Optimal Sampling for a Loan Portfolio

Alex Kisselev (Pine River Cap Management) ⋄ Pak-Wing Fok (University of Delaware)
Kai Yang (George Washington University) ⋄ Dubravka Bodiroga (George Washington University)
Erblin Mehmetaj (George Washington University) ⋄ Diego Torrejon (George Mason University)
Le Zhang (George Washington University) ⋄ Michael Libman (BancLab)

June 14, 2014

1 Problem Statement

Suppose a client has a small collection of loans (a portfolio) with specific characteristics, and wants a measure of the portfolio's credit risk. The client wants us to use a much larger historical database, called "universe" and pull out a sample of loans that matches his/her portfolio composition, but consists of much greater number of loans. Figures 1.1 and 1.2 represent the composition of the client's and universe data, respectively. However, the greater population of loans we sample, the smaller precision when it comes to matching sample and client's data. The first part of the paper explains how to minimize the cost of choosing a larger sample size. First, we establish the necessary notation.

Suppose we are given a portfolio of loans (it could come from a client, for example) where each loan is uniquely assigned to $\nu \in \mathbb{N}$ categories. For each category $K \in \{1, 2, \ldots, \nu\}$ there are m_K possible "buckets" or sub-categories. In other words, for each ν-tuple $(i_1, i_2, \ldots, i_\nu)$, $1 \leq i_K \leq m_K$ there is a corresponding number of loans $n^*(i_1, i_2, \ldots, i_\nu)$, which we will call the *portfolio frequency.* From the portfolio frequencies, we can quickly construct the *portfolio distribution*

$$p_{i_1,\ldots,i_\nu} = \frac{n^*(i_1, \ldots, i_\nu)}{S}, \qquad S \equiv \sum_{i_\nu=1}^{m_\nu} \ldots \sum_{i_1=1}^{m_1} n^*(i_1, \ldots, i_\nu). \tag{1}$$

For example, loans could be classified according to loan size and state of origination corresponding to $\nu = 2$. Loan sizes could be from \$0-\$100,000; \$100,000-\$250,000; \$250,000 - \$500,000; \$500,000-\$750,000 and \$750,000+ forming $m_1 = 5$ subcategories (see figure 1.1). The loans could originate from $m_2 = 3$ different states and for each pair (i_1, i_2), $1 \leq i_1 \leq 5$, $1 \leq i_2 \leq 3$ there is a corresponding loan frequency for the client portfolio.

Suppose further that we have access to a historical database of loans; typically the total number of loans in the database is much larger than the total number in the portfolio (see Figure 1.2). These loans are also assigned to ν categories in the same way as above. In other words, for each ν-tuple $(i_1, i_2, \ldots, i_\nu)$, there is a corresponding number of loans within the database ("universe") $u(i_1, i_2, \ldots, i_\nu)$, which we call the *database frequency.*

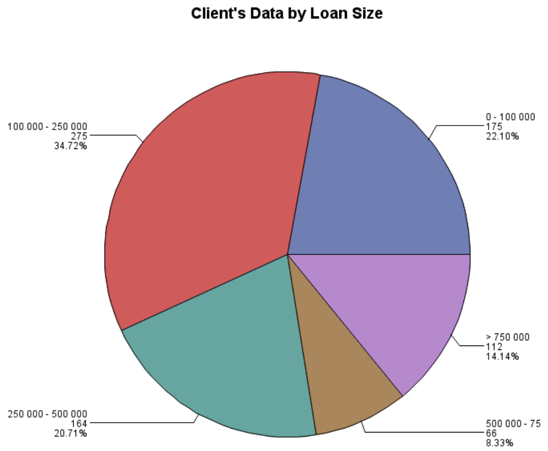


Fig. 1.1

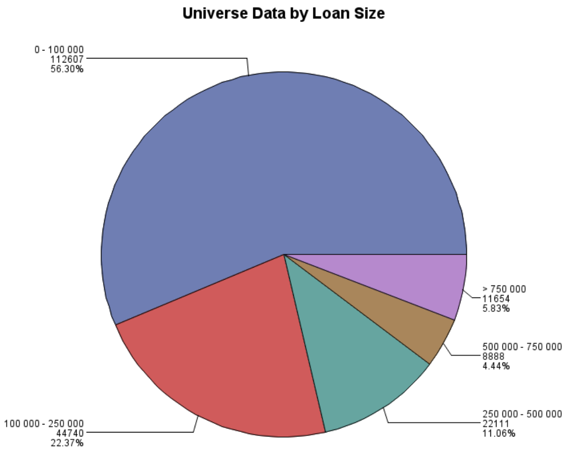


Fig. 1.2

Given a total sample size N, the problem is to select a subset of loans from the database $(n_1, n_2, \ldots, n_\nu)$ – which we call the *sample frequency* – (clearly $0 \leq n_i \leq u_i$) to match the distribution of the portfolio as closely as possible. From this subset, we can construct the *sample distribution*

$$q_{i_1,\ldots,i_\nu} = \frac{n(i_1, i_2, \ldots, i_\nu)}{N}, \qquad N \equiv \sum_{i_\nu=1}^{m_\nu} \cdots \sum_{i_1=1}^{m_1} n(i_1, \ldots, i_\nu). \tag{2}$$

As an example, suppose $\nu = 1$ and our data $(n_1^*, \ldots, n_5^*)$ is given in Table 1:

i	n_i^*	p_i (%)	u_i	n_i	q_i (%)
1	175	22.1	112,607	18,209	22.1
2	275	34.7	44,740	28,614	34.7
3	164	20.7	22,111	17,064	20.7
4	66	8.3	8,888	6,867	8.3
5	112	14.1	11,654	11,654	14.1
Total	$S = 792$	100	200,000	$N = 82,408$	100

Table 1: Example of perfectly matching the database distribution to the portfolio distribution, given $N = 82,408$.

If we are required to draw a total of $N = 82,408$ samples from the database, then the solution is easy. We draw $\{18209, 28614, 17064, 6867, 11654\}$ samples from each category in the database and we are able to match the loan distribution exactly: $q_i = p_i$ for $i = 1, \ldots, 5$ (see Figure 1.3). The problem becomes more interesting and difficult if N is larger, say $N = 99,999$. In the case where $N = 82,408$, we used all the database loans corresponding to $i = 5$. If N is increased beyond 82,408, there are not enough samples in the database to accommodate $p_5 = 0.141$. The problem that this paper will address is how to find the "best" sample frequencies in this case.

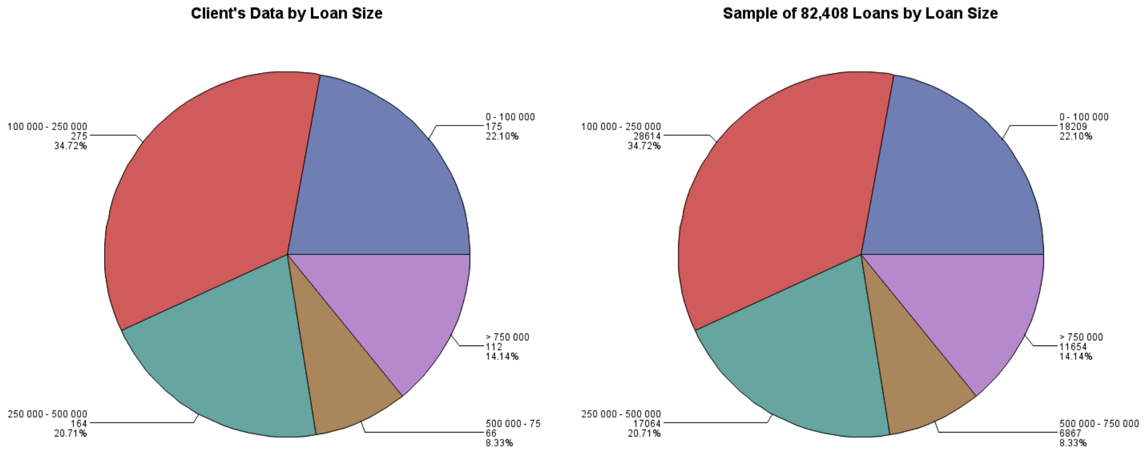


Fig. 1.3

2 Single Category Objective Functions ($\nu = 1$)

Suppose the client wants us to choose a sample so that data is a match based on the loan size category. For this category we are given desired loan fractions $\{p_i\}$ for $i = 1, \ldots, m$, database frequencies u_i, $i = 1, \ldots, m$, and a total sample size N. Then the optimal distribution $(q_1, \ldots, q_m)$ is found by minimizing

$$\Pi(\mathbf{q}) = \sum_{i=1}^{m} (p_i - q_i)^2, \tag{3}$$

subject to the constraints

$$0 \leq q_i \leq u_i/N, \tag{4}$$

$$\sum_{i=1}^{m} q_i = 1, \tag{5}$$

and the database sample frequencies are $n_i = q_i \times N$ for $i = 1, 2, \ldots, m$.

This minimization problem is easily solved using Matlab's `fmincon` command.This command uses iteration to test all the percentages until it finds the ones that are using the data from Table 1. When $N = 99,999$, the optimal database sample sizes are given in the following Table 2.

i	n_i	q_i (%)
1	22,740	22.7
2	35,366	35.4
3	21,351	21.4
4	8,888	8.9
5	11,654	11.7
Total	99,999	100

Table 2: Optimal sample sizes when $N = 99,999$.

Since we obatined the optimal distribution, we can use the SAS procedure `surveyselect` to obtain the optimal sample. The procedure `surveyselect` uses the desired percentages as the input, randomly selects the specified percent of samples from each category, and outputs a table representing the new sample. The proximity of this sample to the client's data can be seen from the Figure 1.4. Although the distribution has slightly changed, we still have significantly larger sample population. Note that we have used all samples in the database for categories 4 and 5.

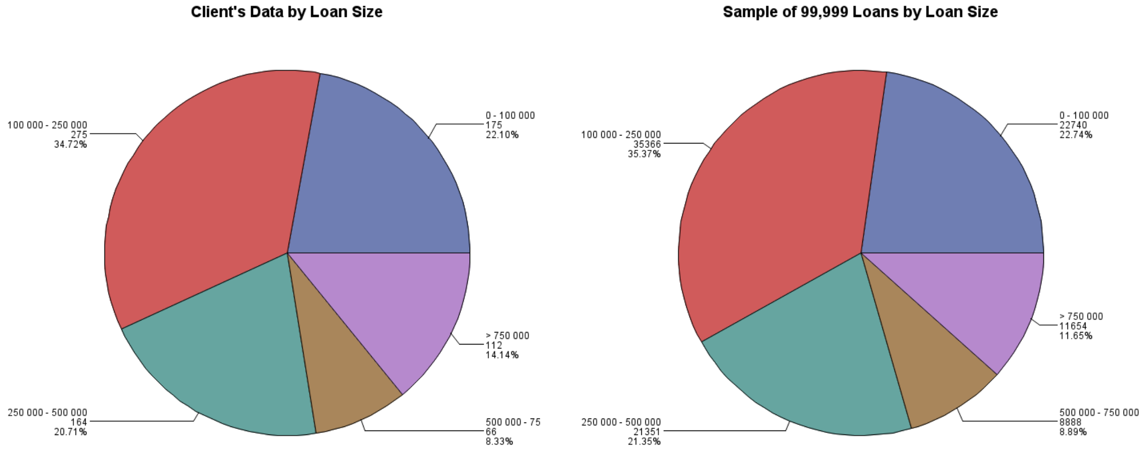


Fig. 1.4

3 Two-category Objective Functions ($\nu = 2$)

In the previous section we saw how to find the best sample fit when the client specifies one important category. Now we expand the problem by involving one additional category. For example, the client wants us to sample loan based on loan size and the state that the loan was issued in (see Table 3).

State	Loan Size 1	Loan Size 2	Loan Size 3
Texas	p_{11}	p_{12}	p_{13}
Montana	p_{21}	p_{22}	p_{23}
Washington	p_{31}	p_{32}	p_{33}

Table 3: Percentages p_{ij} specified by the client, by two categories

Suppose we are given a joint loan distribution $\{p_{ij}\}$ $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$. Then given a total sample size N and database frequencies u_{ij} we wish to minimize

$$\Pi_1(\mathbf{q}) = \sum_{j=1}^{m_2} \sum_{i=1}^{m_1} (p_{ij} - q_{ij})^2, \tag{6}$$

subject to the constraints

$$
\begin{aligned}
0 \leq q_{ij} &\leq u_{ij}/N, & (7)\\
\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} q_{ij} &= 1, & (8)
\end{aligned}
$$

from which we construct the database samples $q_{ij} \times N$.

Alternatively, we may be given weaker data in the form of loan marginal distributions $\{Q_1, Q_2, \ldots, Q_{m_1}\}$ and $\{R_1, R_2, \ldots, R_{m_2}\}$, in which case we should find $\{q_{ij}\}$ to minimize

$$\Pi_2(\mathbf{q}) = \sum_{i=1}^{m_1}\left(Q_i - \sum_{j=1}^{m_2} q_{ij}\right)^2 + \sum_{j=1}^{m_2}\left(R_j - \sum_{i=1}^{m_1} q_{ij}\right)^2, \tag{9}$$

subject to the constraints

$$
\begin{aligned}
0 \leq q_{ij} &\leq u_{ij}/N, & (10)\\
\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} q_{ij} &= 1. & (11)
\end{aligned}
$$

This second approach does not have a unique solution, but hybrid approaches involving minimization of combining Π_1 and Π_2 will give us an optimal solution. For example, we can minimize

$$\Pi_3(\mathbf{q}) = \sum_{i=1}^{m_1}\left(Q_i - \sum_{j=1}^{m_2} q_{ij}\right)^2 + \sum_{j=1}^{m_2}\left(R_j - \sum_{i=1}^{m_1} q_{ij}\right)^2 + \beta \sum_{j=1}^{m_2}\sum_{i=1}^{m_1}(p_{ij} - q_{ij})^2 \tag{12}$$

where $\beta(*)$ (small value) will serve the purpose of a correction term to the flexibility of the second cost function. If the client desires to match the exact $p'_{ij}s$ in the chosen sample, we can set β be a small number close to zero, which will get us a closer match. However, if the client does not care about the exact $p'_{ij}s$ but wants to have specific marginal distributions, than β could be set equal to zero.

In order to test our approach, we chose to sample from the universe data based on loan size and state code categories. After creating the code in Matlab and sampling in SAS, we obtained a sample of 10,000 loans that closely matches the client's distribution (see figure 1.5).

4 Hazard Function Problem

The second part of the paper deals with the risk of the loan to either default or prepay. After choosing a good sample, the client wants us to produce some measure of risk of the resulting portfolio that we sampled out of the universe database. The second part of the paper will demonstrate how to find the hazard function which is a good measure of the risk.

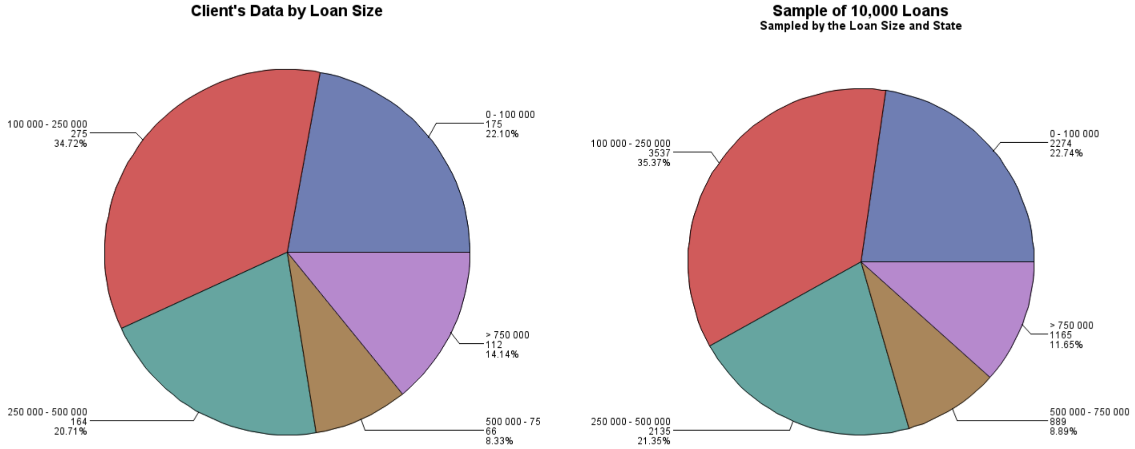


Fig. 1.5

The hazard function is the probability of failure, in a given year i, given that the subject has survived up till year i. This function is a measure of risk, since the greater the hazard is in year i, the greater the risk of failure or death is in that year. However, for our model we are rather interested in finding the hazard function $\{\lambda_i\}$ as the probability of the risk that the loan will die when it is of age i.

In this section we create the best possible hazard function of the loan portfolio with the composition specified by the client. Denote the hazard function by $\{\lambda_i\}$ for $i = 1, \ldots, 13$, denoting the age of the loan. Suppose that the client is interested in a single category of the loan, such as *loan size*, and suppose that the loan size category has m different sections. The client specifies that the desired portfolio composition based on the category is $\underline{p} = (p_1, p_2, \ldots, p_m)$, where $p_k's$ are the percentages corresponding to each category section k. Then we can find the hazard function $\{\lambda_i\}$ for this particular composition.

We derive the hazard function by following these steps:

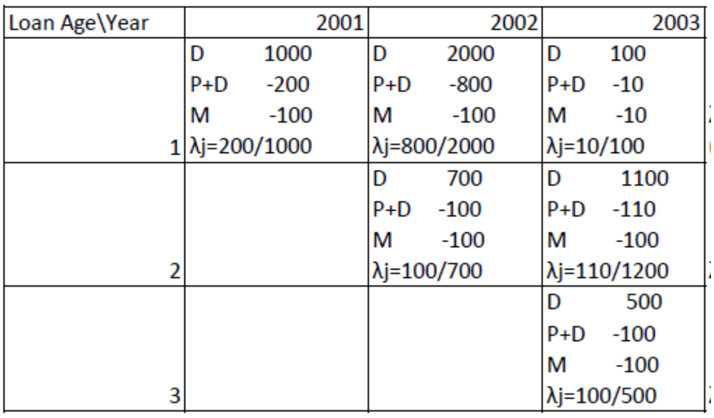
1. The client specifies $\underline{p}$ and the sample size N.

2. We use the optimization problem to generate N samples from the universe, that represent the specified portfolio composition to the best precision.

3. We analyse the sample by year and by loan age. For example, Figure 1.6 presents a nice scenario of how we calculated it the hazard function.P denotes the number of loans prepaid each year, D denotes the number of loans disbursed each year, P+D denotes number of loans prepaid or defaulted each year, and M denotes number of loans that matured each year. Then $\lambda_1 = \frac{200+88+10}{1000+2000+100}$, $\lambda_2 = \frac{100+110}{700+1100}$, $\lambda_3 = \frac{100}{500}$. In general case, we let
$w_j = \#$ of deaths in year j and $d_j = \#$ of disbursments in year j. Then

$$\lambda_i = \frac{\sum_{j=i}^{M} w_j}{\sum_{j=i}^{M} d_j},$$

where M is the number of years we would like to consider. In our model, $M = 13$ since we have data for period 2001-2013.



Loan Age\Year	2001	2002	2003
1	D 1000 P+D -200 M -100 λj=200/1000	D 2000 P+D -800 M -100 λj=800/2000	D 100 P+D -10 M -10 λj=10/100
2		D 700 P+D -100 M -100 λj=100/700	D 1100 P+D -110 M -100 λj=110/1200
3			D 500 P+D -100 M -100 λj=100/500

Fig. 1.6

4. Use SAS to select 1000 different samples from the universe data.

5. For each sample, computer the hazard rates depending on the loan age.

6. Calculate the 10th and 90th percentiles and denote them as the error bounds for our hazard rate.

5 Hazard Rates Results

Suppose the client would like us to find the death risk of the loan sample that we have sampled based on the loan size category. However, since our optimization formula can give us different samples, we would like to estimate the error. We create a loop in SAS that will compute the hazard rates, $\{\lambda_i\}$, 1,000 times (see Figure 7) and output the average values, 10^{th} and 90^{th} percentiles. We will use the 10^{th} and 90^{th} percentiles to determine the error bounds, i.e. we are disregarding the outliers of the individual distributions. In the Figure 1.8 we can see the averaged hazard rates for each loan age, together with the error bar with lower and upper ends representing the 10^{th} and 90^{th} percentile of 1,000 computations. Note that we used a sample of 10,000 loans since choosing a large sample relative to our universe to compute the error will not produce desired results. Choosing a large sample will inevitably lead to the selection of the same data multiple times, but we would like to minimize this by choosing a smaller sample.

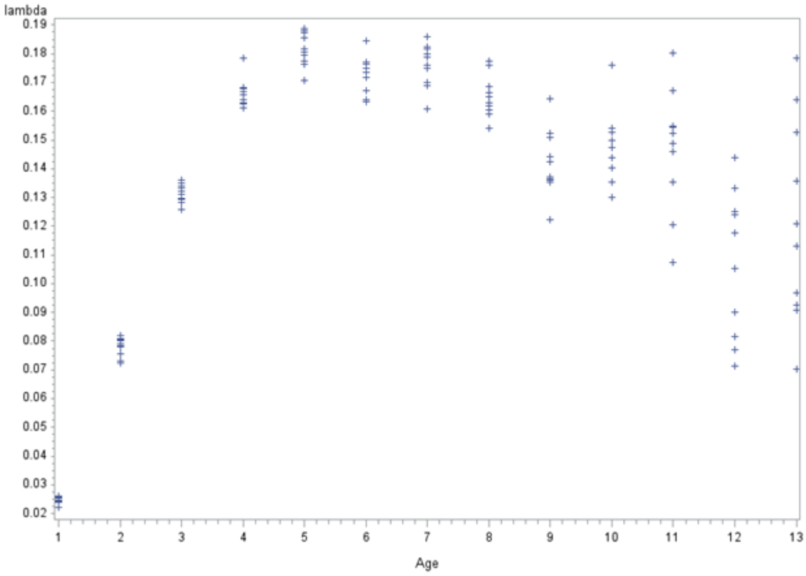


Fig. 1.7

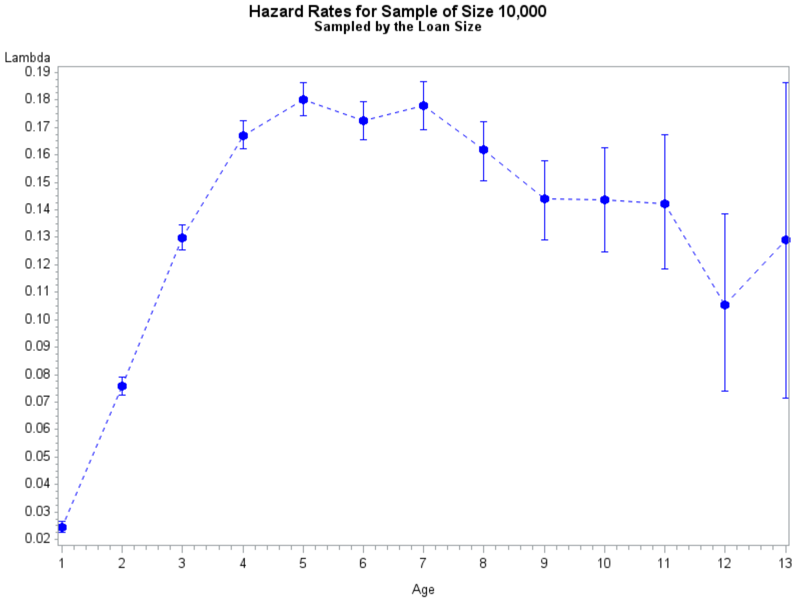


Fig. 1.8

Notice that the error is getting greater as age increases. This makes sense since as the loans get older, more loans die and mature and leave us with much smaller number of loans in the sample. The smaller number of loans, the smaller the precision, and thus the greater the error. The same analysis can be done with the region characteristic,

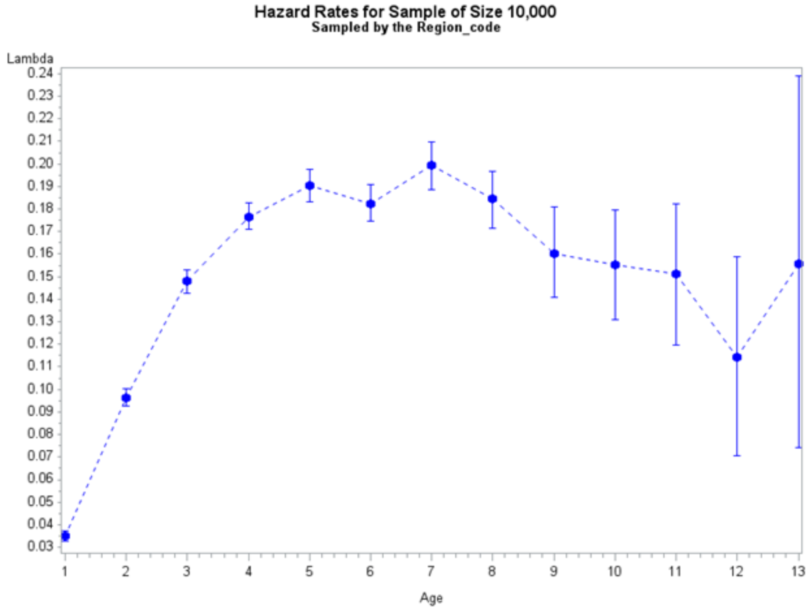


Fig. 1.9

We can also repeat the same process to model the hazard rates for the sample that we chose based on two categories. Choosing a sample based on two categories as opposed to only one category will produce a better match sample to the clients data. Thus the following graph, for a sample based on the loan size and state categories, is a better prediction of the hazard rates for the client's portfolio.

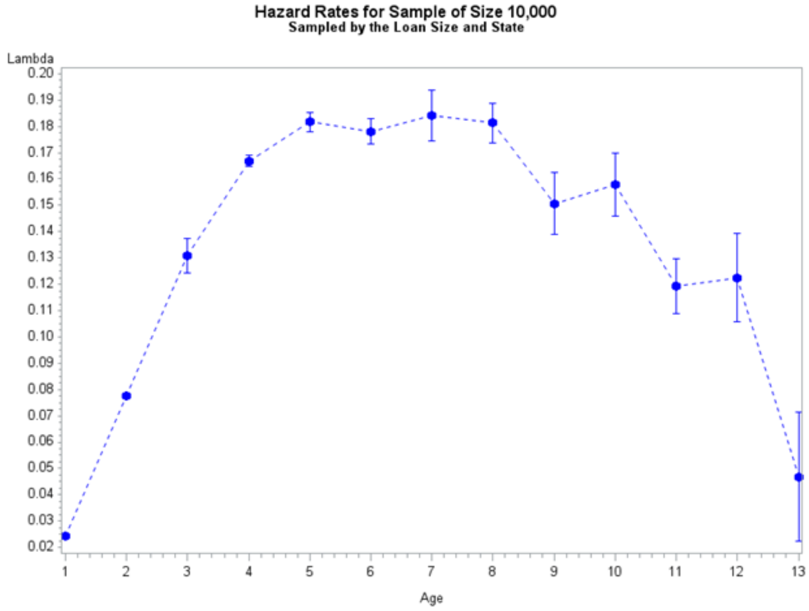


Fig. 1.10

We could repeat the same analysis with three different characteristics. In this case, the characteristics are loan size, region, and maturity of the loan.

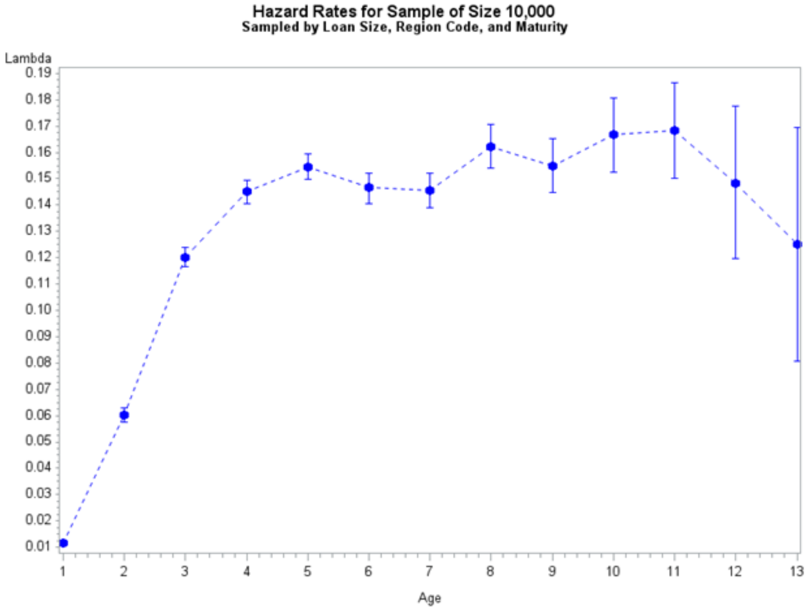


Fig. 1.9

Our models show that young loans have small probability to default or prepay, a greater probability of death at ages 5 to 8, and then the risk decreases for the rest of their life. This result makes sense in practice. Young loans are not likely to default because bad loans are usually not issued, and not likely to prepay since the cost to prepay is high while other interest rates in the industry is approximately at the same level. The risk increases with time since time presents more opportunities to prepay and economy changes might influence the loan to prepay. After year 8, the loan is not likely to default since a business will usually either fail at early stage or survive. Prepayment is also less likely since a big portion of the loan is paid off.

6 Future Work

We have implemented a code to match a portfolio with five distinct characteristics. The next thing in our analysis will be to derive results for such scenario. We will also look at solving the inverse problem, i.e. given a hazard function, we will need to provide the client with a suitable portfolio. Then we investigate if there is a "diversification benefit" of the loan portfolio. Given

$$\begin{aligned}
&(p_1, p_2, p_3, \ldots, p_m) \to \underline{\lambda}^* \\
&(1, 0, 0, \ldots, 0) \to \underline{\lambda}_1 \\
&(0, 1, 0, \ldots, 0) \to \underline{\lambda}_2 \\
&\qquad\quad \vdots \\
&(0, 0, 0, \ldots, 1) \to \underline{\lambda}_m,
\end{aligned}$$

we investigate if the superposition holds, i.e. if $\underline{\lambda}^* \approx p_1\underline{\lambda}_1 + p_2\underline{\lambda}_2 + \cdots + p_m\underline{\lambda}_m$. If this is indeed the case, then it implies no diversification benefit.