-\lambda t^p), \quad t \geq 0,$$

where $\lambda > 0$ and $p > 0$ are the scale and shape parameters. Furthermore, the hazard function for the two-parameter Weibull distribution looks like the following:

$$h(t) = \lambda p t^{p-1},$$

which is constant if $p = 1$, increasing if $p > 1$ and decreasing if $p < 1$. [17].

2.10.6 Fitting a Weibull survival model with covariates

When a Weibull model is fitted a survival curve is estimated for every observation given the covariates of the observation. The shape parameter p is kept constant for all values of the covariates, but the scale parameter λ is reparameterized given the covariates. The scale parameter λ can be expressed as:

$$\lambda = \exp(\beta_0 + \beta_1 X).$$

Where X is the vector of covariates. Knowing the survival function $S(t)$ and λ given X , the survival function with Weibull distribution given the time t and the covariates X can be expressed as [16]:

$$S(t, X) = \exp(-\exp(\beta_0 + \beta_1 X)t^p), \quad t \geq 0.$$

2.10.7 Expected lifetime and median lifetime

To find the expected lifetime the probability density function must be found. Knowing the hazard and survival function the probability density function can be found. The probability density function $f(t)$ is defined as:

$$f(t) = S(t)h(t).$$

When the probability density function is found the expected time can be calculated. The expected time $E[t]$ is given by:

$$E[t] = \int_{-\infty}^{\infty} t f(t) dt.$$

The median lifetime can easily be found directly from the survival curve, the time point where $S(t, X) = 0.5$ is the estimated median survival time. This makes the estimated median lifetime easy to find and interpret [16].

2.10.8 Conditional survival analysis

Typically, survival analysis is a study where objects survival time, or time to an event is investigated. This is often carried out from the start of the objects lifetime. However, the survival probability evolves over time and most often decreases with increased survivorship. This makes it interesting investigating the change in survival probability over the objects lifetime, given that the object haven't had an event up to a certain point. To analyze the change in survival distribution as a progress under an objects lifetime, the method goes under the name conditional survival analysis. Suppose that the lifetime from an objects T with survivor function $S(t) = P(T \geq t)$. Then, the t -year conditional survivor distribution for the object who have survived for $t_0 \geq 0$ years, the probability that it lives additional t years, $S(t|t_0) = P(T \geq t + t_0 | T \geq t_0)$ is given as:

$$S(t|t_0) = \frac{S(t + t_0)}{S(t_0)}, \quad t \geq 0.$$

$S(t|t_0)$ is called the conditional survivor function for an object without an event occurring for t_0 years [18].

2.10.9 Schoenfeld's global test

To test the proportional hazards assumption of Cox's regression model, Schoenfeld's global test is useful. All other methods up to Schoenfeld's global test are based on graphical methods and are therefore subjective. Hence, for cases where the violation of the proportional hazard assumptions is marginal, the graphical methods might be inadequate to detect any violation and the Schoenfeld's global test can be a good complement. More theory regarding the Schoenfeld's global test can for the interested one can be found following the citation [19].

3 Method

3.1 Description of the production line

This work was carried out at [REDACTED], which produces [REDACTED]. More specifically the thesis was made at a section in the factory called [REDACTED]. This section assembles the products before sending them to the next section of the factory. The production line consists of many different assembly lines, involving smaller assembly lines which assists the main line with specific assembly tasks.

In the production line, many industrial robots are working together. The robots are performing a various set of tasks. This includes processes that involves gluing, welding or handling the products in different ways. Each robot are individually programmed and each robots movements are individual. In this thesis, one robot type is evaluated which reduces the number of robots in the analysis to one hundred and eleven.

3.1.1 Robots in the production

The robots that are analyzed are industrial robots from [REDACTED]. There are more than [REDACTED] robots that work in the production line. A typical industrial robot working in the production has six axis with one engine on each axis. In Figure 1, an example of a six axis industrial robot can be seen [20]. This is just an example of how a six axis industrial robot could look. It was chosen as the robot to display since it resembles the robots in the production line and therefore gives a good sensation of what the industrial robots in the factory looks like.

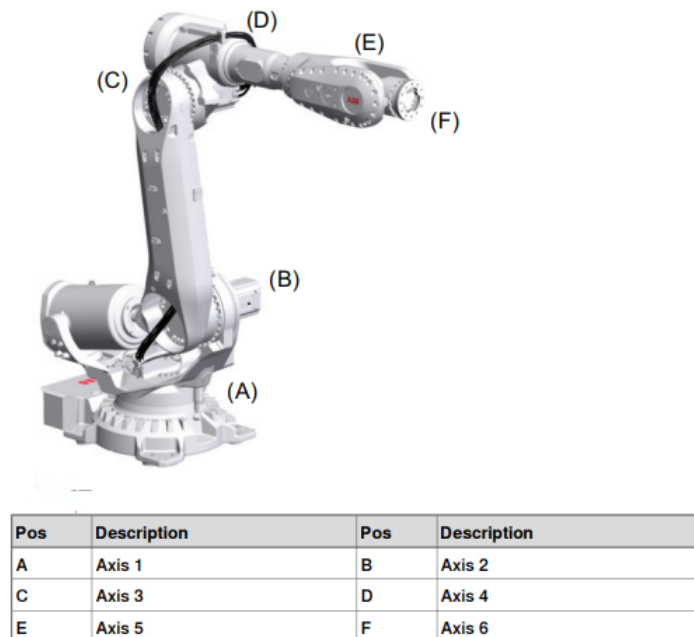


Figure 1: An example of a six axis industrial robot.

In the production line, maintenance or services regarding industrial robots are planned when a robot sends a warning signal and there are one mandatory service planned each year. One hour of production is important for the production rate. Since maintenance of the robots can make certain parts of the production line stand still, the production can gain from maintenance work with longer foresight. It is small margins in the production, if it is a stop in the production line due to something else than a robot malfunctioning, the errant part of the production line might need resources. In this case, valuable prevention work can't be done and the same applies for maintenance or services for robots. It becomes a negative loop, where more things in the production have to be fixed without really have the time to

plan thoroughly.

Depending on which axis the gearbox failure occur, different amount of time is needed to change the gearbox.

- **Axis 6:** [REDACTED]
- **Axis 3:** [REDACTED]
- **Axis 1 and 2:** [REDACTED]

The same applies for engines:

- **Axis 6 and 4:** [REDACTED]
- **Axis 5:** [REDACTED]
- **Axis 2 and 3:** [REDACTED]
- **Axis 1:** [REDACTED]

It exists a stock of spare parts on the factory for urgent breakdowns for both engine and gearbox changes. If a robot sends a warning signal, ordering of spare parts are done. Changing a gearbox costs about [REDACTED] changing an engine. Buying a new robot costs [REDACTED]. From November 2015 until this analysis was carried out, a total of approximately [REDACTED] has been spent on engine changes, with a total of [REDACTED] being changed and [REDACTED].

The gearbox is checked on the robot service once each year. If the gearbox oil is glittery or dark the gearbox needs to be changed. If a gearbox is changed because of this, it is categorized as a failure in this work. Except this check, the only way to know that a gearbox on a robot needs to be changed is through a warning signal sent by the robot or an acute robot breakdown. The same applies for engines, though in this case, oil leakage is checked for during the service. If oil leakage can be seen, the engine is replaced. This as well is labelled as a robot failure in this work. Meaning that if a carried out service detects one of these indications the robot is considered to have failed. However, this is just one approach. Glittery oil in the gearbox or oil leakage can also be seen as a censored variable since it doesn't have to mean that a robot with these indications of failure detected will lead to an actual failure.

3.2 Description of the data

The final data set is a combined data set which originates from different data sources. Below, the data sources are explained.

- **CBM:** Data set from [REDACTED] containing a calculated value for the strain on the different axis, the time the robot has been turned on and the procent of the turned on time it has been moving. The variables that measures the stress of the different axis of a robot gives a measurement of how heavily burdened an axis have been. The variables are summarized values of twelve measurement parameters where each parameter have a score. These parameters have been calculated as a ratio compared to corresponding data base median. Corresponding data base median means [REDACTED] data base of how robots of the same type is runned, this is not only robots in the factory but robots in different factories of the same type. Unfortunately, the only available variable for the stress on the different axis for this study was the summarized values. This means that the covariate for stress on different axis in the data set could have contained considerably more information if this study would have accessed all twelve variables.
- **Robot loads:** Every robot has a corresponding file from [REDACTED] containing information about robot loads. For example the force needed to move in a certain direction and the position the robot proceeds from. The information for each robot are summarized into one data set containing information for each robots "robot load". Each row in the data set corresponds to one robot where the robot ID-number also can be seen.
- **Engine changes:** A list with information regarding engine changes with date and time for the changes.

- **Gearbox changes:** A list with information regarding gearbox changes with date and time for the changes.
- **Robots:** A list of all robots installation date.

3.3 Data handling

Following actions is carried out handling the data:

1. Manually typing in values from "RobotLoad" files into Excel-file "RobotData".
2. Merging data sets taking "RobotID" as the linking key: Merging "RobotData" with "CBM", "RobotLoads", "EngineChanges", "GearboxChanges" and "Robots".

Operations made on the merged data set "RobotData" containing all information needed:

1. Removing engine and gearbox failures which were not the first failure in the studied time period for one robot.
 - The time will only be modelled until the first failure. Risk and life expectancy for robots that have been repaired will not be analyzed since they are considered "dead" after their first failure. Since the number of repaired robots are too few and occur at different time points, meaning that the time is too short to analyze for many of the repaired robots.
2. Mark the robots that had an engine or gearbox failure and add a column containing each robot's response "Y", i.e. "0" or "1".
 - "0": No event, no gearbox or engine failure.
 - "1": Engine or gearbox failure.
3. Calculate the time in number of days from installation to failure.
 - Or installation date until data extraction date for survived robots.
4. Calculate the time a robot has been active/moving during the studied period by multiplying "DutyFactor" with "DutyCounter" to a new column called "ActiveTime".
 - "DutyFactor": Procent of the time the robot has been moving since installation.
 - "DutyCounter": Time measure which counts the time for which the robot were switched on since installation.
5. Calculate the usage time until failure by multiplying "DutyCounter" with "DutyFactor" and the percentage of the studied time period until the first failure occurred for the robot and add the values to a new column called "ActiveTimeUntilFailure".
6. Remove uninformative columns and rows containing NA-values.
7. Remove robots of other type than the one analyzed.
8. Normalize the covariates in the data set using R-function "scale".

One more procedure remains, that is creating one final data set for the logistic classification and one data set for the survival analysis. The data sets are identical with two differences:

1. The column "ActiveTime" is an explanatory variable in the logistic data set. This column does not exist in the survival data set since the time is to be modelled here.
2. The survival data set have "ActiveTimeUntilFailure" as the response together with the response "Y" which will be treated as the censoring variable.
 - The "ActiveTimeUntilFailure" column does not exist in the logistic data set.

3.4 Description of final data

In Table 2, it can be seen what variables the worked data set contains.

Table 2: Variables in the worked data set.

*Explanatory variable only in the logistic data set.

**Response variable together with Y_{event} only in the survival data set.

Variable name	Explanation
Robot Number	Serial number of the robot
Mass	Robots total handling weight (kg)
CoGz	Deviation from robots CoG in z direction (mm)
CoGL	Deviation from robots CoG in L direction (mm)
Jx0	Force needed moving in x direction ($kg * m^2$)
Jy0	Force needed moving in y direction ($kg * m^2$)
Jz0	Force needed moving in z direction ($kg * m^2$)
J5	Force needed moving axis 5 ($kg * m^2$)
J6	Force needed moving axis 6 ($kg * m^2$)
Frame Mass	Load on frame (kg)
Frame CoGx	Frame deviation from robots CoG in x direction (mm)
Frame CoGy	Frame deviation from robots CoG in y direction (mm)
Frame CoGz	Frame deviation from robots CoG in z direction (mm)
Lower Arm Mass	Load on lower arm (kg)
Lower Arm CoGx	Lower arm deviation from robots CoG in x direction (mm)
Lower Arm CoGy	Lower arm deviation from robots CoG in y direction (mm)
Lower Arm CoGz	Lower arm deviation from robots CoG in z direction (mm)
Upper Arm Mass	Load on upper arm (kg)
Upper Arm CoGx	Upper arm deviation from robots CoG in x direction (mm)
Upper Arm CoGy	Upper arm deviation from robots CoG in y direction (mm)
Upper Arm CoGz	Upper arm deviation from robots CoG in z direction (mm)
Strain Axis 1	Stress on Axis 1
Strain Axis 2	Stress on Axis 2
Strain Axis 3	Stress on Axis 3
Strain Axis 4	Stress on Axis 4
Strain Axis 5	Stress on Axis 5
Strain Axis 6	Stress on Axis 6
Y_{event}	Response regarding robot incident (1 if incident, otherwise 0)
Active Time*	Active usage time
Active Time Until Failure**	Active usage time until failure

3.5 Logistic Regression

After structuring and choosing the covariates that are possible to make a model with, the modelling can start. In Table 3 the structure of the final data for logistic regression can be seen, where X_1, X_2, \dots, X_n is the covariates and Y_{Event} is the response variable. Y_{Event} is 1 if the robot has failed and 0 if it has not failed during the studied time period.

Table 3: Structure of the final data set for logistic regression.

X_1	X_2	...	X_n	Y_{Event}
-1.35	0.76	...	1.58	1
1.42	-0.53	...	-1.68	0
0.24	0.04	...	1.51	1
⋮	⋮	⋮	⋮	⋮

The first step is to make a model with all covariates and some chosen interaction effects, this model is called the full model. It is hard to know which interactions to bring into the full model and it is to many variables to check every interaction. The method used to choose interactions is to reason from knowledge

about the different variables, use domain knowledge and ask robot experts about their opinion. Hopefully this strategy is good enough to bring the most important interactions into the full model.

The following interactions are checked in the full model for both logistic regression and survival analysis. The interactions can be seen in Table 4.

Table 4: Chosen interactions.

Interactions
Mass*J5
Mass*Jx0
Jz0*J5
Jx0*J5

When the full model is made, bidirectional step-wise selection is used to find which subset of variables and interactions that give the best model. AIC-score is used as the evaluation metric in the step-wise selection and the model with the best AIC-score is considered the best model. A criteria for the model selection is that the model should be hierarchical. This means that if a covariate is part of an interaction, the covariate should also be included by itself in the model.

When the covariates and interactions for the best model is found, LOOCV is performed. The model is trained using all observations except one in the data, the model then predicts the probability of an event occurring given the covariates X for that observation. If the estimated probability $p(X) > 0.5$, the robot is assumed to have failed during the studied period. This is performed for all observations. The accuracy of the model can be calculated by comparing the predicted outcome to the actual outcome. This can be done without the risk of the model being over-fitted, since it has not been trained on the observation it is predicting. The accuracy is giving a good measurement about how well the model can be generalized to new data. If the model can predict the outcome well for unseen data, it indicates that the coefficients in the logistic model are weighting the importance of the covariates in a good way.

To get a more precise picture of how well the model is predicting, a confusion matrix is made. This makes it possible to evaluate how good the specificity, precision and sensitivity is for the model which gives a more comprehensive evaluation of the model than just the accuracy.

When the model is evaluated with LOOCV, a final model is made including all observations. Using this model, the probabilities for failure during the studied time period is estimated. The estimated probabilities are put in a list with the corresponding robot id, this list is ordered by the estimated probability. This gives an overview of the risk for different robots. The result can then be analysed and used for supervision. For example, robots with high risk might be a good idea to exchange with a more powerful robot type the next time the robot is changed.

3.6 Survival analysis

To model the time to event, a survival analysis is performed. The data is on a bit different format in survival analysis than it was in the logistic regression. In Table 5, the structure of the data for survival analysis is presented, X_1, X_2, \dots, X_n is the covariates and Y_{Time} and Y_{Event} is the response variables. Y_{Event} is 1 if the robot failed and 0 if it did not fail during the studied time period. Y_{Time} is the number of active hours the robot moved until it failed, if the the robot did not fail during the studied time Y_{Time} is the number of hours the robot worked until the data was extracted.

Table 5: Structure of the final data set for survival analysis.

X_1	X_2	...	X_n	Y_{Event}	Y_{Time}
1.90	-0.64	...	-1.47	1	6000
-1.48	0.92	...	-1.02	0	5000
1.71	0.11	...	1.60	1	6500
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots

When the data is on the desired format, the modelling can start. The first model made is a Cox PH model with all possible covariates and some interactions. Just like when making the logistic regression model, choosing what interactions to include is not easy and all possible interactions can't be tested since the data consists of too many covariates. The interactions tested is the same as for logistic regression, and are chosen using domain knowledge and what experts think is reasonable.

From the full model, bidirectional stepwise feature selection is performed and the evaluation used is AIC-score. As for logistic regression a criteria for the model is that the model should be hierarchical, keeping a covariate by itself if it's part of an interaction in the model. The best hierarchical model found regarding AIC-score is kept for further investigation.

The model assumption of proportional hazard is tested using Schoenfeld's global test and looking at scatter plots of the Schoenfeld's residuals. If the Schoenfeld's global test is significant and some pattern or bias can be seen in the scatterplot of the Schoenfeld's residuals the β -values are probably time dependent. This means that the hazard rate between two observations differ for different points in time. If this is the case other models need to be considered.

To be able to make predictions about survival times longer than the studied time in the data, a parametric model is made and a distribution that is similar to the Kaplan-Meier curve for the data is searched. Different distributions are tested and evaluated using the AIC-score along with plots of the estimated distributions compared to the Kaplan-Meier curve. The distributions tested are:

- Exponential distribution
- Gamma distribution
- Log-Normal distribution
- Log-Logistic distribution
- Weibull distribution

Parametric models with the same covariates found for step wise selection in the Cox PH model and the different distributions are tested and evaluated with AIC-score. The best parametric model found is then used to calculate the expected median survival time from the time point of the installation. The estimated median life time is chosen instead of the expected lifetime since it is easy to interpret and calculate directly from the survival curve without calculating the probability density function.

The estimated median lifetime of the robots from installation date is of interest for new robots. For robots already in use, the estimated remaining median lifetime is of interest. The remaining lifetime is calculated using conditional survival analysis, taking into consideration the information that the robot has already survived for a certain time.

4 Results

The results presented below are all obtained with the statistical programming language R, version 4.2.3. Results regarding the logistic regression is presented under section 4.1 and results regarding the survival analysis is presented under section 4.2.

4.1 Logistic regression classification

The first model made is the full model containing all covariates described in Table 2 and some chosen interactions described in Table 4. In the full model using 0.05 as confidence level two variables, StrainAxis5 and UsageTime, show a significant impact on the survival probability. The estimated β -values and p-values for the significant covariates in the full model can be seen in Table 6.

Table 6: Significant covariates in full logistic regression model.

Covariate	Estimated β -value	P-value
StrainAxis5	0.969	0.022
UsageTime	0.908	0.031

A correlation plot of the covariates in the full model can be seen in Figure 2.

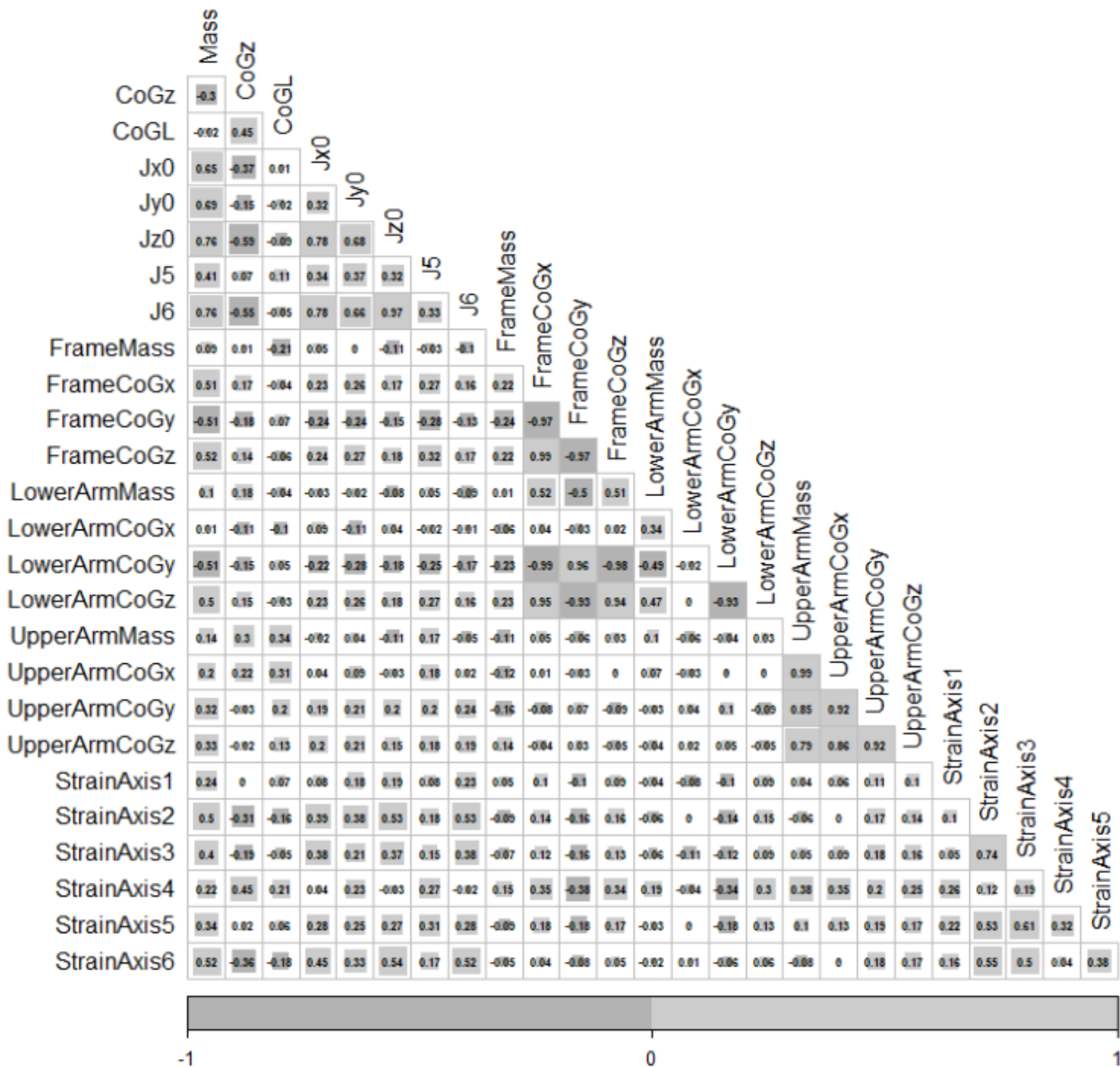


Figure 2: Correlation between covariates in the full logistic model.

After bidirectional elimination a better model than the full model regarding AIC score is found. The AIC-score of the full and final model can be seen in Table 7.

Table 7: AIC scores logistic regression models.

Model	AIC score
Full model	163.11
Final model	130.79

A correlation plot of the covariates in the final model can be seen in Figure 3.

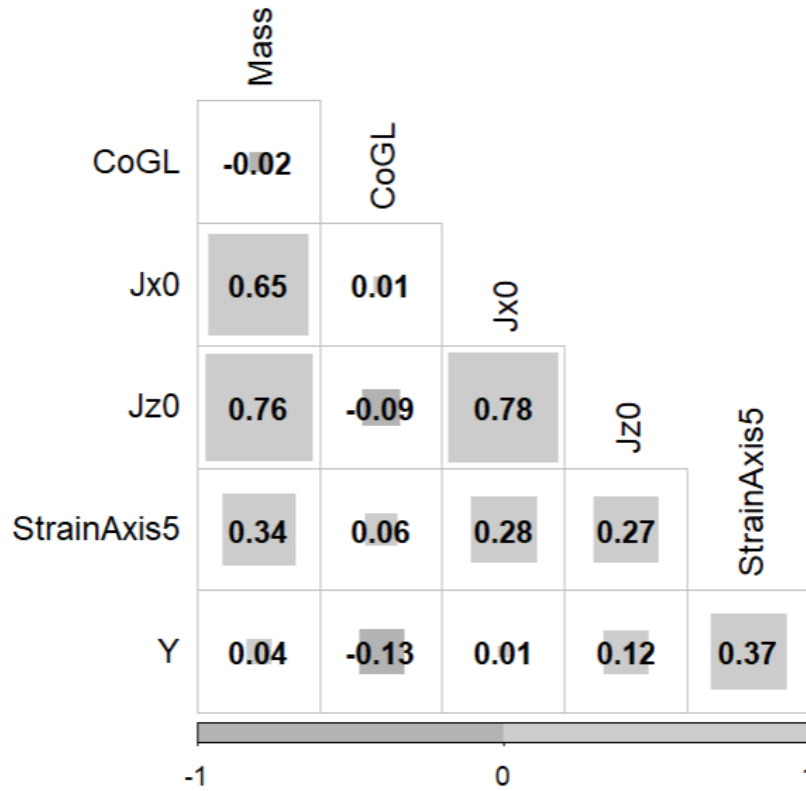


Figure 3: Correlation between covariates in the final logistic model.

When the covariates of the best model is found, the logistic model is retrained one time for each observation in the data and LOOCV is performed. A confusion matrix of the result can be seen in Table 8. The accuracy, sensitivity, precision and specificity can be seen in Table 9.

Table 8: Confusion matrix logistic regression from LOOCV.

	Predicted Negative	Predicted Positive
Actual Negative	51	14
Actual Positive	20	26

Table 9: Performance of logistic regression model, evaluated with LOOCV.

Measure	Performance
Accuracy	0.694
Sensitivity	0.785
Precision	0.588
Specificity	0.565

For the final model a summary showing the covariates, corresponding estimated β -values and p-values in the final model can be seen in Table 10.

Table 10: Covariates of final logistic regression model.

Covariate	Estimated β -value	P-value
Mass	-0.899	0.034
CoGL	-0.395	0.122
Jx0	-0.749	0.181
Jz0	1.825	0.008
StrainAxis5	1.069	$5 \cdot 10^{-4}$
Active Time	0.813	0.005
Jx0*Jz0	-0.382	0.242

When the final model is trained, the model estimations of the probability of failure are calculated for all robots. These probabilities indicate the risk for specific robots, an example with ten robots and their estimated probability of failure is presented in Table 11.

Table 11: Estimated probability of failure with the final logistic model during the studied time period. 10 out of 111 randomly selected robots displayed.

Robot-ID	Estimated probability of failure
██████████	0.536
██████████	0.880
██████████	0.137
██████████	0.180
██████████	0.301
██████████	0.369
██████████	0.214
██████████	0.929
██████████	0.997
██████████	0.953

4.2 Survival analysis

4.2.1 Cox PH model

The first model when performing the survival analysis is a full Cox-PH model, containing all covariates is described in Table 2. The same interactions as in logistic regression are chosen and can be seen in Table 4. At a chosen significance level of 0.05, eight variables show a significant impact on the survival time and can be seen in Table 12. The AIC score for the full model is 358.0151 and for the final model 337.7969 This can be seen in Table 13.

Table 12: Significant covariates in full survival analysis model.

Covariate	P-value
J5	0.043
FrameCoGy	0.024
FrameCoGz	0.017
UpperArmMass	0.031
UpperArmCoGx	0.029
UpperArmCoGy	0.032
StrainAxis2	0.006
StrainAxis6	0.003

Table 13: AIC sores survival analysis model.

Model	AIC score
Full model	358.015
Final model	337.797

The final model, after performing step-AIC contain fourteen covariates and one interaction. These covariates and the interaction effect can be seen in Table 14.

Table 14: Covaraites of final survival analysis model.

Covariate	Estimated β -value	P-value
Mass	0.012	0.229
CoGz	-0.006	0.022
CoGL	-0.011	0.014
Jx0	-0.031	$4*10^{-5}$
J5	0.014	0.005
J6	0.077	0.023
FrameCoGy	-0.100	0.005
FrameCoGz	-0.123	0.003
UpperArmMass	3.769	0.034
UpperArmCoGx	-3.403	0.034
UpperArmCoGy	0.732	0.038
StrainAxis1	-0.809	0.038
StrainAxis2	-1.062	0.004
StrainAxis5	1.152	0.001
Mass*J6	$-2*10^{-4}$	0.042

To check the assumption regarding proportional hazards, the Schoenfeld's global test is carried out together with visual inspection of the scatter plots of the Schoenfeld's residuals. The plotted Schoenfeld's residuals against time for the covariate *Mass* can be seen in Figure 4. The resulting plots for the remaining covariates in the final model can be found in Appendix A.1. From the scatter plots it's hard to conclude that there are any bias or trend for the Schoenfeld's residuals over time.

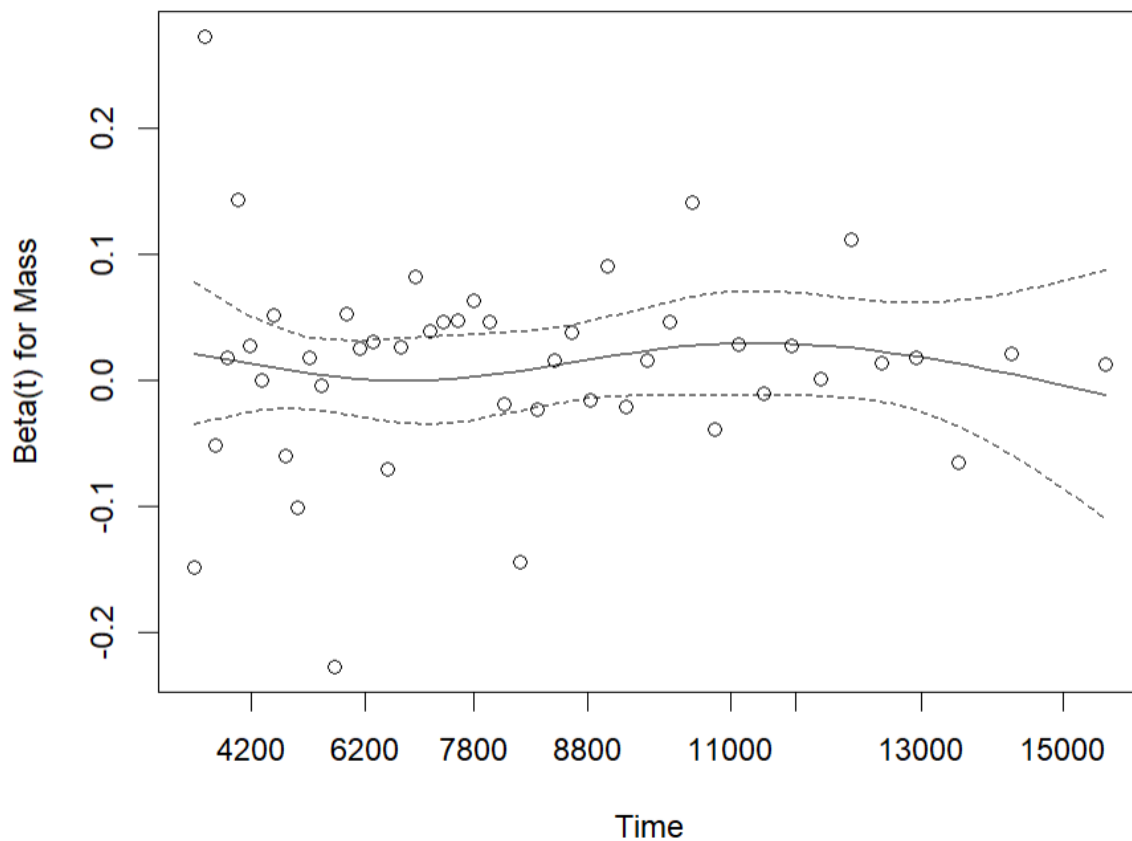


Figure 4: The Schoenfeld residual plot for the covariate Mass. The x-axis show the active usage time for the robots. The y-axis show the value of the Schonefeld residuals, the axis is denoted Beta(t) for Mass since it show if the Beta-value needs to change for different times. The line show the time component of Beta and the dotted lines show a 95% confidence interval for the estimation. If zero is within the confidence interval for all time points, this indicate that the Beta-value does not change with time.

To complement the information from the scatter plots, the Schoenfeld’s Global test is used. The test returned a p-value of 0.06892 indicating that the null-hypothesis can’t be rejected at a significance level of 0.05 and hence, the assumption of proportional hazards will be assumed. Meaning that the β -values in the final model can be assumed to be constant over time. The output of the Schoenfeld’s Global test of the final model can be seen in Table 15.

Table 15: Output of the performed Schoenfeld's global test.

Covariate	P-value
Mass	0.189
CoGz	0.151
CoGL	0.475
Jx0	0.450
J5	0.225
J6	0.210
FrameCoGy	0.712
FrameCoGz	0.969
UpperArmMass	0.642
UpperArmCoGx	0.451
UpperArmCoGy	0.071
StrainAxis1	0.208
StrainAxis2	$8 \cdot 10^{-4}$
StrainAxis5	0.024
Mass*J6	0.232
GLOBAL	0.069

4.2.2 Parametric model

To check which parametric distribution that fitted the data best, the AIC-scores is examined as well as looking at the estimated survival curves for each chosen distribution. The following distributions are tested:

- Weibull distribution
- Exponential distribution
- Log-logistic distribution
- Gamma distribution
- Log-normal distribution

The following AIC-scores for each distribution are obtained and can be seen in Table 16. The Weibull distribution got the lowest AIC-score which suggests that the Weibull distribution fits the data best.

Table 16: AIC-scores for each distribution.

Distribution	AIC-score
Weibull distribution	966.054
Exponential distribution	1013.189
Log-logistic distribution	971.340
Gamma distribution	972.852
Log-normal distribution	981.040

In Figure 5, the Kaplan-Meier curve together with the estimated Weibull survival curve where the covariates equal zero can be seen. The Kaplan-Meier curve plotted together with the other distributions can be found in Appendix A.1.

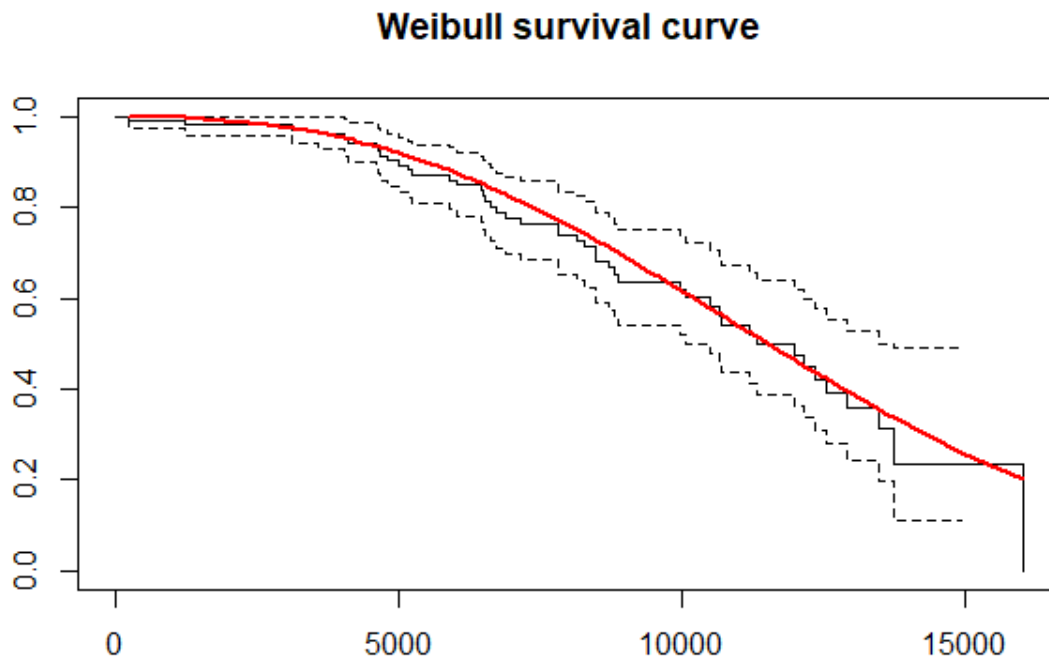


Figure 5: The Kaplan-Meier survival curve (black step function) and estimated Weibull survival curve (red smooth function) when all covariates are assumed to be zero. The dashed lines are the marked 95-percent confidence interval of the Kaplan-Meier curve.

In addition, the corresponding hazard function to the Weibull survival curve with the covariates equal zero can be seen in Figure 6. The corresponding hazard functions for the other distributions can be found in Appendix A.1.

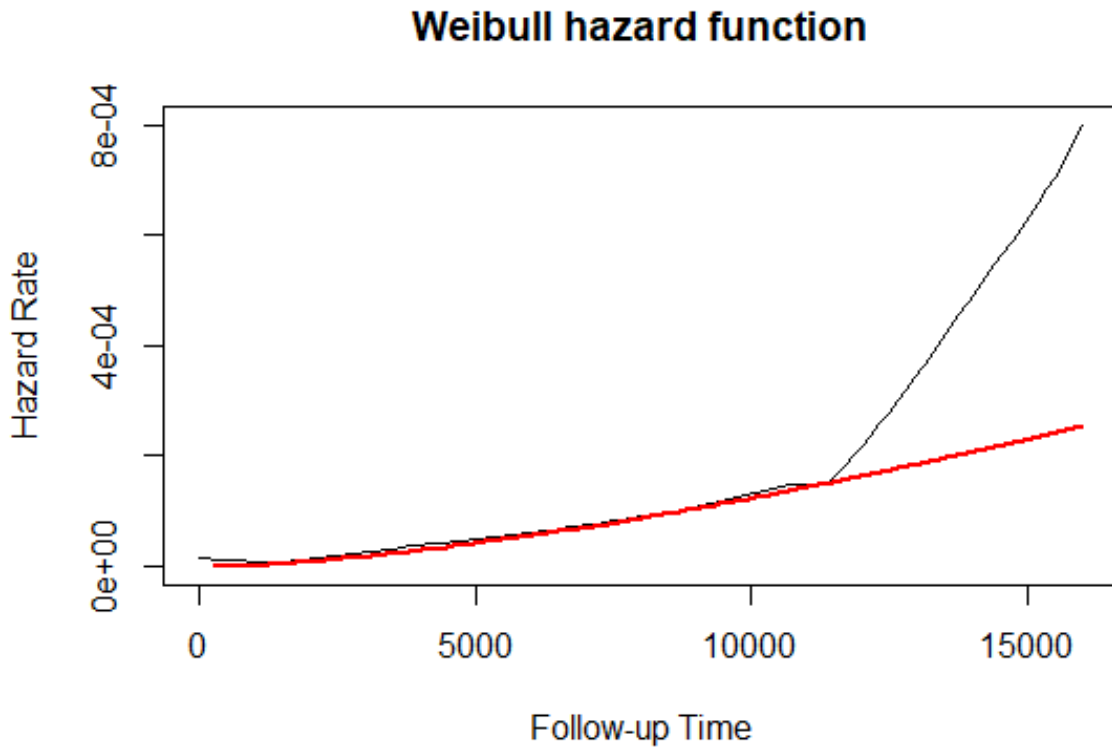


Figure 6: The hazard function from the data (black and not smooth function) to the estimated Weibull hazard function (red smooth function) when all covariates are assumed to be zero.

The final model with Weibull distribution is fitted. The obtained shape parameter is 1.080. The estimated β -values from the final model can be seen in Table 17.

Table 17: Covariates of final Weibull distribution model.

Covariate	β -value
Mass	-0.002
CoGz	0.002
CoGL	0.003
Jx0	0.009
J5	-0.004
J6	-0.021
FrameCoGy	0.028
FrameCoGz	0.034
UpperArmMass	-1.151
UpperArmCoGx	1.044
UpperArmCoGy	-0.228
StrainAxis1	0.193
StrainAxis2	0.332
StrainAxis5	-0.354
Mass*J6	$7 \cdot 10^{-5}$

4.2.3 Expected robot lifetimes

The estimated median lifetimes for the robots from installation date given the covariate values for the robots are calculated. The same applies for the estimated median conditional lifetime. This mean that the robots that have failed during the studied period will have zero in expected conditional life time. Example of the output for ten robots can be seen in Table 18.

Table 18: Estimated number of active hours from installation date for different robots and estimated number of remaining active hours from the end of the studied period. Example with 10 of 111 randomly selected robots displayed.

Robot-Id	Estimated median life time	Conditional estimated median life time
██████████	7650	0
██████████	9491	0
██████████	12570	16070
██████████	17496	0
██████████	9850	0
██████████	12288	0
██████████	13715	17240
██████████	12804	0
██████████	14245	16230
██████████	15252	16390

The final predicted median lifetime is compared to the actual lifetime of the robots that have failed during the studied period. The Pearson correlation between the predicted median and the actual lifetime is 0.49. In Figure 7 the actual survival time is plotted against the median value of the predicted survival curve. The data for the plot has been censored so only the robots that have failed during the studied period are represented in the plot.

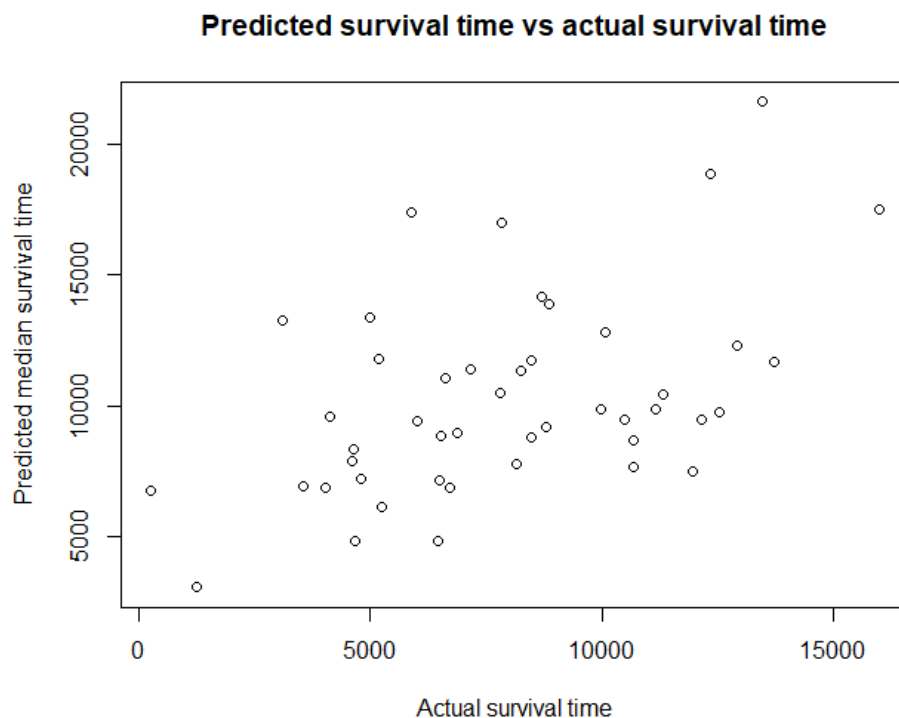


Figure 7: Predicted median lifetime against actual lifetime, both in active working hours, for censored data and the final survival analysis model.

A comparison between the logistic regression model and the final parametric survival model is also made, the Pearson correlation between these are -0.35. The comparison is plotted in Figure 8. The "Predicted risk" axis show the estimated risk of failure from the logistic model and the other axis show the median lifetime from the survival curve made with the final survival model.

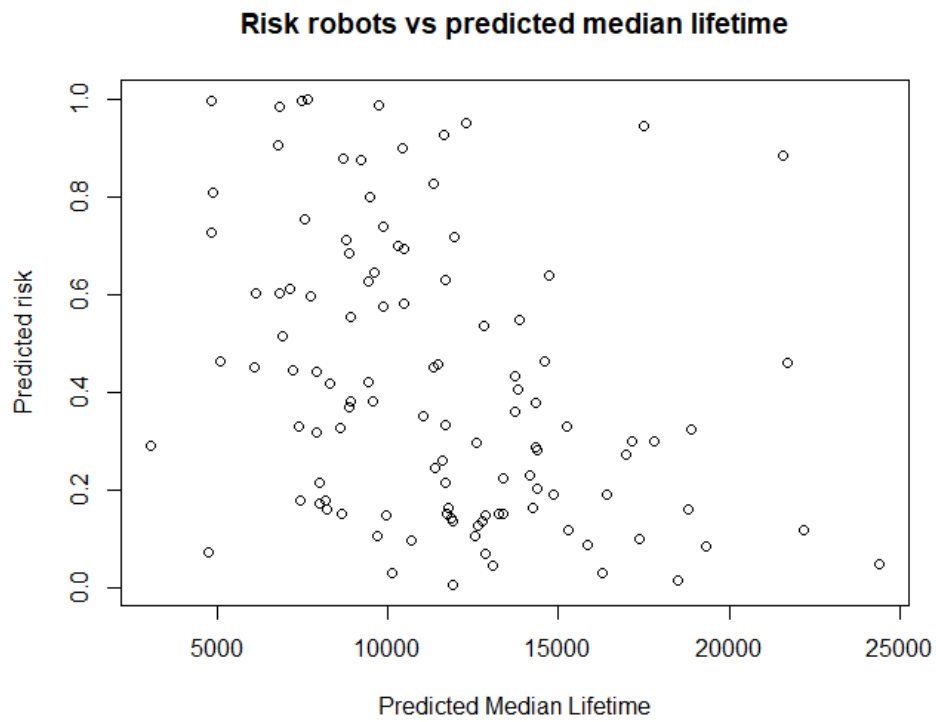


Figure 8: Comparison between the estimated risk for a nine year period the logistic model and the predicted median lifetime of active working hours from the final survival model.

5 Discussion

5.1 Data limitations

The ability to extract information from a data set is limited by the data itself. If the data is bad it is only possible to extract limited amount of information. This being said, the expectations on the data regarding the industrial robots was not the highest. Ideally, a data set with all robots in the production would have been used. This however was not possible and hence, the 111 robots at hand were the ones to be focused on. The data did not contain so much information about the movements of the robot as planned for in the initial state of the thesis work. The covariates in the data mostly describe the forces needed to move in different directions from the robots initial state. The data itself makes it hard to answer some questions. One initial question of interest was how robot movements effect the lifetime and how robot movements could be changed to increase the lifetime. The covariates in the used data set are to the most part hard to change, for example the mass of the object the robot operates can not be changed even if mass effects the robot lifetime. Therefore, data limitations made the thesis focus more on predictions rather than interpretation of important factors of robot lifetime, than what would have been the case if the data was as first thought.

The final data set contains data from different sources, merged together into one set with as much explanatory power as possible. It was not an easy task choosing what to keep in the final data set and it was more than one time the final data set changed. The work was extensive putting together a sufficient data set and it would not have been possible without help from employees having access to various data sources.

Again, the expectations was not the highest and it was hard knowing what useful information that could be extracted. Nevertheless, the results obtained exceeded the initial expectations and important information could be extracted which made conclusions possible. This itself shows great potential for further investigations. However, the limitations where always present and the analysis carried out in this thesis have greater potential with more and better data.

Data that contains movement patterns from the robots, information regarding specific differences in stress between the robots axis or a sufficient observation number for different robot types would open up a whole new world of possibilities. More precise conclusions as well as better predictions would be possible if more time and effort would have been put on extracting relevant data from the robots. This is something to keep in mind and something that always was present in the thought during the thesis work. The limitation of the rather small data set at hand is important to take into account and this could affect the results obtained.

5.2 Logistic regression

The first model trained is the logistic regression model. The model estimates the probability that a robot has failed during a studied time period. This model answer research question one:

1. *What is the risk of robot failure during a time period and how can it be estimated?*

The data that the logistic model is trained on has usage time as a covariate, the usage time can be seen as a ratio that show how much of the studied time the robot has been working. This is an important difference from the data used for the survival analysis and makes the models answer different questions.

The answer of the first question could be of interest when new robot investments are made, ██████ might want to know that a new investment will work over a defined time period. The logistic model gives good information about the probability that a robot will work over the period. If the robot will survive, the defined time period might be of bigger interest than how many active working hours the robot is expected to have. Because after a sufficiently long time period, the production is expected to change so much that the robots will have to be exchanged anyways. If the risk of failure is high for a robot performing a specific task in the production line, the robot can be exchanged with a more powerful robot type.

In the full logistic model, two covariates showed significant effect on the probability of failure. The significant covariates are usage time and the strain in axis 5. Since the data is normalized before modelling, the β -coefficients can be interpreted as the effect the covariates have in the model. From this, it can be seen that usage time and strain in axis 5 have a similar effect on the probability of failure in the model. Still, the effects are hard to interpret since some covariates in the model are correlated. The effect of a covariate might seem small when looking at the p - and β -value, since the effect of that covariate is already explained through other covariates.

The final model is the best model found regarding AIC score. Still, some of the covariates in the final model is correlated which makes the effect of different covariates hard to interpret. Especially the variables Jx0, Jz0, J6 and Mass are all correlated with each other. The variables Jx0, Jz0 and J6 all indicate the moment of inertia needed to move in different directions, it is logical that these variables are correlated with each other and the mass. To get a model that is easy to interpret and make conclusions about what is effecting the probability of failure, it should have been better to avoid inclusion of correlated covariates in the models. Instead new covariates could have been calculated, combining information from correlated covariates. This was not done and it is therefore hard to draw conclusions from the final model. The reason why this is not done is that most of the covariates are hard to change even if their effect is known. It is for example hard to change the mass of the object that is lifted even if it's known that mass is effecting the probability of failure. Instead, more focus is put on trying to predict the probability of failure and less on trying to find the cause of the failures.

The accuracy of the model is evaluated with LOOCV, the score is assumed to give a fair evaluation of how well the model perform without over-fitting. The accuracy of 0.6937 together with the sensitivity, specificity and precision, indicate that the model contains some information, even if it is not great. Higher accuracy was obtained regarding high risk robots, showing that the model predicts the risk of failure for stressed robots in a efficient way. For example, robots with an estimated risk of less than 0.1, fail only 20% of the time and robots with an estimated risk over 0.9 fail 83% of the time. A result that is satisfying since the high- and low risk robots arguable are the most important robots to classify correctly, especially the high risk robots. This shows that the model is a good classifier and gives good valuable information regarding important high risk observations.

5.3 Survival analysis

The second model is the survival analysis model. The model estimates survival functions for the robots from which the expected and median lifetime can be obtained. These models can answer research question two and three:

2. *What is the lifetime for new robots and how can it be estimated?*
3. *What is the remaining lifetime of robots already in use and how can it be estimated?*

In the data used for survival analysis, both the categorical value indicating if the robot has failed during the studied time period and the active time the robot has been moving during the time period are response variables. The active time is of interest since it is what can be assumed to make the robots fail rather than the time they have been standing in the factory. To know for how long the robots will work is of interest both when new robot investments are made and to get an overview of the robots in the factory and for how much more time ██████████ can count on their service.

The first model fitted is a full Cox-PH model. In the model, eight covariates show a significant impact on the robot lifetime. As in the logistic regression case, it is hard to interpret the model since some covariates are highly correlated. The best Cox-PH model regarding AIC score found is considered the final Cox-PH model. As in the logistic model, some covariates are correlated but no action is taken since the model aims to predict the robot survival times rather than being used to interpret the importance of different covariates.

The Cox-PH model assume proportional hazard, this assumption needs to be tested. The scatter plots of residuals against time can be difficult to interpret since it's hard to say if there are any pattern or if the residuals are random. Because of the uncertainty regarding the scatter plots, the Schoenfeld's global residual test is used as a complement to test the proportional hazard assumption. It's not unexpected

that individual covariates have a p -value smaller than 0.05 since many covariates are tested. Hence, it's better to check the Global test to check the proportional hazard assumption for the model. The Global test could not reject the null hypothesis of proportional hazard and hence the assumption is made.

The best parametric model found uses Weibull distribution, this model also assumes proportional hazard. The shape parameter that is used to fit the survival curve for every robot is 1.08. This means that the hazard is increasing over time, meaning that is is more likely for a robot to fail after more time. In Figure 6 it can be seen that the Weibull hazard function seems to fit the non-parametric hazard function well for quite a long time. However, after some time point it doesn't seem to fit the data so well anymore. The worse fit in the end of the studied time period isn't a big concern since few events can change the non-parametric hazard very much in this region.

From the Weibull model, the median estimated lifetime is calculated. This shows which robots are estimated to fail after longer respectively shorter time. The estimated median survival time is plotted against the actual survival time in Figure 7 and the correlation is calculated. This plot and the correlation of 0.49 both indicate that the predictions contain some information.. Even if some pattern can be seen, the estimated median does not perfectly predict the actual survival time. This is not surprising since all robots are not expected to fail at the median point of their survival curve. How well the model actually performs is hard to say. The evaluation techniques for these kinds of models are good for comparisons between models but it is hard to evaluate the model performance as well as for example a classification model.

Also, the remaining estimated median lifetime is calculated for the robots that have not failed yet. This calculation is taking into account the conditional probability that the robots have survived until the time point of the end of the studied time period. The estimated median lifetime is chosen instead of expected lifetime since it is easier to calculate directly from the survival function and since it easy to interpret. The remaining lifetime of the robots that have already failed is zero. In reality some robots have been repaired and it would be interesting to get an estimate of their remaining lifetime as well. This was one of the delimitations of this thesis, but could probably be estimated using some other approach.

The survival model and the logistic model are compared in Figure 8. The models are as described in the report fitted on a bit different data sets so the predictions are not expected to be very similar. Some pattern is expected between the predicted probability of survival from the logistic model and the estimated median survival time from the survival model. It can be seen in the plot that the robots with the highest predicted risk is the ones with the shortest predicted lifetime and that the correlation between them is -0.35. This is expected and shows that the two models have found some similar patterns regarding how the covariates are effecting risk of failure.

6 Conclusions

The aim of this thesis was to estimate the risk of failure for robots in [REDACTED] factory and estimate the lifetime of the robots. The aim was that the estimation should be used as help in decision making. This could for example help to plan when robots should be exchanged and to know which robots that have a high risk of failure and therefore should be exchanged with a more powerful robot type. The aim was also to find which factors that cause the failures and with this knowledge provide guidelines for how to for example change the movements of robots to increase their lifetime.

The first research question was:

1. *What is the risk of robot failure during a time period and how can it be estimated?*

The risk of robot failure during a time period is estimated with logistic regression. The output of the logistic regression model is the estimated probability that a robot have failed during the studied time period. This is an advantage for logistic regression compared to other classification models, since the probability is of interest rather than the classification. The model seems to give useful information about the risk for the robots.

The second question research question was:

2. *What is the lifetime for new robots and how can it be estimated?*

This is estimated using a parametric survival model with Weibull distribution. Survival analysis is often used in medical research but also works well for robots lifetimes. The advantage of parametric survival models is that survival curves for the robots can be estimated. From the estimated survival curves the expected survival time and the estimated median survival time can be calculated. In this thesis, the estimated median survival time has been used rather than expected survival time since the median is easier to get directly from the survival curve without having to calculate the probability density function of survival. This gives concrete information about for how long time the services of different robots can be expected to work from their installation date.

The third research question was:

3. *What is the remaining lifetime of robots already in use and how can it be estimated?*

To estimate this, the same parametric survival model with Weibull distribution is used as when to answer the second question. The difference is that the conditional probability of surviving until the end of the studied time period needs to be taken into account. The estimated median remaining lifetime can easily be calculated from the estimated survival function. This can be used to see when the robots in use now are expected to fail and be needed to be exchanged or repaired.

The fourth and final question was:

4. *What factors are affecting the robot lifetime?*

Due to data limitations, the aim to explain what factors are making the robots increase the risk of failure was dropped. This since what is explained in the data is very hard to change in practice. If the data included covariates that could be changed in practice, for example the speed the robot is moving, it would be interesting to know how the covariates effected the risk. If a covariate was increasing the risk of failure it could be changed by the robot programmer to decrease the risk. This was the initial plan, however the data planned to use was not accessible in the end. The data that was accessible mostly contained covariates that can not be changed in practice. For example the mass of the lifted object, making the understanding of what factors are effecting the risk less interesting.

Despite the flaws in the data, valuable information was extracted which indicates potential regarding used methods. It also signals that further investigations can be a source of value addition for the

operation. Better methods for collecting and storing data would pave the way for deeper and more precise evaluations which for sure would develop the organization.

7 Recommendations

This thesis aim to show the possibilities that lies in analyzing data and extract information from the business to improve and optimize activities. Improvement areas is found as well as areas which are worth digging deeper into and investigate further. The core message will be that data management is the fundamental part which enables future data analysis. A standardized procedure to collect and store data, making it easy to access data from the whole factory will be important for future analysis to be carried out.

Investigating how movement patterns for different robots affect the wear would be interesting to analyze. Trying to find the optimal movements for industrial robots to minimize the stress and maximize life time. This also opens up the possibility for evaluating how different movements affect the different axis on a robot and if a certain axis shows significant stress over others due to specific movements.

Another recommendation for further research would be evaluating distinctions between different robot types. Examine if diverse robot types function better for a specific task than others or if a specific robot type breaks down considerably quicker than others performing a task. This would open the possibility for optimizing the factory with the right robot type at the right place performing the right task.

Another question that would be interesting to research is how to estimate the remaining lifetime of robots that has been repaired and for example had a motor change. This was not applicable in this study due to lack of data and the method chosen. Since the company has many repaired robots in the factory it would be interesting to investigate their risk of failure.

If the probability of failure for repaired robots was known, it would be interesting to use this information to compare the cost of repairing or to buy a new robot. To know the estimated cost of the different options would help the organization in the decision making and since the cost of robots are high, better decisions regarding this could potentially save much money.

References

- [1] [REDACTED]. [REDACTED]. Accessed: 2023-02-14.
- [2] Sravanthi Kuchipudi and Subba Reddy Tatireddy. Applications of big data in various fields. *International Journal of Computer Science and Information Technologies*, 6(5), 2015.
- [3] Morteza Ghobakhloo. Industry 4.0, digitalization, and opportunities for sustainability. *Journal of Cleaner Production*, 252(2), 2020.
- [4] [REDACTED]. [REDACTED]. Accessed: 2023-03-27.
- [5] Riccardo Pinto and Tania Cerquitelli. Robot fault detection and remaining life estimation for predictive maintenance. *Procedia Computer Science*, 151:709–716, 2019.
- [6] Corbinian Nentwich, Sebastian Junker, and Gunther Reinhart. Data-driven models for fault classification and prediction of industrial robots. *Procedia CIRP*, 93:1055–1060, 2020.
- [7] Gareth James, Daniella Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer Nature, New York, 2 edition, 2021.
- [8] Charles Lindsey and Simon Sheather. Variable selection in linear regression. *The Stata Journal*, 10(4):650–669, 2010.
- [9] Alexander Etz. Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, 1(1):60–69, 2018.
- [10] Yuanqing Xia Haomiao Zhou, Zhihong Deng and Mengyin Fu. A new sampling method in particle filter based on pearson correlation coefficient. *Neurocomputing*, 216:208–215, 2016.
- [11] Karla Brkić Alan Jović and Nikola Bogunović. *A review of feature selection methods with applications*. 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015.
- [12] Brian Ramsay Anca Ralescu, Sofia Visa and Esther Van Der Knaap. Confusion matrix-based feature selection. *CEUR Workshop Proceedings*, 710:120–127, 2011.
- [13] Sari Kajava Mikaela Mughal Pekka Matilainen Salla Ruuska, Wilhelmiina Hämäläinen and Jaakko Mononen. Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. *Behavioural Processes*, 148:56–62, 2018.
- [14] Ron P. Podhorodeski and Paul Sobejko. A project in the determination of the moment of inertia. *International Journal of Mechanical Engineering Education*, 33(4):319–338, 2005.
- [15] Roman Tkachenko Michal Gregušand Natalya Shakhovska Ivan Izonin, Bohdan Ilchyshyn and ChristineStrauss. Towards data normalization task for the efficient mining of medical data. *2022 12th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 480–484, 2022.
- [16] David G. Kleinbaum and Mitchel Klein. *Survival Analysis, A Self-Learning Text*. 1996 Springer Science+Business Media, Inc, 233 Spring Street, New York, NY 10013, USA, 2 edition, 2005.
- [17] Saad J. Almalki and Saralees Nadarajah. Modifications of the weibull distribution: A review. *Reliability engineering and System Safety*, 124:32–54, 2014.
- [18] Ho Yun Lee Sin-Ho Jung and Shein-Chung Chow. Statistical methods for conditional survival analysis. *Journal of biopharmaceutical statistics*, 28(5):927–938, 2018.
- [19] Roshini Sooriyarachchi and WWM Abeysekera. Use of schoenfeld’s global test to test the proportional hazards assumptions in the cox proportional hazards model: An application to a clinical study. *Journal of The National Science Foundation of Sri Lanka*, 37(1):41–51, 2009.
- [20] Production specification irb 6700. <https://search.abb.com/library/Download.aspx?DocumentID=3HAC044265-001&LanguageCode=en&DocumentPartId=&Action=Launch>. Accessed: 2023-04-26.

Appendix

A.1 Additional Figures

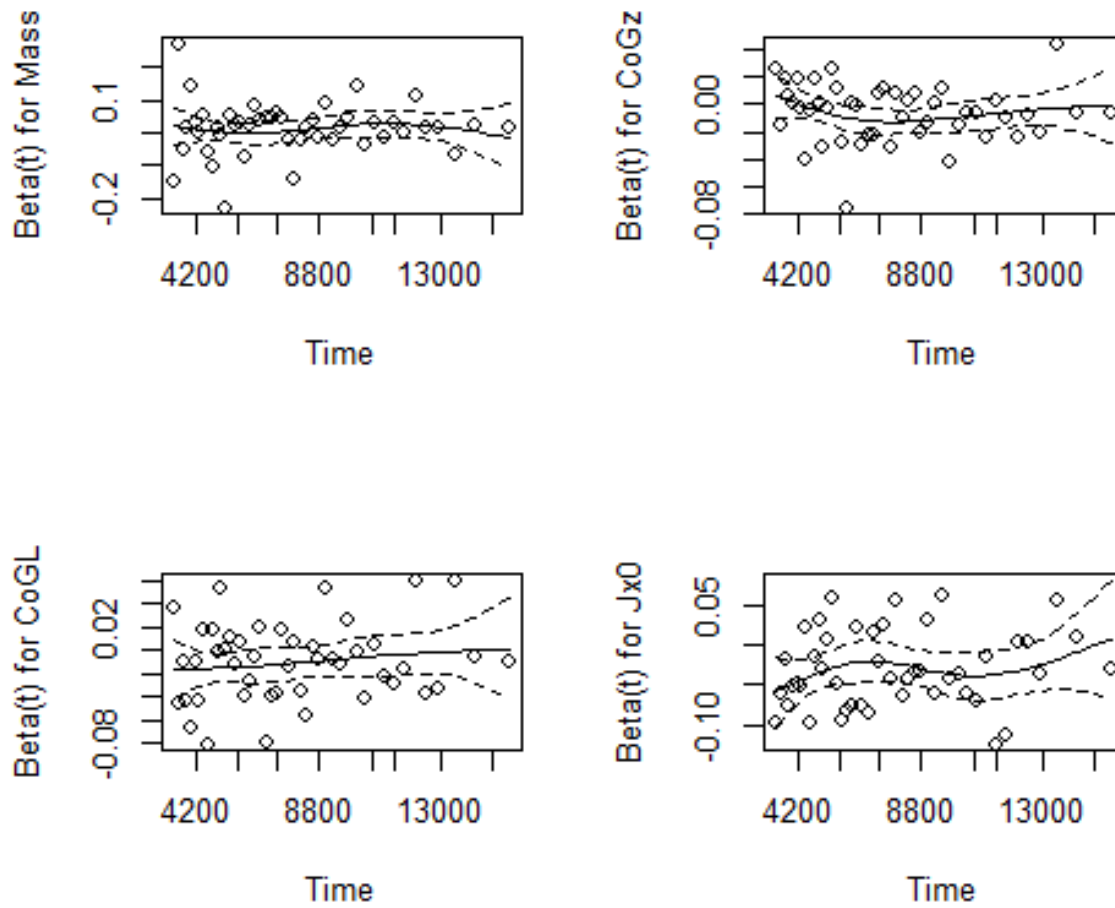


Figure A.1: The Schoenfeld residual plot for covariates Mass, CoGz, CoGL and Jx0.

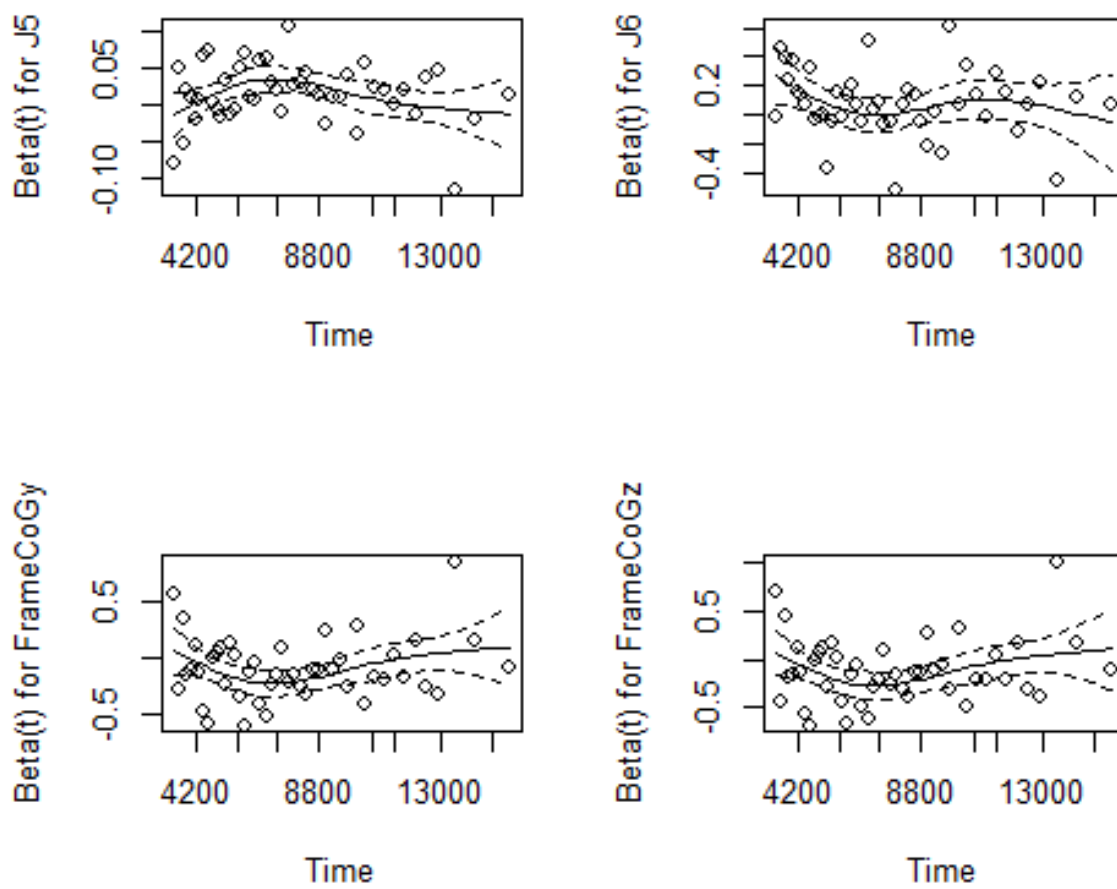


Figure A.2: The Schoenfeld residual plot for covariates J5, J6, FrameCoGy and FrameCoGz.

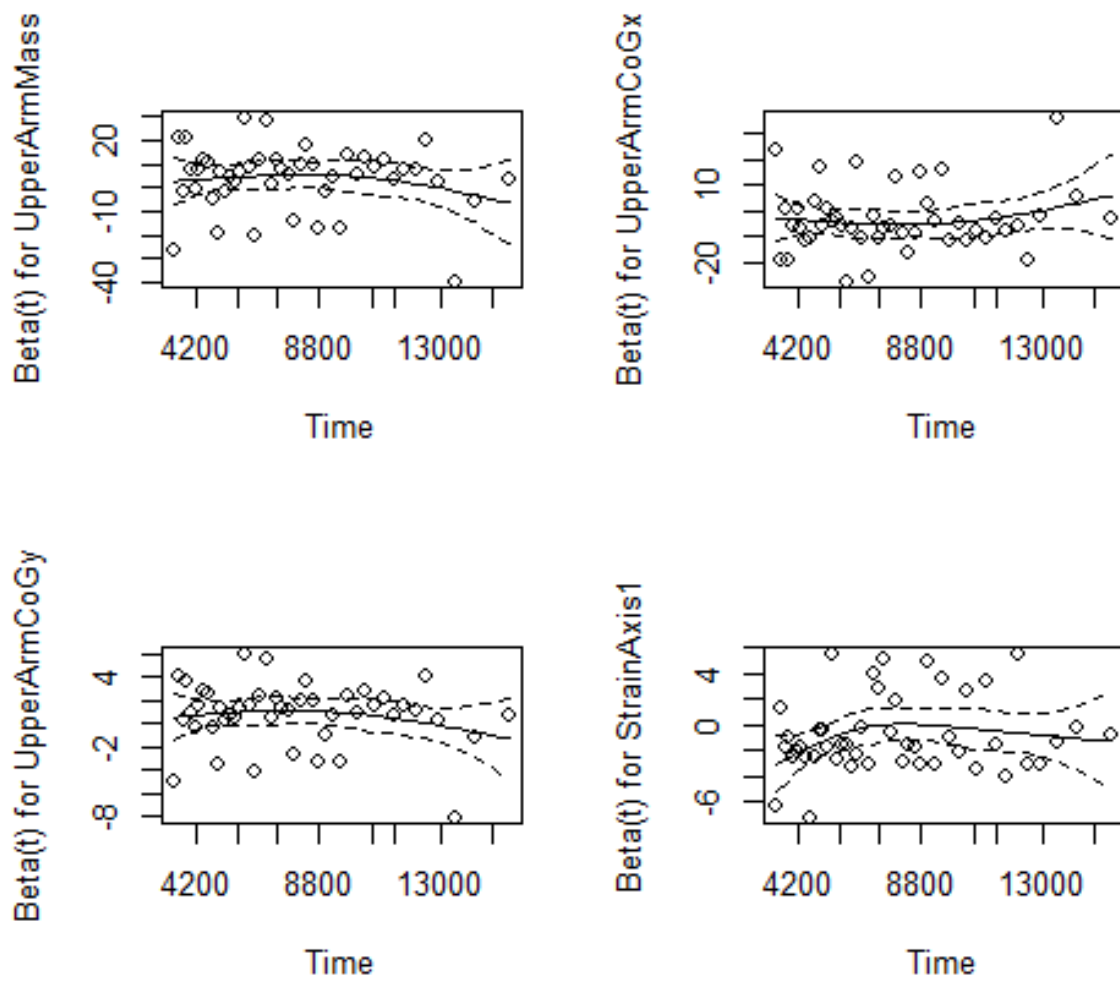


Figure A.3: The Schoenfeld residual plot for covariates UpperArmMass, UpperArmCoGx, UpperArmCoGy and StrainAxis1.

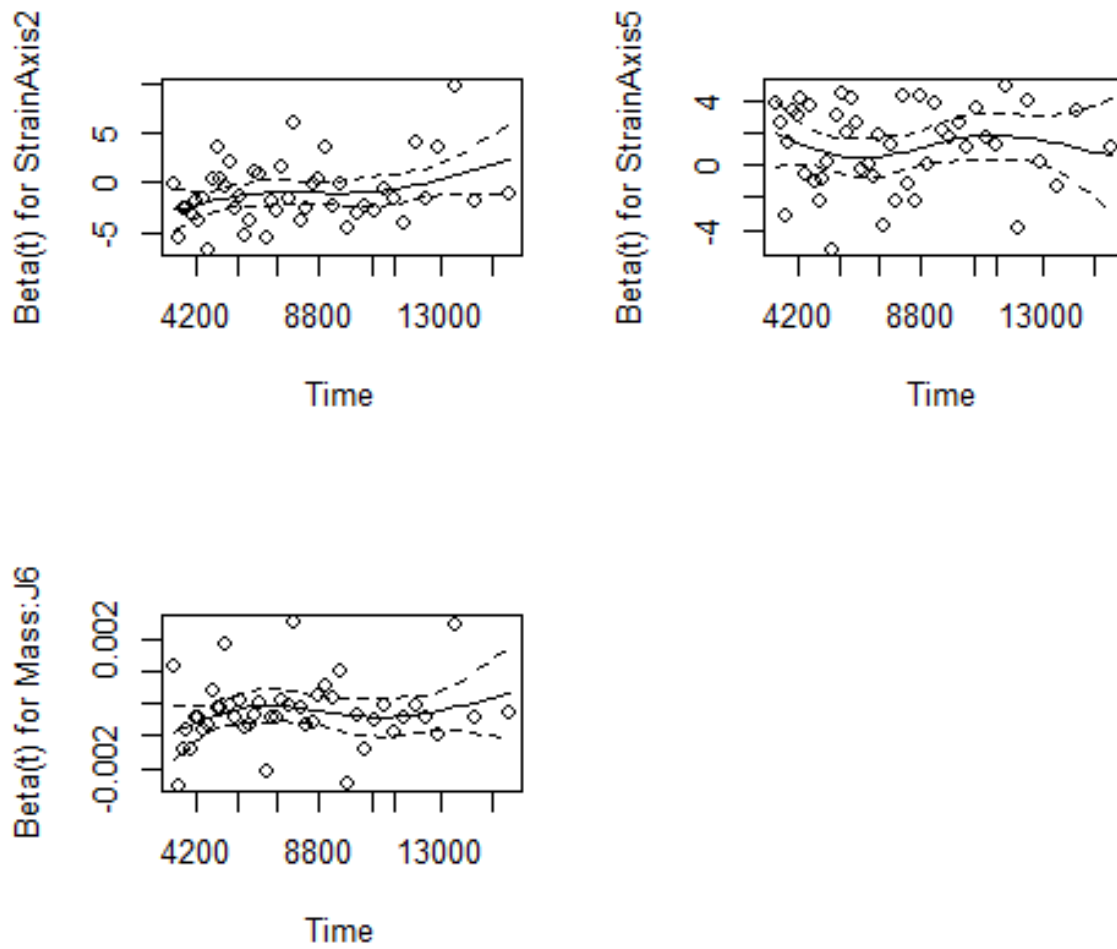


Figure A.4: The Schoenfeld residual plot for covariates StrainAxis2, StrainAxis5 and Mass*J6.

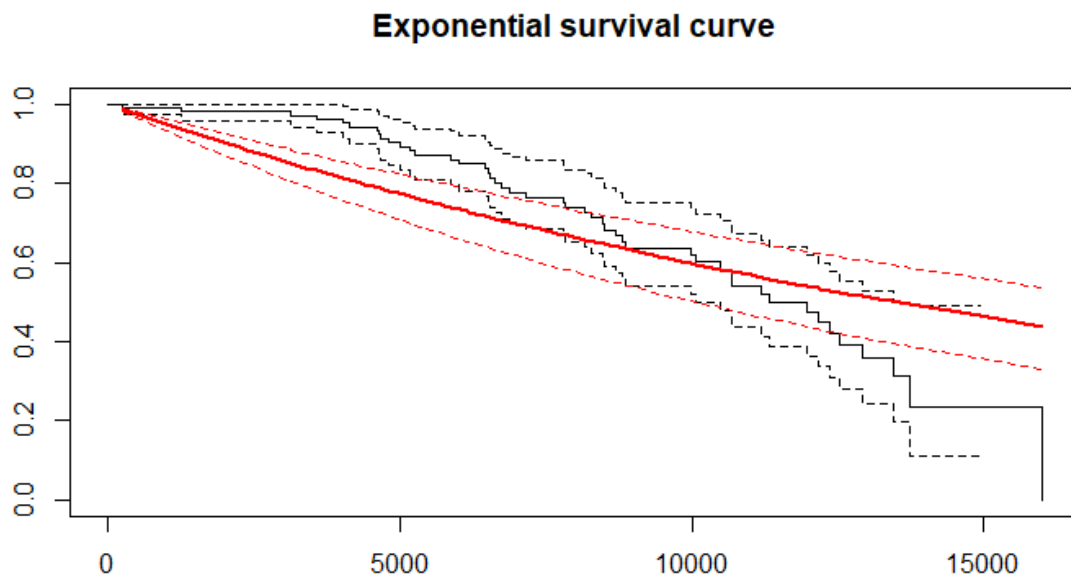


Figure A.5: The Kaplan-Meier curve and estimated Exponential survival curve.

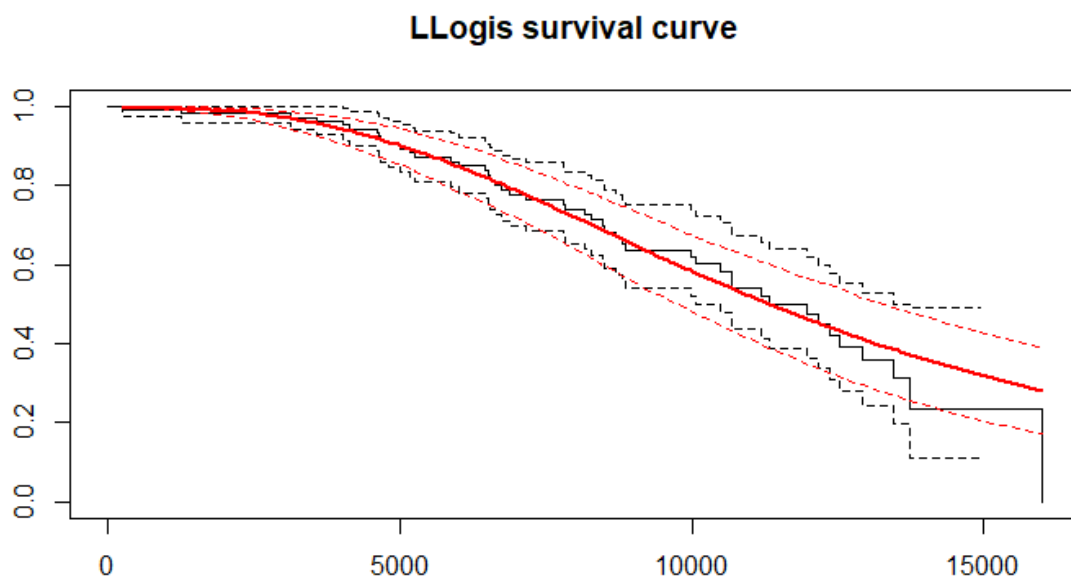


Figure A.6: The Kaplan-Meier curve and estimated Log-logistic survival curve.

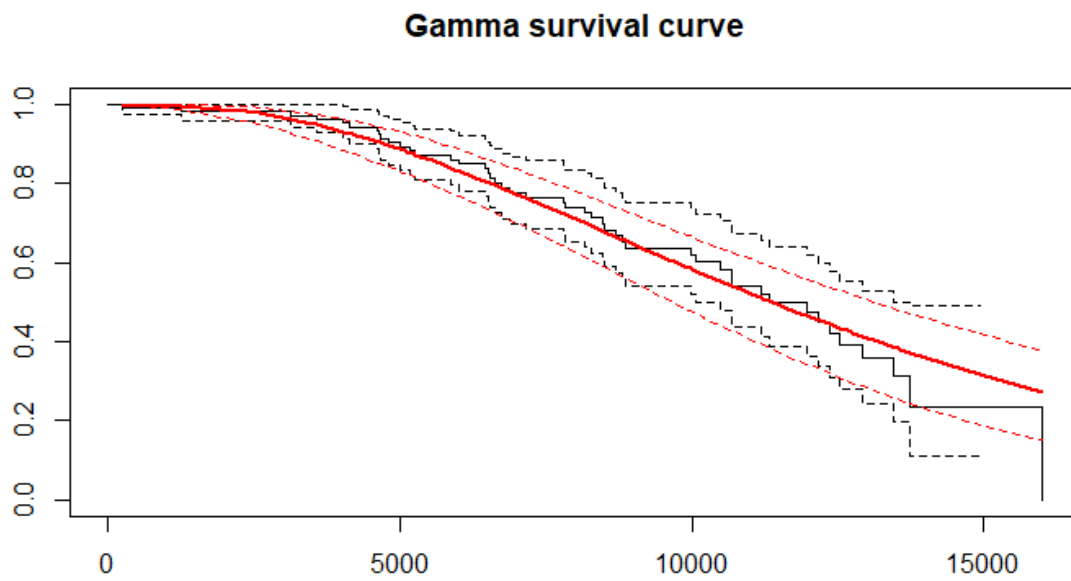


Figure A.7: The Kaplan-Meier curve and estimated Gamma survival curve.

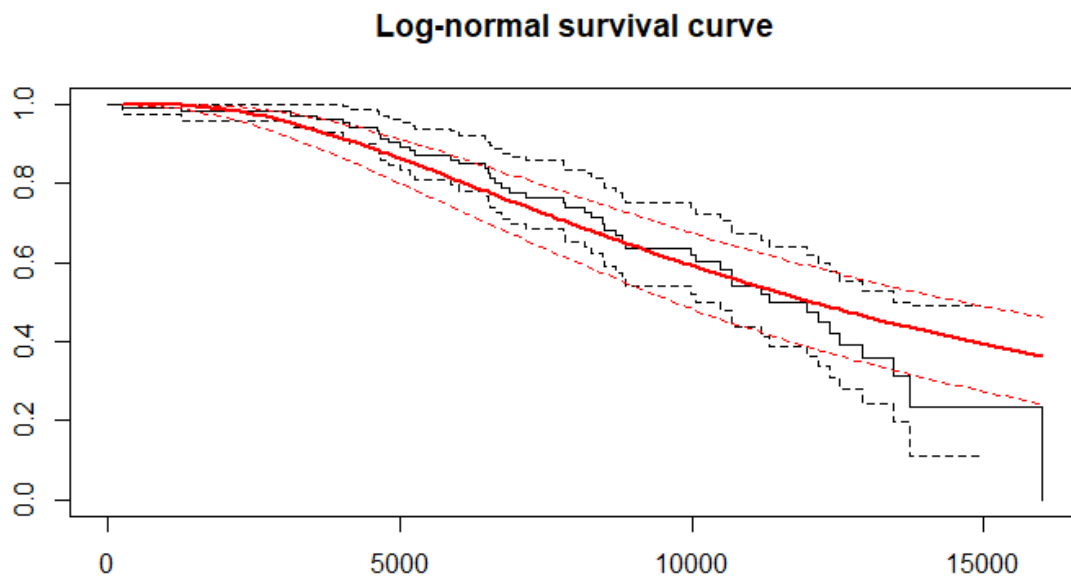


Figure A.8: The Kaplan-Meier curve and estimated Log-normal survival curve.

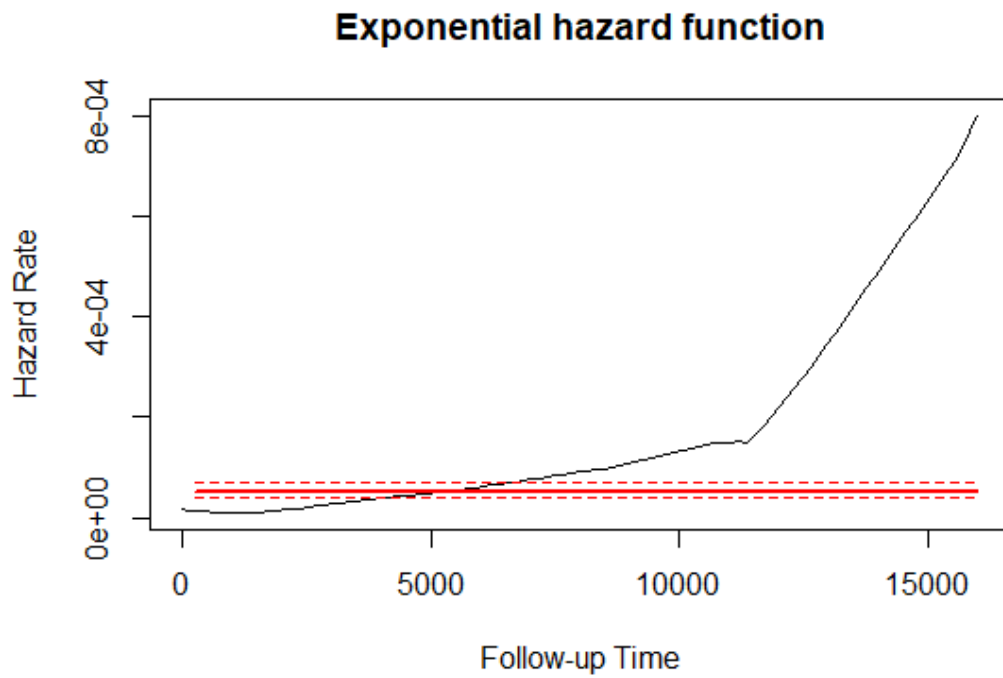


Figure A.9: The corresponding hazard function to the Exponential survival curve.

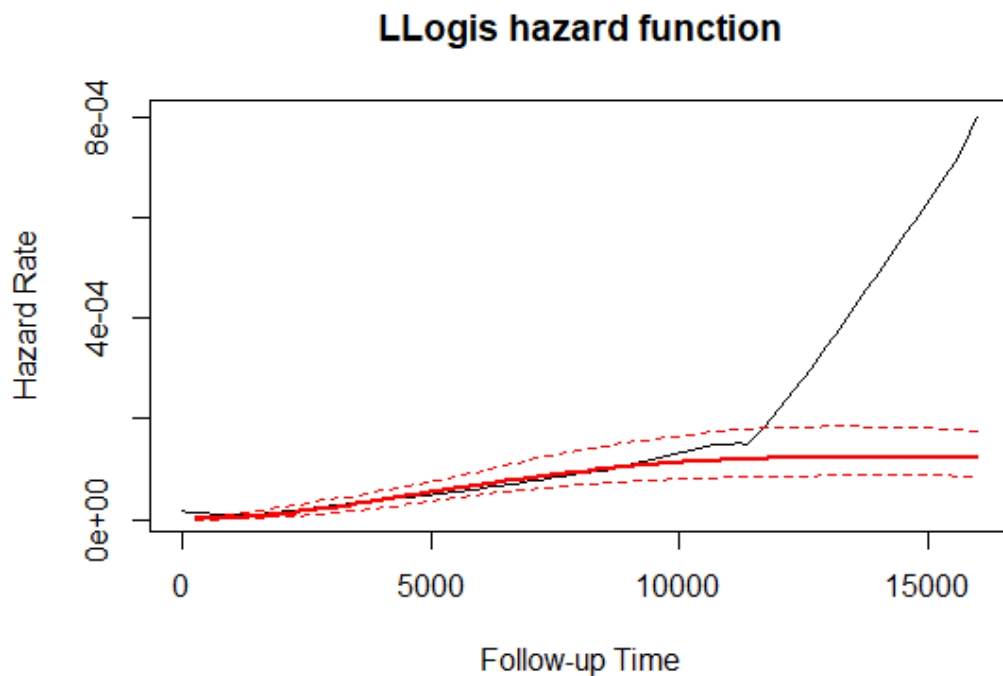


Figure A.10: The corresponding hazard function to the Log-logistic survival curve.

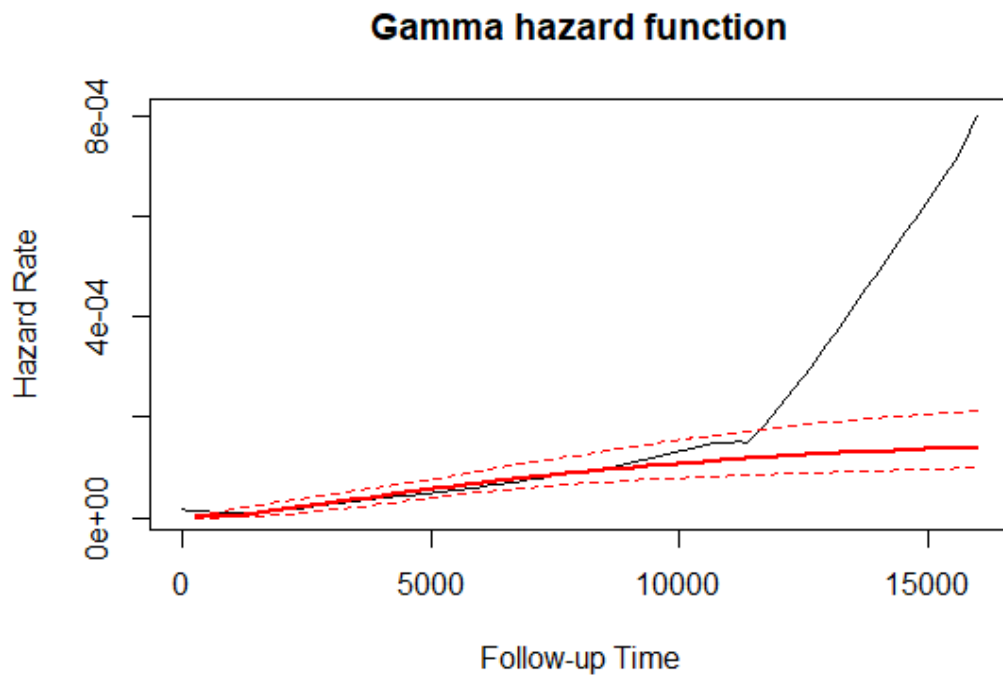


Figure A.11: The corresponding hazard function to the Log-Gamma survival curve.

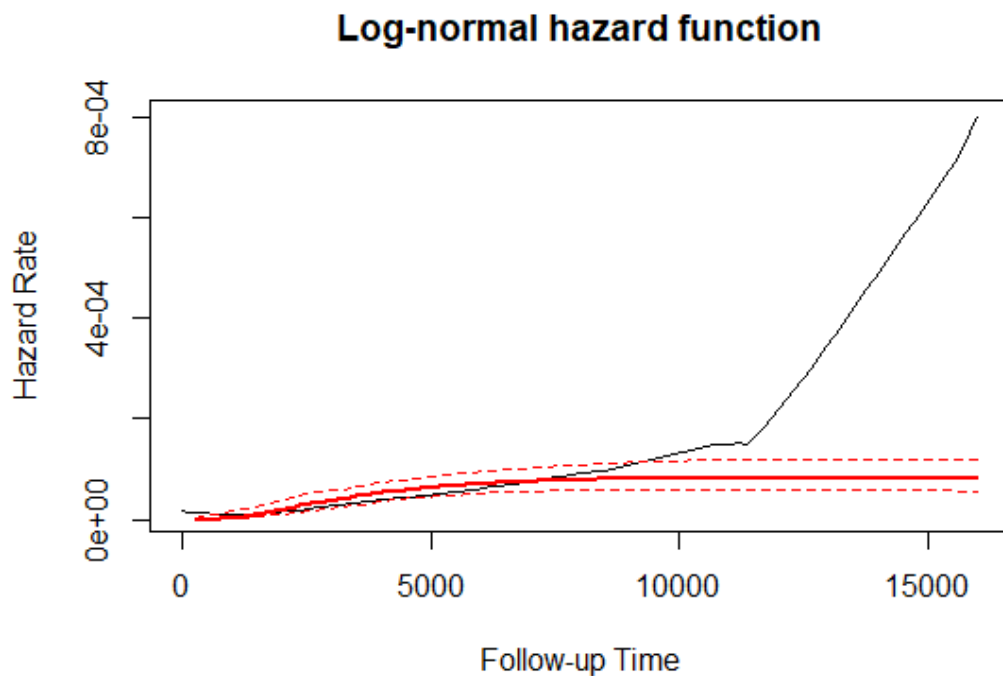


Figure A.12: The corresponding hazard function to the Log-normal survival curve.