Enhancing IRB LGD Modeling with Survival Analysis

A Framework for Extrapolating Incomplete Recoveries

Andrija Djurovic

www.linkedin.com/in/andrija-djurovic

LGD Modeling and Incomplete Recoveries

- In IRB Loss Given Default (LGD) modeling, all defaults must be considered, including those with ongoing or incomplete recovery processes, as they contribute essential information for estimating long-run average LGD.
- Incomplete recovery cases often reflect recent or more complex defaults, and excluding them can result in loss of relevant data and potential underestimation of LGD.
- To incorporate incomplete recoveries, institutions must estimate future recoveries using data from similar past cases with completed recovery processes.
- The inclusion of estimated components means that the long-run average LGD is not entirely objective but combines observed and forecasted elements.
- When adjusting for incomplete recoveries, institutions should:
  - Include all observed recoveries and costs to date.
  - Estimate future recoveries and costs, whether from collateral realization or unsecured processes, within the expected maximum recovery timeframe.
- This presentation focuses on the use of survival analysis for extrapolating incomplete recoveries. For further details, as well as other applications of survival analysis in LGD modeling, practitioners can refer to Witzany et al. (2012) and Joubert et al. (2018).

Survival Analysis

Survival analysis is a statistical method that focuses on analyzing the time until an event of interest occurs. It accounts for censored data, where the event has not yet occurred for some subjects during the observation period. These characteristics make survival analysis a well-suited method for IRB LGD modeling.

While it can be applied to various aspects of LGD modeling, this presentation will focus on its use for extrapolating incomplete recoveries. Practitioners should note that the presented framework offers fundamental elements for applying survival analysis in LGD modeling and can be extended to different process designs. Accordingly, the presentation will use the Kaplan-Meier estimator as a basis for demonstrating its application in extrapolating incomplete recoveries.

The Kaplan-Meier estimator (also known as the product-limit estimator) is given by:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{w_i}{n_i}\right)$$

where:

- $\hat{S}(t)$ is the estimated survival function at time t;
- t_i are the ordered time points at which events occur;
- w_i is the number of events at time t_i;
- n_i is the number of individuals at risk just before time t_i.

Although the following slides present a simplified study of the application of survival analysis for extrapolating incomplete recoveries, they still illustrate the core steps needed to perform this exercise. Practitioners are encouraged to consider this as a foundation and to adapt this flexible approach to various extrapolation process designs.

Simulation Study

Simulation Design

Assume that the overall modeling dataset consists of only three defaulted facilities. Two of the three facilities have completed recovery processes, while one facility is still undergoing an incomplete recovery process.

Based on this dataset, the goal of this simulation study is to estimate future recoveries and, consequently, the LGD value of the facility with the incomplete recovery process.

The presented dataset summarizes the main elements required for applying survival analysis for this purpose, where:

- *id* denotes the facility identification number;
- *tid* is the time in default;
- *ead* is the exposure at the moment of default;
- *cf* represents cash flows for each time-in-default period.

Dataset Overview

```
##   id tid ead cf
##    1   0 250  0
##    1   1 250 20
##    1   2 250 20
##    1   3 250 20
##    1   4 250 20
##    1   5 250 20
##    1   6 250 20
##    1   7 250 20
##    1   8 250 20
##    1   9 250 20
##    1  10 250 20
##    2   0 200  0
##    2   1 200 10
##    2   2 200 10
##    2   3 200 10
##    2   4 200 10
##    2   5 200 10
##    3   0  80  0
##    3   1  80 10
##    3   2  80 10
##    3   3  80 10
##    3   4  80 10
##    3   5  80 10
```

Survival Dataset Preparation

A critical aspect of using survival analysis for LGD modeling is preparing the dataset. Unlike traditional datasets used for LGD modeling, survival datasets require specific adjustments and additional inputs. The following steps outline the key procedures for preparing a survival dataset. Practitioners should note that, depending on the design of the recovery process and the nature of incomplete recoveries, these steps may need to be further adapted to incorporate all relevant inputs. For example, different approaches may be required for extrapolating incomplete recoveries depending on whether the underlying collateral has been realized. However, such considerations are beyond the scope of this presentation, which focuses on the basics of survival analysis for LGD modeling.

1. Add a censoring indicator equal to 1 for all time in default periods with observed cash flows.
2. Add the unrecovered amount (under the cash flow column) for each facility in the modeling dataset. For facilities with a complete recovery process, the time in default for the unrecovered amount should be set to the maximum recovery period (MRP); for facilities with an incomplete recovery process, it should be set to the last observed period.
3. For the new row representing the unrecovered amount per facility, assign a censoring indicator of 0, as defined in step 1.
4. To comply with the requirement of calculating LGD as a default-weighted average, add weights for each facility in the modeling dataset based on the ratio of cash flows to the exposure at the moment of default. The sum of the weights calculated in this way should equal the total number of facilities in the modeling dataset.
5. Add an incomplete recovery indicator equal to 1 for all facilities with an incomplete recovery process. This step is optional and intended to simplify the identification of such facilities.

Note that step 2 can be further adjusted for facilities with complete recovery processes in cases where the final time in default exceeds the MRP. However, for this simulation study, it is assumed that both periods are equal to 10.

Survival Dataset Preparation cont.

Final Dataset

```
##   id tid ead  cf censoring incomplete weights
##    1   0 250   0         1          0   0.000
##    1   1 250  20         1          0   0.080
##    1   2 250  20         1          0   0.080
##    1   3 250  20         1          0   0.080
##    1   4 250  20         1          0   0.080
##    1   5 250  20         1          0   0.080
##    1   6 250  20         1          0   0.080
##    1   7 250  20         1          0   0.080
##    1   8 250  20         1          0   0.080
##    1   9 250  20         1          0   0.080
##    1  10 250  20         1          0   0.080
##    1  10 250  50         0          0   0.200
##    2   0 200   0         1          0   0.000
##    2   1 200  10         1          0   0.050
##    2   2 200  10         1          0   0.050
##    2   3 200  10         1          0   0.050
##    2   4 200  10         1          0   0.050
##    2   5 200  10         1          0   0.050
##    2  10 200 150         0          0   0.750
##    3   0  80   0         1          1   0.000
##    3   1  80  10         1          1   0.125
##    3   2  80  10         1          1   0.125
##    3   3  80  10         1          1   0.125
##    3   4  80  10         1          1   0.125
##    3   5  80  10         1          1   0.125
##    3   5  80  30         0          1   0.375
```

Simulation Study Results

The following table presents the simulation results of the survival analysis.
After preparing the datasets in the previous steps, the observed LGD values are shown in the survival column in the results below. Additionally, for each time in default, the same column reports the corresponding observed LGD values.

```
## Call: survfit(formula = km ~ 1, data = db.3, weights = db.3$weights,
##     type = "kaplan-meier", conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1   3.00   0.255    0.915  0.0546       0.8080        1.000
##     2   2.75   0.255    0.830  0.0805       0.6722        0.988
##     3   2.49   0.255    0.745  0.1024       0.5442        0.946
##     4   2.24   0.255    0.660  0.1226       0.4198        0.900
##     5   1.98   0.255    0.575  0.1416       0.2974        0.853
##     6   1.35   0.080    0.541  0.1507       0.2455        0.836
##     7   1.27   0.080    0.507  0.1600       0.1932        0.821
##     8   1.19   0.080    0.473  0.1695       0.1406        0.805
##     9   1.11   0.080    0.439  0.1791       0.0877        0.790
##    10   1.03   0.080    0.405  0.1888       0.0346        0.775
```

Inputs for Extrapolating Incomplete Recoveries

Based on these results, it is possible to extrapolate the recoveries - and consequently estimate the LGD - for the facility (id = 3) with an incomplete recovery process. Specifically, the observed LGD values reported in the survival column, along with the time in default indicator (tid and time), are the main inputs used for this purpose. The following slide demonstrates this process.

Extrapolation of Incomplete Recoveries

When extrapolating for incomplete recoveries, practitioners should start with the data available up to the last observed period for the analyzed facility. For this specific example, and the facility with `id = 3`, that is period 5 (`tid = 5`), and up to that point, the observed loss is equal to 37.5% ($\frac{30}{80}$).
Although the standard terminology refers to the extrapolation of the incomplete recovery process, within the survival analysis framework, practitioners directly estimate the loss. In this context, the remaining loss is essentially equal to the observed loss up to the last observed period, multiplied by the ratio between the segment loss at the moment of MRP and the segment loss at the same observation period as that of the analyzed facility. Formally, this calculation is expressed as:

$$LGD_{f,MRP} = LGD_{f,t} \cdot \frac{LGD_{s,MRP}}{LGD_{s,t}}$$

where:

- $LGD_{f,MRP}$ denotes the loss extrapolated up to MRP for the analyzed facility;
- $LGD_{f,t}$ is the observed loss up to the last observed period for the analyzed facility;
- $LGD_{s,MRP}$ denotes the segment loss at MRP;
- $LGD_{s,t}$ is the segment loss at the same period as the last observed for the analyzed facility.

For this specific simulation study, the elements and results of the above formula are as follows:

- $LGD_{f,t} = \frac{30}{80} = 0.375$;
- $LGD_{s,MRP} = 0.405$;
- $LGD_{s,t} = 0.575$;
- $LGD_{f,MRP} = LGD_{f,t} \cdot \frac{LGD_{s,MRP}}{LGD_{s,t}} = 0.2641$.

Conclusions

- For LGD modeling, survival analysis is a well-suited and flexible framework that can be used in different aspects of the modeling process.
- A straightforward use of survival analysis in LGD modeling is for the extrapolation of incomplete recoveries.
- Although other methods used for extrapolation often estimate recoveries and then translate them into losses, survival analysis directly projects losses, as demonstrated in the previous slides.
- The main advantages of using survival analysis for this purpose are:
  1. It utilizes all observed information, including facilities with incomplete recovery processes.
  2. When extrapolating losses for facilities with incomplete recovery processes, it does not distort the calibration.
  3. It is flexible enough to incorporate different business inputs, especially those specific to the recovery process, such as the split between collateral realization and cash recoveries.
  4. It can incorporate additional drawings.
  5. It can introduce conservatism through the confidence level of the survival curve.
  6. It can also be extended to incorporate different covariates if deemed necessary for the recovery process.
- As with any other statistical method used in IRB modeling, different designs rely on certain assumptions. Therefore, it is up to practitioners to engage in critical thinking when designing the overall process and to justify the reasonableness of these assumptions.