PwC – Digital Intelligence

Task 6: Predictive & Statistical Modelling

a) **Do you have all the information available (e.g. regarding claims data, assumptions for distributions or the collective risk model) that is required to fit the data to the collective risk model of each risk type (i.e. WIS, non-WIS) and, thus, to derive the aggregate loss distributions? If not, which information is potentially missing? If any, apply expert judgment and explain your decisions you made.**

Ans: At first, it is not clear if the number of sites for WIS (4) and non-WIS (1) has always been constant. This is essential to calibrate the frequency distributions and to predict the expected number of claims for the exposure insured in year 2021.
Regarding the claims data, the following should be mentioned:
–Censored data: No information is provided if the given data set has been censored, for example, not reported claims due to deductibles in the contract, or too low claim sizes. This may impact the fitting approach for the frequency and severity distributions.
–Reported Data: No information is provided if the given claims data is paid or reported and if any corrections for time-lag in reporting are required. Regarding the assumptions for the collective risk model (for each site), we need the following additional information:
–Independence of WIS and non-WIS claims amounts:
By assumption, we only know that within each risk group all claims are independent. However, since it might be assumed that the claims from the WIS site do not influence the claims on the non-WIS site (and vice versa), the above assumption can be assumed.
–Independence of WIS and non-WIS claims frequency:
By assumption, we might assume that the claims frequency for WIS and non-WIS are independent (this is necessary for later purposes; see part (g)).
–Identically distributed WIS and non-WIS claims amounts:
This is a common assumption in insurance practice since insurer intend to form homogeneous collectives. In our case, this assumption is reasonable.

b) **Are the assumptions made for each of the claims' frequency and severity distributions reasonable? How could you validate it?**

Ans:

Poisson distribution:
Reasonableness: The Poisson distribution is common in actuarial practice to model the claims frequency. Therefore, it is reasonable to assume this distribution.
Validation of the assumption: Commonly empirical insurance count data is over-dispersed while the Poisson distribution assumes equi-dispersed data. Calculation of the dispersion based on the data (empirical variance/ empirical mean): if its ratio equals 1, then the assumption is

adequate; if, however, its ratio is greater than 1, then a Negative Binomial distribution might be a reasonable distribution to model the claims frequency. If the ratio is lower than 1, the binomial distribution may be used to model under-dispersed data.
Pareto distribution:
Reasonableness: The Pareto distribution with two parameters (scale and shape parameter) is used in insurance practice to model long tailed claims. As this applies for both cover groups (outliers can be observed in the data), this distribution is adequate. However, commonly in actuarial science, attritional and large losses are modelled separately. One may apply the log-normal distribution for attritional losses and the Pareto distribution for large losses. This requires splicing (separation) of (already very short) data sets in two sub data sets as well as decision making on a threshold that is separating the dataset into an ordinary part modelled by an exponentially behaving distribution (i.e. log-normal) and a heavy-tailed, sub-exponentially behaving part where log-claims are modelled by an exponentially behaving random variable (i.e. Pareto).
Validate of the assumption: 1. Calculation of the empirical Mean Excess function and compare its shape with the shape of the theoretical Mean Excess function of the Pareto distribution.

2. Plot the empirical quantiles
and the theoretical quantiles (with estimated parameters) in a scatter plot: if both lie on the 1st bisector →good fit; otherwise: expert judgement is required (to find a suitable distribution).

c) What challenges are you most likely facing given the data? How would you characterize the dataset of each type of site?

Ans: The fitting of the distributions to the claims data may be challenging due to –Outliers/tails: For both groups, there are claims amounts that are far above the average claims experience (e.g. 230.5 mCHF (WIS) and 125 mCHF (Non-WIS)).
–Skewness: Does the skewness of the data fits the assumed claims distribution?
Characterization of the dataset for each group: There is little data for both sites such that a stable estimation of the unknown parameters of the distributions is hardly possible (e.g. for Non-WIS there are only three observed claims within the observation period to estimate two unknown parameters).

d) How could you validate the provided data set regarding its accuracy and completeness?

Ans: Regarding accuracy and completeness, the following methods can be applied:
–Check the quality of the dataset within the database from which data is provided, i.e. analyze claims reporting processes, perform consistency checks with other systems, reconciliation of provided claims data with original database and those reported at end of prior reporting period or perform interviews with claims handler, reserving actuary, underwriter, etc.
–Check if there were manual adjustments made to the claims data and investigate the reasons for it.

–Reconcile the data provided with paid/reported claim amounts posted to the accounting system and P&L.
–Check whether claims were assigned to the appropriate accident years in the provided tables (i.e. accident year vs accounting year).
–Check, if indeed claims of most recent accident years are fully settled or paid out to the insured and no adjustments (i.e. IBNR factor) to provided claims data is required.
–Benchmarking of claims data to other similar exposures observed in the market.

e) Which parameters do you obtain when you are fitting the claims frequency and severity distributions for each site to the provided data set? Check out Jakob's hints to get you started.

Ans:

Estimation of claims frequency:
Let $i \in \{\text{W, N-W}\}$. Assume that $N^i \sim \text{Poi}_{\lambda_i}$, where $\lambda_i > 0$ is unknown. In view of $\mathbb{E}[N^i] = \lambda_i$, an estimator $\hat{\lambda}_i$ for the claims frequency of i is given by:

$$\hat{\lambda}_i := \frac{\#\text{claims of } i \text{ during observation period}}{\#\text{years in observation period}}.$$

- WIS: $\hat{\lambda}_{\text{W}} = \frac{7}{5} = 1.4$.

Remark: We emphasize that the claims in the table below are from all four WIS sites. That is, the estimator $\hat{\lambda}_{\text{W}}$ for λ_{W} is the frequncy for all of the four sites. Here it does not matter, as long the number of sites are constant. Otherwise one would estimate the unknown parameter lambda per one site and then scale up to four sites.

- non-WIS: $\hat{\lambda}_{\text{N-W}} = \frac{3}{5} = 0.6$.

Estimation of scale parameter:
Let $i \in \{\text{W, N-W}\}$. Assume that $X_1^i \sim \text{Par}_{t_i,\alpha}$ with $\alpha = 2.5$, where t_i is unknown (note all claims within each site are assumed to be identically distributed). In view of $\mathbb{E}[X_1^i] = t_i \cdot \frac{\alpha}{\alpha-1}$ (see hint) and $\alpha = 2.5$ (by assumption), an estimator $\hat{t}_i$ for t_i is given by

$$\hat{t}_i := \frac{1}{5}\sum_{j=1}^{5} y_j^i \cdot \frac{\alpha - 1}{\alpha} = \frac{3}{5} \cdot \frac{1}{5}\sum_{j=1}^{5} y_j^i.$$

where y_j^i denotes the indexed claims for i and year j derived by $y_j^i = (1 + 0.03)^{2021-j} \cdot x_j^i$ for $j \in \{2016, \ldots, 2020\}$ (with x_j^i denoting the reported claim for i and year j).

- WIS: $\hat{t}_{\text{W}} = \frac{3}{5} \cdot 63.9 = 38.3$.
- non-WIS: $\hat{t}_{\text{N-W}} = \frac{3}{5} \cdot 28.5 = 17.1$.

Year	x_j^{W}	y_j^{W}	Year	$x_j^{\text{N-W}}$	$y_j^{\text{N-W}}$
2016	3.1	3.6	2016	4.5	5.2
2016	2.1	2.4	2016	0	0
2016	10.5	12.2	2016	0	0
2017	2.0	2.3	2017	0	0
2018	0	0	2018	125.3	136.9
2018	0	0	2018	0	0
2019	230.5	244.5	2019	0.4	0.4
2019	51.0	54.1	2019	0	0
2020	0.5	0.5	2020	0	0

Table 1: (Indexed) Historical reported loss data in mCHF per Year for WIS and non-WIS. Note: All numbers are rounded to one decimal place.

f) What is the expected value of the claims for the current year 2021 at the individual (i.e. for each site) as well as at the aggregate level? Do you need a simulation tool or can the expected values be derived analytically? Jakob has your back, check out his hint.

Ans:

Maintain the notation from (e). Let $S^i := \sum_{k=1}^{N^i} X_k^i$ for $i \in \{\text{W}, \text{N-W}\}$, and note that S^i corresponds to the total claims amount for coverage group i. By means of Wald's equation, we get

$$\mathbb{E}[S^i] = \mathbb{E}[N^i] \cdot \mathbb{E}[X_1^i] \quad \text{for any } i \in \{\text{W}, \text{N-W}\}.$$

In view of the estimated paramters from (e) as well as $\alpha = 1.5$, we have

$$\hat{\mathbb{E}}[S^{\text{W}}] = \hat{\mathbb{E}}[N^{\text{W}}] \cdot \hat{\mathbb{E}}[X_1^{\text{W}}] = \hat{\lambda}_{\text{W}} \cdot \hat{t}_{\text{W}} \cdot \frac{\alpha}{\alpha - 1} = 1.4 \cdot 38.3 \cdot \frac{2.5}{1.5} = 89.5$$

as well as

$$\hat{\mathbb{E}}[S^{\text{N-W}}] = \hat{\mathbb{E}}[N^{\text{N-W}}] \cdot \hat{\mathbb{E}}[X_1^{\text{N-W}}] = \hat{\lambda}_{\text{N-W}} \cdot \hat{t}_{\text{N-W}} \cdot \frac{\alpha}{\alpha - 1} = 0.6 \cdot 17.1 \cdot \frac{2.5}{1.5} = 17.1.$$

Therefore, the estimated expected total claims amount $S := S^{\text{W}} + S^{\text{N-W}}$ on an aggregated level is given by

$$\hat{\mathbb{E}}[S] = \hat{\mathbb{E}}[S^{\text{W}}] + \hat{\mathbb{E}}[S^{\text{N-W}}] = 89.5 + 17.1 = 106.6.$$

In particular, no simulation is needed to calculate the expected value of claims for year 2021 for each site and the aggregate claims amounts.

g) Approximate the Value at Risk at level 80% of the (total) aggregate loss distribution for both risk types using the standard normal distribution (or "CLT") to obtain an estimate of the equalization reserves to be set up by the insurer. What now? Jakob tells you in some hints.

Ans:

In this part, we are looking for the $z \in \mathbb{R}$ for which applies

$$\mathbb{P}[S \leq z] = 0.8.$$

Denoting by $\Phi_{0,1}^{-1}(0.8)$ the 80%-quantile of the standard normal distribtion $\Phi_{0,1}$ and applying the Central Limit theorem (CLT), we have

$$\begin{aligned}
\mathbb{P}[S \leq z] &= \mathbb{P}\left[\frac{S - \mathbb{E}[S]}{\sqrt{\mathbb{V}\text{ar}[S]}} \leq \frac{z - \mathbb{E}[S]}{\sqrt{\mathbb{V}\text{ar}[S]}}\right] \\
&\approx \Phi_{0,1}\left(\frac{z - \mathbb{E}[S]}{\sqrt{\mathbb{V}\text{ar}[S]}}\right) \overset{!}{=} 0.8 \\
&\Leftrightarrow z = \Phi_{0,1}^{-1}(0.8) \cdot \sqrt{\mathbb{V}\text{ar}[S]} + \mathbb{E}[S]
\end{aligned}$$

Take into account that the CLT may be applied since $\alpha = 2.5 > 2$ (in this case, the second moment of the two-parameter Pareto distribution exists). In view of

$$\begin{aligned}
\mathbb{V}\text{ar}[S^i] &= \mathbb{E}[N^i] \cdot \mathbb{V}\text{ar}[X_1^i] + \mathbb{E}[X_1^i]^2 \cdot \mathbb{V}\text{ar}[N^i] \\
&= \mathbb{E}[N^i] \cdot (\mathbb{E}[(X_1^i)^2] - \mathbb{E}[X_1^i]^2) + \mathbb{E}[X_1^i]^2 \cdot \mathbb{V}\text{ar}[N^i] \\
&= \mathbb{E}[N^i] \cdot \mathbb{E}[(X_1^i)^2]
\end{aligned}$$

for any $i \in \{\text{W,N-W}\}$ (by Wald's equation; note that $\mathbb{E}[N^i] = \mathbb{V}\text{ar}[N^i]$) and thus

$$\begin{aligned}
\hat{\mathbb{V}\text{ar}}[S] &= \hat{\mathbb{V}\text{ar}}[S^{\text{W}}] + \hat{\mathbb{V}\text{ar}}[S^{\text{N-W}}] \\
&= \hat{\mathbb{E}}[N^{\text{W}}] \cdot \hat{\mathbb{E}}[(X_1^{\text{W}})^2] + \hat{\mathbb{E}}[N^{\text{N-W}}] \cdot \hat{\mathbb{E}}[(X_1^{\text{N-W}})^2] \\
&= \hat{\lambda}_{\text{W}} \cdot \hat{t}_{\text{W}}^2 \cdot \frac{2.5}{2.5 - 2} + \hat{\lambda}_{\text{N-W}} \cdot \hat{t}_{\text{N-W}}^2 \cdot \frac{2.5}{2.5 - 2} \\
&= 10268.2 + 877.2 = 11145.4
\end{aligned}$$

by part (e) (N^{W} and $N^{\text{N-W}}$ are assumed to be independent (see part (a)), the estimated 80%-quantile of the aggregate loss distribution is given by

$$\hat{z} = \Phi_{0,1}^{-1}(0.8) \cdot \sqrt{\hat{\mathbb{V}\text{ar}}[S]} + \hat{\mathbb{E}}[S] = 0.84 \cdot \sqrt{11145.4} + 106.6 = 195.3.$$

h) Quantify the impact on the expected aggregate loss for year 2021 if the claims frequency for both sites immediately doubles?

Ans: In view of part (f), doubling the (estimated) claims frequencies for each cover group leads to a doubling of the (estimated) expected aggregate loss.

i) How could the modelling approach be further improved?

Ans: The following approaches can be followed:

– If the dataset is sufficient, one might model attritional losses using a lognormal and large losses with a Pareto distribution. However, for this subdivision a threshold has to be defined which delimit attritional from large claims which can be derived with appropriate statistical methods.

– One could also test other distributions for the claims amounts and compare the goodness-of fit (e.g. by means of the method illustrated in part (b)).

j) Could you apply Machine Learning methods to this example? Explain.

Ans: Due to the limited/small amount of data, an application of Machine Learning methods is quite unstable and therefore not appropriate.

Bonus Question: Without performing any simulations, is it possible to estimate the confidence interval at the 80% level for the expected aggregate loss (sound statistical approximations can be applied)?

Ans: The mentioned confidence interval at level 0.8 might be obtained using similar arguments as in the proof of part (g). More precisely, it can be shown that the confidence interval for the expected aggregate loss is around the estimated expected aggregate loss ($\hat{E}[S] = 106.6$) as calculated in part (f).