ANALYSIS OF BOOTSTRAP TECHNIQUES FOR LOSS RESERVING

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

By

Taryn Ruth Chase

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

Major Department:
Statistics

May 2015

Fargo, North Dakota

NORTH DAKOTA STATE UNIVERSITY

Graduate School

Title

ANALYSIS OF BOOTSTRAP TECHNIQUES FOR LOSS RESERVING

By

Taryn Ruth Chase

The Supervisory Committee certifies that this disquisition complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Tatjana Miljkovic, Department of Statistics (advisor)

Rhonda Magel, Department of Statistics (co-advisor)

Seung Won Hyun, Department of Statistics

Verena Theile, Department of English

APPROVED:

May 7, 2015 — Dr. Rhonda Magel

Date — Department Chair

ABSTRACT

Insurance companies must have an appropriate method of estimating future reserve amounts. These values will directly influence the rates that are charged to the customer. This thesis analyzes stochastic reserving techniques that use bootstrap methods in order to obtain variability estimates of predicted reserves. Bootstrapping techniques are of interest because they usually do not require advanced statistical software to implement. Some bootstrap techniques have incorporated generalized linear models in order to produce results. To analyze how well these methods are performing, data with known future losses was obtained from the National Association of Insurance Commissioners. Analysis of this data shows that most bootstrapping methods produce results that are comparable to one another and to the trusted Chain Ladder method. The methods are then applied to loss data from a small Midwestern insurance company to predict variation of their future reserve amounts.

ACKNOWLEDGMENTS

Thank you to my advisor, Dr. Tatjana Miljkovic, for her continuous help and support. I would also like to thank the Midwestern insurance company for providing original data to be analyzed.

TABLE OF CONTENTS

ABSTRACT iii

ACKNOWLEDGMENTS iv

LIST OF TABLES vii

LIST OF FIGURES viii

LIST OF APPENDIX FIGURES ix

1. INTRODUCTION 1
   - 1.1. Insurance Reserves 1
   - 1.2. Chain Ladder Method 2

2. LITERATURE REVIEW 7

3. METHODOLOGY 10
   - 3.1. Generalized Linear Model Approach 10
   - 3.2. General Example Solving GLM Model 12
   - 3.3. Bootstrap Technique 14
   - 3.4. Bootstrap of Predictions 19
   - 3.5. Overview of Methods to be Performed 20

4. NAIC ANALYSIS 21
   - 4.1. Data Description 21
   - 4.2. Results 22

5. ESTIMATING LOSSES FOR A SMALL INSURANCE COMPANY 28

6. CONCLUSION 34

REFERENCES 36

APPENDIX 38
A.1. Boxplots of Residuals from NAIC Data 38
A.2. Histograms of Total IBNR values from Midwestern company data 41

LIST OF TABLES

Table		Page
1:	Incremental Claim Amounts for Auto Liability Insurance Line	2
2:	Cumulative Claim Amounts for Auto Liability Insurance	3
3:	Predicted Cumulative Claim Amounts	5
4:	Predicted Incremental Claim Amounts	5
5:	General Incremental Data	13
6:	GLM Fitted Incremental Data	14
7:	Development Factors of Table 2	15
8:	Fitted Cumulative Claims Associated with Table 2	15
9:	Example NAIC Table of Known Losses	22
10:	Example of Known Reserve Values	23
11:	Comparing Average Residuals	26
12:	Predicted IBNR Values	29
13:	Total IBNR Values from Bootstrap Methods	30
14:	Prediction Errors (%)	33

LIST OF FIGURES

Figure Page

1: Residual plots of NAIC triangles using Chain Ladder Bootstrap Method 24

2: Residual plots of NAIC triangles using Gamma GLM Method 25

3: Residual plots of NAIC triangles using Bootstrap on Prediction with Chain Ladder 25

4: Predictive Distribution of Total Reserves from Chain Ladder Bootstrap 31

5: Predictive Distribution of Total Reserves from Pred Bootstrap with Inv Gauss 32

LIST OF APPENDIX FIGURES

Figure Page

A1: Residual plots of NAIC data using Gamma Bootstrap Method 38

A2: Residual plots of NAIC data using Bootstrap on Prediction with Gamma GLM 39

A3: Residual plots of NAIC data using Bootstrap on Prediction with Inv Gauss GLM 39

A4: Residual plots of NAIC data using Inverse Gaussian GLM Method 40

A5: Predictive Distribution of Total Reserves from Gamma Bootstrap 41

A6: Predictive Distribution of Total Reserves from Pred Bootstrap with Chain Ladder 42

A7: Predictive Distribution of Total Reserves from Pred Bootstrap with Gamma GLM 42

1. INTRODUCTION

1.1. Insurance Reserves

Loss reserving is essential for insurance companies to meet future liabilities. Insurance losses are not fully developed at the moment of a claim. Claims mature as payments are made and more information is obtained about the true value of the loss. Premiums collected from a calendar year may not be enough to cover that year's claim amounts for various reasons: a claim could be filed years after the accident/incident occurred, certain claims require multiple payments throughout the years (such as disability), or it could take years for a claim to be settled due to liability issues. For these reasons, along with many others, insurance companies must set aside reserves to cover each year's liability. An appropriate method of estimating what the reserve amount should be is necessary. Many methods have been developed and implemented, but this paper will focus on analyzing various Bootstrap and General Linear Model techniques. These will be described fully in later sections.

To demonstrate how to employ these methods claim amount data will be used, but the processes could be extended to claim count data. Claim amount data is usually presented in a triangular form as shown in Table 1, which displays incremental claims made, in thousands, for an Auto Liability line of insurance for the years 2002-2011 . This data was obtained from a small Midwestern insurance company. This form of data can be referred to as a 'loss triangle' or a 'run-off triangle'. The rows represent the year of origin, or accident year, that a claim occurred in. Claim amounts are shown for each development year that occurs after the accident year. Table 1 is considered fully developed after ten years. Different lines of insurance may develop faster or slower depending on the nature of the insurance. Also, development periods may be represented in other quantities, such as quarters or months, as needed.

The elements in Table 1 will be referenced as $X_{i,j}$, the incremental claim amount from accident year i and development year j. Thus $X_{2,3} = 1219$ corresponds to a claim made in 2005 for an accident that occurred in 2003 . It should be noted that it is possible to have negative values in an incremental loss triangle. Negative values occur from salvages or subrogations. When a company replaces customer property they assume ownership of the damaged property. If the

Table 1: Incremental Claim Amounts for Auto Liability Insurance Line

	Development Year									
Accident Year	1	2	3	4	5	6	7	8	9	10
2002	3556	2389	1107	732	272	160	-14	30	86	1
2003	3495	2267	1219	562	53	130	42	10	54	
2004	4556	2068	1463	148	295	242	107	43		
2005	4139	2042	467	441	831	-252	89			
2006	3771	1496	1545	409	457	-203				
2007	3624	1841	277	450	121					
2008	3817	1727	1232	-42						
2009	4524	1888	1960							
2010	3816	604								
2011	4299									

damaged property can be reconditioned and sold, part of the payments that have been made are offset. This type of recovery is called salvage. Subrogation occurs when the insurance company pays for the customer's loss and then recovers payments from a third party who is determined at fault for the loss event. These recuperations are recorded as negative payments.

Loss triangles also display the total amount paid during a calendar year. Each calendar year is shown as a diagonal going from lower-left to upper-right. For example, the total claims paid in 2011 would be the sum of the diagonal beginning at $X_{10,1}$:

$$4299 + 604 + 1960 + (-42) + 121 + (-203) + 89 + 43 + 54 + 1 = 6926$$

Loss reserving techniques try to predict the lower triangle of loss data. These claims are often Incurred But Not Reported (IBNR) where a claim has occurred but has not yet been filed, or Reported But Not Settled (RBNS) where a claim is known but not completely paid. (Kaas, Goovaerts, Dhaene, and Denuit 2009). Loss reserving is trying to predict values of $X_{i,j}$ where calendar year, k, is greater than the last accident year given in the data. The calendar year can be determined by $k = i + j - 1$. This thesis will focus not only on predicting the future values of Table 1 but also determining variability measures for the loss predictions.

1.2. Chain Ladder Method

One of the most commonly used methods of estimating loss reserves is the Chain Ladder Method. It is a very well established actuarial practice that does not require any statistical modeling. This method uses historical loss data patterns to predict future development patterns. It

assumes that the columns of a loss triangle are proportional and approximately the same percentage of claims will be made in similar development years. A full description of the methods and assumptions can be found in IBNR Techniques (Kaas et al. 2009) but a brief description and example will be given next.

The chain ladder method can be performed using incremental data as shown in Table 1, but it is common for loss data to be shown in a cumulative form. The cumulative data takes the same triangular shape but the elements represent the total amount of claims paid for an accident year up to a particular development year. Table 2 displays the cumulative loss data for the Auto Liability Insurance given in Table 1. Each element of the table is calculated using Equation (1) with $Y_{i,j}$ representing the cumulative claim value of accident year i and development year j. i and j are from 1 to m, where m is the number of development years.

$$Y_{i,j} = \sum_{s=1}^{j} X_{i,s} \tag{1}$$

Table 2: Cumulative Claim Amounts for Auto Liability Insurance

	Development Year									
Accident Year	1	2	3	4	5	6	7	8	9	10
2002	3556	5945	7052	7784	8056	8216	8202	8232	8318	8319
2003	3495	5762	6981	7543	7596	7726	7768	7778	7832	
2004	4556	6624	8087	8235	8530	8772	8879	8922		
2005	4139	6181	6648	7089	7920	7668	7757			
2006	3771	5267	6812	7221	7678	7475				
2007	3624	5465	5742	6192	6313					
2008	3817	5544	6776	6734						
2009	4524	6412	8372							
2010	3816	4420								
2011	4299									

Each empty element of Table 2 can be predicted by multiplying the latest claim amount by a development factor. These development factors are calculated as proportions of successive development years.The process for creating the development factor for the jth development year is shown in Equation (2).

$$f_j = \frac{\sum_{t=1}^{i-1} Y_{t,j+1}}{\sum_{t=1}^{i-1} Y_{t,j}} \tag{2}$$

The future claim amounts can then be calculated for $j >= k - i + 1$, as

$$Y_{i,j} = f_{j-1} * Y_{i,j-1} \tag{3}$$

Using the above formula, each calendar year must be predicted in order. It is not possible to find elements for a calendar year if the previous diagonal has not yet been estimated because those numbers will be needed in the calculations. It should be noted that by using Equations (2) and (3) the Chain Ladder method is equivalent to performing a weighted least squares regression with weights of 1/x.

To demonstrate the use of Equations (2) and (3), calculations are shown for finding the first five development factors and predicted cumulative amounts associated with completing the 2012 calendar year diagonal.

Development Factors:

$f_1 = \frac{5945+5762+6624+6181+5267+5465+5544+6412+4420}{3556+3495+4556+4139+3771+3624+3817+4524+3816} = 1.4624$

$f_2 = \frac{7052+6981+8087+6648+6812+5742+6776+8372}{5945+5762+6624+6181+5267+5465+5544+6412} = 1.1964$

$f_3 = \frac{7784+7543+8235+7089+7221+6192+6734}{7052+6981+8087+6648+6812+5742+6776} = 1.0561$

$f_4 = \frac{8056+7596+8530+7920+7678+6313}{7784+7543+8235+7089+7221+6192} = 1.0460$

$f_5 = \frac{8216+7726+8772+7668+7475}{8056+7596+8530+7920+7678} = 1.0019$

Predicted Cumulative Amounts:

$Y_{10,2} = 4299 * 1.4624 = 6287$

$Y_{9,3} = 4420 * 1.1964 = 5288$

$Y_{8,4} = 8372 * 1.0561 = 8842$

$Y_{7,5} = 6734 * 1.0460 = 7044$

$Y_{6,6} = 6313 * 1.0019 = 6325$

In this way the lower triangle of Table 2 can be estimated. For simplification these calculations can be completed using statistical software. Table 3 shows the cumulative claim amounts for the Auto Liability Insurance with the predicted values in bold.

From Table 3 the total claim amounts that will be need to be paid in future development years are estimated. For example, the company can expect to pay an estimated total of $5,966,000 for all claims from 2010. It is also beneficial to calculate the incremental claim amounts because

Table 3: Predicted Cumulative Claim Amounts

	Development Year									
Accident Year	1	2	3	4	5	6	7	8	9	10
2002	3556	5945	7052	7784	8056	8216	8202	8232	8318	8319
2003	3495	5762	6981	7543	7596	7726	7768	7778	7832	**7833**
2004	4556	6624	8087	8235	8530	8772	8879	8922	**9000**	**9001**
2005	4139	6181	6648	7089	7920	7668	7757	**7783**	**7851**	**7852**
2006	3771	5267	6812	7221	7678	7475	**7527**	**7552**	**7618**	**7619**
2007	3624	5465	5742	6192	6313	**6325**	**6369**	**6390**	**6446**	**6447**
2008	3817	5544	6776	6734	**7044**	**7058**	**7106**	**7130**	**7193**	**7193**
2009	4524	6412	8372	**8842**	**9249**	**9267**	**9331**	**9362**	**9444**	**9445**
2010	3816	4420	**5288**	**5585**	**5842**	**5853**	**5894**	**5914**	**5965**	**5966**
2011	4299	**6287**	**7522**	**7944**	**8310**	**8326**	**8383**	**8411**	**8485**	**8486**

this will display how much the company must pay in any development year, not just the total. The incremental values are given in Table 4, with the predicted values in bold. It can be seen that the last development year contains a value of 1 for all accident years. This coincides with the assumption that this particular line of insurance is fully developed after ten years because the amount being settled has significantly decreased.

Table 4: Predicted Incremental Claim Amounts

	Development Year									
Accident Year	1	2	3	4	5	6	7	8	9	10
2002	3556	2389	1107	732	272	160	-14	30	86	1
2003	3495	2267	1219	562	53	130	42	10	54	**1**
2004	4556	2068	1463	148	295	242	107	43	**78**	**1**
2005	4139	2042	467	441	831	-252	89	**26**	**68**	**1**
2006	3771	1496	1545	409	457	-203	**52**	**25**	**66**	**1**
2007	3624	1841	277	450	121	**12**	**44**	**21**	**56**	**1**
2008	3817	1727	1232	-42	**310**	**14**	**49**	**24**	**62**	**1**
2009	4524	1888	1960	**470**	**407**	**18**	**64**	**31**	**82**	**1**
2010	3816	604	**868**	**297**	**257**	**11**	**40**	**20**	**52**	**1**
2011	4299	**1988**	**1235**	**422**	**366**	**16**	**58**	**28**	**74**	**1**

The chain ladder method is simple to implement and produces results that are easy to understand. However, there are certain drawbacks to this method. The reserve predictions can be unstable if the loss data given is volatile. For this reason it is important that the data being analyzed is a good representation of how the data will act in the future, and that the data is not unstable. Another disadvantage to the chain ladder method is that there is no way to estimate the

variability of the predictions; all that is given is the point estimates of the reserves. This thesis will focus on methods that produce variability estimates so confidence limits can be predicted as well.

2. LITERATURE REVIEW

Predicting future claims accurately is vital for all insurance companies. Many methods have been developed and studied with the purpose of obtaining more accurate predictions. Most of these methods are compared to the results produced by the Chain Ladder method. The Chain Ladder method is a deterministic algorithm that is simple to implement but it is not a model based on mathematical statistics where calculations are estimating parameters of a statistical model (Martínez-Miranda, Nielsen, and Verrall 2012). For this reason many stochastic methods have been developed.

A general framework for creating a stochastic method is presented by England and Verrall (2001). The method suggested attempts to smooth the Chain Ladder development factors to get "best estimates" and variability/precision estimates. The smoothing factor is subject to the judgment of the statistician performing the algorithm. This makes the method time consuming and non-consistent. There were also many drawbacks, one of which was that only positive values would be predicted which will not work for every data set. To create a better stochastic method, models using Bayesian statistics were introduced (England and Verrall 2002). These methods also focus on creating a best estimate but they stress that understanding variability is important for accurate model assumptions. Prediction error is a good estimate of variability and should be analyzed.

One method of producing prediction error that has become more popular is to use bootstrap techniques to estimate future claims. Bootstrapping repeatedly resamples the data to create a distribution to estimate bias and variance of a parameter of interest. This method connects to the jackknife technique but is more widely applicable (Efron 1979). Bootstrap techniques were extended to claims reserving because prediction errors could be calculated with a spreadsheet instead of statistical software packages.

To estimate future claims using the bootstrap technique, residuals are repeatedly resampled. Residuals are found using fitted values obtained from the Chain Ladder Method. The Pearson residual was found to be most appropriate for loss data because of its simple form (England and Verrall 1999). After the bootstrap distribution has been created the prediction error can be calculated by including estimation variance and process variance. It has also been suggested that data

be sampled from the process distribution during the bootstrap procedure to provide realizations from the whole predictive distribution (England 2002).

Many variations on the original bootstrap procedure have been introduced to account for different data distributions. Instead of using the Chain Ladder method to calculate fitted and predicted values, these new methods are using results from Generalized Linear Model (GLM) techniques (Wüthrich and Merz 2008). Generalized Linear Model estimates can be calculated for multiple distributions but the Gamma and Poisson distributions are most commonly used for loss reserving. It has been found that the over-dispersed Poisson GLM will produce estimates that are identically to those of the Chain Ladder Method (Wüthrich and Merz 2008). The estimates can be found using matrix algebra if statistical software is not available (Shapland and Leong 2010).

Some methods change the steps of the original bootstrap procedure in an attempt to obtain a more robust prediction error. Pinheiro, Silva, and Centeno suggest a method in which the bootstrap procedure is performed twice in one iteration; once to find the estimated claim values and again to calculate prediction error with new values that are not seen in the original bootstrap (2003). This method was not determined to be more accurate than the original bootstrap and it was also found to be greatly affected by the process distribution chosen.

Recently there has been a focus on the assumptions that should be assessed when using bootstrap techniques. Bootstrapping produces predictive distributions of losses that are of interest to insurers. Some problems that can occur are model inadequacy due to predictions of never-before-observed calendar years, and over-parameterization (Barnett and Zehnwirth 2008). The one-step ahead prediction errors should be analyzed to determine if problems are arising in the bootstrap model. Also the estimates of the last development year should be validated as plausible from the predictive distribution.

Other stochastic methods have found to be comparable to bootstrapping. The EM algorithm can be used as a tool to detect unobserved risks in finite mixture models. This method sees losses as distributions when making predictions. It has been shown that using Monte Carlo based methods via the EM algorithm produces results that are similar to the bootstrap technique (Rempala and Derrig 2005). Another method proposes using Bayesian estimation implemented with Markov Chain Monte Carlo (MCMC) techniques to predict future claims and obtain prediction errors. The Bayesian method produced similar results to the bootstrap technique. However the bootstrap

method is easier to implement and is more amenable to manipulation. Even so, it was found that the Bayesian method was more likely to produce accurate results over time (England and Verrall 2006).

A method that combines Bayesian techniques with bootstrap methods was developed in recent years. The bootstrap produce allows samples to be generated without distribution assumptions so no parametric assumptions are made for the MCMC Bayesian procedures. From this the algorithm will produce not only the point estimates of the future claims but also an accurate empirical approximation of the entire distribution of claims. When compared to the original bootstrap method, this Bayesian technique was found to be comparable and producing accurate results (Peters, Wüthrich, and Shevchenko 2010).

One of the most recent methods of claims reserving that has been developed is much simpler than those with Bayesian or bootstrap techniques. The Double Chain Ladder method uses a simple regression approach to predict future claims. Unlike the original Chain Ladder method, the Double Chain Ladder applies the chain ladder algorithm on both the claim amount data and the claim count data. This method also divides the predicted claims into Reported But Not Settled (RBNS) and Incurred But Not Reported (IBNR). This distinction could be very important for various insurers but is not the focus of many loss reserving techniques. It has been claimed that this method is better than other stochastic methods because it is based on quatities that have a real interpretation in the context of the insurance data (Martínez-Miranda et al. 2012).

This paper will focus on various bootstrap methods. Most of the bootstrap methods that have been analyzed use current insurance data which makes them unable to validate their results with actual observations. Various bootstrap methods will be compared using past data so the accuracy of these methods can be determined.

3. METHODOLOGY

When performing predictions on loss triangles, most often the main goal is to obtain the total amount of claims for each accident year. It is not as important to know the claim amounts for each combination of accident year and development year as it is to know the total for each year. It is also valuable to be able to predict an upper bound on each year's total, so some variability measures must be estimated along with the parameter estimate. Many stochastic reserving techniques have been implemented to create such variability measures. Methods that will be analyzed in later chapters are using generalized linear models and bootstrapping techniques. A description of the methods used for these stochastic reserving techniques will be given in this chapter followed by an overview of the methods that will be implemented in this paper.

3.1. Generalized Linear Model Approach

One stochastic approach to loss reserving is to treat the incremental claim amounts as the response of a Generalized Linear Model (GLM) that uses a logarithmic link function. If the incremental claim amounts for accident year i and development year j are denoted as $X_{i,j}$, then the form of the model proposed would be as follows:

$$E(X_{i,j}) = m_{i,j} \tag{4}$$

$$Var(X_{i,j}) = \phi E(X_{i,j}) = \phi m_{i,j} \tag{5}$$

$$log(m_{i,j}) = \eta_{i,j} \tag{6}$$

$$\eta_{i,j} = c + \alpha_i + \beta_j \qquad \alpha_1 = \beta_1 = 0 \tag{7}$$

From Equations (4) - (7) the generalized linear model is shown to have a variance proportional to its mean, thus an over-dispersed Poisson error distribution is used. It is known that this is an over-dispersed Poisson because of the Poisson distribution's relation to the exponential family. All members of the exponential family will have variances that are proportional to some function of the mean. In this model, accident year and development year are treated as factors with α_i for accident year i and β_j for development year j. The scale parameter ϕ is estimated when the model

is fit. This model is easy to implement with statistical software and produces reserve estimates that are identical to those given by the chain ladder method. This model also works with a small number of negative incremental values, unlike other proposed methods such as those that use the log-normal class.

Another model that is important to understand is one in which the incremental claim amounts are modeled as Gamma response variables. This model also uses a logarithmic link function and a linear predictor. The main difference is that the variance is proportional to the mean squared. Thus, this model can be represented by Equations (4) - (7), if the variance formula in Equation (5) is replaced by Equation (8) below. This model should produce reserves that are similar to the Chain Ladder method, but not identical.

$$Var(X_{i,j}) = \phi E(X_{i,j})^2 = \phi m_{i,j}^2 \tag{8}$$

This paper proposes that another model could be used to form incremental claim amounts. Over-dispersed Poisson and Gamma distributions are commonly used but Inverse Gaussian also has a similar shape. The Inverse Gaussian distribution is also skewed to the right but it has a thicker tail than the Gamma distribution. This could produce similar results and should be explored.

The Inverse Gaussian model commonly uses the link function $1/\mu^2$ instead of the log link function used in the previously described models. Other than this difference, the Inverse Gaussian model is very similar to the Poisson and Gamma models. The Inverse Gaussian model can be described by Equations (9) - (12). The variance is proportional to the mean cubed and a linear predictor is used.

$$E(X_{i,j}) = m_{i,j} \tag{9}$$

$$Var(X_{i,j}) = \phi E(X_{i,j})^3 = \phi m_{i,j}^3 \tag{10}$$

$$\frac{1}{m_{i,j}^2} = \eta_{i,j} \tag{11}$$

$$\eta_{i,j} = c + \alpha_i + \beta_j \qquad \alpha_1 = \beta_1 = 0 \tag{12}$$

The main advantage of using stochastic GLM models, such as the three presented, is that reserve variability estimates can now be calculated. Since loss reserving problems are predicting future claim amounts, the prediction error, or root mean square of prediction, is used as the variability estimate. The calculation of the mean square error of prediction is given by Equation (13).

$$E[(X_{i,j} - \hat{X}_{i,j})^2] \approx Var(X_{i,j}) + Var(\hat{X}_{i,j}) \tag{13}$$

This form of mean square error is valid for models using over-dispersed Poisson, Gamma, or Inverse Gaussian distributions. It can be seen that the mean square error of prediction is the sum of two parts, the variability in the data, or process variance, and the variability due to estimation. The process variance has already been defined in Equations (5), (8), and (10) for over-dispersed Poisson, Gamma, and Inverse Gaussian models respectively but for simplicity a general form of the process variance can be given.

$$Var(X_{i,j}) = \phi m_{i,j}^{\rho} \tag{14}$$

Equation (14) represents the process variance with $\rho = 1$ for the over-dispersed Poisson model, $\rho = 2$ for the Gamma model, and $\rho = 3$ for the Inverse Gaussian model.

Estimate variance is usually found with statistical software but can be calculated using the delta method. Equation (15) shows the estimation variance for Poisson and Gamma models.

$$Var(\hat{X}_{i,j}) \approx \left| \frac{\delta m_{i.j}}{\delta \eta_{i,j}} \right|^2 Var(\eta_{i,j}) = m_{i,j}^2 Var(\eta_{i,j}) \tag{15}$$

Now both parts of the mean square error are defined and the prediction error can be found by taking the square root of the mean square error. An example of how to use the GLM models described to solve loss reserving problems is given next.

3.2. General Example Solving GLM Model

Statistical software packages can be used to solve for the coefficients in the GLM models described in the previous section but for a better understanding of the methods, the coefficients can be solved using matrix algebra. To demonstrate how these calculations can be performed for an over-dispersed Poisson model, a small loss triangle is given in Table 5.

Table 5: General Incremental Data

	1	2	3
1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$
2	$X_{2,1}$	$X_{2,2}$	
3	$X_{3,1}$		

The above losses are then transformed using the log-link function which would give the natural log of each of the values. Then the model is specified using Equation (7) to create a system of equations.

$ln(X_{1,1}) = c$

$ln(X_{2,1}) = c + \alpha_2$

$ln(X_{3,1}) = c + \alpha_3$

$ln(X_{1,2}) = c + \beta_2$

$ln(X_{2,2}) = c + \alpha_2 + \beta_2$

$ln(X_{1,3}) = c + \beta_3$

The coefficients of the model can be solved by writing the system of equations in matrix notation, Y = X*A, and solving using orthogonal decomposition or other matrix methods. The coefficients are solved such that the difference between the actual log incremental values and the fitted log values is minimized.

$$Y = \begin{bmatrix} ln(X_{1,1}) \\ ln(X_{2,1}) \\ ln(X_{3,1}) \\ ln(X_{1,2}) \\ ln(X_{2,2}) \\ ln(X_{1,3}) \end{bmatrix} \quad X = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \quad A = \begin{bmatrix} c \\ \alpha_2 \\ \alpha_3 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

X is called the design matrix and will have dimensions n x p, where n is the number of known elements in the loss triangle and p is the number of coefficients that are being predicted. After solving for the coefficients, the log fitted values can be calculated.

$ln(m_{1,1}) = \hat{c}$

$ln(m_{2,1}) = \hat{c} + \hat{\alpha_2}$

$ln(m_{3,1}) = \hat{c} + \hat{\alpha_3}$

$ln(m_{1,2}) = \hat{c} + \hat{\beta_2}$

$ln(m_{2,2}) = \hat{c} + \hat{\alpha_2} + \hat{\beta_2}$

$ln(m_{1,3}) = \hat{c} + \hat{\beta_3}$

These results can then be exponentiated to solve for the fitted incremental claim values and can be placed in the triangular form.

Table 6: GLM Fitted Incremental Data

	1	2	3
1	$m_{1,1}$	$m_{1,2}$	$m_{1,3}$
2	$m_{2,1}$	$m_{2,2}$	
3	$m_{3,1}$		

The estimated values of the coefficients can also be used to predict the log values for the lower triangle. The predicted incremental values are then found by exponentiating the predicted logs.

$ln(m_{3,2}) = \hat{c} + \hat{\alpha_3} + \hat{\beta_2}$

$ln(m_{2,3}) = \hat{c} + \hat{\alpha_2} + \hat{\beta_3}$

$ln(m_{3,3}) = \hat{c} + \hat{\alpha_3} + \hat{\beta_3}$

This method can be used for Gamma and Inverse Gaussian models as well and, can be used to find fitted values for the upper triangle as well as predicted values for the lower triangle.

3.3. Bootstrap Technique

Bootstrapping is a form of nonparametric Monte Carlo methods that estimate the distribution of a population by resampling from the data (Rizzo 2008). These methods are often done when the sample is the only information available and the distribution is unknown. The objective is to repeatedly resample from the data itself to get a better idea of the distribution or the parameter of interest. When bootstrapping, resampling can be done on the data or on the residuals, depending on the type of problem. If the original data is not a good representation of the actual distribution, the samples obtained by bootstrapping will not become close to the actual values. Therefore it is essential to have a sample that is indicative of the population.

In loss reserving problems the prediction error needs to be estimated and bootstrapping is an appropriate method to do this. In these types of problems it is common to resample, with replacement, the residuals of the upper triangle of claims and create pseudo data sets. In order to find residual values, fitted values must first be calculated. Most bootstrapping techniques use the chain ladder method for fitting and predicting. These fitted values will be different than those found using the GLM approach (except for over-dispersed Poisson model).

Given an upper triangle of observed cumulative claim amounts and the chain ladder development factors, the fitted values can be found. First, the final diagonal of the observed amounts will remain the same for the fitted values. The remaining fitted values are then found backwards by recursively dividing the fitted value at time t by the development factor from time *t-1*. An example of finding fitted values of Table 2 is given next with Tables 7 and 8.

Table 7: Development Factors of Table 2

$f_1 = 1.462$	$f_2 = 1.196$	$f_3 = 1.056$	$f_4 = 1.046$	$f_5 = 1.002$
$f_6 = 1.007$	$f_7 = 1.003$	$f_8 = 1.009$	$f_9 = 1.000$	$f_{10} = 1.000$

Table 8: Fitted Cumulative Claims Associated with Table 2

	Development Year									
Accident Year	1	2	3	4	5	6	7	8	9	10
2002	4214	6163	7374	7788	8146	8162	8218	8246	8318	8319
2003	3968	5803	6943	7333	7670	7685	7738	7764	7832	
2004	4560	6669	7978	8426	8814	8831	8892	8922		
2005	3978	5817	6960	7350	7689	7704	7757			
2006	3860	5644	6753	7132	7461	7475				
2007	3266	4776	5714	6035	6313					
2008	3644	5329	6376	6734						
2009	4785	6998	8372							
2010	3022	4420								
2011	4299									

From the fitted cumulative data, fitted incremental data is easily found by differencing. Once the fitted incremental values have been found, the residuals can be calculated. It is important to use an appropriate residual formula. England and Verrall (1999) propose using either Deviance residuals or Pearson residuals. Pearson residuals are more commonly used because the form is easily inverted to solve for pseudo fitted values.

Pearson residuals:

$$r = \frac{X - m}{\sqrt{m}} \tag{16}$$

where X and m are the observed and fitted incremental claim amounts, respectively. These residuals are considered "unscaled" because they do not contain the scale parameter ϕ. This scale parameter is not needed in the bootstrap calculations but is used when finding the process error so it should be estimated. One estimation of ϕ that is consistent with residuals that are being used is:

$$\phi^* = \frac{\sum r^2}{n - p} \tag{17}$$

with n as the number of data points in the sample, or the number of claims in the upper triangle, and p as the number of parameters to estimate. The sum is then taken over the n residuals that are calculated from the upper triangle.

Once n residuals have been resampled, with replacement, and placed in the upper triangular form a bootstrap data sample is created. The pseudo values for the claim amounts, X^*, are found by solving for X in Equation (16) and using the resampled Pearson residuals, r^*, and the fitted incremental values. In this way bootstrapping gives a new upper triangle of incremental payments.

$$X^* = r^*\sqrt{m} + m \tag{18}$$

Now the prediction error can be calculated in the same way as was described in section 3.1 as the square root of mean square error. It has been shown that mean square error can be broken up into the sum of the process variance and the estimation variance. Using bootstrap methods these variances can be found with simple calculations that do not require advanced statistical software. This is one main advantage of using bootstrapping techniques; instead of trying to find the prediction error with analytic calculations, it can be found with simulation and easily calculated in a spreadsheet.

The bootstrap process variance can be found by multiplying the total reserve amount for each accident year by the estimated scale parameter, ϕ^*. The accident year totals, R, are found by calculating the chain ladder estimates of future reserves on the original observed claim amount data. The estimation variance uses the standard error found for the accident year reserves over

all bootstrap samples. The estimation variance is scaled to take degrees of freedom into account which will make the results comparable to those in section 3.1.

Thus the bootstrap estimates are given by:

$$Var(X_{i,j})_B = \phi^* R \tag{19}$$

$$Var(\hat{X_{i,j}})_B = \frac{n}{n-p}(SE(R))^2 \tag{20}$$

$$PE_B = \sqrt{\phi^* R + \frac{n}{n-p}(SE(R))^2} \tag{21}$$

Now that the basics of the bootstrapping technique are understood the procedure can be defined.

Steps to perform bootstrap procedure:

1. Compute development factors for the cumulative claim amounts
2. Use development factors to obtain fitted cumulative values for the upper triangle
3. Obtain fitted incremental values by differencing
4. Calculate Pearson residuals for each value in the upper triangle by using the observed and fitted incremental values
5. Begin iterative bootstrap loop to be repeated N times
   (a) Resample with replacement n residuals to form a new upper triangle of residual values
   (b) Create pseudo data of incremental claims using the resampled residuals and the fitted incremental values
   (c) Create pseudo data of cumulative claims that is associated with the pseudo incremental claims
   (d) Use the Chain Ladder method on the pseudo cumulative data to estimate future cumulative claims
   (e) Obtain the future incremental claims by differencing
   (f) Sum the predicted incremental claims by accident year to obtain yearly reserve estimates

(g) Store the results and return to the beginning of the bootstrap loop

The stored results will create the predictive distribution. These values can be compared against Chain Ladder results to determine bias and the prediction error can be computed using Equation (21).

England (2002) proposed some changes to the bootstrap procedure. It is suggested that the residuals be scaled to account for degrees of freedom before sampling from them. This would be done by calculating the residuals as

$$r' = \sqrt{\frac{n}{n-p}} x \frac{X-m}{\sqrt{m}} \tag{22}$$

Then step 4 of the bootstrap procedure would be done with Equation (22) instead of Equation (16).

It is also suggested that an additional step be added in the iterative loop to draw a random observation from the underlying process distribution. This would take place right after step (e) of the bootstrap loop. For each cell in the lower triangle, a random observation is drawn from the process distribution with a mean value given as the estimated incremental claim amount and a variance found using Equation (14) along with the estimated scale parameter in Equation (17). Then the accident year reserves in step (f) are found using the values simulated from the process distribution.

The process distributions that have been discussed are over-dispersed Poisson and Gamma. England points out some drawbacks to using over-dispersed Poisson that can be found in Addendum to "Analytic and bootstrap estimates of prediction errors in claims reserving" (2002). One of these drawbacks is that observations will always be multiples of ϕ. For this reason the Gamma distribution is preferable as the process distribution. This allows the advantage of the observations being on a continuous scale and not just being multiples of ϕ. This will change the shape of the predictive distribution but the first two moments will remain the same.

This thesis will explore the differences between using the bootstrap procedure as defined by steps 1 - 5 and the procedure that involves sampling from the process distribution. Both methods will be implemented and compared.

3.4. Bootstrap of Predictions

A further addition to the bootstrap procedure was introduced by Pinheiro, Silva, and Centeno (2003). This method proposes that the fitted and predicted values found using the GLM method be used in the bootstrap procedure instead of those found with the Chain Ladder method. This should not change the reserve estimates by much as these two methods are considered comparable to one another. It was also suggested that an extra step be added to the bootstrap loop to better estimate the prediction error. The new step would take place inside the bootstrap loop and after step (f).

(g) Resample again from the residuals but now take a sample to form a lower triangle

(h) Create pseudo reality of incremental predicted values using the resampled residuals and the predicted lower triangle values

(i) Obtain prediction errors using the new pseudo data of predicted values

(j) Store the results and return to the beginning of the bootstrap loop

As shown in the steps above, this method places residuals calculated from the upper triangle into the lower triangle and uses predicted values to create pseudo data. The residuals are assumed to be independent and identically distributed so it is argued that they could be used to repopulate the lower triangle. This second resample of the residuals is not used to calculate the predicted reserve values, it is only used for prediction error.

This thesis is focused on the reserve estimates produced by different bootstrap procedures so it is of interest to alter the method proposed by Pinheiro. Instead of adding steps (g) through (j) to the existing bootstrap loop, this thesis proposes to replace steps (a) through (f) and use the values found by simulating the lower triangle to estimate the reserve values. This new process can be defined in the steps given below.

Steps to perform Bootstrap of Predictions:

1. Obtain fitted incremental values for the upper triangle
2. Obtain predicted incremental values for the lower triangle
3. Calculate Pearson residuals for each value in the upper triangle by using the observed and fitted incremental values

4. Begin iterative bootstrap loop to be repeated N times

   (a) Resample, with replacement, residuals to form a new **lower** triangle of residual values

   (b) Create pseudo data of incremental claims using the resampled residuals and the **predicted** incremental values

   (c) Sum the pseudo predicted incremental claims by accident year to obtain yearly reserve estimates

   (d) Store the results and return to the beginning of the bootstrap loop

For this new method the fitted and predicted values from steps (1) and (2) can be calculated using either the Chain Ladder method or the GLM approach. It is not known how well this method will perform because the prediction bootstrap has previously only been used for prediction error.

3.5. Overview of Methods to be Performed

In order to compare various reserve estimates found from different methods each process will be implemented using the same data. As a baseline comparison, the Chain Ladder method will also be applied to the data. The methods from this chapter that will be compared are: the GLM approach with Gamma response variables, GLM approach with Inverse Gaussian response variables, Bootstrap technique with Chain Ladder fitted/predicted values, Bootstrap technique with Gamma process distribution, Bootstrap of predictions with Chain Ladder fitted/predicted values, Bootstrap of predictions with Gamma GLM fitted/predicted values, and Bootstrap of predictions with Inverse Gaussian GLM fitted/predicted values. Each bootstrap loop will go through 1,000 iterations. The over-dispersed Poisson GLM method will not be implemented because the reserves that are produced are identical to the Chain Ladder method.

It should be noted that heteroscedasticity is a common problem among all loss reserving methods. The Chain Ladder method described in this paper has been developed to adjust for possible heteroscedasticity but it may not work in all cases. The Generalized Linear Model approaches do not fully address heteroscedasticity so it could be present in the results. This issue should be studied further but for the purpose of this thesis the results from all methods will be compared to one another without considering heteroscedasticity.

4. NAIC ANALYSIS

To determine the effectiveness of each method described in Chapter 3, data with known reserve values will be analyzed. The data was collected by the National Association of Insurance Commissioners (NAIC) and retrieved from the Casualty Actuarial Society (CAS) website (www.casact.org). NAIC is the U.S. standard-setting and regulatory support organization for the insurance market. The members of NAIC are the chief insurance regulators from the 50 states, the District of Columbia, and five U.S. territories. These members are elected or appointed state government officials and their departments. NAIC establishes insurance standards and best practices. The goals of the NAIC are to protect the public interest, promote competitive markets, facilitate fair and equitable treatment of insurance consumers, and promote the reliability, solvency and financial solidity of insurance institutions (www.naic.org).

4.1. Data Description

The National Association of Insurance Commissioners compiles annual statements based on insurance companies' annual financial data. NAIC can only compile data from companies under its regulation. These annual statements contain various schedules and exhibits, such as: income statement, cash flow, underwriting and investment exhibits, number of policies exhibit, property reinsurance (schedule F), 10-year losses by line (schedule P), life reinsurance (schedule S), and premiums written by state (schedule T). Using NAIC data, industry leaders determine market share, conduct market research, and monitor industry trends. Also, NAIC and CAS have made loss data available so that methods of loss reserving can be analyzed.

For the purpose of this paper, data was obtained from Schedule P which gives 10-year loss expenses and loss analysis by line of insurance. Data from 140 companies was available for auto liability insurance. The data for each company includes cumulative paid losses for the accident years 1988-1997. An example of an NAIC data table is shown in Table 9.

The cumulative losses for each accident year are known for ten development years. Thus the data given is of the form of Table 3, a completed table. However, unlike a predicted loss table, the losses in the lower triangle are known instead of estimated. These known future losses can be used to compare with results from various estimation processes.

Table 9: Example NAIC Table of Known Losses

	Development Year									
Accident Year	1	2	3	4	5	6	7	8	9	10
1988	463	903	1659	2190	2301	2331	2347	2349	2367	2366
1989	471	1305	1820	1900	2089	2170	2169	2246	2233	2225
1990	493	1297	2004	2383	2604	2625	2627	2627	2664	2664
1991	469	1323	2284	2816	2954	3083	3174	3177	3177	3177
1992	956	2240	3265	3740	3809	3897	3933	3982	4005	4007
1993	883	2390	4208	4942	5458	5588	6711	6693	6693	6693
1994	1035	2967	5510	7678	7906	7943	8155	8136	8140	8208
1995	900	2984	5481	6055	6389	6404	6409	6441	6455	6474
1996	1412	4449	5908	7591	8063	8358	8464	8467	8468	8470
1997	1782	4819	6429	7273	8410	8545	8547	8530	8577	8600

Most of the methods being used in this paper require incremental losses to be greater than zero. For this reason tables showing negative or zero values have been disregarded. There are 12 tables that have appropriate forms and can be analyzed.

4.2. Results

The accuracy of the each of the methods described in Chapter 3 can be determined by comparing the estimated IBNR values to the true losses given in the NAIC data. As previously mentioned, IBNR values are the total incurred but not reported amounts. These are found by summing all of the predicted incremental future claims for each accident year. Thus the IBNR values represent the total predicted reserve amounts for each year. Since the first accident year is always fully known, the IBNR value will always be zero. For this reason only the IBNR values for accident years two through ten will be discussed (1989-1997).

In order to compare predicted IBNR values with the known losses provided in the NAIC data, the reserve amounts for each accident year must be calculated. The known reserves will consist of all losses that occurred in the lower triangle of each loss table. These are found from the cumulative NAIC loss tables by subtracting the loss of the last diagonal of the upper triangle from the last development year loss. This process can be expressed by Equation (23). Reserve values for Table 9 are given in Table 10 as an example. This calculation is done for each of the 12 NAIC loss tables.

$$R_i = Y_{i,10} - Y_{i,10+1-i} \tag{23}$$

Table 10: Example of Known Reserve Values

Accident Year	Reserves
1989	-8
1990	37
1991	3
1992	110
1993	1235
1994	530
1995	993
1996	4021
1997	6818

For each method of estimation that was performed, predicted IBNR values for each accident year were found for each of the 12 companies. These predictions can then be compared against the actual reserve values to analyze accuracy. Commonly used residual values (observed - expected) will not be a good indicator of how well each technique is performing because the NAIC data was gathered from companies of several sizes, giving different orders of magnitude in the triangles being analyzed. To account for the different loss magnitudes the residuals are found by dividing the difference between the observed and expected IBNR values by the observed value. This adjustment will allow comparison of the residuals over all NAIC triangles.

$$AdjRes = \frac{ObsIBNR - PredIBNR}{ObsIBNR} \tag{24}$$

Once the adjusted residuals have be calculated the distribution of residuals for each accident year can be seen with boxplots like those in Figure 1. Figure 1 shows that the residuals seem to have less spread as accident year increases. This is likely caused by the number of loss values that are required to be estimated for each IBNR value. For the 1989 accident year, one loss value is estimated, cell $X_{2,10}$, and is used in calculating the predicted IBNR value. Whereas for the 1993 accident year there are five estimated values, cells $X_{6,6}$, $X_{6,7}$, $X_{6,8}$, $X_{6,9}$, and $X_{6,10}$. These values are then summed to find the predicted IBNR value for that accident year. The increasing number of cells used to find each year's IBNR value leads to smaller variation around the true value.

In Figure 1 each boxplot has a mean value that is close to zero. This suggests that the Bootstrap technique with Chain Ladder estimates produces results that are similar to the observed IBNR values for NAIC data. In fact most of the methods performed on NAIC data produced box-

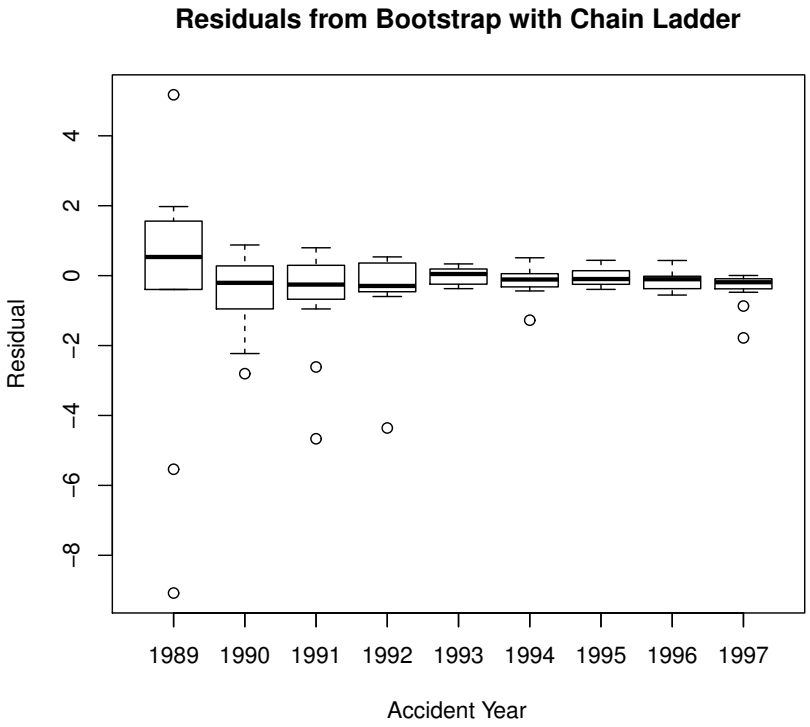

Figure 1: Residual plots of NAIC triangles using Chain Ladder Bootstrap Method

plots that resemble those in Figure 1. This similarity shows that all of the methods are comparable to one another, at least when it comes to residual values.

For comparison Figure 2 shows the boxplots for the Gamma GLM method and Figure 3 shows the boxplots for Prediction Bootstrap with Chain Ladder estimates. Since most of the boxplots look very similar, not all of the graphs are necessary to include. The boxplots from the other methods can be found in the Appendix.

Comparing Figures 1 - 3 it can be seen that the three methods given produce similar residual values. The Gamma GLM method has a larger spread as the residuals are from -10 to 4, and the other methods are between -8 and 4. This is not a large difference but should be noted. Also the Gamma GLM method is not using bootstrapping techniques so only one set of estimated IBNR values is calculated. The bootstrapping techniques find 1000 simulated IBNR estimates and then takes the average value to obtain the predicted IBNR for each year. This difference could account for the slight difference in spread of the residuals.

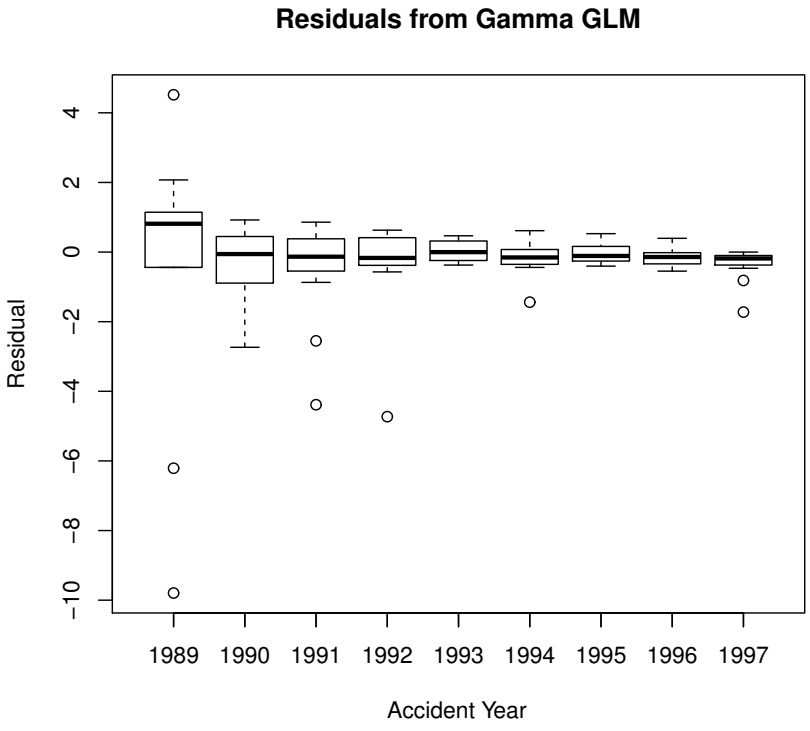

Figure 2: Residual plots of NAIC triangles using Gamma GLM Method

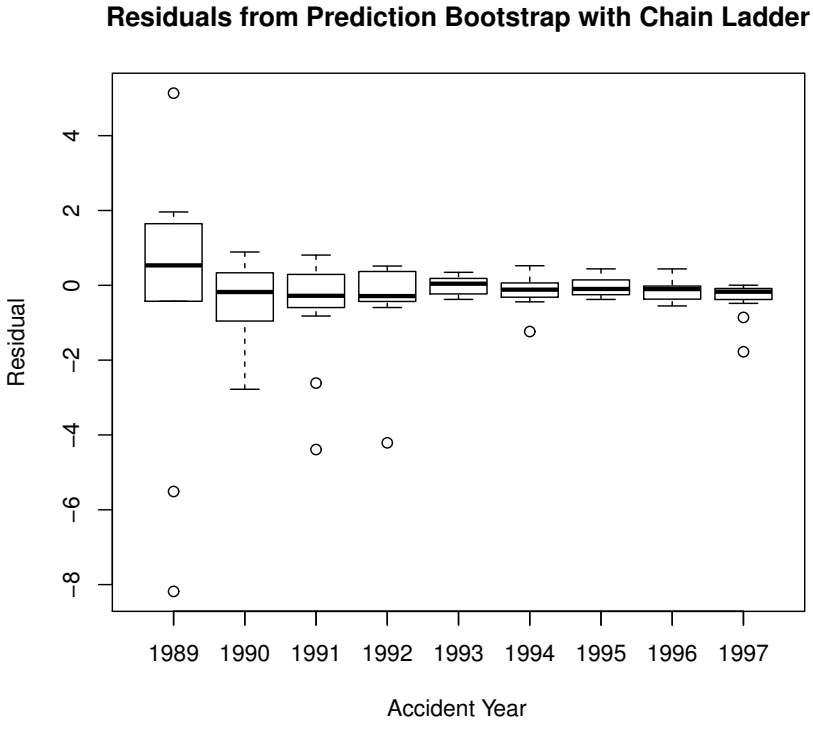

Figure 3: Residual plots of NAIC triangles using Bootstrap on Prediction with Chain Ladder

In order to compare how all of the methods are performing, the average adjusted residual was found for each accident year. Table 11 shows the average residual values for each of the methods described in Chapter 3. The average residual values were also found for the basic Chain Ladder method because the values found using this technique are widely trusted. In Table 11 "CL Bootstrap" is the Bootstrap method that uses the Chain Ladder method for fitted and predicted values, and "Pred Bootstrap" is the Bootstrap of predictions using Chain Ladder predicted values.

Table 11: Comparing Average Residuals

Accident Year	Chain Ladder	Gamma GLM	Inv Gaus GLM	CL Bootstrap
1989	-0.35290	-0.55669	-0.10387	-0.41968
1990	-0.34669	-0.30328	-0.25494	-0.45987
1991	-2.42109	-2.89670	-2.360672	-2.73079
1992	-0.35161	-0.36259	-0.29968	-0.40853
1993	0.00860	0.03060	0.00606	-0.01388
1994	-0.15003	-0.18022	-0.00904	-0.16689
1995	-0.04108	-0.03476	0.02267	-0.04959
1996	-0.15384	-0.15879	-0.08413	-0.16054
1997	-0.36078	-0.35558	-0.07236	-0.36743
Accident Year	Gamma Bootstrap	Pred Bootstrap	Pred w/ Gamma	Pred w/ Inv Gaus
1989	-0.44192	-0.32388	-0.55313	-0.10452
1990	-0.45363	-0.43376	-0.42241	-0.37743
1991	-2.77969	-2.67634	-3.11963	-2.77993
1992	-0.40076	-0.38971	-0.40021	-0.37784
1993	-0.01785	-0.01420	0.00917	-0.02839
1994	-0.16610	-0.15987	-0.19019	-0.03346
1995	-0.04956	-0.04516	-0.03733	0.00837
1996	-0.16283	-0.15779	-0.16223	-0.09513
1997	-0.36589	-0.36373	-0.35287	-0.08051

Analysis of Table 11 shows that all of the methods produce similar average residual values for each accident year. All of the methods produce residuals that are close to zero, meaning the methods are predicting IBNR values close to the observed values. The largest absolute residual appears in the 1991 accident year for all methods. The smallest residual appears in the 1993 accident year for all methods except the last, Prediction Bootstrap with Inverse Gaussian GLM fitted values. The bootstrap methods over estimate the IBNR values for all accident years since all of the residual values are negative, whereas the methods without bootstrapping have at least one positive residual value.

With this comparison table it is reasonable to claim that all methods analyzed will be able to adequately predict future IBNR values. The Gamma GLM, with no bootstrapping, has some of the highest residual values followed by the Prediction Bootstrap that uses Gamma GLM fitted values. These large residuals suggest that a Gamma GLM model is least appropriate for modeling NAIC data. The Inverse Gaussian GLM model and the Prediction Bootstrap using Inverse Gaussian have the smallest residual values. Inverse Gaussian is not a common distribution used in loss reserving but it appears to be the most appropriate for predicting NAIC data.

5. ESTIMATING LOSSES FOR A SMALL INSURANCE COMPANY

A small insurance company in the Midwest provided loss data for an autoliability line of insurance that was presented in Tables 1 and 2 in incremental and cumulative form, respectively. The goal of this paper is to predict reserve amounts based on the provided data that will adequately cover the future losses. It has been shown that each method introduced and analyzed with NAIC data produces estimates that are comparable to actual observed losses. The predicted reserves for the auto liability data from the small company will be found using each method.

In order to perform the methods that use a GLM approach, with or without bootstrapping, all of the incremental values must be positive. As seen in Table 1 the small insurance company data has negative incremental values. In *Bootstrap Modeling: Beyond the Basics*, Shapland and Leong propose several methods for dealing with negative values in a GLM model (2010). The method that will be implemented for the Midwest company data is one in which the absolute value of the largest negative incremental claim is added to all incremental claim amounts. Once the predicted values are found the value that was added is subtracted to return the losses to their original form. This approach was used for the data in Table 1 by adding 253 to each value. The absolute largest negative is 252 but in order to have all of the values be non-negative, 253 must be added. This was done for each method that uses GLM fitted or predicted values.

Once all of the methods have been performed on the small company data, the predicted IBNR values are found and compared. Table 12 shows the predicted IBNR values for each method along with the predicted total reserve amount. For the methods that use bootstrapping, the predicted IBNR values are the average amount over the 1000 iterations.

Analyzing Table 12, the various methods can be compared. The predictions for the first few accident years seem to vary widely depending on method, but the later accident years are more consistent. This agrees with what was found when analyzing the boxplots in the previous chapter. Also the methods that involve using Gamma GLM produce estimates that are the least similar to the other methods. The Gamma GLM models produce the lowest estimates. Since these models were found to have some of the largest residual values in the previous chapter, these estimates are probably not producing the most accurate predictions. It appears that using these predicted values would lead to under-estimation, compared to the other models, and the company would not have

Table 12: Predicted IBNR Values

Accident Year	Chain Ladder	Gamma GLM	Inv Gaus GLM	CL Bootstrap
2003	0.94	-17.09	0.92	6.01
2004	79.10	100.64	71.34	110.44
2005	94.91	-36.36	98.41	143.81
2006	143.80	-63.01	154.02	212.96
2007	133.90	-251.01	166.79	207.15
2008	459.49	13.90	505.49	567.19
2009	1073.29	1350.23	901.52	1260.96
2010	1546.01	566.33	1972.33	1695.36
2011	4186.88	3924.46	3947.89	4413.05
Total	7718.32	5588.10	7818.72	8616.93
Accident Year	Gamma Bootstrap	Pred Bootstrap	Pred w/ Gamma	Pred w/ Inv Gaus
2003	4.86	5.76	-15.92	12.77
2004	110.08	109.92	98.01	91.20
2005	149.70	144.91	-45.43	122.30
2006	203.03	220.41	-64.59	180.33
2007	206.71	217.52	-241.75	195.28
2008	546.10	566.31	20.50	549.84
2009	1285.70	1234.89	1343.40	939.20
2010	1663.76	1756.12	582.26	2011.97
2011	4374.85	4355.53	3929.01	3970.84
Total	8544.78	8611.38	5605.49	8073.73

enough reserves set aside for future loss. It is not recommended that the company use these values for reserve estimation.

The method that produces the largest reserve values is the Bootstrap with Chain Ladder fitted values. The Prediction Bootstrap with Chain Ladder fitted values also produces large results that are similar to the original Bootstrap approach. This suggests that using the Chain Ladder method combined with some form of bootstrapping will produce high reserve values. The total value is almost $1,000,000 higher than that found using only the Chain Ladder method. Even though the values are high, they could still be accurate. As seen in Table 11 the residual values for these methods were not much larger than the other methods so there is no reason to believe that the estimates produced will be inadequate.

The Inverse Gaussian GLM method produces IBNR predictions that are closest to those produced by the Chain Ladder method. This method was very accurate when predicting the NAIC data. For these reasons the predictions found using this method would be appropriate for the company to use when planning for future reserves. The estimates found using Prediction Bootstrap

with Inverse Gaussian GLM fitted values are also similar and were also found to be accurate in the previous chapter. These values could also be used as predictions of future reserves and the variance estimates can then be used to find confidence limits for the reserve values.

The bootstrap techniques perform 1000 iterations to simulate a distribution of the data. Instead of looking at the distribution for each accident year it is common to focus on the distribution of the total reserve value. This value should cover all future losses for each accident year through the final development year. A summary of the distributions from the bootstrap methods performed is given in Table 13. From this, predictions of the future total can be made.

Table 13: Total IBNR Values from Bootstrap Methods

Method	CL Bootstrap	Gamma Bootstrap	Pred Bootstrap
Mean	8616.93	8544.78	8611.38
Standard Dev	1712.90	2788.25	1267.47
Coeff of Var	0.199	0.326	0.147
Skewness	0.163	0.334	-0.047
50th Percentile	8524.06	8233.85	8631.09
75th Percentile	9534.70	10082.99	9464.86
95th Percentile	11351.12	13590.91	10653.78
Method	Pred w/ Gamma	Pred w/ Inv Gaus	
Mean	5605.49	8073.73	
Standard Dev	1229.80	1382.71	
Coeff of Var	0.219	0.171	
Skewness	0.032	0.065	
50th Percentile	5592.26	8043.88	
75th Percentile	6454.32	9057.25	
95th Percentile	7611.99	10370.22	

The mean total value is the predicted total given in Table 12 and the standard deviation was found over the 1000 estimated totals produced by bootstrapping. The predictions in Table 13 seem similar for each bootstrap method besides the Prediction bootstrap using Gamma GLM fits. It has already been seen that methods using the Gamma GLM approach do not seem to be appropriate for this data.

The Gamma Bootstrap represents the original bootstrap approach using a Gamma process distribution. This method has the largest standard deviation and is the most skewed. This method therefore produces the largest estimates for the percentiles. All of the other methods have skewness values that are very small and would suggest that the distribution could be approximately symmetric.

The percentiles are important when considering predictions because companies want to know with a high level of certainty that they will be able to cover all future losses. They can make decisions about how much to set aside based on the average value and the 75th and 95th percentiles. This company should have at least $8,500,000 as total reserves for this insurance line, based on all of the averages but could have around $11,000,000 to make sure they cover the 95th percentile of most of the bootstrap methods.

A histogram provides a visualization of the distributions created for the total reserves. Figures 4 and 5 show the distributions for the Bootstrap method using Chain Ladder fitted values and the Prediction Bootstrap using Inverse Gaussian GLM fitted values. The Chain Ladder Bootstrap had the largest average predicted total and one of the largest standard deviations. The Inverse Gaussian Prediction Bootstrap produced results similar to the basic Chain Ladder method and was found to be one of the most accurate with NAIC data. Histograms from the other three bootstrap methods are available in the Appendix.

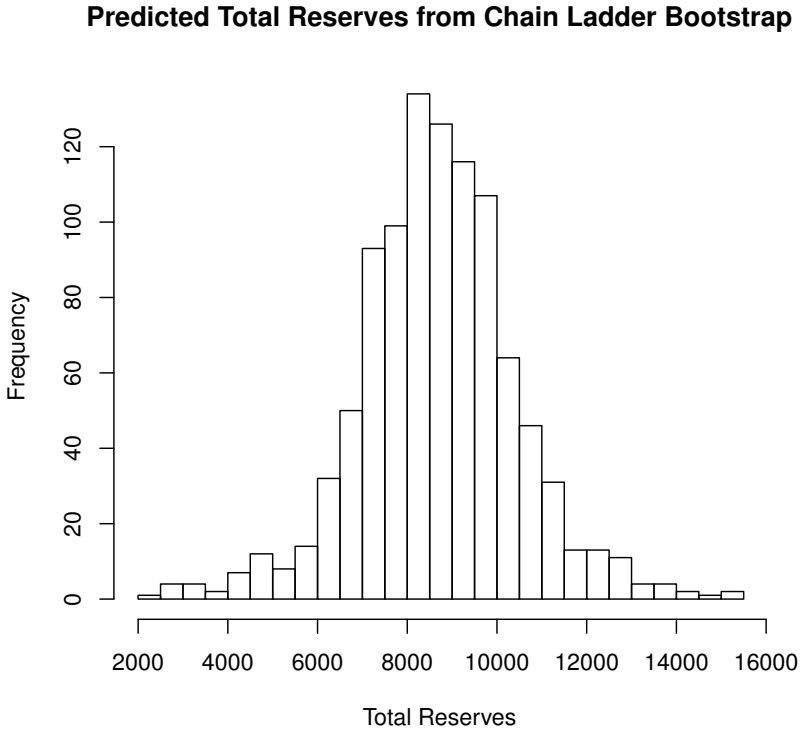

Figure 4: Predictive Distribution of Total Reserves from Chain Ladder Bootstrap

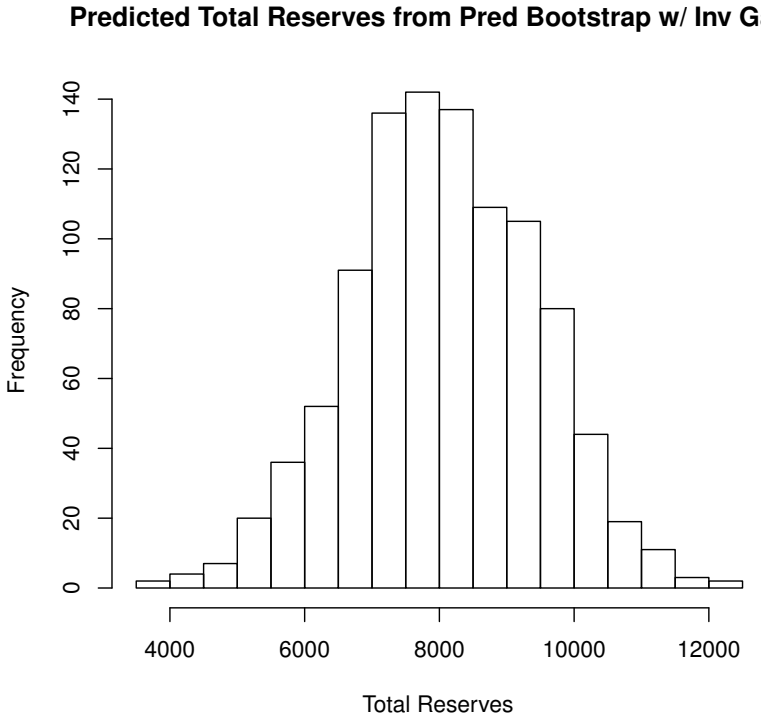

Figure 5: Predictive Distribution of Total Reserves from Pred Bootstrap with Inv Gauss

Comparing Figures 4 and 5, both methods produce total reserve values that have approximately symmetric distributions. The Chain Ladder Bootstrap appears to have a peaked distribution compared to the flat distribution from the Prediction Bootstrap using Inverse Gaussian. This is caused from the larger standard deviation in the Chain Ladder Bootstrap. Graphs like these could be useful when the small company is deciding how much total reserves to set aside because they can see where their amount falls on the graph and how often this was the simulated total.

Aside from average predicted IBNR values and their distributions, the prediction error must be considered when determining which method to use for the estimations. The prediction error for each bootstrap method can be found using Equation 21. A small prediction error is desirable because it indicates that the method is predicting future claims well. The prediction errors for each bootstrap method are shown in Table 14 as a percentage. This percentage is found by dividing the prediction error by the estimated IBNR value for that accident year. In this way the errors from the various methods can be compared.

The percent of prediction error decreases for later accident years. This decrease is consistent with the spread of the residuals becoming smaller for later accident years in the NAIC boxplots. The bootstrap methods that use the Chain Ladder fitted and predicted values have some of the smallest prediction errors. The Prediction Bootstraps that use a GLM approach for fitted values have the largest prediction errors for early accident years, but these decrease quickly for the last accident years. The model that uses values from Gamma GLM has the highest prediction errors which is to be expected as this method has been shown to be the least accurate in the analysis so far.

Table 14: Prediction Errors (%)

Accident Year	CL Bootstrap	Gamma Bootstrap	Pred Bootstrap
2003	431	1058	432
2004	202	327	221
2005	169	280	182
2006	139	240	144
2007	137	231	140
2008	93	148	106
2009	66	104	75
2010	64	97	63
2011	50	66	43
Total	34	46	29

Accident Year	Pred w/ Gamma	Pred w/ Inv Gaus
2003	1153	1421
2004	307	327
2005	717	290
2006	549	227
2007	165	244
2008	2156	107
2009	56	77
2010	107	45
2011	31	31
Total	31	25

From all of the analysis done on the small company data it has been found that most of the bootstrap methods produce results that are comparable to one another. These results have similar distributions and variation estimates. The variance estimates can be used to find percentiles which will allow the company to see an upper bound on the IBNR predictions. From the Prediction Errors it appears that using the Chain Ladder method for fitted/predicted values when performing bootstrapping will give the most accurate estimates.

6. CONCLUSION

Accurate predictions of future losses are essential for insurance companies to set premium rates. This thesis studies various loss reserving methods in an attempt to predict future reserves for a small Midwestern insurance company. The deterministic Chain Ladder method is easy to perform and produces accurate and trusted results but does not give estimates of variability. The bootstrap technique was introduced as a way to produce variability estimates. This technique creates pseudo data sets by repeatedly resampling from residuals that are found from Chain Ladder or Generalized Linear Model fitted values. The pseudo data sets create a predicted distribution from which reserve estimates and their variability can be found.

Various bootstrap methods were implemented on fully known loss data from the National Association of Insurance Commissioners. The results of the bootstrap methods were then compared to the actual observed losses. It was found that the bootstrapping done with GLM estimates produced reserves that were least similar to those found using the Chain Ladder method. The Gamma GLM residuals were the largest which implies that these results are the least accurate. The Inverse Gaussian GLM residuals were the smallest, indicating that this is the best bootstrap method for this data. Inverse Gaussian has not often been used in loss reserving but as these results show, should be studied further. Also, the two bootstrap techniques that use Chain Ladder fitted values had comparable residuals and are still considered accurate.

The bootstrap methods were then applied to the Midwestern insurance company data. The predicted reserves varied for the different methods with the Gamma GLM producing the smallest reserves values and the Chain Ladder bootstrap producing the largest. It was concluded that the Gamma GLM bootstrap is the least accurate and would lead to under-estimation of reserves leaving the insurance company unable to pay future liabilities. The other bootstrap methods produced similar total reserve values with approximately symmetric distributions whose standard deviations vary depending on the method. The distributions can be used to find upper percentiles of the total reserves to give the company a better idea of the predicted total reserve amount.

Although the bootstrap methods produced similar estimates for the Midwestern company data, prediction error must be considered when determining which estimates to use. The methods that use GLM fitted values had much higher prediction error percents that didn't decrease as fast as

the other methods. This suggests that using bootstrap techniques with Chain Ladder fitted values will be more accurate. Now the Midwestern company can use the average total reserve values along with the upper percentiles from the Chain Ladder Bootstrap and Prediction Bootstrap using Chain Ladder to get an estimate of future reserves and how much the reserves are expected to vary.

REFERENCES

About the NAIC. National Association of Insurance Commissioners. (n.d.). Retrieved from http://www.naic.org/index.htm (Last accessed April 20, 2015).

Barnett, G., & Zehnwirth, B. (2008). The Need for Diagnostic Assessment of Bootstrap Predictive Models. UNSW Australian School of Business Research Paper No. 2008ACTL04. Available at http://dx.doi.org/10.2139/ssrn.1134607 (Last accessed April 20, 2015).

Casualty Actuarial Society. Loss reserving data pulled from NAIC schedule P. Retrieved from http://www.casact.org/research/index.cfm?fa=loss reserves data (Last accessed April 20, 2015).

Efron, B. (1979). Bootstrap Methods: Another Look At The Jackknife. The Annals of Statistics, 7(1), 1-26.

England, P. (2002). Addendum to "Analytic and bootstrap estimates of prediction errors in claims reserving". *Insurance: Mathematics and Economics*, 31(3), 461-466.

England, P., & Verrall, R. (1999). Analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance: Mathematics and Economics*, 25(3), 281-293.

England, P., & Verrall, R. (2001). A Flexible Framework for Stochastic Claims Reserving. *Proceedings of Casualty Actuarial Society*, LXXXVIII, 1-38.

England, P., & Verrall, R. (2002). Stochastic Claims Reserving in General Insurance. *British Actuarial Journal*, 8(3), 443-518.

England, P., & Verrall, R. (2006). Predictive Distributions of Outstanding Liabilities in General Insurance. *Annals of Actuarial Science*, 1(2), 221-270.

Kaas, R., Goovaerts, M., Dhaene, J., & Denuit, M. (2009). IBNR Techniques. In *Modern actuarial risk theory using R* (2nd ed., pp. 265-291). Berlin: Springer-Verlag.

Martínez-Miranda, M., Nielsen, J., & Verrall, R. (2012). Double Chain Ladder. *Astin Bulletin*, 42(1), 59-76.

Peters, G., Wüthrich, M., & Shevchenko, P. (2010). Chain ladder method: Bayesian bootstrap versus classical bootstrap. *Insurance: Mathematics and Economics*, 47, 36-51.

Pinheiro, P., Silva, J., & Centeno, M. (2003). Bootstrap Methodology in Claim Reserving. *Journal of Risk and Insurance*, 70(4), 701-714.

Rempala, G., & Derrig, R. (2005). Modeling Hidden Exposures in Claim Severity Via the EM Algorithm. *North American Actuarial Journal*, 9(2), 108-128.

Rizzo, M. (2008). Bootstrap and Jackknife. In *Statistical computing with R* (pp. 183-211). Boca Raton: Chapman & Hall/CRC.

Shapland, M., & Leong, J. (2010). Bootstrap Modeling: Beyond the Basics. Casualty Actuarial Society E-Forum, Fall 2010. Retrieved from http://www.casact.org/pubs/forum/10fforum/ (Last accessed April 20, 2015).

Wüthrich, M., & Merz, M. (2008). Generalized Linear Models and Bootstrap Methods. In *Stochastic claims reserving methods in insurance* (pp. 201-255). Chichester, England: John Wiley & Sons.

APPENDIX

A.1. Boxplots of Residuals from NAIC Data

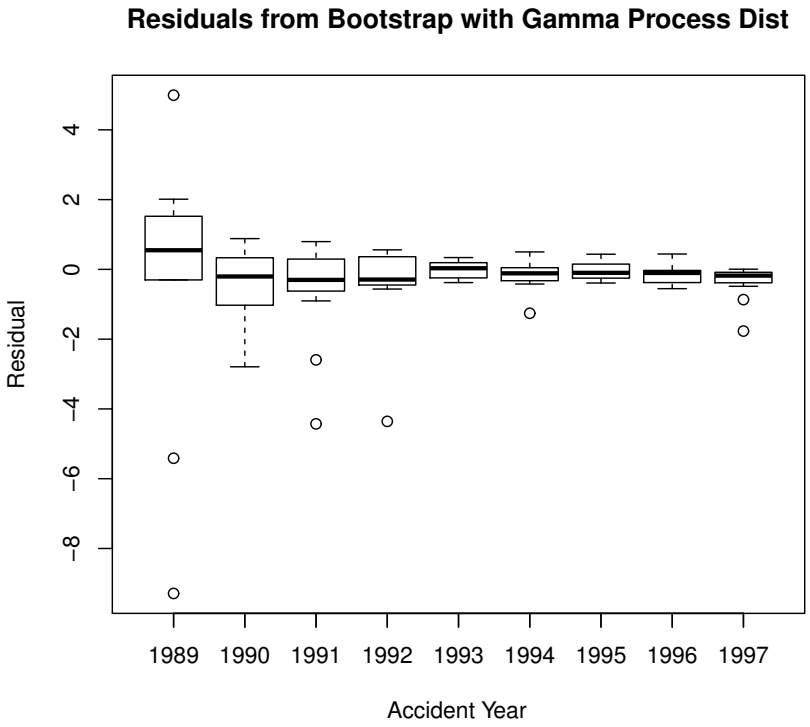

Figure A1: Residual plots of NAIC data using Gamma Bootstrap Method

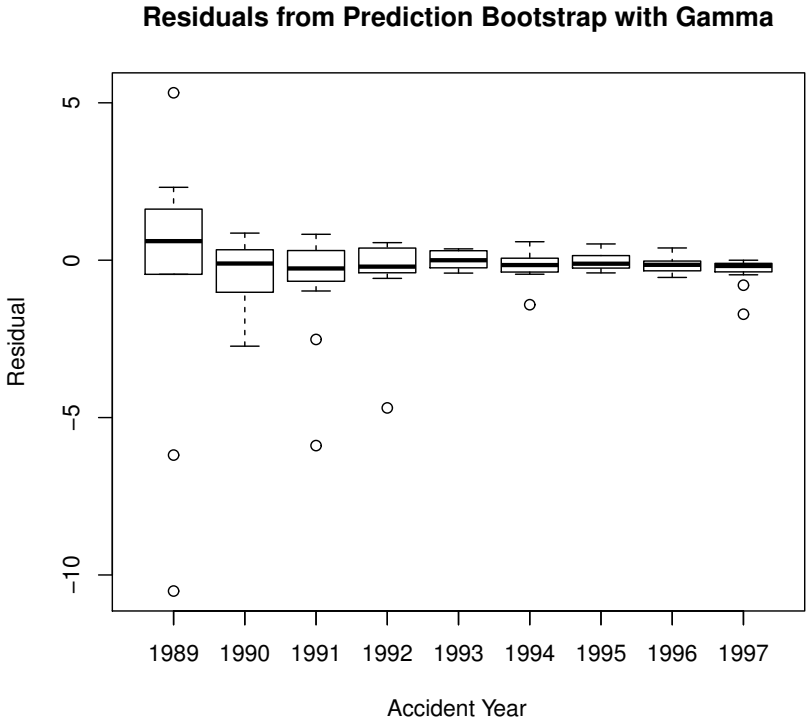

Figure A2: Residual plots of NAIC data using Bootstrap on Prediction with Gamma GLM

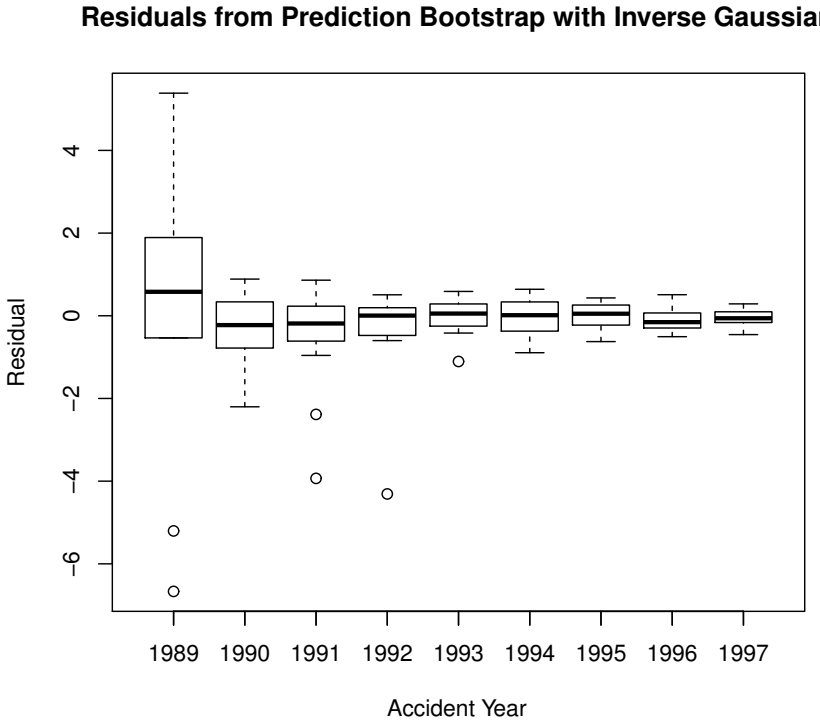

Figure A3: Residual plots of NAIC data using Bootstrap on Prediction with Inv Gauss GLM

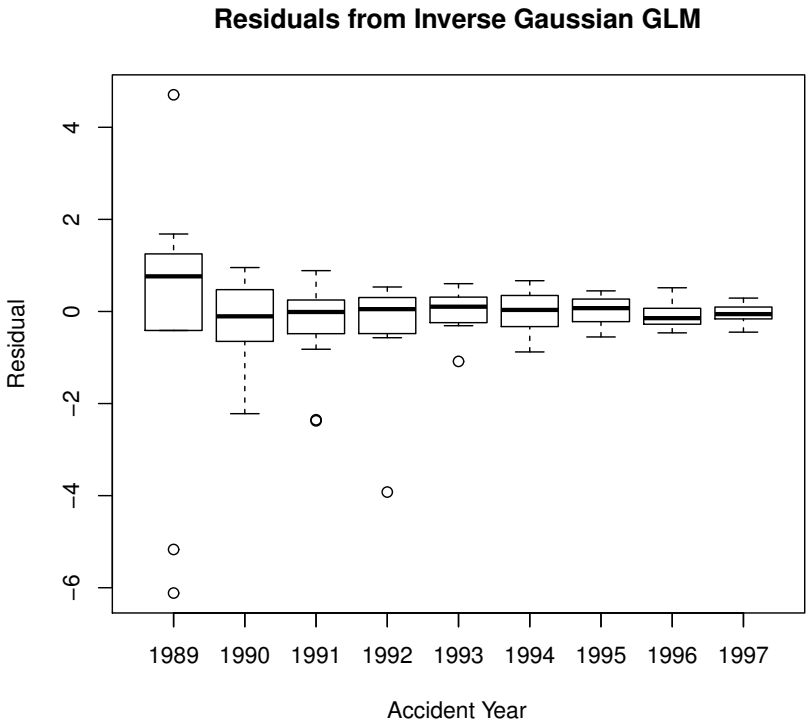

Figure A4: Residual plots of NAIC data using Inverse Gaussian GLM Method

A.2. Histograms of Total IBNR values from Midwestern company data

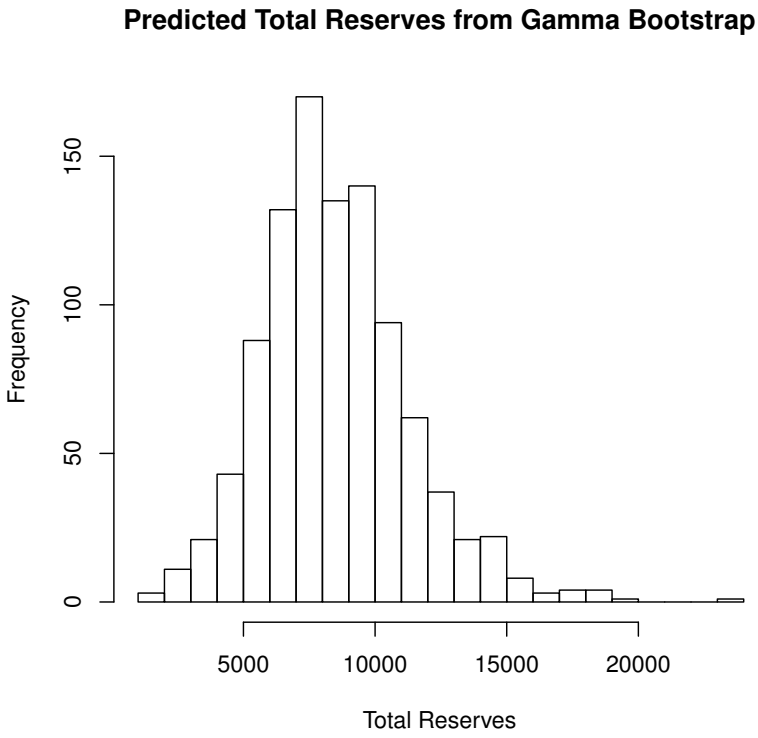

Figure A5: Predictive Distribution of Total Reserves from Gamma Bootstrap

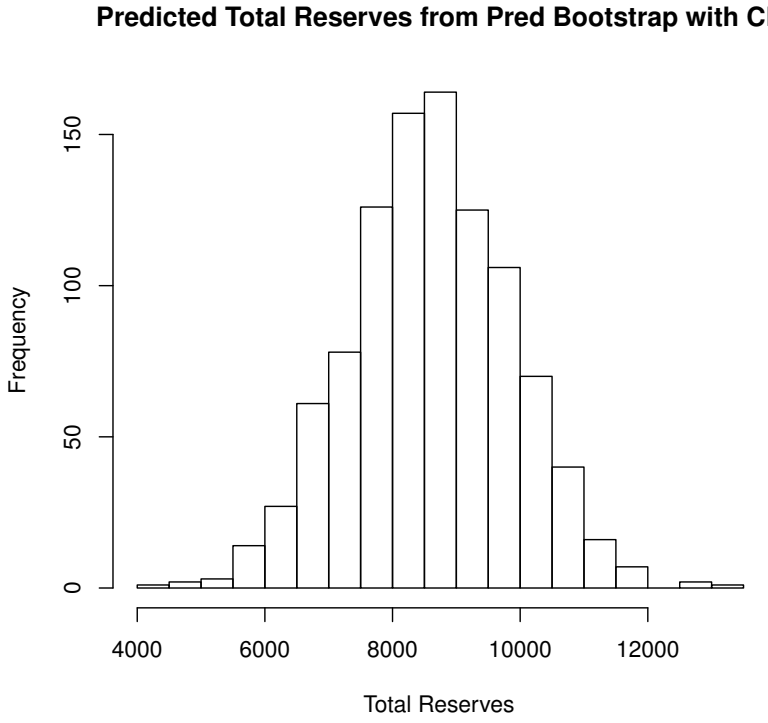

Figure A6: Predictive Distribution of Total Reserves from Pred Bootstrap with Chain Ladder

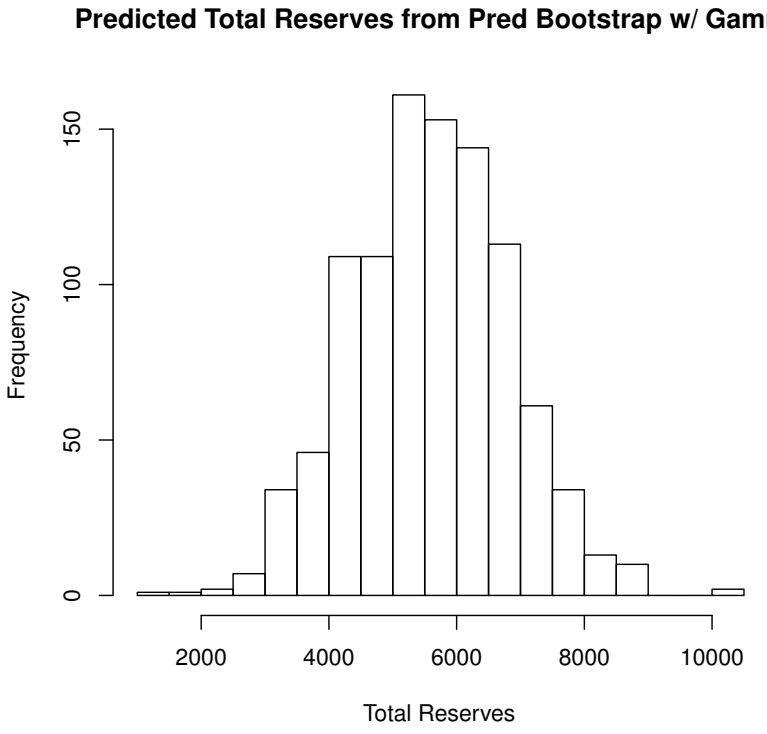

Figure A7: Predictive Distribution of Total Reserves from Pred Bootstrap with Gamma GLM