ECO6936: Empirical Specification

Proposed Title: Modeling default probabilities in peer-to-peer services

Nigar Sultana

07/17/2020

1 Economic Model

Since credit performance data are recorded monthly, I will use discrete-time survival analysis for modeling default in peer-to-peer lending services. The discrete-time survival analysis adopts hazard as log odds (logit), while continuous survival analysis estimates hazard as instantaneous change in the event occurrence rate. In the discrete model, the event time refers to units of measuring the passage of time between the initial time and the time when an individual experience a target event (i.e., in the context of loan default, the event of interest is default). Also, hazard function assess whether and, when event occur.

1.1 Formal Model

I have adopted the discrete model and notation formulated by Singer and Willet (Singer and Willett, 1993).

Consider a homogeneous population of borrowers, each at risk of experience a single target event—default the loan. I assume that, for each individual person, the target event is nonrepeatable; once it happen, it cannot happen again. Since each individual can experience the target event only once, event occurrence is inherently conditional. In order to record event occurrence in discrete-time intervals, the continuous time get divided into an infinite sequence of adjacent time periods $(0, t_1], (t_1, t_2], ..., (t_{j-1}, t_j], ..,$ and so forth. Here, the jth period begins immediately after time t_{j-1} and ends at, time t_j. For example, if time is measured in moths, an event occurring any time after t_1 (the last day of month 1) and up to including t_2 (the last day of month 2) is categorized as happening during the 2nd time interval.

The Discrete-Time Hazard Function : Suppose, T is a discrete random variable which indicate the time period j when the event occurs for a randomly selected individual from the borrowers. Also, let's introduce the P predictors, Z_p $(p = 1, 2, ...\mathbf{P})$ into the definition. The individual i's values for each of the **P** predictors in time period j can be denoted as the vector $z_{ij} = [z_{1ij}\ z_{2ij,...,}z_{\mathbf{P}ij}]$. Now, the discrete-time hazard, h_{ij}, can be defined as conditional probability as follows:

$$h_{ij} = Pr\ [T_i = j|\ T_i \geq j, \mathbf{Z}_{1ij} = z_{1ij}, \mathbf{Z}_{2ij} = z_{2ij}, ...\mathbf{Z}_{Pij} = z_{Pij}]$$

Here, h_{ij} defines the probability that individual i who can be distinguished by his or her predictor values $z_{ij} = [z_1 ij\ z_2 ij, ..., z_{\mathbf{P}} ij]$—experiences the event in time peeriod j, given that he or she survived through all prior events.

2 Statistical Model

Cox (1972) proposed that, since h_{ij} are probabilities, they can be reparameterized to have a

logistic dependence on the predictors and the time periods. Thus, such a model represents the log-odds of event occurrence as a function of predictors. In order to assume that the prediction are linearly associated, Paarsch and Golyaev (2016) and Singer and Willett (1993) incorporated the logistic transformation in the discrete time hazard model, which I have used to estimate the hazard probabilities.

$$log_e \left(\frac{h_{ij}}{1 - h_{ij}} \right) = (\alpha_1 D_{1ij} + \alpha_2 D_{2ij} + ... + \alpha_J D_{Jij} \tag{1}$$

In this written form, it is assumed that the predictors are linearly associated with the logistic transformation of hazard. The conditional log-odds that the event will occur in each time period j (given that it did not occur before, is a linear function of a constant term, α_j (specific to period j), and of the values of the predictors period j multiplied by the appropriate slope parameters. Also, the discrete time hazard model contains no single stand-alone intercepts (Singer and Willett, 1993). The alpha parameters $[\alpha_1, \alpha_2, ..\alpha_j]$ represent multiple intercepts one per time period,. When all the values of the covariates $\mathbf{Z}_1 - Z_P$ are set to zero, the $[\alpha_1, \alpha_2, ..\alpha_j]$ represent the population baseline logit-hazard function. It captures tie period by time period conditional log-odds that the individual whose covairate values are all zero (the baseline group) will experience the event in each time period, conditional on they not already done so.

3 Creating Person period Data Set

Discrete event history data are typically in the format of person-oriented data set in which each person in the sample has one record (line) of data. In order to conduct the requisite logistic regression analyses, the person oriented data set must be converted into a new person-period data set where each person has multiple records (lines of data), one per time period

observation. I have added "duration-mnths" variable as a duration variable with a month basis scale which ranges from 1 to 40 (each scale represents one month). Therefore, respective dummy time variable has been computed to indicate month: 40 time dummy variables to indicate each of 40 months.

In R, the person-period data matrix can be generated by applying the function dataLong() in the R package **discSurv** (Welchowski and Schmid, 2015). The general interface of the function is:

```
dataLong(dataSet, timeColumn,censColumn,timeAsFactor = TRUE)
```

Table 1: Person-Period Data Matrix

obj	timeInt	y	sub_grade	annual_inc	purpose	duration_mnths	loan_outcome_2
1	1	0	B	39600	home_improvement	9	0
1	2	0	B	39600	home_improvement	9	0
1	3	0	B	39600	home_improvement	9	0
1	4	0	B	39600	home_improvement	9	0
1	5	0	B	39600	home_improvement	9	0
1	6	0	B	39600	home_improvement	9	0
1	7	0	B	39600	home_improvement	9	0
1	8	0	B	39600	home_improvement	9	0
1	9	0	B	39600	home_improvement	9	0

The table 1 shows the first 9 observation of 8 columns of person-period data matrix. Three new columns have been added; **obj**, which is an identifier of the individual, **timeInt**, which contains the discrete time valuesand y, which contains the binary response variables. In the table, the first individual (**obj=1**) had an event after 9 months (duratuion-mnths = 9 and loan-outcome is zero which means fully paid). Accordingly, the person-period data matrix for the first individual has nine rows, where each row corresponds to one time interval (timeInt = 1, . . . , 9). The corresponding vector of responses is y= (0, 0, 0, 0, 0, 0, 0, 0, 0). The values of the covariates remain constant over time and are therefore the same in each row.

4 Empirical Specification

Model A : Model A is the baseline model where it includes no intercept term. It contains time indicators as predictors, in my case D_1 through D_{40}, as a group:

$$log_e\ (h_{ij}) = \ [\alpha_1 D_{1ij}\ + \alpha_2 D_{2ij} + ... + \alpha_{40} D_{Jij}]. \tag{2}$$

Model B : Model B includes the time-invariant predictor group, the influence of which is captured by the slope parameter β_1:

$$log_e\ (h_{ij}) = \ [\alpha_1 D_{1ij}\ + \alpha_2 D_{2ij} + ... + \alpha_4 0 J D_{Jij}] + \beta_1\ \text{group}_{ij}. \tag{3}$$

References

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological) 34*(2), 187–202.

Paarsch, H. J. and K. Golyaev (2016). *A gentle introduction to effective computing in quantitative research: What every research assistant should know*. Mit Press.

Singer, J. D. and J. B. Willett (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of educational statistics 18*(2), 155–195.

Welchowski, T. and M. Schmid (2015). discsurv: Discrete time survival analysis. *R package version 1*(1), 1.