Loss Simulation Model Working Party
Parameterize and Test Model

Presented by Joseph O. Marker

Marker Actuarial Services, LLC

and

University of Michigan

2008 CAS Spring Meeting

Simulation Model

The Simulator

Simulates individual claims with report and accident dates.

By claim, simulates opening, interim, and closing transactions.

Saves the transactions to a file.

Summarizes to accident period x development age.
(Report period option also available)

Key Parameters of Simulator

Exposure by time period

Frequency distribution

Report Lag

Intervaluation times

Settlement Lag

Size of Loss

Probability of closure without payment.

Questions:

First Question:
Given a set of detailed data, is the model "rich enough" to simulate
  the data?

Translates to:
Can we fit the data well to the Simulator's supported forms of the
  model parameters and distributions?

Second Question:
Can we test the model using testing methods previously developed
  by the LSMWP?

Fitting the data

Look at now at the first question.
  Can we fit the data well to the distributions supported by the
  model?

An anonymous Source supplied Automobile data.

Students at Ball State University:
  Did the work to summarize the data.
  Parameterized the Simulation model.

Fitting the data (continued)

Limitations:

Students learned predictive modeling while doing the project.

Students did not know a statistical language such as "R".

Model fitting using maximum likelihood was done in Excel .

Due to software limitations, some data was grouped before fitting.

One group of 3 students learned "R" and verified the Excel calculations using R.

Fitting the data (continued)

First step was Data summary:
  Loss information was summarized to a "flat file" at valuation date
  12/31/2006.
  Risk characteristics ("policy coding") were available for each claim.
  Exposure and frequency information was not available.
  Lines Automobile Collision and BI Liability were available.

Model variables investigated:
  Report Lag
  Settlement Lag
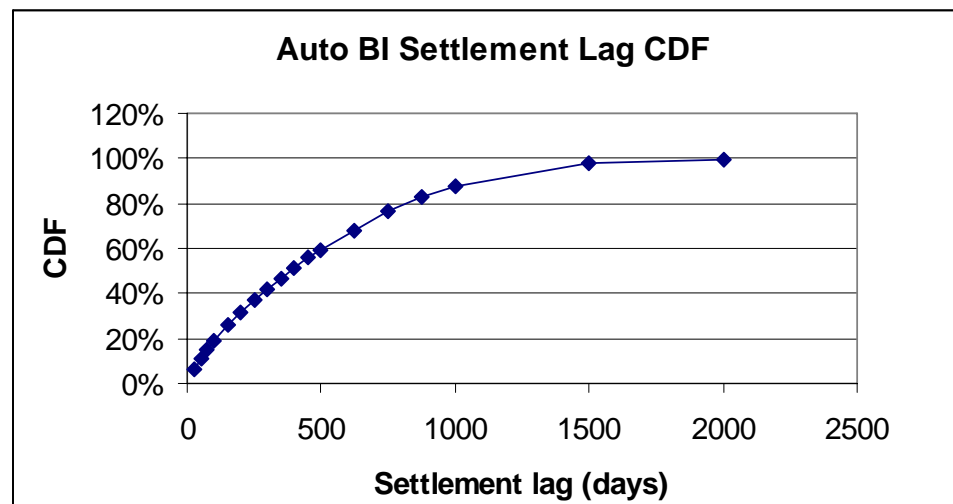 Claim Size, including probability that size = 0.

Used covariates State and Accident Year in the modeling.

Empirical Data: Auto BI lags

Report Lag: Mean = 12 days, Std. Dev = 65 days.
Accident year 2006 excluded to prevent report date truncation.

Settlement Lag: Mean = 485 days, Std. Dev = 418 days.

Auto BI Settlement Lag CDF

Auto Bodily Injury Claim Size

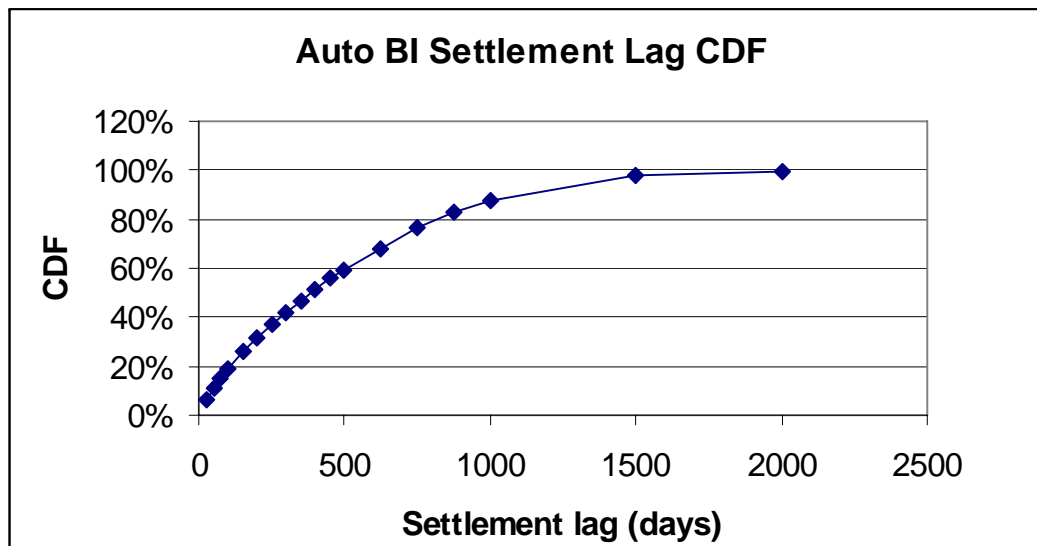| Size range	Number of Claims	Distribution
1 - 499	1,259	4.6%
500 - 1499	2,392	8.7%
1500 - 2499	1,345	4.9%
2500 - 3499	1,496	5.4%
3500 - 4999	1,784	6.5%
5000 - 9499	6,281	22.8%
9500 - 24499	8,672	31.4%
24500 - 49999	2,610	9.5%
50000 - 99999	1,276	4.6%
100000 - 199999	439	1.6%
200000 - 299999	53	0.2%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

Claim size distribution
(given that that a claim has non-zero value)

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

Empirical Data: Auto Collision lags

Report Lag: Mean = 3.6 days, Std. Dev = 22 days.
Accident year 2006 excluded.

Settlement Lag: Mean = 25 days, Std. Dev = 43 days.

Auto Collision Claim Size

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%

Claim size distribution
(given that a claim has non-zero value)

Fitting the Model.

Simulation Model supports the distributions:
  Weibull, Exponential, and Lognormal for lags.
  Weibull, Lognormal, and Pareto for claim size.

These models are not sufficient unless covariates are used.
Logical covariate choices:
  State (i.e. jurisdiction) – categorical.
  Accident Year – categorical or numeric.

Why pick State?
  Legal system varies by state.
  Large states New Jersey (NJ) and Pennsylvania (PA), are "No-Fault" States.

Use GLMs

Report Lag for Auto BI:

Not very interesting because report lag is so short for Auto.

Report lag (days) ~

Lognormal[$\mu = b_{\text{STATE}} - 0.02039$ (acc year $-$ 1995), $\sigma = 1.3508$]

Here b_{STATE} ranges from about 0.77 to 1.39.

Settlement Lag: for Auto BI:

Settlement Lag ~

Weibull [$\mu = b_{\text{STATE}} - 0.04521$ (acc year $-$ 1995), $\sigma = 0.93249$]

Here b_{STATE} ranges from about 5.56 to 6.68.

Model Claim Size X for Auto BI:

Multi-stage process:

1. Model $X \mid X>0$.

2. Model $\Pr[X=0]$ = Product of two factors:

   2a. Model $\Pr[X=0 \mid$ Claim is settled] and

   2b. Model $\Pr[$ Claim is settled $\mid$ Accident Year]

   Note: Step 2 assumes no unpaid claims of size 0.

1. *Model Y = X / X>0.*

Model selected:

$Y \sim$ Lognormal
[$\mu = b_{\text{STATE}} + 0.01915$ (acc year $-$ 1992),
$\sigma = 1.451282$].

Here b_{STATE} ranges from 7.97 to 9.34.

Model must recognize censoring by policy limit.

2. *Model Pr [X = 0].*

2a. Model Pr[X=0 | Claim is settled]
Logistic with

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * (\text{state} = \text{NJ or PA}) .$$

Result: $p = 0.483$ for NJ or PA, $p = 0.289$ otherwise.

2b. Model Pr[Claim is settled | Accident Year]
Logistic with

$$ln\left(\frac{p}{1-p}\right) = 15.187 - 1.0032 \, (\text{Acc year} - 1991)$$

Note: $p > .99$ for all years 2001 and prior.

2. *Model Pr [X = 0] (cont.).*

Then Pr[X=0] is the product of the two probabilities.

Pr[X=0] varies by accident year and
by whether the state is in { NJ,PA }.

Correlation.

Simulator allows correlation between settlement lag and claim size.

Estimate by fitting joint distribution to bivariate normal.

Correlation coefficient $\rho = 0.4289$.

Parametric survival models

These eliminate the need to construct likelihood functions.

Example: Auto BI claim size using software "R":

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

This fits *ln*(size) to Gaussian distribution.
"amtincd.event" indicates whether size is censored at policy limit.

Comments on Parameterizing

Paper on LSMWP website describes the modeling.

Comments on the modeling:
  This data is not useful for report lag modeling.
    Need a longer-tail line.
  This is not a loss development model,
    all data is at one valuation date.
  Auto Collision not discussed due to time limitation.
  Lognormal distribution with deductible
    was used to model Collision size of loss.

Simulation using selected model

Can the Simulator do this simulation?

We used only distributions supported by the Simulator,
and fit those using GLMs.

However, the means vary by the
covariates Accident Year and State.

State is a category variable.
Accident year is numeric

Simulation using selected model (cont.)

Simulator can support "types" within each "line".

How can we use this feature to include covariates?

Covariates were important to finding a decent fit.

How to test the model?

LSMWP set up testing methods.

These methods apply to loss development triangles.

This model is not a loss development model, but rather a model using values as of one valuation date.

How can this be tested?

We need volunteers to help with this.

Benefits of Volunteering:

Addresses an important problem to the CAS

Helps test new actuarial techniques

Helps experienced actuaries develop new skills to use in analyzing data.

Develops working relationships between academia and private companies.

References:

Parameterizing the Loss Simulation Model,
Available on the CAS LSMWP Working Party site.

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

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

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