University of Central Florida

Department of Economics

ECO6936: Final Draft

Modeling Default Probabilities in Peer-to-Peer Services

Nigar Sultana

July 31, 2020

1 Introduction

More than a decade ago, peer-to-peer (P2P) lending services began to connect borrowers and lenders directly through online platforms for micro-credit funding in the absence of traditional banking systems. This idea of lending (or borrowing) money outside the conventional financial system quickly garnered attention, and people began using the P2P service because of its lower transaction and intermediation costs compared to the traditional institutions—the key drivers of interest margins; see Maudos and De Guevara (2004). Moreover, the P2Ps operated outside banking regulatory systems using an online automated system, which implied not carrying loans on their books (thus avoiding liabilities for loans), which helped to minimize the operating cost of P2P services; see Serrano-Cinca (2015). Consequently, these lower costs are transferred across both the demand and supply network of the market. The benefits to the borrowers revolved around the speed of funding, higher funding rate, ease of the application process, reasonable interest rates, and the promise of non-collateralized loans often at competitive interest rates. On the other hand, lenders gained by receiving returns above market rates, diversifying their risk through a variety of transactions, and having the freedom to choose borrowers that match their preferences. The P2P lending platform does not, however, participate in lending decisions or collects deposits. Basically, it sets the rates and terms and enables the transactions. Since 2005, when the first P2P lending service company *(Zopa)* was founded, the global P2P lending market has valued 67.93 billion USD in 2019. Furthermore, it is projected to reach 558.91 billion USD by 2027—a cumulative annual growth rate of 30.14 percent.

Although P2P is growing fast, in addition to its highlighted benefits, there is a high exposure of lenders to default, because of asymmetric information. From the record of the P2P lending platform, lenders have little information concerning the borrowers. On the other hand, a borrower has near-complete information about his or her capability of loan repayment. Typically, registered borrowers post their funding requirements on the platform

and provide relatively limited information for due diligence purposes; see Cummins (2019). After that, a P2P lending platform uses a proprietary credit scoring mechanism to collect and score prospective borrowers individually or as a pool; see Cummins (2019). Then the potential loan requests are offered to the prospective lenders through the platform. Lenders select the loans they want to invest in based on their risk tolerance, investment portfolio goals, and time horizon. As one can see, throughout the process it is difficult for a lender to distinguish between high probability-of-default and solvent borrower; see Serrano-Cinca (2015). Given this fact, an investor who is trying to secure a high return on this platform, confronted with a critical question: how risky is the borrower?

In order to help an investor with this question, P2P services rate each requested loan with a grade that takes into account the borrower's credit history. This grading system represents the relative risk of default among different grades, that is, a lower grade represents a higher chance of default. As discussed earlier, P2P services also provide interest rates corresponding to each grade that, like the grading system, reflects the credit history of the borrower. Therefore, the lower the grade, the higher the default risk and, consequently, the higher the interest rate. An investor is, however, provided with a wide range of additional information about the borrower (for example, loan purpose, income verification, annual income). The contribution of this additional information to the information asymmetry, of course, requires an in-depth analysis. On this note, my goal is to study, whether this additional information—besides the grading system—can help the investors to have a more holistic view of each borrower and to take a more educated decision on whether or not to grant the loan. In other words, whether it can reduce the information asymmetry.

For my analysis, I used data from *Lending Club*, the biggest P2P company in the United States. The sample contains loans funded from 2012 to 2015 including two maturity periods: 36 months and 60 months. Loans with 36 months maturity period considered for the analysis because the loan funded for 60 months maturity period in December 2015, is still unknown. Discrete-time survival analysis was performed for explaining loan default. My

analysis suggests that the grade assigned by the P2P lending company is the best predictor of default. Other variables, for example, loan purpose, verification status, annual income, duration of several financial activities, last highest Fico score can, however, play significant roles to predict default.

The remainder of this paper is organized as follows. In section 2, I present a review of the empirical literature concerned with P2P lending. In section 3 describe the economic and statistical model, while in section 4 provide a discussion on the data and variables stored by the *Lending Club*. An exploratory data analysis is presented in section 5. while in section 6, the empirical results, and interpretations of the analysis. Finally, my conclusion are summarized in Section 7.

2 Literature Review

The idea of using survival analysis in the context of credit risk was first introduced by Narain (1992). He argued that the survival analysis can be used in any credit operation that has predictor variables and the time to event is of interest. Banasik (1999) compared the performance of exponential, Weibul, and Cox model with logistic regression and found that survival analysis methods perform competitively—sometimes superior to the logistic regression model. Cao (2009) used three different approaches to model the probability of default in consumer credits and personal credits by applying survival analysis; these approaches are Cox proportional hazard model, generalized linear model, and a random design nonparametric regression model. These models work within continuous time specification while discreteness is common in duration models in econometrics, which is why I incorporated it into my analysis.

Han and Hausman (1990) specified and estimated a flexible parametric proportional hazard model—based on the Cox model—that takes into account the discrete nature of the

duration data. Additionally, based on the concepts of Han and Hausman's paper, Paarsch and Golyaev (2016) specified a model that incorporates the time-varying covariates. Also, Singer and Willett (1993) provided a practical illustration for fitting the discrete-time hazard model by using standard logistic regression software. For my analysis, I adopted the formal model from Singer and Willett (1993).

Several kinds of research have emerged in response to the growing popularity of P2P lending services, especially on the basis of survival analysis. Byanjankar (2017) proposed a survival analysis approach to predict survival probabilities of loans in European peer-to-peer lending platforms at different time periods. Also, Đurović (2017) examined the relationship between loan characteristics and the probability of default in the P2P market, and Emekter (2015) investigated whether the higher interest rate is enough to compensate incremental credit risk. Around the same time, Serrano-Cinca (2015) found the factors explaining default risk are: interest rate, loan purpose, loan grade, income, credit history, and borrowers' indebtedness. Furthermore, Chen and Han (2012) conducted a comparative study of P2P services in the United States and China and found that two kinds of information, *hard* and *soft* information, have significant influences on lending outcomes in both countries. All these papers significantly contributes to my analysis by giving crucial insights about the potential variables that can help to predict the occurrence of default in P2P lending services.

3 Economic and Statistical Models

Survival analysis consists of a set of statistical approaches used to investigate the time it takes for an event of interest to occur. For example, modeling the duration of unemployment or the time to retirement. Most often in these analyses, it is assumed that the survival time is given by a random variable measured on a continuous scale. In practice, however, duration is often measured in days, years, or months, which are discrete.

Since credit performance data are recorded monthly, I used 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 an 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 experiences a target event (for example, in the context of loan default, the event of interest is the default). Also, hazard function assesses whether and when an event occurs.

3.1 Formal Model

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 j^{th} 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.

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 k predictors, Z_k $(k = 1, 2, ...K)$ into the definition. The individual n's values for each of the k predictors in time period j can be denoted as the vector $z_{nj} = [z_{1nj},\ z_{2nj}, ..., z_{Knj}]$. Now, the discrete-time hazard, h_{nj}, can be defined as conditional probability as follows:

$$h_{nj} = Pr\ [T_n = j |\ T_n \geq j, Z_{1nj} = z_{1nj}, Z_{2nj} = z_{2nj}, ...Z_{Knj} = z_{Knj}]$$

Here, h_{nj} defines the probability that individual n who can be distinguished by his or her predictor values $z_{1nj}, z_{2nj}, ..., z_{Knj}$—experiences the event in time period j, given that he or she survived through all prior events.

3.2 Statistical Model

Cox (1972) proposed that, since h_{nj} 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. The proposed population discrete-time hazard model by Singer and Willet is:

$$h_{nj} = \frac{1}{1 + \exp[-(\alpha_1 D_{1nj} + \alpha_2 D_{2nj} + \cdots + \alpha_J D_{Jnj}) + (\beta_1 Z_{1nj} + \beta_2 Z_{2nj} + \cdots + \beta_K Z_{Knj})]} \tag{1}$$

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 adopted to estimate the hazard probabilities.

$$\log\left(\frac{h_{nj}}{1 - h_{nj}}\right) = (\alpha_1 D_{1nj} + \alpha_2 D_{2nj} + \cdots + \alpha_J D_{Jnj}) + (\beta_1 Z_{1nj} + \beta_2 Z_{2nj} + \cdots + \beta_K Z_{Knj}). \tag{2}$$

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; see (Singer and Willett, 1993). The α parameters $[\alpha_1, \alpha_2, \ldots \alpha_J]$ represent multiple intercepts one per time period. When all the values of the covariates are set to zero, the $[\alpha_1, \alpha_2, \ldots \alpha_J]$ represent the population baseline logit-hazard function. It captures

tie period by time period conditional log-odds that the individual whose covariate values are all zero (the baseline group) will experience the event in each time period, conditional on they not already experienced it before.

3.3 Parameter Estimation

The estimators of discrete-time hazard model in equation (1) can be fitted by the event history data of randomly sampled n individuals. The method of maximum likelihood can be used to obtain the estimators for the parameters $[\alpha_1, \alpha_2, \ldots, \alpha_J,]$ and $[\beta_1, \beta_2, \ldots, \beta_K,]$ in equations (1) and (2). Singer and Willet specified the likelihood function as follows:

$$\mathcal{L} = \prod_{n=1}^{N} \prod_{j=1}^{j_n} h_{nj}^{ynj} \ (1 - h_{nj})^{(1-y_{nj})} \tag{3}$$

Equation (3) represents the likelihood function for the discrete-time hazard process where y_{nj} indicates the data, and h_{nj} specifies the hazard probability parameters. The maximum likelihood estimates of $\alpha_1, \alpha_2, \ldots, \alpha_J, \beta_1, \beta_2, \ldots, \beta_K$ can be obtained by maximizing the likelihood in equation (3), under the logistic reparameterization in equation (2).

4 Data Description

The borrowers' loan data were recorded in the csv format each year. For the purpose of this project, I have selected the completed loans (that is, either fully paid or defaulted), which were issued between January 2012 and December 2015. The data comprise of demographic and financial information of the borrowers and the corresponding loan information. A data dictionary is provided in a separate file in the data frame. In total, there were three data frames, that is, one for each year.

4.1 Data Preparation

I conducted the analysis in R (Team, 2013). First, I imported the data frames in R and combined them; the combined data frame had 844,907 rows and 150 variables. The variables can be classified into three types: numerical, categorical, and date. Moreover, these variables included missing observations, formatting problems, and multicollinearity issues. After the cleaning and transformation process, the data frame reduced to 67 variables. Though the data consisted of 36-month and 60-month loan data, for the simplicity of my analysis, 36-month loans have been selected. The 60-month loans have been excluded because the loan status information that funded in December 2015 cannot be known until December 2020. Finally, there were 589,636 loans for 36-month period, and 82,774 loans defaulted from them. The comprehensive data-wrangling procedure is described in Appendix A.

4.2 Variables

The outcome variable in my analysis is the loan default event among the borrowers. I consider a loan has defaulted if the borrower missed four consecutive monthly payments, i.e., four months passed since the last payment. The *loan outcome* variable indicated the default status by coding 0 for fully paid and 1 for default. The predictive variables can be grouped into four categories: borrower characteristics, duration of several financial activities, loan description, and other accounts' information

The detail description of predictive variables under these four categories are provided in Appendix B.

5 Exploratory Data Analysis

In this section, I present an exploratory data analysis to examine the underlying structure of the data set. The main objective is to obtain an essential understanding of the data, and investigate whether the variables follow any pattern.

Annual Income

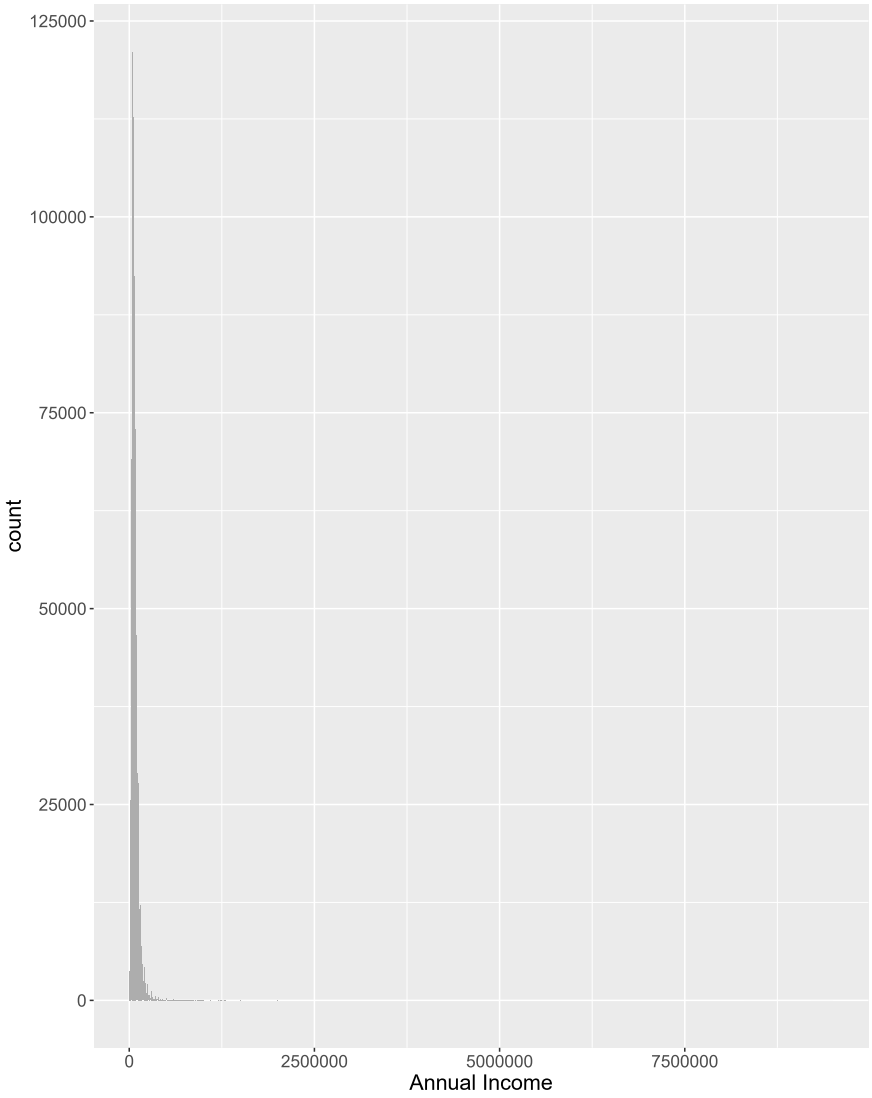

Figure 1: The Histogram of Annual Income

From the distribution of annual income in figure 1, one can see that the income

is skewed to the right. The summary statistics of Table 1 below, shows that the maximum income is 9,500,000 USD while the median income is 65,000. Detail inspection in the spreadsheet has revealed that the there are 38 borrowers with more than 1 million USD annual income. It might be the reason that the investors took some loan from the *Lending Club* platform to assess the procedures before making investment decision.

Table 1: Summary Statistics of Annual Income

Min	1st Qu.	Median	Mean	3rd Qu.	Max
0	45600	65000	75323	90000	9500000

Employment Length

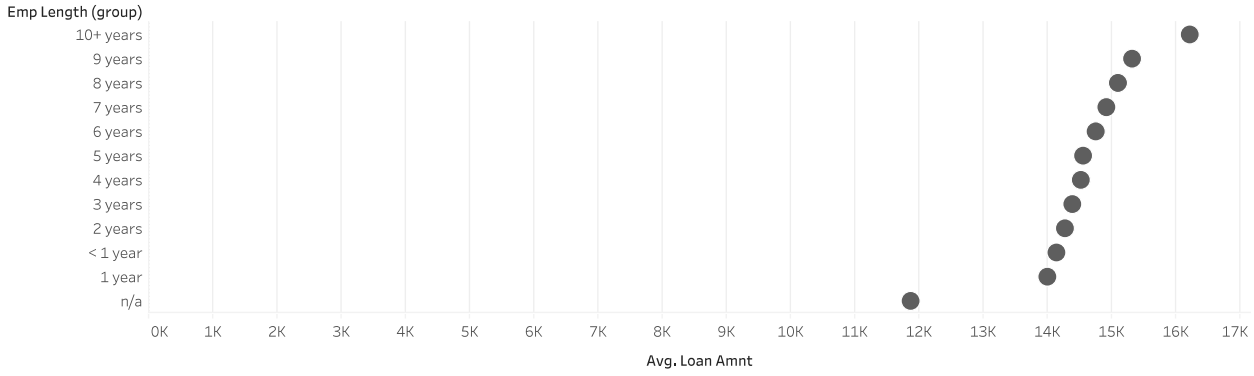

Figure 2: Average Loan amount for Each Employment Length

Average loan amount increases gradually with the employment length. Some observations in the variable were recorded as n/a by Lending Club . I have considered n/a as a category.

Loan Amount

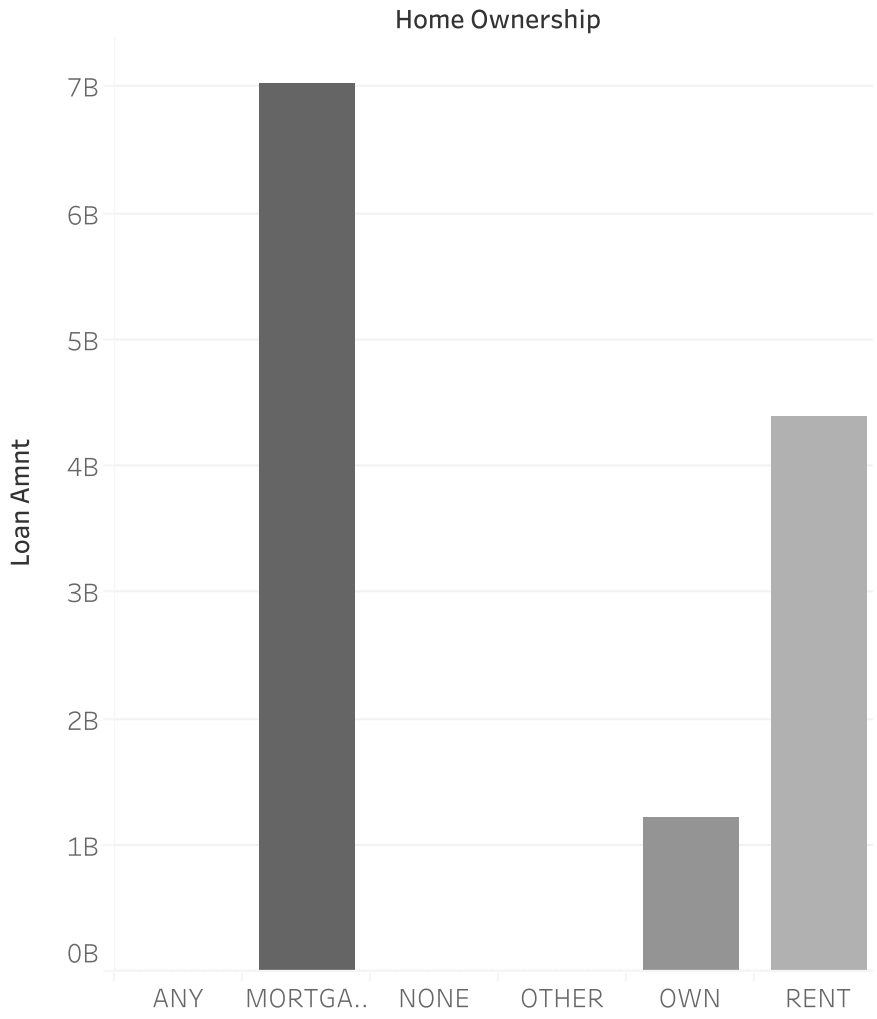

Figure 3: Sum of Loan Amount for Each Home Ownership

Borrowers who pay mortgage took the highest loan amount. Borrowers who have home-ownership status labeled with any, none, and other took very negligible amount of loan compared to the mortgage, own, and rent.

Loan Purpose

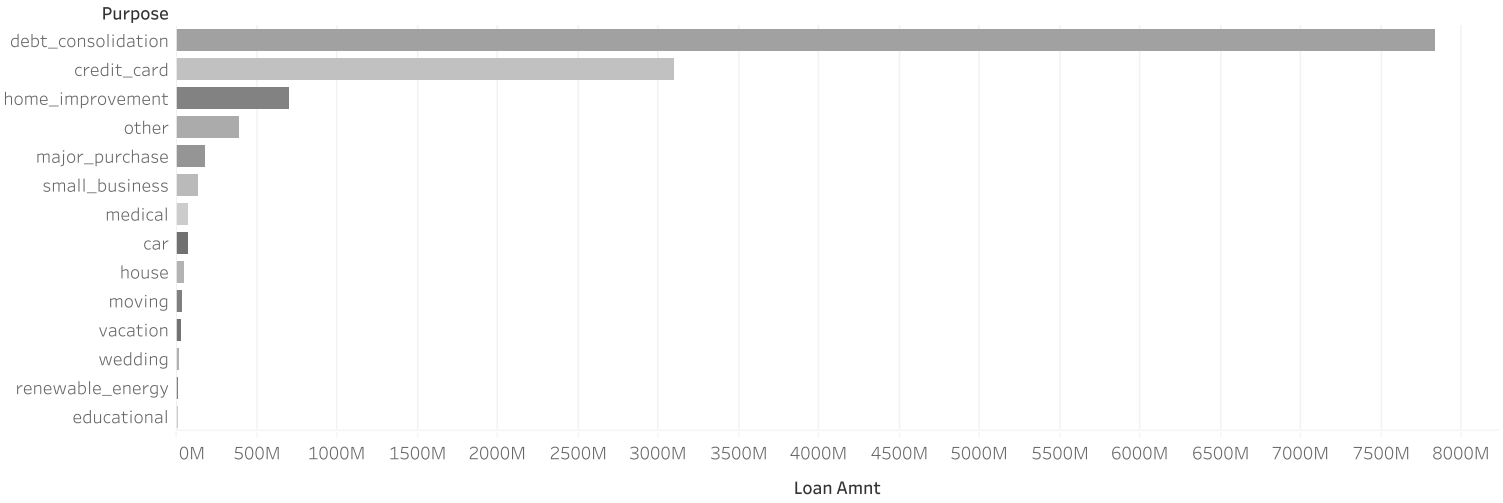

Figure 4: Loan Amount by Purpose

Most of the loan was taken for debt consolidation purpose.

Verification Status

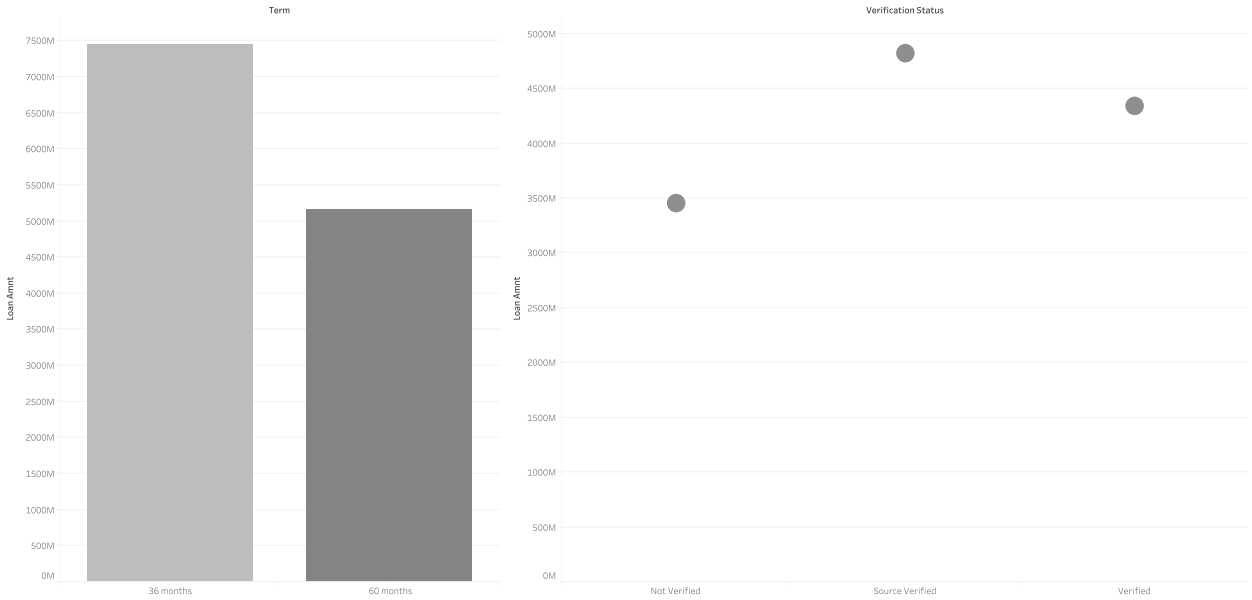

Figure 5: Total Loan Amount and Verification Status

Total loan amount is higher for 36 months term. Also, most of the loan was given for verified source borrower.

Loan Amount and Annual Income

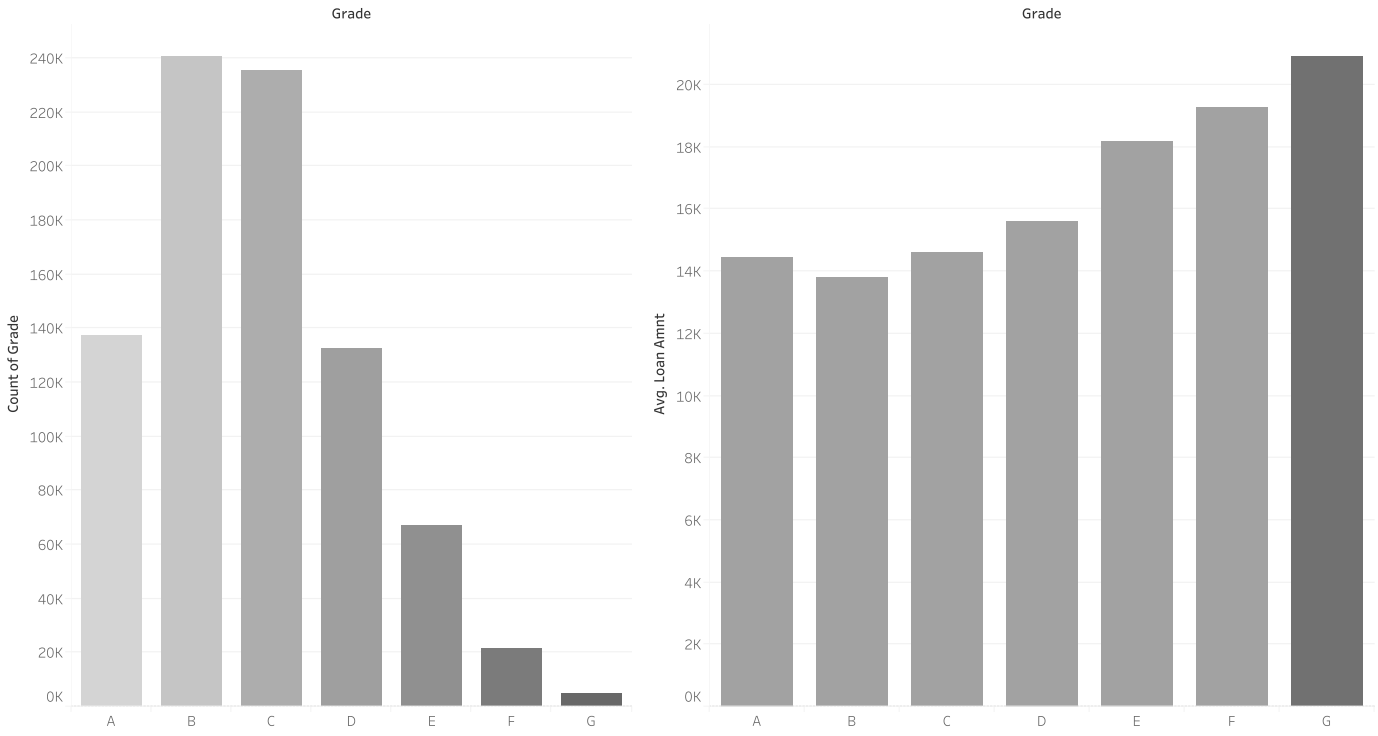

Figure 6: Loan Amount and Annual Income

The left side of the graph shows the total number of borrowers according to their grades. The G grade has lowest number of borrowers. On the right side of the graph, we can see the average loan amount for each grade where G grade borrowers took the highest average loan amount. Therefore, it can be concluded that the borrowers with grade G took the highest loan amount.

Time Series Analysis: Interest Rate

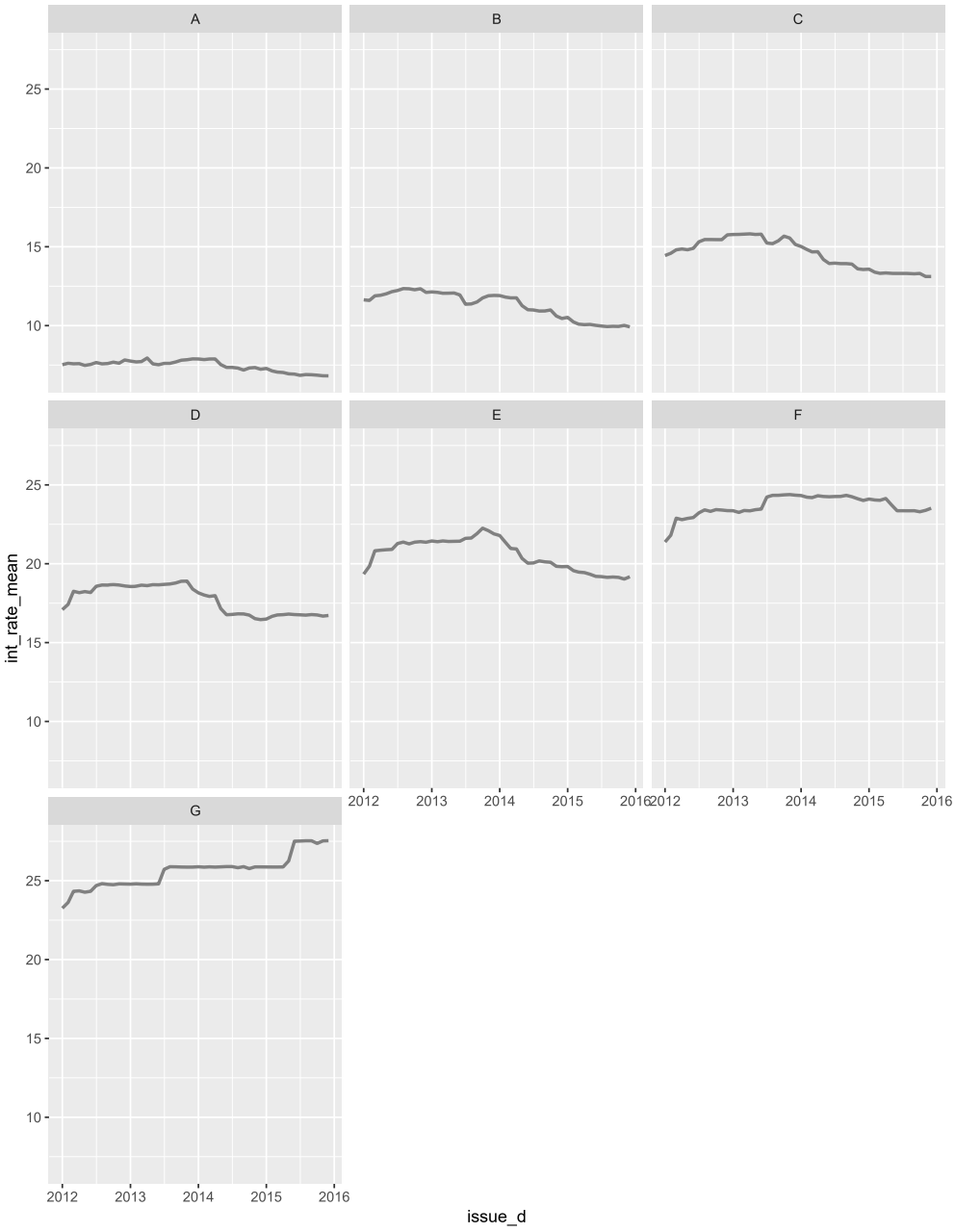

Figure 7: Time Series Analysis of the Interest Rate

From grade A to E, the trend is similar. The average interest rate of grade G has increased over time.

Time Series Analysis: Loan Amount

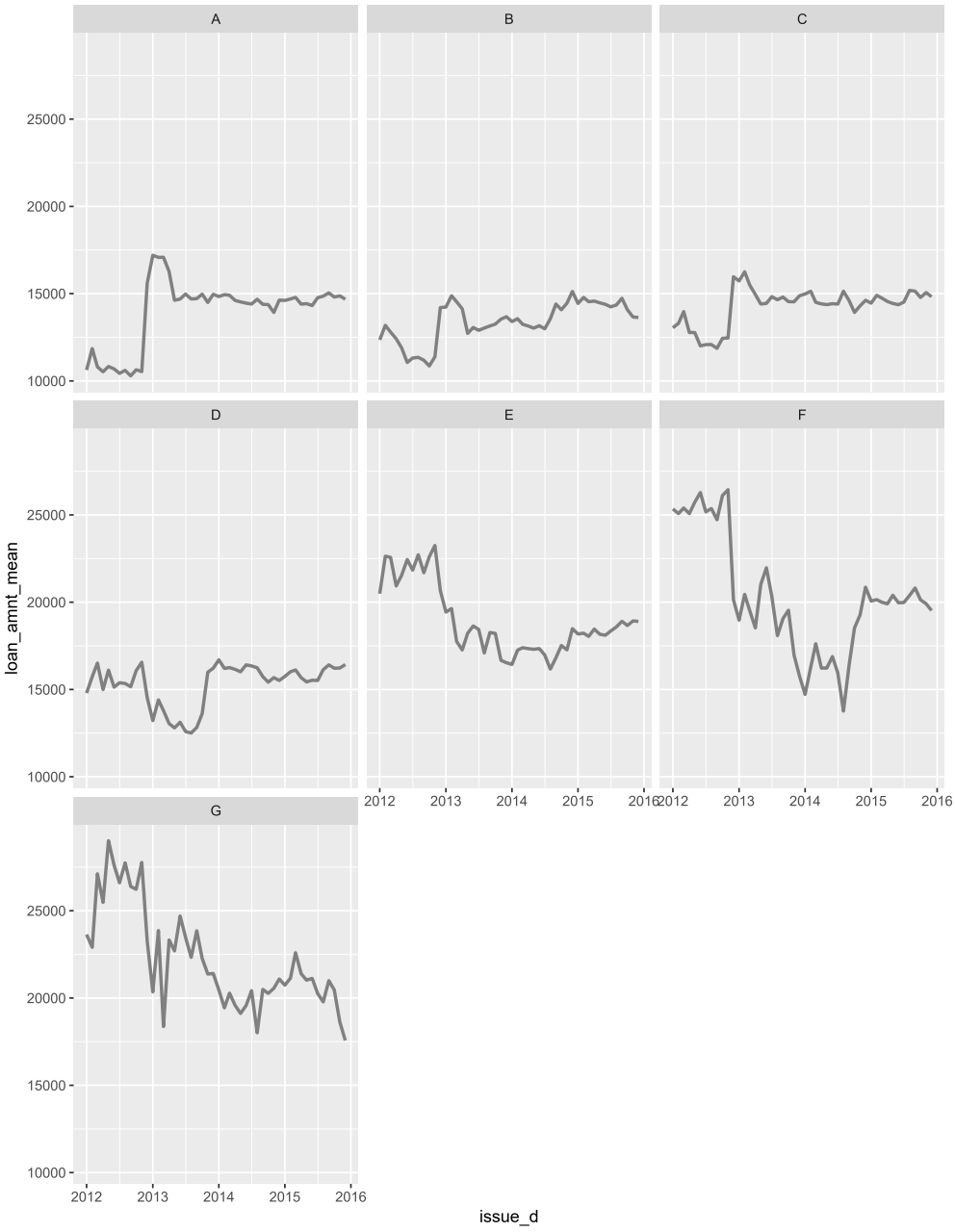

Figure 8: Time Series Analysis of the Loan Amount

Average loan amount fluctuates over time. The trend is similar for grade A to C; there is a spike in 2013. The fluctuation is moderate for grade D and E compared to others. There is a high fluctuation for grade F and G over the time.

6 Empirical Result

In order to use standard logistic regression analysis in the discrete-time hazard model, data need to be transformed before analysis. Typically, discrete event history data are in the format of person-oriented data set in which each person in the sample has one record (line) of data. Before conducting 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 added "duration-quarters" variable as a duration variable with a quarter basis scale which ranges from 1 to 14 (each scale represents three months). Therefore, respective dummy time variable has been computed to indicate month: 14 time dummy variables to indicate in total 42 months.

In R, the person-period data matrix can be generated by applying the function dataLong() in the R package *discSurv*; see Welchowski and Schmid (2015).

Table 2: Person-Period Data Matrix

obj	timeInt	y	loan-amnt	grade	purpose	durtion-quarter
1	1	0	4800	B	home-improvement	3
1	2	0	4800	B	home-improvement	3
1	3	0	4800	B	home-improvement	3
2	1	0	12000	B	debt-consolidation	7
2	2	0	12000	B	debt-consolidation	7
2	3	0	12000	B	debt-consolidation	7
2	4	0	12000	B	debt-consolidation	7
2	5	0	12000	B	debt-consolidation	7
2	6	0	12000	B	debt-consolidation	7
2	7	0	12000	B	debt-consolidation	7

Table 2 collects the first ten observation of seven columns of person-period data

matrix. Each record contains information on three types of variable: (a) the time indicators, (b) the predictores, and (c) the event indicators. The time duration has been taken in four months period. Three new columns have been added; obj, which is an identifier of the individual, timeInt, which contains the discrete time values and y, which contains the binary response variables. In the table, the first individual (obj=1) had an event in 3rd quarter (duratuion-quarter = 3 and loan-outcome is zero which means fully paid). Accordingly, the person-period data matrix for the first individual has three rows, where each row corresponds to one time interval (timeInt = 1,2,3). The corresponding vector of responses is y= (0, 0, 0). The values of the covariates remain constant over time and are therefore the same in each row.

6.1 Empirical Model

After examining several models, I present four discrete-time hazard models to fit the *Lending Club* data. The variables included in the models are described in Table 11 under Appendix B. Model A describes the conceptual predictor in the analysis, that is, time. The model answers the most fundamental question: "what is the risk of event occurrence in each time period?" Model B adds variables related to borrowers' characteristics and duration of borrowers' loan and financial activities to model A. Afterward, loan description variables are added to model B to formulate model C. Finally, all variables from four categories are comprised in model D.

Every model is included within the next model, for example, model A is nested within model B, and so on. Since all the four models are nested within one another, I compared the goodness-of-fit statistics of these four models using the likelihood ratio test. If added predictors are not associated with risk, the extended model will fit no better than the reduced model; if added predictors are associated with risk, then fit will improve; see Willett and Singer (1993). Overall, model D has improved the fit of the hazard model with

a chi-square goodness-of-fit statistic of 619.88 and a p-value of ($p < 0.001$).

6.2 Result Analysis

Fig 9 displays the survival curves for each loan grade. The probability of survival is higher for grade A borrower. the survival of borrower gradually decreases with increased grade. Survival function can be useful for lenders, because it shows the probabilities of default at a certain point of time.

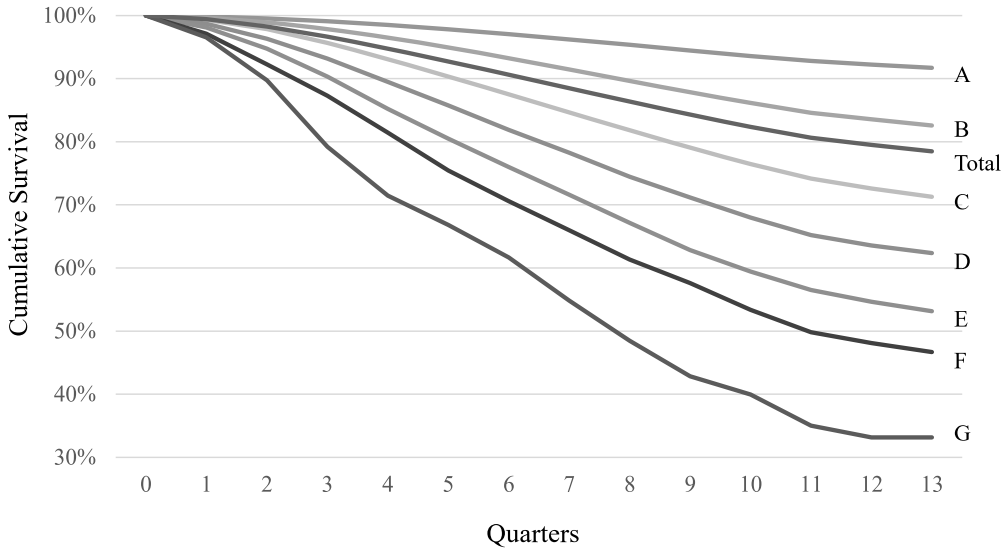

Figure 9: Survival Function for Grade

Now, I present interpretations for the significant variables according to p-values ($p < 0.001$) from the logistic regression of model D. The variables are described orderly according to the categories which I defined in Section 4. The overall empirical results are provided in Table 11 under Appendix C.

6.2.1 Borrower Characteristics

Grade:

Table 3: Coefficient Estimate and Odds-Ratio for Each Grade

Grade	Estimate	Odds Ratio	p-value
B	0.30	1.38	0.0000
C	0.48	1.62	0.0000
D	0.68	1.97	0.0000
E	0.79	2.20	0.0000
F	0.95	2.58	0.0000
G	1.44	4.22	0.0000

Table 3 shows the coefficient estimate and odd-ratio of grade B to grade G while keeping all other variables constant. The grade A used as a reference category for all other grades, hence no coefficient. In other words, the time indicator estimate represents the log-odds of defaulting for a borrower in the grade A category. The odd ratio for grade B is 1.38, which means compared to a grade A borrower, a grade B borrower has 38% higher odds of default. The odd ratio increases as I go along the grade sequence that reflects the fact that a lower grade represents a higher risk of default.

Loan Purpose:

There are eleven types of self-reported loan purposes. Table 4 shows the types that are significant in predicting default according to p-values ($p < 0.001$). The reference category used for estimating coefficients is car loan. The odd-ratio column represents the odds of default for the borrowers taken loan for these seven purposes compared to the loan taken for car purpose.

Table 4: Coefficient Estimate and Odds-Ratio for Loan Purpose

Purpose	Estimate	Odds-Ratio	p-value
Debt-Consolidation	0.29	1.34	0.0000

Credit Card	0.48	1.61	0.0000
Small business	0.95	2.58	0.0000
Vacation	-0.17	0.84	0.0035

The loan funded for small-business has 158 percent higher odds of default than the loan given for car purpose. The possible reason could be because small businesses have less annual incomes and minimum government supports compared to the regular-sized business or corporation in case of business failure. According to the Small Business Administration(SBA), 20 percent of small businesses fail in the first year, 50 percent become bankrupt after five years, and only 33 percent make it to ten years or longer.

On the other hand, debt-consolidation indicates that the borrower took the loan to pay debts that have been combined from several creditors into one monthly payment, which is risky, hence shows higher odds. Also, the loan took for credit card purposes reveals that the borrower was unable to pay the credit card bill regularly. As a result, loans taken for both of these purposes are highly likely to default. On the contrary, vacation shows to be a more secure purpose to fund loan compared to the others in table 4. In summary, more risky loan purposes have higher odds of default. Therefore, besides grade, an investor must take into account the loan purpose factor.

Verification-Status and Initial-List-Status:

There are three categories in verification-status: source verified, verified, and not verified. Source verified category indicates if the income source was verified, and verified category indicates if the income was verified by *Lending Club*. The borrowers who had not verified their income by *Lending Club* were enlisted in the not-verified category. The reference group used for estimating coefficients is not verified group. Also, the initial-list-status has two categories: f And w. The f category used as reference group.

Table 5: Coefficient Estimate and Odds-Ratio for Verification-Status and Initial-List-Status

Variables	Estimate	Odds-Ratio	p-value
Verification-Status: Source-Verified	0.06	1.06	0.0000
Verification-Status: Verified	0.04	1.04	0.0000
Initial-List-Status: w	-0.04	0.96	0.0000
Annual Income	-0.00		0.0000

In Table 5, the source-verified and verified incomes have 6 percent and 4 percent higher odds of default compared to the not-verified source respectively. Also, borrower with "w" initial-list-status has 4 percent lower odds of default than the borrower with "f" initial-list-status.

Annual Income

Although the estimate of annual income indicates a value very close to zero, it is significant according to the p-value. Exploratory data analysis showed that the annual income ranges between zero to nine million USD. As a result, a very small coefficient can multiply out to have a large effect ($1.239 \times 10^{-6} \times 10^{6} = 1.239$), Typically, it is caused by the wide variety of units of variables used for analysis. Overall, the higher annual income enable the borrower to fully pay the loan, consequently, lower the risk of default.

6.2.2 Duration of Financial Activities

Delinquency and Public Record:

Duration related to the borrowers financial activities represents their experience with debts and behavior toward financial crisis. The coefficients estimates of the variables under the duration of financial activities category are given in table 6. Here, one of the most interesting observation is delinquency record. It records the number of incidences of

delinquency in the borrower's credit file for the last two years; delinquency refers to the incident when a borrower is late on his/her payment for more than 30 days. It shows a negative coefficient -0.02 with p-value 0.0011. It represents that the borrowers with multiple delinquency showed resiliency toward financial crisis. In other words, despite having record of multiple missing payments, the borrowers paid back consistently. As a result, the log-odds of default is decreased with the increasing delinquency records. Though it has a relatively less significant p-value, I discuss it here to shed light on the fact that sometimes human perception can be wrong.

Table 6: Coefficient Estimate for Duration of Borrower's Financial Activities

Variables	Estimate	p-value
Delinquency 2 Years	-0.02	0.0011
Number-of-Account-Opened-Past-12-Months	0.01	0.0001
Number-of-Account-Open-Past-24-Months	0.02	0.0000
Inquiry-Last-6-Months	0.05	0.0000
Month-Since-Recent-Bankcard-Account-Opened	-0.001	0.0008
Public-Record	-0.04	0.0000

Inquires and The Number of Opened Accounts:

Typically, a credit inquiry occur when people apply for a loan and permit the lenders to check credit report. If a borrower takes numerous loans in a short time, it could be interpreted as his/her financial hardship. Subsequently, it is possible that he/she will struggle to manage all the repayments, which can lead to loan default. Therefore, multiple loan applications may resulted in several loan inquiries against a borrower, and it can increase the log-odds of default. The explanation is also applicable for public-record.

The number-of-account-opened-past-twelve-months and the number-of-account-open-past-twenty four-months have positive coefficients. The possible explanation can be that the

borrower who opened multiple accounts in last 12 to 24 months was unable to manage the different types of credit accounts and pay the debt. The last 24 months opened account information is, however, more significant than the last 12 months; a larger temporal window of opened accounts provides more information on the borrower. On the other hand, months-since-recent-bankcard-account-opened has a negative coefficient as the higher number shows a borrower's good performance with managing the debt and, consequently, longer time interval between successive opening of credit accounts.

6.2.3 Loan Description

Table 7: Coefficient Estimate for Loan Description

Variables	Estimate	p-value
Funded-Amount-Investment	0.0005	0.0000
Debt-to-Income-Ratio	0.01	0.0000
Fico-Range-Low	0.001	0.0001
Total-Payment	-0.001	0.0000
Last-Fico-Range-High	-0.01	0.0000
Last-Payment-Amount	-0.001	0.0000
Total-Open-to-Buy-Revolving-Bankcards	0.000	0.0005
Total-High-Credit-Limit	-0.000	0.0000
Total-Bank-Account-Limit	-0.000	0.0000
Total-Installment-High-Credit-Limit	-0.000	0.0000

Funded-Amount-Investment:

It represents the total amount of loans committed by investors. Interestingly, the positive coefficient can be attributed to risky behaviors of investors, that is, invests in risky loans for a higher interest rate. Also, a positive coefficient for debt-to-income-ratio indicates

that borrowers with increased debt are more likely to default, which intuitively makes sense. Moreover, the fico-range-low indicates a borrower's lowest FICO score, which illustrates that the borrower may have a hard time paying off his/her loans in the past; hence, it increases the risk of default by contributing a positive coefficient.

Credit Management Performance:

Total-payment, last-fico-range-high, last-payment-amount, total-high-credit-limit, total-bank-account-limit, and total-installment-high-credit-limit are related with borrower's credit management performance. The better a borrower manages his/her credit, the less riskier he/she is, hence lower the log-odds of default. Note that, all of these variables have very small coefficients. I took only three digits after the decimal, which is why some coefficient estimates show zero. Also, these variables range from zero to millions. As a result, a very small coefficient can have a large effect.

6.2.4 Different Account's Information

Table 8: Coefficient estimate for different account's information

Variables	Estimate	p-value
Total Account	0.01	0.0000
Number-of-Currently-Revolving-Trades	-0.03	0.0000
Number-of-Installment-Account	-0.01	0.0000

Though total accounts are considered as an added advantage in the FICO score, it is arguable whether a high number of total accounts represents an important attribute of a less risky borrower. One can make such argument from Table 8 where the positive coefficient of total accounts represents a high risk with more total accounts.

On the other hand, with an installment account, a borrower can pay a fixed payment

for a fixed time, which represents the borrower's sincerity toward loan repayment. Also, a borrower with a revolving trade account is allowed to make a payment that is the same every month until the loan is paid in full. The flexible characteristics of these accounts help the borrower to fully pay back his/her loan. Therefore, a higher number of these accounts related to a less risky borrower.

7 Conclusion

P2P lending companies offer a platform to connect lenders and borrowers with a common shared interest, that is, buying or selling money. Unlike conventional banking systems, however, this platform is online and the process is automated. Like many online businesses, of course, it capitalizes on technologies that reduce the operational cost, which translates to less interest margin. Consequently, this leads to improved efficiency, which is a very important attribute of a market where money is bought and sold. Information asymmetry, however, is a crucial challenge for P2P lending companies where lenders have limited knowledge of the prospective borrower.

Although P2P sites provide a lot of information to reduce information asymmetry, generally, investors assume the credit risk only by considering the grade provided by the P2P services, which translates into an interest rate. In my analysis, I found that grade is indeed the most significant indicator in predicting default. Besides grade, however, loan purpose is also a factor that can explain default—vacation is the least risky loan purpose and small business is the riskiest. Annual income, inquiries in the last six months, public record, number of accounts opened in the past twenty-four months, debt to income ratio, last highest FICO score, total bank account limit, total payment, number of installment account, number of currently revolving trades are also relevant variables.

In summary, I learned that sometimes lenders may take irrelevant or insignificant

information into account when deciding to lend (that is, factors with a higher p-value). Also, some factors may be counter-intuitive in real life (that is, showing opposite sign of coefficient), which can be also attributed to personal bias. Information asymmetry, however, can be reduced if one can educate investors with appropriate and instructive information.

APPENDIX

In this appendix, I document the development of the data set used, data wrangling process as well as the transformation used in obtaining the final data set. The comprehensive data wrangling procedure is described in section A. Description of the variables used in the models are provided in section B. Finally, the empirical results are provided in Table 11 under section C.

A Data Wrangling

A.1 Deleted Variables

First, I replaced the blank cells of the data frame with the "NA" values. Then, I removed those variables which does not have any other values rather than NA. In addition to that, I removed several variables for different reasons which are given in the following:

Pymnt-Plan: Payment plan variable has same categorical value for all the observation. Since it has the same categorical value for all the observation, the variable is not going to offer any significant information about the borrowers. Therefore, I decided not to include this in my data frame.

Policy-Code: Same numerical value for all observation.

Application-Type: 844,394 borrowers have individual application type. The joint application type is insignificant compared to individual. Since it has almost same categorical value for all the observation, the inclusion of this variable will not provide any significant insight.

Url: The url variable does not have any relevant information regarding the default. As a result, I deleted the url variable.

Desc: Description variable records the borrowers' description for the loan purpose. There are other variable which also records the borrower's loan purpose in a short term. From my analysis, I noticed that the loan purpose variable encrypts the essence of the loan description. Hence, I selected loan-purpose variable to include as a features, and delete the description variable.

Zip-Code: There are two variables related to the location of the borrowers; zip code and state. I used the state as the location of the borrowers and removed the zip-code variable.

Title and Emp-title: The emp-title variable records the employment description of the borrowers. The loan title provided by the borrower is registered in the title variable. These two variable have large share of unique categorical values. That's why, I deleted these variable.

Grade: Lending Club assigns loan grade from A to G for the borrowers. Sub-grade is classified into 5 sub-group for each of the grade. For example, grade A is divided into A1 to A5 sub-group. Overall, there are 35 subgroup. Since both variables are related to each other, I excluded the "sub-grade" variable from the model.

Recoveries and Collection-Recovery-Fee: The recoveries and collection-recovery-fee variables document post charge off gross recovery. Both of the variables have almost same values for each observation.

Variables with Similar Observation: I excluded "collections-12-mths-ex-med","out-

prncp-inv", "hardship-flag", "out-prncp" variables for having almost identical value for each observation.

Multicolinearity: Since, all the loan is funded by the investors, the "loan-amnt" is highly correlated with the "funded-amnt" (0.99) and "funded-amnt-inv" (0.99). As a result, I chose to omit the "funded-amnt" from the model. The variable "installment" has strong correlation with "loan-amnt" and "funded-amnt-inv". Borrowers pay high installment for high loan amount. Therefore, I excluded the "installment" in the model. Similarly, I omitted the "num-sats" and "total-rec-prncp" variable due to the strong correlation (0.96) with the "open-acc" and "total-pymnt-inv" respectively.

A.2 More than Sixty Percent Missing Values

I removed 28 of 31 variables with more than sixty percent missing values. The variable "mths-since-last-major-derog" keeps record of months since most recent ninety days or worse rating, and the variable "mths-since-last-record" registers the number of months since last public record. The NA values might indicate that there was no record of missed payment. I replaced the missing values of these two variables with 0.

Table 9: Variables for which More than Sixty Percent Missing Values

Variable name	Count of Missing Values
mths-since-last-record	711494
mths-since-last-major-derog	623184
annual-inc-joint	844396
dti-joint	844398
e open-acc-6m	823535
open-act-il	823535
open-il-12m	823535

open-il-24m	823535
mths-since-rcnt-il	824097
total-bal-il	823535
il-util	826290
open-rv-12m	823535
open-rv-24m	823535
max-bal-bc	823535
all-util	823535
inq-fi	823535
total-cu-tl	823535
inq-last-12m	823535
mths-since-recent-bc-dlq	637275
mths-since-recent-revol-delinq	553957
deferral-term	840158
hardship-amount	838931
hardship-length	840158
hardship-dpd	840158
orig-projected-additional-accrued-interest	839545
hardship-payoff-balance-amount	838931
hardship-last-payment-amount	838931
settlement-date	825230
settlement-percentage	825230
hardship-type	840159
hardship-reason	840158
hardship-status	840158
hardship-start-date	840158
hardship-end-date	840158
hardship-loan-status	840161

settlement-status	825230
settlement-term	825230
verification-status-joint	844396
next-pymnt-date	823516

A.3 Handling the Missing Values

There are 61 variables with less than sixty percent missing observations.

Replacing Missing Values of Numerical Variables: I replaced the missing values of the numerical variables in two ways; replace with median, and replace with zero

Replacing with Median:

funded-amnt
loan-amnt
funded-amnt-inv
annual-inc
dti
fico-range-low
tot-hi-cred-lim

Replacing with Zero:

mths-since-last-delinq
mths-since-last-record
mths-since-last-major-derog
installment
inq-last-6mths
open-acc
pub-rec
revol-bal
total-acc
total-pymnt
total-pymnt-inv
total-rec-prncp
total-rec-int
last-pymnt-amnt
last-fico-range-high
last-fico-range-low
acc-now-delinq
tot-coll-amt

acc-open-past-24mths

bc-util

chargeoff-within-12-mths

delinq-amnt

mo-sin-old-il-acct

mo-sin-old-rev-tl-op

mo-sin-rcnt-rev-tl-op

mo-sin-rcnt-tl

mort-acc

mths-since-recent-bc

mths-since-recent-inq

num-accts-ever-120-pd

num-actv-bc-tl

num-actv-rev-tl

num-bc-sats

num-bc-tl

num-il-tl

num-op-rev-tl

num-rev-accts

num-rev-tl-bal-gt-0

num-sats

num-tl-120dpd-2m

num-tl-30dpd

num-tl-90g-dpd-24m

num-tl-op-past-12m

pct-tl-nvr-dlq

percent-bc-gt-75

pub-rec-bankruptcies

tot-cur-bal

total-rev-hi-lim

avg-cur-bal

bc-open-to-buy

total-bal-ex-mort

total-bc-limit

total-il-high-credit-limit

revol-util

Categorical Variables: The emp-length variable has 43,723 missing values. The *Lending Club* has labeled the missing observation as n/a. I kept those observation and use the label n/a as a category. The other categorical variables have two missing values for the same observations. I removed those two observation.

Date Variables: The "last-pymnt-d" and "last-credit-pull-d" have some missing observations. "Last payment date" records the date when last payment was received. I replaced the missing value of "last-pymnt-d" with the maximum date. The "Last-credit-pull-d" registers the last date of credit inquiries. I chose to replace the missing observations with the median.

B Variables Description

Table 10: Coefficient Estimate for Different Account's Information

Variables	Description
Borrower's Characteristics	The variables that pertains to borrower's demographic information are included in this group
Grade	*Lending Club* assigned loan grade
Purpose	A category provided by the borrower for the loan request.
Verification-Status	Indicates if income was verified by LC, not verified, or if the income source was verified
Initial-List-Status	The initial listing status of the loan. Possible values are w, f
Annual Income	The self-reported annual income provided by the borrower during registration
Duration of Several Financial Activities	This group contains variables that record the duration of borrower's several financial activities
Delinq-2Yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
Mths-Since-Last-Record	The number of months since the last public record.
Mths-Since-Last-Delinq	The number of months since the last delinquency.
Mths-Since-Last-Major-Derog	Months since most recent 90-day or worse rating
Acc-Open-Past-24Mths	Number of trades opened in past 24 months.
Mo-Sin-Old-ill-Acct	Months since oldest bank installment account opened
Mo-Sin-Old-Rev-Tl-Op	Months since oldest revolving account opened
Mo-Sin-Rcnt-Tl	Months since most recent account opened
Mths-Since-Recent-Bc	Months since most recent bankcard account opened.

Mths-Since-Recent-Inq	Months since most recent inquiry.
Num-Tl-30Dpd	Number of accounts currently 30 days past due (updated in past 2 months)
Num-Tl-90g-Dpd-24m	Number of accounts 90 or more days past due in last 24 months
Num-Tl-Op-Past-12m	Number of accounts opened in past 12 months
Inq-Last-6Mths	The number of inquiries in past 6 months
Loan Description	Loan description variables document the borrower's activities related to the loan
Dti	A ratio calculated using the borrower's total monthly debt payment on the total debt obligations.
Revol-Util	Revolving line utilization rate
Fico-Range-Low	The lower boundary range of he borrower's fico score
Total-Pymnt	Payments received to date for total amount funded
Last-Pymnt-Amnt	Last total payment amount received
Last-Fico-Range-High	The upper boundary range of he borrower's fico score
Tot-Cur-Bal	Total current balance of all accounts
Avg-Cur-Bal	Average current balance of all accounts
Bc-Open-To-Buy	Total open to buy on revolving bankcards.
Tot-Hi-Cred-Lim	Total high credit/credit limit
Total-Bc-Limit	Total bankcard high credit/credit limit
Total-Il-hgh-Credit-Limit	Total installment high credit/credit limit
Funded-Amnt-Inv	The total amount committed by investors for that loan at that point in time.
Other Accounts Information	Variables under other account information document the count of the borrower;s account outside the loan activities
Pub-Rec-Bankruptcies	Number of public record bankruptcies
Revol-Bal	Total credit revolving balance

Total-Acc	The total number of credit lines currently in the borrower's credit file
Acc-Now-Delinq	The number of accounts on which the borrower is now delinquent.
Bc-Util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
Delinq-Amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
Mort-Acc	Number of mortgage accounts.
Num-Accts-Ever-120-Pd	Number of accounts ever 120 or more days past due
Num-Actv-Bc-Tl	Number of currently active bankcard accounts
Num-Actv-Rev-Tl	Number of currently active revolving trades
Num-Il-Tl	Number of installment accounts
Percent-Bc-Gt-75	Percentage of all bankcard accounts > 75percent of limit.
Pub-Rec	Number of derogatory public records

C Empirical Result

Table 11: Regression Results of Models

Variable	Model A	Model B	Model C	Model D
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
factor(timeInt)1	-5.16(0.02)	-6.21(0.05)	-0.79(0.18)	-0.71(0.18)
factor(timeInt)2	-4.44(0.01)	-5.48(0.04)	0.93(0.18)	1.00(0.18)
factor(timeInt)3	-4.12(0.01)	-5.15(0.04)	2.00(0.18)	2.07(0.18)
factor(timeInt)4	-3.94(0.01)	-4.96(0.04)	2.84(0.18)	2.91(0.18)
factor(timeInt)5	-3.88(0.01)	-4.89(0.04)	3.47(0.18)	3.55(0.18)
factor(timeInt)6	-3.83(0.01)	-4.82(0.04)	4.03(0.18)	4.10(0.18)

factor(timeInt)7	-3.84(0.01)	-4.83(0.04)	4.41(0.18)	4.49(0.18)
factor(timeInt)8	-3.81(0.01)	-4.78(0.04)	4.73(0.18)	4.81(0.18)
factor(timeInt)9	-3.89(0.01)	-4.85(0.04)	4.84(0.18)	4.91(0.18)
factor(timeInt)10	-3.92(0.01)	-4.88(0.04)	4.90(0.18)	4.98(0.18)
factor(timeInt)11	-4.05(0.02)	-5.00(0.04)	4.79(0.18)	4.87(0.18)
factor(timeInt)12	-4.47(0.02)	-5.42(0.05)	4.34(0.18)	4.42(0.18)
factor(timeInt)13	-4.55(0.05)	-5.60(0.06)	4.02(0.18)	4.10(0.19)
factor(timeInt)14	-1.69(0.17)	-2.85(0.18)	7.06(0.26)	7.13(0.27)
grade B		0.66(0.01)	0.29(0.02)	0.30(0.02)
grade C		1.11(0.01)	0.45(0.02)	0.48(0.02)
grade D		1.38(0.02)	0.64(0.02)	0.68(0.02)
grade E		1.60(0.02)	0.74(0.02)	0.79(0.02)
grade F		1.74(0.03)	0.89(0.04)	0.95(0.04)
grade G		2.00(0.08)	1.35(0.10)	1.44(0.10)
purpose credit-card		0.08(0.04)	0.22(0.04)	0.22(0.04)
purpose debt-consolidation		0.16(0.04)	0.30(0.04)	0.29(0.04)
purpose educational		-5.08(32.06)	-5.88(225.48)	-5.99(225.48)
purpose home-improvement		0.03(0.04)	0.08(0.05)	0.08(0.05)
purpose house		0.15(0.06)	0.19(0.07)	0.18(0.07)
purpose major-purchase		0.06(0.05)	-0.06(0.05)	-0.06(0.05)
purpose medical		0.13(0.05)	0.08(0.05)	0.07(0.05)
purpose moving		0.15(0.05)	0.07(0.06)	0.05(0.06)
purpose other		0.01(0.04)	-0.01(0.04)	-0.02(0.04)

purpose renewable-energy	0.14(0.12)	0.11(0.14)	0.08(0.14)
purpose small-business	0.28(0.05)	0.23(0.05)	0.22(0.05)
purpose vacation	-0.03(0.06)	-0.16(0.06)	-0.17(0.06)
purpose wedding	-0.25(0.09)	0.28(0.10)	0.25(0.10)
verification-statusSource Verified	0.11(0.01)	0.05(0.01)	0.06(0.01)
verification-status:Verified	0.11(0.01)	0.05(0.01)	0.04(0.01)
initial-list-status w	0.03(0.01)	-0.06(0.01)	-0.04(0.01)
delinq-2yrs	0.03(0.01)	-0.003(0.01)	-0.02(0.01)
mths-since-last-delinq	-0.001(0.0002)	-0.0002(0.0002)	-0.001(0.0002)
mths-since-last-record	0.001(0.0001)	0.0000(0.0001)	0.001(0.0002)
inq-last-6mths	0.03(0.004)	0.05(0.005)	0.05(0.005)
annual-inc	-0.0000(0.0000)	-0.0000(0.0000)	-0.0000(0.0000)
mths-since-last-major-derog	0.001(0.0002)	-0.0004(0.0002)	-0.0002(0.0002)
acc-open-past-24mths	0.06(0.002)	0.03(0.002)	0.02(0.002)
mo-sin-old-il-acct	-0.0003(0.0001)	-0.0000(0.0001)	-0.0002(0.0001)
mo-sin-old-rev-tl-op	-0.001(0.0000)	0.0003(0.0000)	0.0001(0.0001)
mo-sin-rcnt-tl	-0.003(0.001)	-0.003(0.001)	-0.002(0.001)

mths-since-recent-bc	-0.002(0.0001)	-0.001(0.0002)	-0.001(0.0002)
mths-since-recent-inq	-0.003(0.001)	0.0000(0.001)	0.001(0.001)
num-tl-30dpd	-0.03(0.05)	-0.03(0.06)	0.03(0.11)
num-tl-90g-dpd-24m	-0.01(0.01)	0.004(0.01)	0.01(0.01)
num-tl-op-past-12m	0.004(0.003)	0.005(0.003)	0.01(0.003)
funded-amnt-inv		0.0005(0.0000)	0.0005(0.0000)
dti		0.01(0.001)	0.01(0.001)
revol-util		-0.001(0.0002)	-0.001(0.0003)
fico-range-low		-0.001(0.0002)	-0.001(0.0002)
total-pymnt		-0.001(0.0000)	-0.001(0.0000)
last-pymnt-amnt		-0.001(0.0000)	-0.001(0.0000)
last-fico-range-high		-0.01(0.0001)	-0.01(0.0001)
tot-cur-bal		-0.0000(0.0000)	0.0000(0.0000)
avg-cur-bal		-0.0000(0.0000)	-0.0000(0.0000)
bc-open-to-buy		0.0000(0.0000)	0.0000(0.0000)
tot-hi-cred-lim		0.0000(0.0000)	-0.0000(0.0000)
total-bc-limit		-0.0000(0.0000)	-0.0000(0.0000)
total-il-high-credit-limit		-0.0000(0.0000)	-0.0000(0.0000)
pub-rec			-0.04(0.01)
pub-rec-bankruptcies			-0.01(0.02)
revol-bal			0.0000(0.0000)
total-acc			0.01(0.001)

acc-now-delinq				-0.06(0.09)
bc-util				-0.0001(0.0003)
delinq-amnt				0.0000(0.0000)
mort-acc				-0.001(0.003)
num-accts-ever-120-pd				-0.01(0.004)
num-actv-bc-tl				0.01(0.004)
num-actv-rev-tl				-0.03(0.002)
num-il-tl				-0.01(0.001)
percent-bc-gt-75				0.0004(0.0002)
Log Likelihood	-414,442.70	-399,167.70	-231,143.60	-230,833.70
Akaike Inf. Crit.	828,913.40	798,437.50	462,415.30	461,821.40

References

Banasik, J., J. N. Crook, and L. C. Thomas (1999). Not if But When will Borrowers Default. *Journal of the Operational Research Society* 50, 1185–1190.

Byanjankar, A. (2017). Predicting Credit Risk in Peer-to-Peer Lending with Survival Analysis. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–8. IEEE.

Cao, R., J. M. Vilar, and A. Devia (2009). Modelling Consumer Credit Risk via Survival Analysis. *SORT: Statistics and Operations Research Transactions* 33, 3–30.

Carmichael, D. (2014). Modeling Default for Peer-to-Peer Loans. *Available at Social Science Research Network* 2529240.

Chen, D. and C. Han (2012). A Comparative Study of Online P2P Lending in the USA and China. *Journal of Internet Banking and Commerce* 17, 12–15.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 34, 187–202.

Cummins, M., T. Lynn, C. Mac an Bhaird, and P. Rosati (2019). "Addressing Information Asymmetries in Online Peer-to-Peer Lending." In Theo Lynn, John G. Mooney, Pierangelo Rosati, and Mark Cummins (Eds.) *Disrupting Finance*, Volume I, pp. 15–31. Cham, Switzerland: Palgrave Pivot.

Emekter, R., Y. Tu, B. Jirasakuldech, and M. Lu (2015). Evaluating Credit Risk and Loan Performance in Online Peer-to-Peer (P2P) Lending. *Applied Economics* 47, 54–70.

Han, A. and J. A. Hausman (1990). Flexible Parametric Estimation of Duration and Competing Risk Models. *Journal of Applied Econometrics* 5, 1–28.

Maudos, J. and J. F. De Guevara (2004). Factors Explaining the Interest Margin in the Banking Sectors of the European Union. *Journal of Banking & Finance* 28, 2259–2281.

Narain, B. (1992). "Survival Analysis and the Credit Granting Decision." In Lyn Thomas, David Edelman, and Jonathan Crook (Eds.) *Readings in Credit Scoring: Foundations, Developments, and Aims*, Volume I, pp. 235–250. New York: Oxford University Press.

Paarsch, H. J. and K. Golyaev (2016). *A Gentle Introduction to Effective Computing in Quantitative Research: What Every Research Assistant Should Know.* Cambridge, Massachusetts: The MIT Press.

Serrano-Cinca, C., B. Gutiérrez-Nieto, and L. López-Palacios (2015). Determinants of Default in P2P Lending. *PLoS ONE* 10, 1–22.

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, 155–195.

R Core Team (2013). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria.* URL https://www.R-project.org/.

Đurović, A. (2017). Estimating Probability of Default on Peer to Peer Market—Survival Analysis Approach. *Journal of Central Banking Theory and Practice* 6, 149–167.

Welchowski, T. and M. Schmid (2015). discsurv: Discrete Time Survival Analysis. *R package version* 1.

Willett, J. B. and J. D. Singer (1993). Investigating Onset, Cessation, Relapse, and Recovery: Why You Should, and How You Can, Use Discrete-time Survival Analysis to Examine Event Occurrence. *Journal of Consulting and Clinical Psychology* 61, 952.