Predictive Modeling: Frequency-Severity Models

Edward (Jed) Frees
University of Wisconsin-Madison, Australian National University

16 November 2024

Basic Terminology

- **Claim**: indemnification upon the occurrence of an insured event.
  - **Loss**: Some authors use "claim" and "loss" interchangeably, others distinguish between the amount suffered by the insured (loss) and the amount paid by the insurer (claim).
- **Frequency**: how often an insured event occurs, typically within a policy contract.
- **Severity**: amount or size of each payment for an insured event.

Sampling

- For each policy i, the observable responses are:
  - N_i: number of claims (events).
  - $y_{ij}, j = 1, \ldots, N_i$: amount of each claim (loss).
  - $S_i = y_{i1} + \cdots + y_{iN_i}$: aggregate claim amount.
- Depending on data availability, responses may include:
  1. Aggregate losses $\{S_i\}$.
  2. Both number and aggregate losses $\{N_i, S_i\}$.
  3. Detailed information about each claim $\{N_i, y_{i1}, \ldots, y_{iN_i}\}$.

Sampling Based Inference

- When individual claim data $\{N_i, y_{i1}, \ldots, y_{iN_i}\}$ is available:
  - Use conditional probability:

$$\mathrm{f}(N, \mathbf{y}) = \mathrm{f}(N) \times \mathrm{f}(\mathbf{y}|N)$$

    where:
    - $\mathrm{f}(N)$ models claim frequency.
    - $\mathrm{f}(\mathbf{y}|N)$ models conditional severity.
- Can use the same strategy when both the number and aggregate losses are available
- Note: No assumption of independence between frequency and severity.
- Other modeling options:
  - **Latent variables**: Affect both frequency and severity.
  - **Copulas**: Model non-linear dependencies.

Generalized Linear Model Strategy

- We now have two dependent variables. Can use the GLM strategy for each. Recall:
- Our typical situation is to consider n observations where, for the ith observation, y_i represents the insurance outcome, and $\mathbf{x}_i$ represents a vector of known rating (explanatory, predictor) variables.
- We choose a distribution that is common to all observations but allow the mean μ_i to vary by i (and sometimes other distribution parameters).
- We typically use a known function $\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$.
  - Here, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ is a vector of $k+1$ parameters.
  - Instead of n unknown means μ_i, we now have only $k+1$ unknown parameters.
  - We typically estimate the parameters using maximum likelihood.

Pricing Using the Mean

- For modeling purposes, let us focus on pricing. Hence, our main interest is the mean.
  - You can think about adding loadings for expenses and risk to this basic quantity to get a price.
  - This is the basis for personal line products, e.g., homeowners, auto.
  - Also provides the foundation for commercial lines (where risk loading and history/credibility take on a greater role).
- For motivation, think about the predictor variables as a single categorical variable:
  - Traditionally, rating variables have been categorical variables.
  - Estimation of the parameters is particularly simple (and thus easy to explain), sometimes only requiring spreadsheets.

Frequency-Severity Models

Two-Part Models

- In a two-part model, one part indicates whether an event (claim) occurs, and the second part indicates the size of the event.
- Let r_i be a binary variable indicating whether or not the i-th subject has an insurance claim, and y_i describes the amount of the claim.
- To estimate a two-part model:
  1. Use a binary regression model with r_i as the dependent variable and $\mathbf{x}_{1i}$ as the set of explanatory variables. Denote the corresponding set of regression coefficients as $\boldsymbol{\beta_1}$.
  2. Conditional on $r_i = 1$, specify a regression model with y_i as the dependent variable and $\mathbf{x}_{2i}$ as the set of explanatory variables. The gamma with a logarithmic link is a typical severity model.

Other Frequency-Severity Models

- For the second form, we have aggregate counts and severities $\{N_i, S_i\}$.
- The two-step frequency-severity model procedure:
  1. Use a count regression model with N_i as the dependent variable and $\mathbf{x}_{1i}$ as the set of explanatory variables.
  2. Conditional on $N_i > 0$, use a GLM with S_i/N_i as the dependent variable and $\mathbf{x}_{2i}$ as the set of explanatory variables.

Tweedie GLMs

- The Tweedie distribution is a Poisson sum of gamma random variables.
- It is used to model "pure premiums," where zeros correspond to no claims, and the positive part is used for the claim amount.
- The Tweedie distribution is a member of the linear exponential family with mean and variance:

$$\mathrm{E}\ S_N = \mu, \quad \mathrm{Var}\ S_N = \phi\mu^p$$

  where $1 < p < 2$.
- With a log-link, we have

$$\mu_i = \exp(\mathbf{x}_{i,T}' \boldsymbol{\beta}_{\boldsymbol{T}}).$$

Comparing the Tweedie to a Frequency-Severity Model

- As an alternative, consider a model composed of frequency and severity components:
  - Use a Poisson regression model for frequency:

$$N_i \sim \text{Poisson}(\lambda_i), \quad \lambda_i = \exp(\mathbf{x}'_{i,F}\boldsymbol{\beta}_F)$$

  - Use a gamma regression for severity:

$$y_{ij} \sim \text{Gamma}(\alpha, \gamma_i), \quad \frac{\alpha}{\gamma_i} = \mathrm{E}\ y_{ij} = \exp(\mathbf{x}'_{i,S}\boldsymbol{\beta}_S)$$

  - The aggregate loss, $S_{N,i} = y_{ij} + \cdots + y_{i,N_i}$, has mean:

$$\begin{aligned} \mathrm{E}\ S_{N,i} &= \mathrm{E}\ N_i \times \mathrm{E}\ y_{ij} \\ &= \exp(\mathbf{x}'_{i,F}\boldsymbol{\beta}_F + \mathbf{x}'_{i,S}\boldsymbol{\beta}_S), \end{aligned}$$

very similar to the Tweedie...

Additional Points of Emphasis

- In the chapter, you will find additional discussion of the concept of **exposure** and how to handle this in a GLM framework.
- Moreover, sometimes we only have available **grouped** data, rather than data based on individual contracts or units of analysis. This represents another complication in actuarial applications of GLM/statistical methodologies. . .
- The chapter also relates frequency-severity modeling to approaches used in related fields.
  - For example, health economists favor **Tobit** models for handling data with lots of zeros. . .

Massachusetts Automobile Claims

- Automobile insurance experience from the state of Massachusetts in 2006.
- Since the dataset represents experience from multiple carriers, the amount of policyholder information may be less comprehensive than typically used by larger carriers employing advanced analytic techniques.
- A random sample of 100,000 policyholders was drawn for the analysis.
- The study includes only bodily injury, property damage liability, and personal injury protection coverages.
  - These are compulsory and thus relatively uniform in Massachusetts.

Number of Policies by Rating Group and Territory

- The distribution of policies is reasonably level across territories.
- In contrast, the distribution by rating group is more uneven; for example, over three-quarters of the policies are from the "Adult" group.

Table 1: Number of Policies by Rating Group and Territory

Rating Group	Terr 1	Terr 2	Terr 3	Terr 4	Terr 5	Terr 6	Total
A - Adult	13905	14603	8600	15609	14722	9177	76616
B - Business	293	268	153	276	183	96	1269
I - Youthful with less than 3 years experience	706	685	415	627	549	471	3453
M - Youthful with 3-6 years experience	700	700	433	830	814	713	4190
S - Senior Citizens	2806	3104	1644	2958	2653	1307	14472

Averages by Rating Group

- The average total loss is 127.48.
- We observe important differences by rating group, where average losses for inexperienced youthful drivers are over 3 times greater than for adult drivers.

Table 2: Averages by Rating Group

Rating Group	Total Loss	Claim Number	Earned Exposure	Annual Mileage	Total Policies
A	115.95	0.040	0.871	12527	76616
B	159.67	0.055	0.894	14406	1269
I	354.68	0.099	0.764	12770	3453
M	187.27	0.065	0.800	13478	4190
S	114.14	0.038	0.914	7611	14472
Total	127.48	0.043	0.870	11858	100000

Averages by Territory

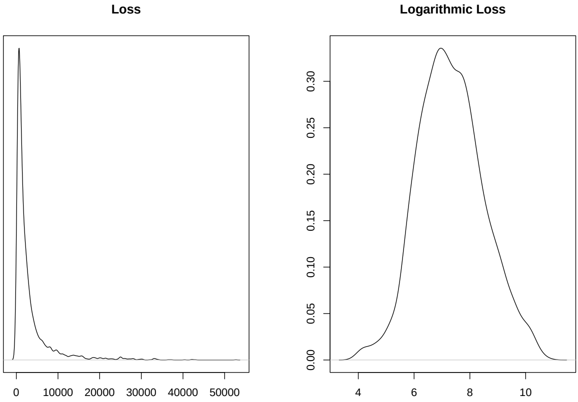
- The average total loss and the number of claims for territory 6 are about twice that for territory 1.

Table 3: Averages by Territory

Territory	Total Loss	Claim Number	Earned Exposure	Annual Mileage	Total Policies
1	98.24	0.032	0.882	12489	18410
2	94.02	0.036	0.876	12324	19360
3	112.21	0.037	0.870	12400	11245
4	126.70	0.044	0.875	11962	20300
5	155.62	0.051	0.866	10956	18921
6	198.95	0.066	0.842	10783	11764
Total	127.48	0.043	0.870	11858	100000

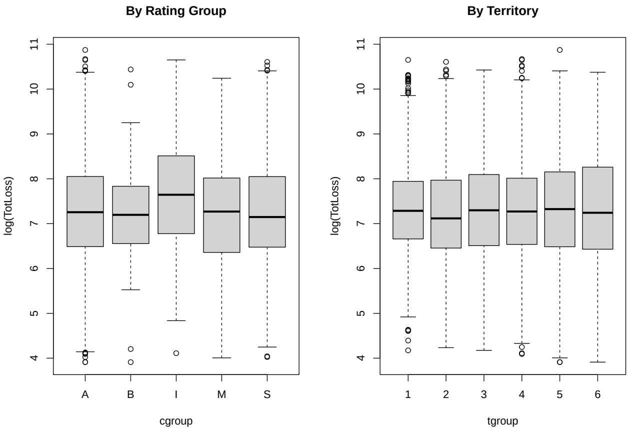
Loss Distribution

- The left-hand panel shows the distribution of losses, and the right-hand panel shows the same distribution on a logarithmic scale.



Logarithmic Loss Distribution by Factor

- The left-hand panel shows the distribution by rating group, while the right-hand panel shows the distribution by territory.



Participants now have an opportunity to explore these data on their own

Enjoy!