DVARA RESEARCH

Characterising the Over-indebted: An Event History Analysis of Financial Diaries

*Sachit Rao**, *Navitha Parthasarathy*†

April 2022

Abstract

Low-income households face vagaries in income and expenses, often requiring the household (HH) to borrow, potentially causing the HH to become over-indebted. Financial diaries capture information on incomes, expenses, loans, and shocks with fine granularity and at frequent intervals of time. In this paper, such a dataset is analysed using the event history analysis approach in order to quantify the risk of HHs becoming over-indebted. Over-indebtedness is defined using two rubrics: monthly debt-to-income ratio and sacrifices made by the HH. The results of this analysis may be used by institutions to customize financial instruments based on the characteristics of a HH.

Dvara Research Working Paper
Series No. WP-2022-04
Version 1.0

*Corresponding author:Sachit Rao, International Institute of Information Technology Bangalore, 26/C Electronics City, Hosur Road, Bangalore, India, 560 100. E-mail:sachit@iiitb.ac.in

†E-mail: navitha.p@iiitb.ac.in

Contents

1 Introduction

The role of formal financial institutions in influencing the well-being of HHs has received sustained attention globally, particularly low-income HHs across the spectrum of economies and in urban and rural settings. An aspect of HH well-being that is in focus is its over-indebtedness and consequently, the drivers that lead to a HH entering this state has been a topic of research. In India and many other countries, microfinance institutions (MFIs) have been identified as being one such driver, (Schicks, 2010; Kamath, Dattasharma & Ramanathan, 2013; Schicks, 2013; Kamath & Ramanathan, 2015; Prathap & Khaitan, 2016). Several policies have been introduced in the Indian context to regulate lending practices of such institutions, (Prathap & Khaitan, 2016), so that, as described in (Schicks, 2013), for the MFI industry, customer protection has become a primary concern.

While this paper does not focus on the policy issues that surround MFIs and its gaps, it attempts, first, to identify over-indebtedness in HHs and second, to characterize such HHs according to their sociodemographic and economic characteristics. The data with which these issues are addressed is from monthly interviews of over 400 HHs for 12 months in Krishnagiri district of the south Indian state of Tamil Nadu; see (Prathap & Khaitan, 2016) for more details on how the HHs were selected. The dataset consists of responses reported by the respondents, on parameters such as, 1. incomes and their sources as well as expenses; 2.

descriptions of shocks and their diverse types; and 3. sources of loans and their contractual agreements; there include several more such parameters in the dataset.

Since the definitions of over-indebtedness in literature are also diverse, this paper uses two definitions of over-indebtedness: 1. from the consumer protection perspective: a HH is over-indebted if it makes sacrifices in order to meet loan repayments; and 2. from the quantitative/scientific perspective: a HH is over-indebted if it spends more than 30 per cent of its monthly income towards loan repayment.

The first definition is borrowed from (Schicks, 2010) and the second from (D'Alessio & Iezzi, 2013); see (D'Alessio & Iezzi, 2013) for other definitions of over-indebtedness in the European context. In (Collins, 2008), 20 per cent is set as the limit for monthly debt-income ratio. These definitions address the mechanisms by which over-indebtedness is identified and forms the first component of focus of this paper. To characterize HHs that become over-indebted - the second component the Event History Analysis (EHA), or Survival Analysis, methodology is adopted in this paper.

The support to use EHA in the financial domain comes from the review and discussions presented in (LeClere, 1999). Briefly, EHA methods use longitudinal data of entities, HHs in this case, collected over a duration of time, rather than a snapshot, in order to highlight the processes that lead to the occurrence of an event, for example bankruptcy in firms or overindebtedness as is the focus of this paper. The Financial Diaries data used in this paper is thus suitable to the application of this approach. These processes are functions of the covariates of the entity, some of which can vary in time. In the case of HHs, examples of covariates are the size of the HH, the number of income earners, loans and shocks, to name a few. Based on the statistical principles of EHA methods, it becomes possible to characterize HHs and their probability of surviving an event, by determining the hazard function.

Arguments that can be offered in favor of applying EHA methods are their ability to consider censored data, multiple states, and repeating events and where not all of which are "absorbing", that is they do not lead to death. See (LeClere, 1999) for types of censoring. In the Diaries dataset, the HHs: have time-varying covariates, such as varying incomes and expenses; display properties of right-censoring and partial left-censoring; and change states during the course of the study, such as becoming overindebted or otherwise, according to one of the definitions. Hence, EHA methods are appropriate to analyze the financial dynamics of HHs.

The particular EHA method that is used in this paper is based on the Cox Proportional Hazards (PH) model, which is a semi-parametric method that does not require the definition of an underlying distribution for HHs; this is in contrast with parametric methods that assume some type of a distribution, for example, exponential, Weibull, or log-normal. Several extensions and variants of the standard Cox PH model, such as the use of stratifications, in order that the proportionality assumption holds, have been proposed. These methods are used in this paper and the results are generated using the statistical software package, R.

With this introduction to the Diaries data, the methodology adopted, and the definitions of over-indebtedness, the contributions of this paper can be summarized as follows: To apply EHA, that considers time-varying attributes of HHs, rather than a snapshot, in order to determine the risk of occurrence of an event, in our case, over-indebtedness of an HH. To demonstrate the flexibility of the EHA technique as it allows for diverse definitions of an event to be applied; permits grouping (stratifying) of HHs that share some similar characteristics, for example, HH size; and handles recurrent events. To identify statistically significant attributes of a low-income HH that influence the risk of over-indebtedness; in turn, these attributes could be used to customize formal financial aid to HHs, for example, dispense low interest rate loans for small amounts, caused by small shocks, and with different terms for loans meant for planned large expenses to be incurred by an HH to meet its social obligations. From a novelty point of view, to the best of the authors' knowledge, EHA methods have not been used in analysing financial behaviour of low-income HHs in the Indian context.

The paper is organised as follows: in Sec. 2, a survey of the literature where EHA methods have been applied in the finance domain is presented; in Sec. 3, the definitions of over-indebtedness are formally defined based on the Financial Diaries; a brief overview of the Cox PH model and the possible covariates that can be selected from the dataset is also provided. Results of application of the Cox PH model are presented with discussions in Sec. 4. Concluding remarks, along with policy implications of the results of this work, are made in Sec. 5.

2 Literature Review

The dataset that is used in this paper is similar to data collected in the form of Financial Diaries - a format of data that has been extensively used globally, and within India, to analyze financial behavior of HHs, for instance, in (Kamath, Dattasharma & Ramanathan, 2013; Kamath & Ramanathan, 2015) in the context of urban poor in India and in (Collins, 2008; Schicks, 2013; Hannagan & Morduch, 2015; Biosca, McHugh, Ibrahim, Baker, Laxton & Donaldson, 2020), which are based on research conducted in South Africa, Ghana, USA, and Scotland, respectively. It should be highlighted that Financial Diaries is also employed in (Kamath, Dattasharma & Ramanathan, 2013), however, the Principal Component Analysis technique is adopted to distinguish the most significant types of expenses made by HHs that are indebted to MFIs as well as those that are not in debt. In addition, these expenses are calculated as a weekly average over the time-period of the study, in contrast with our approach, where, we do not perform any such averaging.

Viswanathan and Shanthi, 2017, also use data from an MFI and build a logistic regression model as well as a neural network model to classify customers as those that pose a credit risk and those that do not. They use about 10 variables, such as HH size, caste and religion, type of house etc., to build their models. However, the dataset does not appear to be longitudinal data, but a snapshot. A similar sort of dataset, the Vietnamese Living Standard Survey, 1998, is used in (Tra Pham & Lensink, 2008) to model the relation – as a probit model – between default and source of credit, for example, formal, informal, and semi-formal. In contrast to characterizing customers or HHs, in (Kumar & Sensarma, 2017), the efficiency of MFIs is modelled using various definitions by following the stochastic distance function approach and panel data from the MIX network dataset for Indian MFIs.

In some of the literature reviewed in (LeClere, 1999), the application of the EHA method to classify entities being studied in the finance domain can be found. For example, in one basis, banks as well as property and life insurers can be classified as being “healthy” or “surviving” or “dead”; an analogous basis is whether these financial entities are solvent or have become insolvent. By defining thresholds on the survival probabilities at the end of a pre-defined time-window, the bank or insurer, is classified as belonging to one of these classes. It is this feature of the EHA method that allows for HHs, which forms the entity of interest in this paper, with particular characteristics, to be classified as being most, or least, at risk of over-indebtedness.

In (Kim & Partington, 2008), the problem of predicting bankruptcy of firms listed in the Australian Securities Exchange is considered. The dataset used consists of data from both non-failed and failed firms between 1986-2008 and where, the number of non-failed firms is an order of magnitude higher than the number of failed firms. The Cox PH model is employed to identify key time-varying covariates that influence probability of bankruptcy. This methodology, and the nature of the dataset, are applicable to analyse the Diaries dataset. A similar problem is tackled in (Gepp & Kumar, 2008), in which it is concluded that the Cox PH model yields comparable results with the other methods when the cost of making Type 1 errors is increased. More importantly, they state “[the Cox PH model] provides more information about the business failure process through the interpretation of the hazard and survival function over time”. In (Gepp & Kumar, 2015), the same dataset is used to compare the Cox PH model with the Decision Trees algorithm. In (Laitinen, 2005), data on businesses in Finland obtained from a credit bureau is used to build a Cox PH model that can predict the time to “first payment default” by the business. By performing stratification of the dataset on size, type of industry, and age, they highlight the result that the Cox model has better forecasting ability in detecting financial distress.

3 Data, Definition, and Methodology

In this section, a brief introduction to the data collected in the Diaries is presented followed by the definitions of over-indebtedness that are used to denote events, as required by the Cox PH model technique.

3.1 Data – Overview of the Financial Diaries

The Financial Diaries data, (Prathap & Khaitan, 2016), is a collection of numerous files (comma separated variable format), each containing some financial attribute of an HH gathered every month and for a total of 12 months; socio-demographic information, such as the HH size, number of income earners, ages of the HH members, etc. are also available. Examples of financial attributes are the expenses incurred by the HH on shocks and their categories - weddings, festivals, etc.; income from various sources labor, business, pension etc.; regular expenses, such as on food items; and loan repayment amounts in that month. There are over 150 such attributes for each HH, which are denoted as covariates in order to apply the Cox PH model.

It is highlighted that excepting for the HH size, which stays constant during the study period, nearly all other covariates are time-varying; for example, for certain HHs, the number of income earners changes month to month and consequently, so do the incomes and expenses. Also, as the data is self-reported, there is a possibility of duplication of data, especially in the expense fields. To the extent possible, such duplicate information is removed. Data is available for (nearly) all HHs every month and thus, the number of missed observations is also small.

3.2 Definition – Events from definitions of over-indebtedness

The Cox PH model and its variants calculate the hazard of an event occurring for a candidate given the historical covariate information for that candidate. In this paper, the two definitions of over-indebtedness, as mentioned in the Introduction, are used to denote these events. For the HHs in the dataset, it is seen that events derived from either of these definitions occur multiple times for an HH in the 12-month period. Hence, the Anderson-Gill (AG) variant of the Cox PH model that handles multiple events per HH is used to determine the hazard functions; see (Therneau,

1996) for theoretical foundations of the A-G model and (Johnson, Boyce, Schwartz & Haroldson, 2004) for its application in wildlife management.

Definition 1: Over-indebtedness as sacrifice.

Viewing over-indebtedness through the consumer protection lens requires the definition of sacrifice. We consider an HH to have made a sacrifice if it spends less than some threshold on regular expenses. *We consider expenses on food and food-related items as being regular*. The threshold is calculated using monthly per capita expense figures for such expenses, as listed in the Level and Pattern of Consumer Expenditure Survey, 2011-2012[1], of the Government of India, for rural Tamil Nadu. The following eleven categories of expenses, with amounts in Indian Rupees in parentheses, are used: 1. Cereals (157.5), 2. Meat (110.5), 3. Beverages (185.3), 4. Vegetables (93.9), 5. Milk (89.2), 6. Spices (78.6), 7. Fruits and Dry fruits (55.6), 8. Pulses (44.8), 9. Oil (41.3), 10. Sugar (12.6), and 11. Salt (2.7), which brings the total expense to Rs. 872. The Survey also pegs monthly per capita expense on Tobacco and Intoxicants at Rs. 57.

Based on these figures, in this paper, the event of sacrifice is said to occur in an HH if the HH spends less than the total amount (Rs. 872) calculated using the per capita expense figures for all the 11 categories of expenses listed above. Thus, using this definition, an example of how the dataset is formatted for one HH to denote the events is shown in (the first 5-6 rows) Table 1. In this table, the column "Counter" denotes the number of fields in which the HH has spent less than the per capita amount. Thus, only if the value of Counter = 11, is the event said to occur. The "Month Start" and "Month End" fields denote the interval of risk of the event occurring for that HH, which is expected to be defined in order to use the A-G variant of the Cox PH model.

Definition 2: Over-indebtedness based on debt-income ratio

With this definition, the HH is said to experience this event if the monthly debt-income ratio is greater than 30 per cent. Based on this definition, for the same HH with 4 members, the events are listed in the last set of rows in Table 1. In these rows, the monthly income includes incomes from all sources, including the so-called irregular income.

From the available data, some of which is presented in Table 1, the points of departure of the two definitions of over-indebtedness can be recognized. The HH experiences more events when the debt-income ratio measure is used when compared with the use of sacrifice. Except for 2 months, 1 and 6, the events using these definitions do not overlap.

Table 1. Occurrence of events based on the two chosen definitions for an HH with 4 members

Definition 1: Over-indebtedness as sacrifice						
Month	Total Expenses	Per capita expenses	Counter	Month Start	Month End	Event
1	210	3488	**11**	0	1	**1**
2	2200	3488	8	1	2	0
…	…	…	…	…	…	…
6	380	3488	**11**	5	6	**1**
…	…	…	…	…	…	…
12	275	3488	9	11	12	0
Definition 2: Over-indebtedness based in debt-income ratio						
Month	Monthly Income	Debt Repayment	Ratio (%)	Month Start	Month End	Event
1	4000	8350	**208.75**	0	1	**1**
2	4000	2450	**61.25**	1	2	**1**
…	…	…	…	…	…	…
5	7700	1450	18.83	4	5	0
…	…	…	…	…	…	…
12	10400	0	0	11	12	0

It is to highlight the influence of these definitions of over-indebtedness on the classification of HHs, that forms one of the main contributions of this paper. This is achieved using the Cox PH model, which is briefly described next.

3.3 Methodology – Application of the Cox PH Model

Given the Diaries data, which is essentially a time-series data for a HH measured at discrete time intervals, EHA allows for time-to-event phenomena to be evaluated as a probability, (Klein & Moeschberger, 2003), of an entity experiencing the event after some time t. The main variables of interest in such studies are the *survival* and *hazard* functions. The former denotes the probability of an individual *surviving* beyond a time t, which in the case of discrete measurements $t_1 < t_2 < \cdots$, is given by $S(t) = \sum_{t_j>t} p(t_j)$ where, $p(t_j)$ is given by the probability mass function of the discrete time variable. The hazard function is given by

$$h(t_j) = \Pr(t = t_j | t \geq t_j) = \frac{p(t_j)}{S(t_{j-1})}, S(t_0) = 1. \tag{1}$$

The discrete survival and hazard functions are related according to

$$S(t_j) = \prod_{t \leq t_j} (1 - h(t_j)). \tag{2}$$

In the Cox PH model, (LeClere, 1999), the hazard function is chosen as

$$h(t_j) = h_0(t_j) exp^{(\beta_1 X_1 + \beta_2 X_2 + \cdots)}, \tag{3}$$

where, $\beta_1, \beta_2, \cdots$ are the regression coefficients; $X_1, X_2, \cdots$ are the covariates; and $h_0(t_j)$ is the baseline hazard function, which is unspecified - hence, the CoxPH model has a semi-parametric form. The term proportional hazards stems from the expression of the ratio of two entities, say i and k, which is of the form

$$\frac{h_i(t_j)}{h_k(t_j)} = exp^{(\beta_1 (X_{1i} - X_{1k}) + \beta_2 (X_{2i} - X_{2k}) + \cdots)}. \tag{4}$$

As can be seen, if the proportional hazards assumption holds, then the hazard ratios are time-invariant and proportional to the values of the respective covariates. The A-G variant of the Cox PH model can handle time-varying covariates in the evaluation of hazard ratios, (Therneau, 1996). Thus, depending on the magnitudes of the coefficients $\beta_1, \beta_2, \cdots$ determined using the time-series data, the relative hazards between the entities i and k can be determined. From [2]: "A positive sign means that the hazard $\cdots$ is higher $\cdots$ for subjects with higher values of that variable." For example, when the variable takes values as positive integers. Thus, from (2), a higher hazard implies a lower probability of survival or experiencing the event.

These results act as the foundations to characterize HHs, based on their covariates, which are at risk of experiencing the events determined using the two definitions of over-indebtedness.

4 Empirical Results

First, the Diaries dataset which consists of data from 411 HHs is split in the ratio 70-30, so that data of 287 HHs form the training set and the remaining 124 form the testing or hold-out set. Within the training set, there are 262 events according to Definition 1 and 995 events according to Definition 2. The inclusion of expenses on Tobacco and Intoxicants in Definition 1 did not significantly increase the number of events. The description of some of the covariates - which are found to be statistically significant - and the values attributed to them, are provided in Table 2. Note that some of the values of these covariates change from month to month.

To highlight the differences between Definitions 1 and 2, the following Cox PH models are built: 1. Stratifying on HH size for both Definitions 1 and 2; and 2. Without any stratification for both definitions. The statistically significant covariates for both models are summarised in Table 3; the results of the two models are discussed in Sections 4.1 and 4.2, respectively.

4.1 Stratification on HH size

Stratification on one of the covariates helps in preserving the proportionality assumption while applying the Cox PH model. Although time-varying covariates can also be selected as stratifying variables, they need to be selected carefully with some modifications. To avoid this issue, the time-invariant covariate of HH size is selected as the stratifying variable. These results can also be used to characterize better the HHs that experience the events as defined.

Table 2. Description of key covariates

Covariate Description (variable name used)	**Value**
Number of income earners, including primary and secondary (*IncEarners*)	0,1,2,…
Number of loans taken from Institutions (*LoanInst*)	0,1,2,…
Flag on irregular income, separate from primary and secondary income (*FlagIrregInc*)	0 (No) and 1 (Yes)
Flag on shock expense greater than 1000 Rs (1000 is the median shock expense) (*FlagShock1000*)	0 (No) and 1 (Yes)
Flag if a loan is taken (*FlagLoanTaken*)	0 (No) and 1 (Yes)
Flag if a HH made a deposit (*FlagSavingsDep*)	0 (No) and 1 (Yes)
Flag if the irregular income of the HH is greater than the regular income (*FlagIrregRegIncDiff*)	0 (No) and 1 (Yes)
Number of HH members under the age of 18 (*AgeUnder18*)	0,1,2,…
HHSize	1 (3 members or fewer) 2 (4 members) 3 (5 members) 4 (more than 5 members)

As can be seen from Table 3, only the covariate defined by the number of income earners (*IncEarners*) is statistically significant, has the same sign, and is common to the events denoted by Definitions 1 and 2. This result indicates the intuitive understanding that an increase in the number of income earners - which leads to increased income in the HH reduces the hazard of either type of event occurring in the HH. The same intuitive explanation can be used to understand the sign of the coefficient for the covariate defining the number of members under the age of 18 (*AgeUnder18*): more the number of dependents, who would typically be under the age of 18, greater is the hazard of over-indebtedness.

Table 3. Regression coefficients of the models with and without stratification on HH size

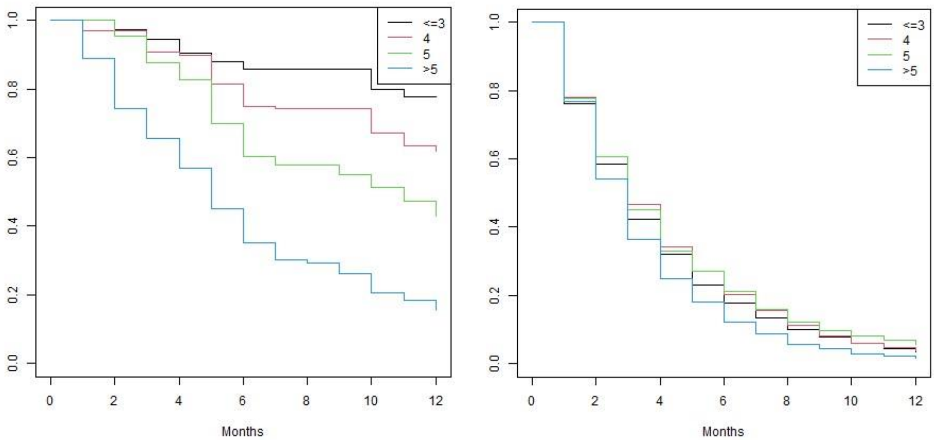
	Definition 1		Definition 2	
Covariate	Coefficient	p-value	Coefficient	p-value
With Stratification				
FlagShock1000	-0.49	0.000581	Violates proportionality	
AgeUnder18	0.12	0.052667	Insignificant	
FlagIrregInc	0.43	0.001581	-0.38	4.56e-09
IncEarners	-0.25	0.000225	-0.18	1.64e-06
LoanInst	Insignificant		0.35	6.25e-15
Without Stratification				
FlagShock1000	-0.51	0.000372	Insignificant	
AgeUnder18	0.12	0.054312	Insignificant	
FlagIrregInc	0.43	0.001646	Insignificant	
HHSize	0.67	<2e-16	Insignificant	
IncEarners	-0.24	0.000439	-0.19	1.85e-07
FlagLoanTaken	Insignificant		0.74	<2e-16
FlagSavingsDep	Insignificant		0.3	3.1e-4
FlagIrregRegIncDiff	Insignificant		-0.53	1.63e-10

Although the covariate for the flag on irregular income (*FlagIrregInc*) is also common, it has a different sign, implying an opposite effect for the two definitions considered. A non-zero irregular income in a month *increases* the hazard of the occurrence of the event given by Definition 1, while it *reduces* the hazard for the other case. The latter result has a simple explanation: the presence of an irregular income increases the overall income of the HH, thus reducing the debt-income ratio in a month. On the other hand, the positive sign of this coefficient for the event given by Definition 1 deserves closer scrutiny. An analysis of a few HHs that experienced this event and also had irregular income revealed that the HH also had other expenses in the same month. For example, one HH spent a considerable sum on house repair while another bore expenses related to injuries. Neither HH actually spent the additional income in their regular expenses, thus leading to the sacrifice in regular expenses. This spending on expenses apart from the regular ones in a month also may have a cascading effect in the later months, leading to further events of sacrifice.

The covariate denoting the occurrence of expenses on shocks greater than 1000 Rs (*FlagShock1000*) also has a negative sign, implying a greater hazard of sacrifice if the shock expense is *less* than 1000 Rs - a seemingly counter-intuitive result. This can be *perhaps* be explained as follows: a HH which expects a larger shock expense, in the form of events or festivals, takes a loan in order to meet this burden without sacrificing regular expenses, whereas a smaller shock expense, which may be unforeseeable, actually leads to the HH making a sacrifice. This counter-intuitive result is supported by observing the correlation between the times of shocks and loans for three HHs, which indicates that shocks and loans are closely correlated in time (the correlation graphs are omitted in the interest of conserving space).

For the event given by Definition 2, while the covariate *FlagShock1000* is significant, it violates the proportionality assumption and is hence not considered. The Schoenfeld residuals are used as a measure of violation of proportional hazards: a significant p-value indicates violation of proportionality. The inclusion of this variable requires special treatment, for example by including its interaction with other covariates or as a non-linear function. Finally, the covariate denoting the number of loans from institutions (*LoanInst*) is positive for the event given by Definition 2, while it is insignificant for Definition 1. This result should also be expected as the number of loans affects the debt-income ratio of HHs, while HHs can experience sacrifice even without having to make loan repayments.

The survival curves of the different strata experiencing the two types of events are shown in Fig. 1; these are derived using the model parameters listed in Table 3. The difference between the survival curves are marked. For nearly all HHs, independent of size, the survival probability curves look the same and importantly, the survival probabilities of experiencing the event according to Definition 2 drop to nearly 0 by the end of the 12-month period. On the other hand, for smaller HHs - three and fewer - the survival probabilities of experiencing the event according to Definition 1 are very high, implying that very few of them actually have to sacrifice their regular expenses. These probabilities also reduce as the HH size increases, with very large HHs being most at risk. The number of HHs within each of these categories of HH size and the number of events given by both definitions, found from the dataset, also support the validity of these models. The number of events given by Definition 1 increases as the HH size increases, while the number of events given by Definition 2 is nearly the same for all HH sizes. An inspection of the dataset reveals that the number of HHs in each HH size category is nearly the same.



(a) Definition 1 **(b) Definition 2**

Figure 1. Survival curves of HHs based on stratifying by HH size

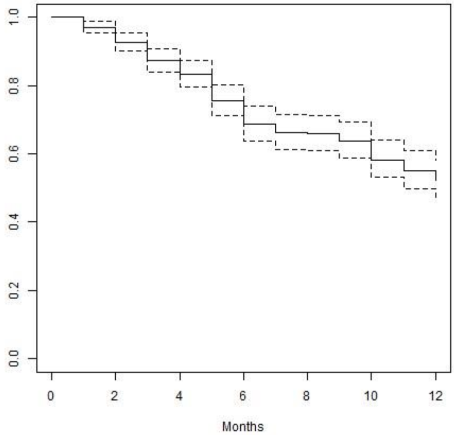
4.2 Without stratification

Since the differences in the survival curves for both events with stratification are stark, models without stratification are also constructed to examine if the two definitions have points of convergence. First, models with the same set of covariates are derived. For this case, models with the following covariates: *FlagIrregInc*, *IncEarners*, and *FlagLoanTaken* (denotes if a loan was taken in a month), are not consistent in the significance levels of the coefficients for both definitions. While these models are not reported here in the interest of conserving space, these results indicate the influence of the definitions of the events in characterising HHs.

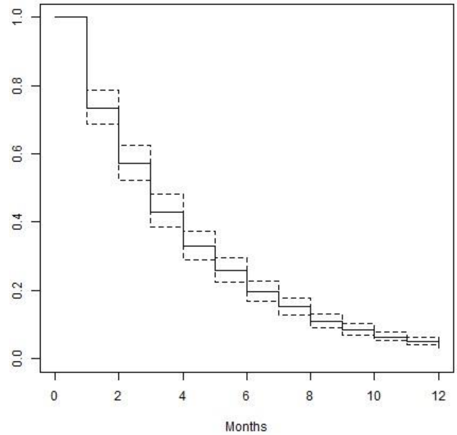
Consequently, different models are constructed for the two event types and their parameters are presented in Table 3. As can be seen, for the event given by Definition 1, the covariates common to with and without stratification, are nearly the same. Thus, their implications remain the same. The covariate of HH size in the model without stratification has a positive coefficient, implying that an increase in HH size increases the risk of sacrifice. This result matches the one obtained with stratification. Hence, from the perspective of Definition 1, stratification, or its absence, yields the same results.

For the case of event given by Definition 2, the model contains two unique covariates: *FlagSavingsDep* and *FlagIrregRegIncDiff*. While the coefficients of the covariates, *IncEarners* and *FlagLoanTaken* have clear explanations, and are the same as with stratification, the two new covariates introduced in this model deserve a closer look. Also note that the HH size is not a significant covariate, as demonstrated by the survival curves in Fig. 1b. The negative coefficient for the covariate *FlagIrregRegIncDiff* implies that income earners in a HH may have to opt for additional, but "irregular", sources in order to minimise the risk of crossing the 30 per cent threshold of debt-income ratio. This result highlights the actions taken by a HH in order to meet its loan repayment obligations. The coefficient for the covariate *FlagSavingsDep* has a positive sign, indicating that a HH that makes a deposit is at higher risk of crossing the 30 per cent threshold. While this result does not have a clear explanation, it should be borne in mind that these numbers are reported by the HHs themselves. Thus, even though a HH may have a higher income in a month, it may choose to set aside some portion as a deposit and report the balance as income that can be used to service debt. No doubt, this issue can be resolved by a careful scrutiny of the reporting practices.

To conclude this section, survival curves obtained with the two models are presented in Fig. 2. As can be observed when Fig. 2a is compared with Fig. 1a, stratification is able to characterise HHs differently, and arguably better, for the event of sacrifice. The event of higher debt-income ratio with and without stratification do not appear distinctly different, as in both cases, the survival probability drops to close to 0.



(a) Definition 1



(b) Definition 2

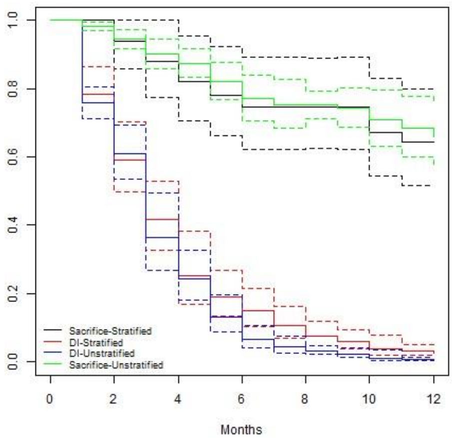
Figure 2. Survival curves of HHs without stratification

4.3 Performance

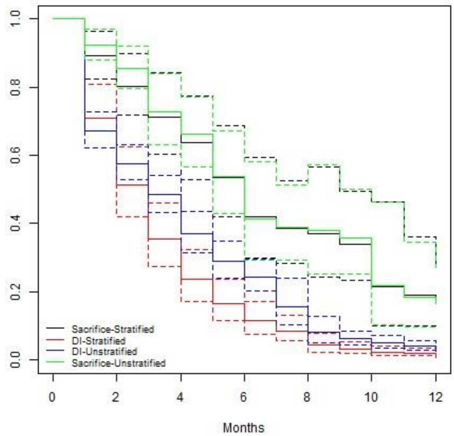
The various models developed are tested using data from HHs in the holdout set. To examine if the models are able to predict risk correctly, one HH from each of the 4 size categories is selected such that, in the 12-month time window, 1. it that has experienced *both* types of events *at least once*; and 2. it did *not* experience *one* of these events. It is highlighted that for some of the HHs, both events occurred multiple times.

The survival curves for a few of these HHs are shown in Fig. 3. The performance of the models derived for the event given by Definition 1 are given by green and black lines, while that according to Definition 2 are given by the blue and red lines, in each of the sub-figures in Fig. 3. It can be observed that for each event type, the survival curves with and without stratification are nearly the same.

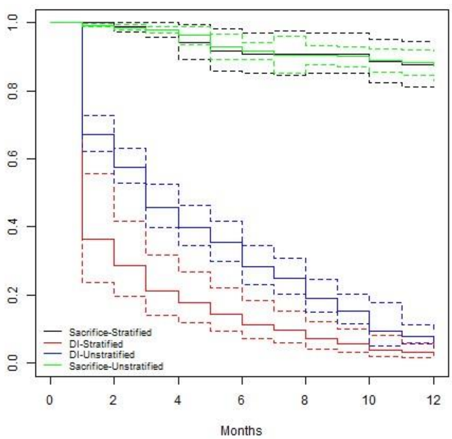
In Fig. 3a, it can be seen that for the smaller sized HH, 3 members, the periods when the HHs do not experience the event given by Definition 1 are captured by the models. The survival probability drops only when the event occurs, but continues to remain over 60 per cent. The model is also able to predict accurately the reduction in survival probability when



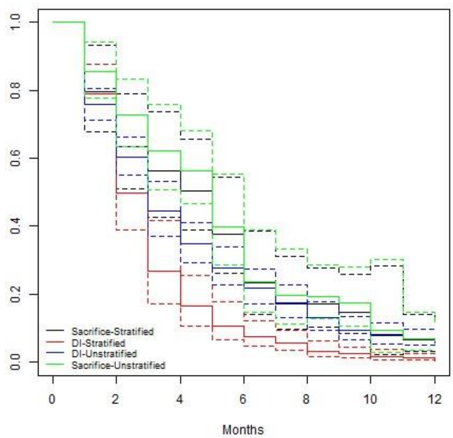
(a) HH Size = 3 (Both events)



(b) HH Size = 8 (Both events)



(c) HH Size = 3 (Only one event)



(d) HH Size = 6 (Only one event)

Figure 3. Survival curves of HHs from the hold-out set

the HH size increases to 8; compare the green and black lines in Figs. 3a and 3b. For the event according to Definition 2, the survival probabilities reduce in each month for both HHs and drops to nearly 0 by the end of the 12-month period - a result that matches the “ground truth” data for these HHs. However, from the results for other HHs (not presented here), the models derived for Definition 1 display a greater reduction in survival probability in the months when these events actually occur, in contrast with the output of models for Definition 2, which show a steady decline in survival probabilities.

Results for two HHs that did *not* experience one of these events are shown in Figs. 3c and 3d. For HHs that did not experience the event given by Definition 2 at all, the survival curves indicate that the survival probabilities ultimately reach close to zero. This result clearly indicates the need to use a different set of features to construct this model, for instance, those that consider interaction between covariates or non-linear functions of the covariates. On the other hand, the survival curves for the event given by Definition 1 are distinct from one another. The survival probability of the HH with 3 members, Fig. 3c, does not reduce to below 80%, which may be considered as the HH never experiencing this event - a result that is supported by the ground truth for this HH. If a threshold of survival probability is set to 50%, then, from the survival curves, it can be said that these HHs never experience this event. Note that the HH with 4 members experiences this event only once, while the other two do not. The model for the HH with size 6, Fig. 3d, is able to predict the drop in survival probability to nearly zero and this result too matches the data for this HH.

To summarise these results, the models for Definition 1 are able to predict accurately the survival probability that also closely matches ground truth, irrespective of the occurrence of the event. They are also able to show the increased survival probabilities with reduction in HH size as well as reductions in these values in the months when they actually have happened. The survival curves for the event given by Definition 2 look similar irrespective of the occurrence of the event. For this case, alternative features will be required to be selected.

5 Conclusions

This paper looked at two disparate definitions of over-indebtedness in low-income HHs and presented time-series-based statistical models in order to identify points of convergence between them. The EHA-based method was adopted to build models so that the probability of risk of a HH becoming over-indebted over a time window can be calculated; these features are in contrast with the results presented in (Viswanathan and

Shanthi, 2017), where, snapshot data is used or in (Kamath, Dattasharma & Ramanathan, 2013) where, a weekly average of some key variables is used to build models. Though the models in this paper were developed mainly on categorical data rather than say, actual income and expense numbers, they considered the time-varying nature of these categorical variables, rather than consider a HHs characteristics measured at a single instant of time; these models also permitted for HHs belonging to different categories to be analyzed distinctly.

These models, which can no doubt be revised, revealed the stark differences between the two definitions of over-indebtedness; the use of such definitions is again novel in the analysis of financial behaviour of low-income HHs. They were also able to characterize those HHs, for instance, ones with fewer members, that may never experience one of these events; or on the other hand, experience over-indebtedness even for small shock expenses. Future work can focus on making the models more detailed, for example, those that consider competing risks, in order to distinguish better, HHs that need financial support from those that need not. Another path is the design of rules of classification based on survival curves.

The results presented in the paper can be used by MFIs to customize financial products based on a historical record of financial behavior by a HH as well as its attributes, for instance, changing number of income earners or variation in regular and irregular income. A financial institution can now view a HH from a consumer protection perspective that can be quantified, as indicated by sacrifices in regular expenses, and develop repayment schedules and terms so that certain HHs - large HHs, for instance - can avoid making such sacrifices. It can also increase the threshold on debt burden of a HH in order that it can meet smaller unplanned expenses, while still being able to borrow larger amounts to meet social obligations, which may be planned well in advance by a HH. The quantified consumer protection metric can complement conventional debt-to-income ratio metrics that may be used regularly to decide on dispensing loans, especially, to low-income HHs.

Notes

1. Level and Pattern of Consumer Expenditure 2011-2012, NSS 68th Round, February 2014, National Sample Survey Office, Government of India
2. http://www.sthda.com/english/wiki/cox-proportional-hazards-model (Accessed on November 17, 2021)

Funding

This work was supported by Dvara Research.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Biosca, O., McHugh, N., Ibrahim, F., Baker, R., Laxton, T. & Donaldson, C. (2020). Walking a tightrope: Using financial diaries to investigate day-to-day financial decisions and the social safety net of the financially excluded. *The ANNALS of the American Academy of Political and Social Science*, *689(1)*, 46–64.

Collins, D. (2008). Debt and household finance: evidence from the Financial Diaries. *Development Southern Africa*, *25(4)*, 469–479.

D'Alessio, G. & Iezzi, S. (2013). Household Over-Indebtedness: Definition and Measurement with Italian Data. In *Bank of Italy Occasional Paper No. 149*.

Gepp, A. & Kumar, K. (2008). The role of survival analysis in financial distress prediction. *International Research Journal of Finance and Economics*, *(16)*, 13–34.

Gepp, A. & Kumar, K. (2015). Predicting financial distress: A comparison of survival analysis and decision tree techniques. *Procedia Computer Science*, *54*, 396–404.

Hannagan, A. & Morduch, J. (2015). Income Gains and Month-to-Month Income Volatility: Household Evidence from the US Financial Diaries. NYU Wagner Research Paper No. 2659883, US Financial Diaries Working Paper.

Johnson, C., Boyce, M., Schwartz, C. & Haroldson, M. (2004). Modeling Survival: Application of the Andersen-Gill Model to Yellowstone Grizzly Bears. *The Journal of Wildlife Management*, *68(4)*, 966–978.

Kamath, R., Dattasharma, A. & Ramanathan, S. (2013). MFI (Microfinance Institutes) Borrowers: Their loans and Repayments - A study in Ramanagaram, India. Working paper no: 397, Indian Institute of Management Bangalore.

Kamath, R. & Ramanathan, S. (2015). Informal businesses and microcredit – Evidence from financial diaries: A study in Ramanagaram, India. *IIMB Management Review*, *27(3)*, 149 – 158.

Kim, M. H. & Partington, G. (2008). The Dynamic Prediction of Company Failure. In *21st Australasian Finance and Banking Conference* (pp. 41–77).

Klein, J. P. & Moeschberger, M. L. (2003). *Survival Analysis*. SpringerVerlag New York.

Kumar, N. & Sensarma, R. (2017). Efficiency of Microfinance Institutions in India: A Stochastic Distance Function Approach. *Journal of Emerging Market Finance*, *16(2)*, 151–168.

Laitinen, E. (2005). Survival analysis and financial distress prediction: Finnish evidence. *Review of Accounting and Finance*, *4*, 76–90.

LeClere, M. J. (1999). The Occurrence and Timing of Events: The Application of Event History Models in Accounting and Finance Research. https://ssrn.com/abstract=181553.

Prathap, V. & Khaitan, R. (2016). When Is Microcredit Unsuitable? Guidelines using primary evidence from low-income households in India. In *Dvara Research Working Paper Series No. 2016-02*.

Schicks, J. (2010). Microfinance Over-Indebtedness: Understanding its drivers and challenging the common myths. Working Papers CEB 10048, ULB – Universite Libre de Bruxelles.

Schicks, J. (2013). The Sacrifices of Micro-Borrowers in Ghana – A Customer-Protection Perspective on Measuring Over-Indebtedness. *The Journal of Development Studies*, *49(9)*, 1238–1255.

Therneau, T. M. (1996). Extending the Cox Model. Technical report no 58, Mayo Clinic, Rochester, Minnesota.

Tra Pham, T. T. & Lensink, R. (2008). Household Borrowing in Vietnam: A Comparative Study of Default Risks of Formal, Informal and Semi–formal Credit. *Journal of Emerging Market Finance*, *7(3)*, 237–261.

Viswanathan P. K. & Shanthi S. K. (2017). Modelling Credit Default in Microfinance—An Indian Case Study. *Journal of Emerging Market Finance*, *16(3)*, 246–258.