

Casualty Actuarial Society E-Forum, Summer 2011



The CAS *E-Forum*, Summer 2011

The Summer 2011 Edition of the CAS *E-Forum* is a cooperative effort between the Committee for the CAS *E-Forum* and various other CAS committees.

This E-Forum includes one call paper and two additional papers. The call for papers was conducted by the CAS Committee on Dynamic Risk Modeling and the Committee on Reserves on the topic “Testing Loss Reserving Methods, Models and Data Using the Loss Simulation Model.” The paper is scheduled to be presented at the 2011 Casualty Loss Reserve Seminar on September 15-16, at the ARIA Resort & Casino in Las Vegas, Nevada.

Committee on Dynamic Risk Modeling

Robert A. Bear, *Chairperson*

Fernando Alberto Alvarado

Morgan Haire Bugbee

Alp Can

Chuan Cao

Patrick J. Crowe

Christopher Diamantoukos

Sholom Feldblum

Stephen A. Finch

Bo Huang

Ziyi Jiao

Steven M. Lacke

Zhe Robin Li

Allen C. Long

Christopher J. Luker

Joseph O. Marker

Douglas W. Oliver

Ying Pan

Theodore R. Shalack

Zhongmei Su

Min Wang

Steve Winstead

Lin James Xia

Wei Xie

Yuanhe (Edward) Yao

Kun Zhang

Barry C. Zurbuchen

Karen Sonnet,

CAS Staff Liaison

Committee on Reserves

Mark R. Shapland, *Chairperson*

Lynne M. Bloom, *Vice Chairperson*

Alp Can

Andrew Martin Chandler

Kofi Gyampo

Aaron M. Halpert

Gloria A. Huberman

Dana F. Joseph

Weng Kah Leong

Jon W. Michelson

Marc B. Pearl

Susan R. Pino

Vladimir Shander

HongTao Wang

Bryan C. Ware

Ernest I. Wilson

Jianlu Xu

Cheri Widowski,

CAS Staff Liaison

CAS *E-Forum*, Summer 2011

Table of Contents

CAS Call Paper on “Testing Loss Reserving Methods, Models and Data Using the Loss Simulation Model”

Loss Simulation Model Testing and Enhancement

Kailan Shang FSA, CFA, PRM, SCJP..... 1-75

Additional Papers

GLM Invariants

Fred Klinker, FCAS, MAAA..... 1-10

The Retrospective Testing of Stochastic Loss Reserve Models

Glenn Meyers, FCAS, MAAA, CERA, Ph.D., and Peng Shi, ASA, Ph.D..... 1-37

***E-Forum* Committee**

Windrie Wong, *Chairperson*

Mark A. Florenz

Karl Goring

Dennis L. Lange

Elizabeth A. Smith, *Staff Liaison*

John Sopkowicz

Zongli Sun

Yingjie Zhang

For information on submitting a paper to the *E-Forum*, visit <http://www.casact.org/pubs/forum/>.

Loss Simulation Model Testing and Enhancement

Kailan Shang FSA, CFA, PRM, SCJP

Abstract. This paper is a response to the Casualty Actuarial Society's call for papers on the topic of "Testing Loss Reserving Methods, Models and Data Using the Loss Simulation Model." Its goal is to test and improve the Loss Simulation Model (LSM). The testing methods used are good sources for analyzing real claim data. A two-state regime-switching feature is also built into the model to add an extra layer of flexibility to describe claim data.

Motivation. The testing and enhancement of the Loss Simulation Model helps improve and refine the model. The test method may also be a good reference for performing tests on real claim data.

Method. Statistical tests are applied to the data simulated by the Loss Simulation Model. Standard distribution fitting methods such as maximum likelihood estimation are used to analyze real claim data. The open-source software LSM is enhanced via programming in Visual Basic.

Results. The LSM is enhanced with two-state regime-switching capability. Testing of the Loss Simulation Model according to the list suggested by the Loss Simulation Model Working Party is conducted. It shows the consistency between model input and model output for the addressed issues except case reserve adequacy.

Conclusions. Categorical variable and two-state regime-switching capability are added to the LSM. Testing of the LSM increases the confidence in the accuracy of this advanced and useful tool.

Keywords. simulation model; loss reserving; regime-switching; copula.

Table of Contents

LOSS SIMULATION MODEL TESTING AND ENHANCEMENT	1
1. INTRODUCTION.....	3
1.1 Research Context.....	3
1.2 Objective.....	3
1.3 Outline	4
2. MODEL TESTING	4
2.1 Negative Binomial Frequency Distribution.....	5
2.2 Correlation.....	7
2.2.1 Clayton Copula	8
2.2.2 Frank Copula	10
2.2.3 Gumbel Copula	12
2.2.4 <i>t</i> Copula.....	14
2.2.5 Correlation between claim size and report lag.....	16
2.3 Severity trend	18
2.4 Alpha in Severity Trend.....	20
2.5 Case Reserve Adequacy Distribution	22
3. REAL DATA AND SIMULATED DATA	25
3.1 Property Line	26
3.2 Liability Line.....	29
3.3 Correlation.....	33
4. MODEL ENHANCEMENT	34
4.1 Two-State Regime-Switching Distribution.....	34
4.2 Testing.....	37
4.2.1 Frequency	39
4.2.2 Severity.....	44
4.2.3 Correlation.....	48
5. CONCLUSION AND FURTHER DEVELOPMENT	53
Acknowledgment	54
APPENDIX A. R CODE	55
A.1 Negative Binomial Frequency Distribution Testing.....	55
A.2 Correlation Test.....	56
A.3 Severity Trend.....	61
A.4 Alpha in Severity Trend.....	62
A.5 Case Reserve Adequacy	63
A.6 Real Claim Data Fitting	64
A.7 Two-State Regime-Switching Feature Testing.....	67
APPENDIX B. QUICK GUIDE FOR TWO-STATE REGIME-SWITCHING	71
5. REFERENCES	75
Biography of the Author.....	75

1. INTRODUCTION

This paper is a response to a call for papers by the Casualty Actuarial Society (CAS) on “Testing Loss Reserving Methods, Models and Data Using the Loss Simulation Model.”

1.1 Research Context

The loss simulation model (LSM) is a tool created by the CAS Loss Simulation Model Working Party (LSMWP)¹ to generate claims that can then be used to test loss reserving methods and models.² The LSMWP paper suggests some model enhancement and additional tests of the LSM.³ Based on the suggested list, additional tests are performed on the simulated results to test the correlation, severity trend, negative binomial distribution for frequency, and case reserve adequacy distribution. Real claim data are used to fit into distributions to determine parameters in LSM. The model is also enhanced by allowing a two-state regime-switching distribution model for both frequency and severity.

1.2 Objective

A. Model Testing

1. Frequency distribution testing

Test the Negative Binomial frequency distribution using various goodness-of-fit testing methods.

2. Test correlation

Test the frequency correlation between different lines for other copula types in addition to the normal copula: Frank, Gumbel, Clayton, and T copula. Those types of copulas are very important to capture the tail risk while the normal copula that has been tested by LSMWP assumes a linear correlation behavior.

Test the correlation between report lag and size of loss under a normal copula.

3. Severity trend and Alpha testing

Apply time series analysis techniques to find the trend and alpha parameters from simulated

¹ For more information about LSM and LSMWP, please visit <http://www.casact.org/research/lsmwp>.

² CAS Loss Simulation Model Working Party Summary Report, pages 4-5,

³ CAS Loss Simulation Model Working Party Summary Report, page 33, The paper addresses the first suggestion about model enhancement and tests 1, 2, 3, 5 of the LSM in the suggestion list.

Loss Simulation Model Testing and Enhancement

data and compared with parameter inputs to check the statistical credibility. Ordinary least square (OLS) method and hypothesis testing are applied to the deterministic time trend model.

4. Case Reserve Adequacy

A 40% time point case reserve adequacy distribution is tested against simulation model input.

B. Real Data and Simulated Data

Marine claim data are used to fit the distribution for frequency and severity using Maximum Likelihood Estimation (MLE) and OLS for trend and seasonality analysis. The correlation between different lines is also estimated. The estimated distribution type and parameters can then be input into Loss Simulation Model (LSM) for simulation and further testing of different reserve methods. This illustrates how to use real data to determine inputs for the LSM. Unfortunately, only final claim data are available and there is no detailed paid loss history. Therefore, the Meyers' Approach⁴ is not applied to test rectangles generated by the simulation model against those from the real data due to the missing details.

C. Model Enhancement

A categorical variable is included to enable setting parameters/distribution type for different states. A two-state regime-switching flexibility is then built in to enable moving from one state to the other state. The transition matrix of states from one period to another is an input table in the user interface. Hopefully, this can add the flexibility to mimic the underlying cycle we normally see in P&C business. The enhancement is intended for frequency and severity distribution. The simulated results based on this enhancement are also tested.

1.3 Outline

The remainder of the paper proceeds as follows. Section 2 will discuss the methodology and results of testing LSM. Section 3 will fit real claim data to distribution and determine trend parameters which are inputs for LSM. Section 4 will present the enhancement being made for the LSM. Section 5 will discuss the conclusion and potential further improvement of the LSM.

2. MODEL TESTING

The LSM is used to simulate claim and transaction data for testing. Once the simulator is run

⁴ CAS Loss Simulation Model Working Party Summary Report, pages. 7-8.

with specified parameters, the relevant R code in [Appendix A](#) is applied to the claim file and transaction file output from LSM. Running R code, process output data and apply statistical tests. A conclusion based on the statistical test results is then drawn for the addressed issues.

2.1 Negative Binomial Frequency Distribution

This test is to check if the simulated frequency result is consistent with the LSM input parameters for negative binomial distribution.⁵

Test Parameters:

- ✓ One Line with annual frequency Negative Binomial (size = 100, probability = 0.4)
- ✓ Monthly exposure: 1
- ✓ Frequency Trend: 1
- ✓ Seasonality: 1
- ✓ Accident Year: 2000
- ✓ Random Seed: 16807
- ✓ # of Simulations: 1000

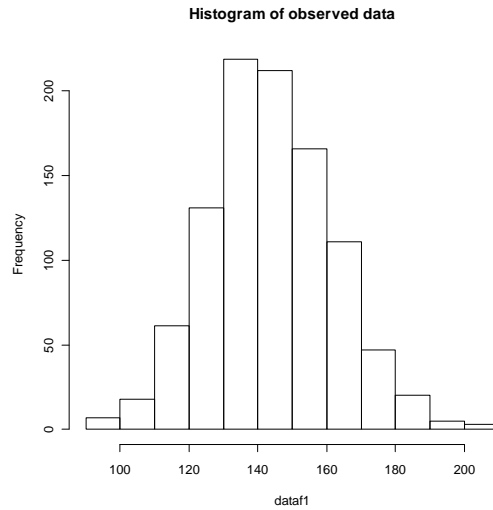
Firstly, we draw a histogram of the simulated frequency data to give an indication of the distribution type.

⁵ Negative Binomial Distribution: “A discrete probability distribution of the number of successes in a sequence of Bernoulli trials before a specified (non-random) number r of failures occurs.”

probability mass function as $\binom{k+r-1}{k} \cdot (1-p)^r \cdot p^k$ p : probability of success, k : number of successes.

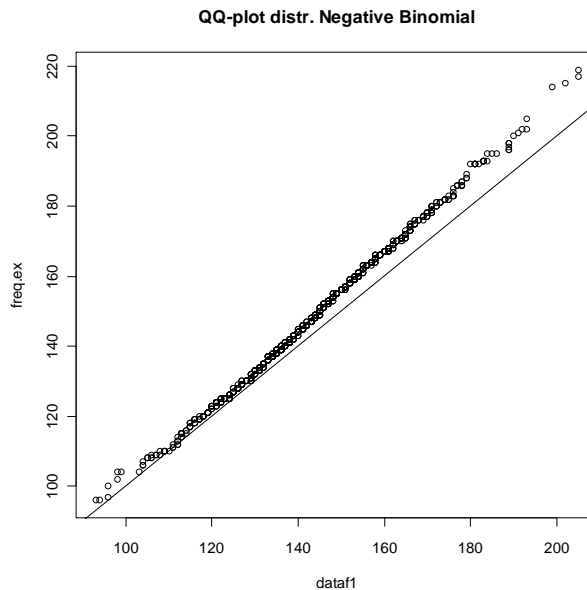
More details can be found at http://en.wikipedia.org/wiki/Negative_binomial_distribution.

Figure 1. Histogram of simulated frequency data (Negative Binomial)



A QQ plot would also be a straightforward way to compare the simulated results with the intended distribution – Negative Binomial (Size = 100, probability = 0.4). From Figure 2, we can see that it is a good fit although the expected frequency distribution in the LSM has a slightly longer tail than the simulated results.

Figure 2. QQ Plot – Simulated results vs. Negative Binomial (size = 100, prob. = 0.4)



Goodness-of-fit test using Pearson's Chi-squared statistic is performed. The results disallow rejecting the null hypothesis that the simulated frequency follows negative binomial distribution.

Goodness-of-fit test for nbinomial distribution

Loss Simulation Model Testing and Enhancement

	X^2	df	$P(> X^2)$
Pearson	197.3816	205	6.360712e-01

In addition, using maximum likelihood (ML) method to fit the negative binomial distribution and calculate the likelihood ratio statistics implies the same conclusion.

Goodness-of-fit test for nbinomial distribution

	X^2	df	$P(> X^2)$
Likelihood Ratio	113.3462	94	0.08499854

Using ML method gives us an estimation of the parameters as follows:

	size	mu
Estimation	117.2378284	144.1840000
Standard deviation	9.5150285	0.5670163

Our LSM inputs (size = 100 and prob = 0.4) imply $\mu = 150$ and variance = 375. The estimated value gives us size = 117 and prob = 0.448. The variance is 321.5. Here $\text{prob} = \text{size}/(\text{size} + \mu)$ and $\text{variance} = \mu + \mu^2/\text{size}$.⁶

We can see that at the significance level of 5%, the confidence interval for size is (98.59, 135.89) which includes the model input size = 100. The mean and variance of the model input and simulated results are also not too far away. Those results together with the goodness-of-fit tests indicate that simulated frequencies are consistent with the negative binomial distribution.

2.2 Correlation

In LSM, there are two places where correlation can be built between variables. One is the correlation between frequencies of different product lines. The other is the correlation between claim size and report lag. The method of modeling correlation in LSM is a copula, which can capture tail risk better than standard linear correlation assumption. Available copula types in LSM include Clayton, Frank, Gumbel, t , and normal copula. A normal copula among different lines' frequencies was tested and summarized in LSMWP paper.⁷

Sections 2.2.1 to 2.2.4 discuss the correlations among frequencies of different lines. Section 2.2.5

⁶ Package *stats* version 2.12.0, R Documentation, The Negative Binomial Distribution.

⁷ CAS Loss Simulation Model Working Party Summary Report, Section 6.2.3, pages 29-33.

discusses the correlation between claim size and report lag. In each section, once the simulator is run with these parameters, the R code in [Appendix A.2](#) is applied to the output claim file and/or transaction file. Running the code produces joint frequencies for two lines of correlated loss size and report lag. Statistical methods are then applied to test the consistency between model inputs and model outputs. Each section contains the model parameters used and a discussion of how well the copula fits the output of the simulation.

2.2.1 Clayton Copula

This test is to check if the Clayton Copula⁸ modeling in LSM is appropriate for correlation between frequencies of different lines.

Test Parameters:

- ✓ Two Lines with annual frequency Poisson ($\lambda = 96$)
- ✓ Monthly exposure: 1
- ✓ Frequency Trend: 1
- ✓ Seasonality: 1
- ✓ Accident Year: 2000
- ✓ Random Seed: 16807
- ✓ Frequency correlation: $\Theta = 5$, $n = 2$ (see footnote 8)
- ✓ # of Simulations: 1000

A simple way to compare is to draw a scatter plot for the intended copula and simulated frequency pairs. Figures 3 and 4 below show that they are of similar patterns.

⁸ Clayton Copula: $C_{\theta}^n(u) = (u_1^{-\theta} + u_1^{-\theta} + \dots + u_n^{-\theta} - n + 1)^{-1/\theta}$ $\theta > 0$. Details can be found on page 153 of Nelsen 2006.

Figure 3. Clayton Copula (5)

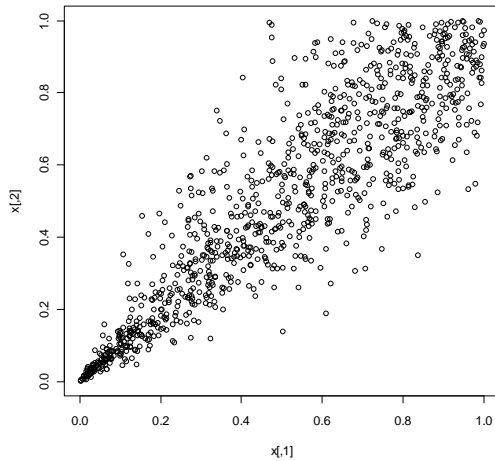
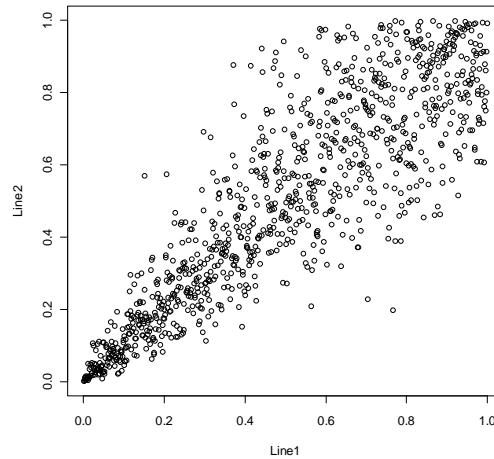


Figure 4. Simulated Frequencies



Clayton copula parameter is then estimated based on simulated frequency data using two methods.

- (1) The estimation is based on the maximum likelihood and a sample of size 998.

	Estimate	Std. Error	z value	$\Pr(> z)$
parameter	4.112557	0.1441209	28.53546	0

The maximized loglikelihood is 822.3826.

- (2) The estimation is based on the inversion of Kendall's tau and a sample of size 998.

	Estimate	Std. Error	z value	$\Pr(> z)$
parameter	4.623835	0.2434634	18.99191	0

We can see that the model parameter Θ as 5 is within 95% confidence interval based on inversion of Kendall's tau but not that for maximum likelihood estimation. This is also consistent with goodness-of-fit test results as below. We use two methods to test whether the correlation between simulated frequencies is consistent with assumed copula.

- (1) Using Maximum Likelihood method for parameter estimation:

Parameter estimate(s): 4.112557

Cramer-von Mises statistic:⁹ 0.03709138 with p -value 0.004950495

(2) Using Inversion of Kendall's tau method for parameter estimation:

Parameter estimate(s): 4.623835

Cramer-von Mises statistic: 0.01276128 with p -value 0.2623762

Based on Inversion of Kendall's tau method, we cannot reject the null hypothesis that the simulated frequencies have a relationship as the Clayton copula with $\Theta = 5$. But using Maximum Likelihood method, it is the opposite conclusion. It would be conservative for us not to reject the null hypothesis given the mixture of statistical test results.

2.2.2 Frank Copula

This test is to check if the Frank Copula¹⁰ modeling in LSM is appropriate for correlation between frequencies of different lines.

Test Parameters:

- ✓ Two Lines with annual frequency Poisson ($\lambda = 96$)
- ✓ Monthly exposure: 1
- ✓ Frequency Trend: 1
- ✓ Seasonality: 1
- ✓ Accident Year: 2000
- ✓ Random Seed: 16807
- ✓ Frequency correlation: $\Theta = 8$, $n = 2$ (see footnote 10)
- ✓ # of Simulations: 1000

A simple way to compare is to draw the scatter plot for the intended copula and simulated frequency pairs. Figures 5 and 6 below show that they are of the similar patterns.

⁹ $S_n^{(k)} = \sum_{i=1}^n \{C_n^{(k)}(\widehat{U}_i^{(k)}) - C_{\theta_n}^{(k)}(\widehat{U}_i^{(k)})\}^2$. Details can be found on page 6 of Kojadinovic and Yan 2010.

¹⁰ Frank Copula: $C_\theta^n(u) = -\frac{1}{\theta} \ln(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1) \dots (e^{-\theta u_n} - 1)}{(e^{-\theta} - 1)^{n-1}})$ $\theta > 0$ Details can be found on page 152 in Nelsen 2006,.

Figure 5. Frank Copula (8)

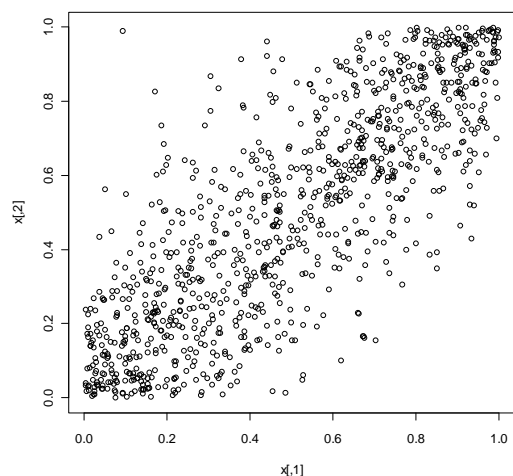
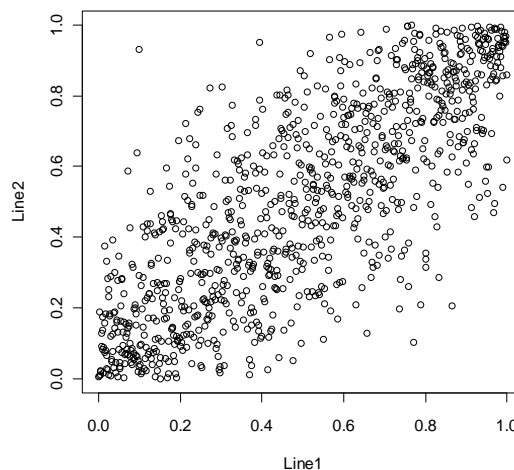


Figure 6. Simulated Frequencies



Frank copula parameter is then estimated based on simulated frequency data using two methods.

(1) The estimation is based on the maximum likelihood and a sample of size 1000.

	Estimate	Std. Error	$\hat{\alpha}$ value	$\Pr(> \hat{\alpha})$
parameter	7.508134	0.2770857	27.09679	0

The maximized loglikelihood is 455.8911.

(2) The estimation is based on the inversion of Kendall's tau and a sample of size 1000.

	Estimate	Std. Error	$\hat{\alpha}$ value	$\Pr(> \hat{\alpha})$
parameter	7.544506	0.3076033	24.52674	0

We can see that the model parameter Θ as 8 is within the 95% confidence interval based on either maximum likelihood or inversion of Kendall's tau.

Goodness-of-fit Test

(1) Using Maximum Likelihood method for parameter estimation:

Parameter estimate(s): 7.508134

Cramer-von Mises statistic: 0.01648723 with p -value 0.3118812

(2) Using Inversion of Kendall's tau method for parameter estimation:

Parameter estimate(s): 7.544506

Cramer-von Mises statistic: 0.01664421 with p -value 0.2029703

Based on those testing results, we cannot reject the null hypothesis that the simulated results are consistent with Frank Copula with Θ equal to 8.

2.2.3 Gumbel Copula

This test is to check if the Gumbel Copula¹¹ modeling in LSM is appropriate for correlation among frequencies of different lines.

Test Parameters:

- ✓ Two Lines with annual frequency Poisson ($\lambda = 96$)
- ✓ Monthly exposure: 1
- ✓ Frequency Trend: 1
- ✓ Seasonality: 1
- ✓ Accident Year: 2000
- ✓ Random Seed: 16807
- ✓ Frequency correlation: $\Theta = 6$, $n = 2$ (see footnote 11)
- ✓ # of Simulations: 1000

A simple way to compare is to draw the scatter plot for the intended copula and simulated frequency pairs. Figures 7 and 8 below show that they are of similar patterns.

¹¹ Gumbel Copula: $C_{\theta}^n(u) = \exp(-[(-\ln u_1)^{\theta} + (-\ln u_2)^{\theta} + \dots + (-\ln u_n)^{\theta}]^{1/\theta})$ $\theta \geq 1$. Details can be found on page 153 in Nelsen 2006.

Figure 7. Gumbel Copula (6)

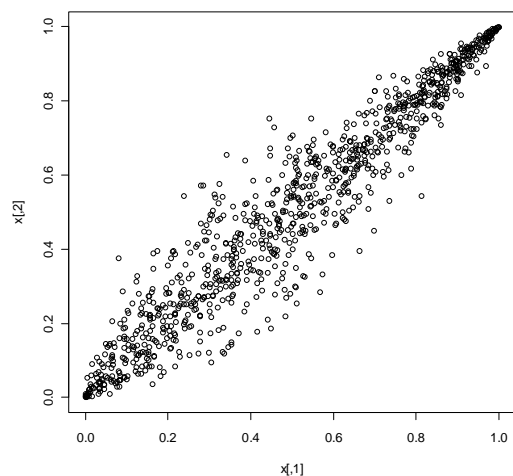
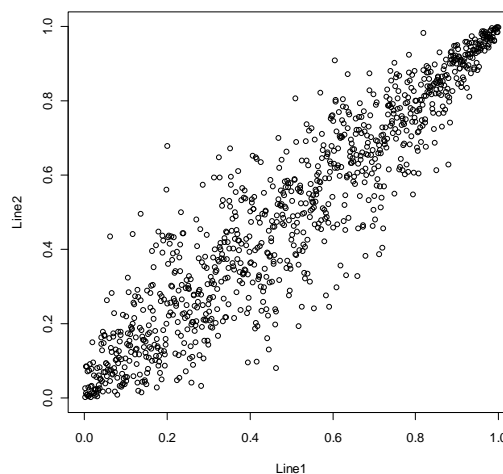


Figure 8. Simulated Frequencies



Gumbel copula parameter is then estimated based on simulated frequency data using two methods.

- (1) The estimation is based on the maximum likelihood and a sample of size 1000.

	Estimate	Std. Error	χ value	$\Pr(> \chi)$
parameter	4.223043	0.1111714	37.98677	0

The maximized loglikelihood is 1038.727.

- (2) The estimation is based on the inversion of Kendall's tau and a sample of size 1000.

	Estimate	Std. Error	χ value	$\Pr(> \chi)$
parameter	4.419024	0.1603205	27.56369	0

We can see that the model parameter Θ as 6 is out of the 95% confidence interval based on either maximum likelihood or inversion of Kendall's tau.

Goodness-of-fit Test

- (1) Using Maximum Likelihood method for parameter estimation:

Parameter estimate(s): 4.223043

Cramer-von Mises statistic: 0.01498423 with p -value 0.1237624

- (2) Using Inversion of Kendall's tau method for parameter estimation:

Loss Simulation Model Testing and Enhancement

Parameter estimate(s): 4.419024

Cramer-von Mises statistic: 0.01063169 with p -value 0.2623762

Based on those testing results, we would reject the null hypothesis that the simulated results are consistent with Gumbel Copula with Θ equal to 6.

2.2.4 t Copula

This test is to check if the t Copula¹² modeling in LSM is appropriate for correlation between frequencies of different lines

Test Parameters:

- ✓ Two Lines with annual frequency Poisson ($\lambda = 96$)
- ✓ Monthly exposure: 1
- ✓ Frequency Trend: 1
- ✓ Seasonality: 1
- ✓ Accident Year: 2000
- ✓ Random Seed: 16807
- ✓ Frequency correlation: v (degree of freedom) = 5, correlation = 0.8, $n = 2$ (see footnote 12)
- ✓ # of Simulations: 1000

A simple way to compare is to draw the scatter plot for the intended copula and simulated frequency pairs. Figures 9 and 10 below show that they are of the similar patterns.

¹² t Copula, or Student t copula, $C_{v,\Sigma}^n(u) = T_{v,\Sigma}(T_v^{-1}(u_1), \dots, T_v^{-1}(u_n))$ v : degree of freedom, Σ : correlation matrix, T : t cumulative distribution function.

Figure 9. t Copula (dof = 5, 0.8)

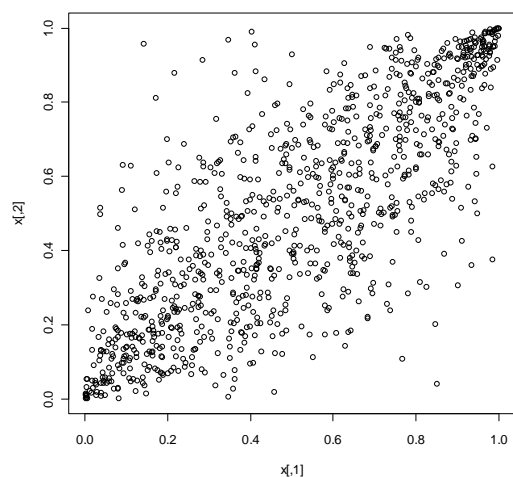
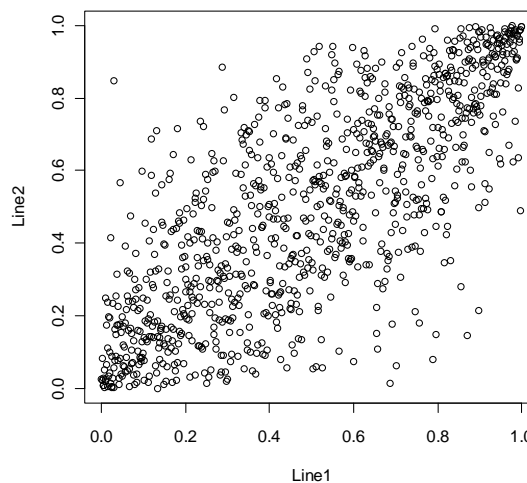


Figure 10. Simulated Frequencies



The t copula parameter is then estimated based on simulated frequency data using two methods.

(1) The estimation is based on the maximum likelihood and a sample of size 1000.

	Estimate	Std. Error	z value	$\Pr(> z)$
parameter	0.7614685	0.01254461	60.70086	0

The maximized loglikelihood is 444.3589.

(2) The estimation is based on the inversion of Kendall's tau and a sample of size 1000.

	Estimate	Std. Error	z value	$\Pr(> z)$
parameter	0.7840726	0.01343576	58.35713	0

We can see that the correlation assumption of 0.8 is within the 95% confidence interval based on inversion of Kendall's tau and within the 99% confidence interval based on maximum likelihood.

Goodness-of-fit Test

(1) Using Maximum Likelihood method for parameter estimation:

Parameter estimate(s): 0.7614685

Cramer-von Mises statistic: 0.04547016 with p -value 0.01485149

(2) Using Inversion of Kendall's tau method for parameter estimation:

Loss Simulation Model Testing and Enhancement

Parameter estimate(s): 0.7840726

Cramer-von Mises statistic: 0.0259301 with p -value 0.04455446

Based on those testing results, it is conservative for us not to reject the null hypothesis that the simulated results are consistent with the t Copula that has correlation = 0.8 and degree of freedom = 5.

2.2.5 Correlation between claim size and report lag

This test is to check if the correlation between claim size and report lag in LSM is appropriately modeled.

Test Parameters:

- ✓ One Line with annual frequency Poisson ($\lambda = 120$)
- ✓ Monthly exposure: 1
- ✓ Frequency Trend: 1.05
- ✓ Seasonality: 1
- ✓ Accident Year: 2000
- ✓ Random Seed: 16807
- ✓ Payment Lag: Exponential with rate = 0.002739726, which implies a mean of 365 days.
- ✓ Size of entire loss: Lognormal with mu = 11.16636357 and sigma = 0.832549779
- ✓ Correlation between payment lag and size of loss: normal copula¹³ with correlation = 0.85, dimension 2 (See footnote 13)
- ✓ # of Simulations: 10^{14}

A simple way to compare is to draw the scatter plot for the intended copula and simulated frequency pairs. Figures 11 and 12 below show that they are of similar patterns.

¹³ Normal Copula, or Gaussian Copula, $C_{\Sigma}^n(\mathbf{u}) = \Phi_{\Sigma}(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_n))$ Σ : correlation matrix

Φ : normal cumulative distribution function. Details can be found on pages 43-54 of Li 2000.

¹⁴ The reason to use 10 simulations instead of 1000 simulations is that 120 claims are expected (Frequency distribution with $l = 120$) in each simulation. The total expected number of pairs of data is 1200 with 10 simulations for correlation analysis.

Figure 11. Normal Copula (0.85)

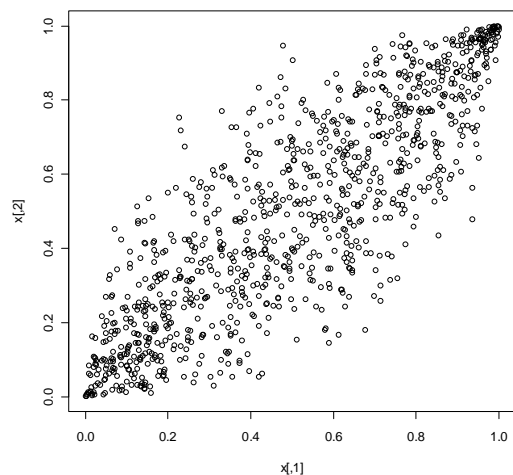
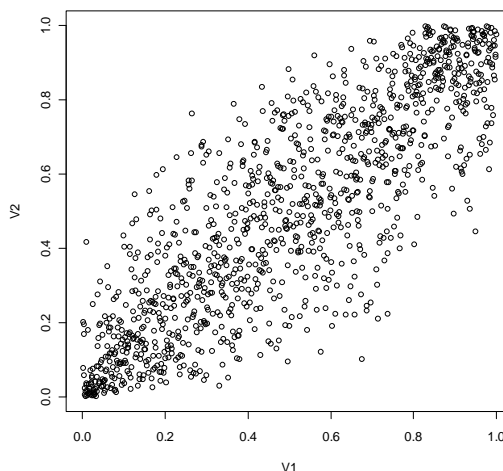


Figure 12. Simulated claim size vs. report lag



The normal copula parameter is then estimated based on simulated frequency data using two methods.

- (1) The estimation is based on the maximum likelihood and a sample of size 1000.

	Estimate	Std. Error	z value	$\Pr(> z)$
rho.1	0.8317376	0.006878922	120.9110	0

The maximized loglikelihood is 694.6756.

- (2) The estimation is based on the inversion of Kendall's tau and a sample of size 1000.

	Estimate	Std. Error	z value	$\Pr(> z)$
parameter	0.8538963	0.007917961	107.8430	0

We can see that the correlation assumption (0.85) is within the 95% confidence interval based on inversion of Kendall's tau.

Goodness-of-fit Test

- (1) Using Maximum Likelihood method for parameter estimation:

Parameter estimate(s): 0.8317376

Cramer-von Mises statistic: 0.06218935 with p -value 0.004950495

- (2) Using Inversion of Kendall's tau method for parameter estimation:

Parameter estimate(s): 0.8538963

Loss Simulation Model Testing and Enhancement

Cramer-von Mises statistic: 0.02898052 with p -value 0.01485149

Based on those testing results, we would reject the null hypothesis at the significance level larger than 1.5% that the simulated results are consistent with Normal Copula that has correlation = 0.85. The difference in the value of correlation coefficients between model input and model output is not small. However, the simulated data still have a strong correlation as intended.

2.3 Severity trend

This test is to check if the severity trend in LSM is modeled as intended.

Test Parameters:

- ✓ One Line with annual frequency Poisson ($\lambda = 96$)
- ✓ Monthly exposure: 1
- ✓ Frequency Trend: 1
- ✓ Seasonality: 1
- ✓ Accident Years: 2000 to 2005
- ✓ Random Seed: 16807
- ✓ Size of entire loss: Lognormal with $\mu = 11.16636357$ and $\sigma = 0.832549779$
- ✓ Severity Trend: 1.5
- ✓ # of Simulations: 300

Figure 13 shows the mean value of loss size over time. There is a clear consistent trend. Figure 14 shows that Seasonal Decomposition of Time Series by Loess (STL),¹⁵ which decomposes a time series into seasonal, trend, and irregular components using loess.¹⁶ It is very obvious there is no seasonality and there exists an upward sloping trend. The residual errors behave like white noise.

¹⁵ Package *stats* version 2.12.0, R Documentation, Seasonal Decomposition of Time Series by Loess. A description of STL is available in Cleveland et al., 1990.

¹⁶ Loess stands for Locally Weighted Regression Fitting.

Figure 13. Mean Loss Size

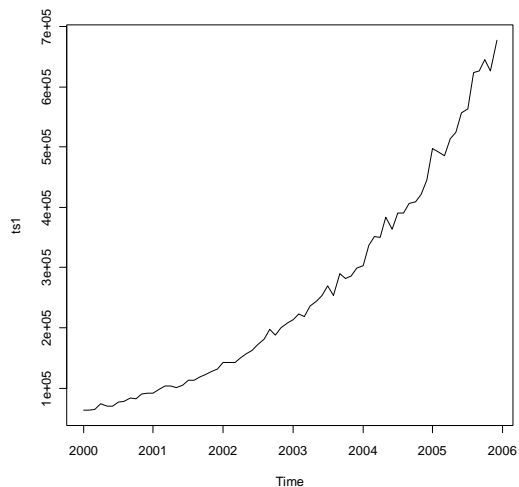
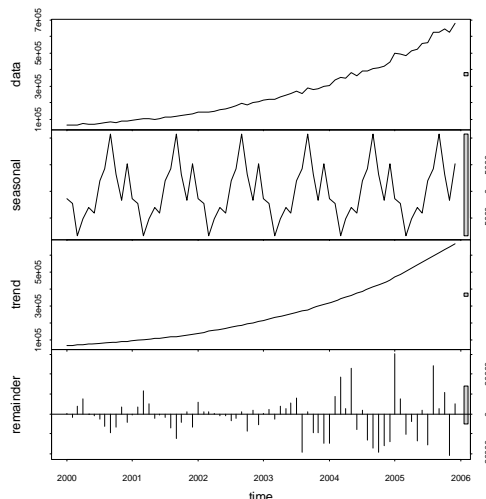


Figure 14. STL



Based on the log of mean loss size, a linear regression that estimates the linear trend factor supports our assumptions.

$$\text{Log}(\text{Mean Loss Size}) = \text{Intercept} + \text{trend} * (\text{time} - 2000) + \text{error term}$$

We get the following results using R.

Residuals:

Min	1Q	Median	3Q	Max
-0.051579	-0.023194	-0.007886	0.023918	0.078750

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	11.034162	0.007526	1466.1	<2e-16
trend	0.405552	0.002196	184.7	<2e-16

Residual standard error: 0.03226 on 70 degrees of freedom

Multiple R-squared: 0.998, Adjusted R-squared: 0.9979

F-statistic: 3.412e+04 on 1 and 70 DF, p-value: < 2.2e-16

We can see that the t test shows that the trend is not equal to 0 at a significant level less than 0.1%. The high adjusted R2 and the F test also show that the trend is obvious. The trend factor

0.405552 is based on the log of the mean loss size and is equivalent to the trend factor of 1.5 for loss size ($\exp(0.405552) = 1.50013$). This is also our model input. Figure 15 shows a good fitting of the regression. Residual graph (Figure 16) shows a white noise pattern. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) also support the existence of linear trend.

Figure 15. Trend fitting

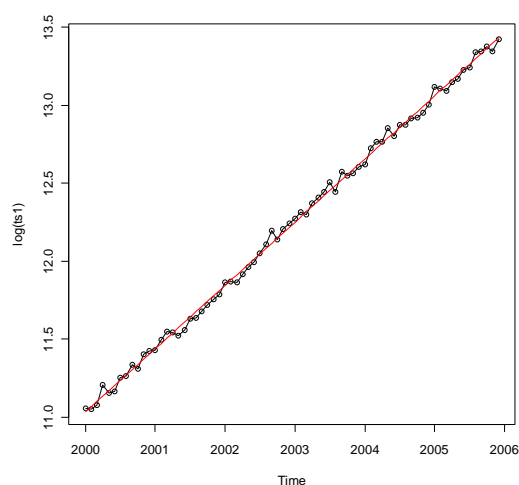
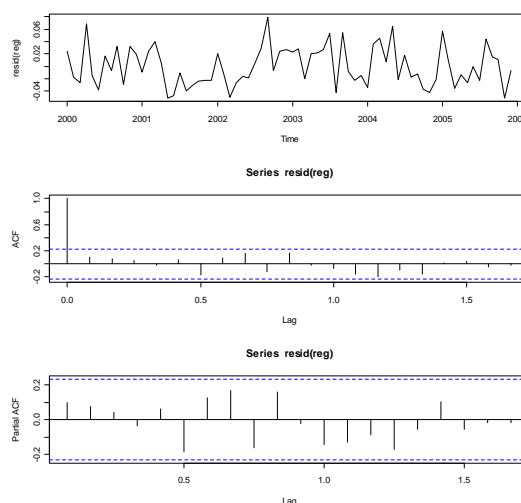


Figure 16. Residual, ACF, PACF



2.4 Alpha in Severity Trend

This test checks if the alpha that determines the persistency of the force of trend for severity in LSM is modeled as intended. As described in LSM,¹⁷ the cumulative trend amounts (cum) are calculated first and then the trend multiplier is calculated as

$$trend = (cum_{acc_date}) \left(\frac{cum_{pmt_date}}{cum_{acc_date}} \right)^\alpha = (cum_{acc_date})^{1-\alpha} (cum_{pmt_date})^\alpha .$$

Test Parameters:

- ✓ One Line with annual frequency Poisson ($\lambda = 96$)
- ✓ Monthly exposure: 1
- ✓ Frequency Trend: 1
- ✓ Seasonality: 1

¹⁷ CAS Loss Simulation Model Working Party Summary Report, pages 66-67.

Loss Simulation Model Testing and Enhancement

- ✓ Accident Years: 2000 to 2001
- ✓ Random Seed: 16807
- ✓ Size of entire loss: Lognormal with $\mu = 11.16636357$ and $\sigma = 0.832549779$
- ✓ Severity Trend: 1.5
- ✓ Alpha: 0.4
- ✓ # of Simulations: 1000

We choose the sample loss payments with report date during the 1st month and payment date during the 7th month.

Therefore, the severity trend multiple is $(1.5^{1/12})^{(1-0.4)} \cdot (1.5^{7/12})^{0.4} \approx 1.122$ for those chosen claims.

The expected loss size is $1.122 \cdot e^{11.166+0.83255^2/2} \approx 112,175$.

The histogram and QQ plot show that the fit is not perfect, but not too far away.

Figure 17. Histogram of severity

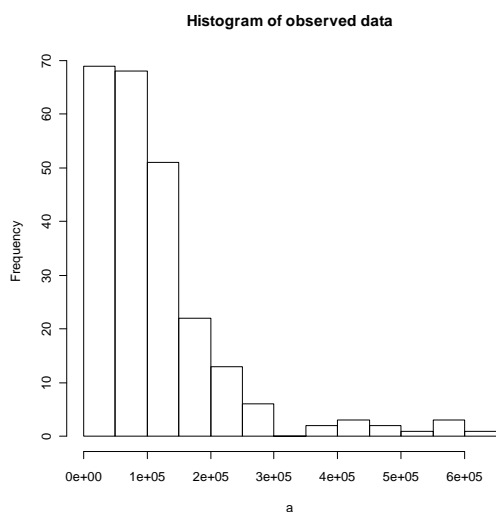


Figure 18. QQ plot of severity

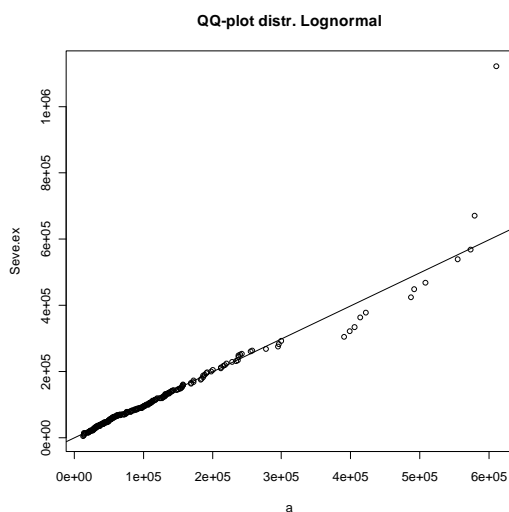
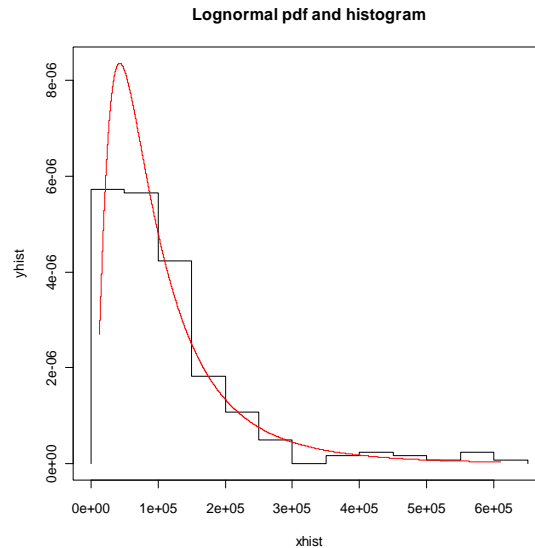


Figure 19. Histogram and fitted probability density function



Maximum likelihood estimation gives us the following fitted parameters and standard deviation. The mean value of severity is 113,346. When only volatility of meanlog estimation is considered, the mean loss derived by model input is within 95% confidence interval.

	meanlog	sdlog
Estimation	11.31595927	0.80279226
Standard Deviation	0.05171240	0.03656619

Results of Kolmogorov-Smirnov test and Anderson-Darling normality test support the lognormal distribution of the sampled payments.

One-sample Kolmogorov-Smirnov test

$D = 0.0405, p\text{-value} = 0.8249$

alternative hypothesis: two-sided

Anderson-Darling normality test

$A = 0.4114, p\text{-value} = 0.3384$

2.5 Case Reserve Adequacy Distribution

In the LSM, the case reserve adequacy distribution parameters are intended to model

Loss Simulation Model Testing and Enhancement

characteristics of an insurer's case loss reserving process. For example, some insurers set a nominal reserve until a claim is investigated while others may set up a formula or "average" reserve initially. The ultimate claim value may be the same in both cases, but the timing and amount of the reserve changes may be quite different. The case reserve adequacy distribution attempts to model this process by generating case reserve adequacy ratio at each valuation date. Case reserve is determined by multiplying the generated final claim amount by case reserve adequacy ratio.

Notice that, for simulated data, the case reserve adequacy parameters do not affect the ultimate claim value. However, in determining LSM parameters from real data where some of the accident years are not fully developed, the case reserve adequacy assumption may be crucial.

This test is to check if the $X\%$ time point case reserve adequacy distribution in LSM is modeled as intended. We choose the 40% time point¹⁸ in this paper.

Test Parameters

- ✓ One Line with annual frequency Poisson ($\lambda = 96$)
- ✓ Monthly exposure: 1
- ✓ Frequency Trend: 1
- ✓ Seasonality: 1
- ✓ Accident Years: 2000 to 2001
- ✓ Random Seed: 16807
- ✓ Size of entire loss: Lognormal with $\mu = 11.16636357$ and $\sigma = 0.832549779$
- ✓ 40% Case Reserve: Lognormal with $\mu = 0.25$ and $\sigma = 0.05$
- ✓ Severity Trend: 1
- ✓ $P(0) = 0.4$
- ✓ Est $P(0) = 0.4$
- ✓ # of Simulations: 8¹⁹

From the test assumption, we know that the mean 40% case reserve adequacy ratio is

¹⁸ The 40% time point is the date that is equal to the 60% Report Date + 40% Final Payment Date.

¹⁹ Similar reason as indicated in footnote 14 as the number of simulated claims is large enough for statistical testing with 8 simulations.

$e^{0.25+0.05^2/2} \approx 1.2856$. The transaction output is used to calculate the case reserve at 40% of payment lag using linear interpolation method. Those values are then used for testing purposes.

The histogram, QQ plot, and probability density function show that the fit is not good.

Figure 20. Histogram of severity

Figure 21. QQ Plot of severity

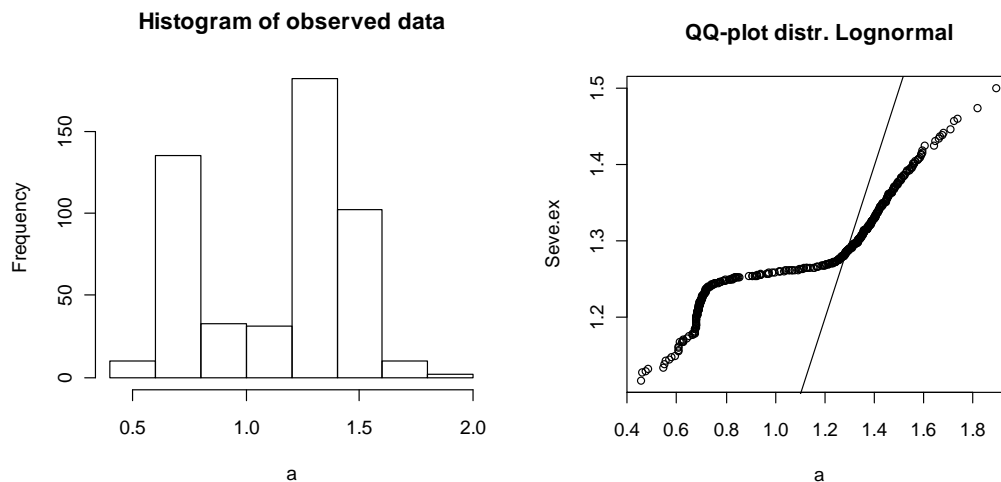
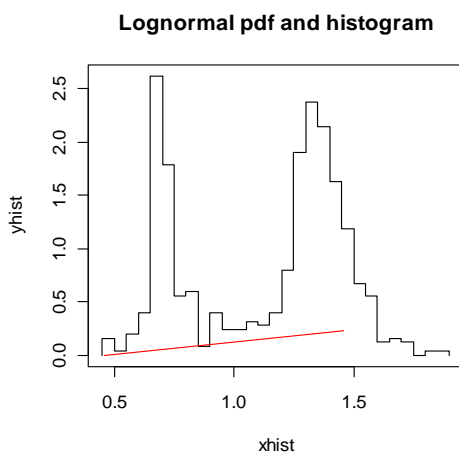


Figure 22. Histogram and fitted probability density function



Maximum likelihood estimation gives us the following fitted parameters and standard deviation. The mean value of severity is 1.141. When only volatility of meanlog estimation is considered, the mean loss derived by model input is within 95% confidence interval.

	meanlog	sdlog
Estimation	0.07973950	0.32269631

Loss Simulation Model Testing and Enhancement

Standard Deviation 0.01435980 0.01015391

Results of Kolmogorov-Smirnov test and Anderson-Darling normality test do not support the lognormal distribution of the sampled case reserve adequacy.

One-sample Kolmogorov-Smirnov test

$D = 0.3869, p\text{-value} < 2.2e-16$

Anderson-Darling normality test

$A = 33.2183, p\text{-value} < 2.2e-16$

Model input and output are not consistent for both the distribution type and the fitted parameters. In the simulation, valuation dates of each claim are generated based on an assumption of waiting period (inter-valuation lag assumption). Before the final payment, case reserve is generated on the simulated valuation dates. Since valuation dates are randomly generated, it often does not coincide with the 40% time point. In those cases, linear interpolation method is used to get case reserve ratio at 40% time point for testing. On the first valuation date, i.e., the report date, a case reserve of 2,000 will be allocated for each claim without any adjustment related to the claim size. If the second valuation date happens after 40% time point, it is clear that linear interpolation method can give us false estimation of what is assumed in the model inputs. Therefore, there is no confident conclusion about whether the model is correct or not.

A way to overcome this is to change the way in which transaction date is determined. In current coding, report date and final settlement date are generated before transaction date and case reserves are generated. We can set a few transaction dates as report date + $X\%$ (payment date – report date) instead of generating them based on waiting period distribution assumption. $X\%$ could be 40%, 70%, and 90% to be consistent with current case reserve adequacy model input setting. In this way, linear interpolation is not needed anymore and the output data we got are also easier for testing the model and reserve methods.

3. REAL DATA AND SIMULATED DATA

Marine claim data are used for distribution fitting, trend analysis, and correlation analysis. Those estimated distributions and parameters could be input for LSM to generate stochastic claim data. Based on those claim data, reserve methods can be tested and evaluated. Unfortunately, paid loss

history data are not available in this example and Meyers' Approach²⁰ cannot be applied due to the lack of details. Below is a snapshot of the claim data used in this section. It has two product lines: Property and Liability. The data period is from 2006 to 2010. The number of accidents is 317 for Property Insurance and 428 for Liability Insurance. All the claims are closed with a final payment.

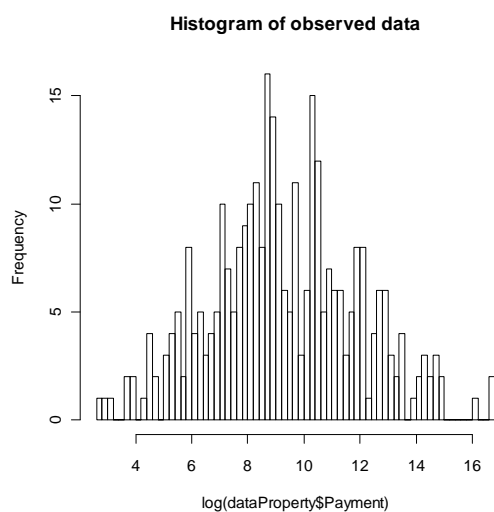
Accident Date	Payment data	Line	Final Payment
12/31/2006	3/30/2008	Property	249
5/1/2006	11/27/2006	Property	16,293
1/22/2010	4/22/2010	Property	65,130
1/22/2006	8/20/2006	Liability	38,544
7/27/2010	2/22/2011	Liability	13,206

3.1 Property Line

Fit the severity

1. Draw a histogram of logarithm of payment to find out the most appropriate claim-size distribution type. Lognormal distribution seems to be a good candidate for describing claim size.

Figure 23. Histogram of Log (Claim Size)



2. Use lognormal distribution fitting for claim size.

	meanlog	sdlog
Estimation	9.2848522	2.6269670

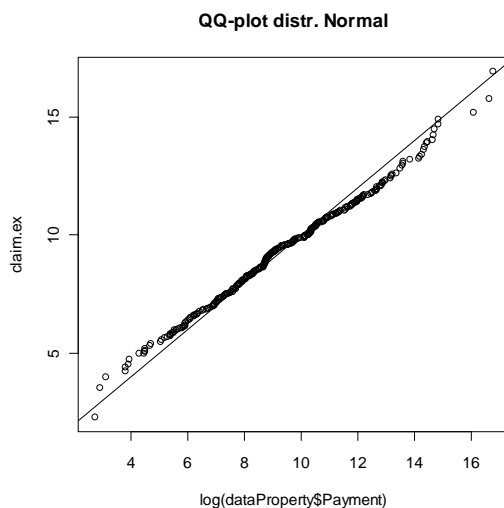
²⁰ CAS Loss Simulation Model Working Party Summary Report, pages 7-8.

Loss Simulation Model Testing and Enhancement

Standard Deviation 0.1484850 0.1049947

3. Use a QQ plot to check the fitting. It is not a perfect fitting but this is probably the best we can achieve.

Figure 24. QQ Plot of Log(Claim Size)



Fit the frequency

4. Draw a time series of frequency data and conduct a Seasonal Decomposition of Time Series. There is no strong evidence of linear trend and seasonality during this period.

Figure 24. Frequency

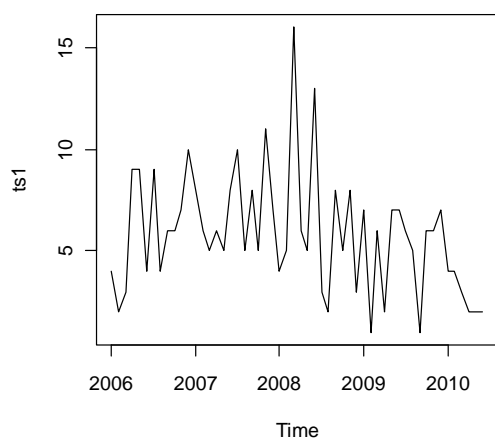
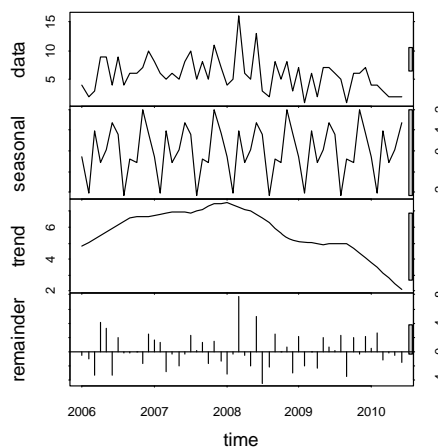


Figure 25. STL



Loss Simulation Model Testing and Enhancement

5. Perform a linear regression for trend analysis.

$$\text{Log(Monthly Frequency)} = \text{Intercept} + \text{trend} * (\text{time} - 2006) + \text{error term.}$$

Residuals:

Min	1Q	Median	3Q	Max
-1.48135	-0.36849	0.04697	0.38654	1.15768

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.93060	0.15164	12.732	<2e-16
trend	-0.14570	0.05919	-2.462	0.0172

Residual standard error: 0.5649 on 52 degrees of freedom.

Multiple R-squared: 0.1044, Adjusted R-squared: 0.08715.

F-statistic: 6.06 on 1 and 52 DF, p-value: 0.01718.

Figure 26. Trend fitting

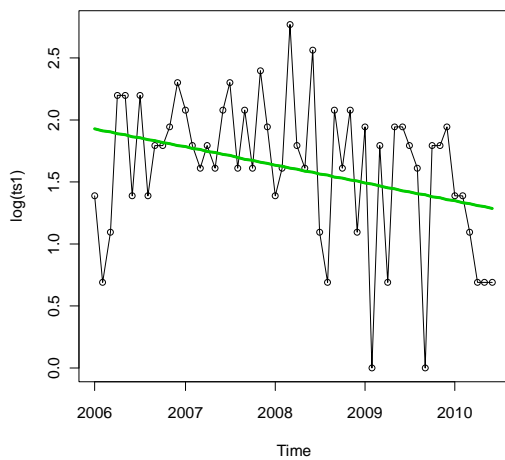
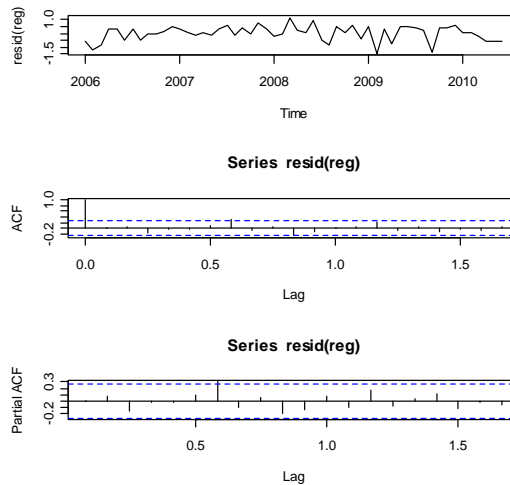


Figure 27. Residual, ACF, PACF



6. Detrend the frequency and fit to the frequency distribution.

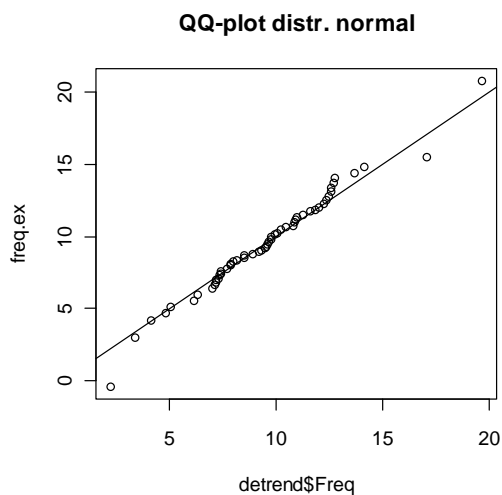
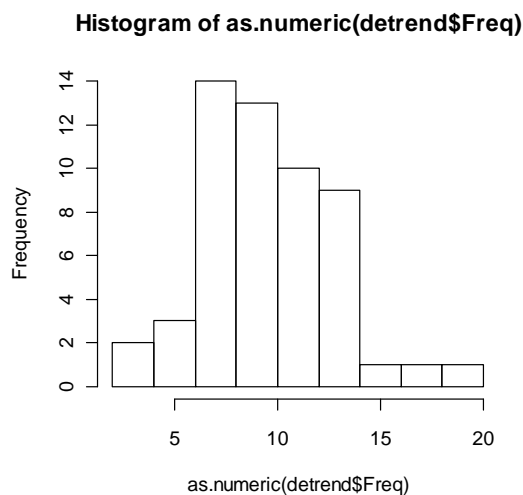
It looks like that lognormal distribution fits the detrended data better.

	meanlog	sdlog
Estimation	9.5539259	3.1311762

Standard Deviation 0.4260991 0.3012976

Figure 28. Histogram of detrend freq.

Figure 29. QQ Plot of detrend freq.



The Kolmogorov-Smirnov test result also supports lognormal distribution assumption.

One-sample Kolmogorov-Smirnov test is as follows:

$D = 0.0814$, $p\text{-value} = 0.8384$.

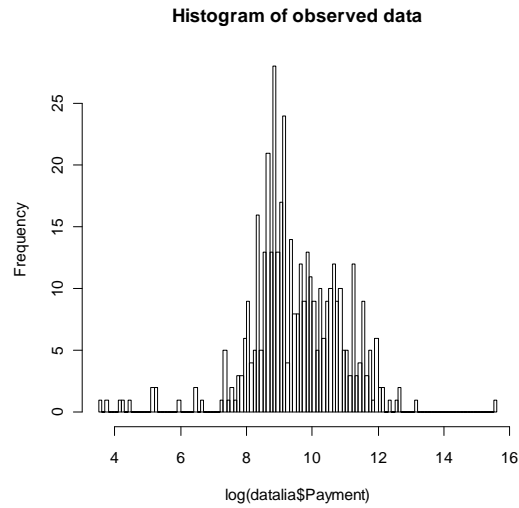
Therefore, we have all the parameters for frequency and severity distribution and trend of frequency for property line.

3.2 Liability Line

Fit the severity

1. Draw a histogram of the logarithm of payments to find out candidates for the distribution type of the claim size. Lognormal distribution seems to be a good candidate for describing claim size.

Figure 30. Histogram of Log (Claim Size)

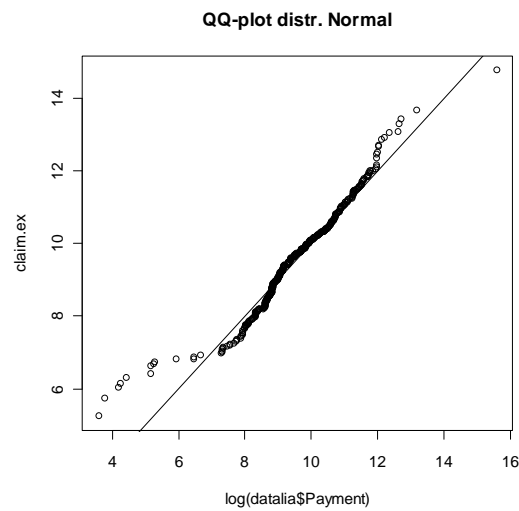


2. Use lognormal distribution fitting for claim size:

	meanlog	sdlog
Estimation	9.50314718	1.42545383
Standard Deviation	0.06890191	0.04872101

3. Use a QQ plot to check the fitting. The fit is not good at the low end.

Figure 31. QQ Plot of Log (Claim Size)



Fit the frequency

4. Draw a time series of frequency data and conduct a Seasonal Decomposition of Time Series. There are no strong evidence of linear trend and seasonality during this period.

Figure 32. Frequency

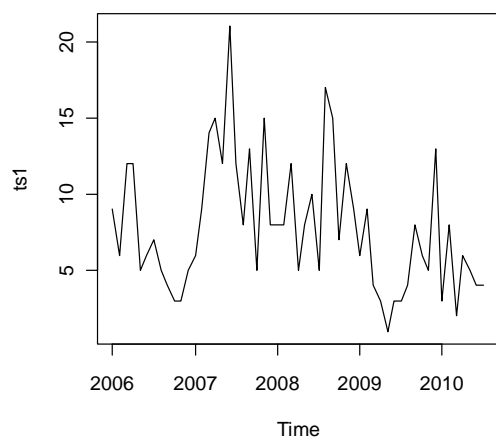
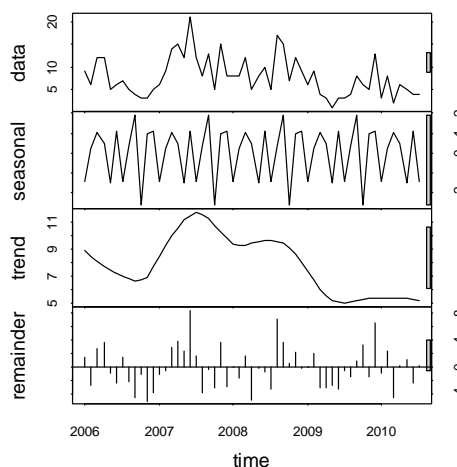


Figure 33. STL



5. Use linear regression for trend analysis.

$$\text{Log}(\text{Monthly Frequency}) = \text{Intercept} + \text{trend} * (\text{time} - 2006) + \text{error term.}$$

Residuals:

Min	1Q	Median	3Q	Max
-1.74504	-0.36590	0.09695	0.42571	1.03941

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.3330	0.2060	11.327	9.03e-16
trend	-0.1357	0.0587	-2.311	0.0247

Residual standard error: 0.5759 on 53 degrees of freedom.

Multiple R-squared: 0.09158, Adjusted R-squared: 0.07444.

F-statistic: 5.343 on 1 and 53 DF, p-value: 0.02472.

Figure 34. Trend fitting

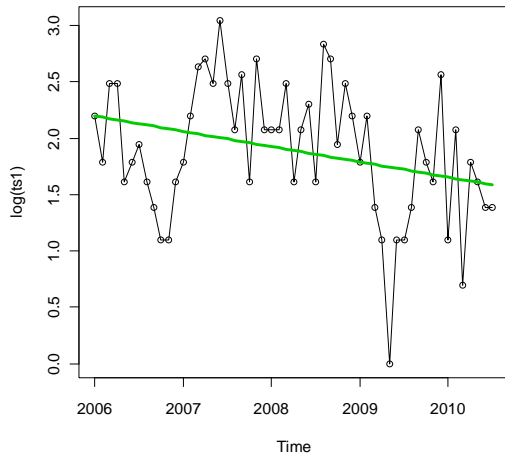
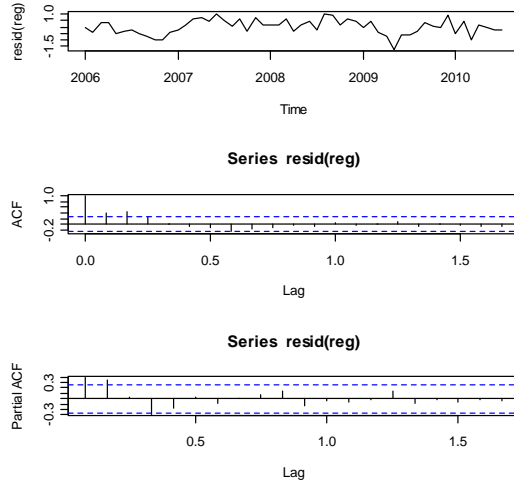


Figure 35. Residual, ACF, PACF



6. Detrend the frequency and fit it to the frequency distribution. It looks like lognormal distribution fits the detrended data better.

	meanlog	sdlog
Estimation	2.35724617	0.38449461
Standard Deviation	0.05184524	0.03666012

Figure 36. Histogram of detrend freq.

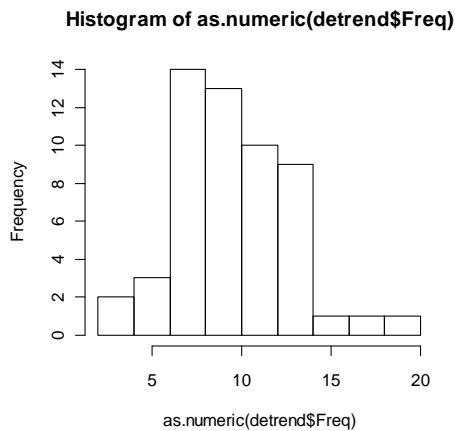
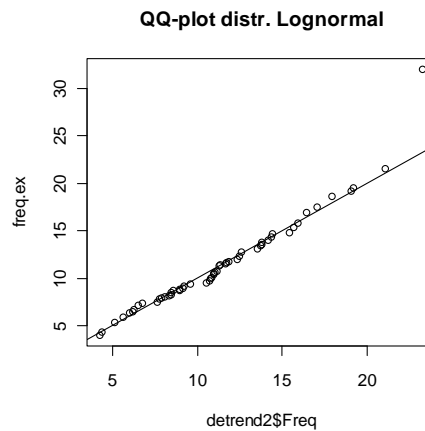


Figure 37. QQ Plot of detrend freq.



The Kolmogorov-Smirnov test also supports the assumption of lognormal distribution.

One-sample Kolmogorov-Smirnov test

$D = 0.0981, p\text{-value} = 0.6293,$

therefore, we have all the parameters for frequency and severity distribution and trend of frequency for liability line.

3.3 Correlation

First, we calculate the correlation coefficient between the two lines' frequencies.

	Line1	Line2
Line1	1.0000000	0.2800634
Line2	0.2800634	1.0000000

The Frank copula parameter is then estimated based on simulated frequency data using two methods. Other types of copula can and should also be used to determine the best fit.

(1) The estimation is based on the maximum likelihood and a sample of size 55.

	Estimate	Std. Error	z value	$\Pr(> z)$
rho.1	1.512390	0.854729	1.769438	0.07682074

The maximized loglikelihood is 1.533443.

(2) The estimation is based on the inversion of Kendall's tau and a sample of size 55.

	Estimate	Std. Error	z value	$\Pr(> z)$
parameter	1.325654	0.918666	1.443020	0.1490148

A simple way to compare is to draw the scatter plot for the intended copula and simulated frequency pairs. The figures below show that they are of the similar patterns.

Figure 38. Frank Copula (1.325654)

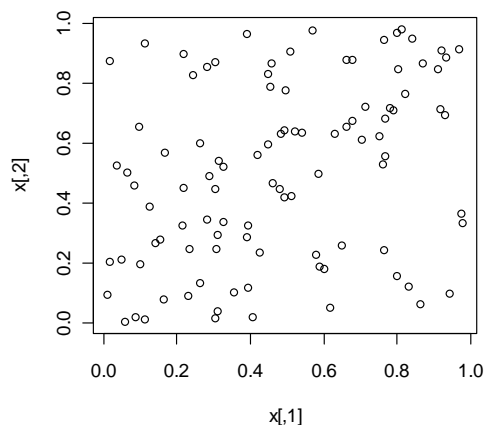
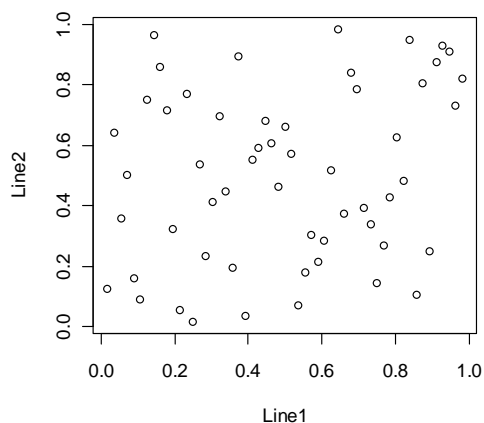


Figure 39. Simulated Frequencies



Goodness-of-fit Test

(1) Using Maximum Likelihood method for parameter estimation:

Parameter estimate(s): 1.512390

Cramer-von Mises statistic: 0.02652859 with p -value 0.3514851

(2) Using Inversion of Kendall's tau method for parameter estimation:

Parameter estimate(s): 1.325654

Cramer-von Mises statistic: 0.02780636 with p -value 0.4009901

Based on those testing results, we would not reject the null hypothesis that the real data are consistent with the Frank copula with parameter 1.325654.

4. MODEL ENHANCEMENT

4.1 Two-State, Regime-Switching Distribution

Sometimes in the real world, one single distribution may not be able to represent the past frequency and severity experience data well. There are normally three reasons behind this:

- (1) Structural change: some exogenous impact causes distribution (distribution type and/or parameters) to change drastically during a time period and last thereafter.
- (2) Cyclical pattern: The business may have some cyclical characteristics. A normal case is the

Loss Simulation Model Testing and Enhancement

underwriting cycle where for a certain period of time, the claim frequencies and/or severities will increase a lot and after that, it will return to a lower level.

- (3) Idiosyncratic risk: The claim data cannot be described by available distribution types. The randomness due to idiosyncratic characteristics makes it hard to fit a certain distribution along the time.

In the LSM, if the structural change is predicted, it can be incorporated by setting frequency/severity trend and even using different severity distributions for different months when the distribution type is expected to change.

However, the current model does not have a direct solution for incorporating the cyclical pattern and idiosyncratic characteristics. In order to add the flexibility of LSM to handle the modeling of them, a categorical variable is included to enable setting parameters/distribution type for different states. For all the variables that are modeled as distribution, two-state regime-switching capability is built in to enable moving from one state to the other state. A two-state, regime-switching model is commonly used in time series analysis. Here state means the status of the object such as frequency and/or severity that is described as a certain distribution.

The user can set two distributions with different parameters and determine the transition probability from one state to another. At the beginning of each month, the model will determine which distribution/state it will be for this month based on the transition matrix.

Let's take frequency distribution as an example to illustrate the process in the model.

Input

- ✓ State 1: Poisson Distribution ($\lambda = 120$)
- ✓ State 2: Negative Binomial Distribution (size = 36, prob = 0.5)
- ✓ Assume the trend, monthly exposure, and seasonality are all 1
- ✓ State 1 persistency: 0.5
- ✓ State 2 persistency: 0.7
- ✓ Seed: 16807

Markov Chain Transition Matrix

State persistency represents the probability that the variable will remain in the same state next

month. Here we assume the transition follows discrete Markov Chain.²¹ It means that the state of next month only depends on the state of the current month but does not depend on the state before the current month. In other words, it is not path-dependent.

Another thing that needs to be determined is the state of the first month. In the current model setting, steady-state probabilities are used. Let's define some variables first:

- ✓ P_{11} : state 1 persistency, the probability that the state will be 1 next month given that it is 1 this month.
- ✓ P_{12} : the probability that the state will be 2 next month given that it is 1 this month.
- ✓ P_{21} : the probability that the state will be 1 next month given that it is 2 this month.
- ✓ P_{22} : state 2 persistency, the probability that the state will be 2 next month given that it is 2 this month.
- ✓ Π_1 : steady probability of state 1.
- ✓ Π_2 : steady probability of state 2.

We have the following relationship held.

$$(\Pi_1 \quad \Pi_2) \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} = (\Pi_1 \quad \Pi_2)$$

$$P_{11} = 1 - P_{12}$$

$$P_{21} = 1 - P_{22}$$

$$\Pi_1 + \Pi_2 = 1$$

We can then derive the steady-state probabilities Π_1 and Π_2 based on state persistencies P_{11} and P_{22} .

$$\Pi_1 = \frac{1 - P_{22}}{2 - P_{11} - P_{22}} = \frac{1 - 0.7}{2 - 0.5 - 0.7} = 0.375$$

$$\Pi_2 = \frac{1 - P_{11}}{2 - P_{11} - P_{22}} = \frac{1 - 0.5}{2 - 0.5 - 0.7} = 0.625$$

Calculation Steps

- (1) Generate uniform random number randf_0 on range $[0,1]$.

²¹ http://en.wikipedia.org/wiki/Markov_chain

Loss Simulation Model Testing and Enhancement

- (2) If $\text{randf}_0 < \Pi_1$, state of first month state is 1, else, it is 2.
- (3) Generate uniform random number randf_i on range [0,1].
- (4) For previous month state I, if $\text{randf}_i < P_{i1}$, then state is 1, else it is 2.
- (5) Repeat step 3 and 4 until the end of the simulation is reached.

Table 1 shows the two-state, regime-switching result for the first simulation.

Table 1. Two-State, Regime-Switching Example

Random Number (RN)	State	Criteria
0.634633548790589	2	$\text{RN} > 0.375$
0.801362191326916	1	$\text{RN} > 0.7$
0.529508789768443	2	$\text{RN} > 0.5$
0.0441845036111772	2	$\text{RN} < 0.7$
0.994539848994464	1	$\text{RN} > 0.7$
0.21886122901924	1	$\text{RN} < 0.5$
0.0928565948270261	1	$\text{RN} < 0.5$
0.797880138037726	2	$\text{RN} > 0.5$
0.129500501556322	2	$\text{RN} < 0.7$
0.24027365935035	2	$\text{RN} < 0.7$
0.797712686471641	1	$\text{RN} > 0.7$
0.0569291599094868	1	$\text{RN} < 0.5$

Based on those generated frequency states, the claim and transaction are populated. This enhancement is intended for frequency and severity distribution although the flexibility is given to all the variables that are modeled as distribution in the LSM.

4.2 Testing

The following model setting is used for testing two-state, regime-switching feature.

Test Parameters:

- ✓ Accident Year: 2000
- ✓ Random Seed: 16807
- ✓ # of Simulations: 300
- ✓ Frequency correlation: Normal Copula with correlation as 95%

Line 1

Annual frequency:

Loss Simulation Model Testing and Enhancement

- ✓ State 1: Poisson ($\lambda = 120$), State 2: Negative Binomial (Size = 36, prob = 0.5)
- ✓ State 1 persistency: 0.15
- ✓ State 2 persistency: 0.9. It is equivalent to $\Pi_1 = 10.53\%$ and $\Pi_2 = 89.47\%$. We can consider state 2 as the long-term normal case while state 1 is the short period where the cases of claim increase a lot compared to state 1.
- ✓ Monthly exposure: 1
- ✓ Frequency Trend: 1
- ✓ Seasonality: 1

Size of entire loss

- ✓ State 1: Lognormal with $\mu = 10$ and $\sigma = 0.832549779$
- ✓ State 2: Lognormal with $\mu = 2$ and $\sigma = 0.832549779$
- ✓ State 1 persistency: 0.3
- ✓ State 2 persistency: 0.8. It is equivalent to $\Pi_1 = 22.22\%$ and $\Pi_2 = 77.78\%$.
- ✓ Severity Trend: 1
- ✓ $P(0) = 0$
- ✓ Est $P(0) = 0$

Line 2

Annual frequency:

- ✓ State 1: Poisson ($\lambda = 120$), State 2: Negative Binomial (Size = 36, prob = 0.5)
- ✓ State 1 persistency: 0.2
- ✓ State 2 persistency: 0.9. It is equivalent to $\Pi_1 = 11.11\%$ and $\Pi_2 = 88.89\%$
- ✓ Monthly exposure: 1
- ✓ Frequency Trend: 1
- ✓ Seasonality: 1

Size of entire loss:

- ✓ Lognormal with $\mu = 10$ and $\sigma = 0.832549779$

- ✓ Severity Trend: 1
- ✓ $P(0) = 0$
- ✓ Est $P(0) = 0$

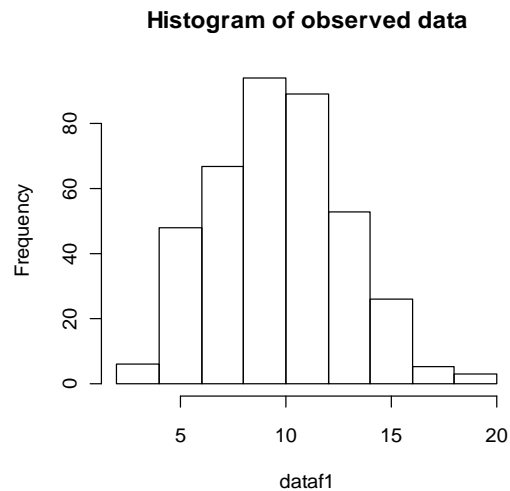
4.2.1 Frequency

We split the claim data according to the state of the monthly frequency and test whether the distribution for each state follows our model assumption.

State 1

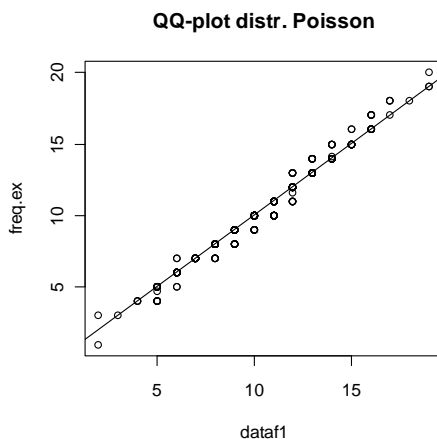
First, we draw a histogram of the simulated frequency data to give intuition of the distribution type.

Figure 40. Histogram of simulated frequency data (State 1)



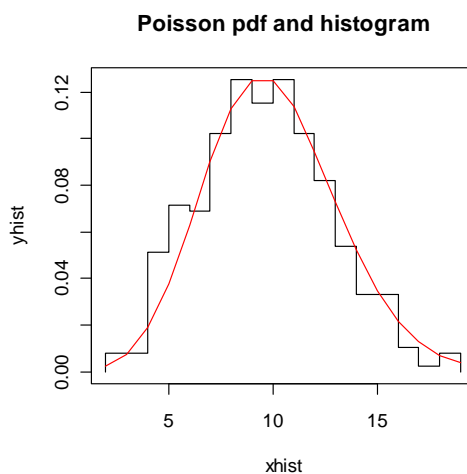
A QQ plot would also be a straightforward way to compare the simulated results with the intended distribution – Poisson ($\lambda = 10$). In Figure 41, we can see that it is a good fit.

Figure 41. QQ Plot – Simulated results vs. Poisson (lambda = 10)



Comparing the probability distribution functions also gives us a vivid illustration of the fit.

Figure 42. PDF – simulated vs. assumption



Goodness-of-fit test using Pearson’s Chi-squared statistic is performed. The results disallow rejecting the null hypothesis that the simulated frequencies follow a Poisson distribution.

Goodness-of-fit test for Poisson distribution

	X^2	df	$P(> X^2)$
Pearson	15.30052	19	0.703315

In addition, using maximum likelihood (ML) method to fit the Poisson distribution and calculate the likelihood Ratio statistics implies the same conclusion.

Loss Simulation Model Testing and Enhancement

Goodness-of-fit test for Poisson distribution

	X^2	df	$P(> X^2)$
Likelihood Ratio	20.27080	17	0.260613

Using ML method gives us an estimation of the parameters as follows:

	lambda
Estimation	10.1329923
Standard deviation	0.1609832

Comparing with our LSM input: $\lambda = 120$, which implies a monthly frequency as Poisson distribution with $\lambda = 10$. We can see that at significance level of 5%, the confidence interval for size is (9.82, 10.45), which includes the model input ($\lambda = 10$).

Two-sample Kolmogorov-Smirnov test

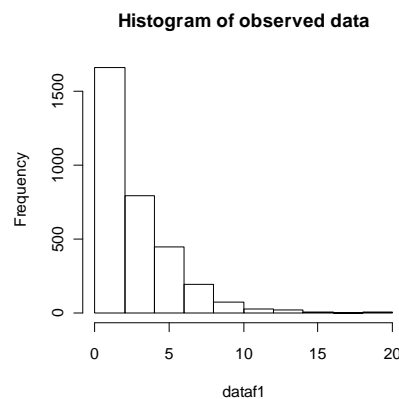
$$D = 0.0411, p\text{-value} = 0.7286$$

The Kolmogorov-Smirnov test also shows a reliable fit. Those results together with the goodness-of-fit tests indicate that simulated frequencies are Poisson distribution.

State 2

Firstly, we draw a histogram of the simulated frequency data to have an indication of the distribution type.

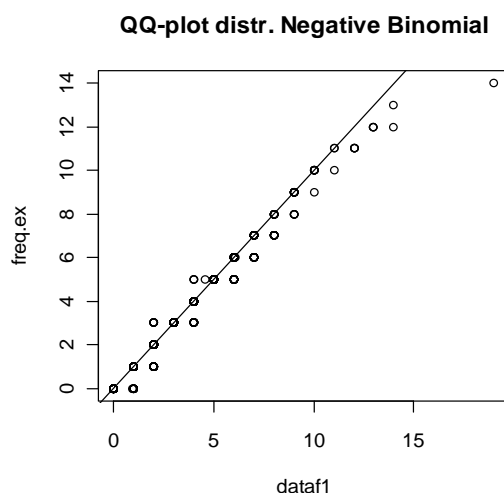
Figure 43. Histogram of simulated frequency data (State 2)



A QQ plot would also be a straightforward way to compare the simulated results with the intended distribution – Negative Binomial (size = 3, prob = 0.5). From the figure below, we can see

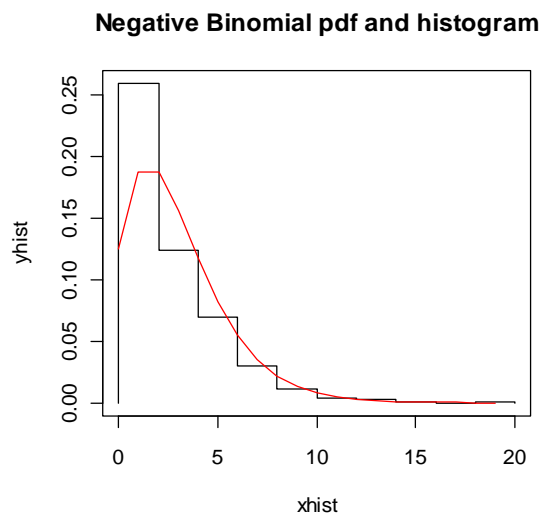
that it is a good fit. The expected frequency distribution in the LSM has a slightly shorter tail than the simulated results.

Figure 44. QQ Plot – Simulated results vs. Negative Binomial (size = 3, prob = 0.5)



Comparing the probability distribution function also shows fit below.

Figure 45. PDF – simulated vs. assumption



Goodness-of-fit test using Pearson's Chi-squared statistic is performed. The results allow us to reject the null hypothesis that the simulated frequencies follow negative binomial distribution.

Goodness-of-fit test for nbinomial distribution

Loss Simulation Model Testing and Enhancement

	X^2	df	$P(> X^2)$
Pearson	30.75979	19	0.042890443

In addition, using maximum likelihood (ML) method to fit the Poisson distribution and calculate the likelihood ratio statistics also implies the same conclusion.

Goodness-of-fit test for Poisson distribution

	X^2	df	$P(> X^2)$
Likelihood Ratio	32.36216	16	0.008968028

Using ML method gives us an estimation of the parameters as follows:

	size	mu
Estimation	2.78375646	3.00250312
Standard deviation	0.14274338	0.04418285

The estimated value gives us size = 2.78 and prob = 0.48. The derived variance is 6.24

Where prob = size/(size+mu) and variance = mu + mu²/size²²

Our LSM inputs of size = 36 and prob = 0.5 implies a monthly frequency as negative binomial distribution with size = 3, prob = 0.5. In comparison to estimated parameters based on simulated frequencies, they are not too far away.

Those results are somewhat consistent with the negative binomial frequency distribution testing results in section 2.1 as the p values are not very high but disallow us rejecting the null hypotheses at low significance level.

Transition Matrix

The implied steady-state probability of the transition matrix is tested against the simulation result. The results and calculation step are shown below. The simulation results show the similar steady-state probability.

²² Package *stats* version 2.12.0, R Documentation, The Negative Binomial Distribution.

Loss Simulation Model Testing and Enhancement

Line 1 Frequency

$$\begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} = \begin{pmatrix} 0.15 & 0.85 \\ 0.1 & 0.9 \end{pmatrix}$$

$$(\Pi_1 \quad \Pi_2) = (10.53\% \quad 89.47\%)$$

Line 2 Frequency

$$\begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} = \begin{pmatrix} 0.2 & 0.8 \\ 0.1 & 0.9 \end{pmatrix}$$

$$(\Pi_1 \quad \Pi_2) = (11.11\% \quad 88.89\%)$$

Non Zero Cases:

State 1: 391

State 1: 410

State 2: 2797

State 2: 2733

Probability of Zero Cases:

State 1: 0.005% (e^{-10})

State 1: 0.005% (e^{-10})

State 2: 0.125 (prob³)

State 2: 0.135 (e^{-2})

Estimated all Cases: Non Zero Cases/ (1 – Probability of Zero Cases)

State 1: 391

State 1: 410

State 2: 3188 (2797/(1-0.125))

State 2: 3161 (2733/(1-0.135))

Total Cases: # of simulations * 12 months = 3600

Steady-state probability (compared with Π_1 & Π_2)

State 1: 391/3600 = 10.86%

State 1: 410/3600 = 11.4%

State 2: 1-10.86% = 89.14%

State 2: 1-11.4% = 88.6%

4.2.2 Severity

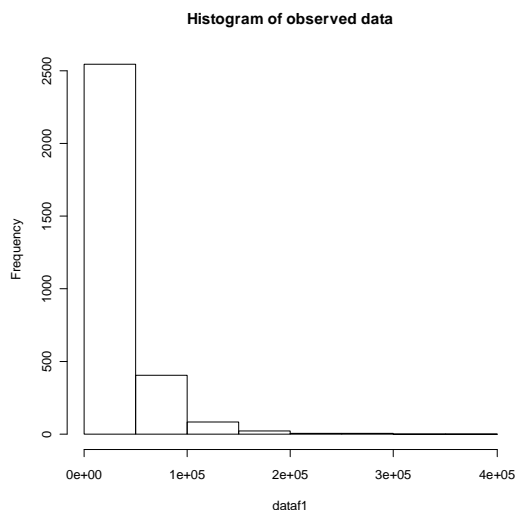
In testing Line 1 severity data, one thing worth noticing is that the size of loss assumption in the LSM is based on report date. Accident date might be a better choice to link size of loss with date of occurrence. For example, the size of loss might be more relevant to the time of catastrophic event like the 2011 Japanese earthquake instead of the time that the loss caused by the event is reported. From the modeling perspective, it also creates difficulties to realize the two-state, regime-switching function as the simulation is looped around each accident date instead of reporting date. In this testing, size of loss assumption is changed to be linked with accident date.

We split the claim data according to the state of the severity and test whether the distribution for each state follows our model assumption.

State 1

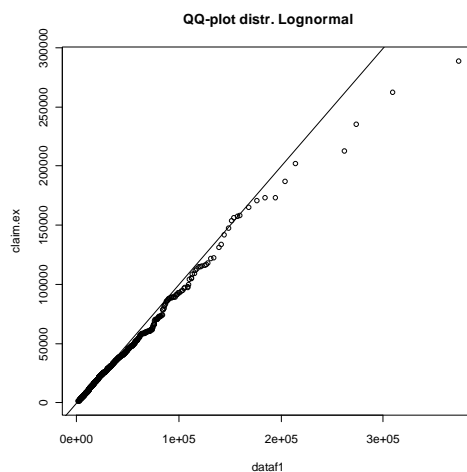
First, we draw a histogram of the simulated severity data to have an indication of the distribution type.

Figure 46. Histogram of simulated severity data (State 1)



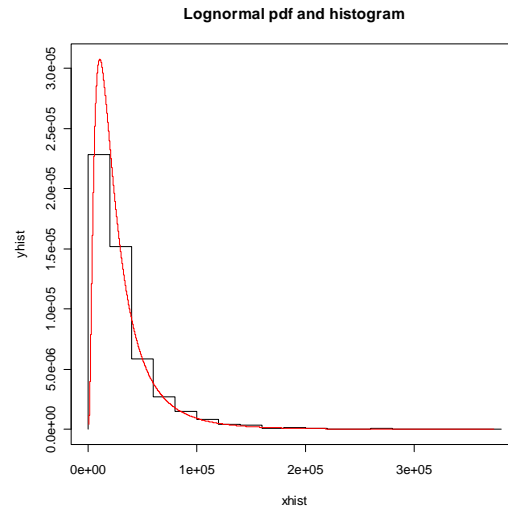
A QQ plot compares the simulated results with the intended distribution – Lognormal ($\mu = 10$ and $\sigma = 0.832549779$). From Figure 47, we can see that it is a good fit although the expected severity distribution as in the LSM has a slightly shorter tail than the simulated results.

Figure 47. QQ Plot – Simulated results vs. Lognormal ($\mu = 10$ and $\sigma = 0.832549779$)



Comparing the probability distribution functions also gives us a vivid illustration of the fit.

Figure 48. PDF – simulated vs. assumption



Using ML method gives us an estimation of the parameters as follows:

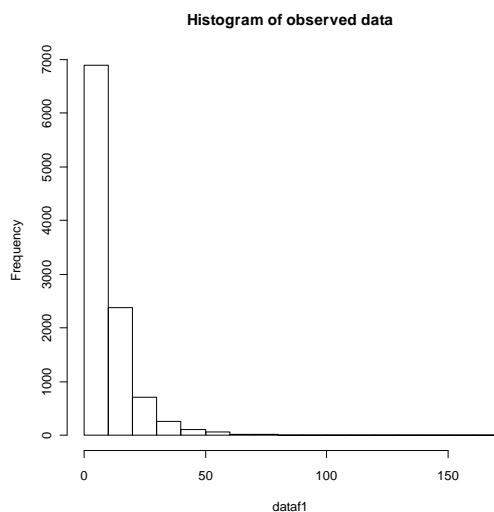
	meanlog	sdlog
Estimation	10.00677788	0.85323121
Standard deviation	0.01536917	0.01086764

Compare with our LSM input: $\mu = 10$ and $\sigma = 0.832549779$. We can see that at significance level of 5%, the confidence intervals for both parameters include the model input.

State 2

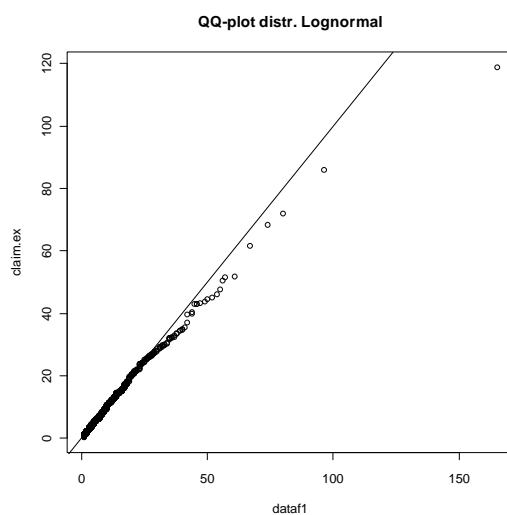
First, we draw a histogram of the simulated severity data to have an indication of the distribution type.

Figure 49. Histogram of simulated severity data (State .2)



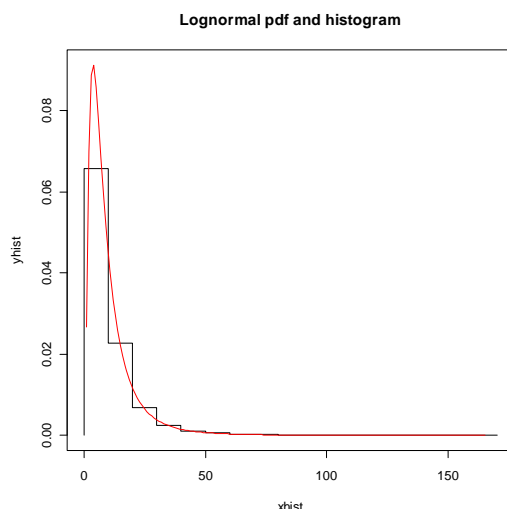
A QQ plot compares the simulated results with the intended distribution –Lognormal ($\mu = 2$, $\sigma = 0.832549779$). From Figure 50, we can see that it is a good fit. The expected severity distribution in the LSM also has a slightly shorter tail than the simulated results as in state 1.

Figure 50. QQ Plot – Simulated results vs. Lognormal ($\mu = 2$, $\sigma = 0.832549779$)



Comparing the probability distribution functions also shows the fit below.

Figure 51. PDF – simulated vs. assumption



Using the ML method gives us an estimation of the parameters as follows:

	meanlog	sdlog
Estimation	2.00714752	0.83957055
Standard deviation	0.00820275	0.00580022

In comparison to our LSM input of $\mu = 2$ and $\sigma = 0.832549779$, we can see at a significant level of 5% that the confidence intervals for both parameters include the model input.

4.2.3 Correlation

Correlation is tested to make sure that the correlation modeling using Copula is not affected by a two-state, regime-switching model. Correlation between frequencies of two lines is chosen for testing. We have four sets of data to test:

- Set 1: Line 1: State 1 and Line 2: State 1
- Set 2: Line 1: State 1 and Line 2: State 2
- Set 3: Line 1: State 2 and Line 2: State 1
- Set 4: Line 1: State 2 and Line 2: State 2

Scatter plots for the intended copula and simulated frequency pairs are shown below. Figures from 52 to 56 below show that they are of similar patterns.

Figure 52. Normal Copula (0.95)

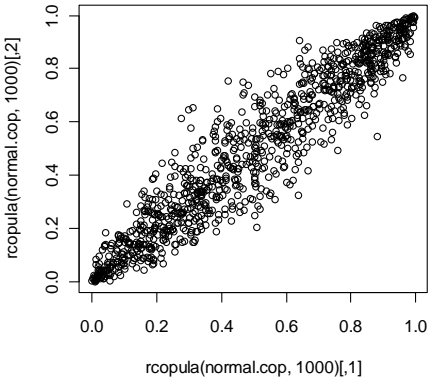


Figure 53. Set 1

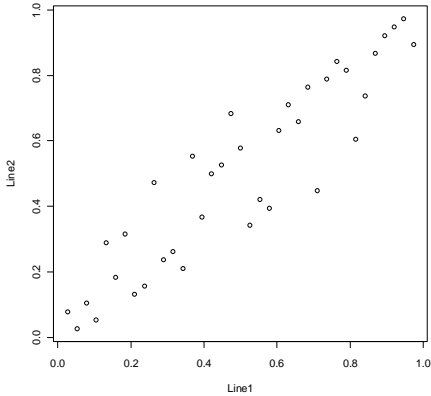


Figure 55. Set 3

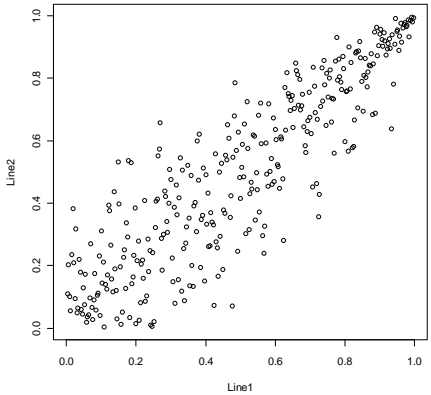


Figure 54. Set 2

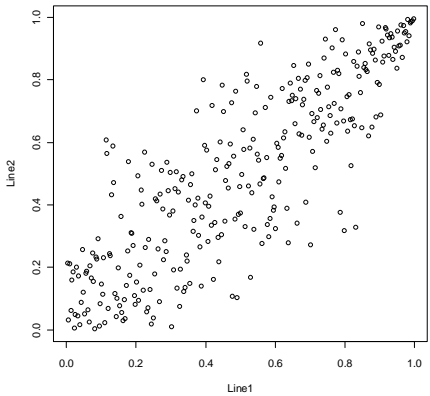
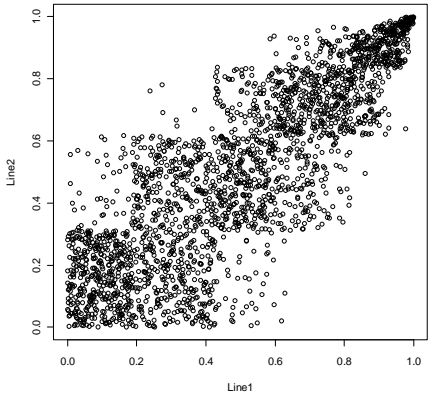


Figure 56. Set 4



For each set, we use maximum likelihood and inversion of Kendall's tau for parameter estimation

and goodness-of-fit test. Below are the results.

Set 1: State 1 for Line 1 and State 1 for Line 2

Normal copula parameter is estimated based on simulated frequency data using two methods.

(1) The estimation is based on the maximum likelihood and a sample of size 37.

	Estimate	Std. Error	χ value	$\Pr(> \chi)$
rho.1	0.9344341	0.01531399	61.01832	0

The maximized loglikelihood is 35.42264.

(2) The estimation is based on the inversion of Kendall's tau and a sample of size 1000.

	Estimate	Std. Error	χ value	$\Pr(> \chi)$
parameter	0.9380688	0.02458959	38.14903	0

We can see that the model parameter 0.95 is within the 95% confidence interval based on either of the two methods.

Goodness-of-fit Test

(1) Using Maximum Likelihood method for parameter estimation:

Parameter estimate(s): 0.9344341

Cramer-von Mises statistic: 0.01936648 with p -value 0.6980198

(2) Using Inversion of Kendall's tau method for parameter estimation:

Parameter estimate(s): 0.9380688

Cramer-von Mises statistic: 0.01821279 with p -value 0.7079208

Kolmogorov-Smirnov test is also done for testing the copula.

Two-sample Kolmogorov-Smirnov test

$D = 0.0423$, p -value = 0.9995

Based on those testing results, we can conclude that the simulated results show the same correlation as defined in model input.

Set 2: State 1 for Line 1 and State 2 for Line 2

Normal copula parameter is estimated based on simulated frequency data using two methods.

(1) The estimation is based on the maximum likelihood and a sample of size 307.

	Estimate	Std. Error	χ value	$\Pr(> \chi)$
rho.1	0.8400551	0.01290163	65.1123	0

The maximized loglikelihood is 183.7114.

(2) The estimation is based on the inversion of Kendall's tau and a sample of size 307.

	Estimate	Std. Error	χ value	$\Pr(> \chi)$
parameter	0.852917	0.01677851	50.83388	0

We can see that the model parameter 0.95 is out of the 95% confidence interval based on either of the two methods.

Goodness-of-fit Test

(1) Using Maximum Likelihood method for parameter estimation:

Parameter estimate(s): 0.8400551

Cramer-von Mises statistic: 0.03961167 with p -value 0.01485149

(2) Using Inversion of Kendall's tau method for parameter estimation

Parameter estimate(s): 0.852917

Cramer-von Mises statistic: 0.03370755 with p -value 0.01485149

Two-sample Kolmogorov-Smirnov test

$D = 0.0213, p\text{-value} = 0.9837$

The testing results show mixed information. One of the possible reasons for this is that we are not using all the simulated data for Set 2 but truncated data. If the number of claim is zero for a particular month, this data is not included in the claim output file from the LSM. Therefore, we are testing against non-zero monthly data only. As state 2 has a 12.5% and 13.5% probability of zero monthly claims for the two lines, respectively, we can see that except for Set 1, all other sets have the similar problem. It is still safe to conclude that high correlation exists as desired by model input.

Set 3: State 2 for Line 1 and State 1 for Line 2

Normal copula parameter is estimated based on simulated frequency data using two methods.

(1) The estimation is based on the maximum likelihood and a sample of size 329.

	Estimate	Std. Error	χ value	$\Pr(> \chi)$
rho.1	0.8644334	0.01056627	81.81065	0

The maximized loglikelihood is 222.0031.

(2) The estimation is based on the inversion of Kendall's tau and a sample of size 329.

	Estimate	Std. Error	χ value	$\Pr(> \chi)$
parameter	0.893593	0.01178312	75.8367	0

We can see that the model parameter 0.95 is out of the 95% confidence interval based on either of the two methods.

Goodness-of-fit Test

(1) Using Maximum Likelihood method for parameter estimation:

Parameter estimate(s): 0.8644334

Cramer-von Mises statistic: 0.07412085 with p -value 0.004950495

(2) Using Inversion of Kendall's tau method for parameter estimation:

Parameter estimate(s): 0.893593

Cramer-von Mises statistic: 0.04756158 with p -value 0.004950495

Two-sample Kolmogorov-Smirnov test

$D = 0.016$, p -value = 0.9996

Similar with Set 2, high correlation exists in the simulated data.

Set 4: State 2 for Line 1 and State 2 for Line 2

Normal copula parameter is estimated based on simulated frequency data using two methods.

(1) The estimation is based on the maximum likelihood and a sample of size 2376.

	Estimate	Std. Error	χ value	$\Pr(> \chi)$
--	----------	------------	--------------	-----------------

Loss Simulation Model Testing and Enhancement

rho.1 0.8114362 0.005444864 149.0278 0

The maximized loglikelihood is 1270.765.

(2) The estimation is based on the inversion of Kendall's tau and a sample of size 2376.

	Estimate	Std. Error	χ value	Pr(> χ)
parameter	0.845676	0.006024305	140.3773	0

We can see that the model parameter 0.95 is out of the 95% confidence interval based on either of the two methods.

Goodness-of-fit Test

(1) Using Maximum Likelihood method for parameter estimation:

Parameter estimate(s): 0.8114362

Cramer-von Mises statistic: 0.5949188 with p -value 0.004950495

(2) Using Inversion of Kendall's tau method for parameter estimation:

Parameter estimate(s): 0.845676

Cramer-von Mises statistic: 0.4380294 with p -value 0.004950495

Two-sample Kolmogorov-Smirnov test

$D = 0.0289$, p -value = 0.1900

Similar with Set 2, high correlation exists in the simulated data.

5. CONCLUSION AND FURTHER DEVELOPMENT

Based on the tests that have been conducted on the LSM, we cannot reject the assumption that model input and output are consistent regarding the following:

- (1) Negative binomial frequency distribution.
- (2) All copula types for frequencies among different lines except Gumbel Copula.
- (3) the correlation modeling between report lag and loss size based on Normal Copula.
- (4) Severity trend.
- (5) Alpha in severity trend.

Loss Simulation Model Testing and Enhancement

Though the statistical test results does not support Gumbel Copula applied to frequencies correlation very well, it is safe to not reject the null hypothesis as at a lower significance level such as 1%; it still passes the goodness-of-fit test.

A case reserve adequacy test shows that the assumption is not consistent with simulation data. This may be caused by the linear interpolation method used to derive 40% time point case reserve. It is suggested revising the way in which valuation date is determined in the LSM. In addition to the simulated valuation dates based on the waiting-period distribution assumption as in the LSM, some deterministic time points can be added as valuation dates. The deterministic valuation dates are interpolated between the report date and the payment date. In the LSM, 0%, 40%, 70%, and 90% time-points, case reserve, adequacy distribution can be input into the model. Therefore, 0%, 40%, 70% and 90% time points may be added as deterministic valuation dates.

Marine claim data are used to fit the distribution for frequency and severity. Trend, seasonality, and correlation analyses are also conducted to determine model parameters. These could be examples of how we use real data to determine appropriate LSM input which can be used for simulation and further testing of different reserve methods. If there are some data about paid loss history of the claims, the LSM can be better utilized to test different reserving methods. This could be an area for further research on the LSM.

Some enhancements have been made to the LSM. In the LSM, size of loss is linked to report date. The accident date might be a better choice for linking the size of loss with date of occurrence as the report lag would only have slight impact in loss size. From the modeling perspective, it also creates difficulties to realize the two-state, regime-switching function as the simulation is looped around each accident date instead of reporting date.

A categorical variable is included to enable setting parameters/distribution type for different states. Two-state, regime-switching flexibility is built in to enable moving from one state to the other state with a specified transition matrix. This, hopefully, can add the flexibility to mimic the underlying cycle we normally see in P&C business. Relevant testing is performed on the simulation data, which shows the consistency between model input and model output.

Acknowledgment

The author acknowledges Robert Bear, Dana F. Joseph, Joe Marker, Bryan Ware, and Kun Zhang for guidance, review, comments, and full support of this research. They helped identify a lot of errors in the earlier versions and made suggestions on the revision. The author is also grateful for the opportunity provided by Casualty Actuarial Society.

APPENDIX A. R²³ CODE

Statistical software R is used for loss simulation testing purpose. Based on the claim and transaction files output from the loss simulation model, R is used to process the data, conduct the statistical test for copula and distribution, and draw graphics for viewing goodness of fit. The R codes are listed below for each test. The input/output directory shall be revised if the codes are to be reused. Lines start with “`##`” is the description of the codes below it.

A.1 Negative Binomial Frequency Distribution Testing

```
## Read raw data (Claim output file)
rawdata<-read.csv("F:/Research/copula/copula test/Negative Binomial Frequency 100
0.4/co.csv",skip=1,header=TRUE)

## Manipulate claim output file to retrieve annual frequency data for each simulation/line
fcn<-function(dataset){
  x<-floor((dataset[4]-20000000)/100)
  return(x)}

## apply fcn which returns the month of accident date
dataindex<-apply(rawdata,1,fcn)
rawdata2<-cbind(rawdata,dataindex)
rawdata3<-aggregate(rawdata2, list(rawdata2$Simulation.No), length)
rawdata4<-rawdata3[,1:2]
dataf1<-rawdata4$Simulation.No
write.csv(datar,"F:/Research/copula/copula test/Negative Binomial Frequency 100 0.4/freq.csv")

##draw histogram
hist(dataf1,main="Histogram of observed data")

##QQPlot
freq.ex<-rbinom(n=1000,size=100,prob=0.4)
qqplot(dataf1,freq.ex,main="QQ-plot distr. Negative Binomial")
abline(0,1) ## a 45-degree reference line is plotted

##Histogram and PDF
h<-hist(dataf1,breaks=10)
xhist<-c(min(h$breaks),h$breaks)
yhist<-c(0,h$density,0)
xfit<-seq(min(dataf1),max(dataf1),by=1)
yfit<-dnbinom(xfit,size=100,prob=0.4)
plot(xhist,yhist,type="s",ylim=c(0,max(yhist,yfit)), main="Negative Binomial pdf and histogram")
lines(xfit,yfit, col="red")

##Goodness of fit test
library(vcd)
gf<-goodfit(dataf1,type="nbinom",par=list(size=100,prob=0.4))
```

²³ R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.

```
summary(gf)
plot(gf)
gf<-goodfit(dataf1,type= "nbinom",method= "ML")
fitdistr(dataf1, "Negative Binomial")
```

A.2 Correlation Test

Correlation among the frequencies of different lines

1. Clayton Copula

```
## Read raw data (Claim output file)
rawdata<-read.csv("F:/Research/copula/copula test/clayton 5/co.csv",skip=1,header=TRUE)
```

```
## Manipulate claim output file to retrieve annual frequency data for each simulation/line
```

```
fcn<-function(dataset){
  x<-floor((dataset[4]-20000000)/100)
  return(x)}
## apply fcn which returns the month of accident date
```

```
dataindex<-apply(rawdata,1,fcn)
rawdata2<-cbind(rawdata,dataindex)
```

```
## 1st month instead of one year occurrences
```

```
rawdata2m<-rawdata2[rawdata2$dataindex==1,]
rawdata3<-aggregate(rawdata2m, list(rawdata2m$Simulation.No,rawdata2m$Line), length)
rawdata4<-rawdata3[,1:3]
data1<-rawdata4[rawdata4$Group.2==1,]
data2<-rawdata4[rawdata4$Group.2==2,]
rawdata5<-merge(data1,data2,by="Group.1")
datar<-cbind(rawdata5$Simulation.No.x,rawdata5$Simulation.No.y)
colnames(datar)<-c("Line1","Line2")
write.csv(datar,"F:/Research/copula/copula test/clayton 5/x.csv")
```

```
## copula test
n<-length(datar[,1])
set.seed(123)
x<- sapply(as.data.frame(datar), rank, ties.method = "random") / (n + 1)
plot(x)
```

```
## Load R packages
library(MASS)
library(methods)
library(mvtnorm)
library(scatterplot3d)
library(mnormt)
library(sn)
library(pspline)
library(copula)
```

```
## Set up copula object for copula distribution and goodness-of-fit test later
clayton.cop <- claytonCopula(6, dim=2)
```

```
## Copula fit with prespecified type.
```

```
fit.clayton<-fitCopula(clayton.cop,x,method="ml")
fit.clayton
fit.clayton<-fitCopula(clayton.cop,x,method="itau")
fit.clayton
```

```
##Copula Goodness-of-fit test
gofCopula(clayton.cop, x, N=100, method = "mpl")
gofCopula(clayton.cop, x, N=100, method = "itau")
```

2. Frank Copula

```
## Read raw data (Claim output file)
rawdata<-read.csv("F:/Research/copula/copula test/frank 8/co.csv",skip=1,header=TRUE)
## Manipulate claim output file to retrieve annual frequency data for each simulation/line
fcn<-function(dataset){
  x<-floor((dataset[4]-20000000)/100)
  return(x)}
## apply fcn which returns the month of accident date
dataindex<-apply(rawdata,1,fcn)
rawdata2<-cbind(rawdata,dataindex)
##1st month instead of one year occurrences
rawdata2m<-rawdata2[rawdata2$dataindex==1,]
##rawdata3<-aggregate(rawdata2, list(rawdata2$Simulation.No,rawdata2$Line), length)
##1st month instead of one year occurrences
rawdata3<-aggregate(rawdata2m, list(rawdata2m$Simulation.No,rawdata2m$Line), length)
rawdata4<-rawdata3[,1:3]
data1<-rawdata4[rawdata4$Group.2==1,]
data2<-rawdata4[rawdata4$Group.2==2,]
rawdata5<-merge(data1,data2,by="Group.1")
datar<-cbind(rawdata5$Simulation.No.x,rawdata5$Simulation.No.y)
colnames(datar)<-c("Line1","Line2")
write.csv(datar,"F:/Research/copula/copula test/frank 8/x.csv")

##copula test
n<-length(datar[,1])
set.seed(123)
x<- sapply(as.data.frame(datar), rank, ties.method = "random") / (n + 1)
plot(x)

##Load R packages
library(MASS)
library(methods)
library(mvtnorm)
library(scatterplot3d)
library(mnormt)
library(sn)
library(pspline)
library(copula)

##Set up copula object for copula distribution and goodness-of-fit test later
frank.cop <- frankCopula(8, dim=2)
```

```
##Copula fit with prespecified type.  
fit.frank<-fitCopula(frank.cop,x,method="ml")  
fit.frank  
fit.frank<-fitCopula(frank.cop,x,method="itau")  
fit.frank
```

```
##Copula Goodness-of-fit test  
gofCopula(frank.cop, x, N=100, method = "mpl")  
gofCopula(frank.cop, x, N=100, method = "itau")
```

3. Gumbel Copula

```
## Read raw data (Claim output file)  
rawdata<-read.csv("F:/Research/copula/copula test/Gumbel 6/co.csv",skip=1,header=TRUE)  
## Manipulate claim output file to retrieve annual frequency data for each simulation/line  
fcn<-function(dataset){  
  x<-floor((dataset[4]-20000000)/100)  
  return(x)}  
## apply fcn which returns the month of accident date  
dataindex<-apply(rawdata,1,fcn)  
##rawdata2<-cbind(rawdata,dataindex)  
##1st month instead of one year occurrences  
rawdata2m<-rawdata2[rawdata2$dataindex==1,]  
##rawdata3<-aggregate(rawdata2, list(rawdata2$Simulation.No,rawdata2$Line), length)  
##1st month instead of one year occurrences  
rawdata3<-aggregate(rawdata2m, list(rawdata2m$Simulation.No,rawdata2m$Line), length)rawdata4<-  
rawdata3[,1:3]  
data1<-rawdata4[rawdata4$Group.2==1,]  
data2<-rawdata4[rawdata4$Group.2==2,]  
rawdata5<-merge(data1,data2,by="Group.1")  
datar<-cbind(rawdata5$Simulation.No.x,rawdata5$Simulation.No.y)  
colnames(datar)<-c("Line1","Line2")  
write.csv(datar,"F:/Research/copula/copula test/Gumbel 6/x.csv")  
  
##copula test  
n<-length(datar[,1])  
set.seed(123)  
x<- sapply(as.data.frame(datar), rank, ties.method = "random") / (n + 1)  
plot(x)  
  
##Load R packages  
library(MASS)  
library(methods)  
library(mvtnorm)  
library(scatterplot3d)  
library(mnormt)  
library(sn)  
library(pspline)  
library(copula)  
  
##Set up copula object for copula distribution and goodness-of-fit test later  
gumbel.cop <- gumbelCopula(3, dim=2)
```

```
##Copula fit with prespecified type.  
fit.gumbel<-fitCopula(gumbel.cop,x,method="ml")  
fit.gumbel  
fit.gumbel<-fitCopula(gumbel.cop,x,method="itau")  
fit.gumbel
```

```
##Copula Goodness-of-fit test  
gofCopula(gumbel.cop, x, N=100, method = "mpl")  
gofCopula(gumbel.cop, x, N=100, method = "itau")
```

4. T Copula

```
## Read raw data (Claim output file)  
rawdata<-read.csv("F:/Research/copula/copula test/t50.8/co.csv",skip=1,header=TRUE)  
## Manipulate claim output file to retrieve annual frequency data for each simulation/line  
fcn<-function(dataset){  
  x<-floor((dataset[4]-20000000)/100)  
  return(x)}  
## apply fcn which returns the month of accident date  
dataindex<-apply(rawdata,1,fcn)  
rawdata2<-cbind(rawdata,dataindex)  
##1st month instead of one year occurrences  
rawdata2m<-rawdata2[rawdata2$dataindex==1,]  
##1st month instead of one year occurrences  
rawdata3<-aggregate(rawdata2m, list(rawdata2m$Simulation.No,rawdata2m$Line), length)  
rawdata4<-rawdata3[,1:3]  
data1<-rawdata4[rawdata4$Group.2==1,]  
data2<-rawdata4[rawdata4$Group.2==2,]  
rawdata5<-merge(data1,data2,by="Group.1")  
datar<-cbind(rawdata5$Simulation.No.x,rawdata5$Simulation.No.y)  
colnames(datar)<-c("Line1","Line2")  
write.csv(datar,"F:/Research/copula/copula test/t50.8/x.csv")
```

```
##copula test  
n<-length(datar[,1])  
set.seed(123)  
x<- sapply(as.data.frame(datar), rank, ties.method = "random") / (n + 1)  
plot(x)
```

```
##Load R packages  
library(MASS)  
library(methods)  
library(mvtnorm)  
library(scatterplot3d)  
library(mnormt)  
library(sn)  
library(pspline)  
library(copula)
```

```
##Set up copula object for copula distribution and goodness-of-fit test later  
t.cop <- tCopula(c(0.8), dim=2, dispstr="un", df=5, df.fixed=TRUE)
```

Loss Simulation Model Testing and Enhancement

```
##Copula fit with prespecified type.
fit.t<-fitCopula(t.cop,x,method="ml")
fit.t
fit.t<-fitCopula(t.cop,x,method="itau")
fit.t
```

```
##Copula Goodness-of-fit test
gofCopula(t.cop, x, N=100, method = "mpl")
gofCopula(t.cop, x, N=100, method = "itau")
```

Correlation between claim size and report lag

```
## Read raw data (Claim and transaction output file)
rawdatap<-read.csv("F:/Research/copula/copula test/copula2/co.csv",skip=1,header=TRUE)
rawdataa<-read.csv("F:/Research/copula/copula test/copula2/to.csv",skip=1,header=TRUE)
```

```
## Manipulate transaction output file to retrieve final payment amount
rawdataa2<-rawdataa[rawdataa$Transaction=="CLS",]
data1<-rawdatap[,c(1,2,3,5)]
data2<-rawdataa2[,c(1,2,3,4,7)]
datan<-merge(data1,data2,by=c("Simulation.No","Occurrence.No","Claim.No"))
```

```
## Translate payment date in terms of years
fcn<-function(dataset){
x<-floor(dataset[5]/10000)-floor(dataset[4]/10000)
y<-floor(dataset[5]/100)-floor(dataset[5]/10000)*100-(floor(dataset[4]/100)-floor(dataset[4]/10000)*100)
z<-dataset[5]-floor(dataset[5]/100)*100-(dataset[4]-floor(dataset[4]/100)*100)
r<-x+y/12+z/365
return(r)}
paymentlag<-apply(datan,1,fcn)
rawdatap2<-cbind(datan,paymentlag)
datar<-cbind(rawdatap2$paymentlag,rawdatap2$Payment)
write.csv(datar,"F:/Research/copula/copula test/copula2/100/x.csv")
```

```
##copula test
n<-length(datar[,1])
set.seed(123)
x<- sapply(as.data.frame(datar), rank, ties.method = "random") / (n + 1)
plot(x)
```

```
##Load R packages
library(MASS)
library(methods)
library(mvtnorm)
library(scatterplot3d)
library(mnormt)
library(sn)
library(pspline)
library(copula)
```

```
##Set up copula object for copula distribution and goodness-of-fit test later
normal.cop <- normalCopula(c(0),dim=2,dispstr="un")
```



```
##Copula fit with pre specified type.
fit.normal<-fitCopula(normal.cop,x,method="ml")
fit.normal
fit.normal<-fitCopula(normal.cop,x,method="itau")
fit.normal
```

```
##Copula Goodness-of-fit test
gofCopula(normal.cop, x, N=100, method = "mpl")
gofCopula(normal.cop, x, N=100, method = "itau")
```

A.3 Severity Trend

```
## Read raw data (Claim and transaction output file)
rawdatap<-read.csv("F:/Research/copula/copula test/strend/co.csv",skip=1,header=TRUE)
rawdataa<-read.csv("F:/Research/copula/copula test/strend/to.csv",skip=1,header=TRUE)
```

```
## Manipulate transaction output file to retrieve final payment amount
rawdataa2<-rawdataa[rawdataa$Transaction=="CLS",]
fcn<-function(dataset){
  x<-floor((dataset[4]-20000000)/100)
  return(x)}

```

```
## apply fcn which returns the month of accident date
dataindex<-apply(rawdatap,1,fcn)
rawdatap2<-cbind(rawdatap,dataindex)
data1<-rawdatap2[,c(1,2,8)]
data2<-rawdataa2[,c(1,2,7)]
datan<-merge(data1,data2,by=c("Simulation.No","Occurrence.No"))
rawdata3<-aggregate(datan, list(datan$dataindex), mean)
##rawdata4<-rawdata3[,c(3,5,6)]
rawdata4<-rawdata3[,c(4,5)]
colnames(rawdata4)<-c("Month","MeanPayment")
write.csv(rawdata4,"F:/Research/copula/copula test/strend/x.csv")
datar<-rawdata4$MeanPayment
```

```
##set up time series
ts1<-ts(datar,start=2000,frequency=12)
plot(ts1)
plot(stl(ts1,s.window="periodic"))
```

```
##linear trend fitting
trend = time(ts1)-2000
reg = lm(log(ts1)~trend, na.action=NULL)
summary(reg)
plot(log(ts1), type="o")
lines(fitted(reg), col=2)
par(mfrow=c(3,1))
plot(resid(reg))
acf(resid(reg),20)
pacf(resid(reg),20)
```

A.4 Alpha in Severity Trend

```
## Read raw data (Claim and transaction output file)
rawdatap<-read.csv("F:/Research/copula/copula test/Alpha/co.csv",skip=1,header=TRUE)
rawdataa<-read.csv("F:/Research/copula/copula test/Alpha/to.csv",skip=1,header=TRUE)

## Manipulate transaction output file to retrieve final payment amount
rawdataa2<-rawdataa[rawdataa$Transaction=="CLS",]
fcn<-function(dataset){
  x<-floor((dataset[4]-20000000)/100)
  return(x)}
## apply fcn which returns the month of accident date
dataindex<-apply(rawdatap,1,fcn)
rawdatap2<-cbind(rawdatap,dataindex)
dataindex2<-apply(rawdataa2[,c(1:4)],1,fcn)
rawdataa3<-cbind(rawdataa2,dataindex2)
data1<-rawdatap2[,c(1,2,8)]
data2<-rawdataa3[,c(1,2,7,8)]
datan<-merge(data1,data2,by=c("Simulation.No","Occurrence.No"))
datam<-datan[datan$dataindex==1,]
b<-datam[datam$dataindex2==7,]
c<-b[b$Payment!=0,]
a<-c$Payment
length(a)

##draw histogram
hist(a,main="Histogram of observed data")
library(MASS)
fitdistr(a, "Lognormal")

##QQPlot
Seve.ex<-(rlnorm(n=1000,meanlog=-0.8726,sdlog=0.9567))
qqplot(a,Seve.ex,main="QQ-plot distr. Lognormal")
abline(0,1) ### a 45-degree reference line is plotted

##Histogram and PDF
h<-hist(a,breaks=10)
xhist<-c(min(h$breaks),h$breaks)
yhist<-c(0,h$density,0)
xfit<-seq(min(a),max(a),by=1)
yfit<-dlnorm(xfit,meanlog=-0.8726,sdlog=0.9567)
plot(xhist,yhist,type="s",ylim=c(0,max(yhist,yfit)), main="Lognormal pdf and histogram")
lines(xfit,yfit, col="red")

##Kolmogorov-Smirnov Tests
ks.test(a,"plnorm", meanlog=-0.8726,sdlog=0.9567)

##Anderson-Darling Test
datas1.norm<-log(a)
library(nortest) ### package loading
ad.test(datas1.norm)
```

A.5 Case Reserve Adequacy

```
## Read raw data (Claim output file)
rawdatap<-read.csv("D:/LS/RS/case reserve/025/to.csv",skip=1,header=TRUE)

## Manipulate transaction output file to retrieve final payment amount
rawdataa<-rawdatap[rawdatap$Simulation.No<101,]

## Calculate the number of days that have passed since Jan 1,2000 until the accident date
x<-(floor(rawdataa[4]/10000)-2000)*365+(floor(rawdataa[4]/100)-floor(rawdataa[4]/10000)*100)*30+rawdataa[4]-
floor(rawdataa[4]/100)*100
rawdatap2<-cbind(rawdataa,x)

## Linear Interpolation of generated case reserves to get 40% time point case reserve
fcn<-function(dataset){
  aa<-dataset$Date
  b<-dataset$Case.Reserve
  bb<-dataset$Case.Reserve
  cc<-dataset$Payment
  count<-length(dataset$Date)
  temp<-0
  for(k in 1:(count-1)){
    bb[k]<-b[k]+temp
    temp<-temp+b[k]
  }
  bb[count]<-cc[count]
  f<-approxfun(aa,bb)
  xmin<-min(dataset[5])
  xmax<-max(dataset[5])
  x<-0.6*xmin+0.4*xmax
  if(cc[count]==0){
    return(0)
  }else{
    return(f(x)/cc[count]/0.6)}
}
rawdata0<-rawdatap2[,c(1,2,6,7,8)]
m<-max(rawdata0$Simulation.No)
a<-matrix(rep(0,m*134),nrow=134,ncol=m)
## Get 40% case reserve for all claims
for(i in 1:m) {
  rawdata00<-rawdata0[rawdata0$Simulation.No==i,]
  rawdata<-as.data.frame(apply(rawdata00,2,abs))
  n<-max(rawdata$Occurrence.No)
  for(j in 1:n) {
    dataset<-as.data.frame(rawdata[rawdata$Occurrence.No==j,])
    a[j,i]=fcn(dataset)
  }
}
a<-as.vector(a)
a<-a[a!=0]

##draw histogram
hist(a,main="Histogram of observed data")
library(MASS)
```

```
fitdistr(a, "Lognormal")
```

```
##QQPlot
```

```
Seve.ex<-(rlnorm(n=1000,meanlog=0.25,sdlog=0.05))  
qqplot(a,Seve.ex,main="QQ-plot distr. Lognormal")  
abline(0,1) ### a 45-degree reference line is plotted
```

```
##Histogram and PDF
```

```
h<-hist(a,breaks=30)  
xhist<-c(min(h$breaks),h$breaks)  
yhist<-c(0,h$density,0)  
xfit<-seq(min(a),max(a),by=1)  
yfit<-dlnorm(xfit,meanlog=0.25,sdlog=0.05)  
plot(xhist,yhist,type="s",ylim=c(0,max(yhist,yfit)), main="Lognormal pdf and histogram")  
lines(xfit,yfit, col="red")
```

```
##Kolmogorov-Smirnov Tests
```

```
ks.test(a,"plnorm", meanlog=0.25,sdlog=0.05)
```

```
##Anderson-Darling Test
```

```
datas1.norm<-log(a)  
library(nortest) ### package loading  
ad.test(datas1.norm)
```

A.6 Real Claim Data Fitting

```
## Read raw data
```

```
rawdata<-read.csv("D:/LS/RS/PL/pl.csv",header=TRUE)  
rawdata1<-rawdata[rawdata$Payment>0,]  
dataProperty0<-rawdata1[rawdata1$Line=="Property",]  
dataProperty<-dataProperty0[,-3]  
datalia0<-rawdata1[rawdata1$Line=="Liability",]  
datalia<-datalia0[,-3]
```

```
##Property
```

```
##draw histogram of claim
```

```
hist(log(dataProperty$Payment),breaks=100,main="Histogram of observed data")  
library(MASS)  
fitdistr(log(dataProperty$Payment), "normal")
```

```
##QQPlot of claim
```

```
claim.ex<-(rlnorm(n=1000,mean=9.285,sd=2.267))  
qqplot(log(dataProperty$Payment),claim.ex,main="QQ-plot distr. Normal")  
abline(0,1) ### a 45-degree reference line is plotted
```

```
rawdata3<-aggregate(dataProperty, list(dataProperty$dataindex), length)  
rawdata4<-rawdata3[,1:2]  
colnames(rawdata4)<-c("tMonth", "Freq")  
summary(rawdata4)
```

```
##set up time series for frequency
```

```
ts1<-ts(rawdata4$Freq,start=2006,frequency=12)
```

Loss Simulation Model Testing and Enhancement

```
plot(ts1)
plot(stl(ts1,s.window="periodic"))

##trend analysis
trend = time(ts1)-2006
reg = lm(log(ts1)~trend, na.action=NULL)
summary(reg)
plot(log(ts1), type="o")
lines(fitted(reg), col=3, lwd=3)

par(mfrow=c(1,1))
plot(resid(reg))
acf(resid(reg),20)
pacf(resid(reg),20)

trendreg<--0.136*rawdata4[1]
detrend<-rawdata4[2]-trendreg

hist(as.numeric(detrend$Freq))
fitdistr(detrend$Freq,"normal")

##QQPlot of detrended frequency
freq.ex<-(rnorm(n=1000,mean=9.554,sd=3.131))
qqplot(detrend$Freq,freq.ex,main="QQ-plot distr. normal")
abline(0,1) ### a 45-degree reference line is plotted

ks.test(detrend$Freq,"pnorm", mean=9.554,sd=3.131)

##Histogram and PDF
h<-hist(detrend$Freq,breaks=15)
xhist<-c(min(h$breaks),h$breaks)
yhist<-c(0,h$density,0)
xfit<-seq(min(detrend$Freq),max(detrend$Freq),length=40)
yfit<-dnorm(xfit,mean=9.554,sd=3.131)
plot(xhist,yhist,type="s",ylim=c(0,max(yhist,yfit)), main="Normal pdf and histogram")
lines(xfit,yfit, col="red")

##Liability
##draw histogram of claim
hist(log(dataalia$Payment),breaks=100,main="Histogram of observed data")

fitdistr(log(dataalia$Payment), "normal")

##QQPlot of claim
claim.ex<-(rlnorm(n=1000,mean=9.5,sd=1.425))
qqplot(log(dataalia$Payment),claim.ex,main="QQ-plot distr. Lognormal")
abline(0,1) ### a 45-degree reference line is plotted

rawdata3<-aggregate(dataalia, list(dataalia$dataindex), length)
rawdata4<-rawdata3[,1:2]
colnames(rawdata4)<-c("tMonth","Freq")
summary(rawdata4)
```

Loss Simulation Model Testing and Enhancement

```
##set up time series
ts1<-ts(rawdata4$Freq,start=2006,frequency=12)
plot(ts1)
plot(stl(ts1,s.window="periodic"))

##trend analysis
trend = time(ts1)-2005
reg = lm(log(ts1)~trend, na.action=NULL)
summary(reg)
plot(log(ts1), type="o")
lines(fitted(reg), col=3,lwd=3)

par(mfrow=c(1,1))
plot(resid(reg))
acf(resid(reg),20)
pacf(resid(reg),20)

trendreg<-0.127*rawdata4[1]
detrend2<-rawdata4[2]-trendreg

##histogram of detrended data
hist(as.numeric(detrend2$Freq))
fitdistr(detrend2$Freq,"lognormal")
fitdistr(detrend2$Freq,"normal")

##QQPlot of detrended frequency
freq.ex<-(rlnorm(n=100,meanlog=2.357,sdlog=0.3845))
qqplot(detrend2$Freq,freq.ex,main="QQ-plot distr. Lognormal")
abline(0,1) ## a 45-degree reference line is plotted

ks.test(detrend2$Freq,"plnorm", meanlog=2.357,sdlog=0.3845)

##Histogram and PDF
h<-hist(detrend2$Freq,breaks=15)
xhist<-c(min(h$breaks),h$breaks)
yhist<-c(0,h$density,0)
xfit<-seq(min(detrend2$Freq),max(detrend2$Freq),length=40)
yfit<-dlnorm(xfit,meanlog=2.357,sdlog=0.3845)
plot(xhist,yhist,type="s",ylim=c(0,max(yhist,yfit)), main="Normal pdf and histogram")
lines(xfit,yfit, col="red")

datar<-cbind(detrend$Freq,detrend2$Freq)
colnames(datar)<-c("Line1","Line2")

##copula test
n<-length(datar[,1])
set.seed(123)
x<- sapply(as.data.frame(datar), rank, ties.method = "random") / (n + 1)
plot(x)

cor(datar)
```

```
##Load R packages
library(MASS)
library(methods)
library(mvtnorm)
library(scatterplot3d)
library(mnormt)
library(sn)
library(pspline)
library(copula)

##Set up copula object for copula distribution and goodness-of-fit test later. Only Frank copula
##is shown here while in real testing different types of copula should all be tested against the data

frank.cop <- frankCopula(6, dim=2)

##Copula fit with pre specified type.
fit.frank<-fitCopula(frank.cop,x,method="ml")
fit.frank
fit.frank<-fitCopula(frank.cop,x,method="itau")
fit.frank

##Copula Goodness-of-fit test
gofCopula(frank.cop, x, N=100, method = "mpl")
gofCopula(frank.cop, x, N=100, method = "itau")
```

A.7 Two-State, Regime-Switching Feature Testing

Frequency

```
## Read raw data (Claim output file)
rawdata<-read.csv("D:/LS/RS/tsw/cc.csv",skip=1,header=TRUE)
## Manipulate claim output file to retrieve annual frequency data for each simulation/line
fcn<-function(dataset){
  x<-floor((dataset[4]-20000000)/100)
  return(x)}
## apply fcn which returns the month of accident date
dataindex<-apply(rawdata,1,fcn)
rawdata1<-cbind(rawdata,dataindex)
rawdata2<-rawdata1[rawdata1$Line==1,]

### State 1 Frequency Testing
rawdatas1<-rawdata2[rawdata2$State==1,]
rawdata3<-aggregate(rawdatas1, list(rawdatas1$Simulation.No,rawdatas1$dataindex), length)
dim(rawdata3)
rawdata4<-rawdata3[,1:3]
dataf1<-rawdata4$Simulation.No

##draw histogram
hist(dataf1,main="Histogram of observed data")

##QQPlot
```

Loss Simulation Model Testing and Enhancement

```
freq.ex<-(rpois(n=1000,lambda=10))
qqplot(dataf1,freq.ex,main="QQ-plot distr. Poisson")
abline(0,1) ### a 45-degree reference line is plotted

##Histogram and PDF
h<-hist(dataf1,breaks=20)
xhist<-c(min(h$breaks),h$breaks)
yhist<-c(0,h$density,0)
xfit<-seq(min(dataf1),max(dataf1),by=1)
yfit<-dpois(xfit,lambda=10)
plot(xhist,yhist,type="s",ylim=c(0,max(yhist,yfit)), main="Poisson pdf and histogram")
lines(xfit,yfit, col="red")

##Goodness of fit test
library(vcd)
gf<-goodfit(dataf1,type="pois",par=list(lambda=10),method="MinChisq")
summary(gf)
plot(gf)
fitdistr(dataf1,"Poisson")
##Kolmogorov-Smirnov Tests
ks.test(dataf1,freq.ex,exact=NULL)

### State 2 Frequency Testing
rawdatas2<-rawdata2[rawdata2$State==2,]
rawdata3<-aggregate(rawdatas2, list(rawdatas2$Simulation.No,rawdatas2$dataindex), length)
dim(rawdata3)
rawdata4<-rawdata3[,1:3]
datafs1<-rawdata4$Simulation.No
dataf1<-c(rep(0,400),datafs1)

##draw histogram
hist(dataf1,main="Histogram of observed data")

##QQPlot
freq.ex<-(rnbinom(n=1000,size=3,prob=0.5))
qqplot(dataf1,freq.ex,main="QQ-plot distr. Negative Binomial")
abline(0,1) ### a 45-degree reference line is plotted

##Histogram and PDF
h<-hist(dataf1,breaks=10)
xhist<-c(min(h$breaks),h$breaks)
yhist<-c(0,h$density,0)
xfit<-seq(min(dataf1),max(dataf1),by=1)
yfit<-dnbinom(xfit,size=3, prob=0.5)
plot(xhist,yhist,type="s",ylim=c(0,max(yhist,yfit)), main="Negative Binomial pdf and histogram")
lines(xfit,yfit, col="red")

##Goodness of fit test
library(vcd)
gf<-goodfit(dataf1,type="nbinom",par=list(size=3,prob=0.5),method="MinChisq")
summary(gf)
plot(gf)
```



```
fitdistr(dataf1, "Negative Binomial")
```

Severity

```
## Read raw data (Claim output file)
rawdatap<-read.csv("D:/LS/RS/tsw/cc.csv",skip=1,header=TRUE)
rawdataa<-read.csv("D:/LS/RS/tsw/tt.csv",skip=1,header=TRUE)
## Manipulate transaction output file to retrieve final payment amount
rawdataa2<-rawdataa[rawdataa$Transaction=="CLS",]
fcn<-function(dataset){
  x<-floor((dataset[4]-20000000)/100)
  return(x)}
## apply fcn which returns the month of accident date
dataindex<-apply(rawdatap,1,fcn)
rawdatap2<-cbind(rawdatap,dataindex)

data1<-rawdatap2[,c(1,2,6,9)]
data2<-rawdataa2[,c(1,2,7,8)]
datan<-merge(data1,data2,by=c("Simulation.No","Occurrence.No"))
datal<-datan[datan$Line==1,]
datam<-aggregate(datal, list(datal $Simulation.No, datal $dataindex), mean)
dim(datam[datam$State==1,])
dim(datam[datam$State==2,])

datal1<-datan[datan$Line==1,]
datans1<-datal1[datal1$State==1,]
datans2<-datal1[datal1$State==2,]

### State 1 Severity Testing
dataf1<-datans1$Payment

##draw histogram
hist(dataf1,main="Histogram of observed data")

##QQPlot
claim.ex<-rlnorm(n=1000,meanlog=10,sdlog=0.83255)
qqplot(dataf1,claim.ex,main="QQ-plot distr. Lognormal")
abline(0,1) ### a 45-degree reference line is plotted

##Histogram and PDF
h<-hist(dataf1,breaks=20)
xhist<-c(min(h$breaks),h$breaks)
yhist<-c(0,h$density,0)
xfit<-seq(min(dataf1),max(dataf1),by=1)
yfit<-dlnorm(xfit,meanlog=10,sdlog=0.83255)
plot(xhist,yhist,type="s",ylim=c(0,max(yhist,yfit)), main="Lognormal pdf and histogram")
lines(xfit,yfit, col="red")

## State 2 Severity Testing
dataf1<-datans2$Payment

##draw histogram
hist(dataf1,main="Histogram of observed data")
```

```
#QQPlot
claim.ex<- (rlnorm(n=1000,meanlog=2,sdlog=0.83255))
qqplot(dataf1,claim.ex,main="QQ-plot distr. Lognormal")
abline(0,1) ### a 45-degree reference line is plotted

##Histogram and PDF
h<-hist(dataf1,breaks=20)
xhist<-c(min(h$breaks),h$breaks)
yhist<-c(0,h$density,0)
xfit<-seq(min(dataf1),max(dataf1),by=1)
yfit<-dlnorm(xfit,meanlog=2,sdlog=0.83255)
plot(xhist,yhist,type="s",ylim=c(0,max(yhist,yfit)), main="Lognormal pdf and histogram")
lines(xfit,yfit, col="red")
```

Correlation

```
## Read raw data (Claim output file)
rawdata<-read.csv("D:/LS/RS/tsw/cc.csv",skip=1,header=TRUE)
## Manipulate claim output file to retrieve monthly frequency data for each simulation/line
fcn<-function(dataset){
  x<-floor((dataset[4]-20000000)/100)
  return(x)}
## apply fcn which returns the month of accident date
dataindex<-apply(rawdata,1,fcn)
rawdata2<-cbind(rawdata,dataindex)
rawdata3<-aggregate(rawdata2, list(rawdata2$Simulation.No,rawdata2$Line,rawdata2$dataindex,rawdata2$State),
length)
rawdata4<-rawdata3[,1:5]
data1<-rawdata4[rawdata4$Group.2==1,]
data2<-rawdata4[rawdata4$Group.2==2,]
rawdata5<-merge(data1,data2,by=c("Group.1","Group.3"))

## Test for Line 1 State 1 and Line 2 State 1. This can be changed to other combinations of states 1&2, ##2&1, and
2&2
rawdata6<-rawdata5[rawdata5$Group.4.x==1,]
rawdata7<-rawdata6[rawdata6$Group.4.y==1,]
datar<-cbind(rawdata7$Simulation.No.x,rawdata7$Simulation.No.y)
colnames(datar)<-c("Line1","Line2")
```

```
##copula test
n<-length(datar[,1])
set.seed(123)
x<- sapply(as.data.frame(datar), rank, ties.method = "random") / (n + 1)
plot(x)
```

```
##Load R packages
library(MASS)
library(methods)
library(mvtnorm)
library(scatterplot3d)
library(mnormt)
library(sn)
```

```
library(pspline)
library(copula)

##Set up copula object for copula distribution and goodness-of-fit test later
normal.cop <- normalCopula(c(0),dim=2,dispstr="un")

##Copula fit with prespecified type.
fit.normal<-fitCopula(normal.cop,x,method="ml")
fit.normal
fit.normal<-fitCopula(normal.cop,x,method="itau")
fit.normal

##Copula Goodness-of-fit test
gofCopula(normal.cop, x, N=100, method = "mpl")
gofCopula(normal.cop, x, N=100, method = "itau")

##K-S test.
normal.fit<-normalCopula(0.95, dim=2)
y<-rcopula(normal.fit,1000)
ks.test(x,y)
```

APPENDIX B. QUICK GUIDE FOR TWO-STATE REGIME-SWITCHING

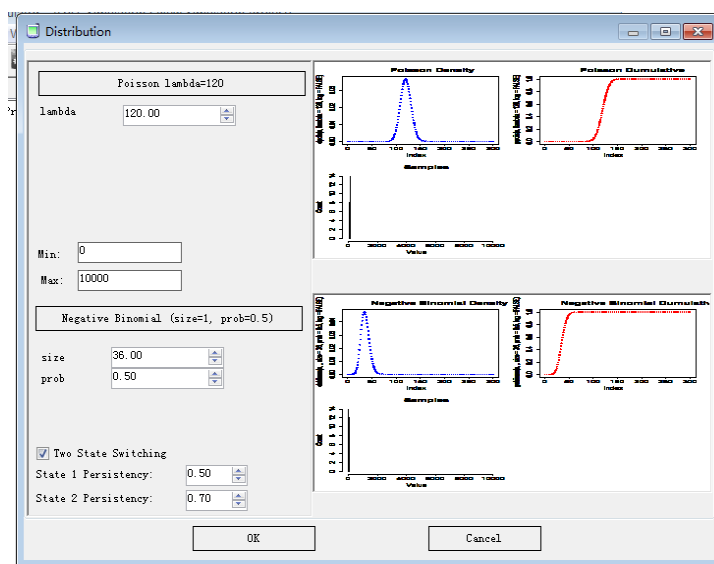
The two-state, regime-switching feature is allowed for all variables that were modeled as distribution in LSM. Below is a short description about the related model input and output.

Model Input

Figure 57 below shows the model input interface of the example in section 4.1. By checking the checkbox “Two-state Switching,” two distribution set up panels will be shown. You would also need to input the State 1 persistency and State 2 persistency. If only one distribution is desired, you could either uncheck the checkbox “Two-State Switching” or input the same distribution type and parameters for the 1st and 2nd distributions. By default, frequency and severity has two-state regime switching. For others like report lag, only one distribution is allowed. Those, however, can be changed. XML import/export setting is also revised for this enhancement.

Loss Simulation Model Testing and Enhancement

Figure 57. Model input of two-state switching feature



Model Output

Claim and transaction output files: A new column “State” is added to record the state of frequency in claim output file and state of severity in transaction output file.

Claim output example snapshot

Simulation 2011/4/10 0:30:14

Simulation	Occurrence Nb	Claim Nb	Accident Date	Report Date	Line	Type	State
1	1	1	2000126	2000701	1	1	2
1	2	1	2000101	2000318	1	1	2
1	3	1	2000106	20010105	1	1	2
1	4	1	2000123	2000823	1	1	2
1	5	1	2000116	2000129	1	1	2
1	6	1	2000223	20000514	1	1	1
1	7	1	2000213	2000327	1	1	1
1	8	1	2000218	2000530	1	1	1
1	9	1	2000223	20010209	1	1	1
1	10	1	2000222	2000823	1	1	1
1	11	1	2000210	2000309	1	1	1
1	12	1	2000326	2000413	1	1	2
1	13	1	2000307	2000614	1	1	2
1	14	1	2000412	2000528	1	1	2
1	15	1	2000422	2000816	1	1	2
1	16	1	2000402	2000626	1	1	2

Loss Simulation Model Testing and Enhancement

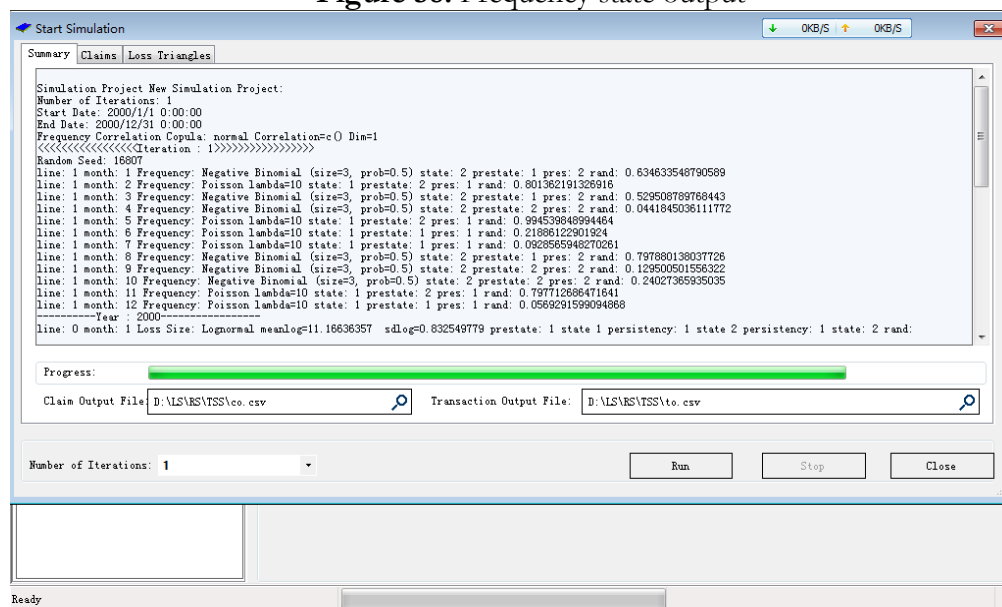
Transaction output example snapshot

Transactions 2011/4/10 0:30:14

Simulation	Occurrence	Claim	Date	Transaction	Case	Reserve	Payment	State
1	1	1	20000701	REP			2000	0 2
1	1	1	20000703	RES		-1999	0	2
1	1	1	20000726	RES		0	0	2
1	1	1	20000804	RES		0	0	2
1	1	1	20000907	RES		0	0	2
1	1	1	20001023	RES		0	0	2
1	1	1	20001108	RES		0	0	2
1	1	1	20010115	RES		0	0	2
1	1	1	20010120	RES		0	0	2
1	1	1	20010501	CLS		-1	2	2
1	2	1	20000318	REP		2000	0	2
1	2	1	20000706	RES		-1998	0	2
1	2	1	20000729	CLS		-2	3	2
1	3	1	20010105	REP		2000	0	2
1	3	1	20010115	RES		-1994	0	2
1	3	1	20010202	RES		0	0	2
1	3	1	20010424	RES		0	0	2
1	3	1	20010515	RES		0	0	2
1	3	1	20010812	CLS		-6	0	2
1	4	1	20000823	REP		2000	0	2
1	4	1	20000907	RES		-1998	0	2
1	4	1	20001129	CLS		-2	0	2
1	5	1	20000129	REP		2000	0	2
1	5	1	20000228	RES		-1992	0	2
1	5	1	20000301	CLS		-8	13	2
1	6	1	20000514	REP		2000	0	1
1	6	1	20000524	RES		-1990	0	1
1	6	1	20000907	RES		0	0	1
1	6	1	20000914	RES		0	0	1

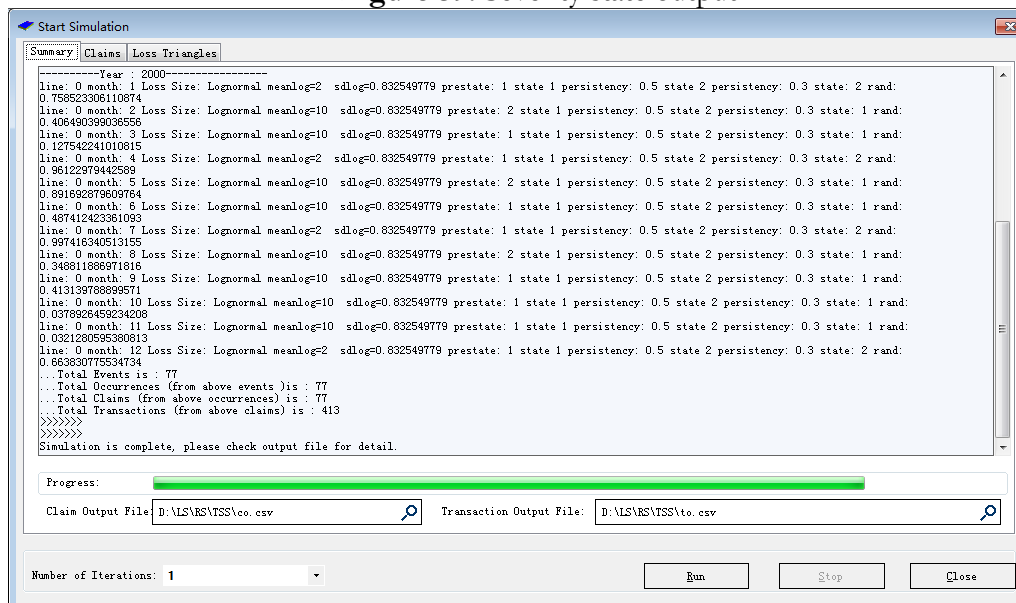
In tab “Summary” of simulation results, the state of each month’s frequency and severity is output for checking and records.

Figure 58. Frequency state output



Loss Simulation Model Testing and Enhancement

Figure 59. Severity state output



5. REFERENCES

- [1] Cleveland, Robert B., William S. Cleveland, Jean E. McRae, and Irma Terpenning, "STL: A Seasonal-Trend Decomposition Procedure Based on Loess," *Journal of Official Statistics* 6, no. 2 (1990): 3-73.
- [2] Genest, Christian, Bruno Rémillard, and David Beaudoin, "Goodness-of-fit tests for copulas: A review and a power study," *Insurance: Mathematics and Economics* 44, no. 2 (1999): 199-213.
- [3] Kojadinovic, Ivan, and Yan Jun, "Modeling Multivariate Distributions with Continuous Margins Using the Copula R Package," *Journal of Statistical Software* 34, no. 9 (2010): 1-20.
- [4] Li, David, "On Default Correlation A Copula Function Approach," *The Journal of Fixed Income* 9, no. 4 (March 2000): 43-54.
- [5] LSMWP, "Modeling Loss Emergence and Settlement Processes-CAS Loss Simulation Model Working Party Summary Report," *Casualty Actuarial Society Forum*, 2010, 4-43.
- [6] Nelsen, R.B., *An Introduction to Copulas*, 2nd ed. (New York: Springer, 2006), 109-155.

Abbreviations and notations

Collect here in alphabetical order all abbreviations and notations used in the paper	
df, degree of freedom	LSM, Loss Simulation Model
LSMWP, Loss Simulation Model Working Party	ML, Maximum Likelihood
MLE, Maximum Likelihood Estimation	OLS, Ordinary Least Square
QQ, Quantile-Quantile plot	RN, Random Number

Biography of the Author

Kailan Shang is a pricing actuary at Manulife Financial in Canada. Before that, he worked in the area of financial risk management in AIA. Years of actuarial and risk management experience has allowed him to get a broad exposure in the fields of economic capital, market-consistent embedded value, financial engineering, dynamic management options and policyholder behavior modeling, product development and management, financial reporting, dynamic solvency testing, and the like.

As an FSA, CFA, PRM, and SCJP, he is also an enthusiast of actuarial research through both volunteer works and funded research program. He participated in the LSMWP and IAA Comprehensive Actuarial Risk Evaluation project and he is now working on the SOA research projects, "Valuation of Embedded Option in Pension Plan" and "Linkage Between Risk Appetite and Strategic Planning."

He can be reached at kailan_shang@manulife.com.

GLM Invariants

Fred Klinker, FCAS, MAAA

Abstract: Those familiar with classic linear regression, as many actuaries are, are aware that, for any regression including an intercept term, there is an exact balance (equality) between (weighted) fitted and (weighted) observed values in aggregate over the whole dataset. Many are also aware that this balance also holds in aggregate over any level of any classification variable appearing in the regression as a main effect. What many may not be aware of is where these balances come from or the fact that they sometimes, but not always, extend to the GLM setting. This paper will discuss the source of the balance conditions in the so-called GLM "Normal Equations". In those cases where balance does not hold, the Normal Equations imply another invariance. The paper will also discuss some applications of these invariants.

Keywords: Generalized Linear Models (GLMs), balance conditions, invariants.

1. INTRODUCTION

Again and again over the years I have heard actuarial students new to GLM modeling express concern that, after fitting a GLM, their mean fitted values sometimes do not match their mean observed values. For anyone raised on classical linear regression, this match between mean fitted and mean observed is taken as a given, at least for regressions including an intercept term. Furthermore, in classical linear regression the match between mean fitted and mean observed holds for each and every level of any classification variable appearing as a model main effect, and this match also frequently appears to vanish in the GLM setting. What is going on? There are invariants for each GLM, depending on the distribution and the link function, but they aren't always the match between mean fitted and mean observed.

The reason has to do with what are called the "GLM Normal Equations". With each GLM combination of assumed distribution of the dependent variable and link function there is associated a set of "Normal Equations", one equation for each model regressor. These can be rewritten in the form of an equality between two quantities, one involving the observed values of the dependent variable, the other where the observed values have been replaced by their fitted values. So these are of the form of an invariance, where the GLM fit preserves some quantity when observed values are replaced by their fitted values. Frequently, this invariance is of the form of an equality between a weighted sum of fitted values and the same weighted sum of observed values. Actuaries would recognize this as the balance condition they have come to expect from classic regression.

Indeed, classic regression, assuming normally distributed errors and an identity link, produces just such balance conditions. So do Poisson count GLMs, assuming Poisson distributed dependent

variables with a log link, and so do logistic regressions, assuming binomially distributed 0/1 dependent variables and a logit link. But, before we become complacent and start to think that these balance conditions are universal, note that balance need not be preserved for severity model GLMs assuming the dependent variable gamma distributed with a log link, which is a very common actuarial model, not only for severities but also for pure premiums and loss ratios. The normal equations for GLMs with gamma distributed dependent variable and log link produce a different set of invariants other than classic balance. Even when balance is not preserved, however, the normal equations can shed some light on the sign and magnitude of the off-balance, as well as some insight as to the source of the off-balance.

It is also sometimes the case that we have available a number of possible weighting variables for our GLMs, and diagnostic residual plots may fail to give clear guidance as to which of these might be preferred. Sometimes the GLM normal equations and their implied invariants will indicate that one weighting variable will come closer to preserving balance than its competitors, and this admittedly extra-statistical, actuarial consideration may be enough to tip the balance in favor of choosing this weight over its competitors.

1.1 Outline of Remainder of Paper

The remainder of this paper proceeds as follows. Section 2 will discuss GLM normal equations and the invariants they imply. Section 3 will show what the normal equations can reveal in an off-balance situation, at least with respect to the very common GLM with gamma distributed dependent variable and log link. Section 4 will show how the normal equations can provide some guidance with regard to choice of weights, and section 5 summarizes.

2. GLM NORMAL EQUATIONS AND THEIR RESULTING INVARIANTS

One solves for the regression coefficients associated with explanatory variables in a regression or GLM by taking partial derivatives of the loglikelihood with respect to each of the regression coefficients and equating them to zero, verifying that each local extremum is indeed a local max, and hopefully a global max as well. These partial derivatives set to zero are the GLM "Normal Equations". They are a set of simultaneous equations, one equation for each regressor, actually one equation for each column of the model design matrix X , where the classic linear regression is written as the matrix expression $Y \sim X\beta$ (Y varies as $X\beta$), where Y is an n -vector of observations, X is an n by p matrix whose columns are the regressor values, for which each row represents one observation, and β is a p -vector of the regression coefficients. If the model includes an intercept term, then X includes a column of all ones to capture that intercept. Each model regressor is

GLM Invariants

included as a column of X where the values in that column are the values of the regressor. If the model includes classification variables, then there are columns in X that are indicator variables for membership in each level of the class variable (1 if in that level, 0 otherwise), etc.

GLMs are based on distributions in the exponential family, which produce loglikelihoods and normal equations of a particularly simple form. Letting x_i represent a column of the model design matrix, the normal equation associated with that x_i is:

$$0 = \sum_j w_j \frac{(y_j - \mu_j)}{V(\mu_j)g'(\mu_j)} x_{ji} \quad (2.1)$$

The sum is over all observations j . w is weight. y is observed value of the dependent variable. μ is fitted value in the original scale of the observation y and relates to the linear predictor via the link function g , where $g(\mu)$ equals the linear predictor, linear in x_i and the other regressors. $g'(\mu)$ is the first derivative of the link function evaluated at the fitted value. $V(\mu)$ is the so-called variance function associated with the distribution being assumed for this particular GLM. It expresses variance of the individual observations y as a function of their expected values. There is a variance function associated with each distribution in the exponential family. For more on GLMs and their associated loglikelihoods and normal equations, one could probably consult any standard text on Generalized Linear Models, but my personal favorite is chapter 2 of McCullagh and Nelder [1].

One can think of the above normal equation (2.1) as an invariance, because one could express this sum as a difference of two sums, one being a weighted mean y , the other being a weighted mean μ . And the normal equation says that these two sums are equal, in other words, here is a value preserved by the fitting process, the same whether we plug observed values or fitted values into it.

Now assume that our GLM includes an intercept term and consider the normal equation associated with the column of X that is all ones, the column representing the intercept term. Then the sum remains over all observations but x_i appears to drop out, because the x_{ji} are all identically one. Next, assume the model includes a classification (as opposed to continuous regressor) main effect where the classification has L levels. This class variable is encoded into the design matrix X via the inclusion of columns for indicator variables for membership in each of the class levels. (For a technical correction to this last statement and how it impacts the following argument, see the appendix.) Let x_i be one of those columns. x_i has elements equal to 1 if the observation is in that level and 0 otherwise. So the above sum becomes a sum over just the observations in that one level, with x_i again appearing to drop out, because it is identically 1 in that level and 0 elsewhere. In all the cases discussed above in this paragraph, the normal equations associated with these variables become:

GLM Invariants

$$0 = \sum_{j \in \text{sub}} w_j \frac{(y_j - \mu_j)}{V(\mu_j)g'(\mu_j)} \quad (2.2)$$

The sum is over a subset of the data, either all the data or just the observations in one level of a class variable appearing as a main effect in the model. The variable x_i appears to have disappeared, but not really; it was just an indicator variable that selected out the subset of data in the sum.

Now suppose that one has made a particularly judicious choice of link function relative to the assumed distribution for the dependent variable such that $V(\mu)g'(\mu)=1$. (Although, admittedly, this is not how one chooses a link. Rather one chooses a link based on some combination of a priori reasoning and empirical evidence that under that link $g(\mu)$ becomes at least approximately linear in the explanatory variables.) Then:

$$0 = \sum_{j \in \text{sub}} w_j (y_j - \mu_j) \quad (2.3)$$

or, equivalently:

$$\langle y \rangle_{\text{sub},w} = \langle \mu \rangle_{\text{sub},w} \quad (2.4)$$

In other words, mean observed y is equal to mean fitted μ , where the mean is taken over a certain subset of the data weighting on w . This is the balance we sought.

When is it the case that $V(\mu)g'(\mu)=1$? For classical linear regression, the dependent variable is assumed normally distributed ($V(\mu)=1$), and the identity link is assumed, $g(\mu)=\mu$, hence $g'(\mu)=1$. Hence the condition is indeed satisfied, and we get our classical balance. The condition is also satisfied for the following distribution/ link function pairs:

- Poisson count models with Poisson distributed dependent variables ($V(\mu)=\mu$) and log link ($g(\mu)=\ln(\mu)$, where \ln is the natural logarithm).
- Logistic regression with binomially distributed dependent variables ($V(\mu)=\mu(1-\mu)$) and logit link ($g(\mu)=\ln(\mu/(1-\mu))$).
- GLMs with gamma distributed dependent variable ($V(\mu)=\mu^2$) and reciprocal link ($g(\mu)=1/\mu$).

For more on distributions, their associated variance functions, and link functions, again see reference [1]. Two important comments at this point: First, from the above, we can expect our Poisson count models and logistic regressions to preserve classical balance. Second, although the above gamma model will also preserve classical balance, it is not usually the case that we will expect reciprocal expectations to be approximately linear in explanatory variables. So, although it would be convenient for purposes of preserving classical balance to adopt a reciprocal link for gamma models, we probably will not usually, because it will probably not preserve linear response. Whether the

GLM Invariants

chosen link preserves linear response can be tested via various GLM diagnostics, including diagnostic plots.

It is common to build gamma distribution models for severity, or pure premium (loss divided by exposure), or loss ratio (loss divided by premium), but assuming a log link to yield a multiplicative model. What are the normal equations, and what GLM invariants are preserved by such gamma models ($V(\mu)=\mu^2$) with log link ($g(\mu)=\ln(\mu)$)?

$$0 = \sum_{j \in sub} w_j \frac{(y_j - \mu_j)}{V(\mu_j)g'(\mu_j)} = \sum_{j \in sub} w_j \frac{(y_j - \mu_j)}{\mu_j} \quad (2.5)$$

or, equivalently:

$$\left\langle \frac{y}{\mu} \right\rangle_{sub,w} = 1 \quad (2.6)$$

Recall that $\langle y/\mu \rangle$ need not equal $\langle y \rangle / \langle \mu \rangle$, depending on the distributions of y and μ . So this identity may not be the classical balance we were hoping for, close perhaps, but not exact. What this says is that the w weighted mean of the ratio y/μ over all the data or over any individual level of any class effect appearing as a main effect in the model equals 1; the GLM fitting algorithm forces these constraints. If, in a given subset of data, there are a few significantly peculiar values of either y or μ , the mean ratio would still be constrained to be 1, but the ratio of means $\langle y \rangle / \langle \mu \rangle$ might be significantly distorted from 1 by the fitting algorithm's attempts to satisfy the constraints. It might prove interesting and possibly an important model diagnostic to drill in to see which individual observations were the greatest source of that discrepancy.

For many of the same situations for which it is common to build gamma models, it is not uncommon to also consider the alternative of a Tweedie distribution ($V(\mu)=\mu^p$) model with log link ($g(\mu)=\ln(\mu)$). p is frequently between 1 and 2, quite often 1.67. (I have heard this more than once at Predictive Modeling and RPM Seminars, but I don't have a reference. My apologies.) A Tweedie with p between 1 and 2 is a compound distribution with Poisson count process and gamma severity. As p tends down to 1 from above, the Tweedie tends towards a pure Poisson process. As p tends up to 2 from below, the Tweedie tends towards a pure gamma severity process. The relevant normal equations are:

$$0 = \sum_{j \in sub} w_j \frac{(y_j - \mu_j)}{V(\mu_j)g'(\mu_j)} = \sum_{j \in sub} w_j \frac{(y_j - \mu_j)}{\mu_j^{(p-1)}} \quad (2.7)$$

, which imply yet another set of invariants, but close to those we have already considered.

3. OFF-BALANCE

As shown in the previous section, for many distribution/ link function combinations the normal equations imply an exact balance between $\langle y \rangle$ and $\langle \mu \rangle$. For other distribution/ link combinations, even when the GLM invariants implied by the normal equations may differ from exact balance, the normal equations may provide some insight into the cause, the sign, and the magnitude of the off-balance.

The case of a gamma model with log link is particularly interesting. Suppose we measure off-balance by the quantity $\langle y \rangle / \langle \mu \rangle - 1$, where $\langle \rangle$ denotes means over subsets weighting on w , but I have suppressed the sub, w subscripts of the previous section. Then:

$$\begin{aligned} \frac{\langle y \rangle}{\langle \mu \rangle} - 1 &= \frac{\langle y \rangle}{\langle \mu \rangle} - \frac{\langle \mu \rangle}{\langle \mu \rangle} = \left\langle \frac{y}{\langle \mu \rangle} - \frac{\mu}{\langle \mu \rangle} \right\rangle = \left\langle \frac{y}{\mu} \frac{\mu}{\langle \mu \rangle} - \frac{\mu}{\langle \mu \rangle} \right\rangle = \left\langle \left(\frac{y}{\mu} - 1 \right) \frac{\mu}{\langle \mu \rangle} \right\rangle \\ &= \left\langle \left(\frac{y}{\mu} - 1 \right) \left(\frac{\mu}{\langle \mu \rangle} - 1 \right) + \left(\frac{y}{\mu} - 1 \right) \right\rangle = \text{Cov} \left[\frac{y}{\mu}, \frac{\mu}{\langle \mu \rangle} \right] \quad (3.1) \end{aligned}$$

On the second line, the second term in the mean vanishes because $\langle y/\mu \rangle$ equals 1 by equation (2.6), and the first term is recognizable as a covariance because the means of both y/μ and $\mu/\langle \mu \rangle$ are 1. Equation (2.6) drives this derivation and is therefore the point of contact in trying to explain off-balance via the normal equations.

This equating of an off-balance to a covariance is interesting. Suppose that for a particular subset of the data (either the whole of the data or a particular level of a particular classification variable in the model) the model is off-balance on the low side, in other words, $\langle \mu \rangle$ less than $\langle y \rangle$. Then that covariance is positive. Then, when $\mu/\langle \mu \rangle$ is on the high side of its mean 1, so is the ratio y/μ , on average. Likewise, on average y/μ is less than its mean 1 when $\mu/\langle \mu \rangle$ is. These observations taken together imply that y grows faster than μ on average in order to yield this behavior of the ratio y/μ , in other words, there is something about the model such that μ is tempered relative to y . On the other hand, when the model (for a particular subset) is off-balance on the high side, in other words, $\langle \mu \rangle$ greater than $\langle y \rangle$, then the relevant covariance is negative, and the above argument cuts the opposite way to imply that there is something about the model that causes μ to be somewhat over-responsive and to grow faster than y on average.

It is also possible, however, that the covariance tells us little about the behavior of the ratio y/μ on average but indicates only that there are some very anomalous values of either y or μ that throw off both the covariance and the approximate balance between $\langle \mu \rangle$ and $\langle y \rangle$. This alternative possibility was already noted in the commentary following equation (2.6). More research is probably

needed to clarify what the covariance result (3.1) is telling us, but it is certainly interesting and suggestive.

While on the subject of off-balance, there is a particularly useful scatterplot for flagging those levels of those classification variables that are more off-balance than one might expect, those levels most in need of investigation and explanation. Each data point in this scatterplot represents one level of one classification variable, and every level of every classification variable in the model has a representative point somewhere in the plot. On the y-axis we plot the ratio $\langle y \rangle / \langle \mu \rangle$, the means taken over the observations in this level of this classification variable. We draw a horizontal reference line at 1 to draw attention to deviations from 1. On the x-axis we plot class level aggregate exposures (on a log scale because these aggregate exposures might vary over a few orders of magnitude). The reason for this x-axis is that, with decreasing aggregate exposures in the level, we expect increasing scatter of the ratio $\langle y \rangle / \langle \mu \rangle$ about its hoped-for value of 1, just due to random fluctuation. We label each point according to the classification variable and level it represents, identify those variables and levels that appear to stand out from the overall pattern, and drill further into them to see if we can understand the greater than expected degree of off-balance.

It was considering just such a scatterplot that led to the investigations that led to equation (3.1) above. Certain extreme levels of one particular classification variable were flagged as excessively off-balance, and, in hindsight, taking into account known issues with the data and the model, it became clearer how under- or over-responsiveness of μ to y (in those particular levels) was responsible for that excessive off-balance.

4. IMPLICATIONS FOR WEIGHTS

Many models are built on dependent variables that are a ratio of aggregate loss to something else. If the denominator is aggregate exposure, the ratio is pure premium; if premium, loss ratio. These models are frequently built not on data at the level of individual risk but rather on data aggregated into cells defined as crossings on all the classification rating variables. The volume of business in these cells can frequently vary by orders of magnitude from one cell to the next, so some form of weighting will be needed, as the dependent variable ratios will tend to be far more volatile in low volume cells than in high. In actuarial circles it is generally assumed that a large volume cell can be treated as a sum of "independent" smaller cells, leading to variance of the dependent variable proportional to the reciprocal of some measure of business volume, which implies weights varying as some measure of business volume. But which measure? Common actuarial intuition and practice would argue for using the quantity already in the denominator of the ratio as the weight as well. It is

useful to see how the normal equations and GLM invariants of this paper bear out this choice.

I have generally found in the past, when I have fit models using each of several candidate volume measures for weights and then examined various residual plots hoping to see in those plots a signature that one particular choice of weight "outperformed" the rest, that rarely, if ever, did the plots clearly indicate one weighting scheme over the others. Some other extra-statistical, actuarial criterion has had to be imposed in order to select one weight over the rest. The normal equations can provide some guidance.

Consider first the case that our choice of distribution/ link function combination is such that equations (2.3) and (2.4) hold. y is a ratio of loss to some denominator, L/D . μ , the fitted ratio, can similarly be thought of as a ratio of fitted loss to the same denominator, \hat{L}/D . In effect, we define the fitted loss \hat{L} as the product μD . If we select the weights w equal to the D , then in equations (2.3) and (2.4) the w and D cancel each other, and these equations say simply that aggregate fitted losses are equal to (in balance with) aggregate observed losses. The equality between aggregate observed and fitted losses is not a statistical necessity (Nothing in the statistical diagnostics argues against a choice of weights other than the D from among a number of reasonable measures of business volume, but only for the w equal to the D do we achieve balance), but it seems a reasonable extra-statistical, actuarial constraint to impose as a means of rationally selecting one weighting scheme over others. Then the aggregate fitted ratio, being the ratio of aggregate fitted losses to aggregate D , equals the aggregate observed ratio, being the ratio of aggregate observed losses to aggregate D . Choice of weights w other than the denominators D in equations (2.3) and (2.4) would result in other "weighted mean fitted ratios" in balance with their corresponding "weighted mean observed ratios", but the interpretation of those "weighted mean ratios" would be far more strained than the interpretation of the more natural weighted mean ratios when w equals D . This is the gist of the usual actuarial intuition regarding weights.

Consider next the Tweedie distribution/ log link normal equations (2.7). If we again select the w equal to the D , w and D again cancel one another, we again have $L-\hat{L}$, but now divided by $\mu^{(p-1)}$:

$$0 = \sum_{j \in \text{sub}} \frac{L_j - \hat{L}_j}{\mu_j^{(p-1)}} \quad (4.1)$$

If this denominator in μ were a constant across the sum, we would again have aggregate fitted loss in balance with aggregate observed loss, but it is not constant. How non-constant is it, because, if close to constant, perhaps aggregate fitted loss and aggregate observed loss may not be far out of balance? First, across much of the data, the range of μ may be relatively modest. Second, in those applications of Tweedie I have seen, p is rarely less than 1.5 or greater than 1.67, so the power of μ

is something like $1/2$ or $2/3$, which further tempers the range of values in the denominator and so brings aggregate fitted and observed losses closer to in balance. In Tweedie/ log link models, I have seen aggregate fitted and observed losses in balance to within a few percent of one another when w is selected equal to D , whereas out of balance by as much as a few tens of percent when another weighting scheme is selected.

Lastly, considering the gamma/ log link model yielding equations (2.5) and (2.6), these look like the Tweedie/ log link case of equation (2.7) but with a p of 2. Hence the exponent on μ in the denominator is 1 rather than the $1/2$ to $2/3$ of the Tweedie case, there is less tempering of the range of the power of μ , and aggregate fitted losses and aggregate observed losses can be more out of balance than in the Tweedie case, even when we select the w equal to the D .

5. SUMMARY AND CONCLUSIONS

Many of us are familiar with the balance between weighted mean fitted values and weighted mean observed values in standard linear regression settings. These same balance conditions extend to many GLM settings with various combinations of assumed distribution of the dependent variable and link function. The source of these balance conditions are the so-called "GLM Normal Equations". Even for those GLMs with distribution/ link function combinations not preserving the usual balance conditions, there is always another GLM invariance implied by the normal equations. The normal equations can also help us to understand the direction and degree of off-balance when off-balance exists as well as understand the consequences to balance or off-balance when choosing a weighting scheme for our weighted GLMs.

Appendix: A Technical Refinement of the Balance Equations Argument for Model Design Matrices of Full Rank

In section 2 of this paper, the argument leading from the normal equations to the balance equations assumed that, for each level of each classification variable appearing in the model as a main effect, the model design matrix included a column equal to the indicator variable for that level of that classification variable. For technical reasons this may not be quite true, and the argument requires a technical refinement, but the conclusion re the balance equations is still true.

What could be wrong with the original argument? Suppose the design matrix includes a column of ones, representing the intercept, and columns for indicator variables for each and every level of a classification variable. Because the sum of all these indicator variable columns reproduces the column of one's (because every observation is in precisely one level), there is a linear dependence

among the design matrix columns, the design matrix is less than full rank, and its inverse is not uniquely defined, which creates problems solving for the regression coefficients. To resolve this issue, many stat packages arbitrarily select one of the levels of the classification variable to serve as the reference level for that classification variable, remove the indicator variable for that level from the design matrix, and peg the regression coefficient and standard error estimate for that level to zero. The resulting reduced design matrix is now full rank and invertible, at the cost of having arbitrarily selected a reference level and removed that indicator variable from the design matrix.

Because the design matrix still includes the column of all ones, the argument of section 2 of this paper establishes that overall balance still holds. Also because of the section 2 argument, balance still holds in each level of the classification variable for which there is still an indicator variable in the design matrix, but does balance still hold for the reference level, given that there is now no indicator variable for that level in the design matrix? One would think so, given overall balance and balance in every other level, that this would imply balance in the reference level as well, and one would be right, for the following reason. Because equation (2.1) holds for each column of the design matrix, it also holds for all linear combinations of those columns:

$$\sum_j w_j \frac{(y_j - \mu_j)}{V(\mu_j)g'(\mu_j)} \sum_i \alpha_i x_{ji} = \sum_i \alpha_i \sum_j w_j \frac{(y_j - \mu_j)}{V(\mu_j)g'(\mu_j)} x_{ji} = \sum_i \alpha_i * 0 = 0 \quad (A.1)$$

where j indexes the observations (the rows of the design matrix), i indexes the variables (the columns of the design matrix), the order of summation can be reversed, and the inner sum of the second expression is zero by equation (2.1). The indicator variable for the reference level of the classification variable in question can indeed be expressed as a linear combination of the other columns of the design matrix, which is why it was declared to be a reference level in the first place. So balance holds for this reference level as well. QED. In fact, equation (A.1) establishes that a normal equation holds not just for any column of the design matrix but also for any other variable that can be expressed as a linear combination of the columns of the design matrix.

6. REFERENCES

- [1] McCullagh, P., and J. A. Nelder, Generalized Linear Models, Second Edition, Chapman and Hall/CRC, 1989.

Author's Biography

Fred Klinker (FCAS, MAAA) is an actuarial manager with ISO's Modeling Division. He is involved with various efforts to introduce modern predictive modeling techniques, especially GLMs, into standard ISO ratemaking. Much of his career prior to ISO was spent as a research actuary involved in financial and statistical modeling at CIGNA P&C, with brief forays into regulation at the Massachusetts Insurance Department and reinsurance at Zurich Re North America, later Converium.

The Retrospective Testing of Stochastic Loss Reserve Models

Glenn Meyers, FCAS, MAAA, CERA, Ph.D.

ISO Innovative Analytics

and

Peng Shi, ASA, Ph.D.

Northern Illinois University

Abstract

Given an $n \times n$ triangle of losses, $X_{AY,Lag}$ ($AY = 1, \dots, n$, $Lag = 1, \dots, n$, $AY + Lag < n + 2$), the goal of a stochastic loss reserve model is to predict the distribution of outcomes, $X_{AY,Lag}$ ($AY + Lag > n + 1$), and sums of losses such as $\sum_{AY=2}^n \sum_{Lag=n+2-AY}^n X_{AY,Lag}$. This paper will propose a set of diagnostics to test the predictive distribution and illustrate the use of these diagnostics on American insurer data as reported to the National Association of Insurance Commissioners (NAIC).

- The data will consist of incremental paid losses for the commercial automobile line of insurance. This data will come from a database containing both the original loss triangles and the outcomes. This database will contain data for hundreds of American insurers, and it will be posted on the Casualty Actuarial Society (CAS) website for all researchers to access.
- The retrospective tests are performed on the familiar stochastic loss reserve model, the bootstrap chain ladder overdispersed Poisson model. The paper will also perform the retrospective tests on a model proposed by the authors.
- The authors' model will assume that the incremental paid losses have a Tweedie distribution, with the expected loss ratio and calendar year trend parameters following an AR(1) time series model. The model will be a hierarchical Bayesian model with the posterior distribution of parameters being estimated by Markov-Chain Monte-Carlo (MCMC) methods.

1. Introduction

In the classic reserving problem for property-casualty insurers, the primary goal of actuaries is to set an adequate reserve to fund losses that have been incurred but not yet developed. In this regard, the reserving actuaries are more interested in a reasonable reserve range rather than a best estimate. Traditional deterministic algorithms are often sufficient for the best estimation of outstanding liabilities, but often insufficient in estimating the downside potential in loss reserves. Over the past three decades, stochastic claims reserving methods have received extensive development, emphasizing the role of variability in claims reserves.

In claims reserving literature, different stochastic methods are proposed to calculate the predictive uncertainty of reserves and, ideally, to derive a full distribution of outstanding payments. The variability of claims reserves could be decomposed into two components, a process error which is intrinsic to the stochastic model and an estimation error that describes the uncertainty in parameter estimates. Both non-parametric and parametric approaches have been discussed along this line of studies. The so-called non-parametric models (various Chain-Ladder techniques among others) are considered by some to be distribution-free and focus on (conditional) mean-squared prediction error to measure the quality of reserve estimates. Parametric models, in contrast, are based on distributional families and thus could lead to a distribution of outstanding claims. Because of the small sample size typically encountered in loss reserving context, the bootstrapping technique and Bayesian method are often involved to incorporate the uncertainty in parameter estimates and thus to provide a predictive distribution for unpaid losses. We refer to Taylor (2000), England and Verrall (2002), and Wüthrich and Merz (2008) for excellent reviews on stochastic loss reserving methods.

With an increasing number of stochastic claims reserving methods emerging in the literature, one critical question to ask is how to evaluate their predictive performance. This question could only be answered based on retrospective tests using the actual realized claims in the lower triangle. Unfortunately, such issue has rarely been addressed in the current literature. Shi *et al.* (2011) is one recent example.

The Retrospective Testing of Stochastic Loss Reserve Models

The goals of this paper are threefold: 1) We will propose a stochastic loss reserving model based on a Tweedie distribution that captures the calendar year trend in claims development. 2) A set of diagnostics will be discussed to test the predictive distribution of outstanding liabilities. The retrospective evaluation will be performed for the proposed method as well as standard formulas. 3) We emphasize the importance of retrospective testing in both loss reserving and risk management practice, and we anticipate that this work will initialize more relevant studies and draw attention from both practitioners and researchers in this perspective.

We note that the sparsity of studies on retrospective tests might be attributed to the unavailability of the data on realized claims. Our access to a rich database from the National Association of Insurance Commissioners (NAIC) provides us an opportunity to perform such evaluation. A great deal of effort has been devoted to the preparation of a quality dataset for loss reserve studies. The detailed summary of the loss reserve dataset is given in Section 2 and the Appendix. We will also post the dataset on the website of the Casualty Actuarial Society (CAS)¹.

The NAIC database contains information on both posted reserves and subsequent paid losses, which allows us to evaluate: 1) the performance of the predictive distribution based on actual losses; 2) the predictive distribution based on posted reserves; 3) the sufficiency of the posted reserves. We will compare the predictive performance between the proposed method and a standard formula. Our analysis will focus on claims-reserve models for a single line of business. It is worth mentioning the emerging reserve studies for dependent lines of business. The retrospective tests for multivariate loss reserving methods could be a direction of future research.

The structure of this article is as follows: Section 2 describes the run-off triangle data from the NAIC and discusses the selection process for the insurers in our analysis. Section 3 presents two stochastic loss reserving method, the chain-ladder over-dispersed Poisson and the Bayesian Tweedie model. Section 4 and Section 5 report the results of retrospective tests for a single insurer and multiple insurers, respectively. Section 6 concludes the paper.

¹ The link for these data is http://www.casact.org/research/index.cfm?fa=loss_reserves_data

2. Data

The claims triangle data used for the retrospective test are from Schedule P of the NAIC database. The NAIC is an American organization of insurance regulators that provides a forum to promote uniformity in insurance regulation among different states. It maintains one of the world's largest insurance regulatory databases, including the statutory accounting report for all insurance companies in the United States.

We consider Schedule P of property-casualty insurers, which includes firm-level run-off triangles of aggregated claims for major personal and commercial lines of business. And the claims are available for both incurred and paid losses². The triangles of paid losses in Schedule P of year 1997 will be used to develop stochastic loss reserving models. Each triangle contains losses for accident years 1988-1997 and at most ten development years. The net premiums earned in each accident year are available for the measurement of business volume. For any insurer, the triangle for a single line of business could be illustrated as in Figure 1. The crosses indicate the data point extracted from 1997 Schedule P.

Figure 1. Schedule P of 1997

Accident Year	Premium	Settlement Lag											
		1	2	3	4	5	6	7	8	9	10		
1988	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	
1989	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	← 1998
1990	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	← 1999
1991	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	← 2000
1992	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	← 2001
1993	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	← 2002
1994	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	← 2003
1995	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	← 2004
1996	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	← 2005
1997	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	xxx	← 2006

To perform the retrospective test, one needs the realized claims in the lower triangle. We square the triangles from Schedule P of year 1997 with outcomes from the Schedule P of

² By “losses” we mean “Incurred net losses and defense & cost containment expenses reported at year end” as specified by the NAIC Schedule P instructions.

The Retrospective Testing of Stochastic Loss Reserve Models

subsequent years. To be more specific, as shown in Figure 1, the losses in accident year 1989 are pulled from the Schedule P of year 1998, the losses in accident year 1990 are pulled from the Schedule P of year 1999, and so on. The overlapping observations from the Schedule P of year 1997 and subsequent years are used to validate the quality of our data. The insurers with inconsistency in the overlapping period are dropped from this study. The detailed process of data preparation can be found in the Appendix. In addition to the actual losses in the lower triangle, the NAIC database provides posted reserves of year 1997. The posted reserves represent the actual amount of fund set by reserving actuaries, based on the predictions from certain claim reserving models, as well as actuarial judgments.

We focus on the run-offs of commercial auto in the retrospective test. Commercial auto is a relatively short tail line and thus the claims are very likely to be closed within ten years. This fact makes the Schedule P data an appropriate first candidate for the retrospective evaluation. The triangles consist of losses net of reinsurance, and quite often insurer groups have mutual reinsurance arrangements between the companies within the group. Consequently, we limit our analysis to single entities, be they insurer groups or true single insurers.

For the retrospective tests, we wanted to test only those insurers we deemed to be “going concern” insurers. Our criterion for selecting insurers was that: (1) earned premium was not subject to wide swings; and (2) the insurers were generally profitable. To implement these criteria we first calculated the coefficient of variation for the earned premium over each of the ten accident years. We then sorted the insurers in increasing order of this coefficient of variation. Then we individually examined the profitability of each insurer, rejecting those insurers that we deemed unprofitable. In the end we selected 50 insurers for this analysis.

Figure 2 shows the earned premiums and cumulative paid losses by accident year for the first insurer we accepted, and Figure 3 shows the earned premium and losses by accident year for the first insurer we rejected. Table 1 gives the Group Codes for all insurers included in this analysis.

The Retrospective Testing of Stochastic Loss Reserve Models

Table 1
Insurer Group Codes

1236	353	14974	21270	1406	914	4839	7080	833	1767
37036	5428	26077	13641	86	1538	38733	2674	388	1759
3492	6947	11037	1090	4731	3240	2623	3034	18767	5185
2500	14176	2135	620	26433	31550	44130	2208	10022	310
2283	1066	8427	10839	19020	26905	671	13528	715	9504

Figure 2

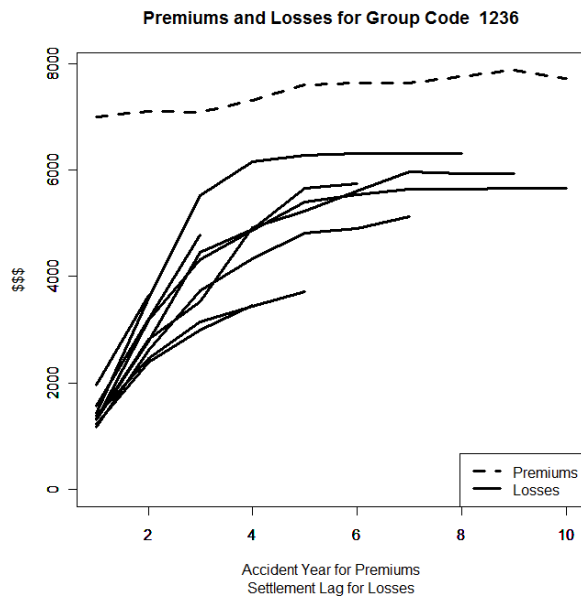
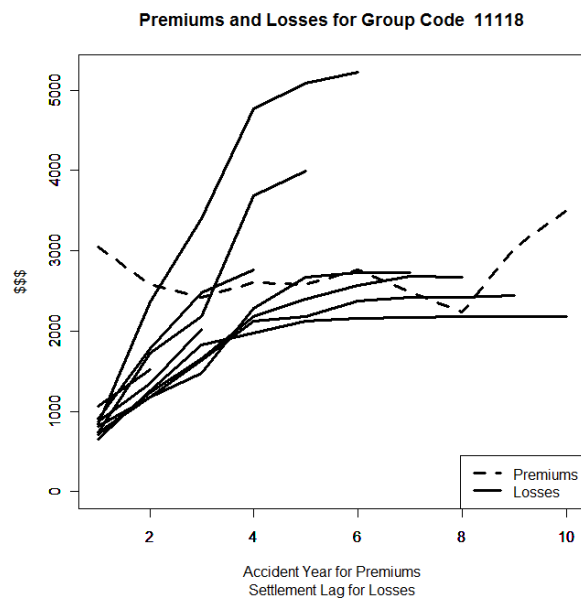


Figure 3



3. Two Loss Reserving Models

Our analysis focuses on incremental paid data. In each run-off triangle, we use $X_{AY,Lag}$ to indicate incremental paid losses for accident years, $AY = 1, \dots, 10$ and settlement lags, $Lag = 1, \dots, 10$. Thus, the paid losses in the upper triangle (training data) and unpaid losses in the lower triangle (test data) could be represented by \mathbf{X}^U and \mathbf{X}^L , respectively:

$$\mathbf{X}^U = \{X_{AY,Lag} : AY + Lag \leq 11\} \text{ and } \mathbf{X}^L = \{X_{AY,Lag} : AY + Lag > 11\}.$$

The retrospective test will be performed for the predictive distributions of elements or functions of elements in set \mathbf{X}^L .

The predictive distribution of outstanding liabilities could be obtained either through bootstrapping techniques or Bayesian methods. In this study, we will propose a Bayesian Autoregressive Tweedie (BAT) model for the prediction of unpaid loss, which is described in the next section. We compare the performance of the proposed method with an industry benchmark, the bootstrap chain-ladder (BCL) model, where the predictive variability of unpaid losses is derived through bootstrapping technique with an over-dispersed Poisson process error. A common thread running through the two models is that they both treat parameter risk by producing simulations of possible parameters for the model (BCL – bootstrap, BAT – Markov Chain Monte-Carlo). Both models treat process risk (BCL – the overdispersed Poisson distribution, BAT - the Tweedie distribution).

3.1 The Bootstrap Chain Ladder (BCL) Model

Bootstrap chain-ladder is simply a chain-ladder algorithm where bootstrapping is employed to accommodate estimation uncertainty. This technique has been applied to both univariate and multivariate loss reserving context; for example, see England and Verrall (2002) and Kirschner et al. (2008). To make this work self-contained, we briefly review the method as follows:

- Apply chain-ladder algorithm to cumulative payments and obtain the fitted incremental payments $\hat{X}_{AY,Lag}$ for $AY + Lag \leq 11$.
- Calculate scale parameter and adjusted Pearson residual

$$\hat{\phi} = \frac{1}{m-p} \sum_{AY+Lag \leq n+1} \left(\frac{X_{AY,Lag} - \hat{X}_{AY,Lag}}{\sqrt{\hat{X}_{AY,Lag}}} \right)^2 \quad \text{and} \quad \hat{R}_{AY,Lag} = \sqrt{\frac{m}{m-p}} \frac{X_{AY,Lag} - \hat{X}_{AY,Lag}}{\sqrt{\hat{X}_{AY,Lag}}},$$

respectively, where $m = n(n+1)/2=55$ and $p = 2n - 1=19$.

- Resample the residuals $\hat{R}_{AY,Lag}^{(s)}$ ($AY + Lag \leq 11$) and create pseudo-triangle by $X_{AY,Lag}^{(s)} = \hat{R}_{AY,Lag}^{(s)} \sqrt{\hat{X}_{AY,Lag}} + \hat{X}_{AY,Lag}$ for $s = 1, \dots, S$.
- Apply chain-ladder algorithm to the cumulative pseudo-payments obtained from $X_{AY,Lag}^{(s)}$ ($AY + Lag \leq 11$) and project the incremental payments in the lower triangle $\hat{X}_{AY,Lag}^{(s)}$ for $AY + Lag > 11$.
- For each cell (AY, Lag) ($AY + Lag > 11$), simulate a payment from a process distribution with mean $\hat{X}_{AY,Lag}^{(s)}$ and variance $\hat{\phi} \hat{X}_{AY,Lag}^{(s)}$, for $s = 1, \dots, S$.

Commonly used process distributions include gamma and over-dispersed Poisson. We report the results based on the latter process error since it is well known that the over-dispersed Poisson model using incremental payments reproduces chain-ladder predictions under certain regularity conditions (see Renshaw and Verrall (1998) and Verrall (2000) for details). Furthermore, a preliminary analysis shows the difference in the predictions based on the two types of process distributions is negligible. We implemented the bootstrap chain-ladder method using the “ChainLadder” package in the statistical computing software R.

3.2 The Bayesian Autoregressive Tweedie (BAT) Model

The objective of this model is given the observed data X^U , predict the distribution of the sum of all amounts in X^L .

The high-level considerations made in formulating this model include:

1. The model should use the reported premiums as a measure of exposure. This consideration has precedent with the Bornhuetter-Ferguson method, but it differs from other popular models such as the chain-ladder. Given that the model uses premiums, it should recognize that competitive conditions in the American insurance industry lead to slowly changing loss ratios over time.
2. As the settlement lag increases, the payments follow no discernable natural pattern other than ultimately, they approach zero.
3. The model should reflect inflationary changes in loss levels by calendar year. This consideration has precedent with other models such as the one proposed by Barnett and Zehnwirth (2000). The model should recognize that inflation can change slowly over time.
4. Process risk is present and important for (AY, Lag) cells with low expected losses. In general, the coefficient of variation of the process risk should decrease as the expected loss increases, but it should never approach zero. Also, the process risk in the later settlement lags should reflect the larger claims that take longer to settle.
5. The model is Bayesian. Loss reserve models tend to have many parameters. As demonstrated by Meyers (2007a), loss reserve models fit by maximum likelihood with a large number of parameters tend to understate the variance of the outcomes. Bayesian approaches will correct for this by incorporating parameter risk into calculating the variance of the outcomes. Other approaches, such as bootstrapping, also incorporate parameter risk.

The Retrospective Testing of Stochastic Loss Reserve Models

The unknown parameters for this model are as follows.

- ELR_{AY} , for $AY = 1, \dots, 10$. These parameters represent the expected loss ratio for accident year AY .
- Dev_{Lag} , for $Lag = 1, \dots, 10$. These parameters represent the paid incremental loss development factors for settlement lag Lag . To prevent overdetermining the model we imposed the constraint that $\sum_{Lag=1}^{10} Dev_{Lag} = 1$.
- CYT_i , for $i = 1, \dots, 19$. These parameters represent the calendar year trend factor. For a given (AY, Lag) cell, we have $i = AY + Lag - 1$. To prevent overdetermining the model we set $CYT_1 = 1$.
- Sev represents the claim severity for claims that settle in the 10th settlement lag. For $Lag < 10$, the claim severity is given by $Sev \cdot (1 - (1 - Lag / 10)^3)$. This expression for the claim severity guarantees that the claim severity increases as the settlement lag increases.
- c represents the contagion parameter as described in Meyers(2007b). Its role is to keep the coefficient of variation of the process risk from decreasing to zero as the expected loss increases. Its precise role will be specified in the likelihood function below.

To allow the $\{ELR_{AY}\}$ parameters to change slowly over time, we impose the following $AR(1)$ structure on the parameters.

$$ELR_{AY} = \mu_A \cdot (1 - \rho_A) + \rho_A \cdot ELR_{AY-1} + \varepsilon_A.$$

From the standard properties of the $AR(1)$ model we have that:

- The long-term average of the ELR_{AY} parameters = μ_A .
- $\text{Corr}(ELR_{AY}, ELR_{AY-k}) = \rho_A^k$.
- $\varepsilon_A \sim N(0, \sigma_A)$.

The Retrospective Testing of Stochastic Loss Reserve Models

The prior distribution of $\{\{ELR_{AY}\}, \mu_A, \rho_A, \sigma_A\}$ takes the form:

$$p(\{\{ELR_{AY}\}, \mu_A, \rho_A, \sigma_A\}) = f(\mu_A) \cdot g(\rho_A) \cdot h(\sigma_A) \cdot \prod_{AY=2}^{10} \Phi(ELR_{AY} - \mu_A \cdot (1 - \rho_A) - \rho_A \cdot ELR_{AY-1} | 0, \sigma_A)$$

where:

- Φ is the standard normal distribution.
- f is a gamma distribution with mean 0.7 and coefficient of variation 0.18.
- g is a uniform (0,1) distribution.
- h is a gamma distribution with mean 0.025 and coefficient of variation 0.5.

We impose a similar structure on $\{CYT_i\}$ with the prior distribution taking the form:

$$p(\{\{CYT_i\}, \mu_c, \rho_c, \sigma_c\}) = f(\mu_c) \cdot g(\rho_c) \cdot h(\sigma_c) \cdot \prod_{i=2}^{10} \Phi(CYT_i - \mu_c \cdot (1 - \rho_c) - \rho_c \cdot CYT_{i-1} | 0, \sigma_c)$$

where:

- Φ is the standard normal distribution.
- f is a gamma distribution with mean 1 and coefficient of variation 0.18.
- g is a uniform (0,1) distribution.
- h is a gamma distribution with mean 0.025 and coefficient of variation 0.5.

The Retrospective Testing of Stochastic Loss Reserve Models

The prior distributions for the remaining parameters were gamma distributions with the parameters given in Table 2. These were derived by fitting a similar model by maximum likelihood to a large number of insurers.

Table 2

Parameter	α	θ	Implied	
			Mean	Std. Dev.
<i>Sev</i>	1.3676	136.248	186.3386	159.3400
<i>c</i>	0.074	0.1391	0.0103	0.0379
<i>Dev</i> ₁	15.81	0.0135	0.2137	0.0537
<i>Dev</i> ₂	42.8538	0.0059	0.2517	0.0385
<i>Dev</i> ₃	56.4944	0.0036	0.2028	0.0270
<i>Dev</i> ₄	30.4528	0.0046	0.1403	0.0254
<i>Dev</i> ₅	10.2309	0.0085	0.0870	0.0272
<i>Dev</i> ₆	5.8094	0.0083	0.0480	0.0199
<i>Dev</i> ₇	3.6954	0.0068	0.0250	0.0130
<i>Dev</i> ₈	2.3934	0.0057	0.0135	0.0087
<i>Dev</i> ₉	1.3559	0.0066	0.0090	0.0077
<i>Dev</i> ₁₀	0.4552	0.0200	0.0091	0.0135

The joint prior distribution for all the parameters is the product of all the individual prior distributions given above.

We used the Tweedie distribution with index $p = 1.67$ to describe the process risk. For a given (AY, Lag) cell, the expected loss is given by:

$$E[X_{AY, Lag}] = Premium_{AY} \cdot ELR_{AY} \cdot Dev_{Lag} \cdot \prod_{i=1}^{AY+Lag-1} CYT_i.$$

The scale parameter for the Tweedie distribution for each (AY, Lag) cell is given by:

$$\phi = \frac{E[X_{AY, Lag}]^{1-p} \cdot Sev \cdot \left(1 - \frac{Lag}{10}\right)^3}{2-p} + c \cdot E[X_{AY, Lag}]^{2-p}.$$

The Retrospective Testing of Stochastic Loss Reserve Models

This expression for ϕ can be explained by noting that the variance for the Tweedie distribution is usually written in the form $\phi \cdot \mu^p$. Substituting $\mu = E[X_{AY,Lag}]$ and the value above for ϕ into the expression for the Tweedie variance yields a variance of $E[X_{AY,Lag}]/(2-p)+c \cdot E[X_{AY,Lag}]^2$. The coefficient of variation squared is then equal to $\frac{1}{E[X_{AY,Lag}] \cdot (2-p)} + c$. This coefficient of variation squared decreases to c as the expected loss, $E[X_{AY,Lag}]$, increases.

The likelihood function for the data³ in the upper triangle is the product of the Tweedie density functions over all the (AY,Lag) cells in the upper triangle, \mathbf{X}^U .

With the prior distribution and the likelihood function specified above, we used the Metropolis Hastings algorithm⁴ to generate a sample of size 1,000 parameter sets from the posterior distribution.

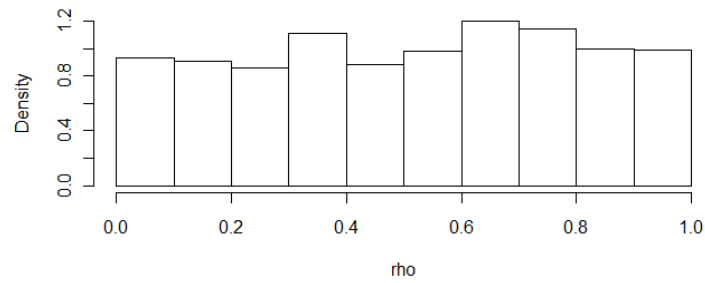
Figures 4 to 14 below graphically show how the data reduces the uncertainty in the range in the parameters by comparing the prior and posterior distributions of the parameters. We produced these plots using the data of the insurer with group code 914.

³ In fitting the data, we dropped all (AY,Lag) cells with negative paid incremental losses.

⁴ See Meyers (2009) for an explanation of the Metropolis Hastings Algorithm. For each parameter, we used a gamma distribution with a shape parameter, $\alpha = 2,000$, for the proposal density function. To obtain convergence and guard against autocorrelation, we ran 50,000 iterations and took a sample of size 1,000 from the last 25,000 iterations.

Figure 4

Prior Distribution of 'rho' for ELR Model



Posterior Distribution of 'rho' for ELR Model

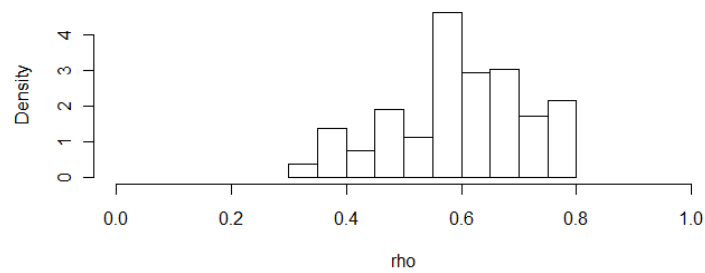
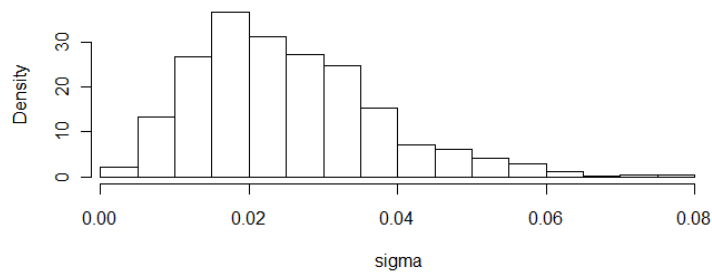


Figure 5

Prior Distribution of 'sigma' for ELR Model



Posterior Distribution of 'sigma' for ELR Model

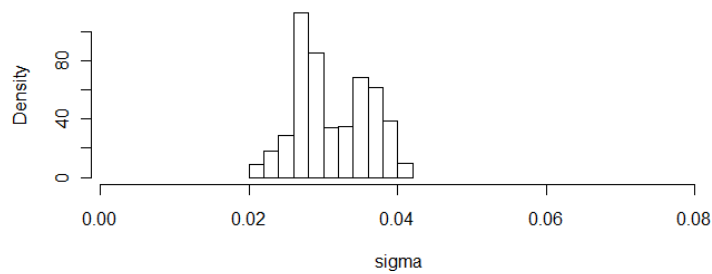


Figure 6

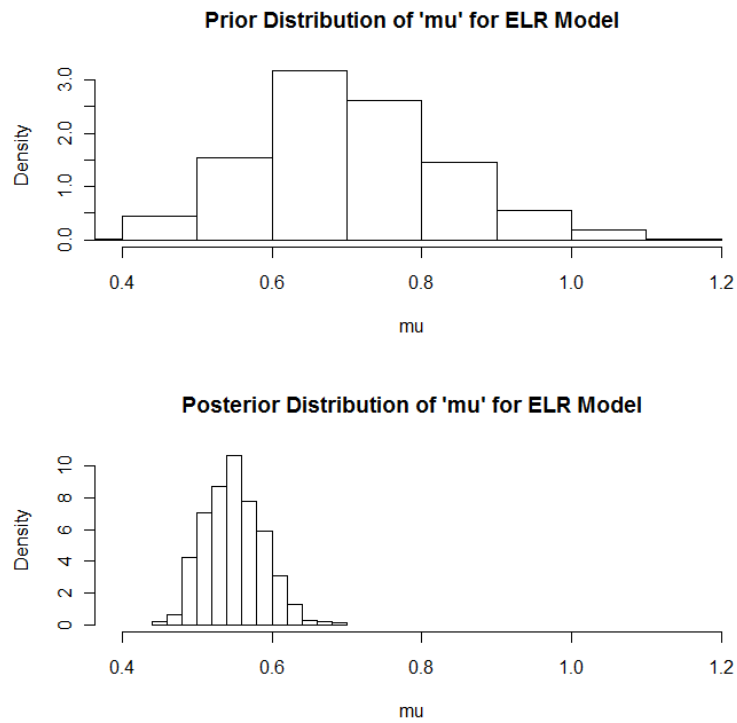


Figure 7

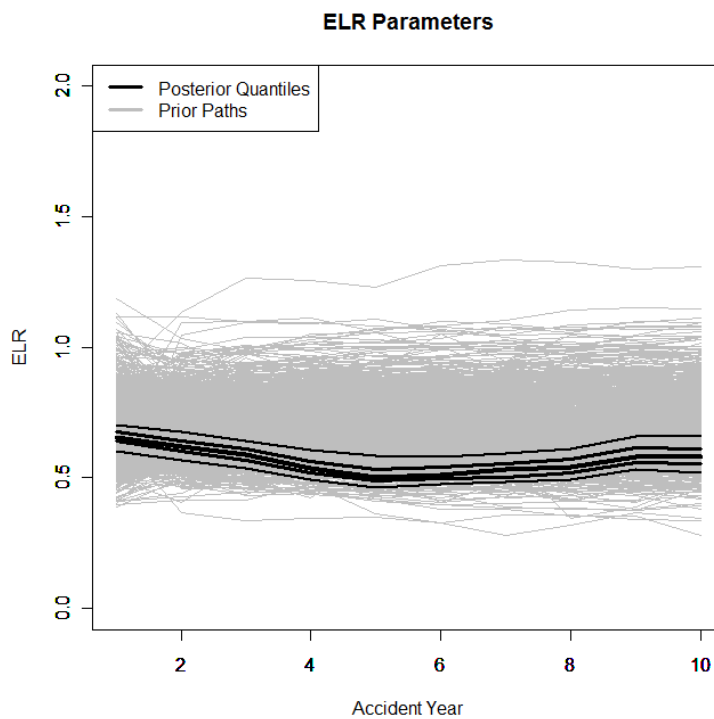


Figure 8

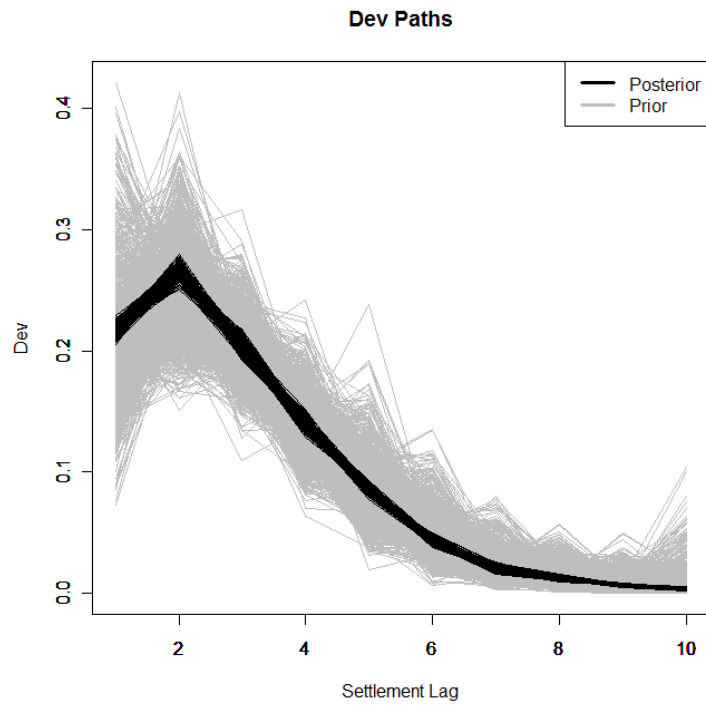


Figure 9

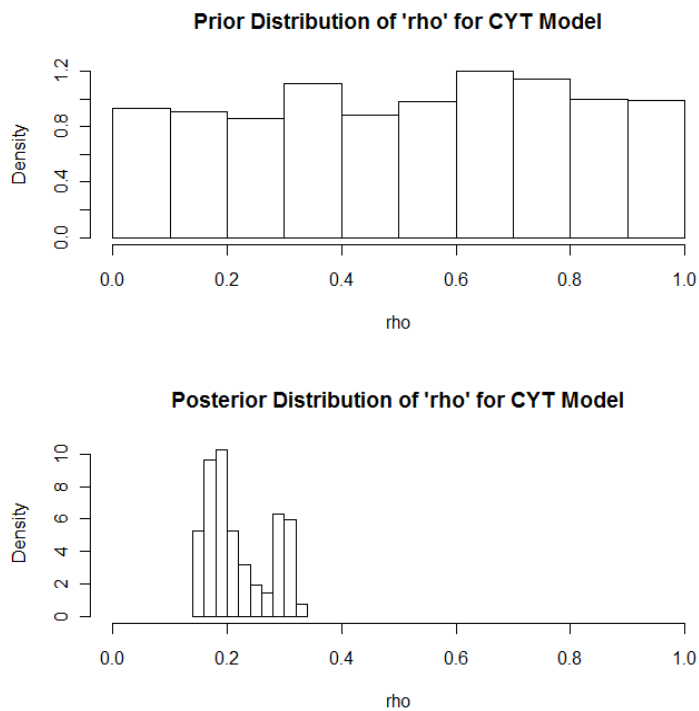


Figure 10

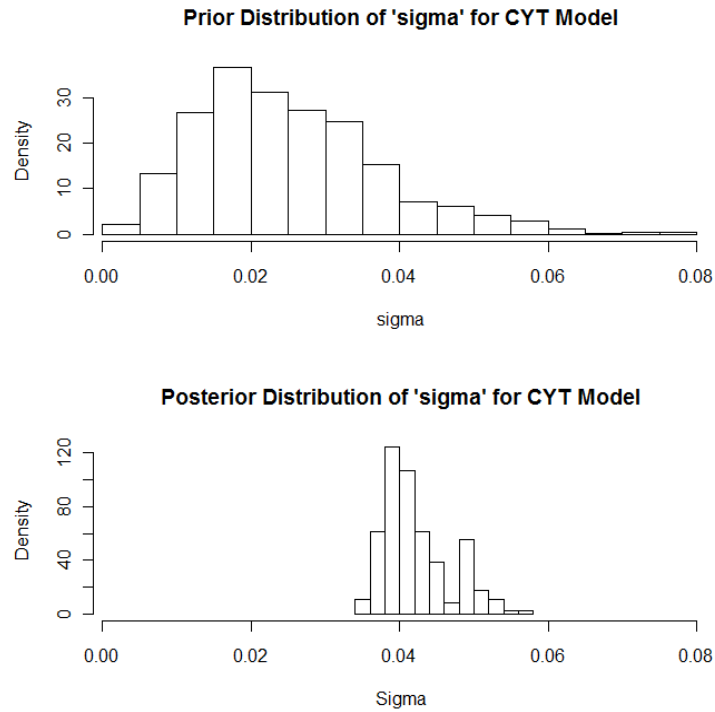


Figure 11

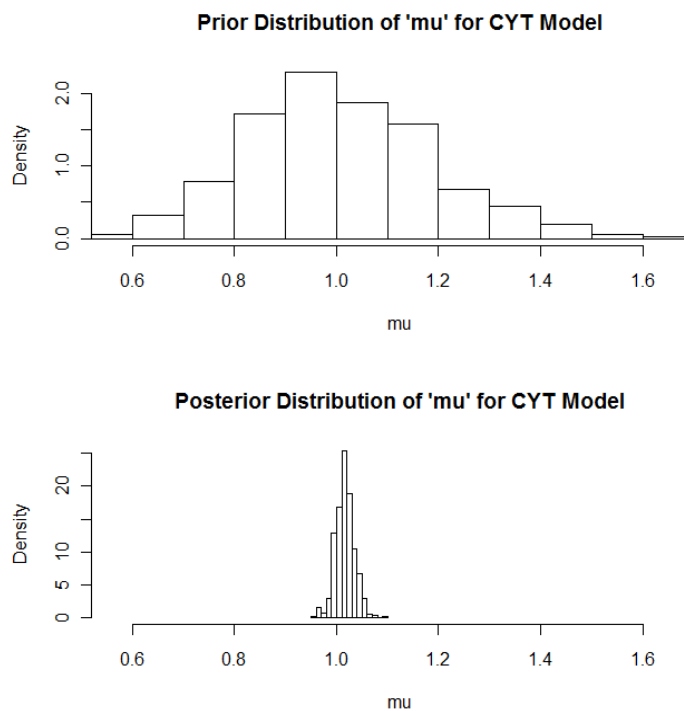


Figure 12

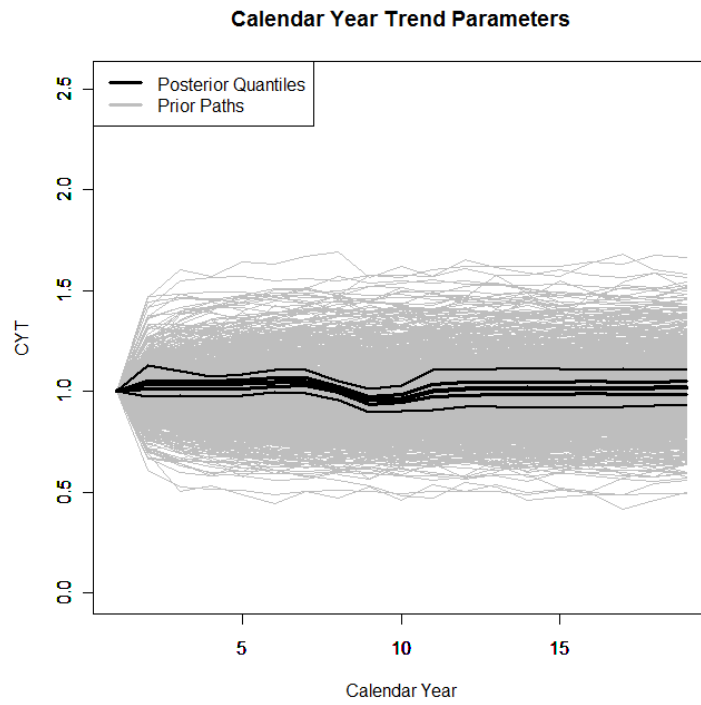


Figure 13

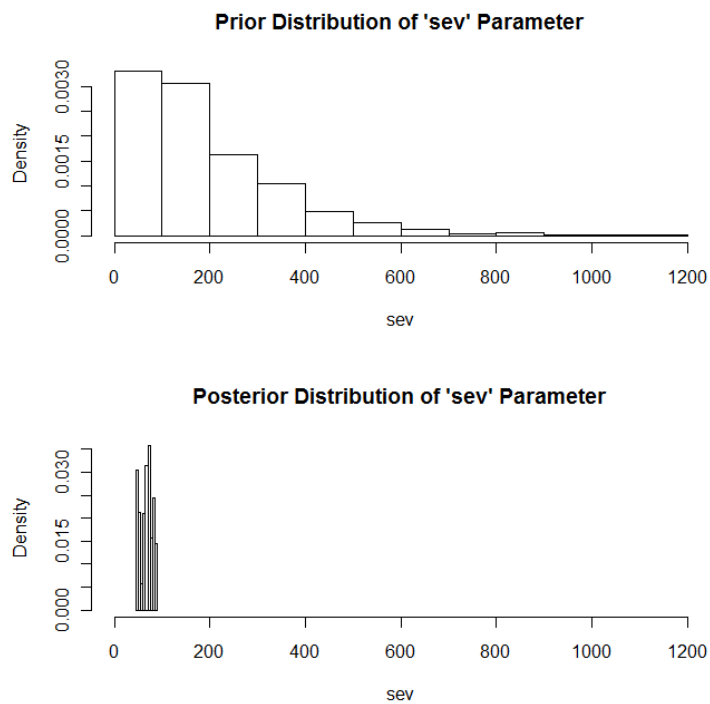
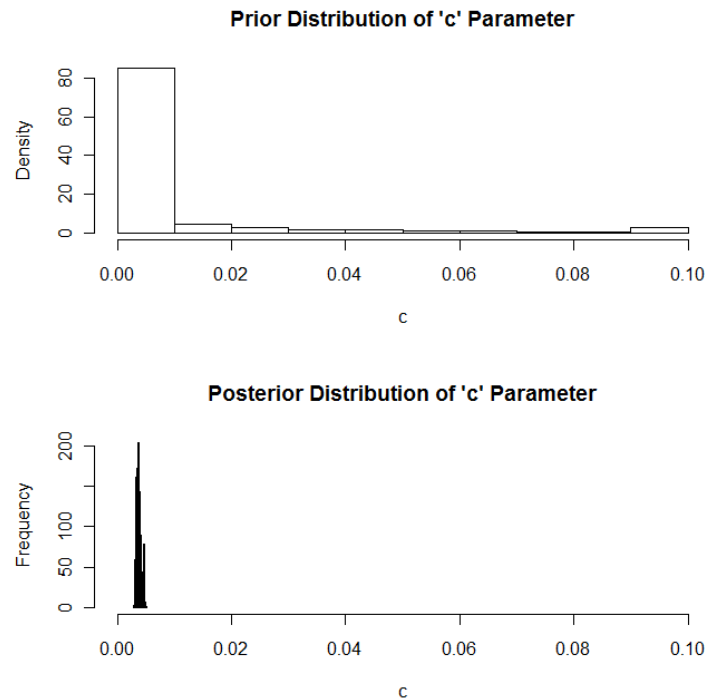


Figure 14



For each of the 1,000 randomly selected parameter sets $\{\{ELR_{AY}\}, \{Dev_{Lag}\}, Sev, c\}$, we then calculated the mean and variance of the Tweedie distribution of $X_{AY,Lag}$ for each (AY,Lag) cell in the lower triangle and then took 10 different random simulations of $\sum_{AY=2}^{10} \sum_{Lag=12-AY}^{10} X_{AY,Lag}$. These simulations produced 10,000 samples of this sum. Given the amount of an outstanding liability, we calculate the cumulative probability by counting the number of simulations that are less than or equal to it.

4. Retrospective Tests for Single Insurers

Loss reserve models are calibrated using the observed run-off triangle and then are used to forecast outstanding liabilities. From the perspective of risk management, a reasonable reserve range is of more interest to reserving actuaries and risk managers. Stochastic claims reserving models achieve this goal by providing a best estimate as well as a variability measure of reserves; for example, the conditional mean-squared prediction error. This paper focuses on testing the predictive distribution of outstanding claims. We emphasize that a fair test should be based on a retrospective evaluation using the realized claims of predictive interests. In this study, the retrospective test will be performed at two levels: individual firm and portfolio of insurers. This section focuses on the tests for single insurers and the next section performs tests for multiple insurers.

At firm level, the retrospective test informs actuaries on the predictive performance of a stochastic claims reserving method for each individual firm. For a specific insurer, we calculate the percentile of realized unpaid losses $x_{AY,Lag}$ for each cell (AY,Lag) in the unobserved triangle, by $p_{AY,Lag} = F(x_{AY,Lag})$, where $F(\cdot)$ denotes the predictive distribution of $X_{AY,Lag}$ derived from a certain stochastic reserving method. All these $p_{AY,Lag}$ ($AY + Lag > 11$) are expected to be a random sample of a uniformly distributed variable on $[0, 1]$, if the model assumptions of the stochastic reserving method are appropriate for the insurer. The uniformity of percentiles could be visualized through graphical tools such as Probability-Probability (PP) plot, or could be easily tested using formal statistics such as a Kolmogorov-Smirnov (KS) test.

We perform the retrospective test for all the insurers in our sample individually. With the BAT model, we observe that for all the insurers, the PP plots for the training data lie within the KS bounds. It was with the test data that the PP plots often deviated outside the KS bounds. The results for the BCL model are similar; i.e., the model fits data well but could produce bad predictions. We demonstrate these analyses with three insurers. The group code for the three insurers are 914, 2674 and 310. We present the following results from the BCL and the BAT model for each insurer: 1) A PP-plot for training data; 2) The percentiles of training data for accident year, settlement lag as well as calendar year; 3) A PP-plot for test data; 4) The

The Retrospective Testing of Stochastic Loss Reserve Models

percentiles of test data for accident year, settlement lag, as well as calendar year. If the model fits well, we should expect the PP-plot to lie along the 45° line, and to see no pattern in the remaining plots by accident year, settlement lag or calendar year. The results are summarized in Figures 15 – 26.

In terms of goodness-of-fit, the PP-plots of training data suggest that both BCL and BAT models fit training data well for all insurers. When examining the test data, the retrospective test shows that the PP plots of both models are within the KS bounds for insurer 914, but outside the KS bounds for insurer 310. For insurer 2764, the BCL model provides better predictive distribution than the BAT model. We attribute such observations to the potential overfitting of the two loss reserving models. Though not reported here, our analysis showed that the loss development of insurer 914 is rather stable over time, while the payments for insurer 2764 and 310 are more volatile from year to year, especially for insurer 310. The higher variability explains the poor predictive performance of both models on insurer 310. Another factor affecting the predictive performance of loss reserving models appears to be an environmental change in the projecting period. Our analysis in the next section shows that the BCL model somehow did a better job in the perceived changing environment.

Figure 15 – BCL Model for Insurer 914 – Training Data

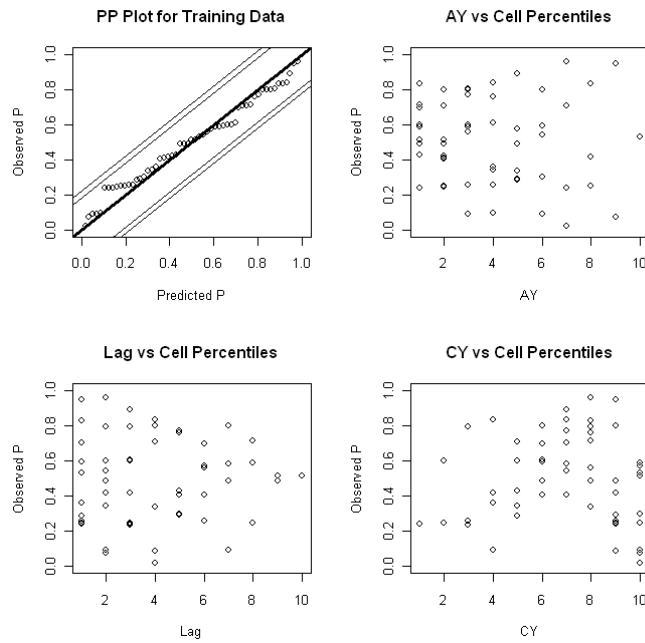


Figure 16 – BCL Model for Insurer 914 – Test Data

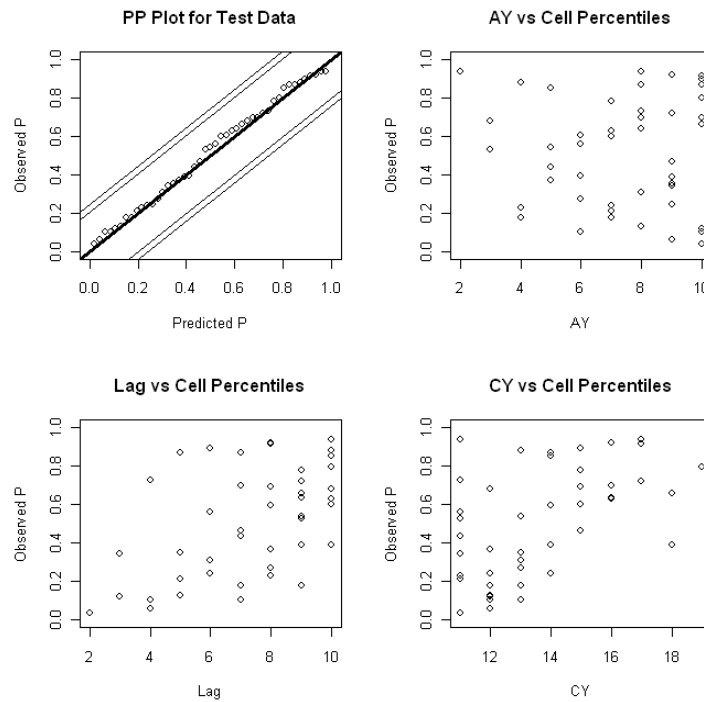


Figure 17 – BAT Model for Insurer 914 – Training Data

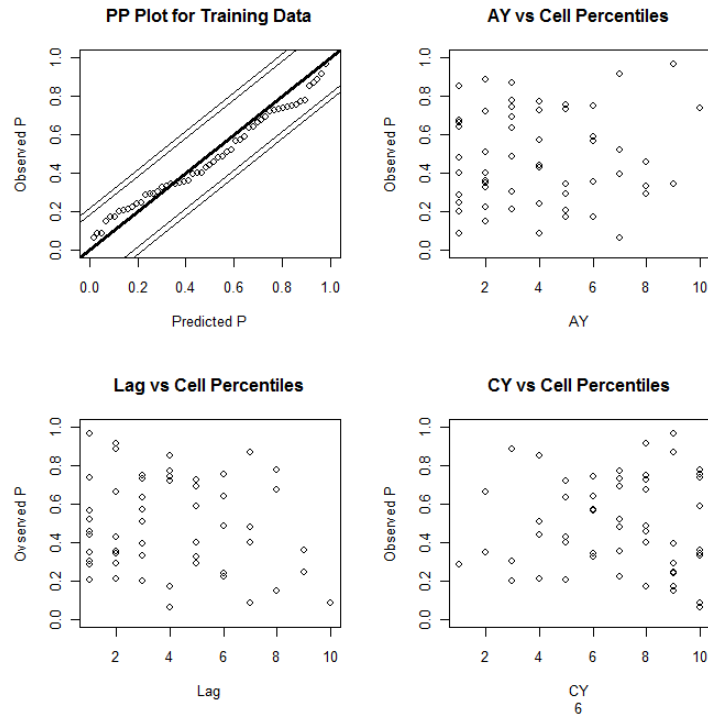
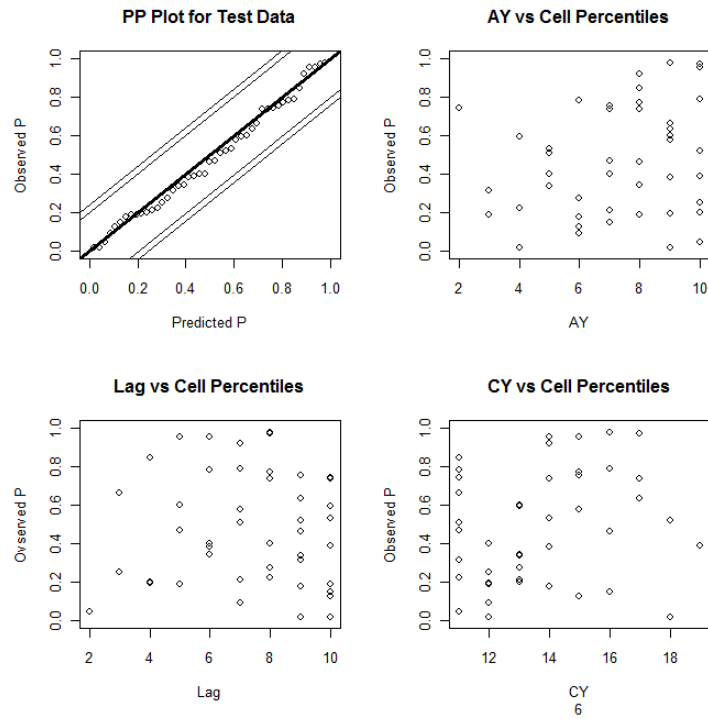


Figure 18 – BAT Model for Insurer 914 – Test Data



The Retrospective Testing of Stochastic Loss Reserve Models

Figure 19 – BCL Model for Insurer 2674 – Training Data

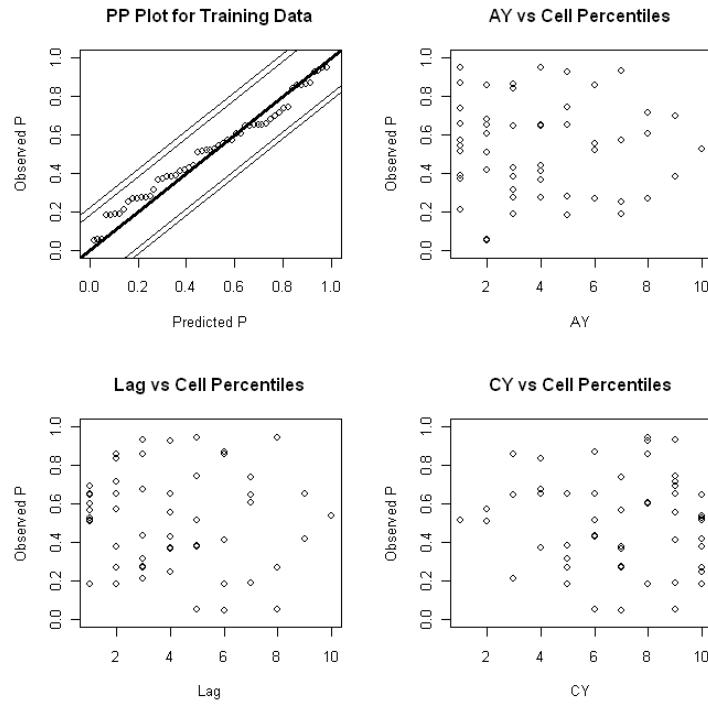


Figure 20 – BCL Model for Insurer 2674 – Test Data

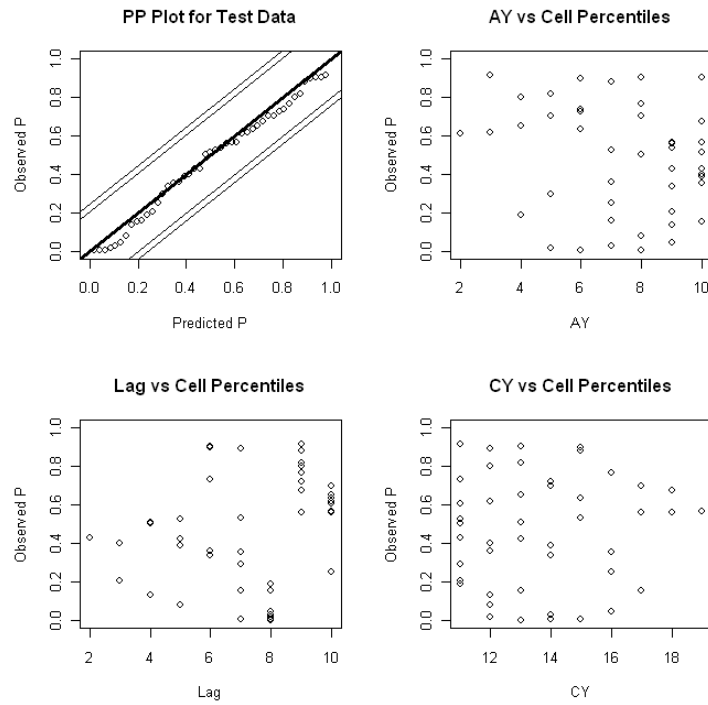


Figure 21 – BAT Model for Insurer 2674 – Training Data

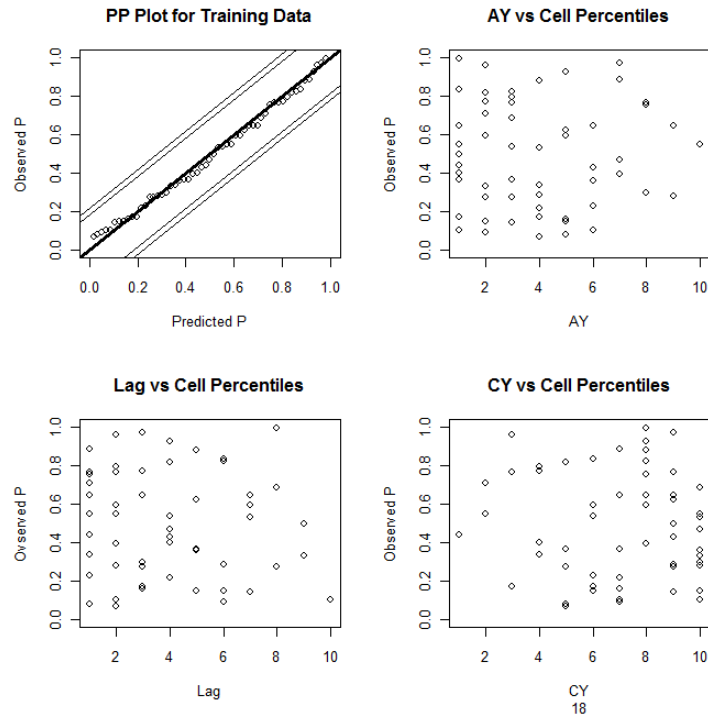


Figure 22 – BAT Model for Insurer 2674 – Test Data

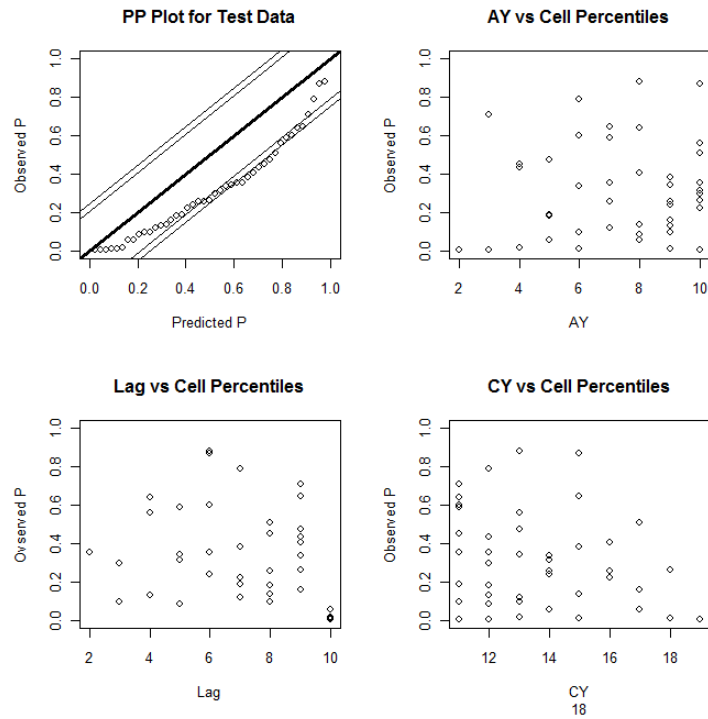


Figure 23 – BCL Model for Insurer 310 – Training Data

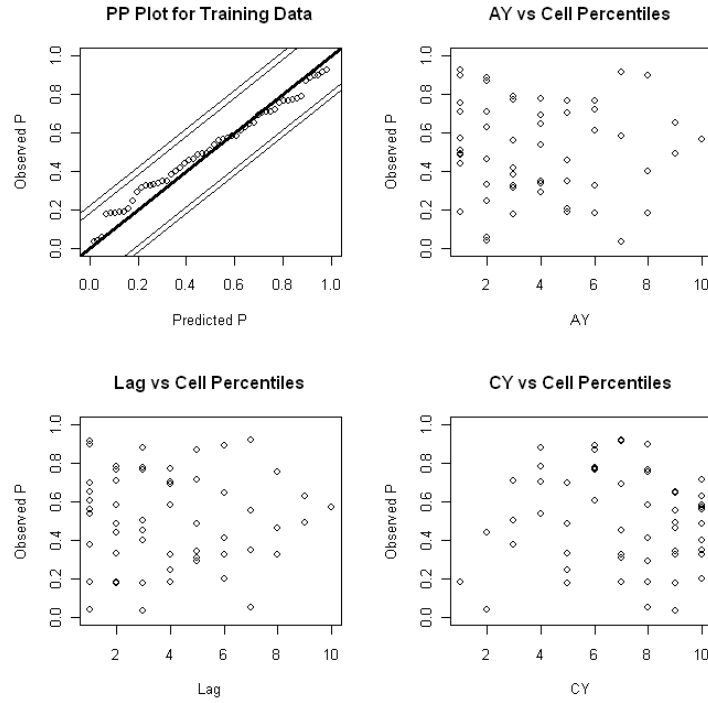


Figure 24 – BCL Model for Insurer 310 – Test Data

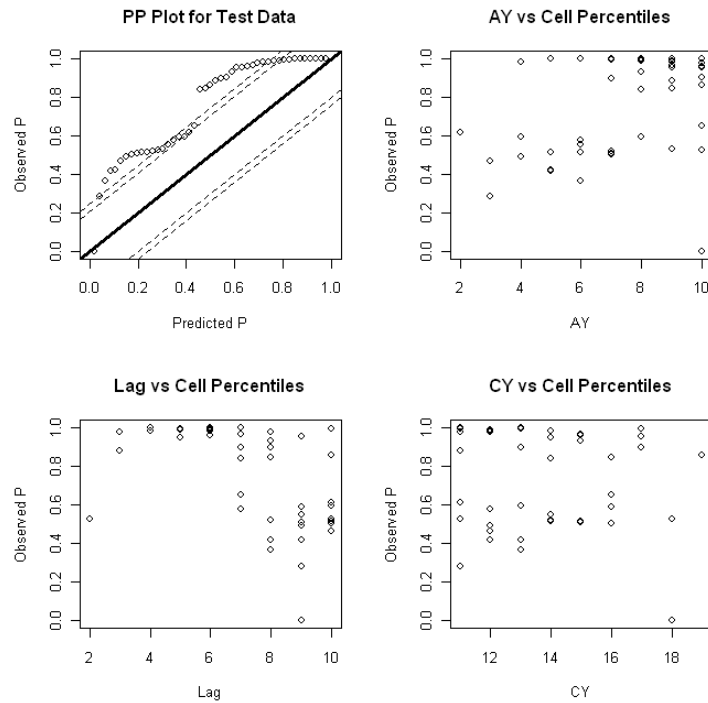


Figure 25 – BAT Model for Insurer 310 – Training Data

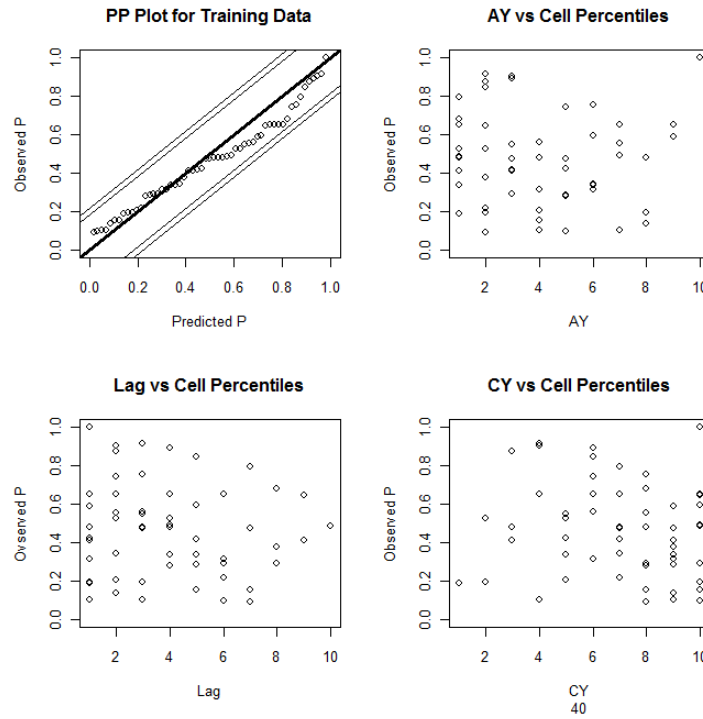
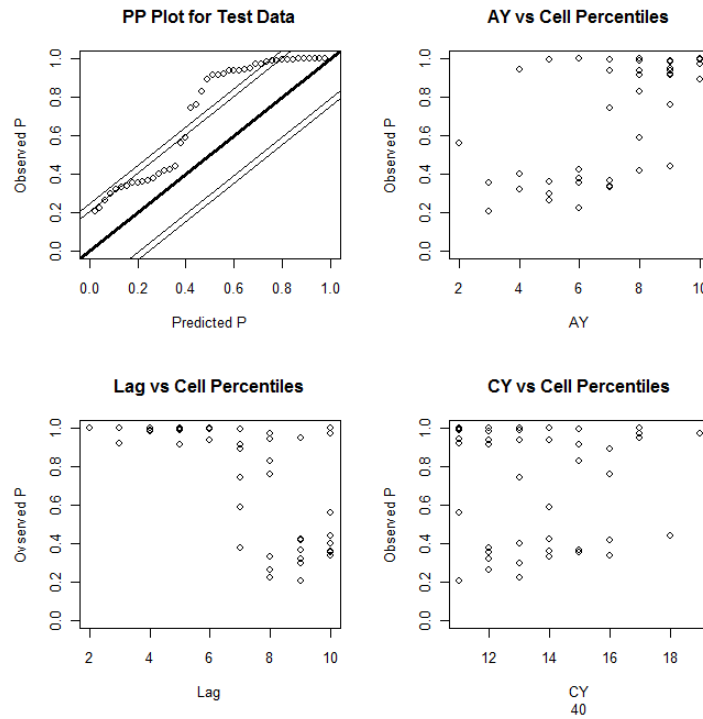


Figure 26 – BAT Model for Insurer 310 – Test Data



5. Retrospective Tests for Multiple Insurers

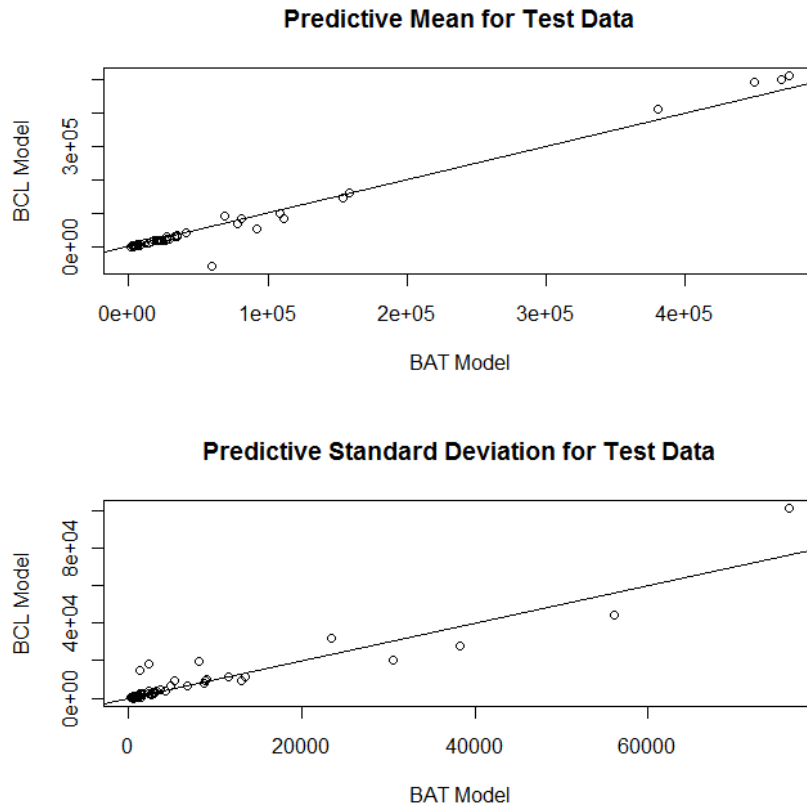
The retrospective test could be performed for a portfolio of insurers as well. At portfolio level, the retrospective test helps detect the potential under or over reserving issue if one single stochastic method is applied to all insurers in the portfolio. The same idea could be generalized to the industry level. Considering a portfolio of N property-casualty insurers, we implement the test using total reserves. Specifically, for the k^{th} ($k = 1, \dots, N$) insurer in the portfolio, we calculate the percentiles of realized total unpaid losses in the lower triangle $p_k^{\text{Tot}} = F(r_k^{\text{Tot}})$. Here $F(\cdot)$ and r_k indicate the corresponding predictive distribution and realized unpaid losses, respectively. Whether the stochastic reserving method is suitable for the insurer portfolio could be answered by examining the uniformity of p_k^{Tot} .

This section compares the predictions of the Bayesian Autoregressive Tweedie (BAT) model and the Bootstrap Chain Ladder (BCL) model. Our data also includes the reserve that each insurer posted in the 1997 Annual Statement. The reserves posted by the insurer differ from the models in that they are not tied to any particular method or model and can reflect insurer judgment. Also, it is not difficult to imagine the various incentives that can influence the judgments in either direction.

Figure 27 compares the predictive means and standard deviations of the total outstanding losses using the BAT and BCL methods. This figure indicates that for the most part, the predictive means are fairly close⁵. There are a noticeable number of instances where the predictive standard deviation is smaller for the BAT model.

⁵ In one case the mechanical application of the BCL model produced a negative mean because of a negative incremental paid loss. Any actuary would reject this result, in practice. The BAT model dropped any cell that contained a negative incremental paid loss.

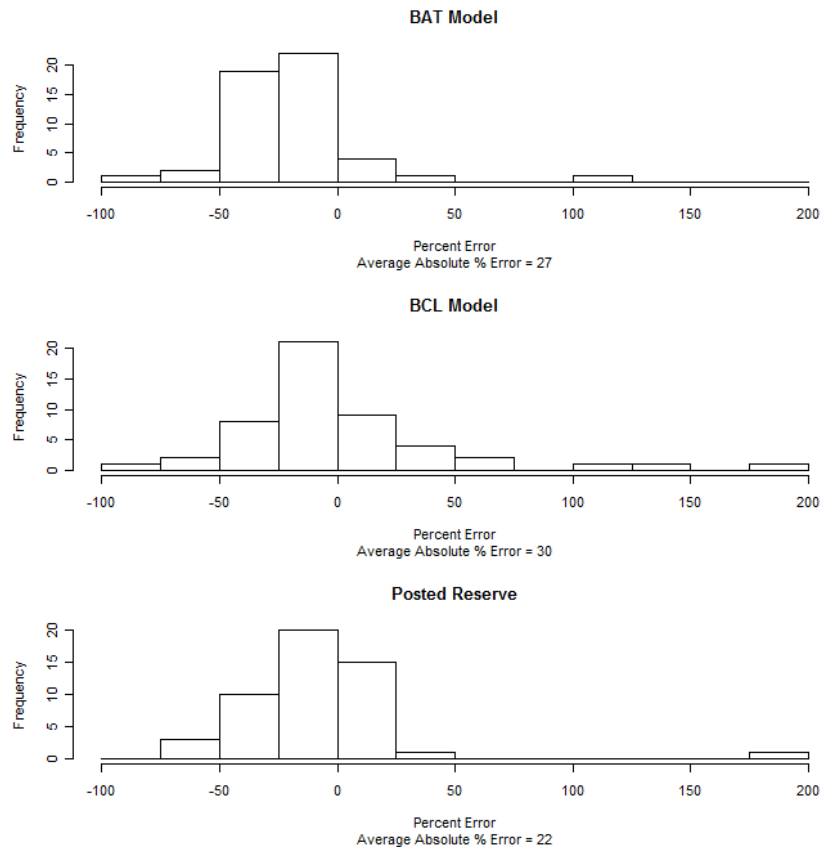
Figure 27



Next, we compare the accuracy of the predictions of the BAT and BCL models with the posted reserves. For both models, we use the predictive means for the test data. Figure 28 compares the percentage error of the three predictions⁶ from the actual outcomes. The mean absolute percentage error was largest for the BCL model, and smallest for the posted reserve. It is worth noting that in most cases, all three estimates predicted losses that were high. It is also worth noting that a previous study of this sort on different data (Meyers 2007c) found that a Bayesian model produced smaller errors than the posted reserve.

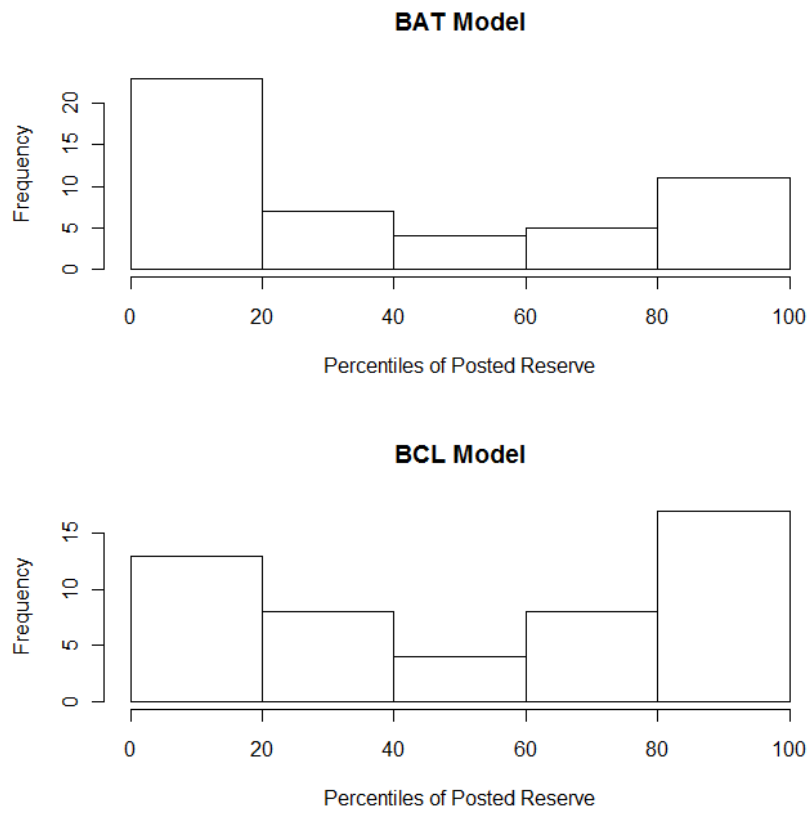
⁶ The BCL model produced one negative and one zero predicted mean. We set the percentage absolute percentage error at -100% and 200% respectively.

Figure 28



When a stochastic loss reserve analysis is performed, a question commonly asked by actuaries is “What percentile should one post a reserve.” While we do not intend to answer that question, we can use the BAT and the BCL models to estimate the percentiles of the actual posted reserve. Figure 29 provides the results. It appears that many insurers post conservative estimates, while many others (correctly as it turns out) posted lower than expected reserves.

Figure 29



The Retrospective Testing of Stochastic Loss Reserve Models

If a loss reserve model is appropriate for all insurers, the predicted percentiles of the data should be uniformly distributed. Figure 30 provides histograms for both models with the training data and Figure 31 provides histograms for both models on the test data. All four histograms indicate non-uniformity of the predicted percentiles. It should come as no surprise that the percentiles tend to be around the middle ranges on the training. Because of the high parameter to data point ratio, we attribute this to overfitting. We interpret the results for the test data as an indication that either: (1) something changed in the environment that resulted in lower claim settlements; or (2) no single model should be expected to apply for all insurers. It appears that, for whatever reason, the BCL did a better job of picking up that environmental change.

Figure 30

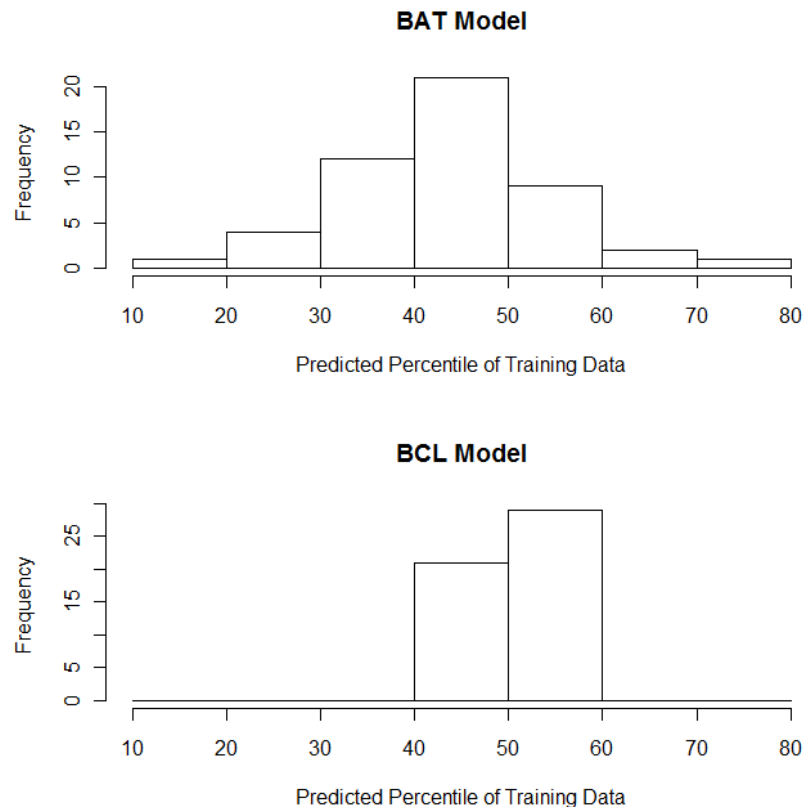
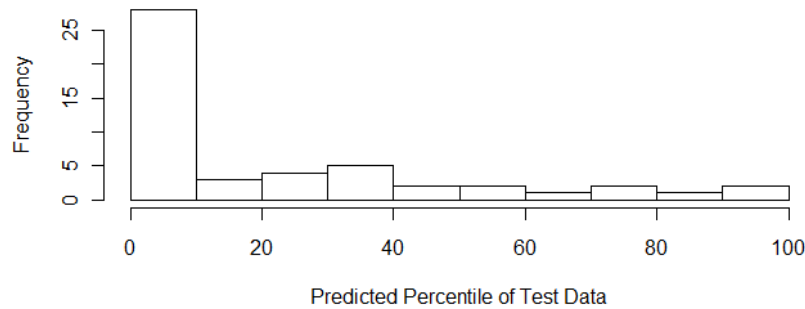
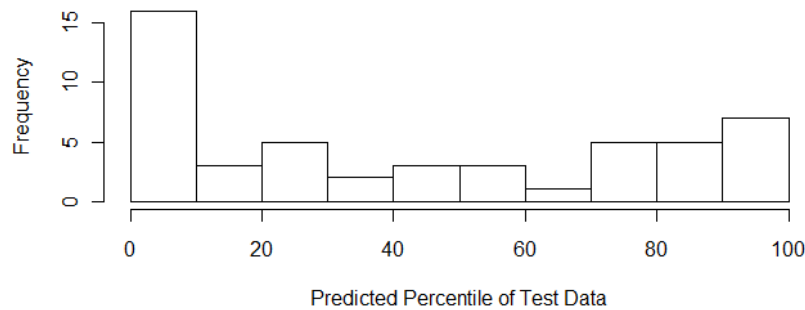


Figure 31

BAT Model



BCL Model



6. Concluding Remarks

The primary purpose of this paper was to introduce a new database that can be used to test predictive distributions from different stochastic loss reserve models. We emphasized the retrospective tests based on realized payments in the projecting periods. We then performed some tests on an established model, bootstrap chain ladder (BCL) model, and a proposed new model, Bayesian Autoregressive Tweedie (BAT) model. At this point in time, we are not ready to declare a winner. These models, and perhaps other models, should be tested on other lines of insurance. And the database is there that will permit further testing.

This particular study suggests that there might be environmental changes that no single model can identify. If this continues to hold, the actuarial profession cannot rely solely on stochastic loss reserve models to manage its reserve risk. We need to develop other risk management strategies that do deal with unforeseen environmental changes.

References:

- England, P. and R. Verrall (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal* 8 (3), 443-518.
- Kirschner, G., C. Kerley, and B. Isaacs (2008). Two approaches to calculating correlated reserve indications across multiple lines of business. *Variance* 2 (1), 15-38.
- Meyers, Glenn G. (2007a). Thinking Outside the Triangle. ASTIN Colloquium 2007. <http://www.actuaries.org/ASTIN/Colloquia/Orlando/Papers/Meyers.pdf>
- Meyers, Glenn G. (2007b). The Common Shock Model for Correlated Insurance Losses, *Variance* 1:1, pp. 40-52.
- Meyers, Glenn G. (2007c). Estimating Predictive Distributions for Loss Reserve Models, *Variance* 1:2, pp. 248-272.
- Meyers, Glenn G. (2009). Bayesian Analysis with the Metropolis Hastings Algorithm, *The Actuarial Review*, November 2009. <http://www.casact.org/newsletter/index.cfm?fa=viewart&id=5866>
- Renshaw, A. and R. Verrall (1998). A stochastic model underlying the chain-ladder technique. *British Actuarial Journal* 4 (4), 903-923.
- Shi, P., S. Basu, and G. G. Meyers (2011). A Bayesian Log-normal Model for Multivariate Loss Reserving. *Submitted for Publication*.
- Taylor, G. (2000). *Loss Reserving: An Actuarial Perspective*. Kluwer Academic Publishers.
- Verrall, R. (2000). An investigation into stochastic claims reserving models and the chain-ladder technique. *Insurance: Mathematics and Economics* 26 (1), 91-99.
- Wüthrich, M. and M. Merz (2008). *Stochastic Claims Reserving Methods in Insurance*. John Wiley & Sons.

Appendix

This appendix describes the data set of loss triangles that we prepared for claims reserving studies. The data cover major personal and commercial lines of business from U.S. property casualty insurers. We extract the claims data from Schedule P – Analysis of Losses and Loss Expenses in the National Association of Insurance Commissioners (NAIC) database.

A.1 Schedule P

NAIC Schedule P contains information on claims for major personal and commercial lines for all property-casualty insurers that write business in U.S. Some parts have sections that separate occurrence from claims made coverages. We focus on the following six lines: (1) private passenger auto liability/medical; (2) commercial auto/truck liability/medical; (3) worker's compensation; (4) medical malpractice – claims made; (5) other liability – occurrence; (6) product liability – occurrence.

For each of the above six lines, the variables to be included in the dataset are pulled from three different parts in Schedule P, including:

Part 1 - Earned premium and some summary loss data

Part 2 - Incurred net loss triangles

Part 3 - Paid net loss triangles

Part 4 - Bulk and IBNR Reserves

A.2 Data Preparation

The triangles consist of losses net of reinsurance, and quite often insurer groups have mutual reinsurance arrangements between the companies within the group. Consequently, we focus on records for single entities in the data preparation, be they insurer groups or true single insurers. The process of data preparation takes three steps:

Step I: Pull triangle data from Schedule P of year 1997. Each triangle includes claims of 10 accident years (1988-1997) and 10 development lags. This data was the training data used for model development.

The Retrospective Testing of Stochastic Loss Reserve Models

Step II: Square the triangles from Schedule P of year 1997 with outcomes from Schedule P of subsequent years. Specifically, the data for accident year 1989 was pulled from Schedule P of year 1998, the data for 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. The data in the lower triangles could be used for model validation purposes.

Step III: We performed a preliminary analysis to ensure the quality of the dataset. An insurer is retained in the final dataset if all following criteria are satisfied: (1) the insurer is available in both Schedule P of year 1997 and subsequent years; (2) the observations (10 accident years and 10 development lags) are complete for the insurer; (3) the claims from Schedule P of year 1997 match those from subsequent years.

A.3 Final Dataset

As a final product, we provide a dataset that contains run-off triangles of six lines of business for all U.S. property casualty insurers. The triangle data correspond to claims of accident year 1988 – 1997 with 10 years development lag. Both upper and lower triangles are included so that one could use the data to develop a model and then test its performance retrospectively. A list of variables in the data is as follows:

Table A.1. Description of Variables

Variable	Description
GRCODE	NAIC company code (including insurer groups and single insurers)
GRNAME	NAIC company name (including insurer groups and single insurers)
AccidentYear	Accident year (calendar year)
DevelopmentYear	Development year (calendar year)
DevelopmentLag	Development year - Incurral year + 1
IncurLoss_	Incurred losses and allocated expenses reported at year end
CumPaidLoss_	Cumulative and paid losses and allocated expenses at year end
EarnedPremD_	Premiums earned at incurral year - direct and assumed
EarnedPremC_	Premiums earned at incurral year - ceded
EarnedPremN_	Premiums earned at incurral year - net
Single	1 indicates a single entity, 0 indicates a group insurer
"_"	Refers to lines of business
B	Private passenger auto liability/medical
C	Commercial auto/truck liability/medical
D	Workers' compensation
F2	Medical malpractice - Claims made
H1	Other liability - Occurrence
R1	Products liability - Occurrence