Parameterizing the Loss Simulation Model

Ball State University Research Course

CAS Loss Simulation Model Working Party

Abstract:

Motivation. The Loss Simulation Model Working Party ("LSMWP") of the Casualty Actuarial Society has developed a prototype Loss Simulation Model. The students from a Ball State University Research Class in Actuarial Science parameterize this model using claim transaction data.

Method. The students use statistical models to fit the data to select distributions, emphasizing the methods covered in the material for Exam 4 from the Spring 2007 CAS Exam Syllabus. The predominant method uses spreadsheet software to construct and maximize the loglikelihood function. A second method uses parametric survival models from the statistical software package "R."

Results. This project provides practical methods for parameterizing report lag, settlement lag, and claim size distributions for the Loss Simulation Model or other predictive models. These methods are accessible even if the modeler does not have access to specialized statistical software. This paper provides methods for determining how model parameters may vary with covariates. Existing statistical methods (e.g., parametric survival models) are suitable for modeling these variables with covariates.

Conclusions. The types of distributions that the Simulation model provides are very useful for modeling lags and claim size on the automobile collision and bodily injury data provided. Based on examining state and loss year as covariates, the model should have the flexibility to allow the parameters to depend on covariates. There is a limited capability in the current model to allow this.

This project demonstrates how the academic community and the actuarial profession can collaborate on research.

Availability. Both Microsoft Excel and the R statistical package were used in the investigation. The R package is free and can be downloaded by accessing the Internet address "http://cran.r-project.org/"..

Keywords. Simulation, Generalized Linear Modeling, Survival Model, Extreme Value, Severity, Personal Automobile, Parameterize.

1. INTRODUCTION

Actuaries can now simulate property-casualty insurance claims and test these models on large data sets due to today's extremely powerful computers. The Loss Simulation Model Working Party ("LSMWP") has developed a prototype Simulation Model (the "Simulator") to make available to its members. A vital step in the modeling process is to parameterize the model using actual insurance data. This is labor-intensive because of the large size of the data sets, the need to manipulate the data to a format usable by software packages, and the sheer volume of analysis. The Ball State University students enrolled in the Spring 2007 Actuarial Science Research Course (the "Research Course") parameterized the model using claim transaction data supplied by an anonymous "Source."

1.1 Research Context

This is a practical paper with two primary purposes:

1) develop methods to determine parameters for the Loss Simulation model, and

2) apply these methods to a set of actual insurance data.

The source data consists of Personal Automobile Bodily Injury and Collision claim transactions. The modeling uses both the free statistical software "R" and spreadsheet-type software Microsoft Excel. The researchers found suitable modeling methods, some developed outside of insurance. For example, parametric survival models are very useful in modeling "time" variables such as report lag and settlement lag. The medical research field employs these models extensively, as do researchers studying machine failure times and Life Insurance actuaries. This paper references work from these fields.

The research focuses on finding models for lag and claim size that follow the distributions programmed into the current Simulation Model. For modeling lags, these are the Weibull, exponential, and lognormal distributions. The Simulator uses the Weibull, lognormal, and Pareto distributions for generating claim sizes. Note that the Weibull and lognormal distribution are used for modeling both lags and claim size. Therefore, many of the techniques for modeling lags apply to modeling claim size.

Whenever possible, this paper references material from the first four CAS Actuarial Examinations.

The research seeks suitable covariates for the models. Obvious examples include state[1] and accident year.

1.2 Objective

The CAS Loss Simulation Model Working Party is developing a Loss Simulator for practicing actuaries. This is the first time that the Simulator has been parameterized using a body of actual insurance data.

Much of the published research assumes that the practitioner has extensive knowledge of statistical modeling, possesses sophisticated statistical software, and/or has a sufficiently powerful computer. This paper addresses the needs of those who may lack one or more of these tools. The research uses spreadsheet software for much of the modeling. The techniques presented are suitable for the university environment, where students may not know a powerful statistical language.

The language "R" was used for the portion of the research using a statistical package. This language is becoming very popular in actuarial research and can be downloaded free of charge.

1.3 Disclaimer

None.

1.4 Outline

The paper proceeds as follows. Section 2 discusses the background, scope and methods for each step in the modeling process. Subsections discuss specific topics:

2.1. Terminology.

2.2. Univariate modeling of lags and claim size using covariates

[1] The term "state" refers to jurisdictions within the United States.

2.3. Correlation between Settlement Lag and Claim Size for Auto BI

2.4. Zero modification of the Claim Size distribution.

2.5. Effect of Deductibles on Collision Losses; Pareto Model

2.6. Interaction of Report Lag and Settlement Lag

2.7. Data Processing

Section 3 presents the results. Section 4 presents the conclusions.

Appendices provide additional details. Appendix A discusses criteria for comparing models. Appendix B discusses the effects of censoring and truncation in the modeling. Appendix C discusses the grouping of data for use in a spreadsheet model. Appendix D shows an example of a model coded in the "R" programming language. Appendix E discusses the results of analyzing univariate distributions of report lag, settlement lag, and claim size using spreadsheet software. Appendix E also discusses how to construct the models. Appendix F contains the Figures in the paper.

2. BACKGROUND, SCOPE, AND METHODS

The "Source" supplied Personal Automobile claim transaction data for the years 1992-2006. This data includes all payment and reserve change transactions. Separate files contain the claim transactions for initial reserve values, reserve modifications, and payments.[2]

This project analyzes the report lag, settlement rate and severity for Auto Bodily Injury Liability and Collision.

<u>Limitations of Scope:</u>

This research fits models for report lag, settlement lag, and claim size, including interactions. This is not the complete set of variables used in the Simulation Model. The claim transaction amounts were summarized to a "life-to-date", or cumulative, basis as of the

[2] More precisely, each of the three transaction types are contained in separate "tables" within a Microsoft Access database.

latest valuation date, 12/31/2006. This research does not therefore include loss development or the timing of multiple claim payments. At the time this research was performed, the premium and exposure data had not been built. Therefore, the research does not address the frequency of claims, which is a variable included in the simulation model.

2.1 Terminology

- Bivariate normal distribution[3] : A distribution with joint density function

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-q/2} \quad , where$$

$$q = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]$$

- Censored data: Data for which the values of a variable are limited by a certain value.

The limiting can be from "above" (or "right"), where the indicated value is a maximum; or from "below" (or "left"), where the indicated value is a minimum. Almost all censoring in insurance loss data is from the right. The "limit" can vary with each observation.[4] As an example, liability claim size is typically right-censored at the policy limit.

- Correlation coefficient: measures how two variables depend on each other.
- Covariate: An independent variable, or predictor, in a regression equation.
- Exponential distribution[5] (with parameter θ):

A distribution with density $f(x) = \frac{e^{-x/\theta}}{\theta}$ for $x > 0$.

[3] See Hogg, McKean, and Craig [2], p. 174

[4] This definition along with Standard deviation, Nested model, Truncation, Covariates, and PP plot are taken from Klugman, Panjer, and Willmot [1], pp. 27, 297, 405-406, 424-425.

[5] This definition, along with those for the lognormal, normal, Pareto, and Weibull distributions are taken from Klugman, Panjer, and Willmot [1], Appendix A.

- Extreme (Minimum) Value or Least Extreme Value distribution (μ , σ)[6] :

A distribution with $F(y) = 1 - exp\ (- exp\ (z))$ and

$$f(y) = \sigma^{-1}\ exp\ (z - exp\ (z)),\ \text{where } z = \frac{y-\mu}{\sigma}\ .$$

An important relationship is that if X is Weibull (θ , τ) , then $Y = ln\ (X)$ follows an Extreme Value distribution with $\mu = \ln(\theta)$ and $\sigma = \tau^{-1}$.

- Lognormal distribution (μ, σ):

A distribution with $f(x) = \dfrac{1}{x\sigma\sqrt{2\pi}}\exp(-z^2/2)$, where $z=\dfrac{\ln x-\mu}{\sigma}$, $x > 0$.

- Nested model: a model that is a "subset" or "special case" of another. For example, a model with fewer covariates is nested within a model with more covariates. An exponential model is a special case of a Weibull model, with $\tau = 1$.

- No-Fault system: A modification of the traditional American legal system whereby the injured party's ability to sue for damages is curtailed in exchange for an ability to receive compensation for injuries without proving the other party was "at fault."

- Normal distribution (μ, σ):

A distribution with $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}}\exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$

- Pareto distribution (α , θ) :

A distribution with $f(x) = \dfrac{\alpha\theta^{\alpha-1}}{(x+\theta)^{\alpha+1}}$

- PP plot ("percent-percent" plot): A plot of the ordered pairs <$G(x)$, $F(x)$> from two cumulative distribution functions. This plot fits into the unit square and is a 45-degree

[6] Tableman and Kim [3], p 58

line if the two distribution functions are equal. In a common application, $F(x)$ is a "fitted" distribution from a model and $G(x) = F_n(x)$ is the empirical distribution of actual values. The model fits well if the graph is close to the 45-degree line connecting <0,0> and <1,1>.

- Report Lag: number of days from the incident to the report date of a claim.

- Settlement Lag: number of days from the report date to the final payment of a claim.

- Standard extreme value distribution: An extreme (minimum) value distribution with $\mu = 0$ and $\sigma = 1$. If a random variable Y has this distribution, then $X = exp(Y)$ has an exponential distribution with $\theta = 1$.

- Truncation: a modification of the data where all values that do not satisfy a certain preset requirement are not included in the data

- Weibull distribution (θ , τ) :

A distribution with $f(x) = \dfrac{\tau (x/\theta)^{\tau} e^{-(x/\theta)^{\tau}}}{x}$ and $F(x) = 1 - exp\left(-(x/\theta)^{\tau}\right)$.

2.2 Using Covariates for Initial Modeling.

The Loss Simulator accepts the following distributions for the lags and claim size:

1) Weibull, exponential, and lognormal for Report Lag and Settlement Lag.
2) Weibull, lognormal, and Pareto for Claim Size.

The Simulator also allows for "zero-modification" of each distribution. A zero-modified distribution is a mixture of an initial distribution and a distribution with all its probability at zero.[7]

A major question in modeling is whether the parameters of a hypothesized distribution vary with a set of covariates (also called predictors). For example, GLMs address whether a

[7] Klugman, Panjer, and Willmot [1], p. 85

function of the mean μ varies linearly with a set of covariates. Parametric survival models handle the situation where the distribution acts like a time variable and may be censored or truncated.

The Loss Simulation Model allows the modeler to vary parameter values on distinct subsets of the data. The modeler can, with enough work, define the subsets by unique values of the predictors and thereby reproduce distributions whose parameters are determined by GLMs, Survival Models, or other methods.

The researchers investigated state and loss year as possible covariates in studying the univariate distribution of the lags and claim sizes. State is a natural covariate because the legal system under which automobile accident damages are determined varies by state. The coefficient of loss year measures trend over time.

There were two main goals for the initial modeling:

determine the best way to categorize states and whether to use loss year (i.e., accident year) as an additional covariate; and

using the covariates from the first step, determine the best parametric model to use from among those supported by the CAS Loss Simulator[8].

Nine states are represented in the data. The largest states are New Jersey and Pennsylvania, which are both "No-Fault" states. Four groupings of state were investigated, ranging from not using state at all to considering each state as its own category. For each grouping, modeling was performed with and without loss year (numeric value of loss year).

The data for studying report lag and settlement lag included loss years 1996 through 2005 only. The year 2006 was excluded because many of its settlement lag values are censored by the valuation date 12/31/2006. Loss years 1992 through 2006 were used for

[8] With statistical software, one can try all combinations of covariates (8) with all model forms (3) for each coverage (2), a total of 48 models in all for each variable. This was not practical for the initial models, which were programmed using Excel spreadsheets. Using spreadsheets, the initial modeling first selected the optimal covariates using one model (8 runs), then selected from among model types (3 runs). For two coverages, the total number of model runs was 22.

claim size modeling.

In calculating maximum likelihood parameter estimates, truncation and censoring apply to some of the variables of interest. Settlement lag for claims open as of 12/31/2006 is censored with a value equal to the difference in days between 12/31/2006 and the report date. For Auto BI, the claim size is censored at the policy limit. For Collision, the claim size is left-shifted and truncated by the amount of the deductible.[9]

This modeling for the lags uses only records where the lag > 0, while the model for claim size uses only records where incurred amount > 0. Section 2.4 describes how to model the probability that the claim size equals zero.

The methods developed can apply to covariates other than state and loss year.

The following subsections discuss the findings from the modeling of each of the lag and claim size variables. Appendix E describes in more detail the modeling and findings from the spreadsheet models. The researchers also ran a model in "R" using the covariates and model types selected from the spreadsheet runs.

In all cases but one, the model using loss year and each individual state as covariates was selected. The only exception was Report Lag for Collision, where the loss year was dropped.

2.2.1 Report Lag for Auto BI

The table below shows the cumulative distribution of report lag in days**.**

Lag	0	5	10	15	90	180	365	Total
Cum #	9,007	28,062	29,623	30,419	32,535	32,902	33,102	33,375
Cumulative Distribution	27.0%	84.1%	88.8%	91.1%	97.5%	98.6%	99.2%	100.0%

[9] See Klugman, Panjer, and Willmot [1], section 11.1, p. 297 for a definition of these terms.

The mean report lag is 12 days and the standard deviation is 65 days. The modeling data includes loss years through 2005, with the claims evaluated as of 12/31/2006. The report lag is right-truncated only if it exceeds the difference in days between the valuation date and the loss date. Since 2005 is the latest loss year used, the smallest possible value of this difference is 365 days. Since over 99% of the claims are reported within one year of the loss date, the possible effect of truncation can be ignored in the modeling. This simplifies the modeling.

One cannot ignore the effect of truncation for a line of business where the report lag is long, such as Medical Malpractice, and where the model uses data from the most recent loss year.

The final spreadsheet model for Report Lag is lognormal with parameters μ and σ = 1.3508. The linear predictors of μ and their coefficients follow:

1	X_{CT}	X_{DE}	X_{KY}	X_{ME}	X_{MD}	X_{OH}	X_{NJ}	X_{PA}	Loss year
0.773	0.282	0.052	1.000	0.071	0.117	0.508	0.617	0.253	-0.02039

The i^{th} fitted value μ_i is calculated $\mu_i = \sum_j X_{ij} b_j$, where the variables X are the ones listed in the top line of the table, and the coefficients b are listed in the second line.

For the fitted model, σ = 1.3508. This model has a total loglikelihood (*ln L*) of -41,904 with AIC = 83,828.

The lognormal model is superior to the Weibull model. The best Weibull model has *ln L* = -49,221 with AIC = 98,462.

As a check, the following parametric survival model in "R" fit the log of the Report Lag to an Extreme Value distribution:

```
temp1 <- Surv(log(data1$report.lag),data1$report.event)

data1$modyr <- data1$lossyear - 1995
```

```
Call:
survreg(formula = temp1 ~ factor(NUM.ST.CD) + modyr, data = data1,
    weights = claim.count, dist = "extreme", x = T, y = T)
```

```
Coefficients:
(Intercept)     factor(NUM.ST.CD)7 factor(NUM.ST.CD)18 factor(NUM.ST.CD)19
 1.350108745         0.126497948         0.006884314         0.140446756
factor(NUM.ST.CD)29 factor(NUM.ST.CD)34 factor(NUM.ST.CD)37 factor(NUM.ST.CD)45
 0.889014350         0.783759088         0.495624055        -0.131643763
 modyr
 -0.025849189

Scale= 1.821328

Loglik(model)= -49221.1   Loglik(intercept only)= -49550.4
        Chisq= 658.62 on 8 degrees of freedom, p= 0
n= 2097
```

A parametric survival model in “R” using the “survreg” command requires a response variable of type “surv,” generated using the function “surv”. The “surv” function requires two vectors as input:

the response variable of interest, and

a corresponding “event” vector that equals 0 if the response variable is censored and equals 1 if the response the variable is uncensored.

In this case, the response variable is *ln* (Report Lag) and the event variable is always 1, since no report lags are censored.

The loglikelihood, scale σ, and AIC match the spreadsheet fit to the Weibull distribution. The coefficients do not appear to match. However, this is mainly because the spreadsheet run uses state indicators for all states except VA, while the R run eliminates the indicator for CT[10].

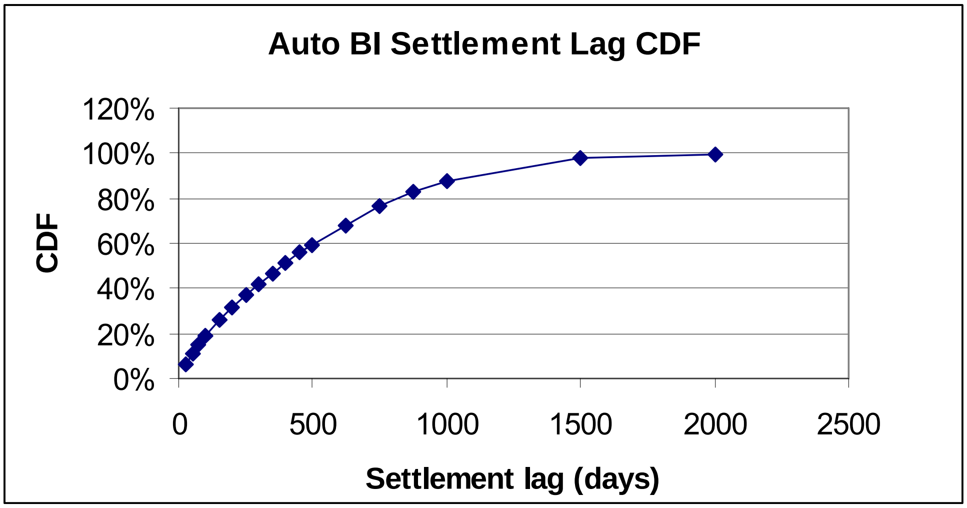
The PP Plot in Figure 1 shows that the fit is poor.

This is not an interesting variable to model, as 84% of the Report Lags are five days or less.

[10] In modeling a categorical variable with *n* categories, only *n*-1 indicator variables are used in the model.

2.2.2 Settlement Lag for Auto BI

Auto BI Settlement Lag has a mean of 485 days and a standard deviation of 418 days. The graph below shows the cumulative distribution. The censored observations (for those claims that have not settled by 12/31/2006) are recorded at their censored values. Therefore, the settlement lag will have a slightly higher mean when all claims are settled. The modeling needs to consider this censoring effect, since 1,306 of the 33,375 Auto BI claims were not settled by 12/31/2006.



The final spreadsheet model is Weibull with parameters σ =0.932492 and **μ**. The linear predictors of **μ** and their coefficients follow:

1	X_{CT}	X_{DE}	X_{KY}	X_{ME}	X_{MD}	X_{OH}	X_{NJ}	X_{PA}	Loss year
5.914	0.395	0.486	1	-0.352	-0.050	0.143	0.741	0.513	-0.04521

This model has a loglikelihood (ln *L*) of -51,211 with AIC = 102,442.

The parametric survival model in R produced the same results. Settlement lag is censored because the latest data is as of 12/31/2006. In this model, the loss year used is the actual loss year minus 1995. The PP Plot and Residual Plot in Figures 3 and 4 suggest that this model works well.

2.2.3 Report Lag for Collision.

The table below shows the cumulative distribution of Collision report lag in days.

Lag	0	5	10	15	90	180	365	Total
Cum #	42,147	113,983	120,081	122,731	125,724	125,790	126,104	126,200
Cumulative Distribution	33.4%	90.3%	95.2%	97.3%	99.6%	99.7%	99.9%	100.0%

The mean report lag is 3.6 days and the standard deviation is 22 days. The modeling data includes loss years through 2005, with the data as of 12/31/2006. Over 99% of the claims are reported within 90 days of the loss date. For the same reasons as given for Auto BI report lag, the effect of truncation can be ignored in the modeling.

The final spreadsheet model is lognormal with parameters σ = 1.0949 and μ. The linear predictors of μ and their coefficients follow:

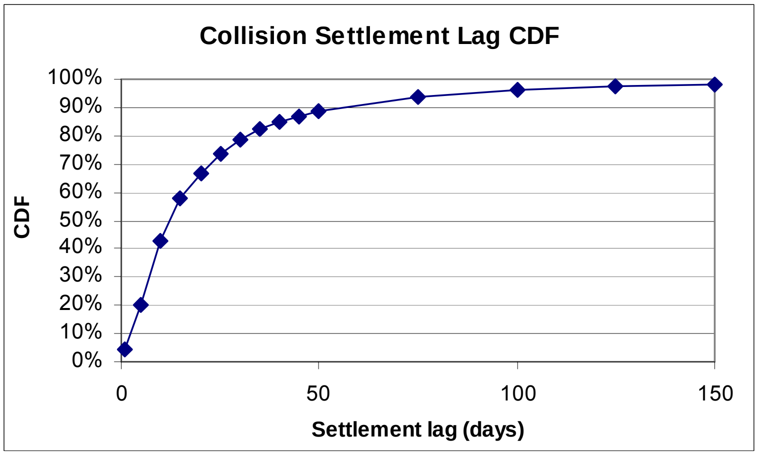
1	X_{CT}	X_{DE}	X_{KY}	X_{ME}	X_{MD}	X_{OH}	X_{NJ}	X_{PA}	Loss year
0.916	-0.097	-0.080	1.000	0.272	-0.140	-0.043	-0.006	-0.064	n/a

This model has loglikelihood (*ln L*) of -131,466, with AIC = 262,950. The loglikelihood is virtually unchanged by the inclusion of loss year as a predictor. Therefore, loss year is dropped from the final model.

The PP-Plot indicates a poor fit. This is not an interesting variable to study, as about 90% of the claims are reported within 5 days of the incident..

2.2.4 Settlement Lag for Collision.

The Collision Settlement Lag has a mean of 25 days and a standard deviation of 43 days. The graph below shows the cumulative distribution. The censored observations (for those claims that have not settled by 12/31/2006) are recorded at their censored values. Therefore, the settlement lag may have a slightly higher mean when all claims are settled. The censoring effect is negligible for collision because only two of the 129,233 claims were not settled by 12/31/2006.



The final spreadsheet model is lognormal with parameters σ =1.110642 and **μ**. The linear predictors of **μ** and their coefficients follow:

1	X_{CT}	X_{DE}	X_{KY}	X_{ME}	X_{MD}	X_{OH}	X_{NJ}	X_{PA}	Loss year
2.965	-0.091	-0.113	1.000	-0.069	-0.278	-0.482	-0.087	-0.148	-0.04405

This model has loglikelihood = -194,502 with AIC = 389,024.

To verify the calculations, a survival model in R was run using the "extreme" (R terminology for Least Extreme Value) distribution. The loglikelihood of -206,393 matches that of spreadsheet Model 1 in Appendix E.

Settlement lag is censored because the latest data is as of 12/31/2006. In this model, the loss year used is the actual loss year minus 1995. The PP Plot and Residual Plot in Figures 7 and 8 suggest that this model works well.

The results of the R model (using the extreme value distribution) follow:

```
Call:
survreg(formula = temp1 ~ factor(NUM.ST.CD) + modyr, data = data1,
    weights = claim.count, dist = "extreme", x = T, y = T)

Coefficients
(Intercept)  factor(NUM.ST.CD)7 factor(NUM.ST.CD)18 factor(NUM.ST.CD)19
 3.10112379          0.41943563          0.01985678          0.25002789
factor(NUM.ST.CD)29 factor(NUM.ST.CD)34 factor(NUM.ST.CD)37 factor(NUM.ST.CD)45
0.37343553          0.34091344          0.32152650          0.44334618
```

```
Modyr
-0.05262430

Scale= 1.130781

Loglik(model)= -206369.3   Loglik(intercept only)= -207379.5
        Chisq= 2020.43 on 8 degrees of freedom, p= 0
n= 7158
```

2.2.5 Auto Bodily Injury Claim Size

The claim size distribution (given that a claim has non-zero value) follows:

Size range	Number of Claims	Distribution
1 - 499	1,259	4.6%
500 - 1499	2,392	8.7%
1500 - 2499	1,345	4.9%
2500 - 3499	1,496	5.4%
3500 - 4999	1,784	6.5%
5000 - 9499	6,281	22.8%
9500 - 24499	8,672	31.4%
24500 - 49999	2,610	9.5%
50000 - 99999	1,276	4.6%
100000 - 199999	439	1.6%
200000 - 299999	53	0.2%
over 299999	0	0.0%
Total	27,607	100.0%

Approximately 9% of the claim amounts are censored at the policy limit.

The spreadsheet model fit both the Weibull and lognormal distributions. The final spreadsheet model is lognormal with parameters σ = 1.451282 and μ. The linear predictors of μ and their coefficients follow:

1	X_{CT}	X_{DE}	X_{KY}	X_{ME}	X_{MD}	X_{OH}	X_{NJ}	X_{PA}	Loss year
8.018	1.321	0.418	0.000	0.148	0.153	-0.047	1.027	0.814	0.01915

This model has loglikelihood = -131,466, with AIC = 262,950. The loss year used is the actual loss year minus 1992.

There is a complication in comparing the results with the results from an R model. The claim sizes for this model are grouped according to the method described in Appendix C. This means that n_i claims in the size interval $(L_i, R_i]$ contribute the amount

$$n_i \ln\left[\ \Phi\left(\frac{\ln R_i - \mu}{\sigma}\right) - \Phi\left(\frac{\ln L_i - \mu}{\sigma}\right)\right] \text{ to the total loglikelihood.}$$

See Appendix B for an explanation.

In running the parametric survival R model, the event variable **amtincd.event** indicates whether the claim size is censored by the policy limit. Both the spreadsheet and R models calculate the loglikelihood the same way for censored amounts.

The R model uses the individual claim sizes rather than grouping them into intervals. This means that the loglikelihood for an individual (uncensored) claim size y_i equals

$$\ln \phi\left[\ \left(\frac{\ln y_i - \mu}{\sigma}\right)\right] - \ln(\sigma)$$

The results for the R model follow:

```
 temp1b= Surv ( log (AMT.INCD), amtincd.event)
data3$modyr = loss year – 1992
Call:
survreg(formula = temp1b ~ factor(NUM.ST.CD) + data3$modyr, data = data3,
    dist = "gaussian", x = T, y = T)

                           Value    Std. Error        z               p
(Intercept)                9.3340    0.07954     117.34         0.00e+00
factor(NUM.ST.CD)7  -0.9087    0.08818     -10.30         6.73e-25
factor(NUM.ST.CD)18 -1.1608    0.20075      -5.78         7.36e-09
factor(NUM.ST.CD)19 -1.1738    0.08562     -13.71         8.88e-43
factor(NUM.ST.CD)29 -0.2957    0.08142      -3.63         2.82e-04
factor(NUM.ST.CD)34 -1.3825    0.12111     -11.41         3.54e-30
factor(NUM.ST.CD)37 -0.5088    0.08139      -6.25         4.08e-10
factor(NUM.ST.CD)45 -1.3269    0.09132     -14.53         7.80e-48
data3$modyr                 0.0195    0.00217       8.95         3.60e-19
Log(scale)                  0.3793    0.00453      83.71         0.00e+00

Scale= 1.46
```

```
Gaussian distribution
Loglik(model)= -47703   Loglik(intercept only)= -48368.9
        Chisq= 1331.75 on 8 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 3
n= 27607
```

The results of the R model are close to those from the spreadsheet. The loglikelihoods are quite different because the loglikelihood calculation for uncensored observations in the spreadsheet uses the cdf, while those in the R model use the density function. The implied annual severity trend is *exp*(.0195) = + 1.97% annually. Figures 9 and 10 show that the model fits fairly well.

2.2.6 Collision Claim Size.

The claim size distribution (given that a claim has non-zero value) follows:

Size range	Number of Claims	Distribution
1 - 149	2,208	1.8%
150 - 249	2,354	2.0%
250 - 349	2,860	2.4%
350 - 549	6,279	5.2%
550 - 999	13,430	11.2%
1000 - 1999	27,861	23.3%
2000 - 2999	18,875	15.8%
3000 - 4999	21,579	18.0%
5000 - 9999	16,694	13.9%
10000 - 25549	7,314	6.1%
over 25500	307	0.3%
Total	119,761	100.0%

The spreadsheet model fit both the Weibull and lognormal distributions. The final spreadsheet model is lognormal with parameters $\sigma = 0.87887$ and μ. The linear predictors of μ and their coefficients follow:

1	X_{CT}	X_{DE}	X_{KY}	X_{ME}	X_{MD}	X_{OH}	X_{NJ}	X_{PA}	Loss year
-43.061	0.252	-0.010	0.118	-0.262	0.061	0.017	0.277	0.226	0.02534

The loglikelihood is -521,780 with AIC = 1,043,581.

The collision loss amounts are grouped as described in Appendix C. Adding the deductible *d* to each grouped loss produces the grouped "ground up" loss and the grouping interval, as described in section 2.5. The likelihood function must account for the fact that ground up loss is truncated at *d*. The ground-up loss X equals the actual loss Y plus the deductible *d*. The likelihood that X lies in the interval (l , u] is given by $\dfrac{F(u|\theta)-F(l|\theta)}{1-F(d|\theta)}$, where *F* is the cdf of the ground up distribution and θ is the set of parameters over which the likelihood function is maximized.

Appendix E shows the detailed results for this modeling.

2.3 Correlation between Settlement Lag and Claim Size

The Simulator allows the modeler to input the correlation between settlement lag and claim size. This section discusses this correlation for Auto BI and provides a practical method to estimate the correlation using a bivariate normal distribution.

The following result makes it easy to calculate the correlation ϱ for a bivariate normal random vector (X , Y):[11]

$$Y \,\big|\, X{=}x \text{ is Normal with mean } \mu_y + \varrho \frac{\sigma_Y}{\sigma_X} (x - \mu_x) \text{ and variance } \sigma_y^2 \,(1 - \varrho^2). \tag{2.3.1}$$

Fitting a bivariate normal distribution requires finding the five parameters μ_x, σ_x, μ_y, σ_y , and ϱ. To fit (X,Y) to a bivariate normal distribution, first fit X and Y separately to normal distributions. This produces the first four of the bivariate normal parameters. Then use (2.3.1) to determine ϱ, recognizing that μ_x, σ_x, μ_y, and σ_y are constants for this calculation.

[11] Hogg and Tanis [5], section 5.6, pp. 305-311

To improve the fit, we introduce covariates $\boldsymbol{T}$, with $T_0 = 1$, $T_1 = X_{NJ}$ (Indicator function for New Jersey), $T_2 = X_{PA}$, and T_3 = Loss year minus 1990[12].

Using maximum likelihood estimation for fitting X to a normal distribution, $X|T$ is normal with mean $\mu_{x|t}$ = 5.5493 + .971195*T_1 + .698352*T_2 - .00753*T_3, and constant standard deviation σ_x = 0.861097. Similarly $Y|T$ is normal with mean $\mu_{y|t}$ = 8.26265 + .77642*T_1 + .563806*T_2 - .014975*T_3, and constant standard deviation σ_y = 1.46849. These two models determine all the parameters of the bivariate normal distribution except the correlation coefficient ϱ.

The loglikelihood function for the joint observation (x,y) is the sum of 1) the loglikelihood function determined by the modeling of X, and 2) the loglikelihood for the conditional (normally distributed) random variable Y|X. The latter is determined from the relationship (2.3.1) above. For individual uncensored observations, this quantity is the sum of the loglikelihoods calculated using two normal density functions.

The discussion above is a simplification of the actual modeling, since the settlement lag is grouped, and both the claim size and settlement lag are subject to censoring.

The settlement lags were grouped into 14-day intervals with an assumed lag equal to that at the midpoint. For example, settlement lags in the interval 1 to 14 days are all assumed to equal 7.5, resulting in $X = \ln(7.5)$. The claim sizes are rounded to the nearest 100 dollars.

The loglikelihood calculation must be modified if the settlement lag is censored. In this case, the cumulative normal cdf is used in the loglikelihood calculation rather than the density. If $X = ln$(settlement lag) is censored at x_i, then we only know that X is somewhere in the interval $[\, x_i, \infty]$. A practical adjustment is to replace x_i with a value whose probability is halfway between $F(x_i)$ and 1. This means finding x_i*, the point for which

$F_X(x_i^*) = \dfrac{1+F(x_i)}{2}$. Use x_i* rather than x_i in calculating the loglikelihood for $Y|X$.

[12] The number of state indicators was reduced from nine to two since this section is illustrative.

The maximum likelihood estimate for the correlation coefficient is ϱ = .4289. See Figures 13 and 14 for a surface plot and contour plot using the joint distribution.

Modeling the relationship between settlement lag and claim size can be tricky if there have been changes in claim reserving practices during the time period studied. The claim size includes both the paid amount and a case reserve, i.e., an amount estimated by claims personnel that will be sufficient to cover the unpaid portion of the claim. A change in reserving practice, for example, would occur when the company sets case reserves using a "formula" and then switches to having all case reserves set by a claim adjuster. The relatively simple models in this section assume that the method of setting case reserves has remained relatively constant.

2.4 Zero modification of the Claim Size Distribution

The models described in Section 2.2 model claim size X, **given that $X > 0$.** This section includes the positive probabilities that $X = 0$ in the modeling. The basic idea is that the claim size X is the mixture of two distributions:

The zero distribution, with $Pr[X=0] = 1$ and

The conditional distribution of $X \mid X > 0$.

The mixing probabilities are $p = \Pr[X=0]$ and $1-p$. (2.4.1)

The entire paper, except for this section, is concerned with modeling the second distribution. This section discusses models to determine p. Much of the modeling includes the variable I_S, the indicator for whether the claim has settled.

We will model $p = \Pr[X=0]$ with the loss year y as a covariate and define $p_y = \Pr[X=0 \mid \text{loss year} = y]$. For this database, $X = 0$ only if I_S = settlement.event = 1 (that is, if the claim has closed). Open claims with zero amounts are very rare for this data. The probability that a claim is settled should decrease with the loss year, since claims that are more recent have had less time to close. We write:

$\Pr[X = 0 \mid y] = \Pr[X=0 \mid I_S = 1]$ times $\Pr[I_S = 1 \mid y]$ in steps.

The two subsections that follow described how each factor on the right-hand side is calculated.

2.4.1 Model the Probability that a Claim is settled, given Loss Year.

The probability that a claim is settled should decrease with the loss year, since losses from more recent years have had less time to settle. We can use logistic regression to model this probability. Letting the loss year *y* be a numeric covariate, the logistic model assumes that $logit(p) = ln\left(\frac{p}{1-p}\right)$ is a linear function $b_0 + b_1{*}y$ of loss year. The inverse of the *logit* function is $ilogit(\text{x}) = \frac{\exp(x)}{1+\exp(x)}$. The fitted probability Pr [$I_S = 1 \mid y$] equals *ilogit* ($b_0 + b_1{*}y$) once *b* is determined by maximum likelihood.

One nice feature of logistic regression is that summarized observations can be used in the modeling. The necessary summarized data is the number of settled claims $n_{y\,1}$ (i.e., the number of claims for loss year *y* with I_S=1) and the number of unsettled claims $n_{y\,2}$ for each loss year *y*. Let $\eta = b_0 + b_1{*}y$ be the linear predictor and n_y be the total number of claims for loss year *y*.

The loglikelihood is given by[13]

$$l\,(b) = \sum_y \left(n_{y1}\,\eta_y - n_y \ \ln(1+e^{\eta_y}) + \ln\binom{n_y}{n_{y1}} \right)$$

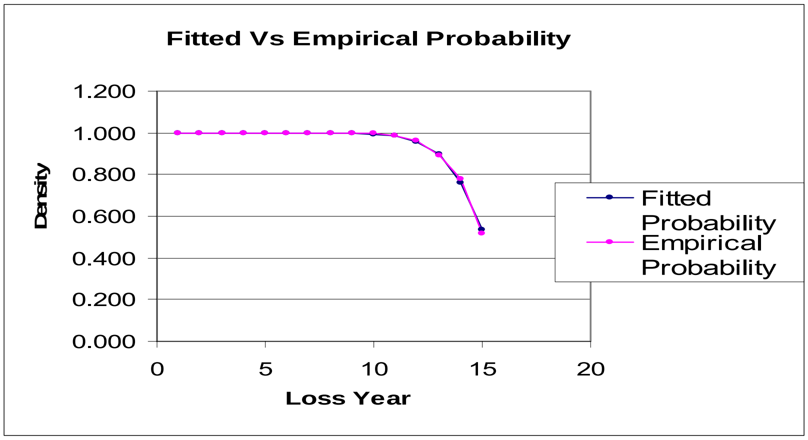
It is generally easier to calculate the loglikelihood directly than it is first to calculate the likelihood, because the latter may involve calculating large factorials, while the loglikelihood can use the simpler *lngamma* function.

[13] See, for example, Faraway [6] pp. 27-28

The following table shows the model data and results.

y	n2	n1	n		η	p_y	
lossyear	# Open	# Settled	# Total	Loglikelihood	Linear predictor	Fitted p	Empirical p
1	0	2131	2131	-0.001	14.183	1.000	1.000
2	0	3228	3228	-0.006	13.180	1.000	1.000
3	0	3238	3238	-0.017	12.177	1.000	1.000
4	0	3082	3082	-0.043	11.174	1.000	1.000
5	0	3432	3432	-0.131	10.171	1.000	1.000
6	1	3763	3764	-1.327	9.167	1.000	1.000
7	1	3366	3367	-1.001	8.164	1.000	1.000
8	16	3370	3386	-17.870	7.161	0.999	0.995
9	9	3174	3183	-2.374	6.158	0.998	0.997
10	13	3149	3162	-3.022	5.154	0.994	0.996
11	43	3185	3228	-3.323	4.151	0.984	0.987
12	124	3006	3130	-3.407	3.148	0.959	0.960
13	375	3077	3452	-4.089	2.145	0.895	0.891
14	724	2551	3275	-8.097	1.141	0.758	0.779
15	1406	1493	2899	-6.424	0.138	0.535	0.515

A graph of fitted versus empirical probabilities follows:



In this model, b_0 = 15.187 and b_1= - 1.0032. The graph shows a reasonably close fit. Loss year 1 is 1992 and loss year 15 is 2006.

2.4.2 Model the Probability that Claim Size = 0 given that the claim is settled.

We again use logistic regression, with the covariates being the indicator variables for state. The summarized data by state is shown in the table below.

In this case, the logistic regression attempts to find probabilities by state x

p_x = Pr { Claim size =0 | State = x and I_S = 1 }.

Models 1 thru 4 were run with different groupings of state:

Model 1 uses indicators for each state.

Model 2 uses indicators for New Jersey and for Pennsylvania.

Model 3 uses one indicator for (Pennsylvania or New Jersey)

Model 4 uses no predictors.

Model 1 is not useful because it has as many predictors as data points. The fitted and empirical probabilities match exactly.

The comparison of nested models shows the following:

Model	Number parameters	Loglikelihood	Likelihood ratio	P-Value	AIC
1	8	-30.677	21.814	0.00057	77.354
2	3	-41.584	2.574	0.10862	89.168
3	2	-42.871	1010.815	0.00000	89.742
4	1	-548.278			1098.557

We select Model 3 to carry forward to the remainder of this section.

x	n1	n2	n		η	p_x	
State	# of zero-size claims	# of claims >0	# in State	Loglikelihood	Linear predictor	Fitted p	Empirical p
6	113	344	457	-5.140	-0.900	0.289	0.247
7	627	1,497	2,124	-4.162	-0.900	0.289	0.295
18	30	63	93	-2.678	-0.900	0.289	0.323
19	955	2,228	3,183	-5.105	-0.900	0.289	0.300
29	8,402	9,153	17,555	-5.796	-0.068	0.483	0.479
34	62	257	319	-10.397	-0.900	0.289	0.194
37	9,707	10,229	19,936	-5.778	-0.068	0.483	0.487
45	454	1,124	1,578	-3.815	-0.900	0.289	0.288
TOTAL	20,350	24,895	45,245				

2.4.3 Put the two models together.

The original goal of the modeling was to develop predictors for the probability that a claim is settled for zero value. This subsection is an example of how to do this given the modeling in subsections 2.4.1 and 2.4.2. Section 2.4.1 produced predicted values for

Pr [$I_S = 1$ | loss year y]. Section 2.4.2 provided predicted values for Pr [Size = 0 | $I_S = 1$ and state x]. Multiplying these quantities gives an estimate of Pr [Size =0 | Loss year y and state x]. These are precisely the mixing probabilities in equations 2.4.1 at the beginning of section 2.4.

The two-way table of these factors follows:

Actual loss yr			1998	1999	2000	2001	2002	2003	2004	2005	2006
Loss yr			**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
	CT	**6**	0.289	0.289	0.288	0.287	0.285	0.277	0.259	0.219	0.154
	DE	**7**	0.289	0.289	0.288	0.287	0.285	0.277	0.259	0.219	0.154
	ME	**18**	0.289	0.289	0.288	0.287	0.285	0.277	0.259	0.219	0.154
STATE	MD	**19**	0.289	0.289	0.288	0.287	0.285	0.277	0.259	0.219	0.154
	NJ	**29**	0.483	0.483	0.482	0.480	0.476	0.463	0.432	0.366	0.258
	OH	**34**	0.289	0.289	0.288	0.287	0.285	0.277	0.259	0.219	0.154
	PA	**37**	0.483	0.483	0.482	0.480	0.476	0.463	0.432	0.366	0.258
	VA	**45**	0.289	0.289	0.288	0.287	0.285	0.277	0.259	0.219	0.154

The probabilities for loss years prior to 1998 match those for year 1998. These values represent probabilities that the claim size equals zero as of the valuation date. To obtain the fitted probabilities that the ultimate claim size equals zero requires a more sophisticated analysis of loss development.

The two-way probability table above assumes that the component conditional probabilities in sections 2.4.1 and 2.4.2 are independent. Whether this is true or not, we have shown that there are covariates that influence the mixing probabilities. In this case, the predictors are 1) the indicator for states NJ or PA and 2) loss year.

2.5 The Effect of Deductibles on Collision Claim Frequency and Severity

The model for collision claim size described in section 2.2.6 is a model for the "ground up" loss distribution W.[14] The collision data is left-shifted and truncated by the amount of the deductible. The algorithm for converting this data to a "ground up" basis follows the "unshifted" approach[15] . This section discusses the nature of the per loss and per payment distributions under various deductibles, which depend on the form of the claim size model. The model selected (Section 2.2.6) for the "ground up" collision loss size is lognormal with

[14] The references frequently use X for the "ground up" distribution. We use W because $\boldsymbol{X}$ refers to covariates in this section.

[15] As described in Klugman, Panjer, and Willmot [1], pp. 341-343

$\mu_i = \sum_j b_j x_{ij}$ and σ constant. The array X_{ij} is the model matrix, with column vectors $X_0 = 1$, X_1 through X_8 set as indicators for eight of the nine states represented in the data, and X_9 the loss year. The lognormal distribution has the following desirable property:

> The j^th moment distribution of a lognormal distribution with parameters μ and σ^2 is also a lognormal distribution with parameters $\mu + j\,\sigma^2$ and σ^2 respectively.[16]

We can now discuss the distributions of Y^P and Y^L for an ordinary deductible *d*, where Y^P and Y^L are the "per payment" and "per loss" variables:[17]

$Y^L = 0$ if $W \le d$, $Y^L = W - d$ if $W > d$.

$Y^P = Y^L \mid Y^L > 0$.

2.5.1 Expected Value by Deductible for given μ.

From the initial modeling described in subsection 2.2.6 and in Appendix E, the lognormal provides a good fit for collision loss size. The expected values of Y^L and Y^P are as follows:

$$E(Y^L) = \exp(\mu + \frac{1}{2}\sigma^2) - \exp(\mu + \frac{1}{2}\sigma^2)\Phi(\frac{\ln d - \mu - \sigma^2}{\sigma}) - d(1 - F_W(d))$$

$$E(Y^P) = E(Y^L)\,/\,(1 - F_W(d))$$

where $F(w) = \Phi(\frac{\ln w - \mu}{\sigma})$.

The formula for $E(Y^L)$ follows from the fact that $Y^L = W - (W \wedge d)$ and the formula for the limited moments of a lognormal distribution[18].

[16] Bickerstaff [4], p 73

[17] Klugman, Panjer, and Willmot [1], p. 116

[18] Klugman, Panjer, and Willmot [1], p. 638-639

For $\mu = 7.7$, the table below shows how different deductibles d affect the values of $E(Y^L)$ and $E(Y^P)$:

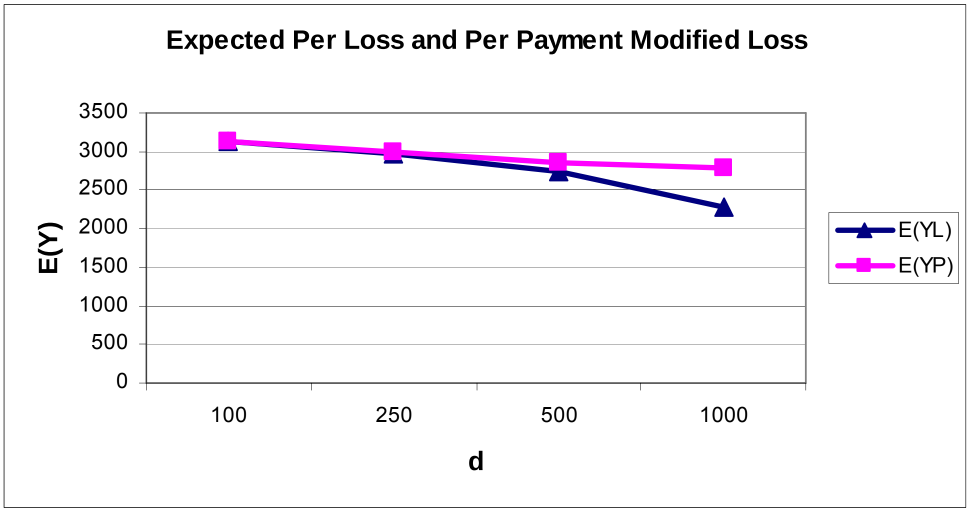
d	100	250	500	1000	μ	7.7
$E(Y^L)$	3124	2975	2730	2285	σ	0.87
$E(Y^P)$	3125	2993	2855	2790		

To show that 7.7 is a representative value of μ, observe that b_0 = -43.06 and the coefficient of loss year is 0.02534 (see section 2.2.6). For loss year 2000, the linear predictor

$$\mu = -43.06 + 0.02534 * 2000 + \text{coefficient of state} = 7.63 + \text{coefficient of state}.$$

The coefficients of the state indicators range from -0.262 to +0.277.

Following is a graph of the data above:



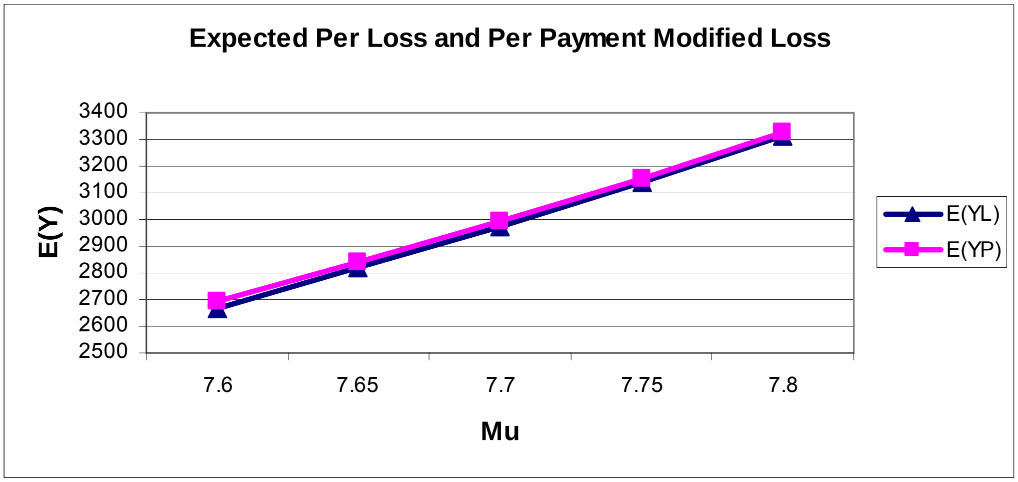
2.5.2 Expected Value for various μ given Deductible *d*.

Recall that μ_i is a linear function of the covariates ($X_{i,j}$; j = 0, 1 2, …*p*), where *p*=9 in this case. The following table shows how different values of μ affect the value of $E(Y^L)$ and $E(Y^P)$ under a fixed deductible of 250:

μ	7.6	7.65	7.7	7.75	7.8	sigma	0.87
$E(Y^L)$	2667.886	2817.389	2974.572	3139.827	3313.566	d	250
$E(Y^P)$	2690.607	2837.852	2992.945	3156.273	3328.241		

Since the coefficient of loss year is 0.025 and the increments in μ in the table are .05, the adjacent entries in each row show the approximate effect of two years' inflation.

In the graph below, E(Y^L) and E(Y^P) are close together because the $250 deductible is much smaller than the average loss.



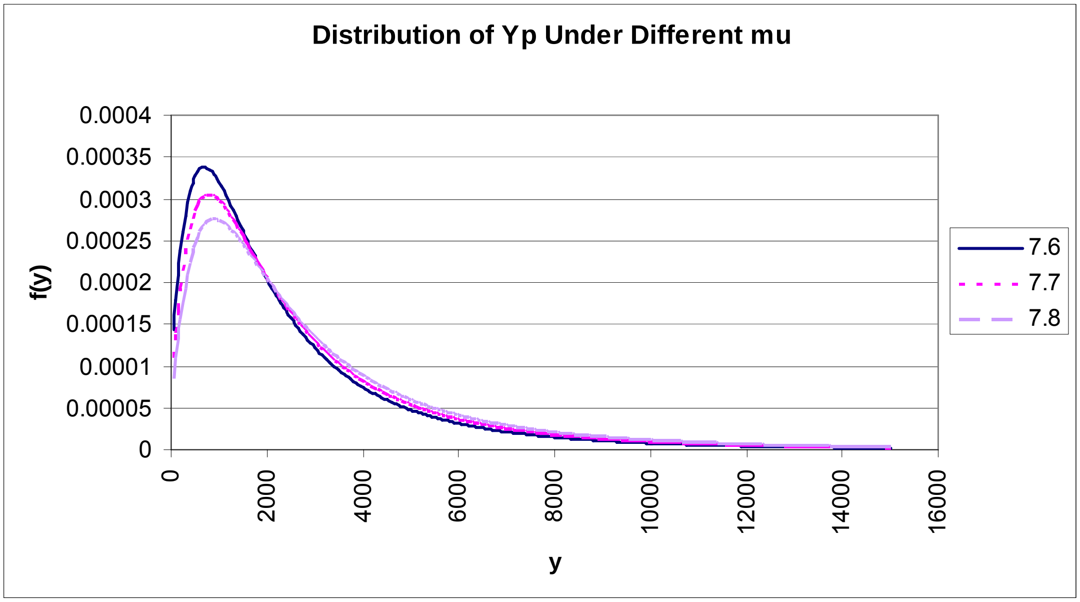
We have discussed the expected values of Y^L and Y^P of a lognormal distribution. We now examine the distributions themselves.

Discrete probability $f_{Y^L}(y) = \Phi\left(\dfrac{\ln d - \mu}{\sigma}\right) = F_W(d)$, if $y = 0$

Density $f_{Y^L}(y) = F_W(y+d) = \dfrac{\exp(-(z^2/2))}{(y+d)\sigma\sqrt{2\pi}}$, if y>0, where $z = \dfrac{\ln(y+d)-\mu}{\sigma}$

$f_{Y^P}(y) = f_{Y^L}(y) \;/\; (1 - F_W(d))$

From this, we can graph the density of Y^P under a fixed deductible of 250:



2.5.3 Effect of Loss Year and deductible on Losses.

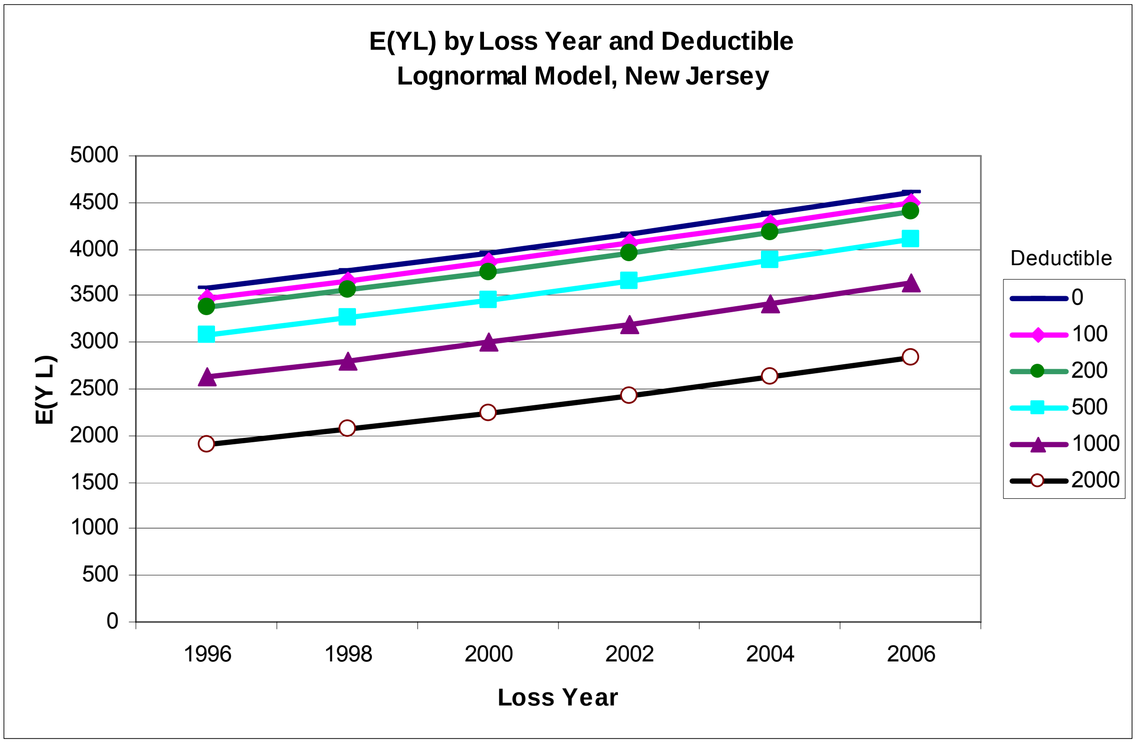
Loss year and deductible are covariates that impact the size of collision losses. Here we single out this impact and give an illustrative example. Consider collision loss in New Jersey. In the model,

$\mu = b_0 + b_{NJ} + b_{year}$ *year, where b_0 = -43.0607, b_{NJ} = 0.277003, b_{year} = 0.025344.

Here μ applies to the lognormal distribution of "ground up" losses, and σ = 0.87.

We show how deductible and loss year together affect the expected value of collision losses. Following are the values of $E(Y^L)$ for loss years 1996 to 2006 under different deductibles, followed by a graph of the same information.

	1996	1998	2000	2002	2004	2006
μ =	7.80308	7.85376	7.90445	7.95514	8.00583	8.05652
0	3574	3760	3956	4161	4378	4605
100	3474	3660	3856	4061	4278	4505
200	3374	3560	3756	3961	4178	4405
500	3079	3264	3459	3664	3880	4107
1000	2623	2804	2995	3196	3409	3633
2000	1907	2069	2242	2426	2622	2830



2.5.4 Pareto Model

The Pareto is one of the severity distributions programmed into the Loss Simulator. Section 2.2 examined distributions that the R statistical language can model in its parametric survival model package. The Pareto distribution is not one of them.

This section discusses the Pareto distribution using the same covariates as for the lognormal model. The modeling is performed by maximizing the loglikelihood using the Excel Solver.

The Pareto distribution has the desirable property that if the "ground up" loss W is Pareto(θ, a), then the "per payment" variable Y^{P} under a deductible d is distributed Pareto(θ+d, a)[19]. From this result, we obtain:

$E(Y^{P}) = (\theta + d) / (\alpha - 1)$ and then

$$\mathrm{E}(Y^{L}) = \mathrm{E}(Y^{P}) * (1\text{-}F_{W}(d)\,) = \mathrm{E}(Y^{P}) * \left(\frac{\theta}{\theta+d}\right)^{\alpha}\ .$$

We developed a similar model as the lognormal to fit the collision losses. The density function of Y^{L} is:

Discrete probability $f_{Y^{L}}(y) = 1 - \left(\frac{\theta}{\theta+d}\right)^{\alpha}$, if $y = 0$

Density $= \dfrac{\alpha\theta^{\alpha}}{(y+d+\theta)^{\alpha+1}}$, if $y>0$

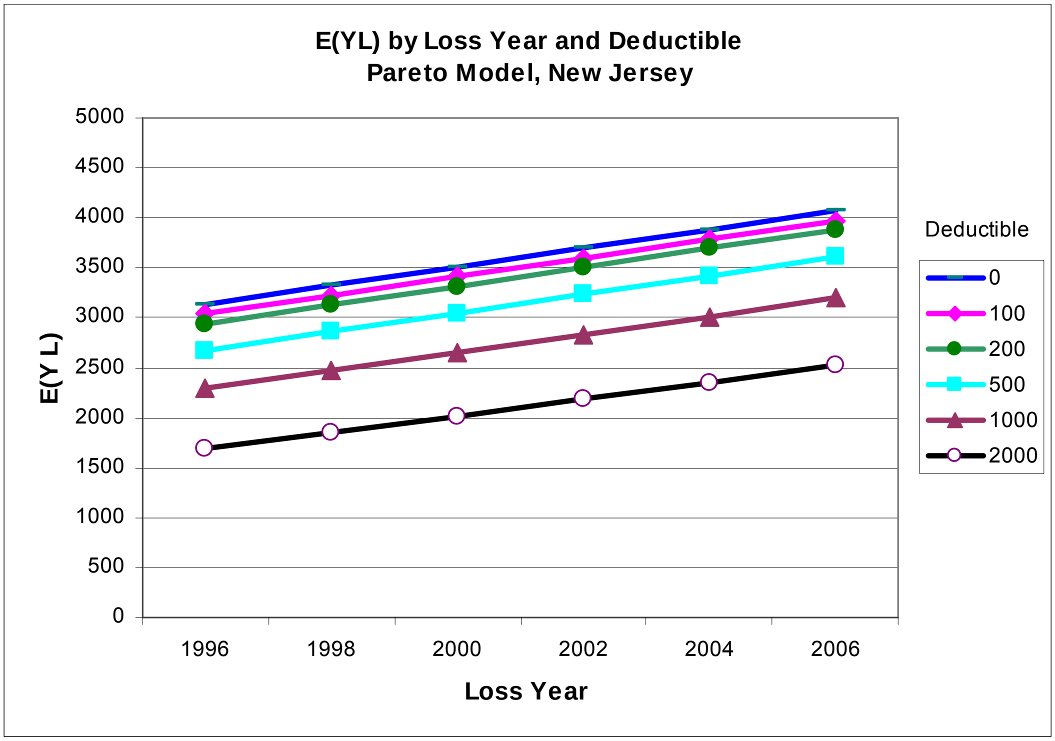
In modeling the Pareto using covariates, the linear predictor $\boldsymbol{\mu} = \boldsymbol{Xb}$ is an estimate of the mean, which equals $\boldsymbol{\theta}/(\alpha\text{-}1)$. Thus, in setting up the Pareto model, one re-parameterizes the distribution as Pareto(μ ,α), with $\boldsymbol{\theta}$ then calculated as $\boldsymbol{\theta} = \boldsymbol{\mu}\ (\alpha\text{-}1)$.[20]

As in section 2.5.3, the table below shows the value of $E(Y^{L})$ for New Jersey from loss year 1996 to 2006 under different deductibles.

Loss year	1996	1998	2000	2002	2004	2006
μ =	3136.04	3323.81	3511.57	3699.33	3887.10	4074.86
θ =	34523.67	36590.71	38657.74	40724.78	42791.82	44858.85
Table of $E(Y^{L})$						
0	3136	3324	3512	3699	3887	4075
100	3038	3225	3413	3601	3788	3976
200	2943	3130	3318	3505	3693	3880
500	2677	2863	3048	3234	3420	3607
1000	2290	2470	2651	2832	3014	3197
2000	1687	1850	2015	2182	2351	2521

[19] Klugman, Panjer, and Willmot [1] See Example 5.2, p. 117

[20] Klugman, Panjer, and Willmot [1], pp. 413-414



2.6 The Interaction of Report Lag and Settlement Lag

This section discusses the correlation between report lag and settlement lag for Auto BI and Collision. For each coverage, the joint distribution of *ln*(report lag) and *ln*(settlement lag) is fit to a bivariate normal, using the procedure described in section 2.3. The first step in this fitting is to determine the lognormal parameters for report lag and settlement lag separately. In doing this, we looked at grouping each variable into 14-day intervals as well as using the variable "as is." After fitting the variables separately, we calculate the correlation coefficient ρ and test its significance.

2.6.1. Report lag and settlement lag for Collision

The normal distributions were fit to the variables X = *ln*(report lag) and Y = *ln*(settlement lag) for both the actual lags ("as is" model) and the lags grouped into intervals with a length of 14 days.

The loglikelihoods for the "as is" models use the density function for uncensored observations. To provide a comparable loglikelihood for the interval model, the loglikelihood was calculated using the density at the midpoint of the interval.[21]

Following are some statistics from the univariate models:

	As is X	Interval X	As Is Y	Interval Y
mu	0.5868	2.1038	2.5511	2.7228
sigma	0.9876	0.4145	1.1476	0.8923
loglikelihood	-181,763	-69,559	-201,171	-168,655
Empirical mean	0.586759		2.5511	

The empirical mean for the "As Is" settlement lag *Y* is censored for all claims still open as of 12/31/2006.

We select the "as is" model for report lag *X* because the mean from the interval model is not close to the empirical mean. When the report lag is grouped, all lags between 0 and 14 days are mapped to 7.5. This overstates the mean report lag, since there is a preponderance of claims with very short report lags. We select the interval model for *Y*.

The value of the correlation coefficient ϱ is 0.05659 for the bivariate normal distribution, using the methods of section 2.3. One way to test for the significance of ϱ is to compare the loglikelihood for the model containing ϱ against a model without ϱ. The latter model is one where *X* and *Y* are independent.

An approximate test relies on the fact that W = (1/2) ln $\left(\frac{1+r}{1-r}\right)$ is approximately normal with mean = (1/2) ln $\left(\frac{1+\rho}{1-\rho}\right)$ and variance 1/(n-3). Note that if ϱ =0, mean(W) = 0. To

[21] The theoretically correct method uses the difference in the cdf between the ending and beginning point of the interval. The density is used here to make the loglikelihood comparable to other models tested.

test the hypothesis that ϱ = 0, construct Z=[*Sqrt*(n-3)/2]*[*ln*(1+r) – ln (1-r)]. For this model Z = 20.365, much larger than the 5% critical value 1.96[22].

2.6.2 Report lag and settlement lag for BI

The normal distributions were fit to the variables X = *ln*(report lag) and Y = *ln*(settlement lag) for both the actual lags ("as is" model) and the lags grouped into intervals with a length of 14 days.

The comments from the previous subsection apply to this section also.

Following are some statistics from the univariate models:

	As is X	Interval X	As Is Y	Interval Y
mu	0.78068	2.20469	5.69340	5.70539
sigma	1.27138	0.67610	1.34862	1.30396
loglikelihood	-49,937	-30,928	-51,319	-50,330

We select the "interval" model for both X and Y.

The value of the correlation coefficient ϱ is 0.00075 for the bivariate normal distribution, using the methods of section 2.3. Using the same test as that used for settlement lag, construct Z=[*Sqrt*(n-3)/2]*[*ln*(1+r) – *ln* (1-r)]. For this model, Z = 0.129649, much smaller than the 5% critical value 1.96[23].

Therefore, the **null hypothesis** that X and Y are independent is **not rejected**.

2.7 Data Processing

The data came in from the anonymous "Source" as a Microsoft Access database with the claim transactions in three separate "tables"[24]: 1) Initial reserve transactions, 2) Reserve

[22] Hogg, McKean, and Craig [2], p. 500

[23] Hogg, McKean, and Craig [2], p. 500

[24] The term "tables" is used for relational databases. If the data were organized as separate "flat files", each table would correspond to a flat file.

changes, and 3) Payments. A fourth table contains descriptive information for each claim. Approximate 30-40% of the entire project consisted of processing the data into a format suitable for the statistical models. Processing this data strained the capacity of the computers processing the information.

To model report lag, settlement lag, and claim size required merging the transaction tables to produce a table of incurred-to-date loss information by claim. In the research class, a subgroup of students learned relational database concepts and performed this processing. This type of work must be "custom-tailored" according to the format of the incoming data.

Important relational database concepts to understand include **primary key**, **inner join**, and **outer join**. The following website contains useful material for understanding relational databases:

http://support.microsoft.com/kb/283698/EN-US/

Follow the directions to download a Word document.

A useful website for understanding "joins" is the following:

http://office.microsoft.com/en-us/access/HA100963201033.aspx

This project does not treat the parameterizing of frequency or exposures. Such work requires processing the premium/exposure transactions. The necessary information was not available at the time the course met. This work would require more computer capacity than handling the loss transactions.

2.7.1 Suggestions for the Data Processing

For this project, the data processing was started by the Source and the primary author during the semester before the start of the class. Much of the "data scrubbing" took place at this time. For projects of this nature, the data processing should take place early and on a "crash" basis. Otherwise, the project may not be completed during the course period because students will be waiting for the data to be cleaned.

3. RESULTS AND DISCUSSION

This project uses actual insurance data to parameterize the report lag, settlement lag, and claim size distributions for the Casualty Actuarial Society's Loss Simulator. Researchers developed models using spreadsheets that optimize[25] likelihood functions. Models were also developed in the statistical language R.

The models were run separately for Auto BI and Auto Collision for each of the variables above. They fit lags to the Weibull, Exponential, and Lognormal distributions, while fitting claim sizes to Weibull, Lognormal, and Pareto distributions. The settlement lags and the Auto BI claim size are censored variables. Parametric survival models in the R language can fit censored variables to all the distributions above except the Pareto.

The models produced good results for settlement lag and claim size, as long as suitable covariates are included. They produced poor results for report lag. The report lags are very short for these two coverages.

Bivariate normal models determine correlation coefficients between settlement lag and report lag, and between settlement lag and claims size. Natural logarithms of all variables were used in these models. The results show a small but significant correlation between report lag and settlement lag for Collision and an insignificant correlation for Auto BI. A significant positive correlation of 0.429 exists between settlement lag and claim size for Auto BI. The Simulator provides for entry of the correlation coefficient.

Section 2.5 presents some of effects of Collision deductibles on the payment severity and mean payments. This section illustrates some of the results of modeling the "ground up" loss distribution as a lognormal variable. This special topic is presented because deductibles truncate and left-shift the loss payments. This is a more complex phenomenon than handling censored data. The language R does not handle this adjustment automatically.[26]

Section 2.4 provides a model for determining the (discrete) probability that claim size equals zero, when only the latest valuation is available for each claim. If the entire valuation history for each claim were used, a more sophisticated model could be developed.

[25] In Microsoft Excel, the "Solver" is the optimizer. See Klugman, Panjer, and Willmot [1], Appendix F, pp. 659-669 for a discussion.

The results support the model types programmed in the Simulator, once covariates are included. For example, the modeling shows that state is clearly an important covariate and that including loss year as a predictor produces an estimated "trend". The PP Plots and Standard Residual Plots provide a way to evaluate the models visually. The Simulator currently allows covariates only in the sense that each level of a categorical predictor (such as state) can be assigned to a different "type", i.e., grouping of data, in the Simulator.

A second goal of the project is to promote understanding of the predictive models presented and to illustrate how to perform predictive modeling without using high-powered software. This is a reason why the models were done in Excel rather than an advanced statistical language.

Modeling using spreadsheets requires a great effort. Some datasets were too large to fit into a spreadsheet, requiring that the observations be grouped. This requires more data processing and more effort in programming the likelihood functions. Had the basic model been programmed in a statistical language, the class could have tested many other covariates beyond state and loss year. Changing covariates requires substantial work in a spreadsheet environment. However, the modeling that was performed using covariates shows the necessity of including these covariates in the Simulator.

This project provides an example of how universities and professional actuarial bodies can collaborate on research.

4. CONCLUSIONS

The models described in this paper provide a way to parameterize the Casualty Actuarial Society's Claim Simulator, which is in the development stage. These models deal specifically with modeling report lags, settlement lags, claims sizes, and their interactions. We found that the Simulator provides a useful set of distributions for simulating these variables. The caveat is that the models need to use covariates (predictors) to refine the estimate of the

[26] SPlus, the commercial available version of S, does handle left-shifting and truncation in the censorReg

means of the distributions. This implies that the Simulator needs to handle covariates.

The models were developed using spreadsheets. This enabled the students to understand thoroughly the calculation of maximum likelihood estimates. Spreadsheets are labor intensive and limited in their capability for developing these models. Further model development should take place using a statistical language. The language "R" performs this task well, is open source, and is available free of charge. Other languages such as SPlus (a commercial counterpart to "R"), SAS, and SPSS can perform this modeling. A subgroup of the students performed modeling in "R", with some source code and results included in this paper.

Parametric survival models and Generalized Linear Models are useful modeling tools for analyzing lags and claim sizes. Logistic regression was used to model the probability that the claim size is zero.

This project goes beyond the typical statistics work at universities because of the large volume of data. The processing of the data uses relational database concepts. Partly because of the data processing requirement, the scope is limited to fitting a subset of the all Simulator's parameters. For example, this paper does not discuss parameters relating to frequency and distribution of exposure. The paper does not analyze loss development. Data from another line of business, such as Medical Malpractice, is needed to provide non-trivial report lags.

A main goal of the research is to explore ways in which the academic community and the learned actuarial bodies (such as the CAS or SOA) can collaborate on research. This is one of the most rewarding results of the project. The CAS supplied the data and considerable consulting on the modeling. Ball State University provided the human resources, the facilities, and academic knowledge.

procedure. This was not tested by the modelers.

REFERENCES

[1] Klugman, Stuart A., Panjer, Harry H., and Willmot, Gordon E. *Loss Models From Data to Decisions, 2nd Edition*, John Wiley & Sons, Inc., 2004.

[2] Hogg, Robert V., McKean, Joseph W., and Craig, Allen T., *Introduction to Mathematical Statistics 6th ed.* Pearson Prentice Hall, 2005.

[3] Tableman, Mara and Kim, Jong Sung. *Survival Analysis Using S: Analysis of Time-to-Event Data.* Chapman and Hall/CRC, 2004.

[4] Bickerstaff, David R., "Automobile Collision Deductibles and Repair Cost Groups: The Lognormal Model," *Proceedings Casualty Actuarial Society 1972*

[5] Hogg, Robert V. and Tanis, Elliott A., *Probability and Statistical Inference, 7th ed.*, Pearson Prentice Hall, 2006.

[6] Faraway, Julian J., *Extending the Linear Model with R*, Chapman and Hall/CRC, 2006.

[7] On-line reference:
http://www.weibull.com/LifeDataWeb/lifedataweb.htm

Abbreviations and notations

AIC, Akaike Information Criteria
CAS, Casualty Actuarial Society
cdf, cumulative distribution function
GLM, generalized linear model

LHS, left-hand side (of an equation)
LSMWP, Loss Simulation Model Working Party
RHS, right-hand size (of an equation)
SOA, Society of Actuaries

Biographies of the Authors

Joseph Marker led the Ball State University Research Class in performing the analysis and writing this paper. He was Visiting Lincoln National Group Distinguished Professor of Actuarial Science at Ball State University during the academic year 2006-2007. He holds a Master's degree in Mathematics from the University of Minnesota and a B.A. in Mathematics from the University of Michigan.

Joe is a Fellow of the Casualty Actuarial Society (CAS) and a member of the American Academy of Actuaries. He is a member of the CAS Loss Simulation Model Working Party (LSMWP) and recently served on the Loss Reserve Variability Working Party. Joe runs the consulting firm Marker Actuarial Services, LLC that consults on Property-Casualty insurance, and had twenty-eight years' prior experience at property-casualty insurance companies. He has authored a paper for the CAS Proceedings and co-authored a paper on Claims-Made Insurance ratemaking that appears on the CAS Examination Syllabus.

The following graduate students from Ball State University constructed the models, built the database, and wrote sections of the paper: **Nii Armah, Florent Chekete, Rong Chen, John Dizer, Daniel Earls, Gregory Faltenovich, Haixia Gu, Blake Hill, David D. Jones, Andrew Kwon, Joseph Poku, Huashi Shao, Eugene Tan,** and **Kejun Xu**. In addition, Haixia Gu, John Dizer, and Andrew Kwon helped organize the work assignments. Florent Chekete and Rong Chen developed the work in the "R" statistical language.

Acknowledgments

The authors acknowledge the assistance of John Emert, Ph.D., Chair of the Department of Mathematical Sciences, and Michael Maggiotto, Ph.D., Dean of the College of Sciences and Humanities, Ball State University, for providing the supporting environment under which this work was performed. Several members of the Casualty Actuarial Society LSMWP acted as consultants, especially Kevin Cleary, Glenn Meyers, Gary Venter, Glen Barnett, and Richard Vaughan. The authors recognize for their help Ball State Faculty members Munni Begum, Mir Ali, Dale Umbach, John Foley, Bart Frye, and Linley Baker. Faculty member Gary Dean provided helpful background material. Susan Bourne, Administrative Coordinator to John Emert, helped the project run smoothly.

Professional actuaries from both the CAS and SOA reviewed the first draft and made helpful suggestions. These include Robert Bear, Stuart Klugman, and Glenn Meyers.

Supplementary Material

The website for the Insightful Corporation contains useful documentation. The company's SPlus software is very close to R. Its resource material can be found at site http://www.insightful.com/insightful_doclib/. Of particular interest are the on-line statistical manuals at

http://www.insightful.com/support/splus60win/statman1.pdf and
http://www.insightful.com/support/splus60win/statman2.pdf

The R statistical modeling package is free and can be downloaded by accessing the Internet address http://cran.r-project.org/ .

Appendix A

Two tests for comparing models are the Likelihood Ratio test and the Akaike Information Criteria (AIC). The Likelihood Ratio test applies to nested models, while the AIC can compare models where one is not a special case of the other.

Likelihood Ratio Test:

For this test, the distributions compared must be nested. The "null hypothesis" H_0 is that the model with fewer parameters is sufficient. Klugman, Panjer and Willmot [1], section 13.4.4, discusses this test. The alternate hypothesis H_1 is that the more complex model is needed.

The likelihood ratio test constructs a test statistic $T = 2\ (ln\ L_1 - ln\ L_0)$, where L is the likelihood associated with each model. If $T >$ a critical value for the Chi-square distribution, the null hypothesis is rejected.[27]

Akaike Information Criterion (AIC)

AIC applies more generally than the likelihood ratio test because it requires neither that the models be nested nor that they have the same error distribution. In the general case,

$$AIC = 2p - 2\ln(L)$$

where p is the number of parameters, and L is the maximum likelihood.

The "preferred" model has the lowest AIC. Note that adding parameters introduces into the AIC a penalty that may offset the gain in loglikelihood. In this way AIC discourages using unneeded parameters[28]. The result is that the AIC method produces an optimal model using the fewest number of parameters.[29]

[27] See Klugman, Panjer, and Willmot [1], page 346.

[28] There is debate over whether the term "2*p*" in the penalty is large enough for datasets with many observations.

[29] See Hogg, McKean, and Craig [2], page 106.

Appendix B

Modeling Censored and Truncated Variables

B. 1. Introduction

Estimation by method of moments and percentile matching is often easy to do, but these estimators tend to perform poorly. The main reason for this is that they use a few features of data, rather than the entire set of observations. It is particularly important to use as much information as possible when the population has a heavy tail.

Another drawback of these methods is that they require that all the observations be from the same random variable. Otherwise it is not clear what to use for the population moments or percentiles. For example if half of the observations have a deductible of 50 and half have a deductible of 100, it is not clear what sample mean should be equated[30].

Of many possibilities, the only one used here is the maximum likelihood estimator. To define the maximum likelihood estimator, let the dataset consist of *n* events A_1 , … ,A_n , where A_j is whatever was observed for the j^{th} observation.

The ***likelihood function*** is

$$L(\theta) = \prod_{j=1}^{n} \Pr(X_j \in A_j \mid \theta)$$

and **the maximum likelihood estimate** of **θ** is the vector that maximizes the likelihood function. [31]

There is no guarantee that the function has the maximum at eligible parameter values. It is possible that various parameters become zero or infinite, the likelihood function will continue to increase. The likelihood function is the probability of obtaining the sample

[30] One way to rectify this drawback is to first determine a data-dependent model such as the Kaplan-Meier estimate, then use percentiles or moments from that model.

[31] Some authors write the likelihood as L(θ|x), where the vector x represents the observed data. Because observed data can take many forms, the dependence of the likelihood function on the data is suppressed in notation.

results that were obtained for the hypothesized type of model, given a particular parameter value ***θ***. One of the major attractions of this estimator is that it is almost always available[32].

It is often easier to maximize the logarithm of the likelihood function. Because it occurs so often, we denote the **loglikelihood function** as $l(\theta) = lnL(\theta)$

B. 2 Complete, Individual Data

When there is no truncation and no censoring, and the value of each observation is recorded, it is easy to write the likelihood and loglikelihood functions:

$$L(\theta) = \prod_{j=1}^{n} f_{x_j}(x_j \mid \theta), \quad l(\theta) = \sum_{j=1}^{n} \ln f_{x_j}(x_j \mid \theta).$$

As an illustration, for the exponential distribution, the value $\hat{\theta}$ that maximizes *l* (*θ*) is determined:

$$l(\theta) = \sum_{j=1}^{n} (-\ln\theta - x_j\theta^{-1}) = -n\ln\theta - n\bar{x}\theta^{-1},$$

$$l'(\theta) = -n\theta^{-1} + n\bar{x}\theta^{-2} = 0,$$

$$n\theta = n\bar{x}$$

$$\hat{\theta} = \bar{x}.$$

B. 3 Complete, Grouped Data

The following material describes how to handle complete, grouped data[33]:

> When data are complete and grouped, the observations may be summarized as follows. Begin with a set of numbers, $c_0 < c_1 < \ldots.. < c_k$, where c_0 is the smallest possible observation, (usually zero), and c_k is the largest possible observation, (usually infinity). For such data, the likelihood function is

[32] See Klugman, Panjer, and Willmot [1], pages 337-339 for a more complete discussion.

[33] Klugman, Panjer, and Willmot [1], p. 341

$$L(\theta)=\prod_{j=1}^{k}[F(c_j\mid\theta)-F(c_{j-1}\mid\theta)]^{n_j}$$

Its logarithm is

$$l(\theta)=\sum\nolimits_{j=1}^{k} n_j \ln[F(c_j\mid\theta)-F(c_{j-1}\mid\theta)].$$

B. 4 Truncated or Censored Data.

Handling censored data is not complicated – simply treat the data as lying in the interval running from the censoring point to infinity. Truncated data, however, presents a great challenge. There are two ways to proceed; one is to shift the data by subtracting the truncation point from each observation, and the other is to accept the fact that there is no information about values below the truncation point.

We have seen that insurance data is often truncated from below and censored from above. Models for time-till-event variables are often called "Survival Models." Examples are report lag and settlement lag. Many of the same techniques apply to claim size distributions, which frequently arise from policies written with a deductible (truncation point) and a policy limit (censoring point).

The observed claim sizes from a policy issued with deductibles are truncated and left-shifted. That is, if *X* is the "ground up" distribution and *X* > *d*, the observed claim amount equals *X-d* rather than *X*.

For example, for a Pareto distribution with parameters **α** and **θ,** the likelihood function using the shifting approach[34] is:

$$L(\alpha)=\prod_{j=1}^{n}\frac{\alpha\theta^{\alpha}}{(\theta+x_j)^{\alpha+1}}$$

$$l(\alpha)=\sum\nolimits_{j=1}^{n}[\ln\alpha+\alpha\ln\theta-(\alpha+1)\ln(x_j+\theta)]$$

[34] Klugman, Panjer, and Willmot [1], pp. 341-343

The likelihood function using the unshifted approach is:

$$L(\alpha)=\prod_{j=1}^{n}[\frac{\alpha\theta^{\alpha}}{(\theta+x_j)^{\alpha+1}}/(\frac{\theta}{\theta+d})^{\alpha}]$$

B.5. Lognormal Log-Likelihood Functions and their Partials [35]

The general log-likelihood function (without the constant) for the lognormal distribution is composed of three summation portions:

$$ln(\mathrm{L})=\Lambda=\sum_{i=1}^{F_e}N_i\,ln\left[\frac{1}{\sigma_T}\phi\left(\frac{\ln(T_i)-\mu'}{\sigma_T}\right)\right]$$

$$+\sum_{i=1}^{S}N_i'\,ln\left[1-\Phi\left(\frac{\ln(T_i')-\mu'}{\sigma_T}\right)\right]$$

$$+\sum_{i=1}^{FI}N_i''\,ln\left[\Phi\left(\frac{\ln(T_{Ri}'')-\mu'}{\sigma_T}\right)-\Phi\left(\frac{\ln(T_{Li}'')-\mu'}{\sigma_T}\right)\right]$$

where:

- F_e is the number of unique exact data points
- N_i is the number of exact observations whose value is T_i.
- μ' and σ_T are the mean and standard deviation of the natural logarithms of the observations (unknown a priori, the are the two parameters to be found)
- T_i is the observed value in i^{th} group of exact observations
- S is the number of unique values for the censored data points
- N_i' is the number of observations whose censored value equals T_i'

[35] See the website [7]. Navigate to the topic "Lognormal distribution – Estimation of Lognormal parameters."

- T_i' is the observed value in the *i*th censored data group
- *FI* is the number of grouping intervals for grouped observations.
- N_i'' is the number of observations in the *i*th grouped data interval
- T_{Li}'' is the beginning of the *i*th interval
- and T_{Ri}'' is the ending of the *i*th interval

The formula above assumes that the dataset groups all the like observations and contains the field *N* with the number of observations. If the data for the first two summations consists of individual observations, then N_i and N_i' all equal one, and the T values are not necessarily unique.

There are several approaches to finding the values of μ and σ that maximize the loglikelihood. An analytical solution can be found by solving for a pair of parameters

(μ', σ_T) so that $\dfrac{\partial \Lambda}{\partial \mu} = 0$ and $\dfrac{\partial \Lambda}{\partial \sigma_T} = 0$.

This project uses numerical methods. In spreadsheet software, such as Excel, one constructs the likelihood function using the formula above, and then uses the "Solver" to maximize this function by changing the variables μ' and σ_T.

In the statistical language R there are two approaches. One approach is to construct the likelihood function as above and then use the "optim" function to maximize its value by changing the two parameters.

Fortunately, the problem above can be solved using "parametric survival" models that are part of the "R" language. This eliminates the need to construct the likelihood function. It does require that an "event" variable be constructed. If censoring is the only modification, the event variable equals 0 if the observation is censored and 1 if the observation is exact.

Appendix C

Grouping of Data.

In performing predictive modeling, one ideally models individual observations. Statistical software, such as "R," handles individual observations well. Much of this project uses spreadsheet software (Microsoft Excel) rather than a statistical package. Excel has a limitation of approximately 65,500 rows[36] (observations). To use Excel for modeling on large datasets requires ones to group the data so that the number of unique combinations of the variable of interest (lag or claim size) and covariates fits within this limitation. For example, in the claim size studies described in section 2.2, the claim sizes (for claims above $0) were grouped as follows:

Claims sized 1-49 were grouped at 25.

Claims sized 50-9499 were rounded to the nearest 100.

Claims sized 9500-99499 were rounded to the nearest 1000

Claims 99500 and above retained their individual values.

Grouping of data for spreadsheets changes the computation of the maximum likelihood function. For exact values from a hypothesized continuous distribution Y, the likelihood for an individual observation uses the density function f_Y. For an observation grouped into an interval $(c_{j-1}, c_j]$ the likelihood function L equals $F_Y(c_j)$-$F_Y(c_{j-1})$.

If all observations are grouped into intervals with endpoints $c_0 < c_1 < c_2 < \ldots . c_k$, the total loglikelihood in the j^{th} interval equals $n_j \, ln[F_Y(c_j)\text{-}F_Y(c_{j-1})]$, where n_j is the number of observations falling into the j^{th} interval.[37] An observation right-censored at c has loglikelihood $ln[\, 1 - F_Y(c)]$.

Grouping and censoring add complication to the modeling. Consider the plot of standardized residuals for a censored observation. All we know about an observation

[36] "Excel 7" was introduced in 2007 and can reportedly accommodate one million rows.

censored at c is that $Y \geq c$. For plotting residuals and PP plots, we substitute y* = $F_Y^{-1}(F^*)$, where F* is randomly chosen from the interval $[F_Y(c),1]$. The rationale follows:

> The variable $F_Y(Y)$ is uniformly distributed on [0,1], since[38]
>
> $\Pr[F_Y(Y) \leq w] = \Pr[Y \leq F_Y^{-1}(w)] = F_Y F_Y^{-1}(w) = w$.
>
> Therefore $F_Y \; \Big| Y>c$ is uniformly distributed in $(F_Y(c), 1]$, showing the reasonableness of the choices F* and y* above.

Another way to handle the plotting for data with censored observations is to use a distribution function derived from Kaplan-Meier estimates.[39] The "PP Plot" in this case would graph the Kaplan-Meier distribution function values (x axis) versus the model probabilities for the uncensored points. This plot for Auto BI settlement lag was virtually indistinguishable from the PP Plot described above.

For portions of this project, the settlement lag and report lag were grouped because grouping them produced more reasonable results than trying to assert that a one-day difference in lag is meaningful.

[37] See Klugman, Panjer, and Willmot [1], page 341.

[38] Here we assume that Y has a continuous strictly increasing cumulative distribution function.

[39] Klugman, Panjer, and Willmot [1], pages 297-303.

Appendix D

Example of a Model in R

This section contains the "R" source code to model the Auto BI size of loss. This model fits natural log of the Claim Size distribution to a Normal (Gaussian) distribution with covariates state (categorical variable) and loss year (numeric variable). The models 1 through 3 refer to three different grouping of state:

Model 1: Separate indicator variables for each state.

Model 2: Separate indicators for NJ and PA, all other states combined.

Model 3: Combined indicator for NJ-PA, all other states combined.

The Claim Size is censored at the policy limits. The model uses the parametric survival modeling procedure "survreg" in R. This procedure performs the correct likelihood calculation for a censored dependent variable.

The R code and the output from Model 1 follow:

```
infileBi <- "  insert name of data file here “

data2BI <-
read.csv(file=infileBi,na.strings="#N/A",colClasses="character")
names(data2BI)
dim(data2BI)

sapply(data2BI,class)

temp1 <- sapply(data2BI,is.factor)
any(temp1)

dataBI <- data2BI

dataBI$AMT.INCD <- as.numeric(dataBI$AMT.INCD)
dataBI$amtincd.event <- as.numeric(dataBI$amtincd.event)
dataBI$lossyear <- as.numeric(dataBI$lossyear)
dataBI$NUM.ST.CD  <- as.numeric(dataBI$NUM.ST.CD)
dataBI$AUTO.USE.CODE  <- as.factor(dataBI$AUTO.USE.CODE)
dataBI$POL.TP.CD <- as.factor(dataBI$POL.TP.CD)
dataBI$LimitType <- as.factor(dataBI$LimitType)
```

```
## removing zeros for logarithmic processing:
## Use only non-zero records

temp1 <- dataBI$AMT.INCD==0
data1 <- dataBI[ ! temp1 , ]

attach(dataBI)
detach(dataBI)
attach(data1)

##  Add state indicator variables to database

StateLabel <- c("CT","KY","ME","MD","NJ","OH","PA","VA")
State_<- factor(data1$NUM.ST.CD,levels =
sort(unique.default(data1$NUM.ST.CD)),labels=StateLabel)
StateCode <- model.matrix( ~ State_ - 1)
data2 <- cbind(data1,StateCode)
detach(data1)
attach(data2)


#Model1 <- StateCode;
OtherStates <- apply(StateCode[,-c(5,7)],1,sum)
OtherStates <- as.numeric(OtherStates)
Model2 <- cbind(StateCode[,c(5,7)],OtherStates)
NJ_PA_State <- apply(StateCode[,c(5,7)],1,sum)
NJ_PA_State <- as.numeric((NJ_PA_State))
Model3 <- cbind(NJ_PA_State,OtherStates)

data3 <- cbind(data2,Model2, Model3)
detach(data2)
attach(data3)


library(MASS)
library(survival)
library(eha)

temp1a <- Surv(AMT.INCD,amtincd.event)
temp1b <- Surv(log(AMT.INCD),amtincd.event)

#temp2a <- Surv(AMT.INCD,Limit.Event)
#temp2b <- Surv(log(AMT.INCD),Limit.Event)


data3$modyr <- lossyear - 1992
NormFit_Model1_Amtincd_BI <-
survreg(temp1b~factor(NUM.ST.CD)+data3$modyr,dist='gaussian',x=T,y=T,da
ta=data3)

StatForLNorm1 <- summary(NormFit_Model1_Amtincd_BI)
StatForLNorm1


#################   Output from R Starts From Here
######################
```

```
> StatForLNorm1

Call:
survreg(formula = temp1b ~ factor(NUM.ST.CD) + data3$modyr, data = 
data3, 
    dist = "gaussian", x = T, y = T)
                      Value Std. Error      z        p
(Intercept)          9.3340    0.07954 117.34 0.00e+00
factor(NUM.ST.CD)7  -0.9087    0.08818 -10.30 6.73e-25
factor(NUM.ST.CD)18 -1.1608    0.20075  -5.78 7.36e-09
factor(NUM.ST.CD)19 -1.1738    0.08562 -13.71 8.88e-43
factor(NUM.ST.CD)29 -0.2957    0.08142  -3.63 2.82e-04
factor(NUM.ST.CD)34 -1.3825    0.12111 -11.41 3.54e-30
factor(NUM.ST.CD)37 -0.5088    0.08139  -6.25 4.08e-10
factor(NUM.ST.CD)45 -1.3269    0.09132 -14.53 7.80e-48
data3$modyr          0.0195    0.00217   8.95 3.60e-19
Log(scale)           0.3793    0.00453  83.71 0.00e+00

Scale= 1.46 

Gaussian distribution
Loglik(model)= -47703   Loglik(intercept only)= -48368.9
        Chisq= 1331.75 on 8 degrees of freedom, p= 0 
Number of Newton-Raphson Iterations: 3 
n= 27607 
```

Appendix E

General description:

The initial univariate models effort considered six response ("dependent") variables using State and Loss Year as covariates ("independent variables"). The six response variables are report lag, settlement lag, and grouped claim size for both Collision and Auto BI. The models were run using software that allows a limited number of observations[40]. This is the reason that the claim sizes are grouped. State was chosen as a covariate because the legal system varies by state, and loss year was chosen as a measure of trend.

To determine optimal sets of covariates, eight models using different combinations were run for each response variable. The models are designated accordingly:

State Grouping used	Loss Year Included	Loss Year Not Included
Each state as a separate level (nine levels)	Model 1	Model 2
NJ, PA, and all other (3 levels)	Model 3	Model 4
NJ/PA combined, all other (2 levels)	Model 5	Model 6
State not used	Model 7	Model 8

State is included as a categorical variable, while Loss Year is a numeric variable. New Jersey and Pennsylvania contain the largest number of records in the database by far.

In the models, the categorical variables are converted to indicator (dummy) variables. We designate the state indicators as X_{CT}, X_{DE}, X_{KY}, X_{ME}, X_{MD}, X_{OH}, X_{NJ}, X_{PA} and X_{VA} , respectively. In fitting the models, the last state indicator variable (X_{VA}) is dropped because the constant 1 is automatically set up as the first covariate. Depending on the model in question, the Loss Year is either the actual loss year or the actual loss year minus a constant.

Models 1 through 8 were fitted using a maximum likelihood estimate to either a Weibull or Lognormal distribution. More precisely, the procedures fit the natural log (*ln*) of the response variable to a "least extreme value" or Gaussian distribution, respectively.

Least Extreme Value Distribution.

[40] The maximum number of observations is 65,535 for the Micosoft Excel software used.

The "least extreme value" distribution is the distribution for the natural log of a Weibull random variable. If we define $\boldsymbol{X}$ to be a 10-column matrix with columns:

$\boldsymbol{X}_0 = 1$;

$\boldsymbol{X}_1$ through $\boldsymbol{X}_8$**:** the first eight state indicators from above;

$\boldsymbol{X}_9$: the loss year;

and define **b** to be a vector length 10, with each respective b_j corresponding to $\boldsymbol{X}_j$, then the coefficients b_0 through b_9 and σ are chosen to maximize the loglikelihood function of **μ** and σ, where $\boldsymbol{\mu} = \boldsymbol{Xb}$. This is accomplished by first standardizing the natural log of the data: $W_i = \frac{\ln(Y_i) - \mu_i}{\sigma}$, where Y is the response variable. The resulting loglikelihood contribution of an observation, for the Least Extreme Value distribution, depends on whether the observation is censored:

$l_i = -\exp(W_i)$ if the observation is censored at Y_i.

$l_i = W_i - \exp(W_i) - \ln(\sigma)$ if uncensored.

$l = \sum n_i l_i$ **,** where n_i is the number of observations if the data is grouped.

<u>Deciding among Models.</u>

In deciding on the best model among the various covariate combinations, we use either the AIC criterion or, if models in question are nested, a likelihood ratio T-test.

We next apply the Weibull, Exponential, and lognormal models using the chosen covariates. The best model type is chosen using the AIC or the likelihood ratio test. The final model uses the best model type with the chosen set of covariates.

The subsection numbers below match those in the Section 2 of the paper.

2.2.1 Report Lag for Auto BI.

Report lag is uncensored. The observations for which report lag = 0 are removed, since the model fits $Y = ln$(Report lag). Models 1-8 are fit using the Weibull distribution, with the coefficients and other results summarized:

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
1	1.218463	1.069081	1.444614	1.295364	1.43744	1.292642	1.971992	1.822052
X_{CT}	0.13164	0.247981						
X_{DE}	0.258141	0.268167						
X_{KY}	1	1						
X_{ME}	0.138545	0.259498						
X_{MD}	0.272092	0.276301						
X_{OH}	0.915405	1.01844						
X_{NJ}	1.020659	1.036112	0.803555	0.808617				
X_{PA}	0.62727	0.644415	0.409933	0.416818				
Loss year	-0.02585		-0.02779		-0.02697		-0.02891	
X_{PA_NJ}					0.616817	0.622545		
Log likelihood	-49221.1	-49240.5	-49238.5	-49261.2	-49357.5	-49378.9	-49526.1	-49550.4
Parameters	10	9	5	4	4	3	2	1
AIC	98462	98499	98487	98530.4	95723	98763.9	99056.2	99102.8

In this model, the Loss Year is the actual Loss Year minus 1995. The coefficient for KY is 1 because there are no Kentucky records. Model 1 is best using the AIC criteria. This model also comes out the best in the "nested model" tests. For example, the likelihood ratio statistic for testing model 1 versus model 3 equals 2*(98499-98442) = 114, with 5 degrees of freedom. This is clearly significant, leading to rejection of Model 3, the less complex model.

We have so far selected the model with each state and loss year as covariates. We now select a distribution type. The table below compares the total loglikelihood and the AIC for the three model types that are currently programmed into the Claim Simulator.

Step 2 -- choose the best model from different distributions:

Models	Total Loglikelihood	AIC
Weibull	-49221	98462
Exponential	-66477	132974
lognormal	-41904	83828

It is important to note that in comparing loglikelihoods between the Weibull and lognormal models, the response variables must be on the same scale. The loglikelihood for the lognormal model above is calculated by fitting *ln*(report lag) to the Gaussian (i.e., Normal) distribution. The Weibull model, therefore, must fit *ln*(report lag) to the Least Extreme Value distribution. The Exponential model is a Weibull model with $\sigma = 1$.

We select the **lognormal model** because it has the lowest AIC. The coefficients of the selected model are given below.

Coefficients of final lognormal model[41] for Report Lag for Auto BI:

1	X_{CT}	X_{DE}	X_{KY}	X_{ME}	X_{MD}	X_{OH}	X_{NJ}	X_{PA}	Loss year
0.773	0.282	0.052	1.000	0.071	0.117	0.508	0.617	0.253	-0.02039

with σ = 1.3508.

2.2.2 Settlement Lag for Auto BI.

Settlement lag has both censored and uncensored values. If a claim is still open at 12/31/2006, the settlement lag is censored at a value equal to the difference between the report date and 12/31/2006. Models 1-8 are fit using the Weibull distribution. The step 1 coefficients and other results are summarized:

The table below summarizes the coefficients and other results for Model runs 1-8:

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
1	5.913837	5.65694	6.078474	5.811978	6.086181	5.809899	6.50678	6.234676
X_{CT}	0.39489	0.580192						
X_{DE}	0.486021	0.502767						
X_{KY}	1	1						
X_{ME}	-0.35192	-0.15005						
X_{MD}	-0.04979	-0.08378						
X_{OH}	0.142887	0.31611						
X_{NJ}	0.741003	0.776737	0.585875	0.619843				
X_{PA}	0.513409	0.534165	0.3582	0.376455				
Loss year	-0.04521		-0.04739		-0.04902		-0.05269	
$X_{PA\ NJ}$					0.047093	0.49835		
Log likelihood	-51211.2	-51483.2	-51391.3	-51691.8	-51586.1	-51908.2	-52083.9	-52446.8
Parameters	10	9	5	4	4	3	2	1
AIC	102442.4	102984.5	102792.5	103391.6	103180.2	103822.4	104171.8	104895.6

In this model, the Loss Year is the actual Loss Year minus 1995. The coefficient for KY is 1 because there are no Kentucky records. Model 1 is best using the AIC criteria. This model also comes out the best in the "nested model" tests. For example, the likelihood ratio

[41] The models exclude records with zero report lag, to avoid complications in taking logarithms. This means that the models overpredict the report lag slightly.

statistic for testing model 1 versus model 3 equals 2*(102792-102442) = 700, with 5 degrees of freedom. This is clearly significant, leading to rejection of Model 3, the less complex model.

We have so far selected the model with each state and loss year as covariates. We now choose a distribution type using the following results:

Models	Total Loglikelihood	AIC
Weibull	-51211	102442
Exponential	-51324	102670
lognormal	-55492	111003

The Weibull (via fitting *ln*(settlement lag) using the full set of covariates to the Least Extreme Value distribution) is the selected type. Coefficients for the final selected Weibull model are those listed in Model 1 table of coefficients above.

2.2.3 Report Lag for Collision.

The table below summarizes the coefficients and other results for Model runs 1-8:

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
1	3.906511	1.633738	7.756	1.471234	10.50934	1.471253	1.476665	1.487037
X_{CT}	-0.34968	-0.35122						
X_{DE}	-0.19921	-0.19879						
X_{KY}	1	1						
X_{ME}	0.314352	0.331701						
X_{MD}	-0.3236	-0.32434						
X_{OH}	-0.17691	-0.166						
X_{NJ}	-0.1167	-0.11678	0.04495	0.04491				
X_{PA}	-0.16016	-0.15996	0.00113	0.00168				
Loss year	-0.00114		-0.00314		-0.00452		0.00184	
X_{PA_NJ}					0.01807	0.018631		
Log likelihood	-157390	-157390	-157450	-157448	-157460	-157455	-157455	-157456
Parameters	10	9	4	3	3	2	2	1
AIC	314801	314797	314909	314901	314927	314914	314913	314914

The loss year variable equals the actual loss year unmodified. Note that the coefficient of loss year is insignificant for all models[42].

[42] Including loss year actually produces slightly lower loglikelihoods. For example, compare Model 5 to Model 6. This is a logical impossibility and is probably related to the optimization routine used in the spreadsheet.

Step 2 -- choose the best distributions using the covariates from Model 2:

Models	Total Loglikelihood	AIC
Weibull	-157389	314796
Exponential	-180703	361424
Lognormal	-131466	262950

Coefficients of final **lognormal** model[43] for Report Lag for Collision:

1	X_{CT}	X_{DE}	X_{KY}	X_{ME}	X_{MD}	X_{OH}	X_{NJ}	X_{PA}	Loss year
0.916	-0.097	-0.080	1.000	0.272	-0.140	-0.043	-0.006	-0.064	n/a

with σ = 1.0949.

2.2.4 Settlement Lag for Collision.

The table below summarizes the coefficients and other results for Model runs 1-8, using the Weibull distribution:

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
1	3.54444	3.243865	3.477206	3.150904	3.474883	3.150773	3.448188	3.144896
X_{CT}	-0.44328	-0.21589						
X_{DE}	-0.02388	-0.03551						
X_{KY}	1	1						
X_{ME}	-0.42354	-0.20413						
X_{MD}	-0.1933	-0.26813						
X_{OH}	-0.10243	0.090488						
X_{NJ}	-0.06988	-0.07354	0.002154	0.018864				
X_{PA}	-0.12179	-0.11536	-0.04993	-0.02294				
Loss year	-0.05262		-0.05349		-0.05314		-0.05288	
X_{PA_NJ}					-0.0298	-0.00696		
Log likelihood	-206369	-207254	-206433	-207362	-206460	-207379	-206466	-207380
Parameters	10	9	4	3	3	2	2	1
AIC	412758.6	414525.7	412873.6	414730.4	412925.4	414762.4	412935.1	414761.1

The loss year variable equals actual loss year minus 1995.

Step 2 -- choose the best model from different distributions:

[43] The models exclude records with zero report lag, to avoid complications in taking logarithms. This means that the models overpredict the report lag slightly.

Models	Total Loglikelihood	AIC
Weibull	-206369	412758
Exponential	-208502	417024
lognormal	-194502	389024

Coefficients of final model for Report Lag for Collision (Model 1 – lognormal)

1	X_{CT}	X_{DE}	X_{KY}	X_{ME}	X_{MD}	X_{OH}	X_{NJ}	X_{PA}	Loss year
2.965	-0.091	-0.113	1.000	-0.069	-0.278	-0.482	-0.087	-0.148	-0.04405

with σ = 1.110642.

2.2.5 Claim Size for Auto BI.

The table below summarizes the coefficients and other results for Model runs 1-8:

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
1	8.018206	8.190601	8.303077	8.407196	8.304788	8.407276	8.840928	8.924454
X_{CT}	1.32073	1.183873						
X_{DE}	0.417502	0.386438						
X_{KY}	0	0.000013						
X_{ME}	0.147621	0.00905						
X_{MD}	0.1525	0.119146						
X_{OH}	-0.00468	-0.15845						
X_{NJ}	1.026733	0.982928	0.772356	0.766713				
X_{PA}	0.813895	0.772005	0.560175	0.555989				
Loss year	0.019146		0.014664		0.663923		0.012215	
X_{PA_NJ}					0.014435	0.6591069		
Log likelihood	-125932	-125971	-126061	-126085	-126117	-126140	-126583	-126599
Parameters	10	9	4	3	3	2	2	1
AIC	251884	251961	252130	252176	252239	252283	253170	253200

Loss year = Actual loss year minus 1995.

Step 2 -- choose the best model from among different distributions:

Models	Total Loglikelihood	AIC
Weibull	-126206	252432
lognormal	-125932	251884

Coefficients of final model for Claim Size for Auto BI (Model 1 – lognormal)

1	X_{CT}	X_{DE}	X_{KY}	X_{ME}	X_{MD}	X_{OH}	X_{NJ}	X_{PA}	Loss year
8.018	1.321	0.418	0.000	0.148	0.153	-0.047	1.027	0.814	0.01915

with σ = 1.451282.

2.2.6 Claim Size for Collision.

All models in this section use the ground-up loss as their response variable. The ground-up loss is not right censored.

Models 1-8 were run fitting *ln*(grouped amt) to the **Normal** distribution:

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
1	-43.06066	1.0541514	-36.92875	7.649928	-32.4874	7.6498	-39.4234	7.831118
X_{CT}	0.252122	-1.189919						
X_{DE}	-0.010184	-0.942964						
X_{KY}	0.11826	-0.804627						
X_{ME}	-0.251515	0.521119						
X_{MD}	0.061232	0.244522						
X_{OH}	0.01697	-0.664438						
X_{NJ}	0.277003	-1.148186	0.255692	0.250249				
X_{PA}	0.225533	-1.140877	0.204388	0.200411				
Loss year	0.025344		0.022294		0.020079		0.023629	
X_{PA_NJ}					0.22447	0.221071		
Log likelihood	-521780	-522581	-521889	-522634	-521951	-522668	-522337	-523110
Parameters	11	10	5	4	4	3	3	2
AIC	1043581	1043788	1043788	1045276	1043910	1045341	1044680	1046225

Step 2 -- choose the best model from different distributions:

Models	Total Loglikelihood	AIC
Weibull	-544376	1088757
lognormal	-521780	1043581

Coefficients of final model for Claim Size for Collision (Model 1 – lognormal)

1	X_{CT}	X_{DE}	X_{KY}	X_{ME}	X_{MD}	X_{OH}	X_{NJ}	X_{PA}	Loss year
-43.061	0.252	-0.010	0.118	-0.262	0.061	0.017	0.277	0.226	0.02534

with σ = 0.87887

The results from the R model are not presented here because R apparently can not account for left-shifted and truncated response variables.

Figures 11 and 12 show that the model appears to fit reasonably well.

Appendix F: Plots for the loss Simulation Model

Figure 1: P-P Plots for Report Lag (Lognormal)

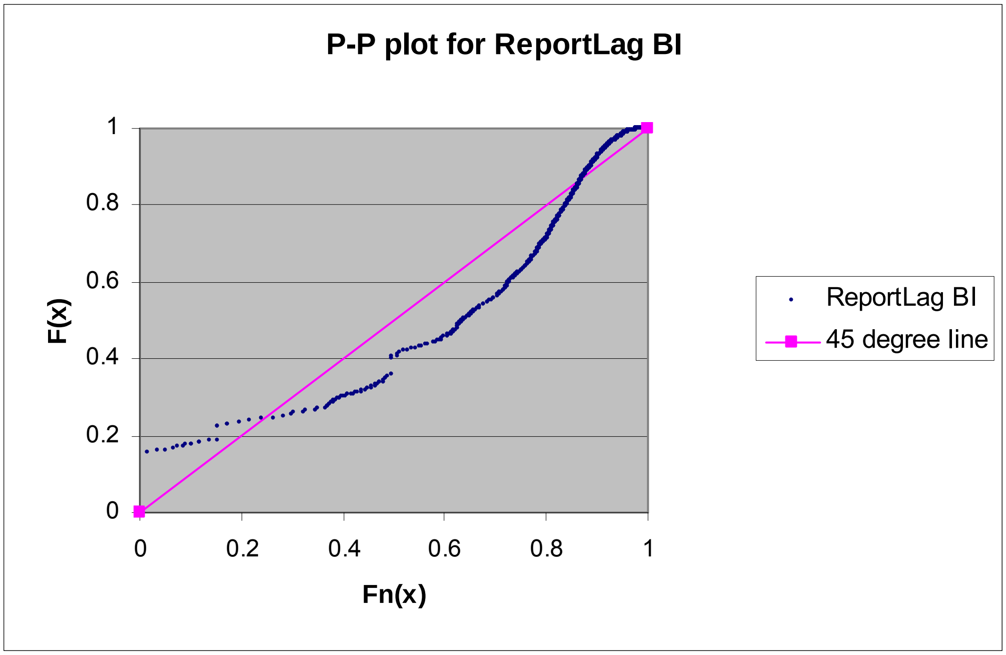


Figure 2: Standardized residual plot for Auto BI (Lognormal)

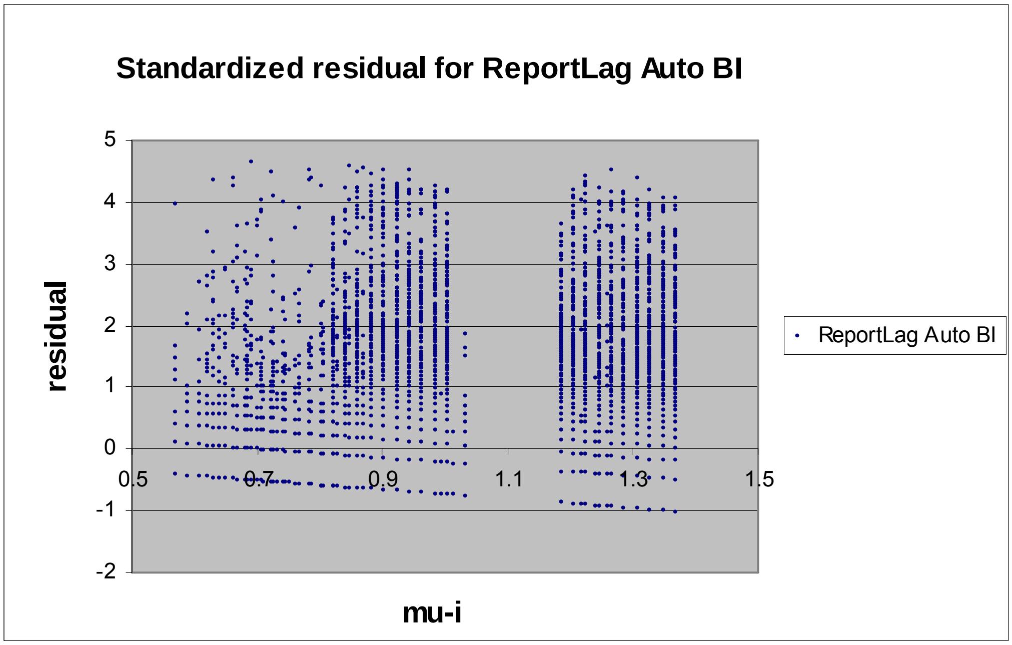


Figure 3: Plots for the Settlement Lag Auto BI (least extreme value)

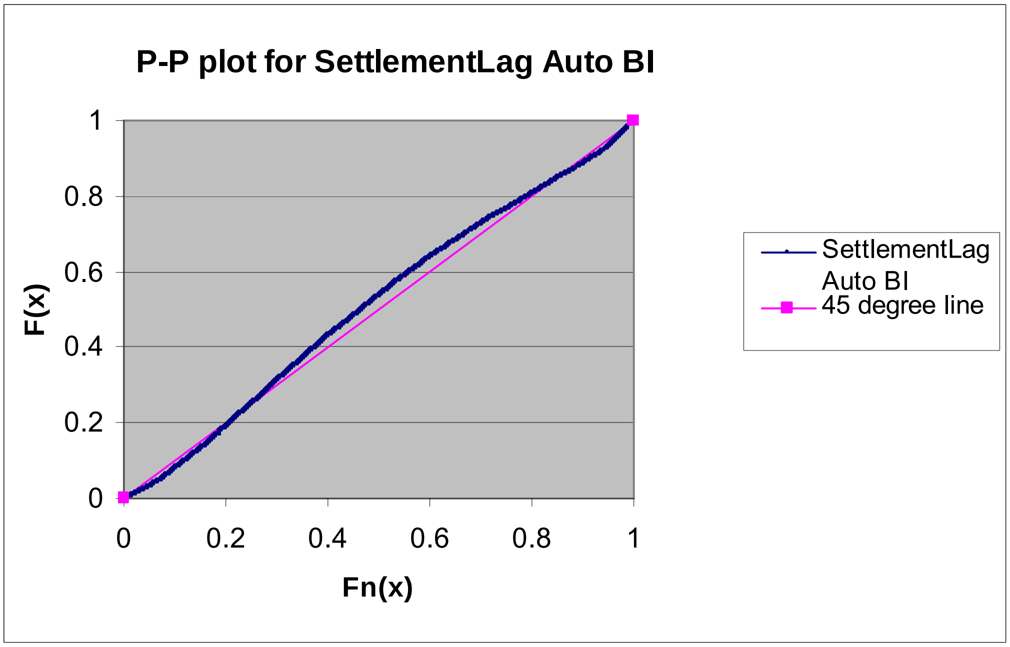


Figure 4: Standardized Residual for Settlement Lag Auto BI

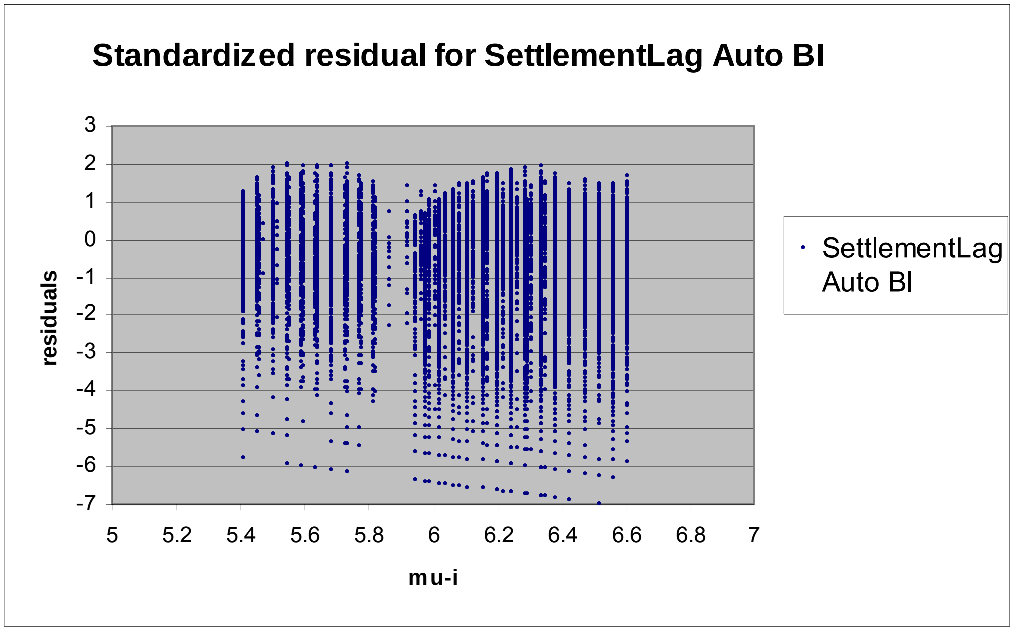


Figure 5: P-P Plot for Report Lag Collision (Lognormal)

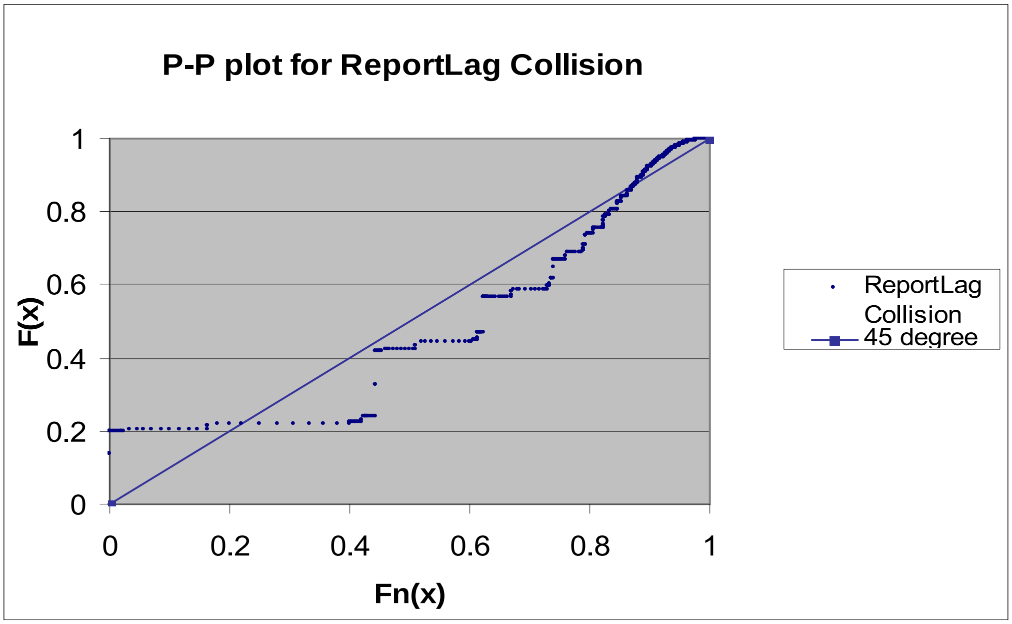


Figure 6: Standardized Residual for Collision (Lognormal)

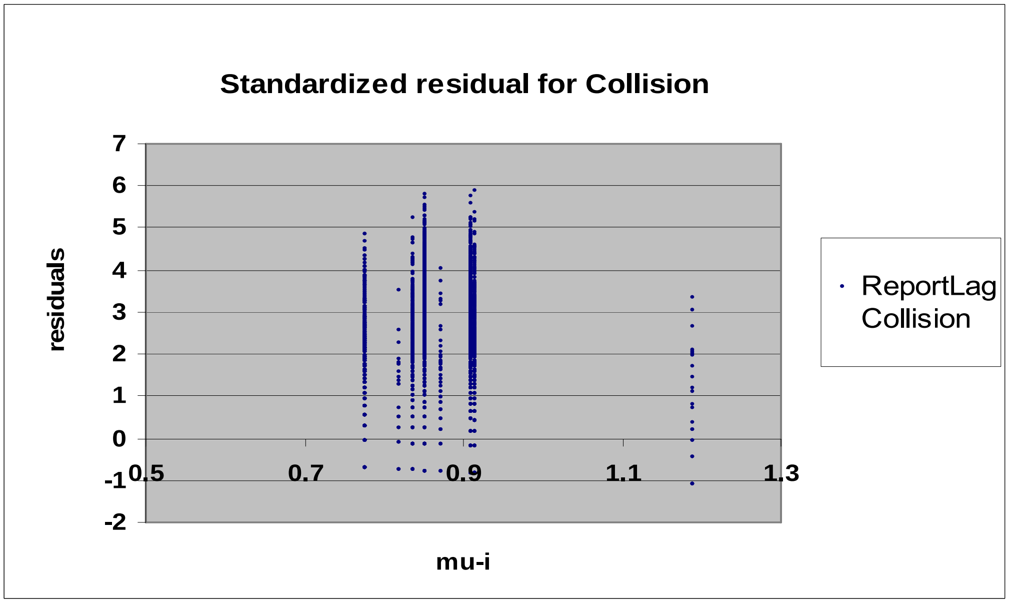


Figure 7: P-P Plot for Settlement Lag Collision (Lognormal)

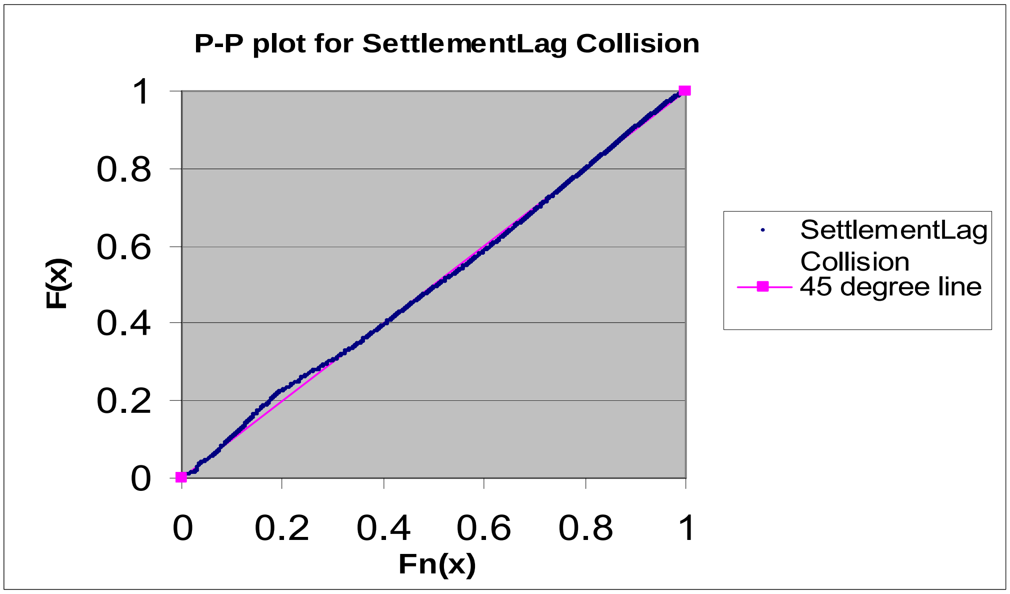


Figure 8: Standardized Residual for Settlement Lag Collision (Lognormal)

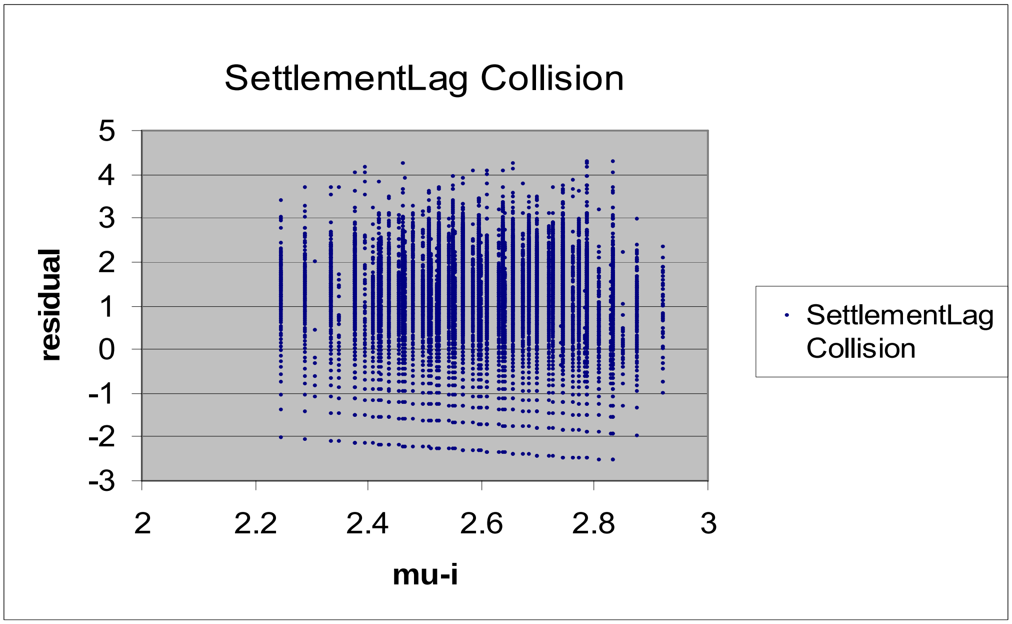


Figure 9: P-P Plot Claim Size for Auto BI (Lognormal)

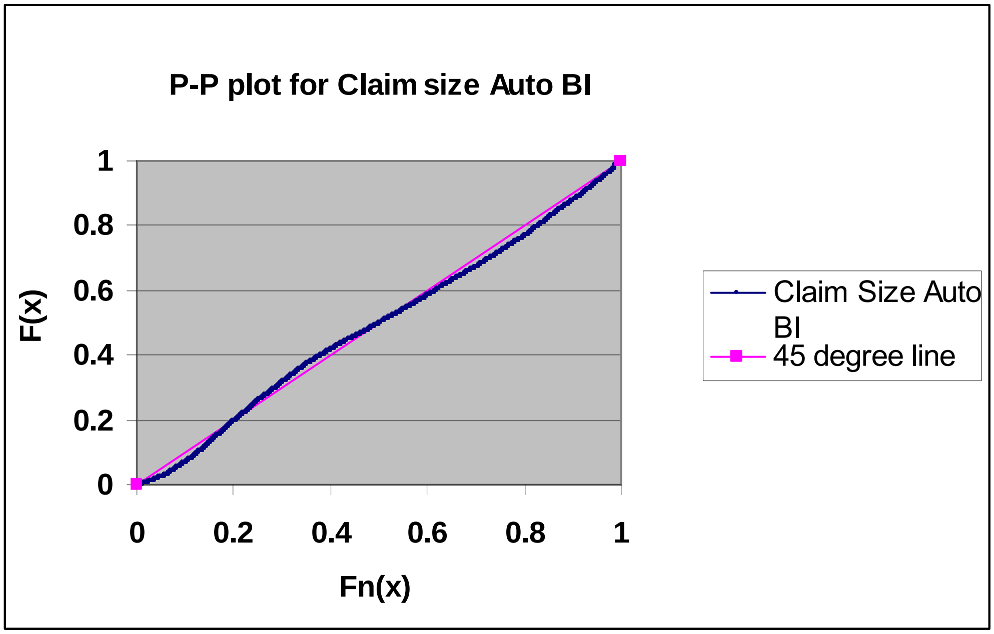


Figure 10: Standardized residual for Claim Size Auto BI

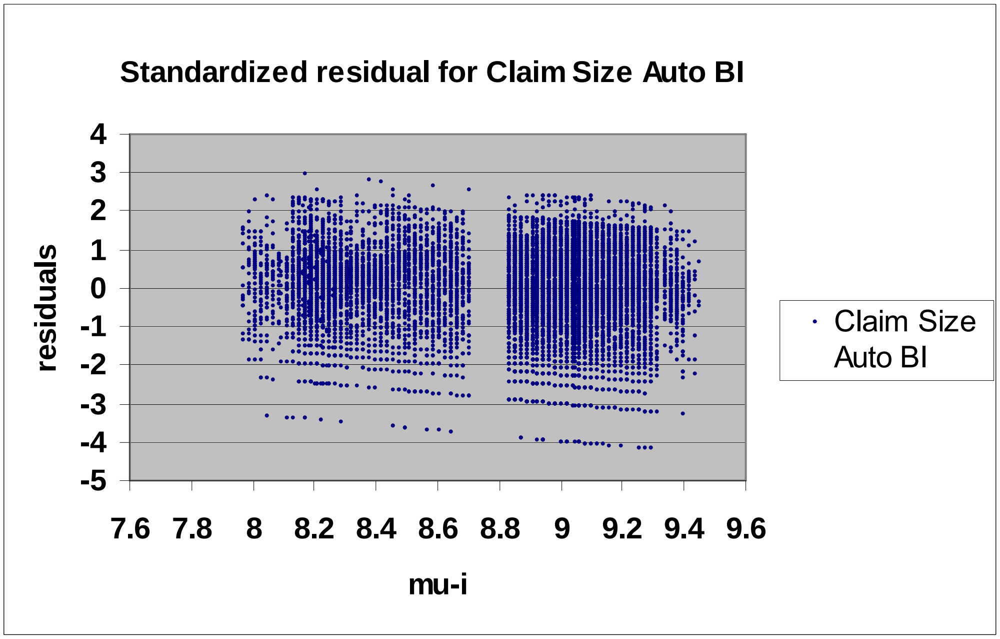


Figure 11: P-P Plot for Claim size Collision (Lognormal)

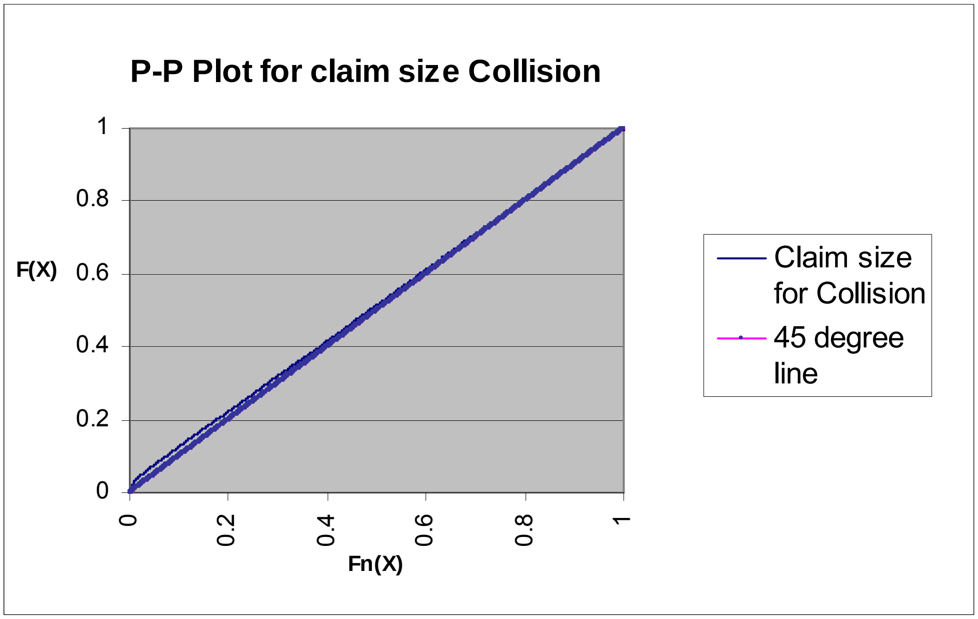


Figure 12: Standardized residual for Claim size Collision (Lognormal)

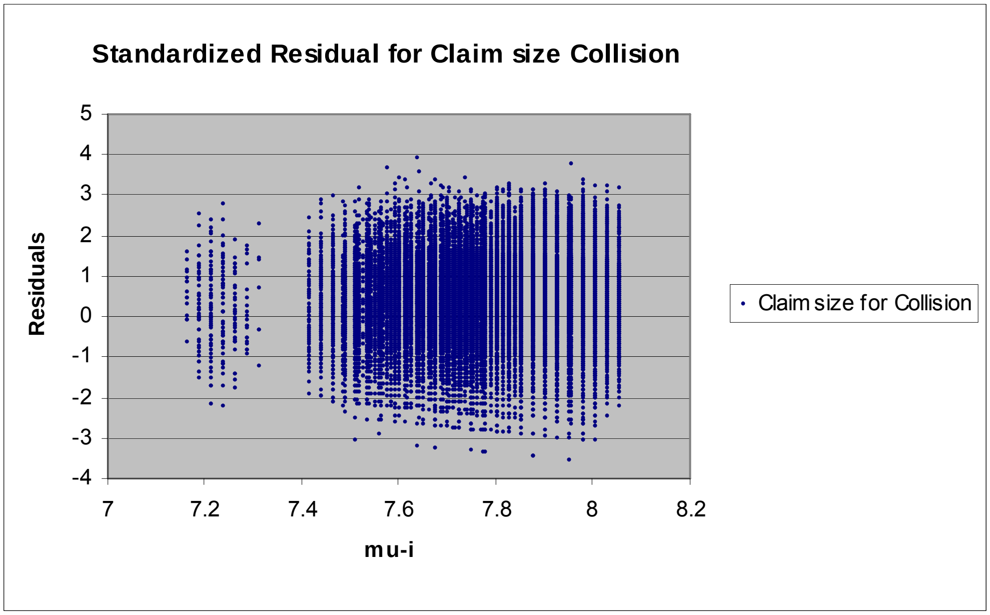


Figure 13: 3D surface plot for Settlement Lag and Claim Size (Auto BI)

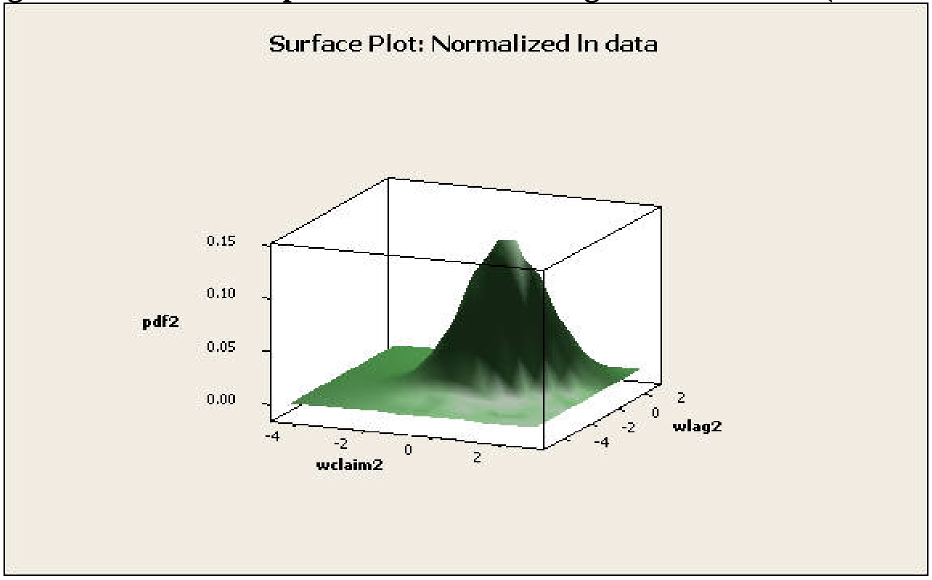
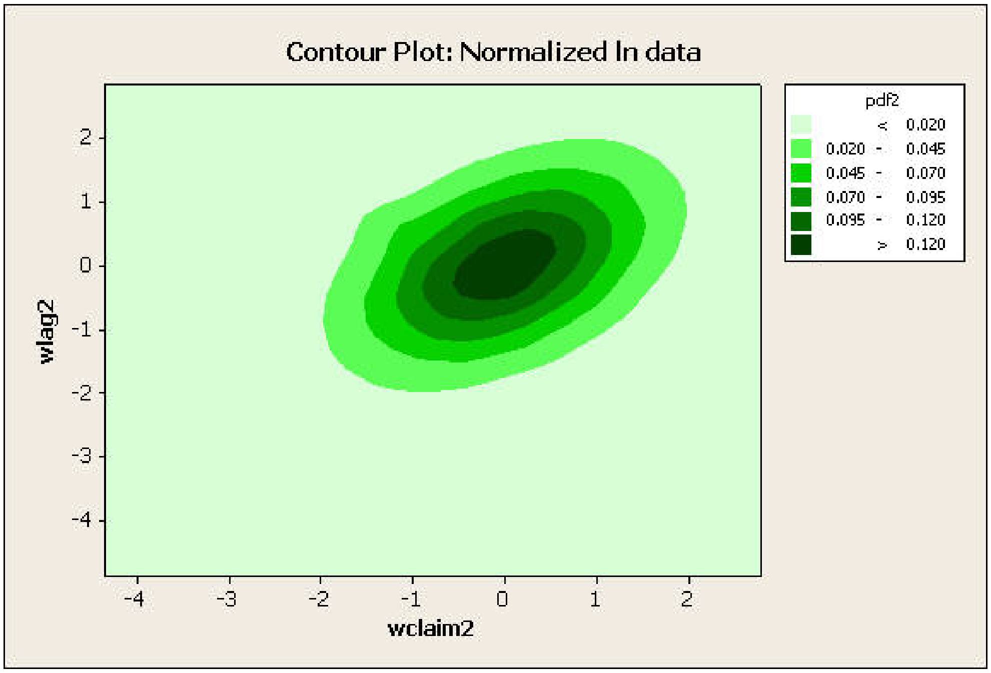


Figure 14: Contour plot for Settlement Lag and Claim Size (Auto BI)


Contour Plot: Normalized ln data
wlag2
wclaim2
pdf2
< 0.020
> 0.120