LOSS RESERVING CHAIN LADDER METHODS APPLIED TO A SMALL MIDWESTERN INSURANCE COMPANY

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Peter Raymond Martin

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

April 2015

Fargo, North Dakota

Title

LOSS RESERVING CHAIN LADDER METHODS APPLIED TO A SMALL MIDWESTERN INSURANCE COMPANY

By

Peter Raymond Martin

The supervisory committee certifies that this thesis complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Tatjana Miljkovic
Chair

Rhonda Magel
Co-Chair

Megan Orr

Indranil Sengupta

Approved:

14 April 2015
Date

Rhonda Magel
Department Chair

ABSTRACT

Estimating future losses is integral to setting aside appropriate reserves in the insurance industry. This study analyzes different Chain Ladder reserving methods based on weighted-least square regression that consider different function of weights. These methods are tested on 78 NAIC fully developed loss triangles. While the CRE Chain Ladder method is selected based on its performance, this method does not work well for a small number of NAIC companies that may have erratic changes in their loss trends. For these outliers, two other methods were explored for the early development years; the nearest neighbor technique and mixture of linear regressions. A recommendation is then made to a small Midwestern insurance company on the best methodology to use for estimating the loss reserves based on the actual data provided. These results can be useful to any other insurance company currently using Chain Ladder methods in loss reserving practices.

ACKNOWLEDGEMENTS

Thank you to both my advisor, Dr. Tatjana Miljkovic, for providing endless help and ideas over the past two years and to the insurance company that provided a triangle of unknown losses for me to analyze.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Reserving is the cornerstone of the insurance industry. An insurance company must set aside enough money to pay all claims, present and future, on the policies currently in force. Inadequate reserves can lead to insolvency and over-adequate reserves can lead to premium rates that are not competitive. Industry best practices attempt to mitigate or spread risk, and therefore new techniques to accurately predict the necessary reserves must be tested repeatedly. Most insurance companies are unwilling to risk changing their methods. The concern is that new techniques will not account for all the nuances and intangible elements contributing to claims and thus fail to improve the accuracy of the reserve predictions. In addiction, this concern discourages companies from testing new techniques due to risk associated with overexposure. This leads to a large proportion of insurance companies continuing to use conventional methods or using external consultants when calculating ultimate losses for an accident year and total losses from a line of insurance.

A loss triangle is a table of claims for one line of insurance over a period of development years. These claims naturally form an upper left triangle of actual incurred losses and a lower right triangle of future losses. Each calendar year, another diagonal of actual losses is added to the triangle and a new accident year is added to the bottom of the triangle. Many lines of insurance take years for their claims to fully develop, such as personal injury or malpractice, while others, such as homeowners liability, fully develop in a year or two. To collect adequate premiums each year, the ultimate losses from that year's policies must be estimated to avoid subsidizing previous years' losses with current year premiums.

Various deterministic and stochastic methods are used to predict these unknown lower triangle losses, with little comparison to the advantages and disadvantages of each. In this thesis a deterministic method of comparing development ratios will be introduced, as well as using weighted least squares regression, the nearest neighbor technique and mixtures of regression. These techniques are compared using a test data set of fully known losses reported by insurance companies from across the country. The techniques discussed in this thesis all use a development factor to estimate each subsequent year's losses. Discussion of these techniques as well as comparison between them using the test data set follows. Ultimately, the best of these techniques will be applied to a single triangle from an actual insurance company, and the estimated losses are discussed.

2. LITERATURE REVIEW

The most basic forms of predicting claim amounts are deterministic, requiring decisions by a person based on experience or expert knowledge to apply the techniques to a triangle of loss data [15, Werner and Modlin 2010]. The resulting prediction's uncertainty cannot be quantified in these cases however. Often, even when applying stochastic methods, adjustments are made, possibly violating assumptions of the stochastic method or invalidating the prediction obtained from the stochastic method [14, Verall 2004].

The most commonly used method in loss reserving is the chain ladder method. The chain ladder method is a distribution-free method, relieving some of the usual assumptions common to most modelling techniques. This method is used by formulating a common ratio of losses between subsequent development years [7, Mack 1993]. The only assumption in the chain ladder method is that subsequent claim years are independent [15, Wuthrich and Merz 2008]. Some variations on the basic chain ladder method [9, Quarg and Mack 2004] can also be used to estimate other values such as reserves and current excess reserves, as well as estimating the standard error of these predictions [10, Schnieper 1991]. Calculating the standard error of the chain ladder method and quantifying the uncertainty with these different variations in the chain ladder method is a helpful way of evaluating the differences between the various methods [7, Mack 1993].

The distribution-free chain ladder method has underlying models that have been the subject of more recent research. These newer models assume claim amounts follow a specific distribution and can lead to the same estimates as the distribution-free chain ladder method. For example, a Poisson model for claim counts can lead to the same expected number of claims as the distribution-free chain ladder estimates [15, Wuthrich and Merz 2008]. Generalized linear models (GLM) have been historically popular in the field of loss reserving, and the increased access to user-friendly statistical software has further bolstered the popularity of methods using GLM [5, Haberman and Renshaw 1996]. Extended Link Ratio techniques, including weighted least squares regression, have been shown to be effective in handling various insurance lines of loss triangle data [1, Barnett and Zehnwirth 2000].

Some methods result in similar reserve estimates while having different theoretical basis. The derived error of prediction of multiple methods is used for comparison [4, England and Verrall 2002], but little attention is paid to the actual future losses once they have developed, namely because most papers use current data, and the true future losses are unknown.

Recent developments in the area of loss reserving have focused on how well the various methods work on large volumes of data as well. Often different techniques work well on a single triangle but are not adequate in large scale application. As a result, improving the model or incorporating more information into the existing model has been proposed [8, Meyers 2012].

Incorporating the analysis of multiple methods and providing an estimate of future losses for each accident year and thus incurred but not reported (IBNR) totals is often ignored in research of new methods. Instead, current research focuses on the viability and theoretical basis of a single or related methods while allowing that adjustments to the theoretical predictions will be made by an actuary with expert knowledge of loss development and ultimate loss factors [4, England and Verrall 2002], rather than attempting to find a method that works well in a pre-determined insurance line. This thesis focuses on using techniques based on fully known commercial auto liability loss data [3, Casualty Actuarial Society] to compare the performance of the estimates, rather than allowing for adjustments to be made post-prediction based on expert knowledge in the field of auto liability.

More recently, Bayesian inference has been used to fit a distribution to the data. These techniques generate a conditional distribution on the known data, often through Markov Chain Monte Carlo simulation, and attempt to explain the uncertainty of the future events. Bayesian inference has been used in many types of actuarial problems, and is well suited in the model-based predictions in this field [11, Scollnik 2001].

Other applications of Bayesian inference focus on the dependence between multiple lines of insurance for a company. This multivariate approach analyzes not just corresponding cells in different triangles, but also the year to year factors related to policy shifts [12, Shi, Basu, and Meyers 2012].

Bayesian techniques can also be applied with general linear mixed models to model outstanding claim counts and amounts. These techniques are applied to data of individual claims and discard the triangular data frame when individual claim amounts are available. These models require much more data concerning the losses from a line of insurance than more conventional methods [6, Jemilohun, Lawl, and Adebara 2013].

While the theoretical derivation of newer and possibly improved techniques is necessary in any field, the application of existing techniques in practice and analysis of the results will show how current methods compare. Ultimately any technique used to forecast necessary reserves will be judged on the accuracy of predictions once the losses are known. That is the aim of this thesis, to apply these methods to a data set of known losses and assesses the accuracy of the resulting estimates.

3. LOSS TRIANGLES

Insurance claims are rarely settled immediately or totally. There is often a lag between a claim and the ultimate development of the payment on the claim. This lag can arise from a number of factors such as continued medical bills or salvage recouped from a car after any investigation is completed. Furthermore, the lag can vary greatly for different lines of insurance. An actuary's job is to identify and estimate these developing claims and ensure enough reserves are available for possible future losses from the current accident year's premiums.

These developing losses naturally form an upper left hand triangle of cumulative incurred claims. Table 3.1 is a simple triangle created to familiarize the reader with the format of the data in this thesis. Each row has one less entry than the previous accident year because the claims have had one less year to develop. The columns correspond to development periods, which in this thesis are always years. However, these periods can vary, and are not always full years.

Table 3.1: A theoretical loss triangle

	Development Year				
Accident Year	1	2	3	4	5
1991	$C_{1,1}$	$C_{1,2}$	$C_{1,3}$	$C_{1,4}$	$C_{1,5}$
1992	$C_{2,1}$	$C_{2,2}$	$C_{2,3}$	$C_{2,4}$	
1993	$C_{3,1}$	$C_{3,2}$	$C_{3,3}$		
1994	$C_{4,1}$	$C_{4,2}$			
1995	$C_{5,1}$				

$C_{i,j}$ denotes cumulative losses from year i at development period j

As these claims develop, a new entry is added to all rows each calendar year. It is necessary to keep claims from different accident years separate to avoid subsidizing from premiums other than the year in which the accident occurred. The simplest method of estimating these future losses, and through them the ultimate losses, uses some information on the ratios from year to year. Equation 3.1 shows how each ratio, r_{ij}, is calculated from a triangle of known losses (Table 3.2). Each r_{ij} is the cumulative losses in year i at development period j. An actuary will pick an appropriate ratio, either one from the known data or some approximation he/she deems sufficient to cover the current exposure from that line [15, Werner and Modlin 2010]. Table 3.3 contains the development ratios by accident and development year for this simple triangle.

$$r_{i,j} = \frac{C_{i,j+1}}{C_{i,j}} \qquad i = \mathit{1{:}4},\ j = \mathit{1{:}4} \tag{3.1}$$

Here the actuary could pick the median ratio, the largest ratio, or any combination of these ratios from each development period. This license with the ultimate reserve forecast is necessary, as an actuary must identify years that do not fit a pattern and should therefore be handled carefully when estimating these reserves. An actuary has expert knowledge of the way these losses can develop as well as access to information surrounding these losses, such as policy or weather behavior associated with an accident year. Shock losses can occur when an unfortunate set of circumstances lead to an inflated number or size of claims. A hail storm can lead to a drastic spike in auto claims, while a severe drought can lead to low yields in agriculture and result in large crop insurance claims. The year to year factors vary heavily, but these shock losses can lead to over-cautious reserves if an actuary does not identify the circumstances surrounding the spike in losses.

Table 3.2: A simple loss triangle

	Development Year				
Accident Year	1	2	3	4	5
1991	400	700	850	930	1000
1992	480	790	1000	1140	
1993	500	950	1190		
1994	570	1050			
1995	600				

Table 3.3: Ratios of losses of the simple example triangle

	Development Period			
Accident Year	1	2	3	4
1991	1.75	1.21	1.09	1.08
1992	1.65	1.27	1.14	
1993	1.9	1.25		
1994	1.84			

An estimation of future losses are represented in Table 3.4.

Table 3.4: Predicted future losses

	Development Year				
Accident Year	1	2	3	4	5
1991	400	700	850	930	1000
1992	480	790	1000	1140	*1231*
1993	500	950	1190	*1356*	*1465*
1994	570	1050	*1313*	*1496*	*1616*
1995	600	*1104*	*1380*	*1573*	*1699*

Predicted future losses (italics) using the second largest ratio for each of the first two development periods and the largest ratio for the last two development periods.

It is easy to see that $C_{5,2} = r_{5,2} * C_{5,1} = 1.84 * 600 = 1104$. Similarly, the rest of the future losses can be filled in, working from left to right across each row. These figures have little theoretical basis however, and rely more on an actuary's intuition and experience. In an effort to reduce this subjective aspect and the unquantified variability therein, a chain ladder method was proposed by Thomas Mack [7, Mack 1993]. A data set that will be used to demonstrate the various methods is introduced in the following section.

4. THE CHAIN LADDER METHOD AND LEAST SQUARES REGRESSION

4.1. Mack's Method

Mack proposed using a common ratio between development years to predict losses [7, Mack 1993]. In general this is referred to as the chain ladder method, with Mack's Method also providing a formula for the standard error of the predictions. There are three underlying assumptions of Mack's Method; 1) independence of accident years, 2) independence of development factors, and 3) the variance of a prediction, $C_{i,k+1}$ is inversely proportional to the previous development period's losses, $C_{i,k}$. The Mack Method estimates a development factor for each development year. Similar to the first method introduced, to obtain an estimate of the future losses, the most recently developed year must simply be multiplied by a development factor to obtain the losses of the next year. This can be extended to fill in all missing entries in a loss triangle. Equation 4.1 shows how we can estimate a future loss, $C_{i,k+1}$, using known losses, $C_{i,k}$, and a development factor, f_k.

$$C_{i,k+1} = C_{i,k} * f_k \tag{4.1}$$

To calculate these development factors using Mack's Method, we sum the losses from subsequent years and find the overall ratio (Equation 4.2). We use Equation 4.2 with the cumulative triangle entries of Table 4.1 to calculate these development factors.

Table 4.1: A simple loss triangle

	Development Year				
Accident Year	1	2	3	4	5
1991	400	700	850	930	1000
1992	480	790	1000	1140	
1993	500	950	1190		
1994	570	1050			
1995	600				

$$f_k = \frac{\sum_{i=1}^{n-k} C_{i,k+1}}{\sum_{i=1}^{n-k} C_{i,k}} \qquad \textit{for k=1:4 and n=5} \tag{4.2}$$

We see the first development factor, $f_1 = \frac{700+790+950+1050}{400+480+500+570} = \frac{3490}{1950} = 1.79$. Continuing with Equation 4.2, all four development factors for this triangle are calculated, with the final results shown in Table 4.2.

Table 4.2: Development factors

$f_1 = 1.79$	$f_2 = 1.25$	$f_3 = 1.12$	$f_4 = 1.08$

All future losses can be estimated using these four development factors and Equation 4.1. Table 4.3 shows all future losses estimated using Mack's Method.

Table 4.3: Estimation of future losses with Mack's Method

		Development Year			
Accident Year	1	2	3	4	5
1991					
1992					1231.2
1993				1332.8	1439.4
1994			1312.5	1470	1587.6
1995		1074	1342.5	1503.6	1392.2

From these estimations, the ultimate losses can be estimated for each accident year, assuming the losses fully develop over five years. The ultimate losses are the final cumulative entry from development year 5 in the triangle. Most triangles take longer than 5 years to develop. Table 4.4 is a 10 by 10 triangle of losses that demonstrates a longer development of losses [13, R Core Team 2013] using Mack's Method.

The losses might still not be fully developed, since the final development factor is still not 1.000. It is however much closer than with the smaller 5 by 5 triangle (Table 4.1). An actuary could include a tail factor to account for any future losses past a triangle's development years. Tail factors are ignored in this thesis, and final year losses are considered ultimate losses.

Mack's Method is but one way to predict these unknown future losses. In the following section least squares regression is introduced to obtain predictions of future losses.

Table 4.4: A triangle from the Chain Ladder Package [13, R Core Team 2013]

	Development Year									
Accident Year	1	2	3	4	5	6	7	8	9	10
1991	5012	8296	10907	11805	13539	16181	18009	18608	18662	18834
1992	106	4285	5396	10666	13782	15599	15496	16169	16704	*16858.0*
1993	3410	8992	13873	16141	18735	22214	22863	22466	*23863.4*	*24083.4*
1994	5655	11555	15766	21266	23425	26803	27067	*27967.3*	*28441.0*	*28703.1*
1995	1092	9565	15836	22169	25955	26180	*27277.9*	*28185.2*	*28662.6*	*28926.7*
1996	1513	6445	11702	12935	15852	*17649.4*	*18389.5*	*19001.2*	*19323.0*	*19501.1*
1997	557	4020	10946	12314	*14428.0*	*16063.9*	*16737.6*	*17294.3*	*17587.2*	*17749.3*
1998	1351	6947	13112	*16663.9*	*19524.7*	*21738.5*	*22650.1*	*23403.5*	*23799.8*	*24019.2*
1999	3133	5395	*8758.9*	*11131.6*	*13042.6*	*14521.4*	*15130.4*	*15633.7*	*15898.5*	*16045.0*
2000	2063	*6187.7*	*10045.8*	*12767.13*	*14958.9*	*16655.0*	*17353.5*	*17930.7*	*18234.4*	*18402.4*

Estimated losses in italics.

4.2. Weighted Least Squares Regression

Least squares regression describes a relationship between a response and a predictor variable. The simple case of a slope only model, where data is fit to the equation $\mathbf{y} = \beta_1 * \mathbf{x} + \boldsymbol{\epsilon}$ will be evaluated. Here $\mathbf{y}$ is the response or dependent variable, the next year's claim amount. The predictor or independent variable is $\mathbf{x}$, the observed claim amount for the most recent year. The slope of the line through the origin that best connects $\mathbf{x}$ and $\mathbf{y}$ is represented by β_1 and is referred to as the development factor. The error term is $\boldsymbol{\epsilon}$. An intercept can be included, β_0, but is assumed to be zero in the slope only model.

In theory, all accident years are independent, all development periods are independent, and the error term, $\boldsymbol{\epsilon}$, is normally distributed with mean 0 and constant variance, σ^2, that is $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$. These assumptions are often violated in an applied setting however. The development years are not independent, because we may be modeling claims that are developing over multiple years. The accident years are also not necessarily independent, since roughly the same cohort of people are insured from year to year, possibly having accidents and incurring claims in multiple years. There is no simple fix for either of these violated assumptions.

Heteroscedasticity is the increasing or decreasing variance of residuals as the fitted value or independent variable increases. Heteroscedasticity thus violates the third underlying assumption of constant variance of the error term in this model. The reason weighted least squares are used in chain ladder modeling is to address the problem of heteroscedasticity. In a general case of a least squares regression model, equal weights are assigned to all data points. Two other cases of weights are considered and the results of the three techniques are compared when applied to a test data set. To differentiate between the methods, equal weights will be called ordinary least squares (denoted OLS). The other two cases will consider weights $\frac{1}{x_i}$ (called the classical ratio estimator and denoted CRE (Knaub 2005)) and $\frac{1}{{x_i}^2}$ (denoted Method 3) . It should be noted that in these models, the β_1 parameter can be referred to as the slope parameter or development factor. The CRE method decreases the variance as the fitted values increase by decreasing the weight of the larger observations, and Method 3 decreases the variance more drastically. Figure 4.1 shows an example of this heteroscedasticity and the changing trends in residual variance as different weights are used. These graphs show only one development period from one triangle from the data from the National Association of Insurance Commissioners (NAIC), as this is a triangle-by-triangle problem that varies for each triangle analyzed. Some triangles might not exhibit heteroscedasticity when using the OLS method. The CRE method could adequately fix the heteroscedasticity, and method 3 might over-correct the problem, creating a decreasing trend in the variance of the residuals.

To obtain an estimate of β_1 a function of the residuals, $F = \sum w_i(y_i - (\hat{\beta}_1 x_i))^2$ is minimized for $i = 1, \ldots, n$. This function represents the deviance from the observations, y_i and the value of the line fit to the data, $\hat{y}_i$. Each w_i represents the weight associated with the i^{th} observation. To minimize the function F and thus minimize the deviance, the derivative of F is taken with respect to β_1 to obtain an explicit equation for parameter estimates of F when the derivative is set equal to zero, as follows:

$$\frac{\delta F}{\delta \beta_1} = \sum x_i w_i (y_i - (\hat{\beta}_1 x_i)) = 0.$$

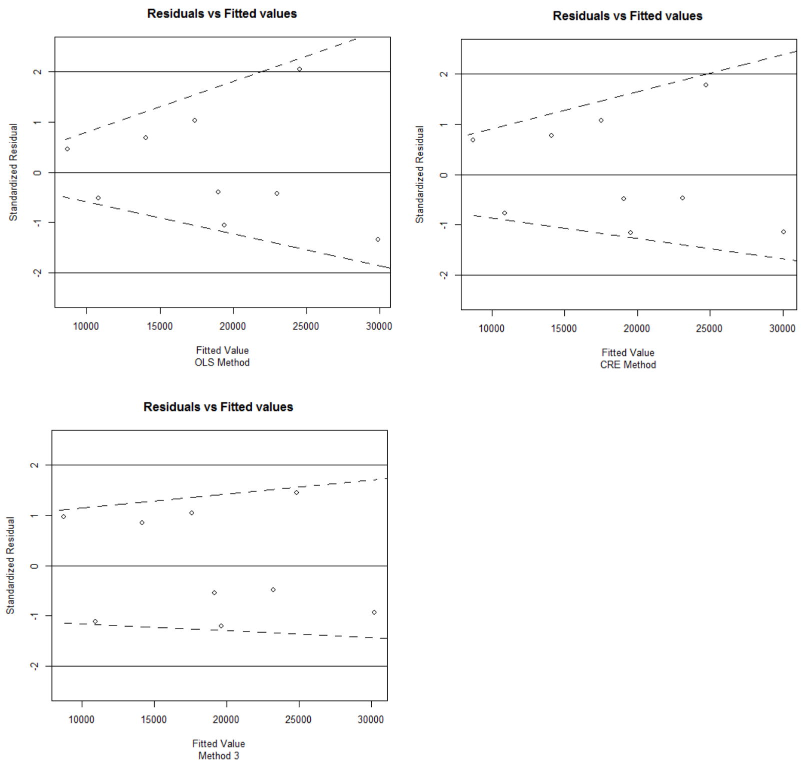

Figure 4.1: An example showing how heteroscedasticity can be addressed

Solving this equation for $\hat{\beta}_1$, the weighted least square estimate of the slope parameter is obtained:

$$\hat{\beta}_1 = \frac{\sum w_i x_i y_i}{\sum w_i x_i^2}. \tag{4.3}$$

Recall the three cases of weights introduced at the beginning of this section. For each weight, a new estimate is obtained by inserting different values for w_i into Equation 4.3, resulting in three estimates of $\hat{\beta}_1$ as

$$\text{OLS: } \hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \qquad \text{CRE: } \hat{\beta}_1 = \frac{\sum y_i}{\sum x_i} \qquad \text{Method 3: } \hat{\beta}_1 = \frac{\sum \frac{y_i}{x_i}}{n}.$$

To obtain development factor $\hat{\beta}_i$, $\mathbf{x}$ will be the j^{th} column, and $\mathbf{y}$ will be the $(j+1)^{th}$ column. The CRE formula for $\hat{\beta}_1$ corresponds to the development factor from Mack's Method

(Equation 4.2) where the notation $C_{\cdot,k} = \mathbf{x}$ and $C_{\cdot,k+1} = \mathbf{y}$. The three cases of weighted least squares are used to reference the different methods in the rest of this thesis.

5. NAIC DATA ANALYSIS

5.1. Data Introduction

Data used in this thesis are extracted from 78 triangles for Commercial Auto Liability, provided by the National Association of Insurance Commissioners (NAIC) as being available on Casualty Actuarial Society (CAS) webpage (www.casorg.com). The NAIC is the national regulatory body system for the United State insurance market consisted of the chief insurance regulators from the 50 states, the District of Columbia and five U.S territories. The members of NAIC, elected or appointed state government officials along with their state insurance department and stuff, regulate the conduct insurance companies and agents in their respective state. For example, in state of North Dakota, the Department of Insurance is located in Bismarck and insurance companies operating in the state must have their rate filings and underwriting policies reviewed and approved by the State Department before the implementation.

The mission of NAIC is to assist state insurance regulators in supporting and improving state regulation of insurance as well as to serve public interest in achieving the fundamental insurance regulatory goals such as protecting public interest, promote competitive markets, facilitate fair and equitable treatment of insurance consumers, promote reliability, solvency of insurance companies.

All insurance companies are required by law to file quarterly and annual financial statement to NAIC. Annual statements are reported in spring of each calendar year and they include a book full of different exhibits some with multiple parts from income statement, cash flow, underwriting and investment exhibits, exhibits of premium and losses, reinsurance exhibits (Schedule F), analysis of losses and loss expenses by line of business (Schedule P), exhibits on premium written (Schedule T), investment exhibits (Schedule D), etc.

The NAIC loss triangles provided on CAS webpage were extracted from Schedule P of the Annual Statement of 158 companies. Schedule P reports Analysis of Losses and Loss Expenses. Glenn G. Meyer, PhD, FCAS and Peng Shi, PhD, ASA coordinated a project between CAS and NAIC with a goal to make loss triangles data available to all interested researchers for purpose of testing various methods for estimating the Incurred But Not Reported (IBNR) losses.

The Schedule P data provided by NAIC include: name and group code of each NAIC insurance company organized, accident year, development year, bulk loss, incurred paid loss, cumulative incurred losses, cumulative paid losses, and posted reserves as of 1997.

The data used in this thesis include cumulative paid losses from Schedule P for 78 out of 158 companies in period 1988-1997 (ten development years) . The triangles showing negative or zero losses are disregarded from the analysis. In addition to the upper triangles, this data also includes the lower triangles. For example, the data from accident year 1989 was pulled from Schedule P of year 1998, the data from accident year 1990 was pulled from Schedule P of year 1999, ……., the data for accident year 1997 was pulled from Schedule P of year 2006. Hence the lower triangles can be used for purpose of model validation. Extensive data validation and quality control measures were performed by project coordinators to insure reliability and quality of the data.

As part of this thesis, several R functions are built to pull together a large volume of data from Schedule P and construct individual loss triangles for 78 NAIC companies. The CAS website provides some R code to summarize these triangles [3, Casualty Actuarial Society], subsidized with my own code to manipulate the data into the triangle format used in this thesis. The code is available per request.

5.2. Assessing Accuracy of IBNR Totals for NAIC Data

From the NAIC database, 78 triangles from different companies were selected from the commercial auto liability line of insurance. The bottom right triangle of losses in these full matrixes are censored to obtain upper left triangles of losses because all losses are known. Then the three techniques are applied and the resulting estimates are compared against the true losses censored from the original matrix. An example of one of these fully known loss triangles is found in Table 5.1.

In much of the analysis, ratios of observed to expected losses are used to account for the different orders of magnitude of the different triangles. Some triangles have losses that range from 10-60, while others have losses in the order of 20,000. This could be the result of clerical policies, reporting losses in the hundreds or thousands, or it could be a result of the companies being different sizes and having different levels of exposure. Comparing the difference between the estimates and the true values is not as meaningful as comparing the ratios because of this discrepancy.

Table 5.1: One NAIC triangle of fully known losses

	Development Year									
Accident Year	1	2	3	4	5	6	7	8	9	10
1988	952	1529	2813	3647	3724	3832	3899	3907	3911	3912
1989	849	1564	2202	2432	2468	2487	2513	2526	2531	2527
1990	983	2211	2830	3832	4039	4065	4102	4155	4268	4274
1991	1657	2685	3169	3600	3900	4320	4332	4338	4341	4341
1992	932	1940	2626	3332	3368	3491	3531	3540	3540	3583
1993	1162	2402	2799	2996	3034	3042	3230	3238	3241	3268
1994	1478	2980	3945	4714	5462	5680	5682	5683	5684	5684
1995	1240	2080	2607	3080	3678	4116	4117	4125	4128	4128
1996	1326	2412	3367	3843	3965	4127	4133	4141	4142	4144
1997	1413	2683	3173	3674	3805	4005	4020	4095	4132	4139

First, the three methods are applied to the NAIC upper left triangles, to ensure adequate fits by analyzing the residuals. Standardized residuals were calculated on the upper left of all NAIC triangles on the elements used in the weighted least squares regression. This means there are nine residuals for the first development period, eight residuals for the second, and so on. Figure 5.1 shows the standardized residuals plotted by development period. Since these techniques are applied to 78 triangles, and each triangle has 44 standardized residuals, we have a total of $78 * 44 = 3432$ residuals plotted on each graph. Almost all residuals have magnitude less than 2.

Investigation into the standardized residuals that were greater than 2 reveals where the majority of these outliers originate. Table 5.2 shows one NAIC triangle and Table 5.3 the standardized residuals of the same triangle.

Before looking at these residuals, it is apparent that the first two accident years have losses that are an order of magnitude larger than the subsequent eight years. If an actuary were trying to estimate necessary reserves for this company, he/she would need to investigate why the losses were so much higher in the first two years. There could have been a policy shift or the lines of insurance could have been re-categorized and resulted in losses being reported through different lines of insurance, or the company might have reduced their exposure to account for losses one tenth or less in the last eight years of the loss triangle. In this case, the residuals were calculated with the OLS method, using equal weights. This is just one example of a triangle with standardized residuals that could indicate outliers. The OLS method was used for this example because it performs poorly and has the most residuals that require inspection with this data.

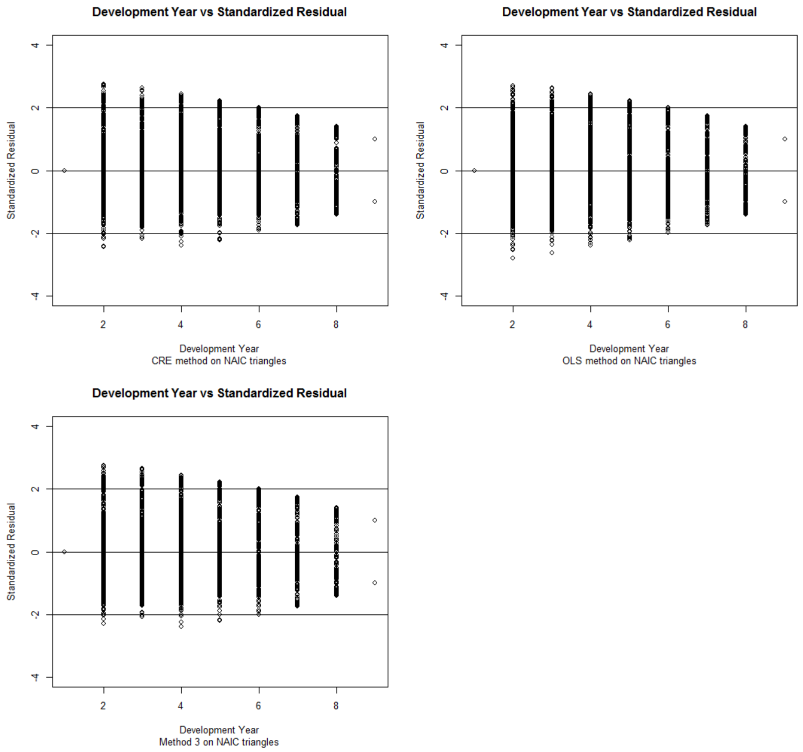

Figure 5.1: Residual plots of NAIC triangles using all three methods

The number of points to model decrease as we move left to right across these triangles. With nine points, the first development period has the most residuals to study, and even these nine do not reveal which method is the best to use in each case. The example from Figure 4.1, page 13, only graphed the first two columns from one triangle, to illustrate a case where it was easy to see the problem of heteroscedasticity and the improvement when different weights were used. Analyzing the residuals has shown us that these methods all seem to be reasonable, while selecting the method that yields homoscedastic residuals improves the validity of our model by satisfying an underlying assumption.

It is important to analyze the predictions and compare them to the known values from the lower right triangles with the NAIC data. Weighted least squares regression provides the equation of a line that is the expected mean of the losses of the next development period. Because the full

Table 5.2: An NAIC triangle with residuals of magnitude greater than 2

	Development Year									
Accident Year	1	2	3	4	5	6	7	8	9	10
1988	128	229	269	308	342	368	391	400	426	432
1989	170	285	323	355	368	396	406	420	420	
1990	10	31	41	60	61	60	60	60		
1991	11	12	13	26	26	26	26			
1992	7	11	11	11	11	11				
1993	11	16	17	17	17					
1994	10	15	17	17						
1995	15	35	40							
1996	11	29								
1997	26									

Table 5.3: OLS standardized residuals from one NAIC triangle with residuals of magnitude greater than 2

	Development Year									
Accident Year	1	2	3	4	5	6	7	8	9	10
1988	-	1.25	2.00	0.97	2.23	0.24	1.71	-1.17	1	-
1989	-	-1.52	-2.14	-1.32	-2.08	0.29	-1.60	1.31	-1	
1990	-	1.63	1.51	1.66	-0.39	-1.87	-0.42	-0.67		
1991	-	-0.82	-0.23	1.35	-0.22	-0.65	-0.18			
1992	-	-0.13	-0.47	-0.16	-0.09	-0.27				
1993	-	-0.35	-0.40	-0.24	-0.15					
1994	-	-0.26	-0.07	-0.24						
1995	-	1.09	-0.07							
1996	-	1.19								
1997	-									

losses from the NAIC data are available, it is possible to assess how these predictions are performing by creating a ratio of the observed losses to the predicted losses. A ratio is used to standardize the error of the predictions, as the order of magnitude of the NAIC triangles varies greatly. If the ratio of observed over predicted losses is greater than one, the observed losses exceeded the predicted losses and therefor the prediction was inadequate. If the ratio is less than one, the prediction was conservative. The expectation is that these ratios will be evenly split, roughly half greater than one and half less than one. However, it is important to note that in application, it wouldn't be prudent to only predict adequate reserves half of the time. At this point however the focus is only with the accuracy of the predictions.

Figure 5.2, Figure 5.3, and Figure 5.4 show ratios of observed over expected losses for all three methods from the NAIC data, sorted by calendar year. The first calendar year, when $x = 1$ on the graph, represents the losses that will manifest next year. A lower right to upper left diagonal is added each additional year's losses are reported.

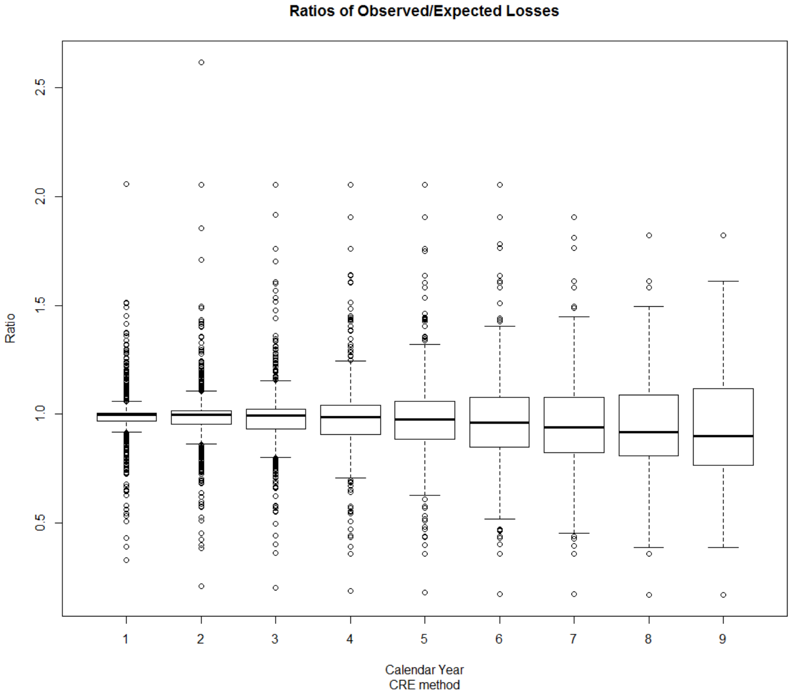

Figure 5.2: Ratios of observed over expected losses by calendar year of all NAIC data; CRE method

The losses corresponding to the first calendar year of losses are from cells [10,2], [9,3], [8,4], . . . , [2,10]; one off the diagonal. The second development year corresponds to the next off-diagonal, consisting of cells [10,3], [9,4], . . . , [3,10]. Table 5.4 shows a triangle of ratios with the first and the fourth calendar year diagonals bold to illustrate how the calendar year losses are found.

From Figure 5.2, the variance of the ratios increases as the calendar year increases. This follows because the second calendar year prediction is using a prediction from the first calendar year as a basis for the second year estimation. Each calendar year adds another source of error, therefore increasing the overall error with each subsequent prediction. These ratios are quite close however, and the first three calendar years are quite well predicted in the vast majority of these NAIC triangles.

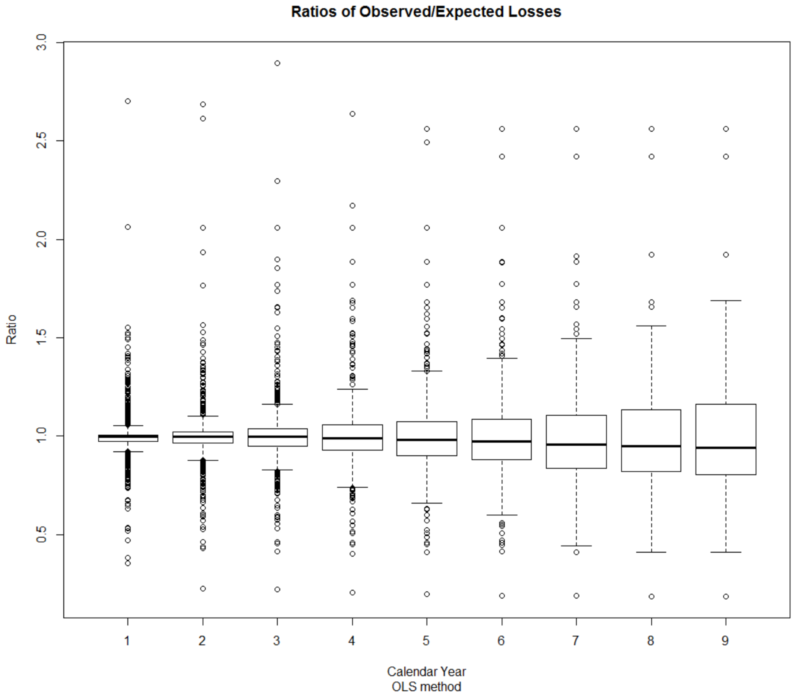

Figure 5.3: Ratios of observed over expected losses by calendar year of all NAIC data; OLS method

Comparing the ratios from the three methods, we see from Figure 5.2, Figure 5.3, and Figure 5.4 that the CRE method seems to have the best ratios of these three methods because of its smaller variance and predictions centered around one. The OLS method has the largest and most frequently outlying ratios, implying the OLS method had the most under-predicted losses. The difference between the CRE method and Method 3 is small, but the CRE method has shorter intervals in the later calendar years, implying more stable predictions when comparing them to the observed losses.

Ultimate losses are also important to gauge. It is important to estimate each yearly loss from each accident year, but the accuracy of our predictions can be summarized by looking at fully developed losses (development year 10). Figure 5.5 shows a boxplot of the ratio of ultimate observed losses and ultimate predicted losses by accident year. Each accident year has one less year of known losses and one more year of predictions which explains the increase in variance of the observed/expected loss ratio. Figure 5.5, Figure 5.6, and Figure 5.7 show box plots of these ultimate observed losses over ultimate predicted losses sorted by accident year for the three methods. The OLS method is performing the worst, the CRE method is the most accurate, and Method 3 has more over-predictions, implying Method 3 has more conservative estimates of ultimate losses for each accident year than the other two methods.

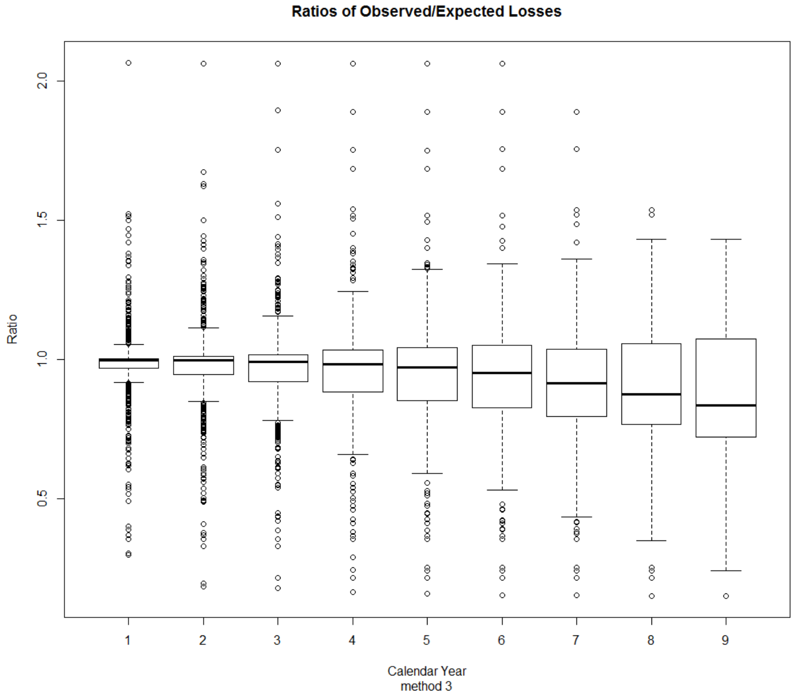

Figure 5.4: Ratios of observed over expected losses by calendar year of all NAIC data; Method 3

When looking at the ratios, the extremely large and small values merit a closer look. The largest ratios, the ratios over 2, were only present in two of the 78 triangles. They were caused by shock losses in the lower right triangle, in triangles that were almost fully developed after two or three years. Since the triangle was almost fully developed, a shock loss means all subsequent predictions will be low, because the triangles are cumulative. The ratios less than .5 were also from a single accident and development year of losses that were surprisingly low (the opposite of a shock loss, although there is no conventional term for that). Each subsequent loss is then affected, creating a row of ratios that are low. The extreme ratios almost exclusively occur in the last two accident years, where only one or two years of developed losses are known.

Table 5.4: Ratios of observed over expected losses

	Development Year									
Accident Year	1	2	3	4	5	6	7	8	9	10
1988	1	1	1	1	1	1	1	1	1	1
1989	1	1	1	1	1	1	1	1	1	**1.000**
1990	1	1	1	1	1	1	1	1	**1.000**	1.000
1991	1	1	1	1	1	1	1	**1.002**	1.002	1.002
1992	1	1	1	1	1	1	**1.000**	1.002	1.002	**1.002**
1993	1	1	1	1	1	**1.008**	1.215	1.220	**1.220**	1.220
1994	1	1	1	1	**0.948**	0.948	0.948	**0.953**	0.953	0.953
1995	1	1	1	**0.787**	0.776	0.776	**0.776**	0.777	0.777	0.777
1996	1	1	**1.227**	1.011	1.024	**1.024**	1.024	1.035	1.035	1.035
1997	1	**1.068**	0.937	0.847	**0.825**	0.825	0.825	0.827	0.827	0.827

Calendar years 1 and 4 diagonals bold.

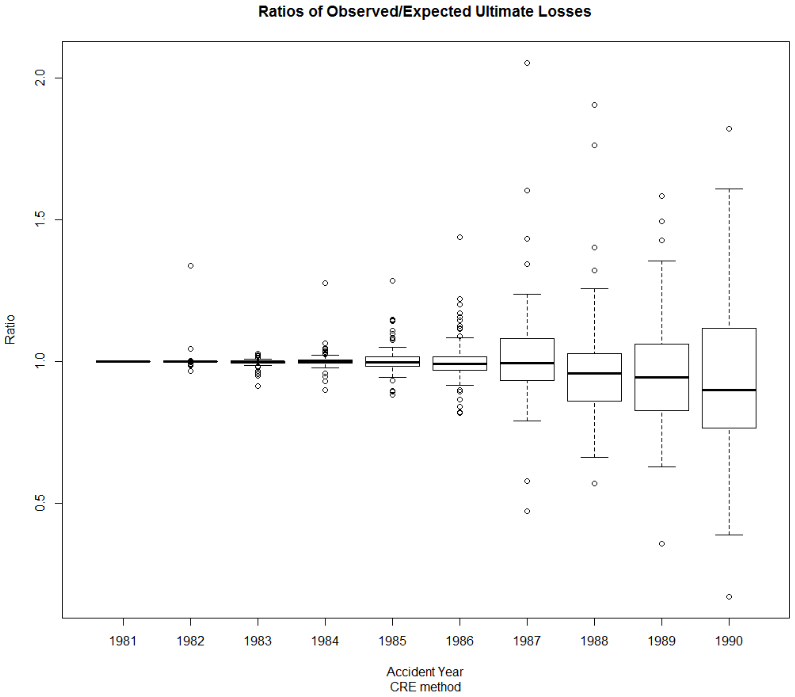

Figure 5.5: Ratios of observed over expected ultimate losses by accident year of all NAIC data; CRE method

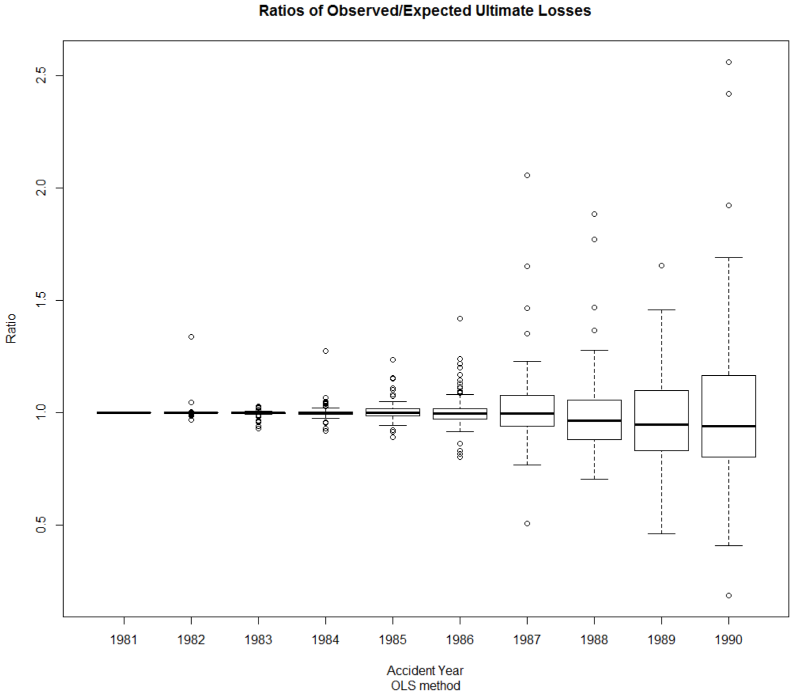

Figure 5.6: Ratios of observed over expected ultimate losses by accident year of all NAIC data; OLS method

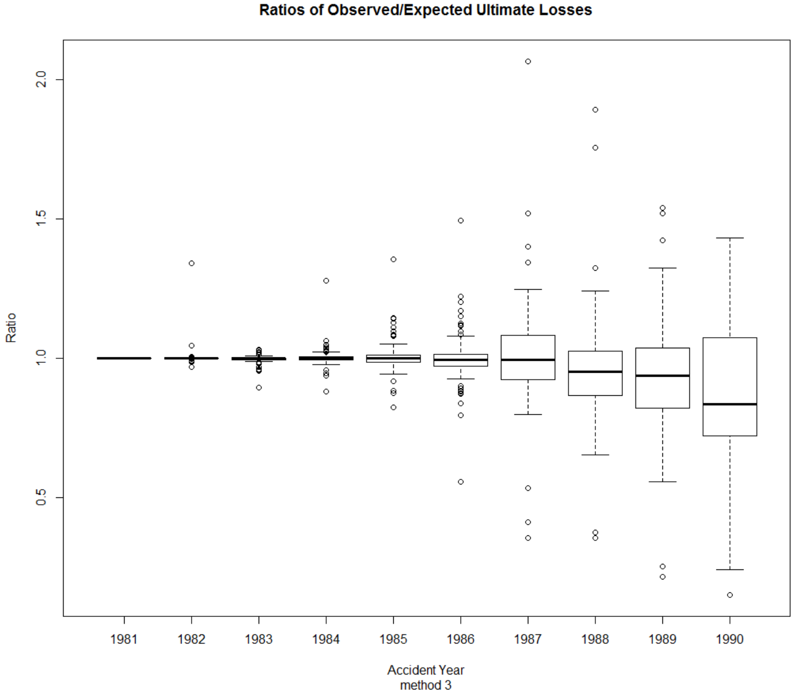

Figure 5.7: Ratios of observed over expected ultimate losses by accident year of all NAIC data; Method 3

5.3. IBNR Totals

Estimating the incurred but not reported losses (IBNR) for each calendar year and the total incurred but not reported outstanding claims for a particular line of insurance allows an actuary to make recommendations for appropriate rates. Table 5.5 shows an example of how IBNR projections are calculated. The last known loss from an accident year is subtracted from the projected ultimate losses and the total is the IBNR losses from that accident year. The first accident year has zero IBNR losses, since it is assumed the losses are fully developed after ten years. The second accident year on the triangle, 1989, has 0.65 IBNR losses. Since the losses in the tenth development year are unknown, they must be estimated. Using the CRE method, ultimate losses are predicted to be 2531.65. Since the triangles are cumulative, subtraction yields the value 0.65 as the yet to be reported losses in the tenth development year. As the accident year increases, more years are unknown, resulting in higher IBNR values each year. The total IBNR from this triangle is found by summing all ten individual accident year IBNR values. Summing the last column of Table 5.5, the total IBNR for this line of insurance is 6576.44. Looking at the cumulative losses both known and projected ultimate losses, 6576.44 is a large sum of future losses for this insurance company to pay out. The IBNR total is larger than any single year's ultimate losses. The importance of accurate reserving is apparent when comparing this total IBNR to the losses of any accident year. A company with inadequate reserves could unknowingly be subsidizing past years' losses with current premiums.

Table 5.5: An example of IBNR totals

	Development Year											
Accident Year	1	2	3	4	5	6	7	8	9	10	Projected Ultimate Losses	IBNR
1988	952	1529	2813	3647	3724	3832	3899	3907	3911	3912	3912	0
1989	849	1564	2202	2432	2468	2487	2513	2526	2531		2531.65	0.65
1990	983	2211	2830	3832	4039	4065	4102	4155			4161.88	6.88
1991	1657	2685	3169	3600	3900	4320	4332				4369.71	37.71
1992	932	1940	2626	3332	3368	3491					3555.40	64.40
1993	1162	2402	2799	2996	3034						3212.87	178.87
1994	1478	2980	3945	4714							5166.53	452.53
1995	1240	2080	2607								3441.64	834.64
1996	1326	2412									4209.55	1797.55
1997	1413										4616.22	3203.22

Calculated using the CRE method.

Since the total losses are known, the projected IBNR total and the actual IBNR total can be compared. From the corresponding full NAIC triangle, the total IBNR was actually 7399. Subtracting the projected total from the known total, $7399 - 6576.44 = 822.56$, the estimated IBNR was 822.56 lower than the actual IBNR value, implying the total was underestimated. Figure 5.8 is a box plot of these differences on all 78 analyzed NAIC triangles.

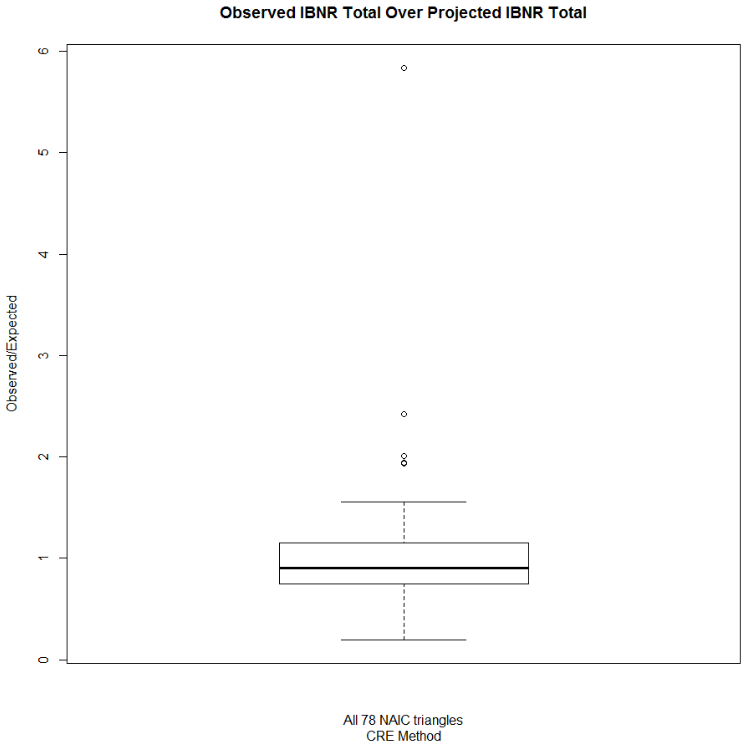

Figure 5.8: Observed-Expected IBNR totals for all NAIC triangles; CRE method

Almost all these ratios fall within the 1.5 IQR whiskers in the box plot graph. Inspection of the largest ratio, 5.84, revealed the triangle had three accident years that were an order of magnitude larger than the other seven, and further, those accident years were all within the lower half of the triangle. These losses were similar to the outliers described earlier in the analysis, and again present a case where the actuary would need to take note of the erratic behavior of the losses and investigate the cause before deciding how to handle them. As a company grows the volume of losses would naturally increase as exposure increases, which is another possible cause of larger values occurring in the latest accident years of this triangle and would also require special handling.

6. OUTLIERS

Closer inspection of the largest and smallest of the IBNR total ratios in Figure 5.8 is warranted. The five largest and five smallest IBNR ratios were analyzed, as well as five triangles with ratios about the median. Visual inspection of graphs of development year one versus development year two, two vs three, and three vs four yielded no clear trends or distinction between the lowest five, the highest five, or the middle five. Two techniques were used to try to obtain better estimates on the triangles of extreme ratios, the nearest neighbor method and mixtures of regression.

If it is possible to improve the earlier calendar year predictions, especially in the later accident years, the compounding of the error will be reduced. The nearest-neighbor grouping technique and a mixture of regression used in the first two to three development periods could possibly improve predictions in the latest three accident years for the first three development years (improving predictions from cells [10,2-4], [9,3-4], and [8,4]) and thus improve predictions all through the last three accident years (the years in which the most predictions and thus the largest IBNR contributions originate). Explanations and examples of each of these procedures follow.

6.1. Nearest Neighbor Method

The nearest neighbor grouping technique can be used to find the best weight to use with weighted linear regression when modeling heteroscedastic small data sets [15, Wuthrich and Merz 2008]. It uses natural breaks in the independent variable to group observations. For the first development period on a triangle, there are nine pairs of points to model. With the nearest neighbor technique, the OLS method is used to calculate residuals for each fitted value. These residuals are then plotted against the indepentend variable, and the observations are grouped. The purpose is to find a weight that will make the variances of each neighborhood of residuals approximately equal. The weight may differ and often does differ between development periods in this technique. Since the number of points decreases, this technique is only applied to the first three development periods on a triangle, hopefully improving the early predictions in the last accident years, as described above. The following example explains the procedure.

Table 6.1 shows the values of x and residuals of the first development period of one NAIC triangle, sorted by increasing x value. These are graphed in Figure 6.1. The residuals clearly exhibit increasing variance as the independent variable increases, implying weighted least squares could possibly fix this heteroscedasticity.

Next, the independent variable must be separated into groups, ideally of similar sizes, creating neighborhoods of x. Consider grouping the four smallest values and the five largest values of x. Then, the mean of each group of x is calculated, as well as the variance of the residuals of each group. Here $\bar{x}_1 = 6428$, $\bar{x}_2 = 11700$, $s_1^2 = 1339881$, and $s_2^2 = 4764481$. To identify the appropriate weights to use, the quantities $\frac{s_1^2}{f(\bar{x}_1)}$ and $\frac{s_2^2}{f(\bar{x}_2)}$ must be equal for some function of the mean of each group. The functions considered are the mean of each group raised to different exponents. In this thesis, the different functions the mean considered are $f(\bar{x}) = \bar{x}^i$ for $i = 0, 0.25, 0.5, 0.75, \dots 4$. Solving the equality $\frac{s_1^2}{f(\bar{x}_1)} = \frac{s_2^2}{f(\bar{x}_2)}$ results in $\frac{s_1^2 f(\bar{x}_2)}{s_2^2 f(\bar{x}_1)} = 1$ or $\frac{s_2^2 f(\bar{x}_1)}{s_1^2 f(\bar{x}_2)} = 1$.

Table 6.1: The independent variable and corresponding residuals

x	4381	5456	7083	8793	9586	9800	11618	12402	15095
Residual	825	-919	1183	1773	-689	-1803	-717	3350	-2087

One of these quantities will have a positive slope and the other will have a negative slope. Considering these reciprocals, $max(\frac{s_1^2 f(\bar{x}_2)}{s_2^2 f(\bar{x}_1)}, \frac{s_2^2 f(\bar{x}_1)}{s_1^2 f(\bar{x}_2)})$ will always be greater than or equal to one. The minimum value will be one and that value of i is the appropriate weight to use. Figure 6.2 graphs the ratio versus the values of i considered. This ratio will be 1 when they are equal and greater than 1 when they are not equal. The graph clearly shows when the ratio approaches 1.

Weighted linear regression is then performed with weights equal to $\frac{1}{x^i}$, for each development period's respective "best weight". For this triangle the suggested weights correspond to $\frac{1}{x^2}$ in the first development period, 1 in the second development period, and $\frac{1}{x}$ in the third development period. After the third development period, the CRE method is applied, because it was shown to have the most accurate predictions among the three methods considered. Total IBNR losses are then obtained once the full triangle is predicted.

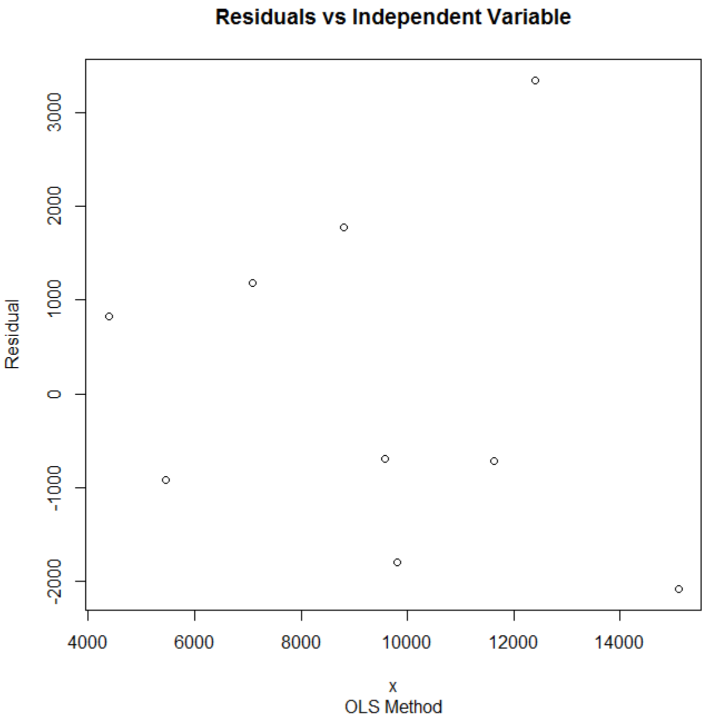

Figure 6.1: Residuals vs Independent Variable; Heteroscedastic

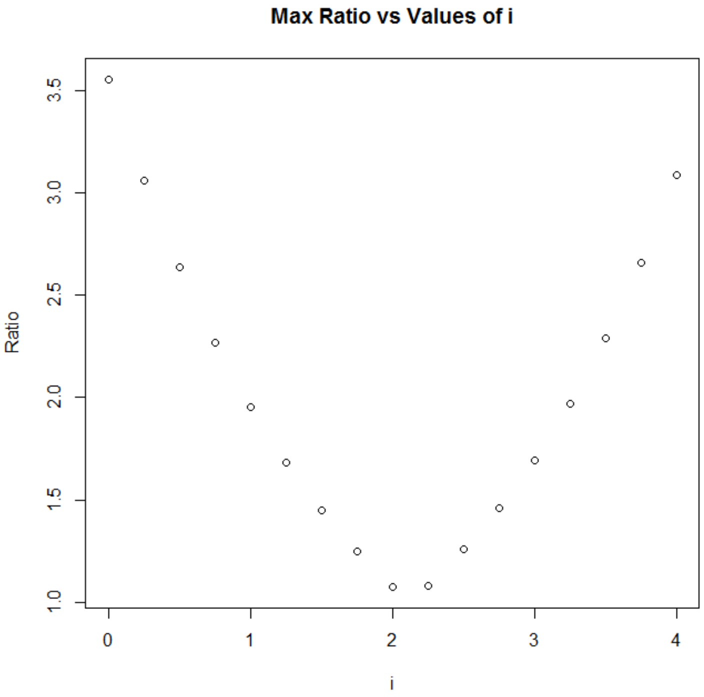

Figure 6.2: $max(\frac{s_1^2 f(\bar{x}_2)}{s_2^2 f(\bar{x}_1)}, \frac{s_1^2 f(\bar{x}_2)}{s_2^2 f(\bar{x}_1)})$ for different values of i

6.2. Mixture of Regressions

The idea behind mixture of linear regressions is that the heterogeneity of data may be associated with different sub-populations rather than a single population. For example, losses from the same development year may be classified by small and large based on their policy structure and size. Thus when analysing a trend in losses for two consecutive development years (by columns) we may observe two distinct groups. In this case a k-component mixture of regression is tested to determine if it provides a better model than a single component weighted-least square regression.

Suppose there are n independent pairs (x_i, y_i) of observations where y_i is the response and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ is the vector of predictors for the ith observation for $i = 1, \ldots, n$. Let $\mathbf{X}$ be the matrix size $n \times p$ with rows corresponding vectors of predictors. Suppose that each observation (x_i, y_i) belongs to one of k groups or components conditional on the membership of jth component where $j = 1, \ldots, k$. Now, the normal regression model can be used as

$$y_i = {\mathbf{x}_i}^{\top} \boldsymbol{\beta}_j + \epsilon_i$$

where ϵ_i is normal random variable with mean zero and variance σ^2 and β_j is the p dimensional vector of regression coefficients. Consider the mixture conditional distribution of $y_i|x_i$ as

$$f_\theta(y_i|x_i) = \sum_{j=1}^{k} \pi_j \phi(y_i|\mathbf{x}_i^{\top} \boldsymbol{\beta}_j, \sigma_j^2)$$

where $\phi(.)$ is Gaussian density with mean $\mathbf{x}_i^{\top} \boldsymbol{\beta}_j$ and variance σ_j^2 and $\boldsymbol{\theta} = (\boldsymbol{\pi}, (\beta_1, \sigma_1^2), \ldots, (\beta_k, \sigma_k^2))$ where $\boldsymbol{\pi}$ represents the vector of mixing probabilities. The goal is to estimate the parameter vector $\hat{\boldsymbol{\theta}}$. This is usually done using the expectation-maximization (EM) algorithm. In this study the estimated parameters of the mixture model are obtained using R package ”mixtool” developed by Benaglia et al. [13, R Core Team 2013] [2, Benaglia, Chauveau, Hunter, and Young 2009].

Once the mixture of regressions model is fit to the data, the BIC of the mixtures model is then compared to the BIC from a single regression model. If the BIC is lower with the single regression model, mixtures are disregarded as this is an indication that the single regression model is a better fit for the data.

It is important to note that due to limited number of observations in a loss triangle, the mixture of regressions is used only for the early development periods. In this case a 2-component model is tested. If a loss triangle contains 10 accident and development years, this model is applicable only for the first 3 development years. Therefore, this methods may impact the first two development factors. For larger size triangles, the mixture model can be applied up to a development year that has minimum 8 accident years (observations).

When these techniques, the nearest neighbor method and mixtures of regression, were performed on the triangles where the CRE method performed the worst as mentioned above, there was no clear improvement of total IBNR estimation with either method. Table 6.2 includes predictions for the worst ten triangles from the CRE predictions (the five highest and five lowest) as well as the five median triangles, triangles with quite accurate predictions with the CRE method. Predictions of total IBNR using all three weighted least squares methods, the nearest neighbor method, as well as mixtures, are included for comparison. There was no distinction between the different predictions on the extreme triangles, and since the considered triangles were already the worst of the CRE predictions, this grouping technique offers no clear improvement on the considered extreme triangles. However, it does show that the nearest neighbor method can improve the predictions on triangles that the CRE method is already providing quite accurate predictions for. Of the five median triangles where the CRE method was providing good predictions, the nearest neighbor method improved four of the five predictions. The improvements were by $3\%, 5\%, 9\%$, and 10%, all significant improvements since losses from a single accident year of a single line of insurance can range well into the hundreds of thousands of dollars.

Inspection of the extreme triangles and the median triangles showed no clear patterns in the data. Observing the full triangle of losses, it is quite apparent where the prediction and the observed values differ. In most cases, one single cell within the lower three lines of the triangle threw off the predictions for that entire accident year, resulting in one or two years of grossly under- or over-predicted losses. Because these are future losses and are not known at the time of modeling, they cannot be accounted for unless an actuary identifies conditions or circumstances that lead to these unfavorable entries in the tables of losses. More detailed knowledge of where or how the losses originate and thus develop would lead an actuary to possibly account for some of these shock losses or unexpectedly low losses.

Table 6.2: IBNR totals using five methods

IBNR Total						
Triangle ID	Actual	OLS	CRE	Method 3	Nearest Neighbor	Mixtures
72	291	**960**	1513	3105	1302	
73	24	103	122	137	**51**	122
8	5269	14206	14676	15175	14143	**11091**
36	572	**1232**	1472	1935	1599	1412
49	309	636	676	710	**619**	940
51	1168	498	603	**727**	588	563
54	1946	918	1001	1109	946	**1228**
47	525	282	261	236	171	**284**
46	868	243	359	504	220	**611**
78	156	0	27	**80**	15	
68	508	**547**	566	587	574	
14	130681	143285	145287	147452	**131129**	
3	89855	96194	99779	104225	**94317**	
28	75433	83231	83577	83900	**75111**	84316
26	17432	19149	19304	19396	**16348**	18591

Most accurate prediction in each row bold.

7. ESTIMATING AUTO LIABILITY IBNR LOSSES FOR A SMALL INSURANCE COMPANY

The ultimate goal of this thesis is to provide predictions and recommendations for a line of commercial auto liability insurance from a smaller Midwestern insurance company. Table 7.1 provides the triangle of cumulative losses obtained from this company. Of the five methods applied to the 78 NAIC triangles, the CRE method was the most precise when predicting the unknown values of losses as well as the total IBNR claims from the lower right triangle. The CRE is thus applied to the real data, resulting in predictions of cumulative losses from each development year and calendar year. Because the first year, 2003, has no losses, the ninth development factor cannot be estimated, and thus the tenth column of losses cannot be projected. From Table 7.2, the seventh and eighth development factors are 1, so the losses are assumed to be fully developed within this triangle.

Table 7.1: Full triangle of a Midwestern insurance company's losses

	Development Year								
Accident Year	1	2	3	4	5	6	7	8	9
2003	0	0	0	0	0	0	0	0	0
2004	5	16	16	16	16	16	16	16	16
2005	191	392	836	854	854	854	854	854	*854*
2006	157	192	192	197	202	204	212	*212*	*212*
2007	219	262	267	292	328	328	*330*	*330*	*330*
2008	167	219	723	935	1023	*1024*	*1032*	*1032*	*1032*
2009	209	779	770	770	*813*	*814*	*821*	*821*	*821*
2010	191	954	829	*906*	*957*	*958*	*965*	*965*	*965*
2011	90	94	*121*	*133*	*140*	*140*	*141*	*141*	*141*
2012	231	*547*	*706*	*771*	*814*	*816*	*822*	*822*	*822*

Predictions using the CRE method in italics.

Table 7.2: Development Factors for the Midwestern company's losses

$f_1 = 2.366$	$f_2 = 1.291$	$f_3 = 1.093$	$f_4 = 1.056$	$f_5 = 1.001$	$f_6 = 1.007$	$f_7 = 1.000$	$f_8 = 1.000$

CRE method.

The IBNR estimated total for the CRE method is 836.38. This is equivalent to a full accident year of ultimate losses. The IBNR totals were also calculated with the OLS method and Method 3, for comparison. The IBNR total from the OLS method was 678.39. Method 3 IBNR total was -3,159.62. This prediction from method three is quite different than the other two, and indicates this triangle might need to be investigated by someone with expert knowledge. Method 3 provided the least weight to large values and the most weight to relatively small values in the triangle of the three methods, and with this triangle having quite varied entries, Method 3 might have inappropriate weights. After discussion of these results with experts from the company, Method 3's results would most likely be disregarded, and the CRE and OLS results would be taken for further consideration.

Figure 7.1 displays standardized residual plots for all three methods. While the residuals do not appear to have any extreme outliers, the residuals from Method 3 are skewed, as the positive residuals are much larger than the negative residuals, likely because the larger losses were given less weight in the model. This indicates again that Method 3 is probably not providing predictions as accurate as the other two methods.

Now applied is the nearest neighbor method, which had better predictions than the CRE method on triangles where the CRE method was providing good predictions already. There are no apparent shock losses present in the upper triangle which is an indication that the CRE method possibly would provide inaccurate predictions. This triangle of real data is therefore assumed to have the "normal" structure that would result in reasonable CRE predictions, and these predictions could possibly be improved with the nearest neighbor method.

Applying the nearest neighbor method to the first three development periods, the three best weights for these development periods were $\frac{1}{x^4}$, $\frac{1}{\sqrt{x}}$, and $\frac{1}{x^{3.5}}$. Table 7.3 shows the predictions using the nearest neighbor technique for the first three development periods and the CRE method for the last five. The estimates are quite similar, and the total IBNR is 838.65, essentially the same as with the CRE method.

Under the assumption that the CRE method is providing somewhat accurate predictions, applying the nearest neighbor method resulted in almost the same predicted values. This suggests the triangle seems to behave quite nicely. Inspection by an actuary would follow, as the Method 3 predictions were so different, before any final recommendations are made.

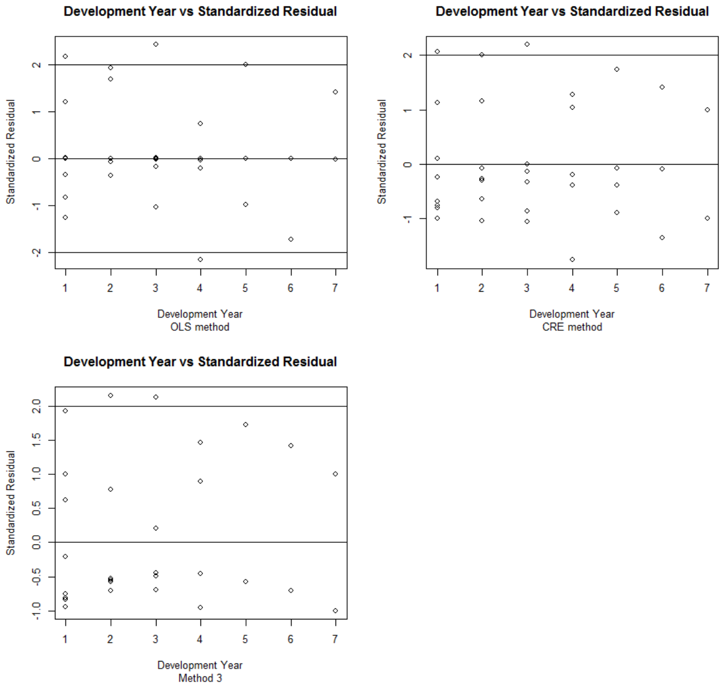

Figure 7.1: Residual plots of the real losses using all three techniques

These results were presented to the company that provided the triangle of losses. In general, this company uses Method 3 to calculate ultimate reserves and uses those ultimate reserves as a starting point for further calculations. Method 3 provided the most conservative projections of the methods used in this thesis. It is up to a company to decide if they want more conservative estimates to prevent over-exposure or more aggressive estimates of reserves, enabling a better annual fiscal report for the company. Regardless of company policy though, the ultimate losses developed in this thesis would directly tie into the calculations for pricing, and are used in conjunction with triangles of the number of claims from a line of insurance to determine the credibility of the triangle of losses. If there are relatively few losses that result in large paid losses, the true ultimate losses could be quite volatile. Again, this is where special handling of these losses and the projections would especially come into play. The company consulted for this project prefers to over-project losses so that losses are paid in full by the end of the year. Their goal is to avoid paying losses from the previous year.

In 2003 the line of business was written through a partner company. Then in 2004 the business was transferred. As the policies were renewed throughout 2004, the claims appeared in this line of losses. The experts from the company would either discount this year or modify their estimates to reflect this transition period of these policies. Modeling this triangle without the year 2003 and 2004 losses would also be taken into account when doing further calculations and making policy recommendations.

Table 7.3: Full triangle of a Midwestern insurance company's losses

	Development Year								
Accident Year	1	2	3	4	5	6	7	8	9
2003	0	0	0	0	0	0	0	0	0
2004	5	16	16	16	16	16	16	16	16
2005	191	392	836	854	854	854	854	854	*854*
2006	157	192	192	197	202	204	212	*212*	*212*
2007	219	262	267	292	328	328	*330*	*330*	*330*
2008	167	219	723	935	1023	*1024*	*1032*	*1032*	*1032*
2009	209	779	770	770	*813*	*814*	*821*	*821*	*821*
2010	191	954	829	*831*	*878*	*879*	*886*	*886*	*886*
2011	90	94	*111*	*111*	*117*	*117*	*118*	*118*	*118*
2012	231	*737*	*867*	*869*	*918*	*920*	*926*	*926*	*926*

Predictions using the nearest neighbor method in italics.

Loss projections such as the ones developed in this thesis are submitted yearly to the governing insurance agencies of every state. A certified actuary must sign off on this company's projections and provide a range of acceptable reserves for the company to hold each year. As this actuary becomes more knowledgeable and comfortable with the company, the range may decrease, but these acceptable reserves still are based on the development of triangles of losses similar to the methods developed in this thesis. Tried and tested methods are preferred from both the company's and the external actuary's standpoints, because the external actuary must sign off on the projections. This results in the continued prevalence of older methods, while some of the more recently developed Bayesian and distribution based methods are not utilized in practice.

8. CONCLUSION

This thesis examined conventional methods and assessed their viability by testing them on 78 NAIC triangles of fully known losses available from the CAS website [3, Casualty Actuarial Society]. Accurate predictions of ultimate losses as well as IBNR losses are important for insurance companies to maintain competitive rates while avoiding over exposure. A deterministic method was introduced as well as the chain ladder method, both of which are quite common in the insurance industry as companies rely on an actuary's expert knowledge to set appropriate reserves and recognize patterns in the data. The company that provided the data for this thesis used ordinary least squares regression to calculate predicted losses. Weighted least squares regression used a more structured theoretical base for the prediction of losses, in an attempt to address concerns regarding the underlying assumptions that are not always satisfied. For example, heteroscedasticity was a common problem and using different weights was introduced as a way to satisfy the assumption of constant variance of the residuals and possibly improve on the predictions when comparing the estimations with the known losses in the NAIC data.

While weighted least squares worked well on many of the triangles, and the CRE method was shown to be the most accurate of the three weighted least squares methods, none of the techniques worked for all 78 triangles analyzed. Consequently, mixtures and the nearest neighbor method were introduced to handle the IBNR predictions with the lowest accuracy. Mixtures were used when comparison of BIC indicated that a mixture model fit the data better than a single regression line, but the results were inconclusive. The nearest neighbor method utilized various weights for each development period, rather than choose one weight for the entire triangle. Again, the estimations for the extreme triangles were imprecise, and no improvement was noted. Of the five median triangles, the most accurate estimates were obtained with the nearest neighbor method in four of the five cases.

The CRE method and the nearest neighbor method were then applied to a triangle of commercial auto liability losses from a smaller Midwestern insurance company. The resulting IBNR predictions from the CRE method, the chosen best of the weighted least squares methods, and the nearest neighbor method, a method that can refine already accurate predictions, were quite close, showing that the predicted future losses will total approximately an entire accident year's worth of losses. The ultimate losses and the subsequent IBNR prediction could provide a starting place for an actuary from the company to set rates for future accident years. Experts from within the company that provided the final triangle analyzed agreed that the predictions from the CRE and nearest neighbor methods seemed reasonable and could be used with other factors when calculating rates or influencing policy decisions.

REFERENCES

[1] G. Barnett and B. Zehnwirth. Best estimates for reserves. *Proceedings of the Casualty Actuarial Society*, LXXXVII:245–321, 2000.

[2] T. Benaglia, D. Chauveau, D. Hunter, and D. Young. mixtools: an r package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6), 2009.

[3] Casualty Actuarial Society. *Loss reserving data pulled from NAIC schedule P*, 2014. Retrieved from http://www.casact.org/research/index.cfm?fa=loss_reserves_data.

[4] P. D. England and R. D. Verrall. Stochastic claims reserving in general insurance. *Institute of Actuaries and Faculty of Actuaries*, 2002.

[5] S. Haberman and A. E. Renshaw. Generalized linear models and actuarial science. *Journal of the Royal Statistical Society, Series D (The Statistician)*, 45(4):407–436, 1996.

[6] V. G. Jemilohun, Y. O. Lawal, and L. Adebara. Statistical analysis of insurance claims reserves in nigeria. *International Journal of Pure and Applied Sciences and Technology*, 18(2):9–22, 1996.

[7] T. Mack. Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, 23, 1993.

[8] G. Meyers. The leveled chain ladder model for stochastic loss reserving. *Casualty Actuarial Society E-Forum*, 2012. Retrieved from http://www.casact.org/pubs/forum/12sumforum/meyers.pdf.

[9] G. Quarg and T. Mack. Munich chain ladder: a reserving method that reduces the gap between ibnr projections based on paid losses and ibnr projections based on incurred losses. *Variance*, 2(2), 2008.

[10] R. Schnieper. Separating true ibnr and ibner claims. *ASTIN Bulletin*, 21, 1991.

[11] D. Scollnik. Actuarial modeling with mcmc and bugs. *North American Actuarial Journal*, 5(2), 2001.

[12] P. Shi, S. Basu, and G. Meyers. A bayesian log-normal model for multivariate loss reserving. *North American Actuarial Journal*, 16(1), 2012.

[13] R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013. Retrieved from http://www.R-project.org/.

[14] R. J. Verrall. Obtaining predictive distributions for reserves which incorporate expert opinion. *Variance*, 1, 2004.

[15] M. Wuthrich and M. Merz. *Stochastic Claims Reserving Methods in Insurance.* John Wiley & Sons Ltd, New York, 2008.