City Research Online

City, University of London Institutional Repository

Citation: Lopes, L R (2021). Predictive Models for Medical Costs in Private Healthcare. (Unpublished Doctoral thesis, City, University of London)

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/26897/

Link to published version:

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

Predictive Models for Medical Costs in Private Healthcare

Leonel Rodrigues Lopes Junior

A Thesis presented to the Faculty of Actuarial Science and Insurance.
In partial fulfilment of the requirements for the Degree of Doctor of Philosophy in Actuarial Science at Bayes Business School (formerly Cass Business School)
City, University of London.

Advisors:
Ben Rickayzen
David Smith
Steven Haberman

London, United Kingdom
September 2021

Contents

List of Figures

List of Tables

Acknowledgements

This thesis would not come to life without the extraordinary guidance from Professor Ben Rickayzen, Professor Steven Haberman and Doctor David Smith. I am so grateful for this incredible opportunity to work with you during all these years. I owe you so much for your valuable advice, support and, above all, patience. Having a Ph.D. degree is already challenging enough, but, as few people know, I also went through some turbulent times while trying to finish this thesis. Despite all of the troubles I put you through, you still got my back all the way and for that I am forever grateful!

I want to thank my friends at Unimed-BH for believing in this research and for giving access to the company's data. Special thanks to Dr Sérgio Bersan, Fernando Biscione, Fernando Coelho, Nelson Otávio and Renato Campos, who moved mountains in order to make this study possible. Also, a tremendous thank you to Gustavo Barreto, who allowed me to finish this work with as much sanity as possible. A particular thank you to Bernardo Lanza, not only for recommending me to this Ph.D. program, but also for the constant support.

Some people change our lives in unimaginable ways. For me, these people are Ana Paula Franco Viegas and Prof. Renato Assunção. I feel so privileged to have crossed paths with them and I will always look up to them with admiration and gratefulness (and I know I should say this to them more often).

I dedicate this work to my mum, dad and brother. Thank you for giving me the strength when I was weak. Thank you for the encouragement when I was hopeless and doubtful. Thank you for the unconditional love.

A special THANK YOU, in capital letters, to the love of my life, my wife Dominika. Your understanding, compassion and indescribable levels of patience were crucial for this achievement. You are the best partner someone could ask for and sometimes I ask myself if I deserve all the love you give me. I promise I will not get myself involved in a project so demanding again...at least in the next couple of months...

My dear friends at the Ph.D. office, thank you for the chats, usually over mugs and mugs of coffee. Peter and McKenzie, thank you for the unforgettable times! To all academic staff in the Department of Actuarial Science and Insurance, thank you for the support and for being fantastic role models. Big thanks to all staff at Business School (formerly Cass) for making my academic life less stressful.

Finally, but by no means less important, a huge thank you to all my friends (you know who you are, so I will not name you here), old ones and new ones, for the words of support, for understanding my absence in many events and for staying by my side, no matter the distance or time-difference.

Certificate of Readiness to be Included in Library

I grant powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Abstract

Crucial insurance operations such as pricing, policy renewal, reserving and underwriting rely heavily upon estimations of future claim amounts. In the healthcare field, specifically, there is also a considerable interest in identifying potential high cost individuals for inclusion in preventive care programs in order to avoid or reduce catastrophic future medical expenditures.

A valuable source of information that could be used as input to predictive models aiming to estimate future medical claim costs is the administrative claims data. The highly detailed information contained in these data come from the invoices that are sent by the healthcare providers to insurers relative to the medical procedures and services provided to the policyholders. So far, however, such data have remained relatively unexplored.

In this thesis, we use a large administrative claims data set (over 795,000 policyholders) from a Brazilian healthcare insurance company in order to build predictive models that are able to extract the most relevant information for claim cost prediction in the policyholder level. We compare traditional models such as multiple linear regression, two-part and frequency-severity, with alternative statistical learning methods that make use of a linear combination of predictors such as ridge regression, lasso and Cubist. We use 10-fold cross-validation to find models that provide a better balance between prediction accuracy and model complexity, making it easier to interpret the outcomes without compromising the accuracy of estimations.

Both lasso and Cubist offer significant improvements over traditional models in terms of accuracy of predictions and in terms of making better use of detailed medical information present in the claims data. In general, all models agree on the relevance of predictors, with previous claim amount being the most important covariate. Other covariates that suggest higher future costs are chemotherapy sessions and hospitalisations related to cancer, diabetes and kidney diseases. Among the variables that suggest lower future cost are physiotherapy sessions and hospitalisations related to pregnancy.

Chapter 1

Introduction

1.1 Background

The prediction of health care costs has been a topic of importance both outside and within the actuarial context. The health econometrics field, for instance, largely utilises statistical models in order to investigate how the characteristics of individuals, including their health status, influence their health care utilisation (Jones, 2000). The purpose of such analyses is to understand what drives the demand for health services and use that knowledge to forecast medical costs, usually at population level. This guides the decisions of policy-makers regarding allocation of medical resources to attempt to cover the predicted demand (Ash et al., 2000). Predictions are made for short terms, such as yearly, as well as for longer terms, such as five, ten or even 50 years (Getzen, 2000).

At a more micro level, insurers and health providers use predictive models for several purposes. Some of these purposes do not involve forecasting the cost amount itself. For instance, insurance companies may be interested in predicting the frequency and duration of hospitalisations for their group of policyholders (Xie et al., 2015). From these results, they are able to define better strategies to select the hospitals that are part of their health provider network and negotiate lower fees with them.

Predictive models can also help insurers identify potential fraudulent claims (Lu and Boritz, 2005). Identifying frauds more efficiently results in lower financial losses due to unnecessary claim payments by the insurer. This amount can be converted into improvements for policyholders, such as better coverage and/or lower prices and ultimately contribute to a fairer and more efficient health system, which benefits both insurers and policyholders (Lieberthal et al., 2018).

Another common use of predictive models is in identifying potential high-cost individuals. It is widely known that a very small proportion of individuals are responsible for a very large proportion of claim costs (Zook and Moore, 1980). Around 50% of these individuals tend to have chronic conditions and around 20% suffer from a very acute condition that results in "catastrophic" medical costs (Schroeder et al., 1979). In order to avoid or reduce these costs, insurers may try to target individuals so they can be enrolled in case or disease management programs (Dove et al., 2003). The earlier this identification occurs, the more efficient the programs are at improving the patient's health and reducing future costs (Rosen et al., 2005).

Crucial insurance operations such as pricing, policy renewal, reserving and underwriting rely heavily upon estimations of future claim amounts. For health insurance in particular, risk adjustment is necessary in order to define fairer payments to healthcare providers (Duncan, 2011). This process relies on statistical modelling to produce estimates of medical costs based on the features of individuals. Cumming et al. (2002) and Winkelman and Mehmud (2007) compare different models typically used by insurers for risk adjustment. These models are usually available in software that can be purchased by the insurer and provide pre-defined algorithms that automatically use the data which has been input to forecast the expected medical expenditures of individuals or groups. They use diagnostic, pharmaceutical and demographic information and group individuals into different categories according to a grouping algorithm. Each model has its own (proprietary) algorithm.

1.2 Challenges of Claim Cost Prediction

Regardless of the purpose of the analysis, modelling cost data is a rather complex task, due to several problems. The distribution of the costs is troublesome, with very long tails and a high proportion of individuals with no cost (Duan et al., 1983). On top of that, the claims costs present high variability as a consequence of price variation[1] and cost sharing[2] (Duncan, 2011). Also, the costs of individuals vary over time i.e. individuals can transition between having large costs in one year to having low or no costs in the following year (Guo et al., 2015).

All of the problems cited above mean that the accuracy of predictions tends usually to be very low. The coefficient of determination (R^2), which is commonly used as a measure of goodness-of-fit and accuracy of predictive models, is rarely higher than 0.20 for healthcare cost predictions at the individual level (Diehr et al., 1999). Newhouse et al. (1989) estimates that, for outpatient procedures, the (theoretical) maximum R^2 possible is 0.48 and 0.15 for total claim costs. Many authors (Cumming et al., 2002; Bertsimas et al., 2008; Duncan et al., 2016) relate the low R^2 values to the influence of very large claims, which are very hard to predict. The size of the residuals resulting from the poor fits of these large claims in the calculation of R^2 offsets the goodness-of-fit achieved by the model for the bulk of the data. For this reason, R^2 would not be the most appropriate measure of accuracy to be used when assessing predictive models for healthcare costs.

1.3 Traditional Models for Claim Cost Prediction

Traditionally, parametric models have been the method of preference for the prediction of claim costs, particularly linear regression models, as concluded by Mihaylova et al. (2011). The reasons behind the preference for linear regression models include easy implementation, interpretability of results and simplicity of the model, which does not require too much expertise to be used (Mihaylova et al., 2011). In their risk adjustment studies, Cumming et al. (2002) and Winkelman and Mehmud (2007) use linear regression models in order to compare the accuracy of each algorithm being tested because this is the standard method adopted in the healthcare insurance market.

Nevertheless, linear regression models have their drawbacks. As discussed previously, the distribution of claim costs is skewed, with long tails, while the inferences made from the results of linear regression models rely on the assumption that the distribution of the response is normal (Basu and Manning, 2009). The relationships among covariates are more likely to be non-linear and transformations (such as log or power transformations) are necessary in order to take these effects into account (Diehr et al., 1999). Also, the large proportion of zero claim individuals cannot be properly accommodated by the linear model (Mihaylova et al., 2011).

A way to overcome these issues is the use of generalised linear models - GLMs (Nelder and Wedderburn, 1972), which allows distributions other than normal to be fitted and linked to a linear combination of covariates. Classical methods used by actuaries to predict claim amounts are directly based on GLMs. This is the case for the two-part model, which, as the name suggests, estimates future claim costs in two stages. The first stage estimates the likelihood of a claim happening and the second stage estimates the total claim amount, conditional on the occurrence of a claim. The frequency-severity model builds upon the two-part model. It independently estimates the number of future claims and the average amount of each claim, with the final outcome being the product of the two estimates. More details about the two-part model and the frequency-severity model are provided in sections 3.4 and 3.5, respectively.

Frees, Gao and Rosenberg (2011) estimate the frequency and amount of healthcare expenditures by type of procedure (i.e., inpatient, outpatient and medical appointments) by fitting a frequency-severity model, a method that will be explained in more detail in future sections. They use variables that capture characteristics of individuals such as demographics, access to health services, socioeconomic status, employment and healthcare insurance coverage, among others, which are extracted from the 2003-2004 Medical Expenditure Panel Survey (MEPS) conducted across the United States. This approach gives more accurate predictions than one-part or two-part models fitted to the same data by the authors. In essence, one-part models attempt to fit the whole distribution of the costs all at

[1] Price variation occurs when the same procedure has a different cost for the insurer due to negotiations. It is common for hospitals to agree on lower prices for some procedures given that the insurer has a large group of policyholders who will potentially look for care in their facilities.

[2] Co-payment and deductibles, known as cost sharing, are used by insurers to reduce healthcare utilisation and deal with moral hazard.

once, whilst two-part models first aim to estimate the likelihood of an individual making a claim and then estimate the average amount of the claim, given that it happened. This directly addresses the problem of zero-cost claims.

Frees et al. (2013) include Gaussian copulas in the frequency-severity model which has been fitted to the same MEPS data used by Frees, Gao and Rosenberg (2011). This feature could capture possible dependencies among the types of medical services; for instance, a medical event might start with hospital outpatient procedures and then be followed by an inpatient hospitalization. The authors conclude that the dependency term is significant, but there is no consensus regarding the model that produces the best results.

1.4 Alternative Methods for Claim Cost Prediction

Recent developments in big-data have allowed more information to be collected, medical records to be digitalised and granular information of administrative medical claims[3] to be stored in a structured way (Groves et al., 2016). This creates opportunities for exploring the potential of such information in providing insights that improve our understanding of factors affecting the utilisation of healthcare services and, ultimately, increase the accuracy of claim cost predictions. On the other hand, the level of detail of information present in the claims data also imposes a challenge in terms of uncovering the most relevant covariates hidden among the large volume of data points.

In parallel, the rise of statistical learning techniques has happened as a response to the need to make sense of the large volume of data being generated. Some of them focus on finding insights in the data to support decision-making, a process called data-mining. Yoo et al. (2012) summarise the main methods for data-mining and their use in healthcare studies. Other methods focus on prediction, which is our interest. Many of these methods move away from the classical regression models, being more flexible when fitting the data, since most of them are based on non-linear, non-parametric methods (Friedman et al., 2001).

Studies that investigate some of these methods for the prediction of healthcare costs include Bertsimas et al. (2008), who use clustering and decision tree algorithms to try to predict the quintile of cost to which the individual would belong in the following year, using claims data (detailed medical procedures, diagnoses and demographics) from over 800,000 insured individuals in the United States. Although these methods are able to find patterns of cost previously unrevealed, they do not advance much in terms of prediction accuracy: their maximum R^2 is 0.18.

Duncan et al. (2016) compare alternative methods to linear regression models fitted to a group of 30,000 commercially insured individuals. They conclude that these models offer a better framework for variable selection and accuracy of predictions than traditional linear models, but there is no unanimity regarding the superiority of a model, with the prediction accuracy being improved up to an R^2 of 0.215. On top of that, the models could not agree on the most significant variables for medical claim cost prediction. Similar conclusions are made by Morid et al. (2017) and Yang et al. (2017), who also explore many statistical methods to fit administrative claims data to forecast medical costs. However, a unanimous conclusion from these authors is that prior costs are the most relevant for predicting future medical claim costs (Winkelman and Mehmud, 2007; Duncan et al., 2016; Morid et al., 2017).

A significant issue faced by the authors is the "black-box" nature of the models that provide the best accuracy (Duncan et al., 2016; Morid et al., 2017; Yang et al., 2017). This is because these models have several hidden steps in order to make them more responsive to the data, which makes it impossible to interpret how each covariate relates to the response (Friedman et al., 2001). More accurate predictions lead to a better management of the risks by the insurer, which has an impact in diverse areas, such as appropriate pricing of future health care plans, better provisioning of resources in the following year, management of the supply of health care services by better knowing the health profile of the insured individuals, cost reduction by targeting individuals suitable for preventive health programs, among other actions. On the other hand, the lack of interpretable results is still a barrier for insurers, who do not know how to incorporate the insights provided by these into their operations (Rioux et al., 2019).

[3] The information contained in the administrative claims data come from the invoices sent by the healthcare providers to insurers in respect of the medical procedures and services used by the policyholders.

1.5 Main Aims and Methodology

In our study, we use a large administrative claims data set (over 795,000 policyholders) from a Brazilian healthcare insurance company in order to find models able to extract the most relevant information for claim cost prediction in the individual level. We compare the performance of traditional models (multiple linear regression, two-part and frequency-severity) with other statistical learning methods that also make use of a linear combination of predictors. We use two categories of methods. One category is the regularised methods, represented in our research by ridge regression and lasso (Hoerl and Kennard, 1970). These methods consist of applying a penalty during the estimation process, making the regression coefficients shrink towards zero. The other category is model-trees, represented by M5 and Cubist (Quinlan, 1992, 1993). These models combine decision-tree techniques to split the data into homogeneous groups with linear regression models, responsible for predictions. These particular models were chosen because they do not move too far away from the traditional linear regression model, which makes interpretation of results feasible, while incorporating better methods for variable selection and coefficient estimation. Additionally, M5 and Cubist offer the possibility of more than one linear model being fitted to different parts of the data, which is more likely to provide more accurate results than trying to fit the entire data set with one equation.

These models are tuned by adjusting some of the parameters that control the complexity of the models. We do that by splitting the data into training and test samples and use 10-fold cross-validation in the training sample in order to find the optimal value for the tuning parameters while trying to avoid over-fitting of the data. We show how we use cross-validation to find models that provide a better balance between prediction accuracy and model complexity, making it easier to interpret the outcomes without significantly compromising the accuracy of predictions. We compare the goodness-of-fit of the models in the training sample and the prediction accuracy in the held-out test sample.

We create over 60 covariates that capture the medical procedures and events such as doctor visits, tests, therapies, emergency care, home care and hospital inpatient admissions. Diagnostic information from the ICD-10 codes of hospitalisations is also used. We test three different ways of grouping these codes: ICD-10 chapters, Charlson comorbities (Charlson et al., 1987) and Global Burden of Diseases causes of death (GBD 2017 Risk Factor Collaborators, 2018). These groupings are an alternative to the software-based grouping algorithms tested by Cumming et al. (2002) and Winkelman and Mehmud (2007). We also include covariates based only on previous claim amount, without medical details, in order to compare the performance of the models with and without more granular medical information.

The new methods are promising in terms of making better use of the large amount of information available in the still relatively unexplored administrative claims data for cost prediction. They also improve the goodness-of-fit and prediction accuracy in comparison to traditional methods.

1.6 Structure of the Thesis

This thesis is structured as follows. In chapter two, we describe the data set and provide more details about the distribution of medical claim costs. In the third chapter, we provide a description of the traditional methods involving linear models, we show how we fit these models to the aggregated data (without the granular medical information) and compare their performances. In chapter four we present the alternative statistical learning methods and compare their results in respect of their application to the aggregated data. The introduction of more detailed data into the models is done in chapter five, where we analyse the relevance of the covariates based on the medical information and compare the performances of traditional and alternative methods when using these data. We present the conclusions of our work and suggestions for further steps in chapter six.

Chapter 2

The Administrative Medical Claims Data

2.1 Descriptive Analysis of the Data

The data analysed comes from Unimed - Belo Horizonte, a large private medical insurance and plans company in Brazil, with over 1.2 million policyholders[4]. Our analysis comprises two calendar years, from 1st January 2016 to 31st December 2017. The claim and policy information from the calendar year 2016 are used to predict the policyholder's medical cost in the calendar year 2017. This means that the independent variables of our models are based on the information related to the calendar year of 2016 and the response variable is the total claim cost that each policyholder experienced in the calendar year of 2017.

We only included in our analysis individuals who were exposed throughout the entire two-year period. In other words, we included in the analysis all policyholders whose policies have a starting date that is not later than 1st January 2016 and that remained active at least until 31st December 2017. This means that all policies with starting date later than 1st January 2016 or with ending date before 31st December 2016 were removed from the analysis. It also means that policies that started before 1st January 2016 are included, provided that they were not canceled before 31st December 2017. By doing this, we aim to reduce the variability in the data caused by individuals who enter and leave the pool of policyholders during the period of analysis. This approach is in line with that adopted by other studies, such as Bertsimas et al. (2008) and Duncan et al. (2016). A total of 795,009 individuals are included in the final data set.

We can organise the information used in the study into three dimensions: demographic, policy design and claim cost information. The demographic information used was age, calculated at the beginning of the observation year (1st January 2016) and gender. 56.1% of policyholders of this group are females, and they are the predominant gender for all age groups from age 15 upwards, which can be seen in Figure 2.1. Also, we have a considerable range of ages being analysed, from new born individuals (age 0) to those over 100 (the maximum age was 107 years old).

Insurers usually collect other demographic information such as marital status, ethnicity, level of education, level of income, region of residency among others. Some of these data were available from this particular company, however it required extra care when extracting the data in order to check for completeness and correctness. Thus, we prioritised policy and claims data, which were readily available and were easier to be extracted and structured for our purposes. Nevertheless, we encourage exploring the use of other demographic data, whenever available.

Regarding policy information, we have features that describe the design of the policy, and reflect the practices and regulation standards of the Brazilian market of private medical insurance and plans. The first feature is the contract type of the policy. There are three types of contract: the employer-based group, which represent policies obtained by employers and are offered as a benefit for employees and their dependants; group by association, which is the one acquired by institutions such as unions and co-operatives to insure their members; and individual contracts, which are purchased by people who

[4] According to the National Regulatory Agency for Private Health Insurance and Plans (ANS) Unimed-Belo Horizonte had 1,266,002 policyholders in December 2019. Source: `http://www.ans.gov.br/perfil-do-setor/dados-e-indicadores-do-setor`.

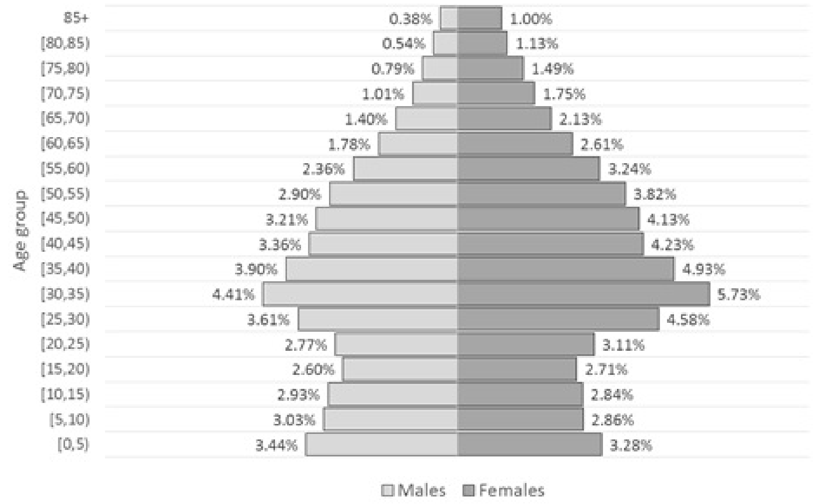


Figure 2.1: Distribution of policyholders analysed by gender and age group

want to improve their own healthcare coverage and that of their dependants. In fact, the concentration of individuals between ages 25 and 40 seen in Figure 2.1 is a consequence of the business strategy of the insurer to sell group type contracts.

Another important policy design feature is plan type. The standard plan type relates to all healthcare plans purchased after the Act 9656, in force since 1998. This Act standardised the medical procedures that must be covered by the healthcare plans being commercialised since that year. Unregulated plans are those sold before the Act 9656. Thus, there is no standard in terms of procedures being covered by them. The coverage and provider network offered by these plans vary greatly and depend on what is established in each contract sold. A particular feature of these plans is that they were not commercialised from the moment the Act 9656 was instated. This explains the fact that the group with this plan type has an average age of 60.5 years, being much older than the groups with standard or restricted plans: 36.3 and 32.5 years, respectively.

Restricted plans offer the same coverage as the standard plans in terms of medical procedures. However, customers of this plan type have access to a more limited health care provider network. Those health care providers receive a lower payment rate from the insurer for the medical procedures related to individuals with restricted plans, which has a direct impact on our cost analysis.

An important feature that has an explicit impact on medical costs is co-payment. For this particular insurer, this is a pre-established fixed fee (not proportional to the cost of procedures[5]) paid by the individual when using a healthcare service covered by the plan. The co-payment is generally charged per procedure, which means that the policyholder pays one fee for each visit to the doctor or medical test made. However, in the case of hospitalisations, the policyholder is charged only one co-payment fee related to that particular hospitalisation. The amount of co-payment varies by the type of procedure and also depends on what is established in the policy. For these reasons, each policyholder pays a different co-payment amount. However, the amount of co-payment paid by each individual is not observable in our data set which is why we include the co-payment information as a policy design feature, indicating whether a policyholder pays co-payment or not.

It is expected that individuals with co-payment (74.7% of the observations in the data set) have lower medical costs than those without it. This happens because they tend to avoid unnecessary medical procedures in order to reduce their expenditure with medical treatments. In this sense, co-payment contributes to reducing the risk of moral hazard, which, in the context of medical insurance, happens when individuals increase their demand for care services because they are being covered by the insurer. Also, those who opt to buy a contract without co-payment (certainly more expensive than one with co-payment) most probably do so because they are aware that they are likely to need to use the healthcare system more frequently, perhaps due to being in poorer health. For a discussion on the topic, we refer to Pauly (1968), who addresses the economic aspects of moral hazard and the use of other methods such as deductibles and coinsurance in order to mitigate it.

[5] Some medical insurers in Brazil might adopt a co-payment that is proportional to the cost of the procedure, but this is not the case for the insurer who provided the data.

When buying a policy, individuals are also allowed to choose the type of hospital accommodation that will be covered by the plan in case of hospitalization. Private rooms in hospitals are, obviously, more expensive than ward rooms, which elevates the observed cost of a hospitalization. As a consequence, plans that offer private room coverage are more expensive. In our data, 42.06% of individuals opted for private hospital accommodation.

For group contracts with less than 30 lives and individual contracts, potential beneficiaries are required to fill in a declaration form regarding their current health state. In some cases, there is also a medical assessment in order to have a better picture of the individual's current health state. Those who declare pre-existing conditions (6.95% of our data set) are expected to experience higher medical costs than those who are not aware of any current sickness.

We include pre-existing conditions as a policy design feature (and not demographic information) because, given the presence of any health condition, the policy is modified in two possible ways. One option is that the policyholder pays a larger premium to include the coverage of medical procedures related to the pre-existing condition. In the second option, the premium remains unchanged, but there is a waiting period of two years for the health conditions of the policyholder. In other words, the medical procedures related to the pre-existing health conditions of the policyholder are not covered by the insurer for two years, starting from the purchase date of the policy. Hence, there is a modification in the policy as a result of pre-existing conditions, which can affect the claim costs. There is a possibility to include which option the policyholder chooses (higher premium or extended waiting period), and this can be an improvement for future work.

The proportion of policyholders for each policy design feature is summarised in Table 2.1.

Table 2.1: Proportion of policyholders by each policy design feature

Feature	**Proportion of policyholders**
Plan type	
Standard	74.52%
Unregulated	4.24%
Restricted	21.24%
Contract type	
Employer	52.42%
Association	23.63%
Individual	23.95%
Ownership	
Owner	59.57%
Dependant	40.43%
Hospital accommodation	
Ward	57.94%
Private	42.06%
Co-payment	
No	25.33%
Yes	74.67%
Pre-existing conditions	
No	93.05%
Yes	6.95%

The policy information variables are described above according to the sequence of choices that each individual makes when purchasing a new health insurance policy in this particular market. Some choices are nested within others. For instance, an individual only chooses to have a policy with co-payment after selecting the plan type. This series of choices reflects a specific behaviour of the policyholder and, consequently, it impacts their claim experience. Using the same example, individuals who choose to have a co-payment do not expect to use medical services very frequently in the near future. Thus, it is more beneficial to purchase a policy with co-payment and lower monthly fees. Such a way of organising the policy information may be helpful for readers who are interested in the behavioural aspects of economic modeling in health insurance.

The third information type, medical claims costs, are the focus of our analyses, since we are interested in predicting the total yearly cost of a policyholder. This cost relates to all medical procedures paid by the medical insurance company to the healthcare providers for the services delivered to its policyholders.

As explained previously, there are individuals who pay the co-payment related to the medical procedures made. Some insurers consider this as a "deductible". Thus, the observed claim cost would be the excess of the co-payment amount. This is not the case in our data set and means that the claim costs observed reflect the total costs of the medical procedures paid by the insurer. In this particular case, the insurer considers co-payment as a revenue from partially sharing the risk with the policyholder and not a deduction in the cost.

Furthermore, it is important to emphasize that the costs are linked to the date when the procedures incurred, not with the date when they were paid. This allows us to have a better picture of the healthcare trajectory of each individual without any disturbance caused by any delays from payment bureaucracy. Finally, all the claims analysed are considered as complete, with no "incurred but not paid" procedures to be included, which is a favourable feature of the data. In general insurance, for instance, it is common to observe incurred claims that take years to be reported to the insurer and, for this aspect, they set the incurred but not reported (IBNR) provisions (Bornhuetter and Ferguson, 1972; Reid, 1978). This is not a concern in our case.

In Brazil, there is no limit in terms of the amount covered. In other words, all costs coming from the covered medical procedures must be paid by the private healthcare company managing the plans. Due to this fact, the distribution of medical costs can display a long tail, related to a few individuals who, most probably, had to remain hospitalized for consecutive months receiving complex treatments. In our data, the maximum cost in the prediction year (2017) is 561,431.20 BRL (£ 136,449)[6].

Another feature of the distribution of the medical costs which contributes to its increased variability is the high proportion of individuals who do not make medical claims. In 2016, 15.7% of policyholders had zero claims. In 2017, this proportion was 16.0%. 10.5% of individuals did not register any medical procedure in either year. The proportion of zeros may seem smaller than what is usually seen in the insurance market. This is because we have excluded the individuals who entered and left the pool of policyholders during the two years of analysis. If we had included the latter, we would have observed that 25.6% of the policyholders did not claim in 2016 (hence, claim cost is zero for that proportion of policyholders) and 25.7% did not claim in 2017.

In addition, small costs with medical appointments and basic medical tests add more fluctuations to the costs distribution. All of this culminates in the highly skewed distribution shown in Figure 2.2. It is important to highlight, however, that the distribution shown is slightly distorted from the distribution of the costs considering all claims of that year. This small distortion is a result of the exclusion of policyholders who enter and leave during the two-year period of the analysis.

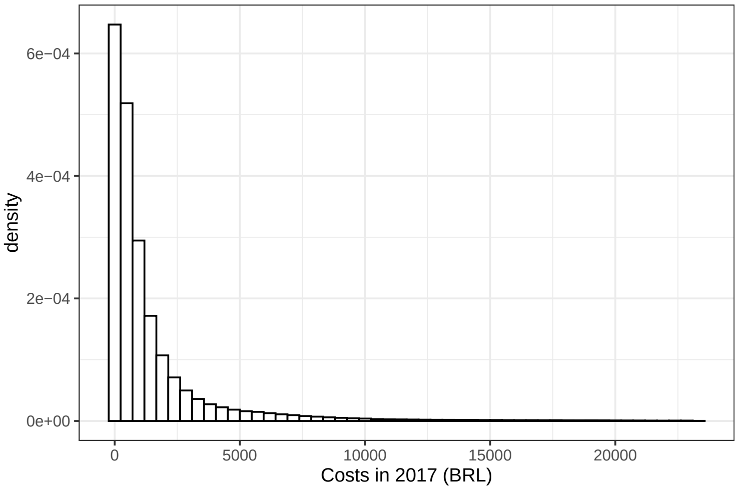


Figure 2.2: Histogram of observed costs in 2017. The costs are in BRL and were capped at 99th percentile of the distribution to allow visualisation

[6] Brazilian Real (BRL) converted to British Sterling Pounds (GBP) using the average of the daily offer exchange rates from 01/01/2017 to 31/12/2017, which was 4.1146 BRL per one GBP. The average of the year period was taken due to the large fluctuations in the exchange rates, which affect the conversion. To illustrate, exchange rates in that year ranged from 3.8033 to 4.4714. The exchange rates relate to the series "Closing quotations of one currency over a period", available in the Central Bank of Brazil web page, which can be accessed at: `https://www4.bcb.gov.br/pec/taxas/port/ptaxnpesq.asp?frame=1`

The cost distribution, when considering the claims from all policyholders, would display a higher proportion of zeros (as detailed in the paragraph above) and a longer tail, which most probably comes from the medical costs of individuals who died during the period of analysis, due to concentration of medical costs close to the end of life. This is discussed in more detail further on in this section. Nevertheless, the great majority of policies that an insurer, who has been operating in the market for many years, is exposed to usually comes from existing contracts, not new ones. And the number of individuals who cancel their policy or die is relatively very small under normal circumstances. Therefore, the distribution of claims analysed should be a good representation of the main volume of claims experienced by the insurer, taking into consideration that our focus is on policyholders who remain during the entire period of analysis.

It is commonly known that age and gender influence the health status of the individual and, consequently, their demand for care. As shown in Figure 2.3, the average costs of males and females are very similar at the early ages, with a small decrease from age 0 to age 10 for both genders - the costs of new-born individuals are generally higher. We observe a steep increase in females' costs around their twenties, mostly due to pregnancy. Males' costs rise more rapidly only from age 50. Around age 65, males' costs surpass females' costs on average, with a peak at age 90 for both genders. This is usually explained by the fact that females seek healthcare services more often and in earlier stages than men, avoiding more severe conditions later in life (Macintyre et al., 1996). Also, the lifetime healthcare costs tend to be largely concentrated close to death and, since mortality of males is higher than mortality of females, it may explain the larger costs of the former group in older ages (Mustard et al., 1998).

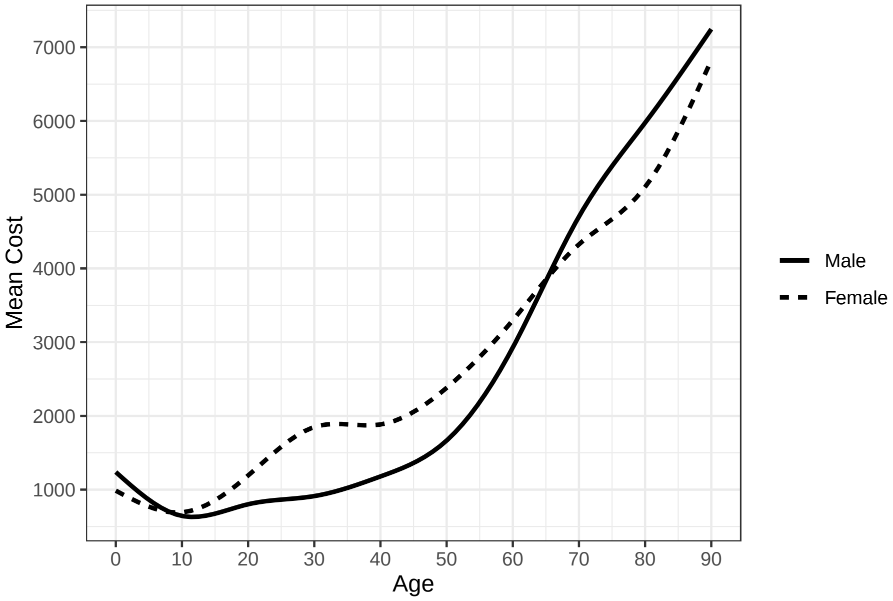


Figure 2.3: Smoothed distribution of the mean cost (in BRL) per policyholder in 2017 by age and gender

Table 2.2 helps us have a better idea of how gender and different policy designs affect the level of medical claim costs. We can see that the mean cost of policyholders with restricted plans is significantly lower than that of individuals with other plan types. The group with unregulated plan type presented the highest mean cost, most probably because a few large claim amounts skewed the mean cost. This conclusion can be reached by observing that the median cost of this group is not the largest among the plan types. Since the median cost is not affected by large claim amounts, we can conclude that there is a distortion in the mean cost of the policyholders with unregulated plan type caused by the large cost amounts. Also, this distortion is more significant in this plan type because the individuals with unregulated plan type represent a relatively smaller group than the individuals with other plan types. Policyholders with employer-based group contracts have lower mean cost than the other two contract types. The policyholders of the employer-based contract type form a younger group than the policyholders of other contract types, which may partially explain the lower mean cost. Another explanation for the lower costs can be the selection effect present in these contracts. These policyholders are fit for work, thus healthier, and so demand fewer medical services. It may also be the fact that 32.44% of the policyholders in the employer-based group have restricted plan types (which is a group that has lower costs). This proportion is relatively high considering that only 12.99% and 4.74% of policyholders with individual and association contracts, respectively, have a restricted plan type.

We can also confirm some other expected results: dependants cost less than contract owners; costs of those with contracts that cover private hospital accommodations are more expensive than those with ward coverage only; individuals who must pay co-payments have costs 36% lower than those who do not pay them; costs of individuals who declared pre-existing health conditions are 66% greater than those who did not declare.

Similar conclusions can be made based on median cost, which is also included in Table 2.2. We can observe that the mean costs are, overall, significantly larger than the median costs, which emphasises the influence of the very large claim amounts when calculating the average cost. We also include information about the inter-quartile range, which brings a fuller picture of the distribution of claim costs by gender and each policy design feature.

Another interesting observation is the proportion of policyholders who do not claim and how it varies according to gender and policy design. This is shown in the column ”proportion of zero costs” of Table 2.2. We observe a lower proportion of zero costs among policyholders who are female, with individual contract types and with pre-existing conditions. A significantly high proportion of policyholders who do not claim is observed among the individuals with unregulated plan type, which is expected due to the differences in coverage of procedures in comparison to restricted and standard plan types.

Table 2.2: Summary statistics of the total cost (in BRL) per policyholder in 2017, by policy design feature

Policy design feature	Mean	Median	Inter-quartile range	Proportion of zero costs
Gender (GENDER)				
Male	1,642.59	398.98	958.43	19.70%
Female	2,266.51	797.80	1,674.19	13.13%
Plan type (PLAN)				
Standard	2,239.63	713.21	1,544.39	14.86%
Unregulated	2,249.47	500.49	1,857.15	37.36%
Restricted	1,075.60	335.63	693.48	15.78%
Contract type (CONTR)				
Employer	1,470.21	449.87	1,068.71	18.55%
Association	2,422.66	757.99	1,782.13	18.56%
Individual	2,712.84	832.33	1,643.63	7.95%
Ownership (OWNER)				
Owner	2,359.94	715.00	1,638.01	14.50%
Dependant	1,452.04	460.69	1,026.79	18.25%
Hospital accommodation (HOSP)				
Ward	1,597.51	489.67	1,102.24	16.21%
Private	2,537.38	789.09	1,770.21	15.75%
Co-payment (COPAY)				
No	2,740.38	867.19	2,042.61	18.14%
Yes	1,739.22	532.31	1,170.95	15.29%
Pre-existing conditions (PRE)				
No	1,905.05	568.59	1,323.76	16.69%
Yes	3,167.95	1,017.96	1,975.53	6.96%

Cross-tables between two or more variables can help us identify potential interactions that could be included in the predictive models. For instance, one could think of a possible interaction between plan type, age group and gender. Table 2.3 shows the distribution of the policyholders by age and gender for each of the plan types. As discussed previously, policies under the unregulated plan type are not commercialised anymore. Consequently, the group of policyholders with this plan type are concentrated in older age groups and are predominantly female. Furthermore, we can observe in Table

2.4 that the mean cost of policyholders by age group and gender has a different pattern in each plan type, which may indicate an interaction between these variables.

Table 2.3: Distribution of policyholders by age group and gender according to plan type

Plan type	Gender	Age group										Subtotal
		[0,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60)	[60,70)	[70,80)	[80,90)	90+	gender
Standard												
	Male	6.56%	5.29%	5.65%	7.92%	6.16%	5.42%	3.49%	1.82%	0.80%	0.11%	43.21%
	Female	6.24%	5.26%	7.26%	11.28%	8.61%	7.67%	5.08%	3.20%	1.84%	0.36%	56.79%
Subtotal age group		12.79%	10.55%	12.91%	19.19%	14.77%	13.09%	8.57%	5.02%	2.64%	0.47%	100.00%
Unregulated												
	Male	0.39%	1.48%	1.93%	1.96%	2.78%	5.52%	7.46%	9.02%	5.08%	0.86%	36.48%
	Female	0.47%	1.58%	2.00%	2.56%	4.72%	9.21%	13.80%	16.77%	10.12%	2.28%	63.52%
Subtotal age group		0.86%	3.07%	3.93%	4.52%	7.51%	14.73%	21.26%	25.79%	15.20%	3.14%	100.00%
Restricted												
	Male	7.02%	6.32%	6.93%	10.18%	8.26%	5.49%	2.28%	0.83%	0.28%	0.03%	47.63%
	Female	6.73%	6.30%	7.67%	11.07%	8.92%	6.23%	3.28%	1.47%	0.63%	0.08%	52.37%
Subtotal age group		13.75%	12.62%	14.60%	21.25%	17.18%	11.72%	5.56%	2.30%	0.90%	0.11%	100.00%

Table 2.4: Mean cost (in BRL) in the prediction year of policyholders by age group and gender according to plan type

Plan type	Gender	Age group										Marginal
		[0,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60)	[60,70)	[70,80)	[80,90)	90+	mean
Standard												
	Male	1,038.82	830.51	958.00	1,161.81	1,579.83	2,346.62	4,089.08	6,098.70	7,673.40	7,571.83	1,865.04
	Female	902.37	1,061.28	1,643.06	2,139.60	2,260.01	3,079.70	4,144.18	5,514.59	6,211.53	8,062.42	2,524.63
Marginal average		972.31	945.53	1,343.39	1,736.28	1,976.27	2,776.21	4,121.75	5,726.31	6,653.92	7,948.03	2,239.63
Unregulated												
	Male	394.91	549.08	589.14	833.95	1,024.35	1,199.48	1,977.21	2,740.58	3,701.36	4,401.18	2,063.25
	Female	419.16	445.34	766.17	945.57	1,358.91	1,489.16	2,129.42	2,776.58	3,780.68	4,596.51	2,356.44
Marginal mean		408.17	495.45	679.39	897.16	1,234.87	1,380.60	2,076.00	2,763.99	3,754.16	4,543.10	2,249.47
Restricted												
	Male	516.29	361.07	534.65	645.57	810.89	1,382.00	2,503.92	4,101.07	5,950.88	6,656.06	870.14
	Female	468.60	480.68	1,076.63	1,222.78	1,241.95	1,785.03	2,520.61	3,368.41	3,533.16	6,704.50	1,262.48
Marginal mean		492.96	420.77	819.25	946.26	1,034.69	1,596.27	2,513.76	3,632.89	4,271.46	6,690.02	1,075.60

Another interesting piece of information is the distribution by age and gender of policyholders according to pre-existing conditions, displayed in Table 2.5 We can see that, as expected, there is a higher concentration of policyholders with pre-existing conditions in older age groups when compared to the group with no pre-existing conditions. Also, the proportion of female policyholders with pre-existing conditions is larger than in the the group with no pre-existing conditions.

In Table 2.6, we can a see that the mean cost by age group and gender of policyholders with pre-existing conditions has a significantly different pattern from the one for policyholders with no pre-existing conditions. This difference is more evident in the first age group. The mean costs of policyholders with pre-existing conditions whose age is between zero and ten years old are significantly higher than the mean costs of policyholders with no pre-existing condition in the same age group.

Table 2.5: Distribution of policyholders by age group and gender according to pre-existing conditions

Pre-existing	Gender	Age group										Subtotal
conditions		[0,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60)	[60,70)	[70,80)	[80,90)	90+	gender
No												
	Male	6.80%	5.60%	5.90%	8.20%	6.51%	5.38%	3.27%	1.82%	0.84%	0.12%	44.44%
	Female	6.49%	5.58%	7.14%	10.69%	8.42%	7.19%	4.75%	3.14%	1.81%	0.36%	55.56%
Subtotal age group		13.28%	11.18%	13.04%	18.89%	14.93%	12.57%	8.03%	4.96%	2.65%	0.48%	100.00%
Yes												
	Male	1.03%	1.96%	3.94%	7.44%	5.82%	6.21%	5.13%	3.20%	1.28%	0.17%	36.19%
	Female	0.85%	1.94%	6.92%	13.20%	9.74%	10.63%	9.31%	6.94%	3.66%	0.62%	63.81%
Subtotal age group		1.89%	3.90%	10.87%	20.64%	15.56%	16.84%	14.43%	10.14%	4.94%	0.79%	100.00%

Some cross-tabulation analyses uncovers some unexpected results. For instance, the great majority of policyholders with pre-existing conditions have policies with co-payment, as shown in Table 2.7. One would expect the opposite, given that these policyholders may use healthcare services more often due to their aggravated health state.

In the initial stages of our analyses, we tested the inclusion of interaction variables in the prediction models. However, they were either not significant, not included in the model because of our variable

Table 2.6: Mean cost (in BRL) in the prediction year of policyholders by age group and gender according to pre-existing conditions

Pre-existing conditions	Gender	Age group										Marginal mean
		[0,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60)	[60,70)	[70,80)	[80,90)	90+	
No												
	Male	853.62	664.60	822.12	989.68	1,331.29	2,031.41	3,568.59	5,180.59	6,503.29	6,446.21	1,562.14
	Female	755.17	869.50	1,461.70	1,883.61	1,968.32	2,706.37	3,635.78	4,767.46	5,508.93	6,985.97	2,179.28
Marginal mean		805.54	766.81	1,172.36	1,495.60	1,690.46	2,417.48	3,608.38	4,918.84	5,824.23	6,850.80	1,905.05
Yes												
	Male	6,333.95	2,424.66	1,293.96	1,489.55	1,805.72	2,775.65	4,507.91	5,724.68	7,190.67	7,977.02	2,964.76
	Female	5,268.66	2,370.44	2,075.02	2,423.93	2,519.06	3,301.04	4,050.85	4,621.70	5,352.53	8,180.45	3,283.18
Marginal mean		5,852.32	2,397.65	1,791.60	2,087.05	2,252.38	3,107.41	4,213.23	4,970.14	5,830.24	8,136.22	3,167.95

Table 2.7: Distribution of policyholders by pre-existing conditions and co-payment

Pre-existing conditions	Copayment	
	No	Yes
No	26.2%	73.8%
Yes	13.6%	86.4%

selection process (which will be explained in the next section) or they did not improve prediction accuracy. Thus, our final models do not include interactions.

2.2 Explanatory Variables

Our data collection was made in two rounds. In the first round, in addition to demographic and policy design features, the monthly claim costs of each policyholder in both years were extracted. Thus, only the amounts in each month of the observation year were known. These data were used to create 14 covariates for the models fitted in chapters three and four of this study. These covariates are described in Table 2.8. Among them, we included two covariates containing demographic information of the policyholders: AGE and GENDER. As was shown by Figure 2.3, these variables influence the health status of individuals and, consequently, the demand for medical care. We expect that claim costs increase, on average, with age and that females have larger costs than males. We also included six covariates related to policy information, since different policy design features lead to different medical costs (Table 2.2).

Because we are investigating the importance of previous cost for prediction, we included six covariates that are based on the medical claim costs of the policyholders in the observation year. The covariate logCOST represents the natural logarithm of the total medical cost in 2016. Taking the log of the previous cost amount reduces the tail of its distribution and is expected to provide a more linear relationship with the future medical cost, our response variable, which is also in logarithmic units. This almost linear relationship between the natural logarithm of the cost in the observation year (2016) and natural logarithm of the cost in the prediction year (2017) can be seen in Figure 2.4 below. The dots that form vertical and horizontal lines close to the axes represent the individuals whose cost was zero in one of the years analysed. With the exception of these individuals, we can observe a positive relationship, close to linear but surrounded by a large cloud of data points.

The frequency of medical claims (FREQ_CLAIM) in the observation year was created according to our definition in section 3.5. For consistency, it makes sense to include this covariate only in the frequency-severity model, as we want to analyse the relationship between previous claim frequency and future claim frequency and size.

The remaining cost-related covariates were created in order to capture utilisation patterns of policyholders and assess their impact on future medical costs. The work done by Bertsimas et al. (2008) inspired the use of the covariates MONTHS_AVG and propMAX_COST. These authors found that individuals with a peak cost pattern tend to have lower future medical costs than individuals whose costs are spread across the observation year. The logic behind it is that peak patterns are related to acute conditions that are likely to be treated quickly and, thus, not repeated in the future. On the other hand, frequent months with medical costs are associated with patients with chronic conditions, who are likely to maintain the demand for care in future periods. Thus, for individuals with similar medical claim amounts in the observation year, we would expect those whose propMAX_COST is

Table 2.8: Description of covariates

Covariates	Description
AGE	Age of the policyholder calculated on 1st January 2016 (observation year).
GENDER	Dummy variable representing the gender of the policyholder: male (0) and female (1).
OWNER	Dummy variable that indicates whether the individual is the owner (0) of the contract or a dependant (1) of the owner.
PLAN	Categorical variable indicating the type of plan purchased by the policyholder: standard (base category); restricted; unregulated.
CONTR	Categorical variable indicating the type of the contract purchased by the policyholder: employer-based group (base category); individual; group by association.
HOSP_ACCOMM	Dummy variable representing the type of accommodation covered by the plan in case of hospitalisation: ward (0) or private (1).
COPAY	Dummy variable that indicates whether the plan has co-payment term (1) or not (0).
PRE_EXISTING	Dummy variable that indicates if the individual declared a pre-existing condition (1) or not (0).
logCOST	Natural logarithm of the total yearly medical claim cost of a policyholder in the observation year. A value of 1 was added before taking the log, allowing the inclusion of individuals whose yearly cost was zero.
logCOST_DEC	Natural logarithm of the total medical cost of a policyholder in December 2016. A value of 1 was added before taking the log, allowing the inclusion of individuals whose cost in this month was zero.
MONTHS_AVG	Number of months that have claim costs above the policyholder‘s average monthly cost in the observation year.
propMAX_COST	Proportion of the maximum monthly cost over the policyholder‘s total claim cost in the observation year.
propLAST_THREE	Proportion of the policyholder‘s cost in the last three months of the observation year over the total yearly cost.
FREQ_CLAIM	Number of claims in the observation year.

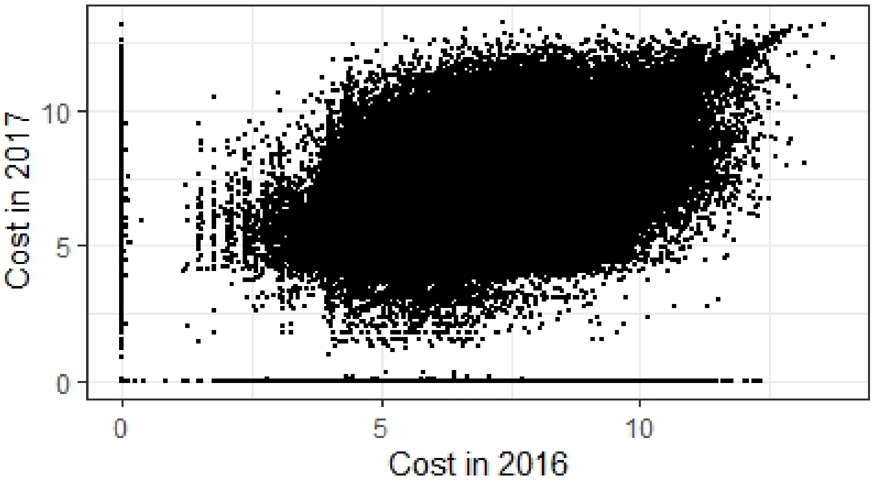


Figure 2.4: Scatter plot of the log-transformed cost in 2016 versus the log-transformed costs in 2017, including zero-cost individuals.

larger will have lower future cost than those whose propMAX_COST is lower. Also, we would expect individuals with larger MONTHS_AVG to have higher costs in the future, as they tend to look for health care more often.

The costs in the last three months of the year was also considered by Bertsimas et al. (2008) as an indicator of larger costs in the future. Thus, we expect that individuals with larger propLAST_THREE will experience higher future costs. Furthermore, we added a variable that measures the cost amount in December: logCOST_DEC. The idea is that larger costs in the last month of the observation year may indicate the beginning of an event that has costs spilled over the next year, increasing the chances of future larger costs.

An evident drawback caused by using explanatory variables based on previous costs is that the modelling of future claim costs is limited to existing policyholders only. This means that the models developed in this study are not appropriate for underwriting or pricing new policies, for example. Models tailored for the latter have many differences compared to the models developed in this study. One obvious difference is that explanatory variables based on the previous experience of each individual, such as previous claim costs, cannot be included in the latter models. Another difference is that the characteristics of the group of new policyholders are likely to be different to the characteristics of existing policyholders, which has a direct impact in their future claiming experience, and the models need to be able to capture this.

There are many models (traditional and new) that can be used for predicting the claim costs of new policyholders. Comprehensive explanations and applications of such methods can be found in Ohlsson and Johansson (2010), Frees, Derrig and Meyers (2014), Frees (2018) and Wuthrich and Buser (2020).

In order to investigate the relevance of claim details in the prediction models (chapter five), a second extraction round was made, with the aim of associating the medical procedures of each claim with the aggregated amounts.

2.3 Extracting Relevant Medical Information From Healthcare Provider Bills

The medical information of the insured lives analysed were extracted from the administrative claims data system of the insurer. These are the claims paid by the insurance company to the healthcare provider (such as hospitals, labs, clinics, etc.) for the medical services supplied to the insured. This means that each piece of information related to the medical procedure and diagnosis of the individual has, necessarily, a cost amount attached to it. It also means that procedures not covered by the healthcare policy or any procedure paid directly out of the insured's pocket is not known by the insurance company and, consequently, not taken into consideration in the analysis.

The claims are sent by the provider as invoices to be paid by the insurance company. These invoices are also known as payment requests. Each request has a unique identification code and it lists all the medical items involved in the claim, as well as the quantity and cost of each item. It also has information about the healthcare professionals involved in that event and identification of the establishment where the care was provided, along with identification of the insured who received the care. Not every single item in the invoice sent by the healthcare provider to the insurance company is paid. An analysis is made to identify particularities such as services not covered, ambiguous charges, discounts that were negotiated with the provider, etc. The cost we analyse is the one after this process, also known as "allowed charges".

The medical item is the most detailed information about the health treatment received by the individual. There are eight types of medical item, which are:

- Doctor fees
  This is the amount paid to the doctor who provided care to the individual, for instance, the visit to the doctor's office, a surgical procedure made, a specific examination such as pregnancy dating scan, among many others. Each service or procedure made by the doctor is counted as one item.

- Medical tests
  These are any types of tests requested by the doctor, from simple blood tests to RX scans and endoscopes. Each test corresponds to one item.

- Medical therapies
  The type of therapy received by the individual, for example, psychotherapy sessions or chemotherapy. Although physiotherapy is not necessarily a medical therapy, rehabilitation sessions with the physiotherapist are also covered by the healthcare plan. Usually each therapy session is considered to be one item.

- Medication
  The medicine administered by the healthcare provider during the care of the patient. Thus, all medications given to the patient during emergency, hospital inpatient or outpatient care are registered. Any prescribed medication bought by the individual in the pharmacy is not covered by the health plan, thus, not included in the data. There are various ways to count medication items. For instance, some medications are administered in pills, others in drops, millilitres, litres, boxes, etc. The names of the medications and the classification groups to which they belong follow the standards of *Brasíndice*, a reference table that lists all medications approved by *Anvisa*, the agency responsible for authorizing and regulating the use and distribution of medications in Brazil.

- Medical supplies
  These are all the supplies and equipment used in the care event. They range from the cotton balls used, gloves or camera lenses in case of video procedures to specific items such as a stent (a very expensive small metal strut implanted into a blocked person's artery to stretch it open).

- Hospital accommodation fees
  The amount paid to the hospital for each day that the patient stayed in the ward or private room during a hospitalisation or emergency event.

- Bundled payment
  In some cases, the services and items involved in an event are bundled together and paid by the insurance company to the healthcare provider as a one-off amount. In those situations, each item is not charged separately but a price is agreed between insurer and provider for that package of services. For instance, a surgery involves several medical items: tests, medications, equipment, supplies, accommodation and other fees. The insurance company can pay an amount to the provider that covers all those items involved in that surgery. With this practice, the insurance company avoids paying more for a claim due to inefficiency of the provider or the charge of extra items that were not necessary for that procedure. Each bundled payment is considered one item. The fees paid to the doctor responsible for that procedure is not included in the bundle.

- Other fees
  These are the various items charged as fees by the provider, such as oxygen supply, ventilation, patient monitoring, ambulance service, among others.

We can aggregate the data extracted at the medical item level of detail and sum the cost amount of all items involved to find the full cost of the claim. We can also add the cost of all the claims to find the total cost of an individual in a specific period. In other words, the deep level of detail allows us flexibility to aggregate and analyse the data as we wish.

The claims are also known as medical events, and can be categorised into six types, as follows:

- Doctor visits
  These are the visits to the doctor's office scheduled in advance by the patient.

- Medical tests and therapies
  These are the medical tests or therapies requested by the doctor whenever necessary. These refer to the ones made in laboratories, clinics or ambulatory (outpatient care). Thus, medical tests or therapies made in urgency/emergency care and during hospitalizations are not considered here.

- Emergency care
  As the name suggests, this is the medical event related to the care provided to individuals in the emergency services room of a hospital. It includes the urgent doctor appointment (the initial medical care given to the patient at the moment of arrival in the urgency or emergency room). This is usually done by a physician on duty and all medical items involved in the care, such as medical tests, medication provided, medical supplies used and any other fees related to the event are included.

- Hospital inpatient
  These are the events related to situations when the individual has to remain in the hospital, usually due to a surgical procedure. Besides the access to all medical items involved in each hospitalization, we also have the ICD codes (diagnoses) informed by the healthcare provider, related to that hospitalization.

- Home care
  The treatment received by the patient in their own home, usually after a hospitalisation or in case further monitoring is necessary.

- Other outpatient services
  These are the medical items (supplies, medications, bundled payments and other fees) occurred in an outpatient service that were not included in the other medical events.

2.4 Data Checking and Cleaning

The addition of the detailed medical information created a few inconsistencies, and corrections were necessary in order to guarantee that the results are reliable. For instance, there were cases where there were hospitalisation costs recorded for a policyholder during the observation year, but the number of hospitalisations and number of days in hospital during that year were zero. Thus we needed to guarantee that the medical information relates to the correct claim amount, which involved removing the inconsistent data points from the analysis.

As stated by Box (1979), model building is an iterative process. In the search for more robust results, we conduct diagnosis checks of the residuals in order to find possible problems either on the data or model specification. Outliers, for example, are known to affect the model estimation, especially when estimators are known to be non-robust in the presence of these observations, which is the case of Ordinary Least Squares[7] (Rousseeuw and Leroy, 2005). This is because, when finding the values for coefficients of the models, the OLS estimator tries to reduce the sum of residuals. When outliers are present, they have much more influence in this sum, forcing the method to prioritise their fitting, resulting in a poorer fitting of the bulk of the data.

In fact, during the diagnostics of traditional models that take into consideration the detailed medical data, we found a set of six observations that unduly influence the residuals, biasing the conclusions regarding the goodness-of-fit of such a model. In Appendix A, we show the large influence of those policyholders on the residuals and describe how each observation deviates from the majority of the cases. This analysis resulted in the removal of five of the cases, which allowed goodness-of-fit and prediction accuracy measures to be more stable.

[7] This method is used to estimate the parameters of the linear regression model, which is explained in the next chapter.

Chapter 3

Traditional Methods for Predicting Medical Claim Amounts

3.1 Introduction

Actuaries have always been interested in understanding and forecasting claim amounts, not only within the health care field but also in other lines of general (Brockman and Wright, 1992; Renshaw and Verrall, 1998) and life insurance (Rioux et al., 2019). Regression models have been a vital component in the arsenal of traditional methods used in order to complete such tasks. Parametric models have been preferred because they allow the investigation of the relationships between the future claim amounts and covariates associated with them.

Furthermore, linearity has been the method of preference because of its easy-to-implement characteristic and a straightforward interpretation of the model (Mihaylova et al., 2011). However, many challenges imposed by the claim costs data need to be overcome in order to turn them into a distribution that can be fitted by a linear model. Improvements have been made involving transformations of variables and the development of generalised linear models that allow distributions other than normal to be fitted.

Extra complications arise from other characteristics of the claim cost distribution, which include the heavy mass at zero as a consequence of a large proportion of policyholders who do not claim during a specified period, usually one year (Duncan et al., 2016). This motivates the development and use of methods that analyse the cost distribution in two parts (Frees et al., 2013). Because one policyholder can make multiple claims, these methods were extended with the creation of frequency-severity models that aim to estimate both the number of claims that a policyholder will make and their amount (Frees, Gao and Rosenberg, 2011). Technology advances allowed insurers to collect and store more and better data (Groves et al., 2016) which can be potentially used for prediction purposes. This created an additional interest in investigating ways to identify which covariates are useful for prediction and how to incorporate them in the traditional methods.

We start this chapter by describing the underlying assumptions and parameter estimation processes of linear regression models and generalised linear models, methods that provide the necessary framework for the building of two-part and frequency-severity models. We then explain the importance of variable transformation in the context of medical claim cost prediction, to help us fulfil the assumptions established when fitting the models. Further issues with collinearity and how to assess it are also tackled. Additionally, we describe the explanatory variables used in the fitting of the models before giving more details regarding our choices for model fitting.

The chapter continues with the analysis of the results, including model coefficients and a comparison of the goodness-of-fit and prediction accuracy of the traditional models fitted in order to investigate which fitted best in our data. We conclude by discussing alternative approaches when fitting traditional methods, the limitations of our approach and summarise the insights provided by our models which can be used by insurers when building and keeping data sets that will later become inputs for the construction of predictive models.

3.2 Linear Models

Linear models are among the most traditionally used methods for prediction and data analysis. The assumed linearity in the predictors makes it simpler to fit and interpret the model output. Given its simplicity, it is usually used as an initial model to be fitted in order to gain insights into the relationship between the response and the explanatory variables before more complex methods are attempted. This is the case in studies such as Frees, Gao and Rosenberg (2011), Frees et al. (2013), Duncan et al. (2016) and Yang et al. (2017). Thus, we are following the steps of these authors and start our analyses with linear models.

In our research, we include within the framework of linear models both multiple linear regression models (also known as Ordinary Least Squares) and generalised linear models. These methods are explained in more detail in the sections that follow.

3.2.1 The Multiple Linear Regression Model

In order to define a multiple linear regression, consider a data set T, with k explanatory variables, one response variable (y_i) and n observations. In our data set, each observation represents a different policyholder and the response variable is the individual's total claim amount in the next year. Thus, the information related to the i^{th} individual is $T_i = (y_i, x_{i1}, x_{i2}, ..., x_{ik})$. Fitting a multiple linear regression model for T_i would mean establishing a linear relationship between the response and the explanatory variables as defined below:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i \tag{3.1}$$

In equation 3.1, the intercept β_0 is the value that the response variable takes when all explanatory variables are set to zero; the parameters $\beta_1, \beta_2, ..., \beta_k$ represent the expected change in the response variable, on average, by varying the value of their respective explanatory variable by one unit. ϵ_i is known as the error term of the model and represents the random error factor that is not explained by the explanatory variables of the model. It is the difference between the observed values and the values estimated by the regression model, as shown in the equation below:

$$\epsilon_i = y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \tag{3.2}$$

The real values of the parameters $\beta_1, \beta_2, ..., \beta_k$ are often unknown. The usual estimation method for finding appropriate estimates for these parameters is the Ordinary Least Squares. The best estimates $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are those that minimise the quadratic form given by:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \tag{3.3}$$

Equation 3.3 is minimised by taking partial derivatives in relation to each of the parameters and equating them to zero.

There are a few assumptions underlying this model, which are necessary for parameter estimation and inference (Frees, 2009). They are as follows:

1. $\mathbf{E}(\epsilon_i) = 0$.
2. $\{x_{i1}, ..., x_{ik}\}$ are non-stochastic variables.
3. $\mathrm{Var}(y_i) = \sigma^2$.
4. $\{y_i\}$ are independent random variables.
5. $\{y_i\}$ are normally distributed.

After finding the best parameter estimates according to the OLS method, we can find the estimated response value for the i-th policyholder, which is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik} \tag{3.4}$$

3.2.2 Variable Transformation

One issue encountered by those modelling medical claim amounts using multiple linear regression is to fulfil the assumptions required in order to fit this model. Specifically, the distribution of the response variable, the next year's medical cost, is clearly not normal, as shown previously in Figure 2.2. Also, it is unlikely that the assumed linearity in the parameters is true for the distribution of costs in its original form. A popular solution to reduce skewness of the data is to transform the response variable to make its distribution approximately normal.

The transformation most commonly used for medical costs is the natural logarithm, which provides more symmetry to the data. When fitting multiple linear regression models to their data, this was one of the approaches tested by Duan et al. (1983), Diehr et al. (1999), Frees, Gao and Rosenberg (2011) and Duncan et al. (2016), to cite a few. However, before we apply this transformation directly into our data, we used another popular method in order to identify the most appropriate transformation of the response variable for our data set: the Box-Cox transformations.

Box-Cox Transformation

The Box-Cox parametric family of transformations (Box and Cox, 1964) is defined as:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log(y), & \text{if } \lambda = 0. \end{cases} \tag{3.5}$$

The task, thus, is to estimate λ, via Maximum Likelihood Estimation, resulting in a distribution of $y^{(\lambda)}$ close to Normal. One limitation of this method, though, is that it is only applicable for positive observations, thus, eliminating the zero-cost individuals from the data. The estimated λ value that maximizes the log-likelihood function for our data-set is -0.05050, which is approximately zero. Figure 3.1 shows the histogram of the transformed costs for $\hat{\lambda} = -0.05050$. The red line represents the values of the normal distribution based on the mean and standard deviation of the transformed data. The distribution seems much closer to normal than the distribution of the original amounts. The skewness of the transformed data is 0.05905, which is very close to zero, the expected value for normally distributed data. However, the positive value suggests that it is slightly skewed to the right, which is confirmed by the histogram in Figure 3.1. The kurtosis of the transformed data is 3.38676, which is close to 3, the expected value for the normal distribution.

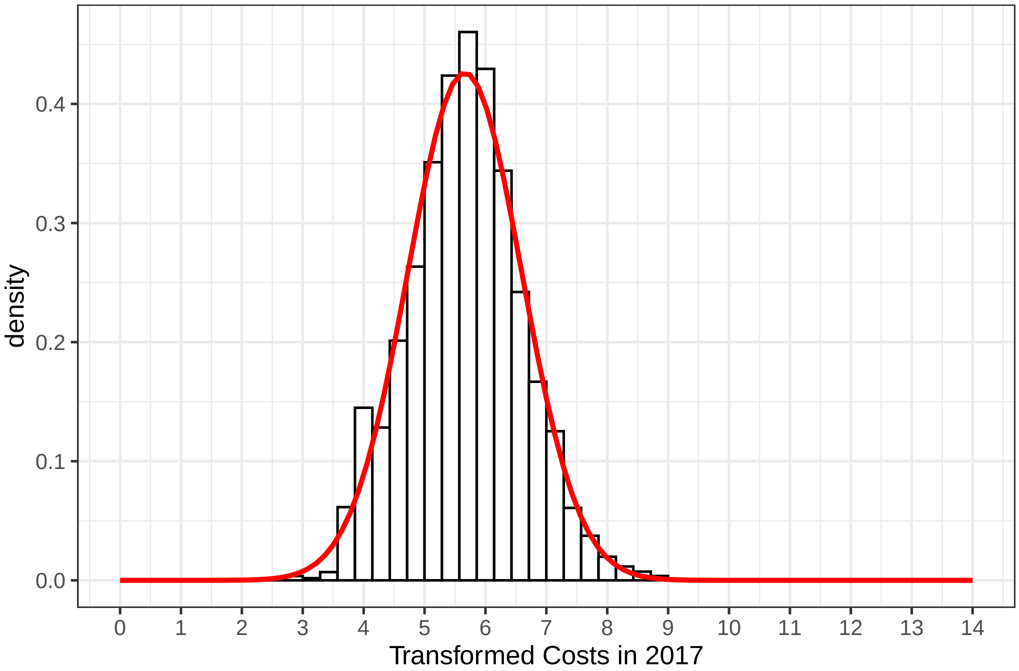


Figure 3.1: Histogram of transformed costs in 2017 based on the Box-Cox estimate of $\hat{\lambda} = -0.05050$.

Due to the higher number of zero claims within the cost distribution, one could consider the transformation $\log(y + c)$, where c is a constant usually equal to 1. This allows zero-cost individuals to be included in the model. For this case, Box and Cox (1964) proposed a modified version of the parametric family of transformations by including a shift parameter λ_2:

$$y^{(\lambda)} = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \text{if } \lambda_1 \neq 0, \\ \log(y + \lambda_2), & \text{if } \lambda_1 = 0. \end{cases} \tag{3.6}$$

For our data, the best estimates for λ_1 and λ_2 were $\hat{\lambda}_1 = 0.12521$ (which is close to zero) and $\hat{\lambda}_2 = 5.61431$ (this value is also relatively close to zero in comparison to the magnitude of claim costs). Applying the estimates to our response variable resulted in a less skewed distribution than the non-transformed distribution. The histogram that displays the distribution of transformed costs using $\hat{\lambda}_1$ and $\hat{\lambda}_2$ is shown in Figure 3.2. We can see, however, that there is a large probability mass on values close to zero, which is separate from the other values of the distribution. It resembles a distribution with two, very distinct parts. We can conclude that this transformation is inappropriate for our model.

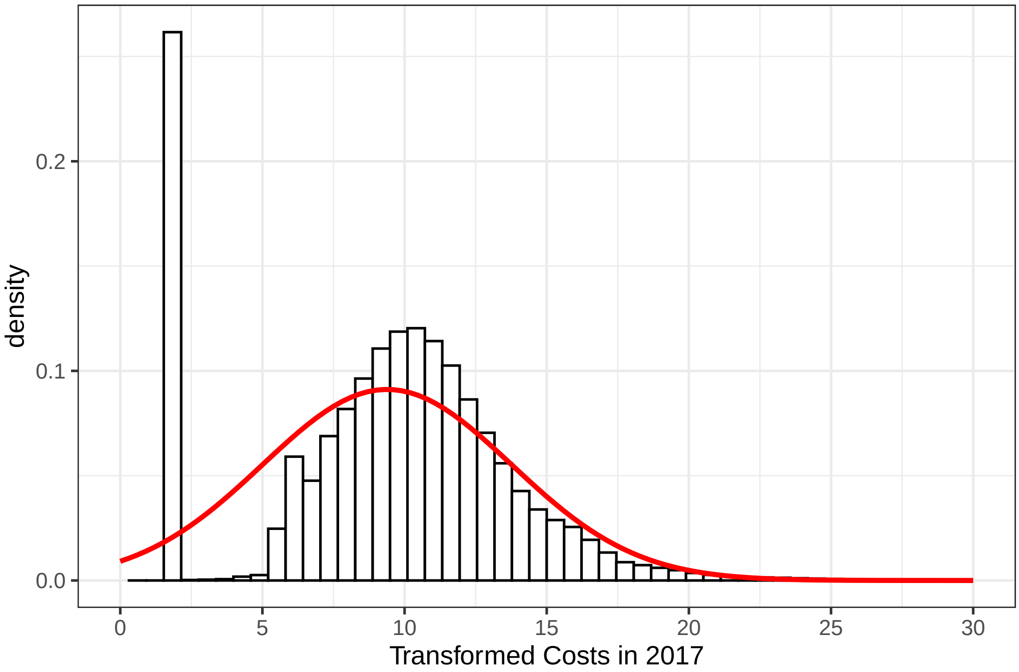


Figure 3.2: Histogram of transformed costs in 2017 based on the best Box-Cox estimates of $\hat{\lambda}_1 = 0.12521$ and $\hat{\lambda}_2 = 5.61431$.

The results displayed in Figures 3.1 and 3.2 suggest that the transformations attempted are not the most suitable for our data. However, the estimates generated are very close to zero, which suggest that the natural log could be a suitable choice for transformation, which is consistent with the literature. Figure 3.3 shows the histogram of the natural logarithm of the positive costs in 2017 (our response variable). The distribution seems more symmetric, with a skewness equal to 0.29691, being slightly closer to zero than the distribution of the data transformed using $\hat{\lambda} = -0.05050$ (Figure 3.1). Despite the fact that its kurtosis is slightly greater than 3 (3.49629), the log transformation still results in a distribution that is the closest to normal among the transformations we tested.

One issue with using log transformation is that, in order to make our model conform to the assumptions of the linear regression model, we are restricting the analysis to individuals who had positive costs in 2017, excluding those with zero costs in that year.

An important consideration that has to be made when working with transformed variables is that predictions are not made in the original scale. Since we wish to compare the predictive performance of different models, including methods that do not require variable transformation (such methods are explained in further sections of this work), we need to re-transform the fitted values back to original scale before calculating any accuracy measures.

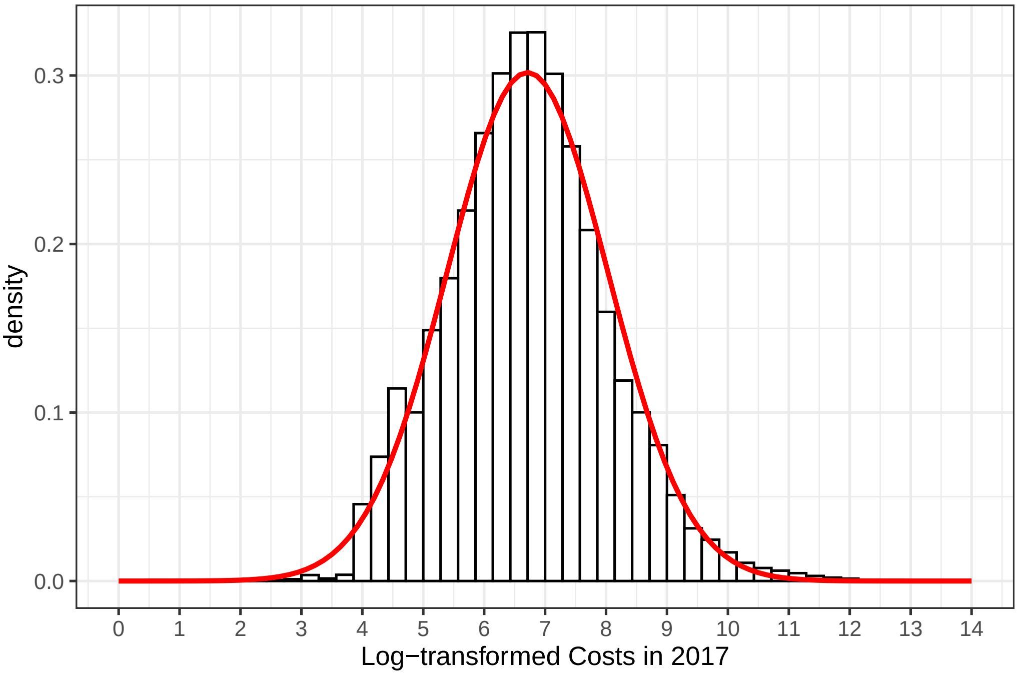


Figure 3.3: Histogram of the distribution of log-transformed positive costs in 2017.

We can define a multiple linear regression model fitted to the natural logarithm of the yearly cost, $l_i = \ln(y_i)$, as:

$$l_i = \log(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i \tag{3.7}$$

In the equation above, if $\log(y_i)$ is assumed to be normally distributed with parameters μ_i and σ^2, then y_i would follow a log-normal distribution with expected value calculated as $\mathbf{E}(y_i) = \exp(\mu_i + \frac{\sigma^2}{2})$. With that in mind, and still holding the assumption that the residuals have mean equal to zero, the re-transformed predicted value of $\hat{y}_i$ is:

$$\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik}) \cdot \exp\left(\frac{\hat{\sigma}^2}{2}\right) \tag{3.8}$$

The real value of σ^2 is unknown, which is the reason why we are using its estimate, $\hat{\sigma}^2$. It is estimated by the mean squared error (Granger and Newbold, 1976; Newman, 1993), defined as:

$$s^2 = \frac{1}{n - (k+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.9}$$

Other ways, not covered, of dealing with $\mathbf{E}(\exp(\epsilon_i))$ are to use a mean or median based approach (Newman, 1993) or smearing (Duan, 1983). So far, we have covered the main ideas of linear regression models, one of the most important and traditional methods for modelling. However, also within the realm of linear models, there is a method that allows linear combinations of parameters to be fitted to distributions other than normal: the generalized linear models (GLMs). These are important as they will provide a framework for other methods to be built on.

3.3 The Generalized Linear Models

GLMs expand the multiple linear regression model by allowing a distribution for the response variable to be different than the normal (Nelder and Wedderburn, 1972). Once again, we use the ideas and definitions available in Frees (2009) in order to describe how GLMs work.

We start by defining the *systematic component* of our model, which is the linear combination of parameters and covariates: $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$. We also define the mean of the response variable as $\mu_i = \mathbf{E}(y_i)$. A link function $g(.)$ is responsible for mapping the systematic component to the mean response. Combining the descriptions above, we can define a GLM as:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} \tag{3.10}$$

The choice of distribution function is made based on the exponential family of distributions, which is defined as:

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + S(y,\phi)\right) \tag{3.11}$$

In equation 3.11, θ is known as the natural parameter and ϕ is a scale parameter. Distributions such as the Gamma, Bernoulli (for binary outcomes), Poisson (for counts) and normal are all part of the exponential family of distributions. This means that they can be generally represented in the form of equation 3.11.

The choices for the link function of a model are usually related to the distribution defined for the response variable. It is also important to observe the domain of the mean response in order to avoid estimated values outside the range of possible values. In our case, for instance, the response variable is the claim cost amount, which cannot take negative values. Thus, it is important that we use link functions that do not lead to negative values. When the link function is defined as the inverse of the mean function, it is known as the canonical link of the model.

The parameters of the model are usually estimated via the Maximum Likelihood method. From equation 3.11, the log-likelihood function can be defined as:

$$L = \sum_{i=1}^{n} \log f(y_i|\theta_i,\phi_i) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi_i} + S(y_i,\phi_i) \right\} \tag{3.12}$$

For the canonical link of a distribution, θ_i is replaced by $(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik})$, creating a direct relationship with the coefficients that we are interested in estimating. The Maximum Likelihood Estimators $(\beta_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik})$ are found by minimising equation 3.12.

The aim of this section is to introduce the ideas involved in the generalized linear models and how they improve upon linear regression models. In the future sections where we describe how we fitted the models into our claims data, we will provide further details regarding the choice of the distribution for the response variable and how we defined the link functions in each case.

Having defined and described linear regression models and GLMs, we can now move to methods that rely on this linear framework in order to fit claim amounts. In particular, we will be discussing methods that aim to fit the distribution in different stages, tackling one of the main challenges arising from the cost data: a large mass at zero-cost.

3.4 Two-Part Model

Two-part models, as the name suggests, split the fitting process of the claim cost distribution into two stages. It comes from the idea that actuaries usually see claims data as having two parts: one part indicating the number of claims that occurred and the other part indicating the amount (or severity) of the claim (Frees, 2009).

The definition of the two-part model adopted is the one used by many authors for predicting healthcare costs, including Buntin and Zaslavsky (2004), Frees, Gao and Rosenberg (2011) and Frees et al. (2013) to cite a few. This is also a popular model among health economists who wish to analyse the healthcare expenditures decomposed in two parts as seen in Jones (2000), Buntin and Zaslavsky (2004) and Mihaylova et al. (2011). According to this definition, the first part of the model estimates the probability that at least one claim occurs during the specified period. In our case, the first part estimates the probability of a policyholder making at least one claim during the year following the observation year. The second part estimates the yearly claim amount conditional on the occurrence of at least one claim. It is usual to fit the two parts independently, with an underlying assumption that the yearly claim amount is not related to claim occurrence. Because they are fitted separately, each part of the model can use a different set of covariates.

There are a few advantages to using this approach in order to predict claim amounts. As explained by Frees (2009), some covariates may be relevant for describing the likelihood of a policyholder making a claim but not be significant for estimating claim amounts and vice-versa. Or the same covariate may have different impacts for claim occurrence and claim amount. The two-part method allows the separate analysis in each context.

As concluded by Mihaylova et al. (2011) in their review of methods used to analyse healthcare costs, two-part models tend to outperform one-stage parametric models in an environment of a large proportion of zeros, which is our case.

In order to define the first stage of the model, suppose that the binary variable r_i indicates that the i-th policyholder makes at least one medical claim in a given year:

$$r_i = \begin{cases} 0, & \text{if policyholder } i \text{ does not claim during the year} \\ 1, & \text{if policyholder } i \text{ makes at least one claim during the year} \end{cases} \tag{3.13}$$

In our study, we fit the probability of occurrence of a medical claim within the GLM framework. We assume that r_i follows a Bernoulli distribution and we use logit as the link function of our model. This approach is commonly used for this purpose, as seen, for instance, in Frees, Gao and Rosenberg (2011) and Frees et al. (2013) and for other investigations involving binary outcomes. The logit function is the canonical link of the Bernoulli distribution and maps the linear combination of covariates to the mean within the range $[0,1]$. For an individual i, the logit function is defined as $\log\left(\frac{\pi_i}{1-\pi_i}\right)$. Thus, by estimating the value of the mean π_i of a Bernoulli distribution, we are estimating a probability, ranging from 0 to 1. Within this context, the estimated probability of occurrence of a medical event for an individual i is:

$$g(\pi_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$$

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik})} \tag{3.14}$$

For the second part, assume that the continuous variable y_i^* represents the yearly claim amount of policyholder i. If $r_i = 0$, then $y_i^* = 0$. In our study, we will use the multiple linear regression model (equation 3.8) on the natural logarithm of the yearly medical costs of individuals with positive claim amounts in the prediction year. Since we intend to make predictions in the original scale of cost amounts, we need to re-transform the fitted value of the second stage, as discussed previously.

The product of first and second parts provides the estimate of the yearly medical claim costs. For a policyholder i, the predicted amount is:

$$\hat{y}_i = \hat{\pi}_i \cdot \hat{y}_i^*$$

$$\hat{y}_i = \left(\frac{\exp(\hat{\beta}_{0,1} + \hat{\beta}_{1,1} x_{i1} + ... + \hat{\beta}_{k,1} x_{ik})}{1 + \exp(\hat{\beta}_{0,1} + \hat{\beta}_{1,1} x_{i1} + ... + \hat{\beta}_{k,1} x_{ik})}\right) \cdot \exp(\hat{\beta}_{0,2} + \hat{\beta}_{1,2} x_{i1} + ... + \hat{\beta}_{k,2} x_{ik}) \cdot \exp\left(\frac{\hat{\sigma}^2}{2}\right) \tag{3.15}$$

Where $\hat{\beta}_{0,1}, \hat{\beta}_{1,1}, ..., \hat{\beta}_{k,1}$ are the maximum likelihood coefficient estimates for the first part of the model and $\hat{\beta}_{0,2}, \hat{\beta}_{1,2}, ..., \hat{\beta}_{k,2}$ are the OLS coefficient estimates for the second part of the model.

3.5 Frequency-Severity Model

As an extension to the two-part model, frequency-severity models take into consideration the feature that several claims might occur during a specified period. For situations where the insurance company records the payment for every single claim, this model can be used to estimate the aggregate loss of the insurer across a group of policyholders or an entire line of business.

According to Klugman et al. (2012), a *collective risk model* can be defined as follows:

$$S = X_1 + X_2 + ... + X_N \ , \quad N = 0, 1, 2, ... \tag{3.16}$$

Where S = 0 when N = 0. The random variables $X_1 + X_2 + ... + X_N$ represent the amount of each claim, and N counts the number of claims paid in the period.

Traditionally, the definition given by equation 3.16 has been used by actuaries when predicting total losses. However, our frequency-severity models are based on the approach followed by Frees, Gao and Rosenberg (2011) in order to estimate individual-level healthcare frequencies and expenditures. They consider that the yearly medical cost of a policyholder is the sum of all of their medical events occurred within that period. Thus, instead of finding the total loss arising from the claims of a group of policies as defined by Klugman et al. (2012), we are estimating the policyholder's yearly cost by aggregating the costs of each of his/her medical events in a year.

The definition in equation 3.16 assumes that the number and amount of medical events are independent and that each claim amount follows the same distribution and also that claim amounts are independent from each other. This is a reasonable assumption, although Frees et al. (2013) found some degree of dependency among frequency and severity of medical events.

We initially need to fit the frequency of medical events in a given year. Because the number of medical events was not available in the data set used, we tried to indirectly infer the value of this variable by using the monthly cost of each policyholder. We counted consecutive months with positive medical costs as one single event. If the months with positive costs are followed by months with no cost, we count them as distinct events. That is to say, if an individual had positive medical costs in all of the 12 months of the calendar year, we count that as one medical event. Consequently, the maximum number of medical events possible is six. We followed this approach for the year 2017, the prediction period, which generated our response variable for the frequency of future medical events. We repeated the same process for the year 2016, which we used as one explanatory variable of the model.

One limitation arises from this approach: the policyholders with positive medical cost in only one month of the calendar year were considered to have one medical event, as well as those with 12 months with positive medical costs. However, the costs of the former are likely to be much lower than the costs of the latter, who experienced a very long medical event. This assumption was made because a medical event can give rise to a series of procedures, such as medical tests, appointments with different physicians, hospitalisations, among others. Those procedures are likely to be spread over more than one month, especially when the event is a consequence of the diagnosis of a more serious health condition.

On the other hand, medical events with very short duration and that repeat over consecutive months are counted as only one event. For instance, imagine a policyholder who sees a doctor with a dermatology specialisation and no further tests were requested, this medical event had duration of only one day. In the next month, if the same policyholder goes to a doctor with a different specialisation, this will be considered by us as being the same event, when it is clearly two separate events. A distinct policyholder, who was hospitalised for two months, for instance, will also have that considered as one event. In this case, this is a better representation of a long medical event, with procedures related to the treatment of the same condition. However, given that the data gathered at this point is aggregated by month and policyholder, this is the best estimation of the frequency of claims we could achieve.

We assumed that the number of medical events of an individual i in a year (N_i) follows a Poisson distribution with mean λ_i. We also used the canonical link function of that distribution, the log-link, in order to fit a GLM for the frequency part. Thus, our model for the frequency of claims is as follows:

$$g(\lambda_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$$

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$$

$$\lambda_i = \exp\left(\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}\right) \tag{3.17}$$

For the severity part, a variable with the average cost per medical event was created as the ratio between the total medical cost of the individual in the prediction year and the number of medical events, as defined above, in the same year. We fitted a multiple linear regression model to the log-transformed cost per claim. In order to assess the model in the original scale of costs, a retransformation was performed (equation 3.8). Thus, the final estimate for the future yearly cost of policyholder i according to our frequency-severity model is:

$$\hat{y}_i = \hat{\lambda}_i \cdot \hat{y}_i^{'}$$

$$\hat{y}_i = \exp\left(\hat{\beta}_0^f + \hat{\beta}_1^f x_{i1} + ... + \hat{\beta}_k^f x_{ik}\right) \cdot \exp(\hat{\beta}_0^s + \hat{\beta}_1^s x_{i1} + ... + \hat{\beta}_k^s x_{ik}) \cdot \exp\left(\frac{\hat{\sigma}^2}{2}\right) \qquad (3.18)$$

Where $\hat{\beta}_0^f, \hat{\beta}_1^f, ..., \hat{\beta}_k^f$ is the set of maximum likelihood coefficient estimates for the frequency part and $\hat{\beta}_0^s, \hat{\beta}_1^s, ..., \hat{\beta}_k^s$ is the set of coefficient estimates defined via OLS for the severity part.

There is an important generalisation of equation 3.16 which consists in modelling the aggregate claim amount using a compound process. Claim arrivals are modelled by a point process $N(t)$, which is a stochastic process that considers the time when each claim arrives t as a random variable. The claim sizes are usually modelled assuming a sequence of independent and identically distributed random variables. There is an extensive literature on the application of this method in the insurance context, including Arjas (1989); Norberg (1993) and more recently Dimitrova et al. (2017); Cai et al. (2019) and Dimitrova et al. (2020)

After defining the models that are fitted to our data, in the following section we describe an analysis we performed in order to check whether there is strong collinearity among our covariates.

3.6 Collinearity Analysis

In regression analysis, highly correlated explanatory variables may contain redundant information and introduce collinearity into the model, which negatively affects its robustness and coefficient estimation in many ways (Hocking, 2003).

According to Frees (2009), coefficients of collinear variables can have inflated standard errors, reducing the significance of the respective covariates (the t-ratio is reduced). Model interpretability is affected in the sense that it is difficult to determine the individual impact of collinear covariates on the response variable, as explained by James et al. (2013), which is aggravated when such variables have opposite signs. In other words, if two collinear covariates with opposite signs contain the same information about the response variable, the effect on the response of one of the covariates might be cancelled out by the other and a better model would be achieved when only one of those covariates is introduced. In some cases, a very important covariate might be assessed as not significant because its contribution is being obscured by a collinear covariate.

Collinearity is an issue likely to occur in models using medical data. Duncan et al. (2016) observed that many medical covariates turned out to be statistically redundant, being rejected by their models. They identified strong collinearity between expenditure variables and Co-existing Condition (CC) codes (binary variables that categorise and simplify diagnosis-related information of individuals). This problem is aggravated for traditional predictive methods such as linear regression models, which do not have inherent mechanisms to avoid or alleviate this problem.

It is also important to remember that, given the nature of administrative claims data, there is no medical or diagnostic claim information without a respective cost amount paid for that event. This creates an environment that facilitates the occurrence of collinearity, as the importance of a piece of medical information may be overshadowed by the related cost amount.

The reasons highlighted above justify the collinearity analysis of the explanatory variables before fitting the models. The most common and simplest method used to identify collinearity is by constructing a correlation matrix of predictors. Each element of the matrix represents the Pearson correlation coefficient between the column variable and the row variable. A high correlation (in absolute value) indicates that including that pair of variables may cause collinearity. However, deciding the threshold for "high correlation" is rather subjective and varies from study to study. A rule-of-thumb proposed by Kuhn and Johnson (2013) is to identify as collinear covariates those whose correlation coefficient is above 0.75 or below -0.75.

Table 3.1 shows the correlation matrix for our numerical covariates. It can be seen that the categorical variables are excluded from this analysis. This is because the Pearson correlation coefficient only makes sense for numerical variables. This coefficient measures how the values of one variable increase or decrease as the values of another variable change. For categorical data, this sense of increase or decrease is not applicable. For instance, in terms of gender, we cannot determine that female is greater than male or vice versa. For this reason, only numerical covariates are chosen. We can observe that none of the coefficients has an absolute value larger than 0.75. Thus, we do not have an apparent evidence of collinearity.

As expected, higher correlations are observed between logCOST and other cost-related covariates (MONTHS_AVG: 0.665, and FREQ_CLAIM: 0.682). This is because these variables are based on cost amount in the observation year, which means they carry similar information. Also, MONTHS_AVG and FREQ_CLAIM have a high correlation coefficient, probably as a consequence of the method we used to estimate the frequency of claims. However, this correlation (0.697) is still below the 0.75 threshold.

Table 3.1: Correlation matrix of numerical covariates

Covariate	AGE	logCOST	logCOST_DEC	propLAST_THREE	propMAX_COST	MONTHS_AVG	FREQ_CLAIM
AGE	1.000						
logCOST	0.158	1.000					
logCOST_DEC	0.170	0.450	1.000				
propLAST_THREE	0.038	0.310	0.461	1.000			
propMAX_COST	-0.041	0.442	-0.034	0.195	1.000		
MONTHS_AVG	0.036	0.665	0.366	0.220	-0.090	1.000	
FREQ_CLAIM	0.001	0.682	0.359	0.256	0.163	0.697	1.000

One down side of the correlation matrix is that it only identifies potential collinearity among pairs of covariates. However, multicollinearity may occur if two or more covariates are linear combinations of another covariate. A popular tool used to identify multicollinearity within the linear regression context is the Variance Inflation Factor, defined by James et al. (2013) as:

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{3.19}$$

Where $R^2_{X_j|X_{-j}}$ is the coefficient of determination of a linear regression model that has X_j as the response variable and the remaining X_{-j} as explanatory variables. Large values of $R^2_{X_j|X_{-j}}$ mean that the linear combination of X_{-j} can sufficiently explain X_j, suggesting that it can be removed from the final model without compromising its predictive power. A common practice is to interpret VIF values above 5 (which occurs when $R^2_{X_j|X_{-j}} > 0.8$) as indicative of some level of multicollinearity, with values above 10 (for values of $R^2_{X_j|X_{-j}}$ greater than 0.9) indicating strong collinearity (Frees, 2009; James et al., 2013).

Since we have a few covariates that are based on cost amounts in the observation year, we conducted a variance inflation factor analysis for all numerical covariates, which is shown in Table 3.2. The covariates are sorted from the largest VIF value to lowest. According to our expectations, logCOST presents the highest value for VIF, but it is still below the threshold of 5. This result suggests that there is no evidence of multicollinearity, so all covariates will be considered as candidates for our models.

Table 3.2: Variance Inflation Factor analysis for numerical covariates

Covariate	**VIF**
logCOST	4.494
MONTHS_AVG	3.326
FREQ_CLAIM	2.384
propMAX_COST	2.215
logCOST_DEC	1.696
propLAST_THREE	1.367
AGE	1.107

3.7 Variable Selection and Model Assessment

3.7.1 Stepwise Variable Selection

Part of the process of building predictive models is identifying which variables will produce a better fit to the data. For this purpose, a stepwise selection method (James et al., 2013) was applied. The Bayesian Information Criterion (BIC), introduced by Schwarz (1978) was the measure of choice to support the decision to include and maintain a covariate in the model. The BIC can be defined as:

$$\mathrm{BIC} = -2 \cdot \mathrm{L}(\hat{\theta}_{\mathrm{MLE}}) + k \cdot \log(n) \tag{3.20}$$

Where $\mathrm{L}(\hat{\theta}_{\mathrm{MLE}})$ is the log-likelihood of the specific function evaluated at the model‘s maximum-likelihood estimator parameters ($\hat{\theta}_{\mathrm{MLE}}$). This represents the improvement in goodness-of-fit provided by the addition of one or more parameters into the model. However, a penalisation term $[k \cdot \log(n)]$, inflates the BIC as more covariates (k) are added. Thus, the lower the level of BIC, the better is the model fit. We chose BIC as the criterion for variable selection because, for large samples such as ours, the penalisation term becomes larger, forcing the production of more parsimonious models, with fewer covariates.

The stepwise method starts by calculating the BIC of a null model, i.e. a model that only has an intercept and no covariates. Then it chooses the covariate that causes the highest decrease in the BIC observed for the null model to be added to the model. In the next step, the method adds another covariate, among the ones remaining, that results in the largest reduction in the BIC of the model with one covariate. The model adds a third covariate, again the one resulting in the lowest BIC. However, after adding that covariate, the method tests whether removing any of the previously added covariates would result in a lower BIC than a model with three covariates. If so, then that covariate is removed, with the possibility of being re-introduced to the model in further steps. This process is repeated until no more reductions in BIC are observed.

3.7.2 Data split

As explained by Friedman et al. (2001), the accuracy of predictions based on the data used to build the model tend to be greater than the accuracy of predictions based on an unseen sample. This can distort our judgement when comparing the predictive accuracy of different models. More complex models, which tend to be more flexible and adapt to the particularities of the data, are more likely to produce lower errors for the data to which they were fitted than less flexible models. To avoid this bias when selecting the best model, and because we have a large amount of data, we randomly split the data set into two samples:

- a **training sample**, which corresponds to 75% of the full data set and is used to build the model. The stepwise selection is conducted in this sample.
- a **test sample**, which corresponds to 25% of the original data set, and which will only be used for model assessment and comparison after the model is selected.

The proportion of the data allocated to each of the samples is somewhat subjective, but we decided to allow more data points to be used for the training of the model in order to improve model robustness.

For each model, variable selection is made by applying a stepwise method based on BIC using the training sample. Goodness-of-fit measures are calculated based on the estimated values in the training sample and prediction accuracy measures are calculated using the data points of the test sample.

Although the test sample does not contain the exact same data points as the training sample, both samples must be similar. For instance, if we have values of covariates in the test sample that lie outside the range of values for that covariate in the training sample, the prediction for those points would be extrapolations of the fitted model, which will cause a reduction in their prediction power. We can see in Table 3.3 a summary of the numerical covariates in the training and test sample. All covariates have very similar values for descriptive statistics in both samples, which indicates that one sample is a good representation of the other. A similar conclusion can be made for the categorical variables based on the results showed in Table 3.4.

Table 3.3: Descriptive statistics of numerical covariates in training and test sample

Covariate	**Sample**	**Descriptive statistics**					
		Min.	p25%	Median	Mean	p75%	Max
logCOST	Training	0.000	5.041	6.321	5.586	7.247	13.728
	Test	0.000	5.039	6.324	5.589	7.249	13.253
logCOST_DEC	Training	0.000	0.000	0.000	1.649	4.410	12.290
	Test	0.000	0.000	0.000	1.652	4.410	11.660
MONTHS_AVG	Training	0	1	3	2.741	4	11
	Test	0	1	3	2.744	4	11
propMAX_COST	Training	0	25	41	43.950	62	100
	Test	0	25	41	43.970	62	100
propLAST_THREE	Training	0	0	9	19.890	31	100
	Test	0	0	9	19.830	31	100
FREQ_CLAIMS	Training	0	1	2	1.973	3	6
	Test	0	1	2	1.973	3	6
AGE	Training	0	21	37	37.680	53	90
	Test	0	21	37	37.590	53	90

Table 3.4: Descriptive analysis of categorical covariates in the training and test samples

Covariate	**Training**	**Test**
GENDER		
Male	43.852%	43.894%
Female	56.148%	56.106%
PLAN		
Standard	74.483%	74.643%
Unregulated	4.268%	4.152%
Restricted	21.248%	21.205%
CONTR		
Employer	52.461%	52.312%
Association	23.620%	23.658%
Individual	23.918%	24.030%
OWNER		
Owner	59.567%	59.561%
Dependant	40.433%	40.439%
COPAY		
No	25.316%	25.382%
Yes	74.684%	74.618%
HOSP_ACCOMM		
Ward	57.956%	57.882%
Private	42.044%	42.118%
PRE_EXISTING		
No	93.056%	93.025%
Yes	6.944%	6.975%

3.7.3 Goodness-of-Fit and Prediction Accuracy Measures

The criteria used by different authors to evaluate model prediction performance usually vary and it is uncommon to see a full overlap of the methods used across the literature. The measures used in this study are based on their popularity among recent studies aiming to predict healthcare costs, such as Cumming et al. (2002), Bertsimas et al. (2008), Frees, Gao and Rosenberg (2011), Duncan et al. (2016) and Morid et al. (2017). Some of these statistics can be used to compare accuracy among other papers while some can only be used to compare models within this study, because they are expressed in the units of the response variable.

We calculated the same statistics in the training and test sample. The results based on the training data are called goodness-of-fit measures and the ones based on the test data points are called prediction accuracy measures.

Coefficient of Determination

The coefficient of determination, commonly known as R^2 can be defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{3.21}$$

This traditional measure, which is intended for measuring goodness-of-fit of Multiple Linear Regression models, is one of the most published among the studies reviewed. Although it allows comparison of different models in the literature, its use in the context of medical costs prediction received criticism from a few authors. The cost outliers, which are generally poorly fitted, inflate the residual sum of squares (numerator of equation 3.21), potentially offsetting the model‘s goodness-of-fit for lower cost individuals (Cumming et al., 2002; Bertsimas et al., 2008; Duncan et al., 2016).

Mean Absolute Error

The Mean Absolute Error can be defined as:

$$\text{MAE} = \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.22}$$

Although very popular, this measure does not allow for comparison among different studies, as it is affected by the variation in the data and the unit of response variable. Thus, it will be used for comparison among models in this study only.

Mean Absolute Proportional Error

In order to allow the comparison of error magnitude for data with different units, the mean absolute proportional error is calculated as the average of absolute ratio of each residual over the mean value of the response variable (Morid et al., 2017). This is represented in the equation below:

$$\text{MAPE} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{\overline{y}} \right| \tag{3.23}$$

One advantage of this measure over R^2 is that it is not based on squares of residuals. This tends to alleviate the impact of very large residuals in the overall accuracy measure.

Quantile-Mean Absolute Deviation

As mentioned above, it is very likely that outliers will be poorly fitted by the model. In order to verify how well the model fits the non-catastrophic medical costs, Duncan et al. (2016) used the quantile-Mean Absolute Deviation, defined as:

$${}_q\text{MAD}_\alpha = \frac{1}{(\alpha \cdot n)} \cdot \sum_{i:|y_i - \hat{y}_i| < q_\alpha} |y_i - \hat{y}_i| \tag{3.24}$$

This measure is similar to the Mean Absolute Error, except it calculates the average absolute deviation for all observations below the α-quantile. In order to stay in line with Duncan et al. (2016), we chose $\alpha = 0.95$, which eliminates the top 5% costliest individuals.

Spearman Rank Correlation Coefficient

The Spearman Rank Correlation Coefficient is a useful measure to identify whether the ranking of model predictions would be a good representation of the ranking of observed medical costs, moving away from judging the model performance based on the pure predictions. It is applicable for situations where the insurer is interested in ranking individuals based on their predicted cost amount for selecting which ones can be included in intervention programs, for instance (Duncan et al., 2016). It is defined as:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} \mathrm{rg}(y_i) - \mathrm{rg}(\hat{y}_i)}{n \cdot (n^2 - 1)} \tag{3.25}$$

Where $\mathrm{rg}(y_i) - \mathrm{rg}(\hat{y}_i)$ is the difference between the rank of the i^{th} observed value and the rank of the i^{th} predicted value.

Gini Statistic

The Gini statistic (also called Gini coefficient) is widely used to measure the level of income or wealth inequality in a population (Frees, 2018). The statistic summarizes the information contained in the Lorenz curve, a method proposed by Lorenz (1905) to investigate the relationship between income and population distributions. The Lorenz curve is represented as a graph where the proportion of the population (sorted according to the level of income, from lowest to highest) is the horizontal axis and the proportion of income or wealth is in the vertical axis. A 45 degree line is observed if both proportions are perfectly equal (in other words, if every individual in the population has exactly the same income). This diagonal is known as the line of equality. The Lorenz curve is the observed distribution of income or wealth in a population.

The Gini statistic is the ratio of the area between the Lorenz curve and the line of equality over the total area under the line of equality. A Gini statistic equal to zero represents total equality while a Gini statistic equal to one represents total inequality (only one individual has all the income / wealth, while the other individuals have none).

In Frees (2018), we can find a modified version of the Lorenz curve that can be useful when comparing two distributions. It is known as performance curve, which is a graph of the distribution of the proportion of claims versus the proportion of premiums. The idea of using the performance curve for insurance applications was introduced by Frees, Meyers and Cummings (2011) and extended by Frees, Meyers and Cummings (2014). This extension is called "ordered" Lorenz curve. The line of equality, in this case, represents the situation where the rank of the proportion of premiums is exactly the same as the rank of the proportion of claims. In other words, the 45 degree line of equality represents the situation where the lowest $s\%$ of the premiums comes from policies responsible for $s\%$ of the total claim amount, for every s between 0% and 100%. Instead of premiums, we are using the Gini statistic based on the performance curve to compare the observed claims distribution with the distribution of the predicted values, which is an application suggested by Frees (2018). A Gini statistic close to zero suggests that the ranking of the predicted claims matches the ranking of the observed claims more closely, which can be interpreted as an evidence of good prediction accuracy. According to Frees, Meyers and Cummings (2011), the Gini statistic based on the performance curve varies between -1 and 1. A negative value indicates that the model tends to underestimate the claim amount, while a positive value indicates overestimation.

In order to calculate the Gini statistic, the first step is to find the cumulative proportion of predicted values $\hat{F}_P(s)$ and the cumulative proportion of observed claim amounts in the prediction year $\hat{F}_L(s)$, which are given by equations 3.26 and 3.27, respectively, adapted from Frees (2018):

$$\hat{F}_P(s) = \frac{\sum_{i=1}^{n} \hat{y}_i I(\hat{y}_i \leq s)}{\sum_{i=1}^{n} \hat{y}_i} \tag{3.26}$$

$$\hat{F}_L(s) = \frac{\sum_{i=1}^{n} y_i I(\hat{y}_i \leq s)}{\sum_{i=1}^{n} y_i} \tag{3.27}$$

Where s represents the possible values between the minimum and maximum predicted values and $I(\cdot)$ is an indicator function, which returns 0 if false or 1 if true. Please note that, in both equations, $I(\cdot)$ depends on the predicted values $\hat{y}_i$. This is because both cumulative proportions are sorted according to the predicted values.

The empirical Gini statistic based on $\hat{F}_P(s)$ and $\hat{F}_L(s)$ is calculated according to equation 3.28 below, which was adapted from Frees (2018):

$$\widehat{Gini} = 1 - \sum_{i=0}^{n-1} \Big(\hat{F}_P(\hat{y}_{i+1}) - \hat{F}_P(\hat{y}_i)\Big) \cdot \Big(\hat{F}_L(\hat{y}_{i+1}) + \hat{F}_L(\hat{y}_i)\Big) \tag{3.28}$$

Where $\hat{F}_P(\hat{y}_0) = 0$ and $\hat{F}_L(\hat{y}_0) = 0$. Also following the presentation of results adopted by Frees (2018), we multiply the result from equation 3.28 by 100, which results in a Gini statistic represented as a percentage.

3.8 Model Fitting

We fitted three different models into the data set aggregated by policyholder. The first model is a multiple linear regression fitted to the individual's log-transformed yearly medical cost (eq. 3.8). As explained previously, only individuals with positive costs in the prediction year are considered for this model. The next model fitted, the two-part model (eq. 3.15), corrects this problem and all policyholders are included in the analysis. Frequency-severity (eq. 3.18) is the third model to be fitted and is also applied to the entire group of policyholders.

In the next section we will analyse the relevance of the coefficients selected in each of the models as well as judge which model provides better goodness-of-fit and prediction accuracy. For completeness, we will include the goodness-of-fit and accuracy measures of all models, but according to our judgement the direct comparison is only appropriate for the two-part and frequency-severity models because they were fitted to the same group of policyholders. The data to which the first model was fitted is a subset of this group, which makes a direct comparison with the other two unfair.

3.9 Results

3.9.1 Coefficient Estimates of The Multiple Linear Regression Model

The coefficient estimates of the multiple linear regression, our initial model, can be found in Table 3.5. The fact that all of the 13 potential covariates were selected by the stepwise method suggests the importance of including variables that reflect demographics, policy design and previous claim costs when modelling future medical costs.

Table 3.5: Coefficient estimates of the multiple linear regression model fitted to policyholders with positive costs in the prediction year

Covariate	**Estimate**	**Std. Error**	**t value**	**Pr(>$\|t\|$)**
(Intercept)	5.19100	0.00790	657.139	<2e-16
AGE	0.01308	0.00008	160.963	<2e-16
GENDERFemale	0.16810	0.00311	53.994	<2e-16
OWNERDependant	-0.01244	0.00340	-3.658	2.54e-04
PLANUnregulated	-0.21890	0.00910	-24.053	<2e-16
PLANRestricted	-0.32140	0.00420	-76.523	<2e-16
CONTRAssociation	0.05067	0.00400	12.674	<2e-16
CONTRIndividual	0.03411	0.00384	8.895	<2e-16
HOSP_ACCOMMPrivate	0.04753	0.00355	13.376	<2e-16
COPAYYes	-0.09501	0.00395	-24.046	<2e-16
PRE_EXISTING	0.08170	0.00580	14.081	<2e-16
logCOST	0.25560	0.00115	221.531	<2e-16
logCOST_DEC	0.06472	0.00074	87.272	<2e-16
MONTHS_AVG	-0.05372	0.00130	-41.451	<2e-16
propMAX_COST	-0.01081	0.00008	-137.211	<2e-16
propLAST_THREE	-0.00069	0.00006	-10.812	<2e-16

Regarding the demographic covariates, we could see that the positive coefficient related to AGE is in accordance to our expectation of higher need for healthcare as individuals get older. Also, the model confirms that, overall, female policyholders tend to utilise medical services more than male individuals.

The coefficients related to policy design covariates also reinforces the impact of different features of insurance contracts on the amount of medical claims. Policyholders with an unregulated plan type are expected to have lower claim amounts than those with standard types (the baseline category for PLAN). This may be because their policies do not cover some of the medical procedures covered by standard plans, causing them to self-insure part of their healthcare services. Policyholders with restricted plans also present, on average, lower future medical costs than those with standard plans, although they are covered for the same medical procedures. This result is a direct consequence of the limitations in the access to a healthcare provider network and price negotiation between provider and insurer. It can also show evidence of self-selection from the perspective of the policyholder: for as long as they consider themselves healthy, they remain with the cheaper plan and lower coverage, since extra coverage is not perceived by them to be necessary.

Policyholders with individual/family contracts tend to have larger costs than those who have employer-based group contracts (the baseline category for CONTR). As explained previously, the decision to buy an individual contract comes entirely from the policyholder, who is more inclined to buy it if they perceive a future need for medical coverage, either because they already have some health condition or they just want extra protection against large medical expenses. On the other hand, employers usually pay fully or partially for the medical insurance policies of their employees and so the purchase of insurance cover does not depend directly from the policyholder's perception of their health state. Thus, there are more policyholders who need less care in the group with employer-based contract type than in the group with individual/family contract type. Also, these policyholders are fit to work and consequently demand less medical care, which characterises a selection of lives that benefits the insurer.

The co-payment feature in the policy also works towards reducing the expected costs of a policyholder, as suggested by the negative sign of the coefficient of COPAY. This can be either because those who decide to have a policy with a co-payment clause consider themselves as healthy individuals with less need for health care, or because the policyholders want to avoid paying for unnecessary medical services, making them claim less than those who do not have a co-payment clause.

The remaining three covariates related to policy design present coefficients that reinforce our expectations. Individuals who declared pre-existing conditions tend to have larger future costs than those who have not declared. Patients who stay in private rooms during hospitalisations are more expensive as well, since the fees for this type of accommodation are higher. Lastly, dependants of policyholders tend to have lower claims than the owners of the policy.

The relatively large positive coefficient of the covariate logCOST emphasizes the relationship between previous and future medical costs. Overall, as we suspected, the larger the previous policyholder's yearly medical cost is, the larger their future yearly cost tends to be. Larger costs in the last month of the observation year (logCOST_DEC) also tend to lead to larger yearly costs in the following year, although its impact is lower than the previous yearly medical costs (logCOST). Undoubtedly, logCOST is at least as large as logCOST_DEC, which emphasises the contribution of these two covariates in the prediction of future medical costs. The model also confirmed that policyholders with costs largely concentrated in one month (propMAX_COST) tend to have lower future costs.

Interestingly, two cost-related covariates had coefficients that suggested an opposite effect on future medical costs than we presumed. MONTHS_AVG was expected to indicate larger future costs, since it was assumed by us and the literature (Bertsimas et al., 2008; Morid et al., 2017) that a larger number of months with claim costs above the average would indicate a chronic pattern, leading to more frequent treatments, leading to a larger yearly medical cost overall. For this model and for our data set, this does not seem to be the case. This result suggests that individuals with more frequent claims tend to have a lower yearly cost than those with less frequent claims. Also, propLAST_THREE was negative, suggesting that a larger concentration of the policyholder's yearly medical cost in the final three months of the observation year leads to lower costs in the following year. Although significant, this covariate's coefficient is very small, and it is likely that removing the covariate from the model would not impair its prediction performance and the value of other coefficient estimates.

3.9.2 Coefficient Estimates of The Two-part Model

We begin by analysing the coefficient estimates of the first stage of the model, which fits the likelihood of a policyholder to have a positive medical claim cost in the prediction year. Their values can be found in Table 3.6 below. The model contains 12 covariates, since propLAST_THREE was not selected by the stepwise method. This result merely confirms that this covariate is not very relevant for claim prediction since the coefficient for this covariate was already very small for the linear regression model fitted previously (Table 3.5).

Table 3.6: Coefficient estimates of the first stage of the two-part model

Covariate	**Estimate**	**Std. Error**	**z value**	**Pr(>$\|z\|$)**
(Intercept)	-0.78121	0.01959	-39.868	<2e-16
AGE	-0.00218	0.00027	-8.043	8.78e-16
GENDERFemale	0.08360	0.00939	8.908	<2e-16
OWNERDependant	-0.10365	0.01048	-9.894	<2e-16
PLANUnregulated	-0.83397	0.02379	-35.063	<2e-16
PLANRestricted	0.23294	0.01230	18.935	<2e-16
CONTRAssociation	0.04910	0.01214	4.044	5.24e-05
CONTRIndividual	0.48780	0.01374	35.501	<2e-16
HOSP_ACCOMMPrivate	0.05370	0.01126	4.770	1.84e-06
COPAYYes	0.12152	0.01294	9.389	<2e-16
PRE_EXISTING	0.30803	0.02400	12.835	<2e-16
logCOST	0.40002	0.00471	84.886	<2e-16
logCOST_DEC	0.14476	0.00370	39.131	<2e-16
MONTHS_AVG	0.36132	0.00687	52.569	<2e-16
propMAX_COST	-0.00546	0.00022	-25.264	<2e-16

The coefficients of some covariates present an opposite effect on the likelihood of making at least one claim than they did on the claim amount. For instance, surprisingly, AGE has a negative coefficient in the first stage of our model, which is counter-intuitive. This may be due to the effects of other covariates. The typical older policyholder may be more likely to have claimed last year, and have had a number of months of treatment last year. Therefore the average older patient is more likely to claim than the average younger patient because these other values are typically higher. Another possible contribution for the negative sign of AGE is the removal of the policyholders who died during the prediction year from the analysis. As discussed previously, the individual's largest medical costs are experienced in their final year of life. Also, the removed individuals died mostly in advanced ages. Thus, the medical costs in the end of life might be large enough to make the AGE coefficient become positive.

Another surprise is the positive coefficient for restricted plan type. However, this can be explained by the fact that this plan type was developed to attract policyholders with lower income. Most of these policyholders probably did not have previous access to the private healthcare providers, being subject to low quality of the services offered by the Brazilian public system. From the moment they are allowed to use a (restricted) private healthcare provider network, they may wish to make the most of this opportunity, which may cause them to be more likely to claim than policyholders with a standard plan type.

Also, we can see that policyholders with a co-payment clause are more likely to claim than those whose policy does not have this feature, which initially goes against our expectations. However, this suggests that, for this particular insurer, the co-payment values established do not stop policyholders looking for healthcare services when needed. In fact, the presence of a co-payment should not have this purpose, since the aim is to avoid unnecessary claims and not avoid policyholders from looking for care. Nevertheless, once they claim, the yearly cost amount tends to be lower than the claim costs of policyholders without co-payment (see Table 3.7). This makes sense if we think that individuals with co-payment policies generally have better health conditions, which means that the medical procedures related to their claims are less complex, hence less expensive, than individuals without co-payment.

Furthermore, MONTHS_AVG has a positive coefficient for the likelihood of making a claim, as opposed to the negative sign observed for the linear regression model fitting the claim amount (Table 3.5). This means that individuals whose claims are scattered throughout the year have a higher chance of making a claim in the next year than individuals whose claims are more concentrated within the observation year.

Overall, the remaining covariates have the same sign on the model related to the likelihood of making a claim (Table 3.6) as they do on the model that predicts the yearly claim amount analysed previously (Table 3.5). The characteristics of policyholders more likely to have positive yearly claim costs are: females; those with individual or group by association types of contract; the patients that stay in private hospital accommodation; policyholders who declared pre-existing conditions; individuals with larger costs in the observation year and individuals with larger costs in the last month of observation year. Policyholders with lower chances of claiming in the following year are: dependants; those who have unregulated plan type policies; those whose costs are largely concentrated in one month.

The coefficient estimates of the second stage, which fits the policyholder's yearly claim amount given that there was at least one claim, can be found in Table 3.7 below. Since we used a linear regression model on the log-transformed cost of the prediction year and the model was applied only to individuals with positive costs in the prediction year, the coefficient estimates of this stage are exactly the same as the estimates for the first model fitted (Table 3.5). Because the interpretation of the coefficients for that model has already been made, we will move forward to the analysis of the coefficients of the frequency-severity model in the following sub-section.

Table 3.7: Coefficient estimates of the second stage of the two-part model

Covariate	**Estimate**	**Std. Error**	**t value**	**Pr(>$\|t\|$)**
(Intercept)	5.19100	0.00790	657.139	<2e-16
AGE	0.01308	0.00008	160.963	<2e-16
GENDERFemale	0.16810	0.00311	53.994	<2e-16
OWNERDependant	-0.01244	0.00340	-3.658	2.54e-04
PLANUnregulated	-0.21890	0.00910	-24.053	<2e-16
PLANRestricted	-0.32140	0.00420	-76.523	<2e-16
CONTRAssociation	0.05067	0.00400	12.674	<2e-16
CONTRIndividual	0.03411	0.00384	8.895	<2e-16
HOSP_ACCOMMPrivate	0.04753	0.00355	13.376	<2e-16
COPAYYes	-0.09501	0.00395	-24.046	<2e-16
PRE_EXISTING	0.08170	0.00580	14.081	<2e-16
logCOST	0.25560	0.00115	221.531	<2e-16
logCOST_DEC	0.06472	0.00074	87.272	<2e-16
MONTHS_AVG	-0.05372	0.00130	-41.451	<2e-16
propMAX_COST	-0.01081	0.00008	-137.211	<2e-16
propLAST_THREE	-0.00069	0.00006	-10.812	<2e-16

3.9.3 Coefficient Estimates of The Frequency-Severity Model

Table 3.8 shows the coefficient estimates of the frequency model. The model contains 12 covariates, since the coefficient estimates of the covariates CONTR and HOSP_ACCOMM were not significant. Thus, they were removed from the model. In order to interpret the coefficients of the model, we need to refer back to the method we used to estimate the frequency of claims for each policyholder: that is, consecutive months with positive costs were counted as one claim. With that in mind, many coefficients that seem odd at first glance start to make more sense. Additionally, some of the covariates will be analysed with their counterparts in the severity model, allowing for a more complete interpretation of the results.

One of the features that contributes to lower frequency of claims is AGE, which has a negative coefficient. This could mean that, as a consequence of the tendency for individuals' health state to deteriorate as they get older, medical procedures are likely to be more complex, which generates either longer events (eg. hospitalisations with long duration) or consecutive events separated by short periods of time (which, with our methodology, means treating them as one claim). In each case, claims will tend to be considered longer and less frequent according to our estimation method. Thus, it makes sense that AGE has a negative coefficient for the future number of claims. To support this argument, we can check the coefficient estimates of the severity model in Table 3.9. AGE was also a selected covariate in this model, with a positive coefficient, which corroborates the idea of longer, more expensive, hence less frequent claims as individuals get older.

Negative coefficients of other covariates may imply a different interpretation. This is the case for dependants and policyholders with unregulated plan type. In both cases, it may just mean that these

Table 3.8: Coefficient estimates of the frequency model

Covariate	**Estimate**	**Std. Error**	**z value**	**Pr(>$\|z\|$)**
(Intercept)	-0.45190	0.00490	-92.263	<2e-16
AGE	-0.00162	0.00005	-32.695	<2e-16
GENDERFemale	0.03546	0.00193	18.337	<2e-16
OWNERDependant	-0.01707	0.00206	-8.275	<2e-16
PLANUnregulated	-0.08963	0.00562	-15.944	<2e-16
PLANRestricted	0.00805	0.00240	3.357	7.88e-04
COPAYYes	0.01142	0.00234	4.892	9.97e-07
PRE_EXISTING	0.02632	0.00351	7.504	6.19e-14
logCOST	0.10110	0.00073	137.776	<2e-16
logCOST_DEC	-0.01023	0.00045	-22.822	<2e-16
MONTHS_AVG	0.07007	0.00084	83.553	<2e-16
propMAX_COST	0.00197	0.00005	37.679	<2e-16
propLAST_THREE	0.00097	0.00004	24.191	<2e-16
FREQ_CLAIMS	0.11040	0.00100	110.200	<2e-16

individuals tend to claim less, on average, than their counter-parts (owners of the policies in the case of dependants and standard plan type policyholders for the unregulated plan type policyholders). In Table 3.9 we can see that PLANUnregulated has a negative coefficient for severity as well, which endorses this interpretation. OWNERDependant was not selected by the stepwise method to be in the severity model, which impacts our interpretation in this case.

The last covariate with a negative coefficient estimate in the frequency model is logCOST_DEC. This suggests that larger costs in the last month of the observation year may be an indication of the beginning of an event that will finish at the start of the following year, and no further claims will be necessary. Alternatively, the positive coefficient of this covariate in the severity model (Table 3.9), may be saying that larger costs in December are an indication of less frequent but longer (and more expensive) claims in the future.

The features of policyholders that are related to more frequent claims in the future (supported by the expected positive coefficient for the claim frequency) are: female individuals; policyholders who declared pre-existing conditions; policyholders with larger number of months with cost above average in the observation year; policyholders with larger costs in the observation year; larger proportion of costs in the last three months of the observation year; frequency of claims in the observation year, which indicates that the individual with frequent claims tend to repeat this pattern in the following year. Also, their claims tend to be of lower amount, as shown by the negative coefficient related to FREQ_CLAIMS in Table 3.9.

Nevertheless, some of the coefficients that imply a positive impact in the expected number of claims are completely the opposite to what we expected. For instance, propMAX_COST was expected to be negative, as acute conditions tend to be treated and not generate further claims. Policyholders of restricted plan types tend to claim more frequently than those with a standard plan type policy. However, their average claim amount is lower, as expected (Table 3.9).

Individuals whose policy contains co-payment tend to claim more (or, in the case of two-part model, more likely to make a claim). However, their claims are less expensive, which emphasises the point we made regarding the better health state of a policyholder whose policy contains co-payment in comparison to those without that clause.

Overall, the covariates of the severity model, that have not been discussed so far, have the expected coefficient size and sign. Being female, having a contract type different than employer-based group, staying in private hospital rooms, declaring pre-existing conditions and having larger yearly costs in the observation year are indicatives of larger average cost per claim.

In 3.9.4 we will show the goodness-of-fit and prediction accuracy measures of each of the models. A comparison between the measures of two-part and frequency-severity will be made, alongside a comparison with results found in the literature.

Table 3.9: Coefficient estimates of the severity model

Covariate	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.83200	0.00752	642.339	<2e-16
AGE	0.01362	0.00008	180.287	<2e-16
GENDERFemale	0.13170	0.00308	42.713	<2e-16
PLANUnregulated	-0.21610	0.00904	-23.905	<2e-16
PLANRestricted	-0.29530	0.00418	-70.706	<2e-16
CONTRAssociation	0.04547	0.00398	11.434	<2e-16
CONTRIndividual	0.03115	0.00375	8.307	<2e-16
HOSP_ACCOMMPrivate	0.04296	0.00353	12.179	<2e-16
COPAYYes	-0.09221	0.00393	-23.459	<2e-16
PRE_EXISTING	0.07077	0.00577	12.266	<2e-16
logCOST	0.24980	0.00119	210.005	<2e-16
logCOST_DEC	0.07307	0.00074	98.864	<2e-16
MONTHS_AVG	-0.04966	0.00139	-35.662	<2e-16
propMAX_COST	-0.01009	0.00008	-126.971	<2e-16
propLAST_THREE	-0.00102	0.00006	-16.016	<2e-16
FREQ_CLAIMS	-0.18090	0.00167	-108.078	<2e-16

3.9.4 Goodness-of-fit And Prediction Accuracy Of Traditional Models

Table 3.10 shows the six goodness-of-fit measures of the three models fitted to the aggregated data. They were calculated based on the data points used to fit the models, i.e., the training sample. The results of the multiple linear regression model are separated from the results of the other two models by a dashed line. This is to reinforce the idea that we should be careful when comparing their goodness-of-fit measures. For instance, the mean absolute error (MAE) of the multiple linear regression is very large when compared to the MAE values of the two-part and frequency-severity models. However, this happened because we restricted this model to the policyholders with positive costs in the prediction year, which makes the average cost in that year become higher. To illustrate this, the average cost in the prediction year of the whole group of policyholders, considering the ones whose cost was zero, is 1,992.84. Removing the individuals who had zero costs in the prediction year makes the average cost rise to 2,372.78. Thus, a fairer way to check whether the MAE of the linear regression model is large or not is to analyse it relative for the average cost of the group to which the model was fitted. This is what the mean absolute proportional error (MAPE) does. We can see in Table 3.10 that the MAPE of the multiple linear regression is slightly lower than the MAPE of the two-part and frequency-severity models, suggesting that its MAE is, in fact, not large in relation to the average cost of the group. A similar logic applies to the larger qMAD$_{0.95}$ measure of the multiple linear regression.

Table 3.10: Goodness-of-fit measures based on training sample

Goodness-of-fit measures	Multiple Linear Regression	Two-part	Frequency-severity
R^2	0.10254	0.11081	0.15838
MAE	1,974	1,696	1,766
MAPE	0.83097	0.85020	0.88523
qMAD$_{0.95}$	1,042	871	958
Spearman Correlation	0.61242	0.70247	0.70110
Gini	7.08884	7.29779	4.67548

Overall, all three models provided a low R^2, which was in fact expected given the issues of fitting medical claim costs and the results found in the literature. Making a direct comparison between the two-part and the frequency-severity models, we can see that the latter had a better performance in terms of R^2 and Gini statistic than the former. However, for the remaining measures, the two-part model performed better than the frequency-severity model.

The MAE and MAPE of the two-part model are 3.96% lower than the ones of the frequency-severity model. The superiority of the two-part model over the frequency-severity model is even higher in terms of $\text{qMAD}_{0.95}$. This measure is 9.08% lower for the two-part model, which means that this model is better at fitting the bulk of the data than the frequency-severity model. On the other hand, the Spearman correlation coefficient values of the two models are very similar, showing no superiority of either model by this criterion.

Based on these measures, we can conclude that the two-part model provided a better fit to the training data than the frequency-severity model, but how would the performance of these models be in a data set different to the one used to fit the models? The answer to this question is shown in Table 3.11, which displays the prediction accuracy measures of the three models in the test sample. Once again, the results from the multiple linear regression are separated by a dashed line to avoid direct comparisons with the other two models.

We can observe initially that, for all three models, the values of all measures based on the test sample are very similar to their counter-parts calculated using the training sample. This shows that the estimations remain consistent despite the change in the data points, and none of the models tends to over-fit the training data. Over-fitting is not necessarily expected in this situation because our models so far are not very flexible (they are based on a linear combination of parameters) and the number of parameters in each of the models is very small relative to the size of the data.

Table 3.11: Prediction accuracy of traditional models based on test sample

Prediction accuracy measures	**Multiple Linear Regression**	**Two-part**	**Frequency-severity**
R^2	0.10241	0.11027	0.15768
MAE	1,971	1,695	1,765
MAPE	0.83298	0.85249	0.88807
$\text{qMAD}_{0.95}$	1,041	872	959
Spearman Correlation	0.61077	0.70032	0.69922
Gini	7.70056	7.93416	5.25542

In line with the goodness-of-fit results, the prediction accuracy of the two-part model is superior to the frequency-severity model. The MAE and MAPE of the two-part model are 4.01% lower than the values produced by the frequency-severity model. The $\text{qMAD}_{0.95}$ is 9.08% lower. The frequency-severity model, however, has a R^2 that is 4.74 percentage points better than the one from the two-part model. In line with the goodness-of-fit measures, we also observe a better performance of the frequency-severity model in terms of Gini statistic. The Spearman correlation coefficient of both models presents no significant difference.

One might expect that the two-part model would have a higher R^2 measure than the frequency-severity model because of its lower MAE, MAPE and $\text{qMAD}_{0.95}$ measures. The worse R^2 is caused by the influence of the very large residuals which resulted from the individuals with large costs in the prediction year. Because the coefficient of determination is based on the squared residuals (eq. 3.21), the very large residuals have a much larger influence in the measure, offsetting the good fit achieved in the bulk of the data which is related to individuals with lower costs. This effect was also pointed out by Cumming et al. (2002) and Bertsimas et al. (2008).

Thus, although the frequency-severity model did a worse job overall, individuals with the largest cost amounts were fitted better by this model than by the two-part model, making its R^2 value higher. This is illustrated by Table 3.12. It shows the MAPE values for each cost decile for the two-part and frequency-severity models.

Each cost decile represents 10% of the total cost in the prediction year of the 198,962 policyholders of the test sample. The policyholders were then sorted according to their yearly cost in the prediction year and allocated to each cost decile until their cumulative sum reaches 10% of the total cost. In other words, the sum of the yearly cost of 125,657 policyholders in the first cost decile is the same as the sum of the yearly cost of the 30,312 policyholders in the second cost decile and so on. Many more policyholders are necessary for the first cost decile because they have the lowest costs of the group - some of them have zero cost in the prediction year. The average cost of the policyholders in the first cost decile is only 314.74 BRL as opposed to 1,304.74 BRL for the second cost decile.

For each cost decile, MAPE was calculated based on the following equation:

$$\mathrm{MAPE}_d = \frac{1}{n_d} \cdot \sum_{i=1}^{n_d} \left| \frac{y_{i,d} - \hat{y}_{i,d}}{\bar{y}_d} \right| \tag{3.29}$$

Where n_d represents the number of individuals in cost decile d, $y_{i,d}$ is the observed yearly cost of the i^{th} policyholder of decile d, $\hat{y}_{i,d}$ is the predicted cost of the i^{th} policyholder of decile d and $\bar{y}_d$ is the average yearly cost of the policyholders in decile d.

We can see in Table 3.12 that the MAPE values of the two-part model are lower (hence, better) than the MAPE values of the frequency-severity model for deciles 10% to 50%. There are 190,412 policyholders in these deciles, which represents 95.70% of the total policyholders in the test sample. On the other hand, the frequency-severity model has lower MAPE in each decile from 60% to 100%, which contain the 4.30% policyholders with the largest costs. We can also observe that both models fit very poorly the low costs of the policyholders in the first decile. There is a large difference between the MAPE values of this decile and the MAPE values of the remaining nine deciles.

Table 3.12: Mean Absolute Proportional Error (MAPE) values in the test sample of the two-part and frequency-severity models for different cost deciles

Cost decile	**Number of policyholders**	**Average cost**	**Two-part**	**Frequency-severity**
10%	125,657	314.74	**2.14957**	2.36274
20%	30,312	1,304.74	**0.80236**	0.91800
30%	17,416	2,270.87	**0.54157**	0.64807
40%	10,476	3,775.05	**0.45638**	0.52989
50%	6,551	6,037.53	**0.51583**	0.54344
60%	4,193	9,431.60	0.61146	**0.59234**
70%	2,357	16,773.96	0.74629	**0.70271**
80%	1,221	32,404.36	0.85181	**0.81189**
90%	563	70,056.85	0.90414	**0.86187**
100%	216	183,601.87	0.94534	**0.90955**

We can observe that the MAPE figures in Table 3.12 are relatively high, particularly taking into consideration the first cost decile, which contains the vast majority of policyholders. Thus, MAPE suggests that the linear regression models used are not very good at predicting claim costs. However, the Gini statistic of both the two-part and frequency-severity models (Tables 3.10 and 3.11) are close to zero, which is a sign of good model fit and prediction accuracy. In fact, this is one of the advantages of the Gini statistic, because it is measuring how closely related the ranks of predicted and observed claim amounts are, instead of how large the residuals are. If the ranking of predicted claim amounts is a good representation of the ranking of observed claim amounts, the insurers could re-scale the the predicted values by a constant (or different constants), for instance, in order to compensate for the large residuals produced.

To conclude, the better performance of the frequency-severity in predicting the costs of the top 4.30% policyholders with largest costs offsets its poorer performance in the remaining data when residuals are squared for the calculation of R^2. Also, it produces a ranking of predicted costs that is more closely related to the ranking of observed claim costs, as demonstrated by the better Gini statistic. However, MAE, MAPE and $\mathrm{qMAD}_{0.95}$ show us the superiority of the two-part model in terms of goodness-of-fit and prediction accuracy for the aggregated data.

3.10 Final Considerations

In this chapter, we described the assumptions and estimation mechanisms behind the traditional methods used by actuaries in order to forecast medical claim amounts. Our approach was limited to the field of linear models, which are very popular, easier to interpret and are able to provide reasonable results both in terms of model inference and prediction accuracy.

Within this context, we applied methods that can be replicated in order to overcome issues that arise from fitting medical costs using administrative claims data. One of them is variable transformation in order to achieve linearity and remove skewness from the data. Box-Cox transformation suggested that, in line with the literature, log-transformation was the most pertinent for our data, but other functions can be applied for different situations and contexts.

The issue with a large mass over zero caused by the high proportion of policyholders who do not claim was addressed by using a two-part model, which allowed us to split the analysis of the claim costs into the likelihood of making a claim - first part of the model - and the claim amount given that there was at least one claim - second part. We discovered that such a division is relevant, because the same covariate can cause different impacts in each part. For instance, we found out that policyholders with a co-payment clause are more likely to make a claim, but their yearly cost tends to be lower than the costs for policyholders who do not have this feature in their contracts.

An extended version of this model, the frequency-severity model, was also applied and we found, among other discoveries, that frequency of claims of a policyholder in the observation year influences their future frequency of claims.

The GLM framework conveniently allowed the use of appropriate distributions for the varying purposes that each part of the models fitted had, i.e., modelling the likelihood of making a claim, the frequency of claims or the claim amount. As in every study, we had to make choices that seemed to fit our data the best, but they are far from being the only choices applicable. For example, the likelihood of making at least one claim was fitted using the logit link function, but Frees (2009) also mentions the probit function as an alternative for estimating the first part of the two-part model. For the number of claims, Poisson was the distribution of choice, but the negative binomial distribution is also a possibility that could be explored (Frees, 2009).

For claim amount we decided to use the multiple linear regression model for log-transformed values of the yearly cost. Other authors (Frees, Gao and Rosenberg, 2011; Frees et al., 2013; Duncan et al., 2016) considered the Gamma distribution within the GLM framework in order to estimate claim amounts. However, because our data comes from a market that does not apply limits for maximum claim amounts, our cost values can be much larger values than the values in the data of the cited studies, which means that our cost distribution have a much longer tail. We did not want to truncate the amounts as done by Duncan et al. (2016), since we are interested in investigating how the models fit these values. Also, in their review of statistical methods for predicting health care costs, Mihaylova et al. (2011) did not recommend the use of the Gamma distribution, since its tail is not long enough to appropriately fit the cost distribution. Thus, we abstained from using it.

Another model traditionally used by actuaries when modelling insurance claim amounts is based on the Tweedie distribution (Tweedie, 1984). This is a mixture of a discrete component and a continuous component and it is also part of the GLM framework. This model assumes that each policyholder has a sequence of independent and identically distributed claims following a Gamma distribution. This is the continuous part. The discrete part is represented by a random variable N which is the number of claims arriving during a determined period and is considered to follow a Poisson distribution (Frees, 2009). Thus, the Tweedie distribution is a Poisson sum of Gamma random variables and has been used by actuaries to model the aggregate loss (or total loss) experienced by the insurance company.

The relevant applications of the Tweedie distribution in claims modelling include Jørgensen and Paes De Souza (1994), Wüthrich (2003), Furman and Landsman (2010) and more recently Halder et al. (2021), among others. Further details about the Tweedie distribution can be found in Frees (2009), Klugman et al. (2012) and Frees, Derrig and Meyers (2014).

Although we focused on the linear regression and GLM in our study, one could also use generalised distributions to model claim severity, given their ability to fit long-tail distributions. In health economics, for instance, the generalised gamma distribution (Stacy et al., 1962) is used to model healthcare costs (Manning et al., 2005). It is considered a flexible approach to modelling claim severity, since it has three parameters: one scale parameter and two shape parameters (Manning et al., 2005; Frees, 2009). Also, many standard distributions such as exponential, log-normal and Weibull are special cases of this distribution, which is helpful when identifying the best model for a particular data set.

The generalised beta distribution of the second kind, also known simply as GB2, is even more flexible than the generalised gamma. It was introduced by Venter (1983) for modelling insurance claims and by McDonald (1984) in the context of modelling wealth distributions. It has four parameters, which makes it a suitable choice for both light-tail or heavy-tail data (Frees, 2018). It can also accommodate negatively skewed data, with a left tail. Cummins et al. (1990) apply this distribution for modelling fire losses. An application in longitudinal data is provided by Sun et al. (2008).

We also confirmed the relevance of potential explanatory variables available in the administrative claims data for cost prediction. All our models agree that female policyholders not only have a higher tendency to claim (or, in the case of frequency-severity models, tend to make more claims than male policyholders) but also have larger claim amounts. Yearly claim costs tend to increase with age, but the two-part and frequency-severity models revealed the rather odd result that the likelihood of making a claim decreases with age. Nevertheless, we reaffirmed the importance of demographic covariates in predictive models.

Likewise, covariates reflecting differences in policy design are key in order to control for variations in medical claim costs that are not directly related to the individual's health state. Still, many studies seem to overlook the importance of these features when predicting claim costs. Our covariates derive from the particularities of the Brazilian market, so modifications are expected and necessary depending on the market needs, geographic region and regulatory obligations. For instance, an insurer may decide to include a categorical covariate that describes different levels of co-payment, instead of a binary covariate indicating the presence of this clause in the policyholder's contract, which was our case. Regardless of the variations that might exist, policy design covariates are relevant for medical claim cost prediction. In fact, other types of information may be relevant for claim cost prediction. For instance, the geographical location of healthcare providers (which, in a majority of cases, may be publicly available) can be important predictors.

Furthermore, previous claims cost amount was a relevant predictor in all the traditional methods tested in this chapter, confirming its impact on the following year's claim amount. Not only that, but covariates capturing utilisation patterns also provided insights about the individual's future costs: larger amounts in the last month of the observation year tend to increase the estimated claim cost of the following year; patients with a concentration of cost during observation year are less likely to claim than patients whose costs were spread over the year. Nonetheless, analysing the concentration of the cost in the final three months did not seem so relevant.

For the aforementioned reasons, an insurer that wants to fit traditional methods to administrative claims data at the individual level should record, ideally in the same data set, demographic features - our study confirmed age and gender, but other possibilities include race, geographic information and income (Duncan, 2011); policy design covariates, controlling for as many factors affecting the claim amount or likelihood of claiming as possible; total claim amount, at least the yearly total, with more periodical subtotals (monthly or daily) allowing for the construction of more covariates that reflect the utilisation pattern of the patient.

One technique applied in regression modelling is the analysis of residuals. In our case, the graphical analysis of the residuals was not very informative, given the amount of data points available. Therefore, in order to improve the visualisation of potential patterns without losing too much information, the plots were made using a randomly selected sample of the data points. For every 50 data points we selected one to be plotted. Thus, the selected sample is 50 times smaller than the training sample. Figures 3.4, 3.5 and 3.6 show the scatter plots of the residuals versus fitted values of the multiple linear regression, two-part and frequency-severity models, respectively.

We can observe that the three graphs look similar, with the variance in residuals displaying an apparent reduction as the size of fitted values increase. Beyond that, the graphs show another limitation of the traditional methods: the maximum fitted values only go up to a fraction of the largest claim sizes, which generates very large residuals. This is not ideal, although it is expected since these models aim to fit the mean of the distribution. For this reason, the residuals appear to have only positive values in the graphs. A good fit would produce a balanced number of positive and negative residuals.

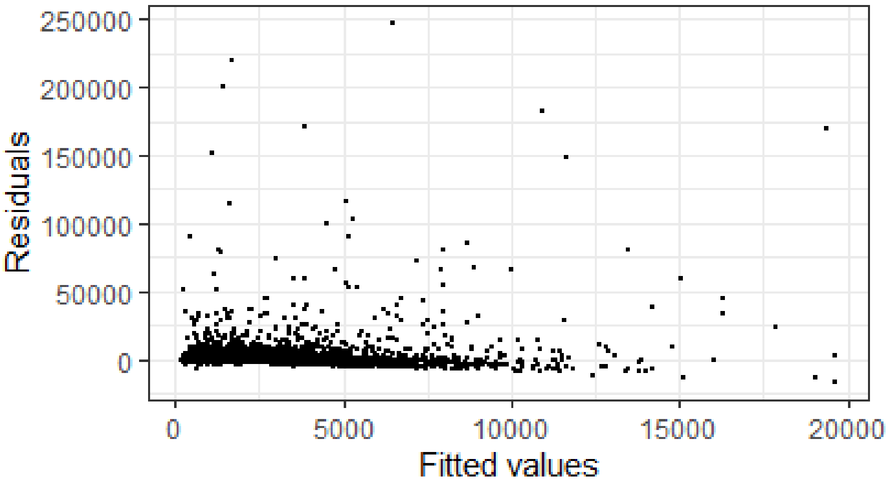


Figure 3.4: Scatter plot of residuals vs. fitted values of the multiple linear regression model.

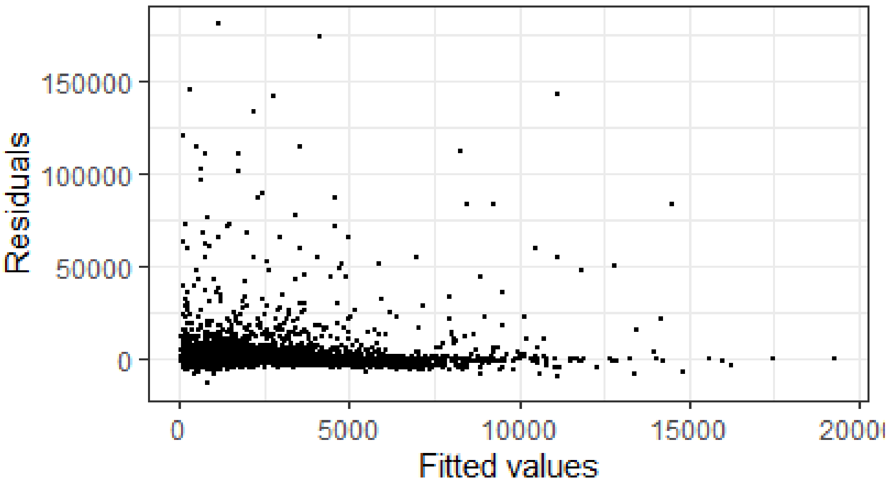


Figure 3.5: Scatter plot of residuals vs. fitted values of the two-part model.

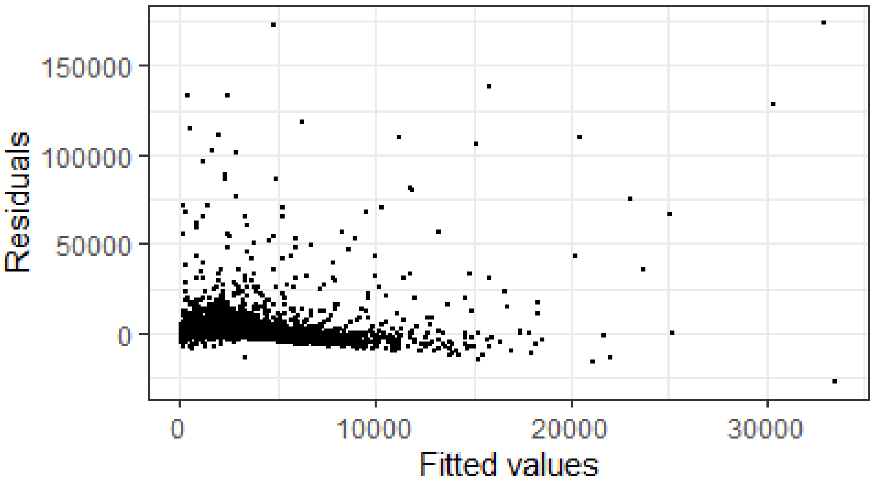


Figure 3.6: Scatter plot of residuals vs. fitted values of the frequency-severity model.

By plotting the residuals against covariates we could identify that some patterns still remain in the residuals, which is not ideal. For instance, the scatter plot of the residuals of the multiple linear model against logCOST in Figure 3.7 show that the variability in residuals increase with size of logCOST. In an ideal situation, there would be no trend, as it would be explained by the model itself. The same pattern is observed for the two-part (Figure 3.8) and frequency-severity (Figure 3.9). In these three graphs, we can also notice a concentration of points for logCOST equal to zero. These points are detached from the main cloud of points, indicating that the models need some extra calibration to consider individuals who made no claims in the observation year.

The plot of residuals against AGE display a less notable pattern, as shown in Figure 3.10 for the multiple linear regression, Figure 3.11 for the two-part model and Figure 3.12 for the frequency-severity model. In all three graphs, the sign of the residuals seem to vary more as age increases, but not as much as observed for logCOST. Also, for all ages, the majority of the residuals are scattered around zero and the points that stand out are very large residuals that are not well fitted. These large residuals, however, do not seem to be concentrated in specific age groups.

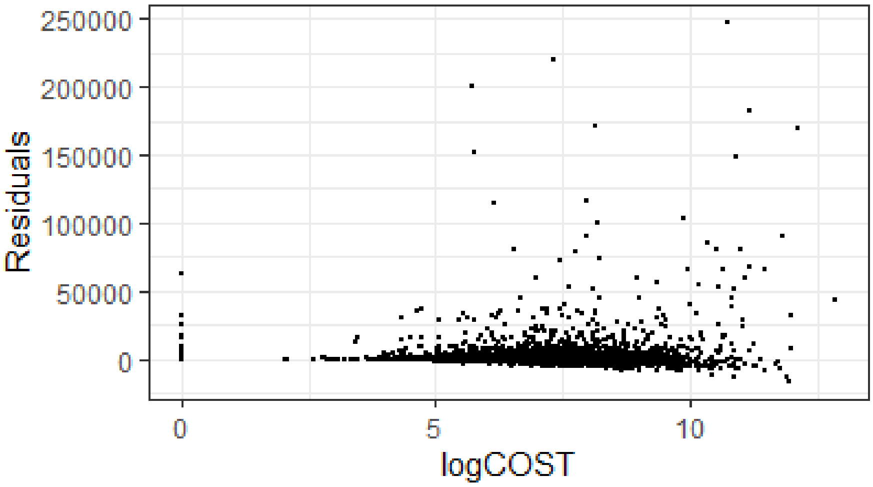


Figure 3.7: Scatter plot of residuals vs. logCOST of the multiple linear regression model.

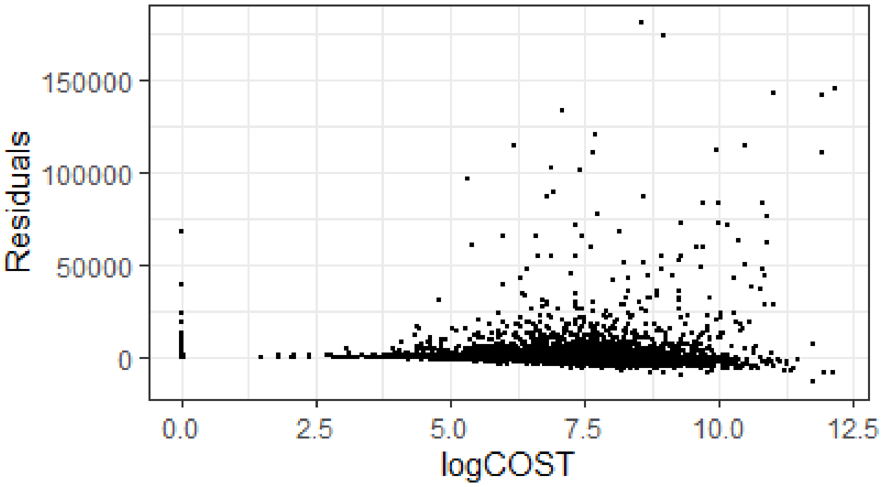


Figure 3.8: Scatter plot of residuals vs. logCOST of the two-part model.

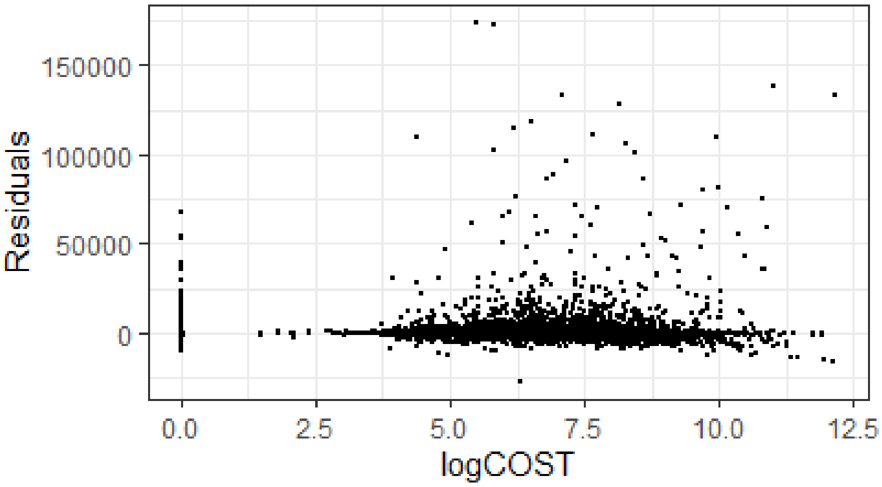


Figure 3.9: Scatter plot of residuals vs. logCOST of the frequency-severity model.

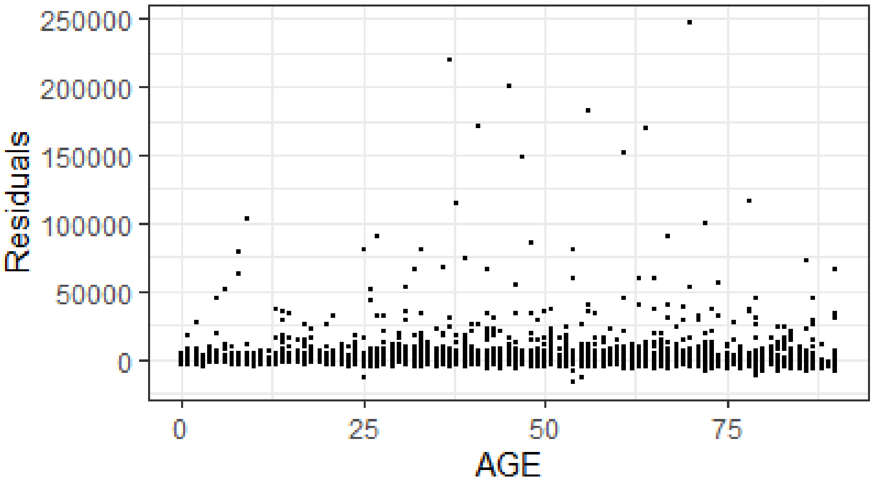


Figure 3.10: Scatter plot of residuals vs. AGE of the multiple linear regression model.

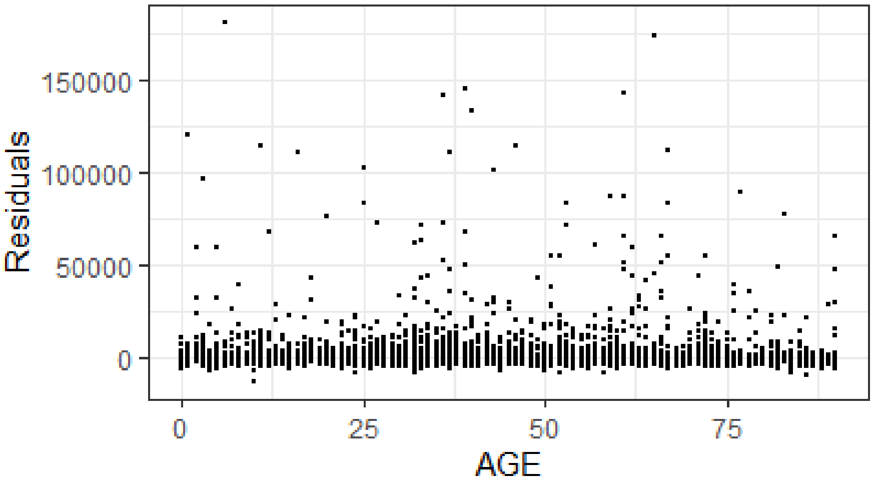


Figure 3.11: Scatter plot of residuals vs. AGE of the two-part model.

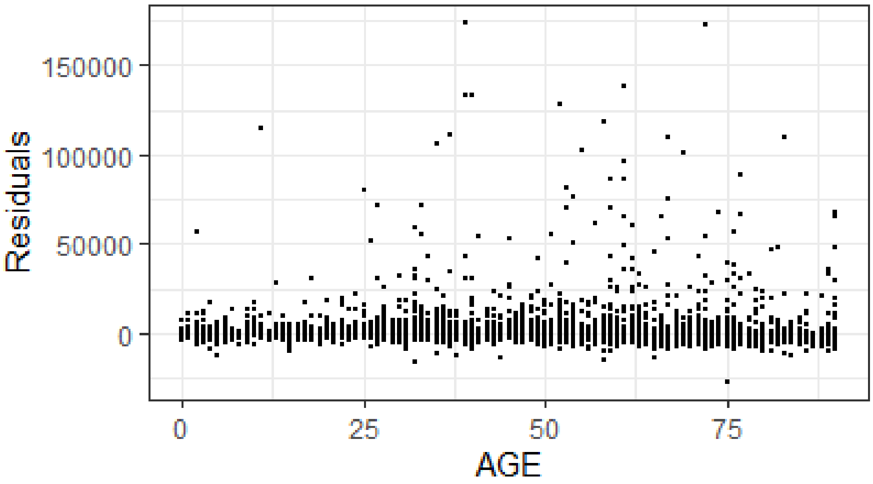


Figure 3.12: Scatter plot of residuals vs. AGE of the frequency-severity model.

In the next chapter, we will describe more recently developed methods that still use ideas from linear models, but incorporate enhanced estimation processes in order to automatically target variable selection and collinearity, among others, with the aim of providing better predictions.

Chapter 4

Statistical Learning Methods for Medical Cost Prediction

4.1 Introduction

In the previous chapter, we have investigated the performance of traditional regression models for the prediction of medical costs of insured individuals. Such models are very popular and have been extensively explored by actuaries in many fields, including health care. On the other hand, the application of more recently developed statistical learning methods on the prediction of medical costs remains limited.

According to James et al. (2013), statistical learning methods can be broadly defined as a series of approaches applied to a data set (or several data sets) with the aim of estimating the *systematic information* that predictors provide about the response variable. This definition assumes that there is a function of unknown form that connects the covariates to the dependent variable. This function represents the systematic information that the statistical learning methods attempt to estimate.

In essence, traditional regression models are considered a statistical learning method. However, alternative statistical learning methods seem to be more promising regarding the task of improving prediction accuracy. The superiority of such methods over traditional regression models has been observed in recent studies. Duncan et al. (2016), Morid et al. (2017) and Yang et al. (2017) directly compared linear regression models with other statistical learning methods for medical costs prediction. In all these studies, linear regression models were outperformed by other statistical learning methods in terms of predictive accuracy.

The increase in prediction accuracy usually comes with a price: the black-box nature of many methods focused on prediction rather than interpretation (Duncan et al., 2016; Yang et al., 2017) makes it harder to measure the impact of a covariate on the future medical costs. In this chapter, we apply certain statistical learning methods to the prediction of next year‘s medical costs. The models that we chose are more transparent in the sense that the contribution of each covariate to the predicted future cost can be directly analysed through their estimated coefficients, just as in traditional regression models.

Specifically, we explore three different methods: regularisation, model trees and rule-based models. We provide a description of these methods and explain how they were fitted to our data set. Given the relevance of model interpretation in our situation, we used cross-validation in order to produce models with reduced complexity without compromising prediction accuracy.

4.2 Regularisation Methods

When fitting the traditional regression models we relied upon the stepwise method preceded by a collinearity analysis. The stepwise method is a discrete variable selection method: each covariate either belongs or does not belong to the final model. According to Friedman et al. (2001) and James et al. (2013), the resulting coefficients, usually estimated by OLS in the case of linear regression, are subject to large variability. In other words, the sizes of the coefficients may vary significantly if the data set used to estimate them changes. Also, the outcome of the stepwise method might also change depending on the data points used, resulting in different covariates chosen for the final model.

As discussed in the previous chapter, collinearity might inflate the coefficients of the linear regression models, resulting in a model with higher prediction error. These issues motivated the creation of an estimation process called *regularisation.* This method introduces a penalisation term that constrains the size of the coefficients during the estimation process. The larger the constraint is, the more the coefficients are shrunk towards zero.

The regularised coefficients are less sensitive to changes in data sets and the inflation caused by collinearity can be controlled by the size of the constraint. By restricting the size of the coefficient estimates, the range of possible values for each coefficient becomes narrower. Consequently, there is less uncertainty surrounding the coefficient estimate and prediction accuracy is expected to be improved.

The regularisation methods differ mainly by the type of constraint function used. More details are provided in the subsections below, with a focus on two regularisation methods within the context of linear models: *ridge regression* and *lasso.*

4.2.1 Ridge Regression

Coefficient estimates of multiple linear regression models are the values that minimise the residual sum of squares, via the Ordinary Least Squares estimation process. Hoerl and Kennard (1970) proposed a penalised sum of squares as a function to be minimised when estimating regularised coefficients, which is shown by equation (4.1) below:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{k}\beta_j x_{ij}\right)^2+\lambda_R\sum_{j=1}^{k}\beta_j^2 \tag{4.1}$$

Where the penalisation is made by the term $\lambda_R \sum_{j=1}^{k}\beta_j^2$, with $\lambda_R \geq 0$. As we can see, the intercept β_0 is not included in the set of coefficients penalised. The tuning parameter λ_R (also known as shrinkage factor) controls the impact of the penalty applied to the estimation of the ridge regression coefficients. When $\lambda_R = 0$, no penalty is applied to the sum in eq. 4.1, and the resulting estimated coefficients are the same as the ones estimated by OLS. As λ_R increases, the penalty applied in the estimation of the coefficients increases, forcing them to reduce towards zero. Consequently, each value of λ_R produces a different set of coefficient estimates. Thus, fitting a ridge regression model relies upon finding an appropriate value for λ_R. This value is usually the one that minimises an error measure, which will be discussed later in this chapter.

Because the penalty term is based on the squared sum of the model coefficients, the scale of the covariates have a large influence on the estimation of ridge coefficients (James et al., 2013). This is not the case for multiple linear regression models, since $x_{ij}\hat{\beta}_j$ remains the same after any linear transformation of the covariates. In order to avoid the scale effect on the estimation of ridge regression coefficients, the covariates are standardised before fitting the model. The value of each observation is divided by the standard deviation of the covariate to which they belong, as shown by the following formula:

$$\tilde{x}_{ij}=\frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2}} \tag{4.2}$$

Although the reduction of the coefficients help to increase prediction accuracy, ridge regression has a drawback. Due to the form of its penalty function $\lambda_R \sum_{j=1}^{k}\beta_j^2$, all coefficients are reduced towards zero, but none of them actually reach zero (except when $\lambda_R = \infty$). This means that the penalty does not result in a reduction in the number of covariates used, leaving all of them in the model, even when the coefficient associated with a covariate is very small.

This can be better explained if we consider that the set of coefficient estimates of the ridge regression $\hat{\beta}_0^r, \hat{\beta}_1^r, ..., \hat{\beta}_k^r$ is the set of values that minimise the sum of squares in equation 4.3 below, subject to the constraint $\sum_{j=1}^{k} \beta_j^2 \leq t$:

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \text{, subject to } \sum_{j=1}^{k} \beta_j^2 \leq t \tag{4.3}$$

Where $t \geq 0$ is a parameter that is specified by the user of the model. There is a one-to-one correspondence between the shrinkage parameter λ_R in equation (4.1) and the constraint t in equation (4.3). In other words, for every possible value of λ_R, there is some t such that equations (4.1) and (4.3) provide the same ridge regression coefficient estimates. It is straight-forward to notice that when $\lambda_R = 0$, t is a value that is not less than the sum of the squared coefficient estimates generated by the Ordinary Least Squares method. As t tends to zero, λ_R tends to infinity, which means that the solution to both equations (4.1) and (4.3) is a set of coefficient estimates $\hat{\beta}_0^r, \hat{\beta}_1^r, ..., \hat{\beta}_k^r$ with values that are very close to zero.

Now imagine that we want to fit a ridge regression to a data set using only two covariates and we want to estimate their coefficients β_1 and β_2. Thus, all possible values for the ridge regression estimates lie within the circle defined by $\beta_1^2 + \beta_2^2 \leq t$. This is represented by the grey circle centered at the origin and with radius equal to t in Figure 4.1. Observe that there are ellipses above the circle. The centre point of each ellipse represents the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ estimated by minimising the residual sum of squares (RSS), via OLS. By applying a penalty in the estimation of the coefficients, we force RSS to increase. As the RSS increases, the estimated coefficients move away from those that minimise RSS. The solutions to the inflated RSS have the shape of ellipses. In other words, all the points that form each of the ellipses surrounding the center point provide the same value of RSS.

The value of RSS increases until it touches the constraint area. The solution provided by ridge regression is the set of coefficients defined by the point where the ellipse is tangent to the constraint area. As we can see, due to the circled shape of the constraint area centered at $(0, 0)$, with no corners or sharp points, the solution for β_1 or β_2 cannot happen in an axis, which means that they cannot be zero.

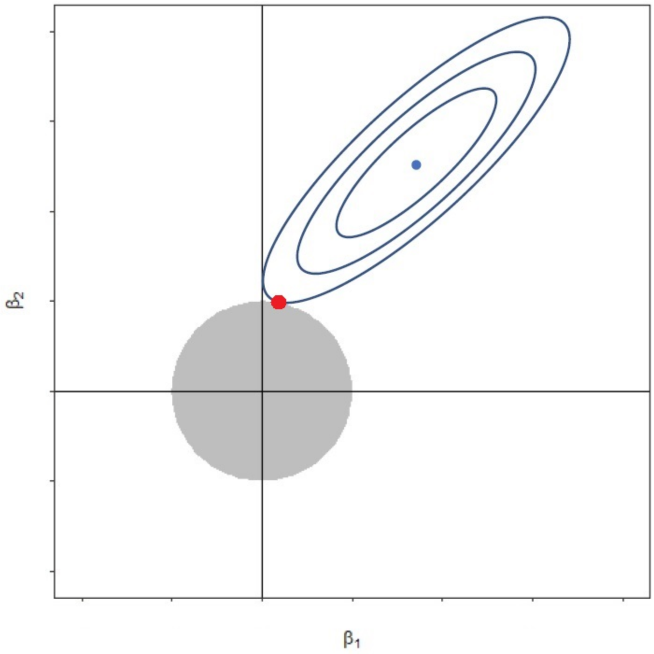


Figure 4.1: Ridge regression constraint function. The red dot represents the location of the ridge estimates. Source: James et al. (2013).

This means that ridge regression does not make variable selection and hence model complexity can be affected when the number of covariates available in the data set increase. This encouraged Tibshirani (1996) to propose an alternative penalty function, giving rise to the lasso estimation method, which will be discussed in the following sub-section.

4.2.2 Lasso

Lasso stands for "least absolute shrinkage and selection operator". Just like ridge regression, lasso also consists of applying a penalty to RSS. However, the penalty term used by lasso is slightly different, as shown by equation (4.4):

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 + \lambda_L \sum_{j=1}^{k}|\beta_j| \tag{4.4}$$

Despite the similarity between ridge regression and lasso regarding the penalty function used, an important aspect makes lasso a distinct coefficient estimation method. This is because the constraint function $\left(\sum_{j=1}^{k}|\beta_j|\right)$ allows some coefficients to be reduced exactly to zero depending on the size of the tuning parameter λ_L. As noted above, this is not true for ridge regression, which shrinks the coefficients towards zero, but not to zero exactly.

As explained by Hastie et al. (2015), the function $\sum_{j=1}^{k}|\beta_j|$ (also referred to as the ℓ_1 norm of β) has the particular property of producing a set of coefficient estimates that contains only a few non-zero values. This does not happen with other $\ell_q, q > 1$. That is the case of ridge regression, which has as the penalty function $\sum_{j=1}^{k}\beta_j^2$, which is the ℓ_2 norm of β. Hence, ridge regression is unable to produce coefficient estimates that are equal to zero.

If we rewrite equation (4.4) into the equivalent equation (4.5) below, we can understand the constraint area formed by lasso in a better way:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k}\beta_j x_{ij}\right)^2 \text{, subject to } \sum_{j=1}^{k}|\beta_j| \leq t \tag{4.5}$$

Hastie et al. (2015) explain that, due to the Lagrangian duality, equations (4.4) and (4.5) are equivalent, with a one-to-one correspondence between t and λ_L. This is also the reason why the ridge regression equations (4.1) and (4.3) are equivalent.

Consider that we want to find the coefficient estimates $\hat{\beta}_0^l, \hat{\beta}_1^l, ..., \hat{\beta}_k^l$ that minimise the residual sum of squares in equation 4.3 subject to the constraint $\sum_{j=1}^{k}|\beta_j| \leq t$. For a situation where only two covariates are used, the constraint function becomes $|\beta_1| + |\beta_2| \leq t$. The area formed by the constraint function of the lasso has a diamond shape, as shown in Figure 4.2. Thus, if the RSS ellipse touches one of the corners of the lasso constraint area, one of the coefficients will equal to zero. In the example shown, $\beta_1 = 0$ for a constraint of size t.

For this reason, lasso is able to perform variable selection and regularisation of coefficients at the same time, which does not happen for ridge regression.

We now need a method to select appropriate values for the estimates of the tuning parameters λ_R and λ_L for ridge regression and lasso, respectively. In the following subsections we describe how we used cross-validation for this purpose.

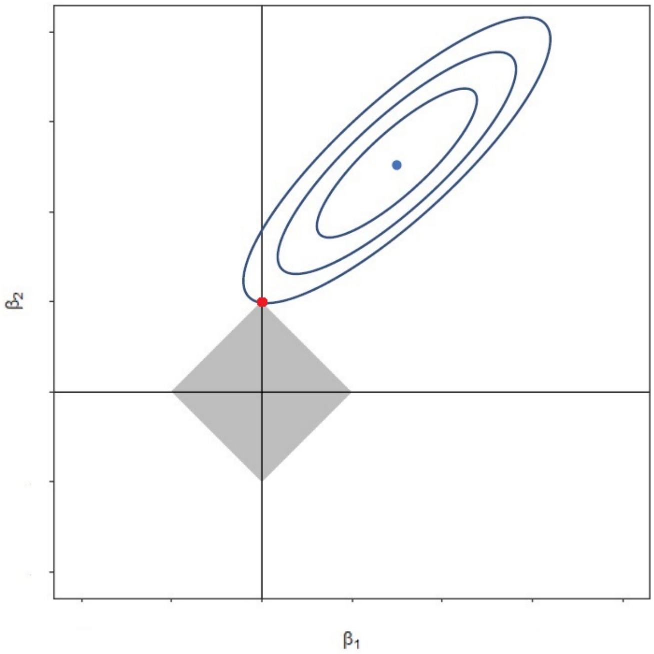


Figure 4.2: Lasso constraint function. The red dot represents the location of the lasso estimates. Source: James et al. (2013).

4.2.3 The K-fold Cross-Validation Method

When fitting the traditional regression models, we allocated 75% of the whole data set for model fitting (training sample) and used the remaining portion (test sample) of the data points for assessing the prediction accuracy of the models in an unseen data set. We judged that this approach was reasonable for fitting the traditional regression models because the estimation of a tuning parameter is not necessary in these methods. In the case of ridge regression and lasso, the estimated coefficients depend directly on the size of the shrinkage factors, which are λ_R for ridge regression and λ_L for lasso. These are the tuning parameters for these models. Other statistical learning methods, such as Cubist, which will be presented further, rely on different tuning parameters. Thus, a method for splitting the data is reasonable in order to assess the estimates of these tuning parameters.

We could follow the same approach and fit ridge regression and lasso models using different values for the tuning parameter in the training sample and choose the tuning parameter that minimises some error measure based on the same sample. However, the level of the tuning parameter varies according to the data used in its estimation. This is where this approach becomes less ideal for ridge regression and lasso. The prediction error estimated in the training sample may be highly variable, which would lead to different levels of the tuning parameter depending on which data points are chosen for the training sample and which are chosen for the test sample (James et al., 2013). Also, the error measure calculated based on the data points used to build the model tends to be lower than the error calculated using unseen data points. Since we are interested in finding models that produce better predictions, it makes more sense to choose a tuning parameter that minimises the error based on a sample other than the training sample.

In an attempt to overcome this issue on the estimation of the tuning parameter, we use cross-validation, a general re-sampling method that can be applied to the fitting of most statistical learning methods. There are different ways of conducting cross-validation (Friedman et al., 2001; James et al., 2013), but our focus will be on the k-fold cross-validation, introduced by Stone (1974).

This method involves randomly splitting the training data set into k different samples, also called folds. In order to find the optimal value of the model's tuning parameter, several potential values are tried. For each potential value, a model is fitted using $k-1$ folds and the prediction error of that particular model is verified in the remaining k-th fold.

The average of the error values across the k folds represents the cross-validation error of each tuning parameter. We can also calculate the estimated standard error related to each tuning parameter, which will give us an idea of how variable the prediction error is for each value of tuning parameter.

The error function used in the cross-validation is the root-mean squared error (RMSE) defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.6}$$

The RMSE has a direct connection to the residual sum of squares, used to estimate the ridge regression (equation 4.1) and lasso (equation 4.4) coefficients. Also, our goodness-of-fit and prediction accuracy measures are based, to some extent, on the difference between observed and fitted values, just as RMSE is. This can be observed in the formulae for the coefficient of determination (eq. 3.21), mean absolute error (eq. 3.22), mean absolute proportional error (eq. 3.23) and quantile-mean absolute deviation (eq. 3.24). For these reasons, we judge that RMSE is a reasonable error function to be used for tuning the models.

The number of folds used is rather subjective, but we followed the number recommended by James et al. (2013) and split the data into $k = 10$ folds. This provides a good balance between speed and accuracy in the estimation of the cross-validation error. Thus, the cross-validation error related to a determined value of tuning parameter λ is defined by:

$$CV_{(\lambda)} = \frac{1}{10}\sum_{i=1}^{10}\text{RMSE}_i^{(\lambda)} \tag{4.7}$$

It is important to note that the folds are created from the same training sample (75% of the original data set) used to fit the traditional models in chapter 3. Consequently, the test sample used to assess and compare the prediction accuracy of ridge and lasso models also contains the same data points as the test sample used for the traditional methods. This was done to allow for more consistency when comparing the performance of different models, including the traditional methods.

To illustrate this process, consider Figure 4.3 below. It shows that the test sample is separated from the training sample before creating the folds used for cross-validation. We split the data into 10 folds of the same size: $F_{(1)}, F_{(2)}, ..., F_{(10)}$. For a particular value of the tuning parameter, we fit a model using the folds $F_{(2)}$ to $F_{(10)}$. Then we compute the RMSE according (eq. 4.6) of the resulting model using fold $F_{(1)}$ (highlighted in black in the second row in Figure 4.3). We fit the model again for the same value of tuning parameter, but this time using the data points from folds $F_{(1)}, F_{(3)}, ..., F_{(10)}$, and compute RMSE using the data points in $F_{(2)}$. We follow this approach until we have ten different RMSE measures for a particular tuning value. We repeat this process for each potential value of tuning parameter which means that each candidate for optimal tuning parameter has ten RMSE measures related to them at the end of the process.

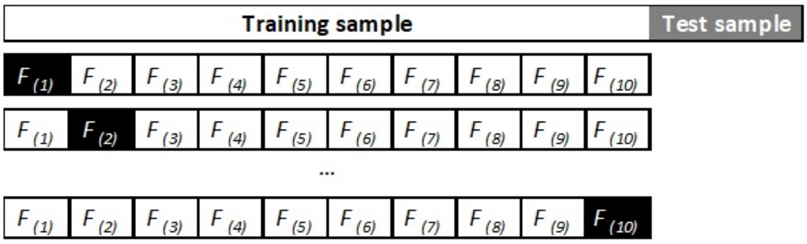


Figure 4.3: Data split in 10-fold cross-validation.

With those ten measures, we can calculate the average and standard deviation of the RMSE related to each value of tuning parameter. We use that to define the optimal value for tuning ridge and lasso according to some criteria, which we will define in the following sections. After choosing the optimal value of the tuning parameter for ridge and lasso, we use it to fit the model using all ten folds (the whole training sample) in order to find the estimated coefficients. Afterwards, the model is applied to the test sample, which we will use to compare the prediction accuracy among ridge and lasso models.

4.2.4 Fitting Ridge Regression and Lasso Models

Because ridge regression and lasso are built within the linear context, we used the log transformation of the yearly cost in the prediction year as our response variable, just as we did for the multiple linear regression model. In these models, we allowed for zero cost individuals to be part of the analysis, so we added the constant 1 to their yearly cost, which makes their log-cost equal to zero.

We used the same 13 covariates, except for FREQ_CLAIMS, that were used to fit the traditional methods in the previous chapter. The covariate representing the frequency of claims was not included because it was built for the frequency-severity model, in order to assess whether the frequency of claims of the observation year is a good predictor of the frequency of claims in the following year. Because ridge and lasso only accept numerical covariates, we transformed the categorical explanatory variables into numerical ones before fitting the models. The description of the covariates inputted can be found in Table 4.1 below.

The goodness-of-fit and prediction accuracy were assessed based on the re-transformed fitted values, i.e., we exponentiated the fitted values and subtracted 1 and compared this result to the observed cost in BRL of each individual. The measures are the same as those used for the traditional methods and their descriptions are found in pages 29 and 30.

Table 4.1: Description of covariates used in ridge and lasso models

Covariates	Description
AGE	Age of the policyholder calculated on 1st January 2016 (observation year).
GENDER	Numerical variable representing the gender of the policyholder: male (0) and female (1).
OWNER	Numerical variable that indicates whether the individual is the owner (0) of the contract or a dependant (1) of the owner.
PLAN	Numerical variable indicating the type of plan purchased by the policyholder: standard (0); restricted(1); unregulated(2).
CONTR	Numerical variable indicating the type of the contract purchased by the policyholder: employer-based group (0); individual(1); group by association (2).
HOSP_ACCOMM	Numerical variable representing the type of accommodation covered by the plan in case of hospitalisation: ward (0) or private (1).
COPAY	Numerical variable that indicates whether the plan has co-payment term (1) or not (0).
PRE_EXISTING	Numerical variable that indicates if the individual declared a pre-existing condition (1) or not (0).
logCOST	Natural logarithm of the total yearly medical claim cost of a policyholder in the observation year. A value of 1 was added before taking the log, allowing the inclusion of individuals whose yearly cost was zero.
logCOST_DEC	Natural logarithm of the total medical cost of a policyholder in December 2016. A value of 1 was added before taking the log, allowing the inclusion of individuals whose cost in this month was zero.
MONTHS_AVG	Number of months that have claim costs above the policyholder‘s average monthly cost in the observation year.
propMAX_COST	Proportion of the maximum monthly cost over the policyholder‘s total claim cost in the observation year.
propLAST_THREE	Proportion of the policyholder‘s cost in the last three months of the observation year over the total yearly cost.

4.2.5 Ridge Regression Results

We tested 100 different candidates for the tuning parameter λ_R ranging from 0.00001 to 10,000. The range is wide enough to cover values for the tuning parameter that are very close to zero, which means a very small penalty is applied, to values that are very large relative to the sizes of data points used in our models. Thus, fitting ridge regression can be thought of as an alternative approach to the 1-stage multiple regression model whose results were presented in section 3.9.1. On that occasion, the final model contained the coefficients which were estimated by the OLS method and the covariates were selected by a stepwise method. In ridge regression, the estimation of coefficients is affected by the size of the shrinkage factor and the covariates in the model are determined in a more continuous way than the stepwise method.

The cross-validation error based on the RMSE for each potential value of λ_R can be seen in Figure 4.4. The red dots represent the 10-fold average of RMSE for the respective value of tuning parameter. The black lines around each red dot represent the one-standard error interval for each tuning parameter. As we can see, the standard errors are very low, which means that there is not much variation in the fitted values for a given λ_R.

We can see that the RMSE is lower, and remains low for very small values of λ_R, increasing slowly as the penalty term increases. The vertical dashed line on the left of the graph in Figure 4.4 is at $\log(\lambda_R^{min}) = -11.04776$ or $\lambda_R^{min} = 0.00002$, which minimizes the cross-validation RMSE for ridge

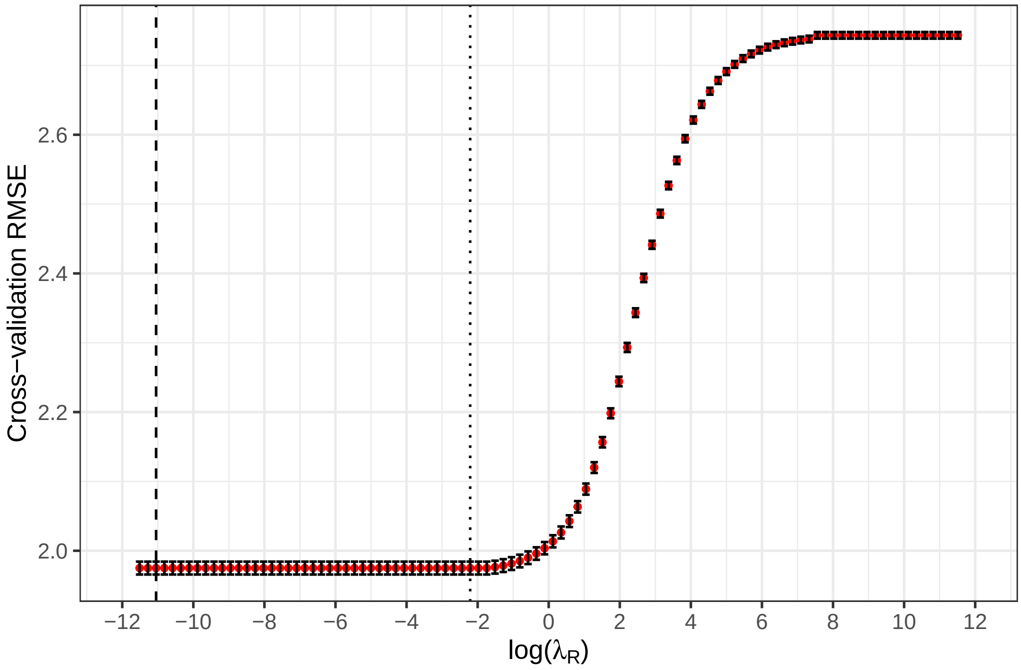


Figure 4.4: Cross-validation RMSE for different values of tuning parameter for ridge regression.

regression. The value is very close to zero, which indicates that the coefficients did not suffer much reduction compared to the OLS estimates in a multiple linear regression model.

We must keep in mind that our aim is to find a model that balances prediction accuracy and model complexity. Selecting the tuning parameter that minimises the cross-validation error may not be the most appropriate approach for that goal. Thus, we can use an alternative method for choosing the optimal tuning parameter that produces a simpler model without compromising accuracy. The *one-standard error criterion* (Kuhn and Johnson, 2013) allows for such balance. According to this criterion, the optimal tuning parameter is the one that produces the simplest model whose cross-validation error is not larger than one standard error of the minimum.

In the case of ridge regression, which does not perform variable selection, this rule may not seem to be very useful. Also, we do not have many covariates in our models yet, as 13 covariates is still considered a small model. However, for models that perform variable selection, it can be helpful in situations where there are too many covariates available and we wish to have a less complex model (fewer covariates) without significantly reducing prediction accuracy. In Figure 4.4, the vertical dotted line positioned at $\log(\lambda_R^{se}) = -2.20955$ (or $\lambda_R^{se} = 0.10975$) indicates the tuning parameter selected according to the one-standard error criterion. It is still a value that is very close to zero, indicating that the best ridge model does not cause much reduction in the multiple linear regression coefficients for the data points used.

We fitted two ridge models according to both tuning parameters. The estimated coefficients of the ridge regression using λ_R^{min} and λ_R^{se} are displayed in Table 4.2. We can see that the coefficients of both models are very similar in size, which means that there is not much reduction from λ_R^{min} to λ_R^{se} penalties. Since ridge regression does not make variable selection, all covariates were included in the final models.

The coefficients of both models are very similar because not much shrinkage is made when increasing the tuning factor from 0.00002 to 0.10975. Based on the size and sign of the coefficients in both ridge models we can observe that, overall, they agree with the results seen in the traditional models. In terms of demographic features, the model shows that, as expected, the yearly medical claim cost increases with age and female policyholders tend to experience larger costs than male policyholders.

Regarding policy design, the coefficients reflect the influence on claim cost amount that we anticipated. Dependants have, on average, lower costs than the owners of the policies. Policyholders with unregulated plan type are expected to have lower costs than those with restricted plan type, which have lower costs than those with standard plans. This effect is similar to what we observed from the traditional methods. A further agreement with the traditional models is that insurers can expect

Table 4.2: Coefficients of ridge regression for two different values of tuning parameter: λ_R^{se} and λ_R^{min}

Covariate	λ_R^{se} = 0.10975	λ_R^{min} = 0.00002
(Intercept)	1.64205	1.58457
AGE	0.00711	0.00650
GENDER (Female=1)	0.18109	0.16265
OWNER(Dependant=1)	-0.08726	-0.08208
PLAN(Restricted=1;Unregulated=2)	-0.09733	-0.08655
CONTR(Association=1;Individual=2)	0.08888	0.08442
HOSP_ACCOMM(Private=1)	0.04099	0.03578
COPAY(Yes=1)	0.09494	0.10703
PRE_EXISTING(Yes=1)	0.19932	0.18870
logCOST	0.51623	0.58745
logCOST_DEC	0.08063	0.06410
MONTHS_AVG	0.21404	0.15918
propMAX_COST	-0.00169	-0.00507
propLAST_THREE	0.00110	0.00126

policyholders with individual contract types to spend more than those with group by association, who spend more than those who have employer-based contract type. Patients who stay in private hospital accommodations when hospitalised and those who declare pre-existing conditions are also expected to have higher costs.

The only coefficient among the covariates representing policy design that has an unexpected sign was the one for COPAY. The multiple linear regression models show that, on average, policyholders with COPAY tend to have lower medical claim amounts (see Table 3.5), although these individuals tend to claim more frequently than those without a co-payment clause (Tables 3.6 and 3.8).

In respect of covariates representing previous cost, we can see that logCOST has the largest coefficient in both models, being the most influential covariate in the model. MONTHS_AVG is also relevant with, an expected, positive influence in the future claim cost. Thus, these models also say that policyholders whose costs are spread out over the year tend to have larger costs than those whose costs are more concentrated. This is confirmed by the negative sign for propMAX_COST. Cost in the last month of the observation year (logCOST_DEC) positively affects future costs, as well as the concentration of costs in the last three months (propLAST_THREE), although the latter has a very small impact due to its small coefficient..

We can now check if there is a significant difference between the goodness-of-fit of ridge models fitted using λ_R^{min} or λ_R^{se}. Table 4.3 shows the six measures calculated based on the data points of the training sample. We also included the measures of the two-part model, which had the best goodness-of-fit and prediction accuracy results among the traditional methods.

Table 4.3: Goodness-of-fit measures of ridge models fitted using λ_R^{se} and λ_R^{min}

Goodness-of-fit measures	λ_R^{se}	λ_R^{min}	Two-part
R^2	0.13232	0.16064	0.11081
MAE	1,504	1,489	1,696
MAPE	0.75406	0.74641	0.85020
$qMAD_{0.95}$	621	617	871
Sp.Cor.	0.69471	0.69663	0.70247
Gini	-8.32949	-9.75504	7.29779

Comparing the two ridge models, we can see that the model fitted using the tuning parameter that minimises the cross-validated RMSE provides slightly better goodness-of-fit than the model tuned with the parameter based on one-standard error criterion for all the measures analysed. λ_R^{min} produces a model whose coefficient of determination is 2.83 percentage points larger than the one produced by λ_R^{se}. Its MAE and MAPE are 1.01% smaller and $qMAD_{0.95}$ 0.56%. Both have very similar Spearman correlation coefficients and Gini statistics, which are negative, indicating that they tend to underestimate the claim amounts. We can conclude, thus, that the model using λ_R^{min} is the best ridge model for our data.

If we compare the goodness-of-fit measures of this model with the ones resulting from the two-part model, we can observe a significant improvement in all of them, except for the Spearman correlation coefficient. The Gini statistic of the two-part model is positive, suggesting a small overestimation of claim amounts, which is the opposite of what is observed in the ridge models. Also, it is slightly closer to zero than the Gini statistics of the ridge models, suggesting a better fit. The selected ridge model (with λ_R^{min}) has a R^2 4.98 percentage points larger than the one from the two-part model and its MAE and MAPE are 12.21% lower (thus, better). A more impressive difference between the models is seen on their qMAD$_{0.95}$ values. This measure is 29.14% lower for ridge model in comparison with the two-part model. This suggests that the ridge model is better at fitting the costs of individuals with lower costs (bulk of the data) than the two-part model.

The same performance can be observed for the prediction accuracy of the ridge models, based on the data points in the test sample, which is shown in Table 4.4. Once again, we included the results for the two-part model for comparison purposes. Regarding the ridge models, the values of all measures calculated using the test sample are very similar to the ones calculated using the training sample, as expected.

Table 4.4: Prediction accuracy measures of ridge models fitted using λ_R^{se} and λ_R^{min}

Prediction accuracy measures	λ_R^{se}	λ_R^{min}	**Two-part**
R^2	0.12837	0.15760	0.11027
MAE	1,501	1,486	1,695
MAPE	0.75497	0.74739	0.85249
qMAD$_{0.95}$	619	615	872
Sp.Cor.	0.69329	0.69509	0.70032
Gini	-7.25026	-8.97265	7.93416

Comparing the accuracy of the ridge models in the test sample, we have that the one tuned using λ_R^{min} has a R^2 2.92 percentage points larger than the model tuned with λ_R^{se}. Also, its MAE and MAPE are 1.00% lower and its qMAD$_{0.95}$ is 0.52% lower. Spearman correlation coefficients and (negative) Gini statistics of both models are very similar. Thus, the ridge model using λ_R^{min} not only has superior goodness-of-fit, but also prediction accuracy.

Once again, the improvement in prediction accuracy observed when comparing its results to the two-part model is impressive. Comparing the ridge model tuned using λ_R^{min} to the two-part model (third and fourth columns of Table 4.4), we can see that its R^2 is 4.73 percentage points higher, its MAE and MAPE are 12.33% lower and, just as observed for goodness-of-fit, qMAD$_{0.95}$ is 29.40% lower. This result is due to the ridge model fitting the low cost individuals much better than the two-part model, as shown in Table 4.5.

Table 4.5: MAPE per cost decile of the ridge model tuned using λ_R^{min} and the two-part model, calculated based on the test sample

Cost decile	**Number of Policyholders**	**Average cost**	**Ridge (λ_R^{min})**	**Two-part**
10%	125,657	314.74	**0.91161**	2.14957
20%	30,312	1,304.74	**0.53079**	0.80236
30%	17,416	2,270.87	**0.52396**	0.54157
40%	10,476	3,775.05	0.58715	**0.45638**
50%	6,551	6,037.53	0.67848	**0.51583**
60%	4,193	9,431.60	0.75755	**0.61146**
70%	2,357	16,773.96	0.82975	**0.74629**
80%	1,221	32,404.36	0.88146	**0.85181**
90%	563	70.056.85	**0.88030**	0.90414
100%	216	183.601.87	**0.89277**	0.94534

As we can see, the improvements in fitting low cost individuals when using the ridge model are shown by the significant improvements in the first two lower cost deciles when compared to the two-part model. For cost deciles 40% to 80%, we can see better MAPE values for the two-part model, but the differences are not as large as that which we observe for the deciles 10% and 20%. Consequently, the qMAD$_{0.95}$ value of the ridge model is lower (thus, superior) than the one from the two-part model. The ridge model still provides better fitting for the individuals in the top 90% and 100% deciles, contributing to a better prediction performance overall.

To conclude, the ridge model tuned with the parameter that minimises the cross-validated RMSE offered superior goodness-of-fit and prediction accuracy than the ridge model fitted using one-standard error parameter. Also, we observe a considerable improvement in accuracy when compared to the two-part model, which provided better performance among the traditional methods. In the following section, we describe the results for the lasso models.

4.2.6 Lasso Results

For lasso, we tried the same 100 values for the tuning parameter that were tested for ridge regression. We can see in Figure 4.5 that the cross-validation RMSE stays roughly the same for very small values of λ_L, but quickly rises as the tuning parameter approaches 1. By the point at which $\log(\lambda_L) = 0.81404$ (or $\lambda_L = 2.25701$), the RMSE reaches its maximum and keeps at this level for all larger values of the tuning parameter. This is because, at this level, the tuning parameter is sufficiently high to shrink all coefficients to zero, producing a null model.

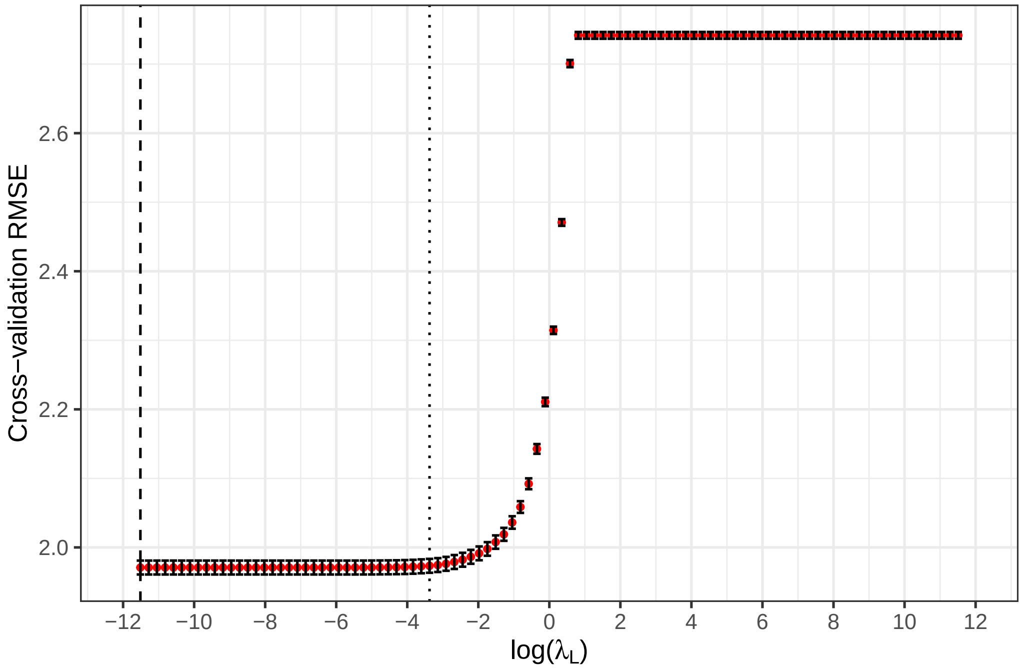


Figure 4.5: Cross-validation error for different values of tuning parameter for lasso.

The dashed vertical line on the far left of the graph in Figure 4.5 is placed on the value of the tuning parameter that minimizes the cross-validation RMSE: $\log(\lambda_L^{min}) = -11.51293$ (or $\lambda_L^{min} = 0.00001$). The dotted vertical line, located more to the centre of the graph, represents the value of the tuning parameter related to the one standard error criterion: $\log(\lambda_L^{se}) = -3.37247$ (or $\lambda_L^{se} = 0.03430$).

Since lasso performs variable selection, it is interesting to observe the way the coefficients are shrunk towards zero, as the value of the tuning parameter increases, until they eventually leave the model. This is represented by the graph in Figure 4.6. The x-axis represents the log of the tuning parameter, the y-axis represents the coefficient size (all of them are standardised to the same scale, allowing for comparison). The vertical dashed and dotted lines are, respectively, λ_L^{min} and λ_L^{se}.

We can see, by the size of the model coefficients, that logCOST is clearly the most influential variable among the covariates. It is the last one to leave the model as the penalty increases, indicating the strong influence of the individual's previous medical costs on their future costs. Its coefficient does not suffer much reduction until it drops steeply for values of λ_L close to one.

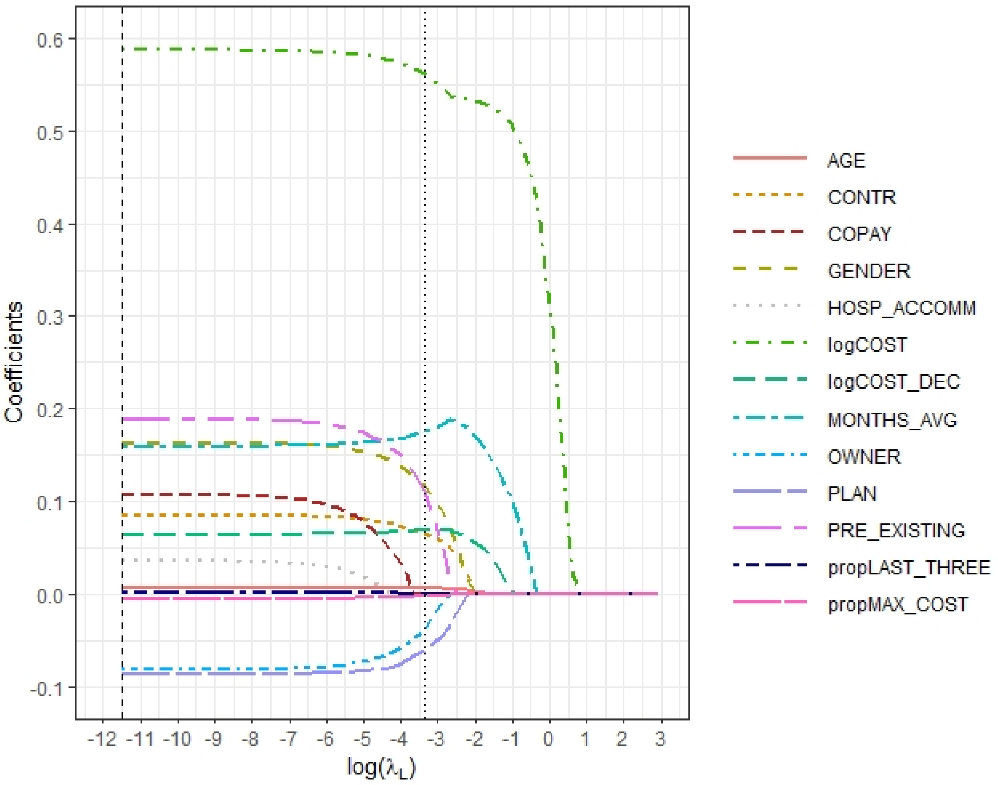


Figure 4.6: Shrinkage of lasso coefficients as the tuning parameter increases. The vertical dashed and dotted lines are, respectively, λ_L^{min} and λ_L^{se}.

The reduction in the coefficients of a few covariates allows the coefficients of others to inflate before they are also shrunk to zero. We can see this behaviour for MONTHS_AVG and logCOST_DEC.

The first coefficient to reach zero is the one from HOSP_ACCOMM (dotted grey line in the graph), followed by propLAST_THREE (dark blue stripped line), which had a very low coefficient even for very small values of λ_L. Before the value for the tuning parameter reaches the one defined by the one-standard error criterion, COPAY also leaves the model. In other words, the lasso model tuned by λ_L^{se} has three fewer covariates then the one fitted using λ_L^{min}. The last three covariates to leave the model are all based on the individual's medical cost in the observation year: logCOST_DEC, MONTHS_AVG and logCOST. This shows the relevance of the previous cost for predicting medical claim costs in relation to covariates containing policy design or demographic information.

Table 4.6 shows the coefficients of the lasso models tuned using λ_L^{min} and λ_L^{se}. As shown previously by the graph in Figure 4.6, the lasso model tuned using λ_L^{min} contains all covariates inputted, while the model tuned with λ_L^{se} does not have HOSP_ACCOMM, COPAY and propLAST_THREE.

By comparing the third columns of Tables 4.2 and 4.6, we can see that the sizes of the coefficients in the ridge and lasso models fitted using the tuning parameter that minimises the cross-validated RMSE are very similar. This happens because not much shrinkage is done. The coefficients that remain in the lasso model tuned using λ_L^{se} reveal an impact of their covariates in the response variable that goes in line with what we have observed in the models fitted so far.

In terms of goodness-of-fit, shown in Table 4.7, we can see that lasso and ridge produce very similar results. In fact, all the measures of the lasso and ridge models tuned with λ_L^{min} and λ_R^{min} (the shrinkage factors that minimise the cross-validated RMSE for lasso and ridge, respectively), have exactly the same values (and very similar Gini statistics, most probably due to approximations). This can be checked by comparing the third and fifth columns of Table 4.7.

By comparing the values of the measures of the two lasso models (second and third columns of Table 4.7), we can see that there is some deterioration in some of the goodness-of-fit measures when fitting the model using λ_L^{se}. However, this deterioration is slightly less than the one observed for ridge models. Specifically, R^2 of the lasso model that uses λ_L^{se} is 2.34 percentage points lower than the one using λ_L^{min}. For ridge, this difference is slightly greater: 2.83 percentage points. MAE and MAPE of the lasso models suffered an increase of only 0.55% when changing λ_L^{min} to λ_L^{se} while the increase for

Table 4.6: Coefficients of lasso models for two different values of tuning parameters: λ_L^{se} and λ_L^{min}

Covariate	λ_L^{se} **= 0.03430**	λ_L^{min} **= 0.00001**
(Intercept)	1.73168	1.60125
AGE	0.00591	0.00646
GENDER (Female=1)	0.11858	0.16483
OWNER(Dependant=1)	-0.03618	-0.08092
PLAN(Restricted=1;Unregulated=2)	-0.06214	-0.08775
CONTR(Association=1;Individual=2)	0.06100	0.08074
HOSP_ACCOMM(Private=1)		0.03663
COPAY(Yes=1)		0.10848
PRE_EXISTING(Yes=1)	0.10493	0.18267
logCOST	0.56185	0.58796
logCOST_DEC	0.06911	0.06436
MONTHS_AVG	0.17354	0.15752
propMAX_COST	-0.00266	-0.00525
propLAST_THREE		0.00119

Table 4.7: Goodness-of-fit measures of lasso, ridge and two-part models, based on the training sample

Goodness-of-fit measures	**Lasso** (λ_L^{se})	**Lasso** (λ_L^{min})	**Ridge** (λ_R^{se})	**Ridge** (λ_R^{min})	**Two-part**
R^2	0.13725	0.16064	0.13232	0.16064	0.11081
MAE	1,497	1,489	1,504	1,489	1,696
MAPE	0.75055	0.74641	0.75406	0.74641	0.85020
$\text{qMAD}_{0.95}$	613	617	621	617	871
Sp.Cor.	0.69444	0.69663	0.69471	0.69663	0.70247
Gini	-6.98140	-9.75450	-8.32949	-9.75504	7.29779

ridge models is 1.01%. The qMAD$_{0.95}$ and Gini statistic of the lasso model based on the one-standard error criterion are, in fact, slightly better than the ones resulting from the lasso model using λ_L^{min}.

These results suggest that, for medical costs data, the penalty function applied when estimating lasso coefficients produce more efficient results than the one applied during the estimation of ridge coefficients. Not only does it produce a simpler model, which depends on fewer covariates, but it also produces a model whose goodness-of-fit does not depart very much from the fitting achieved by the model without much shrinkage.

Similar conclusions can be drawn regarding prediction accuracy for the test sample, as shown in the Table 4.8. Not much deterioration in accuracy is seen comparing the measures of lasso models fitted using λ_L^{se} with the measure of the model fitted using λ_L^{min}. In fact, we see small improvements in terms of qMAD$_{0.95}$ and Gini statistic.

Table 4.8: Prediction accuracy measures of lasso, ridge and two-part models, based on the test sample

Prediction accuracy measures	**Lasso** (λ_L^{se})	**Lasso** (λ_L^{min})	**Ridge** (λ_R^{se})	**Ridge** (λ_R^{min})	**Two-part**
R^2	0.13229	0.15760	0.12837	0.15760	0.11027
MAE	1,494	1,486	1,501	1,486	1,695
MAPE	0.75149	0.74739	0.75497	0.74739	0.85249
qMAD$_{0.95}$	611	615	619	615	872
Sp.Cor.	0.69307	0.69509	0.69329	0.69509	0.70032
Gini	-6.18467	-8.97213	-7.52026	-8.97265	7.93416

In Table 4.9 we can see that lasso tuned according to λ_L^{se} shrinkage factor is superior to the ridge model using λ_R^{min} and the two-part model for fitting the costs of policyholders in 10% to 30% cost decile. On the other hand, the model has worse accuracy among the models shown in Table 4.9 for cost deciles 40% to 80%. Its performance for the top two deciles lies between the ridge model using λ_R^{min} and the two-part model.

Table 4.9: MAPE per cost decile of lasso (λ_L^{se}), ridge (λ_R^{min}) and two-part models, based on the test sample

Cost decile	**Number of Policyholders**	**Average cost**	**Lasso** (λ_L^{se})	**Ridge** (λ_R^{min})	**Two-part**
10%	125,657	314.74	**0.86487**	0.91161	2.14957
20%	30,312	1,304.74	**0.50522**	0.53079	0.80236
30%	17,416	2,270.87	**0.52130**	0.52396	0.54157
40%	10,476	3,775.05	0.60037	0.58715	**0.45638**
50%	6,551	6,037.53	0.69730	0.67848	**0.51583**
60%	4,193	9,431.60	0.77624	0.75755	**0.61146**
70%	2,357	16,773.96	0.84588	0.82975	**0.74629**
80%	1,221	32.404.36	0.89388	0.88146	**0.85181**
90%	563	70.056.85	0.89763	**0.88030**	0.90414
100%	216	183.601.87	0.91214	**0.89277**	0.94534

Among the regularisation models tried for prediction of medical claim costs, both ridge regression and lasso tuned according to the shrinkage factor that minimises the cross-validated RMSE provided the best goodness-of-fit and prediction accuracy. So far, we have not used many covariates to build our predictive models, but for situations where the number of covariates is increased, lasso has the advantage of seeking a simpler model, by leaving the more relevant covariates and reducing the coefficients of less important covariates to zero. This is made without a significant loss in accuracy. For this reason, we can conclude that lasso tuned according to the one standard error rule is the best choice when the number of covariates is large, which will be the case in the following chapter. For simpler data sets, both ridge and lasso provide similar results, with the advantage of improving the performance seen from traditional models.

One disadvantage of all the models tried so far (traditional and regularised linear models) is that they all attempt to fit the whole universe of observations into one single equation. This may not be the best approach when predicting medical costs at the individual level. There might be distinct groups within the whole sample of individuals to which different models would be more appropriate.

For instance, take into consideration the covariate that represents the total cost in the previous year, logCOST. We have learned from the models that this is the most influential covariate for cost prediction. Overall, the larger the total medical cost of an individual is, the greater their next year's total cost tends to be. However, the size of the impact of previous costs into future costs may vary for different groups.

This has motivated us to explore a class of models that aim to split the data into more homogeneous groups and create different predictions for each of them. They are known as *tree-based models* and could be useful in identifying important groups of individuals, such as those who are more likely to be high cost in the next year. They can also provide more insights into variable importance whilst increasing accuracy of predictions. Such models are presented in the following section.

4.3 Tree-based and Model Tree Methods

A tree-based model aims to create splits of the whole data set in order to produce more homogeneous groups. The final settlement bears resemblance to a tree: the "root" contains all the original data points and the "leaves" are the more homogeneous groups created. There are two types of leaves: the terminal leaves, which are the groups that do not generate more splits, and the intermediate leaves (or nodes), which are those linking the root to the terminal leaves. For each split, the method tries to find one covariate and one value that would partition the data into two groups. The covariate and value selected are those that result in the highest reduction of a specific error function.

The original tree-based methods are described as Classification And Regression Trees - CART (Breiman et al., 1984). For classification problems, where the response variable is categorical, the predicted category is the most frequent among the categories in the terminal leaf to which an observation belongs. For regression situations, where the response variable is continuous, the prediction could be either the mean or median of the group.

Since our investigation involves predicting medical costs, which is a continuous variable, this leads to a regression problem. If we were using the CART method, our predictions would be the mean or median of the group to which an individual belongs, which would be too simplistic (Kuhn and Johnson, 2013). Given the large variability and heterogeneity in the data, it may be hard to find enough homogeneous groups that contain individuals with similar average next year‘s cost and categorise them (Bertsimas et al., 2008). An alternative and more robust solution that combines the ideas of tree-based models and linear regression is the *model tree* (Kuhn and Johnson, 2013). One of the most popular model tree methods is called M5 and is covered in the next section.

4.3.1 The M5 Model Tree

This model was introduced by Quinlan (1992). The main feature of M5 is that the leaves (intermediate and terminal) of the tree produced have multiple linear regression models fitted to each of them, which are used to make the predictions. This is an improvement over CART, which simply uses the average of the observations in the terminal leaves as predicted values.

The idea and theory behind the M5 model tree is briefly explained in Quinlan (1992), but there are some crucial details left out by the authors, perhaps due to the commercial purposes of the algorithm at that time. Their lack of information motivated the work published by Wang and Witten (1996), in which they describe a "rational reconstruction" of the M5 model tree. In their work, whenever there is a lack of detail that is crucial for the understanding of the M5 model tree in Quinlan (1992), the authors propose calculations that would be the most logical to carry out with the method. The combination of the information from both papers has allowed us to provide the description of the M5 method below.

Tree growth: the method starts by calculating the standard deviation of the response variable of the observations in the training sample, which we call T. Thereupon, several "tests" are made with the aim to split T into two subsets, T_1 and T_2. In each test, an explanatory variable is selected and binary splits of T based on the possible values of this variable are made. Then the method proceeds to another explanatory variable and this recurrent procedure is performed over the whole space of the available covariates. For every test, the standard deviation of the response variable in each of the two resulting subsets is calculated. The idea is that the standard deviation acts as an error estimate: the greater the standard deviation, the greater is the error involved in the prediction. This provides the possibility of estimating the expected error reduction resulting from each test, which can be determined as:

$$\Delta \text{error}_j = \text{sd}\left(T_j\right) - \sum_{i=1}^{2} \frac{|T_{j,i}|}{|T_j|} \cdot \text{sd}\left(T_{j,i}\right) \tag{4.8}$$

where $\text{sd}\left(T_j\right)$ is the standard deviation of the parent leaf and $\text{sd}\left(T_{j,i}\right)$ is the standard deviation of the observations of the two child leaves of set T_j; $|T_j|$ and $|T_{j,i}|$ are the number of observations in the parent leaf and the number of observations in each of the two child leaves, respectively.

The chosen split is the one that implies the maximum error reduction as measured by equation (4.8). This method is repeated recursively until a leaf contains too few cases or until the expected error reduction is minimal, which implies that another split is unjustifiable.

It is not very clear from Quinlan (1992) how many cases would be considered "too few" or which level of expected error reduction would be regarded as "minimal". Given this limitation, Wang and Witten (1996) have suggested that a leaf would not be split if the standard deviation of the response variable of the observations belonging to that leaf is less than 5% of the standard deviation of the whole initial set (the root node of the tree). They also have observed that the results do not vary much by choosing different thresholds. The explanation for defining the number of cases in a leaf that would be too few will be given in the next steps.

Fitting linear regression models: After growing the tree, the method starts fitting multiple linear regression models in the intermediate nodes using the Ordinary Least Squares estimation method. The only candidates of explanatory variables used in a model are those chosen by the splitting tests or linear regression models in the sub-tree starting from that particular node. For this reason, the terminal leaves of the initial tree do not have linear regression models attached to them, as clarified by Wang and Witten (1996). Instead, only a constant (the median of the observations in that particular leaf) is used as the predicted value.

Simplifying the models: after fitting the linear regression models in each internal node of the tree, the algorithm estimates an accuracy error of each node by calculating the average of the absolute residual. However, this is likely to underestimate the accuracy of the model if it was applied in a data set different than the one used to build it. For this reason, the algorithm multiplies the mean absolute residual by a correction factor, as described by the formula below:

$$\text{Corrected Mean Absolute Residual} = \frac{(n+v)}{(n-v)} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \right) \tag{4.9}$$

where n is the number of observations in a particular node and v is the total number of parameters of the linear model in that node. This multiplicative factor increases the estimated error as a penalty for the excessive number of parameters in the linear model. In each node, parameters are dropped from the corresponding linear model as long as the estimated accuracy error of that node reduces. The exclusion of parameters may increase the average absolute residual, but it also reduces the multiplicative factor, potentially making the overall error estimate decrease.

It may be the case that all parameters are dropped and only a constant (the median of the group) remains in that node. Wang and Witten (1996) use the multiplicative factor $\frac{(n+v)}{(n-v)}$ to define the minimum number of observations in a node that allows for another split. They state that, if $n = v$, the multiplicative factor is infinite (since this makes the denominator equal to zero). To avoid this issue, there should be at least two observations in a node resulting from a split. If there is only one observation in a node, then $n = v = 1$ (the model in that node will be only a constant, which makes $v = 1$) and consequently the multiplicative factor is infinite. Thus, if there are three observations in a node, no further splits are made.

Pruning: after all the models are simplified, the tree is pruned starting from the nodes that generate the terminal leaves. The algorithm checks what produces a lower estimated accuracy error measured by equation (4.9): the linear regression model in that node or the sub-tree starting from that node. Wang and Witten (1996) have worked out that the estimated accuracy error of the sub-tree is the weighted sum of the estimated accuracy errors of the two child nodes, where the weights are the proportion of observations in each child node. If the linear model provides a lower error, the sub-tree is pruned back to that node which now becomes a terminal leaf, with the related linear regression. This is done recursively from the bottom to the root of the tree.

To illustrate this step, consider the diagram displayed in Figure 4.7. On the left-hand side the estimated accuracy error of T_3 is compared to the weighted sum of estimated accuracy errors of T_5 and T_6. Suppose that there are 100 observations in T_3, which are split into 40 observations in T_5 and 60 observations in T_6. Then the accuracy error of T_5 is multiplied by 40% and the one of T_6 by 60% and added together. If this sum is larger than the accuracy error of T_3, then T_5 and T_6 are eliminated from the tree and T_3 becomes a terminal leaf. The same is done for T_1 and its child nodes T_3 and T_4.

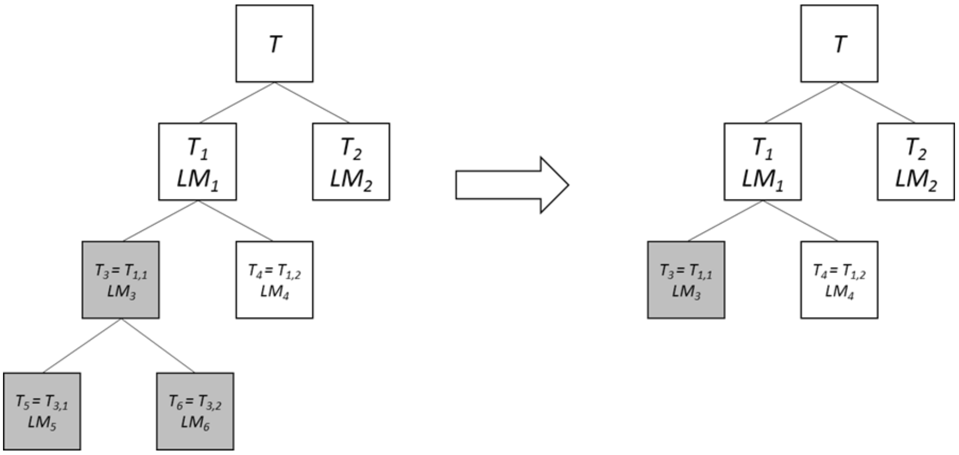


Figure 4.7: Pruning example. The tree on the left-hand side represents the tree after all the splits were made and the linear models simplified. The tree on the right-hand side represents the pruned tree.

Smoothing: after pruning the tree, the model is ready to be used for making predictions. Imagine that we want to estimate the response value of an observation in a new data set. The model identifies the terminal leaf to which this observation belongs and uses the linear model related to that leaf to estimate the response value. After doing this, the algorithm smooths out the predicted value by carrying out a weighted average of that prediction and the estimated value of the previous node. This is done for all nodes along the path from the terminal leaf to which the observation was allocated back to the root of the tree. Assume that we want to smooth out the predicted value in leaf $T_{j,i}$, backed up to its parent leaf T_j. This is made by the formula in equation (4.10).

$$\mathrm{SV}\left(T_j\right) = \frac{n_{j,i} \cdot \mathrm{PV}\left(T_{j,i}\right) + k \cdot \mathrm{PV}\left(T_j\right)}{n_{j,i} + k} \tag{4.10}$$

Where $\mathrm{SV}\left(T_j\right)$ is the smoothed predicted value in the parent node (T_j), $n_{j,i}$ is the number of observations in $(T_{j,i})$; $\mathrm{PV}\left(T_{j,i}\right)$ is the predicted value in the child leaf $(T_{j,i})$; $\mathrm{PV}\left(T_j\right)$ is the predicted value in the parent node (T_j); and k is a smoothing constant, with default value of 15 (Quinlan, 1992). The smoothing process has a higher impact in two situations. The first is when the number of observations in a terminal leaf is not very large. We can see in equation 4.10 that the smoothed value is a weighted average and as $n_{j,i}$ gets smaller, relatively less weight is placed on the predicted value of the terminal leaf. Consequently, the predicted value of the parent node has relatively more weight. The second situation is when the model in the terminal leaf is very different from the models in the intermediate nodes that lead to that leaf. In that case, for the same individual, the predicted values in the terminal and parent leaves may be too different, given that the predictive models relative to that individual are also very different. Thus, smoothing finds a middle value between these two predictions. In fact, smoothing is an optional step, which means that it can either be performed or not be performed at all when tuning the model.

4.3.2 Fitting The M5 Model Tree

In order to fit M5 Model Tree to our medical costs data, we have used the package RWeka (Hornik et al., 2009) in R. We can tune M5 model by adjusting three *hyper-parameters* or tuning parameters. One of them controls whether the predictions will be smoothed out or not. Another hyper-parameter allows the user to decide whether the tree will be pruned or not. Finally, we can choose the minimum number of observations in a leaf, which works as a trade-off: the lower the value of this hyper-parameter is, the larger will be the resulting tree. Thus, the last two hyper-parameters are directly related to the complexity of the resulting M5 model.

The covariates inputted into the model are the same as the ones described in Table 4.1. Because we want more accurate predictions, it makes sense to tune the model to provide smoothed predictions. Also, because we aim to produce a simpler model, allowing for easier interpretation, it is more reasonable to decide to prune the tree. Finally, we would have to decide on the minimum number of observations in a leaf. For that task, once again, we used a 10-fold cross-validation using the Root Mean Squared Error (equation 4.6) as the error function. The process is the same as the one explained to tune the penalty factor for ridge regression and lasso.

Figure 4.8 shows the results of the cross-validation performed for the M5 model. We tested 50 potential candidates for the minimum number of observations per leaf. The values ranged from four (the default in RWeka package) to 5,000. The model that generated the lower cross-validation error was the one having a minimum of 104 observations per leaf. Increasing the minimum number of observations per leaf would cause not only the average RMSE to increase, but also the estimated standard error of the predictions. For values larger than 2,700, the cross-validation RMSE is practically stable, with a large standard error around the predictions, which made it sensible to stop testing larger values for this hyper-parameter.

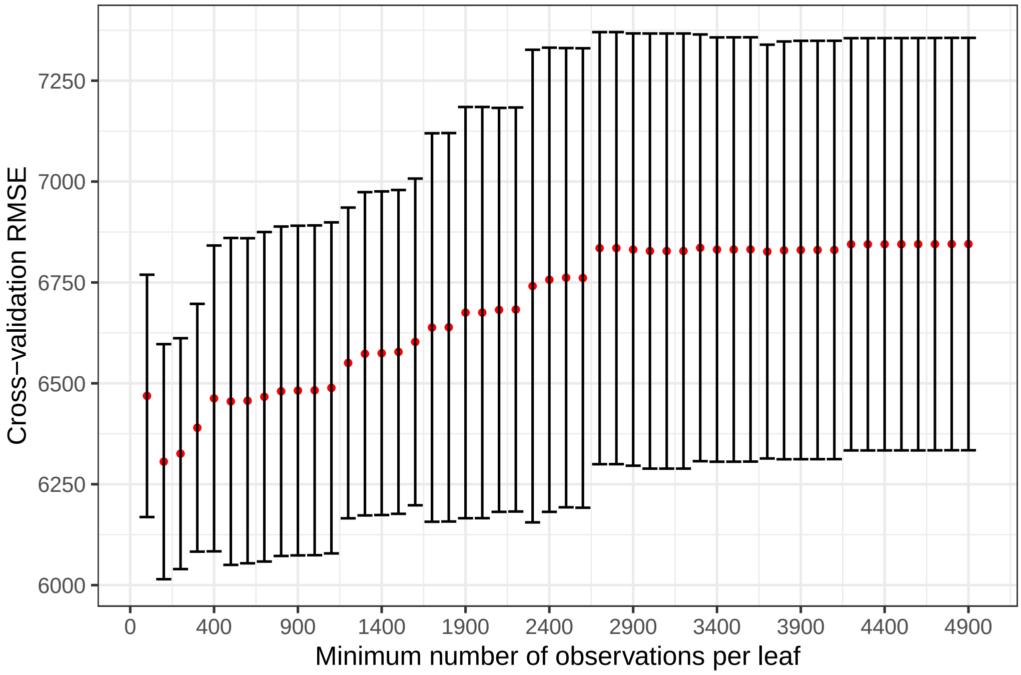


Figure 4.8: Cross-validation error for different values of the minimum number of observations per leaf

The wide confidence intervals show that the M5 model is highly sensitive to the change in the data set, which is not desirable when fitting predictive models. Another critical disadvantage regarding the M5 model fitted to our data set is the very large tree size. By setting the minimum number of instances per leaf to 104, the resulting pruned tree had 61 terminal leaves, which makes the model become extremely complex.

Given the large standard error estimated by the cross-validation, the one standard error rule leads us to a model with minimum number of observations per leaf equal to 1,700. The resulting tree is reduced to 22 terminal leaves, which can be found in Appendix B. This tree size is still relatively large (Appendix B.1) and the 22 linear models contain all of the covariates available (Appendix B.2). M5 does not perform variable selection, which compromises the interpretation of the model.

In terms of prediction accuracy in the test sample, we can see from Table 4.10 that the M5 model has the highest R^2 and lowest Gini measures among the best models tested so far. This is evidence of improvement of the fitting for high cost policyholders. On the other hand, all other measures are worse than the ones from lasso tuned according to the one-standard error criterion. The values of MAE and MAPE for the M5 model are still better than the ones observed for the two-part model, however, its $\mathrm{qMAD}_{0.95}$ is the largest among the models used for comparison. This shows that the complexity resulting from the structure of the M5 model does not lead to a superior prediction accuracy than the much simpler model. Thus, we judge that it is not useful for predicting medical claim costs using administrative claims data.

Table 4.10: Prediction accuracy in the test sample of M5 model with a minimum of 1,700 observations per leaf.

Prediction accuracy measures	**M5**	**Lasso** (λ_L^{se})	**Two-part**
R^2	0.25950	0.13229	0.11027
MAE	1,664	1,494	1,695
MAPE	0.83499	0.75149	0.85249
$\mathrm{qMAD}_{0.95}$	888	611	872
Sp.Cor.	0.66640	0.69307	0.70032
Gini	-0.75162	-6.18467	7.93416

In the face of this situation, we looked for a model closely related to M5 with a built-in mechanism that contributes to model simplification and perhaps improved predictive performance: the *Cubist* model, which is presented in more detail in the section below.

4.3.3 A Rule-based Method: Cubist

Cubist (Kuhn and Quinlan, 2018) is a rule-based model that improves upon the M5 model tree in an attempt to produce simpler models without loss of accuracy. A rule represents the path along the tree, from the root to the terminal leaf to which an observation belongs. It also contains the linear model in the terminal leaf. An example of a rule would be:

Conditions: logCOST <= 5 & AGE < 25 & GENDER = male

Linear Model: COST_2017 = 1,000 +200 logCOST_DEC +10 AGE

The initial stages of Cubist are the same as for the M5 model tree: the tree is grown, linear regression models are fitted to the intermediate leaves, the models are simplified and the tree is pruned. The differences between M5 and Cubist models start from the smoothing stage. According to Kuhn and Johnson (2013), instead of using equation (4.10), Cubist uses the following formula to smooth out predictions:

$$\mathrm{SV}\left(T_j\right) = a \cdot \mathrm{PV}\left(T_{j,i}\right) + (1-a) \cdot \mathrm{PV}\left(T_j\right) \tag{4.11}$$

where $\mathrm{SV}\left(T_j\right)$ is the smoothed predicted value in the parent node T_j, $\mathrm{PV}\left(T_{j,i}\right)$ is the predicted value in the child leaf $(T_{j,i})$, $\mathrm{PV}\left(T_j\right)$ is the predicted value in the parent leaf (T_j) and a is the smoothing coefficient, which is calculated as follows:

$$a = \frac{\mathrm{Var}\left(e_{T_j}\right) - \mathrm{Cov}\left(e_{T_j}, e_{T_{j,i}}\right)}{\mathrm{Var}\left(e_{T_j} - e_{T_{j,i}}\right)} \tag{4.12}$$

where e_{T_j} is the set of residuals of the parent node and $e_{T_{j,i}}$ is the set of residuals of the child node. The formula for the smoothing coefficient uses the variation of residuals in order to determine the weights for the smoothing process. If the variance of residuals from the parent node is too large, it is an indication that the fit in that node is not as good as the fit in the child node. This inflates the value of a, which gives more weight to the predictions from the child node in equation (4.11).

The initial number of rules is equal to the number of terminal leaves of the tree. Instead of keeping this as the final set of rules, the model evaluates the possibility of combining multiple rules into one single rule or even removing entire rules from the model. Each condition that forms a path in a particular rule is deleted in turn, starting near the terminal leaves, and the estimated accuracy error (equation 4.9) calculated without that condition. If the resulting accuracy error is lower, then the condition is permanently deleted. This is done for all conditions of a rule and for the entire set of initial rules, as long the estimated accuracy error is reduced.

According to Quinlan (1987), pruning or combining rules leads to a reduced model, facilitating the interpretation of the results. One issue regarding this approach is that an observation can now be covered by multiple rules, not just one as in the initial set of rules. The predicted value for an observation covered by several rules is the mean of the predictions given by each rule. If the goal is to define homogeneous and mutually exclusive groups to which an observation belongs, then this may represent a problem. Another limitation is that it is not possible (in most cases, especially for models with a large number of rules) to build back a tree from the rules given by the model, in the event that we want to know which conditions or entire rules were pruned away.

Cubist also allows the increase in prediction performance by combining several Cubist models that are fitted to the residuals of the previous models. Such feature is known as *committees*. As described by Kuhn and Johnson (2013), each new committee model is fitted to the adjusted response of the previous committee model according to the following equation:

$$\hat{y}_{(m)} = y - \left(\hat{y}_{(m-1)} - y\right) \tag{4.13}$$

where $\hat{y}_{(m)}$ is the adjusted response value used to build the m^{th} committee model and $\hat{y}_{(m-1)}$ is the value predicted by the previous committee model. The idea is that, if a model underestimates the value of an observation, their response will be inflated and a new prediction will be based on this inflated amount. It is hoped that this new prediction will have a higher value than the one predicted by the previous model, which may be closer to the actual value observed.

The Cubist package implemented in R allows the user to define the number of committees to be fitted to the data. The rules of each committee model are printed separately, so we can interpret each one separately. Although committees may improve prediction accuracy and reduce model variance, they generate several models which contain several rules and linear models with many parameters. Interpreting the results of so many models becomes burdensome and the improvement in accuracy may not compensate for the increase in complexity. Since model interpretability is important for us, we are not using committees to boost prediction accuracy in this study.

Another mechanism implemented in Cubist uses the information of similar observations (or neighbours) in order to enhance the prediction for a particular data point. This method is called *instance-based adjustment* (Quinlan, 1993) where instance is just another name for observation. The adjustment proposed is given by the following formula:

$$\frac{1}{K}\sum_{l=1}^{K} w_l \left[t_l + \left(\hat{y}_i - \hat{t}_l\right)\right] \tag{4.14}$$

where K represents the number of neighbours selected. Neighbours are the data points that are most similar to the observation for which we want to make predictions. $\hat{y}_i$ is the value predicted by the model for the i^{th} observation, t_l is the response value of the l^{th} neighbour, $\hat{t}_l$ is the value predicted by the model for the l^{th} neighbour and w_l is the weight based on the distance between the l^{th} neighbour and the predicted value $\hat{y}_i$.

The definition of distance used by Cubist is called *"Manhattan distance"* or *"city block"* (Kuhn and Johnson, 2013). Assume that we have two observations x and x'. The Manhattan distance between them is the sum of the absolute difference between all explanatory variables related to those observations, as shown in the equation below:

$$D(x, x') = \sum_{d=1}^{P} |x_d - x'_d| \tag{4.15}$$

where P is the number of all explanatory variables available in the set, not just those selected when building the tree. In cases where we have many explanatory variables, the ones that are irrelevant for predictions can act as confusing factors. Thus, this adjustment can actually be not beneficial at all. The Cubist package allows the user to define the number of neighbours that should be used in the adjustment of predictions, from 0 to 9 neighbours. The number of neighbours seems restricted, but remember that this adjustment is made for each prediction. If we have hundreds of thousands of observations in the data set (which is our case), it requires a large computational power to process all the adjustments.

Finally, the weights w_l are defined as follows:

$$w_l = \frac{1}{D_l + 0.5} \tag{4.16}$$

The larger the distance between the neighbour and the observation of interest (D_l), the lower is the weight related to that adjustment. The maximum weight possible is 2, which happens when (D_l) is zero. Thus, the neighbours that are more similar to a particular observation are more relevant to the adjustment made.

The size of improvement achieved from instance-based adjustment varies from data set to data set, as tested by Quinlan (1993) and Kuhn and Johnson (2013). From the perspective of ease of interpretation, it makes the model results less transparent, as we cannot simply apply the linear regression models and split conditions from the rule to find the same estimated value given by the R package output. For these reasons, we have decided not to use instance-based adjustment in our analyses.

4.3.4 Fitting Cubist

There are three hyper-parameters that can be used to tune the Cubist model. The first one is the maximum number of rules produced by the model, which is directly related to the complexity of the model. The second is the number of model committees used. The final one is the number of neighbours used in the instance-based adjustment.

In order to facilitate the interpretation of results, we set the number of committees and neighbours to zero. We used 10-fold cross-validation with RMSE as the loss function to find the optimum number of rules for our Cubist model. The potential number of rules ranged from 1 to 50, and the cross-validation results can be found in Figure 4.9 below.

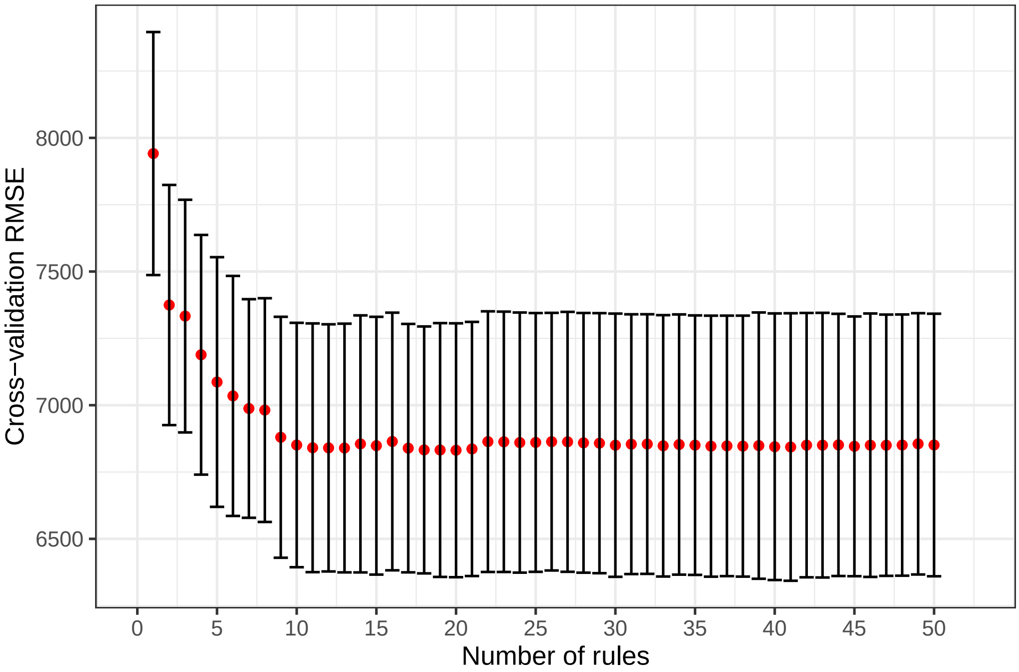


Figure 4.9: Cross-validation error for the maximum number of rules in Cubist

As can be seen, the cross-validated RMSE slowly reduces as the maximum number of rules increase. There is a small jump when increasing the number of rules from 8 to 9, followed by a slight decrease until the model has 11 rules. From that point onwards, there is no significant change in RMSE when varying the number of rules. We see two small peaks for models with 16 and 22 rules. For a number of rules larger than that, RMSE is practically unchanged, meaning that the model achieved the best predictive performance possible, and making the model more complex will not be reflected in more accurate forecasts for medical costs.

We can also observe that the response variable was kept in the original scale for Cubist. We attempted to fit the model for the log-transformed medical cost in the prediction year, in line with what has been done for the models previously fitted. However, the cross-validated RMSE function did not display a decreasing pattern as observed for the Cubist model fitted to the original cost scale. Figure 4.10 shows that the cross-validation RMSE starts low for just one rule (in this case, just a multiple linear regression is fitted) and it has the lowest error when the model has two rules. There is a steep rise in the RMSE when the maximum number of rules is increased to three. The error eventually reduces at a slow pace, but it is still higher than error of one rule model. This suggests that the log-transformed medical costs might be reasonably fitted to a multiple linear regression model, and there is no significant benefit in increasing the flexibility of the model with the Cubist method.

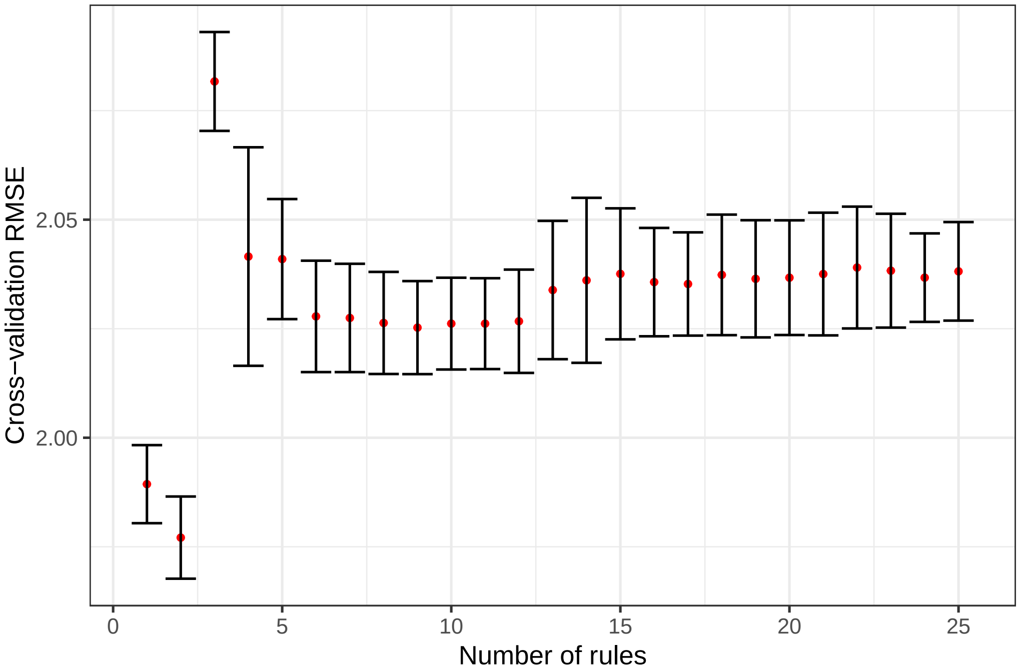


Figure 4.10: Cross-validation error for the maximum number of rules in Cubist model fitted to the log of the medical costs in 2017

Returning to Figure 4.9, the value for the maximum number of rules that minimizes the cross-validation RMSE is 20. According to the one-standard error criterion, we could also choose the model with only four rules. This model, however, is rather too simple, and may not produce satisfactory results in terms of prediction and insights related to variable importance. We may, thus, look for another value for the minimum number of rules that does not produce a very simple model (four rules) or an overly complex model (20 rules), but still provides a reasonable prediction accuracy compared to the accuracy of the model that minimises the cross-validated RMSE. In fact, we can observe that the average RMSE practically does not change for models built with a maximum number of rules between 11 and 20. The average RMSE of the model with 11 rules is 6,841, which is only 0.14% lower than the average RMSE of the model with 20 rules (6,831).

Thus, based on the graphical representation of the cross-validated RMSE, the model with 11 rules is our candidate for a model that potentially provides a better balance between complexity and accuracy. Thus, we will be comparing the performance of three different Cubist models: four, 11 and 20 rules. In the next section, we will compare the goodness-of-fit and prediction accuracy of these models and analyse the rules and linear models produced by the optimal model based on that comparison.

4.3.5 Cubist Results

We start by investigating how the goodness-of-fit measures of Cubist change with the increase in the number of rules. Table 4.11 shows these measures for Cubist models fitted using four (the one-standard error rule), 20 (number of rules that minimise the cross-validated RMSE) and 11 rules (number of rules that is an intermediate between model complexity and accuracy). We can see that Cubist with four rules already has a significantly larger R^2 than the lasso model fitted using the one-standard error rule and the two-part model, whose measures are also displayed in Table 4.11. On the other hand, the Spearman correlation coefficient is the lowest among the models. Its Gini statistic is negative and very far from zero, being the worst Gini statistic among the models fitted so far. Also, its MAE and MAPE are equivalent to the lasso model. Furthermore, its $\mathrm{qMAD}_{0.95}$ value is marginally larger than lasso's value, which shows that Cubist with four rules is not as efficient in fitting the costs of policyholders which are below the top 5% of the data. Thus, the larger R^2 comes from the ability of Cubist to fit the individuals in the top 5% largest costs, reducing the error for this group, which improves the impact of the squared residuals in the overall R^2 measure.

Table 4.11: Goodness-of-fit measures of Cubist models fitted with different number of rules, based on the training sample.

Goodness-of-fit measures	**Cubist** **4 rules**	 **11 rules**	 **20 rules**	**Lasso (λ_L^{se})**	**Two-part**
R^2	0.20614	0.29253	0.31461	0.13725	0.11081
MAE	1,499	1,446	1,417	1,497	1,696
MAPE	0.75175	0.72477	0.71058	0.75055	0.85020
$\mathrm{qMAD}_{0.95}$	631	612	599	613	871
Sp.Cor.	0.67767	0.69666	0.70268	0.69444	0.70247
Gini	-19.04112	-10.60308	-11.36506	-6.98140	7.29779

By analysing the measures of the Cubist model with 11 rules, we can see that adding seven rules to the Cubist model with four rules resulted in a significant improvement in terms of goodness-of-fit. There is a substantial increase of 0.08639 percentage points in R^2; a 3.58% decrease (hence, improvement) in terms of MAE and MAPE, making them the best measures in comparison to measures of previously fitted models, such as lasso and two-part, which are shown in Table 4.11. The 3.15% decrease in $\mathrm{qMAD}_{0.95}$ makes this measure equivalent to the one from lasso. There is an increase of Spearman correlation coefficient of 2.81%, which makes it become similar to the coefficient of lasso and two-part models. We can also notice a significant improvement in terms of Gini statistic. It reduced 44.32% relative to the Gini statistic of Cubist model with four rules, however, it is still worse than the Gini statistic of the lasso model.

Such a considerable improvement in goodness-of-fit is not observed when comparing the measures of the Cubist with 20 rules to the ones of the Cubist with 11 rules. The nine extra rules only cause an increase of 2.21 percentage points in R^2, 1.96% decrease in MAE and MAPE, 2.13% decrease in $\mathrm{qMAD}_{0.95}$ and 0.86% increase in Spearman correlation coefficient. Furthermore, there is a small deterioration in terms of Gini statistic. Thus, although the Cubist model with 20 rules has the best goodness-of-fit measures among all models tried, it is possible to obtain a much simpler Cubist model, such as the one with 11 rules, whose goodness-of-fit measures depart only modestly from the best achievable.

We can draw the same conclusions based on the prediction accuracy measures calculated using the data points from the test sample, which are displayed in Table 4.12. We can see that the prediction accuracy measures based on the test sample data points are very similar to the goodness-of-fit ones based on the training sample for all models compared. The Cubist with 11 rules is the best candidate of the model that balances complexity and prediction accuracy.

Nevertheless, lasso is much simpler than Cubist with 11 rules and still provides a similar prediction accuracy in terms of $\mathrm{qMAD}_{0.95}$. By comparing the prediction accuracy of the models for individuals in different cost deciles, which is shown in Table 4.13, we can see that lasso has the best performance when considering the policyholders in the cost decile 10%. For this reason, we observe a low $\mathrm{qMAD}_{0.95}$ for lasso in Table 4.12 is a consequence of this performance. On the other hand, Cubist provides the

Table 4.12: Prediction accuracy measures of Cubist models fitted with different number of rules, based on the test sample.

Prediction accuracy measures	**Cubist** **4 rules**	**11 rules**	**20 rules**	**Lasso (λ_L^{se})**	**Two-part**
R^2	0.20290	0.28110	0.28504	0.13229	0.11027
MAE	1,496	1,444	1,423	1,494	1,695
MAPE	0.75241	0.72636	0.71587	0.75149	0.85249
$qMAD_{0.95}$	630	612	599	611	872
Sp.Cor.	0.67521	0.69446	0.70106	0.69307	0.70032
Gini	-18.55179	-10.21668	-10.89581	-6.18467	7.93416

best performance for deciles 20% and 30% and the top deciles 90% and 100%. Among the models tested, Cubist is the only model to provide MAPE values lower than 0.8 for the two top deciles. This shows the advantage of Cubist over the other models. Since it splits the data into different groups, the policyholders whose costs are in the top of the distribution have different linear models which make and improve their predictions. However, the trade-off is a loss in prediction accuracy for low cost individuals.

Table 4.13: MAPE per cost decile of Cubist with 11 rules, lasso (λ_L^{se}) and two-part models, based on the test sample.

Cost decile	**Number of Policyholders**	**Average cost**	**Cubist 11 rules**	**Lasso (λ_L^{se})**	**Two-part**
10%	125,657	314.74	1.04551	**0.86487**	2.14957
20%	30,312	1,304.74	**0.47876**	0.50522	0.80236
30%	17,416	2,270.87	**0.45371**	0.52130	0.54157
40%	10,476	3,775.05	0.57628	0.60037	**0.45638**
50%	6,551	6,037.53	0.69490	0.69730	**0.51583**
60%	4,193	9,431.60	0.77481	0.77624	**0.61146**
70%	2,357	16,773.96	0.84636	0.84588	**0.74629**
80%	1,221	32,404.36	0.88498	0.89388	**0.85181**
90%	563	70,056.85	**0.78927**	0.89763	0.90414
100%	216	183,601.87	**0.71919**	0.91214	0.94534

We now analyse the conditions defined by the 11 rules of the Cubist model chosen. As seen in Table 4.14, only three covariates are used to define the 11 groups of the model: AGE, which is present in three rules; logCOST, used to define nine groups and logCOST_DEC, used in six rules. This reflects the relevance of the previous cost amount in splitting policyholders into more homogeneous groups, which results in enhanced prediction accuracy.

The rules are ordered according to the average cost in the prediction year of the group defined by the conditions of each rule. The first group contains younger policyholders (AGE ≤ 27), whose costs in the observation year are lower than 645.50 BRL (logCOST ≤ 6.47). The groups that follow include older policyholders and those with larger costs in the observation year. The covariate logCOST_DEC appears in rule four. In fact, the group formed by the last rule, whose average cost in the prediction year is the highest among all groups, is simply defined by logCOST_DEC > 9.64 (or $15,367$ BRL).

Overall, Cubist says that groups formed by older policyholders, whose costs in the entire observation year and in the last month of observation year were larger, tend to experience larger average costs in the prediction year. As well known in literature, this result suggests that age and previous cost are paramount for claim cost prediction. Gender, although relevant for the linear models, is not significant for defining the homogeneous groups. It is not a surprise that age and previous claim amounts are crucial for predicting medical claim costs, but the result from Cubist is important as a confirmation that they are the most relevant among the covariates tested. Also, it confirms that previous cost is a very good indicator of the health status of the policyholder, most probably due to the fact that, for a claim to exist in the administrative claims data, there must be an amount that was paid by the insurer. This amount is a very good proxy for the future costs of the individual.

The conditions shown in Table 4.14 also reveal the mechanism used by Cubist to improve prediction

Table 4.14: Conditions of the Cubist model with 11 rules

Rule	Condition	Number of observations	Average cost in prediction year
1	AGE <= 27 logCOST <= 6.47	133,971	495.45
2	logCOST <= 6.25	287,517	677.85
3	logCOST <= 8.87	573,348	1,600.87
4	logCOST_DEC <= 4.81 AGE >27	320,300	1,678.79
5	AGE >53 logCOST <= 8.87	131,966	3,032.45
6	logCOST_DEC >4.81 logCOST <= 8.87	88,502	3,470.43
7	logCOST >8.87 logCOST <= 10.51	20,212	7,971.62
8	logCOST_DEC <= 5.39 logCOST >10.51	754	12,380.19
9	logCOST_DEC >5.6 logCOST >8.87 logCOST <= 10.51	7,849	12,946.39
10	logCOST_DEC <= 9.64 logCOST >10.51	2,209	36,337.85
11	logCOST_DEC >9.64	523	65,534.27

accuracy. Observe that many policyholders in the first group are also part of the second group. Furthermore, both groups are part of the third group. As explained previously, the final predicted value of individuals who belong to more than one group is the average of the predictions given by the linear models of the groups that contain them. The average of the predicted values from different linear models is more likely to be closer to the individual's actual claim amount than the prediction made by only one linear model. Thus, the creation of more rules increases the chances that an individual is allocated to multiple groups. Consequently, several predictions are made for that individual, allowing the model to take the average of the predictions rather than consider only the prediction of a single linear model. This process increases the overall accuracy of Cubist model.

The heat-map in Figure 4.11 shows which covariates were used in the linear models of each of the 11 groups formed by Cubist. The horizontal axis represents the rule number and the vertical axis contains the covariate name. The coefficient size is on the log-scale. This means that we have applied the log of the absolute value of each coefficient of each model. This allows the graph to be more informative, since some coefficients are much larger than others. We multiply this result by the sign of the coefficient in the linear model. Thus, coefficients with negative signs in the linear models are also represented in the heat-map as negative, allowing us to interpret the impact of these covariates on the future yearly cost. The shades in blue represent the size of the coefficients of the each covariate in each linear model. The lighter the colour is, the larger the size of a positive coefficient is. Darker cells mean that those are largely negative coefficients. The grey cells indicate that the covariate was not present in the linear model.

The covariates are sorted according to the number of linear models they are part of and the size of their coefficients in those models. Thus, we can see that logCOST, AGE and propMAX_COST are present in all linear models. Also, the size of their coefficients increases with the rule number, indicating more relevance in determining the future cost as the model targets groups of policyholders that tend to cost more in the following year.

COPAY and HOSP_ACCOMM are the least important covariates, with COPAY appearing in the models of rules seven and nine and HOSP_ACCOMM in the linear model of rule six. In a sense, this agrees with lasso λ_L^{se}, because these covariates were not selected by that model (see Table 4.6). Nevertheless, Cubist does not select OWNER and PRE_EXISTING in any of the linear models. Thus, the model is suggesting that, although important, policy design features are secondary covariates, and that future claim costs depend mostly on the age of the individuals and their level of healthcare utilisation measured by total yearly costs and amount of costs in the end of observation year.

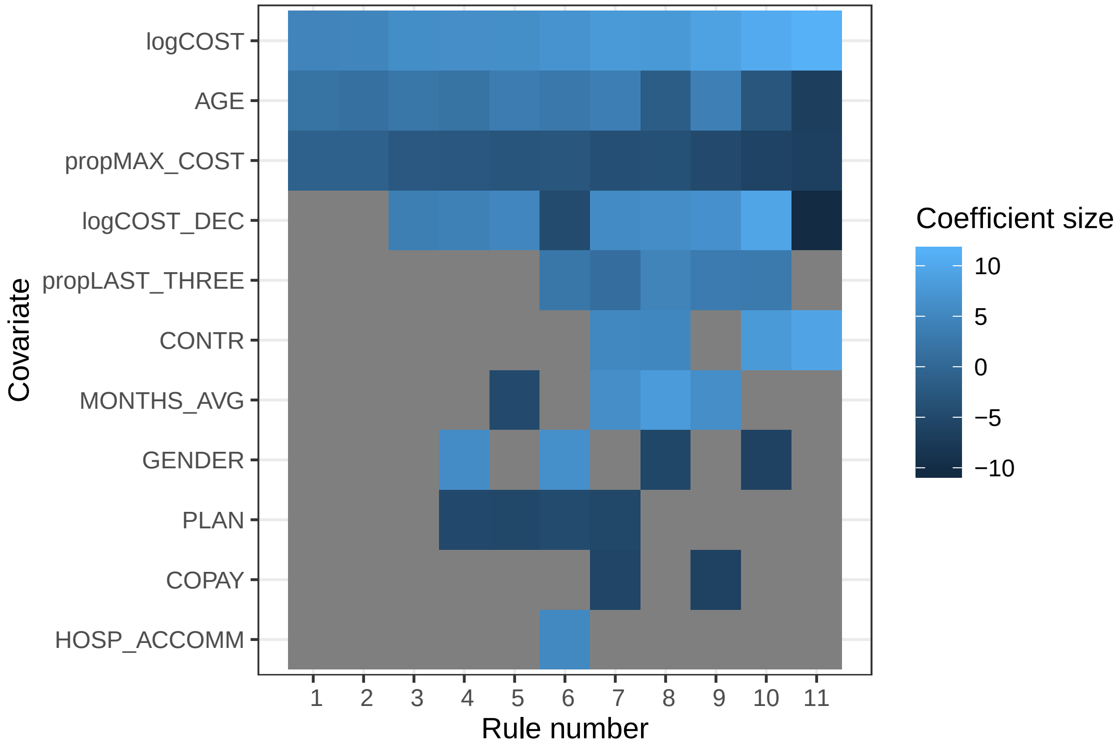


Figure 4.11: Heat-map of linear model coefficients of Cubist with 11 rules.

Generally, the effects of the chosen covariates in the response variable happen as expected and as observed for previous models. For instance, logCOST, propLAST_THREE and CONTR are always positive in all models where they appear, while propMAX_COST, PLAN and COPAY are always negative. The coefficients of other covariates varied depending on the group to which the models were fitted. For instance, AGE had positive coefficients in the linear models of all groups, except for the last two.

The close connection between AGE, logCOST and propMAX_COST as the most relevant covariates in the Cubist model may be a suggestion that the interaction between these covariates is important. When building traditional predictive models, interaction covariates based on these three variables (or combinations of two of these variables) may be added as extra covariates in the model. However, these effects may not be so straight-forward. For instance, the heat-map shows that the coefficients of AGE and propMAX_COST are largely negative, most probably to compensate for the very large positive coefficient of logCOST. This may cause an ambiguous effect when adding an interaction term between these covariates. Consequently, the interactions may not improve the fitting, and may be left out of the model, as observed in our initial analyses.

The complete output of the models are available in Appendix B: split conditions and coefficients of linear models of Cubist with four rules are described in Appendix B.3. Cubist model with 11 rules is found in Appendix B.4 and Cubist with 20 rules is found in B.5.

4.4 Final Considerations

In this chapter, we have described statistical learning methods that build upon the linear framework of traditional models in order to improve the model fitting process. The regularisation methods, ridge regression and lasso, incorporate a penalty factor, controlled by a shrinkage parameter, into the coefficient estimation process, limiting the size of the model coefficients.

For low values of the shrinkage parameter, both models provide the same results in terms of goodness-of-fit and prediction accuracy. However, because the penalty function of lasso allows for coefficients to be zero, a simpler model was achieved by dropping the not so relevant covariates COPAY, HOSP_ACCOMM and propLAST_THREE. This simplification does not impair the accuracy measures of the model in comparison to the less regularised lasso and ridge. In fact, it allows low cost policyholders (those in the first cost decile) to have fitted values with lower prediction error. Since these policyholders form the bulk of the data (in our data, 63% of policyholders are in the lowest cost decile), insurers can make use of lasso to improve predictions of low-cost individuals.

Lasso's ability to keep in the model only the most relevant covariates is a feature that will be useful for our further investigations, where we explore the relevance of detailed medical information from the claims. As will be explained in the next chapter, this will involve the creation of many more covariates. Keeping all of them in the model may only increase the complexity without benefiting prediction accuracy. For this reason, lasso is preferable over ridge regression.

A third regularisation model, elastic-nets, is a hybrid method that blends the penalty functions from lasso and ridge regression in the coefficient estimation process. An extra tuning parameter is incorporated, which controls the weight of each penalty function (ridge or lasso) that is applied during estimation. We refer the reader to James et al. (2013) and Friedman et al. (2001) for further explanations of this method.

The drawback of lasso and ridge regression is that both do not improve significantly the fitting of the tail of the distribution compared to the traditional methods fitted in the previous chapter. That is reflected in the significantly lower $\text{qMAD}_{0.95}$ of these models in comparison to the values of the traditional methods, without a significant improvement in R^2.

"Divide-and-conquer" methods that split the data into more homogeneous groups allow different portions of the data to be fitted separately, allowing policyholders with larger costs to have different forecast functions than the remaining ones. Tree-based methods are applied for categorical response variables, which is not the case for our study, since we have a continuous variable that we aim to predict. Thus, we initially have adopted a regression tree method called M5, which incorporates linear models in each of the groups defined, responsible for producing the predictions. Although this model is not useful for our data, it does reveal the path to the rule-based method Cubist, which turns out to be the best among all models tested in terms of prediction accuracy.

Cubist depends on the parameter that defines the maximum number of rules of the model. We show how to use cross-validation to choose a value for this parameter that balances model complexity and prediction accuracy. Consequently, we are able to find a model with 11 rules whose prediction accuracy measures do not vary much from the model with 20 rules that results in the lowest cross-validated error. We chose the 10-fold cross-validation method, but this can vary depending mostly on the size and type of data available. An alternative would be the leave-one-out cross-validation method, explained in James et al. (2013). As the name suggests, for each candidate value for the model's tuning parameter, the model is fitted to all data points of the training sample, except for one, used to calculate the error measure (RMSE, for instance) for that model. This is repeated for all data points in the training sample: at each loop, a different data point is left aside and used to calculate the error measure of the model trained using the remaining data points. Then the average of the error measures for each tuning candidate is taken, and the appropriate value is chosen. For situations where data volume is not so large, this can be a good alternative. However, our training sample contains 596,045 data points. If we were to apply leave-one-out cross-validation in 100 candidates for shrinkage factor in lasso, for instance, this would mean that $596,045$ x $100 = 59,604,500$ models would have to be fitted, which is computationally demanding. The 10-fold cross-validation provides an approximation that we judge to be satisfactory.

Undoubtedly, the optimal number of rules vary from data to data. However, the cross-validated error should behave in a similar way, allowing for a reasonable number of rules to be chosen based on the graphical representation of the cross-validation error. We choose RMSE due to its connection to the error measures used when assessing performance of the models; however, analysts may try different functions and check how the results differ.

A common result from the models fitted in this chapter is that covariates based on previous cost are the most dominant for prediction. We are able to see that not only on the size of their regularised coefficients in the lasso and ridge regression models, but also in the use of logCOST to split all groups in the selected Cubist models. Additionally, Cubist uses this covariate in all linear models fitted, along with logCOST_DEC, responsible for splitting the group of policyholders with largest average cost in the following year. This result supports the idea that the claim cost amounts reflect the health status of the policyholder. In a situation where further details regarding the medical procedures involved in the claim are unknown, claim amounts are very powerful as a predictor of the individual's future claim costs.

So far we have used past cost data aggregated by policyholder in order to fit our models. On top of demographic features, we also have included policy design covariates to better control for cost variations that do not depend directly on a person's health state. The superiority of cost amount covariates as predictors, however, motivates us to explore whether we could use more detailed information about the medical events in order to improve the predictions achieved so far with the aggregated data. In the following chapter, we will build covariates that break down the total medical costs into the related medical procedures and clinical information about the medical events and investigate how to improve the results from lasso, Cubist and the traditional methods.

Chapter 5

Using Detailed Medical Information From Administrative Claims Data For Predicting Future Costs

5.1 Introduction

In the previous chapters, we have fitted predictive models to covariates based on previous medical costs of insured individuals, as well as demographic and policy information. While previous cost variables are found to be the most relevant for the medical cost predictions, the administrative claims data are far from being fully explored.

These variables represent only the most aggregated level of medical costs that the claims data allow. More detailed information regarding the medical events can be incorporated into the predictive models in order to improve their accuracy. Additionally, diagnostic information is also reported by healthcare providers, and is another potential contribution to the task of improving prediction accuracy.

One of biggest challenges faced by analysts when incorporating these covariates in predictive models lies in translating the raw data from claims into meaningful covariates that allow insights to be made whilst increasing accuracy (Duncan, 2011). Finding the relevant information is also a point of great interest (Duncan et al., 2016; Morid et al., 2017), which brings us back to relying on models with built-in variable selection engines.

In this chapter, we explain how our data set relates to the data sets of other studies and we show how we organised and structured the claims data in order to create informative covariates.

Lasso and Cubist models are the best candidates among the models fitted so far. We now fit these models to the newly created covariates in order to find which claim features are more relevant for cost prediction. We compare them with the multiple linear regression model traditionally used to fit claims costs.

We conclude by observing that the cost-related covariates remain the most important. The improvement in prediction accuracy given by the more detailed covariates is not very significant, but the insights added by them might be of great value for decision makers working in insurance companies.

5.2 The Structure of Medical Data

In this section, we discuss the structure of medical information used for claim cost prediction. Because the type of information and the level of detail available for analysis vary largely depending on the source of the data used, we will briefly cover below the different structures and data sources found in the literature of medical cost prediction.

Surveys are typical sources of information for studies that focus on predicting healthcare costs of a population as a whole. This population may include insured individuals. An example of a publicly available survey used in a few studies is the Medical Expenditure Panel Survey (MEPS). According to the Agency for Healthcare Research and Quality (2019), this large-scale survey collects information from a sample of households representing the population of the United States every year. In addition to that, information regarding the healthcare utilisation (type of services, frequency and cost) of survey participants is gathered from their medical providers.

Examples in the literature regarding medical cost prediction include Frees, Gao and Rosenberg (2011), who use data from MEPS for years 2003 and 2004 in order to estimate the frequency and amount of in-patient admissions and out-patient visits. Frees et al. (2013) use MEPS for years 2005 to 2007 in order to fit multivariate two-part models that take into account possible dependencies among different types of medical events. The data available in the survey allow the authors to group the medical costs into five different categories of event: doctor visits, hospital out-patient services, emergency room, hospital in-patient admissions and home care.

One disadvantage of surveys is that they usually lack detailed information related to the medical procedures and diagnoses involved in treatments received by the individuals. In other words, the data describing the health status of individuals in surveys are limited. In both studies mentioned above, explanatory variables used in the models include self-rated physical and mental health, which reflects the perception of surveyed individuals about their own health status. These may be a misrepresentation of their true state of health, either because they are not able to categorise their status as accurately as a medical professional or because they may not wish to share that they have a specific health condition.

By contrast, administrative claims data can be seen as a potential source of detailed medical information that can be used to draw a better picture of the health status of insured individuals. This is because the data are built from the bills sent by healthcare providers to the insurance company for reimbursement of the services and treatments provided to policyholders.

In a fee-for-service system, healthcare facilities and professionals are paid for each service or procedure conducted. Given that health providers have a great interest in being reimbursed, they tend to fill in the reimbursement forms with information as complete and accurate as possible, which increases the quality of the data used in the models (Duncan, 2011).

The bills received by the insured companies are analysed and the data stored, allowing for a historical overview of the medical services demanded by each insured individual. From those bills, the insurance company can extract information related to diagnosis, medical procedure, medication and supplies used during the treatment, facility (hospital, laboratory, home care, etc) and medical professional responsible for the treatment. The healthcare system of a few countries also cover pharmacy claims, such as in the United States.

This information comes as a list of procedures, medications and/or supplies. Each represents a medical item. For instance, if a policyholder sees a doctor and takes a blood test, there are two items in the invoice received by the insurer. Each item in the bill is listed according to a coding system. There is a coding system for medical procedures, one for medications, one for medical supplies, one for payment of the services conducted by the medical professionals, and one for the diagnoses involved in the treatment, among others. Such level of detail is needed for reimbursement purposes but it is unlikely to be useful for cost prediction. Thus, groupings of medical items are usually made in order to make more sense of the data. These groupings vary from study to study. An example is the classification of medical procedures used by Duncan et al. (2016), which categorises the claim costs into four types of medical events: professional services, pharmacy costs, out-patient, and in-patient services. They also counted the number of visits to primary care provider and emergency room.

Bertsimas et al. (2008) follow a different approach and did not split their data into categories of medical event. Instead, the authors create more than 1,000 explanatory variables based on groupings of diagnoses and procedures codes. More specifically, they classify over 22,000 individual procedures into 180 groups and over 45,000 prescribed drugs into 336 groups. They also create over 700 risk measures reflecting many aspects of an individual's health, such as diabetic patients with foot ulcers, patients with emergency room visits but no office-based encounters, etc. Such tasks require the skills of expert medical professionals and the details of all variables created are not available.

In our data set, the claim costs are separated into six different types of medical events that resemble the categories used by Duncan et al. (2016) and MEPS survey cited above. We have used these categories because they are the standard classification of medical events adopted by the market to which the insurance company that provided the data belongs. In the data system of the insurer, the data is structured and stored in different sets according to these categories. The event types are:

- **Doctor visits**: the scheduled encounters made in the physician's office.
- **Emergency care**: this category combines the costs of all procedures that happened in the emergency room, including the initial contact with the physician on duty.
- **Hospital in-patient admissions**: this category takes into account all the medical procedures that happen during hospitalisations and last longer than 24 hours.
- **Home-care services**: this category comprises the costs of nurse visits and administration of medications made to patients that stay at home.
- **Medical tests and therapies**: this category comprises the additional procedures prescribed by doctors in order to investigate the health state of their patients or, in the case of therapies, repeated procedures necessary as the treatment of certain diseases (such as chemotherapy, radiotherapy, etc). It is important to note that the medical tests that occurred within hospitalisations and emergency events are not considered here, since they are taken into account in the classification of their respective events.
- **Other ambulatory procedures**: this category comprises the remaining procedures and medical items that do not fit into any of the other five event types.

We have created covariates that capture the cost information from these medical events. Six of them contain the proportion of the individual's yearly medical cost that refers to the particular medical event. The proportion of costs in each of the six event types of an average individual can be found in Table 5.1 below.

Covariate	propVISITS	propEMERGENCY	propINPATIENT	propHOMECARE	propTESTS_THERAPIES	propOTHER
Description	Doctor visits	Emergency	Hospital in-patient	Home-care	Medical tests or therapies	Other ambulatory procedures
Average proportion	41.2%	17.2%	6.5%	0.5%	28.1%	6.5%

Table 5.1: Average proportion of yearly medical costs by type of medical event

We can see that, on average, an individual is expected to spend just over 40% of their yearly costs on doctor visits and around 28% on medical tests or therapies. Obviously, each individual will have a different proportion of the yearly cost for each of the medical event types reflecting the procedures provided to them during that year. The idea is that, since these events are interconnected, the difference in the proportions may indicate a pattern of health-care utilisation that is likely to result in higher or lower claim costs in the following year. For instance, individuals with an accident type of pattern are likely to have their yearly cost largely concentrated in hospital in-patient procedures, while those that need continuous care are more likely to have their costs more spread out into events such as doctor visits, medical tests and therapies.

Other covariates related to the type of events and that were included in our models are: number of doctor visits (N_VISITS); number of emergency events (N_EMERGENCY); number of in-patient hospitalisations (N_INPATIENT); and duration of hospitalisations, in days (DAYS_INPATIENT). We also included variables that could provide more information regarding the complexity of hospitalisations. An indicator of complexity is whether a particular hospitalisation involved admission to an intensive care unit (ICU). The covariates COST_ICU and DAYS_ICU contain, respectively, the cost of all procedures that occurred during a hospitalisation in an intensive care unit and the number of days that the patient stayed in that ICU.

The spectrum of medical tests and therapies is very broad. In our data set, there are 1,599 different procedures identified as medical tests and 374 procedure code for therapies. Each of these procedures has a different purpose and may reflect a different health status of an individual or they might be related to more serious conditions that may result in higher costs. For these reasons, 12 distinct (and reasonably homogeneous) categories were created to group all the medical tests and therapies. For each category, we have created a covariate that contains the log of the total yearly cost of procedures that belong to that category and that are not made during in-patient hospitalisations or emergency care events. These covariates are listed in Table 5.2, which also contains a description of each category.

Covariate	Description
TT_ENDOSCOPY	Tests and therapies involving endoscopic procedures.
TT_PHYSIOTHERAPY	Physiotherapy sessions, usually for rehabilitation purposes.
TT_PHONOAUDIOLOGY	Phonoaudiology sessions.
TT_HEMATOLOGY	Procedures related to hematology and hemotherapy.
TT_IMAGE	Medical tests involving imaging, such as rx scans, tomography, ultrasound, among others.
TT_LAB	Clinical tests carried out in laboratories.
TT_MONITORING	Medical tests with the aim of evaluating and monitoring diseases or conditions, e.g. electrocardiograms.
TT_NUCLEAR	Procedures related to nuclear medicine.
TT_CHEMOTHERAPY	Chemotherapy sessions.
TT_RADIOTHERAPY	Radiotherapy sessions.
TT_ASSORTED	Broad range of therapy procedures that would not belong to any other group.
TT_SPECIFIC	Group of medical tests that are only requested in very specific cases. This category exists mainly for claim management purposes.

Table 5.2: Description of covariates related to the categories of medical tests and therapies.

The covariates presented so far cover the information regarding medical events that will be inputted into our models. However, the administrative claims data also contain some potentially valuable information directly related to the illnesses and health conditions of the claimants: the diagnosis information. In the following sub-section, we explain how they were incorporated into our models.

5.2.1 Diagnosis Information in Administrative Claims Data

While procedure, pharmacy and professional claims codes vary largely by country or even by insurance company (Duncan, 2011), diagnosis codes usually comply with an internationally accepted coding system, the International Statistical Classification of Diseases and Related Health Problems - ICD (World Health Organization, 2004).

ICD is a coding protocol that aims to standardise health-related conditions and causes of death. According to World Health Organization (2018), the international classification has been in constant development since its creation in 1893 (at which time it was known as the International List of Causes of Death). The aim of ICD codes is to facilitate the gathering of statistical morbidity and mortality information of an entire population or specific groups of people. This information can also be used to make inferences about a certain population at different points in time or comparisons of different populations at national or international levels.

Studies investigating medical cost prediction that used the 9th revision (ICD-9 or its clinical modification ICD-9-CM) in their models include Bertsimas et al. (2008) and Duncan et al. (2016). The 10th revision is currently used to code diagnoses in more than 100 countries (World Health Organization, 2018), which is also the case for our data set. This creates the opportunity for this study to be one of the first in claim cost prediction to use the ICD-10 codes for diagnoses. The 9th and 10th versions present major differences in structure, which will be briefly explained below, based on the information provided by World Health Organization (2011).

In the 10th revision, each code represents a category that can be formed from up to 10 subcategories. The codes have three or four digits, starting with a letter and followed by two or three integer numbers, ranging from A00.0 to Z99.9. In the 9th revision, all codes were numeric. The third numeric digit represents a subcategory, which provides more detail about the classification. All codes are organized into 22 chapters, displayed in Table 5.3 below. The ICD chapters are the most broad level of detail found in the classification, as they only indicate either the nature of the condition (Chapters

I to V and XV to XXI) or the body system affected by the condition (Chapters VI to XIV).

Chapter	Description
I	Certain infectious and parasitic diseases
II	Neoplasms
III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	Endocrine, nutritional and metabolic diseases
V	Mental, Behavioral and Neurodevelopmental disorders
VI	Diseases of the nervous system
VII	Diseases of the eye and adnexa
VIII	Diseases of the ear and mastoid process
IX	Diseases of the circulatory system
X	Diseases of the respiratory system
XI	Diseases of the digestive system
XII	Diseases of the skin and subcutaneous tissue
XIII	Diseases of the musculoskeletal system and connective tissue
XIV	Diseases of the genitourinary system
XV	Pregnancy, childbirth and the puerperium
XVI	Certain conditions originating in the perinatal period
XVII	Congenital malformations, deformations and chromosomal abnormalities
XVIII	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	Injury, poisoning and certain other consequences of external causes
XX	External causes of morbidity
XXI	Factors influencing health status and contact with health services
XXII	Codes for special purposes

Table 5.3: Description of the 22 ICD-10 chapters. Adapted from “ICD-10 Version: 2016”, available at https://icd.who.int/browse10/2016/en.

One limitation of our data set is that the ICD-10 codes reported by health-care providers are relative to hospital admissions. Thus, we only have the diagnosis information for individuals who are hospitalised during the observation year. Also, the providers usually fill in this information at the moment when the patient is being admitted to the hospital. Thus, the diagnosis reflects the health state of the patient as perceived by the doctor at the initial moment of the hospitalisation, not at the final moments. The classification of the diagnosis at discharge is usually more specific, making use of more detailed levels of the ICD-10 codes, which is only possible after more tests and medical procedures are done. The classification of the diagnosis at admission tends to be more broad and sometimes it is not much related to the real conditions of the patient, given the lack of proper assessment when reporting the ICD code. Also, during the hospitalisation, the patient can develop complications that are classified with different diagnosis codes, which makes the information collected at discharge more complete and more accurate regarding the health condition of the patient than the information collected at admission.

Since there are over 14,000 ICD-10 codes, it becomes impracticable to create one covariate for each code as there are not enough observations for each code to provide sufficient credibility to the results. For this reason, we built one covariate for each of the 22 chapters of the ICD-10 coding system. Each covariate contains the log of the total yearly cost of hospitalisations that occurred due to a disease or condition that belongs to the chapter that the covariate represents. By doing this, we hope that the complexity of hospitalisations involving diagnoses belonging to the same chapter can be indirectly captured by how costly they were.

Thus, we added 45 new explanatory variables to the 13 covariates (Table 4.1) used to fit our models in the previous chapters. From that total, 22 are related to the ICD-10 chapters capturing the diagnoses involved in the policyholders' hospitalisations (Table 5.3); 12 of them represent the categories of medical tests and therapies done (Table 5.2); five covariates describe the proportion of individual's yearly cost in observation year in each medical event type (Table 5.1. We kept propOTHER as the baseline type); the remaining six covariates are N_VISITS, N_EMERGENCY, N_INPATIENT, DAYS_INPATIENT, COST_ICU and DAYS_ICU. We fitted a two-part model, which provided the best performance among the traditional methods tested, using the stepwise method described in 3.7.1 to select the relevant covariates. We also used cross-validated RMSE to tune lasso and Cubist models.

5.3 Results of Predictive Models Using ICD-10 Chapters

5.3.1 Two-part Model

Table 5.4 shows the coefficient estimates of the 26 selected covariates for the first part of the two-part model fitted to the data considering the ICD-10 chapters. From the original covariates, only propLAST_THREE is judged not significant and is left out of the model. Overall, the effects of the covariates on the response variable are similar to what was observed in the two-part model in chapter 3 (Table 3.6).

Table 5.4: Coefficient estimates of the first part of two-part model - ICD-10 chapters

Covariate	Estimate	Std. Error	z value	Pr(> $\|z\|$)
(Intercept)	-0.74515	0.01988	-37.474	<2e-16
AGE	-0.00311	0.00028	-11.129	<2e-16
GENDERFemale	0.03867	0.00949	4.073	4.64e-05
OWNERDependant	-0.10353	0.01051	-9.849	<2e-16
PLANUnregulated	-0.82919	0.02429	-34.141	<2e-16
PLANRestricted	0.21630	0.01232	17.553	<2e-16
CONTRAssociation	0.05241	0.01220	4.296	1.75e-05
CONTRIndividual	0.48321	0.01377	35.087	<2e-16
HOSP_ACCOMMPrivate	0.06425	0.01133	5.672	1.41e-08
COPAYYes	0.14517	0.01301	11.157	<2e-16
PRE_EXISTING	0.30464	0.02409	12.648	<2e-16
logCOST	0.20366	0.00758	26.875	<2e-16
logCOST_DEC	0.10631	0.00371	28.680	<2e-16
MONTHS_AVG	0.19170	0.00980	19.561	<2e-16
propMAX_COST	-0.00256	0.00039	-6.589	4.42e-11
propVISITS	0.00697	0.00039	18.032	<2e-16
propTESTS_THERAPIES	0.00496	0.00048	10.297	<2e-16
propEMERGENCY	0.00546	0.00042	12.915	<2e-16
propINPATIENT	-0.00334	0.00061	-5.480	4.25e-08
propHOMECARE	0.03085	0.00392	7.873	3.46e-15
N_VISITS	0.21923	0.00488	44.893	<2e-16
N_EMERGENCY	0.12744	0.00781	16.324	<2e-16
TT_SPECIFIC	0.01921	0.00466	4.124	3.73e-05
TT_PHYSIOTHERAPY	-0.03412	0.00804	-4.246	2.17e-15
TT_IMAGE	0.02669	0.00373	7.150	8.69e-13
TT_LAB	0.05896	0.00402	14.676	<2e-16
ICD_XV	-0.07685	0.00848	-9.068	<2e-16

Among the new covariates introduced, N_VISITS has a large coefficient, which may suggest that individuals who visit the doctor regularly tend to maintain that behaviour, which increases the chances of future claims. All five covariates related to the proportion of the yearly costs in the observation year were significant. Interestingly, policyholders with the greater proportion of the costs coming from in-patient hospitalisations are less likely to claim in the following year.

Four categories of the medical tests and therapies are chosen, with an emphasis on physiotherapy sessions (TT_PHYSIOTHERAPY) which has a negative coefficient. It suggests that individuals with larger costs related to physiotherapy events tend to have lower costs in the future. Physiotherapy claims usually appear during the final stages of recovery from an injury, which means that individuals are likely to demand less medical treatment in the near future than they currently do. Only one ICD-10 chapter is selected, ICD_XV, related to pregnancy. Pregnant women who deliver in the observation year are less likely to need as much health care in the following year, which might explain the negative coefficient associated with that covariate. However, the newborn baby will incur costs for the insurer in the near future.

The second part of the model has many more covariates than the first part, 41 in total, whose coefficient estimates are shown in the Table 5.5. MONTHS_AVG is not selected by the stepwise method, but the remaining original covariates are retained, with coefficients similar to what is observed in the model fitted without detailed covariates (3.7).

Table 5.5: Coefficient estimates of the second part of two-part model - ICD-10 chapters

Covariate	**Estimate**	**Std. Error**	**t value**	**Pr(> $\|t\|$)**
(Intercept)	5.35300	0.00817	655.437	<2e-16
AGE	0.01099	0.00009	125.543	<2e-16
GENDERFemale	0.14620	0.00314	46.576	<2e-16
OWNERDependant	-0.01513	0.00334	-4.530	5.91e-06
PLANUnregulated	-0.21040	0.00894	-23.535	<2e-16
PLANRestricted	-0.33970	0.00413	-82.228	<2e-16
CONTRAssociation	0.05013	0.00392	12.778	<2e-16
CONTRIndividual	0.02825	0.00376	7.507	6.04e-14
HOSP_ACCOMMPrivate	0.05855	0.00350	16.751	<2e-16
COPAYYes	-0.06222	0.00389	-16.007	<2e-16
PRE_EXISTING	0.08045	0.00569	14.145	<2e-16
logCOST	0.14260	0.00166	85.899	<2e-16
logCOST_DEC	0.03915	0.00075	52.031	<2e-16
propMAX_COST	-0.00346	0.00008	-45.200	<2e-16
propLAST_THREE	0.00048	0.00006	7.469	8.08e-14
propVISITS	-0.00332	0.00009	-36.033	<2e-16
propTESTS_THERAPIES	-0.00307	0.00013	-23.988	<2e-16
propEMERGENCY	-0.00300	0.00011	-26.759	<2e-16
propINPATIENT	-0.00665	0.00020	-33.136	<2e-16
propHOMECARE	0.01354	0.00027	50.946	<2e-16
N_VISITS	0.04238	0.00051	83.678	<2e-16
N_EMERGENCY	0.04787	0.00111	42.955	<2e-16
N_INPATIENT	0.09044	0.00854	10.592	<2e-16
DAYS_INPATIENT	0.01034	0.00049	20.978	<2e-16
COST_ICU	0.01852	0.00194	9.545	<2e-16
DAYS_ICU	0.01062	0.00185	5.744	9.28e-09
TT_ASSORTED	0.00838	0.00224	3.746	1.80e-04
TT_PHYSIOTHERAPY	-0.00440	0.00121	-3.649	2.63e-04
TT_PHONOAUDIOLOGY	0.02996	0.00397	7.544	4.55e-14
TT_HEMATOLOGY	0.03676	0.00947	3.883	1.03e-04
TT_IMAGE	0.02250	0.00085	26.460	<2e-16
TT_LAB	0.01976	0.00101	19.549	<2e-16
TT_CHEMOTHERAPY	0.17970	0.00586	30.641	<2e-16
ICD_II	0.00815	0.00225	3.616	2.99e-04
ICD_V	0.03094	0.00501	6.179	6.46e-10
ICD_VI	0.02577	0.00387	6.662	2.71e-11
ICD_VII	0.02546	0.00486	5.243	1.58e-07
ICD_XI	-0.01390	0.00186	-7.458	8.79e-14
ICD_XIV	0.00802	0.00177	4.534	5.78e-06
ICD_XV	-0.06960	0.00186	-37.371	<2e-16
ICD_XIX	-0.00988	0.00236	-4.183	2.88e-05
ICD_XXI	-0.02183	0.00409	-5.338	9.39e-08

The coefficient related to propHOMECARE suggests that larger concentration of costs in the medical event type tends to lead to larger costs in the following year. This makes sense since patients usually receive home care for an extended time, as a consequence of chronic conditions, not acute ones. Seven types of medical tests and therapies are selected. Once again, TT_PHYSIOTHERAPY has a negative coefficient. On the other hand, TT_CHEMOTHERAPY has a large positive coefficient, indicating high future costs, which is expected due to the complexity of the treatment and may even indicate cancer remission or a need for other expensive care services, such as hospitalisation.

Nine ICD-10 chapters are selected, including ICD_XV, also selected in the model for part one. Once again, it has a negative coefficient, along with ICD_XI (digestive system), ICD_XIX (injury and poisoning) and ICD_XXI (factors influencing health status).

5.3.2 Lasso

We have performed a 10-fold cross-validation in order to define the appropriate size of the shrinkage factor for the lasso model. Once again, the error measure used is the RMSE and 100 potential values for the size of the shrinkage factor, ranging from 0.00001 to 100,000, have been tested.

The results of the cross-validation for lasso considering ICD-10 chapters as diagnosis covariates can be seen in Figure 5.1. We can see a very similar behaviour to what was observed for the lasso model fitted to the data without the medical details (Figure 4.5). RMSE is the lowest for small values of the shrinkage factor and the error does not change much until the shrinkage factor reaches values close to zero, where the RMSE has a steep increase until its maximum is reached and then remains constant, meaning that all coefficients were reduced to zero (null model).

The value that minimises RMSE is $\lambda_L^{min} = 0.00001$ (in the graph, $\log(\lambda_L^{min}) = -11.51292$ - striped vertical line on the left). The value based on the one-standard error criterion is $\lambda_L^{se} = 0.01707$, or $\log(\lambda_L^{min}) = -4.07023$, represented by the dotted line in the graph.

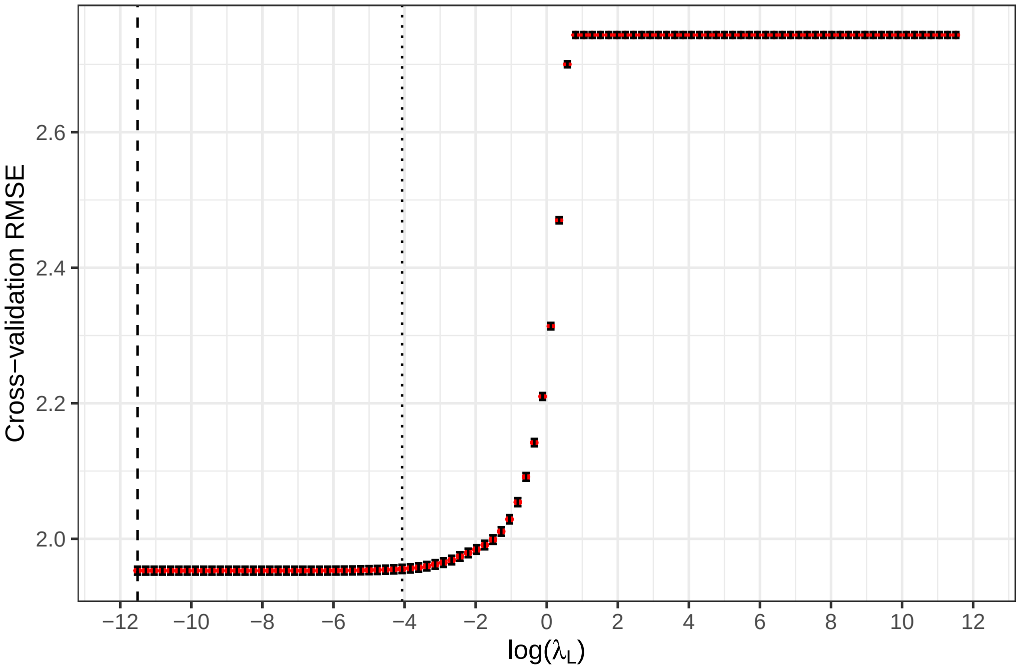


Figure 5.1: Change in the cross-validated RMSE with the increase of shrinkage factor.

The coefficient estimates of the lasso model are found in Table 5.6. The model tuned using λ_L^{min} contains all the covariates. When using the larger shrinkage factor, λ_L^{se}, the model is left with 23 covariates. From the original set of covariates used, HOSP_ACCOMM is the only one that is not selected. Their coefficients agree with the ones from lasso previously fitted (4.6), where logCOST has the largest positive coefficient.

Lasso tuned with λ_L^{se} confirms the importance of some of the new covariates previously selected by the two-part models. ICD_XV was the only ICD-10 chapter selected, and its coefficient is also negative. Chemotherapy costs remain as an important predictor of larger future costs, along with TT_SPECIFIC and more generic medical tests categories TT_IMAGE and TT_LAB.

The proportion of the costs related to in-patient hospitalisations and homecare are also relevant, with the former negatively impacting the amount of the policyholder's future yearly cost and the latter affecting it positively. Both results are observed in the two-part model. The sparse model also confirms the importance of the number of doctor encounters (N_VISITS), the number of emergency events (N_EMERGENCY) and both the frequency of in-patient hospitalisations (N_INPATIENT) and the number of days in the hospital (DAYS_INPATIENT).

Table 5.6: Coefficient estimates of Lasso models fitted using λ_L^{min} and λ_L^{se} - ICD-10 Chapters

Covariate	$\lambda_L^{min} = 0.00001$	$\lambda_L^{se} = 0.01707$	**Covariate**	$\lambda_L^{min} = 0.00001$	$\lambda_L^{se} = 0.01707$
(Intercept)	1.56378	1.68903	TT_HEMATOLOGY	0.00720	
AGE	0.00541	0.00472	TT_IMAGE	0.02987	0.02254
GENDER	0.13046	0.10547	TT_LAB	0.04859	0.04378
OWNER	-0.08778	-0.06224	TT_MONITORING	0.00353	
PLAN	-0.08148	-0.06905	TT_NUCLEAR	-0.03026	
CONTR	0.07704	0.06939	TT_CHEMOTHERAPY	0.08883	0.01075
HOSP_ACCOMM	0.04078		TT_RADIOTHERAPY	-0.04996	
COPAY	0.13007	0.05898	ICD_I	0.02132	
PRE_EXISTING	0.18764	0.14876	ICD_II	0.03281	
logCOST	0.54722	0.54339	ICD_III	0.04576	
logCOST_DEC	0.05340	0.05535	ICD_IV	0.01519	
MONTHS_AVG	0.02831	0.08369	ICD_V	0.06235	
propMAX_COST	-0.00661	-0.00285	ICD_VI	0.04596	
propLAST_THREE	0.00113	0.00067	ICD_VII	0.04657	
propVISITS	0.00793	0.00384	ICD_VIII	0.03281	
propTESTS_THERAPIES	0.00292		ICD_IX	0.02909	
propEMERGENCY	0.00435		ICD_X	0.03254	
propINPATIENT	-0.00849	-0.00683	ICD_XI	0.00624	
propHOMECARE	0.01422	0.00952	ICD_XII	0.02053	
N_VISITS	0.02512	0.02688	ICD_XIII	0.02847	
N_EMERGENCY	0.03073	0.02935	ICD_XIV	0.02647	
N_INPATIENT	0.03880		ICD_XV	-0.05945	-0.06577
DAYS_INPATIENT	0.00342	0.00065	ICD_XVI	0.00923	
COST_ICU	0.00429		ICD_XVII	0.02728	
DAYS_ICU	-0.00492		ICD_XVIII	0.02461	
TT_ASSORTED	0.01765		ICD_XIX	0.00617	
TT_ENDOSCOPY	-0.00177		ICD_XX	0.02174	
TT_SPECIFIC	0.00736	0.00247	ICD_XXI	-0.00771	
TT_PHYSIOTHERAPY	-0.01283		ICD_XXII	0.07549	
TT_PHONOAUDIOLOGY	0.01918				

5.3.3 Cubist

The cross-validated RMSE for Cubist is displayed in Figure 5.2. We can see that the average quickly drops as the number of rules increases, until it reaches the minimum with only seven rules. From that point, the RMSE remains virtually unchanged, but there is an increase in the one-standard error interval for models with 25 rules or more. The addition of the detailed covariates contributes to a simplification of the model, in terms of the number of rules, if compared to the Cubist model fitted to the 13 initial covariates (4.9). In that context, the minimum RMSE was reached for a model with 20 rules. By using more detailed information, Cubist requires fewer rules in order to achieve lower prediction error.

The model with three rules is the one defined by the one-standard error criterion. This number of rules is very small, resulting in a model that is too simple. However, unlike what has been done for the original covariates, we judge that a model with a number of rules between three and seven is not necessary. The model that minimises the RMSE is already simple enough, thus, we judge that a simpler model is not needed. Also, we can observe in Figure 5.2, that choosing a number of rules that is fewer than seven results in larger value of the cross-validated RMSE.

The conditions resulting from the model with seven rules are displayed in Table 5.7. Despite the addition of 45 new covariates, only AGE, logCOST and logCOST_DEC remain as necessary for the splits. This confirms the superiority of previous cost amounts over the detailed medical information for defining more homogeneous groups of policyholders.

The groups of the first two rules are very broad, and target a very large number of policyholders. In fact, the first group is a subset of the second, highlighting Cubist's strategy of allocating the same individuals to different groups, so that the average of their predicted values can be used as the final prediction, thereby increasing the overall accuracy. This also happens for group five, which is a subset of group three, and group four, a subset of group six. The costs in the last month of the observation year (logCOST_DEC) is the main feature that determines the group to which the individual belongs. Individuals with larger costs in that month tend to have a larger average yearly cost in the prediction year.

Nonetheless, many of the more detailed data incorporated into Cubist are present in the linear models within the groups created. Figure 5.3 shows the heat-map of the coefficients of all covariates

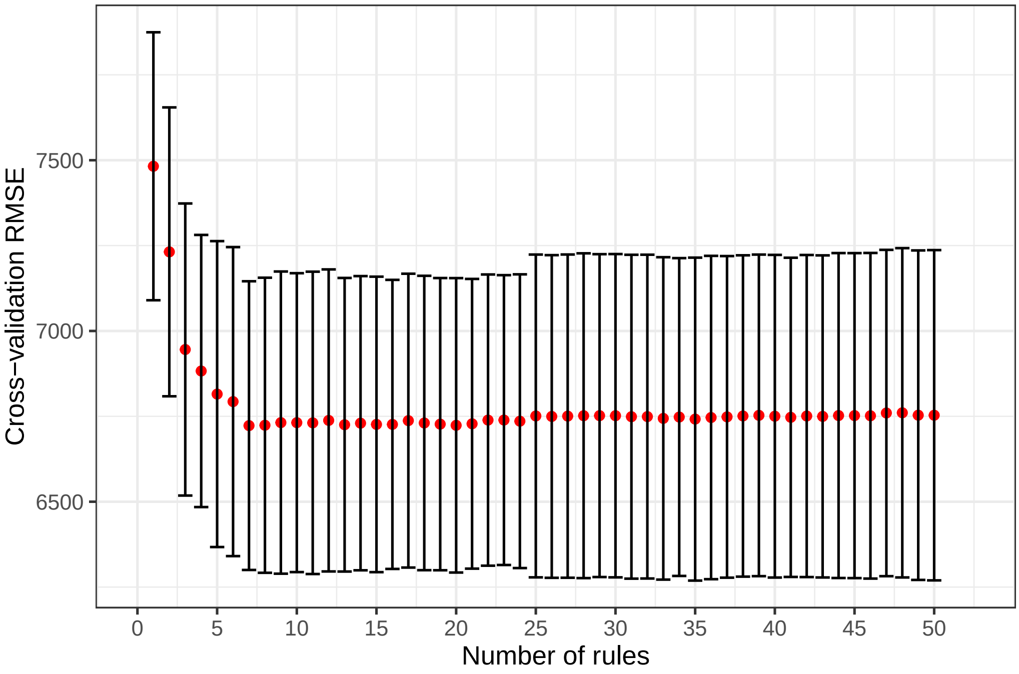


Figure 5.2: Change in the cross-validated RMSE with the increase of the maximum number of rules of Cubist model.

Table 5.7: Conditions of Cubist with seven rules - ICD-10 Chapters

Rule	**Condition**	**Number of observations**	**Average cost in prediction year**
1	AGE <= 53 logCOST <= 8.87	441,382	1,172.85
2	logCOST <= 8.87	573,348	1,600.87
3	logCOST >8.87 logCOST <= 10.51	20,212	7,971.62
4	logCOST_DEC <= 5.39 logCOST >10.51	754	12,380.19
5	logCOST_DEC >5.61 logCOST >8.87 logCOST <= 10.51	7,849	12,946.39
6	logCOST_DEC <= 9.64 logCOST >10.51	2,209	36,337.85
7	logCOST_DEC >9.64 logCOST >10.51	278	106,404.23

used in the linear regression models of the Cubist model with seven rules. The complete output of this model is included in the Appendix C.2, while the simpler model, with three rules, can be found in Appendix C.1.

To some extent, Cubist agrees with lasso and two-part models when it comes to the relevance of a few covariates for cost prediction. For instance, lasso (λ_L^{se}) selects propINPATIENT and propHOMECARE as the most relevant among the five covariates defined with the proportion of the yearly cost in each medical event type. These two are within the top covariates most often used in the linear models of Cubist with seven rules.

From the original set of covariates, logCOST and logCOST_DEC are shown as relevant in the heatmap, appearing in six out of seven linear models. TT_CHEMOTHERAPY, which has a large positive coefficient for the second part of the two-part model and was selected by lasso λ_L^{se}, is also among the most used covariates in the linear models. On the other hand, ICD-10 chapters do not appear in

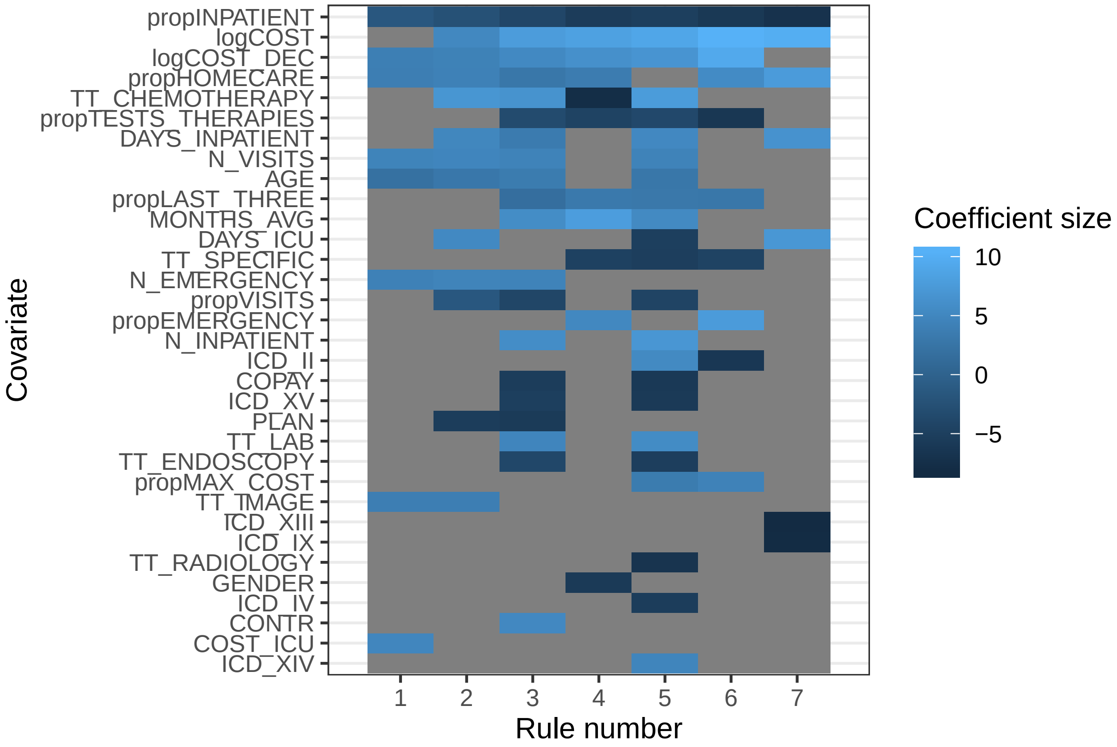


Figure 5.3: Heat-map of coefficients from the linear models in Cubist with seven rules - ICD-10 chapters.

many models. Chapters II (neoplasms) and XV (pregnancy) are present in two linear models each. Chapters IX (injury and poisoning) and XIII (diseases of musculoskeletal system) are in the linear model of the last rule, which is related to the group with the largest average cost in the prediction year, both with negative signs.

Goodness-of-fit and Prediction Accuracy Measures

The goodness-of-fit measures of all models fitted to the data considering ICD-10 chapters can be found in Table 5.8. As expected, the two-part performs poorly in comparison with lasso and Cubist models in all measures except Spearman correlation coefficient. The two lasso models provide very similar results in all measures except for R^2 and Gini statistic. In terms of R^2, the lasso tuned using λ_L^{min} improves upon lasso fitted with a larger tuning parameter (λ_L^{se}) by 2.94 percentage points. Nevertheless, the Gini statistic of the lasso model using λ_L^{se} is closer to zero, indicating better fitting. Also, this model is much simpler than the one fitted using λ_L^{min}, since it has many fewer covariates and there is practically no loss in goodness-of-fit considering MAE, MAPE and Spearman correlation coefficient. In fact, qMAD$_{0.95}$ of lasso that used λ_L^{se} is slightly better than the measure for the lasso λ_L^{min}. Thus, we can select lasso λ_L^{se} as the best among the lasso models in terms of goodness-of-fit.

Table 5.8: Goodness-of-fit measures of two-part, lasso and Cubist models for ICD-10 chapters, based on the training sample

Goodness-of-fit measures	**Two-part**	**Lasso**		**Cubist**	
		λ_L^{min}	λ_L^{se}	**3 rules**	**7 rules**
R^2	0.11440	0.15871	0.12933	0.28059	0.32610
MAE	1,723	1,480	1,480	1,486	1,435
MAPE	0.86400	0.74213	0.74223	0.74486	0.71947
qMAD$_{0.95}$	857	609	601	655	623
Sp.Cor.	0.71557	0.70856	0.70682	0.67363	0.68782
Gini	-3.85425	-14.39344	-10.59153	-10.35869	-12.90198

As we have anticipated, Cubist with seven rules is significantly better than Cubist with three rules in all goodness-of-fit measures, except for the Gini statistic. Its R^2 is 4.55 percentage points better, its MAE and MAPE are 3.42% superior, its $\mathrm{qMAD}_{0.95}$ is 4.82% lower (thus, better) and its Spearman correlation coefficient is 2.11% larger. Thus, seven rules is the obvious choice among the two Cubist models.

It is interesting to observe how the Gini statistic of the Cubist model deteriorates when the number of rules increases from three to seven. This can be evidence that, in order to improve the overall goodness-of-fit (and prediction accuracy), Cubist focuses on improving the fitting of larger claims, which represent only a small percentage of the data points (policyholders). This is because the residuals of very large claims tend to proportionally cause more impact on the overall fitting than the residuals of small claims. Consequently, the fitting of larger claims benefits from the increase in the number of rules, while the smaller claims (the bulk of the data) suffer a small deterioration, affecting the ranking of fitted claims in comparison to the observed claim amounts. Since the Gini statistic focuses on the rank of the data, this measure is negatively affected by the fitting strategy, adopted by Cubist, of focusing on the tail of the distribution.

The conclusions taken from the goodness-of-fit measures also apply to the prediction accuracy measures based on the test sample, shown in Table 5.9. Comparing the two lasso models, the model tuned using λ_L^{se} has worse R^2, but the values of the remaining measures are similar to the values of lasso λ_L^{min}, with an additional advantage of having a Gini statistic closer to zero and the best $\mathrm{qMAD}_{0.95}$ among the models considered in this section. Once again, Cubist with seven rules overcomes the prediction accuracy of Cubist with three rules in all measures, except for the Gini statistic. The two-part model shows a big disadvantage in terms of prediction accuracy when compared to lasso and Cubist, except for the Spearman correlation coefficient and Gini statistic.

Table 5.9: Prediction accuracy measures of two-part, lasso and Cubist models for ICD-10 chapters, based on the test sample

Prediction accuracy measures	**Two-part**	**Lasso**		**Cubist**	
		λ_L^{min}	λ_L^{se}	**3 rules**	**7 rules**
R^2	0.16612	0.15352	0.11891	0.26907	0.29986
MAE	1,719	1,479	1,479	1,484	1,436
MAPE	0.86459	0.74385	0.74410	0.74644	0.72251
$\mathrm{qMAD}_{0.95}$	857	607	599	654	623
Sp.Cor.	0.71326	0.70688	0.70528	0.67121	0.68498
Gini	-3.28282	-13.62226	-9.79274	-9.27465	-12.40097

The best Cubist model (with seven rules) outperforms our choice for lasso (λ_L^{se}) by a significant margin in terms of R^2, MAE and MAPE. On the other hand, lasso provides the lowest $\mathrm{qMAD}_{0.95}$ and highest Spearman correlation coefficient and a Gini statistic that is closer to zero. This is a consequence of the much better fitting provided by lasso for individuals in the lowest cost decile in comparison with the fitting provided by Cubist. This is represented in Table 5.10. Both models perform similarly for individuals in cost decile 20%, but for all other deciles, particularly in the top deciles, which contain the most expensive individuals, Cubist does a much better job than lasso. Thus, we consider Cubist with seven rules as being the optimal choice of predictive model, since it is able to make better use of the data provided in order to provide improved prediction accuracy.

One possible problem with the diagnosis data used so far is that the chapters of ICD codes may not provide an insightful interpretation of the diagnoses since the codes are categorised according mainly to the systems of the human body. For this reason, we have decided to try an alternative grouping system for the diagnosis data that are more related to clinical conditions of individuals. Such a grouping is described in sections 5.4 and 5.6.

Table 5.10: MAPE per cost decile of lasso and Cubist models for ICD-10 chapters, based on test sample

Cost decile	**Number of Policyholders**	**Average cost**	**Lasso (λ_L^{se})**	**Cubist 7 rules**
10%	125,657	314.74	**0.82882**	1.11130
20%	30,312	1,304.74	**0.53301**	0.53427
30%	17,416	2,270.87	0.53204	**0.46748**
40%	10,476	3,775.05	0.58910	**0.54229**
50%	6,551	6,037.53	0.66928	**0.65997**
60%	4,193	9,431.60	0.74838	**0.73632**
70%	2,357	16,773.96	0.82655	**0.82335**
80%	1,221	32,404.36	0.88547	**0.87090**
90%	563	70,056.85	0.90627	**0.77710**
100%	216	183,601.87	0.92199	**0.70233**

5.4 The Charlson Comorbidities

Charlson et al. (1987) have developed and validated a method to classify clinical conditions that largely affect the mortality risk of patients. The initial purpose of this classification was to rank individuals into different risk scores. In order to do this, the authors define 17 categories of comorbid diseases based on the conditions described in medical records of a group of patients who were hospitalised at New York Hospital in 1984. A different weight is assigned to each of the comorbidities. An index score is created for each patient as a result of the sum of the weights of each comorbidity affecting that patient. Thus, a large index indicates that the patient is afflicted by several comorbidities or a comorbidity that results in further health complications, such as AIDS or tumor, both having the highest weights among the comorbidities defined.

A few years later, Deyo et al. (1992) link the Charlson comorbidities to the ICD-9-CM codes, allowing administrative claims data to be used on studies with clinical purposes. As an example, D'Hoore et al. (1996) use the Charlson index as an exploratory variable for models developed to predict the mortality of patients admitted to hospital with ischemic heart disease. They conclude that the index is strongly associated with in-patient death, being a relevant risk factor for short-term mortality.

The index is also relevant for assessing the health care costs during the final days of life of hospitalised patients with advanced cancer (Zhang et al., 2009). In that study, the authors use the comorbidity index to separate the patients into two groups with a similar health profile. A more recent study has used the Charlson index in the development of machine learning algorithms aiming to predict early death of elder cancer patients (Sena et al., 2019).

The creation of the ICD-10 codes has imposed a need to adapt the Charlson comorbidity index to this new coding system of diagnoses. Motivated by that, Halfon et al. (2002) and Sundararajan et al. (2004) have translated Deyo's classification from ICD-9-CM to ICD-10 independently, resulting in some differences between them. In order to create a unique classification for the comorbidities, Quan et al. (2005) have unified the definitions for both ICD-9 and ICD-10 coding systems. We include, in Appendix D, the list of ICD-10 codes that compose each of the 17 comorbidities defined by Charlson. This classification is also implemented in the **icd** R package (Wasey and R Core Team, 2019).

The comorbidities defined by Charlson provide a few advantages over other grouping systems used in the literature. As mentioned, the risk score based on the comorbidities has been a relevant predictor of death and an important factor for analysing medical costs of patients in their final moments of life. Because it has only 17 comorbidities, the Charlson classification targets the more severe conditions which are likely to incur higher health care costs. Other groupings of diagnoses have a broader coverage of diseases, but they are more likely to be redundant and less useful for cost prediction. For instance, Duncan et al. (2016) have used the Hierarchical Condition Categories from the Centers for Medicare and Medicaid Services model (CMS-HCC), which has 189 categories. However, due to the redundancy of many categories, the authors have narrowed down the number of categories to 83. The description of the CMS-HCC model can be found in Pope et al. (2004).

In our study, we do not use the comorbidity index as a single measure of the policyholder's health state. Instead, we have created 17 explanatory variables, one for each comorbidity defined by

Charlson. For each individual, these covariates contain the log-transformed cost of hospitalisations that occurred in the observation year due to a particular comorbidity. This approach allows the models to select which comorbidities are more related to future medical costs. Also, using the cost of the hospitalisations related to each comorbidity (and not just a dummy variable that indicates whether or not a patient has that disease) can enable us to place increased weight on the more complex procedures that resulted from the treatment of that comorbidity. In other words, two policyholders may have been admitted to hospital due to the same comorbidity, but the hospitalisation where the costs are larger probably indicates that this individual is in a worse health state than the one whose hospitalisation costs are lower. This brings different implications for the prediction of future costs.

An additional covariate (COMORBID) was created as the total number of comorbidities identified for an individual during hospitalisations that occurred in the observation year. The 17 covariates and the description of the respective comorbidity are found in Table 5.11.

Covariate	Comorbidity
MI	Myocardial infarction
CHF	Congestive heart failure
PVD	Peripheral vascular disease
STROKE	Cerebrovascular disease
DEMENTIA	Dementia
PULMONARY	Chronic pulmonary disease
RHEUMATIC	Rheumatic disease
PUD	Peptic ulcer disease
LIVER_MILD	Mild liver disease
DIABETES	Diabetes without chronic complication
DIABETES_CX	Diabetes with chronic complication
PARALYSIS	Hemiplegia or paraplegia
RENAL	Renal disease
CANCER	Any malignancy, except malignant neoplasm of skin
LIVER_SEVERE	Moderate or severe liver disease
METS_TUMOR	Metastatic solid tumor
HIV	AIDS/HIV

Table 5.11: Covariates related to the Charlson comorbidities

A total of 5,871 individuals have been hospitalised due to the occurrence of at least one of the 17 comorbidities defined. This represents 8.8% of individuals that were hospitalised in the observation year, meaning that the great majority of hospitalisations have been due to conditions other than the ones defined by Charlson. However, these comorbidities reflect either complex chronic conditions or very complex acute conditions which distinguish them from less severe illnesses. Thus, if a hospitalisation that contains one of these categories is judged as significant by the models, it may provide better predictions for those individuals than the model considering more general and broad classification of diagnosis, such as the ICD-10 chapters. That is why we are using the Charlson classification for diagnosis instead of the ICD-10 chapters. In the following sections, we show the results for two-part, lasso and Cubist models fitted to these new data and will analyse the outcomes of all models fitted.

5.5 Results of Predictive Models Using Charlson Comorbidities

5.5.1 Two-part Model

Using the stepwise approach, a total of 24 covariates have been selected for the first part of the two-part model. Their estimated coefficients can be found in Table 5.12 below. We can observe that none of the covariates related to the Charlson comorbidities appears in the model, thus, not being relevant for predicting the likelihood of a policyholder to claim in the following year.

From the original set of covariates, GENDER and propLAST_THREE are not significant in this model. The effects of the other covariates in the response variable are similar to those observed in the first part of the two-part model fitted using the ICD-10 chapters for the diagnosis data (Table 5.4).

Table 5.12: Coefficient estimates of the first part of two-part model - Charlson comorbidities

Covariate	**Estimate**	**Std. Error**	**z value**	**Pr(> $\|z\|$)**
(Intercept)	-0.73667	0.01967	-37.457	<2e-16
AGE	-0.00294	0.00028	-10.584	<2e-16
OWNERDependant	-0.09529	0.01035	-9.207	<2e-16
PLANUnregulated	-0.82653	0.02427	-34.061	<2e-16
PLANRestricted	0.21553	0.01232	17.496	<2e-16
CONTRAssociation	0.05384	0.01220	4.414	1.01e-05
CONTRIndividual	0.48743	0.01375	35.450	<2e-16
HOSP_ACCOMMPrivate	0.06279	0.01132	5.544	2.95e-08
COPAYYes	0.14397	0.01301	11.064	<2e-16
PRE_EXISTING	0.30584	0.02407	12.704	<2e-16
logCOST	0.20521	0.00756	27.139	<2e-16
logCOST_DEC	0.10722	0.00370	28.948	<2e-16
MONTHS_AVG	0.19568	0.00979	19.980	<2e-16
propMAX_COST	-0.00260	0.00039	-6.698	2.11e-11
propVISITS	0.00691	0.00039	17.859	<2e-16
propTESTS_THERAPIES	0.00494	0.00048	10.247	<2e-16
propEMERGENCY	0.00540	0.00042	12.771	<2e-16
propINPATIENT	-0.00430	0.00059	-7.241	4.45e-13
propHOMECARE	0.03106	0.00396	7.852	4.11e-15
N_VISITS	0.21729	0.00487	44.627	<2e-16
N_EMERGENCY	0.12350	0.00778	15.880	<2e-16
TT_SPECIFIC	0.01976	0.00465	4.248	2.16e-05
TT_PHYSIOTHERAPY	-0.03227	0.00802	-4.023	5.74e-05
TT_IMAGE	0.02556	0.00370	6.913	4.74e-12
TT_LAB	0.05768	0.00401	14.388	<2e-16

The second part of the two-part model using Charlson comorbidities has 33 covariates, whose coefficient estimates are shown in Table 5.13. Three covariates related to the Charlson comorbidities have been selected: myocardial infarction (MI), renal disease (RENAL) and any comorbidity (COMORBID). Hospitalisations associated with myocardial infarction usually involve highly complex and expensive surgical procedures, which occurred in the observation yer. Because we have retained in our analysis only individuals who were alive during the entire prediction year, it means that patients who went through such hospitalisations are alive and most probably healthier than before. Thus, such a highly expensive medical procedure is unlikely to happen in the following year, which may explain the negative sign associated with the coefficient of MI. Also, this classification related to myocardial infarction is likely to arise mostly from acute cases, which are less likely to incur costs as large as those from chronic conditions. On the other hand, hospitalisations related to renal diseases are related to a chronic condition and, most probably, continuous treatments that follow these types of hospitalisations are likely to occur in the near future.

Once again, chemotherapy sessions predicts larger costs in the following year. Image and laboratory tests, although being very common medical procedures, appear as relevant in this model, as well as in all other models fitted using these data. Yearly costs highly concentrated in home care also predict larger costs in the future. Thus, overall, the model confirms the importance of many of the covariates introduced, while highlighting the most important among the Charlson comorbidities.

Table 5.13: Coefficient estimates of the second part of two-part model - Charlson comorbidities

Covariate	**Estimate**	**Std. Error**	**t value**	**Pr(> $\|t\|$)**
(Intercept)	5.34700	0.00818	654.057	<2e-16
AGE	0.01124	0.00009	128.821	<2e-16
GENDERFemale	0.13970	0.00313	44.616	<2e-16
OWNERDependant	-0.01244	0.00334	-3.722	1.97e-04
PLANUnregulated	-0.21160	0.00895	-23.638	<2e-16
PLANRestricted	-0.34000	0.00414	-82.209	<2e-16
CONTRAssociation	0.05156	0.00393	13.129	<2e-16
CONTRIndividual	0.03212	0.00377	8.528	<2e-16
HOSP_ACCOMMPrivate	0.05710	0.00350	16.317	<2e-16
COPAYYes	-0.06404	0.00389	-16.457	<2e-16
PRE_EXISTING	0.08038	0.00570	14.114	<2e-16
logCOST	0.14510	0.00166	87.397	<2e-16
logCOST_DEC	0.03977	0.00075	52.823	<2e-16
propMAX_COST	-0.00360	0.00008	-47.064	<2e-16
propLAST_THREE	0.00051	0.00006	7.933	2.14e-15
propVISITS	-0.00331	0.00009	-35.891	<2e-16
propTESTS_THERAPIES	-0.00297	0.00013	-23.389	<2e-16
propEMERGENCY	-0.00292	0.00011	-26.008	<2e-16
propINPATIENT	-0.00747	0.00020	-38.069	<2e-16
propHOMECARE	0.01336	0.00027	50.200	<2e-16
N_VISITS	0.04214	0.00050	84.513	<2e-16
N_EMERGENCY	0.04518	0.00111	40.659	<2e-16
N_INPATIENT	0.06568	0.00829	7.923	2.32e-15
DAYS_INPATIENT	0.01156	0.00044	25.979	<2e-16
COST_ICU	0.01674	0.00197	8.479	<2e-16
DAYS_ICU	0.00839	0.00182	4.609	4.05e-06
TT_ASSORTED	0.00917	0.00224	4.095	4.22e-05
TT_PHONOAUDIOLOGY	0.03056	0.00398	7.686	1.53e-14
TT_IMAGE	0.01984	0.00085	23.388	<2e-16
TT_LAB	0.01777	0.00099	17.864	<2e-16
TT_CHEMOTHERAPY	0.16830	0.00580	29.005	<2e-16
MI	-0.03098	0.00598	-5.180	2.22e-07
RENAL	0.13130	0.01009	13.014	<2e-16
COMORBID	0.03598	0.00212	16.968	<2e-16

5.5.2 Lasso

Once again, the error measure used for cross-validation is the RMSE and 100 potential values for the size of the shrinkage factor, ranging from 0.00001 to 100,000, have been tested. The results can be seen in Figure 5.4. They show very similar behaviour to what was observed for the model using ICD-10 chapters (5.1). The value that minimises RMSE for this is $\lambda_L^{min} = 0.00016$. In the graph, the striped vertical line highlights this value ($\log(\lambda_L^{min}) = -8.72191$). The tuning factor based on the one-standard error criterion is $\lambda_L^{min} = 0.01707$, with $\log(\lambda_L^{min}) = -4.07023$ represented by the dotted line in the graph.

Based on those values for the tuning parameter, we fitted two lasso models. Their estimated coefficients are shown in Table 5.14. So far, lasso models tuned using the parameter that minimises RMSE produce models with all covariates inputted. It was not the case for the set of covariates that considers Charlson comorbidities for diagnosis. The resulting model fitted with λ_L^{min} does not have two of the comorbidities: peripheral vascular disease (PVD) and cerebrovascular disease (STROKE).

The more regularised model, fitted using λ_L^{se} has 23 covariates, two of them related to the Charlson comorbidities, RENAL and COMORBID, which are also relevant in the second part of the two-part model. Chemotherapy, a strong predictor according to other models, is not significant in this one. Overall, the model confirms the dominance of previous costs (logCOST) as a predictor, and reinforces the effect of some covariates on the response variable.

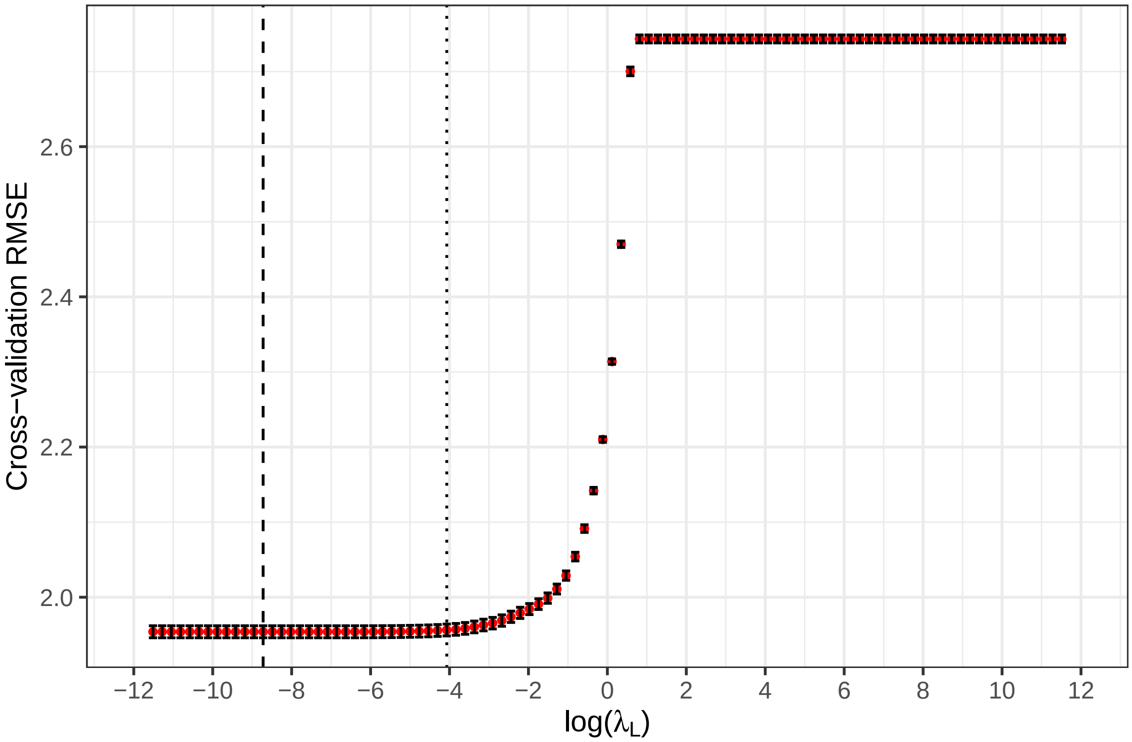


Figure 5.4: Change in the cross-validated RMSE with the increase of shrinkage factor.

Table 5.14: Coefficient estimates of Lasso models fitted using λ_L^{min} and λ_L^{se} - Charlson comorbidities

Covariate	$\lambda_L^{min} = 0.00016$	$\lambda_L^{se} = 0.01707$	**Covariate**	$\lambda_L^{min} = 0.00016$	$\lambda_L^{se} = 0.01707$
(Intercept)	1.55854	1.68184	TT_PHYSIOTHERAPY	-0.00952	
AGE	0.00560	0.00492	TT_PHONOAUDIOLOGY	0.01937	
GENDER	0.12400	0.09906	TT_HEMATOLOGY	0.00747	
OWNER	-0.08464	-0.06010	TT_IMAGE	0.02721	0.02040
PLAN	-0.08185	-0.06888	TT_LAB	0.04633	0.04179
CONTR	0.07908	0.07092	TT_MONITORING	0.00773	0.00148
HOSP_ACCOMM	0.03894		TT_NUCLEAR	-0.02737	
COPAY	0.12794	0.05797	TT_CHEMOTHERAPY	0.08440	
PRE_EXISTING	0.18779	0.14890	TT_RADIOTHERAPY	-0.04817	
logCOST	0.54699	0.54490	MI	-0.01777	
logCOST_DEC	0.05439	0.05612	CHF	0.02240	
MONTHS_AVG	0.03226	0.08591	PVD		
propMAX_COST	-0.00652	-0.00287	STROKE		
propLAST_THREE	0.00114	0.00068	DEMENTIA	-0.06031	
propVISITS	0.00781	0.00376	PULMONARY	0.02816	
propTESTS_THERAPIES	0.00282		RHEUMATIC	0.01096	
propEMERGENCY	0.00431		PUD	-0.02534	
propINPATIENT	-0.00815	-0.00797	LIVER_MILD	0.06056	
propHOMECARE	0.01417	0.00931	DIABETES	0.04402	
N_VISITS	0.02485	0.02649	DIABETES_CX	-0.06291	
N_EMERGENCY	0.02819	0.02695	PARALYSIS	0.01892	
N_INPATIENT	0.09457		RENAL	0.10352	0.00667
DAYS_INPATIENT	0.00570	0.00067	CANCER	0.00161	
COST_ICU	0.00687		LIVER_SEVERE	-0.07726	
DAYS_ICU	-0.00582		METS_TUMOR	-0.00125	
TT_ASSORTED	0.01865		HIV	0.03557	
TT_ENDOSCOPY	0.00060		COMORBID	0.03082	0.02121
TT_SPECIFIC	0.00811	0.00287			

5.5.3 Cubist

We have also used 10-fold cross-validation to find the appropriate number of rules of the Cubist model fitted to the covariates including Charlson comorbidities. The results of this process are shown in Figure 5.5. We can see that the average RMSE rapidly decreases as the number of rules goes from two to three. Then the average RMSE slowly decreases until it reaches its minimum for a model with eight rules. From that point onwards, the average RMSE remains virtually stable and an increase in the one-standard error interval is observed for models with 21 rules or more.

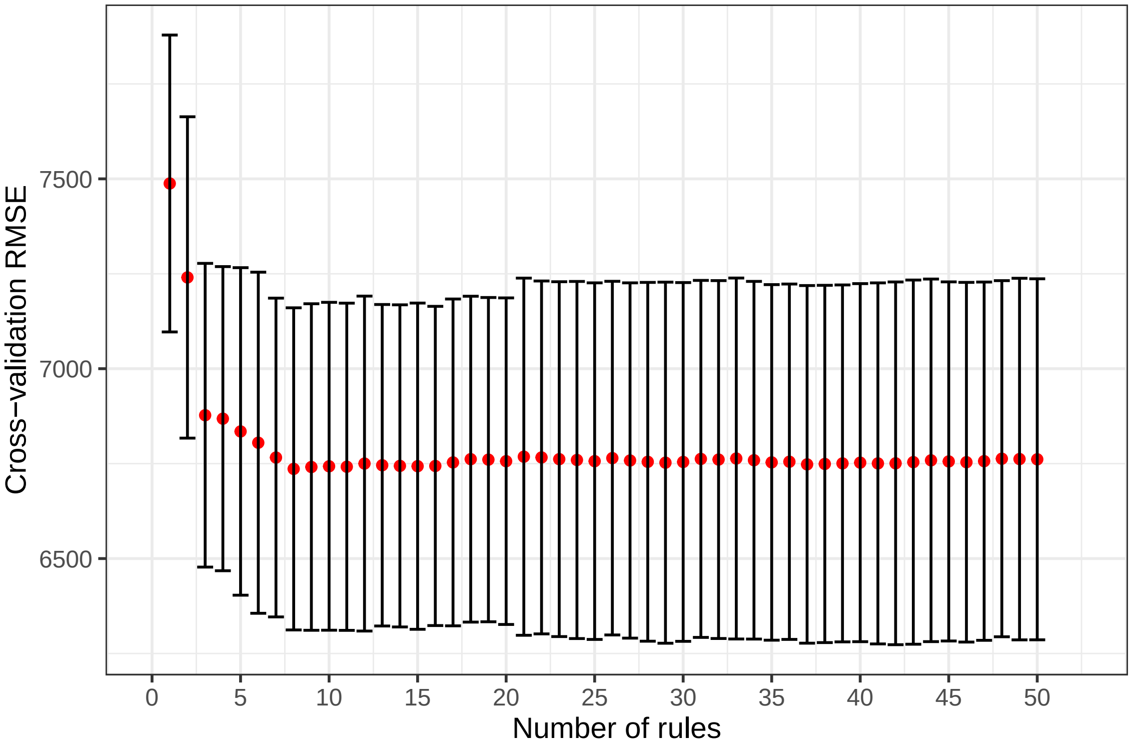


Figure 5.5: Change in the cross-validated RMSE with the increase of the maximum number of rules of Cubist model.

The number of rules that minimises the average RMSE for Cubist fitted to Charlson data (eight rules) is similar to that of the model fitted to ICD-10 chapters (seven rules). This emphasises the ability of Cubist to build simpler models from the more detailed claims data. The model with three rules is the representative of the one-standard error rule. Its output can be found in the Appendix E.1, while the complete output for the model with eight rules is found in Appendix E.2.

The conditions related to the eight rules produced by Cubist are found in Table 5.15. Once again, Cubist uses only logCOST, logCOST_DEC and AGE to make the splits. In fact, some of the groups defined by this model are exactly the same groups defined by the Cubist with seven rules fitted to the ICD-10 data. Specifically, the second group and groups four to eight are the same as the groups two to seven, respectively, in the ICD-10 model (Table 5.7).

Figure 5.6 shows the heat-map of the coefficients of covariates used in the linear regression models of the Cubist model with eight rules. It establishes that logCOST is the most relevant, by being present in all models, with relatively large coefficients. The proportion of yearly costs in in-patient and home care medical events are, once again, among the most important covariates. Other relevant covariates include doctor visits (both in terms of proportion of the yearly cost related to that event type - propVISITS - and number of encounters - N_VISITS), logCOST_DEC and chemotherapy sessions. From the set of covariates related to Charlson comorbidities, renal diseases (RENAL) confirms its relevance, along with COMORBID.

Table 5.15: Conditions of Cubist with eight rules - Charlson comorbidities

Rule	Condition	Number of observations	Average cost in prediction year
1	logCOST_DEC <= 4.81 AGE <= 53 logCOST <= 8.87	384,852	962.28
2	logCOST <= 8.87	573,348	1,600.87
3	AGE >53 logCOST <= 8.87	131,966	3,032.45
4	logCOST >8.87 logCOST <= 10.51	20,212	7,971.62
5	logCOST_DEC <= 5.39 logCOST >10.51	754	12,380.19
6	logCOST_DEC >5.61 logCOST >8.87 logCOST <= 10.51	7,849	12,946.39
7	logCOST_DEC <= 9.64 logCOST >10.51	2,209	36,337.85
8	logCOST_DEC >9.64 logCOST >10.51	278	106,404.23

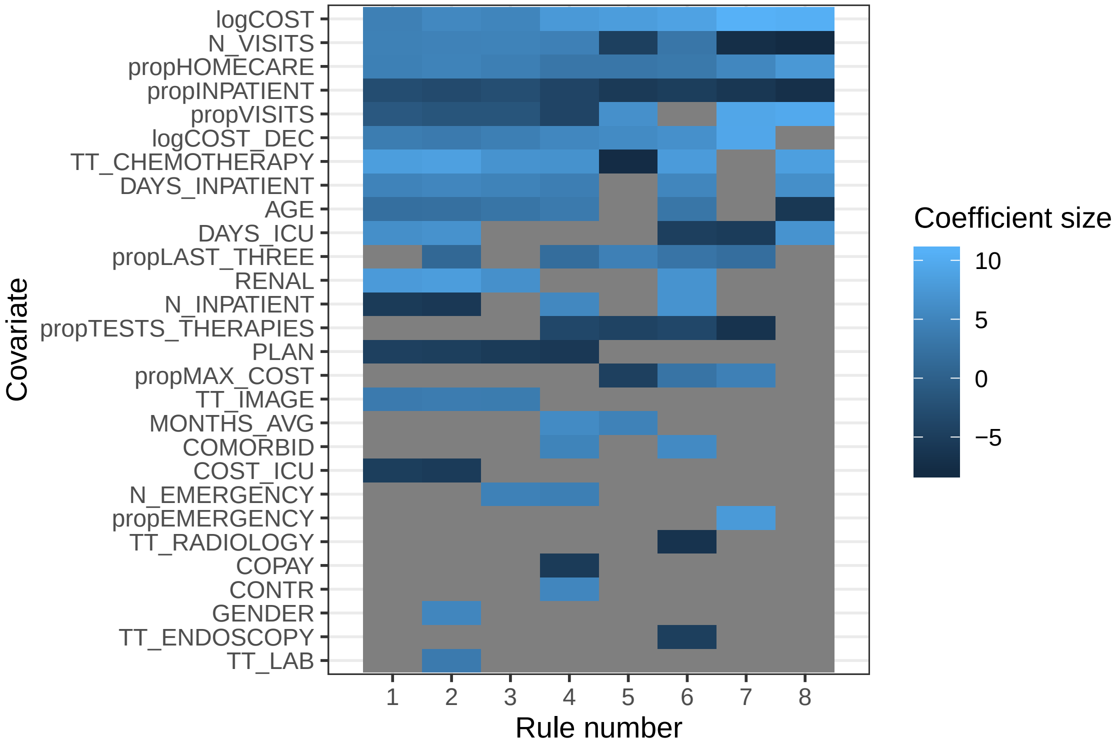


Figure 5.6: Heat-map of coefficients from the linear models in the Cubist model with eight rules.

Goodness-of-fit and Prediction Accuracy Measures

In terms of goodness-of-fit (Table 5.16) and prediction accuracy (Table 5.17), the models perform as observed for the ICD-10 data. As expected, the two-part model provides the worst measures (based both on training and test sample) in comparison with other methods, except for the Spearman correlation coefficient and Gini statistic.

Lasso λ_L^{se} provides better qMAD$_{0.95}$ and Gini statistic but worse R^2 than lasso λ_L^{min}, both in terms of goodness-of-fit and prediction accuracy. Both models have similar performance according to the other measures. Cubist with eight rules, once again, shows its superiority, yielding a significantly better R^2, MAE and MAPE than both lasso models and the two-part model. It also largely improves upon the goodness-of-fit and prediction accuracy from Cubist with three rules. However, its qMAD$_{0.95}$ values (based on training and test samples) are lower than the ones from lasso models. Its Gini statistic is also slightly worse, which is in line with what was observed for the ICD-10 chapters. This confirms the trade-off from using Cubist over lasso. There is a significant gain in the fitting of individuals with larger costs; however, there is a reduction in accuracy for individuals with lower costs.

Table 5.16: Goodness-of-fit measures of two-part, lasso and Cubist models for Charlson comorbidities, based on the training sample

Goodness-of-fit measures	**Two-part**	**Lasso**		**Cubist**	
		λ_L^{min}	λ_L^{se}	**3 rules**	**8 rules**
R^2	0.10350	0.15740	0.13055	0.28472	0.32681
MAE	1,727	1,481	1,481	1,483	1,434
MAPE	0.86455	0.74258	0.74262	0.74341	0.71898
qMAD$_{0.95}$	857	610	602	649	621
Sp.Cor.	0.71463	0.70762	0.70599	0.66563	0.69279
Gini	-4.17339	-14.61230	-10.64222	-10.88468	-13.41145

Table 5.17: Prediction accuracy measures of two-part, lasso and Cubist models for Charlson comorbidities, based on the test sample

Prediction accuracy measures	**Two-part**	**Lasso**		**Cubist**	
		λ_L^{min}	λ_L^{se}	**3 rules**	**8 rules**
R^2	0.09152	0.15870	0.11995	0.27279	0.29729
MAE	1,730	1,477	1,480	1,480	1,436
MAPE	0.86741	0.74433	0.74433	0.74475	0.72242
qMAD$_{0.95}$	857	608	600	648	621
Sp.Cor.	0.71253	0.70609	0.70460	0.66277	0.69012
Gini	-4.01373	-14.07565	-9.82971	-10.06959	-12.68341

By using the Charlson comorbidities, we could establish that hospitalisations related to renal diseases or any of the Charlson comorbidities are related to larger future costs. However, there is no improvement over the goodness-of-fit and prediction accuracy measures provided by the models using ICD-10 chapters (see Tables 5.8 and 5.9). This is because the majority of the explanation of future costs comes from covariates related to the types of events rather than the diagnoses. Additionally, the Charlson comorbidities only covered the diagnoses of 8.8% of the patients hospitalised, leaving the majority of the diagnoses information unused. To overcome these problems, we have tried a different classification system, used in the Global Burden of Disease study, which is described in the following section.

5.6 The Global Burden of Disease Study

The Global Burden of Disease (GBD) Study is an international project conducted by the Institute for Health Metrics and Evaluation (IHME). It aims to describe the causes of mortality and morbidity across 195 countries. Because IHME uses data since 1990, it is possible to track the changes in prevalence and incidence of conditions that are responsible for the death or impairment of a population over the years. The IHME does this by gathering data from all over the world, standardising them, and making estimations by age, sex, region (country level is the most common definition for region, but in some cases the data allow for sub-national locations) and year.

The study provides measures that reflect not only the prevalence of a specific disease, injury or risk factor, but also the impact that they have in the population in terms of years of life lost or years of life lived with disability. The most recent study, GBD 2017, has given rise to a series of five publications, which cover age-specific mortality and life-expectancy (Dicker et al., 2018; Roth et al., 2018), incidence and prevalence of diseases or injuries (James et al., 2018), healthy life expectancy (Kyu et al., 2018) and assessment of clusters of risk factors (GBD 2017 Risk Factor Collaborators, 2018).

The 282 causes of death or disability used by the study are classified into four different levels. The most detailed classification, level four, defines all 282 causes. These causes are grouped into 169 different categories, which compose level three. Level two classifies the causes into 21 groups and level one, the most broad level, allocates them into only three categories: (1) Communicable, maternal, neonatal, and nutritional diseases; (2) Non-communicable diseases; and (3) injuries.

The data related to the causes of death and disabilities are collected by IHME as ICD codes, which, as explained previously in this chapter, is the standardised and validated coding system for diseases and injuries. After cleaning and treating the data, the IHME associates each ICD-10 code into one of the causes defined. The group of ICD-9 and ICD-10 codes that compose each condition is described by Global Burden of Disease Collaborative Network (2018). Based on that, we have categorised the ICD-10 codes present in our hospitalisations data according to the level two classification of GBD causes. We have selected this level because level one only has three categories, which is too broad, while levels three and four have too many categories, and many of them would have an insufficient number of observations that would allow meaningful analyses to be made. We provide the mapping of 12,473 ICD-10 codes into the level two causes in Appendix F. 74% of hospitalisations are related to one of the causes, a proportion that is much larger than the 8.8% from the Charlson classification.

Table 5.18: Proportion of hospitalisations of policyholders in the data set analysed and proportion of deaths in Brazil by GBD level 2 cause

Covariate	**Cause**	**Proportion of hospitalisations**	**Proportion of deaths in Brazil (2016)**
NON_COM	Other non-communicable diseases	21.75%	3.26%
CARDIO	Cardiovascular diseases	17.19%	28.34%
DIGESTIVE	Digestive diseases	17.03%	5.29%
MATERNAL	Maternal and neonatal disorders	15.62%	1.89%
NEOPLASM	Neoplasms	10.90%	17.98%
CRD	Chronic respiratory diseases	6.33%	5.30%
TROPICAL	Neglected tropical diseases and malaria	2.70%	0.58%
RESP_INFEC	Respiratory infections and tuberculosis	1.57%	6.74%
DIABETES	Diabetes and kidney diseases	1.39%	6.73%
SKIN	Skin and subcutaneous diseases	1.11%	0.50%
ENT_INF	Enteric infections	1.01%	0.51%
NEURO	Neurological disorders	0.99%	6.47%
SUBSTANCE	Substance use disorders	0.94%	0.93%
MUSCULO	Musculoskeletal disorders	0.79%	0.24%
OTHER_INFECT	Other infectious diseases	0.21%	0.45%
UNINTENT	Unintentional injuries	0.16%	3.00%
NUTRITION	Nutritional deficiencies	0.14%	0.53%
HIV_STI	HIV/AIDS and sexually transmitted infections	0.08%	1.30%
SELF_HARM	Self-harm and interpersonal violence	0.05%	6.31%
TRANSP_INJ	Transport injuries	0.02%	3.64%
MENTAL_DIS	Mental disorders	0.01%	0.00%

Using this mapping, we have created 21 covariates that reflect the log-transformed hospitalisation costs of each individual that were related to each cause. The number of covariates is similar to the number used for diagnoses grouped into ICD-10 chapters (22 covariates) and Charlson comorbidities (17 covariates). They are shown in Table 5.18, along with the proportion of hospitalisations related to each cause. We also include the proportion of deaths observed in Brazil in 2016[8] (the observation year of our study). The top five causes (neoplasms, maternal and neonatal disorders, digestive diseases, cardiovascular diseases and other non-communicable diseases) are responsible for 82.49%[9] of the hospitalisations of the policyholders in our data set in 2016. These five causes comprise a diverse set of conditions from congenital birth defects, such as down syndrome, to urinary diseases and male/female infertility. Cardiovascular diseases and neoplasms are also the top two causes of death in Brazil for the same year.

A total of 59 covariates are now used for the two-part, lasso and Cubist models, where 21 of them represent the GBD causes. The results are discussed in the following section.

5.7 Results of Predictive Models Using GBD Classification of Causes

5.7.1 Two-part Model

A total of 27 covariates have been selected for the first part of the two-part model. They are shown in Table 5.19. Overall, the covariates have coefficients that relate to the likelihood of claiming as expected. Among the GBD causes of death, digestive diseases and maternal and neonatal disorders are the only relevant ones. Their coefficients indicate that hospitalisations related to these causes contribute to reducing the likelihood of policyholders claiming in the following year. As explained previously, claims related to complications during maternity are unlikely to happen in consecutive years, thus, explaining the negative coefficient. Digestive diseases may be related to acute conditions that are treated or cured, and unlikely to demand prolonged treatment.

Other covariates that decrease the chances of medical claims are physiotherapy sessions, larger concentration of costs in one single month (propMAX_COST) and in hospitalisations (prop_INPATIENT). These effects have been observed for the models using ICD-10 chapters and Charlson comorbidities. Among the covariates that are related to increased likelihood of claims, we highlight, in accordance with previous models, the relevance of N_VISITS, N_EMERGENCY, propHOMECARE, TT_IMAGE and TT_LAB.

For the second part of the two-part model, 39 covariates have been chosen. Their coefficients are displayed in Table 5.20. Among the GBD causes of death, seven covariates are included as relevant. Digestive diseases and maternal and neonatal disorders, whose coefficients indicate lower yearly cost amounts, are already present in the first part of the model. However, this model judges as relevant the hospitalisation causes related to diabetes and kidney diseases (DIABETES), neoplasms, neurological disorders, respiratory infections and tuberculosis, and substance use disorders. Although the covariate DIABETES encompasses "diabetes and kidney disorders", the hospitalisations of the policyholders that fit into this category are more likely to be related to kidney diseases rather than diabetes. This is because hospitalisations are rarely caused by diabetes itself. Also, taking into consideration the relevance of renal conditions in the models fitted to Charlson comorbidities, we can conclude that this category is mostly related to hospitalisations caused by kidney diseases.

We can also observe that chemotherapy procedures largely contribute to increased costs in the following year, while physiotherapy, once again, is related to lower future costs. Covariates related to the number of doctor visits, emergency events and in-patient hospitalisations contribute to larger costs.

[8] Values based on GBD 2017 data available for public access via the GBD Results Tool at: `http://ghdx.healthdata.org/gbd-results-tool`

[9] This proportion is based on the hospitalisations for which one of the 21 causes was mapped.

Table 5.19: Coefficient estimates of first part of the two-part model - GBD

Covariate	**Estimate**	**Std. Error**	**z value**	**Pr**($>\|z\|$)
(Intercept)	-0.74493	0.01988	-37.462	<2e-16
AGE	-0.00311	0.00028	-11.134	<2e-16
GENDERFemale	0.03854	0.00949	4.060	4.90e-05
OWNERDependant	-0.10362	0.01051	-9.859	<2e-16
PLANUnregulated	-0.82949	0.02429	-34.152	<2e-16
PLANRestricted	0.21609	0.01232	17.537	<2e-16
CONTRAssociation	0.05249	0.01220	4.303	1.69e-05
CONTRIndividual	0.48318	0.01377	35.086	<2e-16
HOSP_ACCOMMPrivate	0.06422	0.01133	5.670	1.43e-08
COPAYYes	0.14524	0.01301	11.162	<2e-16
PRE_EXISTING	0.30433	0.02409	12.634	<2e-16
logCOST	0.20306	0.00758	26.794	<2e-16
logCOST_DEC	0.10623	0.00371	28.658	<2e-16
MONTHS_AVG	0.19187	0.00980	19.579	<2e-16
propMAX_COST	-0.00253	0.00039	-6.531	6.55e-11
propVISITS	0.00697	0.00039	18.023	<2e-16
propTESTS_THERAPIES	0.00498	0.00048	10.331	<2e-16
propEMERGENCY	0.00545	0.00042	12.897	<2e-16
propINPATIENT	-0.00276	0.00063	-4.360	1.30e-05
propHOMECARE	0.03078	0.00391	7.878	3.32e-15
N_VISITS	0.21984	0.00489	44.977	<2e-16
N_EMERGENCY	0.12841	0.00781	16.431	<2e-16
TT_SPECIFIC	0.01905	0.00466	4.090	4.32e-05
TT_PHYSIOTHERAPY	-0.03549	0.00805	-4.411	1.03e-05
TT_IMAGE	0.02644	0.00373	7.082	1.42e-12
TT_LAB	0.05903	0.00402	14.692	<2e-16
DIGESTIVE	-0.03296	0.00877	-3.758	1.72e-04
MATERNAL	-0.08186	0.00861	-9.513	<2e-16

Table 5.20: Coefficient estimates of the second part of two-part model - GBD

Covariate	**Estimate**	**Std. Error**	**t value**	**Pr(> $\|t\|$)**
(Intercept)	5.35200	0.00817	655.349	<2e-16
AGE	0.01102	0.00009	125.849	<2e-16
GENDERFemale	0.14680	0.00314	46.797	<2e-16
OWNERDependant	-0.01506	0.00334	-4.510	6.47e-06
PLANUnregulated	-0.21080	0.00894	-23.581	<2e-16
PLANRestricted	-0.34000	0.00413	-82.312	<2e-16
CONTRAssociation	0.05042	0.00392	12.854	<2e-16
CONTRIndividual	0.02825	0.00376	7.509	5.99e-14
HOSP_ACCOMMPrivate	0.05830	0.00350	16.679	<2e-16
COPAYYes	-0.06203	0.00389	-15.959	<2e-16
PRE_EXISTING	0.08039	0.00569	14.136	<2e-16
logCOST	0.14200	0.00166	85.524	<2e-16
logCOST_DEC	0.03905	0.00075	51.904	<2e-16
propMAX_COST	-0.00347	0.00008	-45.348	<2e-16
propLAST_THREE	0.00048	0.00006	7.527	5.20e-14
propVISITS	-0.00330	0.00009	-35.783	<2e-16
propTESTS_THERAPIES	-0.00302	0.00013	-23.599	<2e-16
propEMERGENCY	-0.00296	0.00011	-26.352	<2e-16
propINPATIENT	-0.00635	0.00020	-31.910	<2e-16
propHOMECARE	0.01357	0.00027	51.059	<2e-16
N_VISITS	0.04269	0.00051	84.375	<2e-16
N_EMERGENCY	0.04770	0.00112	42.792	<2e-16
N_INPATIENT	0.08205	0.00831	9.874	<2e-16
DAYS_INPATIENT	0.00976	0.00049	19.959	<2e-16
COST_ICU	0.01578	0.00192	8.205	2.3e-16
DAYS_ICU	0.01079	0.00185	5.842	5.17e-09
TT_ASSORTED	0.00821	0.00224	3.672	2.41e-04
TT_PHYSIOTHERAPY	-0.00502	0.00120	-4.200	2.67e-05
TT_PHONOAUDIOLOGY	0.02993	0.00397	7.536	4.84e-14
TT_HEMATOLOGY	0.03475	0.00947	3.670	2.42e-04
TT_IMAGE	0.02237	0.00085	26.323	<2e-16
TT_LAB	0.01966	0.00101	19.453	<2e-16
TT_CHEMOTHERAPY	0.17920	0.00584	30.671	<2e-16
DIABETES	0.08869	0.00710	12.493	<2e-16
DIGESTIVE	-0.01557	0.00184	-8.446	<2e-16
MATERNAL	-0.06944	0.00179	-38.697	<2e-16
NEOPLASM	0.01233	0.00229	5.375	7.66e-08
NEURO	0.04441	0.00755	5.881	4.08e-09
RESP_INFEC	0.02647	0.00604	4.386	1.16e-05
SUBSTANCE	0.06597	0.00894	7.376	1.63e-13

5.7.2 Lasso

The results of the 10-fold cross-validation for the RMSE of the lasso model can be seen in Figure 5.7. The value that minimises the average RMSE is $\lambda_L^{min} = 0.00010$, represented in the graph by the striped vertical line over $\log(\lambda_L^{min}) = -9.18708$. The one-standard error shrinkage factor is $\lambda_L^{se} = 0.01707$, shown in the graph by the dotted line over the value $\log(\lambda_L^{se}) = -4.07023$.

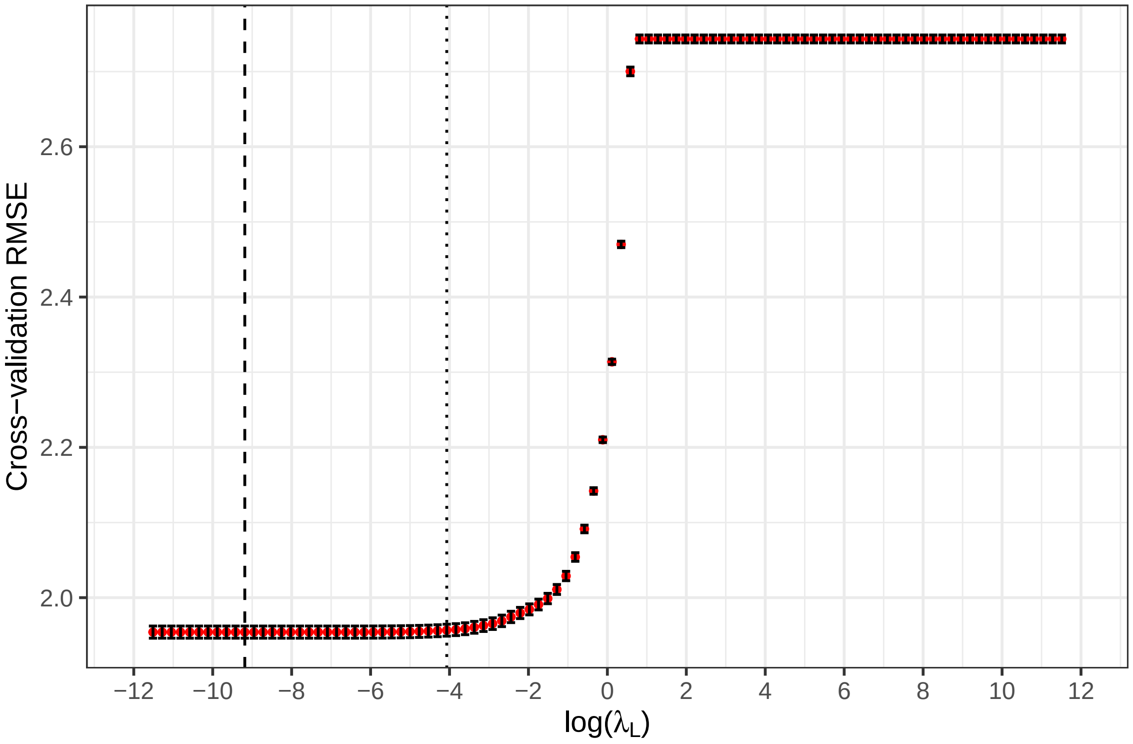


Figure 5.7: Change in the cross-validated RMSE with the increase of shrinkage factor - GBD.

Table 5.21 shows the coefficients of both lasso models fitted using the values for λ_L^{min} and λ_L^{se} determined via cross-validation. While the model tuned with the value that minimises the average RMSE has all covariates inputted, the more sparse model has only 25 covariates. Among them, maternal and neonatal disorders (MATERNAL) and diabetes and kidney diseases (DIABETES) are the only GBD causes relevant. Previous cost, represented by logCOST, continues to be the most powerful predictor while the other selected covariates are in line with the sparse lasso models fitted to ICD-10 chapters and Charlson.

Table 5.21: Coefficient estimates of lasso models fitted using λ_L^{min} and λ_L^{se} - GBD

Covariate	$\lambda_L^{min} = 0.00001$	$\lambda_L^{se} = 0.01707$	**Covariate**	$\lambda_L^{min} = 0.00001$	$\lambda_L^{se} = 0.01707$
(Intercept)	1.56469	1.68895	TT_PHONOAUDIOLOGY	0.01947	
AGE	0.00541	0.00472	TT_HEMATOLOGY	0.00695	
GENDER	0.13077	0.10542	TT_IMAGE	0.02978	0.02252
OWNER	-0.08736	-0.06222	TT_LAB	0.04865	0.04376
PLAN	-0.08192	-0.06905	TT_MONITORING	0.00406	
CONTR	0.07718	0.06942	TT_NUCLEAR	-0.03145	
HOSP_ACCOMM	0.04036		TT_CHEMOTHERAPY	0.09055	0.01072
COPAY	0.12999	0.05901	TT_RADIOTHERAPY	-0.04740	
PRE_EXISTING	0.18777	0.14869	CARDIO	0.00465	
logCOST	0.54642	0.54341	CRD	-0.00088	
logCOST_DEC	0.05345	0.05536	DIABETES	0.07635	0.00047
MONTHS_AVG	0.02867	0.08370	DIGESTIVE	-0.01708	
propMAX_COST	-0.00652	-0.00285	ENT_INF	0.02834	
propLAST_THREE	0.00113	0.00067	HIV_STI	0.02740	
propVISITS	0.00786	0.00384	MATERNAL	-0.08214	-0.06513
propTESTS_THERAPIES	0.00284		MENTAL_DIS	0.01193	
propEMERGENCY	0.00429		MUSCULO	-0.01505	
propINPATIENT	-0.00678	-0.00684	NEOPLASM	0.01820	
propHOMECARE	0.01435	0.00952	NEURO	0.03122	
N_VISITS	0.02591	0.02688	NON_COM	-0.00536	
N_EMERGENCY	0.03129	0.02935	NUTRITION	0.01034	
N_INPATIENT	0.10814		OTHER_INFECT	0.00456	
DAYS_INPATIENT	0.00432	0.00067	RESP_INFEC	0.03046	
COST_ICU	0.00365		SELF_HARM	0.10540	
DAYS_ICU	-0.00338		SKIN	0.00863	
TT_ASSORTED	0.01730		SUBSTANCE	0.04210	
TT_ENDOSCOPY	-0.00142		TRANSP_INJ	-0.17560	
TT_SPECIFIC	0.00751	0.00247	TROPICAL	-0.00262	
TT_PHYSIOTHERAPY	-0.01307		UNINTENT	0.00471	

5.7.3 Cubist

The results of the 10-fold cross-validation conducted to define the optimal number of rules for the Cubist model using GBD data are displayed in Figure 5.8. The number of rules that minimises the average RMSE is eight. This is exactly the number of rules that minimises the RMSE for the Charlson data. From that point onwards, the average RMSE increases with the number of rules, until 25 rules, when it drops and remains approximately constant for larger numbers of rules. The model with three rules is the choice according to the one-standard error criterion. Its output can be seen in Appendix G.1. The full output of the model with eight rules is found in Appendix G.2.

The conditions of Cubist with eight rules are listed in Table 5.22. In line with what was seen for models using ICD-10 and Charlson data, only AGE, logCOST and logCOST_DEC remain as covariates used for defining the homogeneous groups. Many groups are similar to the ones defined for the other data sets. Groups four, five and six of this model are the same for the Charlson data (Table 5.15) and are equivalent to groups three, four and five, respectively, from the ICD-10 chapters (Table 5.7). The covariate used to define the group with largest average costs in the following year is logCOST_DEC, showing the relevance of this information in determining the amount of future claims of policyholders.

The heat-map in Figure 5.9 highlights the most important covariates used in the linear models of the eight groups. Not surprisingly, logCOST is the most relevant, being present in all linear models. propHOMECARE and propINPATIENT maintain their relevance, as well as N_VISITS and logCOST_DEC. Maternal and neonatal disorders are relevant in two of the models, while neoplasms is present in one model. Thus, although relevant among the GBD causes, they are less important than other covariates that describe the medical events and procedures.

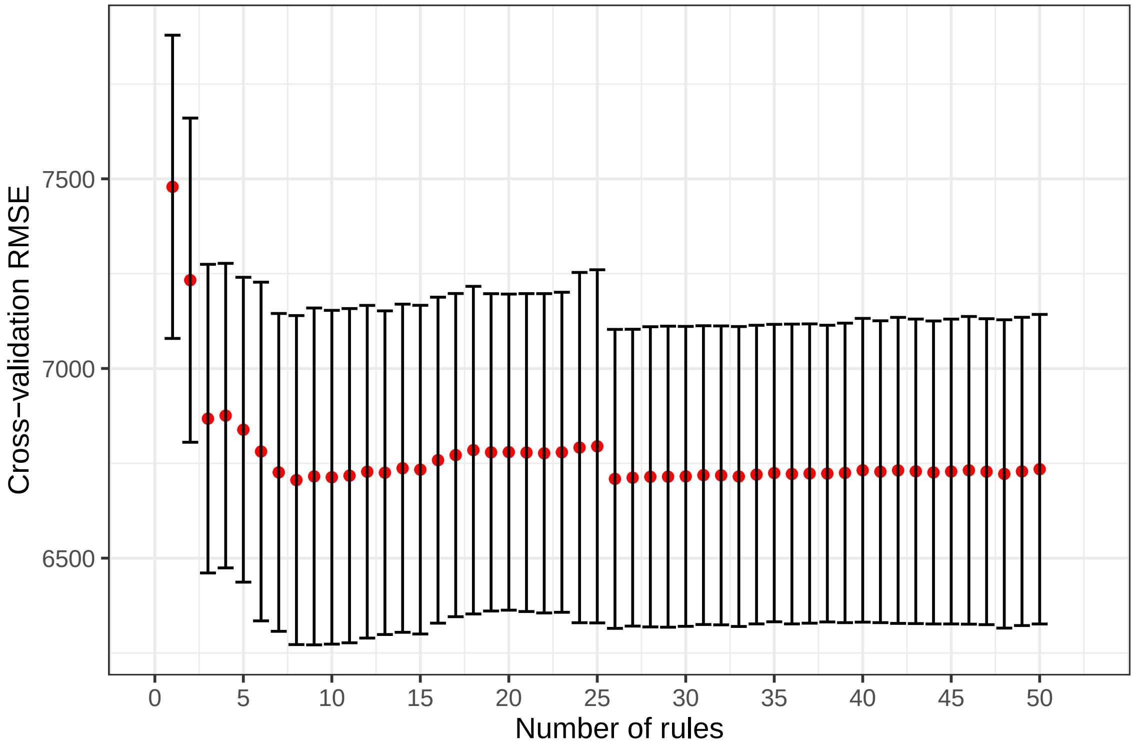


Figure 5.8: Change in the cross-validated RMSE with the increase of the number of rules - GBD.

Table 5.22: Conditions of Cubist with eight rules - GBD

Rule	**Condition**	**Number of observations**	**Average cost in prediction year**
1	AGE <= 27 logCOST <= 8.87	193,389	792.16
2	logCOST_DEC <= 4.81	494,812	1,332.73
3	logCOST <= 8.87	573,348	1,600.87
4	logCOST >8.87 logCOST <= 10.51	20,212	7,971.62
5	logCOST_DEC <= 5.39 logCOST >10.51	754	12,380.19
6	logCOST_DEC >5.61 logCOST >8.87 logCOST <= 10.51	7,849	12,946.39
7	logCOST_DEC >5.39 logCOST_DEC <= 9.64 logCOST >10.51	1,455	48,753.02
8	logCOST_DEC >9.64	523	65,534.27

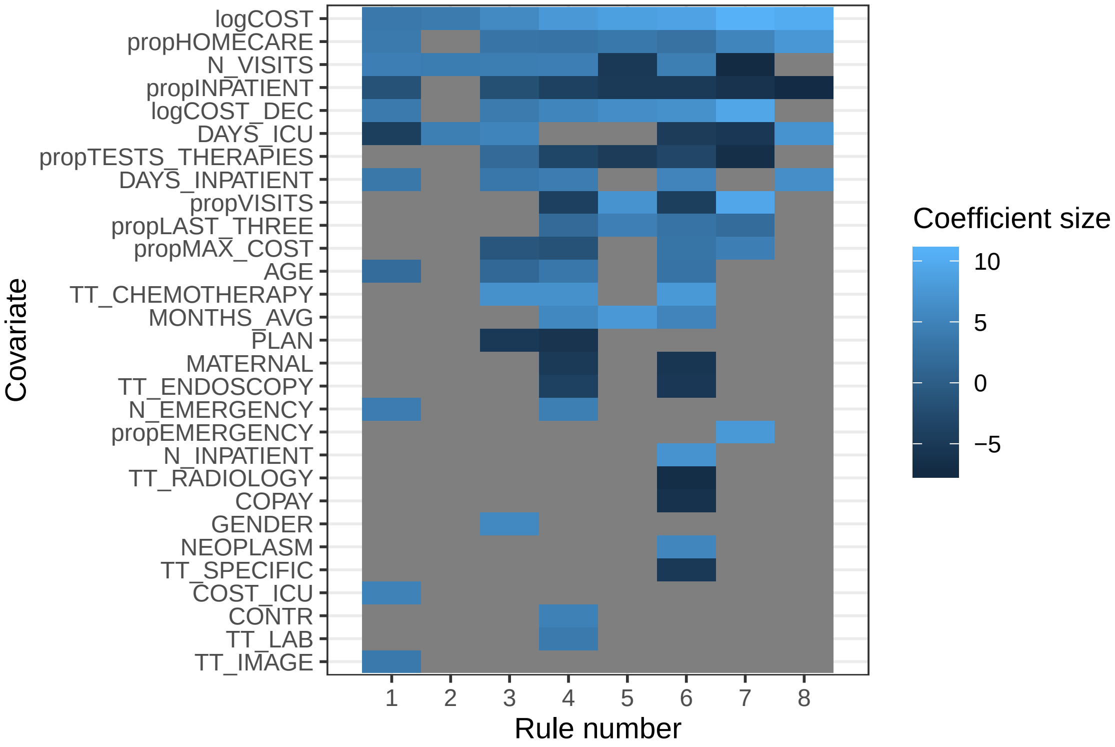


Figure 5.9: Heat-map of coefficients from the linear models in the Cubist model with eight rules - GBD.

Goodness-of-fit and Prediction Accuracy Measures

In terms of goodness-of-fit (Table 5.23) and prediction accuracy (Table 5.24), the models perform as previously observed for ICD-10 chapters and Charlson comorbidities, with one important difference. Cubist with eight rules is not only superior according to R^2, MAE and MAPE (the two-part model still wins in terms of Spearman correlation coefficient and Gini statistic), but its $\text{qMAD}_{0.95}$ is practically the same as the one from lasso λ_L^{se}, the lowest (therefore, best) possible. This has not been the case for Cubist models fitted previously, as we observed a worse fitting for the lower cost individuals as a trade-off for a better fitting for high cost individuals.

A comparison of MAPE measures, based on the test sample, by cost decile between lasso (λ_L^{se}) and Cubist with eight rules is found in Table 5.25. Although lasso fits the first cost decile better than Cubist, the difference is not as large as observed for ICD-10 chapters (Table 5.10). Cubist provides better fit for individuals in cost deciles 20% and 30%, while lasso fits better for cost deciles 30% to 80%. The top two cost deciles are fitted significantly better by Cubist, highlighting the advantage of this model over lasso for predicting values of high cost policyholders.

Table 5.23: Goodness-of-fit measures of two-part, lasso and Cubist models for GBD causes of death

Goodness-of-fit measures	Two-part	Lasso λ_L^{min}	Lasso λ_L^{se}	Cubist 3 rules	Cubist 8 rules
R^2	0.12783	0.15456	0.12914	0.28632	0.31242
MAE	1,724	1,481	1,480	1,485	1,435
MAPE	0.86333	0.74278	0.74226	0.74430	0.71953
$\text{qMAD}_{0.95}$	857	609	601	655	603
Sp.Cor.	0.71554	0.70855	0.70682	0.67311	0.69810
Gini	-3.83784	-14.37934	-10.59581	-10.59171	-15.19607

Table 5.24: Prediction accuracy measures of two-part, lasso and Cubist models for GBD causes of death

Prediction accuracy measures	**Two-part**	**Lasso**		**Cubist**	
		λ_L^{min}	λ_L^{se}	**3 rules**	**8 rules**
R^2	0.14970	0.15519	0.11829	0.27396	0.28263
MAE	1,725	1,477	1,479	1,483	1,435
MAPE	0.86499	0.74321	0.74411	0.74618	0.72214
$qMAD_{0.95}$	857	607	599	654	600
Sp.Cor.	0.71323	0.70684	0.70528	0.67056	0.69737
Gini	-3.24508	-13.55878	-9.78980	-9.70787	-14.48046

Table 5.25: MAPE per cost decile of lasso and Cubist models for GBD conditions, based on the test sample

Cost decile	**Number of policyholders**	**Average cost**	**Lasso (λ_L^{se})**	**Cubist 8 rules**
10%	125,657	314.74	**0.82883**	0.84370
20%	30,312	1,304.74	0.53269	**0.46439**
30%	17,416	2,270.87	0.53205	**0.51750**
40%	10,476	3,775.05	**0.58911**	0.63501
50%	6,551	6,037.53	**0.66928**	0.73239
60%	4,193	9,431.60	**0.74838**	0.78936
70%	2,357	16,773.96	**0.82655**	0.85520
80%	1,221	32,404.36	**0.88557**	0.88699
90%	563	70,056.85	0.90623	**0.78682**
100%	216	183,601.87	0.92231	**0.71024**

5.8 Final Considerations

After translating many medical features from claims data into applicable covariates for predictive models, we have assessed and compared the performance of two-part, lasso and Cubist and explored the usefulness of these expanded sets of explanatory variables for the prediction of future costs.

The two-part model, the representative of the traditional methods, does not benefit from the inclusion of extra covariates, and offers only poor goodness-of-fit and poor accuracy measures when compared with the machine learning methods. The shrinkage mechanism of lasso reveals that it is possible for a more parsimonious model with increased accuracy to be attained. However, this model loses its accuracy for outliers (policyholders in the top cost deciles), which is where Cubist outperforms the other models tested.

We have demonstrated the necessity of the grouping of medical items present in the invoices sent to insurers in order to reduce the number of covariates and make sense of the data. The models created show the relevance of the proportion of costs in each type of event. Higher concentrations of in-patient hospitalisations are usually reflected in lower costs in the following year, while home care indicates larger costs. It is an example of how the models indirectly capture the health care needs of policyholders: acute (hospitalisations, which have short duration and unlikely to occur in the future) versus chronic (home care services, which have longer duration and aim to improving the patient's life given the presence of a certain condition).

The frequency of doctor visits, emergency events and in-patient hospitalisations are also relevant and indicate larger future costs. In terms of the type of tests and therapies, image and lab procedures are present in all models, most probably because they represent the most common tests requested by physicians when investigating a condition of the patients. In other words, encounters with doctors are usually followed by lab and/or image tests, thus, they are all relevant as a whole. Additionally, we have found that physiotherapy sessions are associated with reduced costs in the future while chemotherapy is related to higher costs.

Regarding the diagnosis, we have tested three different classification methods: ICD-10 chapters, Charlson comorbidities and causes from the Global Burden of Disease. In terms of accuracy, the three methods provide similar results for all models. This exposes, to a certain degree, a limitation of our data: we only have diagnosis coming from in-patient hospitalisations (which are relatively infrequent when compared to emergency events and visits of the doctor). Thus, we are only able to feed the models data related to the diagnosis that caused hospitalisations, which only partially reveals the health conditions of the policyholder. For this reason, most of the explanation of future costs (considering only detailed medical information) is driven by covariates related to medical procedures and event types. Better results may be achieved for data sets that also include information of ICD-10 codes for out-patient procedures.

Nevertheless, the diagnoses reveal that hospitalisations related to pregnancy (chapter XV from ICD-10 or maternal and neonatal from GBD) indicate lower future costs, while renal conditions (renal diseases according to Charlson comorbidities; diabetes and kidney diseases according to GBD) contribute to increased costs. From the model outputs, the classification according to GBD causes offer broader coverage of the codes available (73% of hospitalisations versus 8.8% from Charlson). Furthermore, because of the international scope of the Global Burden of Diseases study, it provides an opportunity for insurers to compare the trends of their group of policyholders relative to the population of a region and act on designing preventive programs for possible cost reduction.

The predictive power of the several medical covariates included in the models is not superior to that of the demographic, policy design and previous cost covariates. In fact, the importance of cost amount in the observation year is confirmed in many ways. The covariate logCOST is present in all of the models, with a relatively dominant coefficient size. Also, covariates derived from the monthly costs contribute to improving the performance of all models. In particular, the cost in the last month of the observation year (logCOST_DEC) is crucial for determining either the likelihood of claims or cost amounts.

The more detailed covariates contribute to a simplification in the number of rules needed by Cubist to reach the minimum cross-validated RMSE. Cubist fitted to aggregated costs needed 20 rules to minimise the error function, while the models that use detailed data needed either seven (ICD-10 codes) or eight (Charlson and GBD). Regardless of the number of rules and covariates, Cubist models use only AGE, logCOST and logCOST_DEC to create the more homogeneous groups of policyholders, with the more detailed covariates being present in the linear models that make the predictions.

We conclude that Cubist using the diagnosis classification according to the level 2 causes of the Global Burden of Diseases is the best among all the models tested. Not only does it provide the best R^2 measures in the training and test samples (a consequence of improved accuracy for large cost individuals), but also it shows comparable efficiency in relation to lasso for the bulk of the data (which is reflected in level of $qMAD_{0.95}$ measures).

Chapter 6

Conclusions and Further Steps

In our study, we show how to use the administrative claims data in order to fit traditional and more recently developed statistical learning methods that use linear models for predicting the future claim amounts in the policyholder level. By starting with the linear regression and generalized linear models, we explore, as an exercise, the challenges and obstacles involved in fitting such models. We are also able to compare their performance with other methods and conclude that there is a significant gain in terms of prediction accuracy when using alternative models such as lasso and Cubist.

The method with the best accuracy performance was Cubist, due to its ability to combine different techniques that culminate in improved goodness of fit and more accurate predictions. It uses a decision-tree approach to split the data into more homogeneous groups; it contains a variable selection algorithm that only retains the more relevant covariates in the model; it uses linear regression in order to make predictions, which facilitates interpretation. An advantage that arises from this combination of approaches is that Cubist is able to separately target groups of individuals with larger future costs, which is generally not well fitted by the other regression models.

Our investigations also confirm the relevance of previous claim costs as a predictor of the policyholder's future costs, which is also observed by Duncan et al. (2016) and Morid et al. (2017). Despite the inclusion of more granular medical information related to the procedures and treatments involved in the claims, total previous costs remains the most influential covariate in the predictive models. As seen in Table 6.1, the prediction accuracy in the test sample of the Cubist model that uses detailed medical information and Global Burden of Diseases 2017 for diagnosis only slightly improves the measures of the Cubist fitted to aggregated data. R^2 and Spearman correlation are approximately the same for both models while MAE and MAPE are reduced by 0.6% and $\mathrm{qMAD}_{0.95}$ decreases by 1.9%. Interestingly, the cost only Cubist model has a better Gini statistic, which is an advantage if the ranking of individuals is important for the insurer. However, this model needs more rules than the one using GBD categorisation in order to achieve such performance.

Table 6.1: Comparison of prediction accuracy (in the test sample) of Cubist models using detailed medical information (from GBD) and aggregated cost only

Prediction accuracy measures	**GBD (8 rules)**	**Cost only (11 rules)**
R^2	0.28263	0.28110
MAE	1,435	1,444
MAPE	0.72214	0.72636
$\mathrm{qMAD}_{0.95}$	600	612
Sp.Cor.	0.69737	0.69446
Gini	-14.48046	-10.21688

The relevance of total costs can be seen in the way Cubist mainly uses the previous total cost and cost in the last month of the observation year in the splitting conditions. It suggests that an insurer that only holds information regarding the total cost of the medical claims can still achieve reasonably accurate predictions. Also, it shows that it is a good idea to use the individual's claim costs in order to split the pool of policyholders into more homogeneous groups when forecasting claim amounts. Insurers usually analyse their policyholders by policy type or line of business, which might be necessary for many purposes such as evaluating the profitability of a specific insurance product, for instance. However, when making claim cost predictions in the individual level, Cubist shows that it is more efficient to use claim amount to stratify the policyholders.

Although less relevant, the covariates related to medical procedures of event types are able to indirectly reflect the health state of the policyholders and are useful in distinguishing between chronic and acute conditions. Acute conditions are captured by a larger proportion of costs in in-patient hospitalisations (propINPATIENT), which has coefficients with a negative sign in our models, indicating the prediction of lower costs in the next year. On the other hand, a larger proportion of costs in homecare services is an indication of a chronic condition that causes recurring costs in the future.

Medical therapies are also relevant in enabling the models to distinguish individuals who are at the recovery stage of a condition, and hence expected to have lower future costs, from those whose conditions are more complex and demand extra treatments in the future. In our models, the former is captured by physiotherapy sessions and the latter by chemotherapy sessions.

The diagnostic information is also important in order to define conditions of patients related to either lower or larger future costs. The categorisation according to the GBD 2017 level 2 causes of death performs best in terms of providing a better understanding of the policyholder's health state. According to this categorisation, we could observe that maternity causes for hospitalisation are related to lower costs in the future. This is an interesting result given that analysts usually overlook such information. The hospitalisations due to chronic conditions related to kidney diseases (included under the classification of DIABETES) are also significant for both the traditional and lasso models. Hospitalisations related to neoplasms are also relevant for predicting larger costs in the future.

Another important result from our investigation is that less complex models, in terms of number of parameters and degree of interpretability, can be fitted without a significant loss in accuracy. By using cross-validation to tune Cubist, we observe that the improvement in accuracy resulting from each added rule decreases quickly, and we reach a point where adding more rules only makes Cubist more complex without any benefit in terms of prediction accuracy. Our best model has eight rules, but this number may change slightly for different data sets or even different years of observation and prediction. The point here is to define, using cross-validation, a reasonable number of rules that balances interpretation and accuracy.

Nevertheless, the relevance of these covariates may vary for different countries or even different insurers in the same region. Legislation changes very often, and insurers may be forced to cover medical procedures or specific medications that were not previously covered. Furthermore, as technology advances, new treatments are developed and may be either a substitute or a complement to existing ones. Usually, newly developed technologies are very costly due to their novelty, limited number of qualified professionals able to use those technologies and the large cost of production of the equipment in the early stages of implementing them. These additions change the claims data over time, which modifies the results from the models.

Unusual events may also contribute to the change in the claim costs for an entire population in a given month or even year and affect the group of policyholders covered by the insurer. A natural disaster may incur a large number of accidents, which leads to a peak in emergency events and hospitalisations. The outbreak of a new virus may mean that several people need hospital care, depending on how severe the effects of the virus are for the individual's health. A current and very relevant example is the COVID-19 pandemic. According to European Centre for Disease Prevention and Control (2020), 42% of confirmed cases in 19 European countries have required hospitalisation. They also estimate that 2% of the confirmed cases represent severe hospitalisations, which are hospitalisations in ICU or that required respiratory support. These rates are likely to be different in the future due to the development of vaccines, treatments, medications and as immunity to the virus is acquired. Such unusual events are not likely to repeat yearly, which means that they need to be properly addressed before being taken into consideration for prediction.

In comparison to the available literature, our approach provides better accuracy than that which has been observed in many other studies. The linear models tested by Cumming et al. (2002) provide R^2 values that range from 0.099 to 0.154. Better measures are possible only by truncating claim amounts or by using the models to provide predictions for the same year of the fitting (observation year equal to prediction year). However, their models have only diagnosis and pharmacy covariates, no cost information is used, which may explain the large difference in accuracy in comparison to our results ($R^2 = 0.283$ for the Cubist model using GBD classification). They also report MAPE equal to 0.980 whereas we reached 0.722, i.e., our Cubist model provides a lower error in relation to the actual mean. The updated version of that investigation (Winkelman and Mehmud, 2007) achieves similar results. Their R^2 measures range from 0.124 to 0.213 for prospective models that do not use prior costs as a predictor. Including prior costs boosts the R^2 measures of their models, which range from 0.184 to 0.276, which is still slightly below the Cubist results.

Worse R^2 values than ours are also obtained by Bertsimas et al. (2008). Their classification trees provide $R^2 = 0.162$ and MAPE $= 0.894$ while their clustering algorithm has only $R^2 = 0.180$ and MAPE $= 0.788$. These results are not as good as ours, especially considering that the number of covariates used in their models is large (1,523 explanatory variables) and the final models are very complex to interpret.

R^2 measures from Duncan et al. (2016) range from 0.154 to 0.215, MAPE range from 0.717 (slightly better than ours) to 0.872 and Spearman Correlation coefficient range from 0.521 to 0.619, while our best Cubist model reached 0.697. Their models, however, are generally more complex than ours. For instance, each tree of their M5 model has five or six nodes, but they use committees (equation 4.13) of 25 trees to enhance the prediction accuracy. Their model with the best R^2 is a random forest, a method developed by Breiman (2001). It builds several uncorrelated regression trees by bootstrapping the training data set. The final outcome is the average of the predictions of each tree. The random forest used by Duncan et al. (2016) had 500 trees, which is significantly more complex than our single Cubist tree. Additionally, they have truncated the costs at $\$50,000$ in order to increase prediction accuracy of the models.

Morid et al. (2017) also use more complex methods and achieve better accuracy than ours. They achieve $R^2 = 0.460$ and MAPE $= 0.650$ by using Gradient Boosting Machine (Friedman, 2001). This method consists of improving the predictions of a model (usually a decision-tree) by sequentially fitting the residuals from the previous model and updating the response variable with those fittings before each interaction. The number of interactions is pre-defined but usually is in the thousands, which contributes to the "black-box" nature of this method. This shows that, in order to significantly improve the prediction accuracy of future claim costs, it is necessary to sacrifice the interpretability of the outcome.

An extension of the analysis that might be done is to assess how much the prediction estimates vary due to the change in the data points. It is important that results remain stable and do not change due to small deviations in the data. One approach for such a task is using a bootstrap method (Efron and Tibshirani, 1986) to randomly generate samples from the original data, with replacement. The bootstrapped samples have the same size as the original sample. For each sample, the same procedures for fitting each model are made. For instance, assume we want to use bootstrap to check how volatile the Cubist results are. For this, we generate 50 bootstrap samples, each with 795,009 observations (the size of our data set). We randomly split each of these samples into training and test samples (hence, we have 50 training samples and 50 test samples). Then we tune the number of rules using 10-fold cross-validation in each of the 50 training samples and choose the best model for each case. The next step is to apply the model in the respective test sample and generate the prediction accuracy measures. We would, therefore, end up with 50 R^2 measures, 50 MAE measures, 50 MAPE measures, 50 $\text{qMAD}_{0.95}$ measures and 50 Spearman Correlation coefficients. We can then calculate the average and standard deviation for each measure and judge whether they vary too much or not. The issue with this approach is the high computational power necessary and the large memory used. However, these problems can be overcome by using cloud solutions.

Further investigations can be done in several ways in order to extend the analyses made in this study or to try to improve the results achieved. One possibility is to explore whether different periods of observation and prediction affect the accuracy of predicted claim costs. This is similar to what Yang et al. (2017) does when comparing the prediction accuracy of different models that use information from the previous three, six, nine or 12 months. The aim of their study is to predict the costs of the top 10% individuals with largest medical claim costs from a pool of policyholders in the United States. Their conclusion is that, as the observation period expands, the accuracy of predictions improve, however, the improvement stalls at 12 months. Since we already use an observation period of 12 months in our study, one could investigate the relevance of information of claims occurred prior to that period for predicting future claim amounts. Also, one could explore how the accuracy of predictions improves when the prediction period is shortened from 12 months to six or three months. Still within the topic of the period for analysis, one could define observation and prediction years based on the policy year instead of calendar years and investigate whether this improves prediction accuracy.

Incorporating interactions between covariates may be another area of investigation. Particularly, the Cubist models tested in this study consistently suggest that previous claim costs and age may form a relevant interaction, even though the relationship between them may not be simple. As a suggestion for future work, one could examine ways of incorporating relevant variables and relationships found by more advanced learning models into more traditional statistical models. A relevant work in this matter is made by Wuthrich (2019), who studies ways of improving traditional generalised linear models by using neural network features.

There are also a few improvements that might be carried out in order to possibly increase the accuracy of the models or to provide better insights from the outcomes. As mentioned previously, the ICD-10 codes used in our study are reported when the patient is being admitted to the hospital. If the ICD codes were available at discharge, the classification of diagnosis might be improved and more complete, potentially resulting in better predictions. Furthermore, diagnosis information for outpatient procedures, emergency admissions and doctor visits should be used if available as they provide more information regarding the health status of the policyholder.

Another possibility would be to improve the estimation of the frequency of claims when using a frequency-severity model. We have used the monthly costs to approximate the number of claims, but an alternative way to do this would be counting the number of medical events for each event type. For instance, one could build six different frequency-severity models, one for each of the six categories used in our research: doctor visits, emergency events, tests and therapies, hospitalisations, homecare and other outpatient events. In this way, there would be six models for each policyholder and the prediction of total cost of a particular individual would be the sum of the six predictions. By modelling each event type separately, we have the chance to better capture its particularities. For example, the frequency of doctor visits tends to be much larger than hospitalisations, with a lower proportion of individuals with zero events. On the other hand, doctor visits tend to have lower costs, given the simplicity of the procedure and fewer resources involved than in hospitalisations. The idea is that the combination of the models may result in a more accurate estimate for the total medical cost experienced by each policyholder.

In this context, one might also take into consideration the dependencies between the different types of medical events. Such consideration is made by Erhardt and Czado (2012) when estimating yearly claim payments for individuals insured by a German health insurance company. Using copulas, they demonstrate how to estimate joint probability functions taking into consideration outpatient, inpatient and dental claims. They observe positive correlations between the different treatments and found that modelling the dependency led to more precise results.

Decomposing the total medical claim costs is in line with the approach chosen by many actuaries when modelling the total loss by peril (cause of the loss) in general insurance. By modelling each peril separately, the analyst can use covariates that are relevant in predicting a particular peril but not so useful for predicting others. In this way, there is a chance of building models tuned specifically for each peril, resulting in better fittings. Consequently, the predicted total loss, which is found by adding the predicted values of each peril, should improve in accuracy as well. Frees et al. (2010) and Frees et al. (2012) have followed this approach and modelled the total losses from homeowners insurance, sold as an all-risk policy[10], by perils such as fire, lightning, hail, liability and others.

In order to be consistent with other studies, we have limited our analyses to the group of policyholders whose policies have been in force throughout the whole two-year period comprising observation and prediction. One could extend these analyses by adding back the policyholders who have died during the prediction year, for instance. Since the majority of healthcare costs of individuals tends to be concentrated in the last months of their lives (Felder et al., 2000), there is a great deal of interest from medical insurers on estimating the expected amount to be spent in the future on procedures related to patients who are dying. However, adding these individuals into the analysis brings more variability to the data as they have different exposure to the rest of policyholders and possibly contribute to making the distribution of the claim costs have a heavier tail.

In terms of alternative statistical learning methods, many other approaches can be investigated. Within the regularisation methods, elastic nets (Zou and Hastie, 2005) could be an option. They combine the penalisation factors of ridge regression (equation 4.1) and lasso (equation 4.4) when estimating the model coefficients. Thus, cross-validation can be used to estimate the optimal level of both factors that, when combined, result in the lowest prediction error. According to Zou and Hastie (2005), this approach may produce more accurate predictions because it combines the effective regularisation of the ridge regression penalty with the variable selection nature of the lasso penalty.

Instead of using claim costs as the response variable, one could create categories for the payments

[10] An all-risk policy means that all risks are covered, except for those specifically excluded in the policy

made relative to a particular claim and fit regression trees (Breiman et al., 1984) to predict the number of payments that will be paid in the future. This approach is followed by Wüthrich (2018) who creates categories that indicate, for a particular year, whether there is a payment or not and whether the claim is still open (i.e. more payments are needed) or already closed. Wüthrich (2018) then uses these results to estimate the number of payments needed in future years relative to claims that have reported but not yet settled, and for claims that have occurred but have not been reported. Insurers can use these as inputs for setting appropriate reserve amounts.

Approaches that have been tried recently in contexts other than medical claims can be adapted into the health care world. For instance, Pesantez-Narvaez et al. (2019) compares logistic regression to XGBoost for predicting the occurrence of motor insurance claims using Telematics data. Although they conclude that logistic regression offer better accuracy than a simpler version of the XGBoost model, it is possible to tune the latter in such a way that various interactions are carried out in order to increase prediction accuracy. However, there is a loss in interpretation of the model output when this approach is used.

Non-parametric methods are increasingly being used for forecasting health care claims. An example of a recent study is Richardson and Hartman (2018) which compares a Bayesian non-parametric model to a linear regression fitted to claim costs related to conjunctivitis and lung transplants. They conclude that the non-parametric method provides better accuracy due to its flexibility when dealing with skewness, heavy tails, outliers and zero claims. Flexibility also leads to improved prediction results in the study conducted by Piontkowski (2020). They use a non-parametric approach to fit a smoothed function into the inpatient costs data of insured males of a German insurance company. The forecasts are carried out using a time-series model.

Generally speaking, highly flexible nonlinear methods tend to provide the best prediction performances among the models tested for forecasting medical claim costs. One of these methods is neural networks (Bishop et al., 1995), which was judged as the most accurate by Morid et al. (2017) and Yang et al. (2017). The name of the method comes from the fact that it mimics the way the human brain learns new content. It contains hidden units that generate several combinations of the covariates from the data and use these combinations as inputs for the predictions. There are several algorithms developed within the context of neural networks. The model used by Yang et al. (2017), for instance, is able to take into consideration not only the medical procedures but also the time when they happened in order to make predictions.

The biggest issue with neural networks, which is previously discussed in this thesis, is the complete lack of model interpretability. Because it contains several layers of hidden units, it is impossible to determine how the inputs interact with each other and with the response variable (Kuhn and Johnson, 2013). However, a possible investigation that could follow our study would be to fit a neural network to our data and use the results as the upper limit for the highest prediction accuracy possible, which can only be reached by letting go of model interpretability. We can then assess how the accuracy of predictions from Cubist compares to the one from neural networks. Perhaps the accuracy measures provided by Cubist and neural networks are similar, which means that much more work is needed in order to understand the complexities of medical claims data and effectively use them to produce highly accurate cost predictions. If the improvement in accuracy from neural networks is, in fact, substantial then efforts should be focused on developing more interpretable methods that are able to approximate the maximum accuracy of predictions currently achieved only by ”black-box” methods.

References

Agency for Healthcare Research and Quality (2019), 'MEPS Survey Background', `https://meps.ahrq.gov/mepsweb/about_meps/survey_back.jsp`. [Online; accessed 13-June-2019].

Arjas, E. (1989), 'The claims reserving problem in non-life insurance: Some structural ideas', *ASTIN Bulletin: The Journal of the IAA* **19**(2), 139–152.

Ash, A. S., Ellis, R. P., Pope, G. C., Ayanian, J. Z., Bates, D. W., Burstin, H., Iezzoni, L. I., MacKay, E. and Yu, W. (2000), 'Using diagnoses to describe populations and predict costs', *Health Care Financing Review* **21**(3), 7.

Basu, A. and Manning, W. G. (2009), 'Issues for the next generation of health care cost analyses', *Medical Care* pp. S109–S114.

Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S. and Wang, G. (2008), 'Algorithmic prediction of health-care costs', *Operations Research* **56**(6), 1382–1392.

Bishop, C. M. et al. (1995), *Neural networks for pattern recognition*, Oxford University Press.

Bornhuetter, R. L. and Ferguson, R. E. (1972), The actuary and IBNR, *in* 'Proceedings of the casualty actuarial society', Vol. 59, pp. 181–195.

Box, G. E. (1979), Robustness in the strategy of scientific model building, *in* 'Robustness in statistics', Elsevier, pp. 201–236.

Box, G. E. and Cox, D. R. (1964), 'An analysis of transformations', *Journal of the Royal Statistical Society: Series B (Methodological)* **26**(2), 211–243.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984), *Classification and regression trees*, CRC press.

Brockman, M. J. and Wright, T. (1992), 'Statistical motor rating: making effective use of your data', *Journal of the Institute of Actuaries* **119**(3), 457–543.

Buntin, M. B. and Zaslavsky, A. M. (2004), 'Too much ado about two-part models and transformation?: Comparing methods of modeling medicare expenditures', *Journal of health economics* **23**(3), 525–542.

Cai, H., Nguyen, T. T., Li, Y., Zheng, V. W., Chen, B., Cong, G. and Li, X. (2019), 'Modeling marked temporal point process using multi-relation structure rnn', *Cognitive Computation* pp. 1–14.

Charlson, M. E., Pompei, P., Ales, K. L. and MacKenzie, C. R. (1987), 'A new method of classifying prognostic comorbidity in longitudinal studies: development and validation', *Journal of Chronic Diseases* **40**(5), 373–383.

Cumming, R. B., Knutson, D., Cameron, B. A. and Derrick, B. (2002), 'A comparative analysis of claims-based methods of health risk assessment for commercial populations', *Final report to the Society of Actuaries* .

Cummins, J. D., Dionne, G., McDonald, J. B. and Pritchett, B. M. (1990), 'Applications of the gb2 family of distributions in modeling insurance loss processes', *Insurance: Mathematics and Economics* **9**(4), 257–272.

Deyo, R. A., Cherkin, D. C. and Ciol, M. A. (1992), ‘Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases’, *Journal of Clinical Epidemiology* **45**(6), 613–619.

D’Hoore, W., Bouckaert, A. and Tilquin, C. (1996), ‘Practical considerations on the use of the Charlson comorbidity index with administrative data bases’, *Journal of Clinical Epidemiology* **49**(12), 1429–1433.

Dicker, D., Nguyen, G., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J. et al. (2018), ‘Global, regional, and national age-sex-specific mortality and life expectancy, 1950–2017: a systematic analysis for the global burden of disease study 2017’, *The Lancet* **392**(10159), 1684–1735.

Diehr, P., Yanez, D., Ash, A., Hornbrook, M. and Lin, D. (1999), ‘Methods for analyzing health care utilization and costs’, *Annual Review of Public Health* **20**(1), 125–144.

Dimitrova, D. S., Ignatov, Z. G. and Kaishev, V. K. (2017), ‘On the first crossing of two boundaries by an order statistics risk process’, *Risks* **5**(3), 43.

Dimitrova, D. S., Ignatov, Z. G., Kaishev, V. K. and Tan, S. (2020), ‘On double-boundary non-crossing probability for a class of compound processes with applications’, *European Journal of Operational Research* **282**(2), 602–613.
URL: *https://www.sciencedirect.com/science/article/pii/S037722171930815X*

Dove, H. G., Duncan, I. and Robb, A. (2003), ‘A prediction model for targeting low-cost, high-risk members of managed care organizations’, *Am J Manag Care* **9**(5), 381–9.

Duan, N. (1983), ‘Smearing estimate: a nonparametric retransformation method’, *Journal of the American Statistical Association* **78**(383), 605–610.

Duan, N., Manning, W. G., Morris, C. N. and Newhouse, J. P. (1983), ‘A comparison of alternative models for the demand for medical care’, *Journal of Business & Economic Statistics* **1**(2), 115–126.

Duncan, I. G. (2011), *Healthcare risk adjustment and predictive modeling*, Actex Publications.

Duncan, I., Loginov, M. and Ludkovski, M. (2016), ‘Testing alternative regression frameworks for predictive modeling of health care costs’, *North American Actuarial Journal* **20**(1), 65–87.

Efron, B. and Tibshirani, R. (1986), ‘Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy’, *Statistical Science* pp. 54–75.

Erhardt, V. and Czado, C. (2012), ‘Modeling dependent yearly claim totals including zero claims in private health insurance’, *Scandinavian Actuarial Journal* **2012**(2), 106–129.

European Centre for Disease Prevention and Control (2020), Rapid Risk Assessment: Coronavirus disease 2019 (COVID-19) in the EU/EEA and the UK– ninth update, 23 April 2020, Technical report, ECDC, Stockholm.

Felder, S., Meier, M. and Schmitt, H. (2000), ‘Health care expenditure in the last months of life’, *Journal of Health Economics* **19**(5), 679–695.

Frees, E. W. (2009), *Regression modeling with actuarial and financial applications*, Cambridge University Press.

Frees, E. W. (2018), ‘Loss data analytics’, *arXiv preprint arXiv:1808.06718* .
URL: *https://openacttexts.github.io/Loss-Data-Analytics/*

Frees, E. W., Derrig, R. A. and Meyers, G. (2014), *Predictive modeling applications in actuarial science*, Vol. 1, Cambridge University Press.

Frees, E. W., Gao, J. and Rosenberg, M. A. (2011), ‘Predicting the frequency and amount of health care expenditures’, *North American Actuarial Journal* **15**(3), 377–392.

Frees, E. W. J., Meyers, G. and Cummings, A. D. (2010), ‘Dependent multi-peril ratemaking models’, *ASTIN Bulletin: The Journal of the IAA* **40**(2), 699–726.

Frees, E. W., Jin, X. and Lin, X. (2013), 'Actuarial applications of multivariate two-part regression models', *Annals of Actuarial Science* **7**(2), 258–287.

Frees, E. W., Meyers, G. and Cummings, A. D. (2011), 'Summarizing insurance scores using a Gini index', *Journal of the American Statistical Association* **106**(495), 1085–1098.

Frees, E. W., Meyers, G. and Cummings, A. D. (2012), Predictive modeling of multi-peril homeowners insurance, *in* 'Casualty Actuarial Society E-Forum, Winter 2011 Volume 2'.

Frees, E. W., Meyers, G. and Cummings, A. D. (2014), 'Insurance ratemaking and a Gini index', *Journal of Risk and Insurance* **81**(2), 335–366.

Friedman, J. H. (2001), 'Greedy function approximation: a gradient boosting machine', *Annals of Statistics* pp. 1189–1232.

Friedman, J., Hastie, T. and Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics New York.

Furman, E. and Landsman, Z. (2010), 'Multivariate tweedie distributions and some related capital-at-risk analyses', *Insurance: Mathematics and Economics* **46**(2), 351–361.

GBD 2017 Risk Factor Collaborators (2018), 'Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990-2017: a systematic analysis for the global burden of disease study 2017.', *The Lancet* .

Getzen, T. E. (2000), 'Forecasting health expenditures: short, medium and long (long) term', *Journal of Health Care Finance* **26**(3), 56–72.

Global Burden of Disease Collaborative Network (2018), 'Global Burden of Disease Study 2017 (GBD 2017) Causes of Death and Nonfatal Causes Mapped to ICD Codes', `http://ghdx.healthdata.org/record/ihme-data/gbd-2017-cause-icd-code-mappings`.

Granger, C. W. and Newbold, P. (1976), 'Forecasting transformed series', *Journal of the Royal Statistical Society: Series B (Methodological)* **38**(2), 189–203.

Groves, P., Kayyali, B., Knott, D. and Kuiken, S. V. (2016), *The 'big data' revolution in healthcare: Accelerating value and innovation*, Center for US Health System Reform Business Technology Office.

Guo, X., Gandy, W., Coberley, C., Pope, J., Rula, E. and Wells, A. (2015), 'Predicting health care cost transitions using a multidimensional adaptive prediction process', *Population Health Management* **18**(4), 290–299.

Halder, A., Mohammed, S., Chen, K. and Dey, D. K. (2021), 'Spatial Tweedie exponential dispersion models: an application to insurance rate-making', *Scandinavian Actuarial Journal* pp. 1–20.

Halfon, P., Eggli, Y., van Melle, G., Chevalier, J., Wasserfallen, J.-B. and Burnand, B. (2002), 'Measuring potentially avoidable hospital readmissions', *Journal of Clinical Epidemiology* **55**(6), 573–587.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015), *Statistical learning with sparsity: the lasso and generalizations*, CRC press.

Hocking, R. R. (2003), *Methods and applications of linear models: regression and the analysis of variance*, John Wiley & Sons.

Hoerl, A. E. and Kennard, R. W. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**(1), 55–67.

Hornik, K., Buchta, C. and Zeileis, A. (2009), 'Open-source machine learning: R meets weka', *Computational Statistics* **24**(2), 225–232.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An introduction to statistical learning*, Vol. 112, Springer.

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A. et al. (2018), 'Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017', *The Lancet* **392**(10159), 1789–1858.

Jones, A. M. (2000), Health econometrics, *in* 'Handbook of Health Economics', Vol. 1, Elsevier, pp. 265–344.

Jørgensen, B. and Paes De Souza, M. C. (1994), 'Fitting tweedie's compound poisson model to insurance claims data', *Scandinavian Actuarial Journal* **1994**(1), 69–93.

Klugman, S. A., Panjer, H. H. and Willmot, G. E. (2012), *Loss models: from data to decisions*, Vol. 715, John Wiley & Sons.

Kuhn, M. and Johnson, K. (2013), *Applied predictive modeling*, Vol. 26, Springer.

Kuhn, M. and Quinlan, R. (2018), *Cubist: Rule- And Instance-Based Regression Modeling.* R package version 0.2.2.
URL: *https://CRAN.R-project.org/package=Cubist*

Kyu, H. H., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A. et al. (2018), 'Global, regional, and national disability-adjusted life-years (dalys) for 359 diseases and injuries and healthy life expectancy (hale) for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017', *The Lancet* **392**(10159), 1859–1922.

Lieberthal, R. D., Ai, J., Smith, S. D. and Wojciechowski, R. L. (2018), Examining predictive modeling based approaches to characterizing healthcare fraud, *in* '7th Annual Conference of the American Society of Health Economists', ASHECON.

Lorenz, M. O. (1905), 'Methods of measuring the concentration of wealth', *Publications of the American statistical association* **9**(70), 209–219.

Lu, F. and Boritz, J. E. (2005), Detecting fraud in health insurance data: Learning to model incomplete benford's law distributions, *in* 'European Conference on Machine Learning', Springer, pp. 633–640.

Macintyre, S., Hunt, K. and Sweeting, H. (1996), 'Gender differences in health: are things really as simple as they seem?', *Social science & medicine* **42**(4), 617–624.

Manning, W. G., Basu, A. and Mullahy, J. (2005), 'Generalized modeling approaches to risk adjustment of skewed outcomes data', *Journal of health economics* **24**(3), 465–488.

McDonald, J. B. (1984), 'Some generalized functions for the size distribution of income', **52**, 647–663.

Mihaylova, B., Briggs, A., O'hagan, A. and Thompson, S. G. (2011), 'Review of statistical methods for analysing healthcare resources and costs', *Health Economics* **20**(8), 897–916.

Morid, M. A., Kawamoto, K., Ault, T., Dorius, J. and Abdelrahman, S. (2017), Supervised learning methods for predicting healthcare costs: Systematic literature review and empirical evaluation, *in* 'AMIA Annual Symposium Proceedings', Vol. 2017, American Medical Informatics Association, p. 1312.

Mustard, C. A., Kaufert, P., Kozyrskyj, A. and Mayer, T. (1998), 'Sex differences in the use of health care services', *New England Journal of Medicine* **338**(23), 1678–1683.

Nelder, J. A. and Wedderburn, R. W. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society: Series A (General)* **135**(3), 370–384.

Newhouse, J. P., Manning, W. G., Keeler, E. B. and Sloss, E. M. (1989), 'Adjusting capitation rates using objective health measures and prior utilization', *Health Care Financing Review* **10**(3), 41.

Newman, M. C. (1993), 'Regression analysis of log-transformed data: Statistical bias and its correction', *Environmental Toxicology and Chemistry: An International Journal* **12**(6), 1129–1133.

Norberg, R. (1993), 'Prediction of outstanding liabilities in non-life insurance1', *ASTIN Bulletin: The Journal of the IAA* **23**(1), 95–115.

Ohlsson, E. and Johansson, B. (2010), *Non-life insurance pricing with generalized linear models*, Vol. 174, Springer.

Pauly, M. V. (1968), 'The economics of moral hazard: comment', *American Economic Review* **58**(3), 531–537.

Pesantez-Narvaez, J., Guillen, M. and Alcañiz, M. (2019), 'Predicting motor insurance claims using telematics data—xgboost versus logistic regression', *Risks* **7**(2), 70.

Piontkowski, J. (2020), 'Forecasting health expenses using a functional data model', *Annals of Actuarial Science* **14**(1), 72–82.

Pope, G. C., Kautter, J., Ellis, R. P., Ash, A. S., Ayanian, J. Z., Iezzoni, L. I., Ingber, M. J., Levy, J. M. and Robst, J. (2004), 'Risk adjustment of medicare capitation payments using the cms-hcc model', *Health Care Financing Review* **25**(4), 119.

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., Saunders, L. D., Beck, C. A., Feasby, T. E. and Ghali, W. A. (2005), 'Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data', *Medical Care* pp. 1130–1139.

Quinlan, J. R. (1987), 'Simplifying decision trees', *International Journal of Man-machine Studies* **27**(3), 221–234.

Quinlan, J. R. (1992), Learning with continuous classes, *in* '5th Australian joint conference on artificial intelligence', Vol. 92, World Scientific, pp. 343–348.

Quinlan, J. R. (1993), Combining instance-based and model-based learning, *in* 'Proceedings of the tenth international conference on machine learning', pp. 236–243.

Reid, D. (1978), 'Claim reserves in general insurance', *Journal of the Institute of Actuaries* **105**(3), 211–315.

Renshaw, A. E. and Verrall, R. J. (1998), 'A stochastic model underlying the chain-ladder technique', *British Actuarial Journal* **4**(4), 903–923.

Richardson, R. and Hartman, B. (2018), 'Bayesian nonparametric regression models for modeling and predicting healthcare claims', *Insurance: Mathematics and Economics* **83**, 1–8.

Rioux, J.-Y., Da Silva, A., Jones, H. and Saleh, H. (2019), The Use of Predictive Analytics in the Canadian Life Insurance Industry, Technical report, Society of Actuaries.

Rosen, A. K., Wang, F., Montez, M. E., Rakovski, C. C., Berlowitz, D. R. and Lucove, J. C. (2005), 'Identifying future high-healthcare users', *Disease Management & Health Outcomes* **13**(2), 117–127.

Roth, G. A., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A. et al. (2018), 'Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017', *The Lancet* **392**(10159), 1736–1788.

Rousseeuw, P. J. and Leroy, A. M. (2005), *Robust regression and outlier detection*, Vol. 589, John Wiley & Sons.

Schroeder, S. A., Showstack, J. A. and Roberts, H. E. (1979), 'Frequency and clinical description of high-cost patients in 17 acute-care hospitals', *New England Journal of Medicine* **300**(23), 1306–1309.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of statistics* **6**(2), 461–464.

Sena, G. R., Lima, T. P. F., Mello, M. J. G., Thuler, L. C. S. and Lima, J. T. O. (2019), 'Developing machine learning algorithms for the prediction of early death in elderly cancer patients: Usability study', *JMIR Cancer* **5**(2), e12163.

Stacy, E. W. et al. (1962), 'A generalization of the gamma distribution', *The Annals of mathematical statistics* **33**(3), 1187–1192.

Stone, M. (1974), ‘Cross-validatory choice and assessment of statistical predictions’, *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(2), 111–133.

Sun, J., Frees, E. W. and Rosenberg, M. A. (2008), ‘Heavy-tailed longitudinal data modeling using copulas’, *Insurance: Mathematics and Economics* **42**(2), 817–830.

Sundararajan, V., Henderson, T., Perry, C., Muggivan, A., Quan, H. and Ghali, W. A. (2004), ‘New icd-10 version of the charlson comorbidity index predicted in-hospital mortality’, *Journal of Clinical Epidemiology* **57**(12), 1288–1294.

Tibshirani, R. (1996), ‘Regression shrinkage and selection via the lasso’, *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Tweedie, M. C. (1984), An index which distinguishes between some important exponential families, *in* ‘Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference’, Vol. 579, pp. 579–604.

Venter, G. (1983), Transformed beta and gamma distributions and aggregate losses, *in* ‘Proceedings of the Casualty Actuarial Society’, Vol. 70, pp. 289–308.

Wang, Y. and Witten, I. H. (1996), Induction of model trees for predicting continuous classes. (Working paper 96/23). Hamilton, New Zealand: University of Waikato, Department of Computer Science.

Wasey, J. O. and R Core Team (2019), *icd: Comorbidity Calculations and Tools for ICD-9 and ICD-10 Codes*. R package version 4.0.6.
URL: *https://CRAN.R-project.org/package=icd*

Winkelman, R. and Mehmud, S. (2007), ‘A comparative analysis of claims-based tools for health risk assessment’, *Society of Actuaries* pp. 1–70.

World Health Organization (2004), *International statistical classification of diseases and related health problems: instruction manual*, Vol. 2, World Health Organization.

World Health Organization (2011), ‘International Statistical Classification of Diseases and Related Health Problems – 10th revision, Volume 2, edition 2010’, `https://www.who.int/classifications/icd/ICD10Volume2_en_2010.pdf?ua=1`. [Online; accessed 22-January-2019].

World Health Organization (2018), ‘International Classification of diseases, 11th Revision (ICD-11)’, `https://www.who.int/classifications/icd/en/`. [Online; accessed 18-January-2019].

Wüthrich, M. V. (2003), ‘Claims reserving using tweedie’s compound poisson model’, *ASTIN Bulletin: The Journal of the IAA* **33**(2), 331–346.

Wüthrich, M. V. (2018), ‘Machine learning in individual claims reserving’, *Scandinavian Actuarial Journal* **2018**(6), 465–480.

Wuthrich, M. V. (2019), ‘From generalized linear models to neural networks, and back’, *Available at SSRN 3491790* .

Wuthrich, M. V. and Buser, C. (2020), ‘Data analytics for non-life insurance pricing’, *Swiss Finance Institute Research Paper* (16-68).
URL: *https://ssrn.com/abstract=2870308*

Xie, Y., Schreier, G., Chang, D. C., Neubauer, S., Liu, Y., Redmond, S. J. and Lovell, N. H. (2015), ‘Predicting days in hospital using health insurance claims’, *IEEE Journal of Biomedical and Health Informatics* **19**(4), 1224–1233.

Yang, C., Delcher, C., Shenkman, E. and Ranka, S. (2017), Machine learning approaches for predicting high utilizers in health care, *in* ‘International Conference on Bioinformatics and Biomedical Engineering’, Springer, pp. 382–395.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.-F. and Hua, L. (2012), ‘Data mining in healthcare and biomedicine: a survey of the literature’, *Journal of Medical Systems* **36**(4), 2431–2448.

Zhang, B., Wright, A. A., Huskamp, H. A., Nilsson, M. E., Maciejewski, M. L., Earle, C. C., Block, S. D., Maciejewski, P. K. and Prigerson, H. G. (2009), 'Health care costs in the last week of life: associations with end-of-life conversations', *Archives of Internal Medicine* **169**(5), 480–488.

Zook, C. J. and Moore, F. D. (1980), 'High-cost users of medical care', *New England Journal of Medicine* **302**(18), 996–1002.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: series B (Statistical Methodology)* **67**(2), 301–320.

Appendix A

Analysis of Outliers That Inflate Residuals

Table A.1 shows the 10 largest residuals in the training sample from the linear regression model (second part of two-part model) using ICD-10 chapters as covariates for diagnosis. If the residual is negative, it means that the model is overestimating the real cost of the policyholder in 2017. The six policyholders with the largest residuals are highlighted in bold. Only six observations are highlighted because of the large gap between the sixth largest residual (Policyholder F) and the seventh largest (Policyholder G). As we can see, just six observations were responsible for 47.9% of the residual sum of squares, which is a substantial impact, particularly considering the training sample size (596,000 observations).

Table A.1: Largest residuals from initial attempt to fit a linear regression model to detailed medical data considering ICD-10 chapters

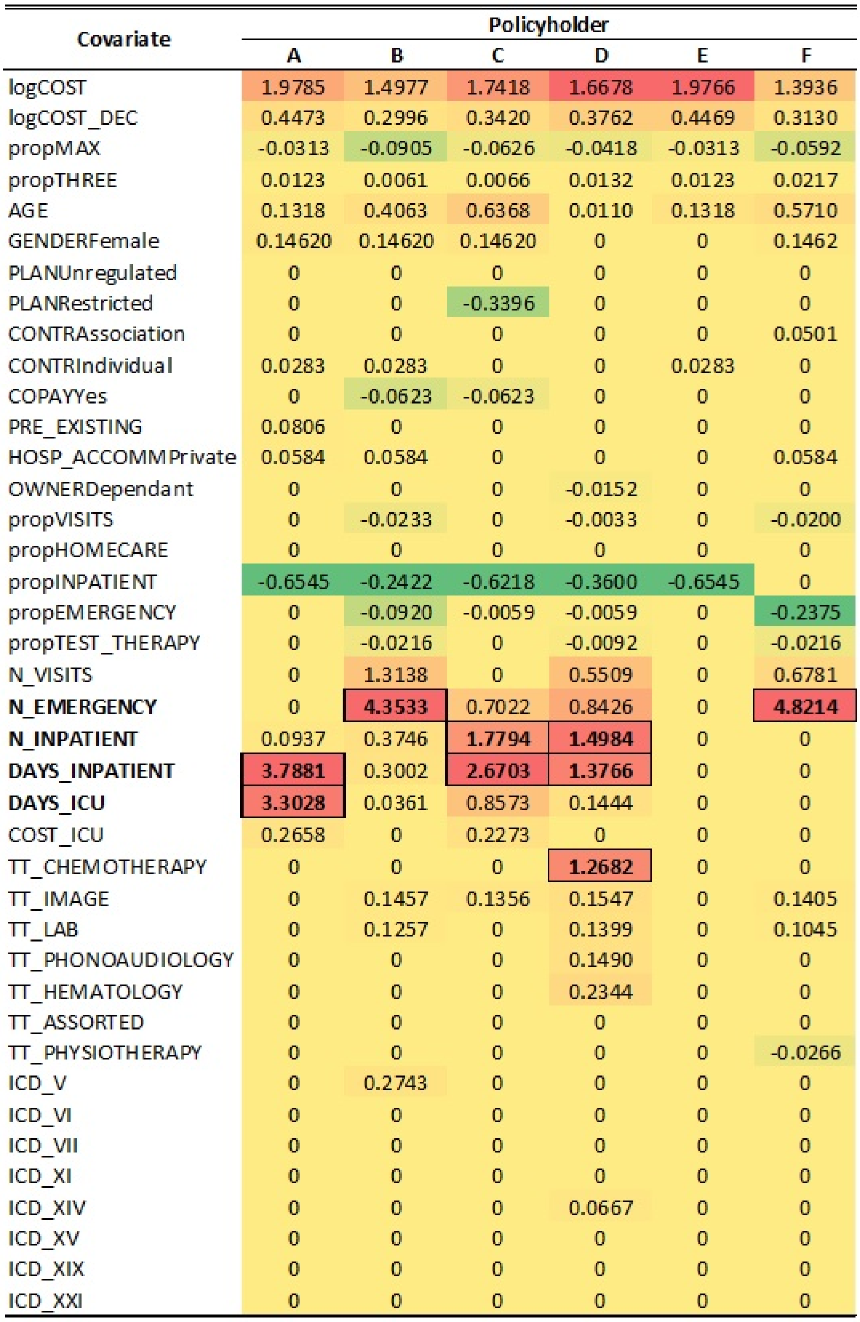
Policyholder	**Observed cost in 2017 (y_i)**	**Estimated cost in 2017 ($\hat{y}_i$)**	**Residual ($y_i - \hat{y}_i$)**	$(y_i - \hat{y}_i)^2$	**Proportion of res. sum of squares**	**Cumulative Proportion**
A	**1,129,801**	**5,578,676**	**-4,448,875**	**1.98e+13**	**30.9%**	**30.9%**
B	**8,212**	**2,472,362**	**-2,464,149**	**6.07e+12**	**9.5%**	**40.4%**
C	**94,033**	**1,251,443**	**-1,157,411**	**1.34e+12**	**2.1%**	**42.5%**
D	**5,168**	**1,138,002**	**-1,132,834**	**1.28e+12**	**2.0%**	**44.5%**
E	**1,117,519**	**2,429**	**1,115,090**	**1.24e+12**	**1.9%**	**46.4%**
F	**28,076**	**1,004,441**	**-976,365**	**9.53e+11**	**1.5%**	**47.9%**
G	545,652	4,062	541,591	2.93e+11	0.5%	48.4%
H	496,483	573	495,910	2.46e+11	0.4%	48.8%
I	490,516	11,970	478,546	2.29e+11	0.4%	49.1%
J	491,015	17,793	473,222	2.24e+11	0.3%	49.5%

The biggest problem with such large residuals is that they super-inflate the residual sum of squares, used in R^2, which makes it even larger than the total sum of squares. Consequently, R^2 becomes negative, misleading the conclusions about the model's goodness-of-fit:

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = \frac{6.40151e+13}{4.16375e+13} = -0.53744$$

By analysing the coefficients of the linear model that give rise to the residuals, we can start investigating the reasons why the model produced estimates that depart so much from the observed values. The table in Figure A.1 shows the value of each coefficient multiplied by the covariate value of each individual ($\hat{\beta}_k \cdot x_i$). Since it is a multiple linear model, we can see the contribution of each covariate in the final prediction for each policyholder.

The cells are coloured according to the size of the contribution of each covariate. Red cells contain the largest values and green cells contain the lowest values. The shades between green and red represent the values between the highest and lowest for that policyholder. We can see that the covariates with the highest values are number of emergency events (N_EMERGENCY), number of in-patient hospitalisations (N_INPATIENT), number of days in the hospital (DAYS_INPATIENT) and number of days in intensive care unit (DAYS_ICU). For policyholder C, cost with chemotherapy sessions is also relevant. Thus, it makes sense to focus on those covariates to check how these policyholders deviate from the group of all policyholders.



Covariate	Policyholder					
	A	B	C	D	E	F
logCOST	1.9785	1.4977	1.7418	1.6678	1.9766	1.3936
logCOST_DEC	0.4473	0.2996	0.3420	0.3762	0.4469	0.3130
propMAX	-0.0313	-0.0905	-0.0626	-0.0418	-0.0313	-0.0592
propTHREE	0.0123	0.0061	0.0066	0.0132	0.0123	0.0217
AGE	0.1318	0.4063	0.6368	0.0110	0.1318	0.5710
GENDERFemale	0.14620	0.14620	0.14620	0	0	0.1462
PLANUnregulated	0	0	0	0	0	0
PLANRestricted	0	0	-0.3396	0	0	0
CONTRAssociation	0	0	0	0	0	0.0501
CONTRIndividual	0.0283	0.0283	0	0	0.0283	0
COPAYYes	0	-0.0623	-0.0623	0	0	0
PRE_EXISTING	0.0806	0	0	0	0	0
HOSP_ACCOMMPrivate	0.0584	0.0584	0	0	0	0.0584
OWNERDependant	0	0	0	-0.0152	0	0
propVISITS	0	-0.0233	0	-0.0033	0	-0.0200
propHOMECARE	0	0	0	0	0	0
propINPATIENT	-0.6545	-0.2422	-0.6218	-0.3600	-0.6545	0
propEMERGENCY	0	-0.0920	-0.0059	-0.0059	0	-0.2375
propTEST_THERAPY	0	-0.0216	0	-0.0092	0	-0.0216
N_VISITS	0	1.3138	0	0.5509	0	0.6781
N_EMERGENCY	0	**4.3533**	0.7022	0.8426	0	**4.8214**
N_INPATIENT	0.0937	0.3746	**1.7794**	**1.4984**	0	0
DAYS_INPATIENT	**3.7881**	0.3002	**2.6703**	**1.3766**	0	0
DAYS_ICU	**3.3028**	0.0361	0.8573	0.1444	0	0
COST_ICU	0.2658	0	0.2273	0	0	0
TT_CHEMOTHERAPY	0	0	0	**1.2682**	0	0
TT_IMAGE	0	0.1457	0.1356	0.1547	0	0.1405
TT_LAB	0	0.1257	0	0.1399	0	0.1045
TT_PHONOAUDIOLOGY	0	0	0	0.1490	0	0
TT_HEMATOLOGY	0	0	0	0.2344	0	0
TT_ASSORTED	0	0	0	0	0	0
TT_PHYSIOTHERAPY	0	0	0	0	0	-0.0266
ICD_V	0	0.2743	0	0	0	0
ICD_VI	0	0	0	0	0	0
ICD_VII	0	0	0	0	0	0
ICD_XI	0	0	0	0	0	0
ICD_XIV	0	0	0	0.0667	0	0
ICD_XV	0	0	0	0	0	0
ICD_XIX	0	0	0	0	0	0
ICD_XXI	0	0	0	0	0	0

Figure A.1: Model coefficients of very large residuals

Table A.2 shows the policyholders with the largest number of events in the emergency room. We can see that policyholders F and B have the largest values, which are much greater than the number of emergency events of the policyholder with the third largest value: 62.

The estimated cost of the policyholders F and B surpass 1,000,000 BRL. No other policyholder in the list has such a large prediction. If we refer back to Figure A.1, we can see that N_EMERGENCY is by far the largest value among the covariates used for prediction for policyholders F and B. This shows the impact of this abnormal number of emergency cases in the cost prediction of these two individuals. If we change the number of emergency events of policyholders F and B to 62, the predicted cost would be 147,375 and 579,263, respectively, which are still very large predictions compared to their observed costs in 2017, but much lower than their predicted costs considering the actual values of emergency events.

Table A.2: Policyholders with largest number of emergency room events.

Policyholder	N_EMERGENCY	Cost 2017	Estimated cost	Residuals
F	**103**	**28,076**	**1,004,441**	**-976,365**
B	**93**	**8,212**	**2,472,362**	**-2,464,149**
	62	11,782	160,996	-149,214
	47	54,432	233,425	-178,993
	42	6,430	114,081	-107,651
	40	2,026	18,633	-16,608
	39	19,633	127,676	-108,043
	38	83,775	84,531	-756
	36	1,018	3,716	-2,698
	33	4,776	10,530	-5,754
	33	4,257	7,446	-3,189

We include Table A.3 that shows the frequency of policyholders by number of events in the emergency room. It illustrates how much the number of events from policyholders F and B depart from the bulk of policyholders.

Table A.3: Distribution of policyholders by range of emergency room events

N_EMERGENCY	Number of policyholders
0	253,196
1 - 5	231,975
6 - 10	13,675
11 - 15	1,460
16 - 20	262
21 - 25	58
26 - 30	13
33	2
36	1
38	1
39	1
40	1
42	1
47	1
62	1
93	1
103	1

Table A.4 lists all individuals who have remained hospitalised in 2016 for longer than 250 days. Policyholder A was hospitalised for the entire year of 2016 in the intensive care unit (ICU). From all individuals who remained in the hospital for longer than 250 days, policyholder A is the only one who has remained in ICU for the entire year. This has made the cost prediction extremely large. Also, policyholder C is hospitalised for 258 days, which is not abnormal since there are 18 other policyholders who have been hospitalised for longer periods. The issue is that the number of hospitalisations for that policyholder is 19, although the number of hospitalisations of individuals who remain in hospital for long periods is not greater than 3. Thus, the combination of large number of days hospitalised plus large number of hospitalisations has made the prediction very large.

Table A.5 lists the 10 policyholders with the largest number of hospitalisations in 2016. As we can see, policyholder C is the fifth in the list. However, this policyholder has also a large number of days in the hospital relative to the others in the list. Policyholder C has almost twice as many days in the hospital than policyholder D, which has the second largest number of days in hospital among the 10 policyholders who had more than 10 hospitalisations. For policyholder C, the number of days in ICU is also very high compared to the other policyholders in the list. Having large values of these three

Table A.4: Policyholders with largest hospitalisation stays

Policyholder	DAYS_INPATIENT	DAYS_ICU	N_INPATIENT	Cost 2017	Predicted cost	Residuals
A	**366**	**366**	**1**	**1,129,801**	**5,578,676**	**-4,448,875**
	366	65	1	460,777	647,246	-186,469
	366	24	1	344,499	211,651	132,849
	366	24	1	168,568	145,554	23,014
	366	18	1	270,753	326,629	-55,875
	366	7	1	322,435	284,822	37,613
	366	5	1	169,663	130,534	39,129
	366	4	1	202,152	316,391	-114,239
	366	2	1	321,357	154,168	167,189
	366	1	1	202,701	310,115	-107,414
	354	19	1	86,693	281,452	-194,758
	349	7	1	180,697	232,951	-52,254
	282	3	3	87,880	83,817	4,064
	278	13	3	277,479	231,092	46,388
	276	3	3	37,952	111,039	-73,087
	273	3	3	36,472	58,058	-21,586
	270	2	2	175	35,077	-34,902
	262	1	1	174,991	58,720	116,270
C	**258**	**95**	**19**	**94,033**	**1,251,443**	**-1,157,411**

covariates makes the predicted cost for policyholder C very inflated.

Policyholder D has the 7th largest number of hospitalisations (16) and the second largest number of days in the hospital among the top 10 policyholders with large number of hospitalisations (133). Although the number of days in intensive care is not very large in comparison to other policyholders in the list, 16 days in ICU is considered a large amount. What drives the overestimation of the cost, however, is the combination of the four covariates: N_INPATIENT, DAYS_INPATIENT, DAYS_ICU and TT_CHEMOTHERAPY. Because the policyholder does not have unique values in one covariate that would justify their omission from the study, we have decided to keep policyholder D in the analysis.

Table A.5: Policyholders with largest number of in-patient hospitalisations

Policyholder	N_INPATIENT	DAYS_INPATIENT	DAYS_ICU	TT_CHEMOTHERAPY	Cost 2017	Predicted cost
	27	123	27	0	13,682	179,785
	23	23	23	0	2,332	30,217
	22	105	22	0	15,704	105,362
	21	21	21	0	10,351	34,999
C	**19**	**258**	**95**	**0**	**94,033**	**1,251,443**
	18	97	18	0	212	36,052
D	**16**	**133**	**16**	**6.87**	**5,168**	**1,138,002**
	13	52	13	0	1,158	31,319
	12	107	12	0	24,030	69,106
	11	102	11	0	538	18,590

Supported by the reasons above, we think that it is more appropriate to remove the observations from the models. Policyholders A and C have been removed for having the number of days in the hospital greater than 250 and number of days in ICU greater than 90. Policyholders B and F have been removed for having a number of events in the emergency room greater than 90. The problem related to policyholder E is an inconsistency in the data: cost of hospitalisations different from zero (propINPATIENT), but number of hospitalisations/days in the hospital is equal to zero. This can be verified from Figure A.1.

Appendix B

Outputs of Models Fitted To Aggregated Data - Chapter 4

B.1 Groups of the M5 model tree tuned with the minimum number of instances per leaf equal to 1,700 (one-standard error rule)

```
logCOST <= 7.263 :
|   AGE <= 31 :
|   |   logCOST_MONTH <= 4.281 :
|   |   |   AGE <= 16 :
|   |   |   |   PLAN <= 1.5 : LM1 (18000/11.672%)
|   |   |   |   PLAN >  1.5 : LM2 (11244/6.376%)
|   |   |   AGE >  16 :
|   |   |   |   CONTR <= 1.5 :
|   |   |   |   |   MONTHS_AVG <= 1.5 : LM3 (15938/11.681%)
|   |   |   |   |   MONTHS_AVG >  1.5 : LM4 (4080/19.233%)
|   |   |   |   CONTR >  1.5 : LM5 (7871/23.91%)
|   |   logCOST_MONTH >  4.281 :
|   |   |   AGE <= 22 :
|   |   |   |   PLAN <= 1.5 :
|   |   |   |   |   propLAST_THREE <= 10.494 : LM6 (35172/20.022%)
|   |   |   |   |   propLAST_THREE >  10.494 : LM7 (45595/30.456%)
|   |   |   |   PLAN >  1.5 :
|   |   |   |   |   AGE <= 13 : LM8 (15161/8.621%)
|   |   |   |   |   AGE >  13 : LM9 (7877/21.833%)
|   |   |   AGE >  22 :
|   |   |   |   propLAST_THREE <= 45.529 : LM10 (36468/31.821%)
|   |   |   |   propLAST_THREE >  45.529 :
|   |   |   |   |   CONTR <= 1.5 :
|   |   |   |   |   |   logCOST <= 6.103 : LM11 (3072/19.035%)
|   |   |   |   |   |   logCOST >  6.103 : LM12 (2796/42.605%)
|   |   |   |   |   CONTR >  1.5 : LM13 (3482/70.464%)
|   AGE >  31 :
|   |   AGE <= 56 :
|   |   |   logCOST <= 6.214 : LM14 (108767/41.715%)
|   |   |   logCOST >  6.214 : LM15 (69224/56.819%)
|   |   AGE >  56 :
|   |   |   logCOST <= 5.583 : LM16 (25798/63.438%)
|   |   |   logCOST >  5.583 : LM17 (38949/97.363%)
logCOST >  7.263 :
|   logCOST <= 8.698 :
|   |   AGE <= 59 : LM18 (80530/75.767%)
|   |   AGE >  59 : LM19 (36689/129.108%)
|   logCOST >  8.698 :
|   |   logCOST <= 9.882 :
|   |   |   logCOST_DEC <= 6.292 : LM20 (15398/109.053%)
|   |   |   logCOST_DEC >  6.292 : LM21 (8141/225.96%)
|   |   logCOST >  9.882 : LM22 (5886/535.47%)
```

B.2 Linear Models in Each Terminal Leaf M5 model tree tuned with the minimum number of instances per leaf equal to 1,700 (one-standard error rule)

```
LM num: 1
COST_2017 =
341.4023 * logCOST
- 179.7681 * logCOST_MONTH
+ 0.1369 * N_MONTH
+ 0.0584 * logCOST_DEC
- 0.0537 * MONTHS_AVG
- 0.0782 * propMAX_COST
+ 0.018 * propLAST_THREE
+ 133.9203 * AGE
+ 0.0322 * GENDER
+ 0.0275 * HOSP
+ 0.0016 * OWNER
- 0.0593 * PRODUTO
+ 46.3411 * CONTR
+ 7.0392

LM num: 2
COST_2017 =
350.1418 * logCOST
- 232.3838 * logCOST_MONTH
+ 0.1369 * N_MONTH
+ 0.0584 * logCOST_DEC
- 0.0537 * MONTHS_AVG
- 0.0782 * propMAX_COST
+ 0.018 * propLAST_THREE
+ 0.3437 * AGE
+ 0.0322 * GENDER
+ 0.0275 * HOSP
+ 0.0016 * OWNER
- 0.0678 * PRODUTO
+ 0.0937 * CONTR
+ 121.1048

LM num: 3
COST_2017 =
111.0543 * logCOST
+ 0.0698 * logCOST_MONTH
+ 0.1369 * N_MONTH
+ 0.0584 * logCOST_DEC
+ 0.1432 * MONTHS_AVG
- 0.0782 * propMAX_COST
+ 0.0398 * propLAST_THREE
+ 0.3956 * AGE
+ 0.1085 * GENDER
+ 0.0275 * HOSP
```

```
- 0.0189 * OWNER
- 0.0954 * PRODUTO
+ 0.1146 * CONTR
+ 181.5716

LM num: 4
COST_2017 =
0.6236 * logCOST
+ 0.0698 * logCOST_MONTH
+ 0.1369 * N_MONTH
+ 0.0584 * logCOST_DEC
+ 0.4831 * MONTHS_AVG
- 0.0782 * propMAX_COST
+ 0.0398 * propLAST_THREE
+ 241.8787 * AGE
+ 236.9236 * GENDER
+ 0.0275 * HOSP
- 0.0189 * OWNER
- 0.1601 * PRODUTO
+ 0.1146 * CONTR
+ 197.7295

LM num: 5
COST_2017 =
178.2319 * logCOST
+ 0.0698 * logCOST_MONTH
+ 0.1369 * N_MONTH
+ 0.0584 * logCOST_DEC
+ 0.1464 * MONTHS_AVG
- 0.0782 * propMAX_COST
+ 0.0727 * propLAST_THREE
+ 236.6095 * AGE
+ 0.1011 * GENDER
+ 0.0275 * HOSP
- 0.0504 * OWNER
- 105.3841 * PRODUTO
+ 210.1767 * CONTR
- 324.5462

LM num: 6
COST_2017 =
0.1177 * logCOST
+ 256.9128 * logCOST_MONTH
+ 353.2715 * N_MONTH
+ 0.0778 * logCOST_DEC
- 51.6925 * MONTHS_AVG
- 0.0639 * propMAX_COST
+ 0.0368 * propLAST_THREE
+ 235.8281 * AGE
+ 0.0476 * GENDER
+ 49.1078 * HOSP
- 0.0362 * OWNER
- 0.0447 * PRODUTO
+ 22.5652 * CONTR
- 449.7188

LM num: 7
COST_2017 =
374.2896 * logCOST
+ 611.1326 * logCOST_MONTH
+ 277.2141 * N_MONTH
+ 19.7231 * logCOST_DEC
- 0.0954 * MONTHS_AVG
- 0.0639 * propMAX_COST
+ 73.8329 * propLAST_THREE
+ 378.8208 * AGE
+ 40.773 * GENDER
+ 84.2758 * HOSP
- 71.3869 * OWNER
- 0.0447 * PRODUTO
+ 0.0103 * CONTR
- 2178.6813

LM num: 8
COST_2017 =
247.0093 * logCOST
+ 0.4036 * logCOST_MONTH
+ 183.0535 * N_MONTH
+ 27.9832 * logCOST_DEC
- 0.1003 * MONTHS_AVG
+ 24.4316 * propMAX_COST
+ 0.0512 * propLAST_THREE
- 87.9627 * AGE
+ 0.0506 * GENDER
+ 0.0735 * HOSP
- 0.0839 * OWNER
- 0.077 * PRODUTO
+ 0.0103 * CONTR
- 352.518

LM num: 9
COST_2017 =
847.6489 * logCOST
+ 0.4036 * logCOST_MONTH
+ 0.828 * N_MONTH
+ 159.5281 * logCOST_DEC
- 0.1003 * MONTHS_AVG
+ 0.0278 * propMAX_COST
+ 0.0658 * propLAST_THREE
+ 874.1211 * AGE
+ 0.0506 * GENDER
+ 0.0735 * HOSP
- 0.1264 * OWNER
- 0.077 * PRODUTO
+ 0.0103 * CONTR
- 2001.362

LM num: 10
COST_2017 =
878.626 * logCOST
- 267.2503 * logCOST_MONTH
+ 0.3522 * N_MONTH
+ 109.7003 * logCOST_DEC
- 0.0808 * MONTHS_AVG
- 0.0687 * propMAX_COST
+ 101.4897 * propLAST_THREE
+ 577.0777 * AGE
+ 305.7901 * GENDER
+ 82.6817 * HOSP
- 0.0069 * OWNER
- 116.0769 * PRODUTO
+ 0.0103 * CONTR
- 1184.8187

LM num: 11
COST_2017 =
13.2691 * logCOST
- 0.0634 * logCOST_MONTH
+ 1.5942 * N_MONTH
+ 0.2956 * logCOST_DEC
- 0.0808 * MONTHS_AVG
- 0.0687 * propMAX_COST
+ 0.9986 * propLAST_THREE
+ 8.1939 * AGE
+ 370.5093 * GENDER
+ 0.1951 * HOSP
- 0.0069 * OWNER
- 0.2182 * PRODUTO
+ 0.0103 * CONTR
+ 481.7162

LM num: 12
COST_2017 =
14.1458 * logCOST
- 0.0634 * logCOST_MONTH
+ 1.5942 * N_MONTH
+ 0.2956 * logCOST_DEC
- 0.0808 * MONTHS_AVG
- 0.0687 * propMAX_COST
+ 0.9986 * propLAST_THREE
+ 1689.8739 * AGE
+ 1339.173 * GENDER
+ 0.1951 * HOSP
- 0.0069 * OWNER
- 0.2182 * PRODUTO
+ 0.0103 * CONTR
- 1140.8784

LM num: 13
COST_2017 =
7.243 * logCOST
- 0.8437 * logCOST_MONTH
+ 860.4294 * N_MONTH
+ 0.2956 * logCOST_DEC
- 0.0808 * MONTHS_AVG
- 0.0687 * propMAX_COST
+ 414.9248 * propLAST_THREE
+ 4.7776 * AGE
+ 2.9984 * GENDER
+ 0.1951 * HOSP
- 0.0069 * OWNER
- 0.2182 * PRODUTO
+ 0.0103 * CONTR
- 612.1737
```

```
LM num: 14
COST_2017 =
755.0341 * logCOST
- 454.7996 * logCOST_MONTH
+ 0.1227 * N_MONTH
+ 0.0685 * logCOST_DEC
- 0.0741 * MONTHS_AVG
+ 22.6435 * propMAX_COST
+ 35.8884 * propLAST_THREE
+ 173.0489 * AGE
+ 70.4014 * GENDER
+ 63.5744 * HOSP
- 54.6354 * OWNER
- 113.2552 * PRODUTO
+ 83.7523 * CONTR
- 11.8717

LM num: 15
COST_2017 =
1289.0073 * logCOST
+ 0.1137 * logCOST_MONTH
+ 134.1364 * N_MONTH
+ 129.2445 * logCOST_DEC
- 0.0836 * MONTHS_AVG
- 92.8677 * propMAX_COST
+ 119.9418 * propLAST_THREE
+ 0.12 * AGE
+ 156.521 * GENDER
+ 228.7829 * HOSP
+ 113.9799 * OWNER
- 183.74 * PRODUTO
+ 39.3166 * CONTR
- 1941.9135

LM num: 16
COST_2017 =
0.5494 * logCOST
+ 528.9961 * logCOST_MONTH
+ 0.3652 * N_MONTH
+ 0.0695 * logCOST_DEC
- 0.1714 * MONTHS_AVG
- 107.2205 * propMAX_COST
+ 0.0526 * propLAST_THREE
+ 310.2363 * AGE
- 316.3228 * GENDER
+ 0.1243 * HOSP
+ 0.0063 * OWNER
- 262.6009 * PRODUTO
+ 98.0569 * CONTR
- 44.586

LM num: 17
COST_2017 =
1343.7621 * logCOST
+ 0.0923 * logCOST_MONTH
+ 282.3968 * N_MONTH
+ 0.0695 * logCOST_DEC
```

```
- 0.1392 * MONTHS_AVG
- 0.1373 * propMAX_COST
+ 90.7692 * propLAST_THREE
+ 1234.0605 * AGE
- 439.9563 * GENDER
+ 0.0961 * HOSP
+ 0.0063 * OWNER
- 410.3239 * PRODUTO
+ 0.0568 * CONTR
- 4320.0221

LM num: 18
COST_2017 =
1277.6459 * logCOST
+ 1709.6704 * logCOST_MONTH
+ 327.3707 * N_MONTH
+ 343.4548 * logCOST_DEC
+ 119.5034 * MONTHS_AVG
- 769.4409 * propMAX_COST
+ 333.6029 * propLAST_THREE
+ 478.028 * AGE
+ 222.784 * GENDER
+ 110.1944 * HOSP
+ 0.0131 * OWNER
- 217.2372 * PRODUTO
+ 0.0527 * CONTR
- 7230.3407

LM num: 19
COST_2017 =
1.9251 * logCOST
+ 2501.3004 * logCOST_MONTH
+ 819.6161 * N_MONTH
+ 338.446 * logCOST_DEC
- 0.0139 * MONTHS_AVG
- 652.2235 * propMAX_COST
+ 408.8281 * propLAST_THREE
+ 1404.4859 * AGE
- 733.0153 * GENDER
+ 0.0578 * HOSP
+ 0.0131 * OWNER
- 568.8956 * PRODUTO
+ 0.0618 * CONTR
- 8037.1877

LM num: 20
COST_2017 =
4476.5379 * logCOST
+ 679.2449 * logCOST_MONTH
+ 575.0414 * N_MONTH
+ 3.0794 * logCOST_DEC
+ 452.8568 * MONTHS_AVG
- 1217.6403 * propMAX_COST
+ 105.9185 * propLAST_THREE
+ 1372.3718 * AGE
- 0.3149 * GENDER
- 0.1126 * HOSP
+ 0.0131 * OWNER
+ 0.1147 * PRODUTO
+ 0.1393 * CONTR
- 15945.337

LM num: 21
COST_2017 =
22184.2834 * logCOST
- 2416.4274 * logCOST_MONTH
- 1.451 * N_MONTH
+ 3.8888 * logCOST_DEC
+ 2510.9793 * MONTHS_AVG
- 11.1873 * propMAX_COST
+ 1010.5977 * propLAST_THREE
+ 1728.7625 * AGE
- 0.3149 * GENDER
- 0.1126 * HOSP
+ 0.0131 * OWNER
+ 0.1147 * PRODUTO
+ 0.1393 * CONTR
- 67737.3013

LM num: 22
COST_2017 =
83477.7053 * logCOST
+ 18.1088 * logCOST_MONTH
- 3.6724 * N_MONTH
+ 7443.7609 * logCOST_DEC
+ 5.0284 * MONTHS_AVG
- 9147.0387 * propMAX_COST
+ 1.4693 * propLAST_THREE
- 3.6345 * AGE
- 0.3149 * GENDER
- 0.1126 * HOSP
+ 0.0131 * OWNER
+ 0.1147 * PRODUTO
+ 0.1393 * CONTR
- 301336.1365
```

B.3 Conditions and Linear Models of Cubist With Four Rules

```
Model:

  Rule 1: [451613 cases, mean 1353.582, range 0 to 534290.3, est err 1060.841]

    if
AGE <= 53
    then
outcome = -434.82 + 210 logCOST - 8 propMAX_COST + 10 AGE
          + 79 logCOST_DEC + 200 GENDER - 48 MONTHS_AVG

  Rule 2: [573348 cases, mean 1600.871, range 0 to 545652.4, est err 1236.440]

    if
logCOST <= 8.87
    then
outcome = -1695.96 + 482 logCOST - 21 propMAX_COST + 26 AGE
          + 130 logCOST_DEC - 141 MONTHS_AVG - 226 PLAN

  Rule 3: [593560 cases, mean 1817.809, range 0 to 545652.4, est err 1666.923]

    if
logCOST <= 10.51
    then
outcome = -56619.681 + 6493 logCOST - 97 propMAX_COST + 1276 MONTHS_AVG
          + 572 logCOST_DEC + 17 AGE + 13 propLAST_THREE + 338 CONTR
          - 436 GENDER - 317 COPAY + 497 PRE_EXISTING

  Rule 4: [22699 cases, mean 11937.669, range 0 to 534290.3, est err 10136.515]

    if
logCOST > 8.87
    then
outcome = -321083.609 + 32211 logCOST - 540 propMAX_COST
          + 3739 logCOST_DEC + 116 propLAST_THREE - 122 AGE + 2669 CONTR
          - 3602 GENDER


Evaluation on training data (596047 cases, sampled):

    Average  |error|           2153.776
    Relative |error|               0.88
    Correlation coefficient        0.42


Attribute usage:
  Conds  Model

   72%    100%    logCOST
   28%    100%    AGE
          100%    propMAX_COST
          100%    logCOST_DEC
           99%    MONTHS_AVG
           65%    GENDER
           38%    propLAST_THREE
           38%    CONTR
           36%    PRE_EXISTING
```

```
36%     COPAY
35%     PLAN
```

B.4 Conditions and Linear Models of Cubist With 11 Rules

```
Model:

  Rule 1: [133971 cases, mean 495.446, range 0 to 149876.9, est err 383.213]

    if
AGE <= 27
logCOST <= 6.47
    then
outcome = -48 + 90 logCOST + 6 AGE - 3 propMAX_COST

  Rule 2: [287517 cases, mean 677.849, range 0 to 496483.2, est err 584.622]

    if
logCOST <= 6.25
    then
outcome = -141.83 + 114 logCOST - 3 propMAX_COST + 4 AGE

  Rule 3: [573348 cases, mean 1600.871, range 0 to 545652.4, est err 1207.399]

    if
logCOST <= 8.87
    then
outcome = -1727.56 + 399 logCOST - 12 propMAX_COST + 12 AGE
          + 39 logCOST_DEC

  Rule 4: [320300 cases, mean 1678.788, range 0 to 496483.2, est err 1308.956]

    if
logCOST_DEC <= 4.81
AGE > 27
    then
outcome = -1639.42 + 432 logCOST - 14 propMAX_COST + 320 GENDER
          - 178 PLAN + 59 logCOST_DEC + 6 AGE

  Rule 5: [131966 cases, mean 3032.453, range 0 to 545652.4, est err 2295.865]

    if
AGE > 53
logCOST <= 8.87
    then
outcome = -1695.96 + 482 logCOST - 21 propMAX_COST + 26 AGE
          + 130 logCOST_DEC - 141 MONTHS_AVG - 226 PLAN

  Rule 6: [88502 cases, mean 3470.430, range 0 to 545652.4, est err 2428.264]

    if
logCOST_DEC > 4.81
logCOST <= 8.87
    then
outcome = -5064.86 + 950 logCOST - 19 propMAX_COST + 12 propLAST_THREE
          + 14 AGE - 120 logCOST_DEC + 503 GENDER + 198 HOSP_ACCOMM
          - 112 PLAN

  Rule 7: [20212 cases, mean 7971.624, range 0 to 345440.5, est err 6026.464]

    if
```

```
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -19516.409 + 2411 logCOST - 55 propMAX_COST + 42 AGE
          + 420 MONTHS_AVG + 266 logCOST_DEC - 226 PLAN + 176 CONTR
          - 316 COPAY + 3 propLAST_THREE

  Rule 8: [754 cases, mean 12380.188, range 0 to 255325.4, est err 12291.881]

    if
logCOST_DEC <= 5.39
logCOST > 10.51
    then
outcome = -24199.149 + 2172 logCOST + 3042 MONTHS_AVG + 88 propLAST_THREE
          - 41 propMAX_COST + 338 logCOST_DEC - 6 AGE + 153 CONTR
          - 235 GENDER

  Rule 9: [7849 cases, mean 12946.391, range 0 to 345440.5, est err 9371.012]

    if
logCOST_DEC > 5.61
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -70401.958 + 7977 logCOST - 142 propMAX_COST + 701 logCOST_DEC
          + 50 AGE + 428 MONTHS_AVG + 23 propLAST_THREE - 518 COPAY

  Rule 10: [2209 cases, mean 36337.848, range 0 to 490154.5, est err 22507.814]

    if
logCOST_DEC <= 9.64
logCOST > 10.51
    then
outcome = -356462.205 + 27925 logCOST + 12881 logCOST_DEC
          - 425 propMAX_COST + 2624 CONTR + 19 propLAST_THREE - 17 AGE
          - 528 GENDER

  Rule 11: [523 cases, mean 65534.273, range 0 to 534290.3, est err 43215.426]

    if
logCOST_DEC > 9.64
    then
outcome = -383696.097 + 78342 logCOST - 33509 logCOST_DEC - 1086 AGE
          - 701 propMAX_COST + 11929 CONTR


Evaluation on training data (596047 cases, sampled):

    Average  |error|           2175.148
    Relative |error|               0.89
    Correlation coefficient        0.43


Attribute usage:
  Conds  Model

   80%   100%    logCOST
   37%   100%    AGE
```

27%	73%	logCOST_DEC
	100%	propMAX_COST
	36%	PLAN
	26%	GENDER
	10%	MONTHS_AVG
	8%	propLAST_THREE
	6%	HOSP_ACCOMM
	2%	COPAY
	2%	CONTR

B.5 Conditions and Linear Models of Cubist With 20 Rules

```
Model:

  Rule 1: [133971 cases, mean 495.446, range 0 to 149876.9, est err 383.213]

    if
AGE <= 27
logCOST <= 6.47
    then
outcome = -48 + 90 logCOST + 6 AGE - 3 propMAX_COST

  Rule 2: [287517 cases, mean 677.849, range 0 to 496483.2, est err 584.622]

    if
logCOST <= 6.25
    then
outcome = -141.83 + 114 logCOST - 3 propMAX_COST + 4 AGE

  Rule 3: [446798 cases, mean 1044.566, range 0 to 496483.2, est err 842.220]

    if
logCOST <= 7.24
    then
outcome = -318 + 219 logCOST + 8 AGE - 5 propMAX_COST - 146 PLAN
          + 39 logCOST_DEC

  Rule 4: [494812 cases, mean 1332.735, range 0 to 496483.2, est err 1038.031]

    if
logCOST_DEC <= 4.81
    then
outcome = -1639.42 + 432 logCOST - 14 propMAX_COST + 320 GENDER
          - 178 PLAN + 59 logCOST_DEC + 6 AGE

  Rule 5: [249686 cases, mean 1454.763, range 0 to 424336.3, est err 1188.504]

    if
logCOST > 5.57
logCOST <= 7.24
    then
outcome = -2956.24 + 534 logCOST + 23 AGE - 12 propMAX_COST
          + 76 logCOST_DEC - 197 PLAN - 62 MONTHS_AVG

  Rule 6: [59418 cases, mean 1461.184, range 0 to 369934.8, est err 966.897]

    if
AGE <= 27
logCOST > 6.47
logCOST <= 8.87
    then
outcome = -1727.56 + 399 logCOST - 12 propMAX_COST + 12 AGE
          + 39 logCOST_DEC

  Rule 7: [56530 cases, mean 2606.388, range 0 to 424336.3, est err 1896.417]

    if
logCOST_DEC > 4.81
```

```
AGE <= 53
logCOST <= 8.87
    then
outcome = -8315.86 + 1394 logCOST - 42 propMAX_COST + 12 AGE
          + 98 MONTHS_AVG + 334 GENDER + 269 HOSP_ACCOMM

  Rule 8: [71044 cases, mean 2846.984, range 0 to 545652.4, est err 2034.483]

    if
logCOST_DEC > 4.81
logCOST <= 8.1
    then
outcome = -5514.64 + 991 logCOST - 12 propMAX_COST - 141 logCOST_DEC
          + 13 propLAST_THREE + 12 AGE + 523 GENDER + 180 HOSP_ACCOMM
          - 104 PLAN

  Rule 9: [53305 cases, mean 4738.375, range 0 to 545652.4, est err 3248.444]

    if
AGE > 53
logCOST > 7.24
logCOST <= 8.87
    then
outcome = -10953.67 + 1530 logCOST - 35 propMAX_COST + 38 AGE
          + 102 logCOST_DEC - 270 PLAN + 5 propLAST_THREE
          + 199 HOSP_ACCOMM

  Rule 10: [12363 cases, mean 4813.252, range 0 to 252244.5, est err 3885.164]

    if
logCOST_DEC <= 5.61
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -16344.581 + 1994 logCOST - 60 propMAX_COST + 49 AGE
          + 245 logCOST_DEC + 186 MONTHS_AVG - 397 PLAN + 343 CONTR

  Rule 11: [20212 cases, mean 7971.624, range 0 to 345440.5, est err 6034.424]

    if
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -17851.601 + 2247 logCOST - 47 propMAX_COST + 409 MONTHS_AVG
          + 32 AGE + 238 logCOST_DEC - 312 COPAY - 155 PLAN + 147 CONTR

  Rule 12: [754 cases, mean 12380.188, range 0 to 255325.4, est err 12291.881]

    if
logCOST_DEC <= 5.39
logCOST > 10.51
    then
outcome = -24199.149 + 2172 logCOST + 3042 MONTHS_AVG
          + 88 propLAST_THREE - 41 propMAX_COST + 338 logCOST_DEC
          - 6 AGE + 153 CONTR - 235 GENDER

  Rule 13: [7849 cases, mean 12946.391, range 0 to 345440.5, est err 9371.012]
```

```
    if
logCOST_DEC > 5.61
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -70401.958 + 7977 logCOST - 142 propMAX_COST + 701 logCOST_DEC
          + 50 AGE + 428 MONTHS_AVG + 23 propLAST_THREE - 518 COPAY

  Rule 14: [1102 cases, mean 15670.787, range 0 to 442991.3, est err 13013.178]

    if
propLAST_THREE <= 12
logCOST > 10.51
    then
outcome = -50011.78 + 1152 propLAST_THREE + 4182 logCOST
          + 2142 logCOST_DEC - 4173 GENDER - 69 propMAX_COST
          + 1083 MONTHS_AVG

  Rule 15: [883 cases, mean 45841.949, range 0 to 389330.6, est err 24434.684]

    if
propLAST_THREE > 12
logCOST_DEC <= 9.64
logCOST > 10.51
logCOST <= 11.47
    then
outcome = -306144.56 + 25465 logCOST + 10195 logCOST_DEC
          - 407 propMAX_COST + 721 CONTR + 15 propLAST_THREE - 10 AGE
          - 340 GENDER + 55 MONTHS_AVG

  Rule 16: [228 cases, mean 68683.102, range 0 to 383465, est err 53075.164]

    if
logCOST_DEC > 9.64
logCOST > 10.51
logCOST <= 12.24
    then
outcome = -134036.618 + 36623 logCOST - 18893 logCOST_DEC
          - 1028 propMAX_COST + 404 propLAST_THREE + 12598 CONTR
          - 408 AGE + 49 MONTHS_AVG

  Rule 17: [391 cases, mean 70072.047, range 0 to 490154.5, est err 34894.543]

    if
logCOST_DEC <= 9.64
logCOST > 11.47
    then
outcome = -367813.467 + 26674 logCOST_DEC + 20094 logCOST
          - 898 propMAX_COST + 6312 CONTR

  Rule 18: [28 cases, mean 80921.094, range 1437.25 to 321303.9, est err 76902.070]

    if
logCOST_DEC > 5.39
logCOST_DEC <= 9.64
AGE <= 11
logCOST > 10.51
logCOST <= 11.47
```

```
    then
outcome = -1869423.614 + 173054 logCOST + 40059 COPAY + 1983 logCOST_DEC
          - 60 propMAX_COST + 125 CONTR

  Rule 19: [20 cases, mean 87974.352, range 651.11 to 490154.5, est err 88569.961]

    if
logCOST_DEC > 5.39
logCOST_DEC <= 5.83
logCOST > 11.47
    then
outcome = 188329.51 - 188303 HOSP_ACCOMM + 447 logCOST + 116 logCOST_DEC
          - 7 propMAX_COST

  Rule 20: [50 cases, mean 278412.625, range 849.28 to 534290.3, est err 75017.117]

    if
logCOST_DEC > 9.64
logCOST > 12.24
    then
outcome = 461678.346 - 3544 propLAST_THREE - 1733 AGE


Evaluation on training data (596047 cases, sampled):

    Average  |error|           2164.521
    Relative |error|               0.88
    Correlation coefficient        0.43


Attribute usage:
  Conds  Model

   74%   100%    logCOST
   34%    75%    logCOST_DEC
   16%   100%    AGE
           7%    propLAST_THREE
         100%    propMAX_COST
          71%    PLAN
          33%    GENDER
          18%    MONTHS_AVG
          10%    HOSP_ACCOMM
           2%    CONTR
           1%    COPAY
```

Appendix C

Conditions and Linear Models of Cubist Fitted To Detailed Medical Data: ICD-10 Chapters

C.1 Cubist With Three Rules

```
Model:

  Rule 1: [573348 cases, mean 1600.8715, range 0 to 545652.4, est err 1204.0887]

    if
logCOST <= 8.87
    then
outcome = -155.98 + 4998 TT_CHEMOTHERAPY + 785 DAYS_ICU
          + 195 DAYS_INPATIENT + 113 propHOMECARE + 97 N_VISITS
          - 21 propINPATIENT + 91 logCOST + 9 AGE - 199 COST_ICU
          - 494 N_INPATIENT + 67 logCOST_DEC - 5 propVISITS
          + 47 TT_IMAGE - 129 PLAN - 4 propTESTS_THERAPIES + 34 TT_LAB

  Rule 2: [20212 cases, mean 7971.6235, range 0 to 345440.5, est err 6147.9111]

    if
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -42834.6407 + 5508 logCOST - 132 propINPATIENT
          - 77 propTESTS_THERAPIES + 343 logCOST_DEC + 394 MONTHS_AVG
          + 17 propVISITS + 16 propMAX_COST + 1912 TT_CHEMOTHERAPY
          - 15 propEMERGENCY + 112 TT_LAB + 72 DAYS_INPATIENT
          - 149 TT_SPECIFIC + 216 DAYS_ICU + 9 propLAST_THREE + 10 AGE
          - 199 COST_ICU - 111 TT_ENDOSCOPY + 178 CONTR - 29 N_VISITS
          - 472 TT_RADIOLOGY - 234 COPAY + 240 N_INPATIENT - 103 ICD_XV

  Rule 3: [2487 cases, mean 44169.9570, range 0 to 534290.3, est err 28970.2988]

    if
logCOST > 10.51
    then
outcome = -350910.427 + 36065 logCOST - 543 propTESTS_THERAPIES
          - 556 propINPATIENT + 2801 logCOST_DEC + 141 propLAST_THREE
          + 414 propHOMECARE - 958 TT_SPECIFIC + 108 DAYS_INPATIENT
          - 1496 TT_CHEMOTHERAPY - 263 DAYS_ICU
```

```
Evaluation on training data (596047 cases, sampled):

    Average  |error|          2252.4041
    Relative |error|               0.92
    Correlation coefficient        0.33


Attribute usage:
  Conds  Model

  100%   100%    logCOST
         100%    propTESTS_THERAPIES
         100%    propINPATIENT
         100%    TT_CHEMOTHERAPY
         100%    DAYS_ICU
         100%    DAYS_INPATIENT
         100%    logCOST_DEC
         100%    propVISITS
         100%    N_VISITS
         100%    TT_LAB
         100%    COST_ICU
         100%    N_INPATIENT
         100%    AGE
          97%    propHOMECARE
          96%    PLAN
          96%    TT_IMAGE
           4%    propLAST_THREE
           4%    TT_SPECIFIC
           3%    propMAX_COST
           3%    MONTHS_AVG
           3%    propEMERGENCY
           3%    TT_ENDOSCOPY
           3%    TT_RADIOLOGY
           3%    ICD_XV
           3%    CONTR
           3%    COPAY
```

C.2 Cubist With Seven Rules

```
Model:

  Rule 1: [441382 cases, mean 1172.8519, range 0 to 496483.2, est err 874.6164]

    if
AGE <= 53
logCOST <= 8.87
    then
outcome = -91 + 87 N_VISITS + 44 propHOMECARE + 7 AGE + 126 COST_ICU
          + 46 logCOST_DEC + 42 TT_IMAGE + 61 N_EMERGENCY
          - 5 propINPATIENT

  Rule 2: [573348 cases, mean 1600.8715, range 0 to 545652.4, est err 1201.8218]

    if
logCOST <= 8.87
    then
outcome = -991.79 + 110 N_VISITS + 128 DAYS_INPATIENT + 151 logCOST
          + 65 propHOMECARE + 17 AGE + 929 TT_CHEMOTHERAPY
          - 12 propINPATIENT + 181 DAYS_ICU - 243 PLAN + 67 logCOST_DEC
          + 89 N_EMERGENCY - 5 propVISITS + 43 TT_IMAGE

  Rule 3: [20212 cases, mean 7971.6235, range 0 to 345440.5, est err 5926.9771]

    if
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -16690.301 + 2140 logCOST - 63 propVISITS - 59 propINPATIENT
          - 30 propTESTS_THERAPIES + 30 AGE + 288 MONTHS_AVG
          + 188 logCOST_DEC + 80 N_VISITS + 110 TT_LAB - 324 PLAN
          + 627 TT_CHEMOTHERAPY - 169 ICD_XV + 80 N_EMERGENCY
          + 163 CONTR + 5 propLAST_THREE + 308 N_INPATIENT
          + 28 DAYS_INPATIENT - 239 COPAY + 16 propHOMECARE
          - 53 TT_ENDOSCOPY

  Rule 4: [754 cases, mean 12380.1885, range 0 to 255325.4, est err 11838.6816]

    if
logCOST_DEC <= 5.39
logCOST > 10.51
    then
outcome = -19737.4187 + 3812 logCOST - 260 propINPATIENT + 2482 MONTHS_AVG
          + 157 propEMERGENCY - 99 propTESTS_THERAPIES + 452 logCOST_DEC
          - 2500 TT_CHEMOTHERAPY + 22 propLAST_THREE - 128 TT_SPECIFIC
          - 373 GENDER + 32 propHOMECARE

  Rule 5: [7849 cases, mean 12946.3906, range 0 to 345440.5, est err 9155.4355]

    if
logCOST_DEC > 5.61
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -61920.8796 + 7071 logCOST - 173 propINPATIENT - 84 propVISITS
          + 779 logCOST_DEC - 43 propTESTS_THERAPIES + 30 propMAX_COST
          + 165 DAYS_INPATIENT + 248 TT_LAB + 1974 TT_CHEMOTHERAPY
          + 18 propLAST_THREE + 1026 N_INPATIENT + 77 N_VISITS
          + 195 MONTHS_AVG - 201 TT_SPECIFIC + 16 AGE - 365 ICD_XV
          - 188 TT_ENDOSCOPY - 894 TT_RADIOLOGY - 168 DAYS_ICU
          - 418 COPAY - 239 ICD_IV + 195 ICD_II
          + 103 ICD_XIV

  Rule 6: [2209 cases, mean 36337.8477, range 0 to 490154.5, est err 21402.3086]

    if
logCOST_DEC <= 9.64
logCOST > 10.51
    then
outcome = -356912.9467 + 30033 logCOST + 1898 propEMERGENCY
          + 11020 logCOST_DEC - 610 propTESTS_THERAPIES
          - 462 propINPATIENT + 71 propMAX_COST + 241 propHOMECARE
          + 16 propLAST_THREE - 542 ICD_II - 102 TT_SPECIFIC

  Rule 7: [278 cases, mean 106404.2344, range 0 to 534290.3, est err 52680.6406]

    if
logCOST_DEC > 9.64
logCOST > 10.51
    then
outcome = -78859.9516 + 19458 logCOST - 1196 propINPATIENT
          + 1925 propHOMECARE - 3370 ICD_IX - 3827 ICD_XIII
          + 579 DAYS_INPATIENT + 1035 DAYS_ICU


Evaluation on training data (596047 cases, sampled):

    Average  |error|          2240.3697
    Relative |error|               0.91
    Correlation coefficient        0.35


Attribute usage:
  Conds  Model

  100%    58%    logCOST
   42%   100%    AGE
    1%   100%    logCOST_DEC
         100%    propINPATIENT
         100%    N_VISITS
          99%    propHOMECARE
          99%    N_EMERGENCY
          97%    TT_IMAGE
          58%    TT_CHEMOTHERAPY
          58%    DAYS_INPATIENT
          57%    propVISITS
          57%    PLAN
          56%    DAYS_ICU
          42%    COST_ICU
           3%    propLAST_THREE
           3%    propTESTS_THERAPIES
           3%    MONTHS_AVG
           3%    TT_ENDOSCOPY
           3%    TT_LAB
```

```
3%     N_INPATIENT
3%     ICD_XV
3%     COPAY
2%     CONTR
1%     TT_SPECIFIC
```

Appendix D

Charlson Comorbidities Mapped Into ICD-10 Codes

Myocardial Infarction
I21.x, I22.x, 125.2

Congestive Heart Failure
109.9, 111.0, 113.0, 113.2, I25.5, I42.0, I42.5-I42.9, I43.x,
I50.x, P29.0

Peripheral Vascular Disease
I70.x, I71.x, I73.1, I73.8, I73.9, 177.1, 179.0, I179.2, K55.1,
K55.8, K55.9, Z95.8, Z95.9

Cerebrovascular Disease
G45.x, G46.x, H34.0, I60.x-I69.x

Dementia
F00.x-F03.x, F05.1, G30.x, G31.1

Chronic Pulmonary Disease
127.8, 127.9, J40.x-J47.x, J60.x-J67.x, J68.4, J70.1, J70.3

Rheumatic Disease
M05.x, M06.x, M31.5, M32.x-M34.x, M35.1, M35.3, M36.0

Peptic Ulcer Disease
K25.x-K28.x

Mild Liver Disease
B18.x, K70.0-K70.3, K70.9, K71.3-K71.5, K71.7, K73.x, K74.x,
K76.0, K76.2-K76.4, K76.8, K76.9, Z94.4

Diabetes Without Chronic Complication
E10.0, E10.1, E10.6, E10.8, E10.9, E11.0, E11.1, E11.6, E11.8,
E11.9, E12.0, E12.1, E12.6, E12.8, E12.9, E13.0, E13.1, E13.6,
E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9

Diabetes With Chronic Complication
E10.2-E10.5, El0.7, E11.2-Ell11.5, E11.7, E12.2-E12.5, E12.7,
E13.2- E13.5, E13.7, E14.2-E14.5, E14.7

Hemiplegia or Paraplegia
G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0-G83.4, G83.9

Renal Disease
112.0, I113.1, N03.2-N03.7, N05.2- N05.7, N18.x, N19.x, N25.0,
Z49.0- Z49.2, Z94.0, Z99.2

Any malignancy, including lymphoma and leukemia,
except malignant neoplasm of skin
C00.x-C26.x, C30.x-C34.x, C37.x- C41.x, C43.x, C45.x-C58.x,
C60.x- C76.x, C81.x-C85.x, C88.x, C90.x-C97.x

Moderate or severe liver disease
185.0, I185.9, I186.4, I198.2, K70.4, K71.1, K72.1, K72.9,
K76.5, K76.6, K76.7

Metastatic solid tumor
C77.x-C80.x

AIDS/HIV
B20.x-B22.x, B24.x

Appendix E

Conditions and Linear Models of Cubist Fitted To Detailed Medical Data: Charlson Comorbidities

E.1 Cubist With Three Rules

```
Model:

  Rule 1: [573348 cases, mean 1600.8715, range 0 to 545652.4, est err 1201.8225]

    if
logCOST <= 8.87
    then
outcome = -87.13 + 4669 TT_CHEMOTHERAPY + 611 DAYS_ICU
          + 171 DAYS_INPATIENT + 110 propHOMECARE + 96 N_VISITS
          + 3046 RENAL - 22 propINPATIENT + 90 logCOST + 9 AGE
          - 176 COST_ICU + 67 logCOST_DEC - 5 propVISITS + 44 TT_IMAGE
          - 128 PLAN - 4 propTESTS_THERAPIES - 250 N_INPATIENT

  Rule 2: [20212 cases, mean 7971.6235, range 0 to 345440.5, est err 6159.1968]

    if
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -40867.631 + 5319 logCOST - 127 propINPATIENT
          - 80 propTESTS_THERAPIES + 348 logCOST_DEC + 27 propVISITS
          + 413 MONTHS_AVG + 2192 TT_CHEMOTHERAPY + 92 DAYS_INPATIENT
          + 12 propMAX_COST + 14 AGE + 252 DAYS_ICU - 148 TT_SPECIFIC
          - 254 COST_ICU + 9 propLAST_THREE + 1650 RENAL - 36 N_VISITS
          + 203 CONTR - 95 TT_ENDOSCOPY - 157 PLAN - 407 TT_RADIOLOGY
          + 128 COMORBID

  Rule 3: [2487 cases, mean 44169.9570, range 0 to 534290.3, est err 28713.7695]

    if
logCOST > 10.51
    then
outcome = -515938.649 + 12829 propVISITS + 51083 logCOST
          - 645 propTESTS_THERAPIES - 581 propINPATIENT - 1898 N_VISITS
          + 2761 logCOST_DEC + 129 propLAST_THREE + 362 propHOMECARE
          + 153 DAYS_INPATIENT - 484 DAYS_ICU + 382 COST_ICU
          - 1094 TT_CHEMOTHERAPY


Evaluation on training data (596047 cases, sampled):

    Average  |error|          2283.8900
    Relative |error|               0.93
    Correlation coefficient        0.33


Attribute usage:
  Conds  Model

  100%   100%    logCOST
         100%    propVISITS
         100%    propTESTS_THERAPIES
         100%    propINPATIENT
         100%    N_VISITS
         100%    TT_CHEMOTHERAPY
         100%    COST_ICU
         100%    DAYS_ICU
         100%    DAYS_INPATIENT
         100%    logCOST_DEC
         100%    PLAN
         100%    RENAL
         100%    AGE
          97%    propHOMECARE
          96%    TT_IMAGE
          96%    N_INPATIENT
           4%    propLAST_THREE
           3%    propMAX_COST
           3%    MONTHS_AVG
           3%    TT_ENDOSCOPY
           3%    TT_SPECIFIC
           3%    TT_RADIOLOGY
           3%    COMORBID
           3%    CONTR
```

E.2 Cubist With Eight Rules

```
Model:

  Rule 1: [384852 cases, mean 962.2833, range 0 to 496483.2, est err 731.1332]

    if
logCOST_DEC <= 4.81
AGE <= 53
logCOST <= 8.87
    then
outcome = -88.68 + 3624 TT_CHEMOTHERAPY + 532 DAYS_ICU
          + 120 DAYS_INPATIENT + 72 propHOMECARE + 84 N_VISITS
          + 2364 RENAL - 16 propINPATIENT + 76 logCOST + 8 AGE
          - 153 COST_ICU + 49 logCOST_DEC + 36 TT_IMAGE
          - 244 N_INPATIENT - 3 propVISITS - 106 PLAN

  Rule 2: [573348 cases, mean 1600.8715, range 0 to 545652.4, est err 1202.8276]

    if
logCOST <= 8.87
    then
outcome = -1495.4401 + 5044 TT_CHEMOTHERAPY + 807 DAYS_ICU
          + 176 DAYS_INPATIENT + 112 propHOMECARE + 224 logCOST
          + 95 N_VISITS - 24 propINPATIENT + 3291 RENAL
          - 213 COST_ICU + 9 AGE - 5 propVISITS - 339 N_INPATIENT
          + 45 TT_IMAGE - 133 PLAN + 35 TT_LAB + 177 GENDER
          + 35 logCOST_DEC + 3 propLAST_THREE

  Rule 3: [131966 cases, mean 3032.4529, range 0 to 545652.4, est err 2228.5847]

    if
AGE > 53
logCOST <= 8.87
    then
outcome = -979.0001 + 109 N_VISITS + 114 DAYS_INPATIENT + 152 logCOST
          + 66 propHOMECARE + 17 AGE - 14 propINPATIENT
          + 963 TT_CHEMOTHERAPY - 242 PLAN + 67 logCOST_DEC
          - 5 propVISITS + 87 N_EMERGENCY + 42 TT_IMAGE + 628 RENAL

  Rule 4: [20212 cases, mean 7971.6235, range 0 to 345440.5, est err 5946.6709]

    if
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -16149.1808 + 2079 logCOST - 60 propVISITS - 59 propINPATIENT
          - 37 propTESTS_THERAPIES + 33 AGE + 327 MONTHS_AVG
          + 191 logCOST_DEC + 78 N_VISITS - 345 PLAN + 62 DAYS_INPATIENT
          + 776 TT_CHEMOTHERAPY + 6 propLAST_THREE + 179 CONTR
          + 20 propHOMECARE + 65 N_EMERGENCY - 241 COPAY
          + 122 COMORBID + 225 N_INPATIENT

  Rule 5: [754 cases, mean 12380.1885, range 0 to 255325.4, est err 11973.9307]

    if
logCOST_DEC <= 5.39
logCOST > 10.51
```

```
    then
outcome = -1529.0713 + 658 propVISITS + 3305 logCOST - 264 propINPATIENT
          - 106 propMAX_COST + 80 propLAST_THREE
          - 71 propTESTS_THERAPIES + 323 logCOST_DEC
          - 2295 TT_CHEMOTHERAPY - 113 N_VISITS + 100 MONTHS_AVG
          + 20 propHOMECARE

  Rule 6: [7849 cases, mean 12946.3906, range 0 to 345440.5, est err 9152.3516]

    if
logCOST_DEC > 5.61
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -57581.717 + 6699 logCOST - 154 propINPATIENT
          + 649 logCOST_DEC - 37 propTESTS_THERAPIES
          + 2612 TT_CHEMOTHERAPY + 164 DAYS_INPATIENT + 15 propMAX_COST
          + 16 propLAST_THREE + 19 AGE + 949 N_INPATIENT
          + 294 COMORBID - 751 TT_RADIOLOGY - 133 TT_ENDOSCOPY
          + 28 propHOMECARE - 119 DAYS_ICU + 936 RENAL
          + 20 N_VISITS

  Rule 7: [2209 cases, mean 36337.8477, range 0 to 490154.5, est err 21283.3145]

    if
logCOST_DEC <= 9.64
logCOST > 10.51
    then
outcome = -494417.0352 + 10694 propVISITS + 42586 logCOST
          + 2197 propEMERGENCY + 11158 logCOST_DEC
          - 779 propTESTS_THERAPIES - 451 propINPATIENT - 1509 N_VISITS
          + 81 propMAX_COST + 189 propHOMECARE - 205 DAYS_ICU
          + 7 propLAST_THREE

  Rule 8: [278 cases, mean 106404.2344, range 0 to 534290.3, est err 52560.4688]

    if
logCOST_DEC > 9.64
logCOST > 10.51
    then
outcome = -207558.2765 + 16505 propVISITS + 31457 logCOST
          - 1209 propINPATIENT - 2760 N_VISITS + 1734 propHOMECARE
          - 382 AGE + 543 DAYS_INPATIENT + 976 DAYS_ICU
          + 4353 TT_CHEMOTHERAPY


Evaluation on training data (596047 cases, sampled):

    Average  |error|              2263.8491
    Relative |error|                   0.92
    Correlation coefficient            0.34


Attribute usage:
  Conds  Model

  100%   100%    logCOST
   46%   100%    AGE
```

```
35%   100%    logCOST_DEC
      100%    propINPATIENT
      100%    propHOMECARE
      100%    N_VISITS
      100%    TT_CHEMOTHERAPY
      100%    DAYS_INPATIENT
       99%    propVISITS
       99%    PLAN
       98%    RENAL
       97%    TT_IMAGE
       88%    N_INPATIENT
       86%    DAYS_ICU
       85%    COST_ICU
       54%    propLAST_THREE
       51%    TT_LAB
       51%    GENDER
       14%    N_EMERGENCY
        3%    propTESTS_THERAPIES
        3%    COST_COMORBID
        2%    MONTHS_AVG
        2%    CONTR
        2%    COPAY
```

Appendix F

Level 2 Classification of Causes from Global Burden of Diseases (2017) Mapped Into ICD-10 Codes

HIV/AIDS and sexually transmitted infections
A50-A58, A60-A60.9, A63-A63.8, B20-B24.9, B63, I98.0, K67.0-K67.2,
M03.1, M73.0-M73.1

Respiratory infections and tuberculosis
A10-A14, A15-A19.9, A48.1, A70, B90-B90.9, B97.4-B97.6, H70-H70.9,
J00-J02.8, J03-J03.8, J04-J04.2, J05-J05.1, J06.0-J06.8, J09-J15.8,
J16-J16.9, J20-J21.9, J36-J36.0, K67.3, K93.0, M49.0, N74.1,
P23.0-P23.4, P37.0, U04-U04.9, U84.3

Enteric infections
A00-A00.9, A01.0-A09.9, A80-A80.9, R19.7

Neglected tropical diseases and malaria
A68-A68.9, A69.2-A69.9, A75-A75.9, A77-A79.9, A82-A82.9, A90-A96.9,
A98-A98.8, B33.0-B33.1, B50-B53.8, B55.0, B56-B57.5, B60-B60.8,
B65-B67.9, B69-B72.0, B74.3-B75, B77-B77.9, B83-B83.8, K93.1, P37.1,
U06-U06.9

Other infectious diseases
A20-A28.9, A32-A39.9, A48.2, A48.4-A48.5, A65-A65.0, A69-A69.1, A74,
A74.8-A74.9, A81-A81.9, A83-A89.9, B00-B06.9, B10-B10.8, B15-B17.9,
B19-B19.9, B25-B27.9, B29.4, B33, B33.3-B33.8, B47-B48.8, B91,
B94.1-B94.2, B95-B95.5, F07.1, G00.0-G00.8, G03-G03.8, G04-G05.8,
G14-G14.6, G21.3, I00, I02, I02.9, I98.1, K67.8, K75.3, K76.3, K77.0,
M49.1, M89.6, P35-P35.9, P37, P37.2, P37.5-P37.9, U82-U84, U85-U89,
Z16-Z16.3

Maternal and neonatal disorders
N96, N98-N98.9, O00-O07.9, O09-O16.9, O20-O26.9, O28-O36.9, O40-O48.1,
O60-O77.9, O80-O92.7, O96-O98.6, O98.8-P04.2, P04.5-P05.9, P07-P15.9,
P19-P22.9, P24-P29.9, P36-P36.9, P38-P39.9, P50-P61.9, P70-P70.1,
P70.3-P72.9, P74-P78.9, P80-P81.9, P83-P84, P90-P94.9, P96, P96.3-P96.4,
P96.8

Nutritional deficiencies
D50.1-D50.8, D51-D52.0, D52.8-D53.9, E00-E02, E40-E46.9, E51-E61.9,

E63-E64.0, E64.2-E64.9, M12.1

Neoplasms
C00-C13.9, C15-C25.9, C30-C34.9, C37-C38.8, C40-C41.9, C43-C45.9, C47-C54.9, C56-C57.8, C58-C58.0, C60-C63.8, C64-C67.9, C68.0-C68.8, C69-C75.8, C81-C86.6, C88-C96.9, D00.1-D00.2, D01.0-D01.3, D02.0-D02.3, D03-D06.9, D07.0-D07.2, D07.4-D07.5, D09.0, D09.2-D09.3, D09.8, D10.0-D10.7, D11-D12.9, D13.0-D13.7, D14.0-D14.3, D15-D16.9, D22-D24.9, D26.0-D27.9, D28.0-D28.1, D28.7, D29.0-D29.8, D30.0-D30.8, D31-D36, D36.1-D36.7, D37.1-D37.5, D38.0-D38.5, D39.1-D39.2, D39.8, D40.0-D40.8, D41.0-D41.8, D42-D43.9, D44.0-D44.8, D45-D47.9, D48.0-D48.6, D49.2-D49.4, D49.6, K62.0-K62.1, K63.5, N60-N60.9, N84.0-N84.1, N87-N87.9

Cardiovascular diseases
B33.2, G45-G46.8, I01-I01.9, I02.0, I05-I09.9, I11-I11.9, I20-I25.9, I28-I28.8, I30-I31.1, I31.8-I37.8, I38-I41.9, I42.1-I42.8, I43-I43.9, I47-I48.9, I51.0-I51.4, I60-I63.9, I65-I66.9, I67.0-I67.3, I67.5-I67.6, I68.0-I68.2, I69.0-I69.3, I70.2-I70.8, I71-I73.9, I77-I83.9, I86-I89.0, I89.9, I98, K75.1

Chronic respiratory diseases
D86-D86.2, D86.9, G47.3, J30-J35.9, J37-J39.9, J41-J46.9, J60-J63.8, J65-J68.9, J70, J70.8-J70.9, J82, J84-J84.9, J91-J92.9

Digestive diseases
B18-B18.9, I84-I85.9, I98.2, K20-K29.9, K31-K31.8, K35-K38.9, K40-K42.9, K44-K46.9, K50-K52.9, K55-K62, K62.2-K62.6, K62.8-K62.9, K64-K64.9, K66.8, K67, K68-K68.9, K70-K70.3, K71.7, K74-K74.9, K75.2, K75.4-K76.2, K76.4-K77, K77.8, K80-K83.9, K85-K86.9, K90-K90.9, K92.8, K93.8, M09.1

Neurological disorders
F00-F03.9, G10-G13.8, G20-G20.9, G23-G24, G24.1-G25.0, G25.2-G25.3, G25.5, G25.8-G26.0, G30-G31.1, G31.8-G31.9, G35-G37.9, G40-G41.9, G61-G61.9, G70-G72, G72.2-G73.7, G90-G90.9, G95-G95.9, M33-M33.9

Mental disorders
F24, F50.0-F50.5

Substance use disorders
F10-F16.9, F18-F19.9, G31.2, G72.1, P04.3-P04.4, P96.1, Q86.0, R78.0-R78.5, X45-X45.9, X65-X65.9, Y15-Y15.9

Diabetes and kidney diseases
D63.1, E10-E11.9, I12-I13.9, N00-N08.8, N15.0, N18-N18.9, P70.2, Q61-Q62.8

Skin and subcutaneous diseases
A46-A46.0, A66-A67.9, B86, D86.3, I89.1-I89.8, L00-L05.9, L08-L08.9, L10-L14.0, L51-L51.9, L88-L89.9, L97-L98.4, M72.5-M72.6

Musculoskeletal disorders
I27.1, I67.7, L93-L93.2, M00-M03.0, M03.2-M03.6, M05-M09.0, M09.2-M09.8, M30-M32.9, M34-M36.8, M40-M43.1, M65-M65.0, M71.0-M71.1, M80-M82.8, M86.3-M86.4, M87-M87.0, M88-M89.0, M89.5, M89.7-M89.9

Other non-communicable diseases
D25-D26, D28.2, D52.1, D55-D58.9, D59.0-D59.3, D59.5-D59.6, D60-D61.9, D64.0, D66-D67, D68.0-D69.8, D70-D75.8, D76-D78.8, D86.8, D89-D89.3, E03-E07.1, E09-E09.9, E15.0, E16.0-E16.9, E20-E34.8, E36-E36.8, E65-E68, E70-E85.2, E88-E89.9, G24.0, G25.1, G25.4, G25.6-G25.7, G72.0, G93.7, G97-G97.9, I95.2-I95.3, I97-I97.9,

I98.9, J70.0-J70.5, J95-J95.9, K43-K43.9, K62.7, K91-K91.9, K94-K95.8, M87.1, N10-N12.9, N14-N15, N15.1-N16.8, N20-N23.0, N25-N28.1, N29-N32.0, N32.3-N32.4, N34-N34.3, N36-N36.9, N39-N39.2, N41-N41.9, N44-N44.0, N45-N45.9, N49-N49.9, N65-N65.1, N72-N72.0, N75-N77.8, N80-N81.9, N83-N83.9, N99-N99.9, P96.0, P96.2, P96.5, Q00-Q07.9, Q10.4-Q18.9, Q20-Q28.9, Q30-Q36, Q37-Q45.9, Q50-Q60.6, Q63-Q86, Q86.1-Q87.8, Q89-Q89.8, Q90-Q93.9, Q95-Q99.8, R50.2, R95-R95.9

Transport injuries
V00-V86.9, V87.2-V87.3, V88.2-V88.3, V90-V98.8

Unintentional injuries
L55-L55.9, L56.3, L56.8-L56.9, L58-L58.9, W00-W46.2, W49-W62.9, W64-W70.9, W73-W75.9, W77-W81.9, W83-W94.9, W97.9, W99-X06.9, X08-X39.9, X46-X48.9, X50-X54.9, X57-X58.9, Y40-Y84.9, Y88-Y88.3

Self-harm and interpersonal violence
U00-U03, X60-X64.9, X66-Y08.9, Y35-Y38.9, Y87.0-Y87.1, Y89.0-Y89.1

Appendix G

Conditions and Linear Models of Cubist Fitted To Detailed Medical Data: Global Burden of Diseases

G.1 Cubist With Three Rules

```
Model:

  Rule 1: [573348 cases, mean 1600.871, range 0 to 545652.4, est err 1203.170]

    if
logCOST <= 8.87
    then
outcome = -147.14 + 4596 TT_CHEMOTHERAPY + 175 DAYS_INPATIENT
          + 110 propHOMECARE + 98 N_VISITS + 401 DAYS_ICU
          - 20 propINPATIENT + 88 logCOST + 1085 DIABETES + 9 AGE
          + 66 logCOST_DEC - 5 propVISITS + 47 TT_IMAGE - 128 PLAN
          - 4 propTESTS_THERAPIES - 94 COST_ICU - 105 MATERNAL
          + 34 TT_LAB

  Rule 2: [20212 cases, mean 7971.624, range 0 to 345440.5, est err 6170.115]

    if
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -43002.511 + 5562 logCOST - 129 propINPATIENT
          - 75 propTESTS_THERAPIES + 350 logCOST_DEC + 26 propVISITS
          + 390 MONTHS_AVG + 2024 TT_CHEMOTHERAPY + 86 DAYS_INPATIENT
          + 10 propMAX_COST - 155 TT_SPECIFIC + 234 DAYS_ICU
          + 10 propLAST_THREE + 12 AGE - 227 COST_ICU - 122 TT_ENDOSCOPY
          - 38 N_VISITS + 178 CONTR + 728 DIABETES - 458 TT_RADIOLOGY
          + 46 TT_LAB - 249 COPAY + 263 N_INPATIENT - 108 MATERNAL

  Rule 3: [2487 cases, mean 44169.957, range 0 to 534290.3, est err 28713.770]

    if
logCOST > 10.51
    then
outcome = -515938.649 + 12829 propVISITS + 51083 logCOST
          - 645 propTESTS_THERAPIES - 581 propINPATIENT - 1898 N_VISITS
          + 2761 logCOST_DEC + 129 propLAST_THREE + 362 propHOMECARE
          + 153 DAYS_INPATIENT - 484 DAYS_ICU + 382 COST_ICU
          - 1094 TT_CHEMOTHERAPY


Evaluation on training data (596047 cases, sampled):

    Average  |error|           2238.702
    Relative |error|               0.91
    Correlation coefficient        0.32


Attribute usage:
  Conds  Model

  100%   100%    logCOST
         100%    propVISITS
         100%    propTESTS_THERAPIES
         100%    propINPATIENT
         100%    N_VISITS
         100%    TT_CHEMOTHERAPY
         100%    COST_ICU
         100%    DAYS_ICU
         100%    DAYS_INPATIENT
         100%    logCOST_DEC
         100%    TT_LAB
         100%    DIABETES
         100%    MATERNAL
         100%    AGE
          97%    propHOMECARE
          96%    PLAN
          96%    TT_IMAGE
           4%    propLAST_THREE
           3%    propMAX_COST
           3%    MONTHS_AVG
           3%    TT_ENDOSCOPY
           3%    TT_SPECIFIC
           3%    TT_RADIOLOGY
           3%    N_INPATIENT
           3%    CONTR
           3%    COPAY
```

G.2 Cubist With Eight Rules

```
Model:

  Rule 1: [193389 cases, mean 792.165, range 0 to 369934.8, est err 574.477]

    if
AGE <= 27
logCOST <= 8.87
    then
outcome = -147 + 84 N_VISITS + 41 propHOMECARE + 7 AGE
          + 32 DAYS_INPATIENT + 118 COST_ICU + 38 TT_IMAGE
          + 41 logCOST_DEC + 59 N_EMERGENCY - 5 propINPATIENT
          + 34 logCOST - 72 DAYS_ICU

  Rule 2: [494812 cases, mean 1332.735, range 0 to 496483.2, est err 1054.183]

    if
logCOST_DEC <= 4.81
    then
outcome = -0 + 60 N_VISITS + 48 logCOST + 77 DAYS_ICU

  Rule 3: [573348 cases, mean 1600.871, range 0 to 545652.4, est err 1196.927]

    if
logCOST <= 8.87
    then
outcome = -1872.26 + 326 logCOST + 71 N_VISITS + 754 TT_CHEMOTHERAPY
          + 147 DAYS_ICU - 8 propINPATIENT - 173 PLAN + 280 GENDER
          + 5 propTESTS_THERAPIES + 48 logCOST_DEC + 30 DAYS_INPATIENT
          + 19 propHOMECARE - 3 propMAX_COST + 4 AGE

  Rule 4: [20212 cases, mean 7971.624, range 0 to 345440.5, est err 5915.767]

    if
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -13713.97 + 1908 logCOST - 63 propVISITS - 57 propINPATIENT
          - 29 propTESTS_THERAPIES + 30 AGE + 258 MONTHS_AVG
          + 184 logCOST_DEC + 83 N_VISITS - 390 PLAN + 59 DAYS_INPATIENT
          + 743 TT_CHEMOTHERAPY - 5 propMAX_COST + 78 N_EMERGENCY
          + 5 propLAST_THREE - 147 MATERNAL + 42 TT_LAB
          + 17 propHOMECARE + 109 CONTR - 54 TT_ENDOSCOPY

  Rule 5: [754 cases, mean 12380.188, range 0 to 255325.4, est err 12495.777]

    if
logCOST_DEC <= 5.39
logCOST > 10.51
    then
outcome = -44645.589 + 1033 propVISITS + 5193 logCOST + 2038 MONTHS_AVG
          - 152 propINPATIENT - 112 propTESTS_THERAPIES
          + 95 propLAST_THREE + 507 logCOST_DEC - 177 N_VISITS
          + 31 propHOMECARE

  Rule 6: [7849 cases, mean 12946.391, range 0 to 345440.5, est err 9165.039]
```

```
    if
logCOST_DEC > 5.61
logCOST > 8.87
logCOST <= 10.51
    then
outcome = -58412.578 + 6798 logCOST - 166 propINPATIENT - 73 propVISITS
          + 759 logCOST_DEC + 23 propMAX_COST - 28 propTESTS_THERAPIES
          + 165 DAYS_INPATIENT + 2214 TT_CHEMOTHERAPY
          + 17 propLAST_THREE + 18 AGE + 1055 N_INPATIENT + 75 N_VISITS
          - 209 TT_ENDOSCOPY - 183 TT_SPECIFIC + 156 MONTHS_AVG
          - 324 MATERNAL - 900 TT_RADIOLOGY - 485 COPAY + 217 NEOPLASM
          - 114 DAYS_ICU + 16 propHOMECARE

  Rule 7: [1455 cases, mean 48753.020, range 0 to 490154.5, est err 25744.354]

    if
logCOST_DEC > 5.39
logCOST_DEC <= 9.64
logCOST > 10.51
    then
outcome = -494210.985 + 10687 propVISITS + 42573 logCOST
          + 2196 propEMERGENCY + 11155 logCOST_DEC
          - 779 propTESTS_THERAPIES - 452 propINPATIENT - 1508 N_VISITS
          + 81 propMAX_COST + 189 propHOMECARE - 204 DAYS_ICU
          + 7 propLAST_THREE

  Rule 8: [523 cases, mean 65534.273, range 0 to 534290.3, est err 36260.074]

    if
logCOST_DEC > 9.64
    then
outcome = -122975.161 + 22663 logCOST - 1410 propINPATIENT
          + 1860 propHOMECARE + 568 DAYS_INPATIENT + 1092 DAYS_ICU


Evaluation on training data (596047 cases, sampled):

    Average  |error|           2228.972
    Relative |error|               0.91
    Correlation coefficient        0.33


Attribute usage:
  Conds  Model

   62%   100%    logCOST
   39%    62%    logCOST_DEC
   15%    62%    AGE
         100%    N_VISITS
          98%    DAYS_ICU
          62%    propINPATIENT
          62%    propHOMECARE
          62%    DAYS_INPATIENT
          47%    propTESTS_THERAPIES
          47%    propMAX_COST
          47%    TT_CHEMOTHERAPY
          46%    PLAN
          44%    GENDER
```

17%	N_EMERGENCY
15%	TT_IMAGE
15%	COST_ICU
2%	propLAST_THREE
2%	propVISITS
2%	MONTHS_AVG
2%	TT_ENDOSCOPY
2%	MATERNAL
2%	TT_LAB
2%	CONTR