Université de Montréal

Statistical Analysis of Machine Learning Estimators of Insurance Premiums

par

Linyan Meng

Département de mathématiques et de statistique

Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès Sciences (M.Sc)
en statistique

Août 2002

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Statistical Analysis of Machine Learning Estimators of Insurance Premiums

présenté par

Linyan Meng

a été évalué par un jury composé des personnes suivantes:

Yves Lepage
(président-rapporteur)

Christian Léger
(directeur de recherche)

Yoshua Bengio
(co-directeur)

Roch Roy
(membre du jury)

Mémoire accepté le:

2 novembre 2002

Acknowledgment

I would like to express my heartfelt gratitude to my supervisor, Professor Christian Léger, for not only his guidance, patient help and constructive criticism, but also his trust, understanding and support. I am lucky to work with an expert like him in the field of statistics, as well as a truly person in life.

I am very indebted to my co–director, Professor Yoshua Bengio, for his continuous counseling, practical assistance and immense help, and to let me be part of his LISA lab. I have learned much from him. All I remember will be good memory.

Thanks are also due to Pascal Vincent and other members at LISA lab, Joumana Ghosn, Nicolas Chapados, Charles Dugas and Réjean Ducharme. I have benefitted greatly from interactions with them. I also wish to thank Jean-François Boudreau and all other friends in Montréal, for their valuable help.

I would like to express my appreciation to Professor Robert Cléroux, Professor Yves Lepage for their excellent lectures and kind help.

Finally, I wish to take this opportunity to express my sincere thanks to Professor Martin Goldstein, Professor Mario Lefebvre at École Polytechnique for their gracious hospitality, encouragement and invaluable help at the beginning of my studies here.

It is true that "thanks" is just a small word. However, I will for life time hold as much appreciation as this nice little word may indicate.

Sommaire

Dans l'industrie de l'assurance, plusieurs éléments affectent les primes versées par les assurés. Basé sur un grand nombre de contrats d'assurance, une compagnie d'assurance désire bâtir un modèle prédictif des réclamations. Ce modèle pourrait par la suite être utilisé pour décider des primes à charger. Plusieurs méthodes de modélisation telles que les modèles linéaires, les modèles linéaires généralisés et les réseaux de neurones peuvent être utilisées. Un critère pour juger de la qualité d'un modèle est la comparaison de la moyenne des réclamations et de la prévision moyenne dans chacun des nombreux sous-ensembles des variables explicatives.

Nous nous concentrons dans ce mémoire sur l'analyse des résidus, utilisée afin de comparer différents modèles. Des tests basés sur la normalité sont effectués sur chaque sous-ensemble disjoint. Puisque certaines réclamations sont très élevées, la distribution des résidus pour tous les modèles possédera des queues épaisses. Même si les tailles échantillonnales sont très grandes, les méthodes de rééchantillonnage seront aussi utilisées. Finalement, puisque beaucoup de sous-ensembles sont étudiés, un grand nombre de tests seront effectués. Les ajustements de Bonferroni seront utilisés pour tenir compte des comparaisons multiples.

Mots clés: Primes d'assurance, test d'hypothèses, bootstrap, ajustement à la Bonferroni, valeurs aberrantes, asymétrie, aplatissement.

Summary

In the insurance industry, many factors affect premiums. Based on a large number of previous contracts, they want to come up with a model to predict claims and therefore decide on the premiums to charge. Many methods can be used, such as linear models (LM), generalized linear models (GLM), and neural networks (NN). One criterion to judge the quality of a model is to compare the average claim to the average prediction in each of a large number of disjoint subsets of the explanatory variables.

This thesis concentrates on data analysis of residuals for comparing different models. We perform traditional normal tests on each disjoint subset. Since some claims are extremely large, the distribution of the residuals of any model will have heavy tails. Even though the sample sizes are very large, bootstrapping will also be used. Finally, we are facing a series of tests as many subsets are involved. Bonferroni adjustments will be made to account for the multiple comparisons.

Keywords: Insurance premium, hypothesis test, bootstrap, Bonferroni adjustment, outliers, skewness, kurtosis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the automobile insurance industry, predicting premiums accurately is the most important part in pricing automobile insurance contracts. Automobile insurance companies need to use the predictions to evaluate their current pricing structure in order to come up with an improved pricing methodology.

Generally, the development of a model that predicts premiums should consider two basic criteria as follows:

- **Minimum MSE**: The Mean-Squared Error of a model should be as small as possible;
- **Fairness**: The mean residual of a model within any subset should be close to zero.

Finding a model with small MSE is well documented and is the subject of many books and papers in regression. This will not be the subject of this thesis. Satisfying the criterion of fairness, however, is complicated due to several factors, including:

- Possibly interesting subsets of insured are numerous. It is difficult to simultaneously test for fairness of all possible subsets;

- Extremely large claims exist and they are hard to predict. This leads to a distribution of residuals of a model having a heavy tail. Therefore, the normal distribution might not provide a good reference for test statistics;

- When more than one subset of insured is involved, differences might incorrectly be declared significant simply due to the number of tests being performed. So we have to take into account the multiple comparisons aspect to correctly conclude about the fairness of the model.

This thesis concentrates on data analysis to evaluate, in data mining context, whether a model is fair or not, that is to evaluate whether the mean residual of a model for each subset is close to zero. As an example, our discussions will be based on a data set from a major North-American insurance company–Université de Montréal study of the pricing of automobile insurance contracts. More information about the joint project will be found in Chapter 2.

1.1 Evaluation of the fairness of a model

The evaluation of the fairness of a model consists of testing that the mean residual for each subset of a model is zero.

There are various ways to split the data into subsets. We can split the data according to explanatory variables as well as dependent variables. Since we can not cover all possible subsets, the evaluation of the fairness of a model will be mainly focused on certain meaningful subsets.

In the case where the subsets are based on the dependent variables, the subsets and the observations are dependent and the variances in each subset might differ considerably. It may therefore be difficult to analyse the data through an ANOVA model. Therefore, we will instead use t-tests for each subset and correct for multiple comparisons.

1.2 Contributions of this thesis

First of all, Bonferroni adjustments are used. The Bonferroni procedure (Miller, 1985), is an old statistical tool which depends solely on this simple probability inequality,

$$P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i). \tag{1.1}$$

The procedure concerns adjusting downward the level of each individual hypothesis test of a particular study to ensure that the probability of incorrectly rejecting at least one of the hypotheses is at most alpha. Because many subsets are involved in evaluating the fairness of a model, hence, we are facing a series of tests, and we will make Bonferroni adjustments to account for multiple comparisons (Hsu, 1996).

Secondly, the bootstrap method is used (Efron & Tibshirani, 1993). The distribution of claims has some very large values which are hard to predict leading to a skewed distribution of residuals, which is far from a normal distribution. Generally, even if the distribution of the observations is not normal, when the sample size is large, we could approximate the distribution of the t-test by the normal distribution because of the central limit theorem. In this study, the total available sample size is 7 418 681, which is really very large. But is it large enough to compensate for the extreme skewness of the distribution of the residuals? Therefore, we will also use the bootstrap to take into account the skewness.

Thirdly, because of the very large sample size, the tests have very high power so that hypotheses of a zero mean residual can easily be rejected without being wildly wrong. We will compute confidence intervals to see just how far from zero is the mean residual of each subset to judge on the fairness of a model. There are also other things that we have learned working in a data mining context. The residuals have very strange distributions: millions of data point are centred on zero while a few points are very large and positive. Many standard graphical methods may not work well to visualize the situation. QQplots work better. Handling data of such extremely high dimension is beyond the ability of some standard statistical software. Therefore, we chose to implement all computations in PLearn, a set of powerful functions in C++, developed by Pascal Vincent, Yoshua Bengio and other members of Lisa lab in Université de Montréal.

1.3 Overview of thesis

The next chapter describes the data and provides information about the model that was selected. Chapter 3 introduces bootstrap methodology for both hypothesis testing and confidence intervals. Chapter 4 provides the necessary background for understanding the Bonferroni procedure and describes Bonferroni adjustment for hypothesis testing, and confidence intervals. The experimental results are reported in Chapter 5.

Chapter 2

Data from the joint project

Based on information from particular insurance contracts between 1990 and 1998, the joint project between a major North-American insurance company and Université de Montréal mainly focused on evaluating and comparing several statistical models for pricing automobile insurance contracts to help the company evaluate its current pricing methodology by comparing it with the pricing structure obtained with these statistical learning algorithms.

Several models were investigated in this study including: a model with a constant, linear model (Draper & Smith, 1998), generalized linear model (McCullagh & Nelder, 1989), greedy multiplicative model, CHAID decision tree, a combination of CHAID and linear model, neural network (Bishop, 1998), softplus neural network (Dugas, Bengio, Belisle, Nadeau & Garcia, 2001), regression support vector machine, and a mixture model. The best model was this mixture model.

We will describe the best model in section 2.2. In the next section, we first describe the data set. Since the goal of this thesis is not the comparison of different statistical learning algorithms, we do not go into all details of the model description and data set preparation.

2.1 Data information

2.1.1 Data preprocessing

The data being studied came from eleven raw data files supplied by the company. We filtered the data in order to suit model development. The preprocessing was based on the following points:

1. Keeping only policies that lasted from the year 1991 to 1998, to ensure that we kept only policies that correspond to similar conditions.
2. Filtering out certain categories of contracts, such as commercial vehicles.
3. Eliminating the policies containing missing values (i.e. the value of some of the input variables was not specified in the data files), to avoid developing complicated learning algorithms.
4. Keeping only policies whose duration was close to one year, to simplify the minimization of the MSE criterion.

After preprocessing, there was a total of 7 418 681 remaining policies, 32 explanatory variables and 5 dependent variables.

Table 2.1 gives the list of explanatory variables; their detailed definitions are in Appendix A

Table 2.1: List of explanatory variables

lease_indicator	policy_start_date	territory
rate_class	claim_rated_scale	fleet_discount
third_party_ext_limit	collision_deductible	comp_deductible
loss_of_use	specified_perils	roadstar_indicator
limited_depreciation	replacement_cost	vehicle_model_age
vehicle_color	drv_is_owner	nyears_since_exam
drv_class	drv_exam_status	drv_age
drv_sex	nyears_since_original_lic	age_last_conviction
drv_last_penalty_pts	drv_num_convictions	drv_total_penalty_pts
age_last_accident	drv_num_accidents	age_last_suspension_end
drv_num_suspensions	drv_suspension_days	

2.1.2 Dependent variables

The dependent variables in this study are 6 Kind Of Loss groups, abbreviated KOL groups, including *bodily injury, property damage, accident death, collision L.O.U., comprehensive,* and *roadstar.* We did not consider the 6th KOL group *roadstar* as it represents very small amounts and displayed unexplained and very peculiar temporal non-stationarities.

The description of the five dependent variables is as follows:

- **Bodily injury**: Claim amount for bodily injury coverage of the other driver;

- **Property damage**: Claim amount for property damage coverage of the other driver;
- **Accident death**: Claim amount for accident and death benefits coverage for the insured;
- **Collision L.O.U.**: Claim amounts for collision and loss of use insurance coverage.
- **Comprehensive**: Claim amounts for comprehensive insurance coverage, i.e., loss or damage to the insured vehicle not covered by collision insurance.

For each dependent variable, the values represent the incurred amount for a particular contract. In principle, the values should be positive, or zero if there was no claim. In fact, there are a few cases of very large positive claims, but for most policies, there was no claimed amount leading to lots of zero in the data set. However, due to accounting practices, there were a few cases of negative values meaning that money was reimbursed to the company.

Figure 2.1 presents the histogram of the distribution of incurred amount for the KOL group *bodily injury* for a subset of size 1 854 670, randomly chosen contracts from the whole population. Because of the size of the data set, it is hard to see that there are extreme values. But consider the QQplot in Figure 2.2. A QQplot(quantile-quantile plot) is a plot of the ordered observations versus the quantiles of a standard normal distribution. If the data are normal, the points should lie close to a line. We notice that there is a very heavy tail to the right side of the distribution.

Figure 2.1: Histogram of incurred amount for bodily injury

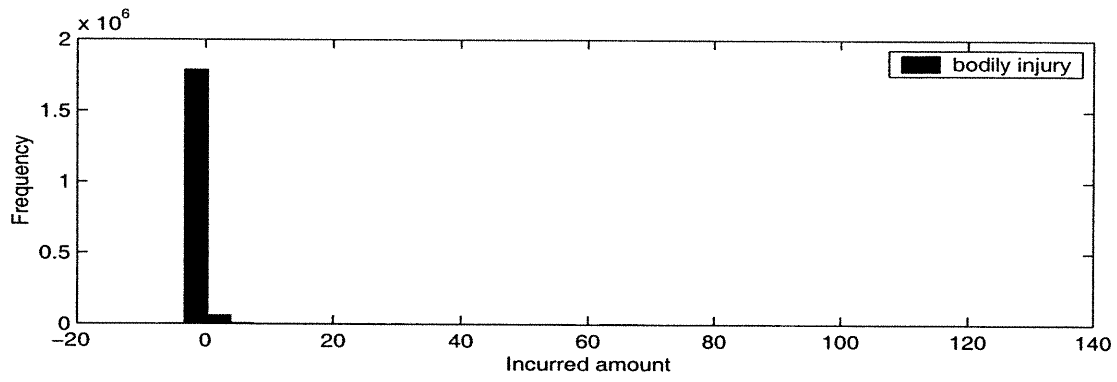
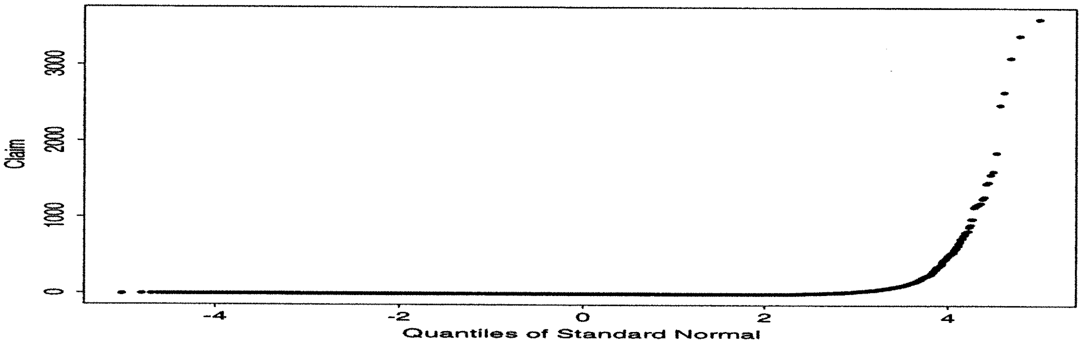


Figure 2.2: QQplot of incurred amount for bodily injury



2.1.3 Data splitting by incurred amount

Due to the peculiar distribution of the incurred amounts, namely mostly zeroes, many small positive amounts, a few large positive amounts, and very few negative values, it turned out that it was a good idea to train separate models on each of those incurred amount "ranges" and to later combine them.

In order to do this, we had to split the data into these four "ranges", whose definition is made a bit tricky as there are 5 different KOL groups (and dependent variables). Specifically, the split was done as follows:

- The **negative** category contains all policies that have at least one negative incurred amount in one of the KOL groups, and all others are 0.
- The **zero** category contains all policies where incurred amounts in all KOL groups are 0.
- The **small positive** category contains policies that have at least a positive incurred amount in one KOL group, but the incurred amounts in all KOL groups are all less than 10 000$.
- The **large positive** category contains all policies that have at least one incurred amount in one KOL group over 10 000$.

The proportion in each category is approximately as follows:

$$\text{Proportion in the category} = \begin{cases} 83.1\% & \text{Zero category} \\ 0.03\% & \text{Negative category} \\ 16.3\% & \text{Small category} \\ 0.57\% & \text{Large category} \end{cases}$$

2.1.4 Training, validation and test set

We previously mentioned that the basic criteria to evaluate candidate models are *minimum MSE* and *Fairness.* Is a fair model with low MSE on a sample which was used to estimate the parameters of the model a good model? Most of the time, we cannot say yes since we do not know the out of sample performance yet. Unfortunately, the out of sample performance, that is the MSE measured outside the sample, would be much worse as the selected model is sometimes far from the true model. This means that we cannot rely on the MSE that was obtained from the sample used to fit the model to choose the best model. We need one more sample to estimate the true MSE.

On the other hand, any class of models has variants inside the class, for example which subset of variables to use. It can be shown that the more variables are used, the more complex the model will be, the more examples (cases) it can fit perfectly (or very well). In statistical learning terminology, the model is said to have more capacity. Too much capacity usually leads to overfitting, i.e., very good estimated MSE within the sample, but a large MSE out of sample. We really need another sample to control the capacity among variants within a class of models.

Therefore, the whole population is split between training set, validation set and test set. The training set is used to estimate the parameters of a particular model in the class, the validation set, a different subset of the data not used to fit the model, is used to select a particular model in the class of models, and the test set, the observations of the data set not used so far, is then used for a final unbiased estimate of performance.

Half of the preprocessed data goes in the training set (3 709 341 policies), a quarter goes in the validation set (1 854 670 policies) and another quarter goes in the test set (1 854 670 policies).

2.2 Model description

As the result of the data preprocessing step previously mentioned, we already obtained a set of size 7 418 681 numerical information of 32 explanatory variables and five dependent variables, *bodily injury*, *property damage*, *accident death*, *collision L.O.U.*, and *comprehensive.* We wish to use these data to develop an empirical predictive model for premiums with low bias and low variance.

We define X to be the $n \times p$ matrix of the form:

$$X_{n \times p} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ & \cdots\cdots & \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

where n is the sample size and p is the number of explanatory variables. Let

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{ip}) \qquad i = 1, 2, \cdots, n \tag{2.1}$$

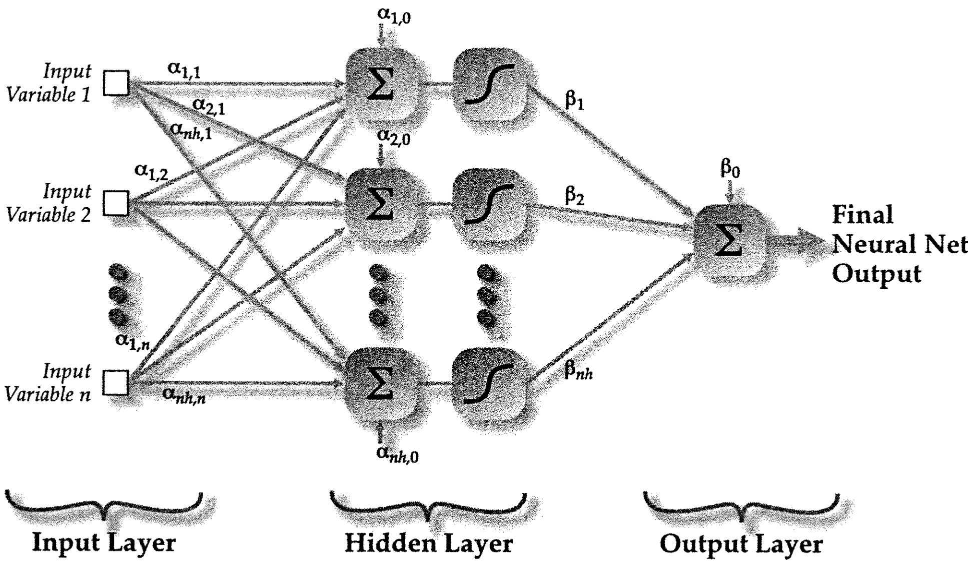
represent the i^{th} vector of explanatory variables associated with i^{th} policy from a sample S, and define

$$Y_{n \times 1} = (y_1, y_2, \cdots, y_n)' \tag{2.2}$$

to be the vector of size n representing the incurred claim amounts for one dependent variable and y_i is the value of the variable associated with the i^{th} policy in the same sample S.

The above notations will be used throughout the following chapters.

Figure 2.3: Neural network



2.2.1 Neural network

Neural networks constitute a large class of learning algorithms remotely inspired by the way the brain computes (Bishop, 1998). They are best known for representing non-parametric models made of non-linear smooth functions. An ordinary neural network consists of the clever combination and simultaneous training of a group of units or neurons that are individually quite simple.

Figure 2.3 illustrates a general ordinary neural network.

Initially, the values of each input variable are linearly combined. Each hidden unit receives a different linear combination of the variables. Within each unit, a non-linear transformation (called a transfer function) is applied to the linear combination. Finally, the results of all hidden units are linearly combined and sent as the output of the network. There could be more than one hidden layer, but for most applications a single one suffices. The output of a neural network can be written in the following form:

$$\hat{Y} = (tanh\,(X\hat{\alpha}))'\,\hat{\beta} \tag{2.3}$$

i.e.,

$$\hat{y}(x_i) = \hat{\beta}_0 + \sum_{j=1}^{n_h} \hat{\beta}_j \cdot \tanh\left(\hat{\alpha}_{j,0} + \sum_{k=1}^{n} \hat{\alpha}_{j,k} x_{ik}\right) \qquad i = 1, 2, \cdots, n \tag{2.4}$$

where $\hat{\beta}$ is a vector parameters with $n_h + 1$ elements, $\hat{\alpha}$ is an $n \times n_h$ matrix of parameters, n_h is the number of hidden units, and the function $\tanh(\cdot)$, the hyperbolic tangent, is the non-linear transfer function.

The procedure to estimate the parameters of a neural network is similar to that for a *Linear model*, but with more complexity and is usually performed by a specific neural network algorithm (Hastie, Tibshirani & Friedman, 2001). Like a linear regression, this model can potentially yield negative premiums in some cases.

A new type of neural network, called softplus neural network, was recently introduced by Dugas et al. (2001) and includes a softplus unit as the final transfer function to ensure that the predicted premium is positive. The output of a softplus neural network can be expressed as follows:

$$\hat{Y} = F\left((\tanh\,(X\hat{\alpha}))'\,\hat{\beta}\right) \tag{2.5}$$

i.e.,

$$\hat{y}(x_i) = F\left(\hat{\beta}_0 + \sum_{j=1}^{n_h} \hat{\beta}_j \cdot \tanh\left(\hat{\alpha}_{j,0} + \sum_{k=1}^{n} \hat{\alpha}_{j,k} x_{ik}\right)\right) \qquad i = 1, 2, \cdots, n \tag{2.6}$$

where $F(\cdot)$ is the softplus function which is actually simply the primitive (integral) function of the *sigmoid* function, that is:

$$\begin{aligned} F(a) &= \log(1+e^a) \\ \text{and} \qquad sigmoid(a) &= \frac{1}{1+\exp(-a)} \end{aligned}$$

The softplus function is convex and monotone increasing with respect to its input and is always strictly positive.

2.2.2 Mixture Models

Mixture models are created by splitting the regression task in a two stage process. The idea is that "large claims" are much less predictable than "small claims" and furthermore because of the heavy tails of their distribution, they would make the estimation of the premium for small claims less stable. In a first stage we thus train separate sub-models (that we call *experts*) for the records associated with different groups of claims grouped by the level of incurred amount, as explained in Section 2.1.3. In the second stage, we have to predict to which group a record belongs, in order to combine the predictions from each of the experts.

One of the advantages of the mixture models proposed here is that each expert can be trained with less data. In fact, no expert needs to be trained for the special group corresponding to zero incurred amount which represents about 83% of the records!

We can write the prediction for the mixture model as follows:

$$\hat{y}(x_i) = \sum_{\mathcal{G}} \hat{y}_{\mathcal{G}}(x_i) P[\mathcal{G}|x_i] \tag{2.7}$$

where $\hat{y}_{\mathcal{G}}(x_i)$ is the prediction associated with the i^{th} policy x_i from a sub-model trained only on data belonging to group $\mathcal{G}$.

This type of model is sometimes called a conditional mixture or a *mixture of experts* (Jacobs, Jordan, Nowlan & Hinton, 1979).

We devise a reliable probabilistic classifier to estimate the conditional class probabilities ($P[\mathcal{G}|\mathcal{X}]$),

$$\mathcal{G} = \begin{cases} 1, & \text{if} \quad A < 0 \qquad\qquad \text{(negative incurred amount)} \\ 2, & \text{if} \quad A = 0 \qquad\qquad \text{(no claim)} \\ 3, & \text{if} \quad 0 < A < \theta \qquad \text{(small incurred amount)} \\ 4, & \text{if} \quad A > \theta \qquad\qquad \text{(large incurred amount)} \end{cases} \tag{2.8}$$

In practice the split is a bit more complex because A is a vector with the values of the incurred amounts for 5 different dependent variables. The actual split was described in Section 2.1.3.

The value of the threshold θ was set to 10 000$ so that roughly 50% of the total expenses were incurred in class 4 and the remaining 50% in the other classes. Negative amounts are due to accounting adjustments often related to accidents incurred in previous years, however their impact is negligible because there are very few of them and the values are very small. In the best performing model, we have actually set their posterior probability to zero, without any measurable deterioration in MSE, thus guaranteeing that all premiums predicted by the model are positive.

For the sub-models, i.e., the *experts*, any of the models mentioned in the beginning of this chapter could be used. However, for the probabilistic classifier, we need a different kind of model, one that can estimate the *class probabilities*, i.e., $P[\mathcal{G} = g|x_i]$. Since $\mathcal{G}$ can take 4 values, the output of the probabilistic classifier consists of 4 probabilities (that always sum to 1). For this purpose we have considered two alternatives: a *constant model* and a *neural network model.*

Constant Class Probability Model

This model assigns a constant probability $P[\mathcal{G} = g|x_i] = p_g$ to each of the 4 amount levels ($g = 1$ to 4). These probabilities are trivially estimated from the average proportions of cases for which $\mathcal{G}$=g ($g = 1$ to 4). This is a reference model.

Neural Network Class Probability Model

This model assigns a probability $P[\mathcal{G} = g|x_i] = p_g(x_i)$ to each of the 4 amount levels ($g = 1$ to 4) according to x_i using a neural network formula (2.4). These $p_g(x_i)$ were required to satisfy:

$$\sum_{g=1}^{4} p_g(x_i) = 1$$

and $p_g(x_i) \geq 0$ for these to represent probabilities. This is achieved by using on the last layer of the neural network a *softmax* transformation which exactly gives us these guarantees. The calculation performed by the neural network is the following:

$$p_g(x_i) \quad = \quad \frac{e^{\hat{y}_g(x_i)}}{\sum_{j=1}^{4} e^{\hat{y}_j(x_i)}} \tag{2.9}$$

where

$$\hat{y}_g(x_i) = \hat{\gamma}_g + \sum_{j=1}^{n_h} \hat{\beta}_{gj} \tanh\left(\hat{\alpha}_{j,0} + \sum_{k=1}^{n} \hat{\alpha}_{j,k} x_{ik}\right) \tag{2.10}$$

and the different $\hat{\gamma}$, $\hat{\beta}$, and $\hat{\alpha}$ are the estimated parameters.

We don't want to train this model to minimize a MSE, but instead we want it to maximize the log-likelihood of the correct class $\mathcal{G}$ given x_i. It means that we are looking for the model that computes probabilities that are as "close" as possible to the "true" probabilities, where closeness is defined in the information-theoretic sense (by the so-called Kullback-Leibler divergence). The parameters γ, β, and α in the above formula are thus obtained by maximizing the log-likelihood criterion:

$$\sum_i \log p_{g_i}(x_i)$$

where the sum runs over the records in the training set, with policy x_i, and g_i is the associated amount level group which depends on the level of the incurred claim amount, as discussed above and in Section 2.1.3. The optimization procedure is similar to that of the *ordinary neural network* and *softplus neural network*, and is based on stochastic gradient descent.

The mixture model can produce negative premiums if and only if the experts can do so. In the model that was finally selected, the experts are softplus neural networks, so the premiums are always positive.

2.3 Computational aspects

Because our data sets are extremely large, it can be quite difficult to use ordinary statistical software, such as Splus. The main problem is memory management, and the need to do some sophisticated modelling (sophisticated at least from a computational point of view), such as neural network.

Also, we need to bootstrap these large data sets about a thousand times to get some of the necessary results. So, we have to consider the computational time, which should be acceptable and as small as possible. The set of programs PLearn has been created to solve these problems.

PLearn is a C++ library, developed by Pascal Vincent and Yoshua Bengio from the LISA lab of Université de Montréal. It uses the object-oriented and operator overloading capabilities of the C++ language to allow library users to express their functions and their optimization as a standard C++ program.

PLearn incorporates many design advantages for large data sets. The most exciting one is its memory allocation design implemented by *VMat*. All computations in this thesis, including bootstrapping, are done in PLearn through the bridge *VMat*. This has saved us time and energy. Without PLearn, it is hard to imagine where and how we could have obtained our results.

Chapter 3

Bootstrap methodology

This chapter introduces some theoretical background and methodology surrounding bootstrap procedures.

3.1 Introduction

Efron (1979) introduced the bootstrap and established a new framework for simulation-based statistical analysis. It enables statisticians to pull more information out of data than any other previously developed statistical method.

The bootstrap addresses the following problem in statistics: how to infer the truth from a finite sample data that is by no means complete. More specifically, it is a technique for estimating sampling distributions conditional on the observed data and generally for use in computing confidence intervals and for making tests of significance.

The basic idea of the bootstrap (Efron & Tibshirani, 1993) or (Davison & Hinkley, 1999), consists of resampling with replacement B times from the observed data bootstrap samples of the same size as the observed data set, to compute the statistic of interest on each bootstrap sample, and then to consider these statistics as a collection of possible values of the statistic of interest. So the essential of the bootstrap consists of using the empirical distribution of the bootstrap replications of a statistic in lieu of the actual distribution of the statistic. In this thesis, the statistic on which the bootstrap will be applied is the mean.

3.2 Bootstrap sampling

Suppose we have an observed data set $x = (x_1, x_2, \cdots, x_n)$, from an unknown probability distribution F, i.e.,

$$x_1, x_2, \cdots, x_n \quad \overset{i.i.d}{\sim} \quad F,$$

we then generate a bootstrap sample $x^* = (x_1^*, x_2^*, \cdots, x_n^*)$ of size n by randomly drawing observations x_i^* with replacement from the empirical distribution $\hat{F}$ of the data,

$$x_1^*, x_2^*, \cdots, x_n^* \quad \overset{i.i.d}{\sim} \quad \hat{F}. \tag{3.1}$$

The empirical distribution function $\hat{F}$ is a simple estimate of the distribution F and is defined to be the discrete distribution that puts probability $1/n$ on each value x_i. In other words, $\hat{F}$ assigns to a set A in the sample space of x its empirical probability

$$Prob_{\hat{F}}\{A\} \quad = \quad \#\{x_i \in A\}\Big/n.$$

The hat symbol indicates quantities calculated from the observed data. We can thus obtain a large number of independent bootstrap samples each of size n,

$$x^{*1}, x^{*2}, \cdots, x^{*B} \tag{3.2}$$

where B is around 1000 and referred to as the number of bootstrap samples.

Suppose that the statistic of interest is

$$\hat{\theta} = s(x).$$

Corresponding to a bootstrap sample $x^* = (x_1^*, x_2^*, \cdots, x_n^*)$, the bootstrap replication of $\hat{\theta}$ is

$$\hat{\theta}^* = s(x^*), \tag{3.3}$$

i.e., the value of the statistic $\hat{\theta}$ computed from x^*.

For example, if $\hat{\theta}$ is the sample mean $\hat{\theta} = \bar{x}$, its bootstrap replication $\hat{\theta}^* = s(x^*)$ will be the mean of the bootstrap sample,

$$\hat{\theta}^* = s(x^*) = \sum_{i=1}^{n} x_i^* \Big/ n.$$

For B bootstrap samples $x^{*1}, x^{*2}, \cdots, x^{*B}$, there will be B bootstrap replications of $\hat{\theta}$,

$$\hat{\theta}^{*1}, \hat{\theta}^{*2}, \cdots, \hat{\theta}^{*B}. \tag{3.4}$$

Its empirical distribution is the basis to compute a bootstrap hypothesis test or constructing bootstrap confidence intervals.

3.3 Bootstrap hypothesis test

Generally, the bootstrap samples as constructed above can be used to construct bootstrap confidence intervals as we shall see in section 3.4. However, in the situation of a bootstrap hypothesis test, the bootstrap sampling procedure needs to be adjusted. This is because, for a hypothesis test, the statistic must be computed from a distribution which satisfies the null hypothesis H_0. The empirical distribution $\hat{F}$ is not an appropriate estimate of F for an hypothesis test, since it does not generally obey H_0. Instead, we have to find an estimate of the distribution of the population that obeys H_0. This could easily be done by transforming the empirical distribution $\hat{F}$ to $\hat{F}_{trans}$ so that H_0 is satisfied. The way to transform the empirical distribution $\hat{F}$ into a distribution $\hat{F}_{trans}$ depends on the null hypothesis H_0. In our problem, we will be testing that the mean of a distribution is 0, i.e., $H_0 : \mu_0 = 0$. Since the mean of the sample is $\bar{x}$, subtracting that mean from each observation will lead to a new distribution $\hat{F}_{trans}$ which satisfies H_0, i.e., its mean is 0.

So let $y_i = x_i - \bar{x}$ and $\hat{F}_{trans}$ be the empirical distribution of the y_i's. The bootstrap samples $y^* = (y_1^*, y_2^*, \cdots, y_n^*)$ can thus be generated from $\hat{F}_{trans}$, i.e.,

$$y_1^*, y_2^*, \cdots, y_n^* \overset{i.i.d}{\sim} \hat{F}_{trans}. \tag{3.5}$$

We are assuming that the test statistic $\hat{\theta}$ is such that if the null hypothesis H_0 is true, its value will be close to 0, and if it is not true, then $\hat{\theta}$ will usually be far away from 0, either positive or negative. For instance, to test the hypothesis that the mean of a distribution is μ, the test statistic $\hat{\theta}$ will be $\bar{x} - \mu$ so that $\hat{\theta}$ will be close to zero if H_0 is true. For a two-sided hypothesis test, if the null hypothesis H_0 is not true, we expect to observe large values of the absolute value of $\hat{\theta}$. To quantify how far from 0 is $\hat{\theta}$ from the original sample x, the Achieved Significance Level, abbreviated ASL as in Efron & Tibshirani (1993), is computed as follows. Suppose that we have observed $|\hat{\theta}|$, the ASL is then equal to the probability of observing at least that large a value when the null hypothesis H_0 is true,

$$ASL = Prob_{H_0}\{|\hat{\theta}^*| \geq |\hat{\theta}|\}$$

In practice, the ASL is usually approximated by the Monte Carlo method as follows:

$$\widehat{ASL} = \#\{|\hat{\theta}^{*b}| \geq |\hat{\theta}|\}\Big/B \qquad b = 1, 2, \cdots, B \tag{3.6}$$

where $\hat{\theta}^{*b}$ is the bootstrap replication of $\hat{\theta}$ for the bootstrap sample y^{*b}.

The smaller the value of ASL, the stronger the evidence against H_0, the larger the value of ASL, the less evidence we have against H_0, e.g. given a significance level α, we reject H_0 if ASL is less than α, and do not reject it if ASL is greater than α.

For convenience, in the following chapter we treat $\widehat{ASL}$ as a *bootstrap p-value* for our experiments, i.e.

$$\begin{aligned}\text{bootstrap p-value} &= \widehat{ASL} \\ &= \#\{|\hat{\theta}^{*b}| \geq |\hat{\theta}|\}\Big/B.\end{aligned} \tag{3.7}$$

Note that if the alternative hypothesis is one-sided, then the evidence against H_0 is measured differently. For instance, if the alternative hypothesis $H_A : \mu < \mu_0$ is satisfied, then we expect small (i.e., negative) values of $\hat{\theta}$, and the ASL is $Prob_{H_0}\{\hat{\theta}^* \leq \hat{\theta}\}$. The other one-sided alternative is treated similarly.

3.4 Bootstrap confidence intervals

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter θ. To construct a confidence interval with exact confidence level $1 - \alpha$, we usually need to know the true distribution of θ, say $J_n(x, F)$ where F is the distribution function of the observations. In practice, F is usually unknown, therefore, $J_n(x, F)$ will be replaced by various acceptable estimated versions.

There are several different approaches for constructing confidence intervals for a statistic of interest $\hat{\theta}$ using bootstrap techniques, including the bootstrap-t, bootstrap percentiles, and BCa methods. We begin by introducing the approximate normal confidence interval.

3.4.1 Approximate normal confidence interval

Consider a sample of size n from a population with unknown distribution F,

$$x_1, x_2, \cdots, x_n \quad \overset{i.i.d}{\sim} \quad F$$

From $x_1, x_2, \cdots, x_n$ we get the empirical distribution $\hat{F}$, the estimate of F. Let $\hat{\theta}_{\hat{F}}$ be the estimate of the parameter θ_F, $\hat{se}_{\hat{F}}$ be the estimate of standard error for $\hat{\theta}_{\hat{F}}$. For many statistics, when the sample size n is large enough, the distribution of $\hat{\theta}_{\hat{F}}$ becomes more and more normal, $\hat{\theta} \dot{\sim} N(\theta, \hat{se}^2)$, i.e. $J_n(x, F)$ is approximately equal to the cumulative distribution function of a normal distribution with mean θ_F and variance $\hat{se}_{\hat{F}}^2$. We have,

$$Z = \frac{\hat{\theta}_{\hat{F}} - \theta_F}{\hat{se}_{\hat{F}}} \quad \dot{\sim} \quad N(0, 1) \qquad \text{when} \quad n \longrightarrow \infty \tag{3.8}$$

In the limit, we can obtain the approximate normal confidence interval with coverage probability $1 - 2\alpha$ for θ_F as follows,

$$Prob_F\{\theta_F \in [\hat{\theta}_{\hat{F}} - z^{(1-\alpha)} \cdot \hat{se}_{\hat{F}}, \quad \hat{\theta}_{\hat{F}} - z^{(\alpha)} \cdot \hat{se}_{\hat{F}}]\} = 1 - 2\alpha$$

which can be written as

$$[\hat{\theta}_{\hat{F}} - z^{(1-\alpha)} \cdot \hat{se}_{\hat{F}}, \quad \hat{\theta}_{\hat{F}} - z^{(\alpha)} \cdot \hat{se}_{\hat{F}}] \tag{3.9}$$

where $z^{(\alpha)}$ is the $100 \cdot \alpha$th percentile point of the standard normal distribution $N(0,1)$.

The approximate normal confidence interval will be better the larger the sample size n is. But how large is large? This depends on the particular statistic and the distribution of the observations F. In the case of interest, the mean, this is inextricably linked to how nonnormal F is (Miller, 1985). This can be evaluated by two statistics: the *skewness* and the *kurtosis* of the distribution F. For a fixed sample size (even very large), the nonnormality of F may be so important that the approximate normal confidence interval will not have a confidence level anywhere close to the claimed one.

3.4.2 Bootstrap-t confidence interval

Bootstrap-t methods estimate the distribution of Z defined in (3.8), directly from the data.

This procedure is based on generating B bootstrap replications of Z to build a table of ordered bootstrap replications. This will estimate the distribution of Z rather than assuming that it is normal. Then it picks up the relative percentile points of the distribution of the B ordered bootstrap replications to construct a bootstrap-t confidence interval similar to the approximate normal confidence interval previously introduced for θ_F. This is done by simply replacing the quantiles of the standard normal distribution z^{α} by the corresponding quantiles of the bootstrap versions of Z as follows:

step 1 Generate B bootstrap samples

$$x^{*1}, x^{*2}, \cdots, x^{*B} \tag{3.10}$$

each of size n from $\hat{F}$, as in (3.2).

step 2 Calculate

$$Z^{*b} = \frac{\hat{\theta}^{*b} - \hat{\theta}}{\hat{se}^{*b}} \tag{3.11}$$

for each bootstrap sample, where $\hat{\theta}^{*b}$ is the value of $\hat{\theta}$ for the bootstrap sample x^{*b}, $\hat{se}^{*}(b)$ is the estimate standard error of $\hat{\theta}^{*}$ for the bootstrap sample x_b^*.

step 3 Sort Z^{*b} such that

$$Z^*_{(1)} \leq Z^*_{(2)} \leq \cdots \leq Z^*_{(B)} \tag{3.12}$$

step 4 Estimate the αth percentile $z^{(\alpha)}$ of Z^{*b} from the empirical distribution of $Z^*_{(b)}$ by

$$\#\{Z^*_{(b)} \leq \hat{z}^{(\alpha)}\}\Big/B \quad = \quad \alpha \tag{3.13}$$

Note:

1. if $B \cdot \alpha$ is an integer, the empirical α quantile is $\hat{z}^{(\alpha)} = Z^*_{(B \cdot \alpha)}$ and the empirical $(1-\alpha)$ quantile is $\hat{z}^{(1-\alpha)} = Z^*_{(B - B \cdot \alpha)}$.
For example, if $B = 1000, \alpha = 5\%$, then $\hat{z}^{(\alpha)} = Z^*_{(50)}$, the $50th$ largest value of Z^*_b, and $\hat{z}^{(\alpha)} = Z^*_{(950)}$, the $950th$ largest value of Z^*_b.

2. if $B \cdot \alpha$ is not an integer, let $k = \lfloor (B+1)\alpha \rfloor$, the largest integer which is less than or equal $(B+1)\alpha$, the empirical α quantile is $\hat{z}^{(\alpha)} = Z^*_{(k)}$, the

empirical $(1-\alpha)$ quantile is $\hat{z}^{(1-\alpha)} = Z^*_{(B+1-k)}$.

step 5 Construct the $(1-2\alpha)$ bootstrap-t confidence interval as follows:

$$[\hat{\theta} - \hat{z}^{(1-\alpha)} \cdot \hat{se}, \quad \hat{\theta} - \hat{z}^{(\alpha)} \cdot \hat{se}] \tag{3.14}$$

Of course, in the case of the mean $\hat{\theta} = \bar{x}$ and an estimate of the standard error is easy to obtain: $\hat{se} = s/\sqrt{n}$ where s is the sample standard deviation and n is the sample size. But when $\hat{\theta}$ is a more complicated statistic, its estimated standard error is difficult to obtain. This is a disadvantage of the bootstrap-t procedure. One option is to use the bootstrap itself to compute an estimate of the standard error of the statistic; this leads to a double bootstrap procedure, one level to compute an estimate of the standard error and a second level to compute the quantiles of the bootstrap distribution of the studentized statistic.

3.4.3 Percentile interval

To conveniently describe the bootstrap percentile interval, we first look at the normal interval in another way. Suppose that $\hat{\theta}$ is distributed according to a normal distribution,

$$\hat{\theta} \quad \sim \quad N(\theta, \hat{se}^2)$$

i.e.,

$$\frac{\hat{\theta} - \theta}{\hat{se}} \quad \sim \quad N(0, 1).$$

The standard normal confidence interval for θ based on $\hat{\theta}$ is:

$$[\hat{\theta}_{normal,lo}, \quad \hat{\theta}_{normal,up}] = [\hat{\theta} - z^{(1-\alpha)} \cdot \hat{se}, \quad \hat{\theta} - z^{(\alpha)} \cdot \hat{se}] \tag{3.15}$$

where, $z^{(\alpha)}$ and $z^{(1-\alpha)}$ are the $100 \cdot \alpha$th and $100 \cdot (1-\alpha)$th percentile points of the standard normal distribution $N(0,1)$. This means that $\hat{\theta}_{normal,lo} = \hat{\theta} - z^{(1-\alpha)} \cdot \hat{se}$ and $\hat{\theta}_{normal,up} = \hat{\theta} - z^{(\alpha)} \cdot \hat{se}$ are the $100 \cdot \alpha$th and $100 \cdot (1-\alpha)$th percentiles of the distribution of $\hat{\theta}$, i.e.,

$\hat{\theta}_{normal,lo} = \hat{\theta}^{(\alpha)} = 100 \cdot \alpha^{th}$ percentile of the distribution of $\hat{\theta}$

$\hat{\theta}_{normal,up} = \hat{\theta}^{(1-\alpha)} = 100 \cdot (1-\alpha)^{th}$ percentile of the distribution of $\hat{\theta}$.

Let G be the cumulative distribution function of $\hat{\theta}$, i.e., $N(\hat{\theta}, \hat{se}^2)$ according to our assumption, then (3.15) can be written as:

$$[\hat{\theta}_{normal,lo}, \quad \hat{\theta}_{normal,up}] = [G^{-1}(\alpha), \quad G^{-1}(1-\alpha)]. \tag{3.16}$$

Inspired by this analogy, let $\hat{G}$ be the cumulative distribution function of $\hat{\theta}^*$, when the number of bootstrap samples B is infinite. We define the $1-2\alpha$ ideal bootstrap percentile interval of the parameter θ by the α and $1-\alpha$ percentiles of $\hat{G}$:

$$[\hat{\theta}_{\infty,lo}, \quad \hat{\theta}_{\infty,up}] = [\hat{G}^{-1}(\alpha), \quad \hat{G}^{-1}(1-\alpha)] \tag{3.17}$$

If we define $\hat{\theta}^{*(\alpha)} = \hat{G}^{-1}(\alpha)$ the $100 \cdot \alpha$th percentile of the cumulative distribution of $\hat{\theta}^*$, like in (3.16), we might write the percentile interval as:

$$[\hat{\theta}_{\infty,lo}, \quad \hat{\theta}_{\infty,up}] = [\hat{\theta}^{*(\alpha)}, \quad \hat{\theta}^{*(1-\alpha)}] \tag{3.18}$$

where

$\hat{\theta}^*$ is the bootstrap replication of the statistic $\hat{\theta}$

$\hat{\theta}^{*(\alpha)} = 100 \cdot \alpha^{th}$ percentile of the empirical distribution of $\hat{\theta}^*$

$\hat{\theta}^{*(1-\alpha)} = 100 \cdot (1-\alpha)^{th}$ percentile of the empirical distribution of $\hat{\theta}^*$.

Since we can only take a finite B in practice, the percentile interval (3.18) is computed by taking the α^{th} and $(1-\alpha)^{th}$ quantiles of the empirical distribution function of the bootstrap replications $\hat{\theta}^*$.

3.4.4 BCa interval

In many cases, there exists a monotone transformation $m(\cdot)$ such that on the new scale, $\hat{\phi} = m(\hat{\theta})$ is more closely approximated by a normal distribution than $\hat{\theta}$ is. That is

$$\frac{\hat{\phi} - \phi}{se_\phi} \quad \dot{\sim} \quad N(-z_0, 1) \tag{3.19}$$

where

$$se_\phi \quad = \quad se_{\phi_0}(1 + a(\phi - \phi_0)).$$

and $\phi_0 = m(\theta_0)$, with θ_0 the true value of the parameter. In this case, when $z_0 \neq 0$ and/or $a \neq 0$, the percentile interval will not do well.

The BCa interval is an improved version of the percentile interval. Its endpoints also depend on percentile of the bootstrap distribution. However, two more parameters, $\hat{z}_0$ and $\hat{a}$, bias-correction and acceleration are imported in the BCa interval. They adjust the bias of the estimator $\hat{\theta}$ and correct the bias of the standard error of $\hat{\theta}$ to suit all θ. Specifically, the bias-correction parameter z_0 measures the discrepancy between the median of $\hat{\theta}^*$ and $\hat{\theta}$, in normal units. The acceleration constant a refers to the rate of change of the standard error of $\hat{\theta}$ with respect to the true parameter value θ.

Again, we let $\hat{G}$ be the cumulative distribution function of $\hat{\theta}^*$, the bootstrap replication of the statistic of interest $\hat{\theta}$. The $1-2\alpha$ BCa interval for θ is given by:

$$[\hat{\theta}_{BCa,lo}, \quad \hat{\theta}_{BCa,up}] = [(\hat{G}^{-1}(\alpha_1), \quad \hat{G}^{-1}(\alpha_2)] \tag{3.20}$$

It also depends on the percentiles of the distribution of $\hat{\theta}^*$. If we define $\hat{\theta}^{*(\alpha)} = \hat{G}^{-1}(\alpha)$, (3.20) becomes:

$$[\hat{\theta}_{BCa,lo}, \quad \hat{\theta}_{BCa,up}] = \left[\hat{\theta}^{*(\alpha_1)}, \quad \hat{\theta}^{*(\alpha_2)}\right] \tag{3.21}$$

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(\alpha)}\right)}\right) \tag{3.22}$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(1-\alpha)}\right)}\right) \tag{3.23}$$

$\Phi(\cdot)$ is the standard normal cumulative distribution function, $z^{(\alpha)}$ is the $100 \cdot \alpha^{th}$ percentile point of the standard normal distribution. As for the constants $\hat{z}_0$ and $\hat{a}$, they are the bias-correction and acceleration adjustments.

The calculation for $\hat{z}_0$ is easy. It is based on the proportion of bootstrap replications $\hat{\theta}^{*b}$ less than the original estimate $\hat{\theta}$,

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}^{*b} < \hat{\theta}\}}{B}\right) \tag{3.24}$$

where, $\Phi^{-1}(\cdot)$ is the inverse function of a standard normal cumulative distribution.

There are several ways to calculate the acceleration adjustment $\hat{a}$. One of the simplest way is through the use of delete-one versions of the statistic $\hat{\theta}$ as follows:

$$\hat{a} = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6\{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2\}^{3/2}} \tag{3.25}$$

where, $\hat{\theta}_{(i)}$ is the i^{th} delete-one statistic of $\hat{\theta}$,

$$\hat{\theta}_{(i)} = s(x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_{n-1}, x_n) \tag{3.26}$$

and $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(i)} \Big/ n$.

For the mean, the formula for calculating $\hat{z}_0$ is as follows,

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\bar{x}^{*b} < \bar{x}\}}{B}\right). \tag{3.27}$$

As for the parameter $\hat{a}$, in fact we do not need to actually compute the delete-one statistic, thus considerably speeding up the computations for large n. Indeed, (3.25) is equivalent to,

$$\hat{a} = \frac{\sum_{i=1}^{n}((x_i - \bar{x})/(n-1))^3}{6\{\sum_{i=1}^{n}((x_i - \bar{x})/(n-1))^2\}^{3/2}}. \tag{3.28}$$

To compute the BCa intervals, we sort the B bootstrap replications $\hat{\theta}^{*b}$ of $\hat{\theta}$,

$$\hat{\theta}^{*1}, \hat{\theta}^{*2}, \cdots, \hat{\theta}^{*B}$$

such that

$$\hat{\theta}^{*}_{(1)} \leq \hat{\theta}^{*}_{(2)} \leq \cdots \leq \hat{\theta}^{*}_{(B)}$$

then,

$$\begin{aligned}
\hat{\theta}^{*\alpha_1} &= \hat{\theta}^{*}_{(B\alpha_1)} = B\alpha_1{}^{th} \text{ largest value of } \hat{\theta}^{*b} \text{ for } b = 1, 2, \cdots, B \\
\hat{\theta}^{*\alpha_2} &= \hat{\theta}^{*}_{(B\alpha_2)} = B\alpha_2{}^{th} \text{ largest value of } \hat{\theta}^{*b} \text{ for } b = 1, 2, \cdots, B.
\end{aligned}$$

3.4.5 Some properties of the confidence intervals

Before discussing some properties of the confidence intervals that we have introduced in the previous sections, we begin by discussing the notion of transformation respecting and the accuracy of confidence intervals.

Consider a parameter θ, and $m(\cdot)$, a monotone function. Then $\phi = m(\theta)$, the transformation of θ, is the parameter in the new scale.

We say that a confidence interval is transformation respecting, if the endpoints of the interval for the parameter ϕ are simply those for the confidence interval for θ mapped by $m(\theta)$. That is:

$$\text{parameter:} \qquad \theta \quad \longrightarrow \quad \phi = m(\theta)$$

$$\text{confidence interval:} \qquad [\hat{\theta}_{lo}, \quad \hat{\theta}_{up}] \quad \longrightarrow \quad [m(\hat{\theta}_{lo}), \quad m(\hat{\theta}_{up})]. \qquad (3.29)$$

Now we discuss the accuracy of a confidence interval. Theoretically, a $(1-2\alpha)$ confidence interval should satisfy

$$Prob\{\theta < \hat{\theta}_{lo}\} = Prob\{\theta > \hat{\theta}_{up}\} = \alpha. \qquad (3.30)$$

For an approximate confidence interval, it is impossible to reach (3.30). Therefore, it is necessary to evaluate its accuracy. Here, we introduce two grade levels as in Hall (1992). We say that a confidence interval is *first-order accurate* if

$$Prob\{\theta < \hat{\theta}_{lo}\} \doteq \alpha + O(n^{-1/2}) \quad \text{and} \quad Prob\{\theta > \hat{\theta}_{up}\} \doteq \alpha + O(n^{-1/2}) \qquad (3.31)$$

where n represents the sample size. A confidence interval is *second-order accurate* if

$$Prob\{\theta < \hat{\theta}_{lo}\} \doteq \alpha + O(n^{-1}) \quad \text{and} \quad Prob\{\theta > \hat{\theta}_{up}\} \doteq \alpha + O(n^{-1}). \qquad (3.32)$$

Obviously, the error of a second-order accurate confidence interval goes to zero at a faster rate than a first-order accurate interval in terms of the sample size n.

We now shortly comment on the properties of the confidence intervals that we described. First of all, the approximate normal interval requires the computation of a standard error estimate. It is first-order accurate and is not transformation respecting.

The bootstrap-t method is particularly applicable to location statistics. It has a good theoretical coverage probability. However, it is difficult to perform bootstrap-t procedure when $\hat{\theta}$ is a complicated statistic, for which there is no simple standard error formula and, in small-sample situation, it tends to be erratic by giving intervals which are too wide and fall outside of the allowable range for a parameter. The bootstrap-t method is not transformation respecting, but is second-order accurate.

The bootstrap percentile interval is less erratic in practice, but has less satisfactory coverage properties since it is only first-order accurate. But it is transformation respecting. Its improved version is the bias-corrected and accelerated interval, abbreviated as the BCa interval.

The BCa interval is a substantial improvement over the percentile method in both theory and practice. They come close to the criteria of goodness, i.e., they closely match exact confidence intervals in the special situations where statistical theory yields an exact answer, but they also give dependably accurate coverage probability being second-order accurate, though their coverage accuracy can still be erratic for small sample size. They are transformation respecting. They are recommended for general use in constructing bootstrap confidence interval, especially for nonparametric problems.

In this study, we are going to construct bootstrap-t intervals, BCa intervals, and approximate normal confidence intervals.

Chapter 4

Bonferroni methodology

4.1 Multiple comparison

When there is more than one hypothesis test being carried out in a single study, we are facing the problem of multiple comparisons as in Miller (1980) and Hsu (1996). On top of the probability of making an error for each individual test, there is also the possibility of making an error in the family of tests. We now consider a family of multiple comparisons consisting of k hypothesis tests. Let A_i be the event that an error of type I occurs for the ith hypothesis test. Assume that

$$P(A_i) = \alpha \qquad i = 1, 2, \cdots, k. \tag{4.1}$$

Then $\cup_{i=1}^{k} A_i$ represents the event that at least one hypothesis test is erroneously declared significant among the k tests. We call this the family error. Clearly, the probability of an error of type I for the family will be inflated.

For example, in the case of $k = 10$ independent hypothesis tests in a family, if we set $\alpha = 0.05$, the probability of getting no significant difference among the 10 tests when the null hypothesis is true for each of them is $(1-\alpha)^{10} = 0.95^{10} = 0.5987$. So the *inflated* α for the family is $1-(1-\alpha)^{10} = 0.4013$. That is, there is a 40% chance to make a type I error in the family rather than the "expected" 5%.

Generally, for a family consisting of k independent hypothesis tests, the probability that no significant difference is detected when the null hypothesis is true is $(1-\alpha)^k$ for tests at the level α, and the *inflated* α is:

$$inflated \quad \alpha = 1-(1-\alpha)^k \tag{4.2}$$

To correct for the multiplicity effect, we would like to keep *inflated* α to be 0.05 for the family. To do that, we might adjust the original α for each test downward. If we make α small enough, the probability that none of the k independent tests is significant will be equal to 0.95, i.e. the *inflated* α is 0.05. In that case α is very small, and $(1-\alpha)^k \doteq 1-k\alpha$. Therefore, the approximate formula for the *inflated* α is:

$$inflated \quad \alpha = 1-(1-\alpha)^k \doteq k\alpha. \tag{4.3}$$

So taking

$$\alpha = \frac{inflated \quad \alpha}{k} \tag{4.4}$$

would lead to an approximate *inflated* α. If we hope to reach *inflated* $\alpha \doteq 0.05$, we might choose $\alpha = 0.05/k$. For example, if $k = 10$, we choose $\alpha = 0.05/10 = 0.005$, then, *inflated* $\alpha = 1-(1-\alpha)^{10} \doteq 0.0489$, which is very close to an overall α of 0.05.

This is the simple idea of Bonferroni adjustment procedure for multiple comparisons.

4.2 Bonferroni's inequality

The discussion in the previous section assumed that the hypothesis tests involved in the multiple comparison procedure were independent. But usually they are not since they usually are based on the same subjects (Miller, 1985), and based on different variables which may not be independent. Hence (4.2) can not be applied exactly.

Let $A_1, A_2, \cdots, A_k$ be $k(k \geq 2)$ random events, we have the well-known Boole's formula:

$$P\left(\cup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i) - \sum_{1 \leq i < j \leq k} P(A_i)P(A_j) + \cdots + (-1)^{k-1} P(\cap_{i=1}^{k} A_i) \quad (4.5)$$

by taking the upper bounds $\sum_{i=1}^{k} P(A_i)$ for (4.5), we get the inequality:

$$P(\cup_{i=1}^{k} A_i) \leq \sum_{i=1}^{k} P(A_i) \quad (4.6)$$

equal if and only if $A_1, A_2, \cdots, A_k$ are disjoint. This is known as Bonferroni's inequality.

According to Bonferroni's inequality (4.6), a family consisting of k hypothesis tests will have a probability of type I error of

$$P(\cup_{i=1}^{k} A_i) \leq \sum_{i=1}^{k} P(A_i) = k\alpha \quad (4.7)$$

equal if and only if $A_1, A_2, \cdots, A_k$ are disjoint.

The probability that none of the family members is significant when the null hypothesis is true for all of them becomes $1 - P\left(\cup_{i=1}^{k} A_i\right)$, which satisfies:

$$1 - P(\cup_{i=1}^{k} A_i) \geq 1 - \sum_{i=1}^{k} P(A_i) \quad (4.8)$$

Bonferroni's inequality (4.6) does not take into account the joint probabilities of errors of the hypothesis tests in the family of tests considered. So when the number of members in the family increases, the Bonferroni procedure might generate spurious conclusion. That is to say that the accuracy of Bonferroni's inequality (4.6) is affected by the number of family members. The larger the number of family members is, the larger the difference between the bound of the inequality (4.6) and the actual probability will be. Therefore the accuracy of Bonferroni's adjustment procedure will diminish.

Often, we do have a choice about the number of members that should be treated as a single family when dealing with multiple comparisons using Bonferroni's method.

In general, it is recommended to use less than 20 members in a family.

Suppose that there are k hypothesis tests in a single study, and that we wish to have a family error equal to at most α,

$$P(\cup_{i=1}^{k} A_i) \leq \alpha.$$

Then, using Bonferonni's inequality, we can safely reject a single hypothesis if its p-value is less than α/k, thereby ensuring that the family wise test is significant at most at the α level.

More specifically, for a multiple comparison family of k members, let $P_i \quad (i = 1, 2, \cdots, k)$ be the observed p-value in the i^{th} hypothesis test. At level α, we reject the hypothesis that all null hypotheses in the family are true if,

$$\min\{P_i\} < \alpha/k \qquad i = 1, 2, \cdots, k. \tag{4.9}$$

Alternatively, we can do the same thing by multiplying the observed p-values from the hypothesis tests by the number of family members k, then compare them with family error α directly. For the i^{th} hypothesis test, $k \times$ observed p-value $=$ $k \times P_i \quad i = 1, 2, \cdots, k$, and if

$$\min\{k * P_i, 1\} < \alpha \qquad i = 1, 2, \cdots, k \tag{4.10}$$

the multiple comparison is significant, i.e., we reject the hypothesis that all null hypotheses in the family are true.

According to (4.9) and (4.10), we might define a Bonferroni p-value for the i^{th} hypothesis test in a multiple comparison as follows:

$$\begin{aligned} \text{Bonferroni p-value} &= \min\{k \times \text{observed p-value}, 1\} \\ &= \min\{k \times P_i, 1\} \qquad i = 1, 2, \cdots, k. \end{aligned} \tag{4.11}$$

From the previous discussion, we have seen that Bonferroni's Method corrects the multiplicity effect through adjusting the type I error of an hypothesis test downward proportionally to the number of family members in the multiple comparison. The same approach can be applied for constructing confidence intervals in a multiple comparison procedure.

The Bonferroni method for adjusting confidence intervals in a multiple comparison procedure is to adjust the percentile points of the distribution of the test statistic. We assume that there are k members in the multiple comparison family in the following discussion.

For the approximate normal confidence interval in section 3.4.1, we adjust the percentile point of the standard normal distribution $N(0,1)$ from $z^{(\alpha)}$ to $z^{(\alpha/k)}$. Then (3.9) becomes,

$$[\hat{\theta}_{\hat{F}} - z^{(1-\frac{\alpha}{k})} \cdot \hat{se}_{\hat{F}}, \quad \hat{\theta}_{\hat{F}} - z^{(\frac{\alpha}{k})} \cdot \hat{se}_{\hat{F}}] \tag{4.12}$$

For the bootstrap-t confidence interval of section 3.4.2, we adjust the empirical α percentile point of statistic Z^{*b} in (3.11) from $Z^*_{(B \cdot \alpha)}$ to $Z^*_{(B \cdot \alpha/k)}$. Thus, (3.14) is changed to:

$$[\hat{\theta} - \hat{t}^{(1-\frac{\alpha}{k})} \cdot \hat{se}, \quad \hat{\theta} - \hat{t}^{(\frac{\alpha}{k})} \cdot \hat{se}] \tag{4.13}$$

where, $t^{(\frac{\alpha}{k})} = Z^*_{(\frac{B \cdot \alpha}{k})}$.

For the BCa interval of section 3.4.4, we adjust α for the two parameters α_1 and α_2 defined by (3.22) and (3.23). The formulas are as follows,

$$\alpha_{Bonf1} = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha/k)}}{1 - \hat{a}\,(\hat{z}_0 + z^{(\alpha/k)})}\right) \tag{4.14}$$

$$\alpha_{Bonf2} = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha/k)}}{1 - \hat{a}\,(\hat{z}_0 + z^{(1-\alpha/k)})}\right) \tag{4.15}$$

The BCa interval in (3.21) is adjusted to:

$$[\hat{\theta}_{BCa,lo}, \quad \hat{\theta}_{BCa,up}] = \left[\hat{\theta}^{*(\alpha_{Bonf1})}, \quad \hat{\theta}^{*(\alpha_{Bonf2})}\right] \tag{4.16}$$

Chapter 5

Experimental results

In this chapter, we only work with residuals of the test set based on the mixture model explained in Section 2.2, which was constructed from the train and validation data sets. We will present the experimental results for the test set.

In the first two sections, we examine distributions for incurred claim amounts and residuals, in Section 5.3, subsets are discussed in more details, Section 5.4 and 5.6 introduce hypothesis tests, Sections 5.5 and 5.7 give the results concerning confidence intervals.

5.1 Claim distribution

We first examine the distribution of the dependent variable. Figure 5.1, the QQplot of the distribution of claims for *bodily injury*, shows that most of the claims are small. This means that there are very small incurred claim amounts for bodily injury associated with most of the policies. The heavy tails at the right-end side tell us that some claims are extremely large. We observe that in the test set, a few claims have very small negative values, due to accounting conventions.

The QQplots of the distribution of claims for the other dependent variables of the test set are listed in Appendix B. Similar behaviours were observed.

5.2 Residual distribution

Residuals are of the form:

$$\text{residual of } i^{th} \text{ policy} \;=\; y_i - \hat{y}_i \tag{5.1}$$

Figure 5.2 shows the QQplot of the distribution of residuals from the mixture model for the test set of the variable Bodily injury. The distribution is positively skewed to the right side, like the distribution of claims, again, due to the presence of extremely large claims. Obviously, the distribution is far from normal.

Table 5.1 gives the skewness and kurtosis statistics for the residuals of the different dependent variables from the test set data fitted to the mixture model. If the distributions were normal, the skewness and kurtosis parameters should both be close to 0 (Miller, 1985).

The QQplots of residual distribution for other dependent variables of mixture model test set are placed in Appendix B.

Figure 5.1: QQplot of claim for bodily injury mixture model test set

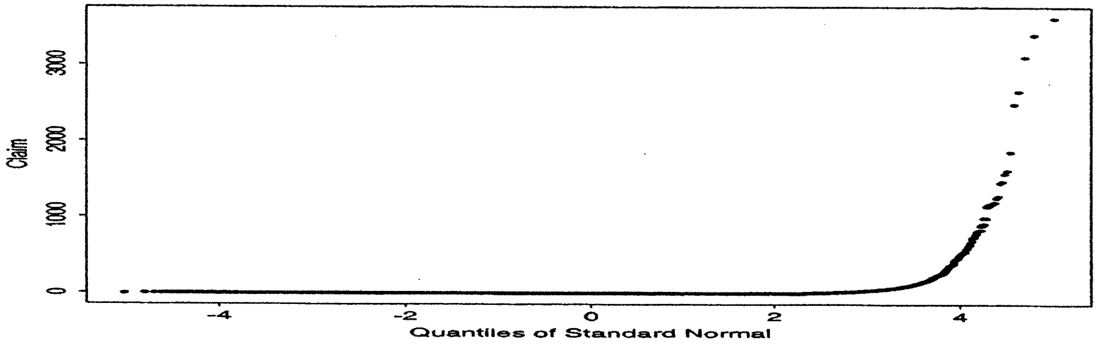


Figure 5.2: QQplot of residual for bodily injury mixture model test set

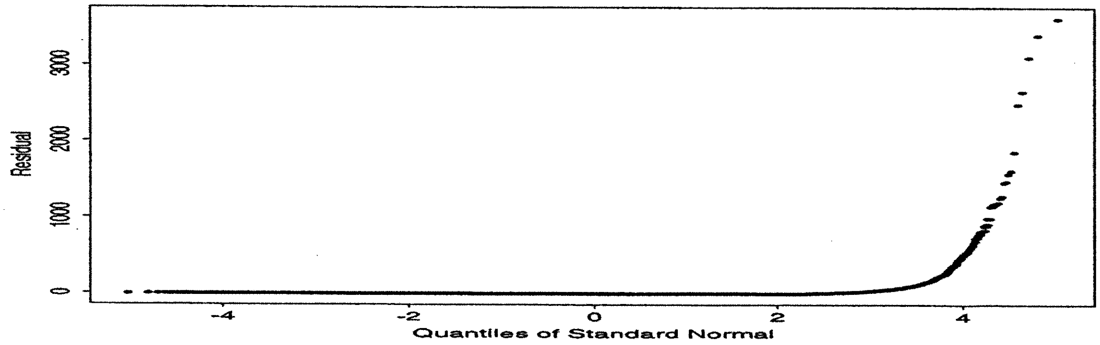


Table 5.1: Skewness and kurtosis statistics for mixture model test set

	skewness	kurtosis
bodily_injury	214	71273
property_damage	29	2749
accident_death	309	152859
collision_lou	14	2333
comprehensive	27	1942
sum	195	61468

5.3 Subsets

5.3.1 Subset definition

Here, we talk in more detail about the subsets which are used for evaluating model fairness in our study.

For a given model, the mean residual could be quite different across different subsets. However, a fair model should such that the mean residual does not vary significantly across different subsets compared with the whole population. This means that the average predictions for a dependent variable within each subset should be statistically close to the average incurred claim amount in that subset for the same dependent variable, i.e. the mean residual of each subset should be close to zero.

There are many ways to split a population into subsets. One might choose explanatory variables to split the data into subsets, such as territory, driver's sex, etc. One might also choose dependent variables to construct subsets. In an ideal situation, the mean residual should be close to zero across all possible subsets.

In this project, we have chosen the predictions of five dependent variables as well as the sum of the five dependent variables to construct different sets of subsets.

We split the population corresponding to the location of the decile of the distribution of predictions. The i^{th} decile of the distribution of predictions is the point immediately above $10i\%$ of the prediction. For example, the 8^{th} decile is the point such that 80% of the predictions are above it. Specifically, the first subset contains the 10% of the customers who are given the lowest predictions by the model, the second subset contains the range 10%-20%, ..., the last subset contains the range

Table 5.2: Subset ranges for bodily injury mixture model test set

	Left end	Right end
subset 1	0.02861	0.0909
subset 2	0.09087	0.1122
subset 3	0.11225	0.1326
subset 4	0.13260	0.1539
subset 5	0.15389	0.1775
subset 6	0.17749	0.2055
subset 7	0.20548	0.2418
subset 8	0.24181	0.2961
subset 9	0.29605	0.3956
subset 10	0.39558	2.9893

90%-100%.

In the thesis, we continue to split the data into the same train, validation, and test subsets that were used in the modelling step of the study. Hence, the test set is of size 1,854,670. Therefore, the ten subsets of the test set is each of size 185,467 exactly.

The subsets ranges obtained from the distribution of predictions of the dependent variable *bodily injury* for the mixture model test set are shown in Table 5.2.

The subset ranges for the other dependent variables as well as their sum are given in Appendix B.

5.3.2 Subset residual distribution

Since we are going to test the fairness of a model through testing that the mean residual of its subsets is zero, it is necessary to examine the distribution of the residuals for each subset.

The QQplots of the distribution, for each subset, of the residuals of the observations of the test set data computed from the mixture model for bodily injury are shown in Figure 5.3. All of the distributions are extremely skewed to the right. This means that the subset residuals have similar distributions to the residuals for bodily injury of the whole test set.

The Qqplots for the subset residuals of the other dependent variables are in Appendix B.

5.4 Normal test

Table 5.3 shows the mean residual for each subset of the mixture model test set. From previous results we have seen that the residual distribution for the whole test set and also for each subset is skewed, so that they are far from the normal distribution. Table 5.4 and Table 5.5 give us the skewness and kurtosis statistics for the whole test set and each subset of the six dependent variables for the mixture model test set.

The test of model fairness for the i^{th} subset can be expressed as follows:

$$H_0: \quad E(\hat{\mu}_i) = 0 \tag{5.2}$$

Figure 5.3: Subset residual distributions for bodily injury mixture model test set

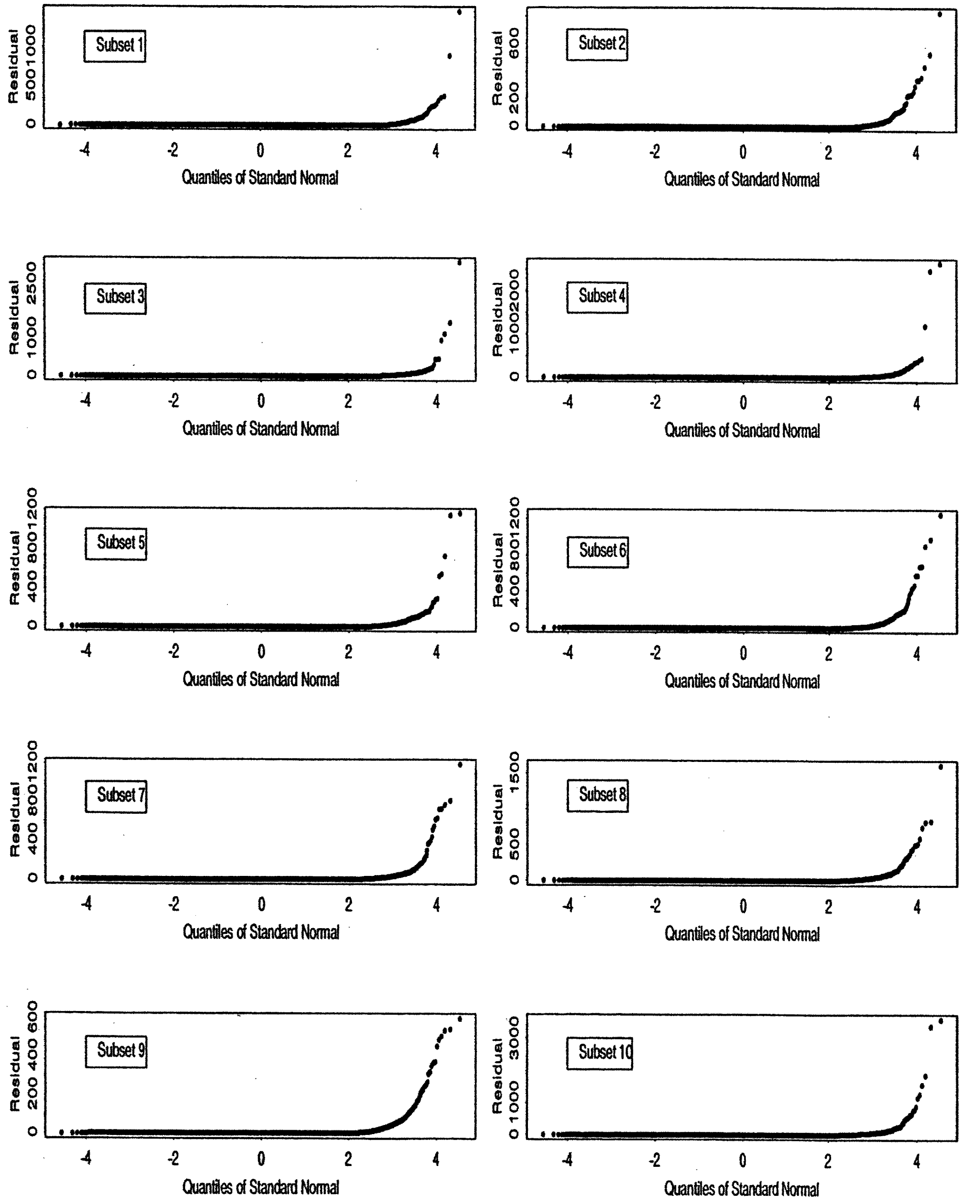


Table 5.3: Mean residuals for mixture model test set

	Bodily	Property	Accident	Collision	Comprehensive	Sum
	Injury	Damage	Death	L.O.U.		
subset 1	0.0074	0.0005	0.0003	-0.0024	-0.0019	0.0147
subset 2	-0.0029	-0.0006	0.0003	-0.0054	-0.0029	-0.0404
subset 3	0.0140	-0.0016	0.0009	-0.0087	-0.0017	0.0088
subset 4	0.0173	0.0007	-0.0007	-0.0059	-0.0002	0.0078
subset 5	-0.0117	-0.0010	-0.0015	-0.0029	-0.0006	0.0050
subset 6	0.0089	-0.0026	0.0100	-0.0018	-0.0032	-0.0260
subset 7	-0.0109	0.0012	-0.0057	-0.0003	-0.0018	0.0038
subset 8	-0.0031	-0.0015	-0.0027	-0.0002	-0.0001	0.0024
subset 9	-0.0403	0.0023	0.0010	0.0016	-0.0020	-0.0026
subset 10	0.0198	0.0039	-0.0069	0.0206	0.0084	0.0100
all	-0.0002	0.0001	-0.0005	-0.0005	-0.0006	-0.0017

The test statistic for the hypothesis H_0 on the i^{th} subset for the test set takes the form:

$$t = \frac{\hat{\mu}_i}{\hat{sd}_i}\sqrt{185467} \tag{5.3}$$

where the standard deviation $\hat{sd}_i$ is defined as follows:

$$\hat{sd}_i = \sqrt{\sum_{j=1}^{185467} \left(y_j - \hat{\mu}_i\right)^2 / 185467} \qquad \text{for test set} \tag{5.4}$$

Since the sample size is large, the normal distribution is used to compute the p-value.

Table 5.6 shows the normal test results for bodily injury of mixture model test set.

We mentioned earlier in Chapter 4 that when there is more than one hypoth-

Table 5.4: Skewness statistics for subsets of mixture model test set

	Bodily	Property	Accident	Collision	Comprehensive	Sum
	Injury	Damage	Death	L.O.U.		
subset 1	202.925	30.739	197.113	0.182	130.317	239.791
subset 2	108.424	22.349	179.518	284.396	51.639	144.380
subset 3	248.135	33.112	130.388	29.249	23.910	110.869
subset 4	225.290	45.321	288.510	3.040	18.707	99.785
subset 5	141.639	44.387	193.311	14.418	20.966	124.155
subset 6	104.104	17.789	250.163	15.226	22.229	102.963
subset 7	99.004	59.675	82.717	13.274	16.816	238.176
subset 8	109.658	25.291	168.708	-1.739	20.078	154.143
subset 9	52.266	18.429	122.263	3.066	23.930	232.815
subset 10	152.636	10.975	104.764	12.100	15.394	168.684
all	214.172	28.677	308.775	14.342	26.836	194.942

Table 5.5: Kurtosis statistics for subsets of mixture model test set

	Bodily	Property	Accident	Collision	Comprehensive	Sum
	Injury	Damage	Death	L.O.U.		
subset 1	54399.7	1667.0	53772.0	-0.98	23551.7	77117.5
subset 2	17278.0	760.7	37035.6	101498.8	5757.0	36429.8
subset 3	75249.5	2366.5	21573.5	2607.7	1073.3	16550.5
subset 4	58571.3	5418.2	101450.0	3118.0	625.0	15135.5
subset 5	27171.1	5273.3	49735.5	358.9	955.3	24035.0
subset 6	14740.5	503.3	70748.5	498.0	1147.0	16373.1
subset 7	13027.6	10240.8	10749.0	257.7	548.8	77106.0
subset 8	18755.9	1499.6	37901.5	1894.4	868.2	34269.2
subset 9	3986.8	747.0	18926.0	2955.9	2540.4	75918.9
subset 10	30978.8	197.3	14430.2	330.0	395.3	37420.3
all	71272.5	2748.8	152858.7	2332.9	1941.5	61468.1

Table 5.6: Normal test for bodily injury of mixture model test set

	Mean residual	$\hat{s.d.}$	t statistic	Normal p-value
subset 1	0.0074	4.8067	0.6622	0.5078
subset 2	-0.0029	3.7882	-0.3312	0.7405
subset 3	0.0140	9.1977	0.6541	0.5131
subset 4	0.0173	9.5340	0.7820	0.4342
subset 5	-0.0117	5.3591	-0.9398	0.3473
subset 6	0.0089	6.5395	0.5846	0.5588
subset 7	-0.0109	6.3609	-0.7358	0.4618
subset 8	-0.0031	7.0470	-0.1889	0.8502
subset 9	-0.0403	5.4024	-3.2129	0.0013
subset 10	-0.0002	15.455	0.5517	0.5812
all	-0.0001	8.0163	-0.0259	0.9793

esis test involved in a single study, it is necessary to adjust for the multiplicity effect, using e.g. the Bonferroni procedure. We are facing exactly this situation here. However, the question is which hypothesis tests should we consider as a single family for which we would like to control the family level? By considering the accuracy of the Bonferroni inequality (4.6), we can not take the entire sixty hypothesis tests as a single family, i.e., six dependent variables (five KOL groups and their sum) for each of ten subsets. Instead, we consider the ten hypothesis tests for one dependent variable as a family, i.e., there are ten family members for each dependent variable.

Table 5.7 presents the p-values of the tests for the mixture model test set. Table 5.8 gives the results of their Bonferroni adjustment. Taking $\alpha = 0.05$, we highlight the significant p-values with a box in Table 5.7 and Table 5.8. For three dependent variables, *bodily injury*, *collision L.O.U.* and *sum*, the normal tests

Table 5.7: Normal p-value for mixture model test set

	Bodily	Property	Accident	Collision	Comprehensive	Sum
	Injury	Damage	Death	L.O.U.		
subset 1	0.5078	0.4803	0.8167	$\boxed{0}$	$\boxed{0}$	0.5759
subset 2	0.7405	0.4692	0.8752	$\boxed{0}$	$\boxed{0}$	$\boxed{0.0035}$
subset 3	0.5131	0.0977	0.6404	$\boxed{0}$	$\boxed{0}$	0.6364
subset 4	0.4342	0.5322	0.8025	$\boxed{0}$	0.7066	0.6482
subset 5	0.3473	0.4110	0.4428	$\boxed{0.0026}$	0.3495	0.7767
subset 6	0.5588	$\boxed{0.0294}$	0.1324	0.1528	$\boxed{0}$	0.0743
subset 7	0.4618	0.4393	$\boxed{0}$	0.8192	$\boxed{0.0448}$	0.8730
subset 8	0.8502	0.3278	0.3488	0.9124	0.9195	0.9144
subset 9	$\boxed{0.0013}$	0.2052	0.7795	0.4618	0.1349	0.9048
subset 10	0.5812	0.0888	$\boxed{0.0492}$	$\boxed{0}$	$\boxed{0}$	0.6708

Table 5.8: Bonferroni normal p-value for mixture model test set

	Bodily	Property	Accident	Collision	Comprehensive	Sum
	Injury	Damage	Death	L.O.U.		
subset 1	1	1	1	$\boxed{0}$	$\boxed{0}$	1
subset 2	1	1	1	$\boxed{0}$	$\boxed{0}$	$\boxed{0.0351}$
subset 3	1	0.9774	1	$\boxed{0}$	$\boxed{0.0002}$	1
subset 4	1	1	1	$\boxed{0}$	1	1
subset 5	1	1	1	$\boxed{0.0263}$	1	1
subset 6	1	0.2935	1	1	$\boxed{0.0001}$	0.7432
subset 7	1	1	$\boxed{0.0002}$	1	0.4476	1
subset 8	1	1	1	1	1	1
subset 9	$\boxed{0.0131}$	1	1	1	1	1
subset 10	1	0.8876	0.4924	$\boxed{0}$	$\boxed{0.0002}$	1

and their Bonferroni adjustment lead to the same qualitative conclusion. For the remaining three dependent variables *property damage*, *accident death* and *comprehensive*, the Bonferroni adjustment leads to one subset becoming non-significant.

We may draw the conclusion that for dependent variables bodily injury, property damage, accident death, and sum over five KOL groups, the mixture model is approximately fair. For collision L.O.U and comprehensive, however, more than half of the subsets are significant; this implies that mixture model gave poor predictions for many subsets of observations.

5.5 Normal confidence interval

Along with point estimators of the mean residuals, we also computed confidence intervals. From (3.9), the $1 - 2\alpha$ normal confidence interval for the mean of the residuals of the i^{th} subset is given by:

$$[\hat{\mu}_i - z^{(1-\alpha)} \cdot \hat{se}_i, \;\; \hat{\mu}_i - z^{(\alpha)} \cdot \hat{se}_i] \tag{5.5}$$

where $\hat{\mu}_i$ is the point estimator of mean residual and $\hat{se}_i$ is its standard error of i^{th} subset of test set, defined in (5.6),

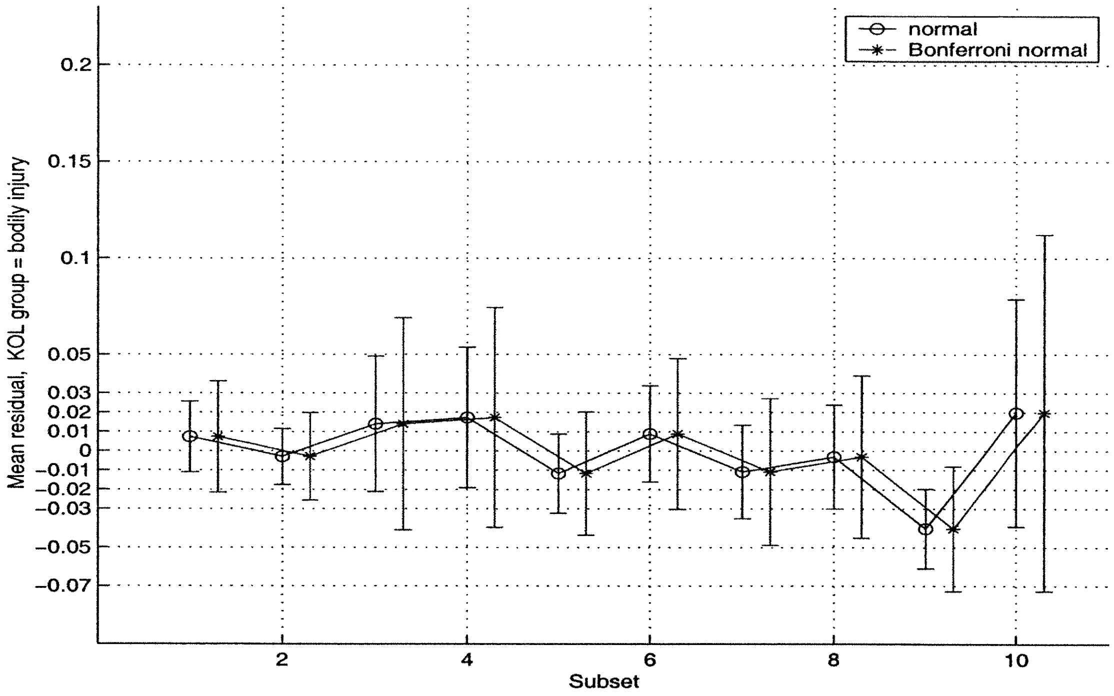
$$\hat{sd}_i = \sqrt{\sum_{j=1}^{185467} (y_j - \hat{\mu}_i)^2 / 185467}$$

$$\hat{se}_i = \hat{sd}_i / \sqrt{185467} \tag{5.6}$$

$z^{(\alpha)}$ is the 100αth percentile point of the standard normal distribution $N(0,1)$.

Table 5.9 shows the 90% normal confidence intervals of bodily injury for the

Figure 5.4: 90% Confidence interval for bodily injury of test set



mixture model test set.

We also report normal confidence intervals based on the Bonferroni procedure. The formula for calculating $1 - 2\alpha$ Bonferroni normal confidence interval for a statistic $\hat{\theta}$ was given in (4.12). For the mean residual of the i^{th} subset, the formula becomes:

$$[\hat{\mu}_i - z^{(1-\frac{\alpha}{10})} \cdot \hat{se}_i,\ \ \widehat{\mu}_i - z^{(\frac{\alpha}{10})} \cdot \hat{se}_i] \tag{5.7}$$

The 90% Bonferroni normal confidence intervals for bodily injury of mixture model test set are listed in Table 5.10.

Figure 5.4 shows the comparison between the normal confidence intervals and the Bonferroni normal confidence intervals for bodily injury of the mixture model test set.

The comparison between the normal confidence intervals and the Bonferroni

Table 5.9: 90% Normal confidence interval for bodily injury of test set

	Mean residual	Left endpoint	Right endpoint
subset 1	0.0074	-0.0110	0.0258
subset 2	-0.0029	-0.0174	0.0116
subset 3	0.0140	-0.0212	0.0491
subset 4	0.0173	-0.0191	0.0537
subset 5	-0.0117	-0.0322	0.0088
subset 6	0.0089	-0.0161	0.0339
subset 7	-0.0109	-0.0352	0.0134
subset 8	-0.0031	-0.0300	0.0238
subset 9	-0.0403	-0.0609	-0.0197
subset 10	-0.0002	-0.0392	0.0788

Table 5.10: 90% Bonferroni normal interval for bodily injury of test set

	Mean residual	Left endpoint	Right endpoint
subset 1	0.0074	-0.0215	0.0361
subset 2	-0.0029	-0.0257	0.0197
subset 3	0.0140	-0.0410	0.0690
subset 4	0.0173	-0.0397	0.0743
subset 5	-0.0117	-0.0437	0.0204
subset 6	0.0089	-0.0302	0.0480
subset 7	-0.0109	-0.0489	0.0272
subset 8	-0.0031	-0.0452	0.0391
subset 9	-0.0403	-0.0726	-0.0080
subset 10	-0.0002	-0.0726	0.1122

confidence intervals for the other dependent variables can be found in Appendix B.

5.6 Bootstrap hypothesis test

We have discussed bootstrap resampling procedures in Section 3.2. For a bootstrap hypothesis test, we need to translate the empirical distribution from $\hat{F}$ to $\hat{F}_{trans}$ such that $\hat{F}_{trans}$ satisfies the null hypothesis. For evaluating model fairness, as mentioned in (5.2), the hypotheses to be tested are,

$$E(\hat{\mu}_i) = 0 \qquad i = 1, 2, \cdots, 10$$

where $\hat{\mu}_i$ is the mean residual of the i^{th} subset.

To construct $\hat{F}_{trans}$, we just need to subtract the observed mean residual $\hat{\mu}_i$ from each residual in the i^{th} subset of size 185467 of the test set,

$$j^{th} \quad \text{central residual} \quad = \quad y_j - \hat{y}_j - \hat{\mu}_i \qquad i = 1, 2, \cdots, 10, \tag{5.8}$$

where, y_j and $\hat{y}_j$ represent the j^{th} observation and prediction in the i^{th} subset, $j = 1, 2, \cdots, 185467$. We resample with replacement from the centred residuals $B = 999$ times for each subset of the test set. We take the size of each bootstrap sample to be the same as the subset, that is 185,467 for the test set.

We compute $\hat{\mu}_i^*$ and $\hat{se}_i^*$ for each subset. The test statistic defined in (5.3), using bootstrap sample b of the i^{th} subset becomes

$$t^{*b} = \frac{\hat{\mu}_i^*}{\hat{se}_i^*}\sqrt{185467}. \tag{5.9}$$

We obtain 999 t^* statistics for each subset,

$$t^{*1}, t^{*2}, \cdots, t^{*999} \tag{5.10}$$

According to the definition of bootstrap p-value in (3.7), the bootstrap p-value for our two sided hypothesis test is:

$$\text{bootstrap p-value} = \#\{|t|^{*b} \geq |t_0|\}\Big/999 \tag{5.11}$$

where t_0 is the absolute value of the observed t statistic of a subset. We compare the 999 absolute values of the above bootstrap t statistics with $|t_0|$ for a subset, and we get the bootstrap p-value for this subset. The bootstrap p-values for the mixture model test set are presented in Table 5.11. The significant p-values are marked with boxes.

Bonferroni bootstrap p-values can be obtained by multiplying the bootstrap p-value with the number of family members and taking the minimum between this number and 1 according to (4.11), i.e.,

$$\text{Bonferroni bootstrap p-value} = \min(10 \times \text{bootstrap p-value}, 1) \tag{5.12}$$

Table 5.12 shows the Bonferroni bootstrap p-values for the mixture model test set. Comparing Table 5.11 and Table 5.12, we observe that the significant results are not changed too much: there is only one significant difference, on subset 2 for the sixth dependent variable sum.

5.7 Comparing normal and bootstrap tests

In this section, we compare the normal and bootstrap hypothesis test procedures. Comparing Table 5.7 and Table 5.11, we can split the sixty p-values into two groups. Group 1 contains combinations of subsets and dependent variables for which the normal p-value and bootstrap p-value are both not significant, or both

Table 5.11: Bootstrap p-value for mixture model test set

	Bodily	Property	Accident	Collision	Comprehensive	Sum
	Injury	Damage	Death	L.O.U.		
subset 1	0.5225	0.5045	0.8338	$\boxed{0}$	$\boxed{0}$	0.6166
subset 2	0.7287	0.4424	0.8699	$\boxed{0}$	$\boxed{0}$	$\boxed{0.0150}$
subset 3	0.5485	0.1051	0.6426	$\boxed{0}$	$\boxed{0}$	0.6236
subset 4	0.4384	0.5415	0.8048	$\boxed{0}$	0.6997	0.6537
subset 5	0.3644	0.4144	0.4595	$\boxed{0.0050}$	0.3604	0.7888
subset 6	0.5506	$\boxed{0.0240}$	0.2102	0.1481	$\boxed{0}$	0.0821
subset 7	0.4394	0.4384	$\boxed{0}$	0.8298	0.0521	0.8899
subset 8	0.8649	0.3233	0.3544	0.8979	0.9209	0.9319
subset 9	$\boxed{0.0020}$	0.1982	0.7678	0.4424	0.1361	0.9229
subset 10	0.6016	0.0871	0.0651	$\boxed{0}$	$\boxed{0}$	0.6747

Table 5.12: Bonferroni bootstrap p-value for mixture model test set

	Bodily	Property	Accident	Collision	Comprehensive	Sum
	Injury	Damage	Death	L.O.U.		
subset 1	1	1	1	$\boxed{0}$	$\boxed{0}$	1
subset 2	1	1	1	$\boxed{0}$	$\boxed{0}$	0.1502
subset 3	1	1	1	$\boxed{0}$	$\boxed{0}$	1
subset 4	1	1	1	$\boxed{0}$	1	1
subset 5	1	1	1	0.0501	1	1
subset 6	1	0.2402	1	1	$\boxed{0}$	0.8208
subset 7	1	1	$\boxed{0}$	1	0.5205	1
subset 8	1	1	1	1	1	1
subset 9	$\boxed{0.0200}$	1	1	1	1	1
subset 10	1	0.8709	0.6507	$\boxed{0}$	$\boxed{0}$	1

zero p-value, which is highly significant. There is a total of 54 p-values in group 1. For the cases in group 1, no matter which hypothesis test we use, it always leads to the same qualitative result.

The other combinations form group 2. Their normal p-values are significant, but not too small. They are subset 9 of bodily injury, subset 6 of property damage, subset 10 of accident death, subset 5 of collision L.O.U, subset 7 of comprehensive and subset 2 for sum with normal p-values of 0.0013, 0.0294, 0.0492, 0.0026, 0.0448, and 0.0035 respectively. Among them, the bootstrap p-values for subset 9 of bodily injury, subset 6 of property damage, subset 5 of collision L.O.U. and subset 7 of comprehensive are similar with normal p-value. However, for subset 10 of accident death, and subset 2 of sum, the bootstrap p-values are quite different from the normal p-values. We list the p-values of subsets in group 2 in Table 5.13.

To investigate what caused the difference between the normal hypothesis test and the bootstrap hypothesis test, we calculate the skewness and kurtosis of the mean of the residuals of the subsets in group 2. The formulas are as follows:

$$Skewness(\text{mean residual}) = Skewness(\text{residual})/\sqrt{185467} \qquad (5.13)$$

$$Kurtosis(\text{mean residual}) = Kurtosis(\text{residual})/185467 \qquad (5.14)$$

Table 5.13: Skewness and Kurtosis for some subsets

	Bodily Injury	Property Damage	Accident Death	Collision L.O.U.	Comprehensive	Sum
subset	9	6	10	5	7	2
N p-value	$\boxed{0.0013}$	$\boxed{0.0294}$	$\boxed{0.0492}$	$\boxed{0.0026}$	$\boxed{0.0448}$	$\boxed{0.0035}$
b p-value	$\boxed{0.0020}$	$\boxed{0.0240}$	0.0651	$\boxed{0.0050}$	0.0521	$\boxed{0.0150}$
skewness	0.1214	0.0413	0.2433	0.0335	0.0390	0.3353
kurtosis	0.0215	0.0027	0.0778	0.0019	0.0030	0.1964

The skewness and kurtosis are also presented in Table 5.13. If the distribution of the mean of the residuals was normal, then both the skewness and the kurtosis of the distribution of the mean would be 0.

Figures 5.5 through Figure 5.10 contain the normal quantile plots of the bootstrap t statistics for those six subsets. A line is added to each plot for the convenience to see how non-normal the distributions of the bootstrap t statistics are.

We notice in Table 5.13 that when the normal p-value and the bootstrap p-value are very similar, the corresponding skewness and kurtosis are not too big, such as subset 6 of property damage, subset 5 of collision L.O.U, and subset 7 of comprehensive. This indicates that the non-normality of the distribution of the mean residuals is not too important, i.e. the approximate normal distribution works well for these subsets. This can be confirmed by Figure 5.5, Figure 5.6 , Figure 5.8 and Figure 5.9. They all approximately show a straight line.

On the other hand, consider subset 10 of accident death and subset 2 of sum in Table 5.13. Their normal p-value and bootstrap p-value are quite different, they are 0.0492 and 0.0651 for subset 10 of accident death, 0.0035 and 0.0150 for subset 2 of sum. These two subsets have a larger value of the skewness and kurtosis of the distribution of the mean, i.e., 0.2433 and 0.0778 for subset 10 of accident death, and 0.3353 and 0.1964 for subset 2 of sum. The corresponding normal quantile plots of the bootstrap t statistics for the two subsets are in Figure 5.7 and Figure 5.10. These two plots are bent up to the left, which means that the distributions are significantly skewed and have longer tails than the normal distribution. Therefore, the bootstrap p-value is more trustworthy than the normal p-value for these two subsets.

Actually, for all subsets in group 2, the bootstrap p-value is more appropriate than the normal p-value since the skewness of these subsets is clearly different from zero (given the sample size), the value corresponding to the normal distribution.

Figure 5.5: QQplot of bootstrap t statistics for bodily injury subset 9

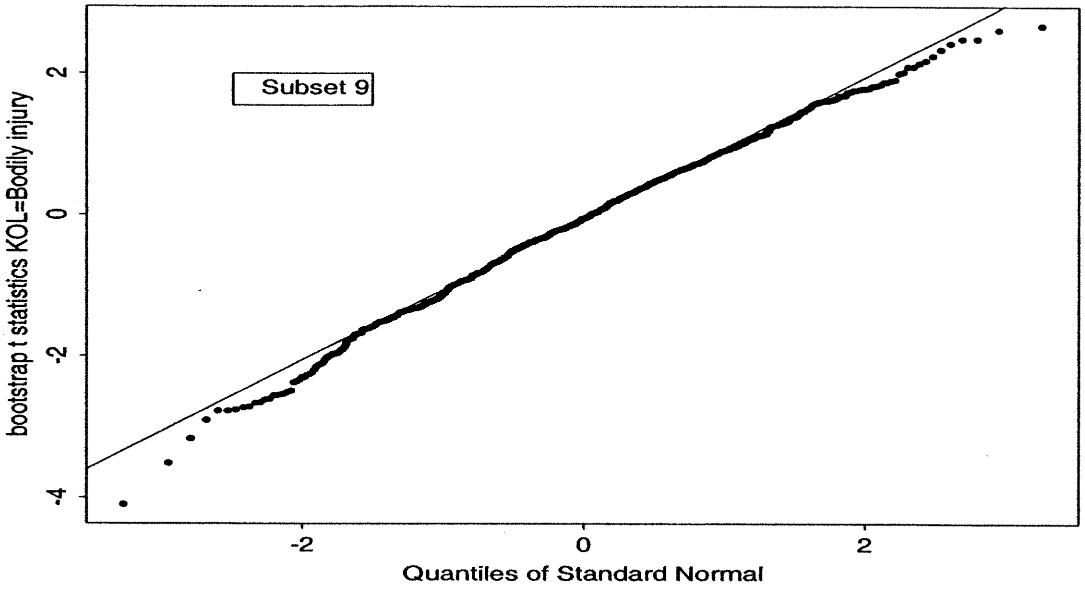


Figure 5.6: QQplot of bootstrap t statistics for property damage subset 6

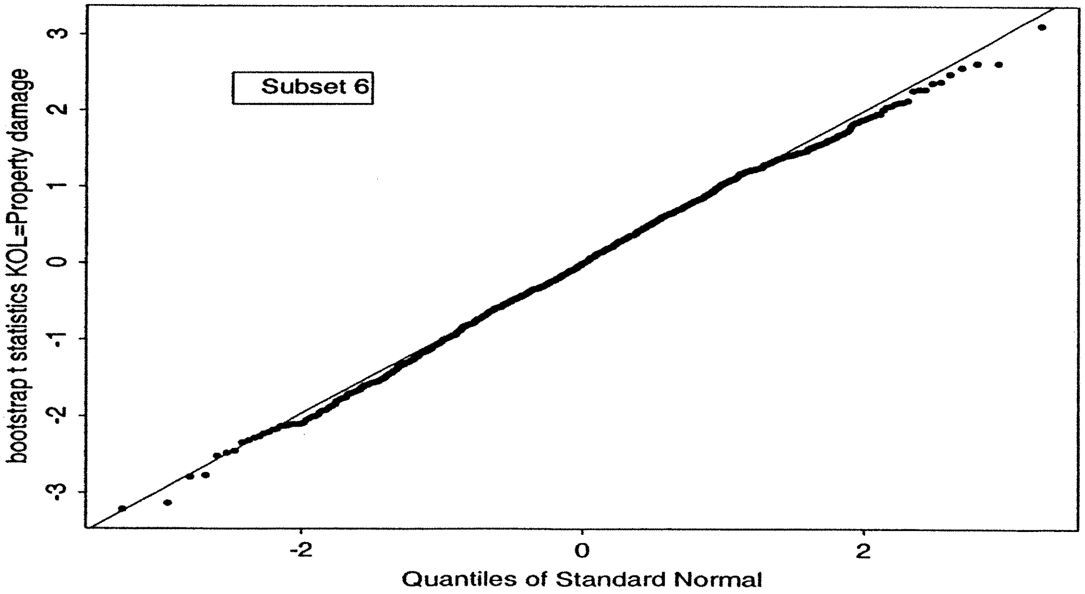


Figure 5.7: QQplot of bootstrap t statistics for accident death subset 10

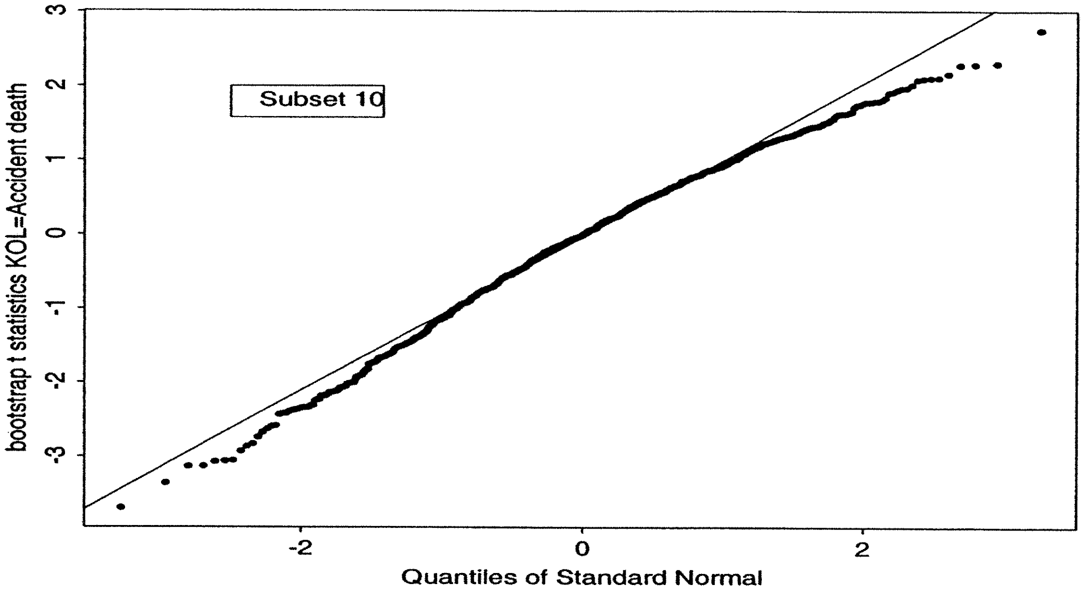


Figure 5.8: QQplot of bootstrap t statistics for collision L.O.U subset 5

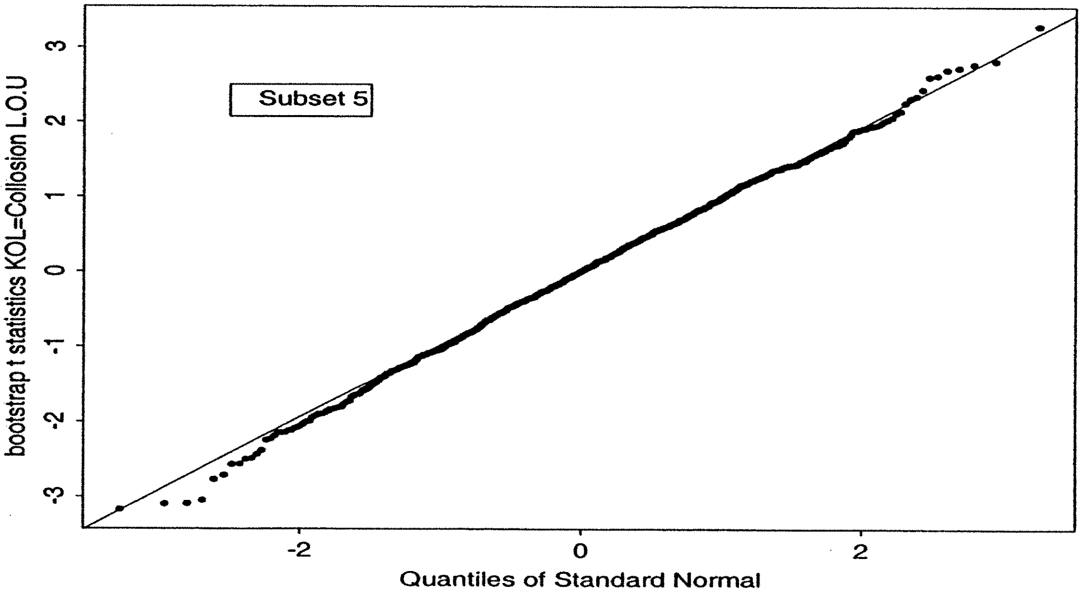


Figure 5.9: QQplot of bootstrap t statistics for comprehensive subset 7

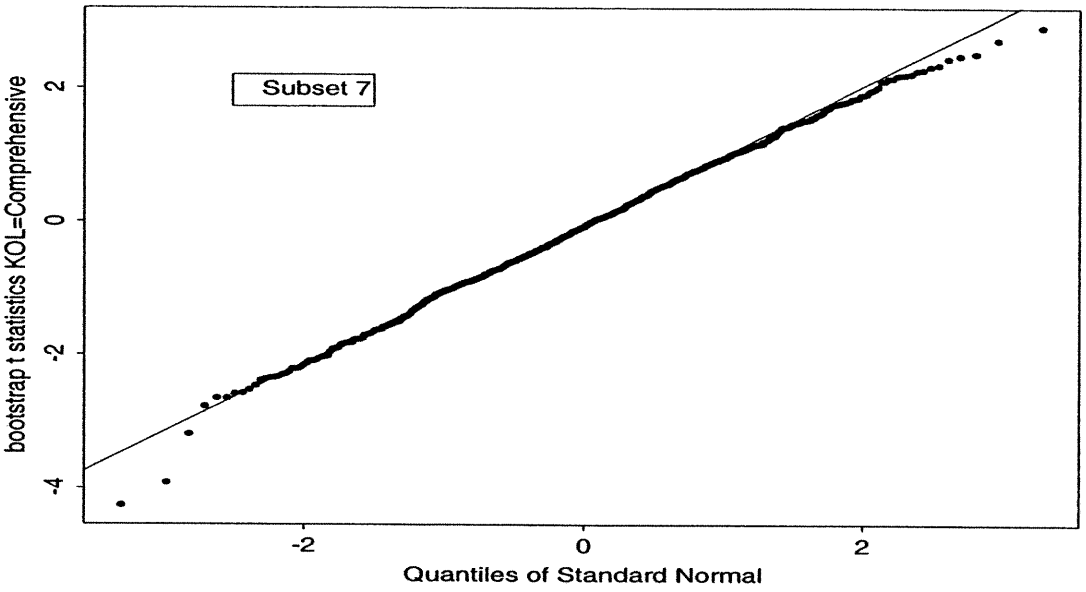
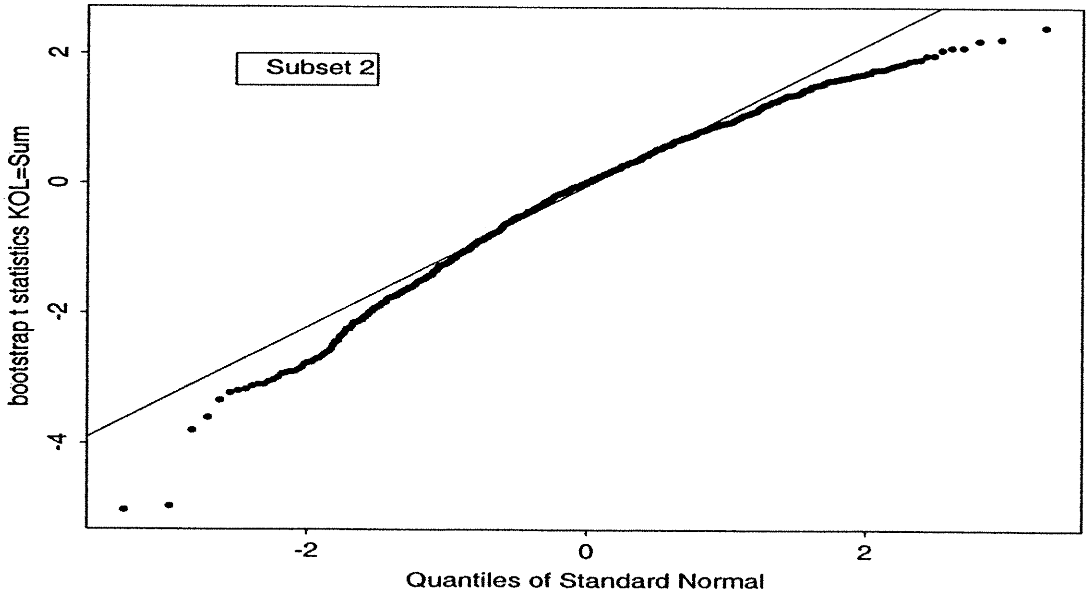


Figure 5.10: QQplot of bootstrap t statistics for sum subset 2



5.8 Bootstrap confidence interval

5.8.1 Bootstrap-t interval

To compute bootstrap-t confidence intervals, the bootstrap samples can be obtained by directly resampling residuals from each subset. Like in bootstrap hypothesis testing, we resample residual $B = 999$ times, each of size 185,467.

To construct $1 - 2\alpha$ bootstrap-t confidence intervals for the mean of the residuals of the i^{th} subset, we first need to calculate bootstrap replications of $\hat{\mu}_i$, and $\hat{se}_i$ for each bootstrap sample. We thus obtain,

$$\hat{\mu}^{*1}, \hat{\mu}^{*2}, \cdots, \hat{\mu}^{*999}$$

$$\hat{se}^{*1}, \hat{se}^{*2}, \cdots, \hat{se}^{*999}.$$

Then we compute the test statistics Z^* defined in (3.11),

$$Z^{*b} = \frac{\hat{\mu}^{*b} - \hat{\mu}_i}{\hat{se}^{*b}}$$

$$Z^{*1}, Z^{*2}, \cdots, Z^{*999}.$$

Ordering the 999 Z^* statistics, we obtain

$$Z_{(*1)} \leqslant Z_{(*2)} \leqslant \cdots \leqslant Z_{(*999)} \tag{5.15}$$

Without resorting to a specific residual distribution for the i^{th} subset, as discussed in Section 3.4.2, by setting $\alpha = 0.05$, the 0.05 and 0.95 percentile points of the Z^{*b} distribution, $z^{(0.05)}$ and $z^{(0.95)}$ are only decided by:

$$k = \lfloor (B+1)\alpha \rfloor = \lfloor 1000 \times 0.05 \rfloor = 50 \tag{5.16}$$

that is:

$$\begin{aligned} z^{(\alpha)} &= z^{(0.05)} &= Z_{(*k)} &= Z_{(*50)} \\ z^{(1-\alpha)} &= z^{(0.95)} &= Z_{(*(B+1-k))} &= Z_{(*950)} \end{aligned}$$

We construct a 90% bootstrap-t confidence interval for the mean of the residuals of the i^{th} subset as follows:

$$(\hat{\mu}_i - Z_{(*950)} \cdot \hat{se}_i, \quad \hat{\mu}_i - Z_{(*50)} \cdot \hat{se}_i) \tag{5.17}$$

where $\hat{\mu}_i$ and $\hat{se}_i$ are the mean residual and its standard error for the i^{th} subset.

The Bonferroni procedure for the bootstrap-t confidence interval is to adjust the α percentile point of Z^{*b} distribution from $z^{(\alpha)}$ to $z^{(\frac{\alpha}{10})}$, 10 is the number of tests that we consider in the family. That is to say the order statistic that must be computed is for

$$k = \lfloor (B+1) \cdot \alpha/10 \rfloor = \lfloor 1000 \times 0.005 \rfloor = 5. \tag{5.18}$$

We thus obtain the $\alpha/10$ and $(1-\alpha/10)$ percentile points of the Z^{*b} distribution:

$$\begin{aligned} z^{(\frac{\alpha}{10})} &= Z_{(k)} &= Z_{(5)} \\ z^{(1-\frac{\alpha}{10})} &= Z_{(B+1-k)} &= Z_{(995)} \end{aligned}$$

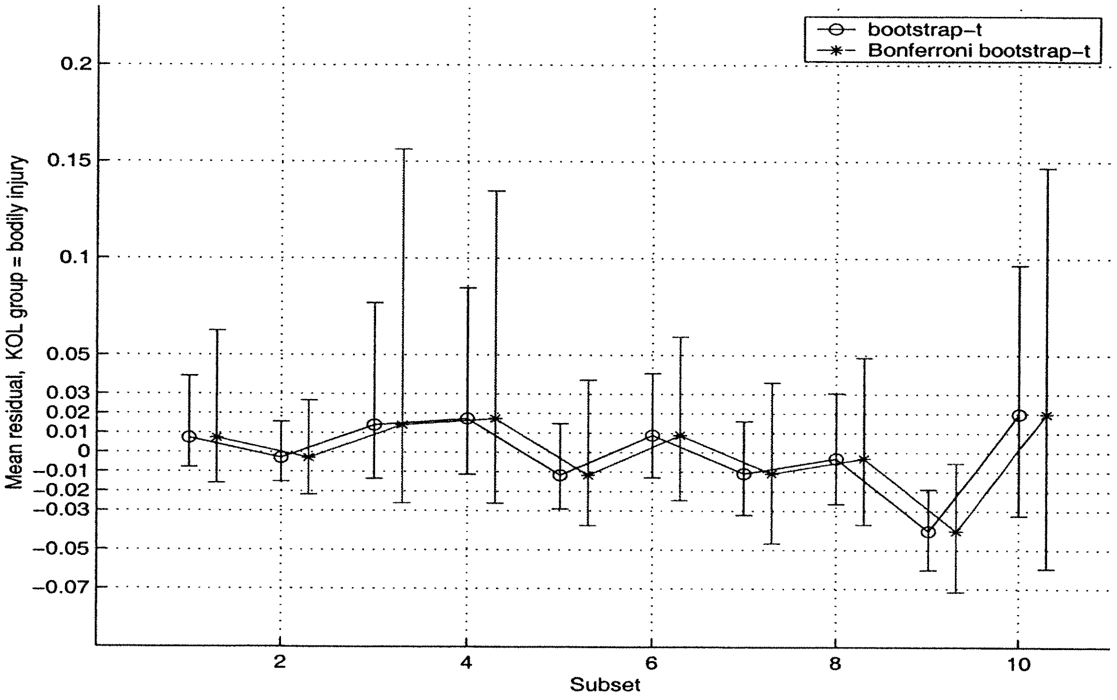
The Bonferroni adjusted 90% bootstrap-t confidence interval is:

$$(\hat{\mu}_i - Z_{(*995)} \cdot \hat{se}_i, \quad \hat{\mu}_i - Z_{(*5)} \cdot \hat{se}_i) \tag{5.19}$$

Table 5.14: 90% bootstrap-t and Bonferroni bootstrap-t of bodily injury test set

		bootstrap-t		Bonferroni bootstrap-t	
	Mean residual	Left end	Right end	Left end	Right end
subset 1	0.0074	-0.0076	0.0390	-0.0157	0.0625
subset 2	-0.0029	-0.0151	0.0157	-0.0217	0.0265
subset 3	0.0140	-0.0135	0.0768	-0.0261	0.1563
subset 4	0.0173	-0.0114	0.0846	-0.0261	0.1348
subset 5	-0.0117	-0.0291	0.0148	-0.0375	0.0372
subset 6	0.0089	-0.0129	0.0407	-0.0246	0.0595
subset 7	-0.0109	-0.0320	0.0160	-0.0467	0.0360
subset 8	-0.0031	-0.0264	0.0304	-0.0371	0.0488
subset 9	-0.0403	-0.0602	-0.0187	-0.0715	-0.0055
subset 10	0.0198	-0.0323	0.0966	-0.0596	0.1470

Figure 5.11: 90% Bootstrap-t and Bonferroni bootstrap-t of bodily injury test set



The resulting bootstrap-t confidence intervals and their Bonferroni adjustments for the bodily injury mixture model test set are presented in Table 5.14.

Figure 5.11 gives the errorbar plot of the confidence intervals in Table 5.14. It shows that the bootstrap-t confidence intervals and Bonferroni bootstrap-t confidence intervals have a similar structure.

The errorbar plots of bootstrap-t and Bonferroni bootstrap-t confidence intervals for the other dependent variables of the mixture model test set listed in Appendix B.

5.8.2 BCa interval

The formulas and parameters for calculating BCa intervals are introduced in Section 3.4.4, e.g., formulas(3.21), (3.27) and (3.25). We are going to construct a $(1-2\alpha)$ BCa interval for mean of the residuals of the i^{th} subset. The statistic $\hat{\theta}$ here is $\hat{\mu}_i$, the mean residual of the i^{th} subset. The BCa bootstrap confidence interval (3.21) becomes

$$(\hat{\mu}_{iBCa,lo}, \quad \hat{\mu}_{iBCa,up}) = (\hat{\mu}_i^{*(\alpha_1)}, \quad \hat{\mu}_i^{*(\alpha_2)}) \tag{5.20}$$

where $\hat{\mu}_i^{*(\alpha_1)}$ represents the α_1th percentile point of the distribution of the bootstrap replicates $\hat{\mu}_i^{*b}$ for $\hat{\mu}_i$. The parameter $\hat{z}_0$ of (3.24) becomes:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\mu}_i^{*b} \geq \hat{\mu}_i\}}{B}\right), \tag{5.21}$$

while (3.22) and (3.23) remain the same:

$$\begin{aligned}\alpha_1 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(\alpha)}\right)}\right)\\ \alpha_2 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(1-\alpha)}\right)}\right)\end{aligned}$$

and (3.25) is now:

$$\hat{a} = \frac{\sum_{j=1}^{185467}((x_j - \hat{\mu}_i)/(185466))^3}{6\{\sum_{j=1}^{185467}((x_j - \hat{\mu}_i)/(185466))^2\}^{3/2}}. \tag{5.22}$$

Table 5.15 shows the 90% BCa for bodily injury of the mixture model test set.

Like for bootstrap-t intervals, the Bonferroni adjustment for the BCa interval consists of transforming the α_1th and α_2th percentile points of the distribution of bootstrap replications for $\hat{\mu}_i$ to α_{Bonf1}th and α_{Bonf2}th percentile points defined in (4.14).

The 90% Bonferroni BCa intervals for bodily injury of the mixture model test set are also listed in Table 5.15.

Figure 5.12 shows the comparison between the the BCa and Bonferoni BCa confidence intervals for the other dependent variables.

Appendix B lists the errorbar plots of BCa and Bonferroni BCa intervals for other dependent variables for the mixture model test set.

5.9 Comparing confidence intervals

In this section, we compare the normal, bootstrap-t and BCa intervals in more detail. Figures 5.13 through 5.18 show the comparison of the three confidence intervals for each of the six dependent variables. Because skewness and kurtosis have a strong effect on the inference procedures, we present in Table 5.16 and

Table 5.15: 90% BCa and Bonferroni BCa for bodily injury of test set

		BCa		Bonferroni BCa	
	Mean residual	Left end	Right end	Left end	Right end
subset 1	0.0074	-0.0062	0.0342	-0.0131	0.0496
subset 2	-0.0029	-0.0163	0.0133	-0.0218	0.0236
subset 3	0.0140	-0.0106	0.0684	-0.0199	0.0851
subset 4	0.0173	-0.0100	0.0705	-0.0206	0.1023
subset 5	-0.0117	-0.0291	0.0135	-0.0371	0.0288
subset 6	0.0089	-0.0119	0.0392	-0.0230	0.0603
subset 7	-0.0109	-0.0316	0.0158	-0.0422	0.0407
subset 8	-0.0031	-0.0267	0.0283	-0.0379	0.0445
subset 9	-0.0403	-0.0583	-0.0172	-0.0692	-0.0048
subset 10	0.0198	-0.0254	0.1003	-0.0455	0.2163

Figure 5.12: 90% BCa and Bonferroni BCa bodily injury of test set

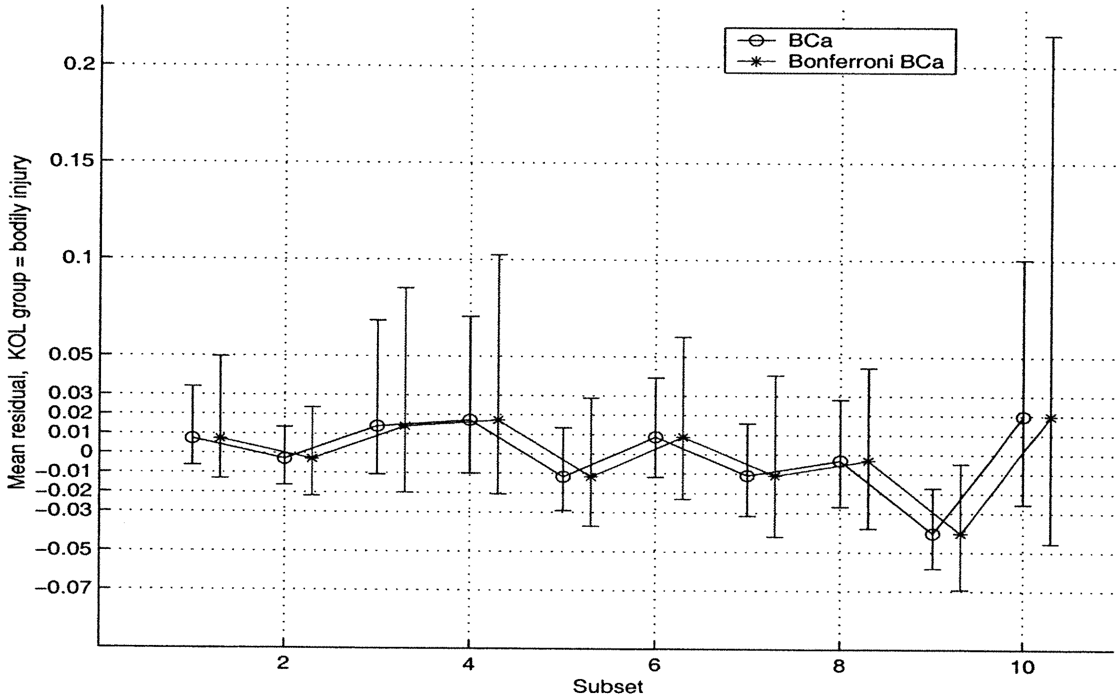


Table 5.17 the skewness and the kurtosis of the distribution of the mean residual for the sixty subsets. The boxed values in these two tables are examples mentioned below.

It is noticeable that the three confidence intervals for most of the sixty subsets are similar, especially when the skewness and the kurtosis of the distribution of the mean are small. For example, subset 6 of property damage has skewness 0.04131 and kurtosis 0.00271, and the normal, bootstrap-t, and BCa intervals are all very similar. This confirms that even though the residual distributions are extremely skewed, the approximate normal distribution for the mean residual which was used to test model fairness is reasonable and acceptable for most of the sixty subsets due to the very large sample size. But when skewness and kurtosis are large, the normal interval is different from the bootstrap-t and BCa intervals. For example, subset 4 of accident death has skewness 0.66993 and kurtosis 0.547, and the three confidence interval are quite different, bootstrap-t interval and BCa interval are skewed strongly, see Figure 5.15.

We also notice that the intervals for bodily injury and sum are much wider than for property damage, accident death, collision L.O.U, and comprehensive, indicating that the variation of the mean residuals in the test set for bodily injury and sum are larger than for the other four dependent variables as illustrated in residual distribution in section 5.3.2. Therefore, the estimated mean residuals for these four dependent variables are more accurate than for bodily injury and sum.

Table 5.16: Skewness of mean residual for subsets of mixture model test set

	Bodily	Property	Accident	Collision	Comprehensive	Sum
	Injury	Damage	Death	L.O.U.		
subset 1	0.4712	0.0714	0.4577	0.0004	0.3026	0.5568
subset 2	0.2518	0.0519	0.4169	0.6604	0.1199	0.3353
subset 3	0.5762	0.0769	0.3028	0.0679	0.0555	0.2574
subset 4	0.5231	0.1052	$\boxed{0.6699}$	0.0071	0.0434	0.2317
subset 5	0.3289	0.1031	0.4489	0.0335	0.0487	0.2883
subset 6	0.2417	$\boxed{0.0413}$	0.5809	0.0354	0.0516	0.2391
subset 7	0.2299	0.1386	0.1921	0.0308	0.0391	0.5531
subset 8	0.2546	0.0587	0.3917	-0.0040	0.0466	0.3579
subset 9	0.1214	0.0428	0.2839	0.0071	0.0556	0.5406
subset 10	0.35444	0.0255	0.2433	0.0281	0.0358	0.3917

Table 5.17: Kurtosis of mear residual for subsets of mixture model test set

	Bodily	Property	Accident	Collision	Comprehensive	Sum
	Injury	Damage	Death	L.O.U.		
subset 1	0.2933	0.0090	0.2899	-5.3e-06	0.1270	0.4158
subset 2	0.0932	0.0041	0.1997	0.5473	0.0310	0.1964
subset 3	0.4057	0.0128	0.1163	0.0141	0.0058	0.0892
subset 4	0.3158	0.0292	$\boxed{0.5470}$	0.0168	0.0034	0.0816
subset 5	0.1465	0.0284	0.2682	0.0019	0.0052	0.1296
subset 6	0.0795	$\boxed{0.0027}$	0.3815	0.0027	0.0062	0.0883
subset 7	0.0702	0.0552	0.0580	0.0014	0.0030	0.4157
subset 8	0.1011	0.0081	0.2044	0.0102	0.0047	0.1848
subset 9	0.0215	0.0040	0.1021	0.0159	0.0137	0.4093
subset 10	0.1670	0.0011	0.0778	0.0018	0.0021	0.2018

Figure 5.13: Comparison of 90% confidence intervals for bodily injury of test set

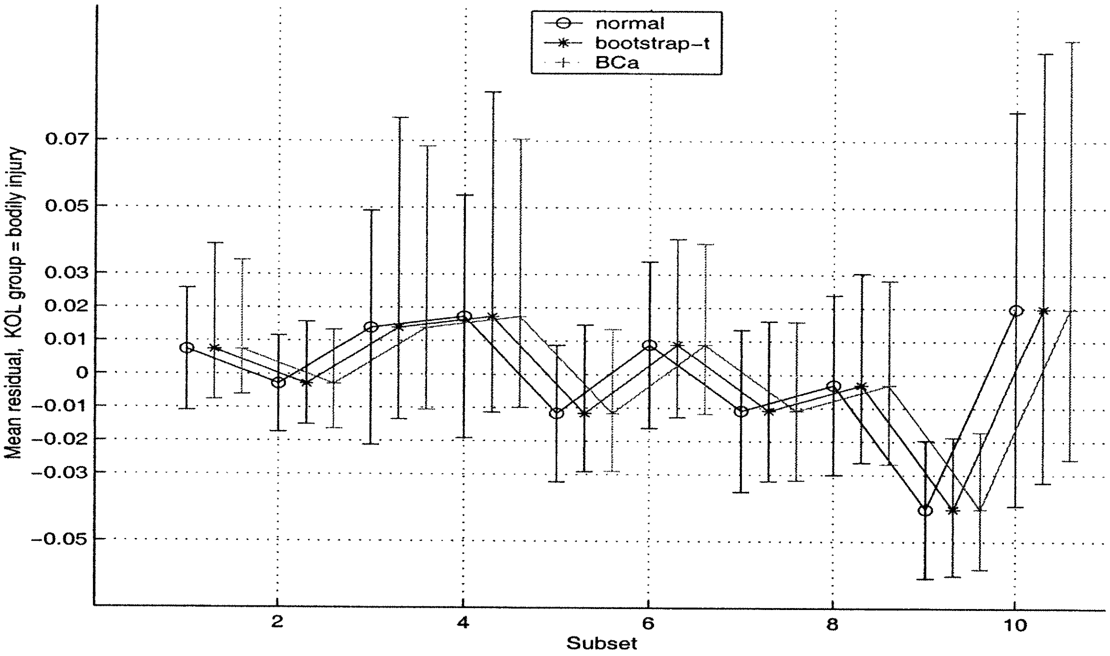


Figure 5.14: Comparison of intervals for property damage mixture test set

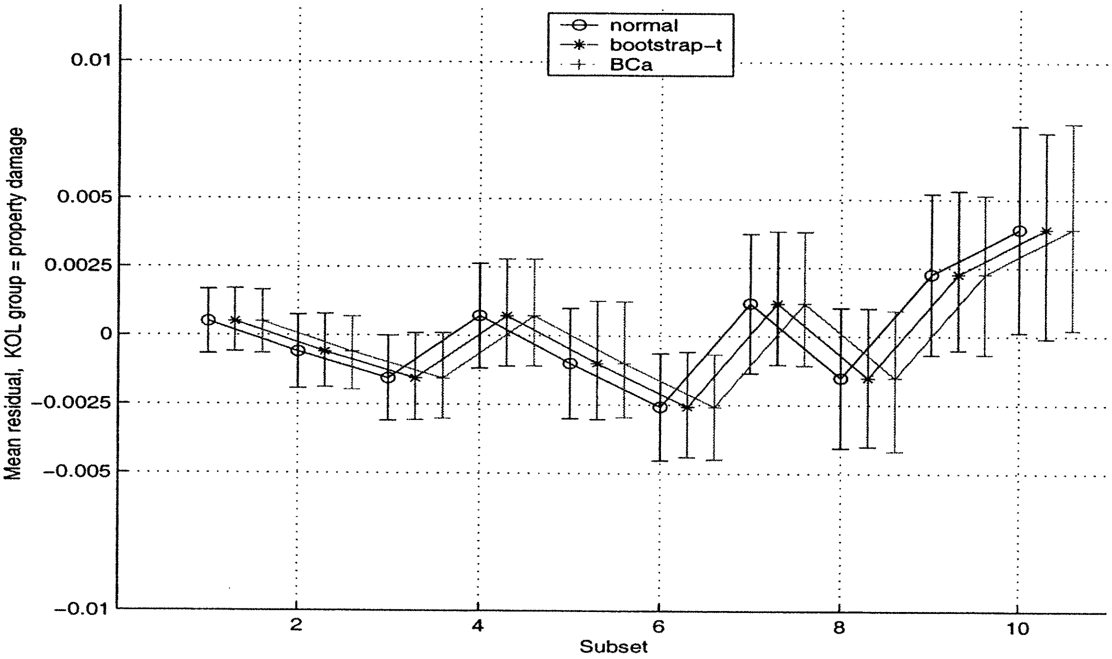


Figure 5.15: Comparison of intervals for accident death mixture test set

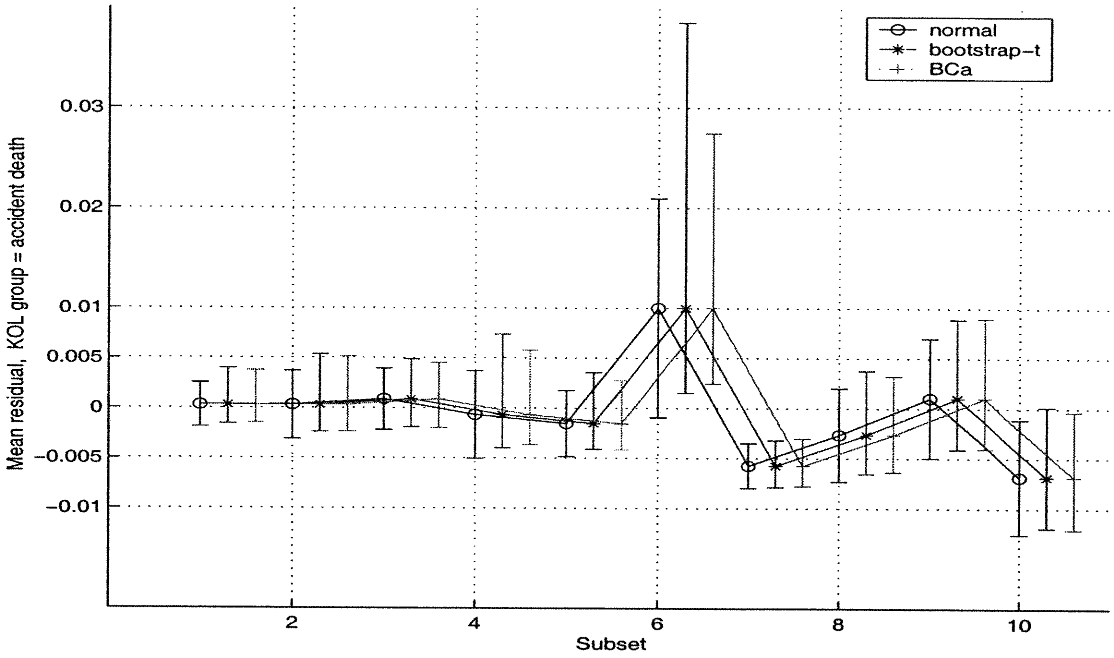


Figure 5.16: Comparison of intervals for collision L.O.U mixture test set

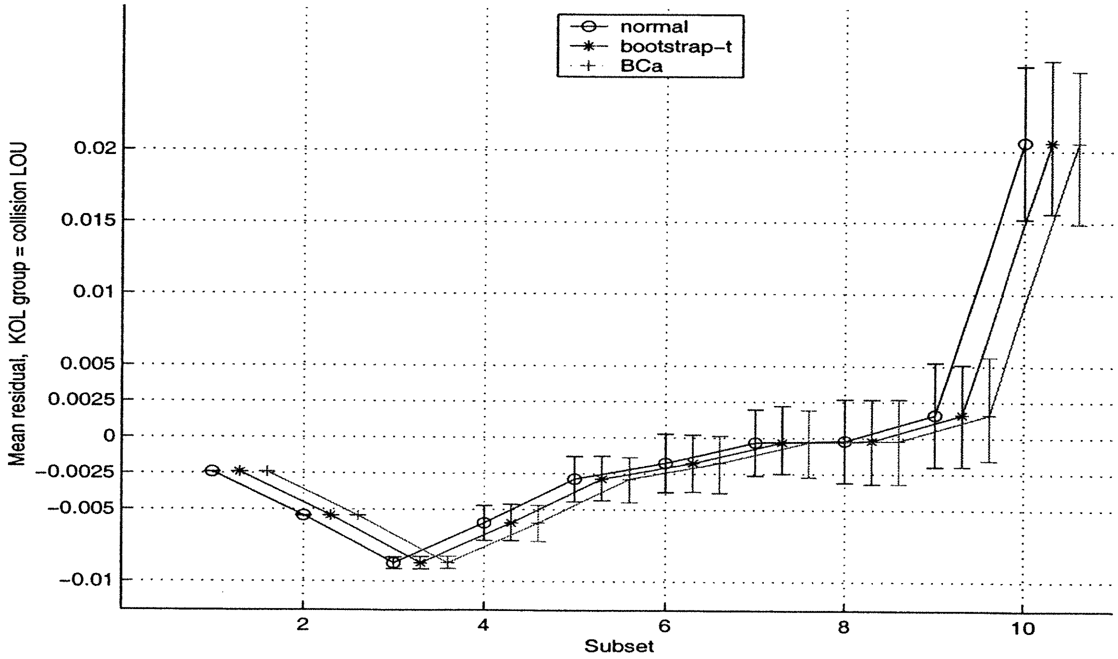


Figure 5.17: Comparison of intervals for comprehensive mixture test set

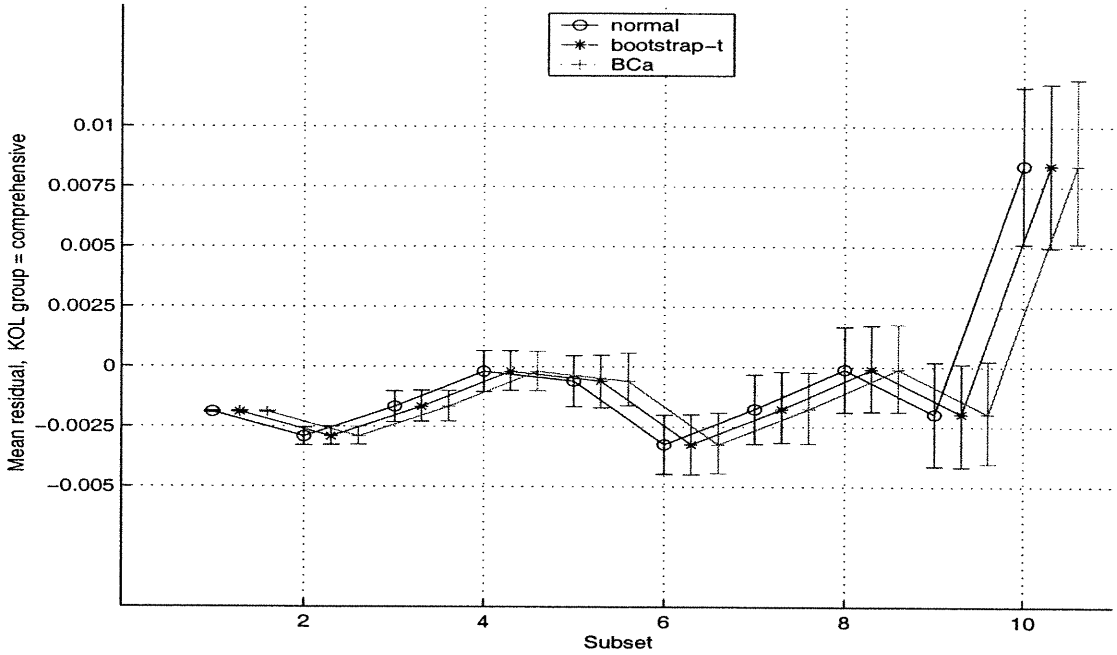
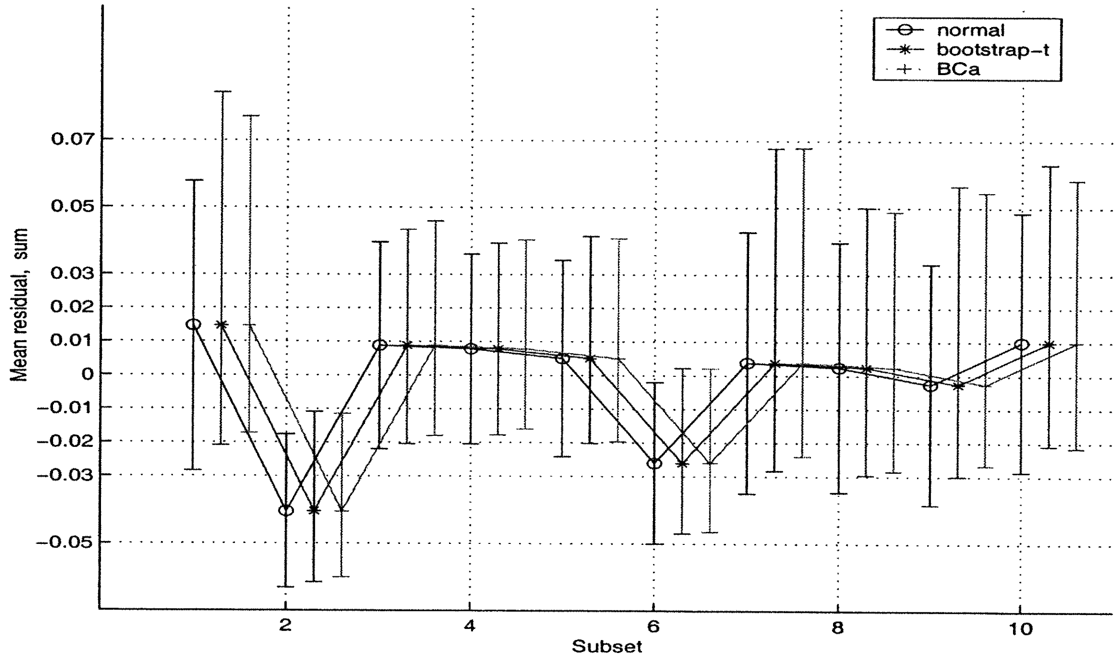


Figure 5.18: Comparison of intervals for sum mixture test set



Chapter 6

Conclusion

In this thesis, we studied the influence of skewness and kurtosis of the residual distribution upon hypothesis testing for evaluating model fairness. We implemented two kinds of hypothesis tests, approximate normal tests and bootstrap tests, followed by Bonferroni adjustments to take into account the multiplicity of tests considered. In conclusion, we summarize our experimental results in two points:

1. Bonferroni procedures showed that the model fairness is adequate for some dependent variables. For others, it is not. For example, for bodily injury, subset 9, the Bonferroni BCa interval for the mean residual is $[-0.0692, -0.0048]$, meaning that on average, each insured in this group would overpay by an amount between 4 and 69 dollars.

2. In large sample circumstances, approximate normal tests could give acceptable results in the middle of the distribution. However, in the tails of the distribution, due to the large skewness and kurtosis, the approximate normal test is not appropriate even in this case where the sample size is close to $200,000$. Bootstrap procedures would give better performance regardless of the residual distribution, especially the BCa interval.

Appendix A

Explanatory variables definition

The definitions of explanatory variables are presented in this appendix. Most of the definitions are easy to understand. However, some of them defined by the company, are measured in some unknown way.

- **lease_indicator** : identifies whether the vehicle is leased or not.
- **policy_start_date**: identifies policy start date.
- **territory**: identifies where the vehicle came from. The company divides the geographic area into territories based on geography and population. Vehicles driven to and from work or school or for pleasure use must be rated in the territory in which they are primarily located when not in use. This is usually the home address.
- **rate_class**: an insurance category, based on how the vehicle is used, to and from work, pleasure use only, business, delivery, etc., which partly determines what the insurance will cost.
- **claim_rated_scale**: shows how claims against the driver's record affect the premium. The place on the Claim_Rated_Scale is a major factor in determining the premiums one pays.
- **fleet_discount**: a scale of discounts and surcharges which takes the combined loss experience of all vehicles in the fleet into consideration.
- **third_party_ext_limit**: identifies Extended Third Party coverage limit amount. Extended Third Party liability gives additional protection if one is sued or found responsible for injuries to others or damage to property as a result of a motor vehicle crash. One can increase his coverage from Third Party Legal liability $200,000 limit to amount ranging from $300,000 to $15 million.
- **collision_deductible**: Collision insurance pays to repair or replace the vehicle if it is damaged as a result of upset or a collision with another vehicle,

a person, or an object, including the ground or highway, or impact with an object on or in the ground. When buying collision insurance, one has to choose a deductible. The deductible is the amount one must pay before the insurance kicks in to pay for the remainder of the repairs.

- **comp_deductible**: comp means comprehensive. Comprehensive insurance covers loss or damage to the vehicle by any cause except loss or damage covered by collision insurance. When buying comprehensive insurance, one has to choose a deductible. The deductible is the amount one must pay before the insurance kicks in to pay for the remainder of the repairs.

- **specified_perils**: this form of insurance provides specific coverage only against fire, lightning, theft(except by an employee or member of the household), windstorm, earthquake, hail, explosion, riot, civil commotion, falling or forced landing of an aircraft or portion of it, rising water or the stranding, sinking, burning, derailment, or collision of a corveyancy transporting the vehicle on land or water. Losses which are not covered include vandalism, malicious mischief or rockchip damage to windshields.

- **loss_of_use**: if one has Collision, Comprehensive or Specified Perils insurance, one can buy a Loss of Use policy to pay for the costs (up to limits chosen by the insured) of substitute transportation while the vehicle is being repaired as the result of an insurable claim.

- **roadstar_indicator**: identifies whether motorists have maintained a 40% discount with the company for more than five years in a row.

- **limited_depreciation**: one can purchase a Limited Depreciation Policy if the vehicle is not more than three model years old, for example, for the calendar year 1997, the following model years are eligible: 1997, or 1998(first model year), 1996 (second model year) and 1995 (third model year), even if one is not the first owner of the vehicle.

- **replacement_cost**: this is for RoadStar only. If one has a claim on a newer vehicle, RoadStar Replacement Cost coverage protects the insured from losing money due to depreciation. Replacement Cost coverage can be purchased for any vehicle that is three model years old or less.

- **vehicle_model_age**:

- **vehicle_color**: identifies the color of the vehicle as stated by the insured when the policy was originally created, or as changed by insured.

- **drv_is_owner**: identifies whether the driver is the owner of the vehicle.

- **nyears_since_exam**: identifies years since the driver passed exam.

- **drv_class**:

- **drv_exam_status**:

- **drv_age**: identifies the age of the driver.

- **drv_sex**: identifies the sex of the driver.

- **nyears_since_original_lic**: identifies years since obtaining original licence.

- **age_last_conviction**: identifies age at the last conviction.

- **drv_last_penalty_pts**: identifies last penalty points of the driver.
- **drv_num_convictions**: identifies times of conviction of the driver.
- **drv_total_penalty_pts**: identifies total penalty points of the driver.
- **age_last_accident**: identifies driver age at last accident.
- **drv_num_accidents**: identifies accident times of the driver.
- **age_last_suspension_end**: identifies driver age at the end of last suspension.
- **drv_num_suspensions**: identifies suspension times of the driver.
- **drv_suspension_days**: identifies suspension days of the driver.

Appendix B

Detailed experimental results

This appendix reports most experimental results in details, including tables and figures already mentioned in Chapter 5. Please note that the residuals were obtained by subtracting predictions from observed claim amounts. In tables with boxes, the boxed values refer to significant results under level $\alpha = 0.05$. The residuals and relative quantities are measured in thousands of dollars.

Figure B.1: QQplot of claim for property damage test set

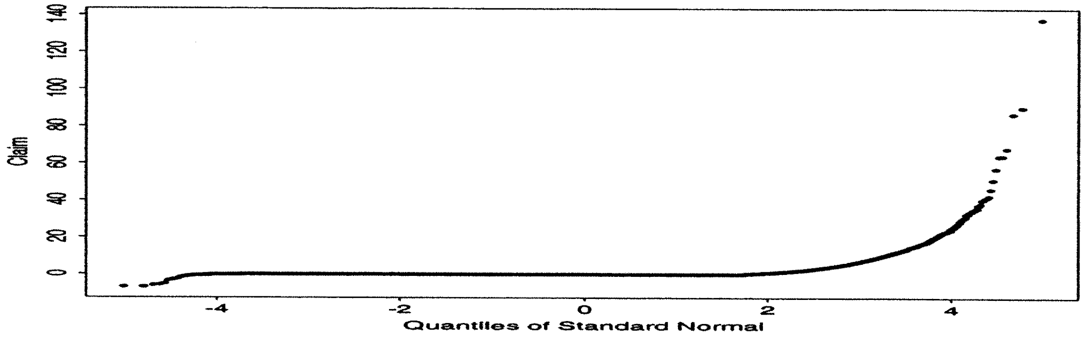


Figure B.2: QQPlot of claim for accident death test set

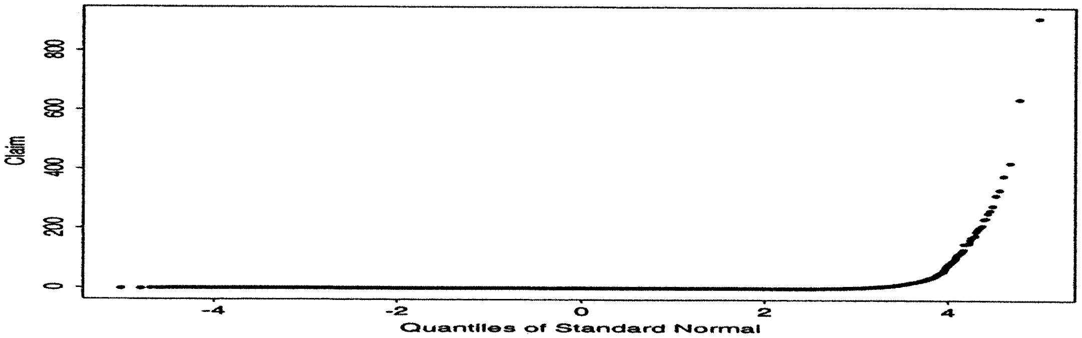


Figure B.3: QQplot of claim for collision L.O.U test set

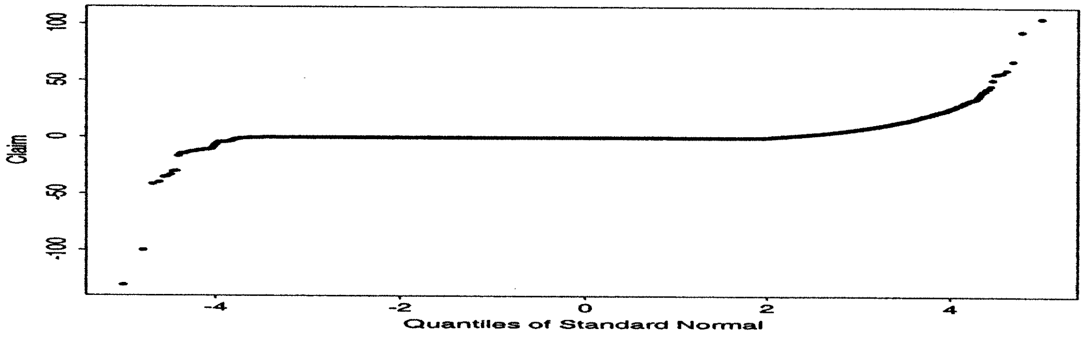


Figure B.4: QQplot of claim for comprehensive test set

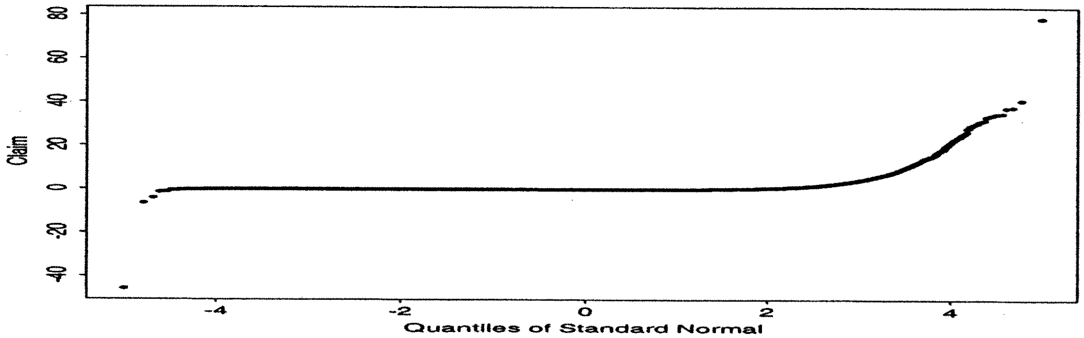


Figure B.5: QQplot of residual for property damage mixture model test set

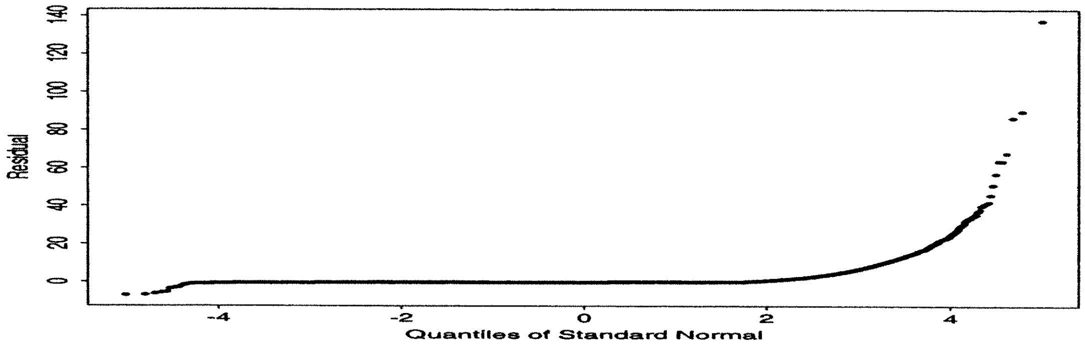


Figure B.6: QQplot of residual for accident death mixture model test set

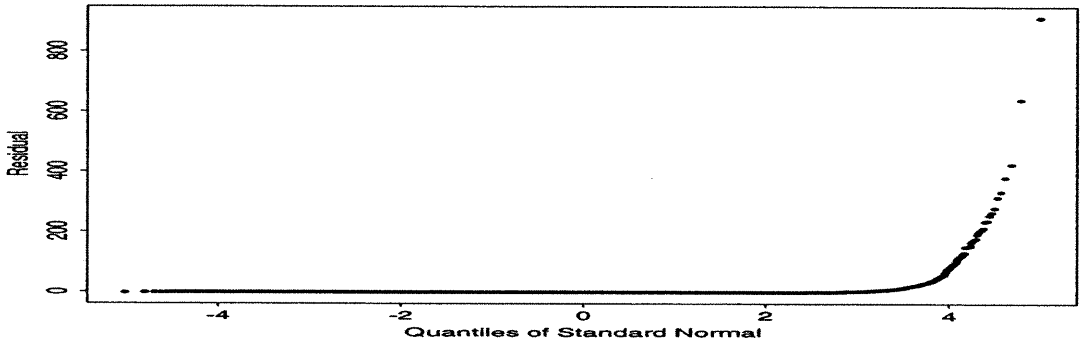


Figure B.7: QQplot of residual for collision L.O.U mixture model test set

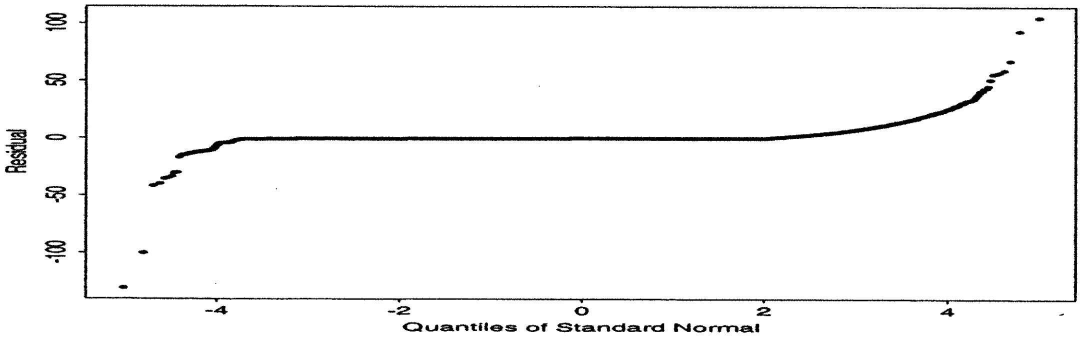


Figure B.8: QQplot of residual for comprehensive mixture model test set

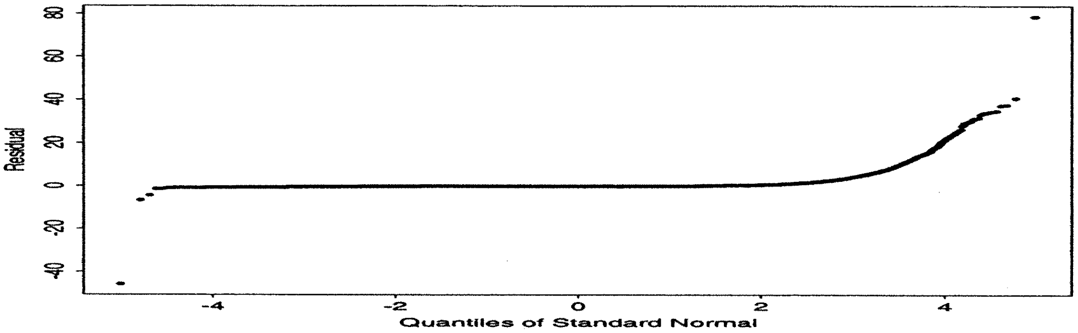


Figure B.9: QQplot of residual for sum mixture model test set

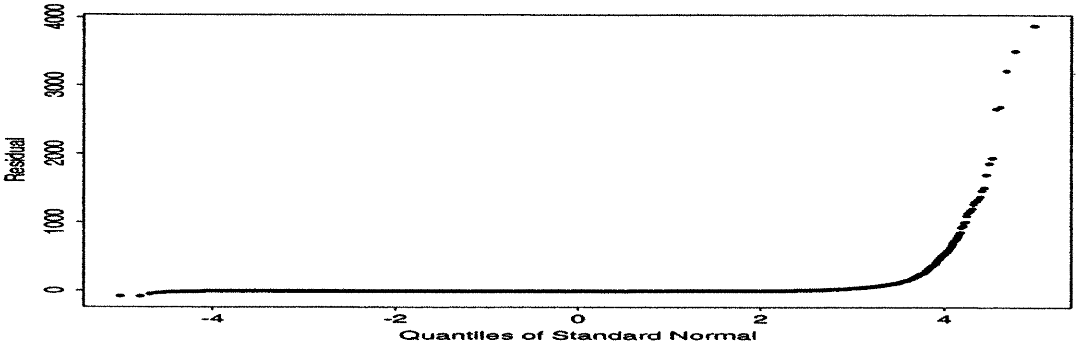


Figure B.10: Subset residual distributions for property damage mixture test set

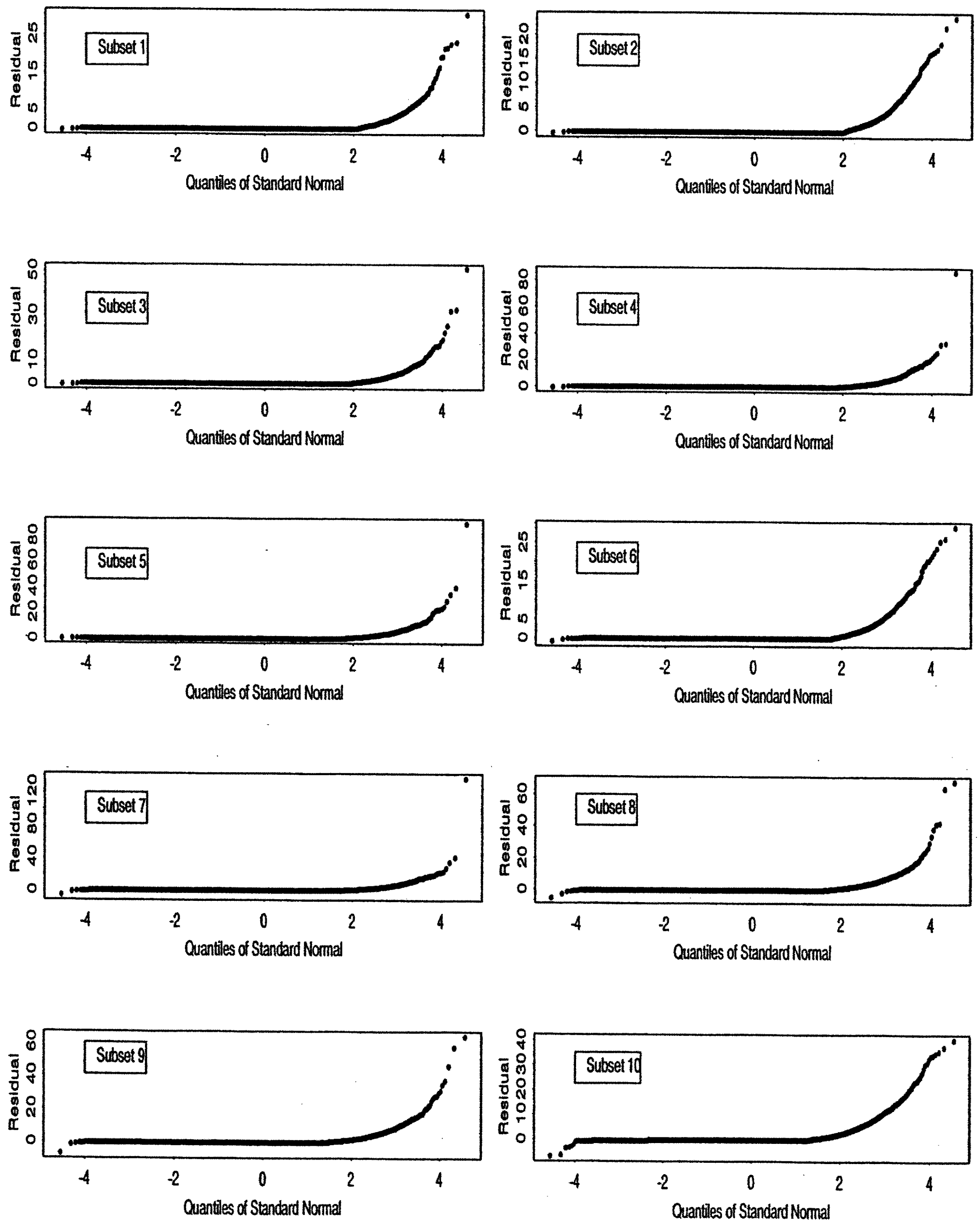


Figure B.11: Subset residual distributions for accident death mixture test set

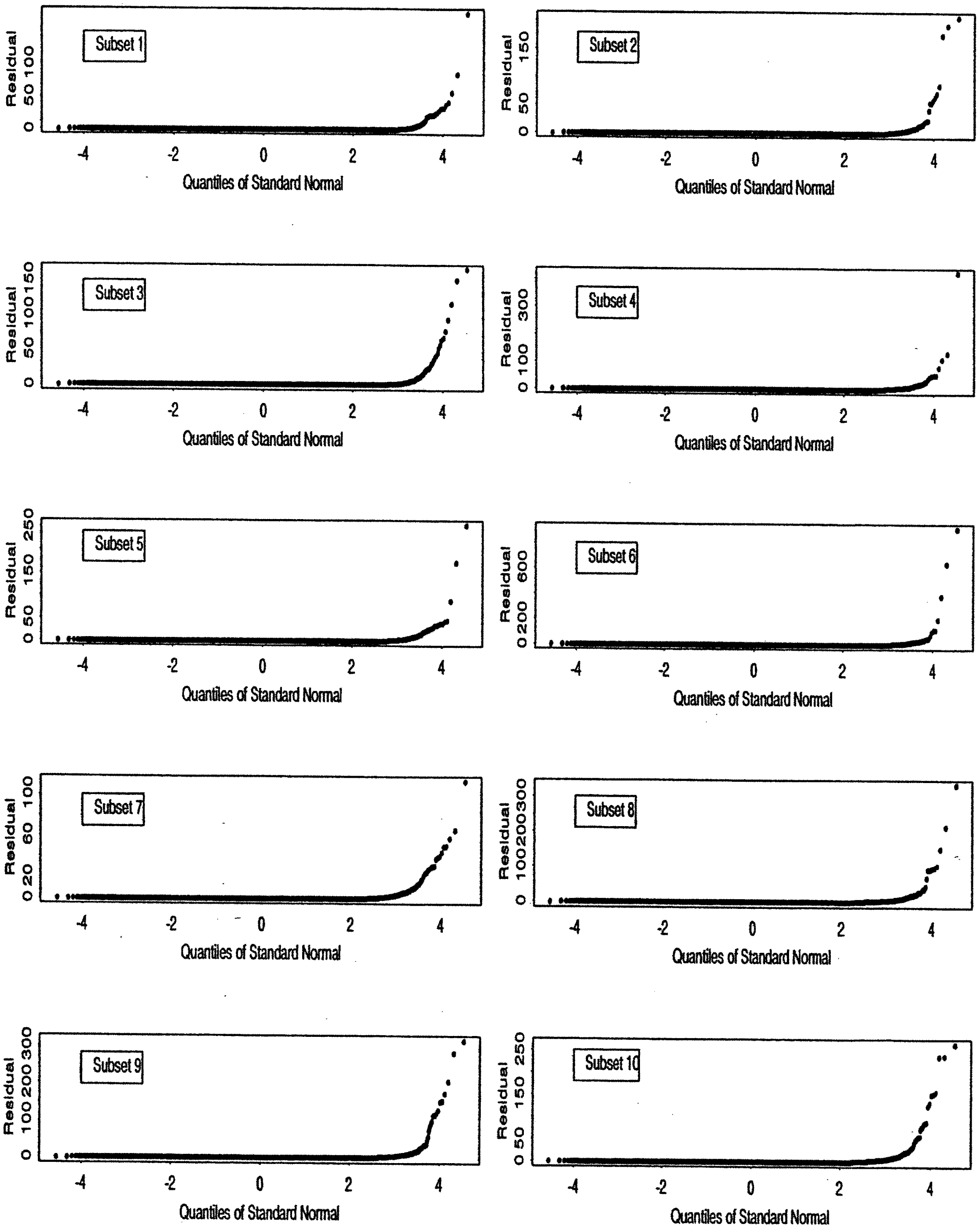


Figure B.12: Subset residual distributions for collision L.O.U mixture test set

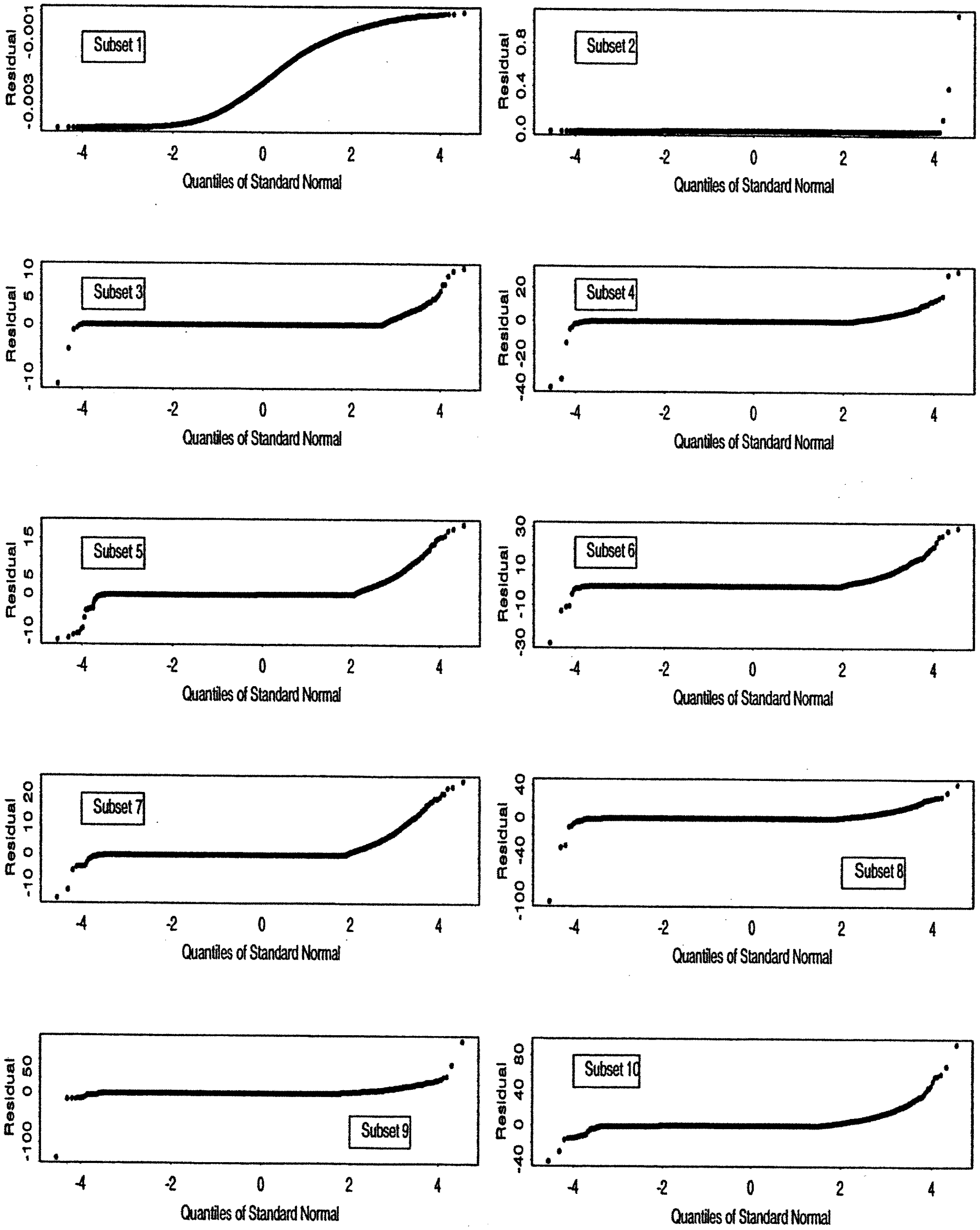


Figure B.13: Subset residual distributions for comprehensive mixture test set

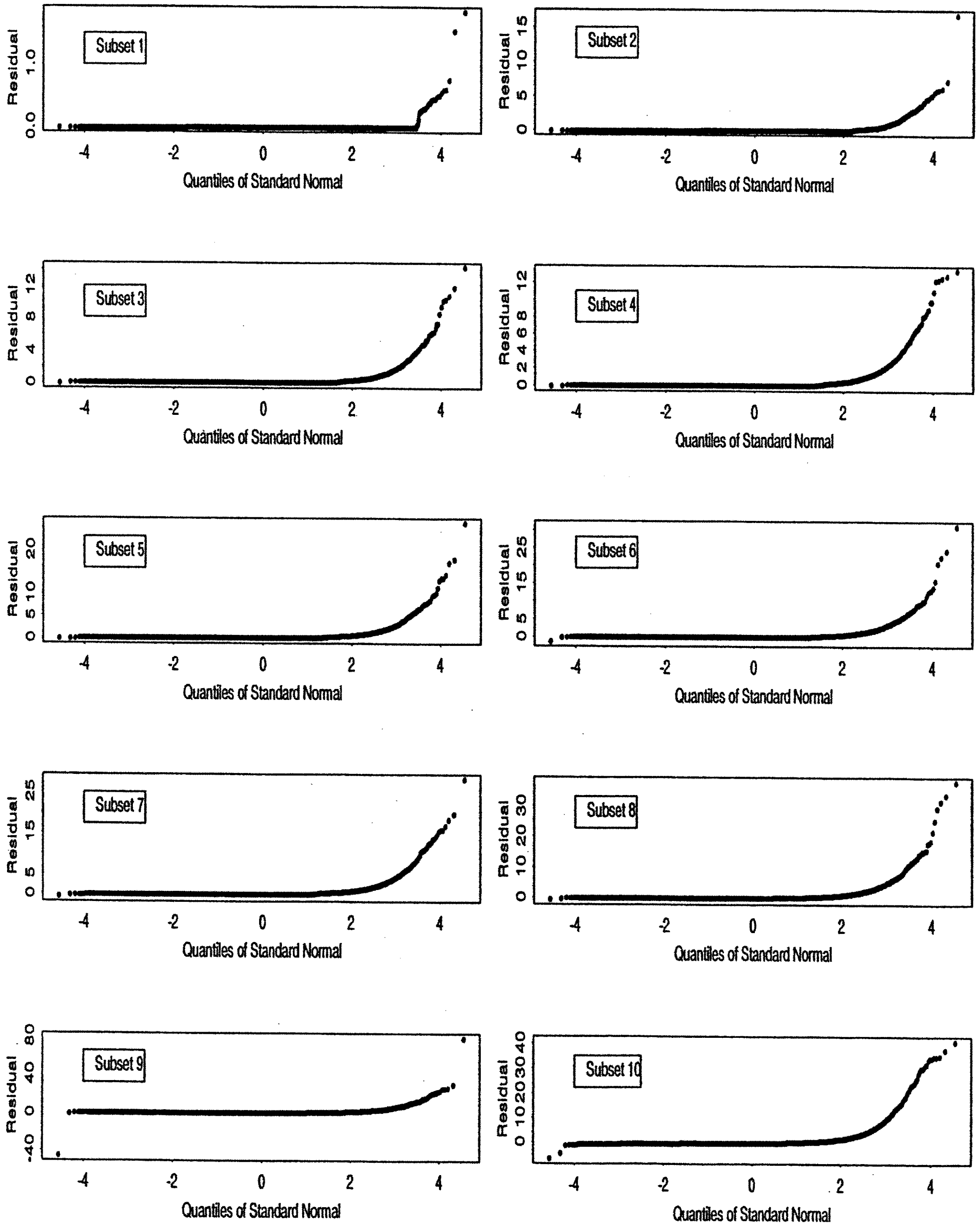


Figure B.14: Subset residual distributions for sum mixture test set

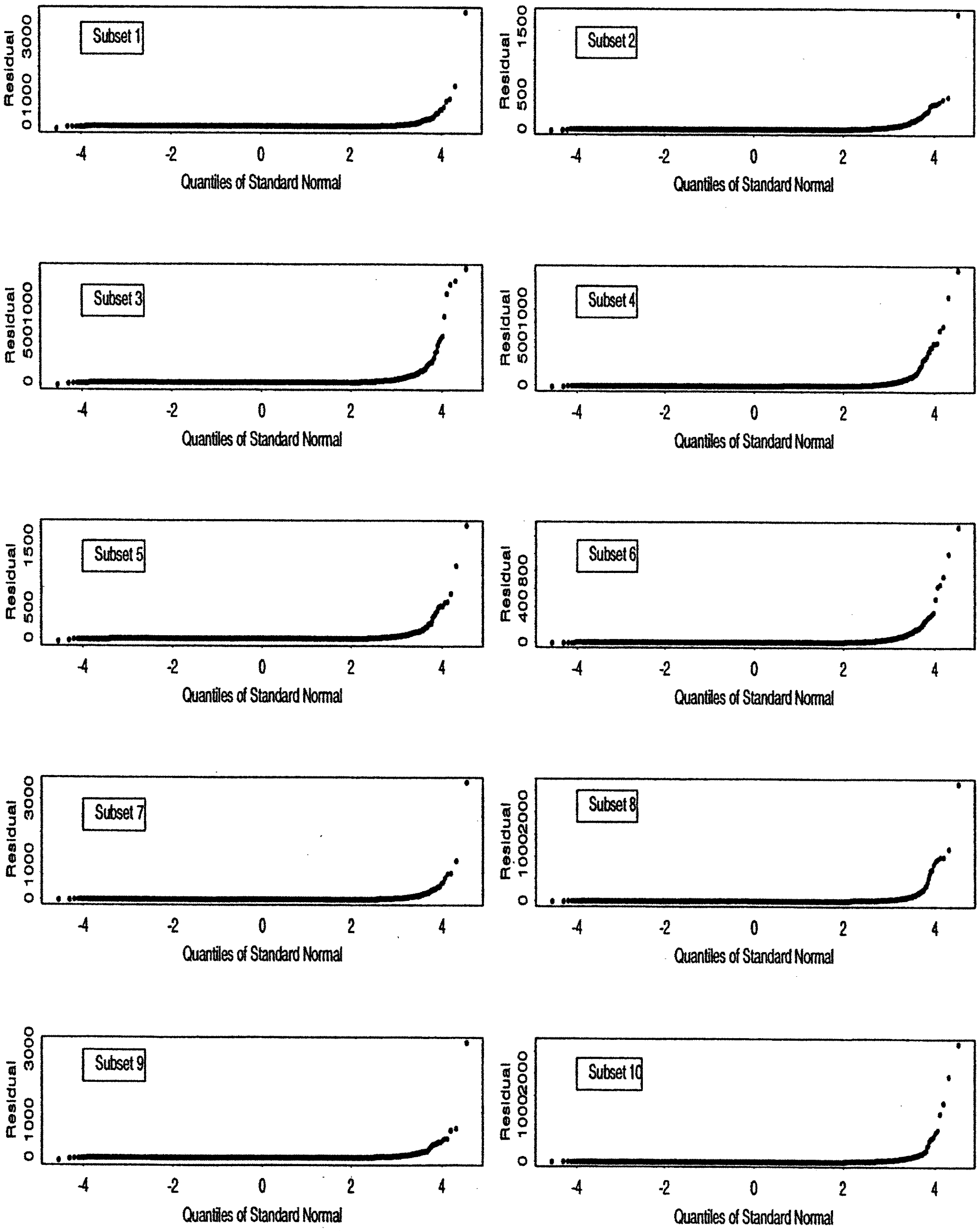


Table B.1: Subset ranges for property demage mixture model test set

	Left end	Right end
subset 1	0.0107	0.0293
subset 2	0.0293	0.0363
subset 3	0.0363	0.0434
subset 4	0.0434	0.0512
subset 5	0.0512	0.0600
subset 6	0.0600	0.0700
subset 7	0.0700	0.0826
subset 8	0.0826	0.0101
subset 9	0.1009	0.0134
subset 10	0.1339	0.7990

Table B.2: Subset ranges of mixture model test set

	Accident death		Collision LOU	
	Left end	Right end	Left end	Right end
subset 1	0.0025	0.0099	0.0004	0.0037
subset 2	0.0099	0.0125	0.0037	0.0078
subset 3	0.0125	0.0150	0.0078	0.0207
subset 4	0.0150	0.0176	0.0207	0.0369
subset 5	0.0176	0.0207	0.0369	0.0509
subset 6	0.0207	0.0244	0.0509	0.0651
subset 7	0.0244	0.0291	0.0651	0.0814
subset 8	0.0291	0.0358	0.0814	0.1037
subset 9	0.0358	0.0472	0.1037	0.1439
subset 10	0.0472	0.3058	0.1439	1.3464

Table B.3: Subset ranges of mixture model test set

	Comprehensive		Sum	
	Left end	Right end	Left end	Right end
subset 1	0.0005	0.0035	0.0490	0.1692
subset 2	0.0035	0.0178	0.1692	0.2201
subset 3	0.0178	0.0313	0.2202	0.2682
subset 4	0.0313	0.0440	0.2682	0.3169
subset 5	0.0440	0.0575	0.3169	0.3705
subset 6	0.0575	0.0720	0.3705	0.4321
subset 7	0.0720	0.0882	0.4321	0.5104
subset 8	0.0882	0.1091	0.5104	0.6242
subset 9	0.1091	0.1441	0.6242	0.8289
subset 10	0.1441	0.9866	0.8289	5.7483

Figure B.15: Confidence interval for property damage mixture model test

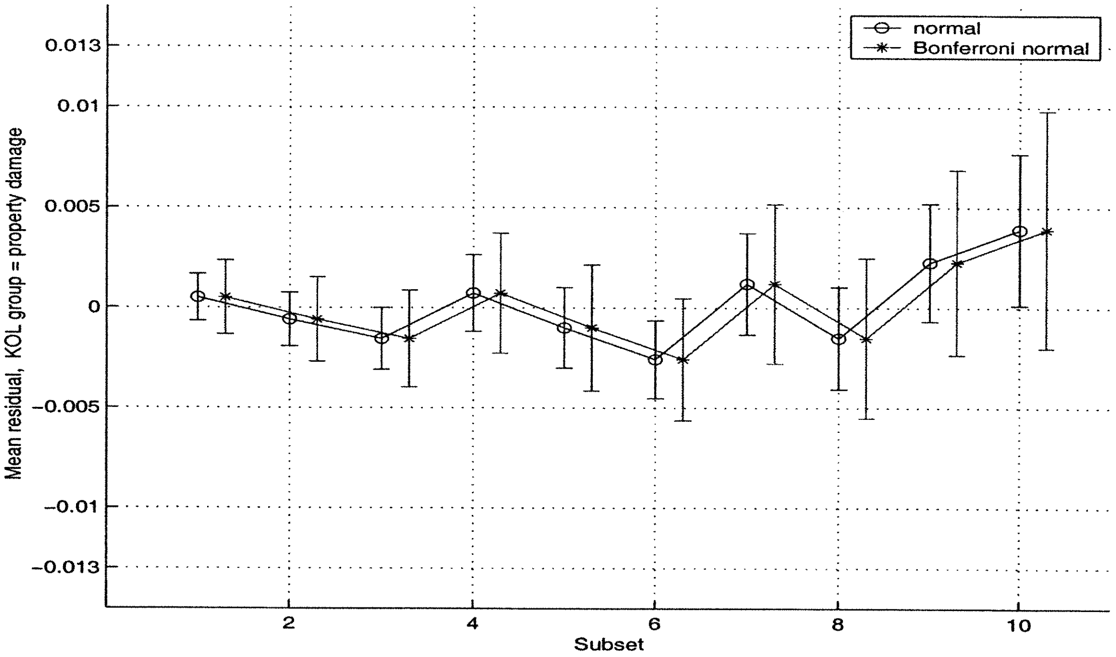


Figure B.16: Confidence interval for accident death mixture model test

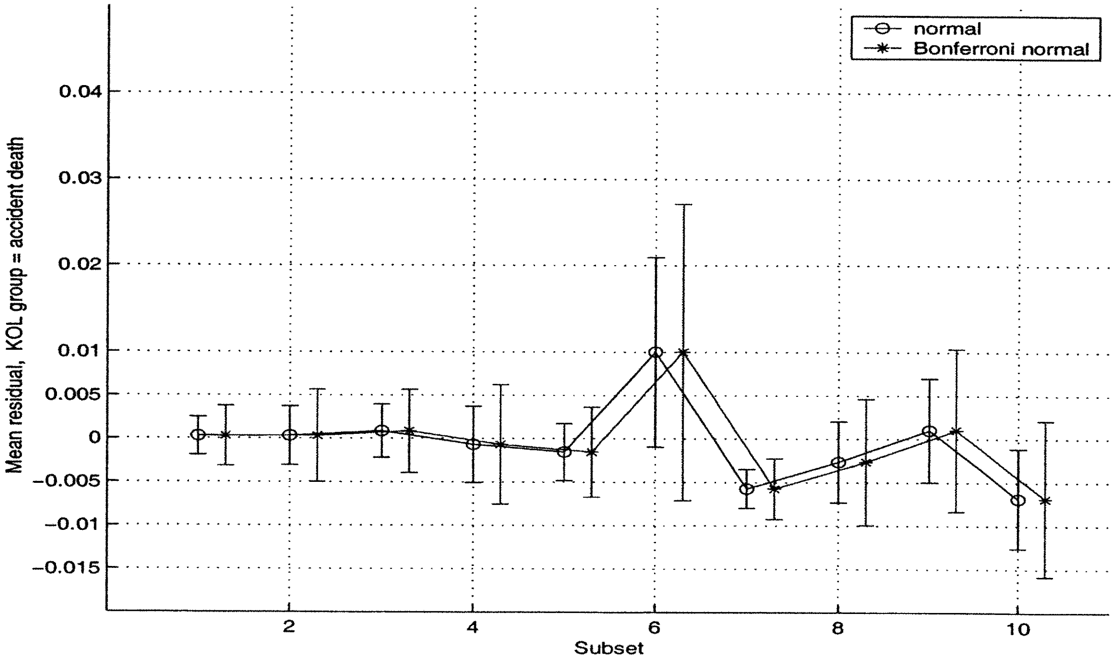


Figure B.17: Confidence interval for collision L.O.U mixture model test

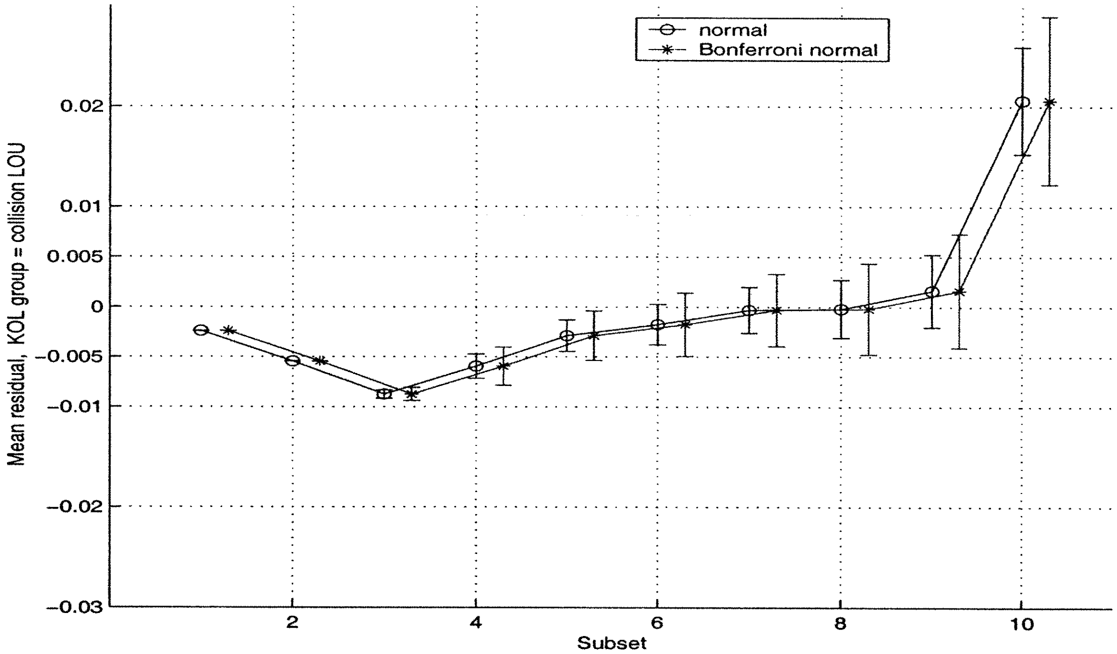


Figure B.18: Confidence interval for comprehensive mixture model test

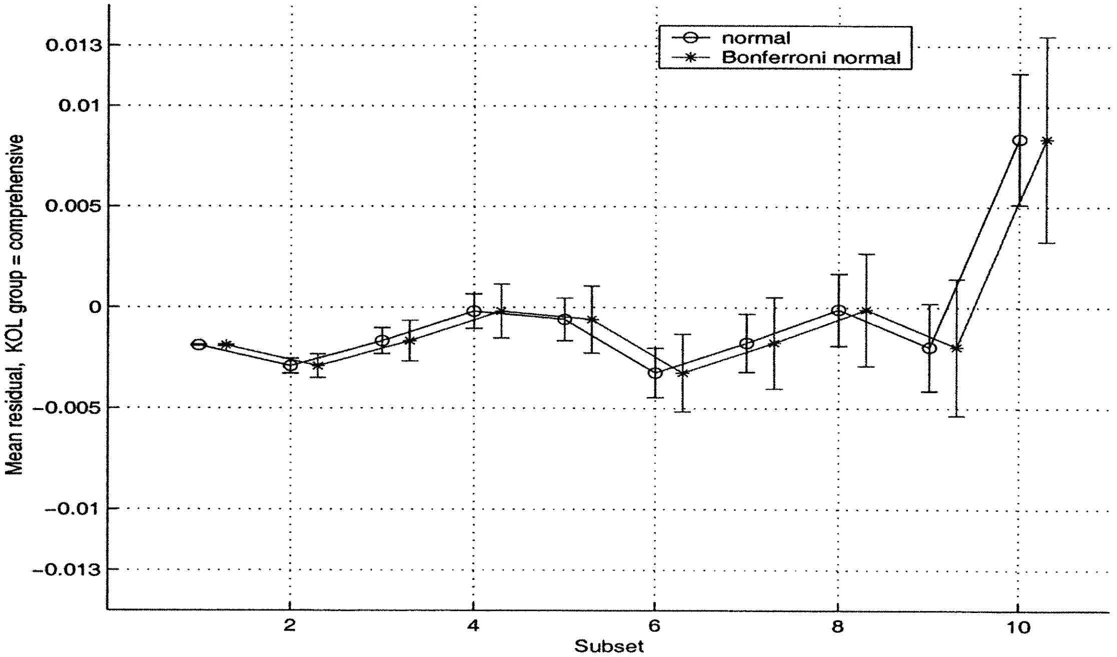


Figure B.19: Confidence interval for sum mixture model test

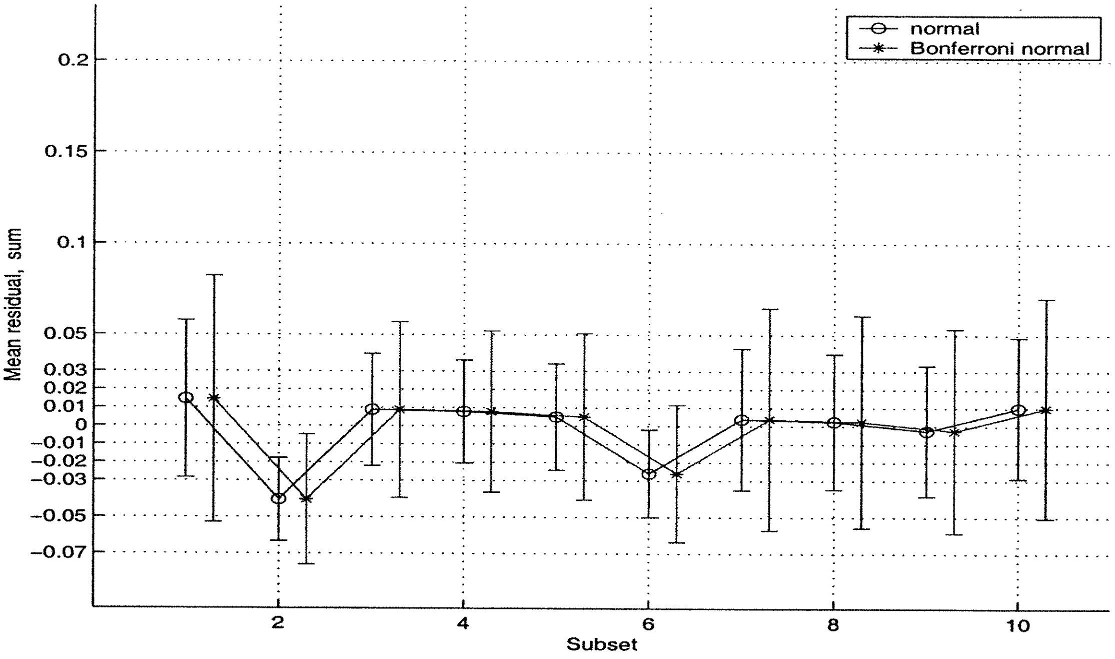


Figure B.20: Bootstrap-t and Bonferroni bootstrap-t for property dam. mix test

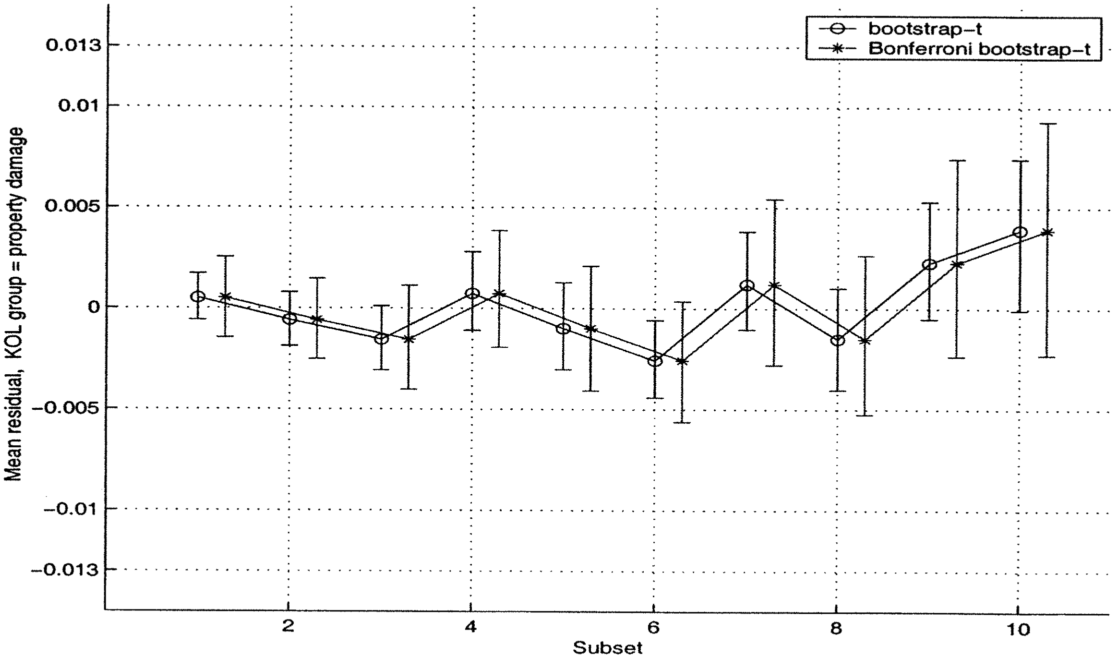


Figure B.21: Bootstrap-t and Bonferroni bootstrap-t for accident death mix test

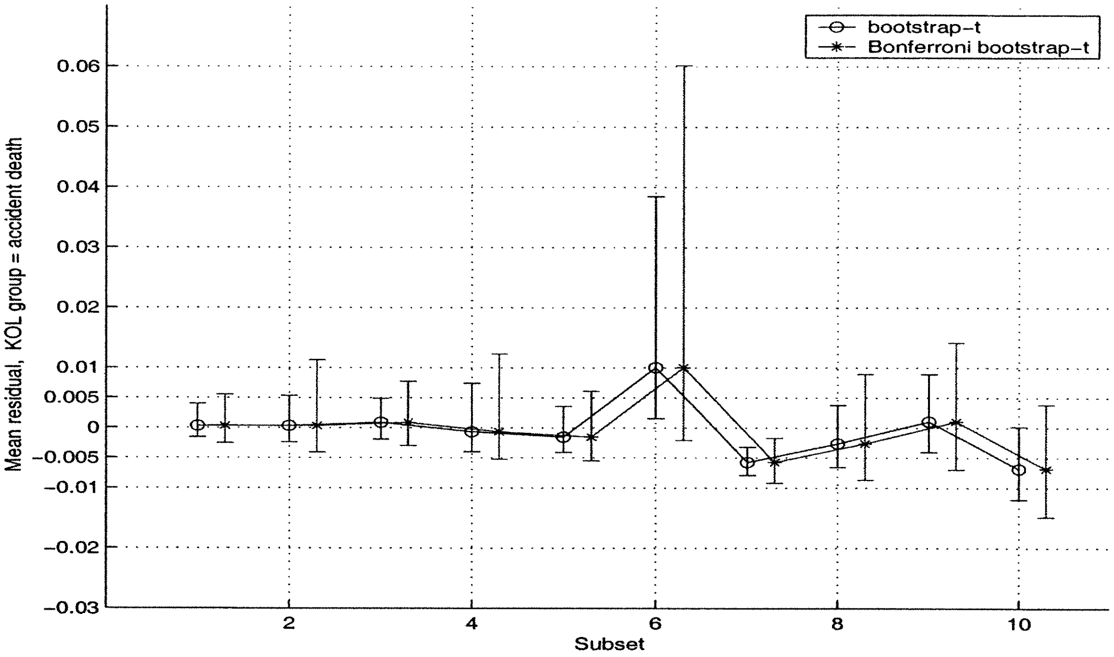


Figure B.22: Bootstrap-t and Bonferroni bootstrap-t for collision L.O.U mix test

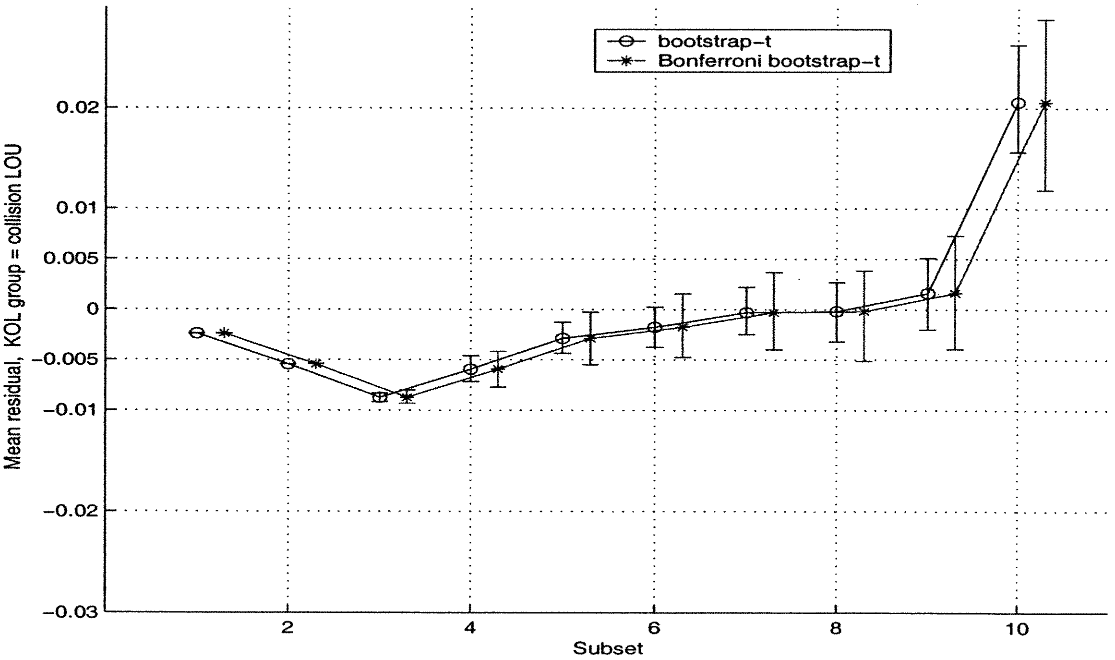


Figure B.23: Bootstrap-t and Bonferroni bootstrap-t for comprehensive mix test

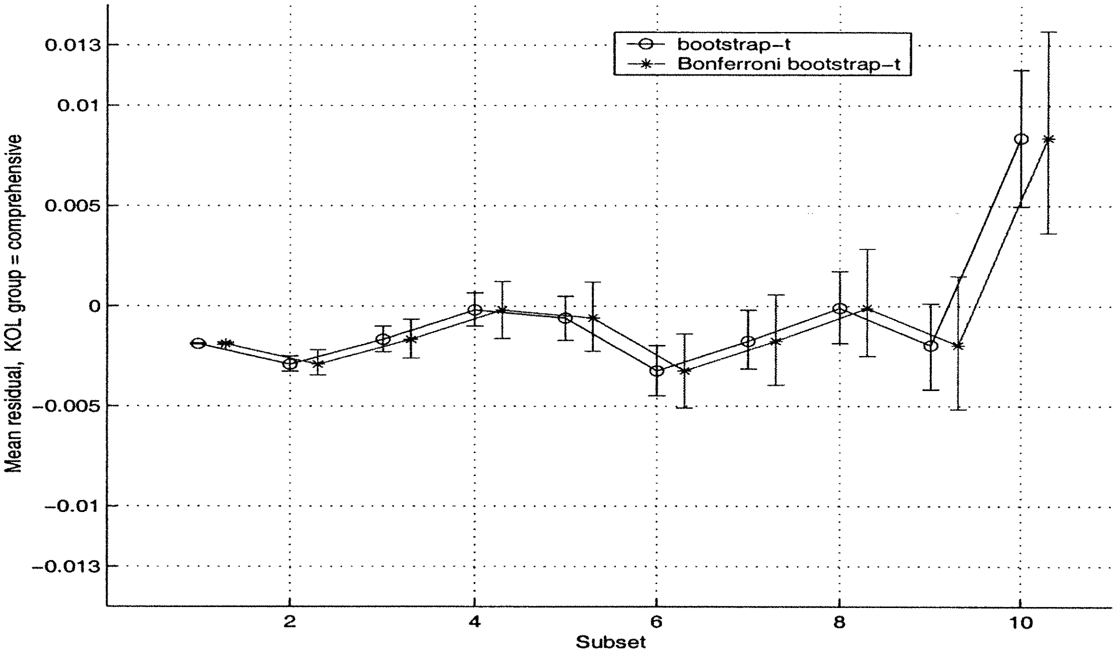


Figure B.24: Bootstrap-t and Bonferroni bootstrap-t for sum mix test

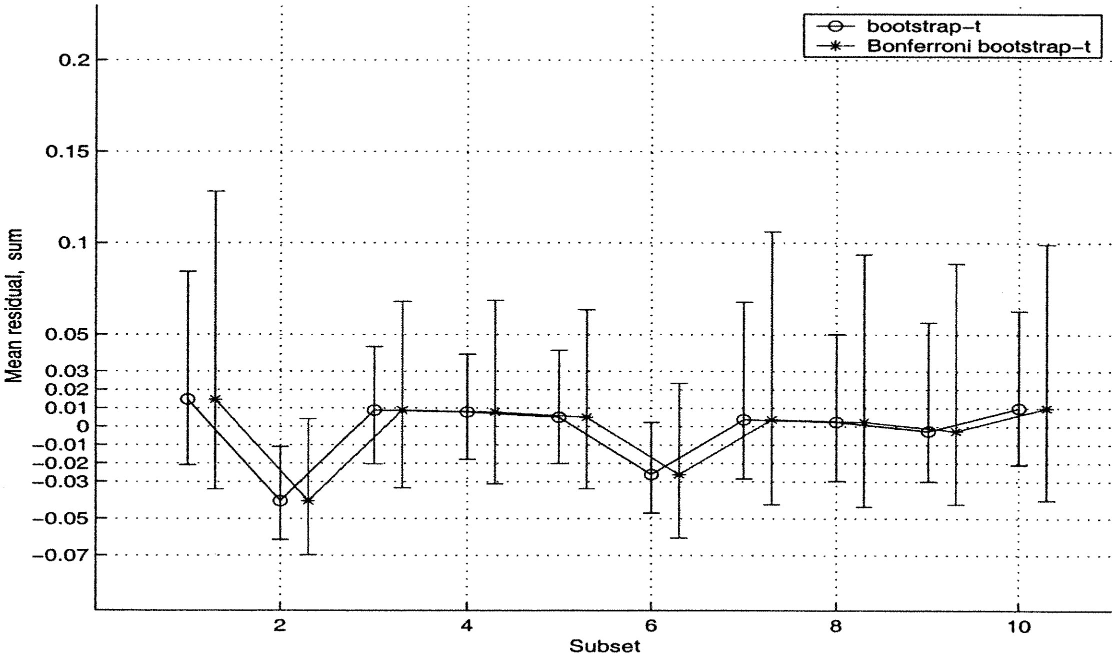


Figure B.25: BCa and Bonferroni BCa for property dam. mixture test set

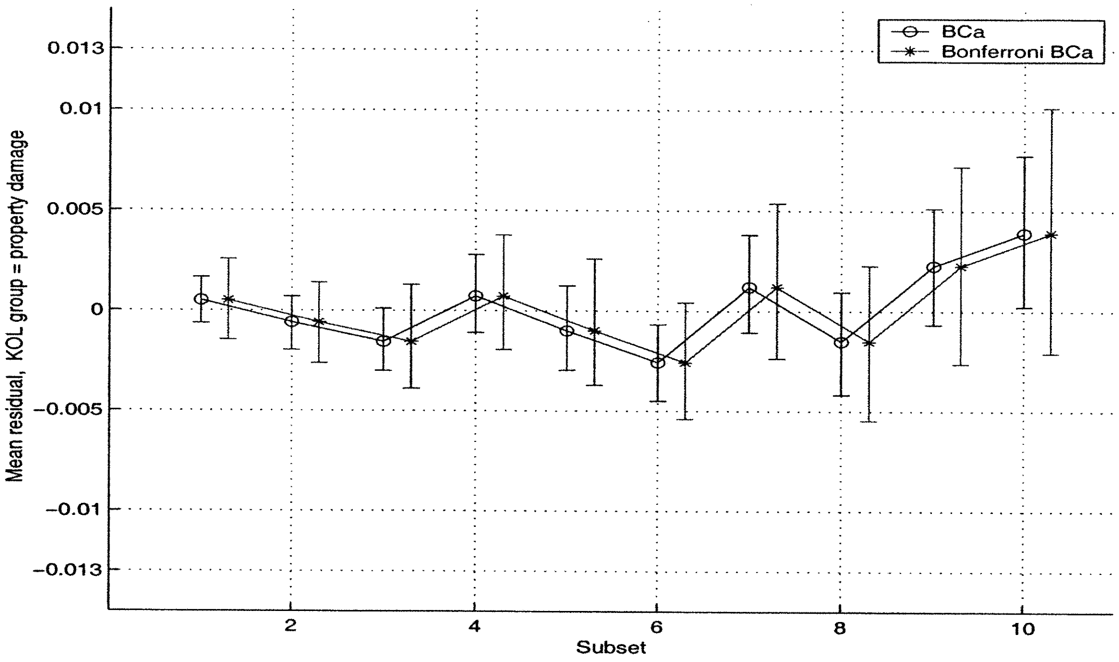


Figure B.26: BCa and Bonferroni BCa for accident death mixture test set

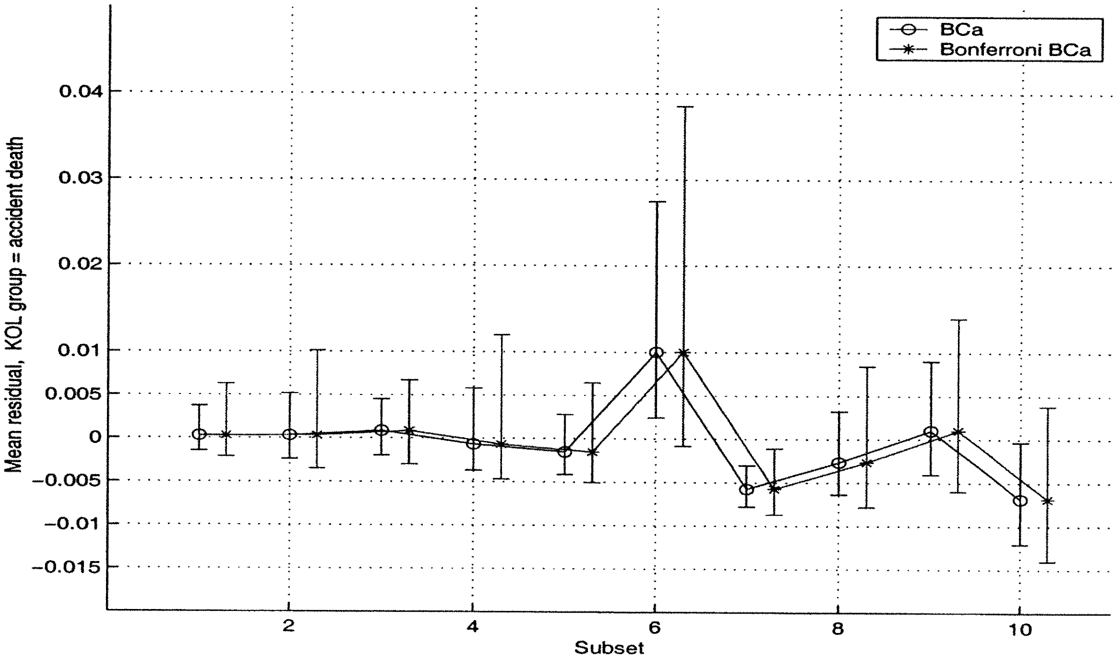


Figure B.27: BCa and Bonferroni BCa for collision L.O.U mixture test set

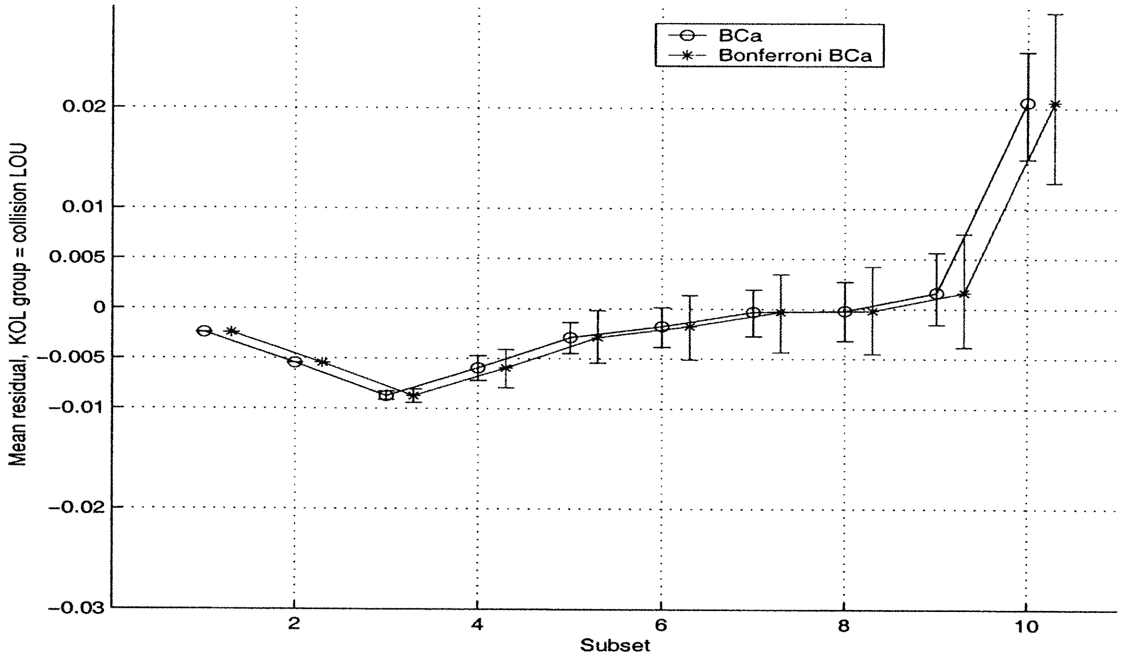


Figure B.28: BCa and Bonferroni BCa for comprehensive mixture test set

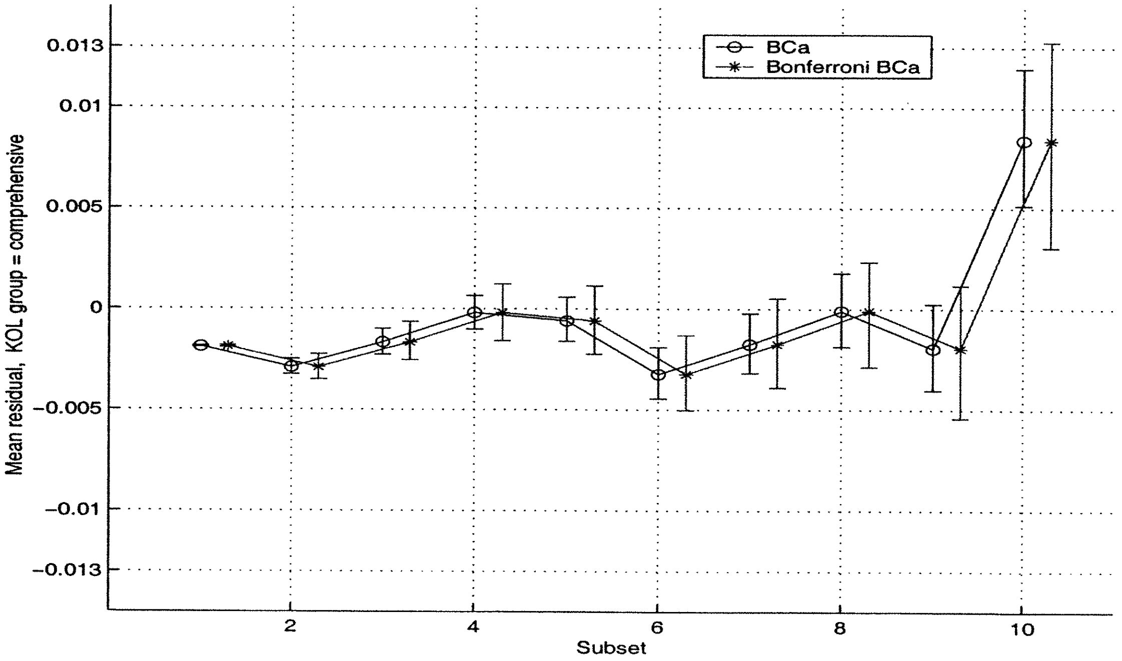
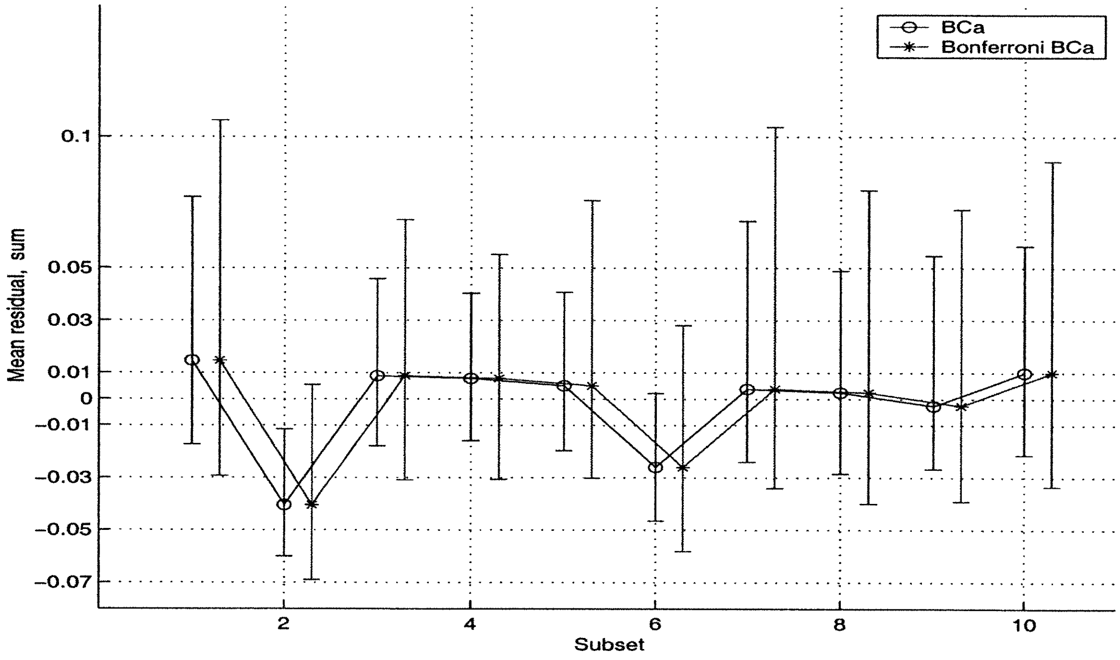


Figure B.29: BCa and Bonferroni BCa for sum mixture test set



Bibliography

BISHOP, M. (1998). *Neural Networks for Pattern Recognition.* Oxford: Clarendon Press.

DAVISON, A. C. & HINKLEY, D. V. (1999). *Bootstrap Methods and their Application.* Cambridge, UK: Cambridge University Press.

DRAPER, N. R. & SMITH, H. (1998). *Applied Regression Analysis.* New York: John Wiley & Sons, Inc.

DUGAS, C., BENGIO, Y., BELISLE, F., NADEAU, C. & GARCIA, R. (2001). Incorporating second-order functional knowledge for better option pricing **13**, 472–478.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26.

EFRON, B. & TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion.* New York: Springer–Verlag.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. (2001). *The Elements of Statistical Learning.* New York: Springer.

HSU, J. (1996). *Multiple Comparisons: Theory and Methods.* London: Chapman & Hall.

JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. & HINTON, G. E. (1979). Adaptive mixture of local experts. *Neural Computation* **3**, 79–87.

MCCULLAGH, P. & NELDER, J. (1989). *Generalized Linear Model.* London: Chapman & Hall.

MILLER, R. G., J. (1980). *Simultaneous Statistical Inference.* New York: Springer–Verlag.

MILLER, R. G., J. (1985). *Beyond ANOVA, Basics of Applied Statistics.* New York: John Wiley & Sons, Inc.