Simulation of Automotive Warranty Data

by

Ronald Boyd Anderson

A thesis
submitted to Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Master of Science
in Statistics and Operations Research.

Victoria University of Wellington
2013

Abstract

This thesis will investigate the prediction of the number of claims in a two dimensional automotive warranty claim model for the case of minimal repair. The method involved fitting marginal distributions for age of claim and mileage of claim seperately. Next, various copulas were fitted to establish the correlation between age and mileage, and assessed for fit. The Gumbel copula is chosen as optimal. From this Gumbel copula, a simulation of warranty claims is undertaken. The method produced a good fit for claim age but performed less well for claim mileage, due to the asymmetry of the correlation between mileage and age. Further research directions to improve the accuracy and usefulness of this model are suggested.

Acknowledgements

I would like to thank my supervisor, A/Prof. Stefanka Chukova for pulling me into this field and recommending that I continue onto a masters degree. Her help in providing important guidance and assistance whenever I needed it, was indispensable. An acknowledgement to Prof. Estate Khmaladze, whose insight helped me overcome a research hurdle. A thank you to Dr. Yuichi Hirose and A/Prof. Megan Clark, for helping me finish up my thesis. Many thanks to Kevin Buckley, whose stalwart efforts helped me get my optimisation and simulation software running on the university high density cluster. A special thank you to Dr. Mark Johnston for his optimism, encouragement and enthusiasm.

In addition, an acknowledgement to Ian Collins and Rachel Hunt, who worked on a course project with me which would eventually lead me into this masters. As well as to the current and former occupants of CO526, Ernie, Jens, Joe, Kathleen, Leala, and Xiaoyu for putting up with me over this year. A thank you to Tim Naylor, family friend, for his advice and to Sima for her help. Finally to my family and friends, thank you for your support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis will investigate the simulation of an automotive warranty process, using a parametric two dimensional model. It will begin with a description of a real automotive warranty database, and follow it through to a parametric model for prediction. The accuracy of the model will be compared to a non-parametric model. The goal of this thesis, is to research and apply the two dimensional parametric model first proposed in Baik, Murthy and Jack (2004) [4] and the likelihood function related to this model derived in Chukova and Hirose (2008) [9]. As it is a parametric model, parameters that relate the model to the real dataset will need to be estimated. The task of fitting and of measuring the quality of such a fit will be addressed. Finally data will be generated and compared to the original database.

1.1 Motivation

Warranty Analysis specialises in the analysis of warranty data, in particular aiming to describe and estimate the reliability characteristics of a product. A warranty is a well-defined legally binding agreement between consumer and manufacturer. To a consumer, a warranty serves as an indicator of the quality of a product and an assurance of a products support under

failure. Warranties also provide a mechanism for manufacturers to track failures, and evaluate the field performance of their products. It is also used as a marketing tool and as a legal guideline for dealing with disputes between consumer and manufacturer.

Warranty cost prediction is of immense interest to manufacturers. The overall economic liability of providing a warranty of a certain type and length to their customers is the motivation for prediction in Warranty Analysis. In particular, the accurate prediction of the total cost or number of failures allows manufacturers to plan ahead financially, while still getting the marketing and customer assurance benefits of a warranty agreement. It also allows manufacturers to predict the impact of extended warranty agreements, in which customers pay an additional fee for greater coverage. The warranty coverage area for automobiles is generally defined by both a time and mileage limit. If a vehicle exceeds either, it is no longer eligible to make a claim. In the US the standard agreement (in 2001) was coverage for 36,000 miles and 3 years.

For cost prediction, assumptions are made about the underlying processes which trigger failures. These assumptions are informed by a combination of common sense and expert opinion. In the case of automotive warranty models in this thesis, some of these fundamental assumptions are: that a vehicle "failure" always results in a claim, that the vehicles-with-claims and vehicles-without-claims come from the same population (i.e. no tangible differences between the two groups of cars), and that vehicles will only leave the warranty coverage due to exceeding either or both of the limits (i.e. the number that are "written off" is negligible). Further assumptions that are model dependent, will be discussed as appropriate later in this document.

There are two general classes of models for cost prediction, parametric and non-parametric. Non-parametric models are distinct from parametric ones in that they do not assume any statistical properties of the automobiles population. Therefore non-parametric models require very few

assumptions about the underlying data and as such have been used to describe a multitude of different products with differing reliability characteristics and warranty conditions. This generality does have its shortcomings, such as the inability to sample or generate new data with a comparable probabilistic structure to the real data.

Parametric models on the other hand, impose underlying statistical structure on the data and as such these models can be used in simulation. Parametric models as their name suggests, have parameters that affect their behaviour. Finding these parameters is often an optimisation problem, and is therefore constricted by the "curse of dimensionality" (the increase in computation time as datasets become multidimensional). The predictive power of this approach is dependent on the model chosen, its parameters, and its applicability to the statistical structure of the data. If a model is chosen incorrectly, or differs too much from the underlying data generating process the results can be meaningless. In general, parametric models make more assumptions on the warranty process. In turn, they allow more flexibility in testing the behaviour and cost of offering different warranty conditions to consumers. It is with this in mind, that one arrives at the motivation for finding parametric models for automotive warranty.

This thesis will consider a certain type of automotive warranty, the non-renewing repair warranty with minimal repair in zero time (repair time negligible). The warranty dataset used is two-dimensional (mileage and age), and these two quantities are not independent of each other. To extract this correlation, a joint distribution function is defined in terms of a copula function based on the marginal distributions for both age and mileage. A hazard function is then derived in terms of this joint distribution. The choice and optimisation of these marginals and copulas will be explored. Due to the fact that warranty data for automobiles is bivariate, two dimensional models are more representative of the underlying data. The model chosen is derived from a 2D non-homogeneous Poisson process (NHPP). The justification for using this model will be outlined in the

next section. The predictive power of this model will be examined, and compared with a non-parametric model. Simulated warranty data generated using the model, will be compared to the real warranty data.

The bivariate failure model used in this thesis was proposed by Baik, Murthy and Jack (2004) [4] and then corrected in [5]. Specifically modelling a warranty process as the result of M identical NHPP-derived processes (with the identical hazard functions), where M is the number of vehicles in the population. Each of these NHPP processes models the claim process of one of these vehicles. Baik, Murthy and Jack (2004) assumes minimal repair and homogeneity in the vehicle population, and in this thesis the same assumptions are made. In the publication 'Warranty Data: An estimation of two-dimensional mean cumulative function" [9], a likelihood function is derived for this 2D-NHPP scenario, in terms of the joint hazard function of the process and marginal distributions. Two copulas are suggested, the Positive Stable copula and the Clayton copula. As an extension this thesis will consider nine copulas in total. Previous attempts to optimise this likelihood function using traditional methods of optimisation were unsuccessful. Ultimately the goal is the prediction of mean number of failures per vehicle over the warranty coverage. This can be directly compared with non-parametric models based on the robust estimator [15], for example for the strata model given in [8]. This thesis aims to use different optimisation heuristics/techniques to optimise the likelihood function derived in [9].

1.2 Related Works

The common threads in the literature on estimation of reliability or failure rate in automotive warranty analysis are in dealing with censored data, and intractable optimisation problems. In this section some important models are discussed. We first consider the one dimensional case, in which reliability is estimated as a function of age or mileage.

One of the most important works on estimating reliability measures for automobile data is Hu and Lawless (1996) [15]. The model they propose is a non-parametric model and estimates the mean cumulative function (MCF) of number of claims as a function of age or mileage. It is entirely dependent on knowing $M(t)$, the number of products (vehicles) in the field that are eligible to make claims. The robust estimator of mean number of claims can be defined in terms of both mileage and age. Corrections to $M(t)$ due to vehicles leaving due to age and mileage accumulation are both accounted for. The strata model proposed in Christozov, Chukova, and Robinson (2010) [8] is an application and extension of [15]. The underlying assumption in this model is that between claims, vehicles accumulate mileage linearly. A vehicle trajectory is therefore piecewise linear. This assumption is also made in this thesis. The stratification approach developed in [8], takes into account the distribution of driving patterns of vehicles. All of these models are one dimensional, in terms of either age or mileage. Further extensions, which allow for calendar time (instead of age) and the estimation of a two dimensional MCF are discussed in [13].

A typical approach in warranty analysis is modelling the claim process as a one dimensional non-homogeneous Poisson process [7]. The work of Lawless and Freddete [12], investigates a recurrent event model as a one dimensional mixed non-homogeneous Poisson process. This method implicitly assumes minimal repair and also uses time as age, rather than calendar time. A key difference in this model with earlier publications, is that homogeneity between products is not assumed, this is captured by allowing random effects into the model (mixed Poisson process). These "random effects" describe the non-homogeneous car populations differing failure rates. A likelihood function is derived and then using likelihood maximisation appropriate parameters are found for a real dataset. The prediction of failures is performed by simulating multiple datasets and calculating the total number of warranty claims. This model is computationally expensive as both the optimisation of the model, and the subse-

quent data simulation are non-trivial problems. This model does not take into account vehicles leaving the warranty region due to mileage accumulation. Another interesting extension of this parametric model is given in Kleyner and Sandborn [20]. The extension is the use of piecewise hazard functions to capture the changing failure rate during product life. The choice of which hazard functions to use, and where the change point between hazard functions is of course problem dependent.

The use of neural networks in warranty cost prediction has been researched by Rai and Singh in [26], and by Xu, et al in [30]. Also I did some research on this topic for the requirements of PGDipSci. In general neural networks are used to extract patterns out of data, for the purposes of classification or regression. They are optimised or "trained" by using a subset of the dataset, and are tested using the another unseen part of the dataset. These are known as the training and test patterns (or datasets) respectively. In Rai and Singh [25], a model used to predict cumulative cost using neural networks is investigated implicitly assuming no censoring. For each vehicle in the field at age i at month n, the cumulative cost of its claims was calculated. In month one, there are only vehicles of age one month that are able to make claims. In month two, there are now vehicles that are one month old and vehicles that are two months old, and so on. These costs are the training and test sets for the neural network. Using the standard back-propagation algorithm, the neural network is trained until it performs well on the training set. It is then tested on the test set, this provides an indicator of its performance on unseen data, and subsequently its predictive power. Neural networks can be retrained as new information or warranty data becomes available, and as such every month this model can be improved. The evaluation of the accuracy is estimated by fitting the neural network multiple times and looking at the sensitivity of the model.

In my project, I extended the work in [25] to use censored data by employing the stratification approach as in [8]. The correction introduced in this approach removed the cars which had left the warranty region due

to mileage. The warranty cost prediction in this approach had a smaller error and was within the 95% confidence interval for the linear regression model.

The most common criticism of neural networks is their tendency to be "overfitted", whilst they perform well on the training data, they are useless for prediction. To limit this problem, experiments to fine tune starting parameters, neural network layout and size, and optimisation strategies are commonly used; Wu and Akbarov [29], used a similar Machine Learning approach. They used prediction using Support Vector Machines for regression as they are not as susceptible to overfitting. Support vector machines fit a hyperplane in a high-dimensional space to either perform classification or regression analysis. One important assumption made about the nature of the warranty process in this approach, is that recent warranty claims are more indicative of failure rate than older claims. This is also an assumption in the neural network model detailed above. As a result these newer claims should have more "weight" in the prediction of future claims and subsequently the overall warranty cost.

The work on a parametric two dimensional model for automotive warranty claims is still somewhat undeveloped. Baik, Murthy and Jack (2004) [4] introduce a model that is an increasing stochastic process generated by an underlying two dimensional non-homogeneous Poisson process. This model will be discussed in the next chapter as the basis for this project. An excellent review of the field of Warranty Analysis can be found in "Warranty Data Analysis: A Review" by Shaomin Wu [28].

Chapter 2

Warranty Prediction: An Overview

In this chapter we discuss the ideas and issues of warranty cost prediction. We will also introduce some important concepts in warranty analysis and statistical modelling which will be used in this thesis. Finally the model that will be used in subsequent chapters is detailed.

2.1 Warranty Agreements

> "A warranty is a seller's assurance to a buyer that a product or service is or shall be as represented. It may be considered to be a contractual agreement between buyer and seller (or manufacturer) which is entered into upon sale of the product or service" [6].

A warranty agreement is a legally binding contract between a consumer and a manufacturer/retailer. It provides an assurance to consumers about the quality of a product. Manufacturers or retailers use warranty agreements as a marketing tool and also use them as a means to track the quality and failure rates of their products. Warranties have become standard prac-

tice in most industries and are even legally required in some. Indeed, it is because of this, that the analysis of warranty data is of immense interest to manufacturers/retailers.

Warranty agreements come in different types depending on the nature of the product, and legal requirements. A warranty is *renewing* if the warranty agreement is extended after a claim is made. Warranties can be *free-replacement* if repair or replacement of the failed product is at no cost to the consumer. Alternatively they can be *rebate* warranties if under failure, the consumer is returned the value of the product (known colloquially as a "Money Back Guarantee"). Warranties can also be multi-dimensional, for example in the automotive industry, both mileage and age are used to define the warranty coverage. In this thesis we consider the 36 month and 36,000 miles warranty agreement.

2.1.1 Repair Types

Usually, after a failure occurs to a product, it must be repaired or replaced. A product can be repairable, non-repairable, or complex (containing both repairable and non-repairable components). If a product is non-repairable, it has to be replaced after failure. Vehicles are complex products, as they have both parts which can be repaired (for example, the transmission) and parts which have to be replaced (for example, headlights). As is often the case in the real world, repair work on a product affects its expected lifetime. Product repair can be

- Complete: a product is repaired to a "good as new " state.
- Minimal: a product is repaired to a state identical to its condition immediately before failure. For example, repairs which do not affect the failure rate of the product (repairing windscreen wipers on a vehicle).

- Imperfect: a product is repaired to a state better or worse than its condition at time of failure.

In some situations, the repair or replacement time required for the rectification of a failure is negligible, and can be ignored as is the case in this thesis. In other cases this repair time or delay has to be factored into any potential warranty model.

2.1.2 Cost Prediction

In general, manufacturers would like to know the expected cost or number of claims per vehicle at an age or mileage of a vehicle's life, this is known as the mean cumulative function (MCF). The MCF is independent of the number of vehicles sold, meaning that the increasing liability of offering a warranty can be easily calculated per vehicle, separate from the sales process. Automotive warranty claims are recurrent events, this means that multiple failures can occur to one vehicle. The goal in this thesis is estimating the likelihood of a failure and subsequently simulate warranty claims to find the expected number of claims per vehicle.

However due to the nature of the claims process, warranty databases are censored. They have both missing and right censored values, this makes analysis and prediction non-trivial. Data censoring is when some values in the database are not known exactly, in the case of automotive warranty they are only known to be above a certain threshold (right censoring). An example of this is a car which fails outside of the warranty coverage due to mileage. The warranty coverage W, is defined as the period in which a vehicle is eligible to make a claim, often measured in calendar time since purchase and a mileage threshold.

In this thesis we will consider the case of automotive warranty which is a *non-renewing repair* warranty with repairs being *minimal* and repair time being *negligible*. We assume that the vehicles with claims and vehicles without claims, come from the same population. In other words, that

there is no difference in these two groups, as they both come from one manufacturing and quality assurance process. We also assume that every failure results in a claim, and therefore claims are indicative of failure rate.

2.2 Survival and Recurrent Event Analysis

Survival analysis (or reliability analysis) is a field that is concerned with the time-to-failure of systems. A product lifetime is the time from sale date to a failure. In this section we will introduce a few important concepts which will be used in this thesis.

2.2.1 Survival and Hazard function

The survival function (also known as the reliability function) is defined as the probability that a product lifetime is greater than a time x, or more formally: If X is the survival time (time to failure or lifetime), with cumulative distribution function (CDF) $F_X(x)$, the survival function $S_X(x)$ is defined as:

$$S_X(x) = P(X > x) = 1 - F_X(x) \ .$$

The *failure rate* of a product is the frequency at which a failure occurs to a product over a fixed time period. The continuous time analog of failure rate is the hazard function $h(x)$ (also known as force of mortality):

$$h(x) = \frac{f(x)}{S_X(x)} \ .$$

The hazard function gives the instantaneous failure rate at a point x, it is not a probability nor a density. However we can think of it as the probability of a failure occurring in the small interval $(x, x + \delta x)$, given that the product has survived till x. The cumulative hazard is the integral of the hazard function.

$$H(x) = \int_0^x h(s)ds \ .$$

The survival function can also be written in terms of the cumulative hazard function as below:

$$S_X(x) = e^{-H(x)} \ . \tag{2.1}$$

The failure rate given by the hazard function determines the expected lifetime of a product. The failure rate can be increasing, decreasing or constant over the lifetime of a product. In general, products have a high failure rate early in a products life, this is called the "infant mortality period". This trend is present in the dataset used in this thesis.

2.2.2 Renewal and Poisson processes

A natural approach of modelling product failures is as a stochastic process, where each failure is a counted event. We will define it formally here. Let X_i be a strictly positive i.i.d. random variable, and consider a sequence of $\{X_1, X_2, X_3, ...\}$. We can think of this sequence as a collection of "failures times". We can then define the number of events that occur before time x as

$$N(x) = \sum_{i=1}^{\infty} 1_{\{X_i<x\}} \ ,$$

where $1_{\{X_i<x\}}$ is the indicator function. This $\{N(x), x \geq 0\}$ is called a renewal process. The expected value of $N(x)$ is dependent on the distribution of X_i, however when no closed form solution exists, simulation can be used to estimate $E\left[N(x)\right]$.

We will introduce a non-homogeneous Poisson process (NHPP) by firstly introducing stationary Poisson process, and then relaxing its assumptions to get NHPP. Assume all non-overlapping increments $N(x+y) - N(x)$ are independent, and at the beginning of the process, no events have occurred ($N(0) = 0$). Next, enforce stationary increments, that is to say, the probability of the number of claims in an interval is only dependent on the length of that interval. This is the form of the Poisson process, $\{N(x), x \geq 0\}$. The expected value of $N(x)$ has a well understood closed form solution (as the

number of events in a time interval is distributed as Poisson). The holding times are exponentially distributed with a rate parameter or intensity, traditionally denoted λ. A Non-homogeneous Poisson process relaxes the restriction on stationary increments, so that intensity $\lambda(x)$ is dependent on x. These processes can be extended to be multi-dimensional, as in the 2D NHPP warranty scenario in this thesis.

2.2.3 Mean Cumulative Function (MCF)

As populations for different vehicles can be of different sizes, an estimate that is independent of population size is desirable. The mean cumulative function $\Lambda(x)$ is the population mean cumulative number (or cost) of claims per vehicle up to a time x. This is a useful indicator of product performance and the main quantity of interest in automotive warranty analysis. The value of the MCF at a time x can be interpreted as the expected total number of claims per vehicle up to a time x. This estimation in combination with information on the sales process (how many vehicles have been sold) gives an estimate of the manufacturers total number (or cost) of claims. In this thesis we will only consider the estimated number of claims, not the cost. Furthermore we will not delve into modelling the sales process.

2.2.4 Lifetime Distributions

Next we provide a brief summary of a few lifetime distributions, to be used in Chapter 3. These distributions are defined only for $x \geq 0$.

Weibull Distribution

Weibull distributions can be used with products with increasing ($\beta < 1$) or decreasing ($\beta > 1$) failure rates. The CDF is given by

$$F(x, \lambda, \beta) = 1 - e^{-(\frac{x}{\theta})^{\beta}} \ , \ \lambda, \beta > 0.$$

where θ is the scale parameter and β is the shape parameter of the distribution.

Exponential Distribution

Exponential is a special case of the Weibull distribution, where $\beta = 1$. The CDF of this distribution is given as

$$F(x, \lambda) = 1 - \lambda e^{-\lambda x} \ , \ \lambda > 0.$$

The exponential function can be used when a constant failure rate λ is required.

Log-normal Distribution

The log-normal distribution has the following form for its CDF:

$$F(x, \mu, \sigma) = \frac{er\left(-\frac{\ln x - \mu}{\sigma\sqrt{2}}\right)}{2} \ ,$$

where $er(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt$ is the complementary error function, μ is the mean and σ is the standard deviation.

2.2.5 Proportional mixture distribution

The proportional sum of cumulative distributions is itself a cumulative distribution, known as a proportional mixture distribution. This mixture distribution for k groups has the form:

$$F_X(x) = \sum_{j=1}^{k} \pi_j F_j(x) \ , \tag{2.2}$$

where π_j are greater than zero and:

$$\sum_{j=1}^{k} \pi_j = 1 \ .$$

This distribution is useful if the data being analysed is thought to be a mixture of samples from different populations with known proportions. We will consider the distribution of the time to first, second, third and fourth claims in Chapter 3.

2.3 Non-parametric warranty models

Non-parametric models are ones that make no assumptions about the underlying probability distributions of product lifetimes. The field of non-parametric warranty models has been successful in cost prediction.

2.3.1 The CR Model

The CR model is based on the robust estimator, first suggested by Hu and Lawless (1996) [15]. The robust estimator is appropriated for the automotive industry and extended in Chukova and Robinson (2005) [10], and is known as the CR-Model. In this model a further assumption is made on the trajectory of the vehicles. This assumption is, from the beginning of a vehicles life, each car accumulates mileage approximately linearly with time. This assumption is confirmed by an analysis of the databases used in the paper. Next we will provide a brief review of the main result of [10]. First we will introduce t as a placeholder for either mileage or age. For the purposes of this model we are only considering discrete time (in this case months). We shall define $n_i(t)$ as the number of warranty claims for vehicle i at time t. The total number of claims of all vehicles at a time t is defined as

$$n(t) = \sum_{i=1}^{M} \delta_i(t) n_i(t),$$

where M is the total number of vehicles in the field, and $\delta_i(t)$ is an indicator function, signalling if the vehicle i is under observation (i.e. it is eligible to

make claims) at time t. Thus the intensity function can be estimated

$$\hat{\lambda}(t) = \frac{n(t)}{\hat{M}(t)},$$

where $\hat{M}(t) = \Sigma_{i=1}^{M} \delta_i(t)$ is the number of vehicles eligible to make a claim at time t. Note that in most automotive warranty agreements, vehicles can leave due to either mileage or vehicle age. Finally the mean cumulative estimator can be defined as below

$$\hat{\Lambda}(t) = \sum_{s=1}^{t} \hat{\lambda}(s)$$

The variance of this mean cumulative estimator is also derived in [15] as

$$V[\hat{\Lambda}(t)] = \sum_{i=1}^{M} \left(\sum_{j=1}^{k} \left[\frac{\delta_i(t_j) n_i(t_j)}{M(t_j)} - \frac{\hat{\lambda}(t_j)}{M} \right] \right)^2,$$

where t_j is a regular time interval such that $0 < t_1 < t_2 < .. < t_{k-1} < t_k = t$, for example days or months. Time in this model can be either mileage, age of the vehicle or even calender/actual time [13]. Due to the nature of failure events, we know $n(t)$ exactly, however the estimation of $M(t)$ is dependent on the vehicles that leave the warranty region due mileage or age.

Mileage Accumulation Rate

In the literature on automotive warranty data, the Mileage Accumulation Rate (MAR) is the amount of mileage accumulated over a fixed time period. In this thesis it is measured as mileage per day. Understanding the way vehicles accumulate mileage is important in the estimation of $M(t)$. First define the nth claim of vehicle i as the age and mileage pair $(T_{i,n}, U_{i,n})$. Also we will define that starting age and mileage as $T_{i,0} = 0$ and $U_{i,0} = 0$ respectively, and the age and mileage cut-off values as T_{max} and U_{max}. We will denote the MAR for vehicle i at claim n as $r_{i,n}$, it has the form of:

$$r_{i,n} = \frac{U_{i,n} - U_{i,n-1}}{T_{i,n} - T_{i,n-1}} \tag{2.3}$$

The MAR over a vehicles entire recorded life is defined as below,

$$\frac{U_{i,max}}{T_{i,max}}, \tag{2.4}$$

where $(U_{i,max}, T_{i,max})$ is the last recorded claim of the vehicle. The empirical distribution of this overall MAR is denoted $F_r(x)$, and estimates the probability of a vehicle having a MAR less than or equal to x. The empirical distribution of $r_{i,n}$ is similarly denoted $F_{r_n}(x)$, and gives the probability of a vehicles nth claim having a MAR less than or equal to x. We will consider these empirical distributions in Chapter 3.

Time as age

The most intuitive way to treat time in the CR-model is the age of the vehicle. In this model we first ignore the vehicles that leave the warranty coverage due to the mileage. The number of vehicles which are eligible to make claims at a time t is therefore

$$\hat{M}(t) = \sum_{i=1}^{M} 1_{\{A_i \geq t\}}$$

where A_i is the censored age of vehicle i. Using this $\hat{M}(t)$ in evaluating the MCF ($\hat{\Lambda}(t)$), produces the following graph plotted in Figure 2.1. However this doesn't take into account vehicles which leave warranty coverage before the age cut-off time due to mileage. If we compensate for these vehicles, we get what is known as the adjusted mean cumulative function. In this case, we must consider the two groups of vehicles, those with claims and those without claims. Let M_1 be the number of vehicles with claims, and M_2 be the number of vehicles without claims, such that $M = M_1 + M_2$. For vehicles with claims, we check if the vehicles age is greater than the time t and also if the estimated MAR is less than the mileage limit U_{max} at time t. The first sum is vehicles which are alive at a time t and have not exceeded the mileage coverage (from our best estimate of their mileage accumulation), and the second sum is for vehicles with no claims. For these

vehicles we estimate the number that do not exceed the mileage coverage by using the empirical distribution of MAR $F_{r_i}(x)$.

$$\hat{M}(t) = \sum_{i=1}^{M_1} 1_{\{A_i \geq t\}} 1_{\{r_{i,max} \leq U_{max}/t\}} + \sum_{i=1}^{M_2} 1_{\{A_i \geq t\}} F_r(U_{max}/t)$$

Time as mileage

Using mileage as time is another way of estimating the mean cumulative function. Similarly in the time as age case, we consider vehicles with claims and vehicles without claims separately. A vehicle with a claim is only part of the population of vehicles able to make claims ($M(m)$), if its mileage is less than the mileage m. We can estimate this vehicles mileage as $r_{i,max}A_i$, and therefore if $r_{i,max} \geq m/A_i$ then the vehicle is able to make claims at a mileage m. For vehicles without claims, we can estimate a vehicles mileage accumulation using the overall empirical MAR distribution. Therefore the adjusted estimate of the number of alive vehicles at mileage m is:

$$\hat{M}(m) = \sum_{i=1}^{M_1} 1_{\{r_{i,max} \geq m/A_i\}} + \sum_{i=1}^{M_2} 1 - F_r(m/A_i)$$

Note in Figure 2.2 that the MCF is roughly linear throughout the mileage lifetime.

Linear Regression for Prediction

A simple prediction method is to attempt linear regression on $\hat{\lambda}(t)$. We shall consider the MCF over a time interval (say one month), this will be denoted $\hat{\lambda}(t_i)$. This is the mean number of claims per vehicle in month i. The assumptions made on $\hat{\lambda}(t_i)$ are that they are independent, approximately linear and that their errors are normally distributed such that.

$$\hat{\lambda}(t_i) = ci + \epsilon_i$$

Using least squares regression, the c parameter is chosen to best fit values of $\hat{\lambda}(t_i)$. This line is then used to predict $\hat{\lambda}(t_i)$ for future months. The MCF

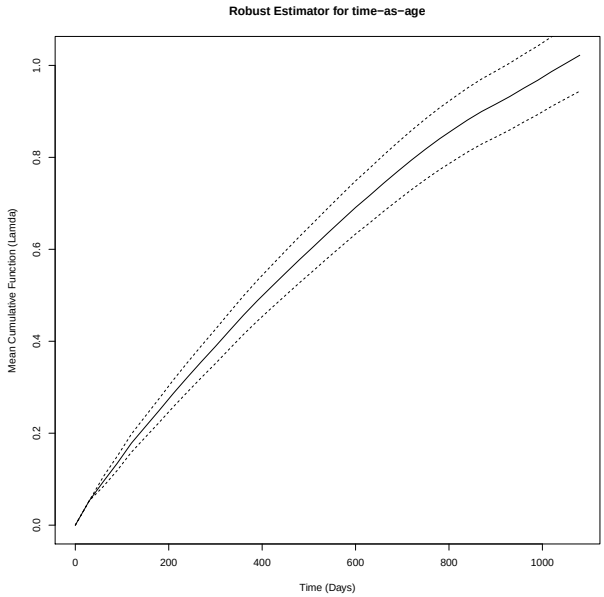

Figure 2.1: Time as age - CR-model example

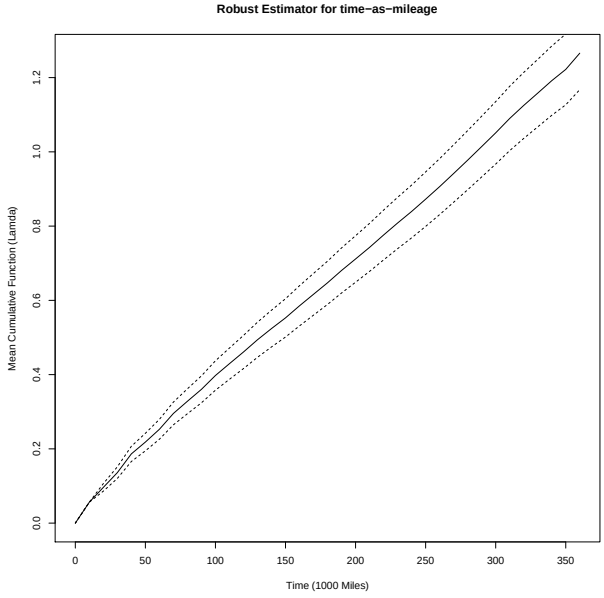

Figure 2.2: Time as mileage - CR-model Estimator example

$\hat{\Lambda}(t)$ is therefore simply the sum of these predicted $\hat{\lambda}(t)$. Unfortunately linear regression outside of the sample range has a large uncertainty. Note in Figure 2.3 how wide the 95% confidence interval of the prediction is, even giving possible negative values for large t. For an in-depth discussion of this method, refer to [13].

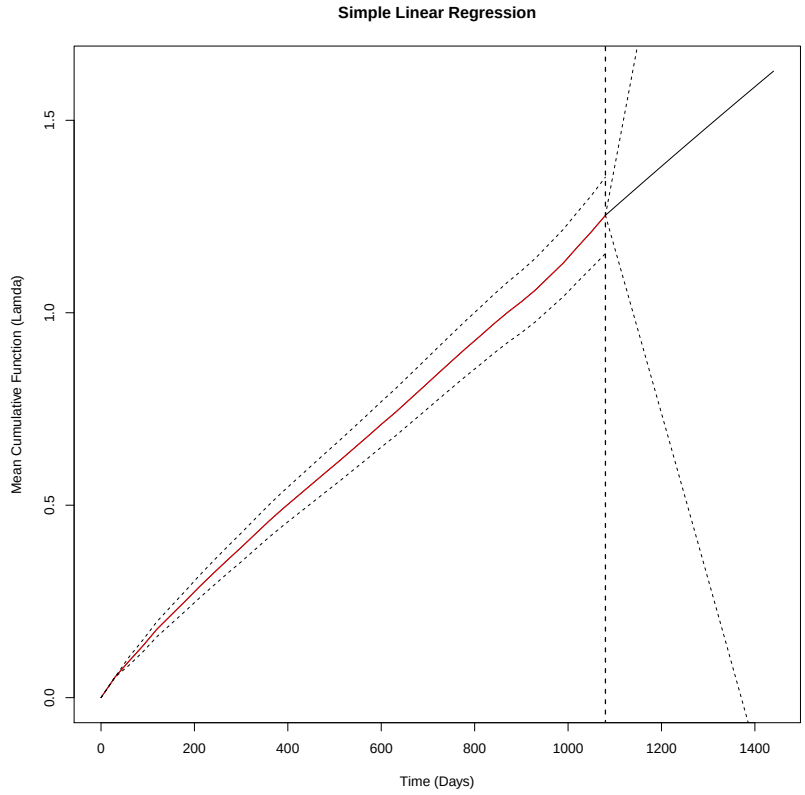

Figure 2.3: Linear regression

The wide confidence interval for the values of $\hat{\Lambda}(t)$ indicates that this method is not appropriate for prediction of the expected warranty cost.

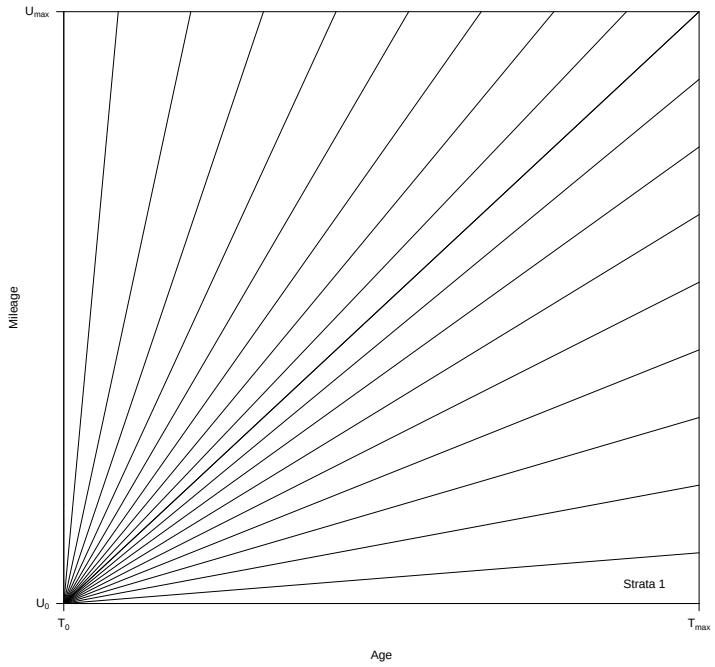

Figure 2.4: Diagram showing strata

2.3.2 Strata model

The strata model uses the same concept of the CR-model but relaxes the assumption of linearity of mileage accumulation. The model is named for the stratification of the warranty region into mileage accumulation rate strata. In the paper [8], $k = 72$ regions are defined for the standard US vehicle warranty, the 36 month or 36,000 mile warranty region. A simplified version of these strata is shown in Figure 2.4, note the position of the strata 1. The warranty region is divided into age-bins of size of one month and mileage-bins of size of 1000 miles. First the cars with claims are used to build a empirical distribution of MARs, counting the number of vehicles which fit into one or more of these strata. Vehicles with highly variable MARs, are known as not stable. The probability of a vehicle having a

MAR in the strata s is therefore

$$p_s = \frac{O_{1,s}}{M_1},\ s = 1, 2, 3..., k$$

where M_1 is the number of vehicles with MAR information (claims) and $O_{1,s}$ is the number of cars with claims with a MAR within the stratum. It follows in the time-is-age case, the number of vehicles in the field able to make claims at time t is:

$$\hat{M}(t) = \left(M - \sum_{i=1}^{t} N_i \right) \left(\sum_{s=1}^{72-t} p_s \right),$$

where N_i is the number of vehicles within age-bin i. Prediction in this method is performed as in the CR-Model, simple linear regression on $\hat{\lambda}(t)$, and provides similar prediction uncertainty. Figures 2.5 and 2.6 show the adjusted and unadjusted strata model for age and mileage. Note that for age, the correction is statistically significant, however for mileage it is not.

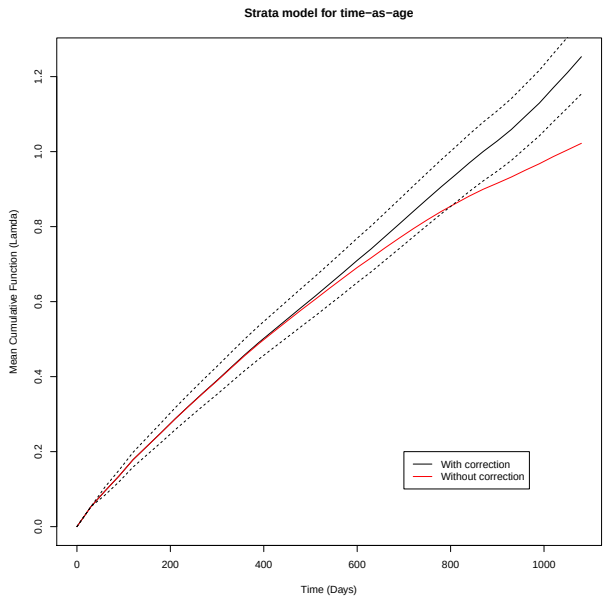

Figure 2.5: Time as age - Strata Model example

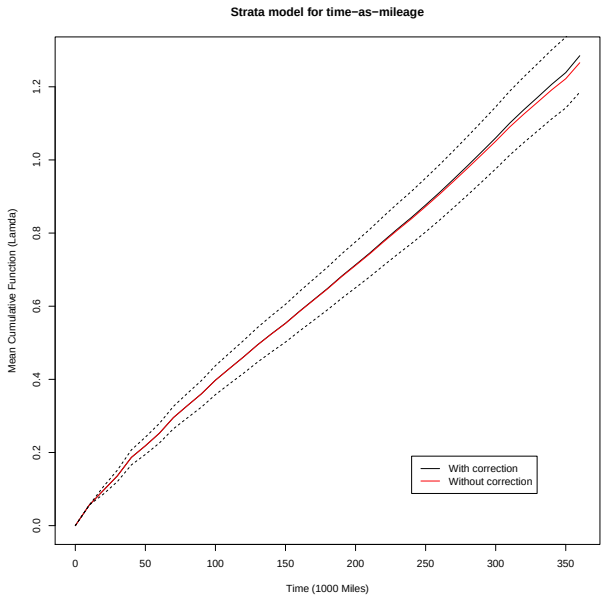

Figure 2.6: Time as mileage - Strata Model example

2.3.3 Neural Networks

The use of neural networks to predict warranty costs due to repairs has been successful for short term predictions as in [25]. A multi-layer perception (MLP) network is a type of neural network, which can be used for regression. Conceptually you can treat a neural network as a black box, with inputs and outputs. For the purposes of warranty analysis, the inputs are age or mileage and the output is the expected cost (or number) of failures per vehicle. The internals of this black box are best described as a weighted connected graph with multiple layers, each fully connected to the next. Each node in this graph has multiple incoming edges and one outgoing edge, all of which are weighted. Similar to a neuron in biology, a node in the neural network will ”fire“ if its inputs meet certain criteria. This criteria is meeting a threshold given by the nodes input weights. The graph weights need to be chosen to give the correct output for a given input. For example, the input ”the age two months“, should output the

mean cumulative number of claims at age two months.

The method to "learn" these weights is known as the *back-propagation* algorithm, an optimisation algorithm. The neural network is presented with a training set, a collection of known ages or mileages and the number of known failures at that age or mileage. This optimisation process minimises the error of the output by changing the weights. Once the error is small enough, it can be used for prediction. To use the trained neural network for prediction, input the age or mileage required and the network will output the estimate. However the accuracy of the prediction has to be estimated by training multiple MLPs, and considering the mean of their outputs.

The most prevalent criticism of the use of neural networks is that the trained network is very much dependent on the quality of the training set. The training set must be indicative of the process being studied, the larger the training set the better. Furthermore, the problem of over fitting is common. A neural network can be too large and as such perfectly predict costs for known data, but perform poorly for unknown data. Finally a major downside to neural networks, is that although they can perform very well in prediction, they do not provide insight into the relationship between the inputs and outputs. Determining what kind of relationship (for example linear, or exponential) the MLP has captured is very difficult. That is to say, it remains a black box.

In my own research for the requirements of a post-graduate diploma I used a MLP to predict number of claims at a time t. I extended the work done in [25] to incorporate the censoring due to age and mileage. This was performed by using the strata model to estimate the mean cumulative function at age t. The 95% confidence interval was generated by optimising 30+ neural networks and comparing their output. Note in Figure 2.7 that it has a much smaller confidence interval than the simple linear regression approach discussed earlier.

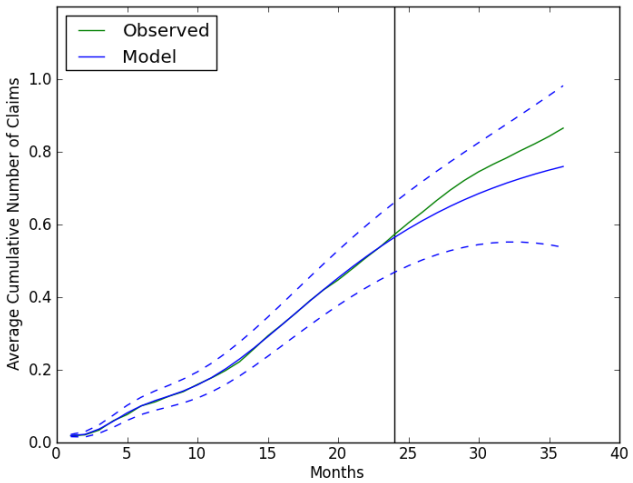

Figure 2.7: Prediction using Neural Networks

2.4 Parametric warranty models

Parametric models are models which assume an underlying distribution of time to failure/claim. In this section we will assume minimal repair, and negligible repair time. We will use the non-homogeneous Poisson process, for both one and two dimensional cases. In this section we will only consider the failure of one vehicle, therefore $E[N(t)]$ is equivalent to $\hat{\Lambda}(t)$ in the previous models.

2.4.1 One Dimensional Case

In the one dimensional case with minimal repair, an analytical model was derived in [7]. The probability of a failure in an increment of time $(t, t+\delta t)$ is proportional to the hazard function. Due to minimal repair this is even true for subsequent failures, as repairing has no effect on failure rate. Now consider the stochastic counting process $\{N(t), t \geq 0\}$ where $N(t)$ is the number of failures up to time t. Under minimal repair, the probability of

n failures in an interval $[0, t]$ is Poisson with intensity $H(t)$, and therefore $\{N(t), t \geq 0\}$ is a non-homogeneous Poisson process.

$$p_n(t) = \frac{e^{-H(t)} H(t)^n}{n!}$$

The expected value of the NHPP process $N(t)$, which is the expected number of claims per vehicle in our model, is simply the integral of the hazard function $h(s)$.

$$E\left[N(t)\right] = \int_0^t h(s)ds$$

As such, this integral can be solved analytically or numerically to get a prediction of the number of failures per vehicle. For more discussion on this model, see [6].

2.4.2 Two Dimensional Case

To understand why the one dimensional case cannot be simply extended to two dimensions we first will consider how claims occur. Consider a vehicle with two claims one at (T_1, U_1) and another at (T_2, U_2). An example of such a vehicle is presented in Figure 2.8, note that $T_2 > T_1$ and $U_2 > U_1$. These claims define the possible region that the vehicle could have moved through. The trajectory the vehicle took to get through these two points is unknown. Consider the three possible trajectories given in Figure 2.9. Each of these trajectories could have generated these two claims. In any case, we know that there are no claims in the regions $[T_0, T_1] \times [U_0, U_1]$, $[T_1, T_2] \times [U_1, U_2]$ and $[T_2, T_{max}] \times [U_2, U_{max}]$. So in the case of two dimensions we introduce a sequence of bivariate pairs $\{(T_n, U_n)\}_1^\infty$ where $T_n < T_{n+1}$ and $U_n < U_{n+1}$, which forms an increasing stochastic process. This process can be described as a stochastic process generated by a two-dimensional non-homogeneous Poisson process $\{N(t, u), t \geq 0, u \geq 0\}$ (see the revised derivation in [5]). What follows is the derivation from [5]. First we assume that (T_1, U_1) is a non-negative bivariate random variable with a distribu-

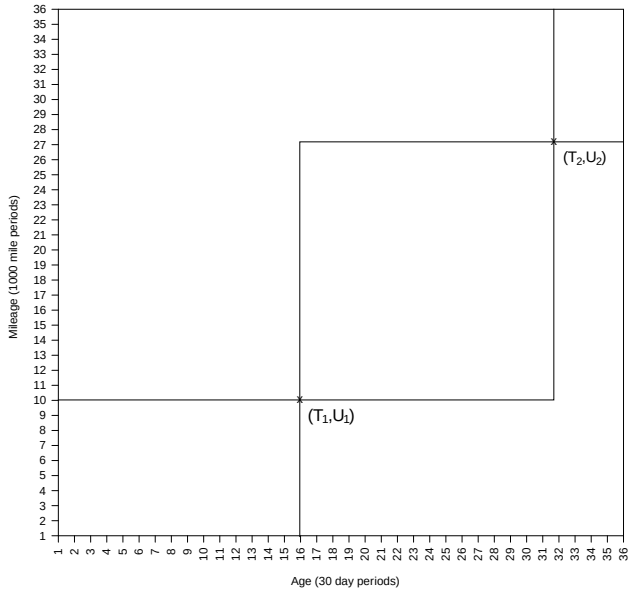

Figure 2.8: Example of claims for one vehicle only

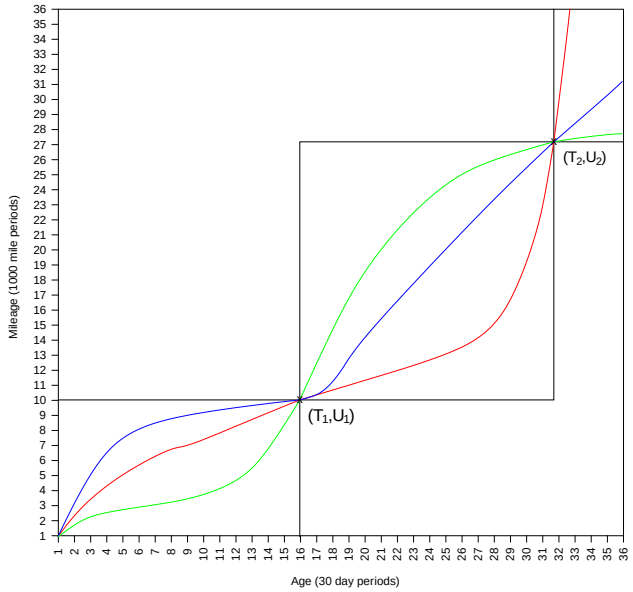

Figure 2.9: Example of three possible trajectories through claims

tion function given by

$$F(t,u) = P(T_1 \leq t, U_1 \leq u) \tag{2.5}$$

We also assume that $F(t,u)$ is differentiable, and therefore its bivariate density function is given as the mixed derivative below

$$f(t,u) = \frac{\partial^2 F(t,u)}{\partial t \partial u} \tag{2.6}$$

The joint hazard function of this joint distribution is therefore

$$h(t,u) = \frac{f(t,u)}{\overline{F}(t,u)} \tag{2.7}$$

Now consider the probability of a failure occurring at $(t, t+\delta t) \times (u, u+\delta u)$ (as in Figure 2.10), under minimal repair

$$P\{N(t, t+\delta t, u, u+\delta u) = 1\} = \lambda(t,u)\delta t \delta u + o(\delta t \delta u), \tag{2.8}$$

the intensity function is therefore

$$\lambda(t,u) = \lim_{\delta t \to 0, \delta u \to 0} \frac{P\{N(t, t+\delta t, u, u+\delta u) = 1\}}{\delta t \delta u} \tag{2.9}$$

Under minimal repair the intensity function is equal to the hazard function ($\lambda(t,u) = h(t,u)$), we will use them equivalently here. Furthermore, the probability of no failure occurring in the region $(t_{i,j}, t_{i,j+1}) \times (u_{i,j}, u_{i,j+1})$ is one minus the probability of a failure occurring in that region and is given by

$$p_0(t_{i,j}, t_{i,j+1}, u_{i,j}, u_{i,j+1}) = 1 - \left[\frac{F(t_{i,j+1}, u_{i,j+1}) - F(t_{i,j}, u_{i,j+1}) - F(t_{i,j+1}, u_{i,j}) + F(t_{i,j}, u_{i,j})}{\overline{F}(t_{i,j}, u_{i,j})} \right] \tag{2.10}$$

In order to consider the probability of n failures in the region $[0,t] \times [0,u]$, we must consider the probability of having $n-1$ failures in a subregion $[0,s] \times [0,r]$ where $s < t$ and $r < u$. Then the probability of one failure occurring at (s,r) ($h(s,r)drds$) and no failure between $[s,t] \times [r,u]$

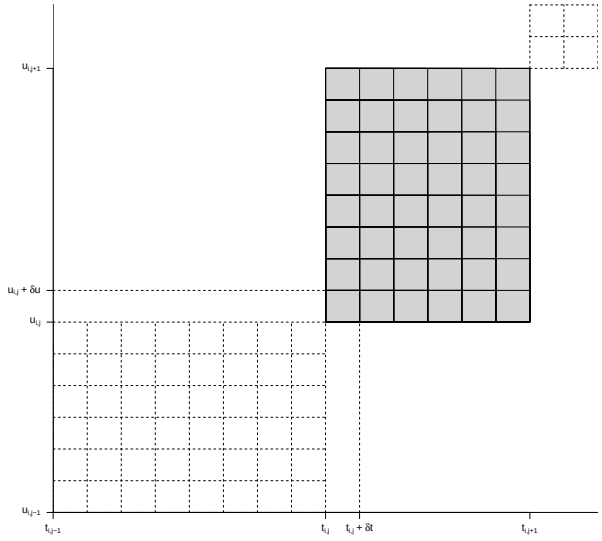

Figure 2.10: Evaluating the hazard function at each point

($p_0(s,t,r,u)$). If we consider all possible values of s and r, it follows that it has the recursive form

$$p_n(t,u) = \int_0^t \int_0^u p_{n-1}(s,r)p_0(s,t,r,u)h(s,r)dsdr \tag{2.11}$$

Therefore the expected number of failures in the region $(0,t) \times (0,u)$ is the sum below

$$E\left[\Lambda(t,u)\right] = \sum_{n=0}^{\infty} np_n(t,u) \tag{2.12}$$

Unfortunately this estimated number of claims $E\left[\Lambda(t,u)\right]$ has no known closed form solution. As such to estimate this quantity we will use simulation, this is described in Chapter 6. We will model the population of M cars as M independent stochastic processes generated by a NHPP with an intensity given by the hazard function $h(t,u)$. Now that we have this stochastic process, we would like a likelihood estimator function to find optimum parameters for this model, given a real dataset.

Likelihood Estimator Function

A likelihood estimator for each vehicle's claims under the above model is derived in Chukova and Hirose [9] for the stochastic process described in the previous section. A summary of the important results are shown here. First consider the probability of there being no claim between $(T_{i,n}, U_{i,n})$ and $(T_{i,n+1}, U_{i,n+1})$, which is given by

$$P\{N_i(T_{i,n}, T_{i,n+1}, U_{i,n}, U_{i,n+1}) = 0\} = \exp\left(-\int_{T_{i,n}}^{T_{i,n+1}} \int_{U_{i,n}}^{U_{i,n+1}} \lambda(s,r)drds\right) \tag{2.13}$$

If vehicle i has n_i claims, then the likelihood of these claims for a given intensity function $h(t,u)$ has the form

$$L_i = \prod_{j=1}^{n_i} h(T_{i,j}, U_{i,j}) \prod_{j=0}^{n_i} \exp\left(-\int_{T_{i,j}}^{T_{i,j+1}} \int_{U_{i,j}}^{U_{i,j+1}} \lambda(s,r)drds\right) \tag{2.14}$$

where $T_{i,n_i+1} = min(A_i, T_{max})$, $U_{i,n_i+1} = U_{max}$, and $T_{i,0} = U_{i,0} = 0$. As each vehicle's failures are independent from each other we can consider the product of individual likelihoods to get the likelihood of multiple vehicles under the model.

$$L = \prod_{i=1}^{M} L_i \tag{2.15}$$

In this chapter we have introduced the building blocks for the model used in this thesis. First we introduced the concept of a mean cumulative estimator, then looked at a one dimensional parametric model for estimating MCF, and finally considered a two dimensional model. This two dimensional model is a stochastic process derived from a NHPP, in addition a likelihood estimator is derived to fit model parameters for given set of observations (claims). In the next chapter we will consider a real warranty dataset and prepare it for analysis. The choice of hazard function will be discussed in Chapter 4.

Chapter 3

The Dataset

In this chapter, we detail the warranty dataset used in this thesis. It is cleaned, analysed and split up into time cuts. The distributions of claim ages and mileages are investigated, and marginal distributions for both are fitted. These marginals will be used in Chapter 5 to model the joint distribution of the number of claims per vehicle. Finally the empirical distribution of MARs for vehicles with claims and the empirical distribution of censoring ages are considered.

3.1 Introduction

In this thesis, real automotive warranty claim data is used to investigate the effectiveness of the 2 dimensional predictive model shown in Chapter 2 (pages 33-37). The goal of this chapter is to investigate the lifetimes of the vehicles as functions of age and mileage seperately (the marginal distributions). The dataset analysed comes from an unnamed car company for one specific model-year of car. This is the same dataset as used in [3], [8], [10], and [13]. As this is a real database, cleaning is required to remove entries which are incorrect, incomplete or the result of "extreme use".

The dataset consists of records of cars with their sale date, vehicle id and numerous other details such as manufacture date, and state of sale.

The dataset also includes records of each claim, including the vehicle the claim was made against, the cost of the repair work, the time the claim was made and the mileage of the vehicle at the time of the claim. Mileage is assumed to be zero at sale date. The mileage data is the only information we have on the driving behaviour of the cars. An analysis of successive claims on a vehicle shows that the mileage accumulation over time is piecewise linear, in other words driving patterns change slightly between claims. In the dataset, 48% of cars have claims, and of those cars, 64% leave the warranty region due to mileage accumulation.

3.2 Data Cleaning

The dataset is cleaned to remove any erroneous cars or extreme claims. This cleaning is identical to most previous works on this data with the following exception, which is also made in [3]. Claims made on the same day are combined into one claim, with a cost totalling the sum of the separate claims. Multiple claims on the same day are likely due to the way the manufacturer/retailer chose to itemise a single claim, and not the result of multiple failures on one day. This cleaning process is performed in the following order:

1. Deleting vehicle records with no sale date.
   As these vehicles have no recorded sale date, they can not be used in modelling since vehicle age cannot be determined.

2. Deleting claim records made before sale date
   These claims are likely made by the manufacturer/retailer and therefore the mechanism for generating these failures is not the same as the bulk of claims in this dataset. The justification of this exclusion is partly due to expert opinion and that the proportion of these claims is relatively small. Furthermore if we did not remove these claims,

we would introduce left-censoring to our data, as we assume all vehicles are sold with a mileage of zero. This cleaning step also agrees with previous papers investigating this dataset.

3. Deleting claim records outside warranty cut-off times
   These claims occur outside the warranty agreement (36,000 miles and 3 years) and therefore will not be included.

4. Deleting claim records with decreasing mileage
   These claims are the result of either input error or possibly even fraud, and are not included.

5. Deleting claim records with extreme usage
   Vehicles with usage claims at greater than 1000 or less than 3 miles per day are excluded. These removals are based on expert opinion as in [13].

Once this cleaning is finalised, each vehicle and its claims are given an age A_i, calculated as time in days since sale date. This allows comparison between vehicles with different sale dates. These claims are plotted in the warranty region 36,000 miles and 3 years as in Figure 3.1. This gives an overall idea of the distribution of claims in the warranty period.

Note the general asymmetric trend of the trajectories of vehicles. It is possible to see that the majority leave the warranty area due to mileage accumulation rather than age. Also note the "infant mortality period" (the high failure rate for young vehicles), in the form of far more failures for smaller ages and mileages. After that period, the failure rate plateaus out, and finally right at the end of the warranty cut-off we see a possible slight increase of claims, the "customer rush" period, in which claims are made to get in before the warranty expires. This picture is somewhat unclear however, so to visualise these claims we will use 2d-bining. 2d-binning is a 2d analogue of histograms. The size of each point (or hexagon) gives the number of claims in that area. In Figure 3.2, single claims are omitted for clarity. Note the absence of the "customer rush" period.

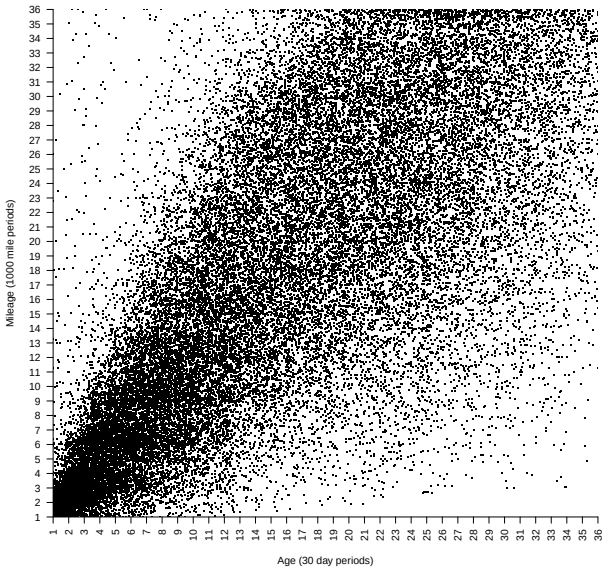

Figure 3.1: All claims after cleaning

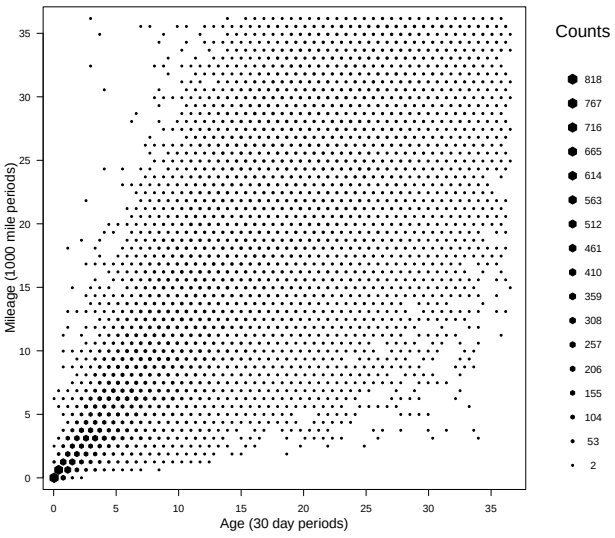

Figure 3.2: 2d-bin graph of data

3.3 Time Cuts

To emulate the accumulation of the data collection process, the dataset is divided to mimic what data points would be available at different calendar times. For a given time cut, say 12 months after first vehicle sale, all vehicles which were sold after the time cut and claims that were incurred after the time cut are removed from the database. The dataset is split into 3 time cuts, 12, 24, and 36 months, and are shown in Table 3.1. These datasets will be used in later chapters, allowing us to predict known numbers of claims (at 36 months) using only 12 or 24 months of claims.

Time Cut (months)	24/10/2001 12 months	24/10/2002 24 months	24/10/2003 36 months
Number of Vehicles	37709	44848	44890
with claims	5855	15661	21736
without claims	31854	29187	23154
Number of Claims	7655	26190	43520

Table 3.1: Table of database time cuts

3.4 Preliminary Analysis

To model the overall two-dimensional joint distribution, we must first understand the behaviour of the marginals. For the purposes of this thesis we are only interested in two dependent random variables *Vehicle Age At Failure* (T) and *Vehicle Mileage At Failure* (U). How are T and U distributed (say with CDFs $F_T(t)$ and $F_U(u)$)? Are they dependent or independent? To compound this problem, both the age and mileage data are censored. The age data is censored as all vehicles are not the same age, and therefore information on their failures is incomplete. Secondly mileage information is only ever available at the time of failure. This implies that for vehicles without failures, we only know that their mileage is greater than or equal

to zero miles. In order to approximate the marginal distributions of number of claims parametrically, we first need to consider the population of vehicles able to make a claim.

The simplest approach is to approximate the survival function using the Kaplan-Meier estimator [19]. Therefore the vehicle population is split into groups. The population able to make a first claim is trivial, it is all the vehicles in the database still within the warranty period. It follows that vehicles able to make a second claim are restricted to the vehicles which have a first claim and so on. It is in this way that the vehicles are grouped, so that distributions of claim at age or mileage is not affected by the number of claims a vehicle has had prior.

3.4.1 Kaplan-Meier (KM) estimator

Fitting the distribution for T

Firstly the KM estimator of the distribution of claims as a function of age t is generated. This is shown in Figure 3.3, note that approximately half of the vehicles do not fail before the 3 years is up, this is consistent with the data in table 3.1. The KM estimator is used to build a censored histogram by using the estimated survival function to calculate densities as described in [16]. This histogram (Figure 3.4) does not immediately suggest a potential parametric distribution. Therefore, the next step is to consider the first, second, third, and fourth or greater claims separately as in Figure 3.5. Note the first claim distribution shows an obvious "infant mortality" period before stabilising. The distribution of second, and third claims appear similar to each other. Using maximum likelihood estimation we fit some common distributions to these groups of claims and compare fits using AIC.

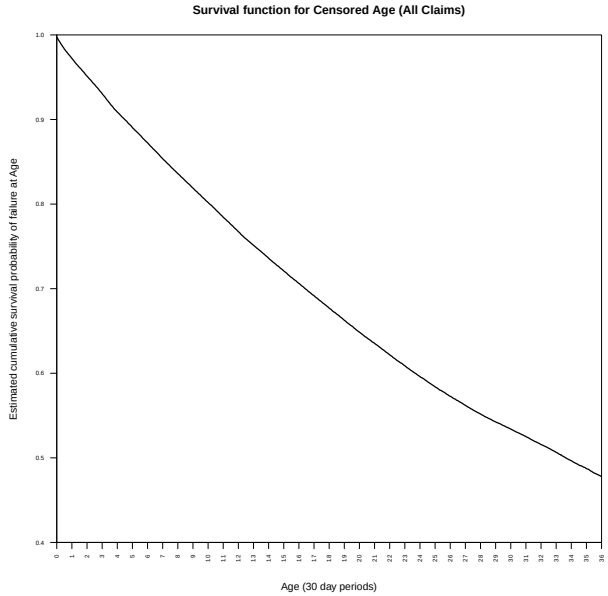

Figure 3.3: KM-estimator of vehicle survival in terms of Age

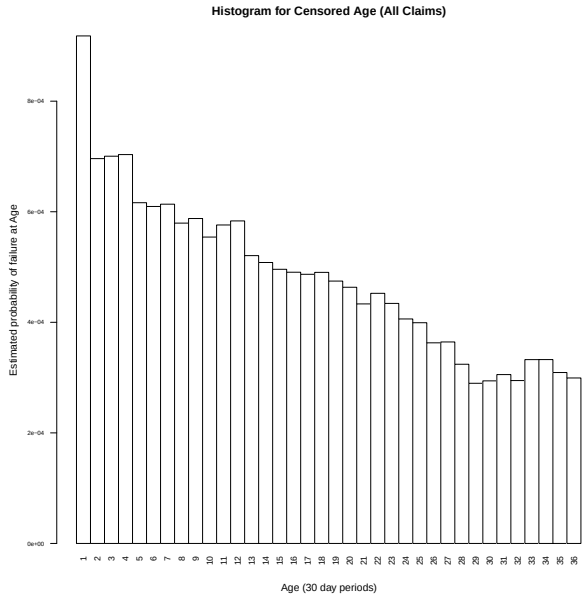

Figure 3.4: Censored Histogram of all vehicle failures

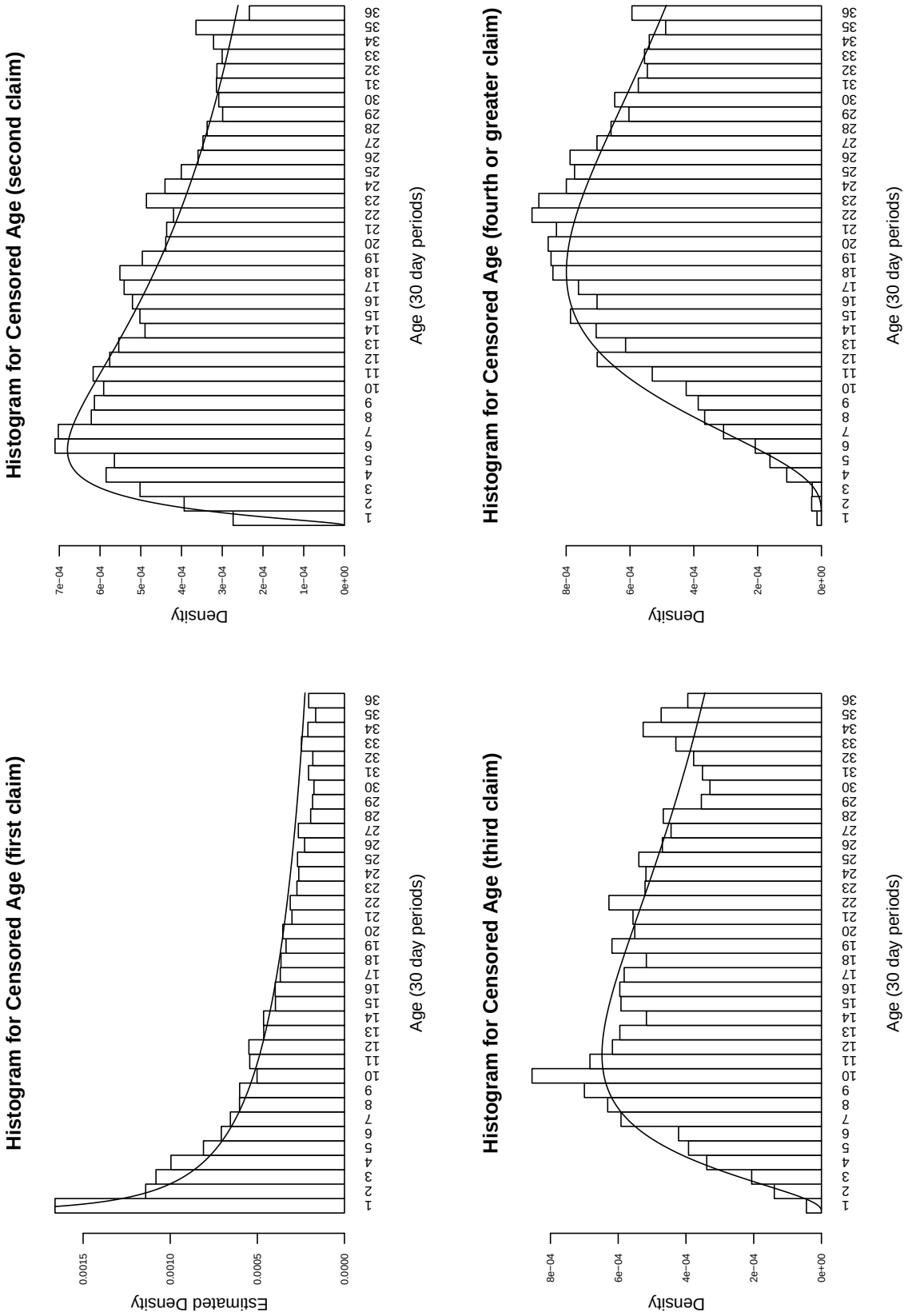

Figure 3.5: First, second, third, and fourth or greater claims

By investigating the distribution of these in isolation, we find that the first claim is likely Weibull, and the rest Lognormal. We consider a proportional mixture of the four distributions, as described in Chapter 2. Specifically the sum of one Weibull distribution with proportion π_1 and three Lognormal distributions with proportion π_2, π_3, and π_4 respectively. These π_i represent the proportion of cars who have i claims.

$$\pi_1 \cdot \left(1 - e^{-\left(\frac{t}{\theta_1}\right)^{\beta_1}}\right) + \pi_2 \cdot \mathcal{N}(t, \mu_1, \sigma_1) + \pi_3 \cdot \mathcal{N}(t, \mu_2, \sigma_2) + \pi_4 \cdot \mathcal{N}(t, \mu_3, \sigma_3) \quad (3.1)$$

where $\mathcal{N}(t, \mu, \sigma)$ is the Lognormal CDF. Plotting the density of this CDF on the censored histogram gives the following fit as seen in Figure 3.6.

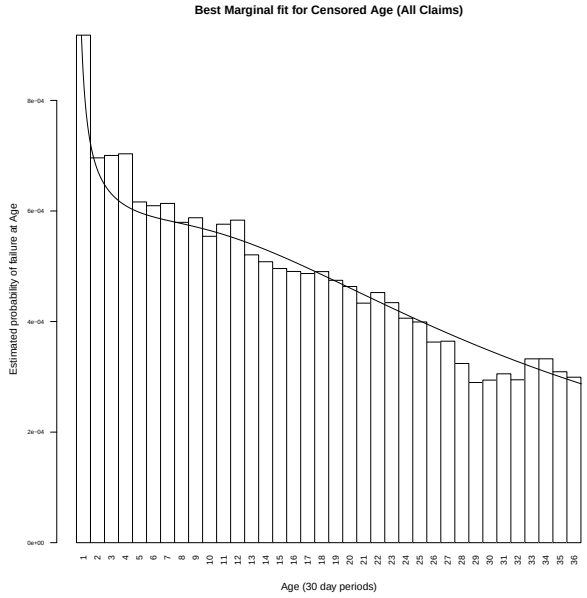

Figure 3.6: Fitted Histogram of all vehicle failures

Fitting the distribution for U

As with the age distribution we start by constructing the KM estimator of the distribution of number of claims as a function of mileage u (see Figure 3.7). Similarly, a censored histogram is constructed to illustrate the density

of the failures in terms of mileage (Figure 3.8). Again the histogram of all claims does not immediately suggest a potential parametric distribution. To further investigate the distribution, the claims are grouped into first claims and second or greater claims. These are fitted by MLE as before, and the two best fits were found to be both Weibull (see Figure 3.9). A mixture distribution is considered:

$$\pi_1 \cdot \left(1 - e^{-\left(\frac{u}{\theta_2}\right)^{\beta_2}}\right) + (\pi_2 + \pi_3 + \pi_4) \cdot \left(1 - e^{-\left(\frac{u}{\theta_3}\right)^{\beta_3}}\right) \tag{3.2}$$

Superimposing the density function of this CDF onto the histogram, shows that the mileage fit is not as precise as the age fit as it deviates far more (Figure 3.10). Furthermore, second or greater claims were so similar in distribution that there was no need to split them into further groupings.

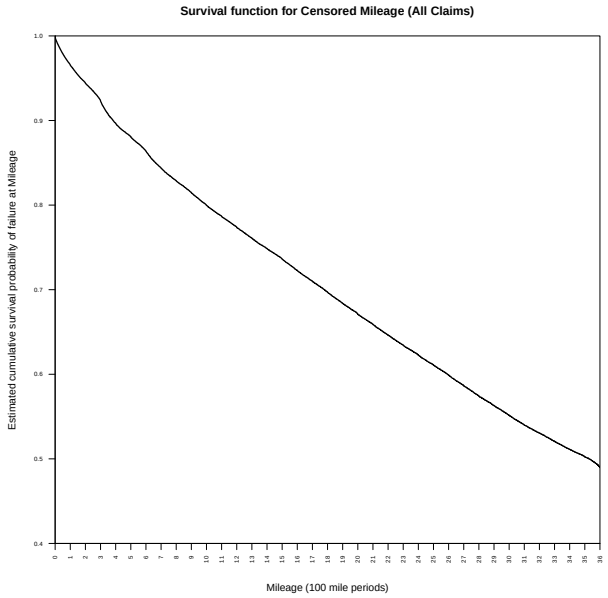

Figure 3.7: KM-estimator of Vehicle in terms of Mileage

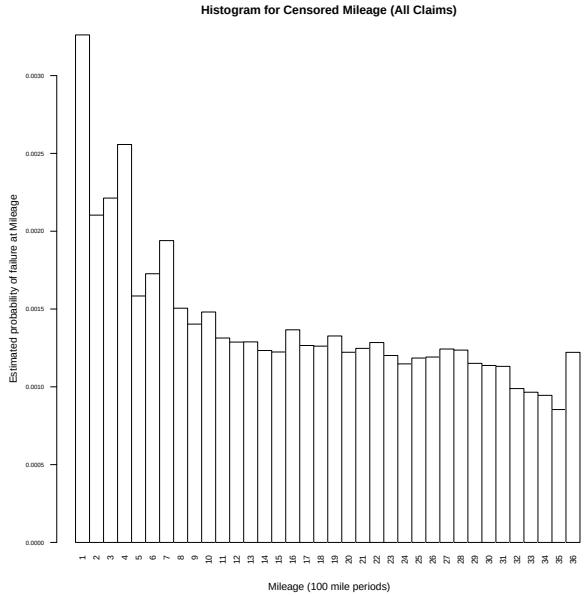

Figure 3.8: Censored Histogram of all vehicle failures

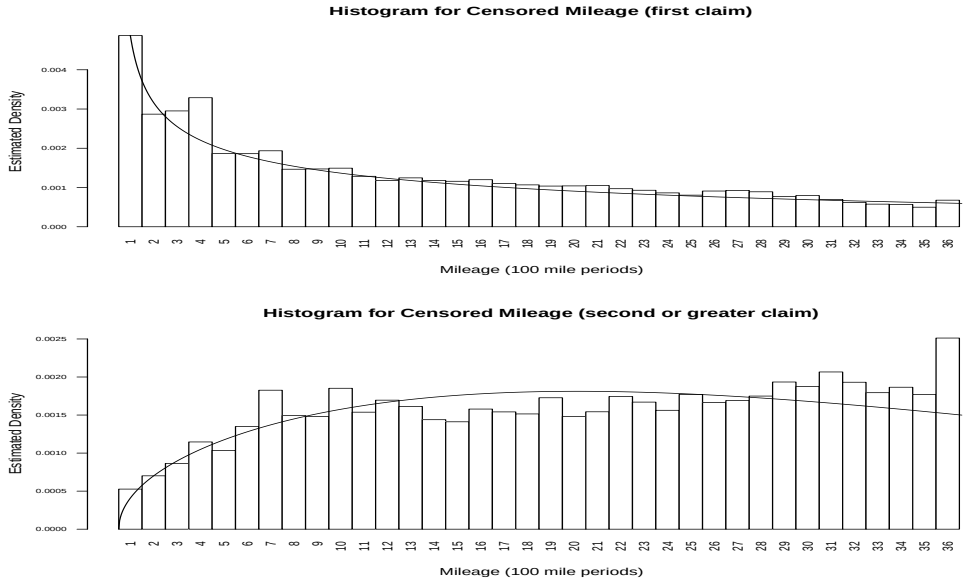

Figure 3.9: First and second or greater claims

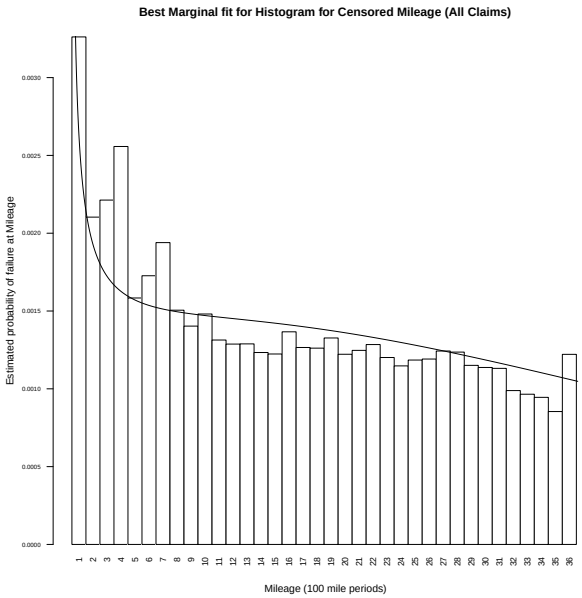

Figure 3.10: Fitted Histogram of all vehicle failures

However this approach is unfortunately flawed, as the true population of vehicles eligible to make a claim is not so straightforward. However it does provide insight into how the distribution of failures changes between first and second claims (and so on) and as such is included here. As this survival function doesn't accurately reflect the population of cars still eligible to make a claim (i.e. $M(t)$), we must consider vehicles leaving the warranty region due to mileage or age. To correctly estimate $M(t)$, we look to the Strata model discussed in Chapter 2. This model makes the assumption of piecewise linearity of mileage accumulation, and has been used on this dataset before as in [10].

3.5 Marginal Distributions for Number Of Claims

In this section we will fit marginal distributions for use in the 2D model defined in Chapter 2. This strata model produces a MCF ($\hat{\Lambda}_X(x)$) which

corrects for vehicles becoming ineligible to make claims due to leaving the warranty region due to age or mileage. Using formula 2.1, we know that the survival function is related to the cumulative hazard function in the following way:

$$F_X(x) = 1 - S_X(x) = 1 - e^{-H(x)}$$

Under the strata model (see page 28), $H(x) = \hat{\Lambda}(x)$ and therefore by fitting a hazard function from a parametric distribution to the strata model MCF we get a parametric distribution for T and U. For both time-as-age and time-as-mileage cases, the MCF estimators are produced using the Strata model, $\hat{\Lambda}_T(t)$ and $\hat{\Lambda}_U(u)$ respectively. Then common distributions are fitted using regression to both $\hat{\Lambda}_T(t)$ and $\hat{\Lambda}_U(u)$ and the best are selected. The best fit for age was found to be a Weibull with $\beta_1 = 0.903$ and $\theta_1 = 864.896$, with $\hat{\Lambda}_T(t) = \left(\frac{t}{\theta_1}\right)^{\beta_1}$. The quality of this fit is demonstrated in Figure 3.11. Mileage was similarly fit with a Weibull with $\beta_2 = 0.911$ and $\theta_2 = 314.295$, with the form $\hat{\Lambda}_U(u) = \left(\frac{u}{\theta_2}\right)^{\beta_2}$. This is shown in Figure 3.12.

Therefore for the marginal distributions of T and U, we choose $F_T(t)$ and $F_U(u)$ to be:

$$F_T(t) = 1 - e^{-\left(\frac{t}{\theta_1}\right)^{\beta_1}} \tag{3.3}$$

and,

$$F_U(u) = 1 - e^{-\left(\frac{u}{\theta_2}\right)^{\beta_2}} \tag{3.4}$$

respectively. Next we briefly the consider the distribution of MARs through the warranty coverage region.

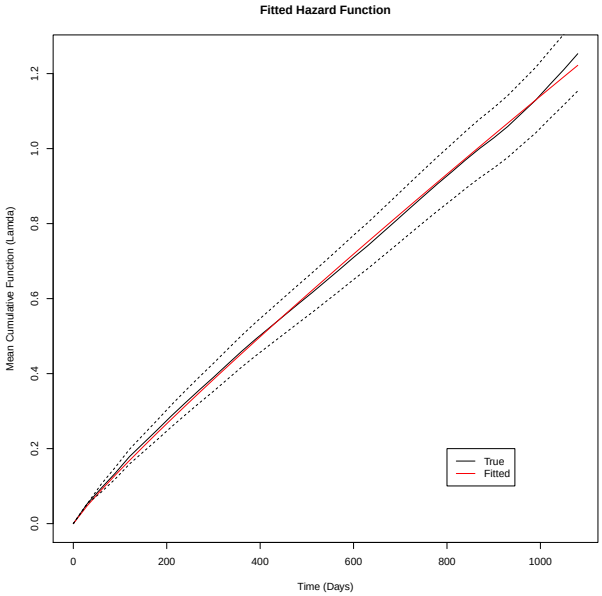

Figure 3.11: Fitted MCF to Strata MCF for Age

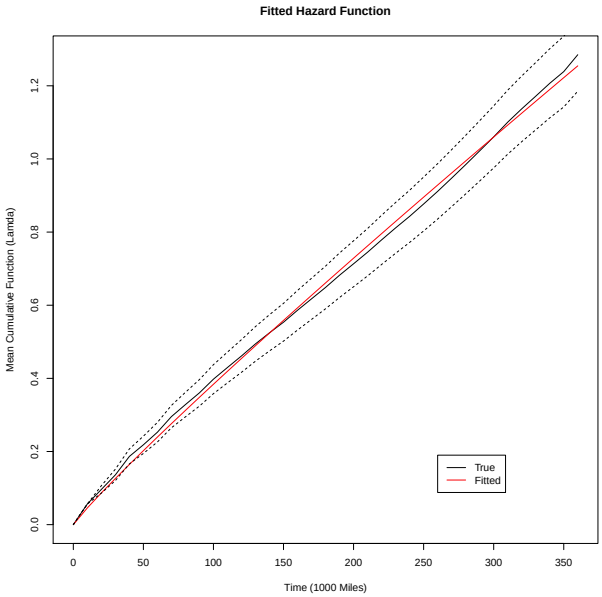

Figure 3.12: Fitted MCF to Strata MCF for Mileage

3.6 Distribution of Mileage Accumulation Rates

The Figure 2.9 shows that a vehicle can take many trajectories through the warranty coverage region. To understand which are the most likely trajectories we consider the distribution of all MARs ($F_r(x)$) for vehicles with claims. The empirical distribution is shown in Figure 3.13. A vehicle which leaves the warranty region due to age must accumulate mileage less than approximately 34 miles per day, anything greater than that and the vehicle will exceed the mileage cutoff before the age cutoff. Note the peak around 36 miles per day, and that the distribution is right skewed (more weight for larger MARs). This is consistent with the majority of vehicles leaving the warranty region due to mileage accumulation. Next we consider the distribution of MAR for first, second and third claims, $F_{r_n}(x)$ (see 2.3).

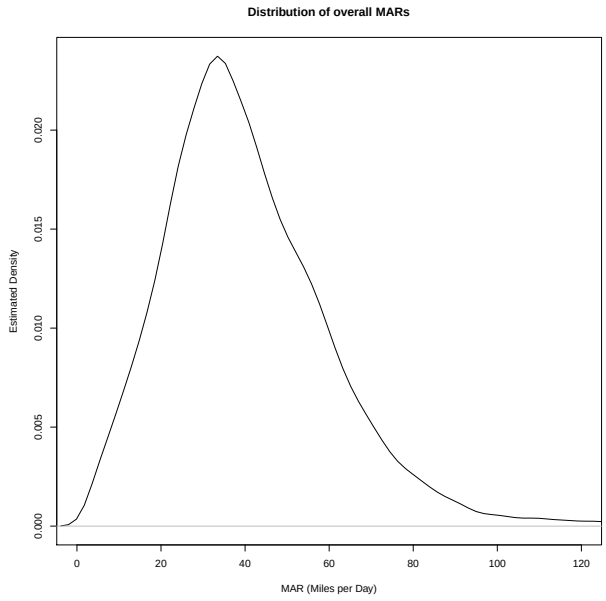

Figure 3.13: MAR distribution for all claims

By comparing the overall MAR with the $F_{r_n}(x)$ for ($n = 1, 2, 3$) as in Figure 3.14, we can see that these distributions are similar in the tails and

also skewed in the same way as the overall MAR. This implies that the trajectories of vehicles are only slightly affected by claims. However this slight difference in distribution for MAR will be considered in the simulation chapter of this thesis. These empirical distributions will be used in the simulation algorithm discussed in Chapter 6.

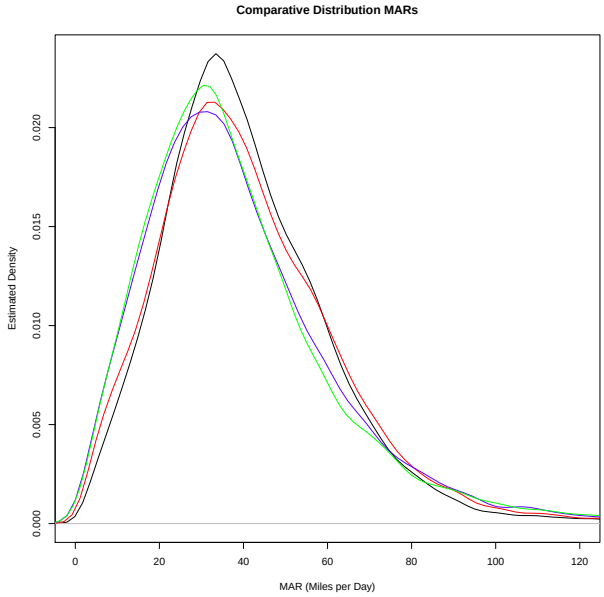

Figure 3.14: MAR distribution comparison

3.7 Distribution of Censoring Ages

Vehicles in the dataset are not all the same age at a time cut, this is because vehicles are sold throughout the warranty period. The distribution of ages (we denote it here by $F_A(x)$ in the dataset is the distribution of the censoring age. As most of the sale process has occurred before 24 months the distribution of ages is relatively stable between 24 and 36 months.

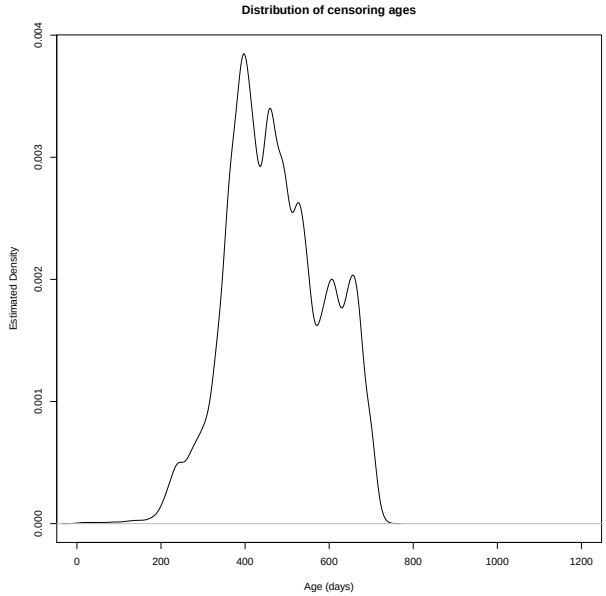

Figure 3.15: Censoring Age for 24 months

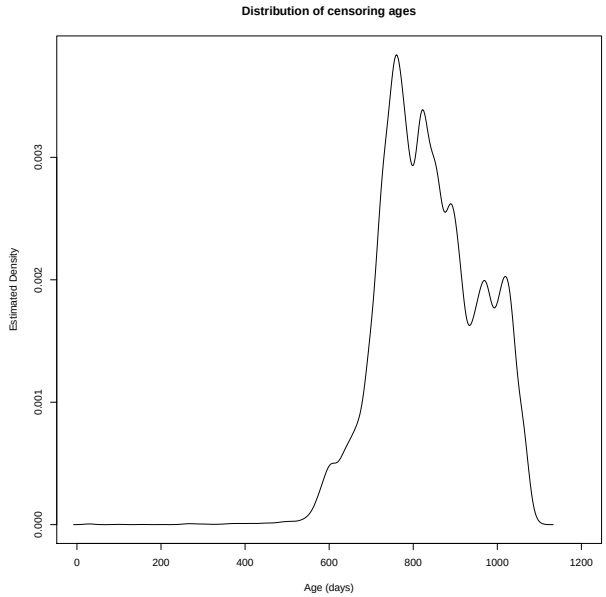

Figure 3.16: Censoring Age for 36 months

Consider Figure 3.15 and Figure 3.16, these distributions are similar as only a negligible number of vehicles are sold in that period (42). A simple time shift (one year in this case) of the 24 month empirical distribution, shows the fit between the two as in Figure 3.17.

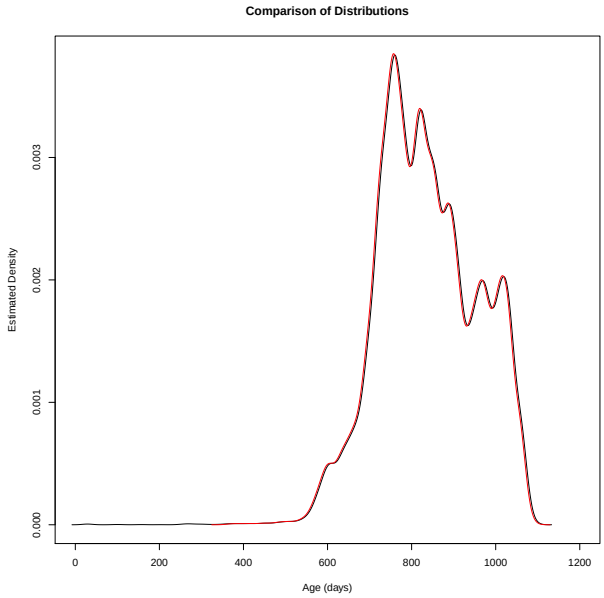

Figure 3.17: Comparison of censoring ages

This chapter has detailed the characteristics that the dataset possesses. We have investigated the distribution of MAR and censoring age, which will be needed in the simulation chapter. Now that the marginals have been decided upon, we consider the use of copulas to "join" these marginals in the next chapter.

Chapter 4

Copulas

In this chapter, copulas are introduced and some examples are presented. These copulas are selected to model the joint hazard function required by the 2D warranty model described in Chapter 2. Finally, the hazard function for each selected Copula is derived.

4.1 What are Copulas?

For multivariate datasets, finding the underlying joint distribution can be analytically difficult and computationally expensive. Informally, copulas are functions that relate joint multivariate distribution functions to their marginal distribution functions. Copulas are a powerful tool, as they allow us to investigate the marginal distribution behaviour in isolation. Marginal distributions are often trivial to fit, and a wealth of distribution-independent univariate goodness-of-fit tests exist such as χ^2 and Kolmogorov-Smirnov tests. In this thesis we will only consider the bivariate copula, for a more complete definition of multivariate copulas and their derivation see [24]. Copulas have found resurging popularity due to the increase in computation power easily available, and their use in modelling financial processes, by brokers and other financial firms.

4.2 Definition

A two dimensional copula $C(x, y)$ is defined as a function with the following properties.

1. The domain of C is $[0, 1]^2$.
2. The range of C is $[0, 1]$.
3. $\forall x, y \, \epsilon[0, 1] : C(x, 0) = C(0, y) = 0$.
4. $\forall x, y \, \epsilon[0, 1] : C(x, 1) = x$ and $C(1, y) = y$.
5. $\forall x_1, x_2, y_1, y_2 \, \epsilon[0, 1]$ such that $x_1 \leq x_2$ and $y_1 \leq y_2$, $C(x_2, y_2) - C(x_2, y_1) - C(x_1, y_2) + C(x_1, y_1) \geq 0$ holds.

These properties are true for continuous bi-variate distributions and therefore must also hold for copulas if they are to be distribution functions. Copula functions are often derived with certain correlation characteristics in mind. For example, radial symmetry in the Gaussian copula forces equal dependence for left and right tails. In the case of the Clayton copula, capturing very strong dependence in the left tail or the Farlie-Gumbel-Morgenstern copula which is only capable of weak dependence modelling.

4.2.1 Sklar's Theorem

Sklar's theorem is an important theorem in relation to the use of Copulas in applied statistics, as it connects copulas to cumulative probability distributions. It is defined as follows: Let $F(t, u)$ be the joint distribution function with margins $F_T(t)$ and $F_U(u)$. Then there exists a copula $C(x, y)$ such that for all (t, u) in the extended real plane $\overline{\mathbf{R}} \times \overline{\mathbf{R}}$.

$$F(t, u) = C(F_T(t), F_U(u)) \tag{4.1}$$

The implication is that one can use marginal distributions and a copula to define the joint distribution. In particular Sklar's theorem states that if $F_T(t)$ and $F_U(u)$ are continuous distributions then the copula C is uniquely defined. An immediate consequence of the above theorem is:

$$C(x, y) = F(F_T^{-1}(x), F_U^{-1}(y)) \tag{4.2}$$

where x and y are in $[0, 1]$. That is to say a joint distribution of inverse marginals is itself a copula, as it meets all of the criteria stated above. Proof of this theorem is found in A. Sklar's 1959 paper [27].

4.3 Properties

A copula

- Is marginally symmetric if $C(x, y) = C(y, x)$ (symmetric across the main diagonal of the unit square)
- Is radially symmetric if $C(x, y) = x+y-1+C(1-x, 1-y)$ (symmetric in the tails)
- Is associative if $C(x, C(y, z)) = C(C(x, y), z)$, meaning multivariate copulas can be constructed from bivariate copulas

4.3.1 Fréchet-Hoeffding bounds inequality

Fréchet-Hoeffding bounds inequality states that for a distribution function $F(t, u)$ with marginals $F_T(t)$ and $F_U(u)$ the following inequality holds

$$max(F_T(t) + F_U(u) - 1, 0) \leq F(t, u) \leq min(F_T(t), F_U(u))$$

As we have defined in equation 4.1, this can be written in terms of a copula C

$$max(x + y - 1, 0) \leq C(x, y) \leq min(x, y) \tag{4.3}$$

These upper and lower bounds are also Copulas and suggest a natural ordering of Copulas. In this thesis we will not be using this theorem, it is included only as an important property.

4.4 Copula types

In this thesis, nine copulas are considered, they are included in this section. The most commonly used copulas come in two types, Archimedian and Elliptical. These two types are distinguished by the means of their construction, and the type of correlation they aim to capture. For each copula, a short description is provided, with its origin, definition and any parameters used to define it, as well as any notable applications. Furthermore to illustrate the dependence structure of the copula, 4000 realisations of each copula are generated and plotted using the R library *Copula*.

4.4.1 Archimedian Copulas

Archimedian copulas have the form

$$C(x, y) = \psi(\psi^{-1}(x) + \psi^{-1}(y)),$$

where $\psi(t)$ and $\psi^{-1}(t)$ are defined as the *generator function* and *inverse generator function* respectively. The class of Archimedian copulas is associative which means that extending these copulas to d dimensions is straightforward. These copulas are popular and widely used due to the few parameters needed to define them (often only one) and most have an elegant, closed, and explicit form. All of the following copulas are marginally symmetric, and unless otherwise stated they only capture positive dependence. Due to this fact, refererences to the left and right tails of a copula, are refering to their behaviour for small values and large values of both x and y.

Product Copula

The product (or independence) copula is the simplest Archimedian copula. As its name suggests it has no dependence between the marginals, and is therefore just the product of the marginal distributions. It has the form:

$$C^{Product}(x, y) = xy \tag{4.4}$$

Its generator function is $\psi(t) = e^{-t}$, and inverse generator function $\psi^{-1}(t) = log(t)$.

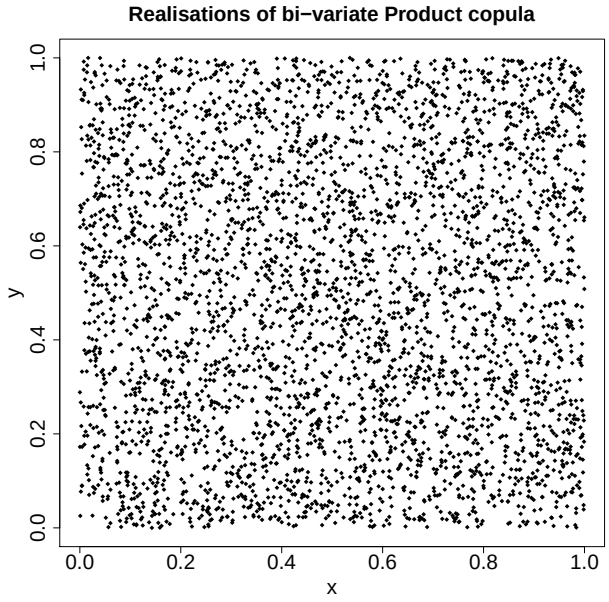

Figure 4.1: Realisations of Product copula

Note the uniform distribution of realisations on the unit square in Figure 4.1. As this copula captures no dependence, in this thesis it will be the "baseline" for comparing the other copulas.

Ali-Mikhail-Haq Copula

The Ali-Mikhail-Haq (AMH) copula can only capture a weak dependence between marginals in the left tail. It was created by Ali, Mikhail, and Haq in the 1978 paper "A class of bivariate distributions including the bivariate logistic" [2]. It is defined by $\psi(t) = \frac{1-\delta}{e^t-\delta}$ and $\psi^{-1}(t) = \log\left(\frac{1-\delta+\delta t}{t}\right)$, and has the form

$$C^{AMH}(x,y) = \frac{xy}{1-\delta\ (1-x)\,(1-y)}, \tag{4.5}$$

where $\delta \in [0,1]$. This parameter determines the strength of dependence, increasing δ increases the dependence between x and y. When $\delta = 0$ the copula reduces to the product copula.

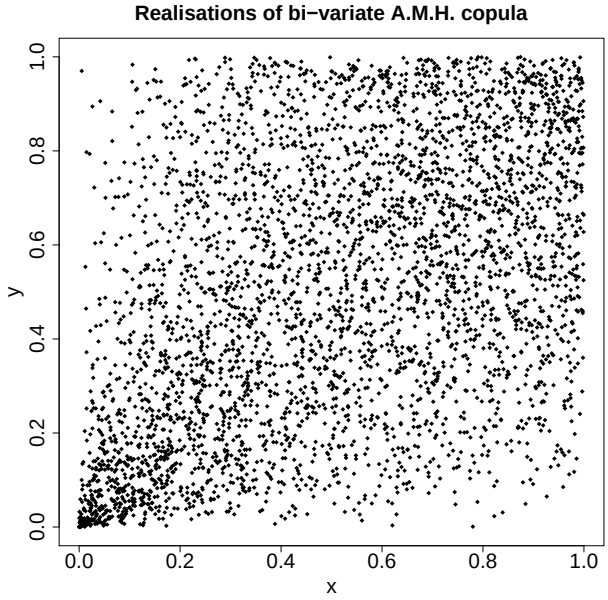

Figure 4.2: Realisations of AMH copula ($\delta = 1$)

It is possible to see the slight dependence for small values of x and y (realisations funnelling outwards) in Figure 4.2 with $\delta = 1$.

Clayton Copula

The Clayton copula was first suggested by D. G. Clayton in 1978 in the journal *Biometrika* [11], as a way to model patient survival times for diseases passed down through families. It is characterised by a very strong correlation in the left tail and decreasing correlation as values of x and y get larger. It cannot capture any negative correlation, and is not radially symmetric. The Clayton copula has been used in epidemiology, risk analysis, and econometrics. It is defined by $\psi(t) = (1 + \delta t)^{-1/\delta}$ and $\psi^{-1}(t) = \frac{1}{\delta}\left(t^{-\delta} - 1\right)$, giving it the form:

$$C^{Clayton}(x, y) = \left(x^{1-\delta} + y^{1-\delta} - 1\right)^{\frac{1}{(1-\delta)}}, \tag{4.6}$$

where $\delta \in (0, 1) \cup (1, \infty)$. As δ increases, so does the correlation between x and y.

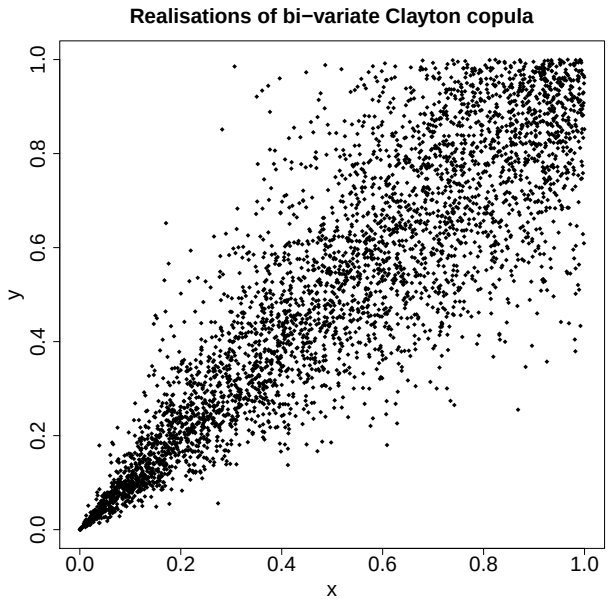

Figure 4.3: Realisations of Clayton copula ($\delta = 5$)

In Figure 4.3, it is possible to see the strong dependence, this is illustrated by the tighter funnelling of the realisations ($\delta = 5$) and how it

changes as x and y increase.

Frank Copula

The Frank copula, developed by M. J. Frank in the article "On the simultaneous associativity of F(x, y) and x + y - F(x, y)", is a marginally and radially symmetric copula (for $\delta > 1$). It has its strongest dependence in the middle of the distribution, and weak dependence in the tails. It has been used in the modelling of price changes in stocks and bonds, due to this behaviour. It is defined by the generator and inverse pair $\psi(t) = -\frac{1}{\delta}\log(1-(1-e^{-\delta})e^{-t})$ and $\psi^{-1}(t) = -\log\left(\frac{e^{-\delta t}-1}{e^{-\delta}-1}\right)$.

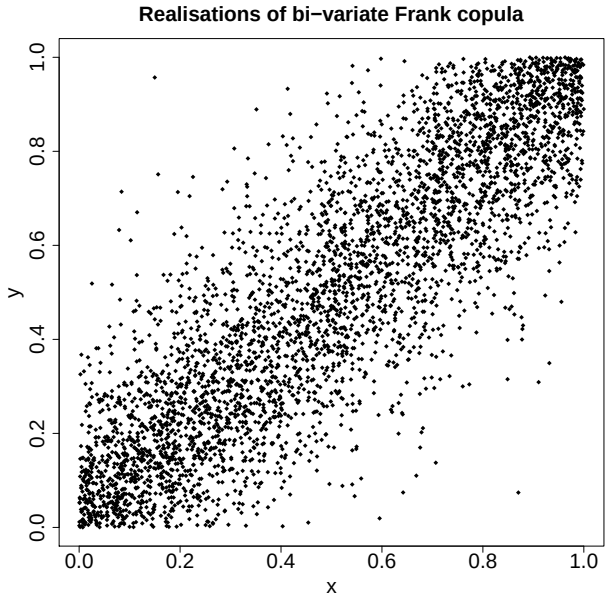

Figure 4.4: Realisations of Frank copula ($\delta = 10$)

The copula therefore has the form:

$$C^{Frank}(x,y) = -\frac{1}{\delta}\log\left(1+\frac{\left(e^{-\delta x}-1\right)\left(e^{-\delta y}-1\right)}{\left(e^{-\delta}-1\right)}\right), \tag{4.7}$$

where $\delta \in (-\infty, \infty)$. For $\delta < 1$ there is negative dependence, and when $\delta = 1$ the copula gives independence. For all $\delta > 1$, there is a strong positive dependence in the middle of the distribution, which increases as δ increases. Notice the strong dependence for the middle range of x and y (most realisations are on the diagonal) in Figure 4.4 with $\delta = 10$.

Joe Copula

The Joe copula was invented by Harry Joe in 1993 with the paper "Parametric Families of Multivariate Distributions with Given Margins " [17]. The copula captures weak dependence in the left tail and strong dependence in the right tail.

$$C^{Joe}(x, y) = 1 - \left((1-x)^{\delta} + (1-y)^{\delta} - (1-x)^{\delta}(1-y)^{\delta}\right)^{\frac{1}{\delta}}, \tag{4.8}$$

where $\delta \in [1, \infty)$. The dependence for smaller x and y increases as δ increases. Its generator and inverse generator functions are $\psi(t) = 1 - (1 - e^{-t})^{1/\delta}$ and $\psi^{-1}(t) = -\log\left(1 - (1-t)^{\delta}\right)$ respectively. Figure 4.5 ($\delta = 4$) emphasises the independence for the left tail and strong dependence in the right tail.

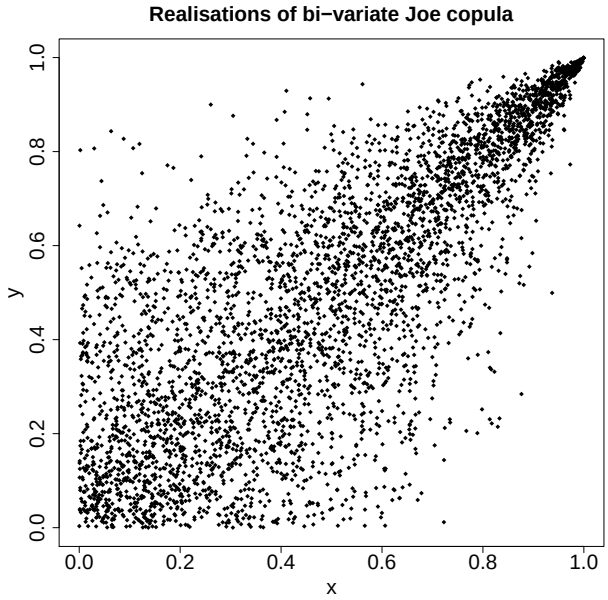

Figure 4.5: Realisations of Joe copula ($\delta = 4$)

Gumbel Copula

The Gumbel copula comes from a 1960 paper written by E. J. Gumbel entitled "Distributions des valeurs extrêmes en plusieurs dimensions" [14]. It has a weak correlation in the left tail and a much stronger correlation in the right tail. It is defined by $\psi(t) = e^{-t^\delta}$ and $\psi^{-1}(t) = (-\log(t))^{1/\delta}$, giving its form as:

$$C^{Gumbel}(x, y) = e^{-\left((-\log(\mathrm{x}))^{\frac{1}{\delta}} + (-\log(\mathrm{y}))^{\frac{1}{\delta}}\right)^{\delta}}, \tag{4.9}$$

where $\delta \in (0, 1)$. As δ tends to 1, the copula approaches independence. As δ decreases, the dependence between x and y increases.

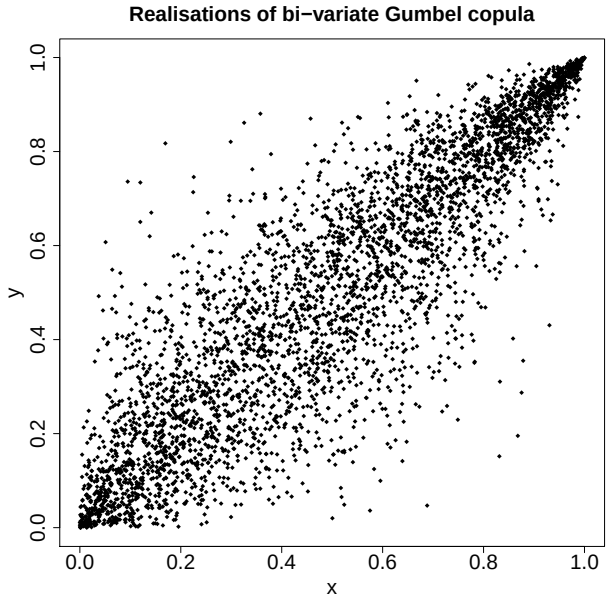

Figure 4.6: Realisations of Gumbel copula ($\delta = 0.25$)

Figure 4.6 ($\delta = 0.25$) shows the strong correlation in the right tail. Note the almost complementary behaviour to the Clayton copula.

4.4.2 Elliptical Copulas

The elliptical copulas are named after the elliptical class of functions they are constructed from. As in equation 4.2, a copula can be defined in terms of inverse marginal distribution functions and a multivariate distribution function. In this thesis we consider two elliptical copulas, the Gaussian copula and the Student-t copula.

Gaussian Copula

The Gaussian copula is arguably the most famous, and until the 2008 Global Financial Crisis, perhaps the most commercially used copula (it has since fallen out of favour). David X. Li popularised its use in financial markets with his paper "On Default Correlation: A Copula Function

Approach" [21] and subsequently it was put into work evaluating risks in bond markets around the world. In the wake the financial crisis, it was seen as one of the many causes of market failure. The trust put in the correlations it could capture and the unpredictability of the financial markets, lead to this situation. This copula is radially symmetric if $\rho > 0$, and allows for both positive and negative correlation. It is constructed from a bivariate standard normal distribution Φ_ρ, and an inverse standard normal distribution Φ^{-1}, and has the following form

$$C^{Gauss}(x, y) = \Phi_\rho \left(\Phi^{-1}(x), \Phi^{-1}(y)\right) \tag{4.10}$$

where $\rho \in [-1, 1]$ is the correlation coefficient. Negative correlation occurs for $\rho < 0$ and positive correlation for $\rho > 0$. Figure 4.7 shows the positive dependence of x and y for $\rho = 0.75$, note that it is symmetric.

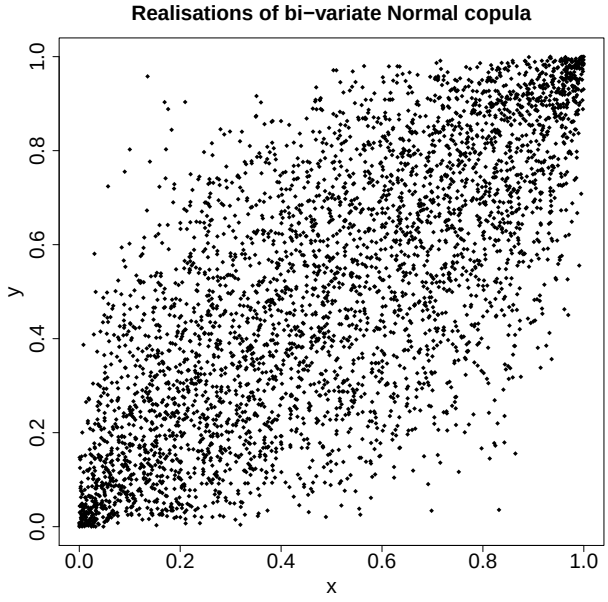

Figure 4.7: Realisations of Gaussian copula ($\rho = 0.75$)

Student-t Copula

The student-t copula is a two parameter copula, and it is also symmetric. Note the Student-t copula allows for less strong dependence at the middle of the distribution in comparison to the Gaussian copula. It has the form:

$$C^{Student}(x, y) = t_{\nu,\rho}(t_\nu^{-1}(x), t_\nu^{-1}(y)) \tag{4.11}$$

where $t_{\nu,\rho}$ is the multivariate student-t CDF, t_ν^{-1} is the inverse univariate Student-t distribution, $\nu \in \{1, 2, 3, \ldots\}$ is the degrees of freedom, and $\rho \in [-1, 1]$ is the correlation coefficient. Unlike the other copulas discussed in this chapter, the parameter ν only takes integer values ≥ 1. In the optimisation chapter of this thesis, it will have to be taken into account.

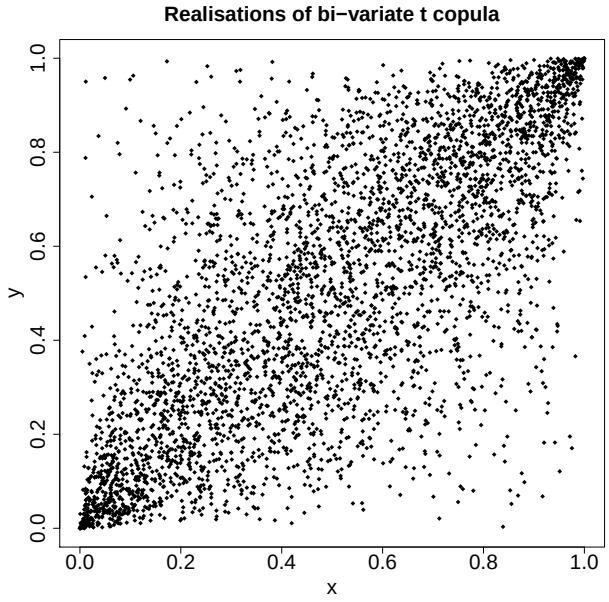

Figure 4.8: Realisations of Student-t copula ($\rho = 0.75$ and $\nu = 5$)

In Figure 4.8, it is clear that there is weaker dependence for middle range values of x and y. These realisations come from a bivariate Student-t copula with $\rho = 0.75$ and $\nu = 5$.

4.4.3 Other Copulas

Farlie-Gumbel-Morgenstern Copula

The Farlie-Gumbel-Morgenstern (FGM) copula comes from the 1956 paper by D. Morgenstern [23]. It only allows for very small correlations in the tails, and as such hasn't many applications. The copula is given by the equation:

$$C^{FGM}(x, y) = xy\left(1 + \delta\left(1 - x\right)\left(1 - y\right)\right) \tag{4.12}$$

where $\delta \in [0, 1]$. For $\delta = 0$ it collapses to the product copula. As δ approaches 1 the correlation increases, though it is still weak in comparison to other copulas. Figure 4.9 shows realisations of the copula with $\delta = 1$, note that even the strongest correlation attainable in this copula is minimal.

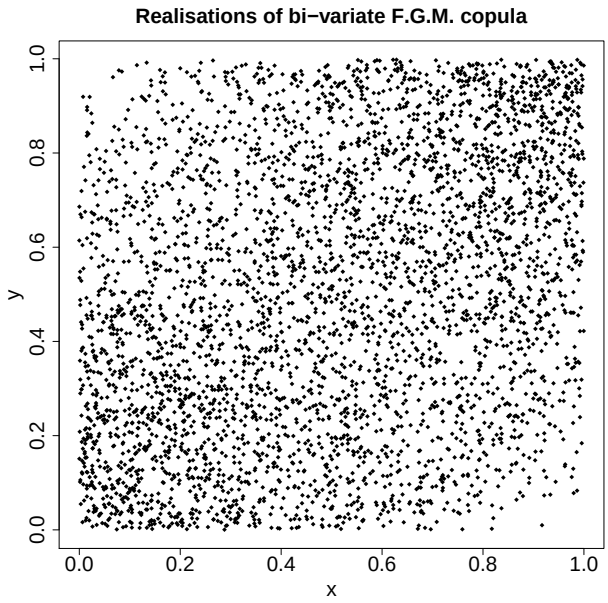

Figure 4.9: Realisations of FGM copula ($\delta = 1$)

4.5 Use in the 2d Warranty Cost Model

In lieu of a known joint distribution for the dataset, we use a copula in conjunction with the marginals estimated in Chapter 3. Which copula is best suited? The likelihood estimator (equation 2.14) given in Chapter 2 requires a joint hazard function. As a copula is just a distribution function, it is possible to derive a joint hazard function for it. In this way we can compare their relative likelihood.

4.5.1 Joint Survival Copula

The two dimensional joint survival copula $\overline{C}(x, y)$ is analogous to the joint survival function $\overline{F}(t, u)$. In two dimensions, the joint survival function is given by:

$$\overline{F}(t, u) = 1 - F_T(t) - F_U(u) + F(t, u) \tag{4.13}$$

Then by substituting in a copula C by Sklar's theorem 4.1 into equation 4.13 we arrive at

$$\overline{C}(F_T(t), F_U(u)) = 1 - F_T(t) - F_U(u) + C(F_T(t), F_U(u)) \tag{4.14}$$

Therefore the survival copula with out marginal distributions [24] is

$$\overline{C}(x, y) = 1 - x - y + C(x, y) \tag{4.15}$$

4.5.2 Joint Hazard Function

From here we derive the two dimensional hazard function $h(t, u)$ for a given copula C. First consider the definition of the bivariate hazard function

$$h(t, u) = \frac{f(t, u)}{\overline{F}(t, u)}, \tag{4.16}$$

where density $f(t,u)$ is $\frac{\partial^2 \overline{F}(t,u)}{\partial t \partial u}$. The mixed derivative of the copula C is simply the copula density. Substituting the joint survival copula in place of the survival function and the copula density leads to

$$h_C(x,y) = \frac{\frac{\partial^2 C(x,y)}{\partial x \partial y}}{1 - x - y + C(x,y)} \tag{4.17}$$

And subsequently inputting the marginals, and using the chain rule, we arrive at an expression for a joint hazard function in terms of C.

$$h_C(F_T(t), F_U(u)) = \frac{\frac{\partial^2 C(F_T(t),F_U(u))}{\partial F_T(t) \partial F_U(u)} \cdot \frac{\partial F_T(t)}{\partial t} \frac{\partial F_U(u)}{\partial u}}{1 - F_T(t) - F_U(u) + C(F_T(t), F_U(u))} \tag{4.18}$$

As before, this hazard function is equal to the intensity function under minimal repair. For each copula a hazard function is derived (in Maple) and simplified where possible. These are displayed in table A.1. Now that these hazard functions are available, the next step is to find the one best suited to the data. The next chapter considers the optimisation of these hazard functions to the data set, ranking the copulas, and the resulting fits.

Chapter 5

Fitting of Copulas

In this chapter we discuss the method used to fit the copulas to the warranty dataset described in Chapter 3. We will use the 36 months time cut and compare the likelihood of the different copulas over that time cut (see section 3.3). Differential evolution optimisation is introduced, and its use is justified. Finally the results of the fit are provided, and the best fit copula is chosen for the simulation of warranty data.

5.1 What defines a good fit?

Goodness of fit is a statistical approach to evaluate how well a statistical model fits a set of observations. Univariate goodness of fit tests such as the Kolmogrov-Smirnov test give an absolute measure of the quality of a fit. Current bi-variate tests however do not provide this absolute measure. There is a two-sample Kolmogorov-Smirnov test (not to be confused with bi-variate KS), but that is not applicable in this thesis. The bi-variate distribution free Kolmogorov-Smirnov test is not as mature and developed as of yet, however work done by Justel, Peña and Zamar[18] has laid the groundwork for its development. As the 2D-KS test is still in its infancy and that simulation is required to approximate the test statistic it is not used here. It is with this mind that we look to Akaike information crite-

rion for model comparison.

5.1.1 Akaike information criterion

Akaike information criterion (AIC) is a measure of relative goodness-of-fit of a model over a set of observations. It was first introduced in the paper "A new look at the statistical model identification" by Hirotugu Akaike in 1974 [1]. It can be used to compare multiple models with different numbers of parameters, in our case the copulas can have one or two parameters.

$$AIC = 2k - \ln L \tag{5.1}$$

Where k is the number of parameters, and $\ln L$ is the log likelihood of a model given the observations. If we consider a set of potential models, the model with the smallest AIC is the best model. To find the log likelihood, and subsequently the AIC of each model we must first find the maximum likelihood of our models.

5.1.2 Maximum likelihood Estimator

Maximum likelihood estimation is a well known method for estimating model parameters. As the name suggest it involves maximising the likelihood of a model given a set of observations by modifying model parameters. As derived in formulae (2.14), the log likelihood estimator for a vehicle i is given by

$$\ln L_i = \sum_{j=1}^{n_i} \ln(h(t_{i,j}, u_{i,j})) - \sum_{j=0}^{n_i} \int_{t_{i,j}}^{t_{i,j+1}} \int_{u_{i,j}}^{u_{i,j+1}} h(s,r) dr ds \tag{5.2}$$

And thus the log likelihood of the model over the entire data set is

$$\ln L = \sum_{i=1}^{M} \ln L_i \tag{5.3}$$

Each evaluation of this function is computationally expensive for two main reasons. First due to the size of the warranty set (containing 44890 vehicles with a total of 43520 claims), and secondly the necessity to perform bivariate integration at every step. Due to this limitation, the use of non-standard optimisation techniques was considered. The method that was decided upon was Differential Evolution.

5.2 Optimisation using Differential Evolution

What follows is an overview of Differential Evolution (DE), for a detailed algorithm description see Appendix C. The general idea of Differential evolution involves evaluating a function to be optimised at numerous points (known as candidate solutions) and using these results to choose new locations in which to search. In the language of Computer Science, Differential evolution (DE) is an *iterative meta-heuristic* method. This means that it is a general method for finding an acceptable solution to an optimisation problem. This is different to the usual goal of finding an optimal or near optimal solution. Meta-heuristics are also characterised by being more abstract than traditional techniques for optimisation in the sense that they involve general heuristics and are not problem specific. They are often inspired by natural phenomena such as flocks of birds (for Particle Swarm Optimisation) and Evolution by Natural Selection (for DE).

DE comes from the family of evolutionary computational techniques, based on ideas of evolution by natural selection. The central idea is to use "evolutionary pressure" on a population of candidate solutions. In this implementation of DE, we use different values of the model parameters as candidate solutions. By comparing the "fitness" or quality of these candidate solutions and combining the best ones, we can maximise the log likelihood function. The choice of DE over standard optimisation techniques is due to the following advantages.

5.2.1 Advantages

- Makes no assumptions about the function being optimised nor the search space
- Minimises the number of function evaluations
- Doesn't use the gradient of the function for optimisation
- Can be easily parallelised to run on multiple processors/cores.

This method however does have many shortcomings.

5.2.2 Disadvantages

- No guarantee of having an acceptable solution on termination. Thus the need for multiple independent runs of the algorithm to verify any results.
- Convergence can be slow, or non-existent.
- Dependent on fine tuning, choosing the starting point and algorithm parameters can drastically affect performance and results

Despite this, DE allows us to find the maximum likelihood of our copulas given the data set. For the purposes of this thesis, the *fitness function* used was the maximum likelihood function $\ln L$. The *population size*, *differential weighting factor*, and *crossover rate* were $NP = 50$, $F = 0.8$, and $CR = 0.5$ respectively. Finally the *termination criteria* was either reaching 2000 generations or no change in fitness for 50 generations, which ever came first. These algorithm parameters were chosen by experimentation and are very much dataset dependent.

The software was written in R and important sections are provided in Appendix B. The optimisation in this thesis was performed by using the

R library *DEoptim*, running in a parallel environment. The parallel environment was the VUW high performance computing facility, using twenty fours cores on each machine. The average run time for each optimisation was two days. For each copula, thirty runs were performed to minimise the chance of a poor optimisation.

5.3 Results

The following results (Table 5.1) were obtained by using differential evolution for optimisation to maximise the log likelihood function described in the previous section. It is possible to see that the Gumbel Copula (with parameter $\delta = 0.214$) is the best fitted copula to our data set. This implies that the Gumbel copula best described the correlation between our two variables Age T and Mileage U.

Copula	$\ln L$	AIC	Parameters
Product	-587465	1174932	-
Ali	-572875	1145752	$\delta = 1.000$
Clayton	-565068	1130138	$\delta = 4.118$
FGM	-583761	1167524	$\delta = 1.000$
Frank	-566861	1133724	$\delta = 13.373$
Joe	-570028	1140058	$\delta = 7.480$
Gaussian	-564643	1129288	$\rho = 0.923$
Student-t	-564626	1129256	$(\nu = 55, \rho = 0.922)$
Gumbel	-564387	1128776	$\delta = 0.214$

Table 5.1: Table of copula likelihoods for 36 months

It suggests that early in a vehicles life, claim age and mileage are heavily correlated, and that this correlation reduces later in a vehicles life. Intuitively, simple graphical comparison between the real warranty claims (as in Figure 3.1) and the (small to mid size) realisations of the Gumbel

Copula (Figure 4.6) show they are very similar in terms of spread. In this chapter we have seen the use of AIC for model selection, and the use of differential evolution for optimisation. We have chosen a Gumbel copula as the best candidate and will discard the other copulas. Now that we have this choice of the Gumbel copula, it will be used in the next chapter on simulation of warranty data.

Chapter 6

Simulation and Prediction

This chapter introduces the algorithm for simulation, and how it is used for prediction. The Gumbel copula optimised over 36 months of data is used to simulate 36 months of data and is compared with the strata model. The fitting process in Chapter 5 is performed at the 24 month time cut, and used to predict 36 months. Finally the 36 month copula is used to predict 48 months of claims.

6.1 Simulation

As there is no closed form solution for $E[N(t,u)]$ we are forced to estimate it with simulation. For each of the M vehicles in the data set, the following algorithm is performed, creating a new data set of simulated claims. Due to the stochastic nature of the simulation, this simulation is performed thirty times, creating thirty different data sets and the result averaged. The following simulation algorithm works by splitting the warranty region into elementary regions in a fine grid or mesh and evaluating the likelihood of a claim at a grid point, using the intensity function at that point. To get appropriate vehicle trajectories and age censoring, we sample from the empirical distribution of mileage accumulation rate and censoring age respectively.

6.1.1 Notation

In the following algorithm, these notations are used.

- i - vehicle i (M vehicles in total)
- n - number of claims for vehicle i
- D - search direction for vehicle i after claim n
- A_i - censoring age for vehicle i
- (t, u) - grid point
- $T_{i,max}$ - maximum age attainable by vehicle i
- $U_{i,max}$ - maximum mileage attainable by vehicle i
- $T_{i,n}$ - age at nth claims by vehicle i
- $U_{i,n}$ - mileage at nth claims by vehicle i

6.1.2 Grid Algorithm

Algorithm 1 Grid Algorithm

Set $i = 1$,
while $i \leq M$ **do**
  Set $n = 0$, $T_{i,0} = U_{i,0} = 0$
  repeat
    Choose a search direction D and define the warranty region
    Divide the region into elementary regions each of area $\delta t \delta u$
    for each grid point (t, u) **do**
      Generate a uniform random variate X
      if $X < h(t, u)\delta t \delta u$ **then**
        $n = n + 1$
        $(T_{i,n}, U_{i,n}) = (t, u)$
        Keep all grid points (t, u) where $(t, u) \geq (T_{i,n}, U_{i,n})$
      end if
    end for
  until entire warranty region is searched ($(t, u) \geq (T_{i,max}, U_{i,max})$)
  Set $i = i + 1$
end while

The algorithm is now detailed more completely below. Initialise:

- The number of claims $n = 0$
- The starting point for each vehicle $T_{i,0} = 0$, and $U_{i,0} = 0$
- The warranty cut off points U_{max} and T_{max} are set to their respective values (36,000 miles and 3 years respectively)

Step 1 Choose a search direction D from the distribution of MAR for first claims ($F_{r_1}(x)$, see section 2.3) and a censoring age A_i from the empirical censoring age distribution $F_A(x)$ (see section 2.4). Use these

two values in combination to construct the searchable warranty region as follows.

First, define the maximum lifetime of the vehicle as the minimum of the warranty cut off and the censoring age.

$$T_{i,max} = min(T_{max}, A_i)$$

Second, the mileage accumulation of the vehicle is estimated by the max lifetime multiplied by the search direction D. For the first claim this takes the following form,

$$M_i = U_{i,0} + D(T_{i,max} - T_{i,0})$$

Finally the maximum attainable mileage for the vehicle is defined as the minimum of the mileage accumulation of the vehicle and the mileage cutoff.

$$U_{i,max} = min(M_i, U_{max}) \tag{6.1}$$

Using $U_{i,max}$ and $T_{i,max}$, the searchable warranty region is defined as $[T_{i,0}, T_{i,max}] \times [U_{i,0}, U_{i,max}]$ as shown in Figure 6.1.

Step 2 Divide the searchable warranty region into a grid of very small elementary regions each with an area of $\delta t \delta u$ as in Figure 6.2. The number of these vertical and horizontal divisions is simply $\frac{T_{i,max}-T_{i,0}}{\delta t}$ and $\frac{U_{i,max}-U_{i,0}}{\delta u}$ respectively.

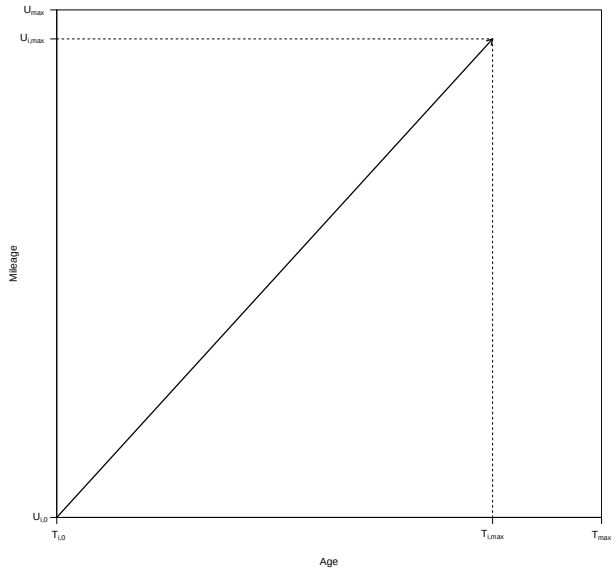

Figure 6.1: Grid algorithm Step 1

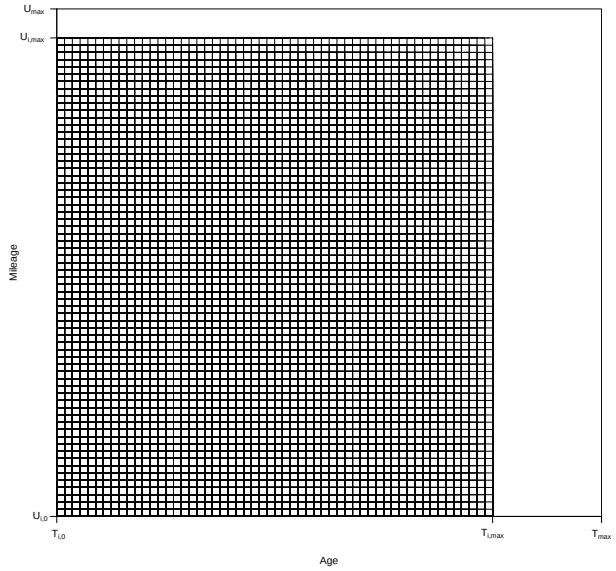

Figure 6.2: Grid algorithm Step 2

Step 3 Evaluate the intensity function $h_C(t,u)$ for each elementary region

The simulation is performed radially outwards in "rings" through the warranty region as described by the numbering given in Figure 6.3. We treat each elementary region as a Bernoulli random variable X with distribution function given below

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq h_C(t,u)\delta t\delta u \\ 0 & \text{otherwise} \end{cases}$$

A grid point (t,u) is defined as $t = T_{i,n} + a\delta t$ and $u = U_{i,n} + b\delta u$, for example in Figure 6.3 the highlighted point is $(t,u) = (T_{i,0}+4\delta t, U_{i,0}+3\delta u)$. For each grid point (t,u), a uniform random variate is generated (X) and compared against the intensity function at that point. More formally the test is $X < h_C(t,u)\delta t\delta u$. If the test is positive, then a claim has occurred at point (t,u), otherwise the algorithm continues to the next grid point. If the entire searchable warranty region is searched in this fashion without a claim, the algorithm is terminated. If a claim is found at say $(t,u) = (T_{i,n} + 4\delta t, U_{i,n} + 3\delta u)$, we need to change the search direction to search for the next claim. First we set $T_{i,n+1} = T_{i,n} + 4\delta t$ and $U_{i,n+1} = U_{i,n} + 3\delta u$, and then continue to Step 4.

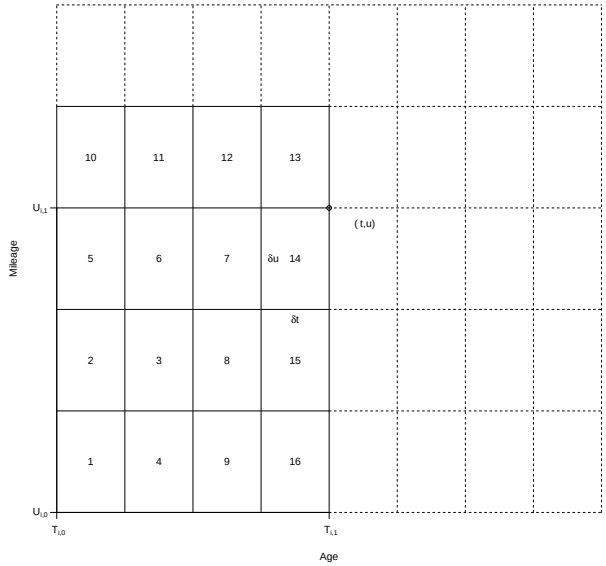

Figure 6.3: Grid algorithm Step 3

Step 4 Increment n and choose a new search direction D (from the MAR distribution of nth claims $F_{r_n}(x)$). The maximum lifetime of the vehicle is not changed, however the maximum attainable mileage is changed. As before, the mileage estimate is calculated $M_i = U_{i,1} + D(T_{i,max} - T_{i,1})$, with the general step being

$$M_i = U_{i,n} + D(T_{i,max} - T_{i,n})$$

and the maximum attainable mileage defined as $U_{i,max} = min(M_i, U_{max})$. This defines a new searchable warranty region $[T_{i,1}, T_{i,max}] \times [U_{i,1}, U_{i,max}]$, note that the region starts from the previous claim outwards as in Figure 6.4. Now the simulation process repeats itself. Go to Step 2 and proceed as before, as shown in Figure 6.5.

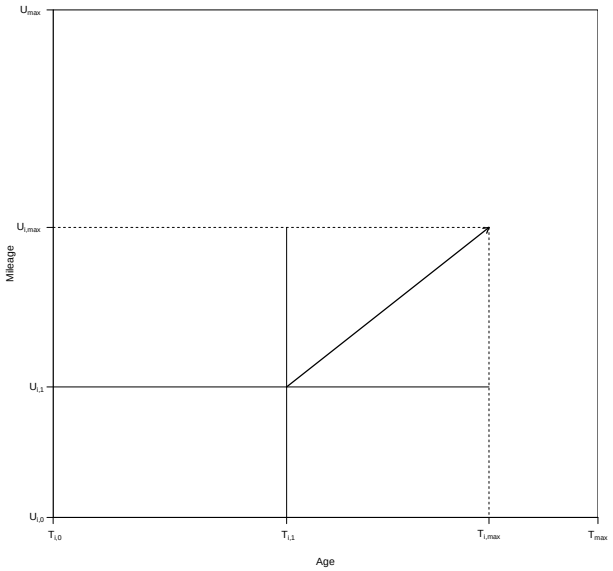

Figure 6.4: Grid algorithm Step 4

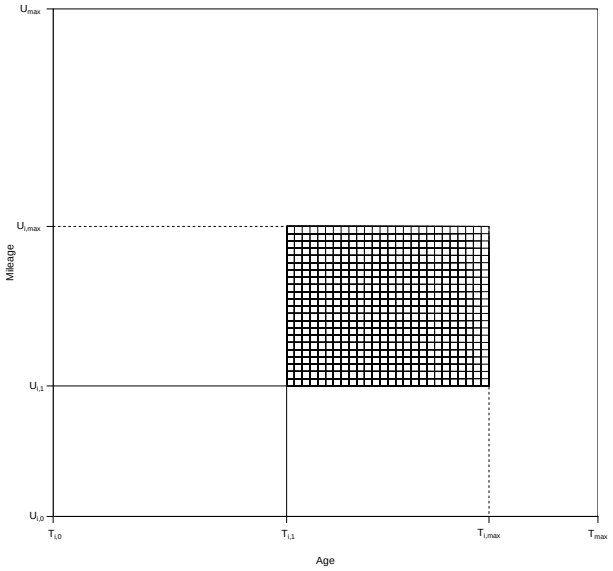

Figure 6.5: Grid algorithm Step 2 for the next iteration

6.2 Prediction

Prediction in this model involves the simulation of data sets using the algorithm in the previous section. After the simulated data set is created, the strata model is used to estimate $\hat{\Lambda}(t)$ at 36,000 miles and 36 months. After thirty are generated, the central limit theorem is used to create a 95% confidence interval for the simulated data MCF. Given the data for 36 months, ideally the simulation process should be able predict within this time space, the results follow.

6.2.1 Prediction of 36 months of claims using all 36 months of data

In this prediction the marginal distributions (Both Weibull from chapter 3) and the Gumbel copula function chosen in chapter 5 are used in the simulation. The goal of this simulation is to test the usefulness of this method as a predictor for mean number of claims. We first consider the case of predicting thirty-six months of data using the intensity function fitted over 36 months of data. The strata model is then calculated over the simulated data set for both age and mileage, and compared with the strata model calculated over the real data. In Figure 6.6, the strata model calculated over the real and simulated data set are compared, note that the simulated values stays within the 95% confidence bounds of the real values. Figure 6.7 again shows $\hat{\Lambda}(t)$ for the real and simulated dataset for mileage, note the divergence for higher mileages. This shows that the simulation of claim age is better than for mileage at high values.

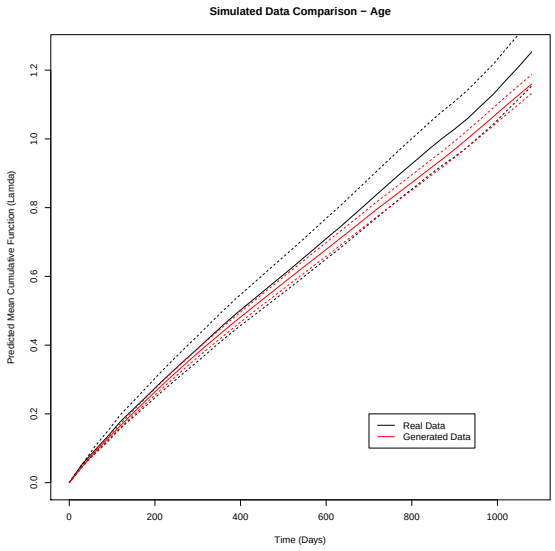

Figure 6.6: Predicted Age MCF for 36 months of claims using all 36 months of data

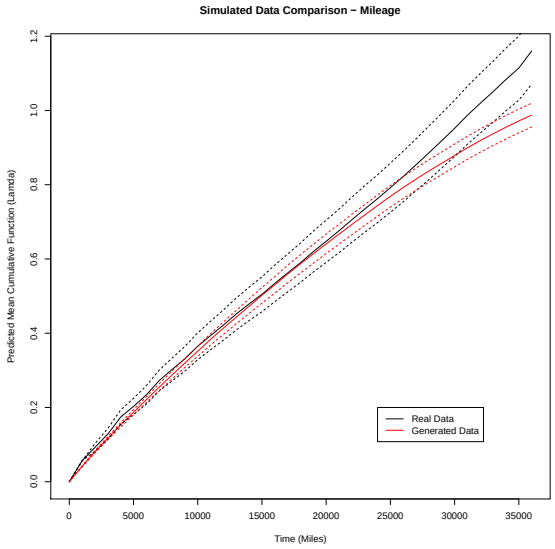

Figure 6.7: Predicted Mileage MCF for 36 months of claims using all 36 months of data

Now that we have an idea of the simulations performance under ideal conditions, we consider the case of predicting thirty six months of data using 24 months of data.

6.2.2 Prediction of 36 months of claims using the first 24 months of data

For the twenty four month case, we must repeat the process of choosing marginal distributions and joint distributions as detailed in the previous chapters on a data set consisting only of records that would have been available twenty four months into the warranty data collection process. This is the twenty four month time cut found in chapter 3. We briefly repeat this process here. First the marginal distributions are selected, the best fit for claim age was Weibull with $\beta_1 = 0.872$ and $\theta_1 = 880.005$. The best fit for claim mileage was again Weibull, now with parameters $\beta_2 = 0.946$ and $\theta_2 = 317.600$.

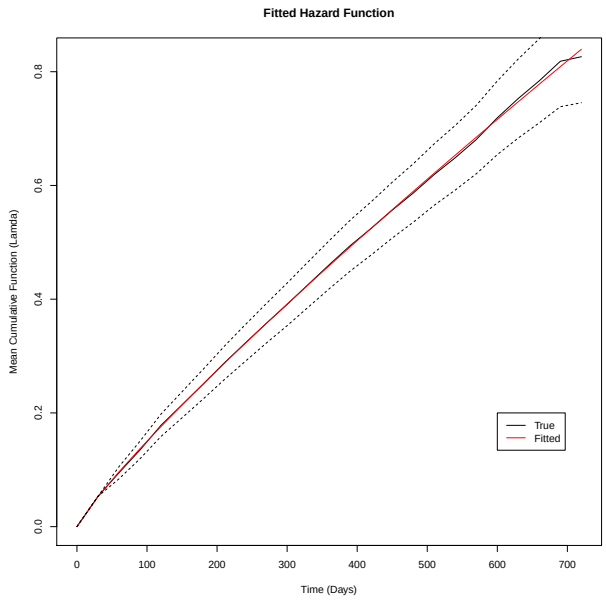

Figure 6.8: Fitted MCF to Strata MCF for age at 24 months

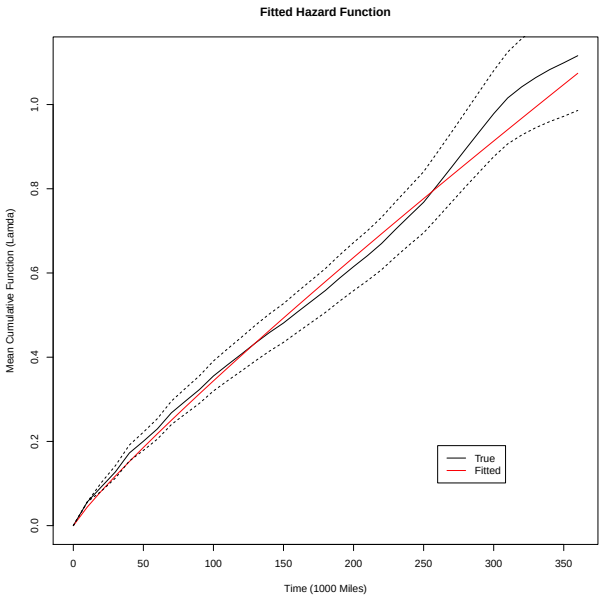

Figure 6.9: Fitted MCF to Strata MCF for Mileage at 24 months

Note the similarity between the marginal distributions for twenty four months, and the ones for thirty six months. These fits can be shown in Figures 6.8 and 6.9, note that the mileage fit is worse than the age fit though still within the 95% confidence interval. Now that we have these marginals, we will find the best fitted copula to join them. The results can be seen in Table 6.1.

As in the thirty six month dataset, the Gumbel copula is still the best fit. We now have all the pieces required to perform the simulation process described in the previous section. In this case, we will predict 36 months and 36,0000 miles of claims and compare these with the real database of thirty six months of data. In figure 6.10 we can see that the prediction of mean number of claims is within the 95% confidence interval for age. However figure 6.11, shows a large deviation at approximately 26,000 miles.

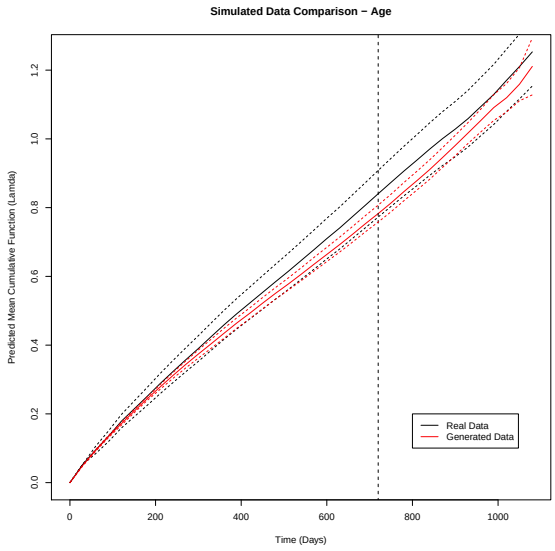

Figure 6.10: Predicted Age MCF for 36 months of claims using the first 24 months of data

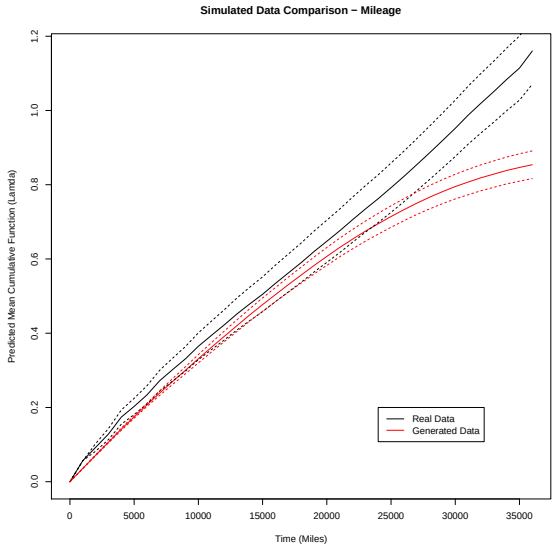

Figure 6.11: Predicted Mileage MCF for 36 months of claims using the first 24 months of data

We will discuss this deviation of the mean number of claims in the next chapter and its implications for this process. Though this does not inspire much confidence in this method to predict number of claims based on mileage accumulation, We attempt to predict at 48 months (and 48,000 miles) and see what the results entail.

6.2.3 Prediction of 48 months of claims using all 36 months of data

As a final test, we use the 36 month time cut to predict a 48 month, and 48,000 mile warranty region. For this section we assume no vehicles are sold after 36 months since the vehicle was first released. This assumption holds for our 24 month time cut of the database but seems unlikely to hold true if a manufacturer offered such an extended warranty agreement (as the sale process may extend further into a vehicles life). The distribution of age censoring is therefore shifted by twelve months, as shown in section 3.7.

Copula	$\ln L$	AIC	Parameters
Product	-349621	699244	-
Ali	-336500	673002	$\delta = 0.999$
Clayton	-331036	662074	$\delta = 3.823$
FGM	-345625	691252	$\delta = 1.000$
Frank	-330894	661790	$\delta = 16.110$
Joe	-332812	665626	$\delta = 10.146$
Gaussian	-329435	658872	$\rho = 0.935$
Student-t	-329368	658740	$(\nu = 20, \rho = 0.936)$
Gumbel	-328850	657702	$\delta = 0.186$

Table 6.1: Table of copula likelihoods for 24 months

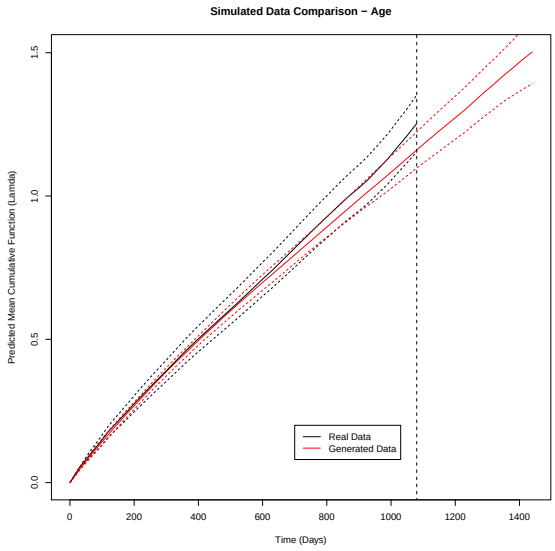

Figure 6.12: Predicted Age MCF for 48 months of claims using all 36 months of data

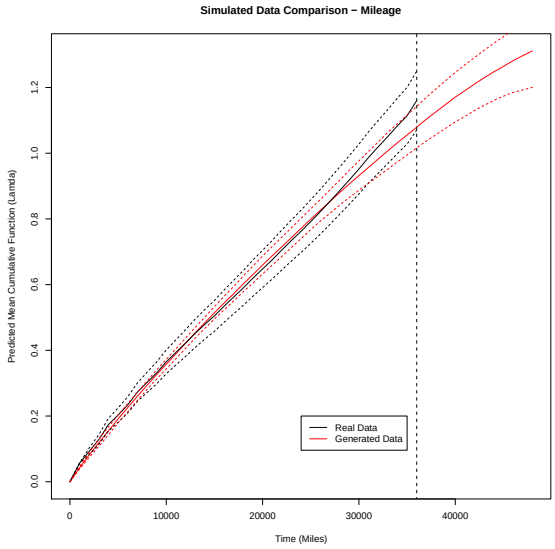

Figure 6.13: Predicted Mileage MCF for 48 months of claims using all 36 months of data

The simulation method is again repeated, however now with $T_{max} = 48$ months and $U_{max} = 48000$ miles. The results are shown above in Figures 6.12 and 6.13, note that the mean number of claims as a function of age and (surprising) mileage performs well for less than 36 months. Indeed the prediction of mileage accumulation seems better than in the 24 month prediction case, this is likely an artefact of the choice of copula and will be discussed in the next chapter. Examples of simulated datasets are presented in Appendix D.

In this chapter we have shown how to predict using the 2D-model from Chapter 2. Furthermore we have shown that under ideal circumstances it performs well for time as age but poorly for time as mileage. It seems obvious that the method used has some very serious and perhaps even fundamental limitations. In the next chapter, we will conclude these findings and make suggestions for further work.

Chapter 7

Conclusion

The key objective for this thesis was to simulate automotive warranty claims in an attempt to predict mean cumulative number of claims as a function of vehicle age and mileage. In this chapter we will discuss the results and the limitations of the method used. Finally further research directions and extensions are described.

7.1 Contributions

In this thesis, we have extended the work done in Chukova and Hirose (2006) and in Baik, Murthy and Jack (2003) and applied it to a real dataset. We cleaned and analysed a real warranty dataset and fitted marginal distributions to the age and mileage of claims. We investigated nine copulas to try and find a hazard function which best reflected the correlation in the dataset. We also created a very simple algorithm for simulating the claims process of the vehicles, and then compared them with the real dataset.

This comparison showed that the method predicted claim age more accurately than claim mileage. This can be seen in the deviation of the MCF (in Figures 6.7, 6.11 , and 6.13) and the difference between the simulated claim plots as seen in Appendix D and the real dataset plot in Chapter 3 (Figure 3.1 on page 42). Indeed it appears that the symmetry of the cop-

ulas used caused the general bulk of the claims to be more symmetric in correlation between age and mileage than in the real data. This suggests a further direction for research, the use of asymmetric copulas. The comparison for the 36 to 48 month prediction showed that the 95% confidence interval of this prediction from this method is narrower than the linear regression example given in Chapter 2. However the "shifting" of the censoring age distribution used to get this result is likely to not be tenable in a real dataset. This is due to the fact that the sales process would likely extend much further into the data collection process were a longer warranty offered.

7.2 Limitations

A major caveat of this research was that the model was only tested on one dataset and therefore its true power for prediction has not been fully investigated. As can be seen in the preceding chapters the method has a few limitations, the first important one being that it does not have a simple closed form of the expected number of claims. This forces us to use numerical or simulation techniques.

Another limitation is the lack of a general two dimensional distribution-free goodness-of-fit test. Comparison of simulated datasets to real datasets is in this thesis performed by comparing the estimated mean cumulative functions. This is unfortunately a poor indicator of their comparative probabilistic structure. Furthermore the use of the strata model (to estimate the mean cumulative function) was less than ideal as it introduced a new layer of "uncertainty". A more elegant way to compare the datasets would have been 2D Kolmogorov-Smirnov or equivalent. However the use of 2D-KS with censored data, seems in itself a large research topic and out of the scope of this thesis.

A final important limitation is the algorithm used to simulate claims. The algorithm described has drawbacks, it is computationally slow, and is

very dependent on grid size. If the grid is too large, the Bernoulli approximation of the probability of each elementary region having a claim breaks down, this can lead to overestimating the number of claims. Also, the way the grid is searched by considering ever increasing "rings',' favours finding claims that occur earlier in age and mileage. It is analogous to breadth first search in Graph theory, finding the closest claims (or nodes) first. This means that the simulation will be slightly biased to claims occurring earlier, and this can be seen in the hex bin graphs in Appendix D.

7.3 Extensions

The obvious extension to this method is to consider, asymmetric copulas, in particular copulas for which $c(x, y) \neq c(y, x)$. We can see the need for asymmetric copulas for this dataset from the plot in Figure 3.1. In the 2008 paper "Construction of asymmetric multivariate copulas" [22], Eckhard Liebscher proposes a method for constructing asymmetric copulas from the family of Archimedian copulas, meaning that the ground work has already been laid to consider this avenue. We believe using this will greatly improve the predictive power of the method for the case of mileage.

As only one dataset is used, the use of other datasets from different years and companies will be needed to verify the assumptions made in this thesis. Another extension, would be considering more potential lifetime distributions and copula functions. The fitting could be improved by using another more suitable optimisation algorithm. As in the end Differential Evolution was overpowered for fitting such a relatively small number of parameters. It did however have the benefit of running on parallelised hardware without much trouble. A further comparison with other two dimensional models is also desirable, including the two-dimensional mean cumulative function in [13].

This model could also be extended to take into account failures that occur due common faults, such has been seen in vehicle recalls in recent

years. Under the condition of a common fault, we can still assume the vehicles trajectories remain independent. In this case the number of estimated claims will be erroneous, and will depend on the nature of the fault. This situation has not been considered in the derivation of 2.14 and therefore is a possible research direction.

Another suggestion is to consider simulation directly from copulas, rather than the discretisation approach used in the grid algorithm. This will allow us to negate the potential bias that the simulation introduced. Some progress has been made in this direction for the prediction of first claims, but not for the subsequent claims. This, in combination with asymmetric copulas, is in this author's opinion, the most promising research direction.

Appendix A

Table of Hazard Functions

Copula	Derived Hazard Function $h_C(F_T(t), F_U(u))$
Product	$\left(\frac{\partial F_T(t)}{\partial t}\frac{\partial F_U(u)}{\partial u}\right)\cdot\frac{1}{(F_T(t)-1)(F_U(u)-1)}$
AMH	$\left(\frac{\partial F_T(t)}{\partial t}\frac{\partial F_U(u)}{\partial u}\right)\cdot\frac{1+\delta^2+\delta F_U(u)+\delta F_T(t)-\delta^2 F_U(u)-\delta^2 F_T(t)+\delta^2 F_T(t)F_U(u)+\delta F_T(t)F_U(u)-2\delta}{(F_U(u)-1)(F_T(t)-1)(\delta F_T(t)+\delta F_U(u)+1-\delta)(\delta-1-\delta F_U(u)-\delta F_T(t)+\delta F_T(t)F_U(u))^2}$
Clayton	$\left(\frac{\partial F_T(t)}{\partial t}\frac{\partial F_U(u)}{\partial u}\right)\cdot\frac{\left(F_T(t)^{1-\delta}+F_U(u)^{1-\delta}-1\right)^{-\frac{1}{\delta-1}-2}F_U(u)^{-\delta}F_T(t)^{-\delta}\delta}{\left(1-F_T(t)-F_U(u)+\left(F_T(t)^{1-\delta}+F_U(u)^{1-\delta}-1\right)^{-\frac{1}{\delta-1}}\right)}$
FGM	$\left(\frac{\partial F_T(t)}{\partial t}\frac{\partial F_U(u)}{\partial u}\right)\cdot\frac{1+\delta-2\delta F_U(u)-2\delta F_T(t)+4\delta F_T(t)F_U(u)}{(F_U(u)-1)(F_T(t)-1)(\delta F_T(t)F_U(u)+1)}$
Frank	$\left(\frac{\partial F_T(t)}{\partial t}\frac{\partial F_U(u)}{\partial u}\right)\cdot\frac{\left(\mathrm{e}^{-\delta}-1\right)\delta^2 xy}{\left(\delta F_T(t)+\delta F_U(u)-\delta+\ln\left(\frac{\mathrm{e}^{-\delta}+xy-x-y}{\mathrm{e}^{-\delta}-1}\right)\right)\left(\mathrm{e}^{-\delta}+xy-x-y\right)^2}$, where $x=\mathrm{e}^{-\delta F_T(t)}$ and $y=\mathrm{e}^{-\delta F_U(u)}$
Joe	$\left(\frac{\partial F_T(t)}{\partial t}\frac{\partial F_U(u)}{\partial u}\right)\cdot\frac{\left(y^\delta+x^\delta-x^\delta y^\delta\right)^{\frac{1}{\delta}}y^{\delta-1}x^{\delta-1}\left(1-y^\delta-x^\delta+x^\delta y^\delta-\delta\right)}{\left(x^\delta y^\delta-y^\delta-x^\delta\right)^2\left(x+y+\left(y^\delta+x^\delta-x^\delta y^\delta\right)^{\frac{1}{\delta}}\right)}$, where $x=1-F_T(t)$ and $y=1-F_U(u)$
Gumbel	$\left(\frac{\partial F_T(t)}{\partial t}\frac{\partial F_U(u)}{\partial u}\right)\cdot\frac{\left(x^{\frac{1}{\delta}}+y^{\frac{1}{\delta}}\right)^{\delta-2}y^{\frac{1}{\delta}-1}x^{\frac{1}{\delta}-1}\left(\frac{1}{\delta}-1+\left(x^{\frac{1}{\delta}}+y^{\frac{1}{\delta}}\right)^{\delta}\right)}{F_T(t)F_U(u)\left(1-F_T(t)-F_U(u)+\mathrm{e}^{-\left(x^{\frac{1}{\delta}}+y^{\delta-1}\right)^{\delta}}\right)}\mathrm{e}^{-\left(x^{\frac{1}{\delta}}+y^{\frac{1}{\delta}}\right)^{\delta}}$, where $x=-\ln\left(F_T(t)\right)$ and y $=-\ln\left(F_U(u)\right)$
Gaussian	$\left(\frac{\partial F_T(t)}{\partial t}\frac{\partial F_U(u)}{\partial u}\right)\cdot\frac{\phi_\rho\left(\Phi^{-1}(F_T(t)),\Phi^{-1}(F_U(u))\right)}{\phi(\Phi^{-1}(F_T(t)))\phi(\Phi^{-1}(F_U(u)))}$
Student-t	$\left(\frac{\partial F_T(t)}{\partial t}\frac{\partial F_U(u)}{\partial u}\right)\cdot\frac{t'_{\nu,\rho}(t_\nu^{-1}(F_T(t)),t_\nu^{-1}(F_U(u)))}{t'_\nu(t_\nu^{-1}(F_T(t)))t'_\nu(t_\nu^{-1}(F_U(u)))}$, where t'_ν is the Student-t density function

Table A.1: Table of hazard functions for chosen copulas

Appendix B

R Code

This appendix contains excerpts from the R source used in optimisation and simulation in this thesis.

B.1 Differential Evolution

This code is dependent on the library *pracma* for integration. The Differential Evolution implementation is provided by the *DEOptim* library which recently has added a parallel implementation. It is dependent on libraries *iterators*, *foreach*, *Rmpi*, and *snow* for parallelisation. It also expects a hazard function h, and that the warranty data is named $alldata$.

The function $ll(x)$ is the log likelihood function lnL. Note that it returns $-lnL$, this is due to fact that most optimisation routines in R focus on minimisation not maximisation.

```
1 ll <- function(x) {  # Likelihood of all vehicles
2   d <<- x[1];
3   total <- 0;
4   for (i in 1:length(alldata)) {
5     total <- total + lli(alldata[[i]]);
6   }
7   if (is.nan(total)) { # If the likelihood is undefined
8     total <- -2147483647 ; # Make it negative infinity
```

```
9   }
10  return(-1*total);  # Minimise instead of maximise
11 }
```

The function $lli(x)$ is the log likelihood function lnL_i, the likelihood of a single vehicle.

```
1  lli <- function(data) { # Likelihood of single vehicle
2    result <-0;
3    if (nrow(data)> 2) {
4      for (k in 1:nrow(data)) {
5        if (data[k,3] == 1) {
6    result <- result + log(h(data[k,2],data[k,1]));
7        }
8      }
9    }
10   for (k in 1:(nrow(data)-1)) {
11     in <- quad2d(h,data[k,2],data[k+1,2],data[k,1],data[k+1,1]);
12     result <- result - in;
13   }
14   return(result)
15 }
```

This next bit of code excerpt shows how to set up the parallel environment, and start the optimisation. Finally, it stops the parallel environment, and saves the results.

```
1  cl <- makeCluster(numSlaves) # Set up parallel environment
2
3  clusterEvalQ(cl, library(pracma,...)) # load any necessary
       libraries
4  clusterExport(cl, ...) # copy any necessary R objects
5
6  registerDoSNOW(cl)
7  # Run parallel DE optimisation of function ll
8  outDEoptim <- DEoptim(ll, c(l), c(u), DEoptim.control(...,
       parallelType = 2))
9
10 stopCluster(cl)
```

```
11
12 save(outDEoptim, file = "outDE.rData")
```

B.2 Grid Simulation

This code is dependent on the libraries pracma (for double integrands), and compiler (for improved speed). The simulation code has been moved to its own code block for ease of reading.

```
1  simulate <- function(N,h,maxage,censoring,ageinc)
2  {
3    maxmileage <- 360
4    claims <- rbind(c(0,0,0),c(0,0,0))
5    cars <- list()
6    for (z in 1:N) { # For each car
7      actualage <- 0
8      dir <- as.double(invmars(runif(1,0,1))) # Choose starting
           direction
9      if (censoring) {
10       actualage <- as.integer(invages(runif(1,0,1)))
11       age <- min(maxage,actualage) # Choose censoring age
12     } else {
13       age <- maxage
14     }
15     totalmileage <- min(360,mar*age) # Find maximum mileage
16
17     # Set up grid and starting points
18     absages <- seq(1,age,ageinc)
19     mileinc <- dir
20     minAge <- 1
21     minMileage <- 1
22     moffset <- 0
23     aoffset <- 0
24     alldir <- c(dir,0);
25
26     for (r in 1:age) {
```

```
27        #
28        # Simulation Code
29        #
30      }
31
32      cars <- c(cars,list(c(z,actualage,list(allmars))))
33    }
34    return(list("claims"=claims, "cars"=cars))
35  }
```

What follows is annotated simulation code for stepping radially outward through the grid and checking each point for claims. Once a claim has been found, the code selects a new direction and changes the search region.

```
1  for (r in 1:age) { # For each small increment of age
2
3    # Define current grid "band"
4    grida <- seq(minAge,r)*ageinc
5    grida <- c(grida,rep(r*ageinc, r-minMileage))
6    gridm <- rep(((r-aoffset)*mileinc)+moffset, r-minAge)
7    gridm <- c(gridm,(seq(r,minMileage)-aoffset)*mileinc+moffset)
8    grid <- cbind(grida,gridm)
9
10   # Evaulate the hazard function on this band
11   checks <- h(grid[,1],grid[,2])*(mileinc)*ageinc;
12
13   # Check if a claim occurs in the band
14   rdms <- runif(length(grid[,1]),0,1)
15   test <- rdms<checks
16   cclaims <- which(test)
17
18   if (length(cclaims) != 0)  { # If a claim occurs
19     if (grid[cclaims[1],1] <= age & grid[cclaims[1],2] <=
         maxmileage) {
20       # and it is within the max age and mileage bounds
21
```

```
22          claims <- rbind(claims,c(grid[cclaims[1],2],grid[cclaims
                [1],1],z))
23          # Add the claim
24
25          # Set the minimum bounds for claims to occur in
26          minAge <- (grid[cclaims[1],1])+1
27          minMileage <- ((grid[cclaims[1],2]-moffset)/mileinc)+1
28
29          # Choose a new direction and continue simulation from
                there
30          mileinc <- as.double(invmars(runif(1,0,1)))
31          alldirs <- rbind(allmars,c(mileinc,minAge))
32          moffset <- grid[cclaims[1],2]
33          aoffset <- grid[cclaims[1],1]
34        }
35      }
36  }
```

The simulation can be called using the following command, with parameters N, m, and a. These are the number of cars to simulate, the maximum age attainable (T_{max} or the time warranty cut off), and the age increment (δt) respectively. It assumes a hazard function is available and defined as $ht(t, u)$, and a mileage accumulation rate distribution to sample from. If age censoring is required, it expects a age censoring distribution to draw maximum ages from $T_{i,max}$.

```
1 results <- simulate(N=n,h=ht,maxage=m,censoring=TRUE,ageinc=a)
```

Appendix C

Differential Evolution Optimisation

C.1 Introduction

Differential evolution is a "metaheutristic" for optimisation, a very general procedure which makes no assumptions about the optimisation problem. It works by iterative improving a population of solutions, employing techniques that are analogous to evolutionary processes. There is no proof of convergence of this algorithm, however with a well chosen search space and algorithm parameters, it has been shown to be very effective in finding global optima. We shall define a *candidate solution* as a vector of size n.

$$\theta_{i,G} = (\theta_{i1,G}, \theta_{i2,G} \ldots, \theta_{in,G})$$

In this thesis the candidate solution is the vector of copula parameters. A population of solutions θ_G is a set of candidate solutions of size NP. During the Gth iteration of this algorithm, a new population θ_{G+1} is created from the previous one. The term *generation* is synonymous with iteration in this procedure, and will be used interchangeably. The algorithm parameters NP, F, and CR are user defined, and will be described in the

following sections.

C.2 The Algorithm

1. Initialise a population of NP solutions θ_i randomly in the search space and set generation counter G to 1.
2. For each solution θ_i:
   - **Mutation Step**
   - **Crossover Step**
   - **Fitness and Selection Step**
3. Check if any solution meets the **Stopping Criteria**, if so go to Step 4. Otherwise, increment G and return to Step 2 with the new population.
4. Terminate and use the best solution of the current generation G

C.2.1 Mutation Step

The mutation step of the algorithm is analogous to genetic mutation in biological evolution. Variety in candidate solutions is introduced in this step, by constructing a *noisy* vector. This noisy vector is the proportional linear combination of three other candidate solutions. It is generated by the following steps

1. Randomly select three solutions from θ_G, defined as $\theta_{j,G}$, $\theta_{k,G}$, and $\theta_{l,G}$.
2. Calculate the noisy vector $v_{i,G}$

$$v_{i,G} = \theta_{j,G} + F \cdot (\theta_{k,G} - \theta_{l,G}),$$

where $F \in [0, 2]$ is an algorithm parameter, the *differential weighting* factor. This weighting factor defines the degree of mutation that is introduced in this step. This process of mutation is self limiting, as over generations the population of solutions converges around optima and thus any linear combination of them is likely to be near the optima is well.

C.2.2 Crossover Step

The crossover step is the algorithm's equivalent of "breeding". The purpose is to combine candidate solutions to produce "offspring", this offspring is known as a trial vector. It is possible to use many different rules for crossover, in this thesis simple binomial crossover is used. The binomial rule randomly selects parts of the candidate vector and noisy vector to create the trial vector, using the following rule

$$u_{iz,G} = \begin{cases} v_{iz,G} & \text{if } U_z \leq CR \\ \theta_{zi,G} & \text{otherwise} \end{cases},$$

where U_z is a uniform random variate and $CR \in [0, 1]$ is the *crossover probability*. The crossover probability determines the proportion of the noisy vector that is introduced into the trial vector. This new trial vector is now compared against its "parent" candidate solution.

C.2.3 Fitness and Selection Step

A fitness function must be defined for this algorithm to work, it needs to measure the quality of a candidate solution in comparison to other candidate solutions. It is named after the concept in biological evolution, which indirectly rates the likelihood of a species propagating their genes under selective pressure (such as natural selection). In the case of this thesis, the fitness function is the log-likelihood function of a copula given the parameters specified by the candidate solution. This fitness function is used to

compare the trial vector $u_{i,G}$ with the candidate solution $\theta_{i,G}$. The fittest solution is added to the next generation, using the rule below.

$$\theta_{i,G+1} = \begin{cases} u_{i,G} & \text{if } f(u_{i,G}) < f(\theta_{i,G}) \\ \theta_{i,G} & \text{otherwise} \end{cases}$$

This selection rule guarantees that the next generation is at least as good as the previous generation. In general however, the next generation will improve on the previous one.

C.2.4 Stopping Criteria

Differential evolution has no natural stopping point (not unlike biological evolution). As such, there are many different termination criteria, here are some common crtieria.

- maximum number of iterations/generations.
- a solution which has a high enough "fitness"
- no change in fitness over multiple generations

If the any of these criteria are met, the algorithm will terminate and return the current "fittest" solution.

Appendix D

Simulated Data Examples

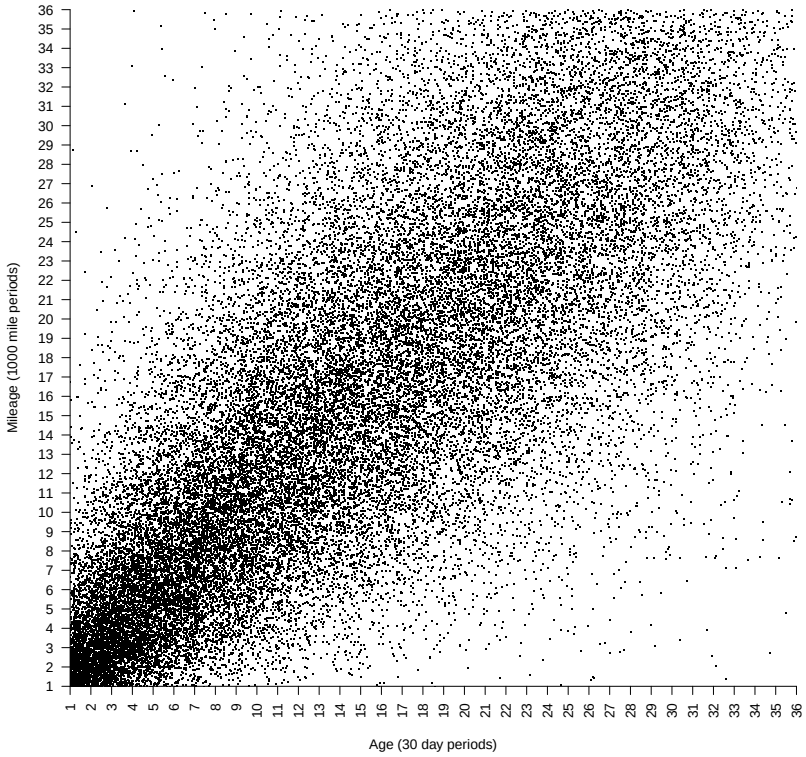

Figure D.1: Example simulated dataset for 36 to 36 months

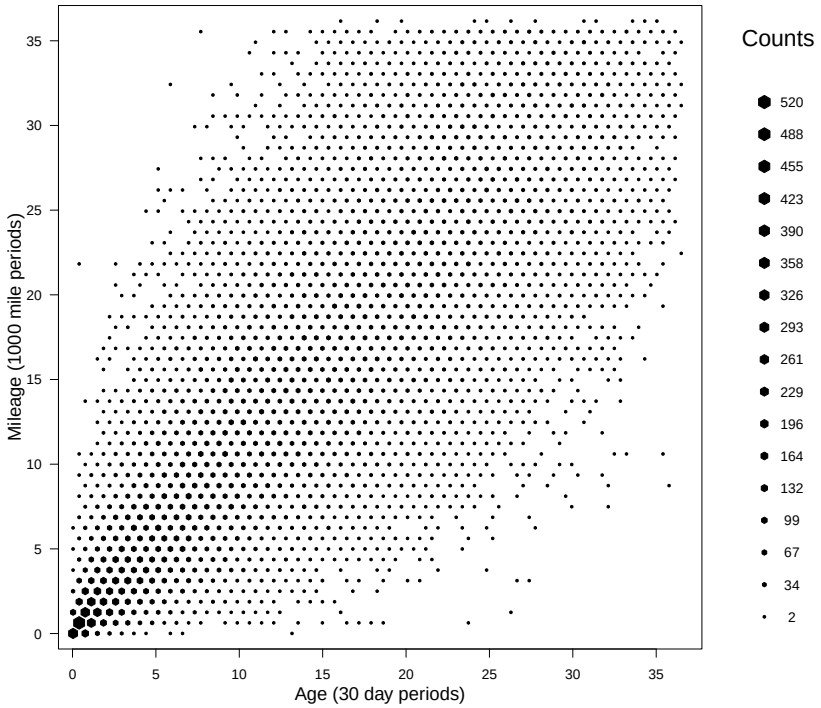

Figure D.2: Hex-bin graph of simulated dataset for 36 to 36 months

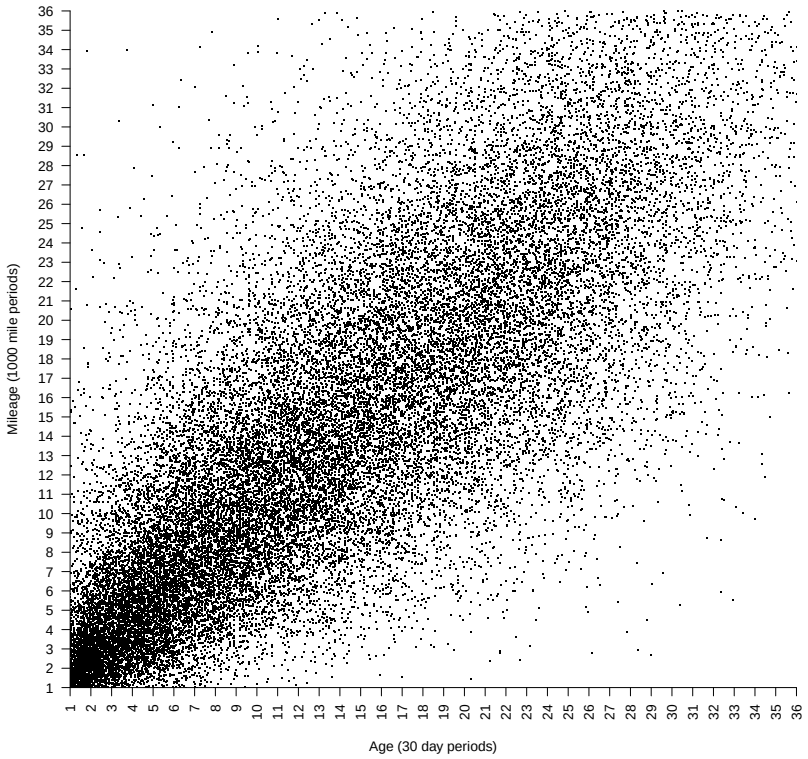

Figure D.3: Example simulated dataset for 24 to 36 months

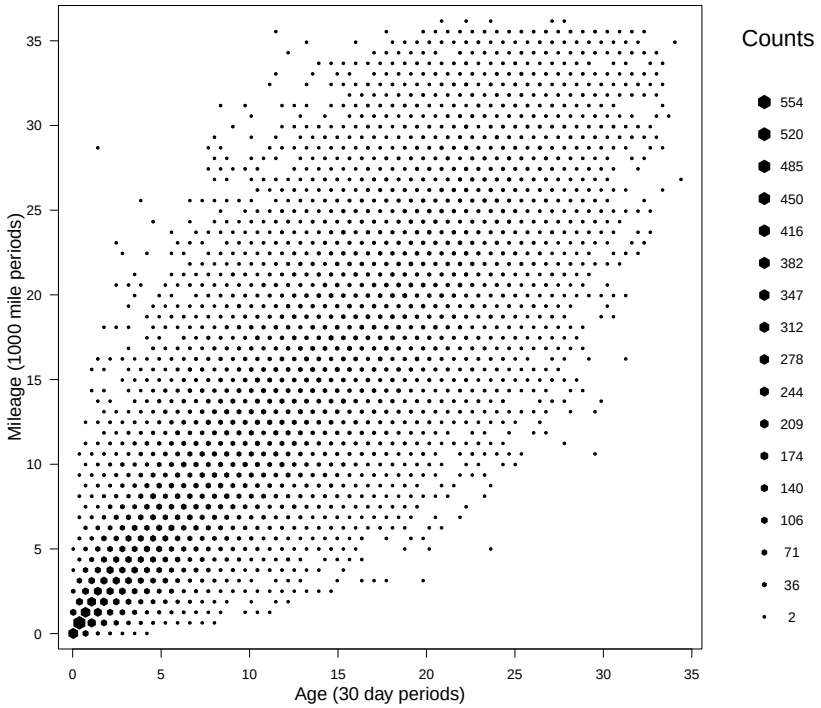

Figure D.4: Hex-bin graph of simulated dataset for 24 to 36 months

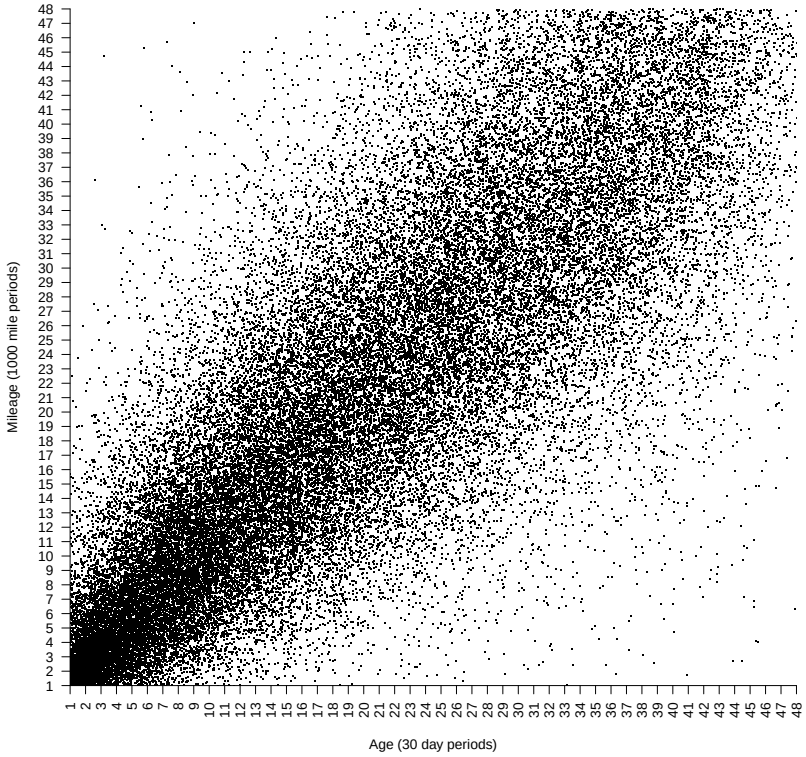

Figure D.5: Example simulated dataset for 36 to 48 months

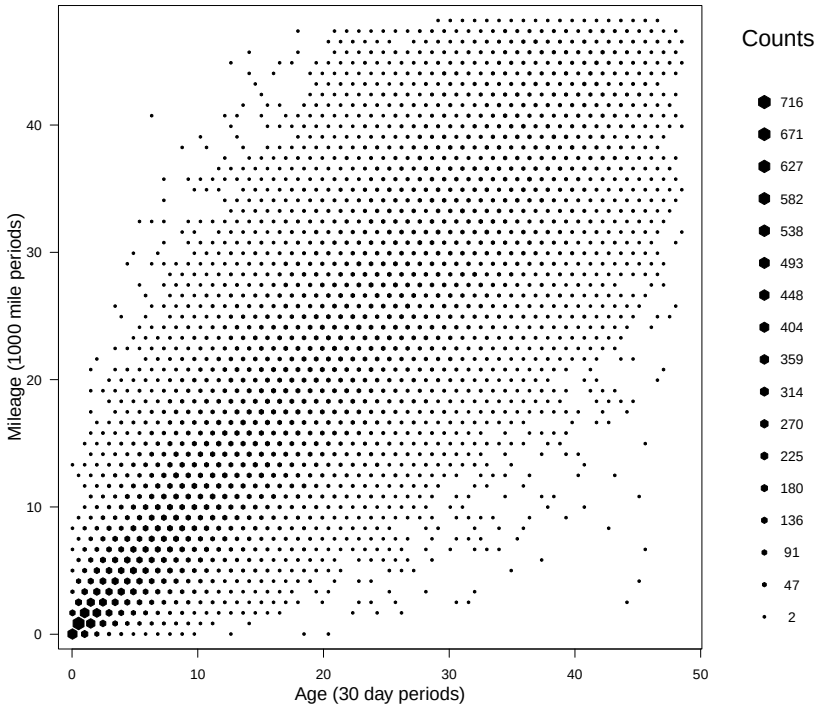

Figure D.6: Hex-bin graph of simulated dataset for 36 to 48 months

Appendix E

Notation

T_{max} Age Warranty Cutoff

U_{max} Mileage Warranty Cutoff

$T_{i,n}$ Vehicle i Age At Failure n

$U_{i,n}$ Vehicle i Mileage At Failure n

$T_{i,max}$ Maximum Vehicle i Age

$U_{i,max}$ Maximum Vehicle i Mileage

$C(x, y)$ Copula distribution function

$c(x, y)$ Copula density function

$h_C(x, y)$ Hazard function using Copula C

$\overline{\mathbf{R}}$ Extended real line

$\overline{\mathbf{R}}^2$ Extended real plane

$\psi(t)$ Copula generator function

$\psi^{-1}(t)$ Inverse copula generator function

$\hat{\Lambda}_T(t)$ Mean Cumulative Function For Age

$\hat{\Lambda}_U(u)$ Mean Cumulative Function For Mileage

Bibliography

[1] AKAIKE, H. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on 19*, 6 (1974), 716–723.

[2] ALI, M. M., MIKHAIL, N., AND HAQ, M. A class of bivariate distributions including the bivariate logistic. *Journal of Multivariate Analysis 8*, 3 (1978), 405 – 412.

[3] ANASTASIADIS, S., ANDERSON, B., AND CHUKOVA, S. Auto warranty and driving patterns. *Reliability Engineering and System Safety 116*, 0 (2013), 126 – 134.

[4] BAIK, J., MURTHY, D., AND JACK, N. Two-dimensional failure modeling with minimal repair. *Naval Research Logistics 51*, 3 (2004), 345–1362.

[5] BAIK, J., MURTHY, D., AND JACK, N. Erratum: Two-dimensional failure modeling with minimal repair which appeared in this journal of april 2004. *Naval Research Logistics 53*, 1 (2006), 115–116.

[6] BLISCHKE, W., AND MURTHY, D. *Product Warranty Handbook*. MARCEL DEKKER Incorporated, 1996.

[7] BLISCHKE, W. R., AND MURTHY, D. N. P. *Reliability: Modeling, Prediction, and Optimization*. John Wiley and Sons, Inc., 2000.

[8] CHRISTOZOV, D., CHUKOVA, S., AND ROBINSON, J. Automotive warranty data: stratification approach for estimating the mean cu-

mulative function. *International Journal of Product Development 12*, 3 (Jan 2010), 254–273.

[9] CHUKOVA, S., AND HIROSE, Y. Warranty data: An estimation of two-dimensional mean cumulative function. *Proceedings of the Asian International Workshop in Advance Reliability Modelling*, 513–520.

[10] CHUKOVA, S., AND ROBINSON, J. Estimating mean cumulative functions from truncated automotive warranty data. *Modern Statistical and Mathematical Methods in Reliability* (2005), 121–135.

[11] CLAYTON, D. G. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika 65*, 1 (1978), 141–151.

[12] FREDETTE, M., AND LAWLESS, J. F. Finite-horizon prediction of recurrent events, with application to forecasts of warranty claims. *Technometrics 49*, 1 (2007), 66–80.

[13] GUAN, T. H. Data mining in automotive warranty analysis. Master's thesis, Victoria University of Wellington, New Zealand, 2010.

[14] GUMBEL, É. J. Distributions des valeurs extrêmes en plusieurs dimensions. *Publ. Inst. Statist. Univ. Paris 9* (1960), 171–173.

[15] HU, X. J., AND LAWLESS, J. F. Estimation of rate and mean functions from truncated recurrent event data. *Journal of the American Statistical Association 91*, 433 (1996), 300–310.

[16] HUZURBAZAR, A. V. A censored data histogram. *Communications in Statistics - Simulation and Computation 34*, 1 (2005), 113–120.

[17] JOE, H. Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis 46*, 2 (1993), 262 – 282.

[18] JUSTEL, A., PEÑA, D., AND ZAMAR, R. A multivariate kolmogorov-smirnov test of goodness of fit. *Statistics and Probability Letters 35*, 3 (1997), 251–259.

[19] KAPLAN, E. L., AND MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*, 282 (1958), pp. 457–481.

[20] KLEYNER, A., AND SANDBORN, P. A warranty forecasting model based on piecewise statistical distributions and stochastic simulation. *Reliability Engineering and System Safety 88*, 3 (2005), 207–214.

[21] LI, D. X. On default correlation: A copula function approach. *The Journal of Fixed Income 4* (2000), 43–54.

[22] LIEBSCHER, E. Construction of asymmetric multivariate copulas. *Journal of Multivariate Analysis 99*, 10 (2008), 2234 – 2250.

[23] MORGENSTERN, D. Einfache beispiele zweidimensionaler verteilungen. *Mitteilungeblatt für mathematische statistik 8* (1956), 234–235.

[24] NELSEN, R. *An Introduction to Copulas (Second Edition)*. Springer Series in Statistics. Springer, 2010.

[25] RAI, B., AND SINGH, N. Forecasting automobile warranty performance in presence of 'maturing data' phenomena using multilayer perceptron neural network. *Journal of Systems Science and Systems Engineering 14* (2005), 159–176.

[26] RAI, B., AND SINGH, N. Automobile warranty forecasting using wavelet transform analysis and neural networks. *Proceedings of the International Conference on Manufacturing Science and Engineering* (2006), 711–716.

[27] SKLAR, A. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris 8* (1959), 229–231.

[28] WU, S. Warranty data analysis: A review. *Quality and Reliability Engineering International 28*, 8 (2012), 795–805.

[29] WU, S., AND AKBAROV, A. Support vector regression for warranty claim forecasting. *European Journal of Operational Research 213*, 1 (2011), 196 – 204.

[30] XU, K., XIE, M., TANG, L., AND HO, S. Application of neural networks in forecasting engine systems reliability. *Applied Soft Computing 2*, 4 (2003), 255 – 268.