Advanced Non–Life Insurance Mathematics

Assignment 1: claim severity modeling

Katrien Antonio and Roel Verbelen, academic year 2016-2017, KU Leuven

Deliverables:

- a detailed report (approach, technicalities, conclusions);
- lay–out, presentation and writing style matter (and will be graded);
- graphs should be integrated in the text;
- programming code follows in an appendix and is submitted as a separate file (`.R` or `.txt`);
- be brief!

Practical matters: hand in your solution through the Assignment tool on TOLEDO. Provide a report (in `pdf`) and programming code (using `R`).

Assignment Questions

In the first part of the assignment, you will fit various parametric models to censored/truncated loss data and compare results. The file `SeverityCensoring.txt` contains information about 9062 claims paid by an insurance company over some observation period.

- The column `claimAmount` shows how much the insurance company paid on each claim. This is the loss data that we will use for model fitting.
- The column `deductible` shows that there is a fixed deductible of 100 EUR for each policy, which means that all the observed claim amounts are truncated from the left at 100.
- The column `rc` shows whether right-censoring is present or not: `NA` indicates that the claim is fully settled by the end of the observation period and therefore the observed claim amount is the full (uncensored) loss associated with this claim. On the other hand, a number in the `rc` column indicates that the claim is not yet fully settled, so the observed claim amount is right-censored.

1. We start by fitting an exponential distribution to the loss data. Compute the log-likelihood function for this model, taking truncation and censoring into account, and maximize it numerically to find the MLE of the rate parameter. Also compute the Akaike Information Criterion (AIC) for this model, for later comparison with other models.

2. Repeat exercise 1 for lognormal, inverse Gaussian and Burr distributions. That is, compute the log-likelihood function (under truncation and censoring) and find the MLEs of the parameters for each of these models. Also compute the AIC for each model.

   Note: For numerical maximization of log-likelihood functions with two or more parameters, you need to specify starting values for the parameters. If the starting values are too far from the true optimal values, the optimization algorithm may lead you to a false local maximum. To avoid this, try to pick "reasonable" starting values whenever you can, e.g. using method of moments. Also: the pdf and cdf of the inverse Gaussian and Burr distributions are provided in the `statmod` and `actuar` packages, respectively. Install and load those packages for this exercise.

3. Next, we want to fit an Erlang mixture distribution with 5 components to the loss data. Recall that this distribution has a pdf of the form

$$f(x;\alpha_1,\ldots,\alpha_5,r_1,\ldots,r_5,\theta)=\sum_{j=1}^{5}\alpha_j\frac{x^{r_j-1}e^{-x/\theta}}{\theta^{r_j}(r_j-1)!},\quad x>0.$$

   Since the usual maximum likelihood approach will not work for mixture distributions, we will make use of the EM (Expectation-Maximization) algorithm to estimate parameters. Thankfully, the `R` code for implementing the EM algorithm for Erlang mixtures was developed in Verbelen et al. (2015) and is available for our use. Download the file `2014-12-16_ME.R` into your working directory and run:

```
source("2014-12-16_ME.R")
loss <- ... ; nrc <- ...
fit.ME <- ME_fit(loss, nrc, trunclower=100, M=5, s=3)
```

   Fill in the dots so that `loss` is the vector of 9062 losses as provided in the `claimAmount` column, and `nrc` is the same as `loss`, but with censored loss amounts replaced by `NA`'s. This will take a few minutes to run, and will produce some warning messages that you can ignore.

   Inspect the object `fit.ME`, identify the MLEs for the five weights $\alpha_1,\ldots,\alpha_5$, the five shape parameters $r_1,\ldots,r_5$, and the scale parameter θ. Also identify the AIC for this model.

4. Plot the Kaplan-Meier estimate of the survival function for the loss data, by installing the package `survival` and then running

```
deds <- ... ; loss <- ... ; full <- ...
fit <- survfit(Surv(deds, loss, full) ~ 1)
plot(fit, mark.time=F, conf.int=F)
```

Fill in the dots so that `deds` is the vector of the 9062 deductibles (all equal to 100) as provided in the `deductible` column, `loss` is the vector of the 9062 losses as provided in the `claimAmount` column, and `full` is a logical vector of length 9062 that has a `TRUE` for non-censored (full) losses and `FALSE` for censored losses.

Note: For more information regarding survival analysis in `R` using the `survival` package, you can check out this online tutorial.

5. Add the plots of the best-fitting (i) exponential, (ii) lognormal, (iii) inverse Gaussian, (iv) Burr and (v) Erlang mixture survival functions to the Kaplan-Meier plot. Recall that we have a left-truncation at 100, so you should plot the curve $\frac{1-F(x)}{1-F(100)}$ for each of the five models, with F denoting the cdf with the best-fitting parameters. Which of the five parametric models seems closest to the Kaplan-Meier estimate?

   Note: The mixed Erlang cdf can be computed as `ME_cdf(x, alpha, shape, theta)`.

6. Compare the AIC values for the five parametric models considered above. Which model gives the best fit according to AIC? Is this consistent with your answer to exercise 5?

In the second part of the assignment, you will use splicing to fit a body-tail combination model to the Secura Re loss data, using a shifted exponential distribution for the body and a Pareto distribution with unit scale for the tail, as demonstrated in class. The data set is available in the file `SecuraRe.txt`. Recall the spliced pdf that was derived in class:

$$f(x) = \begin{cases} \frac{n-k}{n} \cdot \frac{\lambda \exp\{-\lambda(x-1{,}200{,}000)\}}{1-\exp\{-\lambda(X_{n-k,n}-1{,}200{,}000)\}} & x \le X_{n-k,n} \\ \frac{k}{n} \cdot \frac{\alpha(x+1)^{-\alpha-1}}{(X_{n-k,n}+1)^{-\alpha}} & x > X_{n-k,n} \end{cases},$$

where n is the sample size and $X_{n-k,n}$ denotes the $(k+1)^{\text{th}}$ largest observation in the data set. We also derived the cdf for this distribution:

$$F(x) = \begin{cases} \frac{n-k}{n} \cdot \frac{1-\exp\{-\lambda(x-1{,}200{,}000)\}}{1-\exp\{-\lambda(X_{n-k,n}-1{,}200{,}000)\}} & x \le X_{n-k,n} \\ 1 - \frac{k}{n}\left(\frac{x+1}{X_{n-k,n}+1}\right)^{-\alpha} & x > X_{n-k,n} \end{cases}.$$

7. Verify the expression for $F(x)$.

8. Use the estimate $\hat{k} = 95$ (provided by extreme value theory) to compute the log-likelihood function for the spliced model, and maximize this function to find the MLEs

for λ and α. As in exercise 2, use "reasonable" starting values for numerical optimization.

9. Plot the empirical distribution function of the loss data, together with the cdf of the spliced distribution. Does the model provide a good fit to the data?

References

Roel Verbelen, Lan Gong, Katrien Antonio, Andrei Badescu, and X. Sheldon Lin. Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *ASTIN Bulletin*, 45(3):729–758, 2015.