Contributeurs

MAS-II: Modern Actuarial Statistics II

aut., cre. Alec James van Rassel

Référence (manuels, YouTube, notes de cours) En ordre alphabétique :
Contributeurs

Contents

A

Prerequisites

Distributions

Context

We typically use 3 types of random variable to describe losses:

Frequency or number of losses	always discrete
Severity or amount of losses (payment)	usually continuous, can be discrete or mixed too
Aggregate or total loss from summing a number (Frequency) of Severity variables	same as the severity

Discrete Distributions

Context

Discrete random variables are usually counting (frequency) variables, meaning their possible values are $\{0, 1, 2, \ldots\}$

Probability Mass Function (PMF)

N is a *discrete random variable* if it has a ***probability mass function*** p_k such that $\boxed{p_k = \Pr(N = k)}$

Definition	Domain	Condition
$p_k = \Pr(N = k)$	$p_k \in [0, 1]$	$\sum_k p_k = 1$

Poisson Distribution

Notation	Parameters	Domain
$N \sim \text{Poisson}(\lambda)$	$\lambda > 0$	$n = 0, 1, 2, \ldots$

$\Pr(N = n)$	$= \dfrac{e^{-\lambda}\lambda^n}{n!}$
$E[N]$	$= \lambda$
$Var(N)$	$= \lambda$

- If N_1 and N_2 are *independent* Poisson r.v., then $N_1 + N_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.
- The $e^{-\lambda}$ term makes the probabilities sum to 1 as the Taylor series for e^{λ} is

$$e^{\lambda} = 1 + \lambda + \frac{\lambda^2}{2!} + \cdots + \frac{\lambda^n}{n!} + \cdots$$

Binomial Distribution

Context

A binomial r.v. N has m *independent* trials each having a probability q of a loss where n is the total number of losses.

Notation	Parameters	Domain
$N \sim \text{Bin}(m,q)$	$q \in (0,1); m \in \mathbb{N}$	$n = 0,1,2,\ldots$

$\Pr(N=n)$	$= \binom{m}{n} q^n (1-q)^{m-n}$
$E[N]$	$= mq$
$Var(N)$	$= mq$

- If N_1 and N_2 are *independent* binomial r.v. with the ***same*** q then $N_1 + N_2 \sim \text{Bin}(m_1 + m_2, q)$.
- The case where $m = 1$ corresponds to a ***Bernoulli*** r.v.

Geometric Distribution

Context

A geometric r.v. N with mean β can be obtained by setting n as the number of years ***before*** the <u>first</u> loss. Given the geometric distribution is memoryless, each year *independently* has a loss with probability

$$\underbrace{\Pr(N=0)}_{\substack{\text{probability of a}\\\text{loss the first year}}} = \frac{1}{1+\beta}.$$

Notation	Parameters	Domain
$N \sim \text{Geo}(\beta)$	$\beta > 0$	$n = 0,1,2,\ldots$

$\Pr(N=n)$	$= \left(\frac{\beta}{1+\beta}\right)^n \frac{1}{1+\beta}$
$\Pr(N \geq n)$	$= \left(\frac{\beta}{1+\beta}\right)^n$
$E[N]$	$= \beta$
$Var(N)$	$= \beta(1+\beta)$

- Like the exponential distribution, the geometric distribution is memoryless:

$$\Pr(N = d + n | N \geq d) = \Pr(N = n)$$

$$E[N - d | N \geq d] = E[N]$$

Negative Binomial Distribution

Context

A negative binomial r.v. N represents the number of years n with no loss *before* the r^{th} year with a loss. We obtain a negative binomial r.v. $N \sim \text{NBin}(r, \beta)$ by summing r iid geometric r.v., $N_1, N_2, \ldots, N_r$, all with the same mean β.

Notation	Parameters	Domain
$N \sim \text{NBin}(\beta)$	$r, \beta > 0$	$n = 0, 1, 2, \ldots$

$\Pr(N = n)$	$= \binom{r+n-1}{r-1} \left(\frac{\beta}{1+\beta}\right)^n \left(\frac{1}{1+\beta}\right)^r$
$\Pr(N \geq n)$	$= \left(\frac{\beta}{1+\beta}\right)^n$
$E[N]$	$= r\beta$
$Var(N)$	$= r\beta(1+\beta)$

- A geometric r.v. is a negative binomial r.v. with $r = 1$.

Distribution	Mean		Variance
Binomial	mq	$>$	$mq(1-q)$
Poisson	λ	$=$	λ
Geometric	β	$<$	$\beta(1+\beta)$
Negative Binomial	$r\beta$	$<$	$r\beta(1+\beta)$

Severity Distributions

Joint Distributions

Conditional Distributions

Aggregate Distributions

Normal, Uniform, Pareto, Exponential, and Gamma

$\boxed{\gamma(1/2) = \sqrt{\pi}}$

Statistics

Mode

Context

The mode is the value that occurs the most often. A non-mathematical example of the concept is looking at the most used letter in the English alphabet. The letter E is the most used letter in the dictionary and as such is the mode of the English language.

In mathematical terms, the mode is the point which maximises the PMF/PDF.

Finding the mode of a continuous r.v. can be done by calculating the derivative of the PDF and finding the point where it equals 0. If the distribution is

- **unimodal**, i.e. it has a hump, then $\boxed{\text{mode} = x \text{ s.t. } f'(x) = 0}$.
- strictly increasing or decreasing, the mode will be one of the 2 extremes.
  - For example, the exponential distribution is strictly decreasing and its mode is always 0.

For discrete variables, there are some ways to simplify it's calculation:

- Using the table function on the calculator and seeing where the probabilities peak.
- Using the algebraic approach of looking at p_k/p_{k-1}.
  - $p_k > p_{k-1}$ iff $p_k/p_{k-1} > 1$.
  - The mode is the largest k s.t. $p_k > p_{k-1}$.

Note In the exam, it's best to use the calculator approach.

B

Introduction to Credibility

Basic Framework of Credibility

Context

The ***limitation fluctuation credibility*** approach, or ***classical credibility*** approach, calculates an updated prediction (U) of the **loss measure** as a weighted (Z) average of recent claim experience (D) and a rate (M) specified in the manual. Thus, we calculate the *premium* paid by the *risk group* as $\boxed{U = ZD + (1-Z)M}$.

Notation

M Predicted loss based on the "***manual***".

D Observe**d** losses based on the recent experience of the risk group.

Z Weight assigned to the recent experience D called the ***credibility factor*** with $\boxed{Z \in [0,1]}$.

U **U**pdated prediction of the premium.

Terminology

Risk group block of insurance policies, covered for a period of time upon payment of a *premium*.

Claim frequency The number of claims denoted N.

Claim severity The amount of the i^{th} claim denoted X_i.

Aggregate loss The total loss denoted S where $S = X_1 + X_2 + \ldots + X_N$.

Pure premium The pure premium denoted P where $P = S/E$ with E denoting the number of exposure units.

Exam tips

Typical questions about this involve being given 3 of M, D, Z, and U then finding the missing one.

Context

With $\min\{D, M\} \leq U \leq \max\{D, M\}$, we can see that the credibility factor determines the relative importance of the claim experience of the risk group D relative to the manual rate M.

If $Z = 1$, we obtain *Full Credibility* where the predicted premium depends only on the data ($U = D$). It follows that with $Z < 1$, we obtain *Partial Credibility* as the weighted average of both D and M.

Full Credibility

Contexte

The classical credibility approach determines the ***minimum*** *data size* required for the experience data (D) to be given ***full credibility***. The minimum data size, or ***standard for full credibility***, depends on the **loss measure**.

Claim Frequency

The claim frequency random variable N has mean μ_N and variance σ_N^2.
If we assume $N \approx \mathcal{N}(\mu_N, \sigma_N^2)$, then the probability of observing claim frequency **within** k **of the mean** is $\boxed{\Pr(\mu_N - k\mu_N \leq N \leq \mu_N + k\mu_N) = 2\Phi\left(\frac{k\mu_N}{\sigma_N}\right) - 1}$.

We often assume that the claim frequency $\boxed{N \sim \text{Pois}(\lambda_N)}$ and then apply the normal approximation to find the standard for full credibility for claim frequency $\boxed{\lambda_F}$. First, we impose that the probability of the claim being with k of the mean must be at least $1-\alpha$. Then, we rewrite $\boxed{\frac{k\mu_N}{\sigma_N} = k\sqrt{\lambda_N}}$ and set $\boxed{\lambda_N \geq \left(\frac{z_{1-\alpha/2}}{k}\right)^2}$ where $\boxed{\lambda_F = \left(\frac{z_{1-\alpha/2}}{k}\right)^2}$.

Claim Severity

We assume that the loss amounts $X_1, X_2, \ldots, X_N$ are independent and identically distributed random variables with mean μ_X and variance σ_X^2. Full credibility is attributed to $\boxed{D = \bar{X}}$ if $\boxed{2\Phi\left(\frac{k\mu_X}{\sigma_N/\sqrt{N}}\right) - 1 \geq 1 - \alpha}$.

Similarly to claim frequency, we apply the normal approximation with $\boxed{\bar{X} \approx \mathcal{N}\left(\mu_X, \sigma_X^2/N\right)}$. Then, we find $\boxed{N \geq \left(\frac{z_{1-\alpha/2}}{k}\right)^2 \cdot \left(\frac{\sigma_X}{\mu_X}\right)^2 = \lambda_F CV_X^2}$ where the ***standard for full credibility for claim severity*** is $\lambda_F CV_X^2$.

Aggregate Loss

For the aggregate loss $S = X_1 + X_2 + \ldots + X_N$, we have $\boxed{\mu_S = \mu_N \mu_X}$ and $\boxed{\sigma_S^2 = \mu_N \sigma_X^2 + \mu_X^2 \sigma_N^2}$.

With the same normality assumptions for the Poisson distributed N, we find $\boxed{\lambda_N \geq \left(\frac{z_{1-\alpha/2}}{k}\right)^2 \cdot \left(\frac{\mu_X^2 + \sigma_X^2}{\mu_X^2}\right) = \lambda_F(1 + CV_X^2)}$ where the ***standard for full credibility for claim severity*** is $\lambda_F(1 + CV_X^2)$.

Note The conditions are the same for the ***Pure Premium*** as for the aggregate loss.

Partial Credibility

The ***credibility factor*** for :

Claim Frequency is $\boxed{Z = \sqrt{\frac{\lambda_N}{\lambda_F}}}$.

Claim Severity is $\boxed{Z = \sqrt{\frac{N}{\lambda_F CV_X^2}}}$.

Aggregate Loss and Pure Premium is $\boxed{Z = \sqrt{\frac{\lambda_N}{\lambda_F(1+CV_X^2)}}}$

Bühlmann Credibility

Context

Buhlmann's approach, a.k.a. the greatest accuracy approach or the least squares approach, estimates the future loss measure X_n

Basic framework

Variance components

Credibility factors

Bayesian Credibility

Basic framework

Premium

Conjugate distributions

Nonparametric empirical Bayes method

C

Linear Mixed Models

Context

What distinguishes a linear mixed model is that it may include both **fixed-effect parameters** and **random effects**. The mix of these gives the linear *mixed* model its name. Fixed-effect parameters describes the relationships of the covariates to the dependant variable for an *entire population*. Random effects are specific to clusters or subjects *within a population*. Random effects are thus directly used in modelling the random variation in the dependant variable at different levels of the data.

Fixed factors are categorical or classification variables for which all levels (conditions) that are of interest are included. Random factors can be thought of as being *randomly sampled* from a population of levels being studied. The text gives as an example the Dental Veneer case study where if we specified the tooth being sampled, selected teeth would become a fixed factor. This would however limit inferences by teeth rather than generalizing to ”teeth within a patient”.

The case studies use 3 types of data:

clustered The dependant variable is measured once per subject (unit of analysis), and the units are grouped into/nested within clusters of units.

- We can have data sets that are two-level (e.g. rat pup data set), three-level (e.g. classroom data), etc.
- For MAS-II, we shouldn’t have beyond three levels.

repeated-measures The dependant variable is measured more than once (on the same unit of analysis) across levels of a repeated-measures factor(s). (e.g. time, measurement conditions, etc.)

longitudinal The dependant variable is measured at several points in time for each unit of analysis.

- ***Clustered longitudinal*** data combines features of both. (e.g. Dental Veneer data set).
- Each unit is measured more than once, but those units of analysis are nested within clusters.

Context

These 3 are ***hierarchical*** data sets as the observations can be placed into levels of a hierarchy in the data.

Generally:

Level 1 most detailed level; subjects, repeated measures on the same unit of analysis.

Level 2 clusters of units, units of analysis.

Level 3 clusters of clusters, clusters of units.

Levels are emphasized in the text because they help to conceptualize LMM as simple models defined at each level of the data hierarchy.

General Theory

Residual Variance Structures

Diagonal

$$R_i = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

- Assumes residuals from the same subject are *independent*.

Compound Symmetry

$$R_i = \begin{bmatrix} \sigma^2+\sigma_1 & \sigma_1 & \dots & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \dots & \sigma_1 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \sigma_1 & \dots & \sigma^2+\sigma_1 \end{bmatrix}$$

- Assumes *equal correlation* between observations from the same individual.
- Good for clustered or repeated measures data.

First Order Auto-Regressive ($AR(1)$)

$$R_i = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \dots & \sigma^2\rho^{n_i-1} \\ \sigma^2\rho & \sigma^2 & \dots & \sigma^2\rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{n_i-1} & \sigma^2\rho^{n_i-2} & \dots & \sigma^2 \end{bmatrix}$$

- Good for longitudinal with *equal time* between observations.

Model Assumptions

1 **Fixed Effects**

2 **Random Effects**

Algorithms

Expectation Maximization

Newton-Raphson

Fisher Scoring Algorithm

Troubleshooting

Hypothesis Testing

Likelihood Ratio Tests

Mixture of Chi Squares
REML

Non-Likelihood Ratio Tests

t-test

approximating df, not n-p

F-test

Context

Degrees of freedom of the numerator correspond

We get that the test statistic $t \approx F_{\text{num. df,den. df}}$ where the numerator df corresponds to the number of parameters being tested and the denominator degrees of freedom is obtained from R.

The particularity of the F-test is that we must make adjustments for it due to:

1. Random Effects
2. Potential correlation between residuals
3. Estimate covariance matrix

We have a few ways of approximating them:

Scatterwhite

- Method used by R.

Kenward-Rogers

- Method used by SAS.

Type *I*

Sequential

Type *III*

Conditional

Using tests:

1. Compute test statistic
   - F-statistic would be too hard to compute, would have to be provided.
   - For t-test, may just give components of the calculation and have us compute t to the compare it to the CV.
   - For both tests, the number of df would have to be provided.
2. Look up critical value table
3. Reject null / keep effects if test statistic > CV

Other tests

Omnibus Wald Test (good) similar to F-test test statistic asymptotically χ^2
Wald z-test (not good) only good asymptotically and breaks in some situations text recommends LRT instead

EBLUPS

Intra Correlation Coefficient

$\boxed{ICC_{\text{whatever}} = \frac{\text{variance in common}}{\text{total variance}}}$.

2 level model $ICC_{\text{group}} = \frac{\sigma^2_{\text{lvl 2}}}{\sigma^2_{\text{lvl 2}}+\sigma^2}$ 3 level model $ICC_{\text{lvl 3 group}} = \frac{\sigma^2_{\text{lvl 3}}}{\sigma^2_{\text{lvl 3}}+\sigma^2_{\text{lvl 2}}+\sigma^2}$ $ICC_{\text{lvl 2 group}} = \frac{\sigma^2_{\text{lvl 3}}+\sigma^2_{\text{lvl 2}}}{\sigma^2_{\text{lvl 3}}+\sigma^2_{\text{lvl 2}}+\sigma^2}$

EBLUPS

EBLUP

E

B Best i.e. lowest variance among all such unbiased estimators

L Linear as functions of $\boldsymbol{y}_i$

U Unbiased with $\text{E}[\hat{\boldsymbol{u}}_i] = \boldsymbol{u}_i$

P

- Typically tedious to calculate so we use computers unless we calculate only for 1 random effect.

Use Buhlmann's formula where:

M Average predicted value from the implied marginal model

$\bar{Y}$ Average observed value from group

σ^2_{HM} $\text{Var}(u_j) = \sigma^2_{int}$

μ_{PV} $\text{Var}(\varepsilon_{ij}) = \sigma^2$

Prediction is for $M + u_j = M + Z_j(\bar{Y} - M)$.

Information Criteria

Context

When comparing 2 nested models, the more complex will be better than the simpler model. While the *Likelihood Ratio Tests* checks if the simpler model is sufficient, it does not enable us to directly compare the 2 models. In addition, with the LRT we are limited to nested models. The AIC and BIC measures permit us to compare several models which don't have to be nested. They do so by adding a penalty to the likelihood for a model's complexity via the amount of parameters it has.

We wish to maximize the likelihood of our observations. As observed for the LRT, maximizing the likelihood is equivalent to minimizing the loglikelihood or a function thereof. Namely, $-2 \times \ell(\theta)$ (a.k.a. the deviance, see *Graphical Tests*). In both cases, we add a penalty to the measure we wish to minimize.

Akaike Information Criteria (AIC)

The AIC penalizes models which have more parameters by adding twice the number of estimated parameters p in the model to twice the negative log-likelihood: $\boxed{AIC = -2\ell(\theta) + 2p}$.

We choose the model with the smallest AIC.

Context

The disadvantage of the AIC lies in that for 2 nested models the probability of choosing the simpler model knowing it's the true model does not tend towards 1 when the number of observations increases towards infinity. We thus consider it an *inconsistent* measure.

In comparison, the BIC **is** a *consistent* measure given its parameters penalty is a function of the number of observations.

That being said, in both cases, the probability of rejecting the simpler model while the true model is somewhere in between tends towards 1.

Bayesian Information Criteria (BIC)

The BIC penalizes more severely models which have more parameters given its penalty is a function of the number of observations n: $\boxed{BIC = -2\ell(\theta) + \ln(n)p}$.

To better understand the difference between the AIC and BIC penalty, we can use log rules to rewrite the measures:

$$\begin{aligned}
AIC &= -2\ln|\mathcal{L}(\theta)| + 2p \\
&= -2\ln|\mathcal{L}(\theta)| + \ln\left(e^{2p}\right) \\
&= -\left[\ln|\mathcal{L}(\theta)^2| - \ln\left|(e^p)^2\right|\right] \\
&= -\ln\left|\frac{\mathcal{L}(\theta)^2}{(e^p)^2}\right| \\
BIC &= -2\ln|\mathcal{L}(\theta)| + \ln|n|p \\
&= -\left[\ln\left|\mathcal{L}(\theta)^2\right| - \ln|n^p|\right] \\
&= -\ln\left|\frac{\mathcal{L}(\theta)^2}{n^p}\right|
\end{aligned}$$

Context

There's no agreement on which is better for LMM and the text tends to build models piecewise, testing between steps with LRT. So, we probably won't use them much.

Fundamentally, the AIC tries to to find the model that best describes the data under the belief that there is no "correct" model. In contrast, the BIC tries to find the "correct" model under the belief that such a model exists.

Intuitively, we may think we'd prefer the AIC given that it's typically unrealistic to believe there exists a "correct" model. However, some feel the BIC often gives better results. *However*, part C on *Graphical Tests* has other information criterion that are more complicated but arguably better.

Notes:

- REML criterion at convergence is the deviance ($-2\ell(\theta)$).
  - Likelihoods are typically < 1 given they're probability densities.
  - Thus, loglikelihoods are typically negative.
  - Thus a positive output suggests they already multiplied by -2.

Graphical Tests

Not heavily tested. Case study will have some graphs and there will be some questions about case study which may need graphs interpretation
marginal residual residual leftover plugging in estimated fixed effects rarely used
conditional (textbook) / response (R) / raw (typical) residuals residual from estimate of everything In LMM, variance of residual ε_{ij} can vary based on other factors So, still not residual we want
standardized / normalized residuals conditional residual / estimated SD for that residual almost always prefer standardized residual
important:

- Use residual plots for normality testing
- raw data plots are useless; ignore them.
- Standardized residuals adjust the data so we can tell if a residual is an outlier because it's from a high variance group or because it's really an outlier.

implied marginal model is LMM w/o random effects but with same variance structure (var Yij same for both)
marginal model is with just same variance for everything

D

Bayesian Analysis and Markov Chain Monte Carlo

E

Statistical Learning

K-Nearest Neighbors

Decision Trees

Principal Components Analysis (PCA)

Clustering