AXIS/ORMIR Student Challenge
Utility for Fitting Loss Distributions

Haoen CUI and Peng JIN

Department of Mathematics
University of Illinois at Urbana-Champaign

May 01, 2018

Agenda

AXIS's Demand

- Distribution Fitting
  - Coverage Modification
  - Changing Volume
  - Binned Data
- Documentation
  - Usable
  - Maintainable
  - Scalable

$\Longrightarrow$

Agenda

- Statistical Models
  - Coverage Modifications of Severity Random Variables
  - Coverage Modifications of Frequency Random Variables
  - CDF Interpolation
  - Modeling Loss Ratio via GLM
- Demonstration
- Documentation
- Conclusion

Coverage Modifications of Severity Random Variables

- Definition
  - L, raw loss size random variable
    - Likelihood: $f_L(l; \boldsymbol{\theta})$
  - Y, coverage modified loss size random variable
    - Likelihood: $f_Y(y; \boldsymbol{\theta})$
    - Dataset: $\mathcal{D}$

- Coverage Modifications
  - Deductible (ordinary/franchise)
  - Limit
  - Per loss / per payment
  - Coinsurance factor α
  - Interest rate r

Data Fitting Procedure for Coverage Modified Severity Random Variables

Parametric Model $f_L(l; \boldsymbol{\theta}) \xrightarrow{\text{modifications}} f_Y(y; \boldsymbol{\theta}) \xrightarrow{\text{generates}} \mathcal{D} = \{y_i\}_{i=1}^{n}$

$\xrightarrow{\text{est methods}} \widehat{\boldsymbol{\theta}}$

$\xrightarrow{\text{evaluate}}$ GoF Stats/Plots, Boot Conf Interval

Coverage Modifications of Frequency Random Variables

- Definition
  - N^L, raw loss count random variable
  - N^P, **deductible** modified **payment** count random variable

Theorem: N^L and N^P are from the same parametric family under conditions

Assume compound frequency model, $N^P = \sum_{i=1}^{N^L} \mathbb{1}\{L > d\}$ where N^L is parametrized by (α, β) and $\nu := \mathbb{P}(L > d)$. Let $G(\cdot)$ denote the probability generating function of a random variable. Then, *under some conditions*,

$$G_{N^P}(z) = G_{N^L}(z; \alpha^*, \beta^*)$$

where change of parameter is given by

$$\alpha = \mathbb{P}(N^L = 0) \qquad \alpha^* = \mathbb{P}(N^P = 0) \qquad \beta^* = \nu\beta$$

Coverage Modifications of Frequency Random Variables

Table 1: Examples for Change of Parameters Under Coverage Modifications

N^L with parameters $\boldsymbol{\theta} = (\alpha, \beta)$	N^P with parameters $\boldsymbol{\theta}^* = (\alpha^*, \beta^*)$
Poisson(λ)	$\lambda^* = \nu\lambda$
Zero-Modified Poisson(p_0^M, λ)	$p_0^{M*} = \frac{1-e^{-\lambda^*}}{1-e^{-\lambda}} p_0^M + \frac{e^{-\lambda^*}-e^{-\lambda}}{1-e^{-\lambda}}$ and $\lambda^* = \nu\lambda$

Data Fitting Procedure for Coverage Modified Frequency Random Variables

Parametric Model $(a, b, 0)$ or $(a, b, 1)$ $\quad p_{N^L}(n; \boldsymbol{\theta})$ and $\nu = \mathbb{P}(L > d)$

$\xrightarrow{\text{modifications}} p_{N^P}(n; \boldsymbol{\theta}^*) \xrightarrow{\text{generates}} \mathcal{D} = \{n_i\}_{i=1}^n$

$\xrightarrow{\text{est methods}} \widehat{\boldsymbol{\theta}}^* \xrightarrow{\text{transform}} \widehat{\boldsymbol{\theta}} \xrightarrow{\text{evaluate}}$ GoF Stats/Plots, Boot Conf Interval

Binned Data and CDF Interpolation

- Definition
  - L, raw loss size random variable
  - Dataset $\mathcal{D} = \{F_n(l_b)\}_{b=1}^{B}$ where $(l_b)_{b=1}^{B}$ is the set of pre-determined break points
- Goal
  - Find $\hat{F} \in \mathcal{F}$ based on data $\mathcal{D}$
  - where $\mathcal{F}$ is constrained such that its members are all valid distributions

Data Fitting Procedure for Global Parametric Family $\mathcal{F}_\theta$

Suppose $\mathcal{F}_\theta$ is a *global* (*not piecewise*) parametric family, then

$$\hat{F}_\theta = \underset{F_\theta \in \mathcal{F}_\theta}{\operatorname{argmin}} d(F_\theta, F_n)$$

where $d(\cdot,\cdot)$ is a distance measure between distributions (e.g., Kullback–Leibler divergence) and F_n is empirical distribution based on the observed dataset $\mathcal{D}$.

Binned Data and CDF Interpolation

In general, we may consider to interpolate the dataset $\mathcal{D}$ to get our estimate $\hat{F}$. However, we will need to ensure the constraint for arbitrary $\mathcal{F}$. There is no equivalent conditions for a general class of functions.

- Polynomials: *The Bernstein Form of a Polynomial* by Cargo and Shisha (1966) provides a sufficient condition for an increasing quantile function modeling by a polynomial of degree n
- Piecewise Pareto: Linear relationship between log(rank) and log(loss)

$$\forall l \geq l_0 \qquad \ln\Big(\mathbb{P}(L > l)\Big) = -\alpha \ln(l) + \alpha \ln(l_0)$$

- Hunter and Drown (2017) implemented some methods to interpolate CDFs
  - R package: `binsmooth`
  - Publication: *Better Estimates from Binned Income Data: Interpolated CDFs and Mean-Matching*, available on arXiv

Modeling Loss Ratio via GLM

- Tweedie distribution (μ, ϕ, p)
  - Poisson(λ) distributed sum of Gamma(shape$=\alpha$, scale$=\theta$) random losses
  - Tweedie mean, $\mu = \lambda\alpha\theta$
  - Tweedie poewr, $p = \frac{1}{\alpha+1} + 1$
  - Tweedie dispersion, $\phi = \frac{\lambda^{1-p}(\alpha\theta)^{2-p}}{2-p}$
- CAS Monograph *Generalized Linear Models for Insurance Rating* suggests to model loss ratio (aka pure premium) using GLM with Tweedie distribution
  - Note that Tweedie GLM implicitly assumes frequency and severity moves in the same direction

Coverage Modifications of Severity Random Variables

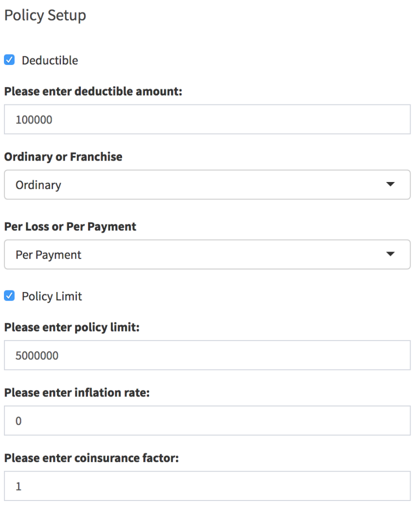

Figure 1: Policy Configuration

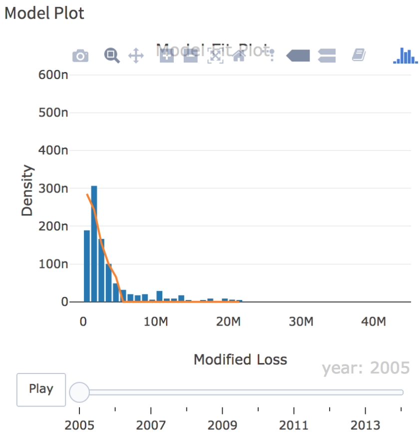

Figure 2: Model Plot with Estimated Parameters

Adding User-Defined Distributions

- Run User-Defined Distribution in *R Code*

```r
# R default: Define Distribution
source(helper_file.R)
dbeta
pbeta
rbeta
# helper.R: Coverage Modification
actuar::coverage(pdf = ddist, cdf = pdist,
                 deductible = deductible, limit = limit,
                 franchise = franchise, inflation = inflation,
                 coinsurance = coinsurance,
                 per.loss = per.loss)
# helper.R: Fit Distribution
fitdistrplus::fitdist(data = temp_data_raw, distr = dist.name,
                      method = method)
```

Adding User-Defined Distributions

- Run User-Defined Distribution in *R Shiny*

```r
# helper.R: Add to the list of distributions
Distribution.list <- list(
  "select.beta"=c(ddist = dbeta, pdist = pbeta,
                 rdist = rbeta, qdist = qbeta,
                 dist.name = "beta", dist.nparams = 2),
  ...)
# App new.R: Add to User Interface
box(title = "Fitting Specification",
    selectInput(inputId = "loss.dist",
                "Select distribution of raw loss random variab
                c("Beta"="select.beta", ...))
```

Summary of Documentations

- *User Manual*: Step-by-step procedure to run the utility
- *Statistics Manual*: Details of statistical assumptions, results, and procedures
- *Maintainability*
    - Users can add new loss distributions
    - Code formatted and commented under Google `R` style
    - Reproducibility ensured by `R` project and `packrat`
- *Scalability*
    - Modular and functional programming
    - Ability to add user-defined loss distributions

Model Constrains and Directions of Future Works

- Credibility framework (e.g., Bayesian)
- From parametric to non-parametric
- Cross Validation

Conclusion

AXIS's Demand

- Distribution Fitting
  - Coverage Modification
  - Changing Volume
  - Binned Data
- Documentation
  - Usable
  - Maintainable
  - Scalable

$\Longrightarrow$

Project Deliverables

- Statistical Models
  - Coverage Modifications
    - Severity Random Variables
    - Frequency Random Variables
  - CDF Interpolation
  - Modeling Loss Ratio via GLM
- Platform
  - Interface: `R Shiny`
  - Development: `R` project and `packrat`
- Documentation
  - User Manual
  - Statistics Manual

References

- G. E. Willmot, H. H. Panjer, S. A. Klugman. (2012). *Loss Models: From Data to Decisions, 4th Edition*.
- M. B. Finan. (2017). *An Introductory Guide in the Construction of Actuarial Models: A Preparation for the Actuarial Exam C/4*. Retrieved From Link.
- D. Bahnemann. (2015). *Distributions for Actuaries*. Retrieved From Link
- M. Goldburd, A. Khare, and D. Tevet. (2016). *Generalized Linear Models for Insurance Rating*. Retrieved From Link