DATA MINING IN AUTOMOTIVE WARRANTY ANALYSIS

by

HER GUAN TEO

A thesis
submitted to the 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
2010

Abstract

This thesis is about data mining in automotive warranty analysis, with an emphasis on modeling the mean cumulative warranty cost or number of claims (per vehicle). In our study, we deal with a type of truncation that is typical for automotive warranty data, where the warranty coverage and the resulting warranty data are limited by age and mileage. Age, as a function of time, is known for all sold vehicles at all time. However, mileage is only observed for a vehicle with at least one claim and only at the time of the claim. To deal with this problem of incomplete mileage information, we consider a linear approach and a piece-wise linear approach within a nonparametric framework. We explore the univariate case, as well as the bivariate case. For the univariate case, we evaluate the mean cumulative warranty cost and its standard error as a function of age, a function of mileage, and a function of actual (calendar) time. For the bivariate case, we evaluate the mean cumulative warranty cost as a function of age and mileage. The effect of reporting delay of claim and several methods for making prediction are also considered. Throughout this thesis, we illustrate the ideas using examples based on real data.

Acknowledgments

I would like to express my sincere thanks to my supervisor, Dr. Stefanka Chukova, for her patience and guidance. She has generously given her time and shared her knowledge with me during my study.

I would also like to thank Dr. Richard Arnold, Dr. Ray Brownrigg, Christopher Ball, and Michelle Li for their useful suggestions and comments. Besides, I am grateful to A/Prof. Megan Clark, all of my lecturers, and the staffs of the School of Mathematics, Statistics and Operations Research, Victoria University of Wellington.

Then, I would like to thank the Public Service Department of Malaysia for funding my study at the Victoria University of Wellington. I am thankful to Mrs. Siti Jalilah Abd Manap, Mrs. Hasmawati Kudin, and Mrs. Nurazrina Ramlan for their assistance and support throughout my study in New Zealand.

Lastly, I extend my special thanks to my parents, my family, and all of my friends for their continuous support and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A *product warranty* is an agreement between the seller and buyer, which establishes a liability between these parties in the event of failure. It specifies the expected performance of the product and the redress available to the buyer if a failure occurs. Here, the *seller* refers to the party responsible for assuring the warranty terms are met, and this is usually the manufacturer or retailer of the product. Then, the *buyer* is normally the ultimate paying consumer [Blischke and Murthy, 1996].

At present, product warranty plays an increasingly important role in the world of businesses and its uses are widespread. However, servicing claims is a cost to the seller. A seller may suffer great losses of millions or even billions of dollars if the true warranty cost is underestimated. On the other hand, overestimating the true warranty cost will make the the seller's products uncompetitive in the market. Therefore, extracting information from warranty data and using it in forecasting warranty cost, as well as in the research and development of the product, are of particular interest to the sellers or manufacturers.

In this thesis, we consider the data mining process in automotive warranty analysis, with an emphasis on modeling the mean cumulative warranty cost or number of claims (per vehicle). Automotive warranty generally guarantees free repairs subject to both age and mileage limits. In

the USA, till recently, the most common limit was 36 months or 36000 miles, whichever comes first. As sales records are retained, vehicle's age is known for all sold vehicles at all time. However, odometer readings are only recorded at the dealerships at the time of a claim, and then posted to the warranty database. Even though it is technically feasible to track mileage accumulation on all vehicles, this is currently not a common practice due to cost and privacy reasons. Therefore, automotive warranty analysis involves two variables (age and mileage), but the information of one of them (mileage) is incomplete [Chukova and Robinson, 2005].

In our study, we deal with the problem of incomplete mileage information explicitly by using a linear approach and a piece-wise linear approach within a nonparametric framework. We explore the univariate case, as well as the bivariate case. For the univariate case, we model the mean cumulative warranty cost and its standard error as a function of age, a function of mileage, and a function of the actual time. For the bivariate case, we model the mean cumulative warranty cost as a function of age and mileage. Besides, we also take into consideration the effect of reporting delay of claim. Several methods for making prediction are also considered.

Next, we provide the objectives and motivations of this research, an overview of warranty concepts, a review of relevant literature, and the organization of the thesis.

1.1 Research Objectives and Motivations

As mentioned earlier, underestimating warranty cost may result in significant losses to the seller, while overestimating warranty cost will make the seller's product uncompetitive. Thus, estimation of warranty cost from available warranty data has become an essential task for the seller. However, the incompleteness and uncleanliness of warranty data create some problems in estimating warranty cost. Therefore, one of the main objec-

tives of our research is to derive an effective and relatively simple method for estimating the warranty cost (more precisely, the mean cumulative warranty cost per vehicle) by utilizing warranty data. In addition, we are also interested in investigating the relationship between the variability of driving pattern and warranty cost, as well as extracting other notable trends in the warranty cost.

1.2 Overview of Warranty Concepts

Now, we briefly discuss some key warranty concepts and terminology.

1.2.1 The Role of Product Warranty

According to Blischke and Murthy [1996], a product warranty serves different purposes from different points of view. They summarized the roles of warranty to the buyer and seller (or manufacturer) as follows.

From the **buyer**'s point of view, the main role of a warranty is *protectional* as it provides a mean of redress in the event of failure. A warranty assures the buyer that a faulty item will either be repaired or replaced at no cost or at a reduced cost. Sometimes, a partial or full reimbursement will be given. The second role of warranty is *informational*, where the warranty terms act as a signal of quality to the buyers. Usually, buyers assume that a product with a longer warranty is often more reliable and has higher quality than one with a shorter warranty.

From the **seller**'s point of view, one of the main roles of warranty is also *protectional*. The warranty terms often specify the conditions of use for which the product is intended. Certain requirements of care and maintenance may also be included in the warranty terms. Consequently, if the warranty terms are violated, the seller only need to provide limited coverage or no coverage in the event of failure. The second role of warranty for seller is *promotional*. Warranty has been used as an effective promo-

tional tool and an instrument to compete with other sellers in the market, since buyers usually infer a more reliable product when a long warranty is provided.

1.2.2 Repairability and Degree of Repair

Repairability is one of the most important features that differentiate the items in warranty analysis. An item can be categorized in terms of repairability as a *repairable* item, a *non-repairable* item, or a *complex* item. In the event of failure, a repairable item (e.g. television) could be repaired or replaced, while a non-repairable item (e.g. frying pan) will require replacement. A complex item may have both repairable and non-repairable components, and it can be considered as a system. An example of complex item is a vehicle, which consists of many systems, subsystems and components.

Pham and Wang [1996] defined the **degree of repair** as a degree to which the ability of the item to function is restored after a repair. They categorized repairs in terms of their impact on the repaired item as follows:

- *Improved Repair* - A repair that makes the item better than when it was initially purchased.
- *Complete Repair* - A repair that resets the performance of the item, so that it operates as a new one upon restart.
- *Imperfect Repair* - A repair that contributes to some noticeable improvement of the item, which makes the performance and expected lifetime of the item as it was at an earlier age. It sets back the clock for the repaired item.
- *Minimal Repair* - A repair that brings the item from a 'down' state to an 'up' state and it has no effect on the performance of the item.

- *Worse Repair* - A repair that worsens the item, which makes the performance of the item as it would have been at a later age. It sets forward the clock for the repaired item.

- *Worst Repair* - A repair that destroys the item accidentally.

Some possible reasons that may lead to the occurrence of worse and worst repair are incorrect assessment of the faulty item, damage cause to the adjacent parts or subsystem of the item while repairing the faulty part, partial or incomplete repair of the faulty part, human errors such as incorrect adjustment and further damage of the item, replacement with faulty, incompatible, or low quality parts, etc. The repair that takes place usually depends on factors like warranty reserves, related costs, reliability and safety requirement, etc.

1.2.3 Dimensionality of Warranty

The *dimensionality* of a warranty refer to the number of variables specified in the warranty terms. The most common warranty is one-dimensional warranty, and the age of the product (or time from purchase) is the most commonly used warranty variable. Another possible variable is the amount of usage measured in miles or hours depending of the type of product. The warranty coverage expires once the preset limit of age or usage is exceeded. Then, *higher-dimensional warranty* involves two or more variables measuring the product service, with guaranteed service amount specified for each variables. Higher-dimensionality warranty is often associated with complex or multi-component item. An example of higher-dimensional warranty is a two-dimensional warranty based on age and mileage, which is commonly used in automotive industry. The warranty coverage expires when the age limit or mileage limit is exceeded, whichever occurs first. Higher-dimensional warranties with more than two warranty variables are used only in a few specialized applications, such as the air-

craft industry which may have age, flight hours, and number of flights as the warranty variables [Blischke and Murthy, 1996].

1.2.4 Base and Extended Warranty

Base warranty coverage is the original warranty coverage provided by the seller at no additional cost, as the cost of base coverage is included in the selling price of the product. The seller may also provide an option of purchasing an *extended* warranty coverage that comes into effect after the base coverage expires [Rai and Singh, 2009]. In our study, we do not consider extended warranty, i.e., claims that occur outside the base warranty coverage would not be considered.

1.2.5 Warranty Reserve

Warranty reserve is the fund, or sum of money, set aside by the seller for the purpose of servicing the warranty claims [Jayaraman, 2008]. The size of warranty reserve depends on the forecasted warranty cost, as well as the other factors like product quality and the number of products under coverage. Thus, accurate forecast of warranty cost is essential to the seller, so that the seller is able to design the warranty reserve and warranty program efficiently to ensure sound cashflow.

1.3 Review of Literature

Here, we provide a brief review of literature that is relevant to our research. Lately, the area of warranty analysis has been a very active research field, and the literature on warranty analysis is very rich. For instance, the general concepts of warranty analysis are given by Blischke and Murthy [1994] and Blischke and Murthy [1996], while Rai and Singh [2009] provide a discussion on the issues, strategies and methods related to the analysis of warranty data.

In our research, we follow closely the ideas of Chukova and Robinson [2005] and Christozov et al. [2008], which are based on the *robust estimator* proposed by Hu and Lawless [1996]. Chukova and Robinson [2005] adopted the robust estimator and a simple linear mileage accumulation model to estimate the number of vehicles that are eligible to generate a claim at a given age or mileage in estimating the mean cumulative warranty cost. Then, Christozov et al. [2008] extended the results of Chukova and Robinson [2005] by allowing for variation in the rate of mileage accumulation over a vehicle's lifetime. They relaxed the linearity assumption for mileage accumulation and proposed instead a piece-wise linear model with nodes occurring at the observed mileages corresponding to warranty claims.

Some other past literature relevant to our study are Lawless et al. [1995] which also dealt with the incomplete mileage information problem by using a simple linear mileage accumulation model, Lawless [1998] which took into account the bias due to reporting delay of claim in the analysis of warranty claims and costs, etc.

1.4 Organization of the Thesis

This thesis is organized as follows:

- In Chapter 2, we briefly discuss the warranty claim process, the structure of warranty database, and the characteristics of automotive warranty data.
- In Chapter 3, we introduce the data mining process.
- In Chapter 4, we present the summary of the datasets that we are going to use.
- In Chapter 5, we introduce the robust estimator which forms the basis of this thesis.

- In Chapter 6, we discuss the models for estimating mean cumulative warranty cost, which use a linear approach in modeling mileage accumulation.
- In Chapter 7, we extend the results of Chapter 6 by using a piecewise linear approach in modeling mileage accumulation.
- in Chapter 8, we propose a bivariate model for estimating mean cumulative warranty cost, using age and mileage as the warranty variables.
- In Chapter 9, we discuss several methods for predicting mean cumulative warranty cost.
- In Chapter 10, we present the conclusions, discussions, and directions of future research of our study.

Throughout the thesis, we will illustrate the ideas using examples based on real data. Microsoft Excel, Microsoft Access, and statistical programming language **R** are used to manage and extract the information we need from the raw data. Then, the models are built in statistical programming language **R**.

Chapter 2

Automotive Warranty Data

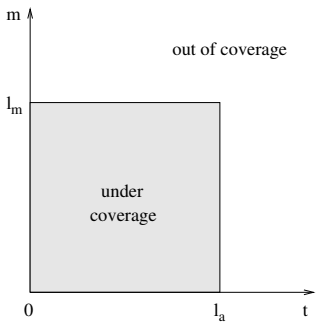

Figure 2.1: Warranty coverage region

Automotive warranties generally guarantee free repairs subject to both age and mileage limits. Let l_a and l_m denote the age limit and mileage limit of a warranty, respectively. In this thesis, we adopt the standard of $l_a = 36$ months and $l_m = 36000$ miles. (Note: Sometimes, we write $36K$ miles for 36000 miles, where K represents 1000.) Figure 2.1 shows the warranty coverage region $[0, l_a) \times [0, l_m)$ for an automotive warranty. Under this policy, the manufacturer agrees to repair a failed vehicle free of charge to the buyer up to age l_a or up to mileage l_m, whichever occurs first. Therefore, only repairs that occur within warranty coverage are included in the

warranty database. In this chapter, we discuss the warranty claim process, the structure of warranty database, and the characteristics of automotive warranty data.

2.1 Warranty Claim Process

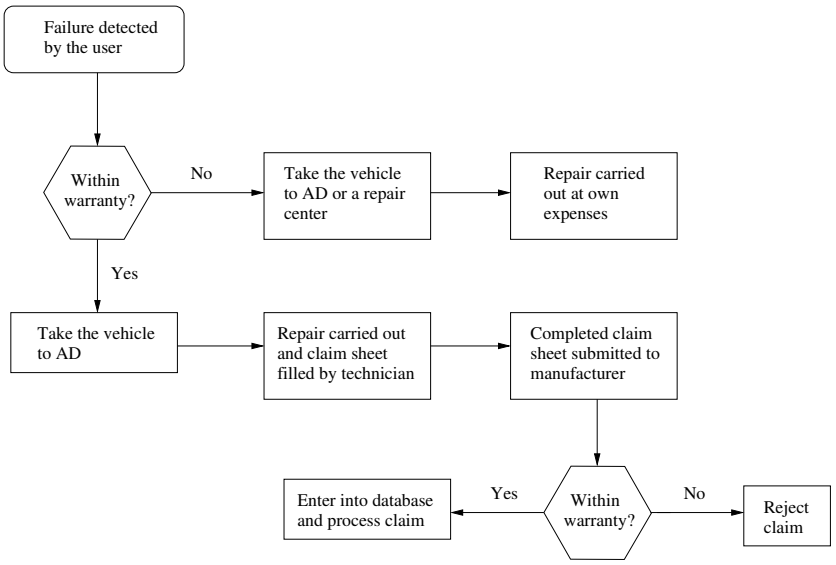

Figure 2.2: Warranty claim process

Figure 2.2 provides a brief overview of the warranty claim process (modified from Rai and Singh [2009]). The process begins with the user detecting the failure. If the vehicle is out of warranty coverage, then it will be taken to the authorized dealership (AD) or other repair center, and the repair is carried out at the owner's expenses. Usually, the latter will be the preferable choice due to the higher cost charged by the authorized dealership. On the other hand, if the vehicle is still within warranty coverage, then it will be taken to the authorized dealership. Then, repair is carried out by the technician and a claim sheet is filled. After that, the completed claim

sheet will be submitted to the manufacturer electronically. Subsequently, the manufacturer will check whether the claim is valid or not. If the claim is found to be invalid, then the claim is rejected. Otherwise, the data is entered into the database and the claim is processed. Note that we refer to the date for which the repair takes place as the *warranty date* and the date for which the claim is posted to the dataset as the *process date*.

2.2 Structure of Automotive Warranty Database

Next, we consider the general structure of an automotive warranty database. In an automotive warranty database, usually the vehicle records and claim records are kept in separate tables, which are linked together by "primary key". The primary key is usually the vehicle's identification number (VID). For illustration, Table 2.1 and Table 2.2 are the sample vehicle record table and sample claim record table, which are linked together by VID.

The purpose of using separate tables is to improve the structure of the database and the efficiency of data management. By using separate tables, there is no need to include all of the details of a vehicle along with each of its claim records. Hence, we can save time and effort to store redundant information. The cost and capacity of data storage can also be reduced. By avoiding redundant information, it will also be easier to update the database, and the chance of making errors can be decreased [Microsoft, n.d.].

VID	Prod. Date	Sale Date	Del. Date	Place of Sale	Owner	Address
000016	21/07/2000	31/07/2000	31/07/2000	New York	John Smith	...
000022	29/09/2000	10/10/2000	10/10/2000	New York	Lisa Johnston	...
000035	12/12/2000	01/01/2001	01/01/2001	Chicago	Mark Hansen	...
000044	31/12/2000	05/01/2001	06/01/2001	Orlando	Paul Lee	...
000064	08/01/2001	04/02/2001	04/02/2001	Memphis	Megan Hall	...
000089	15/01/2001	01/03/2001	05/03/2001	Dallas	Tony Adams	...

Table 2.1: Sample vehicle records

Claim ID	VID	Warr. Date	Proc. Date	Mileage	Cost	Type
000908	000016	16/02/2001	18/02/2001	8045	20	P001
000911	000022	22/04/2001	30/04/2001	8006	12	A022
010299	000044	25/04/2001	26/04/2001	6111	18	A014
011234	000016	10/05/2001	14/05/2001	10250	22	P007
011308	000146	10/11/2001	15/11/2001	7500	15	M122
011309	000146	10/11/2001	15/11/2001	7500	21	P012

Table 2.2: Sample claim records

In the vehicle record and claim record tables, usually each column corresponds to a data field or variable, and each row corresponds to a case or observation (i.e., a vehicle). Some common data fields for vehicle record include VID, production date, sale date, delivery date, place of sale, name and contact of the vehicle's owner, etc. On the other hand, some usual data fields for claim record include claim ID, VID, warranty date, process date, accumulated mileage, total cost of claims, type of claim, and so on.

Since vehicle is a complex system that consists of many sub-systems and components, and each sub-system or component has its own claims, the size of automotive warranty database is usually large and getting access to the information that we want can be a challenging task. Fortunately, using the techniques of data mining and appropriate computer software, we are able to extract the information needed for our analysis. We will discuss the process of data mining in the next chapter.

2.3 Characteristics of Automotive Warranty Data

Automotive warranty data is a form of field data that provides important information on the reliability of the vehicles. It helps the manufacturers to avoid the need and cost of running expensive laboratory tests. Besides, the information obtained from warranty data may be more reliable than that obtained through laboratory tests, because warranty data captures the actual usage behavior and environmental exposure of the vehicles that are

difficult to simulate in the laboratory [Rai and Singh, 2003]. Thus, if this data is well-analyzed, it will be very useful to the planning and decision making of the manufacturers.

However, automotive warranty data has two undesirable characteristics: *incompleteness* and *uncleanliness*. Due to these shortcomings, Rai and Singh [2009] described the warranty data as not always perfect for statistical analysis.

2.3.1 Incompleteness

First of all, as the automotive warranty dataset only includes those repairs that occur within the warranty coverage, no information beyond the warranty limits is available. This creates a form of incompleteness, and the automotive warranty data is said to be *right-truncated* at the age and mileage limits [Rai and Singh, 2009].

Secondly, the odometer readings are only observed at the time of the claim for a vehicle with at least one claim, whereas the mileage information for a vehicle without claims is completely unknown. These result in incomplete mileage information, which can be challenging to analyze. For instance, we may not know whether a vehicle is still under warranty coverage at a given time and hence is eligible to generate claims or not. A vehicle that is still within the age limit may have reached the mileage limit already and hence out of coverage, but this is unknown.

Besides, reporting delay of claim might occur, where there is a delay between the times when a warranty event (repair) occurs and when the claim is posted to the database [Chukova and Robinson, 2005]. As a result, the dataset used in analysis may not include all valid claims, and hence it is incomplete.

Note that automotive warranty data may also be incomplete due to missing warranty data, where vehicles are lost due to an accident or theft [Rai and Singh, 2009].

2.3.2 Uncleanliness

Automotive warranty data is often known to be messy and unclean for reasons such as inaccurate reporting of failures, unintended data entry error, etc.

Inaccurate reporting of failures is closely related to the type of failure mode. As according to Rai and Singh [2009], vehicle failures experienced by users can be mainly classified into two categories: hard failures and soft failures. Hard failures are those that make the vehicle inoperative until repaired (e.g., engine does not start), and they are usually reported immediately. On the other hand, soft failures are those that degrade performance but the vehicles can still be operated (e.g., unusual engine noise). For this type of failure, the users may report it immediately, or later at a convenient time. The latter gives rise to inaccurate reporting of failures and the failures are only known to have occurred prior to the reported time. As a result, the actual time and mileage at the time of failure is unknown. Such data is said to be *left-censored*.

Uncleanliness of automotive warranty data may also arise due to *data masking*, which is a practice for protecting data privacy and proprietary. Data masking prevents the exposure of sensitive or confidential information to the company's competitors and any unauthorized user. Some common data masking methods includes data substitution, numerical manipulation, data shuffling, etc [Wikipedia, n.d.]. As a result of data masking, automotive warranty data may sometimes contains some inconsistent or impossible data. For example, a vehicle may have accumulated unusually large mileage that is not compatible with its age, say 1500 miles while it is 1 day old.

Unclean warranty data may potentially hide the inherent failure patterns and such data can be misleading. Thus, it is essential to screen warranty data before undertaking any statistical analysis.

Chapter 3

Data Mining Process

Data mining is the process of extracting or "mining" knowledge from large amount of data [Han and Kamber, 2001]. Nowadays, due to the development of information technology and the growth of data collection, data mining has become an increasingly important instrument for extracting useful information from raw data. In 1996, *Cross-Industry Standard Process for Data Mining* (CRISP-DM) was developed by analysts from Daimler-Chrysler, SPSS, and NCR. CRISP-DM provides a standard process model that is non-proprietary and freely available for data mining [Chapman et al., 2000]. According to CRISP-DM, the process of data mining can be divided into the following six phases:

- Phase 1: Business understanding
- Phase 2: Data understanding
- Phase 3: Data preparation
- Phase 4: Modeling
- Phase 5: Evaluation
- Phase 6: Deployment

Figure 3.1 shows the CRISP-DM data mining model and the relationship between the six phases. Note that, if needed, we may have to return to the previous phase of the process.

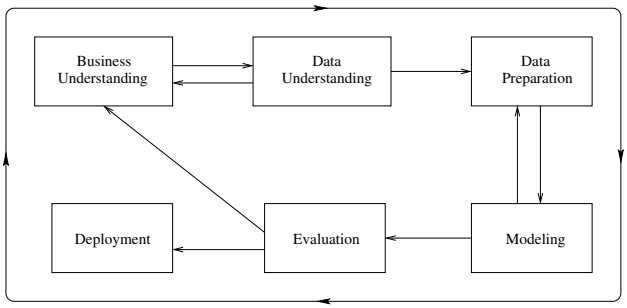

Figure 3.1: CRISP-DM data mining model

3.1 Phase 1: Business Understanding

In the first phase, we have to assess the business goals of the client and find out all of the details about the resources, constraints, requirements and other related factors. For instance, we may want to know about the human resources, technical support, data, computing software and hardware that are available, as well as the schedule of completion and the expectations of the client. Next, we have to determine the objectives of the data mining process based on the business goals and the requirements of the client, and then preparing a preliminary strategy to achieve these objectives [Chapman et al., 2000]. In an automotive industry, a business goal might be "reducing the number of claims and the associated costs". Then, the possible objectives of the data mining process would be:

- describing the patterns of claims,
- estimating the expected number of claims or cost of claims for a vehicle,

- predicting the number of claims or cost of claims at a given vehicle age,
- investigating the association between different types of claims,

and so on.

3.2 Phase 2: Data Understanding

In this phase, the first thing we need to do is to acquire the data. We can either collect the data on our own or make use of secondary data from certain sources. After that, we need to explore the data in order to understand the format, structure, and quantity of the data. Descriptive statistics of the data and its pictorial representation will be useful here. Then, we have to evaluate the quality of the data (e.g., errors, missing data, etc) and check whether the data is appropriate for our analysis.

3.3 Phase 3: Data Preparation

The aim of this phase is to prepare the initial raw data and convert it to a form suitable for the use in the subsequent phases. This phase usually consists of data selection, data cleaning, data creation and so on. First of all, we need to select the relevant subsets of data that are needed for our analysis, and clean the data. Data cleaning usually involves the dealing of missing data, outliers, and impossible data.

There are several methods to deal with missing data. The simplest method is to discard the records with missing values. Alternatively, we can replace the missing values with some constant such as the mean, mode, or any constant specified by the analyst. Another popular method is replacing the missing values with values determined by some statistical techniques like regression. For more details, see Han and Kamber [2001] and Larose [2005].

An outliers is an observations that is significantly different from the rest of the data and do not follow the general behavior of the data. For example, an outlier vehicle may have a mileage accumulation rate of 500 miles per day, while the rest of the vehicles have mileage accumulation rate between 20 and 50 miles per day. Outliers represent the extreme or rare objects in the population that can happen by chance, or they can be caused by measurement or typing errors. Usually, outliers will be eliminated from the data and a note is made in the report. However, outliers may sometimes be of particular interests and they may lead to interesting findings. Further analysis can be performed on outliers and this is referred to as *outlier mining*, which we will not cover here. See Han and Kamber [2001].

Often, there are also some impossible data or data that are not consistent with common sense in the data, which may occur due to human errors. For example, in an automotive warranty data, there may be a claim that occurs when the vehicle is 1 day old but has an accumulated mileage of 6000 miles. Another example will be a claim with warranty date before the sale date, i.e., repair took place before the vehicle is sold. Note that this type of claim may also be incurred by the retailer, where failure occurs at the retail outlet or during pre-sale testing by customer before the vehicle is sold. This type of claim is known as the *zero claim*, which we will not consider in our study. Impossible data should be corrected, if possible. Otherwise, we may consider removing these data. Note that, outliers or impossible data may also be caused by data masking (see Chapter 2.3.2).

After the data is cleaned, we may want to transform certain data variables or create some new variables (from the existing ones), if needed. For example, a new variable "the number of days from the sale date until the warranty date" will need to be calculated for modeling automotive warranty claims. Finally, the data are ready to be used in the subsequent phases.

3.4 Phase 4: Modeling

In this phase, we need to select and construct the appropriate model to extract the information "mined" under the data in order to accomplish the objectives in Phase 1. In our study, our main objective of the data mining process is to estimate the mean cumulative warranty cost (or number of claims) per vehicle for vehicle with typical driving pattern. We will introduce two types of nonparametric univariate models. The first model will adopt a linear approach in modeling mileage accumulation, while the second model will use a piece-wise linear approach. Besides, a bivariate model will also be introduced. We will discuss these models in details in the later sections.

3.5 Phase 5: Evaluation

In this phase, we need to evaluate the performance of the models, examine the results obtained, and review the whole data mining process. We will need to evaluate the accuracy of the estimates produced, the advantages and disadvantages of the models, and any relevant finding during the data mining process. We also have to check that whether our objectives are achieved. Finally, we need to make a decision on the use of the data mining results. If the results are appropriate, we may continue to the deployment phase. Otherwise, we may finish the data mining process here, or revise the data mining process, or start a new data mining process [Chapman et al., 2000].

3.6 Phase 6: Deployment

The final phase of the data mining process is deployment, i.e., to apply the selected model, produce a final report, and make use of the results [Chapman et al., 2000]. For warranty data mining, the results can be used

to design the warranty programs, to improve the product reliability and quality, etc.

Chapter 4

The Datasets

Throughout this thesis, we will illustrate the models introduced using real automotive warranty datasets. The main dataset that we used is called Dataset 2001. It contains the records from 22 May 2000 (the first sale date) up to 24 October 2003 for vehicles sold mainly in years 2000 and 2001. The other datasets are called Dataset 1998, Dataset 1999, Dataset 2000, and Dataset 2006. Datasets 1998 - 2000 contain the records for vehicles sold mainly in years 1998, 1999, and 2000 respectively. Dataset 2006 contains the records for vehicles sold mainly in years 2005 and 2006. In addition to the vehicle and claim records, Dataset 2006 also includes several odometer readings for each vehicle, which are not related to the time of making a claim.

Before we use the above datasets in our analysis, we need to clean these datasets. In our study, we remove the following data:

- Records of those vehicles with missing sale date. (Note: The whole records for these vehicles, including their claim records, are removed.)
- Claims occurred with decreasing accumulated mileage. (Note: Only the claim records are removed, not the whole records of the corresponding vehicles. Similarly for the following claims.)
- Claims occurred before sale date. These claims may be incurred by

the retailers or manufacturers. We will only consider those claims incurred by the buyers (or on behalf of the buyers), which occur on the sale date or after the sale date.

- Claims occurred outside the base warranty coverage (with $l_a = 36$ months and $l_m = 36000$ miles).

In addition, we also remove those claims that occurred at unusual or extreme usage rate, i.e., with a mileage accumulation rate (MAR) of less than 3 miles per day or more than 1000 miles per day. For example, if a claim occurs at a vehicle's age of 10 days and accumulated mileage of 15000 miles, then we say that this claim occurs at a mileage accumulation rate of 1500 miles per day, and we would remove it.

Table 4.1 shows the summary of Dataset 2001 up to four different "cuts" in time: 1 January 2001, 1 January 2002, 1 January 2003, and the actual "cut-off" date 24 October 2003 (the latest process date of a claim). We assume that this actual "cut-off" date is the "current date". By using this dataset, we will analyses the warranty cost on one major system of the vehicle, which is not identified here but referred to as "System P". For convenience, we will call the claims corresponding to this system as P-claims.

Note that the vehicle's age is given by the number of days from the sale date until the current date inclusively (i.e., vehicle's age $=$ current date $-$ sale date $+ 1$). So, a vehicle is 1 day old on its sale date. Then, the median mileage accumulation rate (MAR) is estimated based on the latest claim for those vehicles with at least one claim. It can be seen that the median mileage accumulation rate is around 40 miles per day over the four time cuts. This is more than the rate of 33 miles per day, which corresponds approximately to reaching the 36000-mile limit in three years. Thus, most vehicles leave the warranty coverage due to mileage accumulation.

Next, Table 4.2 shows the summary of Datasets 1998, 1999, 2000, and 2006. The start date refers to the first sale date in the dataset and the end date refers to the "cut-off" date of the dataset (the latest process date of a

claim). We will regard the end date as the current date for each of these datasets.

	01/01/2001	01/01/2002	01/01/12003	24/10/2003
Number of vehicles sold	16764	44761	44879	44890
Number of vehicles with claims	1669	12628	18882	21736
Number of claims	2554	25518	46820	59144
Total cost of claims	86122	751145	1464578	1953220
Number of vehicles with P-claims	48	508	974	1247
Total number of P-claims	50	579	1166	1510
Total cost of P-claims	14512	123292	222825	264539
Median vehicle age (days)	92	322	687	983
Median MAR (miles per day)	40	42	41	38
Median reporting delay (days)	11	8	8	8

Table 4.1: Summary of Dataset 2001

	Dataset 1998	Dataset 1999	Dataset 2000	Dataset 2006
Start date	03/09/1997	18/08/1998	20/08/1999	21/07/2005
End date	31/10/2003	31/10/2003	24/10/2003	22/10/2008
Number of vehicles sold	40048	44755	34807	7440
Number of vehicles with claims	23941	24307	18394	6527
Number of claims	77719	74736	54656	38534
Total cost of claims	2583429	2652406	1838376	1350269

Table 4.2: Summary of Datasets 1998, 1999, 2000 and 2006

Chapter 5

The Robust Estimator

In this section, we introduce the *robust estimator* proposed by Hu and Lawless [1996], which forms the basis of the models that we will introduce. The robust estimator can be used to estimate the rate and mean functions of a recurrent event process without any strong assumptions, and it is robust against the departure of Poisson assumption. Let $n_i(t)$ be the total warranty cost (or number of claims) for vehicle i at time t. It will be convenient but not restrictive to think of time as discrete, that is $t = 1, 2, \ldots$ Also, let $N_i(t)$ be the accumulated warranty cost (or number of claims) up to and including time t for vehicle i. Note that "time" here can be either age or mileage of the vehicle, not necessarily the calendar time.

Suppose M vehicles have been under observation and their records are included in the warranty database. Let τ_i, $i = 1, 2, \ldots, M$, be the time that vehicle i has been under observation, that is from the vehicle's sale date until the time it is out of warranty coverage or until the "cut-off" date of the dataset. We call τ_i the *observation time* of vehicle i. Its precise definition will depend on whether "time" is age, mileage, or actual time. Note that τ_i's may not be known exactly but only approximately. For example, for "time" is age case, the observation time is given by $\tau_i = \min(a_i, l_a, y_i)$, where a_i is the current age of vehicle i (on the "cut-off" date), l_a is the age warranty limit, and y_i is the age at which vehicle i exceeds (or would ex-

ceed) the mileage limit l_m. Usually, y_i is not known. For "time" is mileage case, the observation time is given by $\tau_i = \min(u_i(a_i), l_m, u_i(l_a))$, where $u_i(a)$ is the mileage for vehicle i at age a and l_m is the mileage warranty limit. Again, $u_i(a)$ is usually unknown. Thus, in both cases, we need some measures of the accumulated mileage in order to estimate τ_i [Hu and Lawless, 1996].

Now, let $\hat{\Lambda}(t)$ be the estimator of $\Lambda(t) = E[N_i(t)]$, the mean cumulative warranty cost (or number of claims). In discrete time case, the incremental rate function is $\lambda(t) = \Lambda(t) - \Lambda(t-1)$ with an initial condition $\Lambda(0) = 0$. Let $\delta_i(t)$ be the indicator of whether vehicle i is under observation at time t and hence eligible to generate a claim. For "time" is age case and "time" is mileage case, we have $\delta_i(t) = I(\tau_i \geq t)$. Then, the total warranty cost (or number of claims) at time t for all M vehicles is given by

$$n(t) = \sum_{i=1}^{M} \delta_i(t) n_i(t). \tag{5.1}$$

It can be noted that $\delta_i(t)$ may be unknown in some cases, but the products $\delta_i(t)n_i(t)$ is always known and is available in the database. Hence, $n(t)$ is also always known. Suppose the observation process is independent of the event (claim) process, then the rate function can be estimated by

$$\hat{\lambda}(t) = \frac{n(t)}{MP(t)}, \tag{5.2}$$

where $P(t)$ is the probability that a vehicle is eligible to generate a claim at time t. This is the robust estimator proposed by Hu and Lawless [1996], and they assumed $P(t)$ is known. But, $P(t)$ is usually unknown and needs to be estimated.

As stated in the paper of Chukova and Robinson [2005], it will be more convenient to think in terms of $M(t) = MP(t)$, that is the number of vehicles that are eligible to generate a claim at time t. Then, from Eq. (5.2), we

get

$$\hat{\lambda}(t) = \frac{n(t)}{M(t)}, \tag{5.3}$$

and consequently the associated mean cumulative function estimator is

$$\hat{\Lambda}(t) = \sum_{s=1}^{t} \hat{\lambda}(s), \quad t = 1, 2, \ldots, \tau_{\max}, \tag{5.4}$$

where $\tau_{\max} = \max(\tau_i)$ for $i = 1, 2, \ldots, M$. Under mild conditions and assuming known $M(t)$, Hu and Lawless [1996] proved the asymptotic normality of $\hat{\Lambda}(t)$, with a standard error estimated by the square root of

$$\widehat{Var}[\hat{\Lambda}(t)] = \sum_{i=1}^{M} \left\{ \sum_{s=1}^{t} \left[\frac{\delta_i(s) n_i(s)}{M(s)} - \frac{\hat{\lambda}(s)}{M} \right] \right\}^2. \tag{5.5}$$

Eq. (5.5) can be written as

$$\widehat{Var}[\hat{\Lambda}(t)] = \sum_{i=1}^{M_1} \left[\sum_{s=1}^{t} \frac{\delta_i(s) n_i(s)}{M(s)} - \frac{\hat{\Lambda}(t)}{M} \right]^2 + M_2 \left[\frac{\hat{\Lambda}(t)}{M} \right]^2, \tag{5.6}$$

where M_1 is the number of vehicles that have had at least one claim (and therefore have some reported mileage history) and M_2 is the number of vehicle without claims, such that $M_1 + M_2 = M$. This expression helps to improve computational efficiency. Note that, unless specified otherwise, the 95% confidence intervals mentioned in this thesis are evaluated using Eq. (5.5) or Eq. (5.6).

Chukova and Robinson [2005] computed $\hat{M}(t)$ as an estimate of $M(t)$ from the warranty data and substituted $\hat{M}(t)$ into Eq. (5.3) and Eq. (5.5) to obtain $\hat{\lambda}(t)$, $\hat{\Lambda}(t)$, and the standard error of $\hat{\Lambda}(t)$. In our study, we follow the same approach using different estimators of $M(t)$, which will be discussed later. Note that the validity of Eq. (5.5), with $M(t)$ replaced by $\hat{M}(t)$, is not proven yet. Thus, we also consider the use of bootstrap method in

estimating the standard error of $\hat{\Lambda}(t)$.

5.1 Properties of the Robust Estimator

In the paper of Hu and Lawless [1996], the proofs for the expected value and variance of the robust estimator are given in sketch. Here, we show in detail the proofs of these properties.

5.1.1 Expected Value of the Robust Estimator

Let us consider the expected values of $\hat{\lambda}(t)$ and $\hat{\Lambda}(t)$. Assume that the observation process is independent of the event (claim) process and $P(t)$ is known. Let $E[n_i(t)] = \lambda(t)$, then $E[\delta_i(t)n_i(t)] = P(t)\lambda(t)$ and

$$\begin{aligned} E[\hat{\lambda}(t)] &= E\left[\frac{n(t)}{MP(t)}\right] \\ &= E\left[\frac{\sum_{i=1}^{M}\delta_i(t)n_i(t)}{MP(t)}\right] \\ &= \frac{MP(t)\lambda(t)}{MP(t)} \\ &= \lambda(t). \end{aligned}$$

Consequently, we have

$$E[\hat{\Lambda}(t)] = E\left[\sum_{s=1}^{t}\hat{\lambda}(s)\right] = \sum_{s=1}^{t}\lambda(s) = \Lambda(t).$$

So, $\hat{\lambda}(t)$ and $\hat{\Lambda}(t)$ are the unbiased estimators for $\lambda(t)$ and $\Lambda(t)$, respectively.

5.1.2 Variance of the Robust Estimator

Next, we consider the variance of $\hat{\Lambda}(t)$. Rewrite $\hat{\Lambda}(t)$ as

$$\hat{\Lambda}(t) = \frac{1}{M}\sum_{i=1}^{M} X_i(t), \tag{5.7}$$

where

$$X_i(t) = \sum_{s=1}^{t} \frac{\delta_i(s)n_i(s)}{P(s)} \quad i = 1, 2, \ldots, M, \tag{5.8}$$

are i.i.d. random variables. Note that $E[X_i(t)] = \Lambda(t)$ since

$$E[X_i(t)] = E\left[\sum_{s=1}^{t} \frac{\delta_i(s)n_i(s)}{P(s)}\right] = \sum_{s=1}^{t} \lambda(s) = \Lambda(t)$$

Now, let denote the covariance of $n_i(u)$ and $n_i(v)$ as $c(u, v) = cov[n_i(u), n_i(v)]$ for $u, v = 1, 2, \ldots, \tau_{\max}$. Then, we have

$$E[n_i(u)n_i(v)] = c(u, v) + \lambda(u)\lambda(v), \tag{5.9}$$

since $c(u, v) = cov[n_i(u)n_i(v)] = E[n_i(u)n_i(v)] - \lambda(u)\lambda(v)$. Consequently, the covariance of $X_i(s)$ and $X_i(t)$ is

$$\begin{aligned} cov[X_i(s), X_i(t)] &= cov\left[\sum_{u=1}^{s} \frac{\delta_i(u)n_i(u)}{P(u)}, \sum_{v=1}^{t} \frac{\delta_i(v)n_i(v)}{P(v)}\right] \\ &= \sum_{u=1}^{s}\sum_{v=1}^{t} E\left[\frac{\delta_i(u)\delta_i(v)}{P(u)P(v)} n_i(u)n_i(v)\right] - \Lambda(s)\Lambda(t) \\ &= \sum_{u=1}^{s}\sum_{v=1}^{t} \left\{E\left[\frac{\delta_i(u)\delta_i(v)}{P(u)P(v)} n_i(u)n_i(v)\right] - \lambda(u)\lambda(v)\right\} \end{aligned} \tag{5.10}$$

Then, the variance of $\hat{\Lambda}(t)$ is given by

$$\begin{aligned} Var[\hat{\Lambda}(t)] &= \frac{1}{M^2} Var\left[\sum_{i=1}^{M} X_i(t)\right] \\ &= \frac{1}{M^2} \times M \times cov[X_i(t), X_i(t)] \\ &= \frac{1}{M} \sum_{u=1}^{t} \sum_{v=1}^{t} \left\{ E\left[\frac{\delta_i(u)\delta_i(v)}{P(u)P(v)} n_i(u) n_i(v)\right] - \lambda(v)\lambda(v) \right\} \\ &= \frac{1}{M} \sum_{u=1}^{t} \sum_{v=1}^{t} \frac{1}{P(u)P(v)} \left\{ E[\delta_i(u) n_i(u) \delta_i(v) n_i(v)] - P(u)\lambda(u)P(v)\lambda(v) \right\} \\ &= \frac{1}{M} \sum_{u=1}^{t} \sum_{v=1}^{t} \frac{1}{P(u)P(v)} cov[\delta_i(u) n_i(u), \delta_i(v) n_i(v)]. \end{aligned} \tag{5.11}$$

Finally, by estimating $cov[\delta_i(u)n_i(u), \delta_i(v)n_i(v)]$ with the following

$$\widehat{cov}[\delta_i(u)n_i(u), \delta_i(v)n_i(v)] = \frac{1}{M} \sum_{i=1}^{M} \left[\delta_i(u)n_i(u) - \frac{n(u)}{M}\right] \left[\delta_i(v)n_i(v) - \frac{n(v)}{M}\right], \tag{5.12}$$

we obtain Eq. (5.5)

$$\widehat{Var}[\hat{\Lambda}(t)] = \sum_{i=1}^{M} \left\{ \sum_{s=1}^{t} \left[\frac{\delta_i(s) n_i(s)}{M(s)} - \frac{\hat{\lambda}(s)}{M} \right] \right\}^2$$

after rearranging. Note that this is the proof of Eq. (5.5) with known $M(t)$. The validity of replacing $M(t)$ by $\hat{M}(t)$ in Eq. (5.5) is not proven yet.

Chapter 6

Modeling Mileage Accumulation: Linear Approach

In this chapter, we briefly review the model suggested by Chukova and Robinson [2005], which is based on the robust estimator proposed by Hu and Lawless [1996]. We call this model the CR-Model. The CR-Model makes an assumption that vehicles accumulate mileage approximately linearly with their age. In this model, an estimate of $M(t)$, $\hat{M}(t)$, is computed and substituted into Eq. (5.3) and Eq. (5.5) to obtain $\hat{\lambda}(t)$, $\hat{\Lambda}(t)$, and the standard error of $\hat{\Lambda}(t)$. We will reproduce some of the results in Chukova and Robinson [2005] by using the same dataset, but with different approaches in preparing the data (see Chapter 4) and with different computing softwares. In our study, we use statistical programming language **R**, while Chukova and Robinson [2005] used *Mathematica*. In this chapter, we will also propose a new model for estimating the mean cumulative warranty cost per vehicle in the actual time case. Besides, we will also consider the use of bootstrap method in estimating the standard error of $\hat{\Lambda}(t)$. Note that, from now on, we shall use t to denote age, m to denote mileage, and x to denote the actual (calendar) time.

6.1 Overview of the CR-Model

6.1.1 "Time" is Age Case

Firstly, we consider the case where "time" is the age of the vehicle. Suppose we ignore the withdrawals from warranty coverage due to exceeding the mileage limit. Then, the estimate of the number of vehicles eligible to generate a claim at the target age t, $t \leq l_a$, is simply the number of vehicles age t or older, that is

$$\hat{M}(t) = \sum_{i=1}^{M} I(a_i \geq t), \tag{6.1}$$

where a_i is the current age of vehicle i (on the "cut-off" date). This is the unadjusted estimator of $M(t)$. To get the true warranty claim rate, we need to adjust for withdrawals from warranty coverage due to exceeding the mileage limit. *Note that all adjustments made here and later will always be to* $\hat{M}(t)$.

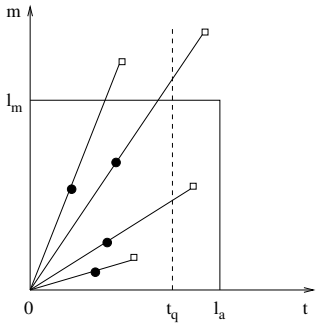

Figure 6.1: "Time" is age case

Recall that the observation time is given by $\tau_i = \min(a_i, l_a, y_i)$, where y_i is the age at which the vehicles exceeds (or would exceed) the mileage limit. Since odometers are not monitored continuously, y_i is usually not known even for vehicles with claims. Thus, for a vehicle with at least one

claim, we simply extrapolate y_i linearly using the age and mileage at the time of the most recent claim. Let $r_i = \beta_i/\alpha_i$, where α_i and β_i are the age and mileage of the vehicle at the latest claim respectively. Then, r_i is the estimated mileage accumulation rate (in miles per day) for vehicle i. Subsequently, at the target age t, vehicle i will contribute to $\hat{M}(t)$ if it is old enough and if its mileage at age t is estimated to have been within the mileage limit l_m. Thus, the contribution of vehicle i to $\hat{M}(t)$ is

$$I(a_i \geq t) I\left(r_i \leq \frac{l_m}{t}\right).$$

Figure 6.1 illustrates the above idea graphically using four hypothetical vehicles. In this figure,

- the large square represents the warranty coverage region,
- the little black circles represent the age and mileage for the latest claim of these vehicles, and
- the little squares represent the extrapolated mileages for these vehicles at their current age (on the "cut-off" date).
- the straight lines represent the trajectories of the vehicles.

It can be seen that two of the vehicles are older than the target age t_q. But, one of them is estimated to leave the warranty coverage due to exceeding the mileage limit before age t_q, and hence it will not contribute to the adjusted (for mileage) value of $\hat{M}(t_q)$.

Now, we need to consider those vehicles that have not experienced a claim. By using the information on the vehicles with claims, we can construct an empirical distribution function for mileage accumulation rate as follows

$$\hat{F}(r) = \frac{1}{M_1} \sum_{i=1}^{M_1} I(r_i \leq r), \tag{6.2}$$

where M_1 is the number of vehicles with claims. Consequently, the probability that a typical vehicle remains in warranty coverage at age t is $\hat{F}(\frac{l_m}{t})$, and hence the contribution to $\hat{M}(t)$ for a vehicle without claims is

$$I(a_i \geq t)\hat{F}\left(\frac{l_m}{t}\right).$$

Every claim has a reporting delay associated with it (perhaps zero), which results in undercounting the number of claims, and the corresponding claim rate and cost. Reporting delay can be estimated by taking the difference between the warranty date and the processing date. Then, we can derive an empirical distribution function for reporting delay as follows

$$\hat{G}(d) = \frac{1}{N_c}\sum_{j=1}^{N_c} I(d_j \leq d), \tag{6.3}$$

where d_j is the delay for claim j and N_c is the total number of claims in the database. If vehicle i, currently at age a_i, had experienced a claim at age $t \leq a_i$, there is a time period of length $(a_i - t)$ for the claim to be posted to the database. The estimated probability that the claim is posted in that length of time is $\hat{G}(a_i - t)$. Subsequently, the contribution to $\hat{M}(t)$ for vehicle i becomes

$$I(a_i \geq t)\hat{G}(a_i - t).$$

This adjustment results in an effect of decreasing $\hat{M}(t)$, and hence increasing the rate function $\hat{\lambda}(t)$.

Note that, by assuming independence between the process of generating reporting delay and that generating mileage accumulation, the adjustment for reporting delay can also be used along with the adjustment for mileage. Table 6.1 shows the contribution of a vehicle to $\hat{M}(t)$ in each case.

Case	Vehicle with claims	Vehicle without claims
Unadjusted	$I(a_i \geq t)$	$I(a_i \geq t)$
Adjusted for mileage	$I(a_i \geq t)I\left(r_i \leq \frac{l_m}{t}\right)$	$I(a_i \geq t)\hat{F}\left(\frac{l_m}{t}\right)$
Adjusted for delay	$I(a_i \geq t)\hat{G}(a_i - t)$	$I(a_i \geq t)\hat{G}(a_i - t)$
Adjusted for mileage & delay	$I(a_i \geq t)I\left(r_i \leq \frac{l_m}{t}\right)\hat{G}(a_i - t)$	$I(a_i \geq t)\hat{F}\left(\frac{l_m}{t}\right)\hat{G}(a_i - t)$

Table 6.1: Contribution to $\hat{M}(t)$ for vehicle i at target age t

Example

In this example, we estimate the mean cumulative cost of P-claims per vehicle, $\Lambda(t)$, up to the age limit of 36 months or equivalently 1095 days (365 days per year) by using Dataset 2001. Note that, in the construction of the empirical distribution function for mileage accumulation rate, $\hat{F}(r)$, we had used the mileage information from all claims, not just those for System P.

Figure 6.2 shows the unadjusted $\hat{\Lambda}(t)$ and Figure 6.3 shows the adjusted for mileage $\hat{\Lambda}(t)$, along with the corresponding 95% confidence intervals (CI). Then, Figure 6.4 illustrates the effect of the adjustment for withdrawals from warranty coverage due to exceeding the mileage limit of 36000 miles. Without further analysis, the bend in the unadjusted curve could be due to a fall in the warranty claim rate with age. But, the adjusted for mileage curve indicates that the "true" rate is not decreasing (or only decreasing slightly), and the bend is caused by vehicles leaving coverage due to mileage.

It can be observed that the difference between the unadjusted curve and the adjusted for mileage curve increases as the vehicle's age approaches the age limit, and the latter goes across the unadjusted 95% confidence interval when the vehicle's age is about 900 days. This indicates that the adjustment for mileage is becoming more statistically significant as the

vehicle's age increases. Based on the empirical distribution function for mileage accumulation rate, $\hat{F}(r)$, we estimate that approximately 2% of the vehicles in our dataset reach the mileage limit in one year, 11% in one and a half years, 29% in two years, 47% in two and a half years, and 62% in three years.

Next, Figure 6.5 illustrates the effect of adjustment for reporting delay. It can be seen that this adjustment is not substantial. It only has very little effects at older vehicle's age ($\geq$ 800 days), as shown in Figure 6.6. Then, Figure 6.7 illustrates the effect of simultaneous adjustment for both mileage and reporting delay. This figure shows a similar pattern as Figure 6.4, and we may conclude that the effect of this adjustment is mainly due to mileage.

Note that the above results replicate the results given by Chukova and Robinson [2005].

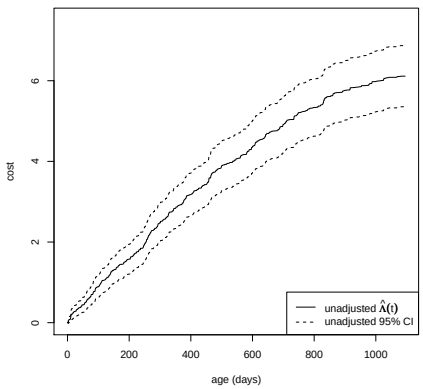

Figure 6.2: Unadjusted $\hat{\Lambda}(t)$

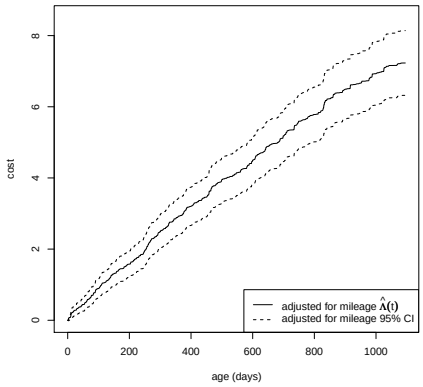

Figure 6.3: Adjusted for mileage $\hat{\Lambda}(t)$

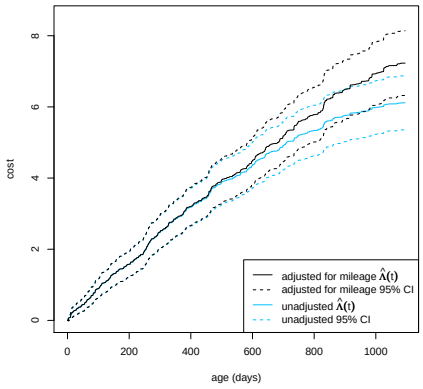

Figure 6.4: Unadjusted $\hat{\Lambda}(t)$ and adjusted for mileage $\hat{\Lambda}(t)$

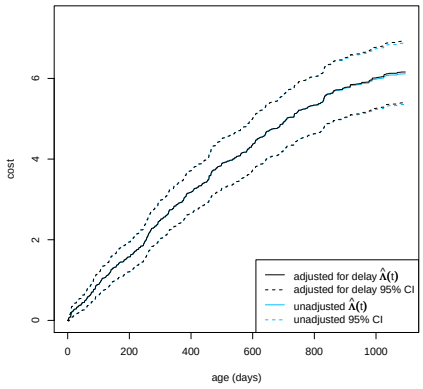

Figure 6.5: Unadjusted $\hat{\Lambda}(t)$ and adjusted for delay $\hat{\Lambda}(t)$

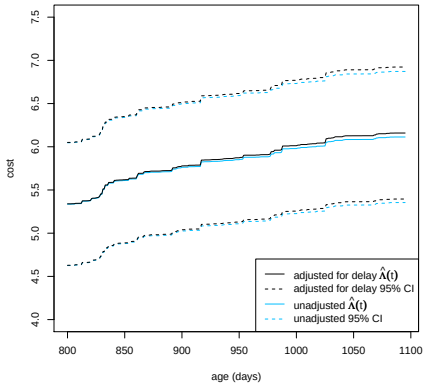

Figure 6.6: Unadjusted $\hat{\Lambda}(t)$ and adjusted for delay $\hat{\Lambda}(t)$, for $t \geq 800$ days

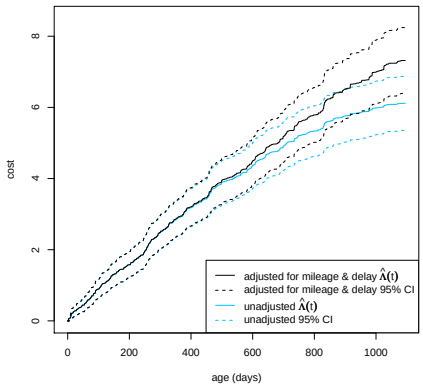

Figure 6.7: Unadjusted $\hat{\Lambda}(t)$ and adjusted for mileage & delay $\hat{\Lambda}(t)$

6.1.2 "Time" is Mileage Case

In this section, analogously to the "time" is age case, we consider the case where "time" is the mileage of the vehicle. Let $\hat{M}(m)$ be the number of vehicles eligible to generate a claim at the target mileage m, $m \leq l_m$. Similar

to the "time" is age case, we consider the same linear mileage accumulation model, and all adjustments will be made to $\hat{M}(m)$.

Firstly, we consider the unadjusted case, where we ignore withdrawals from warranty coverage due to exceeding the age limit l_a. For a vehicle with at least one claim, its current mileage can be estimated by $a_i r_i$, where a_i is the current age (on the "cut-off" date) and r_i is the mileage accumulation rate based on the latest claim. Hence, the contribution to $\hat{M}(m)$ for the vehicle can be estimated by

$$I\left(r_i \geq \frac{m}{a_i}\right).$$

Then, the contribution for a vehicle with no claims is given by

$$1 - \hat{F}\left(\frac{m}{a_i}\right),$$

where $\hat{F}(r)$ is the empirical distribution function for mileage accumulation rate as given in Section 6.1.1. The above expression represents the likelihood that the vehicle has reached m miles. Subsequently, to adjust for withdrawals from warranty coverage due to exceeding the age limit l_a, we simply replace a_i in the unadjusted case by $\min(a_i, l_a)$, the minimum of the vehicle's current age and the warranty age limit.

As in Figure 6.1, Figure 6.8 illustrates the idea for the "time" is mileage case. It can be seen that three of the vehicles are estimated to have exceeded the target mileage m_q at their current ages. But, one of them is estimated to have exceeded the age limit before mileage m_q, and hence it will not contribute to the adjusted (for age) value of $\hat{M}(m_q)$.

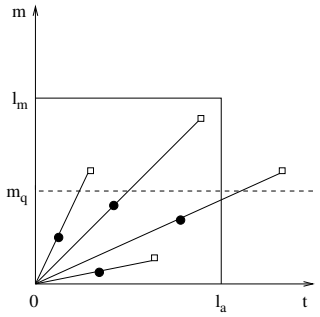

Figure 6.8: "Time" is mileage case

For the "time" is mileage case, the adjustment for reporting delay of claim is more complicated. Firstly, we consider the case for vehicles with at least one claim. If vehicle i, now at age a_i, has experienced a claim at mileage m, its age at mileage m is estimated to be $(\frac{m}{r_i})$ and there is a time period of length $(a_i - \frac{m}{r_i})$ for the claim to be posted. Therefore, to adjust for reporting delay, we multiply the initial contribution by $\hat{G}(a_i - \frac{m}{r_i})$, where $\hat{G}(d)$ is the empirical distribution function for reporting delay as given in Section 6.1.1. Next, for vehicles with no claims, no mileage accumulation rate is available and so we have to consider all previous ages. Suppose mileage m is attained at age j, where $j \leq a_i$, then the mileage accumulation rate is $(\frac{m}{j})$. Let us define the empirical probability mass function of the mileage accumulation as

$$\hat{f}(m,j) = \begin{cases} 1 - \hat{F}(m) & \text{if } j = 1; \\ \hat{F}\left(\frac{m}{j-1}\right) - \hat{F}\left(\frac{m}{j}\right) & \text{if } j = 2, 3, \ldots \end{cases} \tag{6.4}$$

If vehicle i had experienced a claim at age j, there is a time period of $(a_i - j)$ for the claim to be posted. Consequently, the contribution to $\hat{M}(m)$ for vehicle i becomes

$$\sum_{j=1}^{a_i} \hat{f}(m,j)\hat{G}(a_i - j).$$

Table 6.2 shows the contribution of a vehicle to $\hat{M}(m)$ in each case. Assuming independence between the process of generating reporting delay and that generating mileage accumulation, the simultaneous adjustment for both withdrawals due to age and reporting delay is shown in the last row of Table 6.2.

Case	Vehicle with claims	Vehicle without claims
Unadjusted	$I(r_i \geq \frac{m}{a_i})$	$1 - \hat{F}(\frac{m}{a_i})$
Adjusted for age	$I\left(r_i \geq \frac{m}{\min(a_i, l_a)}\right)$	$1 - \hat{F}\left(\frac{m}{\min(a_i, l_a)}\right)$
Adjusted for delay	$I\left(r_i \geq \frac{m}{a_i}\right) \hat{G}\left(a_i - \frac{m}{r_i}\right)$	$\sum_{j=1}^{a_i} \hat{f}(m, j)\hat{G}(a_i - j)$
Adjusted for age & delay	$I\left(r_i \geq \frac{m}{\min(a_i, l_a)}\right) \hat{G}\left(a_i - \frac{m}{r_i}\right)$	$\sum_{j=1}^{\min(a_i, l_a)} \hat{f}(m, j)\hat{G}(a_i - j)$

Table 6.2: Contribution to $\hat{M}(m)$ for vehicle i at target mileage m

Example

In this example, we estimate the mean cumulative cost of P-claims per vehicle, $\Lambda(m)$, up to the mileage limit of 36000 miles by using Dataset 2001.

Figure 6.9 shows the unadjusted $\hat{\Lambda}(m)$ and Figure 6.10 shows the adjusted for age $\hat{\Lambda}(m)$, along with the corresponding 95% confidence intervals (CI). Then, Figure 6.11 and Figure 6.12 illustrate the effect of the adjustment for withdrawals from warranty coverage due to exceeding the age limit of 36 months. Unlike the adjustment for mileage in the previous example, this adjustment for age has little impacts here. This is because the adjustment for age does not begin until the oldest vehicle exceeds the age limit, which is three years from the first sale in our dataset. Also, there is relatively few vehicles, only about 38%, that are estimated to leave warranty coverage due to age (based on the empirical distribution function for mileage accumulation rate).

Next, Figure 6.13 shows the effect of the adjustment for reporting delay

and Figure 6.14 illustrates the effect of simultaneous adjustment for both age and reporting delay. Both of these adjustments also have no significant effects.

Note that the above results replicate the results given by Chukova and Robinson [2005].

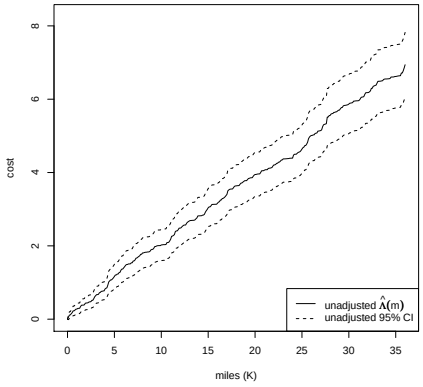

Figure 6.9: Unadjusted $\hat{\Lambda}(m)$

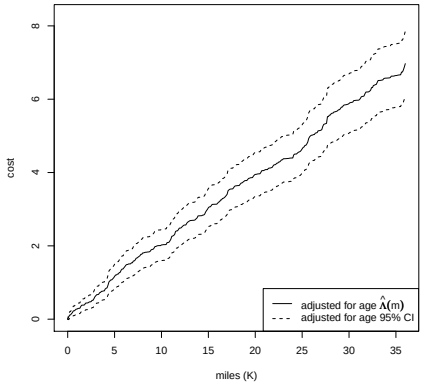

Figure 6.10: Adjusted for age $\hat{\Lambda}(m)$

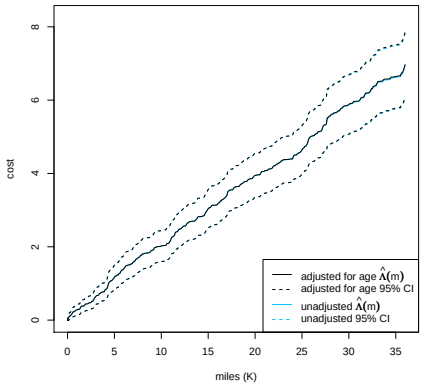

Figure 6.11: Unadjusted $\hat{\Lambda}(m)$ and adjusted for age $\hat{\Lambda}(m)$

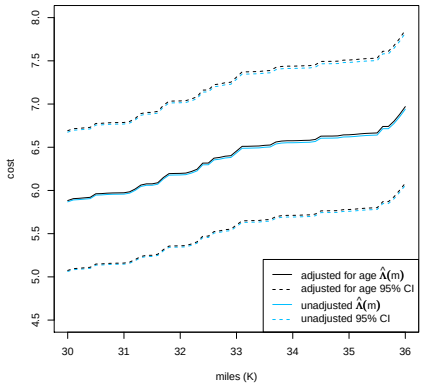

Figure 6.12: Unadjusted $\hat{\Lambda}(m)$ and adjusted for age $\hat{\Lambda}(m)$, for $m \geq 30K$ miles

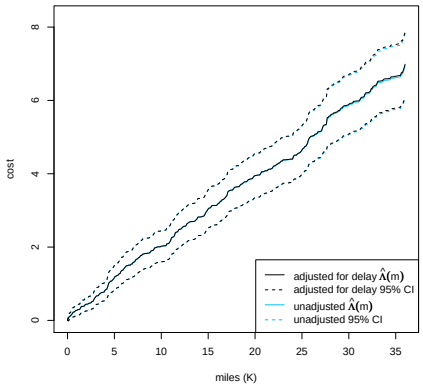

Figure 6.13: Unadjusted $\hat{\Lambda}(m)$ and adjusted for delay $\hat{\Lambda}(m)$

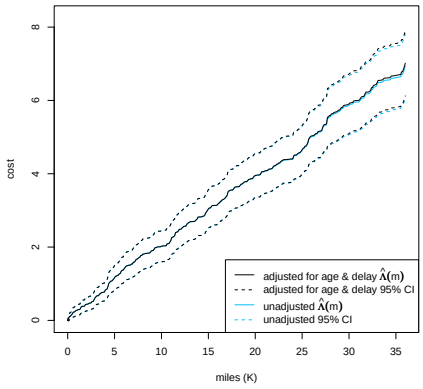

Figure 6.14: Unadjusted $\hat{\Lambda}(m)$ and adjusted for age & delay $\hat{\Lambda}(m)$

6.2 New Model: Actual Time Case

Sometimes, it may be of interest to analyze the data with respect to the actual (calendar) time, instead of usage measures like age or mileage. For example, in designing the warranty programs and warranty reserves, we may want to know the mean cumulative warranty cost (or number of claims) for a batch of vehicles sold after a certain date. Here, we develop a new model for estimating the mean cumulative warranty cost per vehicle in the actual time case, $\Lambda(x)$. Again, all adjustments will always be made to $\hat{M}(x)$.

Let X denote the current time (the "cut-off" date). Suppose we ignore the withdrawals from warranty coverage due to mileage, then the estimate of the number of vehicles eligible to generate a claim at the target time x, $x \leq X$, is simply the number of vehicles sold before or at time x and still within the age limit, i.e.,

$$\hat{M}(x) = \sum_{i=1}^{M} I(s_i \leq x) I(z_i \leq l_a), \tag{6.5}$$

where s_i is the sale date and z_i is the age at the target time of vehicle i. This is the unadjusted estimator of $M(x)$ (or we can say that this is the adjusted for age estimator, since we have taken into account the age limit).

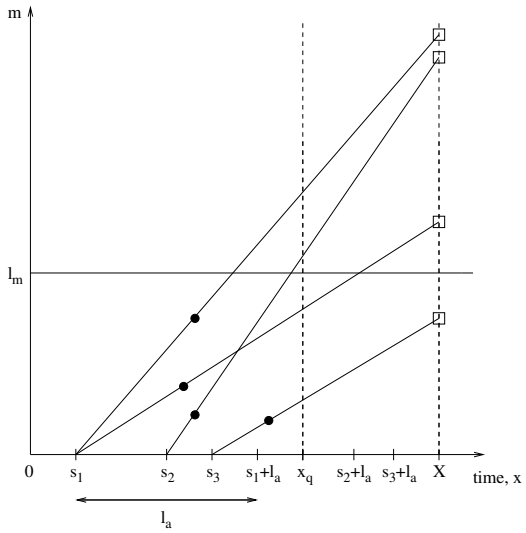

Figure 6.15: Actual time case

To get the true warranty claim rate, we need to adjust for withdrawals due to mileage. We use a similar approach as in the previous two cases, and all adjustments will be made to $\hat{M}(x)$. Then, for a vehicle with at least one claims, its contribution to $\hat{M}(x)$ is given by

$$I(s_i \leq x)I(z_i \leq l_a)I\left(r_i \leq \frac{l_m}{z_i}\right).$$

For a vehicle with no claims, its contribution is given by

$$I(s_i \leq x)I(z_i \leq l_a)\hat{F}\left(\frac{l_m}{z_i}\right),$$

where $\hat{F}(r)$ is the empirical distribution function for mileage accumulation rate as given in Section 6.1.1.

Figure 6.15 illustrates this idea above graphically using four hypothetical vehicles. In this figure,

- the large square represents the warranty coverage region,
- the little black circles represent the time and mileage for the latest claim of the four vehicles, and
- the little squares represent the extrapolated mileages of the four vehicles at the current time X.
- the straight lines represent the trajectories of the vehicles.

It can be seen that only the vehicle sold at time s_3 is under warranty coverage at the target time x_q, and will contribute to $M(x_q)$. The other three vehicles are all out of warranty coverage at time x_q. Both of the vehicles sold at time s_1 leave warranty coverage due to age, and one of them is also estimated to have exceeded the mileage limit before time x_q. The vehicle sold at time s_2 is still within the age limit, but it is estimated to have exceeded the mileage limit before time x_q.

Now, we consider the effect of reporting delay of claim. If vehicle i had experienced a claim at time x, then there is a time period of length $(X - x)$ for the claim to be posted to the database. The estimated probability that the claim is posted in that length of time is $\hat{G}(X - x)$, where $\hat{G}(d)$ is the empirical distribution function for reporting delay as given in Section 6.1.1. So, we can adjust the contribution of vehicle i by multiplying it with $\hat{G}(X - x)$. Then, by assuming independence between the process of generating reporting delay and that generating mileage accumulation, the adjustment for reporting delay can also be used along with the adjustment for mileage. Table 6.3 shows the contribution of a vehicle to $\hat{M}(t)$ in each case.

Case	Vehicle with claims	Vehicle without claims
Unadjusted	$I(s_i \leq x)I(z_i \leq l_a)$	$I(s_i \leq x)I(z_i \leq l_a)$
Adjusted for mileage	$I(s_i \leq x)I(z_i \leq l_a)I\left(r_i \leq \frac{l_m}{z_i}\right)$	$I(s_i \leq x)I(z_i \leq l_a)\hat{F}\left(\frac{l_m}{z_i}\right)$
Adjusted for delay	$I(s_i \leq x)I(z_i \leq l_a)\hat{G}(X-x)$	$I(s_i \leq x)I(z_i \leq l_a)\hat{G}(X-x)$
Adjusted for mileage & delay	$I(s_i \leq x)I(z_i \leq l_a)I\left(r_i \leq \frac{l_m}{z_i}\right) \times \hat{G}(X-x)$	$I(s_i \leq x)I(z_i \leq l_a)\hat{F}\left(\frac{l_m}{z_i}\right) \times \hat{G}(X-x)$

Table 6.3: Contribution to $\hat{M}(x)$ for vehicle i at target time x

Example

In this example, we estimate the mean cumulative cost of P-claims per vehicle, $\hat{\Lambda}(x)$, for Dataset 2001 from 22 May 2000 ($x = 1$, the first sale date) until 24 October 2003 ($X = 1251$, the current date or "cut-off" date).

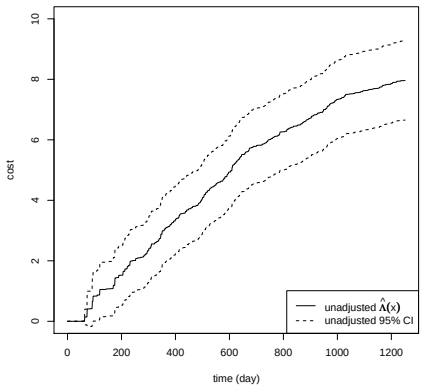

Figure 6.16: Unadjusted $\hat{\Lambda}(x)$ and 95% CI's

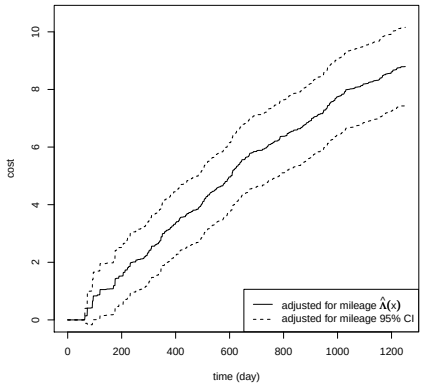

Figure 6.17: Adjusted for mileage $\hat{\Lambda}(x)$ and 95% CI's

Figure 6.16 shows the unadjusted $\hat{\Lambda}(x)$ and Figure 6.17 shows the adjusted for mileage $\hat{\Lambda}(x)$, along with the corresponding 95% confidence intervals (CI). In both figures, we see that the mean cumulative cost of P-

claims is initially zero, and then increases starting from $x = 57$ or 17 July 2000 (for which the first P-claim occurred).

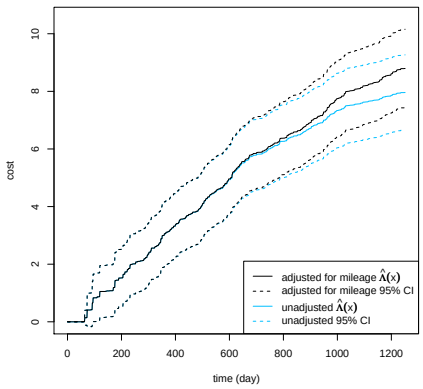

Figure 6.18: Unadjusted $\hat{\Lambda}(x)$ and adjusted for mileage $\hat{\Lambda}(x)$

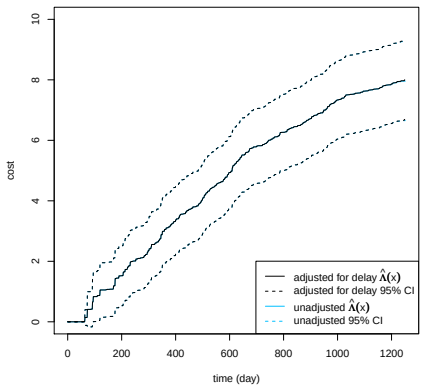

Figure 6.19: Unadjusted $\hat{\Lambda}(x)$ and adjusted for delay $\hat{\Lambda}(x)$

Figure 6.18 illustrates the effect of the adjustment for withdrawals from warranty coverage due to exceeding the mileage limit of 36000 miles. Without further analysis, the bend in the unadjusted curve could be due to a fall in the warranty claim rates over time. But, the adjusted for mileage curve indicates that the "true" rate is only slightly decreasing. However, unlike the adjustment for mileage in the example for "time" is age case, this adjustment is less significant as the adjusted for mileage curve still falls within the unadjusted 95% confidence interval and there is a substantial overlap of the two corresponding 95% confidence intervals. Nevertheless, this adjustment is becoming more significant, and we would expect the adjusted curve to go across the unadjusted 95% confidence interval as the time increase.

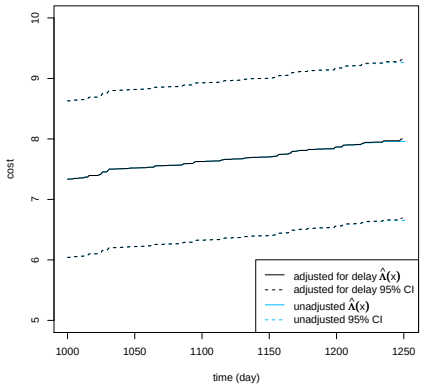

Figure 6.20: Unadjusted $\hat{\Lambda}(x)$ and adjusted for delay $\hat{\Lambda}(x)$, for day $x \geq 1000$

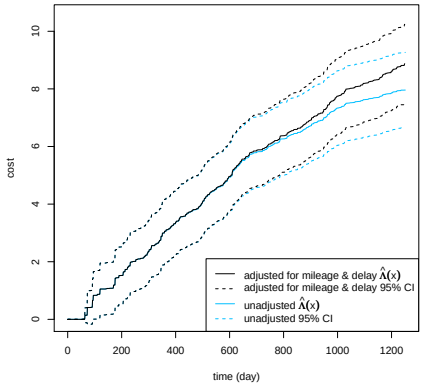

Figure 6.21: Unadjusted $\hat{\Lambda}(x)$ and adjusted for mileage & delay $\hat{\Lambda}(x)$

Next, Figure 6.19 illustrates the effect of adjustment for reporting delay. It can be seen that this adjustment is not substantial. Figure 6.21 illustrates the effect of simultaneous adjustment for both mileage and reporting delay. This figure shows a similar pattern as Figure 6.18, and we may conclude that the effect of this adjustment is mainly due to mileage.

6.3 Bootstrap Estimate of Standard Error

In this section, we contribute to the study of Chukova and Robinson [2005] by using bootstrap method in estimating the standard error of the mean cumulative function $\hat{\Lambda}(u)$, where u can be age t, mileage m, or the actual time x. The bootstrap method was proposed by B. Efron in 1979 [Efron, 1979]. Two major uses of the bootstrap are: the estimation of the standard error of a statistic and the construction of a confidence interval. There are two types of bootstrap:

- *nonparametric bootstrap* method, where we resample the observations *with replacement*.

- *parametric bootstrap* method, where we build a theoretical model using estimated parameters and resample from that distribution.

Here, we consider the nonparametric bootstrap method.

Suppose we have a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_M)$, where $x_i, i = 1, 2, \ldots, M$, represents an observation. Let θ be the statistic of interest, which can be evaluated as a function of $\mathbf{x}$, say $\hat{\theta} = s(\mathbf{x})$. Then, the bootstrap algorithm for estimating the standard error of $\hat{\theta}$ is given as follows [Efron and Tibshirani, 1993]:

1. Draw B independent *bootstrap samples* $\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_B^*$ with replacement from $\mathbf{x}$.

2. Evaluate the *bootstrap replication* of $\hat{\theta}$ corresponding to each bootstrap sample,

$$\hat{\theta}_b^* = s(\mathbf{x}_b^*), \qquad \text{for } b = 1, 2, \ldots, B. \tag{6.6}$$

3. Estimate the standard error of $\hat{\theta}$ using the standard deviation of the bootstrap replications. The result is called the *bootstrap estimate of standard error*, denoted by $\widehat{se}_B$.

An usual question is: how many bootstrap samples should we use? According to Efron and Tibshirani [1993], for estimating standard error, the number B will normally be in the range 25-200 (much larger values of B are needed for bootstrap confidence intervals).

There are many methods for constructing bootstrap confidence intervals. Here, we consider two types of $100(1-\alpha)\%$ confidence intervals:

- The *standard bootstrap confidence interval*

$$\hat{\theta} \pm z_{1-\alpha/2} \times \widehat{se}_B,$$

  where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution.

- The *percentile confidence interval*, where the $100(\alpha/2)$th and $100(1-\alpha/2)$th percentiles of the bootstrap replications are taken as the lower and upper limits of the confidence interval.

If the bootstrap distribution is assumed to be normal, then the standard bootstrap confidence interval is valid, and the standard bootstrap confidence interval and the percentile confidence interval will nearly agree. If the bootstrap distribution is not normal, then the two types of confidence intervals differ and the percentile confidence intervals is preferable [Efron and Tibshirani, 1993].

Figures 6.22 - 6.27 show the 95% confidence intervals evaluated using Eq. (5.5), the 95% standard bootstrap confidence intervals, and the 95% percentile confidence intervals for the following six different estimates computed in our examples: unadjusted $\hat{\Lambda}(t)$, adjusted for mileage $\hat{\Lambda}(t)$, unadjusted $\hat{\Lambda}(m)$, adjusted for age $\hat{\Lambda}(m)$, unadjusted $\hat{\Lambda}(x)$, and adjusted for mileage $\hat{\Lambda}(x)$. Here, we use $B = 1000$ and each observation consists of the details of a vehicle, including its claim records. In each case, we see that the three types of confidence intervals roughly agree and the differences between these confidence intervals are not significant. These results suggest that Eq. (5.5), with $M(t)$ replaced by $\hat{M}(t)$, works well for evaluating the standard error. However, a mathematical proof is still required, and we hope to achieve this in the future. Note that the 95% percentile confidence intervals are (usually) asymmetric, while the other two types of 95% confidence intervals are always symmetric.

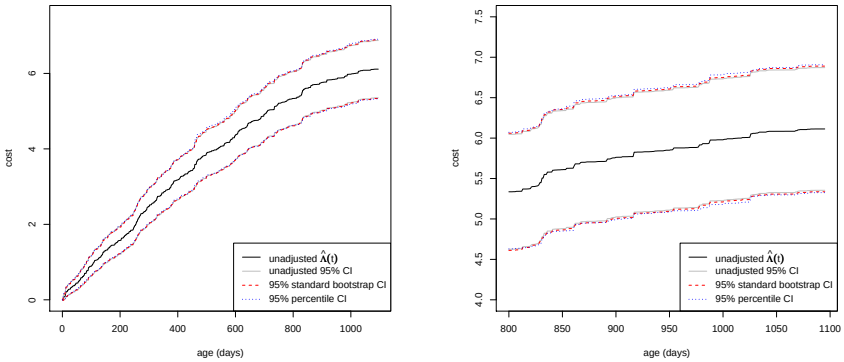

Figure 6.22: Unadjusted $\hat{\Lambda}(t)$ and 95% CI's

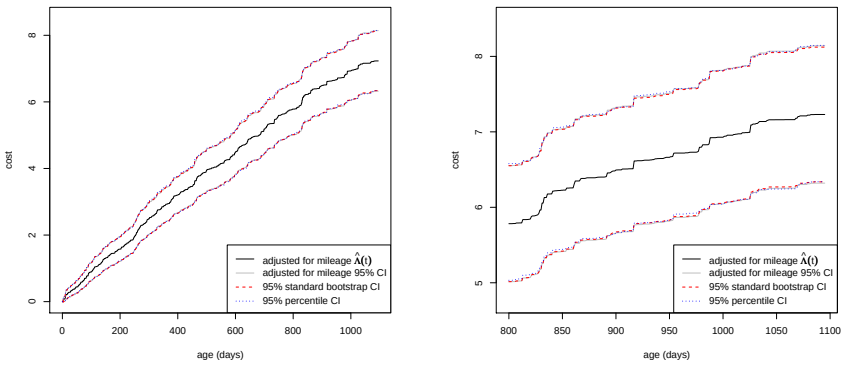

Figure 6.23: Adjusted for mileage $\hat{\Lambda}(t)$ and 95% CI's

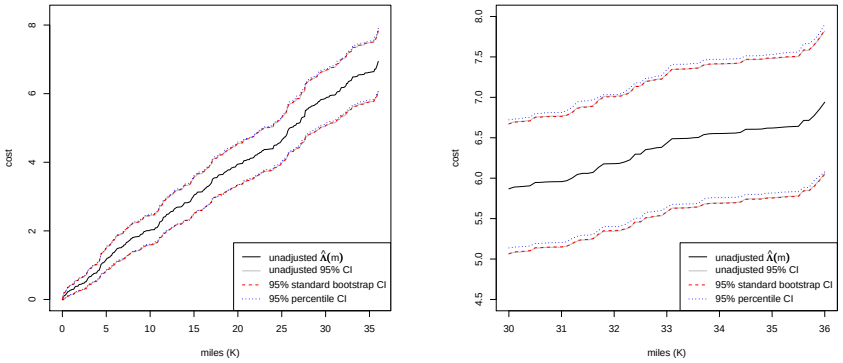

Figure 6.24: Unadjusted $\hat{\Lambda}(m)$ and 95% CI's

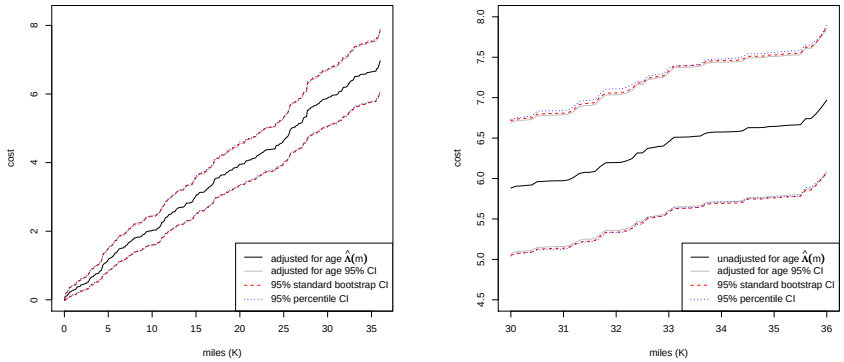

Figure 6.25: Adjusted for age $\hat{\Lambda}(m)$ and 95% CI's

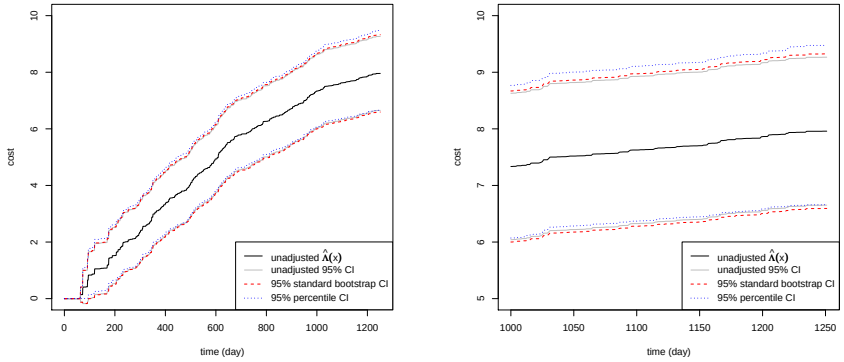

Figure 6.26: Unadjusted $\hat{\Lambda}(x)$ and 95% CI's

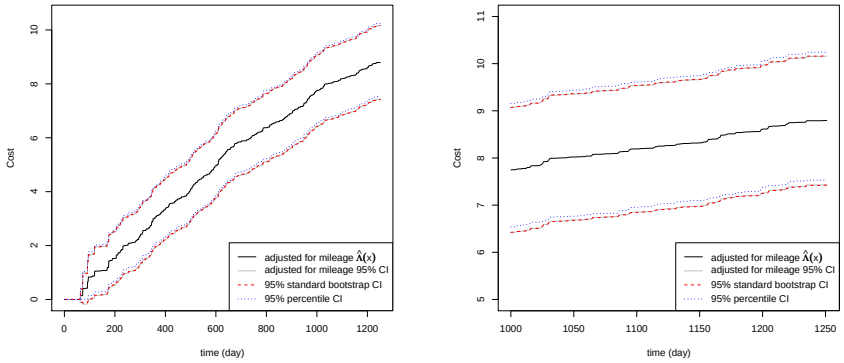

Figure 6.27: Adjusted for mileage $\hat{\Lambda}(x)$ and 95% CI's

6.4 Summary and Discussions

The models introduced in this chapter are based on the assumption that vehicles accumulate mileage approximately linearly with their age. We had considered "time" is age case, "time" is mileage case, and "time" is actual (calendar) time case. Our findings can be summarized as follows:

- In the "time" is age case, the adjustment for mileage is becoming statistically significant as the vehicle's age increases. This is shown by the increasing difference between the unadjusted curve and the adjusted for mileage curve of the mean cumulative warranty cost as the vehicle's age increases.

- In the "time" is mileage case, the adjustment for age is not statistically significant. This is because the adjustment for age does not begin until the oldest vehicle exceeds the age limit, which is three years from the first sale in our dataset. Also, the majority (62%) of the vehicles in the dataset are estimated to leave coverage due to mileage, instead of age.

- In the actual time case, as in the "time" is age case, the adjustment for mileage is becoming statistically significant as the vehicle's age increases. Nevertheless, the effect of the adjustment for mileage in the actual time case is less pronounced compared to the "time" is age case.

- The adjustment for reporting delay has little impact in each of the three cases.

In addition, we also used bootstrap method in estimating the standard errors of the mean cumulative warranty cost. The results suggest that Eq. (5.5), with $M(t)$ replaced by $\hat{M}(t)$, works well for evaluating the standard error. However, a mathematical proof is still required.

Chapter 7

Modeling Mileage Accumulation: Piece-Wise Linear Approach

In this chapter, we review the model proposed by Christozov et al. [2008], which is also based the robust estimator of Hu and Lawless [1996] and is an extension of the CR-Model. We will call this model the CCR-Model. This model relaxes the linearity assumption on mileage accumulation and allows for variation in the mileage accumulation rate over a vehicle's lifetime. Instead of using only the last claim to estimate the mileage accumulation rate, the CCR-model adopts a piece-wise linear approach, which uses all of the claims in the warranty database to characterize driving pattern (or mileage accumulation pattern). However, the CCR-Model does not take into account the effect of reporting delay of claim. Here, we will reproduce some of the results in Christozov et al. [2008] by using the same dataset, but with different computing softwares (we used statistical programming language **R**, while Christozov et al. [2008] used Microsoft Excel) and with different approaches in preparing the data (see Chapter 4). We will also develop a new model for the estimating the mean cumulative warranty cost per vehicle in the actual time case. Besides, we will also estimate the standard errors of the mean cumulative functions by bootstrap method. Note that all adjustments will always be made to $\hat{M}(t)$, $\hat{M}(m)$, or

$\hat{M}(x)$.

7.1 Grouping the Vehicles with Claims

Before we consider the CCR-Model, we first divide the vehicles with claims into several groups based on the their observed driving pattern (or mileage accumulation pattern), as in Christozov et al. [2008]. To measure the variability of driving pattern of a vehicle, we partition the warranty coverage region into strata, as shown in Figure 7.1, and then assign the vehicle to a particular group associated with the number of strata its trajectory goes through during its warranty life. We assume the trajectory of the vehicle to be piece-wise linear between its consecutive claims, as shown in Figure 7.2.

Theorem 7.1. *A vehicle is said to be stable with respect to a certain range, if the trajectory of this vehicle remains within this range throughout its warranty life [Christozov et al., 2008].*

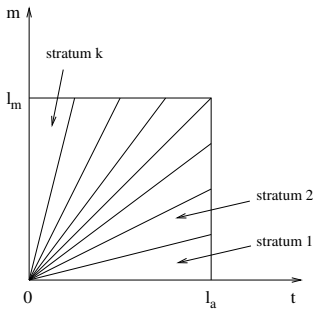

Figure 7.1: Warranty coverage region with strata

Let $\Pi_k, k > 0$, denote the strata partition of the warranty coverage region with k strata. For example, the partition of the warranty coverage region in Figure 7.1 is Π_8. Then, a vehicle is said to be k_i*-stable*, if the

claims of this vehicle are spread within exactly k_i strata of Π_k, where k_i is a positive integer number such that $k_0 = 1 < k_1 < k_2 < \ldots < k$. This means that this vehicle is stable with respect to an aggregated stratum consisting k_i strata of Π_k. We denote the size of this group of vehicles by $M_1^{k_i}$. In addition, let $O_{1,s}^{k_i}, s = 1, 2, \ldots, k$, be the number of k_i-stable vehicles, for which claims fall into stratum s of Π_k, such that

$$O_{1,1}^{k_i} + O_{1,2}^{k_i} + \ldots + O_{1,k}^{k_i} = M_1^{k_i}. \tag{7.1}$$

These counts $O_{1,s}^{k_i}$ will be used to estimate the strata distribution, which reflects the proportion of vehicles within each strata. Note that the counts $O_{1,s}^{k_i}$ may not be an integer. In order to estimate $O_{1,s}^{k_i}$, we need to take into account the contribution of each k_i-stable vehicle to the number of vehicles in each of the k_i strata for which the vehicle is k_i-stable. We assume this contribution to be uniform over the corresponding k_i strata. This means that each k_i-stable vehicle contributes a fraction of $1/k_i$ to each of these k_i strata. For example, Figure 7.2 shows the trajectory of a 3_8-stable vehicle, which goes through strata 2, 3 and 4. So, the contribution of this vehicle to the number of vehicles in each of these strata is $1/3$.

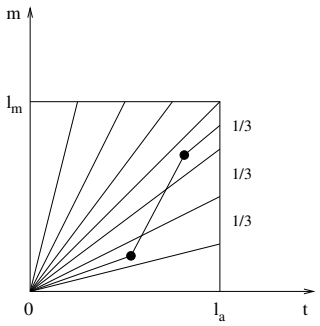

Figure 7.2: Trajectory of a 3_8-stable vehicle

By identifying the groups of vehicles with claims as above, we partition the set of all M_1 vehicles with claims into non-overlapping groups, called

driving pattern groups (DPG's). In each of these DPG's, the variability of driving pattern follows a similar pattern. Let k_d be the size of the largest number of stratum needed to describe the stable driving patterns of the vehicles with claims effectively. Then, we will have several groups of vehicles, such that the first group consists of M_1^1 1-stable vehicles, the second group consists of $M_1^{k_1}$ k_1-stable vehicles, ... and the last group consists of $M_1^{k_d}$ k_d-stable vehicles. Besides that, there will also be another group of vehicles, say of size $M_1^{k_d+1}$, with *unstable* driving pattern. The vehicles in this group are said to be unstable, as their driving trajectories during their warranty lives go through more than k_d strata of Π_k. For this group of unstable vehicles, we simply apply the simple linear mileage accumulation model used in Chapter 6, i.e., we assume that the trajectories of these vehicles are approximately linear, determined by their latest claim. Further, each unstable vehicle is assigned to a single stratum based on its latest claim.

At the end, the whole set of vehicles with claims is partitioned into non-overlapping groups, such that

$$M_1 = M_1^1 + M_1^{k_1} + \ldots + M_1^{k_d} + M_1^{k_d+1}. \tag{7.2}$$

In addition, the number of vehicles with claims within the stratum s is given by

$$O_{1,s} = O_{1,s}^1 + O_{1,s}^{k_1} + \ldots + O_{1,s}^{k_d} + O_{1,s}^{k_d+1} \tag{7.3}$$

for $s = 1, 2, ..., k$. Note that the notion of unstable cars is not important. We can build a model by using a total of k DPG's, i.e., $k_d = k$. As a result, the number of unstable vehicles will be equal to zero.

7.2 Estimating the Strata Distribution

By using the strata counts $O_{1,s}, s = 1, 2, ..., k$, given by Eq. (7.3), we can compute the strata distribution $\mathbf{p} = (p_1, p_2, \ldots, p_k)$, where p_s = Prob(a vehicle with claims belongs to stratum s) by

$$p_s = \frac{O_{1,s}}{M_1}, \quad \text{for } s = 1, 2, ..., k. \tag{7.4}$$

We assume that the strata distribution is time independent, i.e., it remains the same over different age intervals. Further, we also assume that the driving patterns of the vehicles with and without claims are probabilistically the same. Thus, the strata distribution can be used as a reasonable representation for the vehicles without claims, and hence the strata distribution describes the set of all vehicles in the database.

Example

Let us divide the warranty coverage region into age-bins with size of one month (each month has equal number of days) and mileage-bins with size of 1000 miles. Also, let us partition the warranty coverage region into $k = 72$ strata, so that each stratum is reasonably narrow. Then, the set of vehicles with claims (or odometer readings) is divided into several non-overlapping driving pattern groups (DPG's) as follows:

- Vehicles with a single claim - group S. Note that multiple claims occurred at the same time and mileage are regarded as a single claim.
- Vehicles with more than one claim
  - Vehicles with all claims within one stratum - group W_1.
  - Vehicles with all claims within three strata - group W_3.
  - Vehicles with all claims within six strata - group W_6.

– Vehicles with all claims spread over more than six strata - group U.
(This is the group of unstable vehicles.)

Next, in order to estimate the strata distribution for the above set of DPG's, we determine the contribution of a vehicle to a stratum in the following way:

- For group S, each vehicle belongs to a single stratum determined by its claim.
- For group W_1, each vehicle belongs to a single stratum determined by its claims.
- For group W_3, each vehicle is uniformly distributed over the three associated strata.
- For group W_6, each vehicle is uniformly distributed over the six associated strata.
- For group U, each vehicle belongs to a single stratum determined by its last claim.

Table 7.1 shows the number of vehicles with claims in each DPG for Datasets 1998 - 2001. Then, Figure 7.3 shows the estimated strata distribution for these datasets, which are asymmetric and skewed to the left.

Dataset	S	W_1	W_3	W_6	U
2001	11175	1619	3171	2230	3541
2000	8833	1557	2866	1957	3181
1999	11446	1888	3816	2719	4438
1998	10381	1936	3905	2840	4879

Table 7.1: Number of vehicles with claims in each DPG for Datasets 1998 - 2001

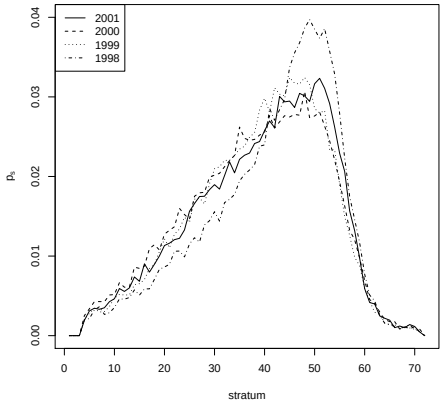

Figure 7.3: Estimated strata distribution for Datasets 1998 - 2001

7.3 Overview of the CCR-Model

Now, our goal is to estimate the mean cumulative warranty cost (or number of claims) per vehicle. Here, we provide an overview of the CCR-Model. We will consider the "time" is age case and the "time" is mileage case.

7.3.1 "Time" is Age Case

Firstly, we consider the case where "time" is the age of the vehicle. Let us define a regular partition $0 = t_0 < t_1 < \ldots < t_{n-1} < t_n = l_a$ of size n such that

$$t_j - t_{j-1} = h_a, \quad j = 1, 2, \ldots, n,$$

where $h_a = l_a/n$ and l_a is the warranty age limit, as in Figure 7.4. If necessary, we can extend this partition beyond the warranty age limit l_a. As the "time" discretization is defined by the step h_a, we assume discrete values $t_j = jh_a$ for $j = 1, 2, \ldots, n, \ldots$, such that $t_0 = 0, t_1 = h_a, \ldots, t_n = l_a, t_{n+1} =$

$l_a + h_a, \ldots$

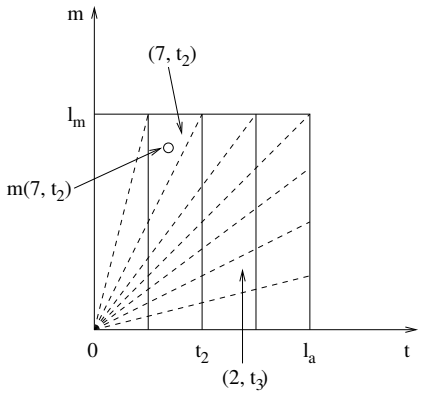

Figure 7.4: Age-bins

Let $N_j^{(t)}$ be the number of vehicles with age within an age-bin $\Delta_j^{(t)} = [t_{j-1}, t_j), j = 1, 2, \ldots, n$. As vehicle's age is known for all vehicles at all time, $N_j^{(t)}$ is always known. If we ignore withdrawals from warranty coverage due to mileage, then the number of vehicles eligible to generate a claim at the target age t_q, $t_q \leq l_a$, is the number of vehicles age t_q or older, i.e.,

$$\hat{M}(t_q) = \sum_{i=1}^{M} I(a_i \geq t_q) = M - \sum_{j=1}^{q} N_j^{(t)}, \tag{7.5}$$

where a_i is the current age of vehicle i (on the "cut-off" date).

Then, to adjust for withdrawals from warranty coverage due to mileage, the estimator $\hat{M}(t_q)$ is adjusted to

$$\hat{M}(t_q) = \left(M - \sum_{j=1}^{q} N_j^{(t)} \right) \left(\sum_{s=1}^{k-q} p_s \right), \tag{7.6}$$

where p_s is the probability that a vehicle belongs to stratum s. For example, consider Figure 7.4 and suppose we want to estimate $M(t_2)$. We need

to estimate the number of vehicle age t_2 or older, which are still under warranty coverage at time t_2. The number of vehicles age t_2 or older is equal to $M - \sum_{j=1}^{2} N_j^{(t)}$. At time t_2, vehicles with driving patterns associated with strata 7 and 8 would have left the warranty coverage, whereas vehicles with driving pattern associated with strata 1 to 6 would still be under coverage. Therefore, the proportion of vehicles that are still under coverage at time t_2 is $\sum_{s=1}^{6} p_s$. Hence, the required estimate is given by

$$\hat{M}(t_2) = \left(M - \sum_{j=1}^{2} N_j^{(t)} \right) \left(\sum_{s=1}^{6} p_s \right).$$

Example I

In this example, we estimate the mean cumulative cost of P-claims per vehicle, $\hat{\Lambda}(t)$, up to the age limit of $l_a = 36$ months by using Dataset 2001. Note that, for convenience, we write $t = t_q = 1, 2, \ldots$ (in months).

Figure 7.5 shows the unadjusted $\hat{\Lambda}(t)$ and Figure 7.6 shows the adjusted for mileage $\hat{\Lambda}(t)$, along with the 95% confidence intervals (CI) evaluated using Eq. (5.5), the 95% standard bootstrap confidence intervals, and the 95% percentile confidence intervals. We can see that the three confidence intervals roughly agree in both cases.

Then, Figure 7.7 illustrates the effect of the adjustment for withdrawals from warranty coverage due to exceeding the mileage limit of 36000 miles. The adjusted for mileage curve indicates that the "true" warranty claim rate is not decreasing (or only decreasing slightly). In addition, the increasing difference between the unadjusted curve and the adjusted for mileage curve indicates that the adjustment for mileage is becoming more statistically significant as the vehicle's age increases.

It can be seen that the adjusted for mileage curve goes across the unadjusted 95% confidence interval when the vehicle's age is about 29 months. Whereas in the example (for CR-Model) of Section 6.1.1, the adjusted for mileage curve goes across the unadjusted 95% confidence interval when

the vehicle's age is approximately 900 days (about 29 months). Overall, the results we obtained are quite similar to the results obtained using the CR-Model.

Note that the above results replicate the results given by Christozov et al. [2008].

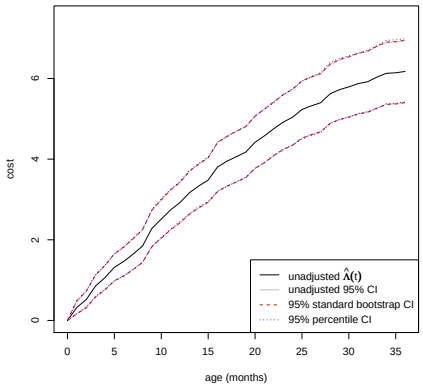

Figure 7.5: Unadjusted $\hat{\Lambda}(t)$ and 95% CI's

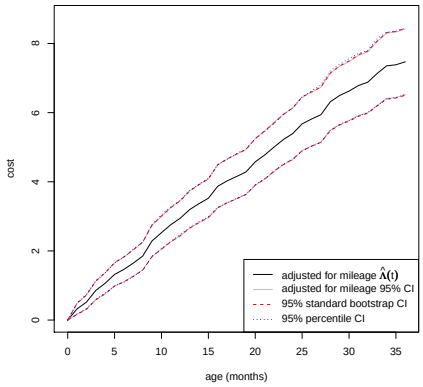

Figure 7.6: Adjusted for mileage $\hat{\Lambda}(t)$ and 95% CI's

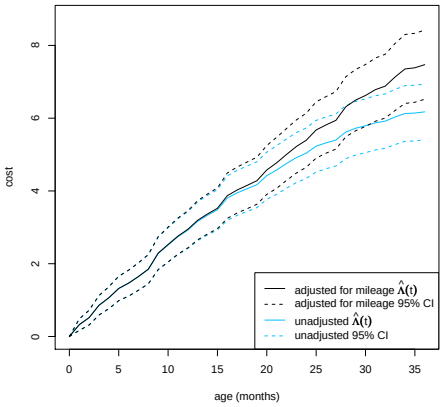

Figure 7.7: Unadjusted $\hat{\Lambda}(t)$ and adjusted for mileage $\hat{\Lambda}(t)$

Example II

Now, by using Datasets 1998 - 2001, we explore the relationship between the variability of driving pattern and the mean cumulative warranty cost (per vehicle). We will consider the adjusted for mileage $\hat{\Lambda}(t)$. Note that, instead of using the cost of P-claims only, we examine the total cost of all claims. Also, for each of the datasets, the set of vehicles with no claims is divided into each DPG according to the proportion of vehicles with claims in each DPG.

Let us consider the following DPG's: S, W_1, W_3, W_6, and U, as already defined. Figures 7.8 - 7.11 show the mean cumulative warranty cost for different DPG's for Datasets 1998 - 2001 respectively. For Datasets 1998 - 2000, we see that group W_1 has the lowest cost, followed by group W_3 and group W_6, while group U have the highest cost. For Dataset 2001, the cost for group W_1 is initially the lowest, but there is a sharp increase in the cost of this group when the vehicle's age exceeds 25 months for some unknown reasons. Due to the lack of information, we are unable to look into this further.

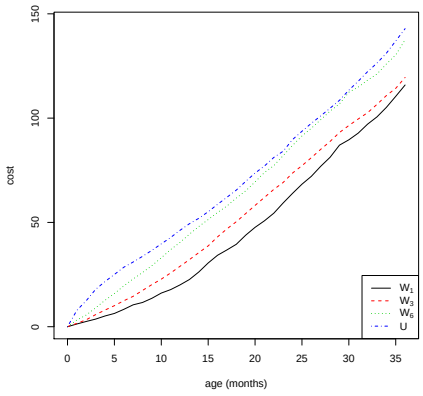

Figure 7.8: Mean cumulative warranty cost for different DPG's for Dataset 1998

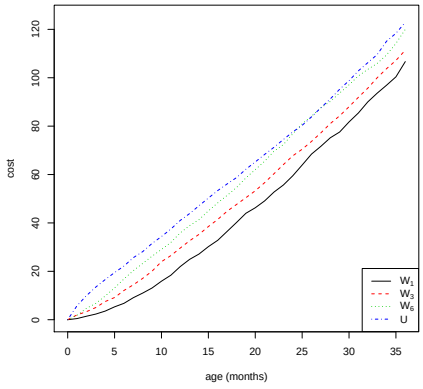

Figure 7.9: Mean cumulative warranty cost for different DPG's for Dataset 1999

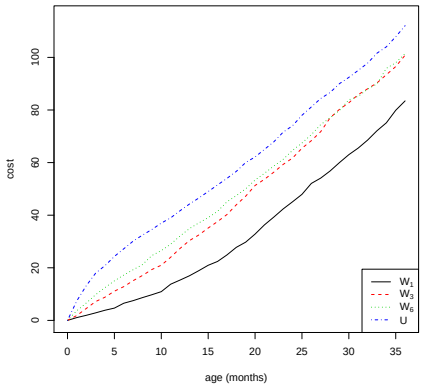

Figure 7.10: Mean cumulative warranty cost for different DPG's for Dataset 2000

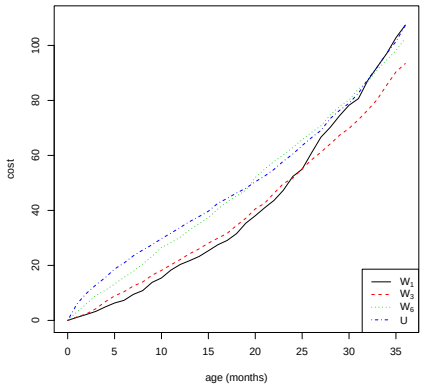

Figure 7.11: Mean cumulative warranty cost for different DPG's for Dataset 2001

Figures 7.12 - 7.15 show the mean cumulative warranty cost for different DPG's at $t = 36$ months for Datasets 1998 - 2001 respectively. Except for Figure 7.15 which corresponds to Dataset 2001, the other figures all demonstrate an upward trend over DPG's with increasing variability of driving pattern. For further analysis, we fit a trend line to each of these graphs (by using simple linear regression). For Datasets 1998 - 2000, the slopes of the trend lines are 9.9098, 5.7303, and 8.6441 respectively. All of these slopes are positive and significant at the 10% level. For Dataset 2001, the slope of the trend line is 0.8840. Even though this slope is also positive, it is not significant at the 10% level.

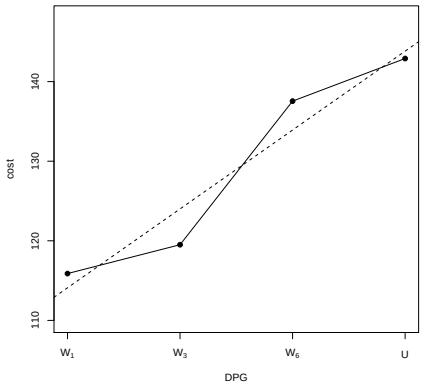

Figure 7.12: Mean cumulative warranty cost for different DPG's at $t = 36$ months for Dataset 1998

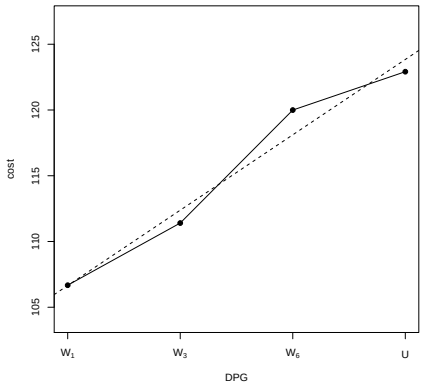

Figure 7.13: Mean cumulative warranty cost for different DPG's at $t = 36$ months for Dataset 1999

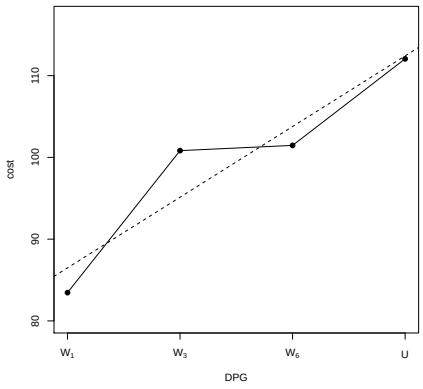

Figure 7.14: Mean cumulative warranty cost for different DPG's at $t = 36$ months for Dataset 2000

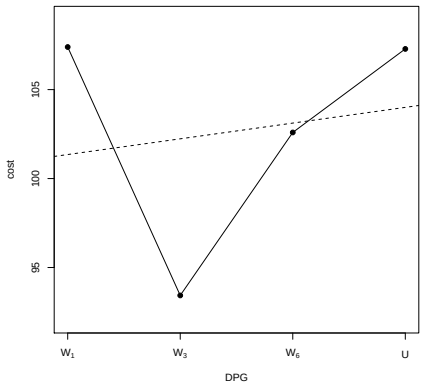

Figure 7.15: Mean cumulative warranty cost for different DPG's at $t = 36$ months for Dataset 2001

Next, we investigate further the relationship between the variability of the driving pattern and the mean cumulative warranty cost by using a different definition of DPG's, with more groups as follows: $S, W_1, W_2, \ldots, W_{10}$, and U (with claims spread over more than 10 strata). Figures 7.16 - 7.19 show the mean cumulative warranty cost for different DPG's at $t = 36$

months for Datasets 1998 - 2001 respectively. All of these figures, including the one for Dataset 2001, demonstrate an upward trend over DPG's with increasing variability of driving pattern. Again, we fit a trend line to each of these graphs. For Datasets 1998 - 2000, the slopes of the trend lines are 3.8506, 2.4027, and 1.9952 respectively. All of these slopes are positive and significant at the 10% level. For Dataset 2001, the slope of the trend line is 0.9112, which is also positive. However, this slope is not significant at the 10% level.

Overall, the above results suggest that a higher variability of driving pattern leads to a higher mean cumulative warranty cost. This is a very interesting observation that requires further study. It suggests that the variability of driving pattern should be taken into account in modeling mileage accumulation.

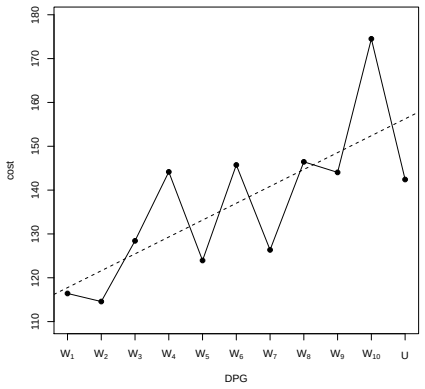

Figure 7.16: Mean cumulative warranty cost for different DPG's at $t = 36$ months for Dataset 1998

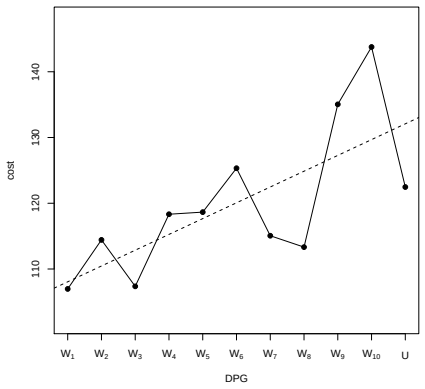

Figure 7.17: Mean cumulative warranty cost for different DPG's at $t = 36$ months for Dataset 1999

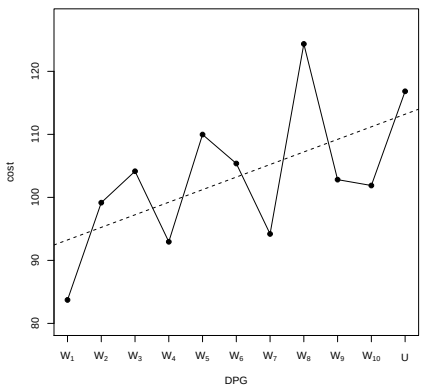

Figure 7.18: Mean cumulative warranty cost for different DPG's at $t = 36$ months for Dataset 2000

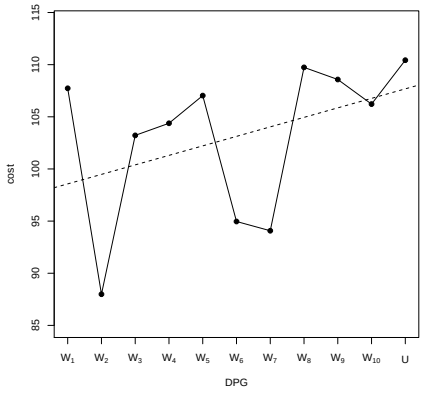

Figure 7.19: Mean cumulative warranty cost for different DPG's at $t = 36$ months for Dataset 2001

7.3.2 "Time" is Mileage Case

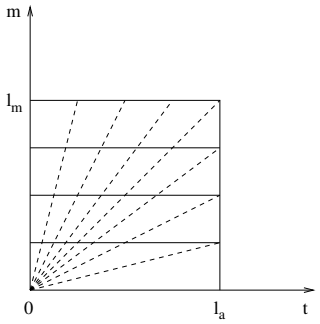

Figure 7.20: Mileage-bins

Next, we consider the case where "time" is the mileage of the vehicle. Let us define a regular partition $0 = m_0 < m_1 < \ldots < m_{n-1} < m_n = l_m$ of size n such that

$$m_j - m_{j-1} = h_m, \quad j = 1, 2, \ldots, n,$$

where $h_m = l_m/n$ and l_m is the warranty mileage limit, as in Figure 7.20. If necessary, we can extend this partition beyond the warranty mileage limit l_m. As the "time" discretization is defined by the step h_m, we assume discrete values $m_j = jh_m$ for $j = 1, 2, \ldots, n, \ldots$, such that $m_0 = 0, m_1 = h_m, \ldots, m_n = l_m, m_{n+1} = l_m + h_m, \ldots$

Then, let $N_j^{(m)}$ denote the number of vehicles with accumulated mileage within a mileage-bin $\Delta_j^{(m)} = [m_{j-1}, m_j), j = 1, 2, \ldots, n$. In addition, consider an age-strata grid (s, t_i), determined by the stratum s for $s = 1, 2, \ldots, k$, and the age-bin $\Delta_i^{(t)}$ for $i = 1, 2, \ldots, n, \ldots$ In order to estimate $\hat{M}(m)$, we need to know $N_j^{(m)}$, $j = 1, 2, \ldots, n$. Since the current mileage is not known exactly even for vehicles with claims, we will estimate $N_j^{(m)}$ by using the strata distribution and the age-strata grid (s, t_i). The idea is similar to the ideas of analyzing group data. For each cell (s, t_i) of the age-strata grid, we identify a typical mileage representation, say $m(s, t_i)$. For example, in Figure 7.4, $m(7, t_2)$ represents the mileage for the cell $(7, t_2)$. Then, the number of vehicles with current mileage equal to $m(s, t_i)$ is estimated by

$$N_{m(s,t_i)} = p_s N_i^{(t)}. \tag{7.7}$$

Subsequently, we estimate $N_j^{(m)}$ by adding up the numbers of vehicles with typical mileage representation that fall within the j^{th} mileage-bin $\Delta_j^{(m)}$, i.e.,

$$N_j^{(m)} = \sum_{m(s,t_i) \in \Delta_j^{(m)}} N_{m(s,t_i)}. \tag{7.8}$$

For the unadjusted case, we first estimate $N_j^{(m)}$ by extending the age-strata grid beyond the warranty age limit l_a to cover all vehicles, including those that had exceeded the age limit. Consequently, the number of vehicles that are eligible to generate a claim at the target mileage m_q, $m_q \leq l_m$,

is given by

$$\hat{M}(m_q) = M - \sum_{j=1}^{q} N_j^{(m)}. \tag{7.9}$$

In order to adjust $\hat{M}(m_q)$ for withdrawals from warranty coverage due to age, we first divide each mileage-bin into two parts, one part for vehicles with age less than the age limit and the other part for vehicles with age larger or equal to the age limit. Then, for a mileage-bin $\Delta_j^{(m)} = [m_{j-1}, m_j)$, let

- $N_{j1}^{(m)}$ be the number of vehicles in this mileage-bin with age less the age limit, and
- $N_{j2}^{(m)}$ be the number of vehicles in this mileage-bin with age larger than or equal to the age limit,

such that $N_{j1}^{(m)} + N_{j2}^{(m)} = N_j^{(m)}$. We can compute $N_{j1}^{(m)}$ by the method for estimating $N_j^{(m)}$ above, but using the age-strata grid within the age-limit only. Subsequently, the adjusted for age estimator for $M(m_q)$ is given by

$$\hat{M}(m_q) = \left(\tilde{M} - \sum_{j=1}^{q} N_{j1}^{(m)} \right) + \left(M - \tilde{M} \right) \sum_{s=q+1}^{k} p_s, \tag{7.10}$$

where $\tilde{M} = \sum_{i=1}^{n} N_i^{(t)}$ is the number of vehicles with age less than the age limit. Note that we have corrected the adjusted for mileage estimator given by Christozov et al. [2008].

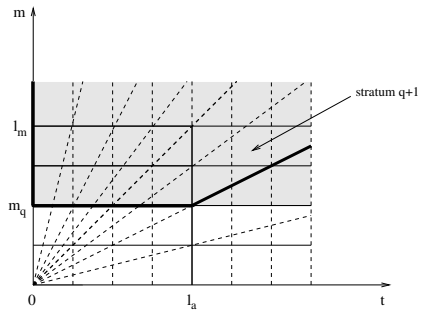

Figure 7.21: Unadjusted $\hat{M}(m)$ and adjusted for age $\hat{M}(m)$

Figure 7.21 illustrates the difference between the unadjusted $\hat{M}(m)$ and the adjusted for age $\hat{M}(m)$. Let m_q denote the target mileage. The unadjusted $\hat{M}(m_q)$ will count all vehicles with accumulated mileage greater than or equal to m_q, including those that had exceeded the age limit, as represented by the shaded region. On the other hand, the adjusted for age $\hat{M}(m_q)$ will count

- the vehicles with accumulated mileage greater than or equal to m_q *and* with age less than age limit, and
- the vehicles with age greater than or equal to the age limit *and* belong to strata $q+1, q+2, \ldots, k$,

as represented by the shaded region surrounded by the dark solid line. The vehicles with age greater than or equal to the age limit *and* belong to strata $q+1, q+2, \ldots, k$ are included, as they would have reached mileage m_q before they exceeded the age limit.

Example I

In this example, we estimate the mean cumulative cost of P-claims per vehicle, $\hat{\Lambda}(m)$, up to the mileage limit of $l_m = 36000$ miles for Dataset 2001. Note that, for convenience, we write $m = m_q = 1, 2, \ldots$ (in unit of 1000's).

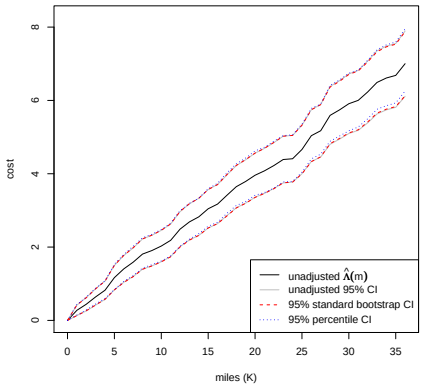

Figure 7.22: Unadjusted $\hat{\Lambda}(m)$ and 95% CI's

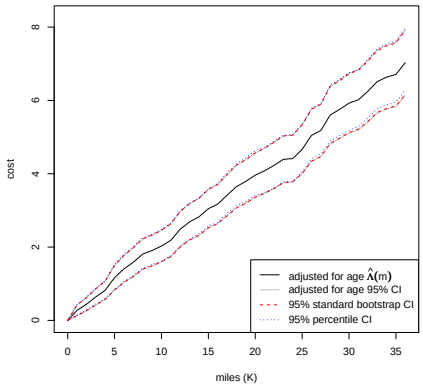

Figure 7.23: Adjusted for age $\hat{\Lambda}(m)$ and 95% CI's

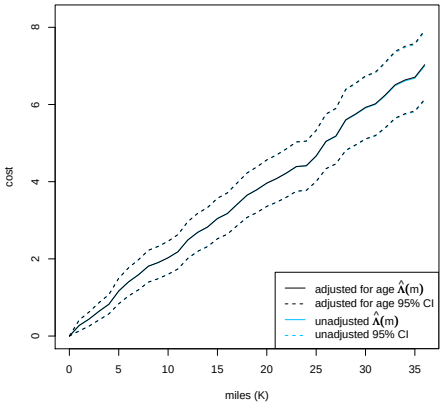

Figure 7.24: Unadjusted $\hat{\Lambda}(m)$ and adjusted for age $\hat{\Lambda}(m)$

Figure 7.22 shows the unadjusted $\hat{\Lambda}(m)$ and Figure 7.23 shows the adjusted for age $\hat{\Lambda}(m)$, along with the 95% confidence intervals (CI) evaluated using Eq. (5.5), the 95% standard bootstrap confidence intervals, and the 95% percentile confidence intervals. We see that the three confidence intervals roughly agree in both cases. Then, Figure 7.24 illustrates

the effect of the adjustment for withdrawals from warranty coverage due to exceeding the age limit of 36 months. It can be seen that this adjustment is not significant, but it is slightly more significant than the corresponding adjustment done in the example (for CR-Model) of Section 6.1.2.

Example II

Now, by using Datasets 1998 - 2001, we explore the relationship between the variability of driving pattern and the mean cumulative warranty cost (per vehicle). We will consider the adjusted for age $\hat{\Lambda}(m)$. Note that, instead of using the cost of P-claims only, we examine the total cost of all claims. Also, for each of the datasets, the set of vehicles with no claims is divided into each DPG according to the proportion of vehicles with claims in each DPG.

Let us consider the following DPG's: S, W_1, W_3, W_6, and U, as already defined. Figures 7.25 - 7.28 show the mean cumulative warranty cost for different DPG's for Datasets 1998 - 2001 respectively. For Datasets 1998 - 2000, we see that group W_1 has the lowest cost, followed by group W_3 and group W_6, while group U have the highest cost. For Dataset 2001, the cost for group W_1 is initially the lowest, but there is a sharp increase in the cost of this group when the vehicle's mileage exceeds $25K$ miles for some unknown reasons. Due to the lack of information, we are unable to look into this further.

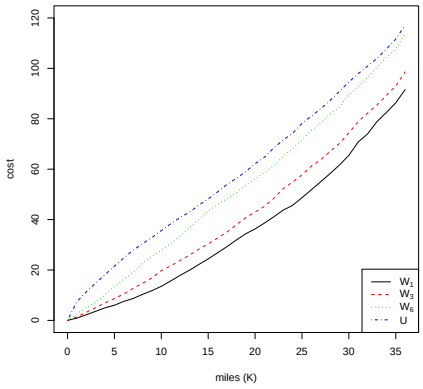

Figure 7.25: Mean cumulative warranty cost for different DPG's for Dataset 1998

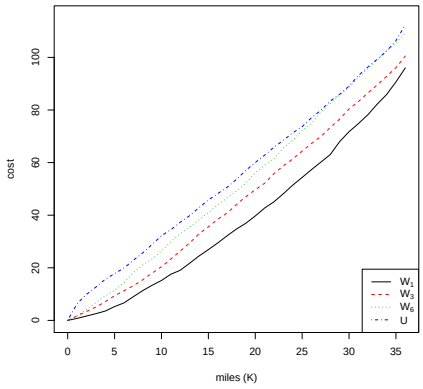

Figure 7.26: Mean cumulative warranty cost for different DPG's for Dataset 1999

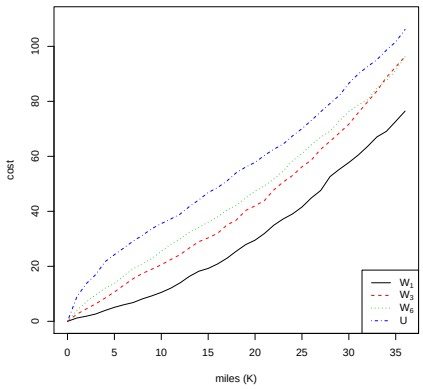

Figure 7.27: Mean cumulative warranty cost for different DPG's for Dataset 2000

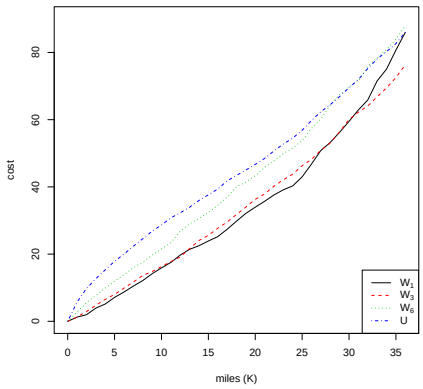

Figure 7.28: Mean cumulative warranty cost for different DPG's for Dataset 2001

Figures 7.29 - 7.32 show the mean cumulative warranty cost for different DPG's at $m = 36K$ miles for Datasets 1998 - 2001 respectively. Except for Figure 7.32 which corresponds to Dataset 2001, the other figures all demonstrate an upward trend over DPG's with increasing variability of

driving pattern. For further analysis, we fit a trend line to each of these graphs (by using simple linear regression). For Datasets 1998 - 2000, the slopes of the trend lines are 9.1855, 5.7994, and 8.9071 respectively. All of these slopes are positive and significant at the 10% level. For Dataset 2001, the slope of the trend line is 1.1768. Even though this slope is also positive, it is not significant at the 10% level.

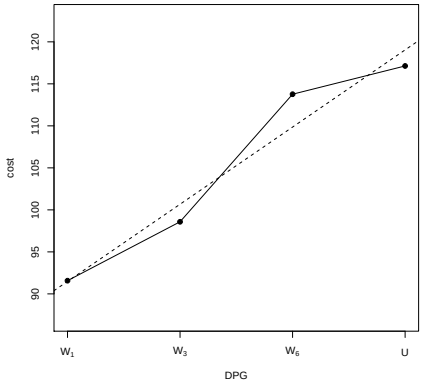

Figure 7.29: Mean cumulative warranty cost for different DPG's at $m = 36K$ miles for Dataset 1998

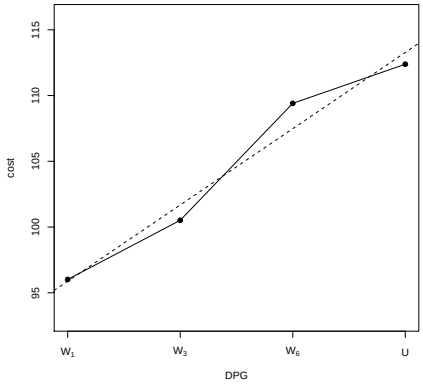

Figure 7.30: Mean cumulative warranty cost for different DPG's at $m = 36K$ miles for Dataset 1999

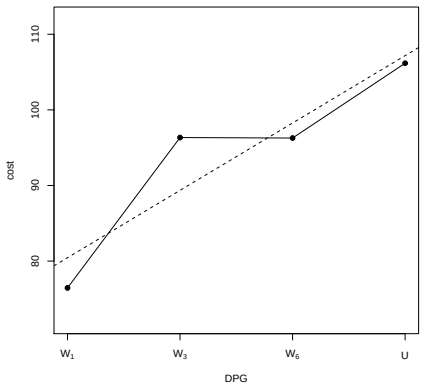

Figure 7.31: Mean cumulative warranty cost for different DPG's at $m = 36K$ miles for Dataset 2000

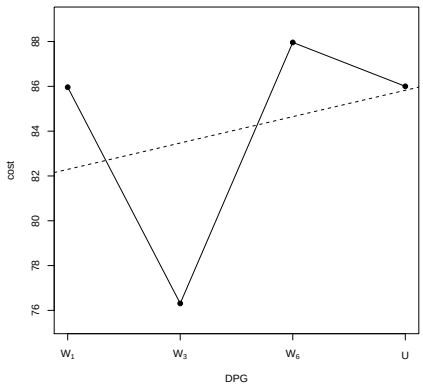

Figure 7.32: Mean cumulative warranty cost for different DPG's at $m = 36K$ miles for Dataset 2001

Next, we investigate further the relationship between the variability of the driving pattern and the mean cumulative warranty cost by using a different definition of DPG's, with more groups as follows: $S, W_1, W_2, \ldots, W_{10}$, and U (with claims spread over more than 10 strata). Figures 7.33 - 7.36 show the mean cumulative warranty cost for different DPG's at $m = 36K$ miles for Datasets 1998 - 2001 respectively. All of these figures, including the one for Dataset 2001, demonstrate an upward trend over DPG's with increasing variability of driving pattern. Again, we fit a trend line to each of these graphs. For Datasets 1998 - 2000, the slopes of the trend lines are 3.2474, 2.2805, and 1.9128 respectively. All of these slopes are positive and significant at the 10% level. For Dataset 2001, the slope of the fitted trend line is 0.4992, which is also positive. However, this slope is not significant at the 10% level.

Overall, the above results suggest that a higher variability of driving pattern leads to a higher mean cumulative warranty cost. Therefore, we should take into account the variability of driving pattern in modeling mileage accumulation. However, more study on this observation are required.

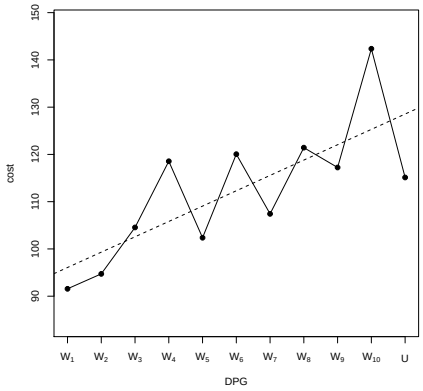

Figure 7.33: Mean cumulative warranty cost for different DPG's at $m = 36K$ miles for Dataset 1998

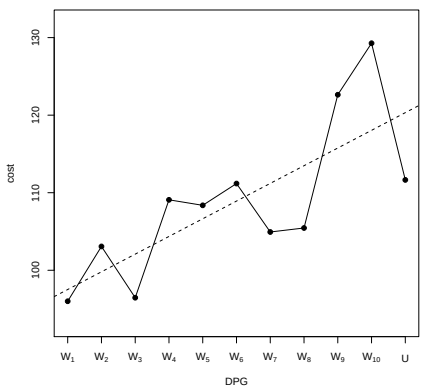

Figure 7.34: Mean cumulative warranty cost for different DPG's at $m = 36K$ miles for Dataset 1999

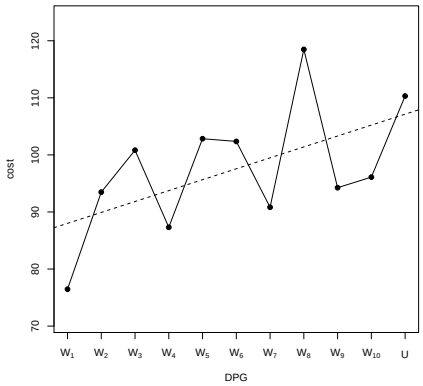

Figure 7.35: Mean cumulative warranty cost for different DPG's at $m = 36K$ miles for Dataset 2000

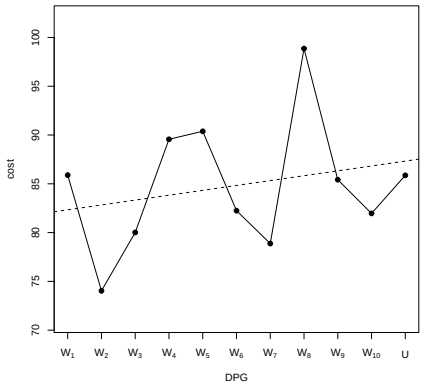

Figure 7.36: Mean cumulative warranty cost for different DPG's at $m = 36K$ miles for Dataset 2001

7.4 New Model: Actual Time Case

Now, we develop a new model for estimating the mean cumulative warranty cost per vehicle in the actual time case, $\Lambda(x)$. For simplicity and better illustration of the idea, the models introduced here are based on the partition we have already defined in our example, Π_{72}. These models can be modified easily for other forms of partition, if required.

Let X denote the current time (the "cut-off" date). Also, let us define a regular partition of time $0 = x_0 < x_1 < \ldots < x_{n-1} < x_n$, such that $X \in (x_{n-1}, x_n]$ and

$$x_j - x_{j-1} = h, \quad j = 1, 2, \ldots, n,$$

as in Figure 7.37. Here, we have $h = 1$ month. Then, let $N_j^{(x)}$ be the number of vehicles sold within a time-bin $\Delta_j^{(x)} = (x_{j-1}, x_j], j = 1, 2, \ldots, n$. As the sale date of a vehicle is always known, $N_j^{(x)}$ is also always known.

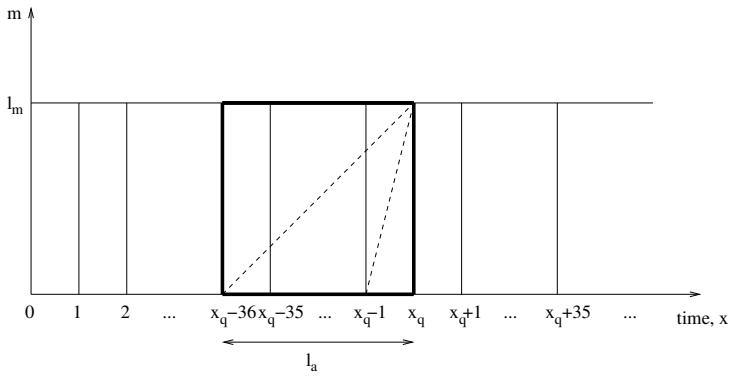

Figure 7.37: Time-bins

If we ignore withdrawals from warranty coverage due to mileage, then the number of vehicles eligible to generate a claim at the target time x_q is simply the number of vehicles sold within the 36 months before x_q, i.e.,

$$\hat{M}(x_q) = \begin{cases} \sum_{i=1}^{q} N_{q-i+1}^{(x)} & \text{if } q = 1, 2, \ldots, 35; \\ \sum_{i=1}^{36} N_{q-i+1}^{(x)} & \text{if } q = 36, 37, \ldots, n. \end{cases} \tag{7.11}$$

To take into account the withdrawals due to mileage, we will make the adjustment for mileage separately for each time-bin. Suppose we make the adjustment for mileage from the *left*-endpoint of each time-bin. This means that all vehicles in a particular time-bin are regarded as having the *oldest* possible age. For illustration, let us consider Figure 7.37 and the first time-bin before the target time x_q, $\Delta_q^{(x)} = (x_{q-1}, x_q]$. A total of $N_q^{(x)}$ vehicles would be sold within this time-bin, and these vehicles would have ages between 0 and 1 month at the target time x_q. To adjust for mileage, we will regard all of these vehicle as being 1-month old. Then, the estimated proportion of these $N_q^{(x)}$ vehicles that are still within the mileage limit at time x_q will be $\sum_{s=1}^{71} p_s$, where p_s is the probability that a vehicle belongs to stratum s.

Similarly, the estimated proportion of the $N_{q-1}^{(x)}$ vehicles sold within

$\Delta_{q-1}^{(x)} = (x_{q-2}, x_{q-1}]$ that are still within the mileage limit at time x_q is $\sum_{s=1}^{70} p_s$; the estimated proportion of the $N_{q-2}^{(x)}$ vehicles sold within $\Delta_{q-2}^{(x)} = (x_{q-3}, x_{q-2}]$ that are still within the mileage limit at time x_q is $\sum_{s=1}^{69} p_s$; etc. Lastly, the estimated proportion of the $N_{q-35}^{(x)}$ vehicles sold within $\Delta_{q-35}^{(x)} = (x_{q-36}, x_{q-35}]$ that are still within the mileage limit at time x_q is $\sum_{s=1}^{36} p_s$. Therefore, the adjusted for mileage estimator for $M(x_q)$ is

$$\hat{M}(x_q) = \begin{cases} \sum_{i=1}^{q} \left(N_{q-i+1}^{(x)} \times \sum_{s=1}^{72-i} p_s \right) & \text{if } q = 1, 2, \ldots, 35; \\ \sum_{i=1}^{36} \left(N_{q-i+1}^{(x)} \times \sum_{s=1}^{72-i} p_s \right) & \text{if } q = 36, 37, \ldots, n, \end{cases} \tag{7.12}$$

where p_s is the probability that a vehicle belongs to stratum s.

We can make different adjustments for mileage to each time-bin, and the resulting estimators for $M(x)$ are slightly different from above. Suppose we make the adjustment from the *right*-endpoint of each time-bin, that is, all vehicles in a particular time-bin are regarded as having the *youngest* possible age. Then the adjusted for mileage estimator for $M(x_q)$ becomes

$$\hat{M}(x_q) = \begin{cases} \sum_{i=1}^{q} \left(N_{q-i+1}^{(x)} \times \sum_{s=1}^{72-i+1} p_s \right) & \text{if } q = 1, 2, \ldots, 35; \\ \sum_{i=1}^{36} \left(N_{q-i+1}^{(x)} \times \sum_{s=1}^{72-i+1} p_s \right) & \text{if } q = 36, 37, \ldots, n. \end{cases} \tag{7.13}$$

Another adjustment, which is more conventional, will be to consider the *midpoint* of each time-bin. Then, the adjusted for mileage estimator for $M(x_q)$ becomes

$$\hat{M}(x_q) = \begin{cases} \sum_{i=1}^{q} \left[N_{q-i+1}^{(x)} \times \left(\sum_{s=1}^{72-i} p_s + \frac{1}{2} p_{72-i+1} \right) \right] & \text{if } q = 1, 2, \ldots, 35; \\ \sum_{i=1}^{36} \left[N_{q-i+1}^{(x)} \times \left(\sum_{s=1}^{72-i} p_s + \frac{1}{2} p_{72-i+1} \right) \right] & \text{if } q = 36, 37, \ldots, n. \end{cases} \tag{7.14}$$

Example I

In this example, we estimate the mean cumulative cost of P-claims per vehicle, $\hat{\Lambda}(x)$, for Dataset 2001 from $x = 1$ (starting at the first sale date, 22 May 2000) until $x = 42$ (which includes the "cut-off" date, 24 October 2003). Note that, for convenience, we write $x = x_q = 1, 2, \ldots, 42$ (month).

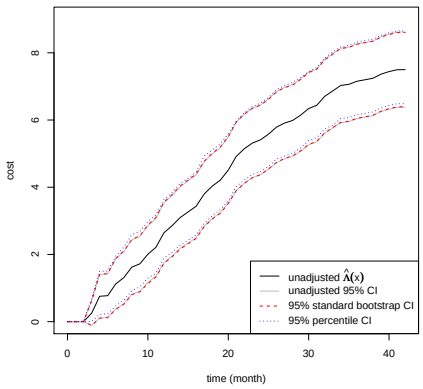

Figure 7.38: Unadjusted $\hat{\Lambda}(x)$ and 95% CI's

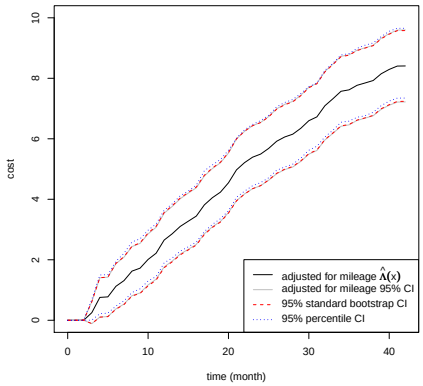

Figure 7.39: Adjusted for mileage $\hat{\Lambda}(x)$ and 95% CI's

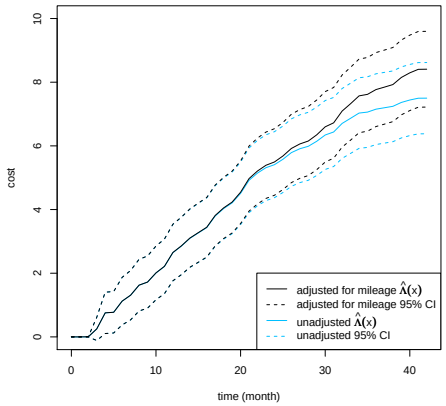

Figure 7.40: Unadjusted $\hat{\Lambda}(x)$ and adjusted for mileage $\hat{\Lambda}(x)$

Figure 7.38 shows the unadjusted $\hat{\Lambda}(x)$ and Figure 7.39 shows the adjusted for mileage $\hat{\Lambda}(x)$, along with the 95% confidence intervals (CI) evaluated using Eq. (5.5), the 95% standard bootstrap confidence intervals, and the 95% percentile confidence intervals. In both cases, we see that the three confidence intervals roughly agree. Then, Figure 7.40 illustrates the effect of the adjustment for withdrawals from warranty coverage due to exceeding the mileage limit of 36000 miles (by adjustment from the midpoint of each time-bin). Compared to the adjustment for mileage in the example for "time" is age case (see Section 7.3.1), this adjustment is less significant here. The adjusted for mileage curve still lies within the unadjusted 95% confidence interval, and there is a substantial overlap of the two corresponding 95% confidence intervals. Nevertheless, this adjustment is becoming more significant, and we would expect the adjusted curve to go across the unadjusted 95% confidence interval as time increases.

Figure 7.41 illustrates the difference between the three adjustments for mileage: from left-endpoint, from right-endpoint, and from the midpoint of each time-bin. It can be observed that the curve for adjustment from the left-endpoint is the highest, the curve for adjustment from the right-endpoint is the lowest, while the curve for adjustment from the midpoint lies between the two curves. Overall, the differences between the three curves are not significant. As a conservative approach, we chose to adopt the adjustment from the midpoint.

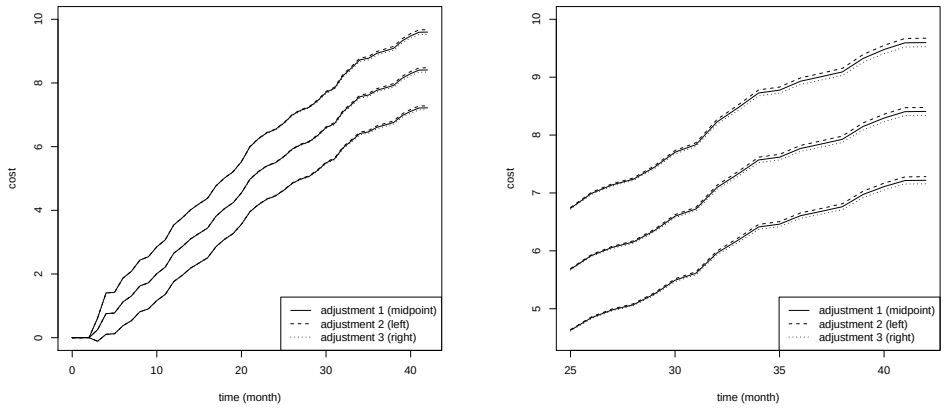

Figure 7.41: Different adjustments for mileage

Example II

Now, by using Datasets 1998 - 2001, we explore the relationship between the variability of driving pattern and the mean cumulative warranty cost (per vehicle). We will consider the adjusted for mileage $\hat{\Lambda}(x)$. Note that, instead of using the cost of P-claims only, we examine the total cost of all claims. Also, for each of the datasets, the set of vehicles with no claims is divided into each DPG according to the proportion of vehicles with claims in each DPG.

Let us consider the following DPG's: S, W_1, W_3, W_6, and U, as already defined. Figures 7.42 - 7.45 show the mean cumulative warranty cost for different DPG's for Datasets 1998 - 2001 respectively. In general, for each these datasets, we see that group W_1 has the lowest cost, followed by group W_3 and group W_6, while group U have the highest cost. For Dataset 2001, unlike in the previous examples, the cost for group W_1 is initially the lowest, and it only overtook group W_3 after $x = 35$.

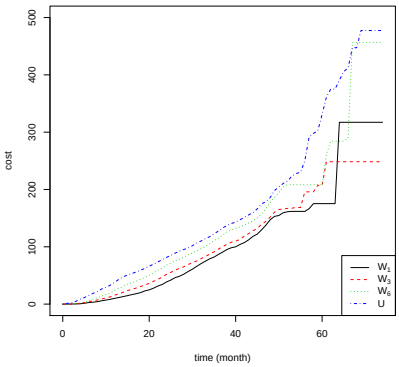

Figure 7.42: Mean cumulative warranty cost for different DPG's for Dataset 1998

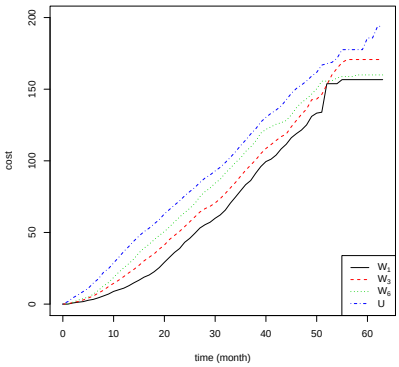

Figure 7.43: Mean cumulative warranty cost for different DPG's for Dataset 1999

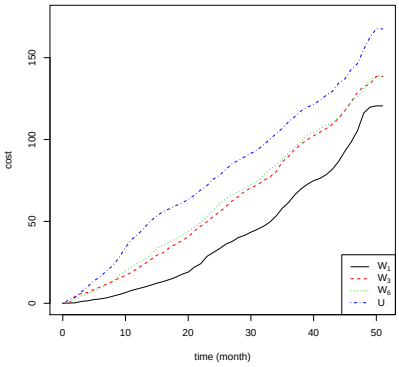

Figure 7.44: Mean cumulative warranty cost for different DPG's for Dataset 2000

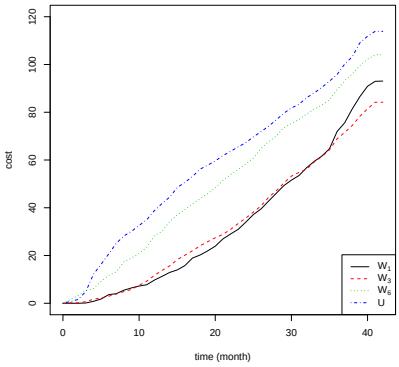

Figure 7.45: Mean cumulative warranty cost for different DPG's for Dataset 2001

Figures 7.46 - 7.49 show the mean cumulative warranty cost for different DPG's at $x = 36$ for Datasets 1998 - 2001 respectively. Note that each of these datasets has different length of time period. So, for consistency, we consider the mean cumulative warranty cost for different DPG's at $x = 36$. All of Figures 7.46 - 7.49 demonstrate an upward trend over DPG's with increasing variability of driving pattern. For further analysis,

we fit a trend line to each of these graphs (by using simple linear regression). For Datasets 1998 - 2001, the slopes of the trend lines are 14.1963, 10.6491, 14.9861, and 9.2276 respectively. All of these slopes are significant at the 10% level.

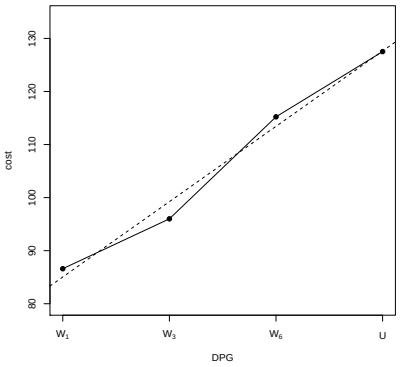

Figure 7.46: Mean cumulative warranty cost for different DPG's at $x = 36$ for Dataset 1998

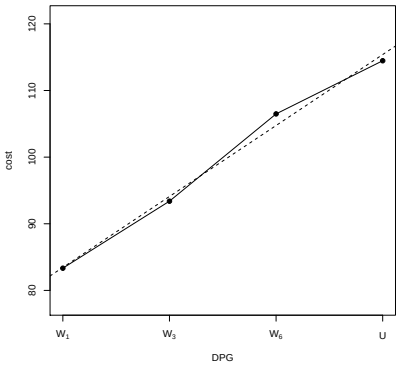

Figure 7.47: Mean cumulative warranty cost for different DPG's at $x = 36$ for Dataset 1999

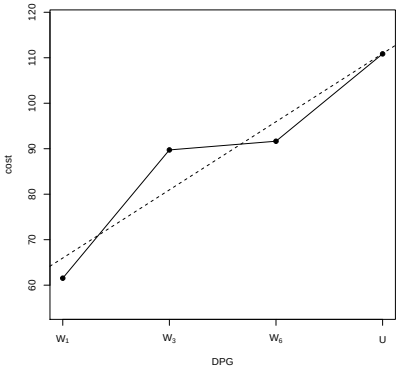

Figure 7.48: Mean cumulative warranty cost for different DPG's at $x = 36$ for Dataset 2000

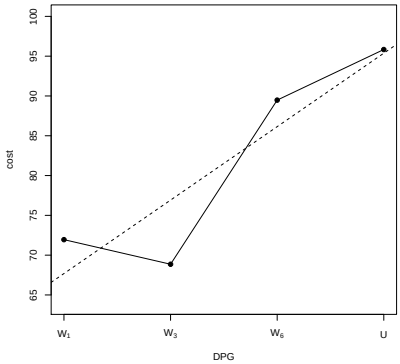

Figure 7.49: Mean cumulative warranty cost for different DPG's at $x = 36$ for Dataset 2001

Next, we investigate further the relationship between the variability of the driving pattern and the mean cumulative warranty cost by using a different definition of DPG's, with more groups as follows: $S, W_1, W_2, \ldots, W_{10}$, and U (with claims spread over more than 10 strata). Figures 7.50 - 7.53 show the mean cumulative warranty cost for different DPG's at $x = 36$ for Datasets 1998 - 2001 respectively. Again, all of these figures demonstrate an upward trend over DPG's with increasing variability of driving pattern. For Datasets 1998 - 2001, the slopes of the trend lines are 4.7529, 3.0578, 3.4205, and 2.6534 respectively. All of these slopes are significant at the 5% level.

Overall, the above results suggest that a higher variability of driving pattern leads to a higher mean cumulative warranty cost. Nevertheless, more study on this observation are required.

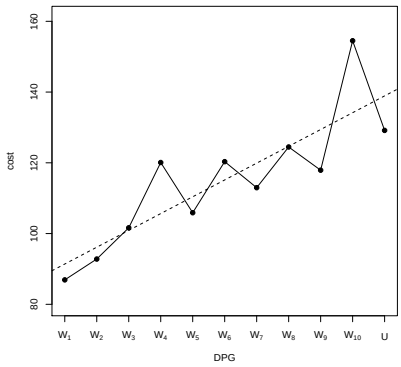

Figure 7.50: Mean cumulative warranty cost for different DPG's at $x = 36$ for Dataset 1998

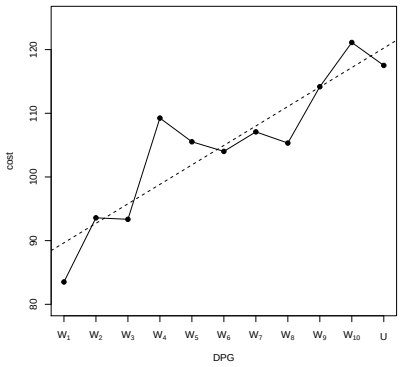

Figure 7.51: Mean cumulative warranty cost for different DPG's at $x = 36$ for Dataset 1999

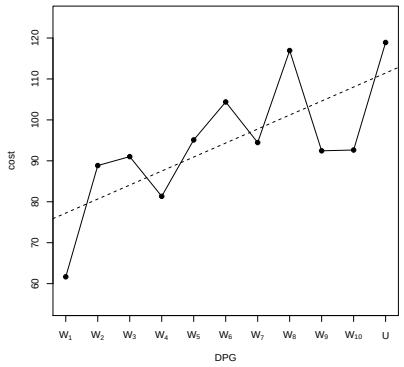

Figure 7.52: Mean cumulative warranty cost for different DPG's at $x = 36$ for Dataset 2000

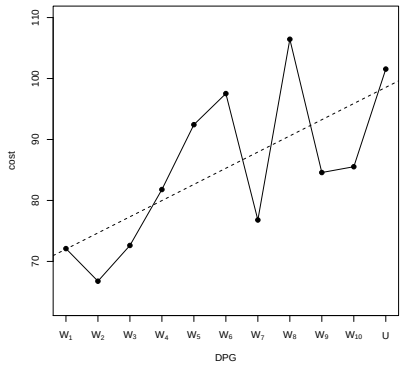

Figure 7.53: Mean cumulative warranty cost for different DPG's at $x = 36$ for Dataset 2001

7.5 Further Investigation

Since the driving pattern of a vehicle is determined by the odometer readings at the time of making a claim, the driving pattern identified may be dependent on the number of claims. Due to our modeling approach, it is possible that a vehicle with more claims is more likely to have higher variability of driving pattern compared to those with fewer claims, and hence has a higher warranty cost. Thus, to investigate further the relationship between the variability of driving pattern and the mean cumulative warranty cost, we examine the mean cumulative costs of all claims per vehicle for different DPG's by using Dataset 2006. In this dataset, each vehicle has some odometer readings which are not related to the time of making a claim. Thus, we will be able to characterize the driving pattern of a vehicle in a better way and to reduce the influence of the number of claims on the determination of driving pattern.

Now, let us consider the adjusted for mileage $\hat{\Lambda}(t)$, adjusted for age $\hat{\Lambda}(m)$, and adjusted for mileage $\hat{\Lambda}(x)$ for the following DPG's: $S, W_1, W_3, W_6,$ and U. Note that, instead of using the vehicles with claims only, we use all

vehicles in computing the strata distribution since all of them have some mileage information. Also, all vehicles are divided into each DPG based on their odometer readings, instead of claims. Figures 7.54 - 7.59 illustrate the results. In each case, we observe that group W_1 has the lowest cost, followed by group W_3 and group W_6, while group U have the highest cost. Then, in Figures 7.55, 7.57, and 7.59, each of the trend lines (fitted by simple linear regression) has slope 29.3359, 18.2823, and 23.9869 respectively. All of these slopes are significant at the 5% level.

The above results support our finding on the relationship between the variability of driving pattern and the mean cumulative warranty cost, where the mean cumulative warranty cost increases as the variability of driving pattern increases. Of course, these results are based on a single dataset only and more study are required.

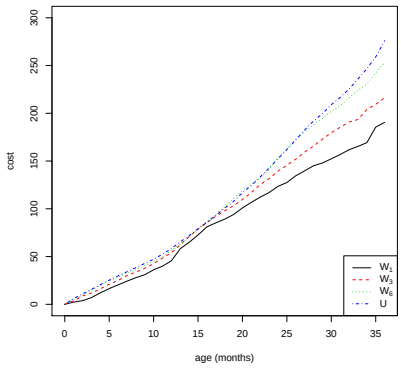

Figure 7.54: Adjusted for mileage $\hat{\Lambda}(t)$ per DPG for Dataset 2006

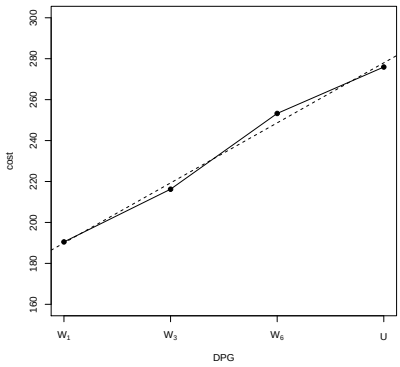

Figure 7.55: Adjusted for mileage $\hat{\Lambda}(t)$ per DPG at $t = 36$ months

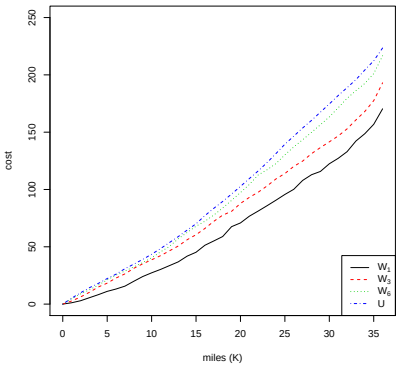

Figure 7.56: Adjusted for age $\hat{\Lambda}(m)$ per DPG for Dataset 2006

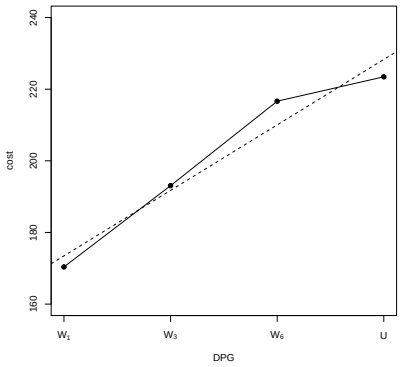

Figure 7.57: Adjusted for age $\hat{\Lambda}(m)$ per DPG at $m = 36K$ miles

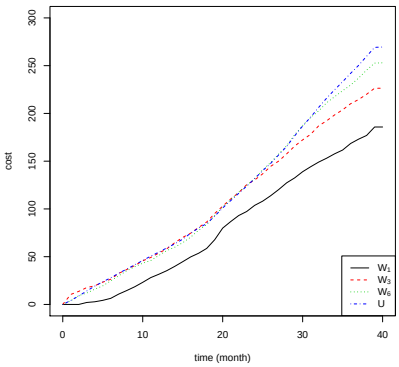

Figure 7.58: Adjusted for mileage $\hat{\Lambda}(x)$ per DPG for Dataset 2006

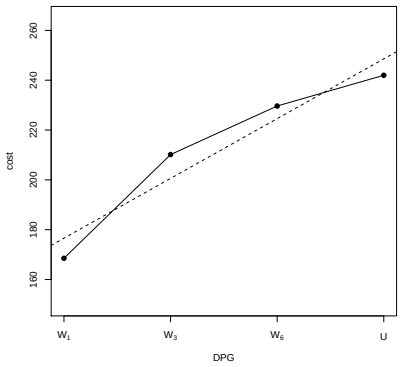

Figure 7.59: Adjusted for mileage $\hat{\Lambda}(x)$ per DPG at $x = 36$

7.6 Summary and Discussions

So far, we had estimated the mean cumulative cost of P-claims per vehicle by using two different approaches in modeling mileage accumulation. In Chapter 6, we made the assumption that vehicles accumulate mileage approximately linearly with their age. On the other hand, in this chapter, we

relaxed the linearity assumption of mileage accumulation and allows for variation in the rate of mileage accumulation over vehicle's lifetime using a piece-wise linear model.

We found that the results produced by these two approaches are quite similar. A possible reason for this similarity is that the driving pattern (or mileage accumulation pattern) for the majority of the vehicles are fairly stable with respect to a reasonably narrow range, and hence a linear approximation for their trajectories is satisfactory. From Table 7.1, we see that about 2/3 of the vehicles from each of the datasets, excluding those in group S, are stable with respect to an aggregated stratum consisting of six strata (with an angle of approximately 0.13 radian or 7.5 degree).

For the models introduced in this chapter, we estimate the mean cumulative warranty cost (or number of claims) per vehicle at "time" $t_j, j = 1, 2, \ldots, n,$ which define the regular partition. In the estimation procedure, the warranty cost at a particular "time" t_j actually includes the warranty cost for the interval $(t_{j-1}, t_j]$. As a result, the precision of the results produced depends on the chosen partition. The narrower the intervals are, the more precise the results are.

By considering the mean cumulative cost of all claims per vehicle for different DPG, we observe that a higher variability of driving pattern tends to result in a higher mean cumulative warranty cost. This is a very interesting finding, which needs to be investigated further. It suggests that we should take into account the variability of driving pattern in modeling mileage accumulation. Note that we had excluded group S in the analysis of this relationship, as it may not be sufficient to characterize a vehicle's driving pattern based on a single claim (or a single odometer reading).

Chapter 8

Estimating Bivariate Mean Cumulative Warranty Cost

In this chapter, we propose a new model for estimating bivariate mean cumulative warranty cost (or number of claims) per vehicle as a function of age and mileage. We assume all vehicles to have some mileage information (odometer readings), which may or may not be observed at the time of making a claim. Let us define a regular partition of age $0 = t_0 < t_1 < \ldots < t_{n-1} < t_n = l_a < \ldots$ and a regular partition of mileage $0 = m_0 < m_1 < \ldots < m_{n-1} < m_n = l_m < \ldots$, such that

$$t_p - t_{p-1} = h_a \quad p = 1, 2, \ldots, \quad \text{and} \quad m_q - m_{q-1} = h_m \quad q = 1, 2, \ldots,$$

where $h_a = l_a/n$ and $h_m = l_m/n$. Together, these two partitions form an age-mileage grid as in Figure 8.1. Then, let $n(t_p, m_q)$ denote the total warranty cost for cell $(t_p, m_q) = (t_{p-1}, t_p] \times (m_{q-1}, m_q]$ of the age-mileage grid for all vehicles. Also, let $N(t_p, m_q)$ be the number of vehicles in cell (t_p, m_q) of the age-mileage grid, i.e., the number of vehicles with age within $(t_{p-1}, t_p]$ and with accumulated mileage within $(m_{q-1}, m_q]$.

Now, let $\Lambda(t_p, m_q)$ be the mean cumulative warranty cost up to and including cell (t_p, m_q), with an initial condition $\Lambda(0, 0) = 0$. Then, the

corresponding rate function is

$$\lambda(t_p, m_q) = \Lambda(t_p, m_q) - \Lambda(t_{p-1}, m_q) - \Lambda(t_p, m_{q-1}) + \Lambda(t_{p-1}, m_{q-1}). \quad (8.1)$$

The rate function $\lambda(t_p, m_q)$ can be estimated by

$$\hat{\lambda}(t_p, m_q) = \frac{n(t_p, m_q)}{\hat{M}(t_p, m_q)}, \quad (8.2)$$

where $\hat{M}(t_p, m_q)$ denote the estimate for $M(t_p, m_q)$, the number of vehicles that are eligible to generate a claim in cell (t_p, m_q). Consequently, the mean cumulative function estimator is given by

$$\hat{\Lambda}(t_p, m_q) = \sum_{u=1}^{p} \sum_{v=1}^{q} \hat{\lambda}(t_u, m_v). \quad (8.3)$$

This is the natural extension of the (univariate) robust estimator in Chapter 5. The mathematical expression for the standard error of $\hat{\Lambda}(t_p, m_q)$ is still not available. Thus, we will evaluate the standard error of $\hat{\Lambda}(t_p, m_q)$ by bootstrap method (see Section 6.3).

Next, we consider the estimation of $M(t_p, m_q)$. We will use both linear and piece-wise linear approaches in modeling mileage accumulation.

8.1 Estimation of $M(t_p, m_q)$: Linear Approach

In this section, we introduce the first method to estimate $M(t_p, m_q)$, the number of vehicles that are eligible to generate a claim in cell (t_p, m_q). This method adopts the simple linear mileage accumulation model in Chapter 6, i.e., we assume that vehicles accumulate mileage approximately linearly with their age.

Let $r_i = \beta_i/\alpha_i$ be the mileage accumulation rate (MAR) of vehicle i, where α_i and β_i are the age and mileage of the vehicle at the latest odometer reading respectively. Once we know r_i for a vehicle, we can extrapolate

the accumulated mileage of the vehicle at a given age easily. For instance, the accumulated mileage for a vehicle with current age a_i and MAR r_i is extrapolated to be $a_i r_i$. Then, let A_i and B_i be the age-bin and mileage-bin that vehicle i belongs to at its current age. If the current age of a vehicle is within the age-bin $(t_{p-1}, t_p]$, then $A_i = p$. Similarly, if the extrapolated mileage of a vehicle is within the mileage-bin $(m_{q-1}, m_q]$, then $B_i = q$. Consequently, the number of vehicles in cell (t_p, m_q) can be estimated by

$$N(t_p, m_q) = \sum_{i=1}^{M} I(A_i = p) I(B_i = q). \tag{8.4}$$

To estimate $\hat{M}(t_p, m_q)$ using linear approach, there is no need to compute $N(t_p, m_q)$ explicitly. Suppose we ignore withdrawals from warranty coverage (due to age or mileage), and whether the trajectory of a vehicle actually goes through the target cell (t_p, m_q) or not, then $\hat{M}(t_p, m_q)$ is simply given by

$$\hat{M}(t_p, m_q) = \sum_{i=1}^{M} I(A_i \geq p) I(B_i \geq q). \tag{8.5}$$

This is the unadjusted estimator for $M(t_p, m_q)$. In order to adjust for withdrawals from warranty coverage and to count only those vehicles that really go through the target cell (t_p, m_q), $\hat{M}(t_p, m_q)$ becomes

$$\hat{M}(t_p, m_q) = \sum_{i=1}^{M} I(A_i \geq p) I(B_i \geq q) I\left(\frac{m_q - 1}{t_p} \leq r_i \leq \frac{m_q}{t_p - 1} \right). \tag{8.6}$$

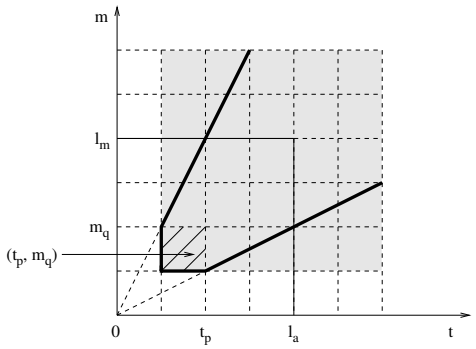

Figure 8.1: Age-mileage grid, unadjusted $\hat{M}(t_p, m_q)$ and adjusted $\hat{M}(t_p, m_q)$

Figure 8.1 illustrates the idea of estimating $\hat{M}(t_p, m_q)$. For the unadjusted case, $\hat{M}(t_p, m_q)$ is simply the number of vehicles with trajectory lies in the shaded region, i.e., the number of vehicles with age larger than t_{p-1} and with mileage larger than m_{q-1}. Then, to take into account both age and mileage warranty limit, and to consider only those vehicles that really go through the target cell (t_p, m_q), $\hat{M}(t_p, m_q)$ counts only those vehicles with trajectory lies in the shaded region surrounded by the dark solid line.

Example I

Let us define a set of age-bins with size of one month and a set of mileage-bins with size of 1000 miles. Then, by using Dataset 2006, we estimate the bivariate mean cumulative cost of P-claims per vehicle, $\Lambda(t, m)$, up to the age limit $l_a = 36$ months and the mileage limit $l_m = 36K$ miles. Note that, for convenience, we write $t = t_p = 1, 2, \ldots$ (in months) and $m = m_q = 1, 2, \ldots$ (in 1000 miles).

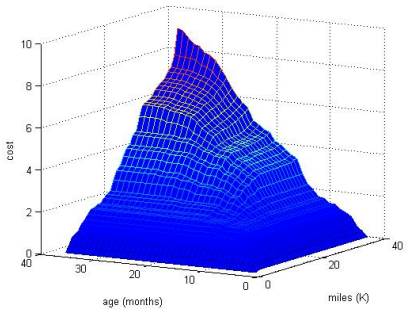

Figure 8.2: Unadjusted $\hat{\Lambda}(t, m)$

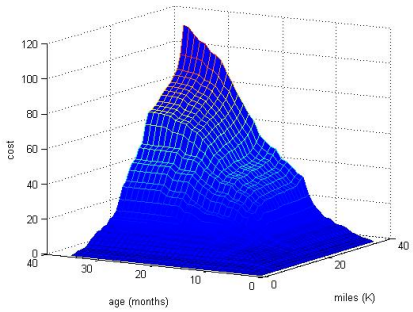

Figure 8.3: Adjusted $\hat{\Lambda}(t, m)$

Figure 8.2 shows the unadjusted $\hat{\Lambda}(t, m)$ and Figure 8.3 shows the adjusted $\hat{\Lambda}(t, m)$. Then, Figures 8.4 shows the comparisons between the unadjusted and adjusted $\hat{\Lambda}(t, m)$. It can be seen clearly that the adjusted $\hat{\Lambda}(t, m)$ is much larger than the unadjusted $\hat{\Lambda}(t, m)$ for all t and m, where the former completely lies above the latter. For $t = m = 36$, the adjusted $\hat{\Lambda}(36, 36)$ is about 12 times of the unadjusted $\hat{\Lambda}(36, 36)$.

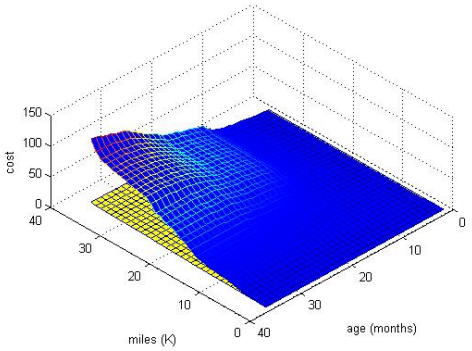

Figure 8.4: Unadjusted $\hat{\Lambda}(t, m)$ (lower) and adjusted $\hat{\Lambda}(t, m)$ (upper)

There is a huge difference between the unadjusted and adjusted results, because the adjusted $\hat{M}(t, m)$ is much lower than the unadjusted $\hat{M}(t, m)$. The unadjusted $\hat{M}(t, m)$ counts all of the vehicles with age larger than t and with mileage larger than m, no matter the vehicles actually go

through cell (t, m) or not. This leads to an overestimation of $M(t, m)$. On the other hand, the adjusted $\hat{M}(t, m)$ only counts those vehicles that actually go through cell (t, m). (See Figure 8.1.)

Example II

Now, we estimate the standard error of $\hat{\Lambda}(t, m)$ by bootstrap method, and then compute the 95% standard bootstrap confidence interval and 95% percentile confidence interval.

Firstly, we consider the unadjusted case. Figure 8.5 shows the unadjusted $\hat{\Lambda}(t, m)$ with its 95% standard bootstrap CI, and Figure 8.6 shows the unadjusted $\hat{\Lambda}(t, m)$ with its 95% percentile CI. It can be seen that the two types of CI's are very similar. For better illustration, Figure 8.7 shows the unadjusted $\hat{\Lambda}(t, 36)$ and the two types of 95% CI's, where we fix m at $36K$ miles. Similarly, Figure 8.8 shows the unadjusted $\hat{\Lambda}(36, m)$ and the two types of 95% CI's, where we fix t at 36 months. From these two figures, we see that the two types of CI's are roughly the same.

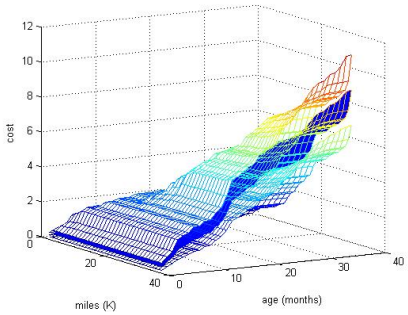

Figure 8.5: Unadjusted $\hat{\Lambda}(t, m)$ and 95% standard bootstrap CI

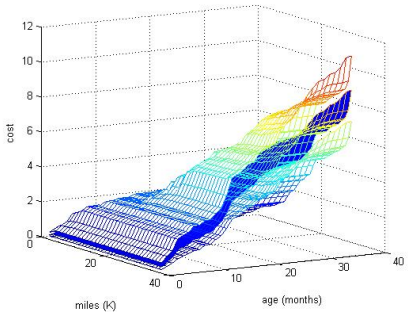

Figure 8.6: Unadjusted $\hat{\Lambda}(t, m)$ and 95% percentile CI

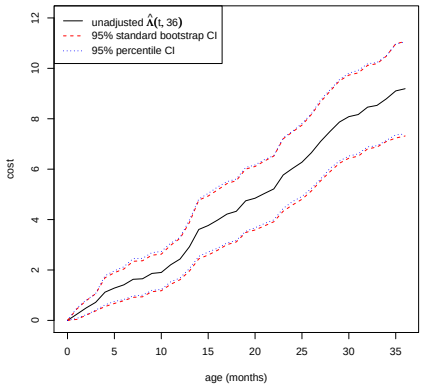

Figure 8.7: Unadjusted $\hat{\Lambda}(t, 36)$ and 95% CI's

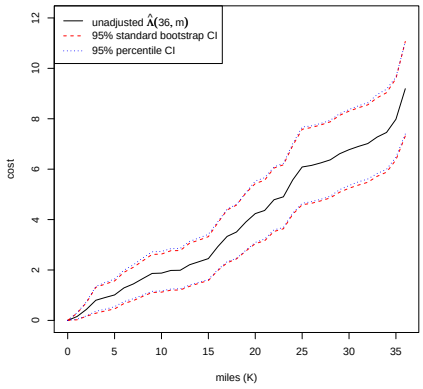

Figure 8.8: Unadjusted $\hat{\Lambda}(36, m)$ and 95% CI's

Next, we consider the adjusted case. Figure 8.9 shows the adjusted $\hat{\Lambda}(t, m)$ with its 95% standard bootstrap CI, and Figure 8.10 shows the adjusted $\hat{\Lambda}(t, m)$ with its 95% percentile CI. Again, it can be seen that the two types of CI's are very similar. For better illustration, Figure 8.11 shows the adjusted $\hat{\Lambda}(t, 36)$ and the two types of 95% CI's, where we fix m at $36K$ miles. Similarly, Figure 8.12 shows the adjusted $\hat{\Lambda}(36, m)$ and the two types of 95% CI's, where we fix t at 36 months. From these two figures, we see that the two types of CI's are nearly the same.

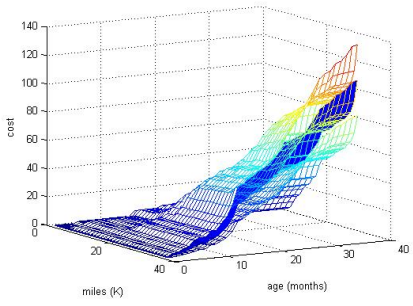

Figure 8.9: Adjusted $\hat{\Lambda}(t, m)$ and 95% standard bootstrap CI

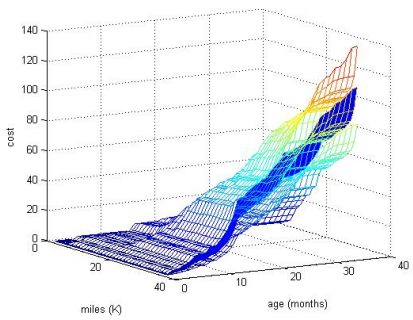

Figure 8.10: Adjusted $\hat{\Lambda}(t, m)$ and 95% percentile CI

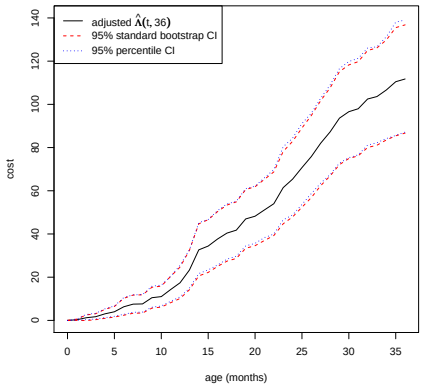

Figure 8.11: Adjusted $\hat{\Lambda}(t, 36)$ and 95% CI's

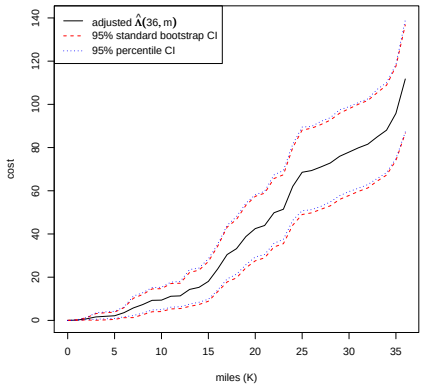

Figure 8.12: Adjusted $\hat{\Lambda}(36, m)$ and 95% CI's

8.2 Estimation of $M(t_p, m_q)$: Piece-Wise Linear Approach

Now, we introduce the second method to estimate $M(t_p, m_q)$, the number of vehicles eligible to generate a claim in cell (t_p, m_q), which uses a piece-wise linear approach in modeling mileage accumulation as in Chapter 7. To begin with, we partition the warranty coverage region into k strata, and evaluate the strata distribution according to the procedure in Chapter 7. Note that, instead of using the vehicles with claims only, we use all vehicles in computing the strata distribution since all of them have some mileage information. Also, all vehicles are divided into each DPG based on their odometer readings, instead of claims.

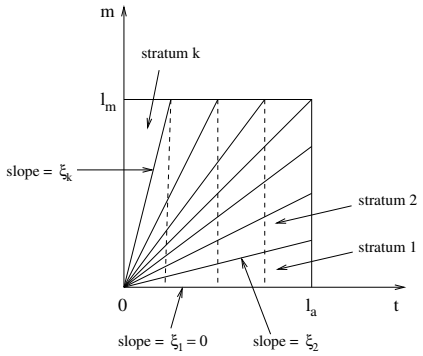

Figure 8.13: Age-strata grid

Let us consider an age-strata grid (s, t_i), determined by the stratum s for $s = 1, 2, \ldots, k$, and the age-bin $(t_{i-1}, t_i]$ for $i = 1, 2, \ldots, n, \ldots$, shown in Figure 8.13. Note that the age-strata grid can be extended beyond the warranty coverage region, if necessary. For each cell (s, t_i) of the age-strata grid, we identify a typical mileage representation, say $m(s, t_i)$. Then, we estimate the number of vehicles with current mileage equal to $m(s, t_i)$ by

$$N_{m(s,t_i)} = p_s N_i^{(t)}, \tag{8.7}$$

where p_s is the probability that a vehicle belongs to stratum s and $N_i^{(t)}$ is the number of vehicles with age within $(t_{i-1}, t_i]$. Subsequently, we estimate $N(t_p, m_q)$ by adding up the numbers of vehicles with typical mileage representation that fall within cell (t_p, m_q), i.e.,

$$N(t_p, m_q) = \sum_{m(s,t_p) \in (t_p, m_q)} N_{m(s,t_p)}. \tag{8.8}$$

If we ignore withdrawals from warranty coverage (due to age or mileage), and whether a vehicle actually goes through the target cell (t_p, m_q) or not,

$\hat{M}(t_p, m_q)$ is simply given by

$$\hat{M}(t_p, m_q) = \sum_{u \geq p} \sum_{v \geq q} N(t_u, m_v). \tag{8.9}$$

This is the unadjusted estimator for $M(t_p, m_q)$.

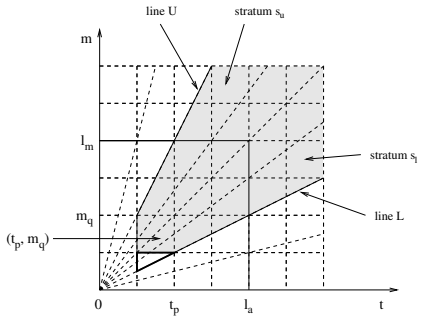

Figure 8.14: Adjusted $\hat{M}(t_p, m_q)$ - Case 1

Next, we consider the adjusted case which is more tricky. In the adjusted case, we only count those vehicles that go through the target cell (t_p, m_q), while they are still under warranty coverage. That is, we count only those vehicles that belong to the shaded area in Figure 8.14 or Figure 8.15. We need to consider two different cases. In Case 1, as illustrated by Figure 8.14, the lines L and U which (partly) define the shaded region, coincide with the boundary lines of the strata. Therefore, $M(t_p, m_q)$ can be estimated by the number of vehicles that are older than age t_{p-1} and belong to the aggregated stratum formed by strata $s_l, \ldots, s_u$, i.e.,

$$\hat{M}(t_p, m_q) = \left(\sum_{i \geq p} N_i^{(t)} \right) \left(\sum_{s=s_l}^{s_u} p_s \right), \tag{8.10}$$

where p_s is the probability that a vehicle belongs to stratum s and $N_i^{(t)}$ is the number of vehicles with age within $(t_{i-1}, t_i]$. Note that the estimator above includes vehicles that belong to the little triangle below the shaded

region. We assume the total number of these vehicles is negligible in our estimation of $M(t_p, m_q)$.

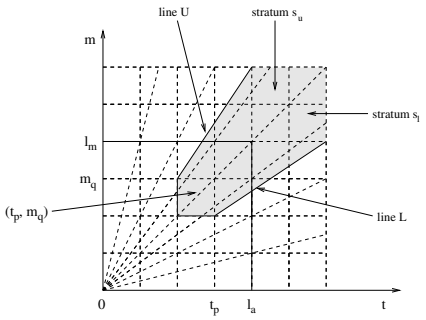

Figure 8.15: Adjusted $\hat{M}(t_p, m_q)$ - Case 2(a)

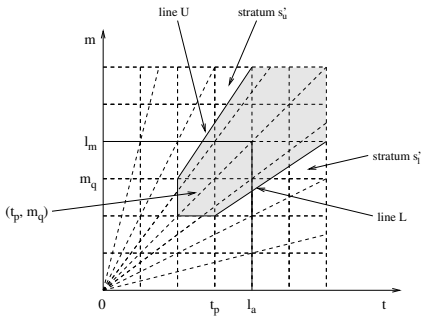

Figure 8.16: Adjusted $\hat{M}(t_p, m_q)$ - Case 2(b)

In Case 2, as illustrated by Figure 8.15, the lines L and U (or either one of them) do not coincide with the boundary lines of the strata. So, to estimate $M(t_p, m_q)$, we first evaluate a lower bound and an upper bound for $M(t_p, m_q)$, and then take the average value of these bounds as an estimate of $M(t_p, m_q)$. We can apply the technique used in Case 1 to obtain the lower and upper bounds for $M(t_p, m_q)$. That is,

- We estimate the lower bound for $M(t_p, m_q)$ by the number of vehicles that are older than age t_{p-1} and belong to the largest aggregated stratum (formed by strata $s_l, \ldots, s_u$) that is inside the shaded region, as shown in Figure 8.15.

- We estimate the upper bound for $M(t_p, m_q)$ by the number of vehicles that are older than age t_{p-1} and belong to the smallest aggregated stratum (formed by strata $s'_l, \ldots, s'_u$) that includes the shaded region, as shown in Figure 8.16.

Note that, the estimate of $M(t_p, m_q)$ obtained by taking the average value of the bounds is only a rough approximation. More precise procedure for estimating $M(t_p, m_q)$ could be developed, but it will require a significant increase in the complexity of the model.

The following proposition summarizes the estimators for $M(t_p, m_q)$ used in both Case 1 and Case 2:

Proposition 8.1. *Let $\xi_1, \xi_2, \ldots, \xi_k$ be the slopes of the boundary lines of the strata, as shown in Figure 8.13. Also, let R_L and R_U be the slopes of the lines L and U, as in Figures 8.14 and 8.15, respectively. Then,*

- *The estimate for $M(t_p, m_q)$ in Case 1 or the lower bound for $M(t_p, m_q)$ in Case 2 is given by*

$$\hat{M}_L(t_p, m_q) = \left(\sum_{i \geq p} N_i^{(t)} \right) \left(\sum_{s=s_l}^{s_u} p_s \right), \tag{8.11}$$

where

$$s_l = \begin{cases} 1 & \text{if } m_{q-1} = 0 \\ \min\{s : \xi_s \geq R_L\} & \text{if } m_{q-1} > 0, \end{cases}$$

$$s_u = \begin{cases} k & \text{if } t_{p-1} = 0 \\ \max\{s : \xi_s \leq R_U\} - 1 & \text{if } t_{p-1} > 0. \end{cases}$$

- *The estimate for $M(t_p, m_q)$ in Case 1 or the upper bound for $M(t_p, m_q)$ in Case 2 is given by*

$$\hat{M}_U(t_p, m_q) = \left(\sum_{i \geq p} N_i^{(t)} \right) \left(\sum_{s=s'_l}^{s'_u} p_s \right), \tag{8.12}$$

where

$$s'_l = \begin{cases} 1 & \text{if } m_{q-1} = 0 \\ \max\{s : \xi_s \leq R_L\} & \text{if } m_{q-1} > 0, \end{cases}$$

$$s'_u = \begin{cases} k & \text{if } t_{p-1} = 0 \\ \min\{s : \xi_s \geq R_U\} - 1 & \text{if } t_{p-1} > 0. \end{cases}$$

- *Consequently, the estimate of $M(t_p, m_q)$, for any of Case 1 and Case 2, is*

given by

$$\hat{M}(t_p, m_q) = \frac{1}{2}\left[\hat{M}_L(t_p, m_q) + \hat{M}_U(t_p, m_q)\right]. \tag{8.13}$$

For example, let us consider Figures 8.15 and 8.16, which involves $k = 8$ strata. Firstly, we obtain the lower bound for $M(t_p, m_q)$. Since t_{p-1} and m_{q-1} are both greater than zero, we have

$$s_l = \min\{s : \xi_s \geq R_L\} = 4 \quad \text{and} \quad s_u = \max\{s : \xi_s \leq R_U\} - 1 = 5.$$

Then, the lower bound for $M(t_p, m_q)$ is given by

$$\hat{M}_L(t_p, m_q) = \left(\sum_{i \geq p} N_i^{(t)}\right)\left(\sum_{s=4}^{5} p_s\right).$$

Next, we obtain the upper bound for $M(t_p, m_q)$. Since t_{p-1} and m_{q-1} are both greater than zero, we have

$$s_l' = \max\{s : \xi_s \leq R_L\} = 3 \quad \text{and} \quad s_u' = \min\{s : \xi_s \geq R_U\} - 1 = 6.$$

Then, the upper bound for $M(t_p, m_q)$ is given by

$$\hat{M}_U(t_p, m_q) = \left(\sum_{i \geq p} N_i^{(t)}\right)\left(\sum_{s=3}^{6} p_s\right).$$

Finally, the estimate of $M(t_p, m_q)$ is given by the average value of $\hat{M}_L(t_p, m_q)$ and $\hat{M}_U(t_p, m_q)$.

Example I

Let us define a set of age-bins with size of one month and a set of mileage-bins with size of 1000 miles. Then, by using Dataset 2006, we estimate

the bivariate mean cumulative cost of P-claims per vehicle, $\Lambda(t, m)$, up to the age limit $l_a = 36$ months and the mileage limit $l_m = 36K$ miles. Note that, for convenience, we write $t = t_p = 1, 2, \ldots$ (in months) and $m = m_q = 1, 2, \ldots$ (in 1000 miles). Also, the strata distribution is constructed by using the definition of driving pattern groups (DPG's) used in Chapter 7, i.e., with 5 DPG's: S, W_1, W_3, W_6, and U.

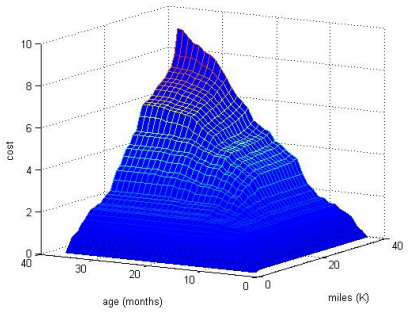

Figure 8.17: Unadjusted $\hat{\Lambda}(t, m)$

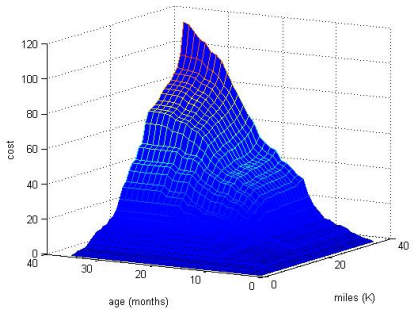

Figure 8.18: Adjusted $\hat{\Lambda}(t, m)$

Figure 8.17 shows the unadjusted $\hat{\Lambda}(t, m)$ and Figure 8.18 shows the adjusted $\hat{\Lambda}(t, m)$. Then, Figures 8.19 shows the comparisons between unadjusted and adjusted $\hat{\Lambda}(t, m)$. It can be seen clearly that the adjusted $\hat{\Lambda}(t, m)$ is much larger than the unadjusted $\hat{\Lambda}(t, m)$ for all t and m, where the former completely lies above the latter. Also, the results we obtained here are quite similar to the results in Example I for linear approach. For better comparison between the results produced by linear and piece-wise linear approaches, Table 8.1 shows the values of unadjusted and adjusted $\hat{\Lambda}(t, t)$ for $t = m = 6, 12, 18, 24, 30, 36$ produced by these two approaches. In both unadjusted and adjusted cases, we see that the estimates produced by these two approaches are very similar.

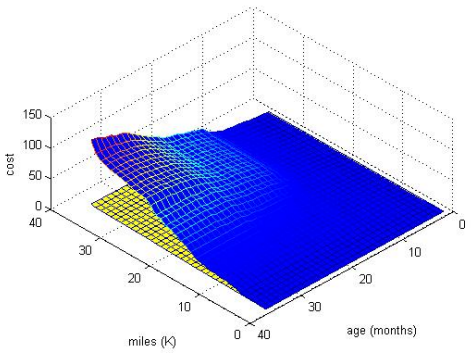

Figure 8.19: Unadjusted $\hat{\Lambda}(t, m)$ (lower) and adjusted $\hat{\Lambda}(t, m)$ (upper)

	Unadjusted $\hat{\Lambda}(t,t)$		Adjusted $\hat{\Lambda}(t,t)$	
t	Linear	PWL	Linear	PWL
6	1.2583	1.2584	3.3958	3.3778
12	1.8266	1.8265	7.5152	7.7210
18	2.9375	2.9393	19.1413	18.9235
24	4.7921	4.7986	44.7756	45.2015
30	6.5526	6.5611	73.2004	74.2236
36	9.1919	9.2010	111.7453	113.4169

Table 8.1: Unadjusted and adjusted $\hat{\Lambda}(t, t)$ for $t = m = 6, 12, 18, 24, 30, 36$, produced by linear and piece-wise linear (PWL) approaches

Example II

Next, we estimate the standard error of $\hat{\Lambda}(t, m)$ by bootstrap method, and then compute the 95% standard bootstrap confidence interval and 95% percentile confidence interval.

Firstly, we consider the unadjusted case. Figure 8.20 shows the unadjusted $\hat{\Lambda}(t, m)$ with its 95% standard bootstrap CI, and Figure 8.21 shows the unadjusted $\hat{\Lambda}(t, m)$ with its 95% percentile CI. It can be seen that the two types of CI's are very similar. For better illustration, Figure 8.22 shows

the unadjusted $\hat{\Lambda}(t, 36)$ and the two types of 95% CI's, where we fix m at $36K$ miles. Similarly, Figure 8.23 shows the unadjusted $\hat{\Lambda}(36, m)$ and the two types of 95% CI's, where we fix t at 36 months. From these two figures, we see that the two types of CI's are roughly the same.

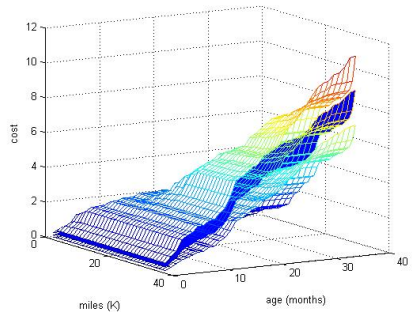

Figure 8.20: Unadjusted $\hat{\Lambda}(t, m)$ and 95% standard bootstrap CI

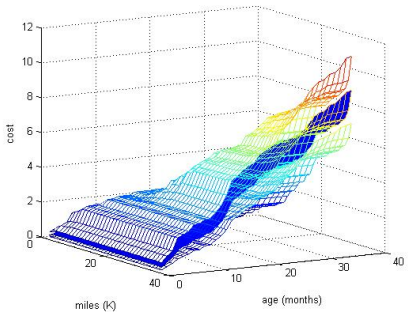

Figure 8.21: Unadjusted $\hat{\Lambda}(t, m)$ and 95% percentile CI

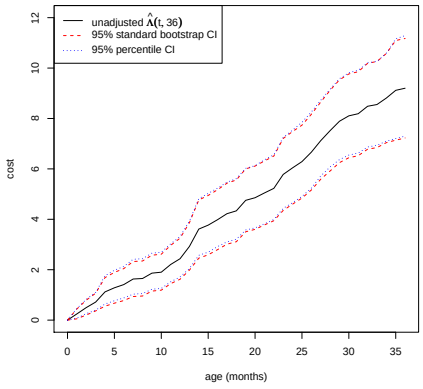

Figure 8.22: Unadjusted $\hat{\Lambda}(t, 36)$ and 95% CI's

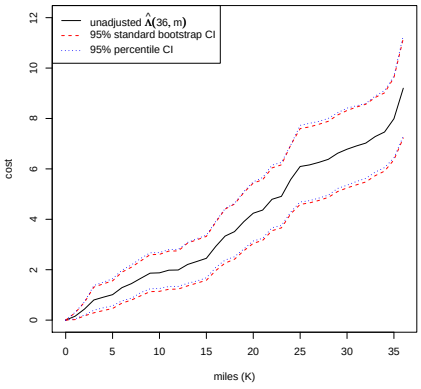

Figure 8.23: Unadjusted $\hat{\Lambda}(36, m)$ and 95% CI's

Next, we consider the adjusted case. Figure 8.24 shows the adjusted $\hat{\Lambda}(t, m)$ with its 95% standard bootstrap CI, and Figure 8.25 shows the adjusted $\hat{\Lambda}(t, m)$ with its 95% percentile CI. Again, it can be seen that the two

types of CI's are very similar. Then, Figure 8.26 shows the adjusted $\hat{\Lambda}(t, 36)$ and the two types of 95% CI's, where we fix m at $36K$ miles. Similarly, Figure 8.27 shows the adjusted $\hat{\Lambda}(36, m)$ and the two types of 95% CI's, where we fix t at 36 months. From these two figures, we see that the two types of CI's are nearly the same.

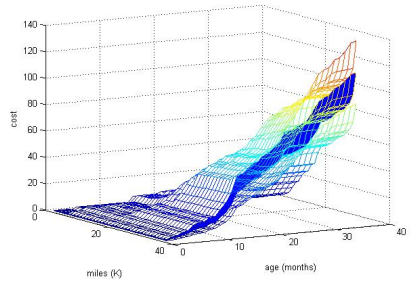

Figure 8.24: Adjusted $\hat{\Lambda}(t, m)$ and 95% standard bootstrap CI

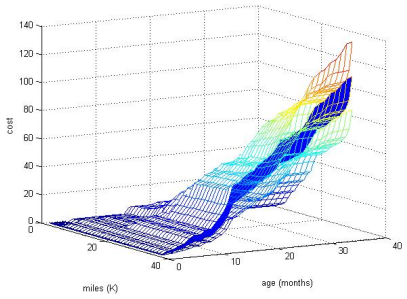

Figure 8.25: Adjusted $\hat{\Lambda}(t, m)$ and 95% percentile CI

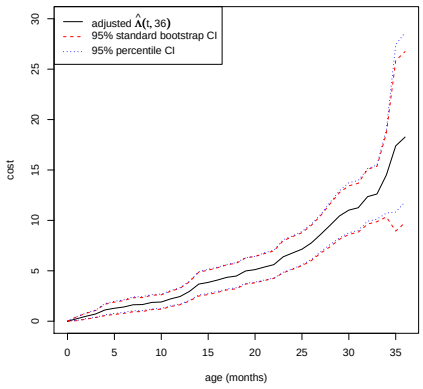

Figure 8.26: Adjusted $\hat{\Lambda}(t, 36)$ and 95% CI's

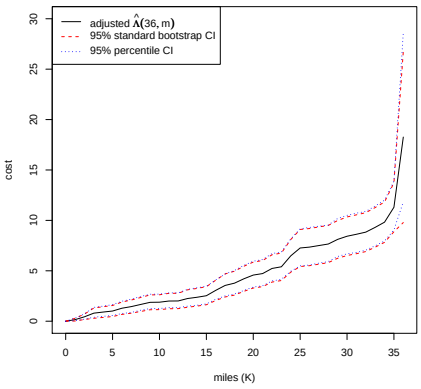

Figure 8.27: Adjusted $\hat{\Lambda}(36, m)$ and 95% CI's

8.3 Univariate Mean Cumulative Warranty Cost for Intervals

In this section, we consider two types of estimators for estimating the univariate mean cumulative warranty cost for intervals as follows:

- the univariate estimators associated with the bivariate model, and
- the direct univariate estimators.

8.3.1 Univariate Estimator associated with the Bivariate Model

Associate with the bivariate rate function $\lambda(t_p, m_q)$, there are two univariate rate functions $\lambda_1(t_p)$ and $\lambda_2(m_q)$, where

- $\lambda_1(t_p)$ is the mean warranty cost per vehicle for age-bin $(t_{p-1}, t_p]$, and
- $\lambda_2(m_q)$ is the mean warranty cost per vehicle for mileage-bin $(m_{q-1}, m_q]$.

These functions are given by

$$\lambda_1(t_p) = \sum_q \lambda(t_p, m_q) \quad \text{and} \quad \lambda_2(m_q) = \sum_p \lambda(t_p, m_q), \tag{8.14}$$

respectively. The above idea is similar to the idea of obtaining the marginal distribution functions from a bivariate distribution function. To obtain $\lambda_1(t_p)$, we sum up $\lambda(t_p, m_q)$ for all m_q. Similarly, to obtain $\lambda_2(m_q)$, we sum up $\lambda(t_p, m_q)$ for all t_p. Then, the corresponding mean cumulative functions are given by

$$\Lambda_1(t_p) = \sum_{j=1}^{p} \lambda_1(t_j) \quad \text{and} \quad \Lambda_2(m_q) = \sum_{j=1}^{q} \lambda_2(m_j), \tag{8.15}$$

respectively. Note that $\hat{\Lambda}_1(t_p)$ is equal to $\hat{\Lambda}(t_p, l_m)$ and $\hat{\Lambda}_2(m_q)$ is equal to $\hat{\Lambda}(l_a, m_q)$, provided no claim has occurred outside the warranty coverage region.

8.3.2 Direct Univariate Estimator

Now, we consider the direct estimators for estimating the univariate mean cumulative warranty cost over an age interval or mileage interval. Let

- $\Lambda(t_p)$ be the mean cumulative function of warranty cost up to and including age-bin $(t_{p-1}, t_p]$, and
- $\Lambda(m_q)$ be the mean cumulative function of warranty cost up to and including mileage-bin $(m_{q-1}, m_q]$.

Also, let $\lambda(t_p)$ and $\lambda(m_q)$ be the corresponding rate functions, $n(t_p)$ be the total warranty cost for age-bin $(t_{p-1}, t_p]$, and $n(m_q)$ be the total warranty cost for mileage-bin $(m_{q-1}, m_q]$.

To estimate $\Lambda(t_p)$ and $\Lambda(m_q)$, we use the robust estimator in Chapter 5, with the intervals taken as the units of "time". Then, for the "time" is age case, the rate function can be estimated by

$$\hat{\lambda}(t_p) = \frac{n(t_p)}{\hat{M}(t_p)}, \tag{8.16}$$

where $\hat{M}(t_p)$ is the estimated number of vehicles that are eligible to generate a claim in $(t_{p-1}, t_p]$. Consequently, the associate mean cumulative function estimator is given by

$$\hat{\Lambda}(t_p) = \sum_{j=1}^{p} \hat{\lambda}(t_j). \tag{8.17}$$

Similarly, for the "time" is mileage case, the rate function can be estimated by

$$\hat{\lambda}(m_q) = \frac{n(m_q)}{\hat{M}(m_q)}, \tag{8.18}$$

where $\hat{M}(m_q)$ is the estimated number of vehicles that are eligible to generate a claim in $(m_{q-1}, m_q]$. Consequently, the associate mean cumulative

function estimator is given by

$$\hat{\Lambda}(m_q) = \sum_{j=1}^{q} \hat{\lambda}(m_j). \tag{8.19}$$

The standard errors of $\hat{\Lambda}(t_p)$ and $\hat{\Lambda}(m_q)$ can be estimated by Eq. (5.5), with $M(t_p)$ replaced by $\hat{M}(t_p)$ and $M(m_q)$ replaced by $\hat{M}(m_q)$, respectively.

We can compute $\hat{M}(t_p)$ and $\hat{M}(m_q)$ by using a linear or piece-wise linear approach in modeling mileage accumulation. Here, we introduce the estimators for $\hat{M}(t_p)$ and $\hat{M}(m_q)$, which are modified from the CR-Model (linear approach) in Chapter 6 and the CCR-Model (piece-wise linear approach) in Chapter 7. We assume all vehicles to have some mileage information.

Linear Approach

Firstly, we consider the linear approach. For the "time" is age case, if we ignore withdrawals from warranty coverage due to mileage, then the number of vehicles that are eligible to generate a claim in $(t_{p-1}, t_p]$, $M(t_p)$, can be estimated by

$$\hat{M}(t_p) = \sum_{i=1}^{M} I(A_i \geq p). \tag{8.20}$$

To adjust for withdrawals due to mileage, $\hat{M}(t_p)$ becomes

$$\hat{M}(t_p) = \sum_{i=1}^{M} I(A_i \geq p) I\left(r_i \leq \frac{l_m}{t_p - 1}\right). \tag{8.21}$$

For the "time" is mileage case, if we ignore withdrawals from warranty coverage due to age, then the number of vehicles that are eligible to gen-

erate a claim in $(m_{q-1}, m_q]$, $M(m_q)$, can be estimated by

$$\hat{M}(m_q) = \sum_{i=1}^{M} I(B_i \geq q). \tag{8.22}$$

To adjust for withdrawals due to age, $\hat{M}(m_q)$ becomes

$$\hat{M}(m_q) = \sum_{i=1}^{M} I(B_i \geq q) I\left(r_i \geq \frac{m_{q-1}}{l_a}\right) = \sum_{i=1}^{M} I\left(r_i \geq \frac{m_{q-1}}{\min(a_i, l_a)}\right), \tag{8.23}$$

where a_i is the current age of vehicle i.

Piece-Wise Linear Approach

Next, we consider the piece-wise linear approach. For the "time" is age case, if we ignore withdrawals from warranty coverage due to the mileage, then the number of vehicles that are eligible to generate a claim in $(t_{p-1}, t_p]$, $M(t_p)$, can be estimated by

$$\hat{M}(t_p) = \sum_{j \geq p} N_j^{(t)}. \tag{8.24}$$

Then, to adjust for withdrawals due to the mileage, $\hat{M}(t_p)$ becomes

$$\hat{M}(t_p) = \left(\sum_{j \geq p} N_j^{(t)}\right) \left(\sum_{s=1}^{k-p+1} p_s\right), \tag{8.25}$$

where p_s is the probability that a vehicle belongs to stratum s.

For the "time" is mileage case, if we ignore withdrawals from warranty coverage due to the age, then the number of vehicles that are eligible to generate a claim in $(m_{q-1}, m_q]$, $M(m_q)$, can be estimated by

$$\hat{M}(m_q) = \sum_{j \geq q} N_j^{(m)}, \tag{8.26}$$

where

$$N_j^{(m)} = \sum_i N(t_i, m_j) \tag{8.27}$$

is the number of vehicles with mileage within $(m_{j-1}, m_j]$. Then, to adjust for withdrawals due to the age, $\hat{M}(m_q)$ becomes

$$\hat{M}(m_q) = \sum_{j \geq q} \tilde{N}_j^{(m)} + \left(M - \tilde{M}\right) \sum_{s=q}^{k} p_s, \tag{8.28}$$

where $\tilde{M}$ is the number of vehicles with age within the age limit, and

$$\tilde{N}_j^{(m)} = \sum_{i=1}^{n} N(t_i, m_j). \tag{8.29}$$

is the number of vehicles with mileage within $(m_{j-1}, m_j]$ and with age within the age limit.

8.3.3 Discussions

Intuitively, one might think that $\hat{\lambda}(t_p) = \hat{\lambda}_1(t_p)$ and $\hat{\lambda}(m_q) = \hat{\lambda}_2(m_q)$. Unfortunately, these are *not* true because

$$\begin{aligned}
\hat{\lambda}(t_p) &= \sum_q \frac{n(t_p, m_q)}{\hat{M}(t_p)} \\
&= \sum_q \left[\frac{n(t_p, m_q)}{\hat{M}(t_p, m_q)} \times \frac{\hat{M}(t_p, m_q)}{\hat{M}(t_p)} \right] \\
&= \sum_q \left[\hat{\lambda}(t_p, m_q) \times \frac{\hat{M}(t_p, m_q)}{\hat{M}(t_p)} \right] \\
&\neq \sum_q \hat{\lambda}(t_p, m_q),
\end{aligned}$$

and similarly for $\hat{\lambda}(m_q)$. Moreover, it can be shown that

$$\hat{\lambda}(t_p) \leq \hat{\lambda}_1(t_p) \quad \text{and} \quad \hat{\lambda}(m_q) \leq \hat{\lambda}_2(m_q), \tag{8.30}$$

since $\hat{M}(t_p) \geq \hat{M}(t_p, m_q)$ and $\hat{M}(m_q) \geq \hat{M}(t_p, m_q)$.

By comparing the definitions of $\hat{\lambda}(t_p)$ and $\hat{\lambda}(m_q)$ with the definitions for $\hat{\lambda}_1(t_p)$ and $\hat{\lambda}_2(m_q)$. We see that $\hat{\lambda}_1(t_p)$ and $\hat{\lambda}_2(m_q)$ will be the more reliable and accurate estimators for the mean cumulative warranty cost. To explain why, let us consider the warranty cost $n(t_p, m_q)$ for cell (t_p, m_q). For $\hat{\lambda}_1(t_p)$ and $\hat{\lambda}_2(m_q)$, $n(t_p, m_q)$ is averaged over the estimated number of vehicles that are eligible to generate a claim in that cell, $\hat{M}(t_p, m_q)$. On the other hand, for $\hat{\lambda}(t_p)$ and $\hat{\lambda}(m_q)$, $n(t_p, m_q)$ is averaged over $\hat{M}(t_p)$ and $\hat{M}(m_q)$ respectively, which include vehicles with trajectories that do not go through cell (t_p, m_q). Hence, the mean cumulative warranty cost is underestimated.

Example I

In this example, we estimate the mean cumulative cost of P-claims per vehicle over an age interval and a mileage interval, $\Lambda(t)$ and $\Lambda(m)$, by using Dataset 2006. We use both linear and piece-wise linear approaches in modeling mileage accumulation. For convenience, we write $t = t_p = 1, 2, \ldots$ (in months) and $m = m_q = 1, 2, \ldots$ (in 1000 miles).

Figure 8.28 and Figure 8.29 show the results for the linear approach. For the "time" is age case, we see that the unadjusted $\hat{\Lambda}(t)$ and the adjusted for mileage $\hat{\lambda}(t)$ are initially similar, but the difference between the two estimates increases as t increases. There is a kink at $t = 22$ months, where the adjusted for mileage $\hat{\lambda}(t)$ starts to increase at a higher rate. For the "time" is mileage case, the unadjusted $\hat{\Lambda}(m)$ and adjusted for age $\hat{\Lambda}(m)$ are nearly identical for all m. The adjustment for age has very little impacts here. This is because the adjustment for age does not begin until the oldest vehicle exceeds the age limit, which is three years from the first sale in our dataset. Also, the majority of the vehicles in the dataset are estimated

to leave coverage due to mileage, instead of age. Based on the empirical distribution of mileage accumulation rate (not shown), approximately 62% of the vehicles in Dataset 2006 are anticipated to leave warranty coverage due to exceeding the mileage limit, before they reach the age limit.

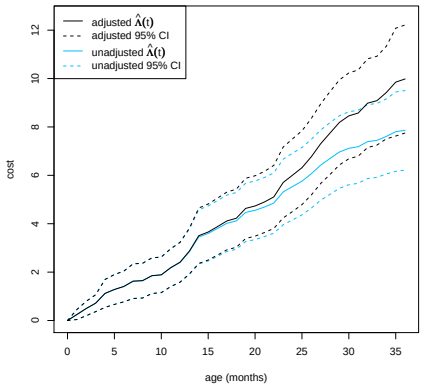

Figure 8.28: Unadjusted $\hat{\Lambda}(t)$ and adjusted for mileage $\hat{\Lambda}(t)$, by linear approach

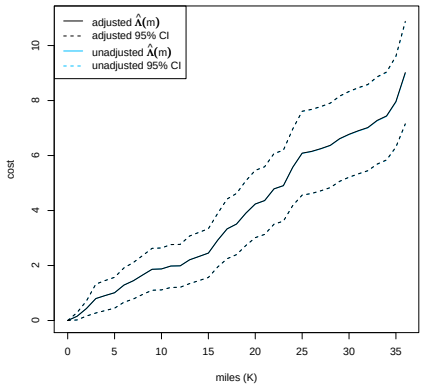

Figure 8.29: Unadjusted $\hat{\Lambda}(m)$ and adjusted for age $\hat{\Lambda}(m)$, by linear approach

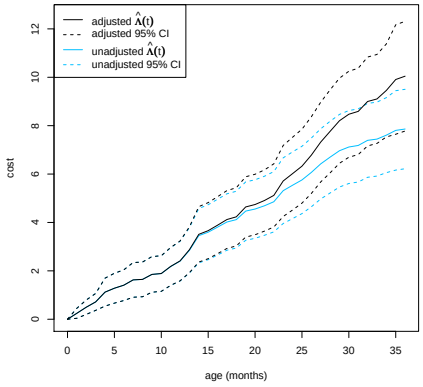

Figure 8.30: Unadjusted $\hat{\lambda}(t)$ and adjusted for mileage $\hat{\lambda}(t)$, by piecewise linear approach

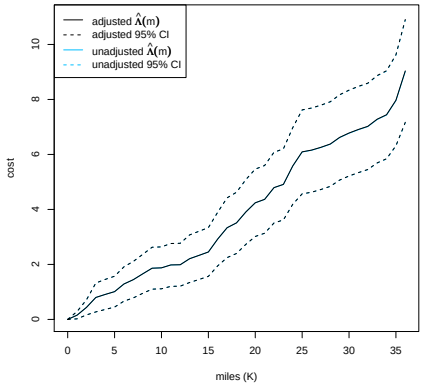

Figure 8.31: Unadjusted $\hat{\Lambda}(m)$ and adjusted for age $\hat{\Lambda}(m)$, by piecewise linear approach

Then, Figure 8.30 and Figure 8.31 show the results for the piece-wise linear approach. It can be seen that the results obtained by linear and piece-wise linear approaches are very similar.

Example II

Now, using the results obtained by linear approach, let us compare the direct univariate estimators, $\hat{\Lambda}(t)$ and $\hat{\lambda}(m)$ from the previous example, with the univariate estimators associated with the bivariate model, $\hat{\lambda}_1(t)$ and $\hat{\lambda}_2(m)$ (following from Example I of Section 8.1).

Firstly, we consider the unadjusted case. Figure 8.32 shows the unadjusted $\hat{\lambda}(t)$ and $\hat{\lambda}_1(t)$, and Figure 8.33 shows the corresponding mean cumulative functions $\hat{\Lambda}(t)$ and $\hat{\Lambda}_1(t)$. It can be seen that $\hat{\lambda}(t)$ and $\hat{\lambda}_1(t)$ are initially similar, with $\hat{\lambda}_1(t)$ being slightly greater than $\hat{\lambda}(t)$. Then, the difference between the two estimates becomes larger at older age. Consequently, $\hat{\Lambda}(t)$ and $\hat{\Lambda}_1(t)$ are also similar initially and then the difference between the two estimates increases as t increases, with $\hat{\Lambda}_1(t)$ being greater than $\hat{\Lambda}(t)$.

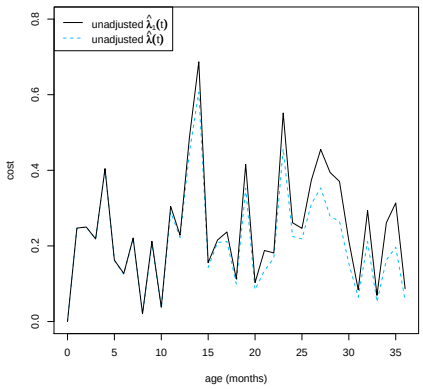

Figure 8.32: Unadjusted $\hat{\lambda}(t)$ and $\hat{\lambda}_1(t)$

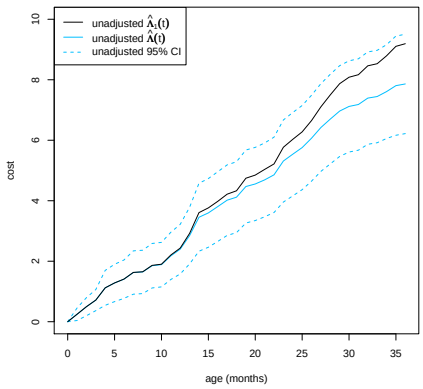

Figure 8.33: Unadjusted $\hat{\Lambda}(t)$ and $\hat{\Lambda}_1(t)$

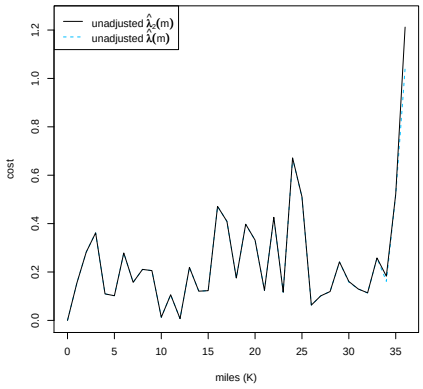

Figure 8.34: Unadjusted $\hat{\lambda}(m)$ and $\hat{\lambda}_2(m)$

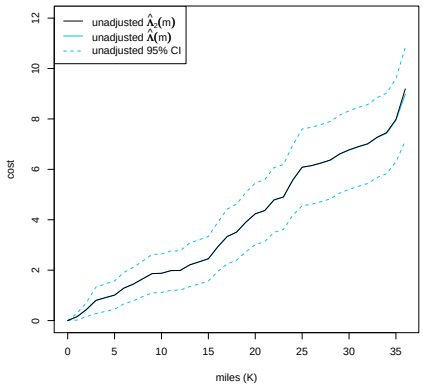

Figure 8.35: Unadjusted $\hat{\Lambda}(m)$ and $\hat{\Lambda}_2(m)$

Then, Figure 8.34 shows the unadjusted $\hat{\lambda}(m)$ and $\hat{\lambda}_2(m)$, and Figure 8.35 shows the corresponding mean cumulative functions $\hat{\Lambda}(m)$ and $\hat{\Lambda}_2(m)$. It can be seen that $\hat{\lambda}(m)$ and $\hat{\lambda}_2(m)$ are very close to each other, but $\hat{\lambda}_2(m)$ is slightly greater than $\hat{\lambda}(m)$. Consequently, $\hat{\Lambda}(m)$ is also slightly greater than $\hat{\Lambda}_2(m)$.

Next, we consider the adjusted case. Figures 8.36 - 8.39 illustrate the results. We observe a similar pattern, where $\hat{\lambda}_1(t)$ is greater than $\hat{\lambda}(t)$, and $\hat{\lambda}_2(m)$ is greater than $\hat{\lambda}(m)$. In addition, the difference between $\hat{\lambda}(t)$ and $\hat{\lambda}_1(t)$, as well as the difference between $\hat{\lambda}(m)$ and $\hat{\lambda}_2(m)$, are much larger than the corresponding differences in the unadjusted case.

From Figures 8.34 and 8.38, we can observe one common feature for all $\hat{\lambda}(m)$ and $\hat{\lambda}_2(m)$, no matter they are adjusted or not. That is, there exists a "spike" near the mileage limit of $36K$ miles, where the warranty cost is much higher. This interesting observation may be attributed to the existence of "customer-rush near the warranty expiration limit", which may result in a relatively high number of claims and hence a high warranty cost near the warranty expiration limit. Such a phenomenon may occur as a result of soft failures, where the vehicle users delay failure reporting until warranty is about to expire [Rai and Singh, 2004]. Note that, this

phenomenon is not detected in "time" is age case, probably because the majority of the vehicles leave warranty coverage due to mileage, instead of age.

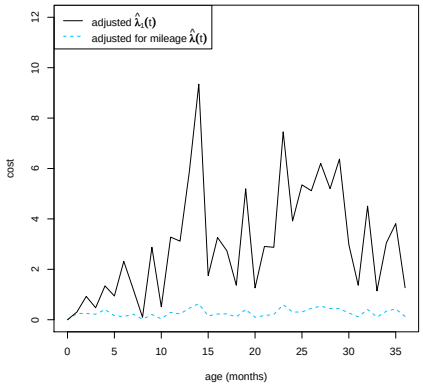

Figure 8.36: Adjusted for mileage $\hat{\lambda}(t)$ and adjusted $\hat{\lambda}_1(t)$

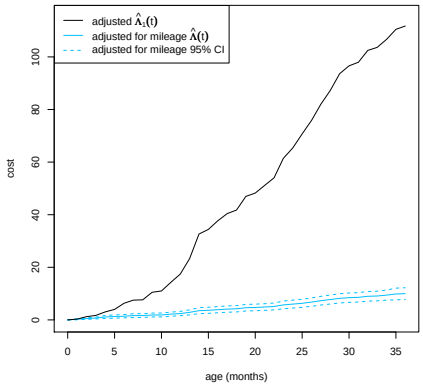

Figure 8.37: Adjusted for mileage $\hat{\Lambda}(t)$ and adjusted $\hat{\Lambda}_1(t)$

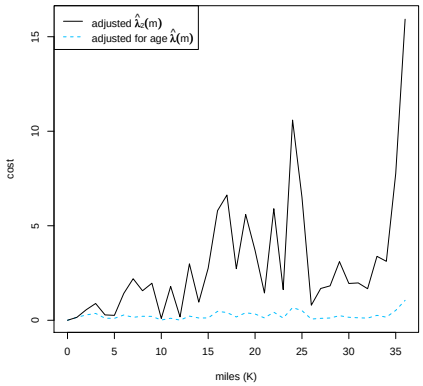

Figure 8.38: Adjusted for age $\hat{\lambda}(m)$ and adjusted $\hat{\lambda}_2(m)$

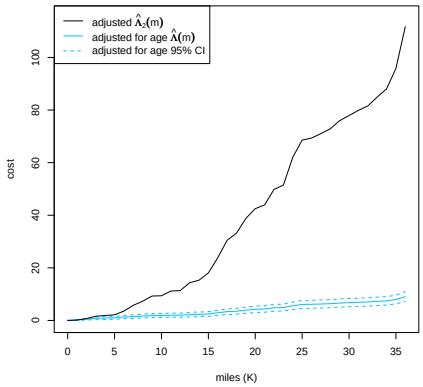

Figure 8.39: Adjusted for age $\hat{\Lambda}(m)$ and adjusted $\hat{\Lambda}_2(m)$

8.4 Summary and Discussions

In this chapter, we had proposed a bivariate model for estimating the mean cumulative warranty cost as a function of age and mileage. To deal with the problem of incomplete mileage information, we considered both linear and piece-wise linear approaches in modeling mileage accumulation. Then, to evaluate the standard error of our estimate, we considered the use of bootstrap method. In addition, we also considered two types of univariate estimator for intervals: one associated with the bivariate model (Section 8.3.1) and one direct estimator (Section 8.3.2).

Our findings can be summarized as follows:

- the results produced by using linear and piece-wise linear approaches in modeling mileage accumulation are not significantly different.
- the results produced by the univariate estimators associated with the bivariate model are very different from the results produced by the direct univariate estimators for intervals. By comparing the definitions of these estimators, we see that the univariate estimators associated with the bivariate model would be more reliable than the direct univariate estimators. This suggests that a direct univariate estimator might not be sufficient for estimating the mean cumulative warranty cost, when the warranty program involves two variables.

In addition, from the univariate results, we also observed the phenomenon of "customer-rush near warranty expiration limit" in "time" is mileage case. We did not detect this phenomenon in "time" is age case, probably because the majority (62%) of the vehicles in our dataset leave warranty coverage due to mileage, instead of age. Of course, the above results are based on a single dataset and more study are required.

The model introduced in this article is simple and straightforward, but it requires all vehicles to have some mileage information. In practice, the mileage information for vehicles with claims are available in the database,

while the mileage information for vehicles without claims are usually unknown. This is one major weakness of the model which we would like to improve.

Chapter 9

Predicting Mean Cumulative Warranty Cost

In this chapter, we consider several simple methods for predicting the mean cumulative function $\Lambda(u), u = 1, 2, \ldots$, where u can be age t, mileage m, as well as the actual time x. Note that, in fact, we are predicting the estimated mean cumulative function $\hat{\Lambda}(u)$.

9.1 Curve Fitting

The first method of making prediction is curve fitting. From our examples, we can see that some of the estimated mean cumulative functions $\hat{\Lambda}(u)$ shows a roughly upward linear trend. So, for these estimates with linear trend, we may try to fit a straight line (with no intercept) of the form

$$y = au. \tag{9.1}$$

Here, y corresponds to $\hat{\Lambda}(u)$, u is the function argument, and a is the parameter to be estimated. We can estimate the parameter a easily by the method of least squares (LS). For instance, Figure 9.1 shows the fitted LS line for the adjusted for mileage $\hat{\Lambda}(t)$, produced by the CR-Model in the

example of Section 6.1.1. The estimated value of a is $\hat{a} = 0.0072$, with sum of squared errors equal to 61.5173. Note that the fitted line seems to have a slope of approximately 1, and so a should also be close to 1. However, this is not true because the graph has actually been rescaled.

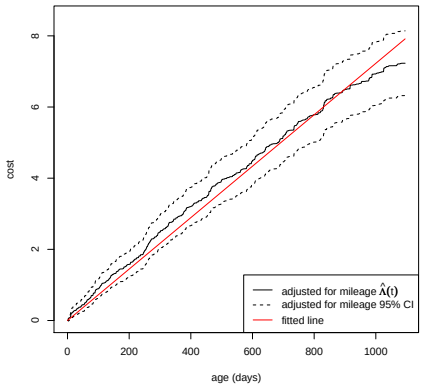

Figure 9.1: Adjusted for mileage $\hat{\Lambda}(t)$ and the fitted LS line

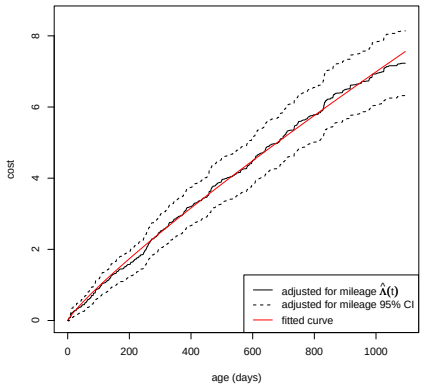

Figure 9.2: Adjusted for mileage $\hat{\Lambda}(t)$ and the fitted LS curve

It can be noted that a straight line does not provide a very good fit in this case, as shown by the large value of sum of squared error. So, instead of using a straight line, we may try to fit a curve. Here, we suggest to fit a curve of the form

$$y = au^b \tag{9.2}$$

i.e., the *power law* model. In Eq. (9.2), y corresponds to $\hat{\Lambda}(u)$, u is the function argument, while a and b are the parameters to be estimated. We can estimate these parameters by the method of least square. Note that the statistical programming language **R** provides a useful built-in function `nlm()` for this purpose. Figure 9.2 shows the fitted LS curve for the adjusted for mileage $\hat{\Lambda}(t)$. The estimated values of a and b are $\hat{a} = 0.0177$ and $\hat{b} = 0.8653$ respectively. The sum of squared errors equal to 9.5168,

which is much lower than that of using a straight line (61.5173). By comparing Figures 9.1 and 9.2, we also see that the fitted curve fits the data better than the fitted line.

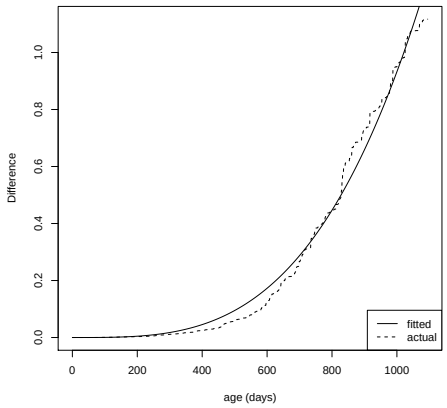

Figure 9.3: Difference between unadjusted $\hat{\Lambda}(t)$ and adjusted for mileage $\hat{\Lambda}(t)$

Sometimes, it may be useful to model the difference between the unadjusted and adjusted $\hat{\Lambda}(u)$. By doing so, we will be able to estimate the adjusted $\hat{\Lambda}(u)$ from the unadjusted $\hat{\Lambda}(u)$, which may be easier to compute. For instance, let us model the difference between the unadjusted $\hat{\Lambda}(t)$ and the adjusted for mileage $\hat{\Lambda}(t)$, produced by the CR-Model in the example of Section 6.1.1, and try to fit a curve of the form of Eq. (9.2). Figure 9.3 shows the difference between these two estimates and the fitted LS curve. The estimated values of a and b are $\hat{a} = 1.178 \times 10^{-10}$ and $\hat{b} = 3.3$ respectively, with sum of squared errors equal to 1.4416. Then, Figure 9.4 shows the fitted adjusted for mileage $\hat{\Lambda}(t)$, which is obtained by adding up the unadjusted $\hat{\Lambda}(t)$ and the fitted difference. On the other hand, Figure 9.5 shows the fitted unadjusted $\hat{\Lambda}(t)$, which is obtained by subtracting the fitted difference from the adjusted for mileage $\hat{\Lambda}(t)$. In both cases, the fitted curve provides a reasonably good fit. This implies that the curve fitted to

the difference between the unadjusted $\hat{\Lambda}(t)$ and the adjusted for mileage $\hat{\Lambda}(t)$ also provides a good fit, and we will be able to derive the adjusted estimate from the unadjusted estimate with less computational effort.

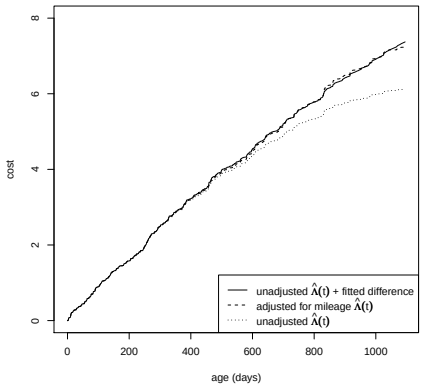

Figure 9.4: Unadjusted $\hat{\Lambda}(t)$ + fitted difference

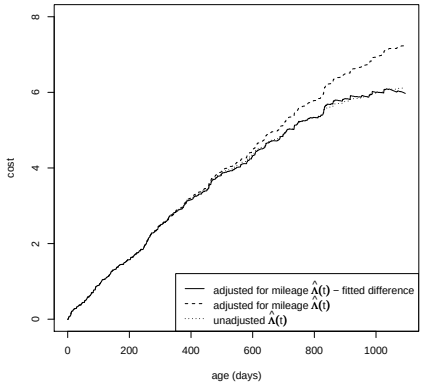

Figure 9.5: Adjusted for mileage $\hat{\Lambda}(t)$ − fitted difference

Curve fitting is a simple and straightforward method for making prediction. However, this method has one major shortcoming: it does not provide coverage probabilities and prediction intervals for the forecasts or predicted values. To overcome this problem, we consider two methods which provide predictive distributions for the forecasts. These methods are *simple linear regression* and *dynamic linear model*.

9.2 Simple Linear Regression

The simple linear regression model has the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{9.3}$$

where

- Y_i is the response variable,

- X_i is the explanatory variable (a known constant),
- β_0 and β_1 are the parameters (i.e. the intercept and slope),
- ϵ_i is a random error term with mean zero and constant variance σ^2.

We assume that ϵ_i and ϵ_j are independent for all $i \neq j$. Usually, we also assume that the errors are normally distributed, i.e., $\epsilon_i \sim N(0, \sigma^2)$.

First of all, we provide some basic facts for simple linear regression. By using the method of least squares, the estimates of the parameters β_0 and β_1 are given by

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{9.4}$$

where $\bar{y} = \frac{1}{n}\sum_i y_i$ and $\bar{x} = \frac{1}{n}\sum_i x_i$. Consequently, the regression line is given by $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, and we define the residuals as $e_i = y_i - \hat{y}_i$, i.e. the difference between the observed value and the fitted value. Then, an unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_i e_i^2}{n-2}. \tag{9.5}$$

Given a value x_* and a fitted regression line $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, the predicted value of y_* is $\tilde{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$, and the standard error of prediction is given by

$$se(\tilde{y}_*) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}}} \tag{9.6}$$

Then, assume that the errors ϵ_i are normally distributed, the $100(1-\alpha)\%$ prediction interval for y_* is

$$\tilde{y}_* \pm t_{\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{XX}}}, \tag{9.7}$$

where $t_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the t-distribution with $n-2$ degrees of freedom.

Now, we consider the application of simple linear regression in predicting $\hat{\Lambda}(u)$. Since $\hat{\Lambda}(u)$ are clearly not independent, it is not appropriate to apply simple linear regression to $\hat{\Lambda}(u)$ directly. Therefore, instead of using $\hat{\Lambda}(u)$, we will work with the estimated rate function $\hat{\lambda}(u)$, which are more likely to be independent. That is, we regard $\hat{\lambda}(u)$ as our response variable Y and fit a simple linear regression model.

Suppose we had estimated $\hat{\lambda}(u)$ and hence $\hat{\Lambda}(u)$ for $u = 1, 2, \ldots, n$, and we would like to predict the value of $\hat{\Lambda}(n+k), k \geq 1$. By using the fitted regression line for $\hat{\lambda}(u)$,

$$\hat{\lambda}_{reg}(u) = \hat{\beta}_0 + \hat{\beta}_1 u, \tag{9.8}$$

the predicted value of $\hat{\lambda}(n+k)$ is given by

$$\tilde{\lambda}(n+k) = \hat{\beta}_0 + \hat{\beta}_1(n+k). \tag{9.9}$$

Consequently, the predicted value of $\hat{\Lambda}(n+k)$ is given by

$$\tilde{\Lambda}(n+k) = \hat{\Lambda}(n) + \sum_{j=1}^{k} \tilde{\lambda}(n+j). \tag{9.10}$$

Further, let L_j and U_j be the lower and upper limits of the $100(1-\alpha)\%$ prediction interval for $\hat{\lambda}(n+j)$ respectively. By assuming independence of $\hat{\lambda}(u)$, the $100(1-\alpha)\%$ prediction interval for $\hat{\Lambda}(n+k)$ is given by

$$\hat{\Lambda}(n) + \sum_{j=1}^{k} L_j \leq \hat{\Lambda}(n+k) \leq \hat{\Lambda}(n) + \sum_{j=1}^{k} U_j. \tag{9.11}$$

Note that the method above requires $\hat{\lambda}(u)$ to exhibit a roughly linear trend. If $\hat{\lambda}(u)$ shows a nonlinear trend, we might include a quadratic term u^2 (and higher-order terms, if necessary). Then, the simple linear regres-

sion would become a multiple linear regression.

Example

Let us consider the adjusted for mileage $\hat{\lambda}(t)$ and $\hat{\Lambda}(t)$, produced by the CCR-Model in Example I of Section 7.3.1. Figure 9.6 shows the graph of $\hat{\lambda}(t)$ and the fitted regression line which is downward sloping. Then, Figure 9.7 shows the graph of $\hat{\Lambda}(t)$ and the fitted curve

$$\hat{\Lambda}_{reg}(t) = \sum_{j=1}^{t} \hat{\lambda}_{reg}(j). \tag{9.12}$$

It can be seen that the curve fits the data very well.

Next, Figure 9.8 shows the forecasts and the corresponding 95% prediction intervals for $\hat{\Lambda}(t)$ for the last 18 month, based on the data of the first 18 months. It can be seen that the forecasts are lower than the actual data, but still lie within the 95% confidence interval for $\hat{\Lambda}(t)$. In addition, the 95% prediction interval gets wider as t increases. This indicates, as expected, that the forecasts further away are less precise.

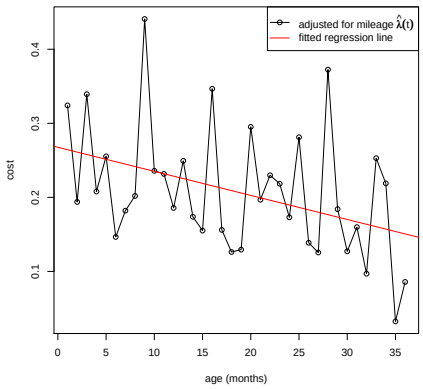

Figure 9.6: Adjusted for mileage $\hat{\lambda}(t)$ and fitted regression line

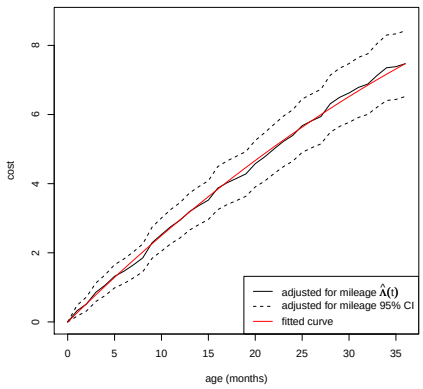

Figure 9.7: Adjusted for mileage $\hat{\Lambda}(t)$ and fitted curve

Figure 9.9 shows the forecasts and the corresponding 95% prediction intervals for $\hat{\Lambda}(t)$ for the last 12 month, based on the data of the first 24 months. It can be seen that the forecasts and the actual data are quite close. With more data in hand, we are able to improve the precision of the forecasts.

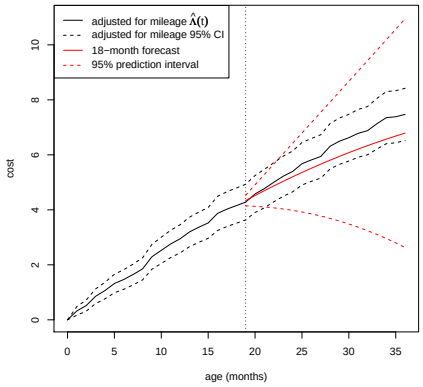

Figure 9.8: 18-month forecast and 95% prediction interval

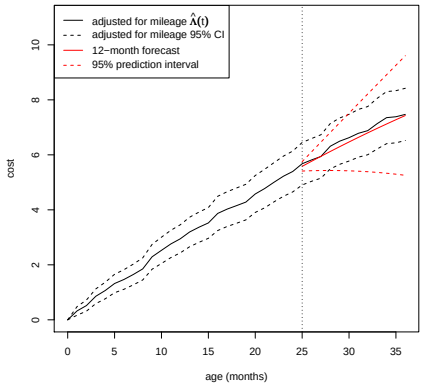

Figure 9.9: 12-month forecast and 95% prediction interval

9.3 Dynamic Linear Model

Dynamic linear model (DLM) is an important class of state space model. In this section, we follow the notations and definitions used by Campagnoli et al. [2009]. Let $Y_1, Y_2, \ldots, Y_t$ denote the observed values of a variable of interest at "times" $1, 2, \ldots, t$, which may be scalars or p-dimensional vectors. Then, the dynamic linear model is defined by the *observation equation*

$$Y_t = F_t\theta_t + v_t, \quad v_t \sim N_p(0, V_t), \quad \text{for } t \geq 1, \tag{9.13}$$

and the *state equation* or *system equation*

$$\theta_t = G_t\theta_{t-1} + w_t, \quad w_t \sim N_q(0, W_t), \quad \text{for } t \geq 1, \tag{9.14}$$

where

- θ_t is a q-dimensional state vector,
- F_t and G_t are known matrices of order $p \times q$ and $q \times q$ respectively,
- $\{v_t\}$ and $\{w_t\}$ are two independent sequences of independent Normal random vectors, with means zero and known variances V_t and W_t, respectively. Note that a dynamic linear model can also be fitted with unknown V_t and W_t from a Bayesian perspective. See Campagnoli et al. [2009] for more details.

In addition, at time $t = 0$, a Normal prior distribution for the state vector is specified by

$$\theta_0 \sim N_q(m_0, C_0), \tag{9.15}$$

and it is assumed that θ_0 is independent of $\{v_t\}$ and $\{w_t\}$.

Here, we consider an example of DLM with $p = 1$ and $q = 2$, called the *linear growth model* or *local linear trend model*. The linear growth model is a polynomial DLM of order 2, and it is defined by

$$\begin{aligned} Y_t &= \mu_t + v_t, & v_t &\sim N(0, V), \\ \mu_t &= \mu_{t-1} + \beta_{t-1} + w_{t,1}, & w_{t,1} &\sim N(0, \sigma_\mu^2), \\ \beta_t &= \beta_{t-1} + w_{t,2}, & w_{t,2} &\sim N(0, \sigma_\beta^2), \end{aligned} \tag{9.16}$$

with uncorrelated errors $v_t, w_{t,1}$ and $w_{t,2}$. In matrix form, the model is given by

$$F = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad \theta_t = \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix}, \quad G = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad W = \begin{bmatrix} \sigma_\mu^2 & 0 \\ 0 & \sigma_\beta^2 \end{bmatrix}. \tag{9.17}$$

In this model, μ_t and β_t are interpreted as the *local level* and *local growth rate* respectively. It is assumed that the current level μ_t changes linearly over time and the growth rate may also evolve. Note that the matrices F_t, G_t, V_t

and W_t are constant, and the system variances σ_μ^2 and σ_β^2 are allowed to be zero.

9.3.1 Kalman Filter

For a given DLM, our main task is to forecast the value of the next state vector θ_{t+1} and the next observation Y_{t+1}, based on the observations up to the current time t, and to update our estimate once new data become available at time $t+1$. This is a *filtering problem* and can be solved recursively by the well-known *Kalman filter*.

Proposition 9.1 (Kalman Filter)**.** *Let $y_{1:t}$ denote the sequence $Y_1, Y_2, \ldots, Y_t$ and let*

$$\theta_{t-1}|y_{1:t-1} \sim N_q(m_{t-1}, C_{t-1}).$$

Then, the following statements hold [Campagnoli et al., 2009].

(i) The one-step-ahead predictive distribution of θ_t given $y_{1:t-1}$ is $N_q(a_t, R_t)$ with

$$a_t = G_t m_{t-1} \quad \text{and} \quad R_t = G_t C_{t-1} G_t' + W_t. \tag{9.18}$$

(ii) The one-step-ahead predictive distribution of Y_t given $y_{1:t-1}$ is $N_p(f_t, Q_t)$ with

$$f_t = F_t a_t \quad \text{and} \quad Q_t = F_t R_t F_t' + V_t. \tag{9.19}$$

(iii) The filtering distribution of θ_t given $y_{1:t}$ is $N_q(m_t, C_t)$ with

$$m_t = a_t + R_t F_t' Q_t^{-1} e_t \quad \text{and} \quad C_t = R_t - R_t F_t' Q_t^{-1} F_t R_t, \tag{9.20}$$

where $e_t = Y_t - f_t$ is the forecast error. (Note: A' denote the transpose of matrix A.)

Often, one might be interested in looking a bit further in the future, say k steps ahead for $k \geq 1$. Then, the predictive distributions for the k-step-ahead state vector θ_{t+k} and observation Y_{t+k} can be computed by using the following proposition.

Proposition 9.2. *Let $a_t(0) = m_t$ and $R_t(0) = C_t$. Then, for $k \geq 1$. the following statements hold [Campagnoli et al., 2009].*

(i) The predictive distribution of θ_{t+k} given $y_{1:t}$ is $N_q(a_t(k), R_t(k))$ with

$$a_t(k) = G_{t+k}a_t(k-1) \quad \text{and} \quad R_t(k) = G_{t+k}R_t(k-1)G'_{t+k} + W_{t+k}. \tag{9.21}$$

(ii) The predictive distribution of Y_{t+k} given $y_{1:t}$ is $N_p(f_t(k), Q_t(k))$ with

$$f_t(k) = F_{t+k}a_t(k) \quad \text{and} \quad Q_t(k) = F_{t+k}R_t(k)F'_{t+k} + V_{t+k}. \tag{9.22}$$

9.3.2 Statistical Programming Language R: Package `dlm`

Now, we briefly introduce some useful functions for the implementation of DLM, provided by the **R** package `dlm`:

- `dlm()`: This function creates a DLM object.
- `dlmModPoly()`: This function creates an n-th order polynomial DLM, and can be used to set up a linear growth model with the argument `order=2` (which is the default).
- `dlmMLE()`: This function returns the maximum likelihood estimates (MLE) of the unknown parameters in a specified DLM for a given dataset.
- `dlmLL()`: This function returns the negative log-likelihood of a specified DLM for a given dataset.

- `dlmFilter()`: This function applies the Kalman filter. Its output includes the data, the specified DLM, the mean and variance of the predictive and filtered distributions for state vector, and the one-step-ahead forecasts. Note that the variances are given in term of their singular value decomposition (SVD).

- `dlmSvd2var()`: This function can be used to reconstruct the variances from their SVD.

- `dlmForecast()`: This function evaluates the mean and variance of the predictive distribution for future states and observations.

For more details, see Campagnoli et al. [2009].

Example

Let the adjusted for mileage $\hat{\Lambda}(t)$, produced by the CCR-Model in Example I of Section 7.3.1, be the variable of interest Y_t. Suppose the variances V and W are known (they were actually estimated) as follows:

$$V = 0.002325481 \quad \text{and} \quad W = \begin{bmatrix} 0.001828051 & 0 \\ 0 & 0.0007354366 \end{bmatrix}.$$

Then, we fit a linear growth model with the above variances. In addition, we initialize the model with

$$m_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \text{and} \quad C_0 = \begin{bmatrix} 1 \times 10^7 & 0 \\ 0 & 1 \times 10^7 \end{bmatrix},$$

which is the default setting for `dlmModPoly()`. Note that we assume V and W are known, but they were actually estimated by using `dlmMLE()`.

Figure 9.10 shows the one-step-ahead forecasts and the corresponding 95% prediction intervals for $\hat{\Lambda}(t)$. It can be seen that the forecasts and the actual data are very close. Note that the prior variance C_0 is very large,

and hence the standard errors of the first few forecasts will also be very large. Here, we left out the first three forecasts to allow the Kalman filter to adjust.

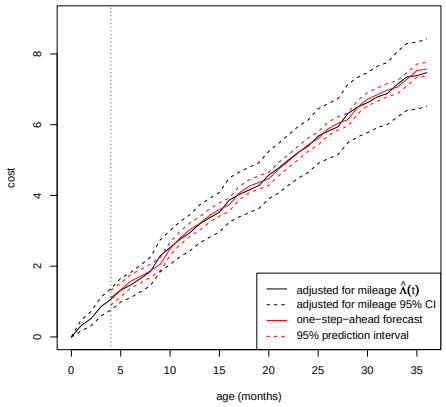

Figure 9.10: One-step-ahead forecast and 95% prediction interval

Next, Figure 9.11 shows the forecasts and the corresponding 95% prediction intervals for $\hat{\Lambda}(t)$ for the last 18 month, based on the data of the first 18 months. Similarly, Figures 9.12 shows the forecasts and the corresponding 95% prediction intervals for $\hat{\Lambda}(t)$ for the last 12 months, based on the data of the first 24 months. In both figures, we can see that the forecasts and the actual data are quite close. Also, the 95% prediction interval gets wider as t increases. This indicates, as expected, that the forecasts further away are less precise.

Then, let us consider the forecasts for the last 6 months produced based on the data of the first 18 months and 24 months. By comparing Figures 9.11 and 9.12, it can be noted that the forecasts based on the first 18 months are closer to the actual data than the forecasts based on the first 24 months. However, the 95% prediction intervals for the forecasts based on the first 18 months are wider than that for the forecasts based on the first 24 months, which indicate a greater level of uncertainty. Thus, in prac-

tice, we would prefer to use the forecasts based on the first 24 months, i.e., based on more data.

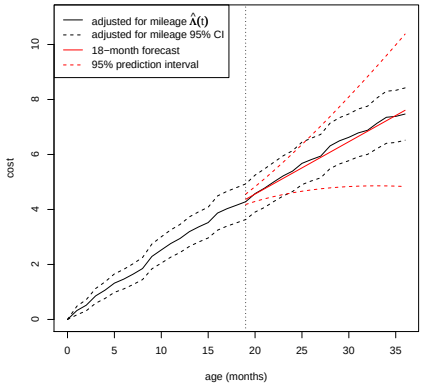

Figure 9.11: 18-month forecast and 95% prediction interval

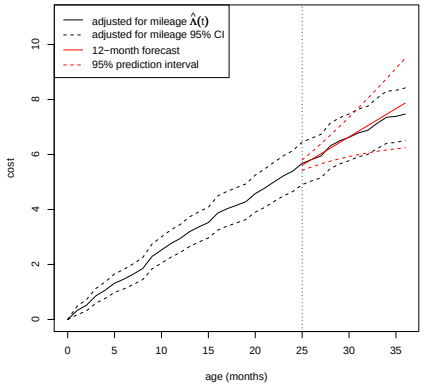

Figure 9.12: 12-month forecast and 95% prediction interval

9.4 Method Comparisons: Simple Linear Regression vs Dynamic Linear Model

Now, let us compare the two methods for prediction: simple linear regression and dynamic linear model. Table 9.1 shows the 12-month forecast and the corresponding 95% prediction intervals (lower limits, upper limits, and interval widths) produced by the two methods in our examples. It can be seen that the forecasts produced by the two methods are both quite close to the actual data. However, the 95% prediction intervals for the simple linear regression approach are wider than that for dynamic linear model, which indicate a greater level of uncertainty.

		Simple Linear Regression					Dynamic Linear Model				
t	$\hat{\Lambda}(t)$	Forecast	Residual	Lower	Upper	Width	Forecast	Residual	Lower	Upper	Width
25	5.6762	5.5814	0.0947	5.4110	5.7518	0.3408	5.6109	0.0652	5.4236	5.7983	0.3748
26	5.8149	5.7648	0.0501	5.4224	6.1073	0.6849	5.8166	-0.0017	5.5477	6.0854	0.5378
27	5.9405	5.9451	-0.0047	5.4289	6.4614	1.0325	6.0222	-0.0817	5.6577	6.3866	0.7289
28	6.3130	6.1224	0.1906	5.4305	6.8142	1.3837	6.2278	0.08520	5.7563	6.6993	0.9430
29	6.4970	6.2966	0.2005	5.4271	7.1660	1.7388	6.4334	0.0637	5.8448	7.0220	1.1772
30	6.6242	6.4677	0.1565	5.4187	7.5167	2.0980	6.6390	-0.0148	5.9243	7.3537	1.4295
31	6.7840	6.6357	0.1483	5.4050	7.8665	2.4615	6.8446	-0.0606	5.9954	7.6938	1.6984
32	6.8809	6.8007	0.0802	5.3860	8.2154	2.8295	7.0502	-0.1693	6.0587	8.0417	1.9830
33	7.1339	6.9626	0.1712	5.3616	8.5636	3.2020	7.2558	-0.1220	6.1147	8.3970	2.2823
34	7.3527	7.1215	0.2313	5.3317	8.9112	3.5794	7.4614	-0.1087	6.1636	8.7592	2.5956
35	7.3852	7.2772	0.1080	5.2964	9.2581	3.9618	7.6670	-0.2818	6.2059	9.1281	2.9222
36	7.4710	7.4299	0.0411	5.2553	9.6046	4.3492	7.8727	-0.4017	6.2418	9.5035	3.2616

Table 9.1: 12-month forecast and 95% prediction intervals

Next, we compare the performances of the two methods by considering three different measures of accuracy: the mean absolute deviation (MAD), mean square error (MSE), and the mean absolute percentage error (MAPE), defined respectively by

$$MAD = \frac{1}{h}\sum_t |e_t|, \tag{9.23}$$

$$MSE = \frac{1}{h}\sum_t e_t^2, \tag{9.24}$$

$$MAPE = \frac{1}{h}\sum_t \frac{|e_t|}{Y_t}, \tag{9.25}$$

where h is the number of terms, $e_t = Y_t - \tilde{Y}_t$ is the residual, Y_t is the data, and $\tilde{Y}_t$ is the forecast. Table 9.2 shows the measures of accuracy for the two methods (based on the 12 month forecast). It can be seen that, in the case of 12-month forecast, none of the two methods is clearly better than the other.

Method	MAD	MSE	MAPE
Simple Linear Regression	0.123093	0.019758	0.018321
Dynamic Linear Model	0.121363	0.026855	0.017367

Table 9.2: Measures of accuracy

The effectiveness of the simple linear approach depends on how well the model fits the data. In our examples, the simple linear regression model fitted based on the data of the first 24 months provides a reasonably good fit, and hence the 12-month forecast produced is quite accurate. But, if the model does not fit the data well, then the results obtained will also be less precise as in the case of 18-month forecast (see Figure 9.8). In addition, the simple linear regression approach assumes that the estimated rate functions $\hat{\lambda}(u)$ are independent (which is difficult to justify). If this assumption fails, then the results produced may not be reliable. Thus, in practice, we would recommend to use dynamic linear model.

9.5 Predicting Bivariate Mean Cumulative Warranty Cost

In this section, we consider one possible method for predicting the (estimated) bivariate mean cumulative warranty cost $\hat{\Lambda}(t, m)$, which is the extension of the simple linear regression approach used in the univariate case. Instead of using simple linear regression, we use *multiple linear regression*.

In general, the multiple linear regression has the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{9.26}$$

where

- Y_i is the value of the response variable,

- $X_{i1}, \ldots, X_{ik}$ are the values of the explanatory variables (known constants),
- $\beta_0, \beta_1, \ldots, \beta_k$ are the parameters,
- ϵ_i is a random error term with mean zero and constant variance σ^2.

We assume that ϵ_i and ϵ_j are independent for all $i \neq j$. Usually, we also assume that the errors are normally distributed, i.e., $\epsilon_i \sim N(0, \sigma^2)$. Equivalently, in matrix form, the model is given by

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times r}\boldsymbol{\beta}_{r\times 1} + \boldsymbol{\epsilon}_{n\times 1}, \quad \boldsymbol{\epsilon}_{n\times 1} \sim N(0, \sigma^2\mathbf{I}), \tag{9.27}$$

where $r = k + 1$, $\mathbf{I}$ is an identity matrix, and $\mathbf{X}$ has full column rank. (For convenience, we will omit the dimensions of these matrices from now on.)

The estimates of $\boldsymbol{\beta}$ and σ^2 are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n-r}\mathbf{Y}'\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{Y} \tag{9.28}$$

respectively. Then, given a (column) vector $\mathbf{x}_*$, the point estimate of the response is $\tilde{y}_* = \mathbf{x}_*'\hat{\boldsymbol{\beta}}$ and the standard error of prediction is given by

$$se(\tilde{y}_*) = \hat{\sigma}\sqrt{1 + \mathbf{x}_*'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*}. \tag{9.29}$$

Consequently, the $100(1-\alpha)\%$ prediction interval for y_* is

$$\tilde{y}_* \pm t_{\alpha/2}\hat{\sigma}\sqrt{1 + \mathbf{x}_*'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*}, \tag{9.30}$$

where $t_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the t-distribution with $n - r$ degrees of freedom.

Now, we consider the application of multiple linear regression in the predicting $\hat{\Lambda}(t, m)$. As similar to the univariate case, we work with the estimated rate function $\hat{\lambda}(t, m)$, instead of using $\hat{\Lambda}(t, m)$. That is, we assume that $\hat{\lambda}(t, m)$ are independent and fit a multiple linear regression model to

$\hat{\lambda}(t,m)$, where the explanatory variables are functions of t and m.

Suppose we had estimated $\hat{\lambda}(t,m)$ and hence $\hat{\Lambda}(t,m)$ for $t = 1,2,\ldots,a, m = 1,2,\ldots,b$, and we would like to predict the value of $\hat{\Lambda}(a+u,b+v)$ for $u,v \geq 1$. Let $\tilde{\lambda}(a+u,b+v)$ be the predicted value given by the fitted multiple regression model

$$\hat{\lambda}_{reg}(t,m) = \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{9.31}$$

Then, the predicted value of $\hat{\Lambda}(a+u,b+v)$ is given by

$$\begin{aligned}\tilde{\Lambda}(a+u,b+v) &= \hat{\Lambda}(a,b) + \sum_{i=1}^{a}\sum_{j=1}^{v}\tilde{\lambda}(i,b+j) + \sum_{i=1}^{u}\sum_{j=1}^{b}\tilde{\lambda}(a+i,j) \\ &\quad + \sum_{i=1}^{u}\sum_{j=1}^{v}\tilde{\lambda}(a+i,b+j). \end{aligned}\tag{9.32}$$

Further, let L_{ij} and U_{ij} be the lower and upper limits of the $100(1-\alpha)\%$ prediction interval for $\hat{\lambda}(a+i,b+j)$ respectively. Then, the lower and upper limits of the $100(1-\alpha)\%$ prediction interval for $\hat{\Lambda}(a+u,b+v)$ are given by

$$\hat{\Lambda}(a,b) + \sum_{i=1}^{a}\sum_{j=1}^{v}L_{i,b+j} + \sum_{i=1}^{u}\sum_{j=1}^{b}L_{a+i,j} + \sum_{i=1}^{u}\sum_{j=1}^{v}L_{a+i,b+j} \tag{9.33}$$

and

$$\hat{\Lambda}(a,b) + \sum_{i=1}^{a}\sum_{j=1}^{v}U_{i,b+j} + \sum_{i=1}^{u}\sum_{j=1}^{b}U_{a+i,j} + \sum_{i=1}^{u}\sum_{j=1}^{v}U_{a+i,b+j} \tag{9.34}$$

respectively.

Example

Let us consider the unadjusted $\hat{\lambda}(t, m)$ and $\hat{\Lambda}(t, m)$ computed in Example I of Section 8.1. Here, we fit the following multiple linear regression model with an interaction term

$$\hat{\lambda}(t, m) = \beta_0 + \beta_1 t + \beta_2 m + \beta_3 (t \times m) + \epsilon. \tag{9.35}$$

Note that the interaction term is regarded as an explanatory variable. Figure 9.13 shows the graph of $\hat{\Lambda}(t, m)$ and the fitted plane

$$\hat{\Lambda}_{reg}(t, m) = \sum_{i=1}^{t} \sum_{j=1}^{m} \hat{\lambda}_{reg}(i, j), \tag{9.36}$$

where

$$\hat{\lambda}_{reg}(t, m) = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 m + \hat{\beta}_3 (t \times m). \tag{9.37}$$

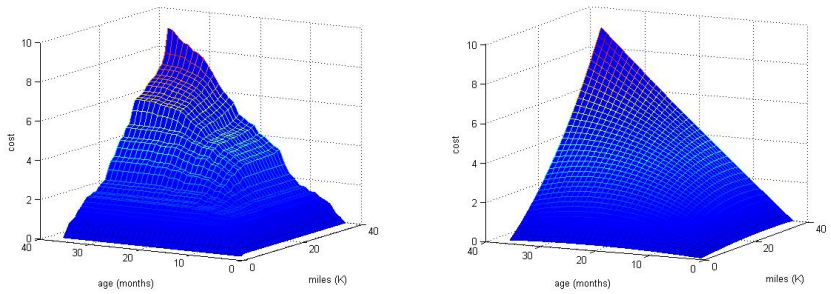

Figure 9.13: Unadjusted $\hat{\Lambda}(t, m)$ and fitted plane

For better illustration, Figure 9.14 shows the unadjusted $\hat{\Lambda}(t, 36)$ and the fitted curve (part of the fitted plane), where we fix m at $36K$ miles. Similarly, Figure 9.15 shows the unadjusted $\hat{\Lambda}(36, m)$ and the fitted curve, where we fix t at 36 months. From these two figures, we see that the fitted

curve provides a reasonably good fit.

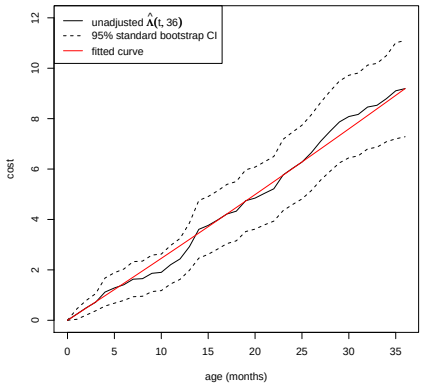

Figure 9.14: Unadjusted $\hat{\Lambda}(t, 36)$ and fitted curve

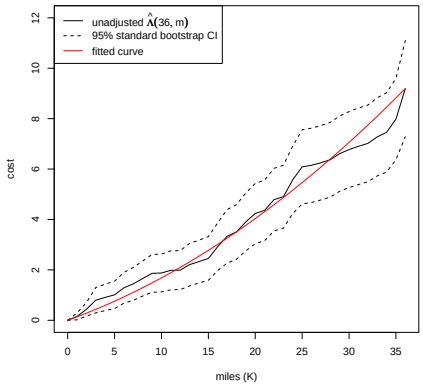

Figure 9.15: Unadjusted $\hat{\Lambda}(36, m)$ and fitted curve

Next, Table 9.3 shows the forecasts and the corresponding 95% prediction intervals for $\hat{\Lambda}(t, t)$ for $t = m = 31, 32, \ldots, 36$, based on the data up to $t = m = 30$. It can be seen that the forecasts are higher than the actual data, and the residual increases as t increases. The 95% prediction interval also gets wider as t increases.

t	$\hat{\Lambda}(t, t)$	Forecast	Residual	Lower	Upper	Width
31	6.7330	7.0272	-0.2942	4.0846	9.9697	5.8851
32	6.9484	7.5591	-0.6107	1.5755	13.5427	11.9672
33	7.2497	8.1560	-0.9063	-0.9676	17.2796	18.2472
34	7.4031	8.8260	-1.4228	-3.5371	21.1890	24.7260
35	7.9421	9.5774	-1.6353	-6.1250	25.2798	31.4047
36	9.1919	10.4193	-1.2274	-8.7230	29.5616	38.2845

Table 9.3: Forecast and 95% prediction interval of $\hat{\Lambda}(t, t)$ for $t = m = 31, 32, \ldots, 36$.

Even though the forecasts computed are not very far from the actual data, the 95% prediction intervals produced are too wide and hence they are not very informative. A better method for predicting the bivariate

mean cumulative warranty cost would need to be developed, and this will be part of our future research.

Remark If the phenomenon of "customer-rush near the warranty expiration limit" exist, certain adjustment needs to be done to the prediction procedure in order to avoid biasness of the results. In this thesis, we do not look into this problem. This is also one of the possible directions for our future research.

9.6 Summary and Discussions

In this chapter, we had considered several methods for predicting the univariate mean cumulative warranty cost. The first method introduced is curve fitting. This method is simple, but it does not provide coverage probabilities and prediction intervals for the forecasts or predicted values. To overcome this problem, we consider two additional methods: simple linear regression and dynamic linear model. The simple linear regression approach assumes that the estimated rate functions $\hat{\lambda}(u)$ are independent. Since it is difficult to justify this independence assumption, we recommend caution. In practice, we would recommend to use dynamic linear model. At the end of this chapter, we also suggested the use of multiple linear regression in predicting bivariate mean cumulative warranty cost. We showed that the performance of this method is not satisfactory, and hence development of a better prediction method is needed.

Chapter 10

Conclusions, Discussions and Future Works

One of the main objectives of our research is to utilize the content of warranty data in estimating and predicting mean cumulative warranty cost per vehicle. So far, in this thesis, we had discussed the structure of automotive warranty database, the characteristics of warranty data, the data mining process, the robust estimator and some of its extensions, prediction, etc.

In our study, we had followed closely the ideas in Chukova and Robinson [2005] and Christozov et al. [2008], which we called the CR-Model and CCR-Model respectively (according to the authors' names). Both of these models are based on the robust estimator, which relies on the assumption that the observation process is independent of the event process. Without requiring any supplementary source of data for mileage accumulation, these models deal with the problem of incomplete mileage information, which typically occurs in automotive warranty analysis by using different approaches in modeling mileage accumulation.

The CR-Model makes an assumption that vehicles accumulate mileage approximately linearly with their age. This model is simple and convenient, but it only uses the last claim to extrapolate mileage accumulation

rate and it does not account for changes in this rate with age. By using different computing softwares (we used statistical programming language **R**, while Chukova and Robinson [2005] used *Mathematica*) and different approach in preparing the data (see Chapter 4), we managed to reproduce some of the results given by Chukova and Robinson [2005] using the same dataset. In addition, we also made our own contribution as follows:

- We proposed a new model for estimating the mean cumulative warranty cost in the actual time case, by using the same linear approach in modeling mileage accumulation. This actual time model will be useful for planning warranty program and warranty reserve.

- We used bootstrap method in the estimation of standard error. The results we obtained suggest that Eq. (5.5), with $M(t)$ replaced by $\hat{M}(t)$, works well for evaluating the standard error. However, a mathematical proof is still required, and we hope to achieve this in the future.

The CCR-Model uses a piece-wise linear approach in modeling mileage accumulation, and it takes into consideration the variability in driving pattern (or mileage accumulation pattern). However, This model does not take into account the effect of reporting delay of claim. By using different computing softwares (we used statistical programming language **R**, while Christozov et al. [2008] used Microsoft Excel) and different approach in preparing the data (see Chapter 4), we managed to reproduce some of the results given by Christozov et al. [2008] using the same dataset. In addition, we also made our own contributions as follows:

- We proposed a new model for estimating the mean cumulative warranty cost in the actual time case, by using the same piece-wise linear approach in modeling mileage accumulation.

- We applied bootstrap method in estimating the standard errors of the mean cumulative warranty cost.

- We investigated the relationship between the variability of driving pattern and the mean cumulative warranty cost. We observed that a higher variability of driving pattern leads to a higher mean cumulative warranty cost. This is a very interesting observation that requires more study. It suggests that we should take into account the variability of driving pattern in modeling mileage accumulation.

Due to our modeling approach, there is some concerns regarding the dependency of the identified driving pattern on the number of claims, as a vehicle with more claims may be more likely to have higher variability of driving pattern compared to those with fewer claims. Thus, we investigate further the relationship between mean cumulative warranty cost and variability of driving pattern by using Dataset 2006. In this dataset, each vehicle has some odometer readings which are not related to time of making a claim. Hence, we will be able to characterize the driving pattern of a vehicle in a better way and to reduce the influence of the number of claims on the determination of driving pattern. The results produced again show that the mean cumulative warranty cost tends to increase as the variability of driving pattern increases. Of course, these results are based on a single dataset and more study are required.

In Chapter 8, we proposed a bivariate model for estimating the mean cumulative warranty cost as a function of age and mileage, by considering both linear and piece-wise linear approaches in modeling mileage accumulation. As the mathematical expression for the standard error of the bivariate estimate is still not available, we used bootstrap method to evaluate the standard error. Besides, we also considered two types of univariate estimator for intervals: one associated with the bivariate model (Section 8.3.1) and one direct estimator (Section 8.3.2). Our findings can be summarized as follows:

- The results produced by using linear and piece-wise linear approaches in modeling mileage accumulation are not significantly different.

- The results produced by the univariate estimators associated with the bivariate model are very different from the results produced by the direct univariate estimators for intervals. By comparing the definitions of these estimators, we see that the univariate estimators associated with the bivariate model would be more reliable than the direct univariate estimators. This suggests that a direct univariate estimator might not be sufficient for estimating the mean cumulative warranty cost, when the warranty program involves two variables.

In addition, from the univariate results, we also observed the phenomenon of "customer-rush near warranty expiration limit" in "time" is mileage case. We did not detect this phenomenon in "time" is age case, probably because the majority of the vehicles in our dataset leave warranty coverage due to mileage, instead of age. Again, the above results are based on a single dataset and more study are required.

In Chapter 9, we considered several methods for predicting the univariate mean cumulative warranty cost. The first method introduced is curve fitting. This method is simple, but it does not provide coverage probabilities and prediction intervals for the forecasts or predicted values. To overcome this problem, we consider two additional methods: simple linear regression and dynamic linear model. The simple linear regression approach assumes that the estimated rate functions $\hat{\lambda}(u)$ are independent. Since it is difficult to justify this independence assumption, we recommend caution. In practice, we would recommend to use dynamic linear model. At the end of this chapter, we also suggested the use of multiple linear regression in predicting bivariate mean cumulative warranty cost. We showed that the performance of this method is not satisfactory, and hence development of a better prediction method is needed.

We have reached the end of this thesis, but our study does not end here. In the future, we like to further our study on the relationship between the variability of driving pattern and the mean cumulative warranty cost, as well as the phenomenon of "customer-rush near warranty expiration

limit". We also like to develop some better methods for predicting the warranty cost in both univariate and bivariate cases. In addition, we also like to improve the current bivariate model to account for vehicles without mileage information and reporting delay of claim, and to derive a mathematical expression for the standard error of the bivariate estimate.

Note that we do not include the **R** programs in this thesis due to the sizes of these programs, but they are available upon request.

Bibliography

W. R. Blischke and D. N. P. Murthy. *Warranty Cost Analysis*. Marcel Dekker, New York, 1994.

W. R. Blischke and D. N. P. Murthy. *Product Warranty Handbook*. Marcel Dekker, New York, 1996.

P. Campagnoli, G. Petris, and S. Petrone. *Dynamic Linear Models with R*. Springer, New York, 2009.

P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. CRISP-DM 1.0 Step-by-step Data Minig Guide. http://www.crisp-dm.org/Process/index.htm, 2000.

D. Christozov, S. Chukova, and J. Robinson. Automotive Warranty Data: Estimation of the Mean Cumulative Function using Stratification Approach. Technical report, SMSCS, Victoria University of Wellington, New Zealand, 2008.

S. Chukova and J. Robinson. Estimating Mean Cumulative Functions from Truncated Automotive Warranty Data. In A. Wilson, N. Limnios, S. Keller-McNulty, and Y. Armijo, editors, *Modern Statistical and Mathematical Methods in Reliability*, pages 121–135. World Scientific, Singapore, 2005.

M. Clark and J. Randal. *A First Course in Applied Statistics: with Applica-*

tions in Biology, Business and the Social Sciences. Pearson Education New Zealand, Auckland, 2004.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics*, 7(1):1–26, 1979.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Acedemic Press, San Diego, 2001.

X. Hu and J. Lawless. Estimation of Rate and Mean Functions from Truncated Recurrent Event Data. *Journal of the American Statistical Association*, 91(433):300–310, 1996.

R. Jayaraman. *On Minimizing Expected Warranty Cost in 1-Dimension and 2-Dimensions with Different Repair Options*. PhD thesis, Department of Industrial Engineering, Texas Tech University, 2008.

D. Larose. *Discovering Knowledge in Data: An Introduction to Data mining*. John Wiley & Sons, New Jersey, 2005.

J. Lawless. Statistical Analysis of Product Warranty Data. *International Statistical Review*, 66(1):41–60, 1998.

J. Lawless, X. Hu, and J. Cao. Methods for the Estimation of Failure Distributions and Rates from Automotive Warranty Data. *Lifetime Data Analysis*, 1:227–240, 1995.

R. Meinhold and N. Singpurwalla. Understanding the Kalman Filter. *The American Statistician*, 37(2):123–127, 1983.

Microsoft. Table that Data: Why Separate Tables? http://office.microsoft.com/training/Training.aspx?AssetID=RP061494331033&CTT=6&Origin=RC061183261033, n.d.

H. Pham and H. Wang. Imperfect Maintenance. *European Journal of Operational Research*, 94:425–438, 1996.

B. Rai and N. Singh. Hazard rate estimation from incomplete and unclean warranty data. *Reliability Engineering and System Safety*, 81:79–92, 2003.

B. Rai and N. Singh. Modeling and analysis of automobile warranty data in presence of bias due to customer-rush near warranty expiration limit. *Reliability Engineering and System Safety*, 86:83–94, 2004.

B. Rai and N. Singh. *Reliability Analysis and Prediction with Warranty Data: Issues, Strategies and Methods*. CRC Press, Boca Raton, 2009.

S. Weisberg. *Applied Linear Regression*. John Wiley & Sons, New Jersey, 3rd edition, 2005.

Wikipedia. Data Masking. http://en.wikipedia.org/wiki/Data_masking, n.d.