5-8-2015

Risk Assessment and Pricing for Group Health Claims

Shujuan Huang
University of Connecticut - Storrs, hshujuan@gmail.com

Risk Assessment and Pricing for Group Health Claims

Shujuan Huang, Ph.D

University of Connecticut, 2015

Abstract:

Risk assessment is essential for insurance pricing and risk management. This study develops several predictive models with data from a major national health insurer. Specifically, four models (lognormal, gamma, log-skew-t, and Lomax) for Episode Treatment Groups based costs are compared using four different metrics (AIC and BIC weights, random forest feature classification, and Bayesian model averaging). Several case studies are provided for illustration. Experimental results show that random forest feature classification is preferred for large data set for its computational efficiency and sufficient accuracy. For small data sets, Bayesian model averaging is recommended for its better accuracy.

Given the target variable is semi-continuous, heavy-tailed and clustered, nine candidate models are investigated including the Tweedie GLM and GAM, several two-part models, quantile regression, and a finite mixture model. A comprehensive model selection strategy and framework are suggested for different goals. A few evaluation mechanisms are investigated, considering measures of distance, effectiveness, distribution similarity, or location. In particular, the minimal distance probability matrix is proposed as a robust model selection technique. A few interesting conclusions are drawn between the transitivity of the matrix of relation and the existence of a single robust best model among candidates.

This research also develops a stop-loss coverage pricing model for self-funded health plans. The formulas that denote the net stop-loss premium are derived and predictive analytics are deployed to capture the relationship between certain characteristics and the target variable. A case study about Specific Stop-Loss (SSL) only coverage is given and future work is summarized.

Keywords: Predictive modelling, Risk assessment, Episode Treatment Groups, Stop-loss pricing, Model averaging, Model selection, Random Forest, Health Insurance Pricing, Tweedie model, Two part model

Risk Assessment and Pricing for Group Health Claims

Shujuan Huang

B.S, University of Electronic Science and Technology of China (2005)

M.S, University of Electronic Science and Technology of China (2008)

M.S in University of Connecticut (2013)

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2015

2015

APPROVAL PAGE

Doctor of Philosophy Dissertation

Risk Assessment and Pricing for Group Health Claims

Presented by

Shujuan Huang, B.S, M.S.

Major Advisor _____

Brian Hartman

Associate Advisor _____

Jeyaraj Vadiveloo

Associate Advisor _____

James G. Bridgeman

University of Connecticut

2015

Acknowledgements

The completion of this dissertation would have been impossible without the help and support from many people. I firstly wish to express my sincerest gratitude to my advisor, Prof. Brian Hartman. Brian has provided tremendous intellectual guidance with seemingly infinite patience to help me accomplish this degree. He always inspires me and shows a great deal of trust in my ability. He left tremendous space for me to explore different ideas, while he was always available when I need discussion, intuition, and advice. He lets me think, lets me grow and makes me a better person. Without his advice and support, I would not have made it.

My sincere thanks also go to my advisory committee Prof. Jeyaraj Vadiveloo for his brilliant ideas, guidance and inspiration. With two years' project experience with Jay, I have learned abundance valuable experiential skills. I would like to thank Prof. James Bridgeman for his inspiration and support on my admission to actuarial program in UConn, my research, and my teaching. His constructive comments, wise suggestions, and kind concerns help me out many times during difficult times, and make me feel warm and encouraged. I would also like to mention the guidance and insights I received from Prof. Emiliano Valdez who was my first year's advisor during my PhD study in UConn. Appreciation also goes to Prof Evarist Giné for his continuous support during my study UConn. His smile is always unforgettable and he will live in our memories.

I am also grateful to Jeffrey Young and Javier Abalo from Cigna, whose insightful advices and support are extremely important for me to complete this dissertation. I have been very fortunate to work with them on our projects. I would like to extend my sincere thanks to Monique Roy for her administrative support. Being a work from home student during the last semester, it would have been impossible for me to keep up with the required paperwork and records without

Monique's help. I would like to thank my fellow researchers Rozita Ramli, Gao Niu, Jiatian Xu, Wenyuan Zheng, Huili Tang, Ushani D. Kariyawasam, Priyantha H Katuwandeniyage. We had many thoughtful discussions in our office, and I have learned a lot from them. Thanks to my best friends Hua Ren, Xiaorong Sun and Ruoyu Dong for their endless support and encouragement.

Finally, and above all, I would like to show my deepest appreciation to my father Tianming Huang, my mother Chaoying Zhu, and my grandma Shuchao Li for their unwavering love, support, and encouragement through all my life. Thank my husband Xi Yun for his firm love, trust, support, and encouragement even in the most difficult days. Their love, expectation, and encouragement are the fuel of my work and life.

Table of Contents

Chapter 1 Introduction

1.1 The Importance of Predictive Modeling in Healthcare

Predictive analytics in healthcare has been gaining popularity as more data have been increasingly available and used in practice. Lab and diagnostic tests now provide terabytes of healthcare-related data. From the health insurer's point of view, predictive modeling can help with cost control, pricing, reserving, risk management, and marketing. More and more insurers are turning to predictive analytics for insight into future or unknown events.

Many literatures have highlighted the importance of these predictive models. As shown in Duncan's (2011) comprehensive review on healthcare risk-adjustment and predictive modeling, models for predicting health costs include the generalized linear model, tree-based models, and artificial neural networks, such as those introduced with Medicaid and Medicare. Dove et al. (2003) describe the development and validation of a predictive model designed to identify and target HMO members who are likely to incur high costs. Frees et al. (2011) model total health expenditures through multiple events using hierarchical models. Frees et al. (2014) review advanced statistical topics that aim to develop fundamentals of predictive modelling and provide corresponding applications in insurance and risk management.

In March 2010, the U.S. Congress passed the Patient Protection and Affordable Care Act (ACA), also called ObamaCare. It increases access to commercial health insurance coverage by restricting insurers from denying coverage, excluding individuals with pre-existing conditions, and varying premiums based on an individual's health status. Meanwhile, it requires all businesses with 50 or more full-time equivalent employees (FTE) to provide health insurance to at least 95% of their FTEs and their dependents up to age 26, or pay a fee beginning in 2015/2016. Health plans will be allowed to adjust premiums based only on individual-versus-

family enrollment (i.e., individual, individual + dependent(s), etc.); geographic area; age (which cannot vary by more than 3 to 1 among adults); and tobacco use (which cannot vary by more than 1.5 to 1). Other factors that insurers traditionally use to calculate accurate premiums, such as health status, use of health services, and gender, will no longer be allowed under the ACA.

Given the timing of the law, ACA-influenced trends will affect how predictive modeling is used to improve pricing and risk management. For example, group health insurance is purchased by an employer and offered to eligible employees (and family members) as a benefit, but related laws and plan details can vary significantly by state and by employer. The ACA requires states to create and operate exchanges for individuals and small businesses to purchase insurance. While premium rates for small employer groups are regulated by federal law, large group health insurance policies are usually underwritten at the time of purchase, with rates adjusted based on employee participation and prior claims experience. Regardless of the size of the business, risk assessment and mitigation techniques are important for health insurance companies. Society of Actuaries(SOA) has completed a few experience studies for group insurance, such as the 1991-92 Group Medical Insurance Large Claims Database, Medical Large Claims Experience Study, Risks & Mitigation for Health Insurance Companies (Rosenblatt and Segal, 2013), Uncertainty in Risk Adjustment (Mehmud and Yi, 2012), A Comparative Analysis of Claims-Based Tools for Health Risk Assessment (Winkelman and Mehmud, 2007), and Cost of the Newly Insured Under the Affordable Care Act (2013). The trends in healthcare research projects of SOA show the desirability and popularity of predictive modelling in healthcare.

1.2 Motivation and Research Goal

1.2.1 Episode Treatment Groups (ETGs)

Symmetry Episode Treatment Groups (ETGs) were introduced and patented by OPTUM as an episode grouper for medical and pharmacy claims. They combine related services into a distinct medically relevant unit describing a complete episode of care, thus applying to diverse groups such as healthcare providers, researchers and administrators. It is worthwhile for a healthcare insurer to investigate ETG-based cost for each patient to project future losses. Symmetry ETGs are currently used by more than 300 healthcare plans and their providers in the United States. In spite of its wide use, how to effectively extract signal from those ETGs for more accurate insurance pricing and better risk management is still an outstanding issue. There is a gap between historical ETG costs and potential losses for our current and future policy holders. However, since ETG data is private and expensive, they are not widely studied in academic literature. The necessity and desirability of ETGs in health plan pricing created an opportunity between us and a major healthcare insurer. This research serves as a starting point and illustration for application of ETGs in healthcare predictive modeling. Research into innovative applications of ETGs is expected to continue to grow.

1.2.2 Claim-based Risk Assessment and Pricing

Claim-based risk assessment in healthcare is the process of determining the relative costs of a person or a group based on their medical history, demographics, regions, etc. The goal of the research is to build and develop innovative and holistic predictive risk models for both existing and prospective customers from the perspective of health insurance companies. Healthcare actuaries can benefit from an in-depth understanding of risk assessment and risk adjustment as it is a key driver of the bottom line of healthcare organizations. In the process of claim-based risk assessment and pricing, the analysts aim at translating risk metrics into monetary quantities in terms of gains and losses, and develop pricing models to improve margins and profit over the predicted risk while assuring market share through acquisition and retention. There are more than

500 covariates in our dataset. Unnecessary predictors will add noise to the estimation of other quantities that we are interested in, and collinearity may be caused by having multiple variables describing the same policyholder features. I aim to interpret the data in the simplest way--redundant predictors should be removed. A few advanced predictive modelling techniques are investigated in this dissertation, which can be used by actuaries to gain competitive advantages in situations with complex data. Each model has its own advantages and disadvantages, making model selection a very interesting, necessary and desirable task. I aim at developing a comprehensive and data-based model selection framework and strategy for practitioners. Through the research, we can gain insight not only on which models perform better in terms of different goals, but also whether the model selection measures themselves are effective and efficient. I plan to explore model (or variable) selection within the model and model selection among different types of models; and finally summarize the model selection strategies and suggestions to data analysts or actuaries in health insurance industry.

For group health insurance coverage, the rating process usually begins with the development of claim costs, incorporating pooled claims and pooling charges. I will especially concentrate on the stop-loss pricing in this dissertation. In health insurance, stop-loss coverage is a policy designed to protect a self-funded employer from catastrophic losses. It usually takes effect after a certain amount has been paid. Employers providing health insurance through a self-insured plan often subscribe to stop-loss policies for risk management. In fact, high claims are very unpredictable and volatile in practice; incorrect pricing of stop-loss coverage can create huge losses. Hereby, we are trying to use more powerful predictive and data mining techniques to capture the relationship between certain characteristics and the target variables compared to traditional actuarial practice. In contrast to study losses only on the aggregate level, one can also investigate them on the individual and group levels, which will provide more detailed information to build the models and improve the pricing.

1.3 Scope of Study

1.3.1 Model Selection and Averaging of Health Costs in Episode Treatment Groups

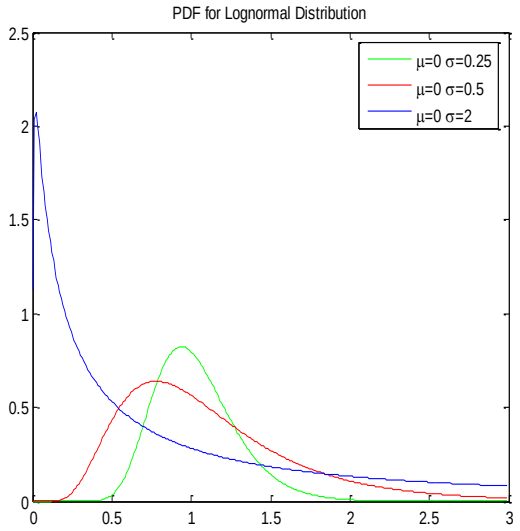
Given the cost of information on each ETG for each policy holder with that major healthcare insurer during 2012, I am trying to make full use of this information for risk assessment, or disease-specific insurance product design and pricing. There are 320 non-routine ETGs in all, such as hemophilia, arterial trauma, eating disorder, and heart failure. Proper model selection for those ETG-based costs is essential to modeling. The optimal model (or model probabilities) can change depending on the disease. It is well recognized that insurance loss distributions are strongly skewed with heavy tails. Fitting an adequate loss distribution to insurance data is a difficult and critical task in actuarial literature. When one model is dramatically better than the others, only knowing the best model will be sufficient. However, when the potential models are very similar in their fit for some data sets, the model averaging techniques enable us to average the fits for a number of models, instead of using only a single best model. A simulation should account for that model uncertainty by drawing a proportion of the simulations from each of the models that fit the data well. It gives the analyst greater insight into the relative merits of the competing models.

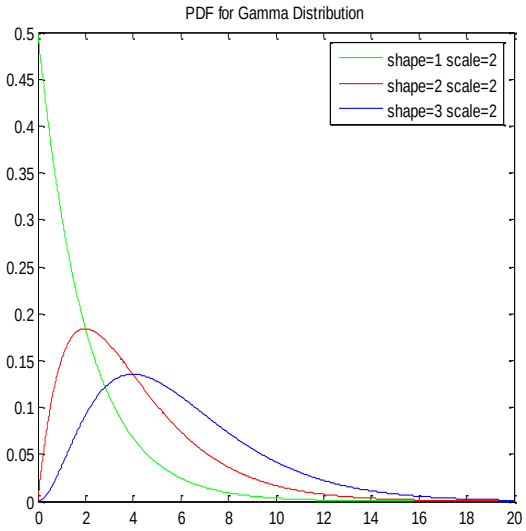
I considered four continuous probability distributions in our paper: lognormal, gamma, Lomax, and log-skew-t. Lognormal and gamma distributions are widely used in numerous fields (see, e.g. Kleiber and Kotz, 2003). Their density functions are summarized in table 1.1. The Lomax distribution is a Pareto distribution that has been shifted so that its support begins at zero (see, Klugman et al., 2012). The log-skew-t distribution is a continuous probability distribution of a random variable whose logarithm is skew-t distributed. The skew-t distribution generalizes the t distribution to allow for non-zero skewness. The skew-t distribution is extensively investigated as a promising candidate for both theoretical and empirical work in actuarial science (see, e.g., Ferreira and Steel, 2007; Jones and Faddy, 2003; Eling, 2012). The density functions of the four

distributions have different shapes and tail thicknesses, but all have been used in business, economics, and actuarial modeling. Hence it is desirable and necessary for us to explore model averaging among them.

Table 1.1 Summary on potential distributions

Distributions	Density	Support
Lognormal	$\frac{1}{x\sqrt{2\pi\sigma}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$	$(0,+\infty)$
Gamma	$\frac{1}{\tau(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$ Where k > 0 shape, θ > 0 scale.	$(0,+\infty)$
Log skew T	If $Y \sim \log skewT \;\; X = ln(Y)$ Then $X \sim skewT$ and $f_Y(y) = \frac{1}{y} f_X(\ln(y))$ $f_{ST}(y;\xi,w^2,\alpha,v) = \frac{2}{w} t_v(x_y) T_{v+1}\left(\alpha x_y \sqrt{\frac{v+1}{v+x_y^2}}\right), x_y = \frac{x-xi}{w}$	$(0,+\infty)$
Lomax	$\frac{\alpha}{\lambda}\left[1+\frac{x}{\lambda}\right]^{-(\alpha+1)}$ Where $\alpha > 0$ shape, $\lambda > 0$ scale	$[0,+\infty)$





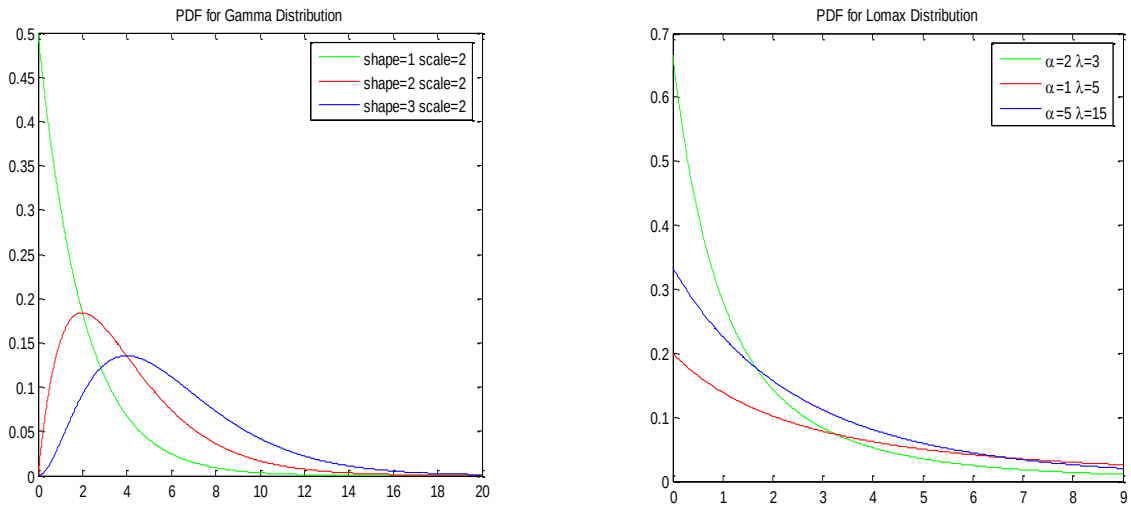


Fig. 1.1 Density functions of lognormal, gamma, log skew T and Lomax distribution

Given all the models under consideration, relying on only one or two measures of model fit would produce more definitive results, as each information criterion is designed to identify the best candidate model in a particular well-defined sense. Alternatively, multiple measures can capture more of the complexity in the model selection problem. Traditionally, we compare log-likelihood based information criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to choose the distribution with the best fit. AIC and BIC are maximum likelihood estimate driven and try to balance good fit with parsimony. BIC generally penalizes free parameters more harshly than AIC, but in our experiments their results are quite similar. However, many studies such as Shtatland et al. (2000) show that information criteria have disadvantages. For example, they assume a single optimal model (AIC- or BIC-optimal) and can be computationally intense when the data sets are large. In addition, Kuha (2004) shows that both AIC and BIC are good approximations of their own theoretical target quantities, but both criteria can still fail in this respect, even in the very simple examples considered in his research. Rather than simply looking for an AIC-optimal or BIC-optimal model, I calculate the AIC and BIC weights. These weights can be easily calculated from the raw AIC/BIC values, and provide a

straightforward interpretation as the probabilities of each model being the best model in an AIC or BIC sense.

Further, Bayesian inference and parallel model selection are studied in this dissertation. Bayesian model averaging can provide the probabilities of each model being the best given the data and all models under consideration, enabling model averaging and providing deeper insights into the relationships between the models. Several methods for calculating these probabilities have been suggested in literature: RJMCMC, the saturation method (Carlin and Chib, 1995) and parallel model selection (Congdon, 2006). In actuarial literature, Hartman and Groendyke (2013) discussed model selection and averaging in the financial risk management context.

In this study, we have more than 33 million observations, but only a Thinkpad with a 2.50 GHz Intel Quad-Core processor and 8 GB RAM is used for experiments. The first three metrics (AIC, BIC weights and Bayesian model averaging) struggle with big data in terms of processing time. Therefore, it is desirable to find a more efficient approach to select the best fitting distribution. Random forests are very popular ensemble learning methods for classification (or regression) in data mining and have been widely used in diverse areas (Liaw and Wiener, 2002). They are also highly efficient when compared to the other three metrics. At the first look, there seems to be no connection between the two, one is for distribution fitting and selection, and another one is for classification. However, if I view all data sets following one distribution as one cluster, selecting the best distribution is equivalent to putting the observations into the correct cluster. Schwartz et al. (2014) used classification techniques (decision trees) to make model recommendations for a common marketing problem (i.e., forecasting repeat purchasing incidence for a cohort of new customers). They showed the method's capability to discriminate among an integrated family of a hidden Markov models and their constrained variants in managerial contexts, even outside of the HMM framework. In this dissertation, I share a similar idea in model selection using

classification techniques, but apply to different problem settings using different techniques. Several case studies will be provided and compared using all four metrics on all potential models.

1.3.2 Risk Assessment and Pricing in Healthcare

Claim-based risk assessment in healthcare aims at determining the relative costs of a person or a group based on their medical history, demographics, regions, etc. Our target variable has a combination of a point-mass at zero and a right skewed distribution. When the number of zeros is more than expected under a standard continuous distribution, the data is said to be semi-continuous and I need to use the models that are tailored for semi-continuous outcomes. As shown in fig 1.2, if I am trying to fit the semi-continuous data by a single distribution on the left, the density curve won't be a good fit. I wish to build the models to be able to capture the density shape like the one on the right in fig.1.2.

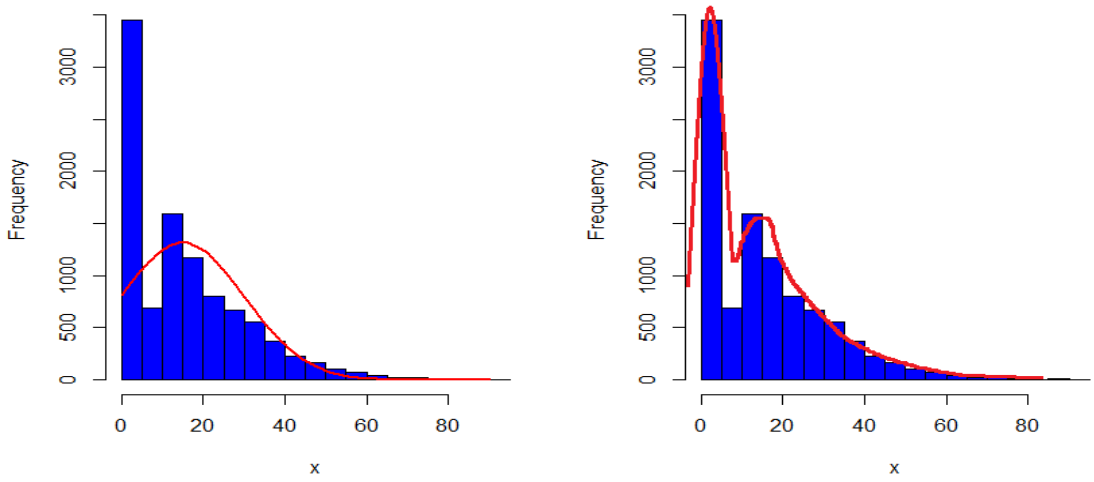


Fig.1.2 Histogram and density for semi-continuous data

Meanwhile, when estimating loss distributions with heavy tails in insurance, it is difficult to find a simple parametric model that fits all claim sizes; thereby large and small losses are usually split in traditional actuarial practice. But that approach involves determining the threshold level between large and small losses, which can be subjective and introduce bias conclusions. In

contrast, a heavy-tailed model is a unified approach to the estimation of loss distribution, where the heavy-tailed distribution is defined by its structure of the decline in probabilities for large deviations. The models are more data-driven and objective than traditional models. In this dissertation, due to the semi-continuous and heavy-tailed properties of the outcome variables, the modeling framework for claim-based risk assessment is summarized in fig. 2.1. In general, there are four types of semi-continuous models discussed in this dissertation: Tweedie, two-part, quantile regression and finite mixture models. The Tweedie and two-part models have their subtype models. Meanwhile, in order to capture the dependence between observations within groups, hierarchical models are investigated. I will discuss why those models apply and their pros and cons in table 1.2.

Table 1.2 Why those model candidates apply

Features of target variable	Model Candidate	Justifications and Comments
Semi-continuous	Tweedie	The Tweedie distribution (Compound Poisson-Gamma Distribution) has nonnegative support and can have a discrete mass at zero with certain parameter values, making it useful to model responses that are a mixture of zeros and positive values. It usually requires fewer parameters than other models, but requires strong distributional assumptions about the target.
	Two-part	Semi-continuous data can be viewed as arising from two distinct stochastic processes: one governing the occurrence of zeros and the second determining the observed value given a nonzero response (Neelon, 2013). The first process is commonly referred to as the binary part of the data, and the second is often termed the “intensity” or “continuous” part. Two-part models explicitly accommodate both data-generating processes. They usually have more parameters, but flexible enough to allow different subsets of variables for two parts respectively.
	Quantile Regression	Regression quantile estimates can be used to construct prediction and tolerance intervals without assuming any parametric error distribution or variance heterogeneity. It would be appropriate to use quantile regression to estimate conditional quantiles for high-dimensional predictor variable when focusing on those quantiles beyond the zero part in this dissertation.

	Finite mixture model	Semi-continuous models can be expressed as at least two-component mixtures in which one component has a degenerate distribution at zero and the other component is a positive support model. In fact, depending on the distribution, we can have more than one component for the positive part, which is very flexible to fit multimodal or heavy-tailed densities in addition to the zero part.
Heavy-tailed	Transformation/ Rescaling	A nonlinear transformation changes (increases or decreases) linear relationships between variables and the correlation between variables.
	GLM	Generalized linear model (GLM) is a generalization of ordinary linear regression . It allows for response variables to have error structures other than a normal distribution. GLM also generalizes linear regression by allowing the linear part to be related to the response variable through a link function. In addition, the magnitude of the variance of each measurement is allowed to be a function of its predicted value.
	GAM	"A generalized additive model (GAM) is a generalized linear model in which the linear predictor depends linearly on unknown smooth functions of some predictor variables, and interest focuses on inference about these smooth functions." – Hastie and Tibshirani (1990). GAMs are extremely flexible models for fitting smooth curves to heavy-tailed data.
	Finite mixture model	A finite mixture model is a convex combination of two or more probability density functions. By combining the properties of the individual probability density functions, mixture models are capable of approximating any arbitrary distribution including heavy tailed distributions.
	Quantile regression	In a heavy-tailed environment, median will be a better measure of location than mean. Quantile regression will be a natural fit that is capable of estimating conditional quantiles for high-dimensional predictor variables in heavy-tailed situations.
Dependence	GLMM	A mixed model is a model that contains fixed and random effects. For analysis of multilevel data, random cluster and subject effects can be added into the regression model to account for the correlation of the data. For our healthcare data, all datasets are collected with an inherent multilevel structure: individual policyholders clustered within accounts. GLMM would be a nature fit which is able to deal with clustered data and parameters which vary by group (account).
	GEE	Compared to conditional GLMM, the GEE estimates are "marginal" in the sense that the parameter estimates themselves are indifferent to the grouped structure of the data.

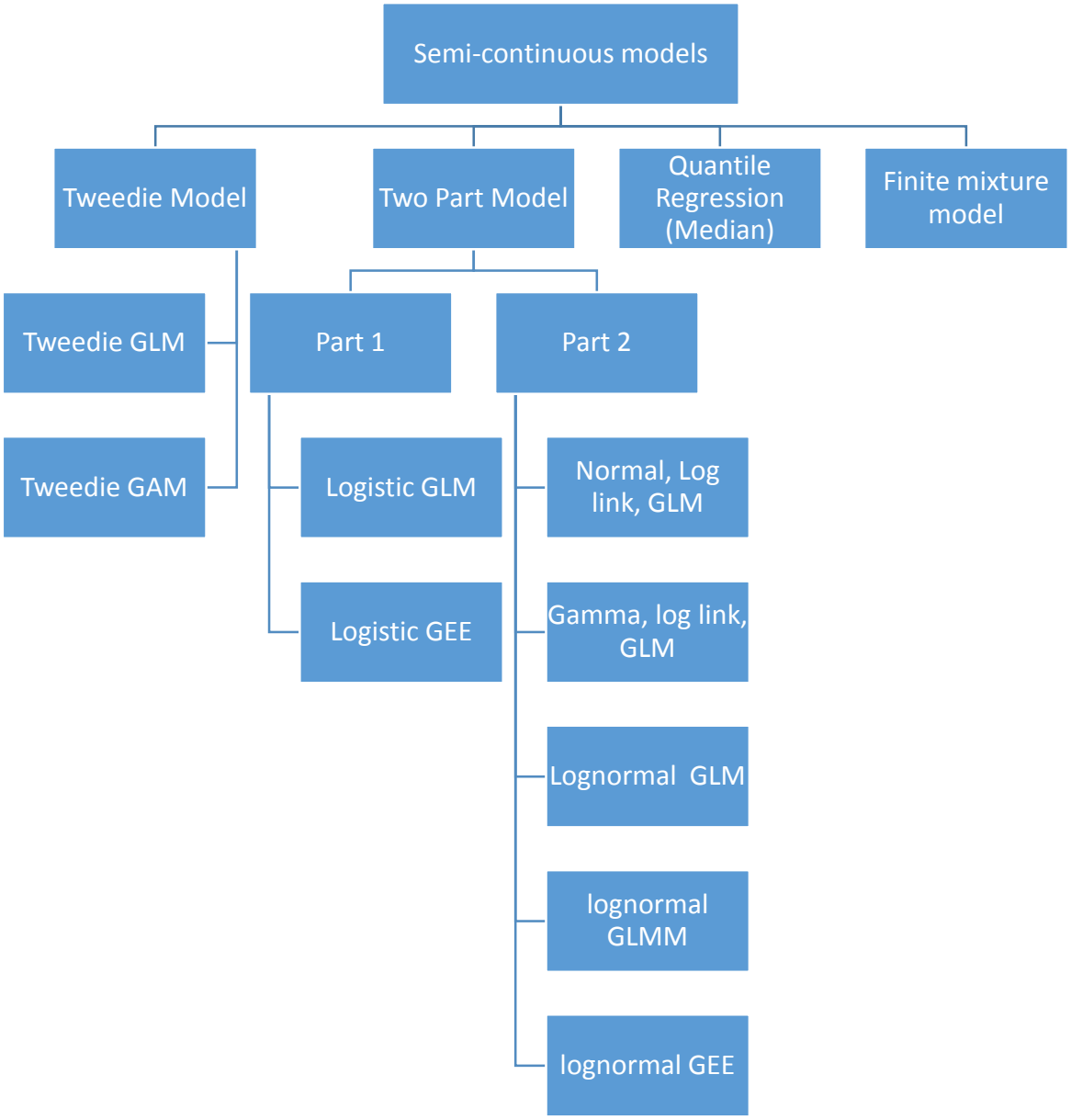


Fig 1.3 Risk Assessment for group health claims modeling framework

1.4 Contributions to Literature

1.4.1 Model Selection and Averaging of Health Costs in Episode Treatment Groups

ETGs (Episode Treatment Groups) were introduced for identifying and classifying an entire episode of care for evidence-based medicine and healthcare management reporting. In spite of

their wide use, how to effectively use ETGs for health plan pricing is still an outstanding and interesting issue from health plan carrier point of view. This research aims at investigating the application of ETGs in health plan pricing and risk management, with a focus on model selection and averaging. The optimal model (or model probabilities) can vary depending on the disease. Insurance loss distributions are commonly skewed with heavy tails. Using lighter-tailed distributions for modeling may significantly bias the results; however, this issue has not been carefully addressed in many situations of actuarial practice. Therefore, in this dissertation, I compare four potential models: lognormal, gamma, log-skew-t, and Lomax; where gamma is the default distribution for positive continuous response variables in practice. However, in my experiments, none of the metrics select the gamma distribution as the best model for any of the 320 different ETGs. Thus, one needs to be cautious in building a gamma model for heavy-tailed data.

In addition to model selection and averaging, this dissertation also contributes by recommending various metrics for different data sizes and goals of the analyst. Four metrics are considered: AIC weights, BIC weights, Bayesian parallel model selection and random forest feature classification. AIC and BIC are commonly used maximum likelihood estimate driven information criteria, and they try to balance good fit with parsimony. BIC generally penalizes free parameters more strongly than AIC, but in our experiments their results are quite similar in most cases. Bayesian parallel model selection yields the probabilities of each model being the best given the data among all models under consideration, enabling model averaging and providing deeper insights into the relationships between the models. Since we have 33 million ETG cost observations from 9 million claimants, I proposed random forest feature classification in order to achieve greater efficiency. In other words, if we treat all the data sets following one distribution as one cluster, selecting the best distribution is equivalent to allocating the observations into the correct cluster. I divided the classification process into three steps: domain specific feature extraction, random forest training

for prediction and random forest model selection. The moment-based features outperform percentile-based features in distinguishing distributions. If I use both moment-based and percentile-based features, we can achieve the lowest out-of-bag error rate and the best performance in distinguishing distributions. Since the random forest model selection is based on the extracted information rather than the original big data sets, it can significantly reduce computing time. The results show that random forest only takes 2 minutes for the whole process, but AIC/BIC needs around 4 hours. Bayesian parallel model selection may need approximately 4 weeks on the same task. Furthermore, the accuracy among the four metrics is compared. On average, the Bayesian approach performs best because it exactly identifies lognormal and log-skew-t distribution, though it is less certain about gamma and Lomax compared to AIC weights. AIC weights also did a good job on average. Random Forest performs a little bit worse than the other two, but it still can identify the model with the best fit. Especially when we need to deal with big data, its efficiency is valuable without losing much accuracy.

1.4.2 Risk Assessment and Pricing

As predictive analytics in healthcare gain popularity in practice, actuaries and other financial analysts are facing challenges and opportunities in the best practice and innovation. A bunch of new advanced statistical and data mining techniques are developing every day in academia and data science industry, while terabytes of information are collected in insurance industry waiting for evaluation and analysis. Traditional actuarial models are more frequently questioned and challenged by predictive analytics; while predictive analytics are questioned by their complicity and interpretability in practice. There seems to always be a gap between what we have and where we want to arrive. Luckily, my research comes from a project with a major national health insurer and I can get access to most recent real industry data. The topics in this dissertation come from practice and go back to practice. Some of the models are actually being used now in practice and

I got a lot of valuable feedback from practitioners. Meanwhile, from research point of view, getting access to data is always a fantastic starting point. Rather than building the models in a simulation world full of assumptions, we get a chance to build a framework of down to earth models with real data from industry, and let them speak for themselves with objective and fact-based evidence.

In particular, the target variable in this study is semi-continuous, heavy-tailed and clustered; hence all the candidate models are tailored for those properties, which are very common for insurance loss data, not just in healthcare, but also in life insurance, property and casualty. In general, four types of semi-continuous models are investigated in this dissertation: Tweedie model, two part model, quantile regression and finite mixture model. Tweedie model includes Tweedie Generalized Linear Model (GLM) and Tweedie Generalized Additive model (GAM). Two part models include generalized linear mixed model (GLMM), generalized estimating equation (GEE), GLM with various choices on distribution assumptions and link functions. Each model has been carefully fitted and diagnosed with appropriate goodness-of-fit statistics. Meanwhile, various variable selection techniques were carefully discussed and compared. To sum up, the first contribution of the second part is a comprehensive claim-based risk assessment analytical framework using real industry data with nice promising model candidates. The comparable results and methodologies took the veil off different linear or nonlinear models, hence inspiring continuous efforts on improving the models theoretically. The results shown in the dissertation can be used as a good reference for data analysts or actuaries in healthcare insurance industry for modelling, risk management or pricing.

In this dissertation, I proposed the minimal distance probability matrix as a powerful and robust model selection technique, where the idea comes from Game Theory. Following the idea of Arrow's “impossibility” theorem, I believe one on one comparison on minimal distance counting will give us the most unbiased and robust information for decision making. Rather than using an

aggregate measure, the benefit of the minimal distance probability matrix lies in the capability to compare and count every prediction on the individual level; in that case a few extremely biased predictions won't distort the overall results like other aggregate distance measures. Meanwhile, I can derive the relation matrix from the probability matrix. Investigating the properties of the relation matrix can help us choose the best model selection strategies. Some interesting conclusions are as follows:

- When the matrix of relation is transitive, there exists a single robust best model based on the individual level absolute distance measure; we can get the unique maximal value of total vote when counting the votes for each model.

- When the matrix of relation is not transitive, there is no single robust best model based on the individual level absolute distance measure, but other alternative strategies can be used such as model averaging.

This dissertation also contributes to the development of an objective and comprehensive model selection framework and strategies for decision making from different perspectives. In general, there are two steps, model selection within the model and model selection among different types of models. I aim at not only selecting models with the best model fit and prediction accuracy, but also investigating the efficiency and effectiveness of evaluation techniques themselves. Different evaluation measures are suggested for different prediction goals and priorities. A few evaluation mechanisms discussed in this dissertation are: measure of distance (such as MSPE, MAPE, and minimal distance probability matrix), measure of effectiveness (lift chart), measure of distribution similarity (such as histograms and Bhattacharyya coefficient) and Measure of location (mean, median, trimmed mean, and Winsorized mean). It is also worth mentioning that most literature says that one cannot use lift charts to measure the accuracy of models that predict continuous

numeric values. In this dissertation, I discussed this issue thoroughly and I believe we still could use lift charts to measure the accuracy of models with continuous numeric values; even though the way to interpret the results will be quite different than those with discrete values. If the target variable is continuous, gains chart provides us with statistics relative to the mean of the target variable. To be specific, first, we need to check the cumulative mean response curve. Only when this curve is monotone decreasing, we have reasons to believe this model will be more effective than no predictive model. Unlike cumulative gains for binary or count data, higher lift doesn't mean the corresponding model is more effective because we want the prediction to be close to the actual, not over-predicting. We only care about whether this curve is monotonically decreasing. Second, we need to check whether the mean predicted response curve and the mean actual response curve. It is suggested that the mean actual response curve of a good predictive model should be monotonically decreasing too; and the closer the two curves, the more effective the model will be.

The third contribution of this part is in stop-loss pricing for self-funded health plans. After a review of traditional actuarial models for stop-loss pricing, I raised my concern about a few highly simplified assumptions in traditional actuarial models; and pointed out that high claims are very unpredictable and volatile in practice; incorrect pricing of stop-loss coverage can create huge losses. In this dissertation, first, the formulas that denote the net stop-loss premium are derived in terms of left censored and shifted variables, as well as limited loss variables. Then, predictive analytics are used to capture the relationship between certain characteristics and the target variable. A case study about SSL only stop-loss insurance is given and future work are summarized. The approach and solutions using predictive modelling will contribute to a more efficient stop-loss healthcare insurance market.

Chapter 2 Modeling Techniques in Predictive Analytics

2.1 Modeling Techniques for Semi-continuous and Heavy-tailed Data

2.1.1 Generalized Linear Models (GLM)

2.1.1.1 Overview

Generalized linear models, introduced by Nelder and Wedderbur (1972), generalizes linear models to non-normal data. Before the generalized models were developed, modeling of right skewed and heavy tailed data typically relied on transformations of the data. The transformations were usually chosen to improve symmetry and normality. Those transformations, however, have implication for the error structures of the models. Moreover, back-transforming estimates may introduce bias. In contrast, generalized linear models apply a transformation, known as the link function, to the mean of the data. Besides, in a GLM, the outcome of the dependent variable *Y* is assumed to be generated from a particular distribution in the exponential family. The mean μ of the distribution depends on the independent variables *X* through the following formula:

$$E[Y] = \mu = g^{-1}(X\beta) \ or \ \ \mathrm{g}(E[Y]) = X\beta \tag{2.1}$$

$$Y \sim Distributions\ in\ exponential\ family$$

where $E[Y]$ is the expected value of Y and $X\beta$ is the linear predictor.

2.1.1.2 Normal GLM with log link vs Lognormal GLM

Normal GLM with log link assumes normal distribution for the response variable and uses log as the link function. The log link function applies to a deterministic component, the mean of the data, not to each data point. In contrast, lognormal GLM assumes the lognormal distribution for the response variable and uses identity link function. In certain cases, we can take the log transformation of each data point and make use of the relationship between normal distribution and lognormal distribution to derive the lognormal GLM.

Let's start with Normal GLM with log link

$$E[Y] = \exp(X\beta) \ or \log(E[Y]) = X\beta, \quad Y \sim Normal \tag{2.2}$$

where $E[Y]$ is the expected value of Y ,and $X\beta$ is the linear predictor. Next, let's move to lognormal GLM. We can treat the logarithm of the response variable as a normal distributed random variable. The mean and variance are estimated on the logarithmic scale, assuming a normal distribution. To convert means and variance for log(Y) into those of Y, we can use the following relationship:

$$E[Y] = exp(\mu)\sqrt{w} \tag{2.3}$$

$$Var[Y] = \exp(2\mu)\, w(w-1) \tag{2.4}$$

$$w = \exp(\sigma^2). \tag{2.5}$$

That's because if the random variable X follows log-normally distribution, then $Y = log(X)$ is normally distributed. Likewise, if Y has a normal distribution, then $X = exp(Y)$ has a log-normal distribution. A random variable which is log-normally distributed has only positive real values support. Be noted that if $X \sim ln\, N(\mu, \sigma^2)$, then

$$E[X] = exp(\mu + \frac{\sigma^2}{2}) \tag{2.6}$$

$$Var[X] = (e^{\sigma^2} - 1)exp(2\mu + \sigma^2) \tag{2.7}$$

2.1.2 Generalized Additive Models (GAM)

2.1.2.1 Overview

Compared to GLM, Generalized Additive Models (GAM) estimates an additive approximation to the multivariate regression function. GAMs were originally developed by Hastie and Tibshirani (1990) to blend properties of generalized linear models with additive models. Rather than fitting multiple variables simultaneously, the algorithm of GAM fits a smooth curve to each variable and then combines the results additively. Stone (1985) indicates that the advantages of an additive

approximation are at least twofold. First, it avoids the curse of dimensionality by estimating the individual additive terms using a univariate smoother, at the price of losing the ability of universal approximation. Second, it can tell how the dependent variable changes with the independent variables from the estimates of the individual terms. In addition, nonparametric regression relaxes the usual assumption of linearity and enables people to uncover structure in the relationship between the independent variables and the dependent variable that might otherwise be missed. The general formula for GAM is shown in formula (2.8)

$$E[Y] = f(X_1, X_2, \dots, X_n) = s_0 + s_1(X_1) + \cdots + s_n(X_n) \tag{2.8}$$

where $s_i(X_i)$, i = 1, ... , p are smooth functions.

These functions are estimated in a nonparametric way. A combination of backfitting and local scoring algorithms is used in the fitting of the model through either Proc GAM in SAS or package *mgcv/gamlss* in R; although other alternative inference methods are available as well. In nature, a spline is a numeric function that is piecewise-defined by polynomial functions. It possesses a sufficiently high degree of smoothness at the places where the polynomial pieces connect. The most commonly used splines are cubic spline. A cubic spline is essentially a connection of multiple cubic polynomial regressions. Michael Clark (2009) explains that we choose points of the variable at which to create sections, and these points are referred to as knots. Separate cubic polynomials are fit at each section, and then joined at the knots to create a continuous curve.

2.1.2.2 Estimation

In this study, estimation of GAM is conducted with a penalized likelihood approach through the mgcv package in R by Wood (2015). Suppose we have the GAM as follows:

$$g(\mu) = X\beta + f(x_1) + f(x_2) \dots + f(x_n) \tag{2.9}$$

But note that each smooth has its own model matrix made up of the bases. So for each smooth covariate we have:

$$f_j = \tilde{X}_j \tilde{\beta}_j \tag{2.10}$$

Given a matrix of coefficients S, penalized likelihood function would be:

$$l_p(\beta) = l(\beta) - \frac{1}{2}\sum_j \lambda_j \beta^T S_j \beta \tag{2.11}$$

where $l(\beta)$ is the usual GLM likelihood function, and λ_j are the smoothing parameters.

Hastie and Tibshirani (2009) explains that λ establishes a trade-off between the goodness of fit and the smoothness, where the smoothing parameters are estimated by cross-validation procedure. Output of GAM is usually separated into parametric and smooth (or nonparametric) parts. In general, we are trying to seek a balance between an undersmoothed fit and an oversmoothed fit (Clark, 2009).

2.1.2.3 GLM vs GAM

GAMs are extremely flexible for fitting smooth curves to data. In most cases, they often achieve results superior to GLMs, at least in perspective of goodness-of-fit. However, they are somewhat hard to evaluate or interpret, because they lack a parametric equation for the results. Indeed, some analysts prefer GLMs to GAMs because GLMs provide better insight for the results, even if the goodness-of-fit of the results of GLMs is lower. Many analysts fit GAMs as a way of determining the correct curve shape for GLMs, and deciding the order of polynomials as suggested by the GAM plot. In practice, it is usually preferable to rely on a simple yet well understood model for predicting future cases, than on a complex model that is difficult to interpret and summarize.

2.1.3 Tweedie Model

The Tweedie distribution, introduced by Tweedie, M.C.K. (1984), has nonnegative support and can have a discrete mass at zero. It is attractive to model responses that are a mixture of zeros and positive values. The Tweedie distribution belongs to the exponential family, so it conveniently fits into the generalized linear models framework. When the Tweedie index is between 1 and 2, Tweedie model is known as compound Poisson exponential dispersion model. It also includes the purely continuous normal and gamma distributions, and the purely discrete scaled Poisson distribution. For any random variable Y that is Tweedie distributed, the variance $\mathrm{var}(\mathrm{Y})$ relates to the mean $\mathrm{E}(\mathrm{Y})$ by the power law as follows

$$var\ (Y) = \emptyset[E(Y)]^p \tag{2.12}$$

where both $\emptyset$ and p are positive constants, $\emptyset$ is the dispersion parameter and p is an extra parameter that controls the variance of the distribution (the Tweedie Index). To be specific, we have

- Normal distribution, $p = 0$
- Poisson distribution, $p = 1$
- Compound Poisson–gamma distribution, $1 < p < 2$
- Gamma distribution, $p = 2$
- Positive stable distributions, $2 < p < 3$
- Inverse Gaussian distribution, $p = 3$
- Positive stable distributions, $p > 3$
- Extreme stable distributions, $p = \infty$
- No Tweedie model exists, $0 < p < 1$

Fig. 2.1 shows simulation of Tweedie model for different values of $p\ (1 < p < 2)$ and $\emptyset$.

For semi-continuous target, the strength of Tweedie model lies in its ability to fit a wide range of distributions of data, especially the dataset with a huge spike in the distribution at 0 where other standard distribution are not able to capture. However, it also has its limitation. It requires strong distributional assumptions about the target. We also need to be careful about the sensitivity of Tweedie index p. And our experiment shows that if we use Tweedie model for prediction, it cannot generate predictions as exactly zero. We will discuss the Tweedie model with more details in chapter 4.

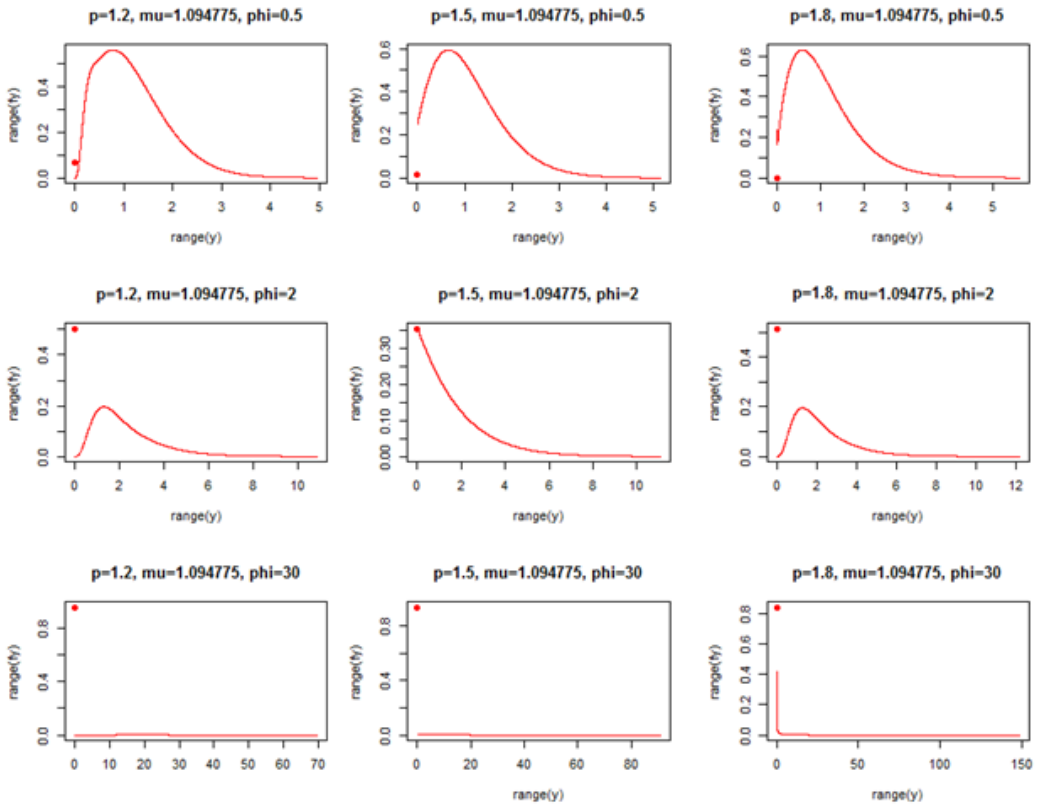


Fig.2.1 Simulation of Tweedie distributions

2.1.4 Two Part Model

This section discusses two part model. We associate the healthcare costs with two components: the event of the positive costs and its amount, if the costs exist. Therefore, the first part is to check whether there will be healthcare cost associated with the participant, and part 2 predicts the

severity, given the cost is greater than zero. In the first part, the logistic regression can be used on the binary target. In the second part, multiple choices are available for the amount target such as lognormal GLM, gamma GLM. Then the prediction of the interval target can be computed from the value model and adjusted by the posterior probabilities of the class target. Or if we don't want to multiply the prediction of the interval target, you can use a filter to reset the predicted values of observations with non-event class prediction to 0. The procedure of two part model is shown in figure 2.2. The transfer function means the probability predicted in the first part can be used as an explanatory variable in the second part.

- Part 1: Will the policy holder has positive healthcare costs (Logistic Regression)
- Part 2: Average healthcare cost per person given the claim occurs (Heavy-tailed amount models)

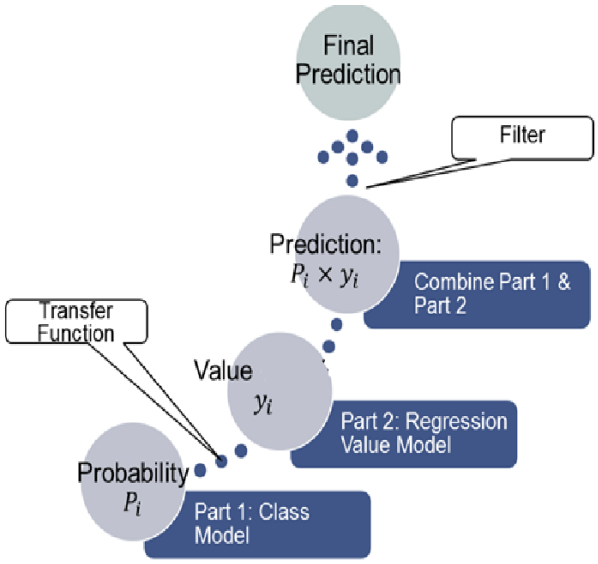


Fig 2.2 Procedure of two-part models

2.1.5 Quantile Regression Forest

2.1.5.1 Quantile Regression Overview

For ratemaking purposes, actuaries are always concerned with the appropriate measure of the center of the cost distribution. Current practice focuses more on the mean regression models and on the expectation of the response. Actually, if one is interested in the center of a distribution, why not use the median instead of the mean to measure the location? Motivated by this question, this study investigates median regression, which is distribution-free and emphasizes the relation between covariates and the median of the response variable.

Quantile regression, introduced by Koenker and Bassett (1978), extends these ideas to the estimation of conditional quantiles, where quantiles of the conditional distribution of the response variable are expressed as functions of observable covariates. Quantiles are inseparably linked to the sample sorting. Therefore, we can define the quantiles through an alternative expedient as an optimization problem. Similar to the sample mean is defined as the solution to the problem of minimizing a sum of squared residuals, Koenker and Hallock (2001) demonstrated that the median can be defined as the solution to the problem of minimizing a sum of absolute residuals. They indicate that, “The symmetry of the piecewise linear absolute value function implies that the minimization of the sum of absolute residuals must equate the number of positive and negative residuals, thus assuring that there are the same number of observations above and below the median.” The prediction from most regression models is a point estimate of the conditional mean of a response given a set of predictors. However, the conditional mean measures only the center of the conditional distribution of the response. A more complete summary of the conditional distribution is provided by its quantiles. The 0.5 quantile (i.e., the median) can serve as a measure of the center. The 0.9 quantile indicates the point where 90% percent of the data have values less than this number.

Quantiles arise naturally in actuarial sciences. For example, one may desire to know a certain level (e.g., 0.9 quantile) of total medical and pharmacy costs, given all the cost information in the past year and other demographic, or risk score information. This is especially important in stop-loss or large deductible health plan. For heavy tailed data, median (0.5 quantile) may be a better measure for the average costs. Recent advances in computing allow the development of regression models for predicting a given quantile of the conditional distribution, both parametrically and non-parametrically. The general approach is called quantile regression, but the methodology (of conditional quantile estimation) applies to any statistical model, like multiple regression, support vector machines, and random forests. The technique and the examples display many of the features common in both machine learning and statistics. Effectively, QR produces the entire conditional distribution of y. With it one can estimate and conduct inference on conditional quantile functions in Extreme Value Theory.

2.1.5.2 Quantile Regression Estimation

Marzban (2003) reviews how the quantile regression estimation performs. Instead of minimizing $\sum_i^n (y_i - (\alpha_0 + \alpha_1 x_i))^2$, QR minimizes $\sum_i^n f(y_i - (\alpha_0 + \alpha_1 x_i))$ to obtain the β^{th} quantile, where

$$f(y-q) = \begin{cases} \beta(y-q) & y \geq q \\ (1-\beta)(q-y) & y < q \end{cases} \tag{2.13}$$

Instead of $E[y|x] = \alpha_0 + \alpha_1 x$, for each quantile, parametric QR defines

$$Q[y|x] = \alpha_0 + \alpha_1 x \tag{2.14}$$

And nonparametric QR gives

$$Q[y|x] = splines \tag{2.15}$$

Especially, QR Forests give

$$Q[y|x] = piecewise\ constant \tag{2.16}$$

Quantile regression can be used to construct prediction intervals. Unlike the confidence interval in most statistical inferences, a 95% prediction interval for the value of Y can be built by

$$l(x) = [Q_{0.025}(x), Q_{0.975}(x)] \tag{2.17}$$

Since quantile regression aims at estimating the conditional quantiles from data, quantile regression can be treated as an optimization problem, similar to the estimation of the conditional mean that is achieved by minimizing a squared error loss function. Let the loss function L_α be defined for $0 < \alpha < 1$ by the weighted absolute deviations

$$L_\alpha(y, q) = \begin{cases} \alpha|y - q| & y \geq q \\ (1 - \alpha)|y - q| & y < q \end{cases} \tag{2.18}$$

While the conditional mean minimizes the expected squared error loss, conditional quantiles minimizes the expected loss $E[L_\alpha]$

$$Q_\alpha(x) = \arg\min_{q} E\{L_\alpha(Y, q)|X = x\} \tag{2.19}$$

A parametric quantile regression can be solved by optimizing the parameters in order to minimize the empirical loss. This can be achieved efficiently due to the convex nature of the optimization problem (Portnoy and Koenker, 1997). Non-parametric approaches (He and Ng 1998) share similar ideas. Chaudhuri and Loh (2002) developed a tree-based approach for estimation of conditional quantile with good performance and easy interpretation.

2.1.5.3 Quantile Regression Forest

Conditional quantiles can be estimated through quantile regression forests, a generalization of random forests. Quantile regression forests demonstrate a non-parametric and accurate way of estimating conditional quantiles for high-dimensional predictor variables. The prediction of random forests can then be treated as an adaptive neighborhood classification and regression procedure (Lin and Jeon, 2006). For quantile regression forests, trees are grown as the same way as the original random forests algorithm. Then the conditional distribution is estimated by the

weighted distribution of observed response variables, where the weights attached to observations are identical to the standard random forests algorithm. The essential difference between quantile regression forests and random forests is that for each node in each tree, random forests keeps only the mean of the observations that fall into this node and neglects all other information. In contrast, quantile regression forests keep the value of all observations in this node (not just their mean), and assesses the conditional distribution based on this information. The estimation can be done via a package quantregForest for R. The package builds upon the excellent R package randomForest (Liaw and Wiener, 2002).

Let's review the details of quantile regression forest approximation. All the following details in this section originated from the Liaw and Wiener (2002). It is known that random forests estimate the conditional mean $E[Y \mid X = x]$ by a weighted mean over the observations of the response variable Y. Similarly, the weighted observations give not only a good approximation to the conditional mean but to the full conditional distribution. The conditional distribution function of Y, given $X = x$, is given by

$$E[y \mid X = x] = P\left(Y \leq y \mid X = x\right) = E\left(1_{\{Y \leq y\}} \mid X = x\right) \tag{2.20}$$

Let's assume θ as the random parameter vector that determines how a tree is grown. The corresponding tree is denoted as $T(\theta)$. Let B be the space in which X lives, that is $X: \Omega \mapsto \mathrm{B} \subseteq \mathbb{R}^p$, where $p \in \mathbb{N}_+$ is the dimensionality of the predictor variable. Every leaf $l = 1,2,\ldots,L$ of a tree corresponds to a rectangular subspace of B. Denote this rectangular subspace by $R_l \subseteq \mathrm{B}$ for every leaf $l = 1,2,\ldots,L$. For every $x \in \mathrm{B}$, there is one and only one leaf l such that $l \in R_l$ (corresponding to the leaf that is obtained when dropping x down the tree). Denote this leaf by $l(x,\theta)$ for tree $T(\theta)$. The prediction of a single tree $T(\theta)$ for a new data point $X = x$ is then obtained by averaging over the observed values in the leaf $l(x,\theta)$. Let the weight vector $w_i(x,\theta)$

be given by a positive constant if observation X_i is part of leaf $l(x, \theta)$ and 0 otherwise. The weights sum to one, and thus

$$w_i(x, \theta) = \frac{1_{\{X_i \in R_{l(x,\theta)}\}}}{\#\{j : X_i \in R_{l(x,\theta)}\}} \tag{2.21}$$

Using random forests, the conditional mean $\mathrm{E}(Y|X = x)$ is approximated by the averaged prediction of k single trees, each constructed with an i.i.d vector $\theta_t, t = 1, \dots, k$. Let $w_i(x)$ be the average of $w_i(x, \theta_t)$ over this collection of trees.

$$w_i(x) = k^{-1} \sum_{t=1}^{n} w_i(x, \theta_t) \tag{2.22}$$

Just as E(Y |X = x) is approximated by a weighted mean over the observations of Y , define an approximation to

E(1 {Y ≤y} |X = x) by the weighted mean over the observations of 1 {Y ≤y} ,

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x) 1_{\{Y_i \leq y\}} \tag{2.23}$$

using the same $w_i(x)$ as for random forests, defined in equation (2.22). This approximation is at the heart of the quantile regression forests algorithm.

The algorithm for computing the estimate $\hat{F}(y|X = x)$ can be summarized as:

1) Grow k trees $T(\theta_t), t = 1, \dots, k$ as in random forest. However, for every leaf of every tree, take note of all observations in this leaf, not just their average.
2) For a give $X = x$, drop x down all trees. Compute the weight $w_i(x, \theta_t)$ of observation $i \in \{1, \dots, n\}$ for every tree as in (2.21). Compute weight $w_i(x)$ for every observation $i \in \{1, \dots, n\}$ as an average over $w_i(x, \theta_t), t = 1, \dots, k,$ as in (2.22)
3) Compute the estimate of the distribution function as in (2.23) for all $y \in R$, using the weight from step 2)

Estimate $\hat{Q}_a(x)$ of the conditional quantiles $Q_a(x)$ are obtained by plugging $\hat{F}(y|X=x)$ instead of $F(y|X=x)$ into $Q_\alpha(x) = \inf\{y: F(y|X=x) \geq \alpha\}$. Other approaches for estimating quantiles from empirical distribution functions are discussed in Hyndman and Fan (1996).

2.1.6 Finite mixture model

Finite mixture model is one of the most flexible mechanisms that can capture excess zeros by adding a degenerate distribution. It provides us a parametric alternative to describe the unknown distribution in terms of mixtures of known distributions. A finite mixture model also enables us to assess the probabilities of events or simulate draws from the unknown distribution the same way we do when our data are from a known distribution. In addition, they provide a mechanism that can account for unobserved heterogeneity in the data. The expression for the density or likelihood of a response value y in a general k-component finite mixture model (Kessler & McDowell, 2012) is:

$$f(y) = \sum_{j=1}^{n} \pi_j(Z, \alpha_j) p_j(y; x'_j \beta_j, \emptyset_j) \tag{2.24}$$

In this model, the parametric distribution p_j are weighted by the mixing probabilities π_j. The component distributions p_j can depend on regressor variables x_j, regression parameters β_j, and possibly scale parameters $\emptyset_j$. The mixing probabilities π_j, which sum to 1, can depend on regressor variable Z and corresponding parameter α_j. Kessler & McDowell (2012) show that these probabilities can be modeled using a logit transform if *k* = 2, and as a generalized logit model if *k* > 2. The component distributions p_j are indexed by *j* because the distributions might belong to different families. For example, to manage overdispersion in a two-component model, we could model one component as a normal (Gaussian) variable and the second component as a variable with a *t* distribution with lower degrees of freedom.

The FMM procedure in SAS offers maximum likelihood estimation for numerous continuous and discrete responses. It uses a dual quasi-Newton optimization algorithm by default, but we also can choose from several other optimization techniques to produce the maximum likelihood estimates.

2.2 Modeling Techniques for Hierarchical Data

2.2.1 Introduction to hierarchical modeling

Generally, we consider a hierarchical model to be a regression in which the parameters are given a probability model (Gelman and Hill, 2006). Hierarchical models are also called as multilevel models for two reasons: first, from the structure of the data; second, from the model itself. They are a direct extension of regression models to process clustered data and parameters which vary by group. In general, healthcare insurance companies provide healthcare insurance plans for not only individuals and families, but also employers and organizations. The datasets we are using in this dissertation are collected with a natural multilevel structure and we are interested in the effects of certain observed group attributes. Obviously, hierarchical models fit.

2.2.2.1 Motivations for multilevel modeling

Gelman and Hill (2006) summarized a few motivations for multilevel modelling as follows:

- Analysis of structured data - Some datasets are collected with an inherent multilevel structure, for example, individual participant within accounts in group health insurance
- To study group level effects - Multilevel models allow us to study effects that vary by group. It allows the estimation of group averages and group-level effects, compromising between the overly noisy within-group estimate and the oversimplified regression estimate that ignores group indicators

- Including predictors at two different levels- We have outcome measurements at the individual level and predictors at the individual and group levels. But in a classical regression it is not possible to include county-level indicators as well along with county-level predictors—the predictors would become collinear. The multilevel model provides a coherent model that simultaneously incorporates both individual- and group-level models.
- Prediction- Classical regression models are commonly used for predicting outcomes for new cases. But what if the data vary by group? Then we can make predictions for new units in existing groups or in new groups. The latter is difficult to do in classical regression. If a model ignores group effects, it will tend to understate the error in predictions for new groups. But a classical regression that includes group effects does not have any automatic way of getting predictions for a new group, because there would not be an indicator for this group school in the model.

2.2.2.2 Complete Pooling, No Pooling and Partial Pooling

Now let's look at how the hierarchical models work and distinguish from its alternatives (Gelman and Hill, 2006). The traditional alternatives to multilevel modeling are complete pooling and no pooling.

First, let's start with complete pooling. There are many modeling approaches for analysis with multilevel structured data. One approach is to take steps for different levels of analyses: model the data at the lowest level and then aggregate the modeling results on higher level units. For example, one may begin with individual participant-level data, estimating an individual-level risk model, computing expected risk of the group by summing up the risks of its members. Unlike the hierarchical models which can take account of the variability at each level of the hierarchy, the complete-pooling analysis ignores any variation in average risk levels between groups. This is

undesirable, particularly since the goal of our analysis was to identify groups with high risk. We do not want to pool away the main subject of our study.

Second, let's look at no pooling. For no pooling, data from different sources are analyzed separately. Problems may arise when sample sizes for some groups are small and when there are interaction between individual and group level predictors. Meanwhile, looking at all the counties together: the estimates from the no-pooling model overstate the variation among accounts and tend to make the individual account look more different than they actually are. Usually, there are two versions of no pooling. One version fits a separate regression model within each group, the other include the group indicators and estimate the model classically.

Finally, let's move to partial pooling. Multilevel/hierarchical model is one kind of partial pooling, which include both individual and group level predictors at once, not into two steps. In the multilevel model, a "soft constraint" is applied to the group predictors because they are assigned a probability distribution.

2.2.2.3 Analysis Framework

All multilevel models are Bayesian in the sense of assigning probability distributions to varying regression coefficients. The distinction between Bayesian and non-Bayesian multilevel models arises only for the question of modeling other parameters, the non-varying coefficients and the variance parameter. Since for our data structure, we have more than 1000 groups and only two levels are needed, mixed-effects model are preferred to Bayesian model. The analysis framework is shown in fig. 2.3.

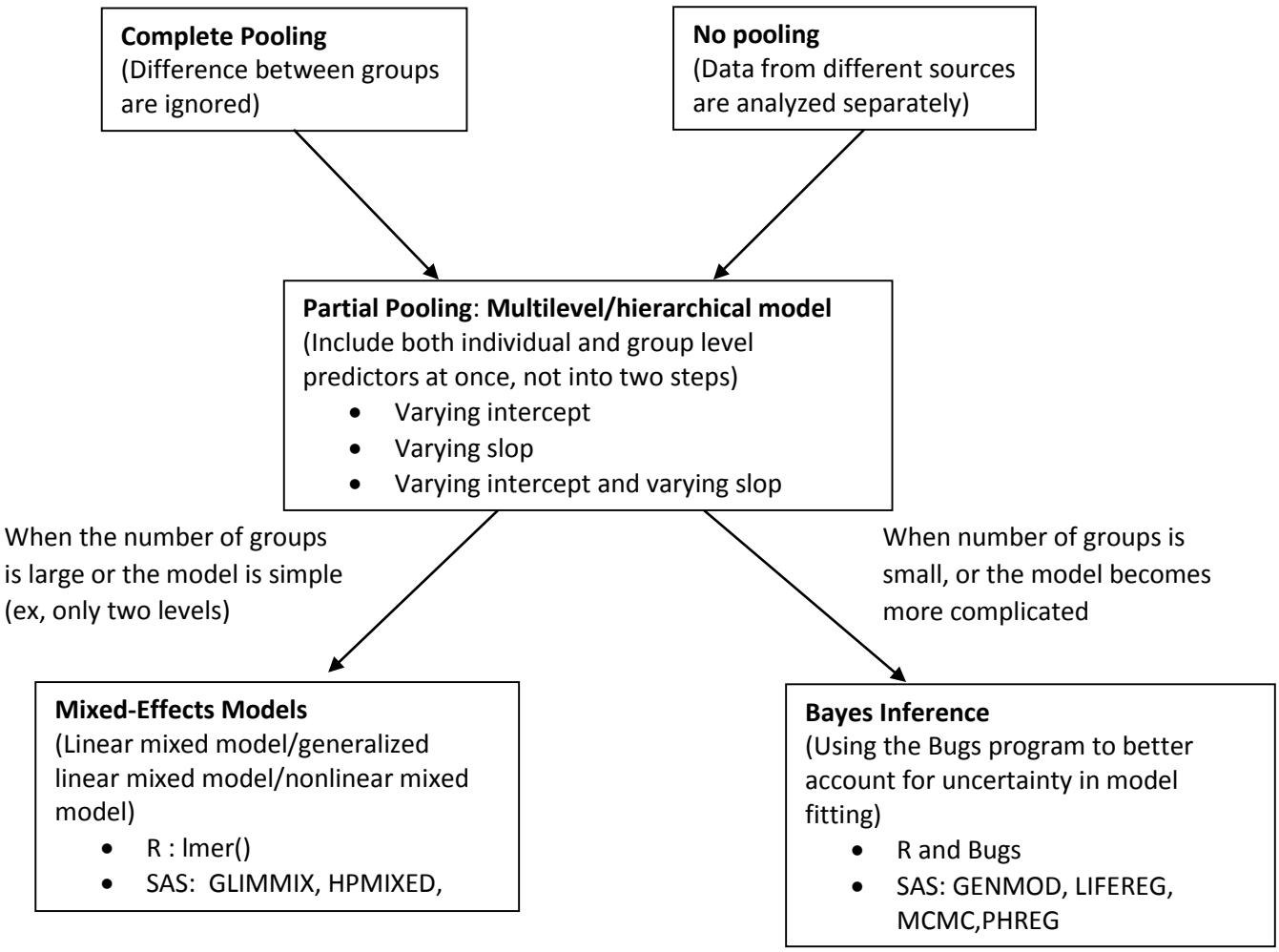


Fig.2.3 Analysis framework for hierarchical model

2.2.2 Generalized linear mixed model (GLMM)

A mixed model[1] contains both fixed and random effects. The regression coefficients that are called random effects are considered outcomes of a random process identified with the model that is predicting them. On the contrary, fixed effects correspond either to parameters that do not vary among groups or to parameters that vary but are not modeled themselves. The hierarchical

[1] Introduction to Generalized Linear Mixed Models, UCLA, http://www.ats.ucla.edu/stat/mult_pkg/glmm.htm

models and mixed models are actually equivalent, two ways to express the same relationships. Hierarchical models suppose that higher level units are drawn from a population and produce posterior estimates of unit effects (Dai, Li, and Rocke, 2006). A mixed-effects model includes both fixed and random effects. The standard linear mixed model (LMM) is thus represented by the following assumptions:

$$Y = X\beta + Z\gamma + \varepsilon \tag{2.25}$$

$$\gamma \sim N(0, G) \tag{2.26}$$

$$\varepsilon \sim N(0, R) \tag{2.27}$$

$$Cov[\gamma, \varepsilon] = 0 \tag{2.28}$$

The matrix *G* is covariance matrix for the random effects and the matrix *R* is covariance matrix for the random errors. A G-side random effect in a mixed model is an element of γ and its variance is expressed through an element in G. An R -side random variable is an element of ε, and its variance is an element of R.

A natural question is when to use fixed effects (in the sense of varying coefficients that are unmodeled) and when to use random effects. It is widely recognized that fixed effects are appropriate if group-level coefficients are of interest, and random effects are appropriate if interest lies in the underlying population. Others recommend fixed effects when the groups in the data represent all possible groups, and random effects when the population includes groups not in the data. Suggested by Brady et al. (2014), we can build the mixed model by the top-down strategy through following steps:

- Step1: Fit a model with all possible fixed predictors and random account specific intercepts.

- Step2: Add possible interaction terms. (Two predictors interact if the effect on the response variable of one predictor depends on the value of the other.)
- Step3: Select a covariance structure for the residuals
- Step4: Reduce the model by removing non-significant fixed effects, and assess model diagnostics.

2.2.3 Generalized Estimating Equation (GEE)

2.2.3.1 Overview

Generalized Estimating Equation (GEE), introduced by Liang and Zeger (1986), is an extension of GLMs. The goal of GEE is to fit a model to correlated and clustered responses. In GEE, the responses Y_1 ,Y_2 , ... , Y_n are assumed to be correlated or clustered. It does not require the homogeneity of variance, but errors are correlated. Compared to mixed effects model, GEE is a marginal model (or called Population average model) that does not require distributional assumptions for the observations, only a regression model for the mean response. That is to say, the model for the mean response only depends on the covariates of interest, and not on any random effects or previous responses. In contrast, mixed effects models are referred to as subject-specific models and the mean response depends not only on covariates but also on a vector of random effects (Fitzmaurice and Molenberghs, 2009). Its form is like a GLM, but full specification of the joint distribution not required, and thus no likelihood function. Summarized by Garrett et.al (2012), a marginal model for longitudinal data has the following three-part specifications:

N = the number of subjects

Y_{ij} = the response variable for the i^{th} subject on the j^{th} measurement

n_i= the number of repeated measurements of the response on the i^{th} subject

The response for the i^{th} subject can be grouped into a n_i × 1 vector

1) The marginal expectation of the response $\mu_{ij} = E(Y_{ij}|X_{ij})$ depends on the covariates X_{ij}, through a link function $g(\mu_{ij}) = X_{ij}^T\beta$
2) The variance of each Y_{ij}, given the covariates, depends on the mean through Y_{ij} according to $VAR(Y_{ij}|X_{ij}) = \emptyset_{ij}v(\mu_{ij})$

   where $\emptyset_{ij}$ is scale and v is variance function.
3) The pairwise (or two-way) within-subject association among repeated response, given the covariates, is assumed to be a function of the means, μ_{ij}, and an additional set of within-subject association parameter, α, of the secondary interest.

$$V_i = V_i(\beta, \alpha, \emptyset) = \emptyset A_i(\beta)^{\frac{1}{2}} R_i(\alpha) A_i(\beta)^{\frac{1}{2}} \quad (2.29)$$

Where A_i is a $n_i \times n_i$ diagonal matrix, that is $A_i = diag(VAR\{y_{ij}\}) = diag\left(v(\mu_{ij})\right), j = 1, \dots, n_i$ and $R_i(\alpha)$ is the $n_i \times n_i$ correlation matrix (a function of α)

There are both random components and systematic components. The link function can be any $g(\mu_i)$, e.g., identity, log, logit, etc. The random component is described by the same variance functions as in the independence case, but the covariance structure of the correlated responses must also be specified and modeled.

Comments:

- V_i is known as a "working" covariance matrix to distinguish it from the true underlying covariance among the Y_i. That is, the term "working" acknowledges our uncertainty about the assumed model for the variances and within subject associations; unless they have been correctly modeled, our model for the covariance matrix may not be correct. The GEE approach allows the modeler to specify an incorrect structure.
- Interpretation of β: in terms of contrasts of the changes in the transformed mean responses in sub-populations.

2.2.3.2 Covariance specification.

The essential idea of GEE is to generalize the usual univariate likelihood equations by introducing the covariance matrix of the vector of response Y_i. It can accommodate a wide variety of correlation structures such as auto-regressive, exchangeable, and unstructured. Here are four correlation structures:

- Independence (correlation between time points is independent)

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Exchangable (Compound Symmetry)

$$\begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

- AutoRegressive Order 1 (AR 1)

$$\begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

- Unstructured

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}$$

$Where\ \rho_{ij} = corr(Y_{ij}, Y_{ik})$ for the i^{th} subject at times *j* and *k*.

2.2.3.3 Parameter Estimation

Unlike maximum likelihood estimation (MLE) or ordinary least squares (OLS), GEE uses quasi-likelihood estimation, but these may coincide at time. The quasi-likelihood estimators are estimates of quasi-likelihood equations, hence being called as *generalized* estimating equations. A quasi-likelihood estimate of β comes from maximization of normality-based log-likelihood without assuming that the response follows normally distribution. In most cases, there are no

closed-form solutions, therefore the GEE estimates are obtained by using an iterative algorithm, iterative quasi-scoring procedure. GEE estimates of model parameters are valid even if the covariance is misspecified (because they depend on the first moment, e.g., mean). However, if the correlation structure is misspecified, the standard errors may not be good, and adjustments based on the data are thereby needed to get more appropriate standard errors. Agresti (2013) indicates that a chosen model in practice could not be exactly correct, but choosing appropriately a working correlation (covariance structure) can help improve efficiency of the estimates.

2.3 Variable Selection and Shrinkage

2.3.1 Variable Selection Review

There are more than 500 covariates in our dataset. It is commonly recognized that unnecessary predictors will add noise to the estimation of other quantities that we are interested in, and collinearity may be caused by having too many variables trying to do the same job. We aim at interpreting the data in the simplest way, that is, redundant predictors should be removed. A few commonly used variable selection methods are reviewed as follows:

- Stepwise selection/Forward selection/Backward elimination.

  They are based on the 'importance' of a covariate defined in terms of the statistical significance of the coefficient of the variable. They do not address overfitting or underfitting problems. But some researchers argue that confidence intervals produced with a stepwise procedure are falsely narrow. (Shtatland et al. 2000, Flom et al. 2007)

- Criterion-based procedures (AIC, AICc, BIC, Adjusted R^2)

  The Akaike Information Criterion (AIC) is an approach of selecting a model from a set of model candidates. The goal is to minimize the Kullback-Leibler distance between the

model and the truth. It's originated from information theory, but we always see it as a criterion that seeks a model that has a good fit to the truth but few parameters (Burnham and Anderson, 2002). It can be defined as:

$$AIC = -2 * (\log likelihood) + 2k \quad (2.30)$$

where K is the number of free parameters in the model.

AICc (corrected AIC) is AIC with a correction for finite sample sizes. It takes into account sample size by increasing the relative penalty for model complexity with small data sets. It is defined as

$$AICc = -2 * (\log likelihood) + 2k * (\frac{n}{n-K-1}) \quad (2.31)$$

where n is the sample size. As n gets larger, AICc converges to AIC. Therefore there's really no harm in always using AICc no matter what sample size is.

- Variable Importance （Decision Tree, Least Angle Regression (LARS), R/Chi square）

  We have a rank a variable importance by decision tree, or least angle regression, or R/Chi Square, but don't know where to cut the models. It depends on what model we are using.

2.3.2 Shrinkage

Similar to variable selection, Goeman at al. (2012) says that the purpose of this shrinkage is to prevent overfitting arising from either collinearity of the covariates or high-dimensionality. Penalized regression is proposed for simultaneous variable selection and coefficient estimation. In penalized regression, we minimize the residual sum of squares subject to the functions of the value of coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficient that are exactly 0 and hence gives interpretable models (Tibshirani, 1996). Typically, there are four types of penalized regression as follows:

Table 2.1 Introduction and Comparison between different penalized regression models

Types	Definition		Tools
Ridge Regression	Ridge regression penalizes the size of the regression coefficients by the sum of squared values of coefficients being less than a constant. $$\min \sum_i \left(y_i - x_i^T \beta\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$		Proc REG in SAS or lm.ridge in MASS package in R
LASSO Regression	The "lasso" minimizes the residual sum of squares subject to the sum of the absolute value of coefficients being less than a constant. $$\min \sum_i \left(y_i - x_i^T \beta\right)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right|$$		SAS GLMSELECT procedure
Elastic Net	Elasticnet is introduced as a compromise between these LASSO and ridge regression, and has a penalty which is a mix of L_1 and L_2 norms. $$\min \sum_i (y_i - x_i^T \beta)^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} \left|\beta_j\right|$$		SAS GLMSELECT procedure
Bayesian LASSO Regression	From a Bayesian perspective, one can think of the penalty as arising from a prior distribution on the parameters. It is suggested that Lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e., double-exponential) priors, horse shoe priors. We call these models as the Bayesian Lasso.		Function Blasso() using R package " monomvn"
	Laplace Prior	$\pi\left(\beta \mid \sigma^2\right) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda \left|\beta_i\right| \sqrt{\sigma^2}}$	
	Horseshoe Prior	$\left(y \mid \theta\right) \sim N\left(\theta, \sigma^2 I\right), \left(\theta_i \mid \lambda_i\right) \sim N\left(0, \lambda_i^2\right)$ $\left(\lambda_i \mid \tau\right) \sim C^+\left(0, \tau\right), \tau \sim C^+\left(0, \sigma\right)$ Where $C^+\left(0, a\right)$ is a standard half-Cauchy distribution on the positive real with scale parameter a	

Comments:

- Cross-validation is a good tool for estimating the best value for λ.

- LASSO (Least absolute shrinkage and selection operator) outperforms Ridge when coefficients are mostly zero, while Ridge outperforms LASSO when lots of small coefficient.
- LASSO does variable selection, while Ridge not.

Chapter 3 Model Selection and Averaging of Health Costs in ETGs

3.1 Data

I am using ETG cost data from a major national health insurer. It has 33 million sample observations from 9 million claimants. Each observation represents the total cost per claimant per year on each ETG. For those policyholders without claim cost on certain ETGs, there is no zero record for them in the data set. There are 347 ETGs in all, including 320 non-routine ETGs, such as AIDS, hemophilia, and personality disorder. Only those non-routine ETGs in this dissertation are considered in this research because we cannot gain much information from those routine ETGs such as routine exam, inoculations, conditional exam, and other preventative services. Basic summary statistics for several randomly selected ETGs are shown in Table 3.1 for illustration. Different ETGs have various claim frequencies, means and standard deviations.

Table 3.1: Dictionary and summary statistics for selected ETGs.

ETG Code	Frequency	ETG description	Mean	Standard deviation
1301	13,534	AIDS	15,570	25,246
1635	2,679	Hyper-functioning adrenal gland	2,035	8,963
1640	1,162	Hypo-functioning parathyroid gland	1,704	6,314
2068	16,554	Agranulocytosis	4,677	17,923
2070	822	Hemophilia	94,343	303,552
2080	944	Anemia of chronic diseases	2,434	10,943
2082	49,409	Iron anemia	1,772	5,208
2394	1,550	Personality disorder	1,718	5,263
3868	42,401	Congestive heart failure	10,870	56,777
4370	50	Lung transplant	461,226	338,683
4744	4,162	Trauma of stomach or esophagus	6,562	10,994
7112	1,668	Juvenile arthritis	7,193	27,44

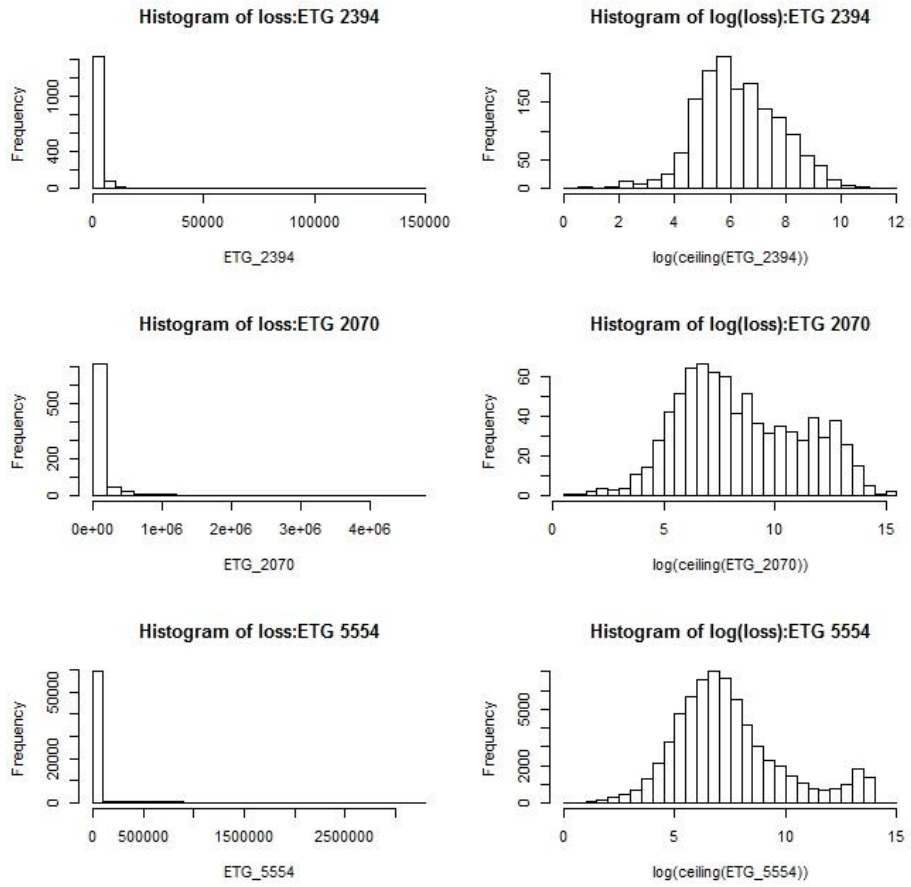


Figure 3.1: Histograms of loss (left panel) and log-loss (right panel) for three ETGs

The histograms of these costs both on the original and log scale give insight into the skewness and tail thickness of the data. Using that information, I could choose plausible candidate distributions. Specifically, lognormal, gamma, Lomax, and log-skew-t distributions are considered in the study. Although almost all the ETGs show similar shape with a heavy tail and right skewness on the original scale, the histograms for those costs on the log scale vary among different ETGs. The histograms for three randomly selected ETGs are shown in Figure 3.1. The total cost per claimant per year on each ETG is on the dollar basis, hence all the values in the data set are positive.

3.2 Model Selection

Proper model selection for those ETG based costs is essential to adequately price and risk management in health insurance. The optimal model (or model probabilities) can change depending on the disease. As discussed in the introduction, model averaging enables us to average the fits for a number of models, instead of picking a single best model. It gives the analysts greater insight into the relative merits of the competing models.

3.2.1 AIC and BIC Weights

Following the recommendations given by Akaike (1978) and Burnham and Anderson (2002), I can compute the change in values of AIC and BIC with respect to those of the best candidate model. In particular, we compute

$$\begin{aligned} w_i^{\mathrm{AIC}} &= \frac{\exp(-\Delta_i^{\mathrm{AIC}}/2)}{\sum_{k=1}^{K}\exp(-\Delta_k^{\mathrm{AIC}}/2)} \quad \text{with} \quad \Delta_i^{\mathrm{AIC}} = \mathrm{AIC}_i - \min\left\{\mathrm{AIC}_1,\ldots,\mathrm{AIC}_K\right\} \\ w_i^{\mathrm{BIC}} &= \frac{\exp(-\Delta_i^{\mathrm{BIC}}/2)}{\sum_{k=1}^{K}\exp(-\Delta_k^{\mathrm{BIC}}/2)} \quad \text{with} \quad \Delta_i^{\mathrm{BIC}} = \mathrm{BIC}_i - \min\left\{\mathrm{BIC}_1,\ldots,\mathrm{BIC}_K\right\}, \end{aligned} \tag{3.1}$$

Where K denotes the number of candidate models. The weights w_i^{AIC} are known as AIC weights or Akaike weights. Similarly, the weights w_i^{BIC} are called the BIC weights. For illustrative purposes, the AIC values and Akaike weights on four models for selected ETGs are provided in Table 3.2.

Table 3.2: Akaike weights and AIC values for the four candidate models and selected ETGs

	Akaike Weights				AIC values			
ETG code	lognormal	gamma	log-skew-t	Lomax	lognormal	gamma	log-skew-t	Lomax
1301	0	0	1	0	288,909	289,613	286,796	287,556
1635	0	0	1	0	44,022	46,907	43,765	43,808
1640	0	0	1	0	18,640	19,920	18,567	18,617
2068	0	0	1	0	286,108	299,983	285,891	285,954
2070	0.882	0	0.118	0	17,897	18,309	17,901	17,930
2080	0	0	0.998	0.002	14,755	15,835	14,684	14,697

2082	0	0	1	0	725,294	760,699	724,756	726,749
2394	0	0	1	0	25,175	26,374	25,144	25,182
3144	0.001	0	0.99	0.009	328	344	315	324
3169	0	0	1	0	2,508,992	2,606,074	2,508,562	2,511,985
3868	0	0	1	0	797,694	837,377	797,623	799,257
4370	0.002	0.087	0.816	0.095	1,416	1,408	1,404	1,408
4744	0	0	1	0	80,732	81,476	80,539	80,580
7112	0	0	0.973	0.027	30,786	32,166	30,766	30,773

For those randomly selected ETGs, the distributions for some ETGs such as ETG-1301 and ETG-3868 are immediately apparent. The log-skew-t distribution is also dominant for ETG-2080 and ETG-3144. It indicates that AIC values and Akaike weights have a strong preference for the log-skew-t distribution for most of these data sets. However, there are exceptions. For ETG-2070, the probability spreads between two distributions: 0.882 probability to lognormal model and 0.118 probability to the log-skew-t. And for ETG-4370, the probability spreads among all four distributions: 0.002 probability to lognormal model, 0.087 probability to gamma distribution, 0.816 distribution to log-skew-t distribution and 0.095 probability to the Lomax distribution.

3.2.2 Bayesian Inference and Parallel Model Selection

The Bayesian approach allows one to learn about the whole distribution of quantities of interest rather than just a point estimate of parameters, which can be very useful in actuarial science. Rather than trying to identify the best model, a parallel model selection method proposed by Congdon (2006) will provide the posterior probabilities of each model being the best, enabling model averaging and providing deeper insights into the relationships between the models. The uncertainty in the model selection process can also be explicitly modeled in the model.

Table 3.3: Prior distribution settings for the candidate models

Candidate Model (*Parameters*)	*Prior Distributions*	*Number of Thinned Samples Per Chain*	*Number of Burn-in Samples Per Chain*
lognormal (μ, τ)	$\mu \sim \text{normal}\,(6,\ 5)$ $\tau \sim \text{gamma}\,(4,\ 4.5)$	30,000	20,000
gamma (τ, ν)	$\tau \sim \text{gamma}\,(2,\ 3)$ $\nu \,\vert\, \omega \sim \text{exponential}\,(\omega)$ $\omega \sim \text{uniform}\,(0.01,\ 10)$	50,000	35,000
log-skew-t $(\alpha, \xi, \nu, \Omega)$	$\alpha \sim \text{normal}\,(50,\ 4)$ $\xi \,\vert\, \theta \sim \text{normal}\,(\theta,\ 7)$ $\nu \sim \text{exponential}\,(0.25)$ $\Omega \sim \text{inverse gamma}\,(6,\ 1)$ $\theta \sim \text{normal}\,(0,\ 5)$	300,000	260,000
Lomax (λ, α)	$\lambda \sim \text{gamma}\,(2,\ 3)$ $\alpha \,\vert\, \omega \sim \text{exponential}\,(\omega)$ $\omega \sim \text{uniform}\,(0.01,\ 10)$	300,000	20,000

The LaplaceDemon package in R is used to perform parallel MCMC algorithms. Several algorithms were tried and compared, such as Hit-and-Run Metropolis (Chen and Schmeiser, 1993), No-U-Turn Sampler (Hoffman and Gelman, 2014; Bai, 2009) and Hamiltonian Monte Carlo (Neal, 2011). Three chains are running in most cases, each in parallel, where a sequence $x_1, x_2, \ldots$ of random elements of some set is a Markov chain if the conditional distribution of x_{n+1} given $x_1, x_2, \ldots\ x_n$ depends on x_n only. The traces of those three MCMC chains initialized with different starting values. When doing model selection, non-informative priors will excessively penalize complex models, so I set the priors to be semi-informative. After that, I look at the data or maximum likelihood estimates (MLEs) of the candidate model parameters and try to find hyperparameters which will put most of the probability mass on a reasonable range around those parameter estimates. The prior distributions for the parameters of the candidate models are given in Table 3.3. The other two important settings are burn-in sample and thinned sample. Burn-in sample refers to the samples after discarding an initial portion of a Markov chain sample so that the effect of initial values on the posterior inference is minimized. The thinned samples were

introduced to reduce sample autocorrelations by keeping every k^{th} simulated draw from each sequence. In fact, the robustness of the priors varies among different distribution and prior choices. For example, our current choice for lognormal distribution is very robust. The priors for Lomax and log-skew-t distributions are relatively robust. They work well for almost all the ETGs, but need longer time to achieve convergence. Therefore, I usually assign larger number of iterations and more burn-in samples for them. The prior for the gamma model had a reasonably large impact on the results. Our current choice of priors is relatively robust and works well for almost all the ETGs.

Parallel model selection is applied to several randomly selected ETGs; the posterior model probabilities are given in Table 3.4. The distributions for some ETGs such as hemophilia, AIDS, and agranulocytosis are immediately apparent. The lognormal distribution is also dominant for lung transplant and many others. For personality disorder, the probability spreads between two distributions: 0.783 probability to lognormal model and 0.217 probability to the log-skew-t.

In addition to the improved understanding of the data, these probabilities can be used for model averaging. When one model is dramatically better than the others, only knowing the best model will be sufficient. When the potential models are very similar in their fit for some data sets, a simulation should account for that model uncertainty by drawing a proportion of the simulations from each of the models that fit the data well. For example, to simulate future ETG cost streams for personality disorder, 78.3% samples can be drawn from lognormal distribution, and 21.7% of the samples drawn from log-skew-t. Under the standard methods, the proper model proportions are unknown.

Table 3.4: Sample posterior model probabilities using Bayesian parallel model selection

ETG Code	ETG description	lognormal	gamma	log-skew-t	Lomax
1301	AIDS	0	0	1	0
1635	Hyper-functioning adrenal gland	0	0	1	0
1640	Hypo-functioning parathyroid gland	0	0	1	0
2068	Agranulocytosis	0	0	1	0
2070	Hemophilia	1	0	0	0
2080	Anemia of chronic diseases	0	0	1	0
2082	Iron anemia	0	0	1	0
2394	Personality disorder	0.783	0	0.217	0
3868	Congestive heart failure	0.450	0	0.550	0
4370	Lung transplant	0.999	0	0.001	0
4744	Trauma of stomach or esophagus	0	0	1	0
7112	Juvenile arthritis	0.999	0	0.001	0

3.2.3 Random Forest

In previous section I present the procedure and highlight the benefit of Bayesian model averaging over traditional methods. Ideally we want to apply the Bayesian approach to all the ETGs cost data (including more than 33 million samples). However, it takes a long time to complete Bayesian inference and model selection on all ETGs. Therefore, it is desirable to find a faster approach for huge data sets. Random forests are an ensemble learning method for classification that works by constructing many decision trees at training time and outputting the class that is the mode of the classes output by the individual trees. It grows a multitude of classification trees and each tree outputs a classification.

We can think of the trees as voting for the classification, and then the random forest chooses the classification with the most votes. As mentioned earlier, if I treat all the ETGs following one distribution as one cluster, selecting the best distribution is equivalent to putting the ETGs into the best cluster. In this case, we can extract some features from each data set and use those features for Random Forest (RF) classification. We do not need to look at each data point in the data set,

just some summary statistics, which can save a lot of time. Our experiments also show that RF is extremely fast when compared to the MLE approach (e.g., the system time for MLE is about 120 times that of RF). We can carry out RF model selection through the following three steps:

• Step 1: Domain Specific Feature Extraction.

We extract 12 features (mean, median, standard deviation, interquartile range, median absolute deviation, 10th, 25th, 75th, 90th percentiles, coefficient of variation, skewness, and kurtosis) from the data set both on the original and log scale. Therefore, we have 24 features in all for Random Forest Model Selection. The information is saved as one row for each dataset and 24 columns for each row. Basically, there are two types of features:

– Moment-based characteristics (e.g., mean, standard deviation, coefficient of variation, skewness, and kurtosis) for raw data and the same measures for log-data.

– Percentile-based characteristics (e.g., 10th, 25th, 50th, 75th, 90th percentiles, median absolute deviation, and interquartile range) for raw data and the same measures for log-data.

• Step 2: Random Forest Training for Prediction.

Create a moderate size data set (e.g., 600 observations for each distribution) with known response variables to train the random forest. Our experiments show that the number of observations can be approximately chosen as the square of the number of variables in random forest to achieve a reasonable out-of-bag error rate. We have 24 covariates here, therefore a dataset with 600 observations will be sufficient.

• Step 3: Random Forest Model Selection.

Apply the trained Random Forest in Step 2 to the original data set with features generated in Step 1.

In Step 1, we first use two groups of characteristics (the moment-based features and percentile-based features) separately and find that the approach based on moment-type features outperforms the percentile-type approach in distinguishing distributions. Furthermore, using both moment-based and percentile-based features one can achieve the lowest out-of-bag error rate and the best performance in distinguishing distributions. These findings are summarized in Table 3.5.

Table 3.5: Performance of moment-based features versus percentile-based features.

Candidate Model Used	Feature Selection	Out-of-bag Error Rate
Lognormal, gamma, Lomax	Moment-based features only Percentile-based features only Both types of features	0.25% 1.00% 0.08%
Lognormal, gamma, Lomax, Log-skew-t	Moment-based features only Percentile-based features only Both types of features	3.53% 13.63% 2.01%

The performance of RF also depends on the difficulty of the tasks. If the clusters have obvious distinguishable features (there is a huge difference between the lognormal, gamma and Lomax distribution), RF would recognize that and the misclassification rate would be very low. But if the clusters are quite similar, then it is more difficult to distinguish the models. The more candidate distributions with similar characteristics, the worse the random forest performs.

Table 3.6 shows the RF classification results on training data and Table 3.7 shows the results on the testing data. Since RF grows many classification trees, we set the number of cases in the training set as 4,000, sample 4,000 cases at random – but with replacement, from the original

data set. Also, we have 24 input variables, usually a number m ≪ 24 is specified such that at each node, m variables are selected at random out of the 24 and the best split on these m is used to split the node. The value of m is held constant during the forest growing. Here we choose the optimal m as 6, which was determined by experimentation.

Table 3.6: Random forest classification results on training data.

Candidate Models Used	Number of Trees in Random Forest	Number of Vars Used at Each Split	Out-of-Bag Error Rate
lognormal, gamma, Lomax, log-skew-t	4,000	6	0.25%
lognormal, gamma, Lomax	4,000	6	0.00%

Table 3.7: Random forest classification results on testing data.

Candidate Models Used	Misclassification R
lognormal, gamma, Lomax, log-skew-t	23.8%
lognormal, gamma, Lomax	1.2%

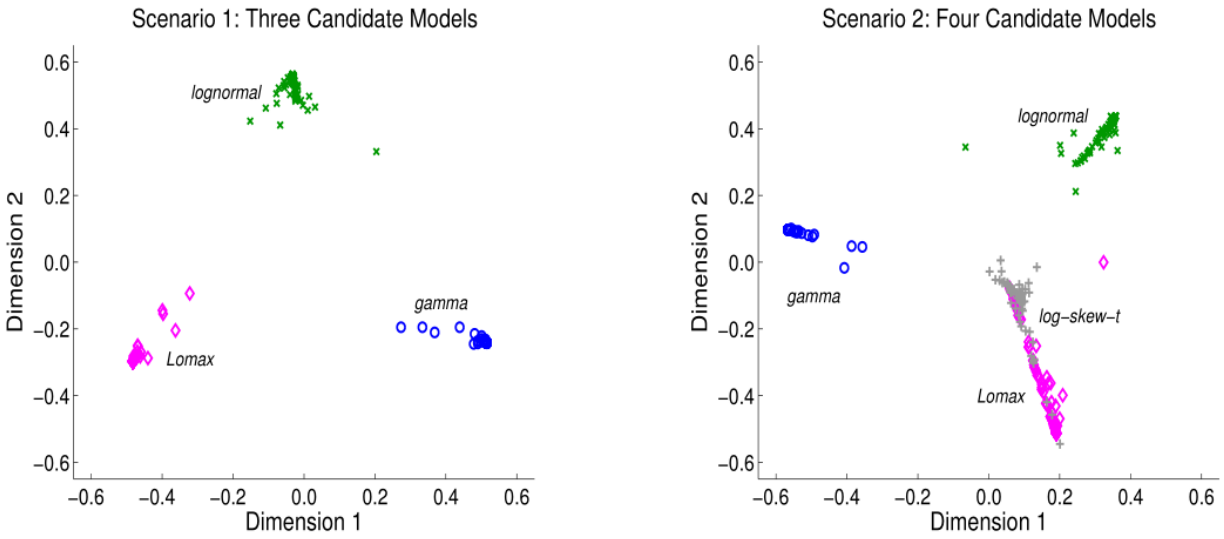


Figure 3.2: Multidimensional scaling plots of proximity matrix for two scenarios

Multidimensional scaling plot is an ordination technique to visualize the level of similarity between individual cases in a data set. It aims to place each object in n-dimensional space such that the

between-object distances are preserved as far as possible. In Figure 3.2, the statistical features of each data set are represented by a point in a two dimensional space. The points are arranged in this space so that the distances between pairs of points relates to the similarities among the pairs of objects. That is, two similar objects are represented by two points that are close together, and two dissimilar objects are represented by two points that are far apart. Table 3.6, 3.7 and fig.3.2 tells us that if only three distributions (gamma, lognormal, Lomax) are considered, they are easily distinguishable. When the log-skew-t distribution is added to the mix, more similarities is introduced because some points with different shapes are close together. Thus, it is clear that the most difficult task would be the classification for all four distributions (lognormal, gamma, log-skew-t, Lomax) because the points from different distributions cannot be easily distinguished.

3.3 Results

To determine how well the metrics work in our settings, I set up a simulation study. To start with, I use the MLE approach to fit four distributions on the same real ETG data. And then I use these MLE-fitted models to simulate four random samples with 600 observations each that follows one of the lognormal, gamma, log-skew-t, and Lomax distributions, respectively. After that I apply the three model selection methodologies (AIC weights, RF, Bayesian) to the simulated data sets and check how accurately each approach identifies the true model. My findings are summarized in Table 3.8.

Table 3.8: Model selection accuracy: AIC weights, Random Forest, Bayesian.

Model selection methodology	Selected distribution	Distribution used to simulate data			
		lognormal	gamma	log-skew-t	Lomax
AIC weights	lognormal	75.81%	0.00%	24.19%	0.00%
	gamma	0.00%	94.42%	5.58%	0.00%
	log-skew-t	0.00%	0.00%	93.91%	6.09%
	Lomax	0.00%	0.00%	27.81%	72.19%
Random Forest	lognormal	99.70%	0.00%	0.10%	0.20%

	gamma	11.30%	62.75%	15.00%	10.95%
	log-skew-t	0.08%	0.03%	67.58%	32.33%
	Lomax	0.03%	0.00%	43.98%	56.00%
Bayesian	lognormal	100.00%	0.00%	0.00%	0.00%
	gamma	1.90%	93.90%	3.14%	1.06%
	log-skew-t	0.00%	0.00%	100.00%	0.00%
	Lomax	0.23%	0.00%	38.54%	61.23%

In each 4×4 matrix in Table 3.8, if the probabilities on the diagonals are close to 100%, the metric accurately selects the true model. From the results, we can observe and compare level of the mode uncertainty and prediction power over different metrics. Though the most computationally intense of the three methods, on an average sense, Bayesian performs best because it exactly identifies lognormal and log-skew-t distributions and it is slightly less certain about gamma and Lomax compared to AIC weights. AIC weights did a good job on average. Random Forest performs slightly more poorly than the other two, but it still can almost surely identify the model with the best fit. Especially when I need to deal with big data sets, its efficiency is valuable without losing much accuracy.

Next, I apply Random Forest and AIC weights metrics to perform the model selection exercise for all 320 ETGs. I did not apply Bayesian parallel model selection in the second step because I only have access to an 8 GB Thinkpad with a 2.50 GHz Intel Quad-Core processor for modeling. Based on our experience, assuming the size of the data set is less than 5000 observations and it can converge, it takes about 2 hours to fit all five distributions for a single ETG. Sometimes it does not converge, and then we need more time to either increase the number of iterations, or recheck the prior distributions. The approximate time to complete Bayesian inference and model selection on all ETGs is 4 weeks. Thus Bayesian parallel model selection does not work well for big data without super computers. Despite the fact that MLE is commonly seen as an efficient method, it still takes about 4 hours in all to finish the model selection for all the ETGs. In contrast, Random

Forest feature classification can be done within 2 minutes. This is explained by the fact that for AIC weights, every observation is used for inference and model selection, while for Random Forest, the model selection is done on the extracted features of the data set, which is a much smaller data set than the original data set. When extracting feature information from the original data set, it also takes a small amount of time. However, compared to the inference and model selection time on each observation, the total time for information extraction and feature classification using random forest is still much less. Table 3.9 shows the speed comparison among all four methodologies.

Table 3.9: Speed comparison (on all 320 ETGs) among four metrics

Model Selection Methodology	Time
Random Forest	~2 minutes
AIC and BIC	~4 hours
Bayesian	~4 weeks

Now I explore how consistent the RF and AIC methodologies are in selecting the same model (for all 320 ETGs). First, in Table 3.10, I only use three distributions (lognormal, gamma, Lomax) as candidates for model selection. Those three distributions have obvious distinguishable features. In the 3 × 3 matrix, RF and AIC agree on all the 197 ETGs model selections on the diagonal. For some ETGs, compared to RF, AIC prefers lognormal to Lomax.

Table 3.10: Comparison of model assignments by RF and AIC for all 320 ETGs
(Three candidate models)

Distribution Selected by RF	Distribution Selected By AIC			RF total
	lognormal	gamma	Lomax	
lognormal	100	11	19	130
gamma	1	5	3	9
Lomax	87	2	92	181
AIC Total	188	18	114	320

Next, in Table 3.10, I use four distributions (lognormal, gamma, Lomax, log-skew-t). AIC has an apparent preference for the log-skew-t distribution because it selects this model for 292 of 320 ETGs. Random forest also selects the log-skew-t distribution for most ETGs, but at the same time it assigns 131 ETGs to lognormal distribution. One common theme is that none of the metrics select the gamma distribution for any ETG. That is understandable because compared to other distributions, gamma is relatively light tailed. Given the heavy tails for most ETG costs, once the log-skew-t distribution is one of the candidates, no metric will select gamma distribution as the best model.

Table 3.11: Comparison of model assignments by RF and AIC for all 320 ETGs
(Four candidate models)

Distribution Selected by RF	Distribution Selected by AIC				RF total
	lognormal	gamma	Lomax	log-skew-t	
lognormal	23	0	2	106	131
gamma	0	0	0	0	0
Lomax	1	0	1	25	27
log-skew-t	0	0	1	161	162
AIC Total	24	0	4	292	320

Chapter 4 Risk Assessment and Pricing for Group Health Claims

In this chapter, I aim to develop claim-based risk assessment models; translate any risk metrics into monetary quantities in terms of gains/losses or profit; and develop models to correctly price in the presence of risk information.

4.1 Data

In this dissertation, claim cost data from a major national health insurer is used for analysis. The target variable is the total medical and pharmacy cost per patient in a year (from July 2011 to June 2012). All the patients are selected to be those who are active on membership from calendar year 2011 to 2012; and we have 967031 samples in all. All the information through March, 2011 is used to model their relationship with the total costs (from July 2011 to June 2012).

4.1.1 Response variable

In our dataset, the dependent variable total medical and pharmacy cost is a continuous variable with 12.38% zeros. The basics statistical measures and quantiles are summarized in table 4.1 and 4.2, respectively. The histograms of the cost and nonzero cost only on both original scale and log scale are shown from fig. 4.1 to 4.4. From those summary statistics and histograms, it is easy to observe that in healthcare, the costs are typically not only right skewed but also heavy-tailed due to a few extremely high-cost patients. As suggested by the histograms on log scale, the log transformation can be used to correct the right skewness, hence improving the fitting. However, the transformation means switching from an additive mean structure to a multiplicative one, meanwhile changing the variance structure. I will discuss more details later when comparing the models. The properties such as the probability mass at zero, the skewness and heavy tail of the distribution give us a hint to use semi-continuous and fat-tailed models in this dissertation.

Table 4.1 Basic Statistical Measures

Location		Variability	
Mean	4915.806	Std Deviation	17711
Median	1104.050	Variance	313677535
Mode	0.000	Range	2631172
		Interquartile Range	3459

Table 4.2 Quantiles

Level	Quantile
100% Max	2,631,172.15
99%	61,781.02
95%	19,374.01
90%	10,604.05
75% Q3	3,723.75
50% Median	1,104.05
25% Q1	265.00
10%	0.00
5%	0.00
1%	0.00
0% Min	0.00

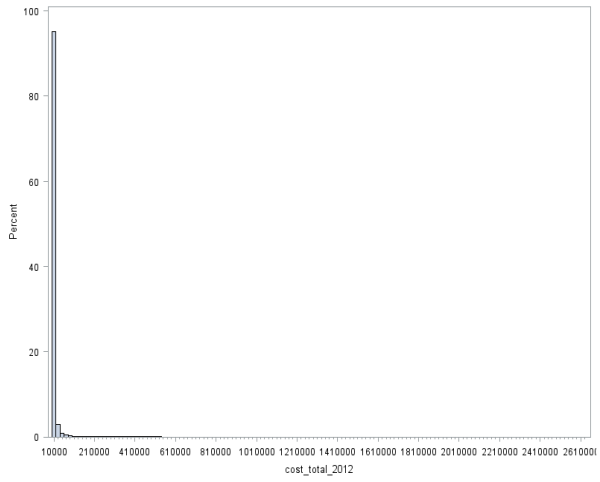


Fig. 4.1 Histogram of response variable

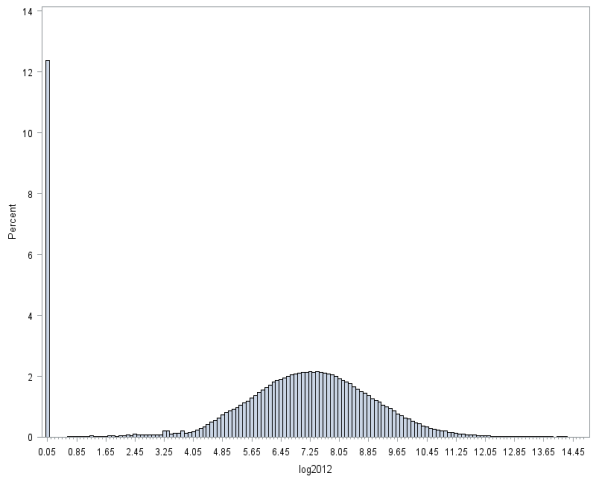


Fig. 4.2 Histogram of response variable on log scale

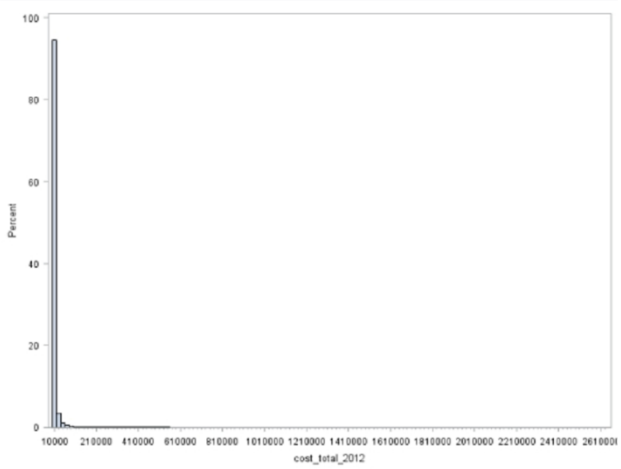


Fig. 4.3 Histogram of nonzero response variable

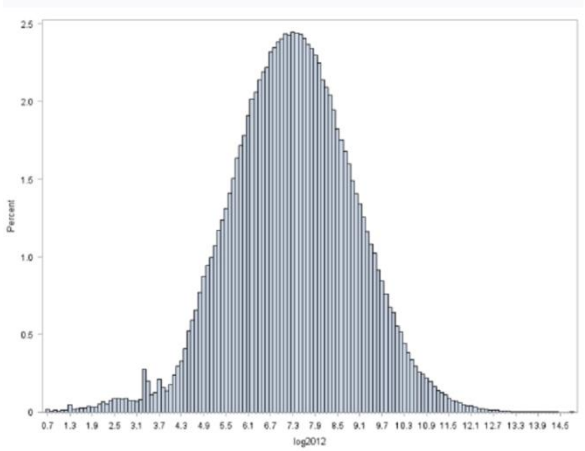


Fig. 4.4 Histogram of nonzero response variable on log scale

4.1.2 Explanatory variables

In summary, there are 582 explanatory variables in our dataset, including both account level and individual level characteristics, such as demographics, ERG risk score, cost summary, ETG codes, STD drug codes (ex.Antineoplastics, Antinauseants, Penicillins). There are 2244 group insurance accounts in all, where the account size ranges from 1 to 30341. In order to prevent unnecessary predictors adding noise to the estimation of other quantities that we are interested in, I will perform variable selection (or shrinkage) in all the models. I aim at interpreting the data in the simplest way--redundant predictors should be removed. Most variables (except those disease or drug codes variables) are listed in table 4.3.

Table 4.3 List of variables

Variable Name	Description
cost_total_2012	Total medical and pharmacy cost per patient in a year (from July 2011 to June 2012)
ETG_COUNT	Number of ETG that the participant has.
MBR_AGE	Member age
PROSPTCV_RISK_CAT	Prospective ERG risk category

PROSPTCV_RISK_NUM	Prospective ERG risk score
SEX_CD	Member gender
SIC_CD	Industry type
account	Group Insurance Account number
chemotherapy	Whether this patient has been given chemotherapy treatment
cost_maint_rx	Sum of expenditures on maintenance dug
cost_specialty_rx	Sum of expenditures on special dug
spfc_count	How many different rx classes does this person have?
gen_brn	Member type: subscriber or dependent
Inpatient	Inpatient indicator
ACCT_SUBTY	Account type
injctbl	Whether this person used injection drug? (1 for yes, 0 for no)
MRKT_SEG_CD	Group insurance type
cost_med	Sum of total medical cost paid by the insurance company between July 2011 and June 2012
cost_phm	Sum of total pharmacy cost paid by the insurance company between July 2011 and June 2012
diag_count	Count of diagnostic disease
std_count	Count of std drug code that this patient had used
lab	Count of lab tests
ov	Count of office visit
log_spe_rx	Log of total specialty drug costs
log_main_rx	Log of total maintenance drug costs
specialty	Whether this patient has used any special drugs
hitech_rad	Count of high-tech radiology this patient has used
dialysis	Dialysis indicator
maint	Maintenance drug indicator
Sic_2d	Two digits industry type code
oon	Out of network indicator
Log2012	Log of variable *cost_total_2012*
logmed	Log of variable *cost_med*
logphm	Log of variable *cost_phm*

Before building any model, I checked the Pearson correlation coefficients between the response variable and all other explanatory variables. All the variables with absolute correlation coefficients greater than 0.1 are shown in table 4.4; it can give us a rough idea about variable importance for the later models. For nonparametric measures of association, we may use other tests such as Spearman rank-order correlation, Kendall's tau-b coefficient or Hoeffding's measure of

dependence. Rather than checking the non-linear relationships through those tests, it would be more appropriated to discuss the details through nonlinear models in the later sections.

Table 4.4 Pearson correlation coefficients (abs>0.1)

variable	cost_total_2012	variable	cost_total_2012
PROSPTCV_RISK_NUM	0.328	fact_196	0.138
cost_phm	0.286	chemotherapy	0.137
cost_med	0.285	d020	0.124
diag_count	0.264	d062	0.121
cost_maint_rx	0.261	etg3881	0.121
spfc_count	0.242	fact_114	0.120
std_count	0.233	std76	0.117
ETG_COUNT	0.222	fact_486	0.116
lab	0.208	std48	0.115
ov	0.205	d064	0.115
log_spe_rx	0.202	std40	0.113
fact_920	0.192	std71	0.111
log_main_rx	0.191	etg5554	0.110
specialty	0.190	std01	0.108
cost_specialty_rx	0.189	d018	0.108
logphm	0.184	std11	0.108
log_spe_ccdr	0.179	etg7122	0.107
cost_specialty_ccdr	0.168	oon	0.107
logmed	0.162	etg1630	0.106
hitech_rad	0.154	homehealth	0.106
fact_487	0.150	d076	0.104
inpatient	0.147	std65	0.103
fact_437	0.141	std58	0.101
MBR_AGE	0.140	d090	0.101
fact_488	0.140	d041	0.100
dialysis	0.140		

4.1.3 Data Partition and Collinearity Check

- **Data Partition**

In predictive modelling, one common strategy to assess model performance is to split the data source into training set, validation set and test set. We often use the training set for preliminary model fitting; and the validation set to prevent a modeling node from over-fitting, and to compare prediction models. The test set is then used for a final assessment of the chosen model. In this dissertation, the percentages for training, validation and test are set to be 50%, 25% and 25%,

respectively. For testing purpose, I also selected a random sample with 6000, 3000 and 3000 observations due to the limit computation power of the computer to perform the experiment. Once the models and programs were developed based on the small sample, we can refit the models using all the observations.

- **Collinearity Check**

When one explanatory variable is nearly a linear combination of other explanatory variables in the model, the affected estimates will be unstable and may result in high standard errors. This problem is called collinearity. It is always suggested to find out which variables are nearly collinear with other variables before modelling. I started with scatterplot matrix. The analysis was performed on 18 continuous variables and count variables. If we plot all variables in one scatterplot, it would be too small and vague for each small figure to be recognized. Hence we plot all the costs related variables within one figure, and other variables in another scatterplot. Actually when the explanatory variables are also right skewed and heavy tailed, a common strategy to improve model fitting is to apply log transformation. Meanwhile, the transformations can help reduce collinearity. For example, if we look at the variable “cost_maint_rx” (total maintenance drug costs) with the variable “cost_phm_june” (total pharmacy costs), the collinearity between these two variables on the original scale (fig. 4.5) is much stronger than that on the log scale (fig.4.6). And fig. 4.6 tells us that there exits collinearity between variable “spfc_count” (Count of specific rx classes that the patient has) and “std_count” (Count of std drug codes that the patient has). We can further check the collinarity by condition index shown in table 4.5.

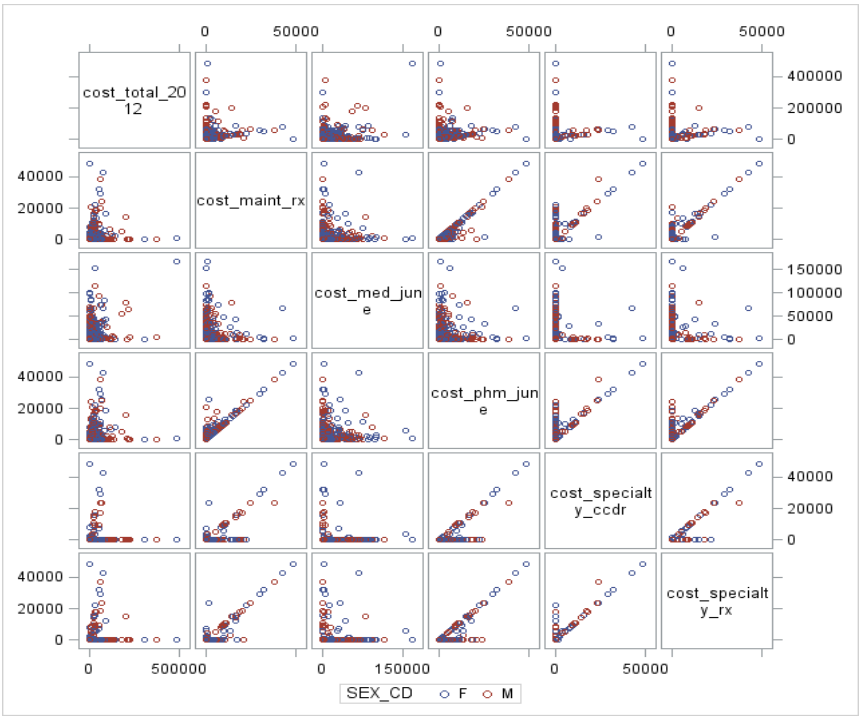


Fig. 4.5 Scatterplot matrix on cost related variables

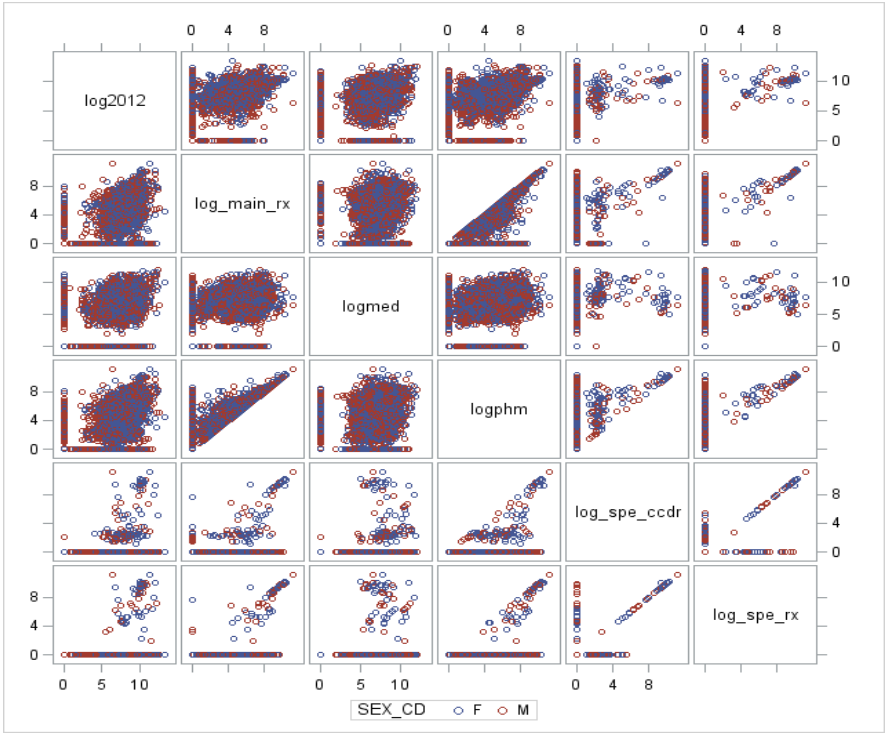


Fig. 4.6 Scatterplot matrix on log cost related variables

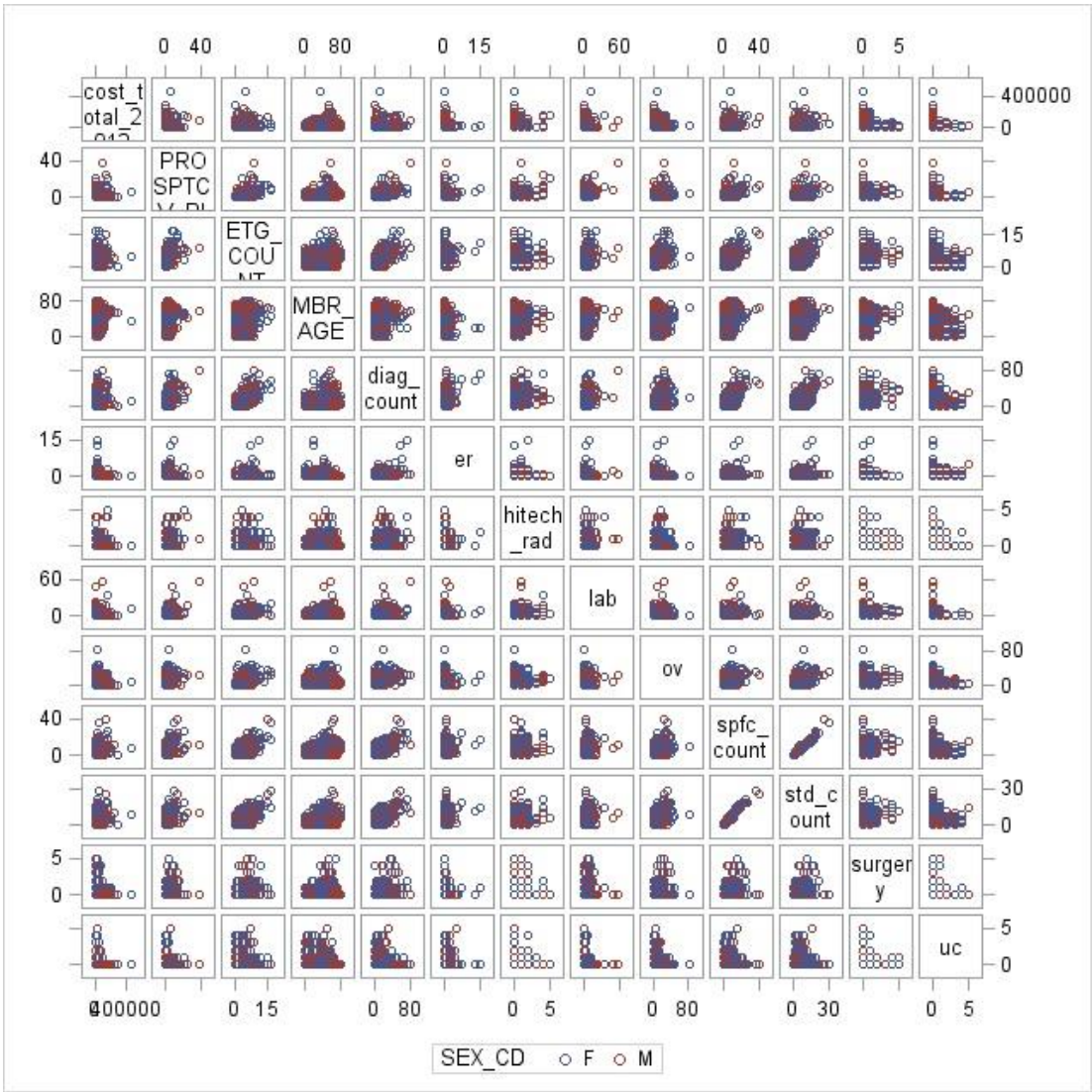


Fig. 4.7 Scatterplot matrix on other continous data

After the initial exploration through scatterplot matrix, I further checked the collinearity by the condition indices through Proc Reg in SAS. Belsey, Kuh, and Welsch (1980) suggest that when this number is around 10, weak dependencies might be beginning to affect the estimates. When this number is greater than 100, the estimates might have a fair amount of numerical error. PROC REG in SAS generates the proportion of the variance of the estimate accounted for by each principal component. A collinearity problem might occur when a component associates with a high condition index contributes strongly (variance proportion larger than about 0.5) to the variance of

two or more variables. In table 4.5, the condition index in the last row is 23.969, which is significantly greater than 10. Meanwhile, variable *spfc_count* and *std_count* both have proportion of variation greater than 0.5. Hereby I believe the evidence of collinearity between these two variables is strong. There are many ways to deal with collinearity such as variable reduction by Principal Components Analysis (PCA), Shrinkage methods, manual variable selection, and variable reduction via partial least squares etc. Those two variables would be closely watched and treated correspondingly in later models.

Table 4.5 Collinearity diagnostics

Collinearity Diagnostics (intercept adjusted)											
Number	Eigenvalue	Condition Index	Proportion of Variation								
			ETG_COUNT	MBR_AGE	RISK_NUM	logphm	logmed	spfc_count	std_count	log_main_rx	log_spe_rx
1	6.828	1.000	0.004	0.002	0.005	0.002	0.005	0.000	0.000	0.003	0.001
2	1.647	2.036	0.002	0.000	0.002	0.000	0.003	0.000	0.000	0.000	0.125
3	1.332	2.264	0.000	0.054	0.001	0.018	0.011	0.001	0.001	0.031	0.000
4	1.155	2.432	0.000	0.164	0.024	0.005	0.009	0.000	0.000	0.001	0.001
5	0.979	2.641	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	0.953	2.677	0.000	0.025	0.001	0.001	0.001	0.000	0.000	0.001	0.000
7	0.848	2.838	0.000	0.281	0.019	0.002	0.014	0.000	0.000	0.000	0.000
8	0.834	2.861	0.003	0.002	0.021	0.004	0.000	0.000	0.000	0.003	0.001
9	0.746	3.026	0.005	0.016	0.009	0.005	0.010	0.000	0.000	0.007	0.000
10	0.593	3.392	0.048	0.011	0.053	0.019	0.155	0.002	0.002	0.020	0.000
11	0.508	3.665	0.013	0.071	0.005	0.004	0.261	0.005	0.005	0.002	0.005
12	0.418	4.039	0.027	0.325	0.386	0.009	0.002	0.002	0.002	0.025	0.007
13	0.350	4.420	0.016	0.000	0.001	0.015	0.341	0.004	0.005	0.106	0.004
14	0.266	5.064	0.203	0.002	0.046	0.003	0.051	0.003	0.003	0.018	0.456
15	0.234	5.397	0.255	0.006	0.399	0.001	0.025	0.001	0.001	0.001	0.391
16	0.183	6.107	0.416	0.022	0.019	0.001	0.001	0.000	0.000	0.032	0.006
17	0.114	7.745	0.006	0.018	0.006	0.905	0.104	0.001	0.001	0.746	0.000
18	0.012	**23.969**	0.001	0.001	0.003	0.005	0.006	**0.979**	**0.980**	0.004	0.003

4.2 Models

This section shows the model fitting and testing results for each model in the analytics framework introduced in section 1.3.2. Later in section 4.3, all the results would be put together for comparison and model selection.

4.2.1 Two Part Model

Two Part model is the most widely used for semi-continuous data. In this project, I use logistic regression to model the binary target (If the total cost is positive, the binary target for logistic regression is 1. If the total cost is zero, the binary target for the logistic regression is 0). For part two, four different options will be discussed and compared later. In short, there are two stages for the model:

- Stage 1: Create a binary indicator variable to show whether target variable is positive or not; fit to the data using logistic regression on the indicator variable.
- Stage 2: Amount models (linear or nonlinear) on nonzero cost.

4.2.1.1 Part 1 Logistic GLM

When the target is a binary variable, the logistic GLM model is almost the default model with good performance in a bunch of cases. In fact, we compared Logistic GEE model, decision tree model and probit model with logistics GLM. Their performances are quite similar in terms of fit statistics and miss-classification rate on hold-out samples. Hereby I simply adopt the simplest logistic GLM in this dissertation.

Let's say the probability of positive cost is p, and then the probability of zero cost is $q = 1 - p$. Then the odds of success are defined as $odds(success) = p/q$. There is a direct relationship between the coefficients and the odds ratios. A logit is defined on the log base of the odds, that

is $logit(p) = \log(odds) = \log(\frac{p}{q})$. Therefore, logistic regression is actually an ordinary regression using the logit as the response variable

$$logit(p) = a + bX \qquad or \qquad \log\left(\frac{p}{q}\right) = a + bX \tag{4.1}$$

This means that the coefficients in logistic regression are in terms of the log odds. Let's take variable *logmed* as an example. For every one unit change in *logmed*, the log odds of positive cost (versus zero cost) increases by 0.1878. Table 4.6 gives the coefficients as odds ratios and fig. 4.8 shows the odds ratio estimates with 95% profile-likelihood confidence limits. An odds ratio is the exponentiated coefficient, and can be interpreted as the multiplicative change in the odds for a one unit change in the predictor variable. For example, for a one unit increase in *logmed*, the odds of cost being positive (versus zero cost) increase by a factor of 1.207. Similarly, we can interpret the results for other explanatory variables.

Table 4.6 Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	-0.4712	0.1451	10.5479	0.0012	0.624
logmed	1	0.1878	0.0186	101.9222	<.0001	1.207
logphm	1	0.2184	0.027	65.2221	<.0001	1.244
ETG_COUNT	1	0.4003	0.067	35.7509	<.0001	1.492
ACCT_SUBTY	1	-2.706	0.3616	56.0145	<.0001	0.067
ACCT_SUBTY	1	0.9135	0.1431	40.7373	<.0001	2.493
ACCT_SUBTY	1	1.2404	0.1845	45.2237	<.0001	3.457
fact_437	1	1.4302	0.3157	20.5184	<.0001	4.18
SEX_CD	1	0.3258	0.0488	44.5383	<.0001	1.385
etg1630	1	-1.679	0.407	17.0185	<.0001	0.187
etg3866	1	-3.8203	1.5587	6.0071	0.0142	0.022
etg5220	1	-4.6851	1.6268	8.2935	0.004	0.009
etg7119	1	-1.8323	0.4794	14.611	0.0001	0.16
fact_114	1	1.4825	0.3241	20.9246	<.0001	4.404
std42	1	-0.8509	0.2118	16.1388	<.0001	0.427

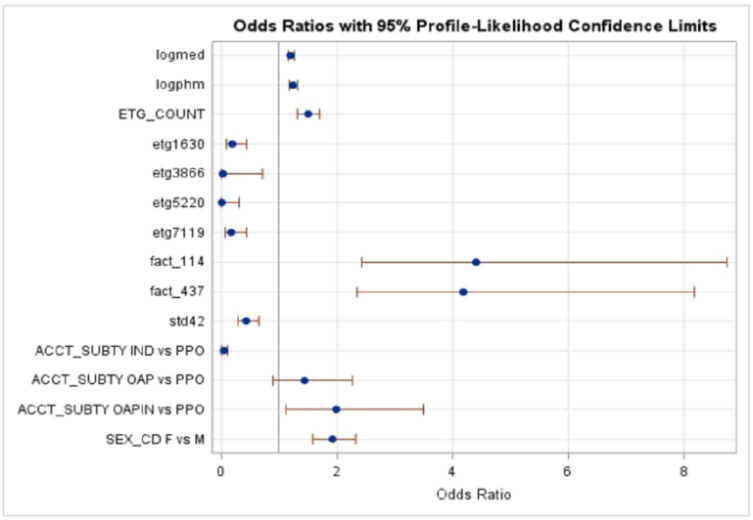


Fig.4.8 Odds Ratio estimates with 95% Profile-likelihood Confidence limits

4.2.1.2 Part 2: Normal GLM with Log Link

In this section, GLM model is fitted with normality assumption on the response variable and log link function by the HPGENSELECT procedure in SAS. The default optimization technique used is a modification of the Newton-Raphson algorithm with a ridged Hessian. Compared to other procedures such as GENMOD, this procedure is mainly designed for large-data tasks in predictive model building, model fitting, and scoring. The stepwise AICc is performed for the variable selection, and both entry significance level and stay significance level are set to be 0.05. In fact, with larger sample sizes, the AIC and AICc are almost the same, as shown in table 4.7. The variables selected and parameter estimates are shown in table 4.9 including parameter estimates, standard errors and P values. The P values indicate that almost all the variables we selected are statistical significant.

Table 4.7 Fit Statistics

-2 Log Likelihood	123331
AIC	123379
AICc	123379
BIC	123536

Table 4.8 Variable Selection information

Selection Method	Stepwise
Stop Criterion	Significance Level
Choose Criterion	AICc
Entry Significance Level (SLE)	0.05
Stay Significance Level (SLS)	0.05

Table 4.9 Parameter Estimates (Two Part Model, Normal with log link)

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	7.171442	0.451093	252.7438	<.0001
logmed	1	0.175654	0.054991	10.2032	0.0014
d018	1	0.866223	0.16589	27.2658	<.0001
d021	1	2.403915	0.368956	42.4512	<.0001
d027	1	3.567859	0.356186	100.337	<.0001
d059	1	1.83174	0.379522	23.2945	<.0001
d066	1	0.764991	0.195846	15.2575	<.0001
d089	1	1.956065	0.2826	47.9096	<.0001
dialysis	1	3.837655	0.308452	154.7952	<.0001
etg1631	1	2.660649	0.405402	43.0728	<.0001
etg1647	1	-1.023005	0.347728	8.6552	0.0033
etg2390	1	5.096248	0.572284	79.3008	<.0001
etg4778	1	2.045116	0.370307	30.5009	<.0001
etg5216	1	3.033728	0.237576	163.0608	<.0001
etg5220	1	2.247404	0.372154	36.4684	<.0001
fact_487	1	0.99291	0.241326	16.9283	<.0001
fact_493	1	0.742112	0.219793	11.4001	0.0007
fact_530	1	0.981075	0.314652	9.7217	0.0018
specialty	1	1.208953	0.15955	57.4147	<.0001
std11	1	0.708388	0.157055	20.3441	<.0001
std23	1	2.182762	0.332788	43.0207	<.0001
std75	1	3.72381	0.529402	49.4771	<.0001
maint 0	1	-0.936439	0.541305	2.9928	0.0836

4.2.1.3 Part 2: Lognormal GLM with Identity link

Recall that the histogram of nonzero target costs on the original scale shown in fig. 4.3 and 4.4, it is highly positively skewed, but more symmetric and normal distributed after the log transformation. It gives us a hint to use lognormal distribution for the dependent variable in GLM. In this model, we use the elastic net method for model selection. Unlike the LASSO method which is upper-bounded by the number of training samples, there is no restriction for the elastic net method, which incorporates an additional ridge regression penalty. The optimal value of criterion is chosen to be the minimal validation ASE (Average Squared Error). Fig 4.9 shows the coefficients progression of all the effects selected, which is plotted as a function of the step

number. In fig.4.9, the minimal ASE occurs at step 249. The fit statistics and parameter estimates are summarized in table 4.10 and 4.11 respectively.

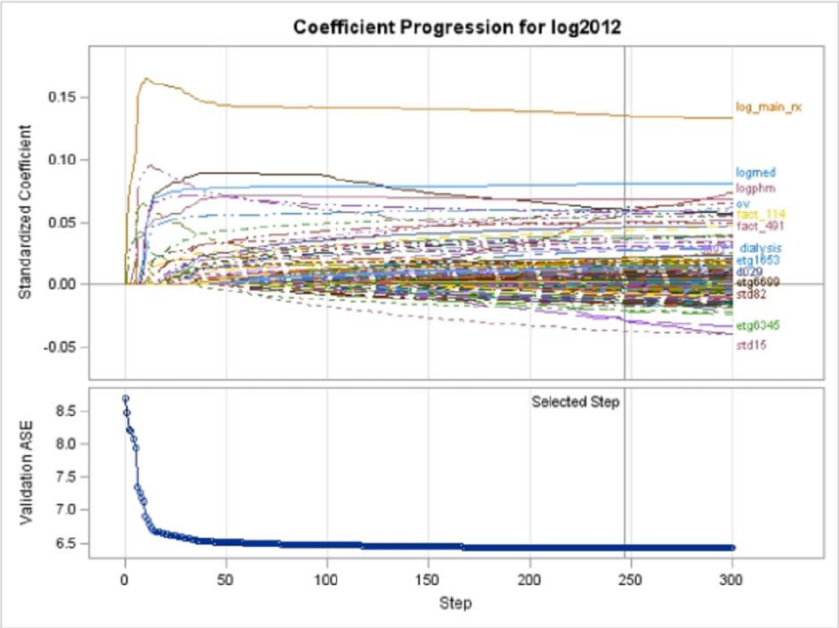


Fig.4.9 Elastic Net Coefficient Progression

Table 4.10 Fit Statistics

Log Likelihood	-8958.91
AIC (smaller is better)	17957.82
AICC (smaller is better)	17957.98
BIC (smaller is better)	18089.11

Table 4.11 Analysis of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > Chi Sq
Intercept		1	4.835	0.217	4.410	5.259	498.420	<.0001
predict		1	0.567	0.277	0.025	1.109	4.210	0.040
RISK_NUM		1	0.095	0.016	0.063	0.127	34.220	<.0001
logmed		1	0.038	0.012	0.014	0.061	9.710	0.002
logphm		1	0.033	0.015	0.004	0.062	4.870	0.027
diag_count		1	0.019	0.005	0.009	0.028	15.200	<.0001
etg2390		1	5.347	1.338	2.724	7.969	15.960	<.0001
log_main_rx		1	0.188	0.018	0.154	0.223	113.270	<.0001
fact_488		1	0.511	0.132	0.252	0.769	14.990	0.000
fact_491		1	0.454	0.125	0.208	0.700	13.120	0.000

fact_493		1	0.691	0.123	0.449	0.933	31.380	<.0001
fact_920		1	0.568	0.133	0.308	0.828	18.300	<.0001
MBR_AGE		1	0.009	0.001	0.007	0.011	58.620	<.0001
ov		1	0.023	0.006	0.011	0.034	15.420	<.0001
oon		1	0.158	0.056	0.049	0.267	8.100	0.004
inpatient		1	-0.159	0.082	-0.319	0.001	3.790	0.052
log_spe_rx		1	0.092	0.021	0.051	0.133	19.240	<.0001
maint	0	1	0.597	0.084	0.433	0.761	51.030	<.0001
SEX_CD	F	1	0.169	0.039	0.092	0.245	18.750	<.0001

4.2.1.4 Part 2: Gamma GLM with log link

Gamma GLM with log link is almost the default model in industry for positive continuous target variable. Gamma distribution has heavier tail than normal distribution, but actuarially not heavy enough to capture those extremely fat-tailed data. Therefore it is necessary for us to test this model through real data in industry, and summarize its pros and cons. Here is the gamma GLM with log link:

$$E[Y] = \exp(X\beta) \ or \log(E[Y]) = X\beta \ , Y \sim Gamma \tag{4.2}$$

Where $E[Y]$ is the expected value of Y, $X\beta$ is the linear predictor (a linear combination of unknown parameters β). The model is fitted using the HPGENSELECT procedure in SAS. The stepwise AICc was performed for the variable selection, and both entry significance level and stay significance level were set to be 0.05. The variables selected and parameter estimates are shown in table 4.14. The two variables *PROSPTCV_RISK_CAT* (prospective ERG risk category) and *Sic_2d* (2 digits industry type) are also selected by this model; but the parameter estimation are not shown here because there are too many levels in those two categorical variables.

Table 4.12 Fit Statistics

-2 Log Likelihood	95520
AIC	95740
AICc	95745
BIC	96462

Table 4.13 Variable Selection information

Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Choose Criterion	AICc
Entry Significance Level (SLE)	0.05
Stay Significance Level (SLS)	0.05

Table 4.14 Parameter Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	4.911849	1.077172	20.7931	<.0001
chemotherapy	1	1.362626	0.26218	27.0118	<.0001
cost_med_june	1	1.24E-05	2.5E-06	24.8682	<.0001
cost_phm_june	1	0.000107	1.15E-05	86.4509	<.0001
d027	1	3.777966	1.238244	9.309	0.0023
etg2390	1	4.483938	1.185819	14.2982	0.0002
etg3143	1	2.043067	1.184566	2.9747	0.0846
etg5216	1	2.587023	0.693548	13.9139	0.0002
fact_487	1	1.326886	0.19645	45.621	<.0001
fact_488	1	0.770892	0.118675	42.1957	<.0001
fact_491	1	0.542382	0.112062	23.4256	<.0001
fact_493	1	0.569474	0.110997	26.3224	<.0001
fact_920	1	0.824272	0.11657	49.9998	<.0001
spfc_count	1	0.047923	0.006622	52.3668	<.0001
std75	1	3.179548	1.191045	7.1265	0.0076

4.2.1.5 Part 2: lognormal GLMM with identity link

In this generalized liner mixed model (GLMM), the distribution of the response variable is assumed to be lognormal. The overall Modeling Framework is set as two levels as follows:

- Level 1: Individual Level

$$Y_{ap} = \partial_a + X_{ap}\beta + \varepsilon_{ap}, \quad \varepsilon_{ap} \sim N\left(0, \sigma^2\right) \tag{4.3}$$

- Level 2: Group level

$$\partial_a = \mu_a + Z_a\gamma + \eta_a, \quad \eta_a \sim N\left(0, \sigma_a^2\right) \tag{4.4}$$

Or equivalently, we can write our model in the following way:

$$Y_{ap} = \partial_a + X_{ap}\beta + \varepsilon_{ap}, \partial_a \sim N\left(\mu_a + Z_a\gamma, \sigma_a^2\right), \varepsilon_{ap} \sim N\left(0, \sigma^2\right) \quad (4.5)$$

Where X_{ap} are individual level predictors available for both current customers and prospective customers,

Z_a : Account level predictors

a : Account index

p : Individual index

∂_a : Random group level intercepts

β : Coefficients for individual level predictors

γ : Coefficients for account level predictors

ε_{ap} : R side covariance matrix in linear mixed model

η_a : G side covariance matrix in linear mixed model.

For the linear mixed model, a common question is that how to specify the fixed-effects design matrix and random-effects design matrix. To put it simply, if we believe a qualitative variable whose levels are randomly sampled from a population of levels being studied, it is appropriate for us to model it as a random effect. In this model, I want to explore the dependence among samples from the same account to improve group insurance pricing, hence building a simplest GLMM with random account intercept only. This model is fitted by Proc GLIMMIX in SAS using Restricted Maximum Likelihood (REML) estimation. REML is used because the full Maximal likelihood estimates of the variance components are generally biased downwards, but this problem can be eliminated by using the REML. Kreft & de Leeuw (1998) explain that the idea of REML is to apply the principle of maximum likelihood to the least-squares residuals. It means that we first remove the effect of the fixed variables, then the distribution of the residuals no longer depends on the estimates of the fixed effects but only depends on the variance component. It factors the likelihood

into two parts: one of which does not depend on β (vector of parameters in the fixed part), but only on v (the vector of all the unknown parameters in variance components) in mixed models. The fit statistics of GLMM is shown in table 4.15. CAIC is short for conditional AIC (conditional on the ML estimates of the fixed effects and empirical Bayes estimates of the random effects), which is used to compare a series of mixed models with different random effects structures. Table 4.16 shows the solution for fixed effects and fig. 4.10 shows the conditional Pearson residual plot for the target variable. The solutions of random effects are not shown here because there are too many account levels in the random parts.

Table 4.15 Fit Statistics

-2 Res Log Likelihood	18067.10
AIC (smaller is better)	18071.1
AICc (smaller is better)	18071.1
BIC (smaller is better)	18081.55
CAIC (smaller is better)	18083.55

Table 4.16 Solutions for Fixed Effects

Effect/Variables	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	4.761	0.2169	5204	21.95	<.0001
predict	0.6446	0.2773	5195	2.32	0.0201
PROSPTCV_RISK_NUM	0.09082	0.01618	5223	5.61	<.0001
logmed	0.03612	0.01212	5209	2.98	0.0029
logphm	0.03182	0.01499	5174	2.12	0.0338
diag_count	0.01488	0.004422	5191	3.37	0.0008
etg2390	5.3493	1.348	5225	3.97	<.0001
log_main_rx	0.1918	0.01774	5223	10.81	<.0001
fact_488	0.5059	0.1329	5221	3.81	0.0001
fact_491	0.4557	0.1263	5225	3.61	0.0003
fact_493	0.6959	0.1242	5225	5.6	<.0001
fact_920	0.5766	0.1337	5225	4.31	<.0001
MBR_AGE	0.009151	0.001167	5211	7.84	<.0001
ov	0.0252	0.005716	5218	4.41	<.0001
oon	0.1576	0.05601	5207	2.81	0.0049
log_spe_rx	0.09203	0.02115	5221	4.35	<.0001
maint	0.6082	0.08401	5219	7.24	<.0001
SEX_CD	0.1665	0.03922	5132	4.24	<.0001

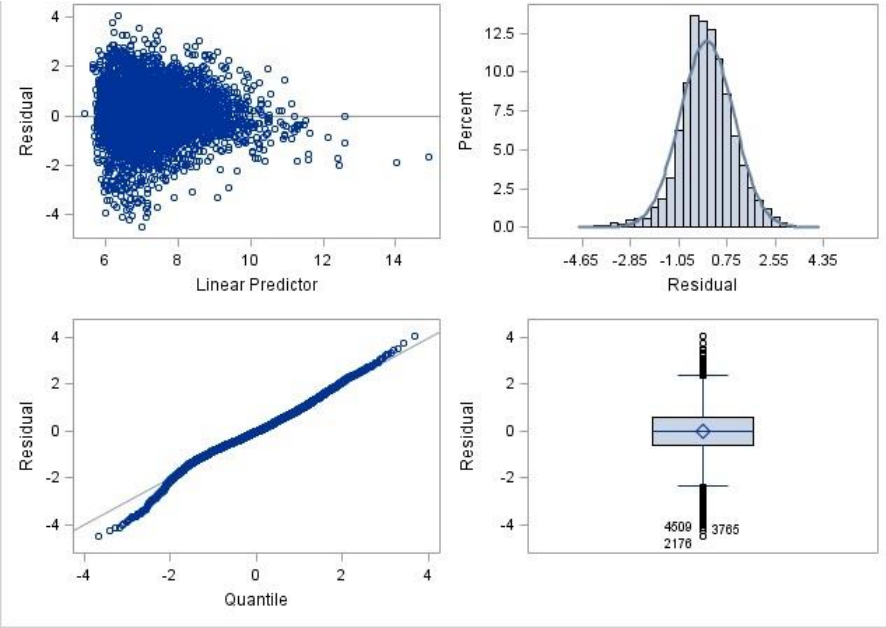


Fig.4.10 Conditional Pearson Residuals for Target variable

4.2.1.6 Part 2: GEE

In general, because the model fit of GEE is really an estimating procedure and there is no likelihood function, we don't need to test for the model fit of the GEE. But we can check the empirical estimate of the standard errors and covariance. I compared the empirical estimates with the model-based estimates. For output, we can still use overall goodness-of-fit statistics such as Pearson chi-square statistic, X^2 and corresponding P value shown in table 4.19, and QIC/QICu shown in table 4.17. The analysis of GEE parameter estimates are shown in table 4.18.

The QIC (Quasi-likelihood under the Independence model Criterion) statistic was introduced by Pan (2001) and further discussed by Hardin and Hilbe (2003). It is analogous to the commonly used AIC (Akaike's Information Criterion) statistic for comparing models fit with likelihood-based methods. Because the generalized estimating equations (GEE) method is not a likelihood-based method, the AIC statistic is hereby not available. First, let's review the formula for AIC:

$$AIC = -2 * \log likelihood + 2k \quad (4.6)$$

Where k is the number of parameters in the model.

And the formula for QIC by Pan (2001) is

$$QIC = -2Q(\hat{\mu}; I) + 2trace(\hat{\Omega}_I^{-1}\hat{V}_R) \quad (4.7)$$

Where I represents the independent covariance structure used to calculate the quasi-likelihood. $\hat{\mu} = g^{-1}(x\hat{\beta})$ and $g^{-1}()$ is the inverse link function. The coefficient estimates $\hat{\beta}$ and robust variance estimator $\hat{V}_R$ are obtained from a general working covariance structure R. Another variance estimator $\hat{\Omega}_I$ is obtained under the assumption of an independence correlation structure.

Table 4.17 GEE fit criteria

QIC	5264.683
QICu	5260

Table 4.18 Analysis of GEE parameter estimates

Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	4.7626	0.3174	4.1406	5.3846	15.01	<.0001
predict	0.6472	0.3922	-0.1215	1.4158	1.65	0.0989
PROSPTCV_RISK_NUM	0.0909	0.0189	0.0538	0.128	4.8	<.0001
logmed	0.0363	0.0147	0.0074	0.0651	2.46	0.0138
logphm	0.0306	0.0178	-0.0043	0.0656	1.72	0.086
diag_count	0.0149	0.0049	0.0053	0.0245	3.05	0.0023
log_main_rx	0.1919	0.0196	0.1534	0.2303	9.78	<.0001
fact_488	0.5051	0.154	0.2033	0.807	3.28	0.001
fact_491	0.4524	0.1286	0.2003	0.7045	3.52	0.0004
fact_493	0.6932	0.1072	0.4831	0.9034	6.46	<.0001
fact_920	0.5727	0.1551	0.2687	0.8768	3.69	0.0002
MBR_AGE	0.0093	0.0012	0.007	0.0116	7.84	<.0001
ov	0.0253	0.0058	0.0139	0.0367	4.34	<.0001
oon	0.1554	0.0519	0.0538	0.2571	3	0.0027
log_spe_rx	0.0921	0.0142	0.0643	0.1199	6.49	<.0001
maint	0.6116	0.106	0.4038	0.8193	5.77	<.0001
SEX_CD	0.1629	0.0395	0.0856	0.2403	4.13	<.0001

Table 4.19 Score Statistics For Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
predict	1	2.62	0.1055
PROSPTCV_RISK_NUM	1	23.14	<.0001
logmed	1	5.78	0.0162
logphm	1	3	0.0831
diag_count	1	8.05	0.0045
log_main_rx	1	65.11	<.0001
fact_488	1	10.12	0.0015
fact_491	1	11.23	0.0008
fact_493	1	27.48	<.0001
fact_920	1	9.45	0.0021
MBR_AGE	1	40.15	<.0001
ov	1	19.02	<.0001
oon	1	9.13	0.0025
log_spe_rx	1	24.94	<.0001
maint	1	31.88	<.0001
SEX_CD	1	14.43	0.0001

4.2.2 Tweedie Model

4.2.2.1 Tweedie GLM

Tweedie GLM is fitted by PROC HPGENSELECT in SAS. Tweedie Index is optimized as 1.698313 and the corresponding dispersion parameter is 16.96895. The stepwise AICc is performed for the variable selection, and both entry significance level and stay significance level are set to be 0.05. The variables selected and parameter estimates are shown in table 4.22.

Table 4.20 Fit statistics

-2 Log Likelihood	99729
AIC	100019
AICc	100026
BIC	100991

Table 4.21 Variable selection information

Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Choose Criterion	AICc
Entry Significance Level (SLE)	0.05
Stay Significance Level (SLS)	0.05
Optimization Technique	Quasi-Newton

Table 4.22 Parameter Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	7.46317	0.79825	87.4116	<.0001
chemotherapy	1	1.421462	0.181596	61.2717	<.0001
cost_med_june	1	5.88E-06	1.45E-06	16.429	<.0001
cost_phm_june	1	6.15E-05	6.61E-06	86.7054	<.0001
logphm	1	0.072551	0.008	82.2482	<.0001
d023	1	0.909706	0.260086	12.234	0.0005
d027	1	3.943903	0.736871	28.6464	<.0001
d054	1	-0.1319	0.043343	9.2611	0.0023
d063	1	-0.72493	0.172762	17.6072	<.0001
d067	1	-0.50255	0.139462	12.9852	0.0003
d077	1	-0.22842	0.060652	14.1828	0.0002
dialysis	1	3.852647	0.662144	33.8543	<.0001
etg1620	1	1.580971	0.631573	6.2661	0.0123
etg1647	1	-0.20155	0.056532	12.7108	0.0004
etg2072	1	1.11362	0.405484	7.5427	0.006
etg2390	1	4.604221	0.636297	52.3592	<.0001
etg2716	1	0.3266	0.126027	6.7159	0.0096
etg3143	1	2.370828	0.828392	8.1908	0.0042
etg3890	1	1.162056	0.289684	16.0918	<.0001
etg4735	1	0.843511	0.287799	8.5902	0.0034
etg4764	1	0.71574	0.231352	9.5712	0.002
etg5214	1	0.659294	0.262156	6.3247	0.0119
etg5216	1	2.186668	0.48947	19.9578	<.0001
etg5218	1	2.005464	0.552685	13.1666	0.0003
etg6022	1	1.159004	0.301412	14.786	0.0001
fact_114	1	0.221796	0.050023	19.6595	<.0001
fact_439	1	0.375079	0.138786	7.3038	0.0069
fact_487	1	1.120499	0.147766	57.5006	<.0001
fact_488	1	0.889279	0.105976	70.414	<.0001
fact_491	1	0.590758	0.094155	39.3668	<.0001
fact_492	1	0.520884	0.188848	7.6077	0.0058
fact_493	1	0.460491	0.096112	22.9556	<.0001
fact_494	1	0.431202	0.084936	25.7737	<.0001
fact_920	1	0.867	0.096098	81.398	<.0001
oon	1	0.159425	0.046935	11.5376	0.0007
ov	1	0.012854	0.00432	8.8525	0.0029
std07	1	0.242032	0.066051	13.4274	0.0002
std15	1	-0.22468	0.065185	11.8807	0.0006

std22	1	-0.15715	0.048449	10.5213	0.0012
std25	1	0.168281	0.049864	11.3891	0.0007
std27	1	0.249985	0.058153	18.479	<.0001
std48	1	0.253085	0.077993	10.5297	0.0012
std58	1	0.209947	0.072028	8.496	0.0036
std75	1	3.651833	0.692234	27.8302	<.0001

4.2.2.2 Tweedie GAM

A number of smooth functions are available for GAM. In our models, both cubic regression splines and thin plate regression splines (TPRS) are tried, the latter being the default for a GAM in this mgcv package in R. As a brief summary, both work well in general. I select cubic regression in this model. It is known that generalized linear models emphasize estimation and inference for the parameters of the model, while generalized additive models focus on exploring data nonparametrically. I used *tweedie.profile* function in R to estimate the optimal Tweedie Index power through MLE. The optimal Tweedie index is 1.787755 (Shown in fig.4.11) and the estimated scale is 27.219.

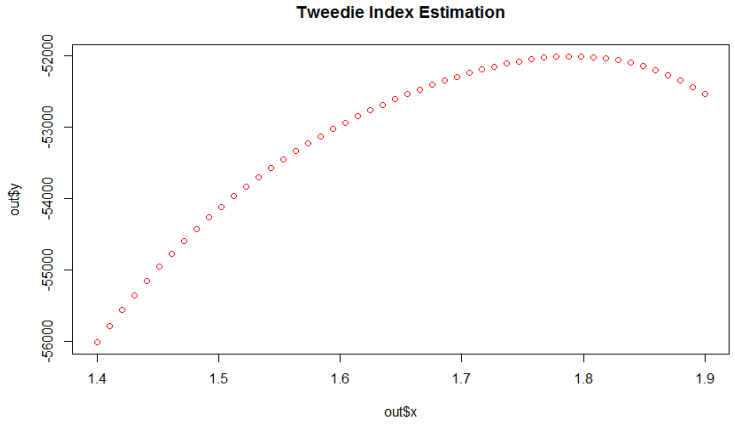


Fig. 4.11 Optimal Tweedie index parameter

The variable selection is done through AIC stepwise model selection using function stepGAIC() in package GAMLSS in R. For models with smoothing additive terms, a term formula might be

$$\sim 1 + x + s(x) \tag{4.8}$$

where x is an explanatory variable.

This means that an explanatory variable could either appear not at all, linearly, or as a smooth function estimated nonparametrically. The 1 in the formula (4.8) gives the option of leaving the term out of the model entirely. In the model, every term is described by such a term formula (4.8). Then the final model will be built up by selecting a component from each formula. The process is repeated until either the AIC criterion cannot be decreased by any of the eligible steps, or until the maximum number of steps has been used. Many of the standard results computed by GAM are similar to those results reported by other linear or nonlinear model fitting procedures. Both predicted and residual values for the final model can be computed, and various graphs of the residuals can be displayed to help us identify possible outliers, etc. The model fitting results are shown in table 4.23 and 4.24, respectively. The effective degrees of freedom for all those four smooth terms are between 6 and 7, which indicate nonlinear relationships.

Table 4.23 Tweedie GAM Parametric coefficients

	Estimate	Std.Erro	t value	Pr(>|t|)	Signif
(Intercept)	0	0	NA	NA	
logmed	0.402139	0.005707	70.462	2.00E-16	***
PROSPTCV_RISK_NUM	0.118163	0.024398	4.843	1.31E-06	***
SEX_CDM	-0.02364	0.008911	-2.652	0.00801	**

Table 4.24 Tweedie GAM Approximate significance of smooth terms

	edf	Ref.df	F	p-value	Signif.
s(logphm)	6.572	7.684	22.919	2.00E-16	***
s(logmed)	6.761	7.859	530.64	2.00E-16	***
s(PROSPTCV_RISK_NUM)	6.592	7.476	33.827	2.00E-16	***
s(diag_count)	6.951	7.852	2.654	0.00713	**

The main results of interest of GAM are how the predictors are related to the dependent variable. The smoothing component plot in fig. 4.12 shows several things. The solid line is the predicted value of the dependent variable as a function of the x axis. The small lines along the x axis are the "rug", showing the location of the sample plots. The default is to plot all four

smooths on the same scale. Accordingly, the results for the *logphm* looks flat because the standard errors on the smooths of other components are relatively large, such as PROSPTCV_RISK_NUM (prospective ERG risk number).

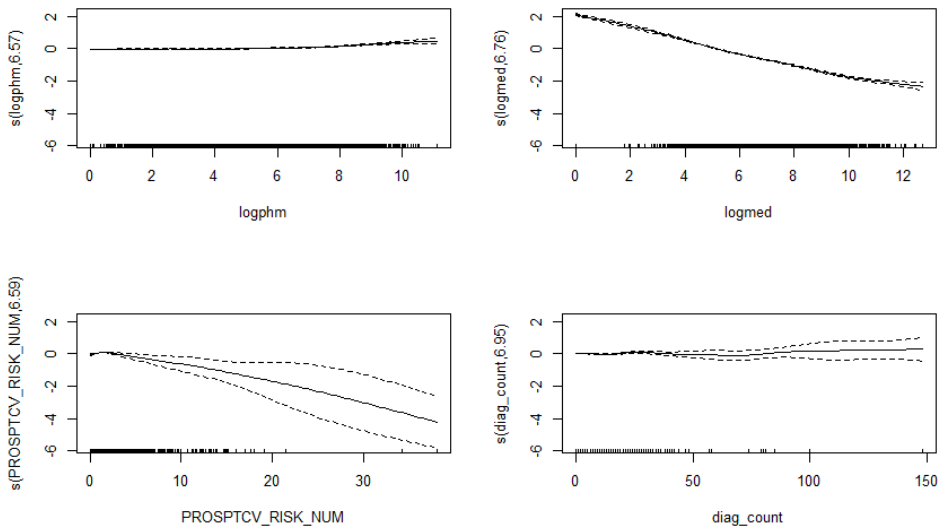


Fig. 4.12 Smoothing component plot

4.2.3 Quantile Regression Forest

The quantile regression forest is fitted using package 'quantregForest' in R developed by Nicolai Meinshausen. The random forest fitted in our model is composed of 1000 decision trees. Number of variable tried at each split is optimized by minimizing out-of-bag (oob) error and our experiment selects 4. Summary of predicted quantile on training samples and hold-out samples are shown in table 4.25 and 4.26, respectively. Actually, the quantiles are quite similar on the training samples and hold-out samples.

Table 4.25 Summary of predicted quantiles on training samples

	quantile= 0.1	quantile= 0.5	quantile= 0.9
Min.	0.00	0.00	1661.00
1st Qu.	0.00	326.40	2910.00
Median	10.79	752.40	4617.00
Mean	567.08	2114.30	9561.00
3rd Qu.	424.84	2095.60	9703.00
Max.	21499.09	64290.60	314563.00

Table 4.26 Summary of predicted quantiles on hold-out samples

	quantile= 0.1	quantile= 0.5	quantile= 0.9
Min.	0.00	18.61	1804.00
1st Qu.	0.00	334.66	2928.00
Median	0.00	703.18	4594.00
Mean	564.40	2092.67	9527.00
3rd Qu.	457.00	2151.43	9994.00
Max.	19925.20	64134.18	312853.00

Meanwhile, quantile regression forest can give us the rank of variable importance. In every tree grown in the quantile regression forest, we randomly permute the values of variable m. First, we put down the OOB cases, and count the number of votes cast for the correct class. Then we subtract the number of votes for the correct class in the variable-m-permuted OOB data from the number of votes for the correct class in the untouched OOB data[2]. The average of this number over all trees in the forest is the raw importance score for variable m, we call this raw importance score as *IncNodePurity* (shown in table 4.27). Equivalently, the variance importance can be visualized in fig.4.12.

Table 4.27 Variable importance rank

	IncNodePurity
cost_med_june	3.6099E+11
cost_phm_june	2.16109E+11
diag_count	1.96567E+11
cost_maint_rx	1.71405E+11
MBR_AGE	1.56508E+11

[2] Leo Breiman and Adele Cutler, Random Forests, https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

spfc_count	91941024055
std_count	64703345904
cost_specialty_rx	24016906106
specialty	23636318927
oon	8033814559
maint	3527891716

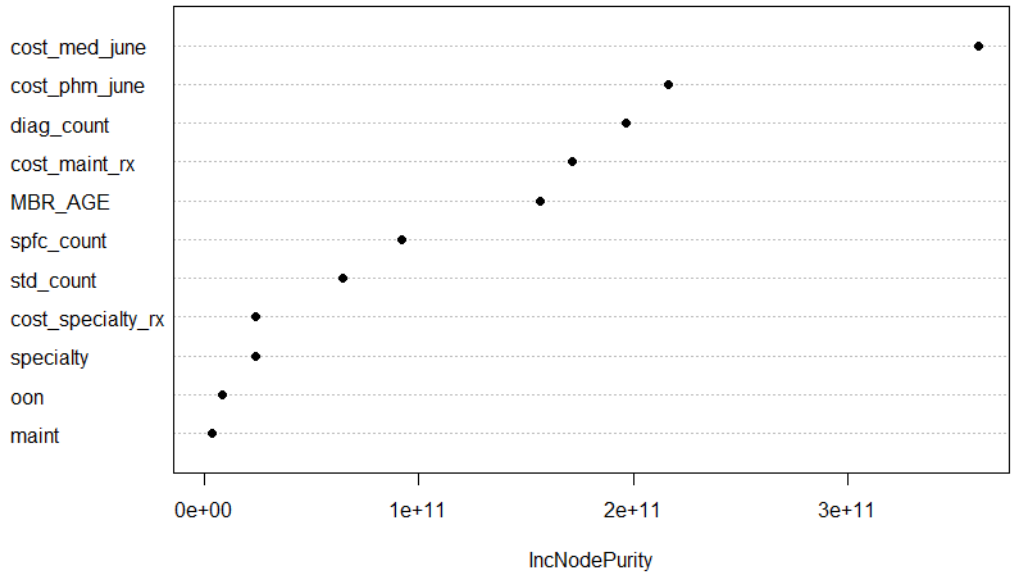


Fig. 4.12 Mean decrease accuracy

4.2.4 Finite Mixture Models

This model is estimated by Proc FMM in SAS using Maximum Likelihood estimation with Dual Quasi-Newton as the optimization technique. There are three components, one in the degenerated distribution at zero, another two components are Weibull distributed with different parameters. I choose Weibull distribution because it is a well-known heavy-tailed distribution with many appealing properties. A mixture of two Weibull components can has a number of parameters which include shape parameters, scale parameters and location parameters in addition to the mixing parameter. It is very flexible and capable of capturing the skewness and

heavy tailed property showed in our response variable. The fit statistics is shown in table 4.28 and the parameter estimates for mixed Weibull model is shown in table 4.29.

Table 4.28 FMM fit statistics

-2 Log Likelihood	100080
AIC (Smaller is Better)	100152
AICC (Smaller is Better)	100152
BIC (Smaller is Better)	100393
Effective Components	3

Table 4.29 Parameter estimates for mixed Weibull model

Component	Effect	Estimate	Component	Effect	Estimate
1	Intercept	6.7691	2	Intercept	7.3787
1	PROSPTCV_RISK_NUM	0.1438	2	PROSPTCV_RISK_NUM	-0.2267
1	logmed	0.00226	2	logmed	-0.1718
1	logphm	0.04086	2	logphm	-0.1918
1	diag_count	0.01381	2	diag_count	-0.069
1	etg2390	4.5771	2	etg2390	-0.6368
1	log_main_rx	0.06318	2	log_main_rx	-0.1429
1	fact_488	0.8375	2	fact_488	-4.22
1	fact_491	0.4509	2	fact_491	-1.1143
1	fact_493	0.5033	2	fact_493	-1.1008
1	fact_920	0.7746	2	fact_920	-0.5224
1	MBR_AGE	0.01185	2	MBR_AGE	-0.0231
1	ov	0.01994	2	ov	-0.1096
1	oon	0.258	2	oon	-0.0885
1	inpatient	0.01634	2	inpatient	0.2384
1	log_spe_rx	0.09201	2	log_spe_rx	-0.1281
1	Scale Parameter	1.3467	2	Scale Parameter	0.05021

4.3 Model Comparison and Selection

4.3.1 Model Comparison and Selection framework

Various types of linear and nonlinear candidate models are fitted in the previous section using real industry data from a major healthcare insurer. To select the best model, correct specification of the models and goodness of fit would be the first concern. Meanwhile, since the ultimate goal

of predictive modelling is to generate the most accurate prediction. The question turns into how to measure the accuracy of the predictions. In general, I can start by checking the loss function (absolute error loss or squared error loss), and maximize the distribution similarities between the prediction and actual values.

The typical loss functions include absolute error loss and squared error loss. Mean squared prediction error and mean absolute prediction error are two widely used measures corresponding to the loss functions. Meanwhile I proposed a methodology to measure individual level minimal distance counting, which measures the error loss from another perspective. What's more, when we are trying to measure the distribution similarity between the predictions and actual values, we need to guarantee that the model can generate higher predictions for those with higher actual values, and low predictions for those with lower actual values. Otherwise, even though the two distributions overlap to a great extent, they may not be the desirable results that we want to see. Gains chart can be a measure of the corresponding predictive effectiveness. The model selection framework is summarized in fig. 4.13. In fact, I am not only aiming at choosing the best model for our healthcare risk assessment, but also discussing whether the evaluation techniques themselves are effective under certain circumstances. Finally, the suggestions on evaluation procedures will be given for future analysis and decision making in section 4.3.6. All the details will be shown and discussed from section 4.3.2 to section 4.3.6.

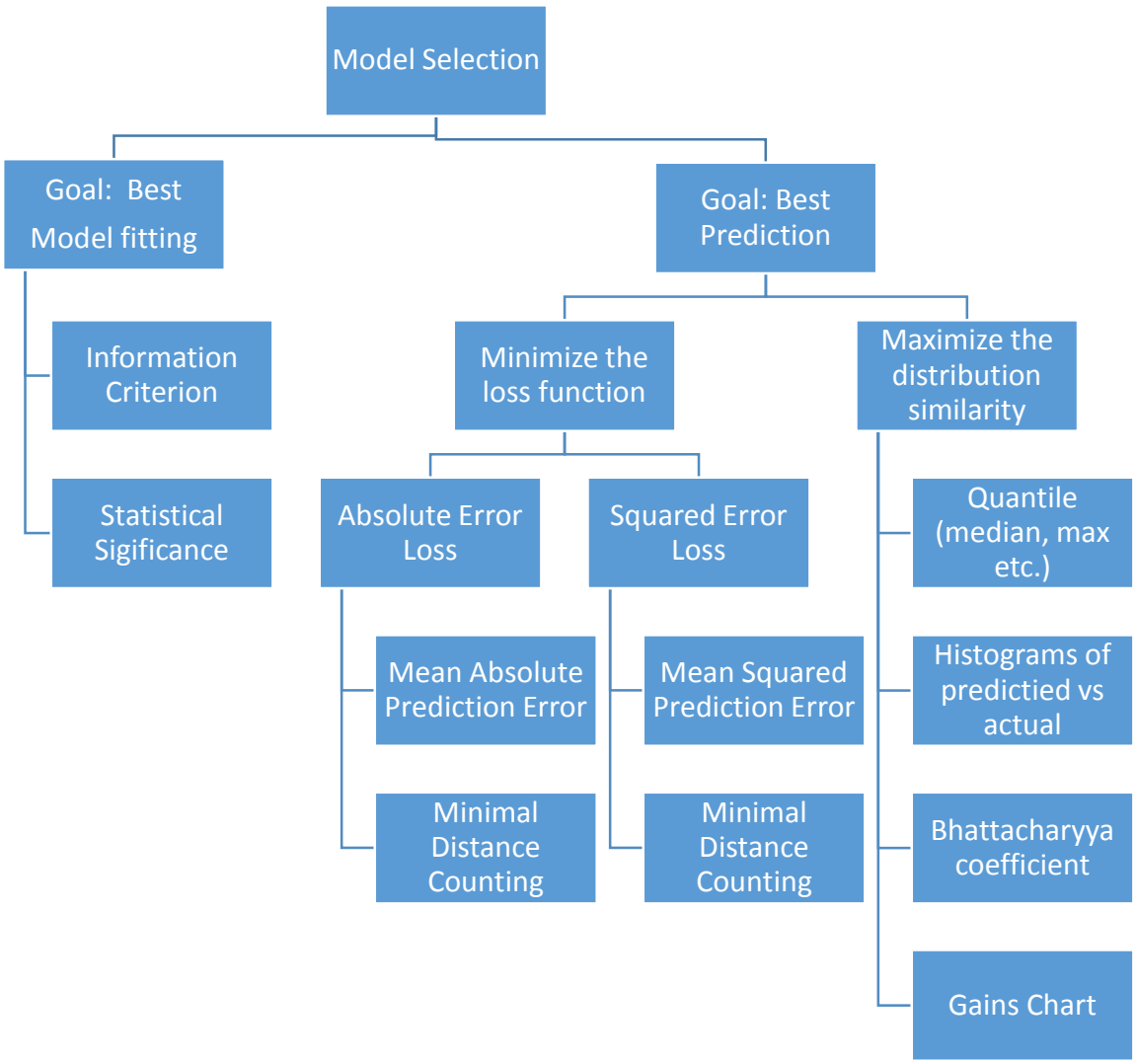


Fig. 4.13 Model Selection Framework

Comments: Minimal Distance Counting Comparison between models will be the same under either the MAPE or MSPE measure.

4.3.2 Fit Statistics

First, let's compare the log likelihood and information criteria AIC, AICc and BIC (the smaller, the better) in table 4.30. For GEE and quantile regression forest, there is no likelihood, hence not showing here. But we should be aware that in genera the ANOVA will do a log-likelihood ratio test to see if the addition of the random doctor effect is significant. Except that we cannot compare

the AIC/BIC between GLM and GLMM. That's because AICs are only comparable and appropriate for nested models, and on the same response variable without different transformations. However, in our experiments, there are different GLM models with different fixed design matrixes, various distribution assumptions and link function options. It is not appropriate to compare the AIC for GLMM with other AICs. Meanwhile, it is not appropriate to compare the AIC/AICc/BIC of lognormal GLM with other GLMs. Because the lognormal GLM is built up through log transformation of the response variable, the target variable is not exact the same as others on the original scale. To sum up, there are a lot constrains when we are trying to compare AIC/BIC among different type of models; they are more appropriate for nested models and exactly same target variable (without variable transformation).

Table 4.30 Fit statistics comparison

Models		-2 Log Likelihood	AIC	AICc	BIC
Two Part Model	Normal, Log link, GLM	123,331	123,379	123,379	123,536
	Gamma, log link, GLM	95,520	95,740	95,745	96,462
	Lognormal GLMM	18,067	18,071	18,071	18,082
	Lognormal GLM	17,932	17,958	17,958	18,089
	Lognormal GEE	NA	NA	NA	NA
Tweedie Model	Tweedie GLM	99,729	100,019	100,026	100,991
	Tweedie GAM	101,415	101,475	101,475	101,675
QR	QRF	NA	NA	NA	NA
FMM	FMM	100,080	100,152	100,152	100,393

4.3.3 Mean, quantiles, MAPE, and MSPE on holdout samples

Now I will compare the mean absolute prediction error (MAPE), mean squared prediction error (MSPE), mean, median and max of the predictions generated by different models. The last model "*mean*" means that without any predictive modeling, we simply use average value of the training data as the prediction for each observation in the test set. For right skewed and heavy tailed data, the mean might be pulled in the direction of the skewness. That is to say, for right skewed data,

the mean will be greater than the median. Meanwhile, since almost all the model candidates are heavy-tailed models tailored for our response variable, there might be extreme predictions given by each model. Extreme values in the tails will significantly distort the mean. But these extreme values won't distort the median because the median is based on ranks. In general, the median can provide a better estimate of location than does the mean for data with extreme values in the tails, but mean is still important. Two common alternative locations measures for mean can be checked as follows:

- Trimmed Mean[3] - similar to the mid-mean except different percentile values are used. A common choice is to trim 5% of the points in both the lower and upper tails, i.e., calculate the mean for data between the 5th and 95th percentiles.
- Winsorized Mean[4] - similar to the trimmed mean. However, instead of trimming the points, they are set to the lowest (or highest) value. For example, all data below the 5th percentile are set equal to the value of the 5th percentile and all data greater than the 95th percentile are set equal to the 95th percentile.

The comparison of all candidate models on mean, median, max, trimmed mean and Winsorized mean are summarized in table 4.31. Meanwhile, MAPE and MSPE are two very important distance measure between the actual value and predictions. They are shown in table 4.32.

Table 4.31 Comparison among Mean, trimmed mean, Winsorized mean, quantiles on holdout samples

Models		Mean	Trimmed Mean	Winsorized Mean	Median	MAX
Actual		5,222.43	2,698.81	3,374.04	1,144.06	278,840.55
Two Part Model	Lognormal GLM	2,927.42	1,461.81	1,725.70	852.73	1,404,973.35
	Normal, Log link, GLM	7,110.11	2,748.35	3,073.40	1,659.64	9,400,778.92

[3]Statistical terms http://www.investopedia.com/terms/t/trimmed_mean.asp
[4] Statistical terms http://www.investopedia.com/terms/w/winsorized_mean.asp

	Gamma, log link, GLM	15,815.10	3,319.97	3,672.89	2,733.55	24,372,890.90
	Lognormal GLMM	3,267.08	1,472.94	1,724.11	871.08	2,428,070.02
	Lognormal GEE	3,289.69	1,472.81	1,723.13	872.17	2,489,127.19
Tweedie Model	Tweedie GLM	11,063.62	3,301.89	3,703.13	2,509.74	16,268,856.80
	Tweedie GAM	4,677.10	3,093.07	3,591.94	1,938.57	616,908.80
QR	QRF	2,077.55	1,434.50	1,687.43	767.59	64,043.54
FMM	FMM	8,129.03	2,334.87	2,771.08	1,449.52	8,919,971.25
mean		5,059.08	5,059.08	5,059.08	5,059.08	5,059.08

Table 4.32 Comparison between MAPE and MSPE

Models		MSPE	MAPE
Two Part Model	Lognormal GLM	750,960,558	4,618.63
	Normal, Log link, GLM	28,614,893,651	8,123.88
	Gamma, log link, GLM	206,215,772,286	16,025.40
	Lognormal GLMM	1,937,205,546	4,961.46
	Lognormal GEE	2,030,107,770	4,985.71
Tweedie Model	Tweedie GLM	87,381,751,375	11,239.60
	Tweedie GAM	310,460,104	4,925.00
QR	QRF	254,402,427	4,106.95
FMM	FMM	27,673,547,537	8,778.59
mean		277,479,437	6,328.53

From the results in table 4.31 and 4.32, surprisingly, the most naive mean model outperforms many other models in terms of mean and MSPE, even though common sense tells us that is not the truth. That is because the extreme values in the tails have distorted the mean. Trimmed mean and Winsorized mean are better measures than mean for right skewed and heavy tailed data. For the calculation of MAPE and MSPE, extreme values in the tails also significantly distort the results, but it is not appropriate to apply the similar trimmed or Winsorized treatment to them because we are not able to tell where the extreme large prediction bias comes from (most time it comes from inaccurate predictions). However, we can deal with it by capping the predictions. In fact, in practice, when the data is scored to give predicting results to each policy holder for risk assessment purpose, we need to deal with those unreasonable extreme predictions carefully. Experienced analysts can tell some extremely large predictions are almost impossible in practice

based on our experience and how medical and pharmacy costs are composed. If we cap all the extreme large predictions by 1.5 times of the maximal value of the training dataset (that is, replace all the predictions greater than 418260.83 by 418260.83 in our study), we can get the updated results in table 4.33. This comparison is more fair and reasonable for those models that might did a good job on 99% predictions, but gave extremely biased results on a few predictions which could hinder and distort the overall results. One may argue that why we are using 1.5 times of the maximal value of the training dataset, not 2 times, or 3 times, or others. It is indeed a very interesting question and we can build a model for this number if longitudinal data over years are available. Either longitudinal data analysis or time series models can help us improve this. For now, we don't have multiple years' data available, we believe 1.5 can be an appropriate starting point and it can be easily modified if more information available.

Table 4.33 Comparison among Mean, MAPE, MSPE, quantiles on capped holdout samples

Models		Mean	Median	Max	MAPE	MSPE
Actual		5,222.43	1,144.06	278,840.55	0	0
Two Part Model	Lognormal GLM	2,598.52	852.73	418,260.83	4,289.72	275,013,586.00
	Normal, Log link, GLM	4,110.11	1,659.64	418,260.83	5,123.89	337,716,369.00
	Gamma, log link, GLM	5,697.88	2,733.55	418,260.83	5,908.26	523,557,417.00
	Lognormal GLMM	2,597.14	871.08	418,260.83	4,291.52	282,351,776.00
	Lognormal GEE	2,599.40	872.17	418,260.83	4,295.42	282,833,143.00
Tweedie Model	Tweedie GLM	5,273.53	2,509.74	418,260.83	5,449.50	407,938,255.00
	Tweedie GAM	4,610.89	1,938.57	418,260.83	4,858.78	266,823,485.00
QR	QRF	2,077.55	767.59	64,043.54	4,106.95	254,402,427.00
FMM	FMM	4,441.90	1,449.52	418,260.83	5,091.45	392,284,234.00
mean		5,059.08	5,059.08	5,059.08	6,328.53	277,479,437.00

4.3.4 Distribution similarity between actual and predicted

4.3.4.1 Histograms

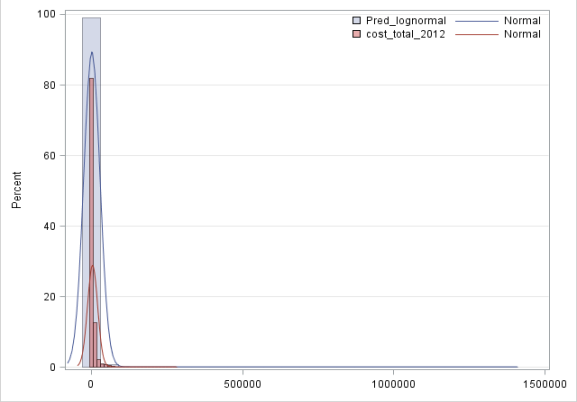
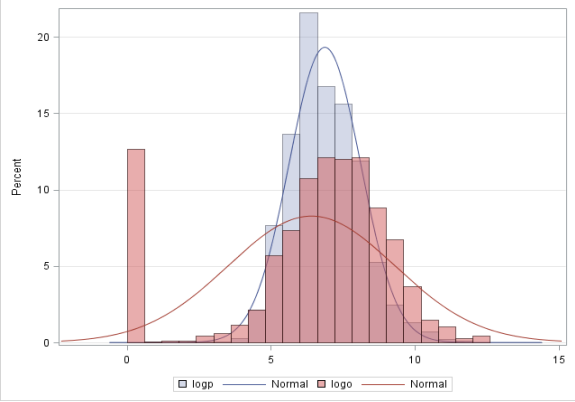
To check the distribution similarity between the predicted and actual values, the most straight forward method is to compare the histograms. When it is difficult to distinguish the two visually

when the data is heavy tailed, histograms on the log scale are compared in table 4.31. Usually, the histogram graphically shows the following:

- Center of the data;
- Spread/ Peakedness of the data;
- Skewness of the data;
- Presence of outliers and multiple modes in the data.

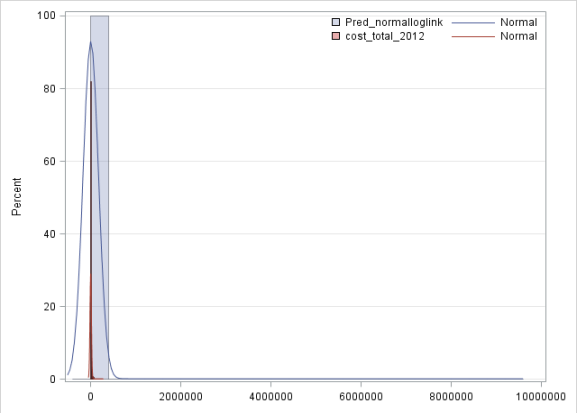
These features provide strong indications of the proper distributional model for the data. In all those figures in table 4.34, the blue histograms denote the predictions, while the red ones denote the actual values. If there is a big area that overlaps between the red and the blue, that means the model performs well. For example, for the two part lognormal GLM, the actual data has a lower peak and heavier tails than the predictions. For quantile regression forest, the predictions are a little bit left skewed compared to the actual on the log scale. Similarly, we can check the performance of other models from the histograms. Overall, the models perform well, even though each model has its flaws, but they are useful.

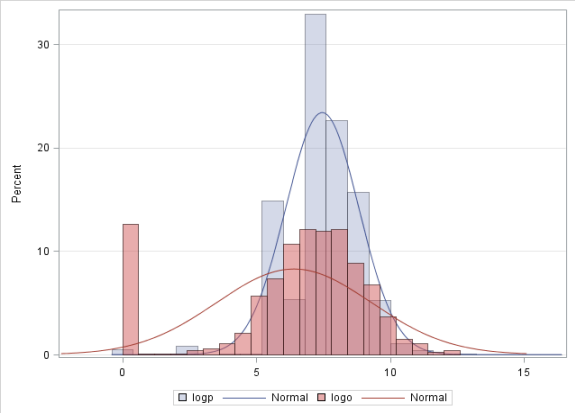
Table 4.34 Histogram Comparison between prediction and actual

Models	Predictions vs Actual	log(Predictions) vs log(Actual)
Two Part Lognormal GLM		

Comments: The actual data has a lower peak and heavier tails than the predictions.

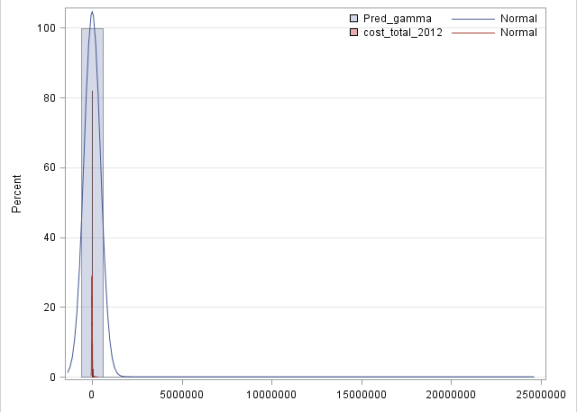
Two Part Normal, Log link, GLM

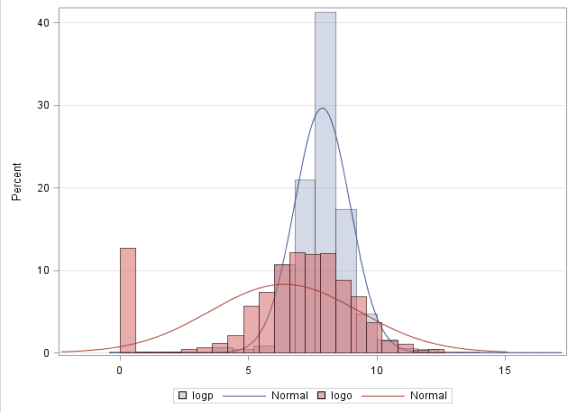




Comments: Again, the actual data has a lower peak and heavier tails than the predictions.

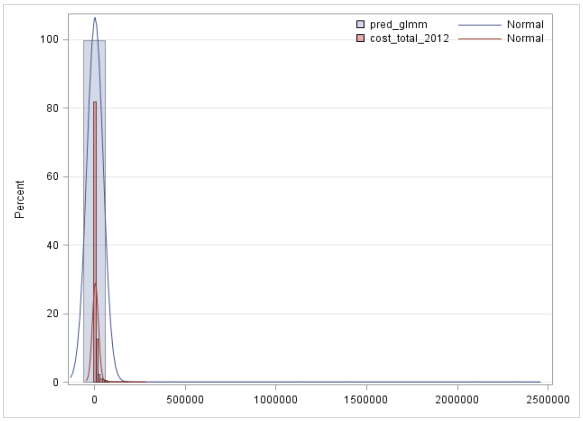
Two Part Gamma, log link, GLM

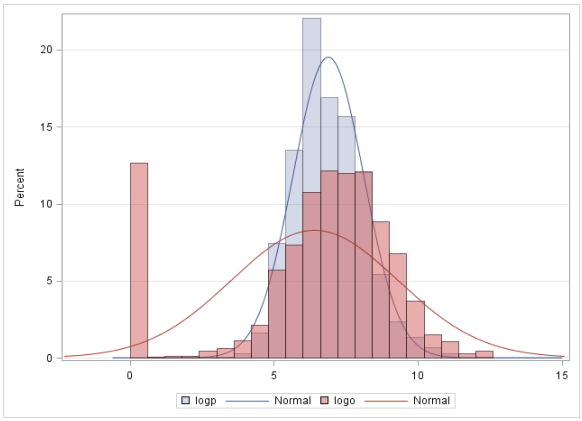




Comments: The actual data has a lower peak and heavier tails than the predictions. In addition, the predictions are more right skewed than the actual values.

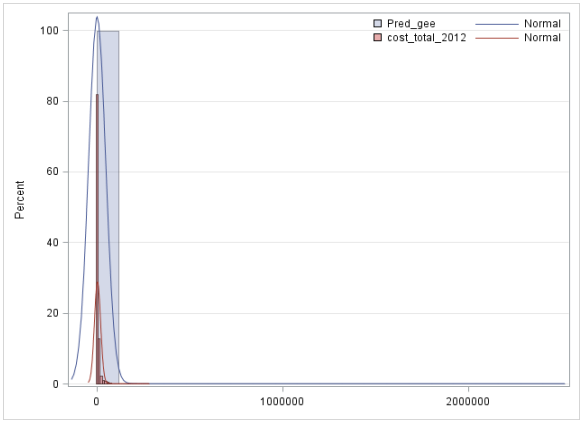
Two Part Lognormal GLMM

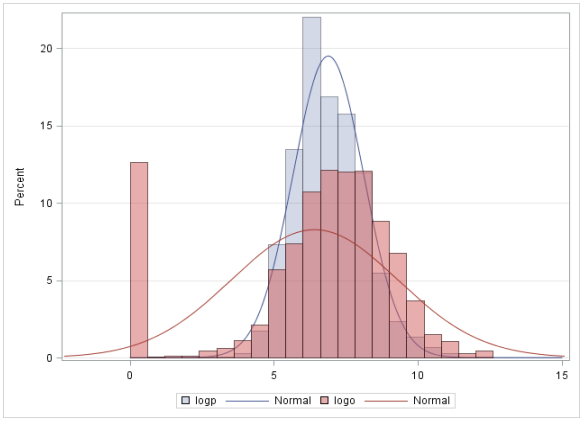




Comments: The predictions appear to be too small on average.

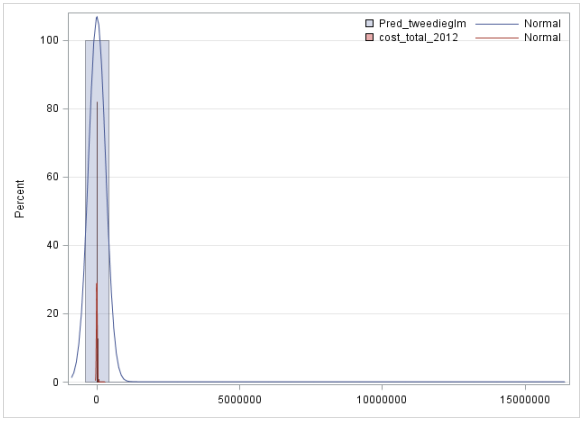
Two Part Lognormal GEE

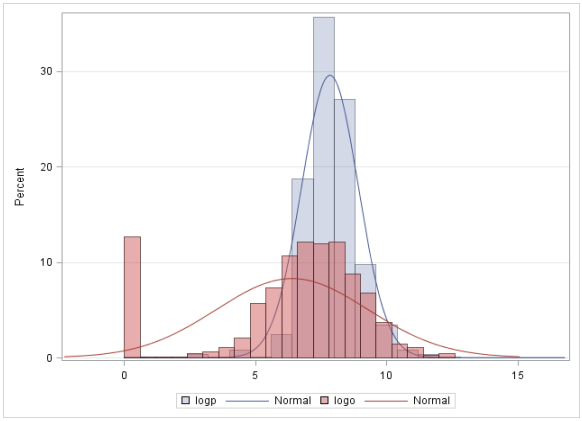




Comments: Again, the predictions appear to be too small on average.

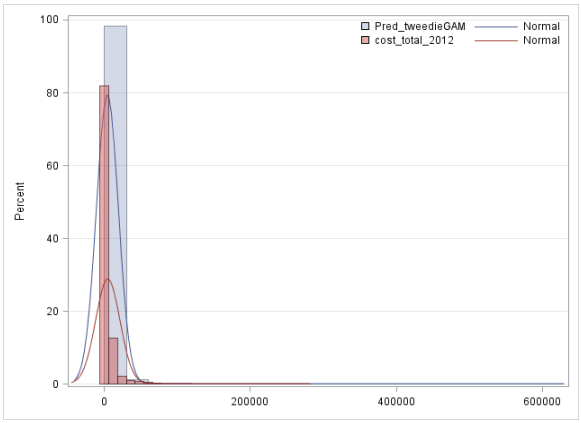
Tweedie GLM

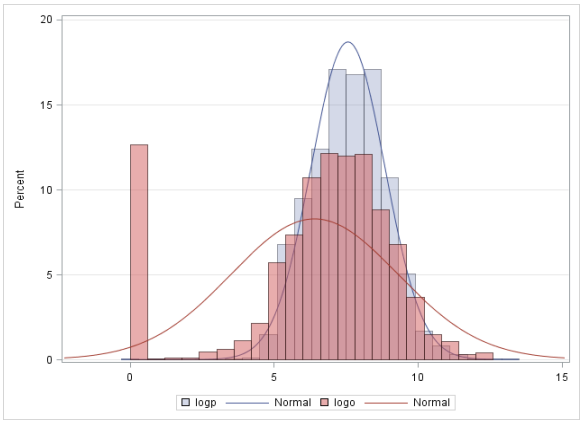




Comments: The predictions appear to be slightly too large on average.

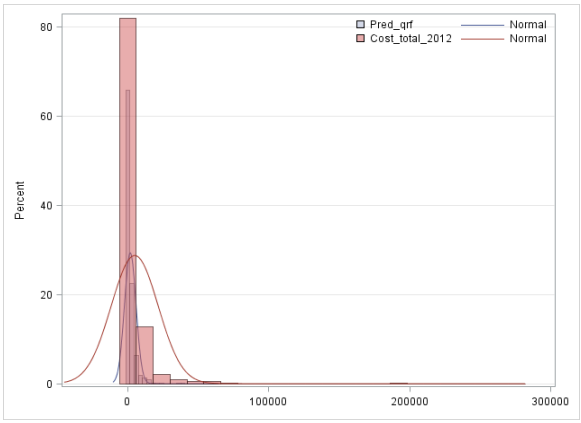
Tweedie GAM

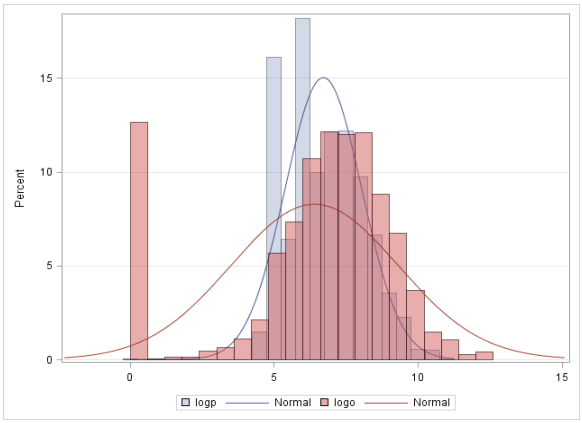




Comments: The actual data has heavier tails than the predictions.

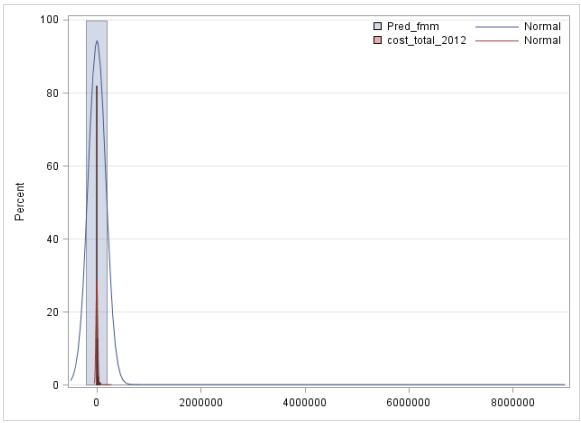
Quantile Regression Forest

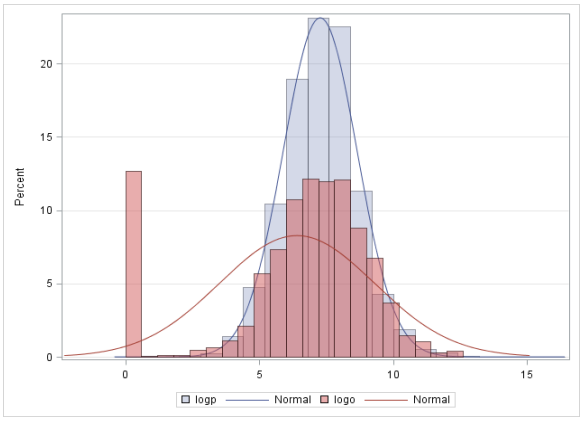




Comments: The predictions are a little bit left skewed compared to the actual.

Finite Mixture Model





Comments: The actual data has a shorter peak and heavier tails than the predictions.

4.3.4.2 Bhattacharyya coefficient

Rather than letting the graphs tell the difference in distribution, I am also trying to find a measure to summarize the distribution similarity into a number. The Bhattacharyya coefficient is such a measure that can give us an objective and fact-based comparison. The Bhattacharyya coefficient, introduced by Bhattacharyya, A. (1943), is an approximate measure of overlap between two statistical samples. Calculating the Bhattacharyya coefficient is a rudimentary form of integration (or summation) of the overlap of the two samples. The interval of the values of the two samples is split into a chosen number of partitions, and the number of members of each sample in each partition is used to calculate Bhattacharyya coefficient,

$$Bhattacharyya = \sum_{i=1}^{n} \sqrt{(\sum a_i \sum b_i)} \tag{4.9}$$

Where a and b are two samples, n is the number of partitions, and $\sum a_i, \sum b_i$ are the total number of samples a and b in the i^{th} partition. In this study, I divide the interval into 5000 subintervals. That is $n = 5000$. Therefore, the value of Bhattacharyya coefficient will be larger if there is a larger area of overlap between the two samples; while it will be zero if there is no overlap at all (due to the multiplication by zero in every partition). The maximal value will be the number of observations in the hold out sample. We can calculate the distribution match rate as:

$$Distribution\ match\ rate = \frac{Bhattacharyya\ coefficient}{\#\ of\ obervations\ in\ the\ sample} \tag{4.10}$$

It ranges from 0 to1.

Table 4.35 Comparison on distribution similarity between predicted and actual

Models		Bhattacharyya coefficient	Distribution Match Rate
Actual		3000.00	100.00%
Two Part Model	Lognormal GLM	2566.19	85.54%
	Normal, Log link, GLM	2432.14	81.07%
	Gamma, log link, GLM	2404.35	80.14%
	Lognormal GLMM	2580.16	86.01%
	Lognormal GEE	2567.15	85.57%

Tweedie Model	Tweedie GLM	2371.82	79.06%
	Tweedie GAM	2549.68	84.99%
Quantile Regression	Quantile regression forest	2491.06	83.04%

From the results shown in table 4.35, the lognormal GLMM produces the highest Bhattacharyya coefficient compared to other candidate, indicating 86.01% overlap between two statistical samples (actual and predicted).

4.3.4.3 Gains Chart for Continuous Data

In data mining, lift chart or gains chart[5] (Berry and Linoff, 1999) is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without predictive models. They are visual aids for measuring model performance. However, in most literatures, lift charts require the predictable attribute to be a discrete value. In other words, we cannot use lift charts to measure the accuracy of models that predict continuous numeric values[6]. Actually, after in-depth investigation, it is not impossible to draw the gains chart for the continuous numeric values; however, the way to interpret the results will be quite different than those with discrete values. In short, if the target variable is continuous, gains chart provides us with statistics relative to the mean of the target variable. Because if we don't build any predictive models, we will tend to use the long term average as the prediction; then gains chart for continuous value is a measure of the effectiveness of a predictive models compared to the mean of the target value. What's more, we can compare the gains chart of the prediction with actual to measure the effectiveness of the predictive models. Let's take the FMM as example, the gains chart is shown

[5]Cumulative Gains and Lift Charts http://www2.cs.uregina.ca/~dbd/cs831/notes/lift_chart/lift_chart.html
[6] https://msdn.microsoft.com/en-us/library/ms175428.aspx

in table 4.36. We can calculate this gains table and lift chart from the following steps. The gains package in R is used for calculation and plotting.

Step 1: Rank the observations by predicted outcome values

Step 2: Create subgroups (often deciles) with similar predicted values. Here we have 3000 observations in all and each subgroup has 300 observations.

Step 3: Calculate and display the average and cumulative actual outcomes by group

Step 4: Calculate and display the average predicted outcomes by group

Step 5: Measure the Cumulative lift index for each group compared to the baseline. In our case study, the baseline will be the mean of all the actual value, that is 5222.43. For example, in table 4.33, $284 = 100 * \frac{14817.49}{5222.43}$. And in this gains table, no matter which model we are looking at, the cumulative lift on the last row will 100.

If we plot the gains table for FMM in table 4.36 in the first chart in fig. 4.14, there are three lines.

- First, we need to check the brown curves. The brown curves plot the cumulative mean response (Column 6 in table 4.36). Only when this curve is monotone decreasing, we have reason to believe this model will be more effective than no predictive model. If without any predictive models, this curve will be horizontal with value 5222.43. However, different from cumulative gains for binary or count data, higher lift doesn't mean this model is more effective. Because we want the prediction to be close to the actual, not over-predicting. If it is not monotone decreasing, we need to double check whether the models are correctly specified.
- Second, we need to check whether the blue curve and the red curve are close enough. The blue curve plots the mean predicted response and the red curve plots the mean actual response. We rank the observations by predicted outcomes; so without doubt, the blue

curve will be monotone decreasing. We need to make sure the red curve is also monotone decreasing, and the closer the two curves, the better the model will be.

Table 4.36: Gains table for FMM

Depth of file	N	Cumulative N	Mean Predicted	Mean Actual	Cumulative Actual	Cumulative Lift
10	300	300	65,020.83	20,569.51	20,569.51	394
20	300	600	5,359.45	9,065.46	14,817.49	284
30	300	900	3,504.68	6,757.29	12,130.75	232
40	300	1200	2,476.28	4,467.10	10,214.84	196
50	300	1500	1,731.28	3,332.47	8,838.37	169
60	300	1800	1,231.65	2,743.81	7,822.61	150
70	300	2100	868.23	1,745.78	6,954.49	133
80	300	2400	588.19	1,369.03	6,256.30	120
90	300	2700	361.57	1,291.54	5,704.66	109
100	300	3000	148.18	882.37	5,222.43	100

Table 4.37: Gains table for QRF

Depth of file	N	Cumulative N	Mean Predicted	Mean Actual	Cumulative Actual	Cumulative Lift
10	300	300	10960.68	18622.39	18622.39	357
20	300	600	3619.74	8131.59	13376.99	256
30	300	900	2172.75	5981.42	10911.8	209
40	300	1200	1449.96	5738.5	9618.48	184
50	300	1500	963.92	4139.57	8522.69	163
60	300	1800	594.89	2817.62	7571.85	145
70	300	2100	409.3	1740.08	6738.74	129
80	300	2400	314.48	2214.9	6173.26	118
90	300	2700	167.97	1357.11	5638.13	108
100	300	3000	121.85	1481.16	5222.43	100

Table 4.38: Gains table for lognormal GLM

Depth of file	N	Cumulative N	Mean Predicted	Mean Actual	Cumulative Actual	Cumulative Lift
10	300	300	19059.91	18590.08	18590.08	356
20	300	600	3441.13	9333.06	13961.57	267
30	300	900	2157.42	7765.98	11896.37	228
40	300	1200	1481.64	5164.21	10213.33	196
50	300	1500	1014.27	4134.64	8997.59	172
60	300	1800	727.6	2161.92	7858.31	150
70	300	2100	545.92	1396.16	6935.15	133

80	300	2400	413.84	1649.86	6274.49	120
90	300	2700	280.69	1216.89	5712.53	109
100	300	3000	151.77	811.55	5222.43	100

Table 4.39: Gains table for Tweedie GAM

Depth of file	N	Cumulative N	Mean Predicted	Mean Actual	Cumulative Actual	Cumulative Lift
10	300	300	24602.76	18730.06	18730.06	359
20	300	600	7277.07	9451.12	14090.59	270
30	300	900	4797.46	6413.89	11531.69	221
40	300	1200	3400.43	4933.55	9882.16	189
50	300	1500	2334.63	4817.27	8869.18	170
60	300	1800	1667.34	2642.09	7831.33	150
70	300	2100	1190.2	2129.94	7016.85	134
80	300	2400	810.36	1090.29	6276.03	120
90	300	2700	457.54	1039.27	5694.16	109
100	300	3000	233.23	976.87	5222.43	100

Table 4.40: Gains table for Normal Log link

Depth of file	N	Cumulative N	Mean Predicted	Mean Actual	Cumulative Actual	Cumulative Lift
10	300	300	51106.38	15270.11	15270.11	292
20	300	600	6125.25	9316	12293.05	235
30	300	900	4212.15	6240.04	10275.38	197
40	300	1200	3237.39	4554.88	8845.26	169
50	300	1500	2081.96	5674.34	8211.07	157
60	300	1800	1486.05	2778.36	7305.62	140
70	300	2100	1231.47	3093.96	6703.96	128
80	300	2400	990.68	2431.37	6169.88	118
90	300	2700	404.87	1725.26	5658.55	108
100	300	3000	217.34	1115.64	5222.43	100

Table 4.41: Gains table for Tweedie GLM

Depth of file	N	Cumulative N	Mean Predicted	Mean Actual	Cumulative Actual	Cumulative Lift
10	300	300	86081.65	20436.13	20436.13	391
20	300	600	6650.69	9473.97	14955.05	286
30	300	900	4546.6	6006.48	11972.19	229
40	300	1200	3470.73	4760.48	10169.27	195
50	300	1500	2779.36	3278.22	8791.06	168
60	300	1800	2270.05	1873.16	7638.07	146
70	300	2100	1875.87	2345.69	6882.02	132

80	300	2400	1445.43	2105.35	6284.94	120
90	300	2700	961.52	957.17	5692.96	109
100	300	3000	554.32	987.69	5222.43	100

Table 4.42: Gains table for GEE

Depth of file	N	Cumulative N	Mean Predicted	Mean Actual	Cumulative Actual	Cumulative Lift
10	300	300	22573.05	18580.42	18580.42	356
20	300	600	3441.95	9432.24	14006.33	268
30	300	900	2181.66	7992.92	12001.86	230
40	300	1200	1499.88	4734.81	10185.1	195
50	300	1500	1030.81	4187.68	8985.61	172
60	300	1800	740.22	2336.32	7877.4	151
70	300	2100	557.01	1292.09	6936.64	133
80	300	2400	424.75	1643.07	6274.94	120
90	300	2700	290.13	1222.08	5713.51	109
100	300	3000	157.47	802.72	5222.43	100

Table 4.43: Gains table for Gamma

Depth of file	N	Cumulative N	Mean Predicted	Mean Actual	Cumulative Actual	Cumulative Lift
10	300	300	133166.5	20147.09	20147.09	386
20	300	600	6412.12	9602.9	14874.99	285
30	300	900	4499.1	6633.85	12127.95	232
40	300	1200	3521.52	3719.99	10025.96	192
50	300	1500	2963.2	2805.14	8581.79	164
60	300	1800	2517.08	3355.2	7710.69	148
70	300	2100	2044.24	2006.37	6895.79	132
80	300	2400	1493.42	1923.63	6274.27	120
90	300	2700	1006.76	1171.32	5707.28	109
100	300	3000	527.07	858.86	5222.43	100

Table 4.44: Gains table for GLMM

Depth of file	N	Cumulative N	Mean Predicted	Mean Actual	Cumulative Actual	Cumulative Lift
10	300	300	22345.56	18632.34	18632.34	357
20	300	600	3443.83	9392.43	14012.39	268
30	300	900	2181.91	7999.69	12008.15	230
40	300	1200	1500.75	4747.56	10193	195
50	300	1500	1029.72	4187.81	8991.97	172

60	300	1800	739.15	2310.18	7878.33	151
70	300	2100	557.2	1287.63	6936.81	133
80	300	2400	424.32	1636.37	6274.25	120
90	300	2700	290.15	1225.8	5713.31	109
100	300	3000	158.2	804.54	5222.43	100

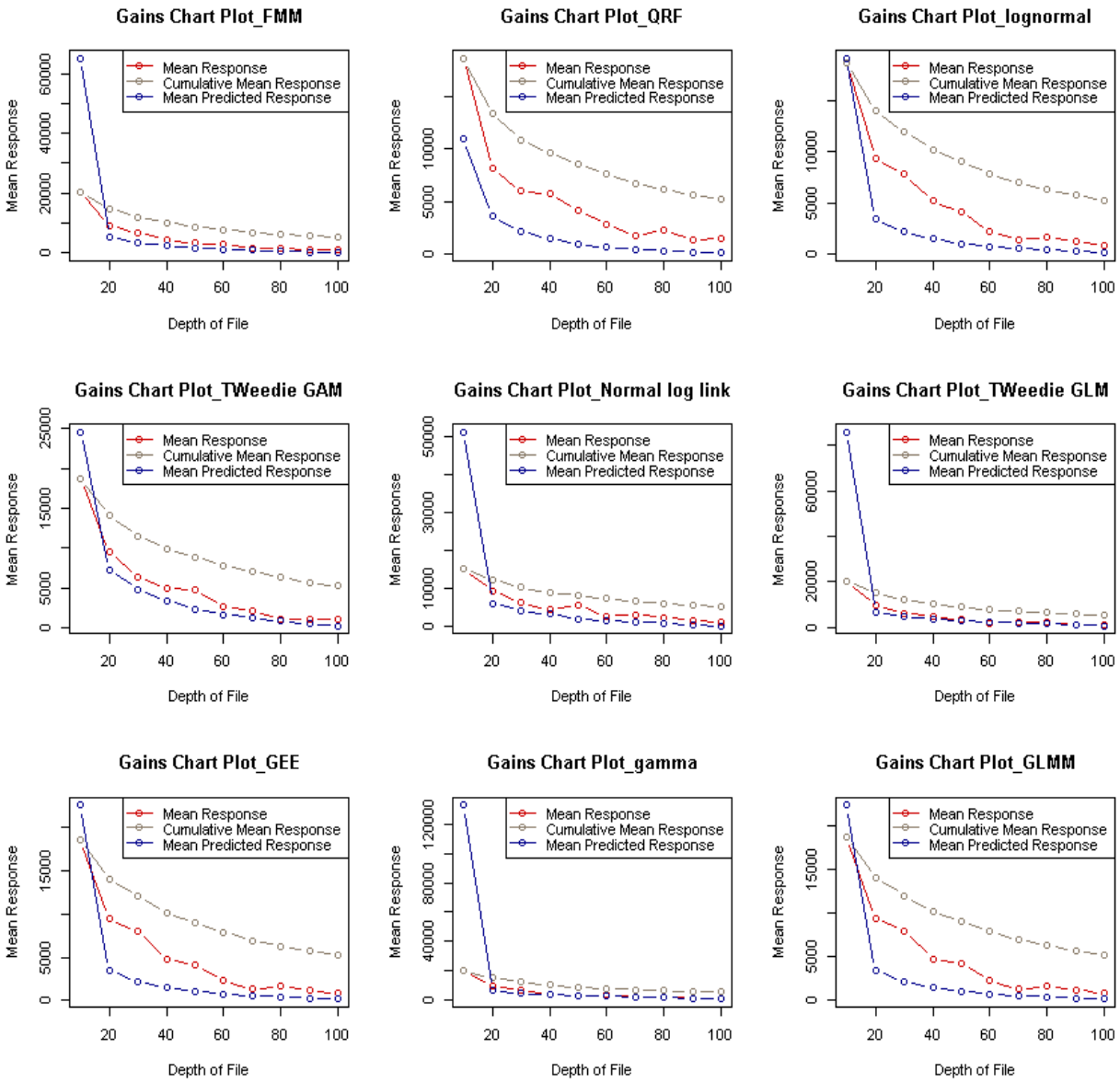


Fig 4.14 Gains Chart for Nine models

4.3.5 Minimal Distance Probability Matrix

4.3.5.1 Definitions

Rather than using an aggregate measure to compare the models, I am trying to come up with an individual level mechanism to compare the predictions for each observation. In that way, we can avoid a few extreme biased predictions distorting the overall results. I call this method as one on one minimal distance probability matrix. Let's define this matrix as follows:

$$M_{N,N} = \begin{pmatrix} p_{1,1} & \cdots & p_{1,N} \\ \vdots & p_{i,j} & \vdots \\ p_{N,1} & \cdots & p_{N,N} \end{pmatrix} \tag{4.11}$$

$$i = 1,2,\dots,N, \quad j = 1,2,\dots,N, \text{ Here } N = 10$$

Where N is the total number of candidate models in our comparison. i *and* j are indexes of different models, $p_{i,j}$ means the wining probability of model i compared to model j in terms of minimal absolute distance. Therefore $p_{i,j} + p_{j,i} = 1$

For example, if we want to compare the predictions given by gamma and GEE models, here are the steps to calculate this matrix:

- Step 1: Given predicted values vector V_{gamma} and V_{GEE}, and actual value vector V . I start with comparing $\left|V_{gamma} - \mathrm{V}\right|$ and $|V_{GEE} - \mathrm{V}|$
    - Initial values: N_{gamma}=0, N_{GEE}=0
    - For i in 1:N
    - if $\left|V_{gamma} - \mathrm{V}\right|_i > |V_{GEE} - \mathrm{V}|_i$ then N_{gamma}=N_{gamma}+1
    - else if $\left|V_{gamma} - \mathrm{V}\right|_i = |V_{GEE} - \mathrm{V}|_i$ then N_{gamma}=N_{GEE}+0.5, N_{GEE}=N_{GEE}+0.5,
    - else N_{GEE}=N_{GEE}+1,
- Step 2: Calculate $p_{gamma,GEE} = \frac{N_{gamma}}{N_{gamma}+N_{GEE}}$, $p_{GEE,gamma} = \frac{N_{GEE}}{N_{gamma}+N_{GEE}}$

Similarly, we can calculate the probability of other elements in this matrix and the result is given in table 4.45.

Table 4.45 One on one minimal distance probability matrix

	Gamma	GEE	lognormal	Normal Log link	GLMM	QRF	Tweedie GAM	Tweedie GLM	FMM	Mean
Gamma	0.500	0.355	0.356	0.445	0.355	0.357	0.412	0.495	0.369	0.752
GEE	0.645	0.500	0.496	0.612	0.462	0.491	0.620	0.643	0.589	0.756
lognormal	0.644	0.504	0.500	0.612	0.491	0.497	0.621	0.643	0.590	0.754
Normal	0.555	0.388	0.388	0.500	0.388	0.390	0.482	0.542	0.446	0.725
GLMM	0.645	0.538	0.509	0.612	0.500	0.493	0.621	0.644	0.590	0.756
QRF	0.643	0.509	0.503	0.610	0.507	0.500	0.618	0.644	0.588	0.758
Tweedie GAM	0.588	0.380	0.379	0.518	0.379	0.382	0.500	0.576	0.440	0.727
Tweedie GLM	0.505	0.357	0.357	0.458	0.356	0.356	0.424	0.500	0.369	0.755
FMM	0.631	0.411	0.410	0.554	0.410	0.412	0.560	0.631	0.500	0.754
Mean	0.248	0.244	0.246	0.275	0.244	0.242	0.273	0.245	0.246	0.500

All the elements on the diagonal will be 0.5 because any model cannot beat itself. As we expected, the mean model will be worse than any other predictive models we built, because if look at the last row, every wining probability is lower than 0.5, in fact, lower than 0.3. If we look at other rows, of example, the first row, the gamma model. Except the last column, all the probabilities are less than 0.5. This should call our attention because in practice, gamma GLM with log link function is almost the default regression model for positive continuous target, but our result show that it performs worse than all other predictive models except no model. Another thing worth to mention is that QRF outperforms all other models using this measure, because if we look the sixth row, almost all the probabilities are great than or equal to 0.5, even though some values are pretty close to 0.5.

We also can derive the matrix of relations from the probability matrix. Let's define the relation matrix as follows:

$$R_{N,N} = \begin{pmatrix} r_{1,1} & \cdots & r_{1,N} \\ \vdots & r_{i,j} & \vdots \\ r_{N,1} & \cdots & r_{N,N} \end{pmatrix} \quad i = 1,2,\dots,N, \quad j = 1,2,\dots,N, \text{ Here } N = 10 \qquad (4.12)$$

$$\text{If } p_{i,j} \geq 0.5 \;\; then \;\; R_{ij} = 1, else \; R_{ij} = 0$$

Where $r_{i,j}$ is a relation between model i and model j that measures the prediction accuracy. If $R_{ij} = 1$ means model model i wins model j. If $R_{ij} = 0$, it means model i loses to model j. The matrix of relation derived is shown in table 4.46

Table 4.46 Matrix of relations

	Gamma	GEE	lognormal	Normal Log link	GLMM	QRF	Tweedie GAM	Tweedie GLM	FMM	Mean
Gamma	1	0	0	0	0	0	0	0	0	1
GEE	1	1	0	1	0	0	1	1	1	1
lognormal	1	1	1	1	0	0	1	1	1	1
Normal	1	0	0	1	0	0	0	1	0	1
GLMM	1	1	1	1	1	0	1	1	1	1
QRF	1	1	1	1	1	1	1	1	1	1
Tweedie GAM	1	0	0	1	0	0	1	1	0	1
Tweedie GLM	1	0	0	0	0	0	0	1	0	1
FMM	1	0	0	1	0	0	1	1	1	1
Mean	0	0	0	0	0	0	0	0	0	1

4.3.5.2 Properties of matrix of relations

- Reflexivity ($R_{ii} = 1$): Yes
- Symmetry($R_{i,j} = R_{j,i}$): No , but $R_{i,j} + R_{j,i} = 1$
- Transitivity(if *a* ~ *b* and *b* ~ *c* then *a* ~ *c*): Not always

  Why? Let's look at one example, suppose we have three distance vectors where each element in the vectors denotes the distance between the predicted and actual, hence the smaller the better.

  A=(3, 2, 1), B=(1,3,2), C=(2,1,3)

However, we have A wins B, B wins C, C wins A

Theoretically, the matrix of relation in this situation doesn't have to be transitive. However, the matrix in our case study is transitive. To prove this, we can try to find the non-zero entries in R^2. If R already has a 1 in each of those positions, *R* is transitive; if not, it's not transitive. Compare the matrix R in table 4.47 and the matrix R^2 in table 4.48, it can be easily figured out that our matrix of relation is transitive.

Table 4.47 Matrix R

1	0	0	0	0	0	0	0	0	1
1	1	0	1	0	0	1	1	1	1
1	1	1	1	0	0	1	1	1	1
1	0	0	1	0	0	0	1	0	1
1	1	1	1	1	0	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	0	0	1	0	0	1	1	0	1
1	0	0	0	0	0	0	1	0	1
1	0	0	1	0	0	1	1	1	1
0	0	0	0	0	0	0	0	0	1

Table 4.48 Matrix R^2

1	0	0	0	0	0	0	0	0	2
6	1	0	4	0	0	3	5	2	7
7	2	1	5	0	0	4	6	3	8
3	0	0	1	0	0	0	2	0	4
8	3	2	6	1	0	5	7	4	9
9	4	3	7	2	1	6	8	5	10
4	0	0	2	0	0	1	3	0	5
2	0	0	0	0	0	0	1	0	3
5	0	0	3	0	0	2	4	1	6
0	0	0	0	0	0	0	0	0	1

Conclusions:

- When a matrix of relation is transitive, there exists a single robust best model based on the absolute distance measure

- When a matrix of relation is not transitive, there is no single robust best model based on the absolute distance measure, but other alternative strategy can be used such as model averaging.
- When the relation of matrix is transitive, we can get the unique maximal value of total vote when counting the votes for each model (shown in table 4.49). Where the vote for each model is calculated as the sum of row of the relation matrix R.

Table 4.49 Total vote rank

	Vote	Rank
Gamma	2	9
GEE	7	4
lognormal	8	3
Normal	4	7
GLMM	9	2
QRF	10	1
Tweedie GAM	5	6
Tweedie GLM	3	8
FMM	6	5
Mean	1	10

4.3.5.3 Why one on one comparison?

In Game theory, Kenneth Arrow's “impossibility” theorem, or “general possibility” theorem (Arrow, K.J., 1950) answers a very basic question in the theory of collective decision-making. Say there are some alternatives to choose among. They could be policies, public projects, candidates in an election, or predictive models for risk assessment here. For example, if we are comparing only two models GLMM and lognormal GLM, the winning case for them will be 1528 and 1472, separately. However, if we are adding the GEE model, even though it only wins 692 cases and it cannot beat neither the GLMM nor lognormal GLM model. At this time Lognormal GLM wins GLMM because many cases once won by GLMM are switched to GEE, even though if we only compare GLM with GLMM, GLMM wins. If there are limit number of models, my method using

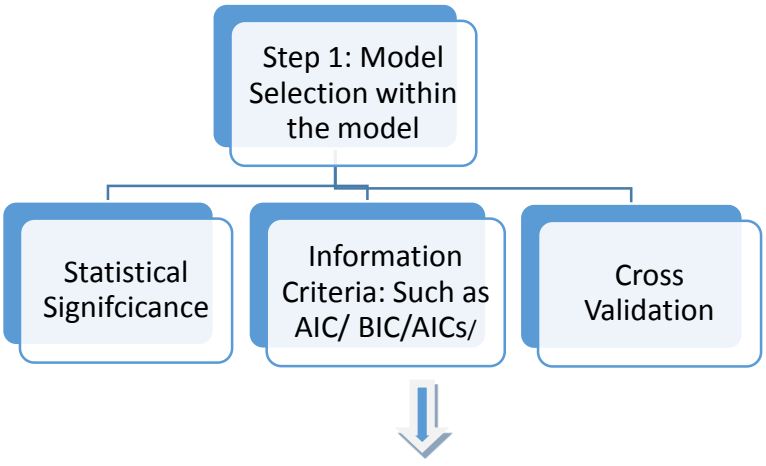
the probability matrix will be the best strategy to select the most robust models. But if practice, we cannot guarantee that all the reasonable models are taken into our consideration. Therefore the initial model selection will be critical. If we believe the first round model selection is appropriate and reasonable, our minimal distance probability matrix will be a powerful tool.

Table 4.50 Application of Arrow's theorem

Obs	GEE	GLMM	Lognormal GLM	Total
1	692	920	1388	3000
2	1385	1615		3000
3		1528	1472	3000
4	1488		1512	3000

4.3.6 Suggestion for Model Selection Strategies and Decision Making

For now, all the candidate models are fitted and summarized. It is the time to integrate the results and make decision. Let's briefly review the model selection process for risk assessment shown in fig.4.15. In general, there are two steps: model selection within the model and model selection among different types of models. In fact, once we are done the experiments and knowing the strength and weakness of each type of model, it is not always necessary to run so many candidate models in the future. The results and methodologies shown in the dissertation can be used as a good reference for those data analysts or actuaries in health insurance industry.



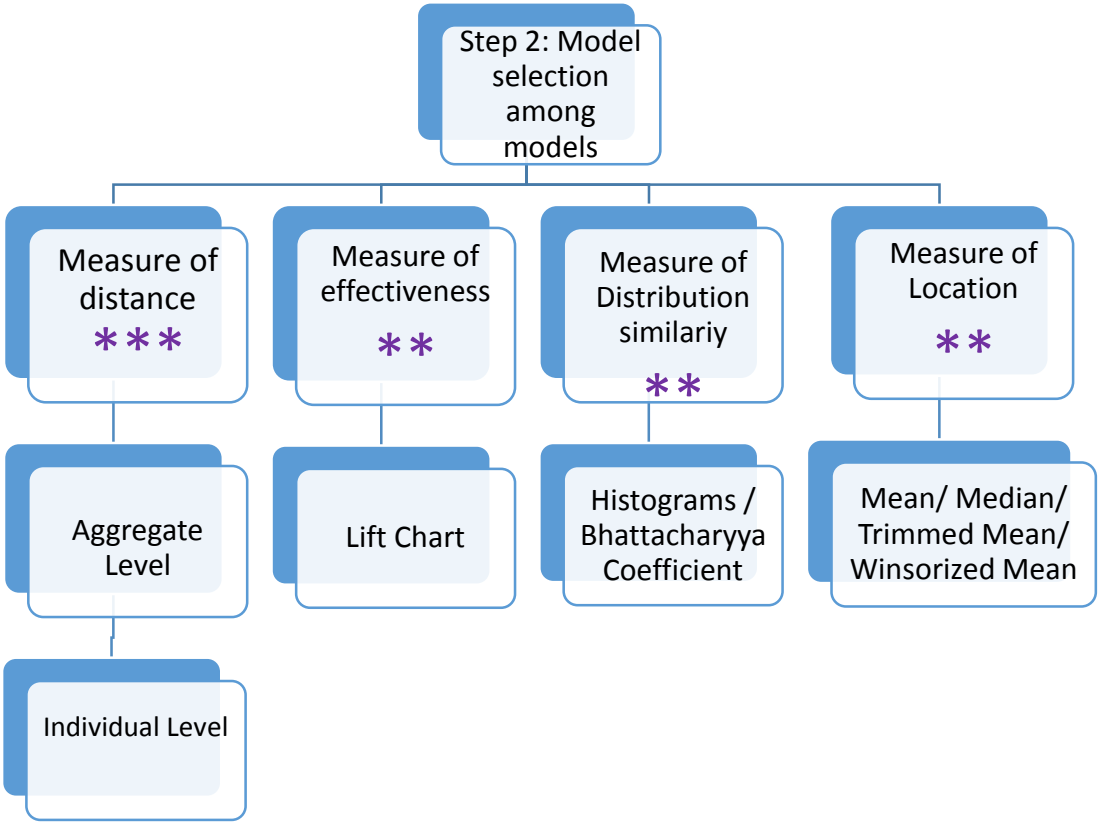


Fig. 4.15 Model selection steps and strategies

- Step 1: Model selection within each model

As I discussed in section 4.3.2, AIC/BIC/AICc are only comparable and appropriate for nested models. They are not appropriate to compare those information criteria for some models with different design matrices, such GLM and GLMM. But they are very critical criteria for model selection (variable selection) with in a model. Meanwhile, we need to check statistical significance and do cross validation in most cases to guarantee the best fit.

- Step 2: Model selection between models

In order to select the best predictive models, we need to compare them from different perspectives. If we can find a model which is the best no matter which measure we are looking at, that would be fantastic. However, in practice, it is not always the case. We may have a bunch

of evaluations techniques available and the conclusions by each measure are not always consistent. At this circumstance, using only one criterion may be too subjective and we should look at different measures based on a prudent attitude.

- Measure of distance (***): Since the goal is to get the most accurate prediction, the minimal distance between the predicted and actual will be the most direct measure. Different norms can be the measures, such as MSPE, MAPE. However, if we only look at one aggregate number, sometimes its value can be distorted by a few extreme predictions, especially when the data is heavy tailed. Hence an individual measure will be desirable and necessary.
- Measure of effectiveness (**)
- Measure of distribution similarity (**)
- Measure of location (**)

To sum up, all the distance measures (aggregate level and individual level) between the predicted and actual select quantile regression forest as the best model, even though it could not outperform other models in terms of measure of effective, location, distribution similarity (except for trimmed mean). If we revisit its histograms on the log scale in table 4.34, its predictions are not as heavy tailed as the actual. In other words, it tends to under-predict those extremely large costs and over-predict those tiny expenses. But if we want to select one model with most accurate predictions, QRF can be suggested because it can get the most close predictions to the actual. However, depending on the goals, the analysts or actuaries can have their own choices or adjustment based on their budget and all other information. Tweedie GAM, lognormal GLM, lognormal GLMM also perform well from different perspectives. In fact, once we are done the experiments and knowing the strength and weakness of each type of model, it is not always necessary to run so many candidate models in the future. I believe the results and methodologies shown in the dissertation can be used as a good reference for data analysts or actuaries in healthcare insurance industry.

Table 4.51 Model Selection Summary

	Specific measure	Winner	Second place	Third place
Measure of distance	MSPE	QRF	Tweedie GAM	Lognormal GLM
	MAPE	QRF	Lognormal GLM	Tweedie GAM
	Minimal distance counting	QRF	Lognormal GLMM	Lognormal GLM
Measure of distribution similarity	Bhattacharyya coefficient	Lognormal GLMM	Lognormal GEE	Lognormal GLM
Measure of effectiveness	Gains Chart	Tweedie GAM	Lognormal GLM	FMM
Measure of location	Mean	Tweedie GLM	Gamma, log link, GLM	Tweedie GAM
	Median	Lognormal GEE	Lognormal GLMM	Lognormal GLM
	Trimmed Mean	Normal, Log link, GLM	QRF	Tweedie GAM
	Winsorized Mean	Tweedie GAM	Gamma, log link, GLM	Normal, Log link, GLM

4.4 Pricing for group health claims

4.4.1 General Introduction

For group health insurance coverage, Bluhm (2012) defines that the “gross premium” represents the cost of coverage to the customer. It is composed of estimated claim cost, plus certain expenses, and less investment credits. The rating process begins with the development of claim costs, which have already been discussed in previous sections of chapter 4, with appropriate reflection of pooled claims and pooling charges. In this section, we will try to briefly review the overall pricing process and concentrate on the pricing of a special product—Stoploss coverage for self-founded health plan.

4.4.2 Manual rate development

According to pricing of group insurance by Bluhm (2012), manual premium rates are the rates that would be charged in the absence of any credibility being given to past claim experience and with no health underwriting of the group. Those manual rates are updated regularly, and are often determined separately for different group size categories and different products. Usually, they are weighted by the experience rating with a group's own experience to establish the gross premium for a particular group. Typically, manual rates will be ultimately adjusted for market strategy considerations in competitive market environment. Manual rates are often used as a reference point for a particular group or individual, and the premium rates for a group may be expressed as a percentage of the manual rates, such as "manual+10%", or "manual-5%". The determination of the appropriate percentage of manual rates (discount or extra charge) involves determining the ratio of the expected claim and the other costs for the particular group to the comparable cost expected under manual rates.

In fact, only rating variables allowed by law will be shown in the manual rate. Therefore, in practice, healthcare insurance companies always have two sets of models. One is for manual rate development or update, and only rating variable allowed by the law will be used such as age, group size, contract type, tier choice. Another set of model is for underwriting or risk management with more factors (such as health condition, group industry type, and claim history) taken into consideration and more complicated model structures. Actually, to correct price, we often need to make a few pricing assumptions including administrative expense, commissions and other sales expenses, taxes, contributions to the surplus (which typically reflect the level of risk and the profit expectation for the assumption of that risk) ,and credit for investment income on asset and cash flow. As predictive modelling gains popularity nowadays, the magic of the second set of model is playing a more and more important role in insurance risk assessment and pricing, and may finally

change the pricing procedures in insurance industry. In this dissertation, without operational cost related information available, I will focus on claim-level cost analysis and prediction. For most regular types of insurance product (group or individual), this dissertation from section 4.1 to section 4.3 carries out comprehensive analysis and suggestions are given. However, for some special product, such as self-funded health plan, we need to do some special treatments on modelling. The details will be discussed in section 4.4.3.

4.4.3 Stop-Loss Pricing for Self-funded Health Plans

4.4.3.1 Introduction

In health insurance, a stop loss policy is a product designed to protect self-funded employer from catastrophic losses. It takes effect after a certain amount has been paid in claims. Employers providing health insurance for their employees through a self-insured plan often subscribe to stop-loss policies for risk management. It is similar to the excess of loss reinsurance from the risk management point of view. We will summarize and compare the two in table 4.52.

Table 4.52 Comparison between stop-loss and excess of loss reinsurance

	Stop Loss Coverage for self-funded Plans	Excess of Loss Reinsurance
Definition	[7]A form of coverage purchased by employers who wanted to self-fund their employee benefit plans, but do not wish to assume all the liability for losses. Under a stop-loss policy, the insurance company will be liable for losses that exceed certain limits called deductibles (Much higher deductible than usual health plans).	[8]A form of reinsurance that indemnifies the ceding company for the portion of a loss that exceeds its own retention
Applications	Health, life	Mostly on Casualty, but also applied to life, health, property

[7] Insurance Glossary http://www.isuparagon.com/what_is_self_funded_plan.aspx
[8] Insurance Glossary http://www.irmi.com/online/insurance-glossary/terms/e/excess-of-loss-reinsurance.aspx?cmd=print

Types	• Specific Stop-Loss (SSL) • Aggregate Stop-Loss (ASL) SSL deductibles are more common than ASL. The size of the specific deductible is a reflection of the groups risk tolerance. Generally, the larger the group, the larger the specific deductible and vice versa. Aggregate Stop Loss limits an employer's liability to overall claim fluctuation.	• Aggregate excess of loss • Per occurrence excess of loss • Per policy excess of loss • Catastrophe excess of loss • Per risk excess of loss
Cedent	Self-funded employer	The primary insurer
Carrier	Reinsurance or Stop Loss Insurance carrier	The Reinsurance Company
Regulation	[9]Regulated Minimum SSL and ASL Deductible varied by states. For Connecticut, Minimum Specific Deductible is $6500, no minimum for ASL.	Federal and State Regulation

4.4.3.2 Review on traditional actuarial models

In practice, the traditional approach to deal with stop-loss policies in actuarial practice is called Loss Elimination Ratio (LER) approach. LER is estimated through either empirical distribution or fitted data approach. Where LER is defined as % losses eliminated due to a specific deductible. That is to say:

$$LER = \frac{Losses \text{ and LAE Eliminated by deductible}}{\text{Total groud-up losses and LAE}} \tag{4.13}$$

The empirical distribution approach is trying to estimate the LER through empirical study of the past claims, that is

$$LER = 1 - \frac{\sum \max[0, (\text{Loss} - \text{Pooling Point})]}{\sum \text{Loss}} \tag{4.14}$$

Or LER can be calculated given a continuous distribution of losses.

[9] Group Size, Minimum Specific Deductible and Aggregate Attachment Point Requirements for Stop-Loss Insurance https://www.starmarkinc.com/email/starmark/S669-329.pdf

$$LER = 1 - \frac{\int_0^a xf(x)dx + a\int_0^\infty f(x)dx}{\int_0^\infty xf(x)dx} \tag{4.15}$$

Where a is the pooling point (or deductible, or attachment point)

To sum up, all the traditional actuarial models are assuming the past claims pattern will continue in the future. A few rating variables are used to enumerate classifications in the first step, and the prices are developed for each category. In fact, high claims are very unpredictable and volatile in practice; incorrect pricing of stop-loss coverage can create huge losses. Therefore, we are trying to use more powerful predictive and data mining techniques to capture the relationship between certain characteristics and the target variable. Aiming at not only the losses on the aggregate level, but also on the individual and group level, we can use as much as information once available (not just the rating variable) to build the model and see how they can improve the pricing. I will introduce the models in section 4.4.3.3 and show it through a case study in sections 4.4.3.4.

4.4.3.3 Formulation

Let Y_{ij} be the eligible claims for individual i in group j in a policy year. We assume there are n_j enrolled employees (and dependent) in group j. Let a_{ij} denotes pooling point of SSL (Can be different for employees because of Lasering[10]) for individual i in group j, and b_j denotes the pooling point of ASL for group j. Since only those amounts that beyond the SSL can be

[10] "Lasering" is a common practice of stop loss carriers that are unwilling to accept risk associated with past high dollar claimants, or place higher SSL polling point for some participants due to high risk concerns.

reimbursed by stop-loss coverage, the SSL reimbursement due for group j can be expressed as follows:

$$\mathrm{Y}_{\mathrm{SSL,j}}=\sum_{\mathrm{i}}^{\mathrm{n_j}}\left(\mathrm{Y}_{\mathrm{ij}}-a_{ij}\mid \mathrm{Y}_{\mathrm{ij}}>a_{ij}\right)=\sum_{\mathrm{i}}^{\mathrm{n_j}}\left(\mathrm{Y}_{\mathrm{ij}}-a_{ij}\right)_{+} \tag{4.16}$$

Where $\mathrm{n_j}$ the number of participants in group j

$\mathrm{Y}_{\mathrm{ij}}-a_{ij}\mid \mathrm{Y}_{\mathrm{ij}}>a_{ij}$ denotes the excess loss variable (It is defined conditional on $\mathrm{Y}_{\mathrm{ij}}>a_{ij}$)

and $\left(\mathrm{Y}_{\mathrm{ij}}-a_{ij}\right)_{+}$ denotes the left censored and shifted variable. (It is defined as:

$\left(\mathrm{Y}_{\mathrm{ij}}-a_{ij}\right)_{+}=\begin{cases}0 & ,\mathrm{Y}_{\mathrm{ij}}<a_{ij}\\ \mathrm{Y}_{\mathrm{ij}}-a_{ij} & ,\mathrm{Y}_{\mathrm{ij}}\geq a_{ij}\end{cases}$, Values below d are nor ignored but are set equal to zero).

Allowable claims for group j would be the amount that subtracts the SSL reimbursement from the total costs:

$$Y_j=\left(\sum_{\mathrm{i}}^{\mathrm{n}}\mathrm{Y}_{\mathrm{ij}}\right)-Y_{SSL,j}=\sum_{\mathrm{i}}^{\mathrm{n}}\left(\mathrm{Y}_{\mathrm{ij}}\wedge a_{ij}\right) \tag{4.17}$$

Where $\mathrm{Y}_{\mathrm{ij}}\wedge a_{ij}$ is a limited loss variable and defined as:

$$Y=\mathrm{Y}_{\mathrm{ij}}\wedge a_{ij}=\begin{cases}X, & \mathrm{Y}_{\mathrm{ij}}<a_{ij}\\ a_{ij}, & \mathrm{Y}_{\mathrm{ij}}\geq a_{ij}\end{cases} \tag{4.18}$$

Where (4.17) is derived from the conclusion $X=\left(X-d\right)_{+}+X\wedge d$

Meanwhile, we have $E\left[\left(X-d\right)_{+}\right]=E\left[X\right]-E\left[X\wedge d\right]$

The ASL Reimbursement due for group j would be:

$$Y_{ASL,j}=\left(Y_j-b_j\right)_{+}=\left(\sum_{\mathrm{i}}^{\mathrm{n}}\left(\mathrm{Y}_{\mathrm{ij}}\wedge a_{ij}\right)-b_j\right)_{+} \tag{4.19}$$

Then the total Stop Loss Reimbursement for group j would be the summation of SSL reimbursement and ASL reimbursement, that is

$$Y_{SL,j} = Y_{SSL,j} + Y_{ASL,j} = \sum_{\mathrm{i}}^{\mathrm{n}} \left(\mathrm{Y_{ij}} - a_{ij} \right)_+ + \left(\sum_{\mathrm{i}}^{\mathrm{n}} \left(\mathrm{Y_{ij}} \wedge a_{ij} \right) - b_j \right)_+ \tag{4.20}$$

Therefore, we can get the net stop-loss premium for group j as

$$\pi(d, a, j) = E[Y_{SL,j}] = E[Y_{SSL,j}] + E[Y_{ASL,j}] \tag{4.21}$$

And the gross stop-loss premium:

$$\delta(d, a) = (1+\rho)\pi(d, a) = (1+\rho)E[Y_{SL}] = (1+\rho)\left(E[Y_{SSL}] + E[Y_{ASL}]\right) \tag{4.22}$$

Where $a = \{a_1, a_2, \cdots, a_n\}, \rho > 0$ is the relative safety loading. We note that δ is a decreasing function on d, a respectively. In this dissertation, I will only focus on the net stop-loss premium later.

Let's look the most general case: Stop-Loss Insurance with Both SSL and ASL

The expected SSL Reimbursement is

$$E\left[Y_{SSL,j}\right] = E\left[\sum_{\mathrm{i}}^{\mathrm{n}} \left(\mathrm{Y_{ij}} - a_{ij} \right)_+\right] = \sum_{\mathrm{i}}^{\mathrm{n}} E\left[\left(\mathrm{Y_{ij}} - a_{ij} \right)_+\right] \tag{4.23}$$

And the ASL Reimbursement is

$$\begin{aligned} E\left[Y_{ASL}, j\right] &= E\left[\left(\sum_{\mathrm{i}}^{\mathrm{n}} \left(\mathrm{Y_{ij}} \wedge a_{ij} \right) - b_j \right)_+\right] = E\left[\sum_{\mathrm{i}}^{\mathrm{n}} \left(\mathrm{Y_{ij}} \wedge a_{ij} \right) - \left(\left(\sum_{\mathrm{i}}^{\mathrm{n}} \left(\mathrm{Y_{ij}} \wedge a_{ij} \right) \right) \wedge b_j \right)\right] \\ &= \sum_{i}^{n} E\left[\mathrm{Y_{ij}} \wedge a_{ij}\right] - E\left[\left(\left(\sum_{\mathrm{i}}^{\mathrm{n}} \left(\mathrm{Y_{ij}} \wedge a_{ij} \right) \right) \wedge b_j \right)\right] \end{aligned} \tag{4.24}$$

Then the total Reimbursement would be

$$\begin{aligned} &E[Y_{SL,j}] = E[Y_{SSL,j}] + E[Y_{ASL,j}] \\ &= \sum_{\mathrm{i}}^{\mathrm{n}} E\left[\left(\mathrm{Y_{ij}} - a_{ij} \right)_+\right] + \sum_{i}^{n} E\left[\mathrm{Y_{ij}} \wedge a_{ij}\right] - E\left[\left(\left(\sum_{\mathrm{i}}^{\mathrm{n}} \left(\mathrm{Y_{ij}} \wedge a_{ij} \right) \right) \wedge b_j \right)\right] \\ &= \sum_{i}^{n} E\left[\mathrm{Y_{ij}}\right] - E\left[\left(\left(\sum_{\mathrm{i}}^{\mathrm{n}} \left(\mathrm{Y_{ij}} \wedge a_{ij} \right) \right) \wedge b_j \right)\right] \end{aligned} \tag{4.25}$$

In practice, the distribution of the random variable of interest depends on certain characteristics of the underlying situation. In chapter 4, we discussed how to get the estimates for $E\left[\mathrm{Y_{ij}}\right]$ given different individual and account level characteristics. Now we only need to focus on estimating $E\left[\left(\left(\sum_{\mathrm{i}}^{\mathrm{n}}\left(\mathrm{Y_{ij}}\wedge a_{ij}\right)\right)\wedge b_j\right)\right]$. Once we fit the distribution of $\mathrm{Y_{ij}}$ from data, we can either solve this analytically if it is possible, or by simulation. In the following section, I will show a case study for SSL only stop-loss insurance.

4.4.3.4 Case Study: SSL only stop-loss insurance

In practice, many large group employers, especially for those with more than 1000 participants, they prefer to buy SSL only policy since they have higher tolerance to risk. Usually they are looking at aggregate stop-loss going forward as just "sleep insurance". Hereby we can treat SSL only policy as $b_j \to \infty$, a special case of SSL & ASL policy. Meanwhile, although the carrier may place a lasering on a member with an ongoing condition, and the specific deductible for him/her may be higher than others. In most cases, the same specific deductible amount is used for all participants in the group and lasoring is only allowed in the first year. In this dissertation, since all the participants are not in their first year of the health plan, we simply assume the pooling point for SSL is the same for everyone in that group, in other words, $a_{1,j} = a_{2,j} = ... = a_{n,j} = a_j$.

For SSL only stop-loss insurance, let's say, for group j, the ASL Reimbursement will be zero and the expectation for SSL reimbursement will be:

$$E\left[Y_{SSL,j}\right] = E\left[\sum_{\mathrm{i}}^{\mathrm{n_j}}\left(\mathrm{Y_{ij}} - a_{ij}\right)_+\right] = \sum_{\mathrm{i}}^{\mathrm{n_j}} E\left[\left(\mathrm{Y_{ij}} - a_j\right)_+\right] \tag{4.26}$$

Then the net stop-loss premium per participant will be:

$$\pi\left(a_j, j\right) = E[Y_{SL,j}]/n_j = \left(E[Y_{SSL,j}] + E[Y_{ASL,j}]\right)/n_j = \sum_{\mathrm{i}}^{\mathrm{n_j}} E\left[\left(\mathrm{Y_{ij}} - a_j\right)_+\right]/n_j \tag{4.27}$$

Where $E\left[Y_{ASL}\right] = 0$

Mathematically, we have

$$\begin{aligned} & E\left[\left(\mathrm{Y_{ij}} - a_j\right)_+\right] \\ &= E\left[\left(\mathrm{Y_{ij}} - a_j\right)_+ \mid \mathrm{Y_{ij}} \geq a_j\right] P\left(\mathrm{Y_{ij}} \geq a_j\right) + E\left[\left(\mathrm{Y_{ij}} - a_j\right)_+ \mid \mathrm{Y_{ij}} < a_j\right] P\left(\mathrm{Y_{ij}} < a_j\right) \\ &= E\left[\mathrm{Y_{ij}} - a_j \mid \mathrm{Y_{ij}} \geq a_j\right] P\left(\mathrm{Y_{ij}} \geq a_j\right) + 0 \\ &= E\left[\mathrm{Y_{ij}} - a_j \mid \mathrm{Y_{ij}} \geq a_j\right] E\left[1_{\{\mathrm{Y_{ij}} \geq a_j\}}\right] \end{aligned} \tag{4.28}$$

Where $\left(\mathrm{Y_{ij}} - a_{ij}\right)_+ = \begin{cases} 0 & , \mathrm{Y_{ij}} < a_{ij} \\ \mathrm{Y_{ij}} - a_{ij} & , \mathrm{Y_{ij}} \geq a_{ij} \end{cases}$

Therefore the net premium per person would be

$$E\left[Y_{SSL,j}\right]/n_j = \sum_{\mathrm{i}}^{\mathrm{n_j}} E\left[\left(\mathrm{Y_{ij}} - a_j\right)_+\right]/n_j = \sum_{\mathrm{i}}^{\mathrm{n_j}} E\left[\mathrm{Y_{ij}} - a_j \mid \mathrm{Y_{ij}} \geq a_j\right] E\left[1_{\{\mathrm{Y_{ij}} \geq a_j\}}\right]/n_j \tag{4.29}$$

In practice, the distribution of the random variable of interest depends on certain characteristics of the underlying situation. Hereby we can approach this by regression models. This is similar to what we did in two stage regression. We need to estimate two parts, $E\left[1_{\{\mathrm{Y_{ij}} \geq a_j\}}\right]$ and $E\left[\mathrm{Y_{ij}} - a_j \mid \mathrm{Y_{ij}} \geq a_j\right]$ separately.

First, we can create an indicator variable $1_{\{\mathrm{Y_{ij}} \geq a_j\}} = \begin{cases} 1 & if \ \ \mathrm{Y_{ij}} \geq a_j \\ 0 & otherwise \end{cases}$, then we can have our regression models on binary targets and get estimation for $E\left[1_{\{\mathrm{Y_{ij}} \geq a_j\}}\right]$. Next we can build the model on the amounts which are beyond the pooling point and get estimation for $E\left[\mathrm{Y_{ij}} - a_j \mid \mathrm{Y_{ij}} \geq a_j\right]$. A lot of predictive model candidates are available to us. For binary target,

we can use logistic (or probit) regression or decision tree; for the amount model, we can use GLM (generalized linear model), GAM (generalized additive model), GLMM (generalized linear mixed model), quantile regression etc as we discussed in previous sections in chapter 4.

Here since stop-loss policy is for group-insurance only, and the goal is to correct price, the goal would be the most accurate net premium for each group. Mathematically, we aim to estimate the most accurate $\sum_{i}^{n_j} E\left[Y_{ij} - a_j \mid Y_{ij} \geq a_j \right] E\left[1_{\{Y_{ij} \geq a_j\}} \right] / n_j$. For a certain group, n_j and a_j are constants. a_j may have different levels for different plans based on the employers' budget and choice on risk tolerance. I look at a few most commonly used values for a_j in practice in this dissertation: 10000, 125000, 150000, and 175000. Rather than building a separate model for each group based on its own claim experience, here I use hierarchical models for stop-loss pricing where both individual level and group level effects are taken into consideration. For illustration, I fit the model on a simple dataset with only two accounts. Both are national accounts with various individual and group level characteristics. To be specific, the modeling process is carried out through the following steps:

- First, we need to estimate the conditional distribution function, which is equivalent to the expectation of the indicator variable mathematically. I generate the shifted and left censored variable given different pooling levels; and mark all the data points that have been censored by an indicator variable (censored 0, uncensored 1). Then the most widely used logistic regression is used to model the binary target. The variables are selected through stepwise AIC by Fisher's scoring optimization.

$$P\left(Y_{ij} \geq a_j \right) = E\left[1_{\{Y_{ij} \geq a_j\}} \right] = f^{-1}(X_{ij}\beta) \tag{4.30}$$

- Second, we will only use the observations that beyond the pooling point to build the generalized linear mixed model with only random account intercept and lognormal assumption on the response variable. The estimation is done through GLIMMIX procedure in SAS by restricted maximal likelihood estimation technique.

$$\begin{aligned} & g(E\left[\mathrm{Y_{ij}}-a_j \mid \mathrm{Y_{ij}} \geq a_j\right])=\partial_j+\mathrm{X_{ij}}\beta \\ & \partial_j \sim N\left(\mu_j, \sigma_j^2\right) \end{aligned} \tag{4.31}$$

- Third, combine the results in step 1 and step 2 together, and sum the predictions for each account and divided by the group size, we can get the net premium estimates for each group

$$\pi\left(a_j, j\right)=E[Y_{SL,j}]/n_j=\sum_{\mathrm{i}}^{\mathrm{n_j}} E\left[\mathrm{Y_{ij}}-a_j \mid \mathrm{Y_{ij}} \geq a_j\right] E\left[1_{\{\mathrm{Y_{ij}} \geq a_j\}}\right]/n_j \tag{4.32}$$

- Fourth, repeat the step 1 and step 2 for another three pooling levels. The variables selected might be different for different pooling levels, but with the same variable selection criteria and techniques. Many variables are proved to be predictive and powerful, such as Group size, Rating Area, Group's industry type, Health status etc.

Finally, the results are summarized in table 4.53. Account A is in the financial service industry with 17693 participants; and account B is in public administration industry with 14489 participants. Since we don't have multiple years' data available, we will test the model on data partition bases, rather than longitudinal data bases. 60% of the data are used for training and 40% are used for validation. The average losses beyond the pooling point on the training dataset are used as the prediction for the testing datasets. There is a lot of unexplained randomness there. The predictions can perform very well if the distribution of the response

variable stays stable on both training and testing by coincidence, but performs poorly if that is not the case. For example, if we look at account B, the calculated net stop-loss premiums from the training set are very close to the values on the test set. It is very difficult for our predictive models to beat the mean model on account B. But if we look at account A, there is a big difference between the training and test sets, our predictive models can win under all cases. Actually, this comparison is a little bit tough for our predictive models since we are using the same year and same account data to test the results, while we believe the volatility will be much serious in practice when we need do pricing for new groups and future years. Another concern is that the results given by the predictive models didn't keep the monotonicity between the pooling point and net premium (The larger the pooling point, the lesser the net premium). We believe we can improve it by credibility pricing which is a common practice in actuarial science, but we are not addressing this issue in this dissertation considering the time limit. In spite of that, our predictive models still win to a large extent (5.5/8=68.75%) on aggregate and it is able to tell us which individuals are at higher risks. That is not possible in our traditional aggregate level models. Some insurance companies maintain various health service programs, such as health coaching. Once our predictive models figure out which patients are at higher risk, insurance carrier may approach them through those programs, thus finally help reduce insurance costs.

Table 4.53 Net stop-loss premium

	Pooling point	Empirical stop loss on Training	Actual Net Premium on testing	Prediction by Models	Winner
Account A	100k	1926.57	2279.91	2190.02	Predictive Model
	125k	1722.21	2073.00	1901.08	Predictive Model
	150k	1556.98	1902.00	2164.68	Predictive Model
	175k	1418.6	1759.42	1769.51	Predictive Model
Account B	100k	2090.22	2261.76	1958.36	Mean Model
	125k	1890.89	2064.28	1776.77	Mean Model
	150k	1729.65	1908.67	1849.42	Predictive Model
	175k	1594.02	1780.16	1595.46	Almost Tie

Chapter 5 Conclusions and future work

Predictive modeling has grown to be a powerful tool in healthcare in terms of cost control, pricing, reserving, marketing and risk management. ETGs (Episode Treatment Groups) were introduced for identifying and classifying an entire episode of care for evidence-based medicine and healthcare management reporting. In spite of its wide use, how to effectively use ETGs for health plan pricing is still an outstanding and interesting issue from the perspective of health plan carriers. This research investigated the application of ETGs in health plan pricing and risk management, with a focus on model selection for ETG-based costs. The best-performed model can vary depending on the disease. Insurance loss distributions are commonly skewed with heavy tails. Using lighter-tailed distributions for modeling may significantly bias the results. Unfortunately, this issue has not been carefully addressed in many situations of actuarial practice. This dissertation compares four potential models (distributions): lognormal, gamma, log-skew-*t*, and Lomax; where gamma is the default distribution for positive continuous explanatory variables in practice. The experimental results show that none of the metrics select the gamma distribution as the best model for any of the 320 different ETGs. Thus, one needs to be cautious in selecting gamma models for heavy-tailed data.

In addition to model selection and averaging, this dissertation also contributes by recommending various metrics for data of different sizes and analysts with different goals. Four metrics are considered in this dissertation: AIC weights, BIC weights, Bayesian parallel model selection and random forest feature classification. AIC and BIC are commonly used maximum likelihood estimate driven information criteria, and both of them try to balance good fit with parsimony. In general, BIC penalizes free parameters more strongly than AIC, but their results are quite similar in most cases experimented in this study. Bayesian model averaging computes the probabilities

of each model being the best given the data among all models under consideration. It enables model averaging and provides deeper insights into the relationships among the models.

Since I have 33 million ETG cost observations from 9 million claimants, the first three metrics (AIC, BIC weights and Bayesian model averaging) struggle with big data in terms of processing time. Hence random forest feature classification is proposed in order to achieve greater efficiency. Random forests treat all the data sets following one distribution as one cluster, and select the best distribution by allocating the observations into the correct cluster. The classification process is divided into three steps: domain specific feature extraction, random forest training for prediction, and random forest model selection. My experimental results show that the moment-based features perform better than percentile-based features in distinguishing distributions. If both moment-based and percentile-based features are used, we can achieve the lowest out-of-bag error rate and the best performance in distinguishing distributions. Since the random forest model selection is based on the extracted information rather than the original big data sets, it has significantly reduced the computing time. Experimental results show that random forest only take 2 minutes for the whole process, but AIC/BIC spend around 4 hours. Bayesian parallel model selection may need approximately 4 weeks on the same task and the same experimental platform. Furthermore, the accuracy of the four metrics is also compared. On average, the Bayesian approach can achieve the highest accuracy because it exactly identifies lognormal and log-skew-*t* distribution though is less certain about gamma and Lomax compared to AIC weights. AIC weights perform well on average. Random Forest performs a little bit worse than the other two, but it can still identify the model with the best fit. Especially when the data volume is huge, its efficiency is promising without losing much accuracy. The future work is to investigate the possible dependence among ETGs, and incorporate ETGs into risk assessment regression framework, as well as disease specific product design and pricing.

In the second part of this study, a claim-based risk assessment in healthcare is conducted with data from a major national health insurer to determine the relative costs of a person or a group based on their medical history, demographics, regions, etc. Given the target variable in this study is semi-continuous, heavy-tailed and clustered, all the candidate models are tailored for those properties, which are very common properties of cost data in actuarial practice. Four types of semi-continuous models are fit in this dissertation: Tweedie model, two part model, quantile regression and finite mixture model. To select the best model, correct specification of the models and goodness of fit is the first concern. Different information criteria tailored for those models and cross validation were tested and compared. In fact, the ultimate goal of predictive modelling is to generate the most accurate predictions. An objective and comprehensive model selection framework for decision making from different perspectives are suggested in the dissertation. I aim at not only choosing the best model under certain goal, but also making sure that the evaluation techniques themselves are appropriate under certain situations. Objective and comprehensive comments are given to each evaluation measure, which is desirable for decision maker when in front of a bunch of evaluation measures and none of the models can win over all measures. Four evaluation mechanisms discussed in this dissertation include measure of distance (such as MSPE, MAPE, and minimal distance probability matrix), measure of effectiveness (lift chart), measure of distribution similarity (such as histograms and Bhattacharyya Coefficient) and measure of location (mean, median, trimmed mean, Winsorized mean). By the measure of distance, quantile regression forest will be the winner, while lognormal GLM and Tweedie GAM also perform well. In contrast, by the gains chart measure, Tweedie GAM, lognormal GEE and lognormal GLM were the top three. Lognormal GLMM is the winner under the measure of distribution similarity. The results shown in the dissertation can be used as a good reference for data analysts or actuaries in healthcare insurance industry for modelling, risk management or pricing.

It is also worth mentioning that this study proposed the minimal distance probability matrix as a model selection technique. Inspired by Arrow's “impossibility” theorem in Game theory, one on one comparison on minimal distance counting can give us the most unbiased and robust strategies for model selection decision making. Rather than using an aggregate measure, the benefit of this individual level evaluation technique lies in take every prediction accuracy into consideration and a few extreme biased prediction won’t distort the overall results like other distance measures. Meanwhile, we can derive the matrix of relations from the probability matrix, and investigate the properties of the matrix relations to help us select best model. Some interesting conclusions between transitivity of the matrix of relation and the existence of a single robust model among candidates.

Last but not least, this dissertation contributes to the stop-loss pricing model for self-funded health plans. After a review of traditional actuarial models and basic information of stop-loss policy, many concerns were raised about a few highly simplified assumptions in traditional actuarial models; and high claims are very unpredictable and volatile in practice; incorrect pricing of stop-loss coverage can create huge losses. Then a new pricing model based on predictive analytics is built. First, the formulas that denote the net stop-loss premium are derived in terms of left censored and shifted variables, as well as limited loss variables for different types of stop-loss policy, such as stop-loss policy with both SSL and ASL, or SSL only stop-loss policy. Then, predictive models are built to capture the relationship between certain characteristics and the target variable. A case study about SSL only stop-loss insurance is given. For future work, we will try to derive more theoretical results or solutions for stop-loss policy with both SSL and ASL, and test the models on a larger scale of data.

The healthcare predictive analytics has become increasingly prominent recently. We can expect more smart technologies, much bigger volume of electronic health records, and more data integration/ fusion/ cleaning/ analysis challenges to rise soon. I believe data science will play a critical role in the development of artificial intelligence applications. More big data techniques will open frontier in insurance risk assessment and pricing, finally leading to a happier and healthier lifestyle.

Bibliography

[1] Agresti, A. (2013). Categorical data analysis. John Wiley & Sons.

[2] Akaike, H. (1978). On the likelihood of a time series model. The Statistician, 27, 217–235.

[3] Alice Rosenblatt, and Sim Segal(2012), Risks & Mitigation for Health Insurance Companies, Sponsored by Society of Actuaries Health Section.

[4] Arrow, K.J.,(1950) "A Difficulty in the Concept of Social Welfare", Journal of Political Economy 58(4), pp. 328–346.

[5] Bai, Y. (2009). Convergence of Adaptive Markov Chain Monte Carlo Methods. PhD dissertation, Department of Statistics, University of Toronto.

[6] Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley and Sons.

[7] Bhattacharyya, A. (1943). "On a measure of divergence between two statistical populations defined by their probability distributions". Bulletin of the Calcutta Mathematical Society 35: 99–109. MR 0010358.

[8] Brady T. West; Kathleen B. Welch; Andrzej T Galecki (2014). Linear Mixed Models: A Practical Guide Using Statistical Software, Second Edition. CRC Press. pp. 56–. ISBN 978-1-4665-6099-4.

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

[10] Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media.

[11] Carlin, B.P. and Chib, S. (1995). Bayesian model choice via Markov Chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B, 57(3), 473–484.

[12] Chaudhuri, P., & Loh, W. Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. Bernoulli, 561-576

[13] Chen, M.H. and Schmeiser, B. (1993). Performance of the Gibbs, Hit-and-Run, and Metropolis samplers. Journal of Computational & Graphical Statistics, 2(3), 251–272.

[14] Congdon, P. (2006). Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. Computational Statistics & Data Analysis, 50(2), 346–357.

[15] Dai, J., Li, Z., & Rocke, D. (2006). Hierarchical logistic regression modeling with SAS GLIMMIX. In Proceedings of the Thirty-first Annual SAS Users Group International Conference. Cary, North Carolina: SAS Institute Inc.

[16] Dave Kessler, Allen McDowell (2012), Introducing the FMM Procedure for Finite Mixture Models, SAS Global Forum 2012

[17] Dove, H.G., Duncan, I., and Robb, A. (2003). A prediction model for targeting low-cost, high-risk members of managed care organizations. The American Journal of Managed Care, 9(5),381–389.

[18] Duncan, I. (2011). Healthcare Risk Adjustment and Predictive Modeling. Actex Publications.

[19] Edward W. Frees; Richard A. Derrig; Glenn Meyers (2014). Predictive Modeling Applications in Actuarial Science: Volume 1, Predictive Modeling Techniques. Cambridge University Press. ISBN 978-1-139-99231-2.

[20] Eling, M. (2012). Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? Insurance: Mathematics & Economics, 51(2), 239–248.

[21] Ferreira, J. and Steel, M.F. (2007). A new class of skewed multivariate distributions with applications to regression analysis. Statistica Sinica, 17(2), 505–529.

[22] Fitzmaurice, G. M., & Molenberghs, G. (2009). Advances in longitudinal data analysis: an historical perspective. Longitudinal Data Analysis, 3-30.

[23] Flom, P. L., & Cassell, D. L. (2007). Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. InNorthEast SAS Users Group Inc 20th Annual Conference: 11-14th November 2007; Baltimore, Maryland.

[24] 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.

[25] Friedrichs and Hense (2007): Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression. Mon. Wea. Rev.

[26] Garrett M. Fitzmaurice; Nan M. Laird; James H. Ware (2012). Applied Longitudinal Analysis. John Wiley & Sons. pp. 344–. ISBN 978-1-118-55179-0.

[27] Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

[28] Goeman, J., Meijer, R., & Chaturvedi, N. (2012). L1 and L2 penalized regression models. cran. r-project.

[29] Hardin, J. W., & Hilbe, J. M. (2003). Generalized estimating equations. Boca Raton, FL: Chapman and Hall/CRC Press

[30] Hartman, B.M. and Groendyke, C. (2013). Model selection and averaging in financial risk management. North American Actuarial Journal, 17(3), 216–228.

[31] Hastie, T. J. and Tibshirani, R. J. (1990). Generalized Additive Models. Chapman & Hall/CRC. ISBN 978-0-412-34390-2.

[32] He, X., Ng, P., & Portnoy, S. (1998). Bivariate quantile smoothing splines.Journal of the Royal Statistical Society: Series B (Statistical Methodology),60(3), 537-550.

[33] Hoffman, M.D. and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in hamiltonian Monte Carlo. Journal of Machine Learning Research, 15(1), 1351–1381.

[34] Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages.The American Statistician, 50(4), 361-365

[35] Ita G G Kreft; Ita Kreft; Jan de Leeuw (1998). Introducing Multilevel Modeling. SAGE Publications. ISBN 978-0-7619-5141-4.

[36] Jeremie Juban , Lionel Fugon, George Kariniotakis (2007): Probabilistic short-term wind power forecasting based on kernel density estimators. European Wind Energy Conference. (QR Forests)

[37] John Bjrnar Bremnes (2004): Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. Mon. Wea. Rev.

[38] Jones, M. and Faddy, M. (2003). A skew extension of the t-distribution, with applications. Journal of the Royal Statistical Society: Series B, 65(1), 159–174.

[39] Kleiber, C. and Kotz, S. (2003). Statistical Size Distributions in Economics and Actuarial Sciences. Wiley.

[40] Klugman, S.A., Panjer, H.H., and Willmot, G.E. (2012). Loss Models: From Data to Decisions, 4th edition. Wiley.

[41] Koenker, R. and Bassett, G. W. (1978), “Regression Quantiles,” Econometrica, 46, 33–50.

[42] Koenker, R., & Hallock, K. (2001). Quantile regression: An introduction. Journal of Economic Perspectives, 15(4), 43-56.

[43] Koenker, R., & Schorfheide, F. (1994). Quantile spline models for global temperature change. Climatic Change, 28(4), 395-404.

[44] Kuha, J. (2004). AIC and BIC comparisons of assumptions and performance. Sociological Methods & Research, 33(2), 188–229.

[45] Liaw, A. and Wiener, M. (2002). Classification and Regression by Random Forest. R news, 2(3), 18–22.

[46] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest.R news, 2(3), 18-22.

[47] Lin, Y., & Jeon, Y. (2006). Random forests and adaptive nearest neighbors.Journal of the American Statistical Association, 101(474), 578-590.

[48] Marzban, C. (2003). Quantile Regression. Applied Physics Lab., Department of Statistics, Univ. of Washington, Seattle, WA, USA, 98195.

[49] Mehmud, S. M., & Yi, R. (2012). Uncertainty in Risk Adjustment. Society of Actuaries,

[50] Michael clark (2009), Generalized additive models-Getting started with additive models in R, Center for social research, University of Notre Dame

[51] Michael J. A. Berry; Gordon S. Linoff (2008). MASTERING DATA MINING: THE ART AND SCIENCE OF CUSTOMER RELATIONSHIP MANAGEMENT. Wiley India Pvt. Limited. ISBN 978-81-265-1825-8.

[52] Neal, R. (2011). MCMC using Hamiltonian dynamics. In Handbook of Markov Chain Monte Carlo (S. Brooks, A. Gelman, G.L. Jones, X-L. Meng, eds.), 113–162. Chapman & Hall/CRC.

[53] Neelon, B. (2013). Two-Part Models for Zero-Modified Count and Semicontinuous Data (Doctoral dissertation, Duke University).

[54] Olivia Parr Rud (2001). Data Mining Cookbook: Modeling Data for Marketing, Risk, and Customer Relationship Management. John Wiley & Sons. pp. 385–. ISBN 978-0-471-43751-2.

[55] Pan, W. (2001). Akaike's information criterion in generalized estimating equations. Biometrics, 57(1), 120-125.

[56] Portnoy, S., & Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators.Statistical Science, 12(4), 279-300.

[57] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[58] Schwartz, E.M., Bradlow, E.T., and Fader, P.S. (2014). Model selection using database characteristics: Developing a classification tree for longitudinal incidence data. Marketing Science, 33(2), 188–205.

[59] Shtatland, E.S., Moore, S., Dashevsky, I., Miroshnik, I., Cain, E., and Barton, M.B. (2000). How to be a Bayesian in SAS: Model selection uncertainty in PROC LOGISTIC and PROC GENMOD. In Proceedings of the 13th Annual NorthEast SAS Users Group Conference, 1–9.

[60] Simon Wood.(2015). Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation. R package, version 1.8-5.

[61] Stone, C.J. (1985),"Additive Regression and Other Nonparametric Models", Annals of Statistics, 13, 689-705.

[62] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

[63] Trevor Hastie; Robert Tibshirani; Jerome Friedman (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Science & Business Media. ISBN 978-0-387-84858-7.

[64] Tweedie, M.C.K. (1984). "An index which distinguishes between some important exponential families". In Ghosh, J.K.; Roy, J. Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. Calcutta: Indian Statistical Institute. pp. 579–604. MR 786162.

[65] William F. Bluhm (2012). Group Insurance. ACTEX Publications. ISBN 978-1-56698-932-9.

[66] Winkelman, R., & Mehmud, S. (2007). A comparative analysis of claims-based tools for health risk assessment. Society of Actuaries, 1-70.

[67] Wood, S. N. (2006). Generalized Additive Models: An Introduction with R. Chapman & Hall/CRC. ISBN 978-1-58488-474-3.

[68] Wood, S. N. (2000) Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society: Series B 62(2),413-428.

[69] Wood, S. N. (2008). "Fast stable direct fitting and smoothness selection for generalized additive models". Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (3): 495–518. doi:10.1111/j.1467-9868.2007.00646.x.

[70] Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. Biometrics, 121-130.