HÖGSKOLAN
DALARNA

MASTER THESIS IN MICRODATA ANALYSIS

Modelling S shaped hazard function
A case of evaluating Volvo Cars training project

Author:
Hana Hozzánková

Supervisor:
Md. Moudud Alam

2014

Business Intelligence Program
School for Technology and Business Studies
Dalarna University

Abstract

The aim of this study is to introduce a new S-shaped hazard function, derive its corresponding likelihood function and then apply it to a real dataset to show advantages of such method. We defined an inovative S-shaped hazard function based on *arctangent* with three additional parameters. Our hazard function opens new possibilities for the measurement of the lock-in effect directly using one of its parameters. Its significance is further tested by a standard Wald test. We derive a log-likelihood function and apply it on a data from Volvo Cars Project. According to our results we conclude that the program was unsuccessful. Surprising shapes of final hazard functions lead us to closer examination of possible limitations. We conclude that our results could be affected by several aspects such as data quality, computational methods and possible violation of independency assumption. Nevertheless, this study works as a useful summary of the newly-defined S-shaped hazard function and highlights its innovative possibility to estimate lock-in effect together with its significance testing.

Keywords: Shape restricted hazard function, Duration data, Length of unemployment, Propensity score matching

Contents

List of Figures

Introduction

Training programs have been held as a tool for integrating the unemployed into the work force through development of their certain skills. After program is finished, the evaluation is in common interest. Different parties that took part of training programs and provided resources want to know if the program was successful and fullfiled the objectives or not. With increasing number of government training programs during last 30 years (Heckman et al., 1999) there have been developed a few approaches for evaluations.

Evaluations of programs differ in two aspects. First, what they evaluate and second, what kind of approach they are using. What can possibly be evaluated in programs?

Firstly, evaluation for training program can measure increasment in salary after one participated in program and then consequently, we are able to count rate of return (how long it will take to return costs of training by increasment in salary). And secondly, we can measure increasment in probability to get hired after participation in training program. In this study, we use the second approach.

Second aspect, in which evaluations differ, is the kind of approach. There are two kinds; descriptive and parametric, by assuming specific hazard function (or density) for duration of unemployment. All kinds often have to face problem of unobservable part of data. In that after one participated in training program, we observed how long it took him/ her to find a job, we are not able anymore to observe his/ her duration of unemployment without participating in training (unobservable data). We estimate these unobservable data in terms to only compare what is comparable. Descriptive methods differ in way how they deal with (estimate) unobservable data. We shortly introduce three descriptive methods.

"The before-after estimation" is the most commonly used evaluation strategy compares a person with himself/ herself. The unobservable data (one's outcome if he/ she would not participate in program) are estimated by taking old data about the same person from history (before he/ she particpated in program). This comparison strategy is based on longitudinal data. (Heckman et al., 1999)

In "The difference-in-differences estimator", we do not compare necessary person with himself/ herself. The estimator is given by difference between (average) gain to

individuals who participated in program and (average) gain to individuals who did not participate in program (under assumption that on average if persons who participate in the program would not participate they would have same gain as those persons who actually do not participate). In terms to find the best match for participating individual from non-participating group, propensity scores can be used, (see e.g. Alam et al. (2012) and Harkman et al. (1996)). Propensity scores match similar individuals from two groups according to their characteristics (gender, region, education,...). By applying propensity scores, we get two balanced groups and by using those groups we are able to evaluate 'pure' effect of training program (we get valid results). Propensity scores reduce individual deviations in characteristics between participants and non-participants.

"The cross-section estimator" compares mean outcomes of participants and non-participants at time t (under the assumption that on average persons who do not participate in the program have the same no- treatment outcome as those who do participate).

The second stream of methods that assumes specific hazard function (or density) for duration of unemployment, is the one we are interested in our study. The shape of hazard function for given data is unknown (shapes of hazard functions depend on the distributional assumptions; common ones are Weibull, Exponential or Log-Logistic; see e.g. Thierry and Sollogoub (1995) and de Koning et al. (1991)). However, the choice is somewhat arbitrary and therefore any closer specification or approach becomes attractive. This is precisely the motivation for our study.

We introduce a brand new shape of parametrizable hazard function- the S-shape (see Figure 3). Our S-shaped hazard function is designed in such way that parameters included in its specification explain some of its features. By varying these paremeters we are able to locate the lock-in effect, compare the exit rates of two different groups (when one is lower, higher and when they are equal), find out about different accelerations for exit rates and most importantly, we can evaluate a given program correctly. These measures cannot be accomplished by using descriptive approach only. Parameters make this hazard function flexible. By maximizing its log-likelihood function, we get optimal values of parameters. Then we carry out statistical tests of the hazard function's

distinguished features of the training and non-training groups via standard Wald test.

In the second part, we apply our S-shaped hazard function on the Volvo data. We compare different results for two groups based on optimized parameters.

The aim of this study is to introduce new S-shaped hazard function, derive appropriate likelihood function, apply this function to real dataset and show the advantages this new type of modelling brings to the field.

1 Statistical Methods

In our study, we have data about people who participated in training program and about people who did not participate in program. As far as we are able to get length of unemployment from both groups (tranining and non-tranining) that follows directly after training, we consider these data as duration data. In these duration data, the duration stands for length of staying unemployed or we can say ability to change one's status from unemployed to employed one. We are interested only in first spell. In other words, only first length of unemployment after training is our interest.

We set T for duration of stay in the state (unemployment). The population is assumed to be homogeneous with respect to the systematic factors, regressor variables, that affect the distribution of T. This means that everyone's duration of stay will be a realisation of a random variable from the same probability distribution.

However, we do not know exactly how true distribution of T looks like. We observed part of population but part of that part are censored data [1] which complicate the situation. Because of censored data it is difficult to get 'pure' probability distribution of T.

We can express the conditional probability that a person leaves the state within an interval dt at or after time t as follows:

$$P(t \leq T < t + dt | T \geq t) \tag{1}$$

where dt is very short period of time. Equation 1 stands for probability that change of state (from unemployed to employed) will happen between time t and $t + dt$, given that state was not changed until time t.

If we divide this probability by dt we get average probability of leaving per unit time period over a short time interval after t, and by considering this average over shorter and shorter intervals we define:

$$\Theta(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt} \tag{2}$$

as the hazard function. It is the instantaneous rate of leaving per unit time period at t (Lancaster, 1990). By rewriting Equation 2 and applying conditional probability

[1] censored data are data expressing that one was still unemployed (state was not changed) at the time when experiment was finished

relationship we get:

$$\begin{aligned}\lim_{dt\to 0}\Theta(t)dt = \lim_{dt\to 0}P(t\le T<t+dt|T\ge t) = \lim_{dt\to 0}\frac{P(t\le T<t+dt, T\ge t)}{P(T\ge t)} = \\ = \lim_{dt\to 0}\frac{P(t\le T<t+dt)}{P(T\ge t)} = \lim_{dt\to 0}\frac{F(t+dt)-F(t)}{1-F(t)}\end{aligned} \tag{3}$$

Dividing by dt and definition of derivative we get:

$$\Theta(t) = \lim_{dt\to 0}\frac{F(t+dt)-F(t)}{dt}\frac{1}{1-F(t)} = F'(t)\frac{1}{1-F(t)} = \frac{f(t)}{\bar{F}(t)} \tag{4}$$

Where $\bar{F}(t)$ is so-called survival function.

$$-\frac{d(ln\bar{F}(t))}{dt} = +\frac{1}{\bar{F}(t)}f(t) = \frac{f(t)}{\bar{F}(t)} = \Theta(t) \tag{5}$$

By using Equation 5:

$$\begin{aligned}\Theta(t) &= -\frac{d(ln\bar{F}(t))}{dt} \\ -\int_0^t \Theta(s)ds &= ln\bar{F}(t) \\ \bar{F}(t) &= e^{-\int_0^t \Theta(s)ds}\end{aligned} \tag{6}$$

Equation 5 shows that the survival function can be derived from hazard function. Next, by putting Equation 5 and Equation 6 together we can also express density function by hazard function as follows:

$$f(t) = \Theta(t)e^{-\int_0^t \Theta(s)ds} \tag{7}$$

We face uncensored and censored data which means our overall likelihood will be comprised of two parts (Lancaster, 1990). Uncensored data mean we know exact real length of state (unemployed) of individual. The individual found job and it took time t. All uncensored data are described by the density function of t. We also have censored data. From censored data it is not possible to get true infromation about length of duration of state. It gives us only hint that we can be sure that the duration is at least of length t. Therefore this probability comes from complementory cumulative distribution function of t, what equals to survival function. Finally, after summarizing

all information we get likelihood function:

$$L(\theta) = \prod_{i=1}^{n} \left[p_\theta(t_i)^{\delta_i} P_\theta(T > t_i)^{1-\delta_i} \right] = \prod_{i=1}^{n} \left[f_\theta(t_i)^{\delta_i} \bar{F}_\theta(t_i)^{1-\delta_i} \right] =$$
$$= \prod_{i=1}^{n} \left[\left(\Theta(t_i) e^{-\int_0^{t_i} \Theta(s)ds} \right)^{\delta_i} \left(e^{-\int_0^{t_i} \Theta(s)ds} \right)^{1-\delta_i} \right] \tag{8}$$

where δ_i is indicator that indicates if data is censored ($= 0$) or uncensored ($= 1$) and θ is vector of parameters.

From Equation 4, we see that by assuming certain distribution of duration time t we get hazard function. It also works backwards. We can define hazard function and then derive density and survival function (see Equation 6 and Equation 7). In Figure 1, you can see few hazard functions derived from different distributions. In our study, the objective is to define our hazard function with S-shape (see Figure 3) and then get likelihood function.

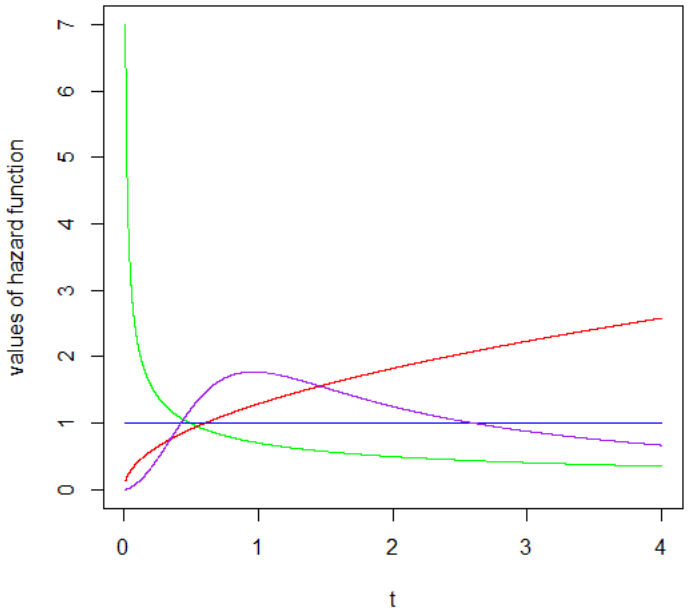

Figure 1: Illustration of four hazard functions given by different distributions: Weibull with $\alpha = 0.5$, $\gamma = 1.4$ (green), Exponential with $\gamma = 1$ (blue), Log- Logistic with $\alpha = 2.7$, $\gamma = 1.9$ (purple) and Weibull with $\alpha = 1.5$, $\gamma = 0.86$ (red)

As far as shapes of hazard function depend on parameters (for example, see red and green curve in Figure 1 that both are derived from Weibull distribution but they differ in parameters) and real shape of hazard function is unknown, it is in our interest to get the optimal shape of assumed hazard function. By optimal we mean the one that is closest to the real one. We identify optimal parameters by Maximum Likelihood Estimation (MLE). Usage of log-likelihood function instead of likelihood function brings simplification into computation and has no impact on results (as far as logarithm is monotonic function).

By putting log on Equation 8 we get log-likelihood:

$$\begin{aligned} logL(\theta) = l(\theta) = \sum_{i=1}^{n} log\left[\{\Theta(t_i)\bar{F}(t_i)\}^{\delta_i}\{\bar{F}(t_i)\}^{1-\delta_i}\right] = \\ = \sum_{i=1}^{n}[\delta_i log\{\Theta(t_i)\bar{F}(t_i)\}] + \sum_{i=1}^{n}[(1-\delta_i)log\bar{F}(t_i)] = \\ = \sum_{i=1}^{n}[\delta_i log\Theta(t_i)] - \sum_{i=1}^{n}\int_0^{t_i}\Theta(s)ds \end{aligned} \tag{9}$$

2 Modelling of S-Shaped Hazard Function

Consistenly with our aim, we introduce here our own S-shaped hazard function.

2.1 Why "S" shape?

It is very reasonable that people who got trainings (notice as first group from now) also got locked-in during that period of time; they could not look for a job and could not be hired immediately after their layoffs, we see the hazard needs to be close to zero at the beginning phase. In short term, trainings postponed their opportunity to get job and so, technically, they have very low probability to get employed in this time. This lock-in effect (very low or close to zero probability to change one's status from unemployed to employed) is essential characteristic for this situation. On the contrary, among people from other group (notice as second group from now) that did not have such limitations (they had higher probability to find job at the beginning phase) we expect they should be more successful in getting job, in short term. This fact should result in difference between hazard function for first group and the hazard function for second group in the beginning phase.

Later after training, we expect enourmous increasing of probability to get a job among the people from first group. It is not only because of releasing them from lock-in phase but also because of development in their qualifications and skills.

Even when shape of the final part of hazard function is ambiguous for us, we propose S-shape as sufficient and elegant one.

Letter "S" describes our function best because the hazard function is at the beginning very close to zero then after lock-in phase there is enourmous increasing and then finally, we expect some stable phase exactly as the letter "S" does.

2.2 Definition of our S-shaped hazard function

As far as we can express log-likelihood function by hazard function (see Equation 9), it is very useful to specify and restrict hazard function (restrictions are based on theory, knowledge and available information).

According to our definition and theory, our function needs to fullfil following condi-

tions. It needs to:

- hold S shape
- be flexible (so it can get closer to different data)
- be non-negative for time $t > 0$ and starts at zero $h(0) = 0$
- be integrable (so we are able to get log-likelihood function- see Equation 9).

Our proposed S-shaped hazard function is as follows:

$$\Theta(t) = \frac{arctan\left(\frac{t-c}{a}\right) - arctan\left(\frac{-c}{a}\right)}{b} \tag{10}$$

that can be also rewrite as:

$$\Theta(t) = \frac{tan^{-1}\left(\frac{c}{a}\right) - tan^{-1}\left(\frac{c-t}{a}\right)}{b} \tag{11}$$

We were motivated to use arctangent function because needed S-shape is directly included in arctangent. Then we modified basic arctangent function by including parameters such that the hazard function fullfils all stated conditions and it also becomes more flexible. Parameters a and c provide flexibility in horizontal way and parameter b provides flexibility in vertical movements. The part " -arctan $\left(\frac{-c}{a}\right)$ " from equation10 ensures it starts at zero, $h(0) = 0$ (see Figure 2). So far as arctangent is increasing monotonic function on real numbers and we set it such that $h(0) = 0$ this implies it is greater than zero for positive time, $t > 0$.

Last condition is to be integrable. After we integrated our hazard we got:

$$\int_0^t \Theta(s)ds = \frac{t\ arctan\left(\frac{c}{a}\right)}{b} + \frac{c-t}{b} arctan\left(\frac{c-t}{a}\right) - \\ -\frac{c}{b} arctan\left(\frac{c}{a}\right) - \frac{a}{2b} ln\left(a^2 + c^2 - 2ct + t^2a^2 + c^2\right)$$

And so, we showed it fullfils all proposed conditions we mentioned above. Now through integrated hazard function, we can get density function and survival function of duration of unemployment and therefore also log-likelihood function.

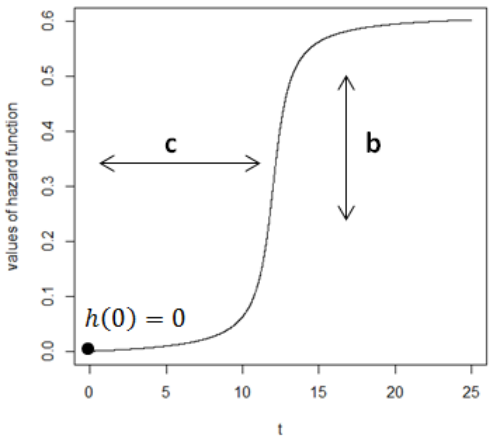

Figure 2: Ilustration of how parameters a, b and c from hazard function affect the shape of our hazard function

Density function:

$$f(t) = \frac{1}{b}\left(arctan\left(\frac{t-c}{a}\right) - arctan\left(\frac{-c}{a}\right)\right)$$
$$exp\left(-\frac{t}{b}arctan\left(\frac{c}{a}\right) + \frac{c-t}{b}arctan\left(\frac{c-t}{a}\right) - \frac{c}{b}arctan\left(\frac{c}{a}\right) - \frac{a}{2b}ln\left(a^2+c^2-2ct+t^2a^2+c^2\right)\right)$$

Survival function:

$$\bar{F}(t) = 1 - F(t) =$$
$$= exp\left(-\frac{t}{b}arctan\left(\frac{c}{a}\right) + \frac{c-t}{b}arctan\left(\frac{c-t}{a}\right) - \frac{c}{b}arctan\left(\frac{c}{a}\right) - \frac{a}{2b}ln\left(a^2+c^2-2ct+t^2a^2+c^2\right)\right)$$

From Equation 9, we get our log-likelihood function:

$$logL(\theta) = l(\theta) = l(a,b,c) = \sum_{i=1}^{n}\left[\delta_i log\left\{\frac{1}{b}\left(arctan\left(\frac{c}{a}\right) - arctan\left(\frac{c-t_i}{a}\right)\right)\right\}\right] +$$
$$+\frac{nc}{b}arctan\left(\frac{c}{a}\right) - \sum_{i=1}^{n}\left[\frac{t_i}{b}arctan\left(\frac{c}{a}\right)\right] -$$
$$-\sum_{i=1}^{n}\left[\frac{c-t_i}{b}arctan\left(\frac{c-t_i}{a}\right)\right] + \sum_{i=1}^{n}\left[\frac{a}{2b}log\left(\frac{a^2+c^2-2ct_i+{t_i}^2}{a^2+c^2}\right)\right] \tag{12}$$

Moreover, parameters included in our hazard function do not only make the hazard more flexible but also fullfil another role. From Figure 2, we can see that parameter c decides about horizontal movement of hazard. Therefore, when we have two hazard

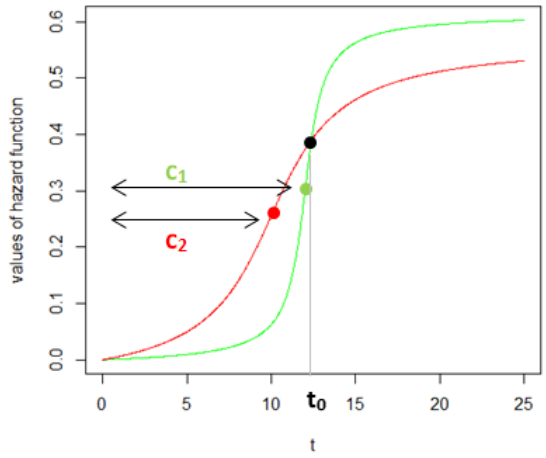

Figure 3: Ilustration for our S-shaped function with two different parameter settings: $a = 0.8$, $b = 5$ and $c_1 = 12$ (green) and $a = 3$, $b = 5$ and $c_2 = 10$ (red)

functions (one about participants and one about non-participants), parameter c enables us to measure lock-in effect for participants (if there is any).

We define lock-in effect as difference between parameter c_1 for participants and parameter c_2 for non-participants (see Figure 3).

$$\text{lock-in} = c_1 - c_2$$

Our hazard function has competitive advantage because until now, there is no other hazard function defined that would be able to measure the lock-in effect in terms of a single paramater.

We can also test if lock-in effect is significant. We use Wald test to test null hypothesis:

$$H_0 : c_1 - c_2 = 0$$

and then we know statistic t follows Normal distribution:

$$t = \frac{c_1 - c_2}{\sqrt{Var(c_1 - c_2)}} \sim N(0, 1)$$

and because:

$$Var(c_1 - c_2) = Var(c_1) + Var(c_2)$$

We want to find variance of c_1 and c_2. This is done through inverse Fisher information matrix. Firstly, we get observed Fisher information matrix:

$$I(\hat{a},\hat{b},\hat{c}) = -\begin{pmatrix} \frac{\partial^2 l}{\partial^2 a} & \frac{\partial^2 l}{\partial a \partial b} & \frac{\partial^2 l}{\partial a \partial c} \\ \frac{\partial^2 l}{\partial b \partial a} & \frac{\partial^2 l}{\partial^2 b} & \frac{\partial^2 l}{\partial b \partial c} \\ \frac{\partial^2 l}{\partial c \partial a} & \frac{\partial^2 l}{\partial c \partial b} & \frac{\partial^2 l}{\partial^2 c} \end{pmatrix}$$

and then inverse of observed Fisher's information:

$$I^{-1}(\hat{a},\hat{b},\hat{c}) = \begin{pmatrix} I^{11} & I^{12} & I^{13} \\ I^{21} & I^{22} & I^{23} \\ I^{31} & I^{32} & I^{33} \end{pmatrix}$$

From inverse Fisher information matrix we can get $Var(\hat{c}) = I^{33}_{\hat{c}}$. And so, after we get Fisher information matrix for optimized parameters $I(a_1, b_1, c_1)$ for training group we can compute inverse of this matrix and look at element I^{33}_{c1} that will tell us $Var(c_1)$ (process for getting $Var(c_2)$ variance of parameter c_2 for nonTraining group is analogous). Then we can test if lock-in is significant at $\alpha = 5\%$ confidence level as:

$$t = \frac{c_1 - c_2}{\sqrt{I^{33}_{c1} + I^{33}_{c2}}} \sim N(0,1)$$

If $t > 1.96$ then we reject null hypothesis H_0 and therefore lock-in effect is significantly different from zero at 5% confidence level (mostly, we expect this results). Otherwise, lock-in effect is not significantly different from zero and we pronounce that even there is difference at the beginning phase in hazards between participants and non-participants but it is not significant difference. Not significant lock-in effect is positive indicator for training program (but this is what we do not expect to happen and therefore it becomes suspisious and it indicates data should go through further examination; possible warning for data quality check).

Also intersection of two hazard functions is the point of interest. If we have case as is shown in Figure 3 and we denote green color is hazard rate for participants and red is hazard rate for non-participants then we see that until time t_0 (time when two hazards equal) hazard rate for participants is lower than for non-participants. And after time t_0, the hazard rate for participants got higher than hazard rate for non-participants. Then we can conclude that training program is successful because despite of the beginning when participants had lower hazard rate than non-participants but then at time t_0 they catch them and after t_0 they got even higher hazard rate than non-participants.

Despite of the program disadvantaged participants in short term, participants gain in long term when they got ahead of non-participants. And so, by putting two hazard functions into equality:

$$h_1(t) = h_2(t)$$

by solving this equation we get estimated time when participants catch non-participants and get ahead of them in terms of hazard rate. The specific position of two hazards depends on parameters a, b, c from hazard for participants and non-participants. One situation is shown in Figure 3 where we conclude success of program, another situation could be if there is no intersection. It would mean that one group never catch another. And if the lower rate is for participants then we propose program as unsuccessful. There is more possible scenarios. It only depends on final parameters for two groups to be compared.

In one's interest could be also to find out when is the highest velocity and acceleration for certain hazard rate (or compare speeds (or accelerations) of two hazards). When does hazard gain the most for unit time? Speed can be examined by taking derivative of hazard function and then looking at maximum (or for comparison just plot two speeds of hazards). Derivative of our hazard function is:

$$v(t) = \frac{dh(t)}{dt} = \left[\frac{1}{b} arctan\left(\frac{t-c}{a}\right)\right]'_t = \frac{1}{ab}\,\frac{1}{1+\left(\frac{t-c}{a}\right)^2} \tag{13}$$

Figure 4 shows comparison between two velocities of hazards from Figure 3.

We see that velocity of hazard for participants reach the highest speed around time $t = 12$ and the highest velocity for participants is almost four times higher than the highest velocity for non-participants. The highest velocity for non-participants is reached approximately at time $t = 10$. From Figure 4, we can see when hazard for participants catches the hazard for non-participants ($t = 12$). Acceleration is derivative from velocity and so it can be derived as:

$$a(t) = \frac{dv(t)}{dt} = \frac{-2}{a^2\, b}\,\frac{(t-c)}{a}\left[1+\left(\frac{t-c}{a}\right)^2\right]^{-2} \tag{14}$$

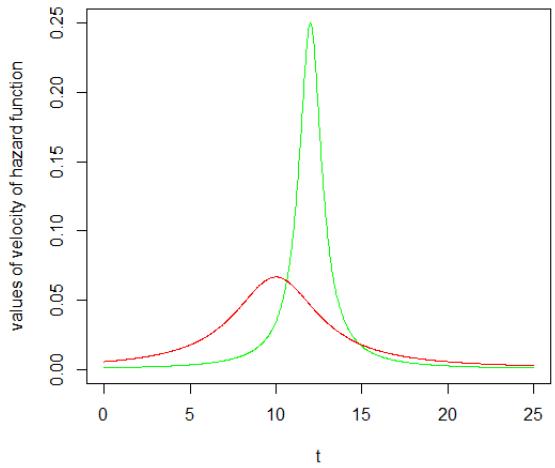

Figure 4: Ilustration of comparison between two velocities of hazards from Figure 3(green for participants and red for non-participants)

Plot for comparison between two accelerations of hazard from Figure 3 is shown in Figure 5. We see from Figure 5 that hazard function for participants reaches much higher acceleration than one for non-participants but then also reaches much lower acceleration what means hazard for non-participants is more stable. Peeks look symmetric (around x axis).

Because of the measurements that are available by using our hazard function (especially the lock-in effect), we find our definition of hazard as very usefull one for evaluation of program.

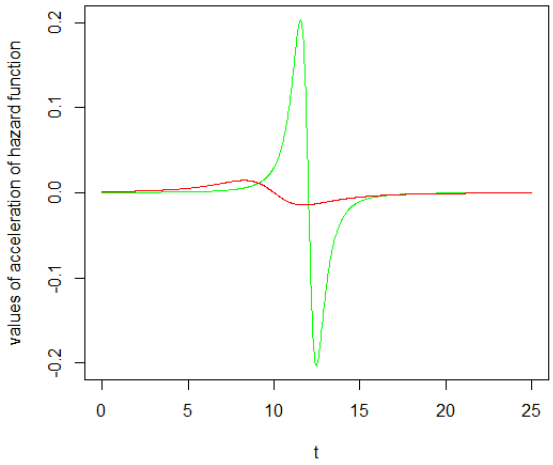

Figure 5: Ilustration of comparison between two accelerations of hazards (green for participants and red for non-participants) from Figure 3

3 Evaluation of Short Term Effect of Volvo Cars Project

In 2008, Volvo company was forced to go through restructuring process due to external factors. Because of the restructuring, 5713 current workers were made redundant. When almost 6000 people are becoming suddenly unemployed in particular country, goverment should face to this problem and take part of responsibility for it. Rate of unemployment is a good marker of social welfare, it shows how well the economy is, therefore it is in goverment's interest to minimize unemployment rate. When Swedish goverment conceded situation as definitive, they decided to make some arrangments to help people who will lose their job in Volvo. Swedish government made a project supported by European Globalization Fund. The aim of the project was to increase chance to get a new job after one was made redundant by Volvo company. The plan was to change qualification of those people through received trainings which should result in developing their new skills, so they should be able to work in different field than car industry. The project started in 2009. Support was offered to all redundant workers (5713) in terms of possibility to get training.

We got records about 4974 people from unemployment office (we lost contact about 739 people). In Sweden, one should announce when he/ she changes employment status in unemployment office. 1182 out of 4974 workers decided to participate in trainings and 3792 people decided not to participate. Starting points for trainings varied among individuals.

The aim of this study is to examine differences in behavior of duration of unemployment between group of redundant people that got trainings and those who did not.

Our goals are to construct two S shaped hazard functions (one for each group) and then according to their specific shapes make conclusions.

We have access to data from unemployment office where one can track development of individual's employment status over time. In subsection 3.1 we will introduce datasets that we worked with, explain important variables and later in subsection 3.2 how we process data in R.

3.1 Data Description

We have two datasets "Inskrivna" and "Sokandekategoribyten". Inskrivna is main dataset with 20834 records about 5487 unique individuals. Number of records vary from 1 to 34 per individual. Inskrivna contains 20 atributes such as: ID of individual, date when period of state begun,home region of individual, home land code of individual, date when period of state finished, state of (un)employment, gender, country of birth, education,... The oldest date from Inskrivna when period of state begun is from 1991-07-31 and the latest true date (that was recorded) when period of state finished is from 2012-01-24. Some records did not contain this information; empty space was present for such records and we filled it with date "2012-02-15" this date is set for end of experiment. We did this correction for both datasets. Inskrivna contained 1305 records without ending date and Sokandekategoribyten also 1305 (probably same records).

Sokandekategoribyten is complementary dataset giving us more records about individual. Atributes in Sokandekateribyten are not identical with those in Inskrivna, so we can get more information about individual by merging these two datasets. We are able identify same individuals through ID. Sokandekategoribyten contains 65278 records also about 5487 individuals (about same ones as Inskrivna). Number of records vary from 1 to 65 per individual. Sokandekategoribyten contains 21 atributes such as: ID of individual, date when period of state begun, state of employment (but different states of (un)employment are defined as in Inskrivna), date when period of state finished, if one participated in vocational training within the field of expertise,... The oldest date from Sokandekategoribyten when period of state begun is from 1987-10-14 and the latest true date (that was recorded) when period of state finished is from 2012-01-24.

3.2 Data Processing

The 'Volvo Project' was already evaluated through comparison of weeks unemployment rates in time, see Alam et al. (2012). Despite of studies differ in kinds of approach how to evaluate program, we tried to process data in way to be as consistent with Alam et al. (2012) as possible. Alam et al. (2012) used descriptive approach while we are assuming specific S-shaped hazard function for data (see Equation 10). We

examine difference between two groups (participants and non-participants) through their S-shaped hazard functions. Description of process how we got final data about participants and non-participants follows.

We have two datasets "Inskrivna" (20834 records) and "Sokandekategoribyten" (65278 records). Each record contains information about individual's certain state of employment. Start date ("INTR_DAT" in Inskrivna and in Sokandekategoribyten) and end date ("AV_DAT" in Inskrivna and in Sokandekategoribyten) are available and so we are able to compute duration of state. We need to clasify one's state of employment as "Employed" or "Unemployed". Classifications for states of employment differ in two datasets.

Consistently with Alam et al. (2012), we define one as employed if in Inskrivna the AVORS code equals to 1 (Permanent employment or business without support), 2 (Temporary employment), 3 (Back to the same employer) or 4 (Employment by Samhall). If AVORS was empty, information about one's employment is NA (not available). Otherwise, we assume one as unemployed. As far as information about employment is crucial for evaluation, we remove records from Inskrivna that do not contain information about (un)employment (AVORS=NA, 1305 records removed). We filter important variables from Inskrivna: "PERSON_ID" (unique ID of individual), "INTR_DAT" (start date for certain state), "AV_DAT" (end date for certain state) and "AVORS" (we transformed original coding from 1, 2, 3, 4,... into 0/1 that corresponds to unemployed/ employed).

Similar process was made with data from "Sokandekategoribyten". In "Sokandekategoribyten", there is SKAT code instead of AVORS and one is defined as employed if SKAT equals to 21 (Part-time employment), 22 (Payroll employment), 31 (Temporary work), 35(Ombytessokande Samhall) or 41 (Looking for change). We checked if there are any records with empty SKAT (no such record) and otherwise one was taken as unemployed. We filter important variables from Sokandekategoribyten: "PERSON_ID" (unique ID of individual), "INSKA_DAT" (start date for certain state), "UTSKA_DAT" (end date for certain state) and "SKAT" (we transformed original coding from 21, 22, 31, 35,... into 0/1 that corresponds to unemployed/ employed). After we filtered important variables from Sokandekategoribyten and we transformed original

SKAT we got consistent information with information from Inskrivna that contains ID, Start date, End date and Employed (0/1). We put information from Inskrivna and Sokandekategoribyten together. If End date was empty we set it as "2012-02-15" which corresponds to end of experiment (we did not observe true end of state).

We do not have information when (date) one participated in training program. We have only list that contains IDs of individuals that participated in program and those who did not participated. Next, we know training programs officially started on January 1st, 2010 (but it virtually started before the oficial starting date,some time in 2009). We denote this week as week 57 because it is 57^{th} week from possibly first layoff. Consistently with Alam et al. (2012), we will examine only individuals that had record with unemployed state that lasted during week 57 and look at its length. By this process we want to catch first spell of unemployment. As far as program was officially launched this time and it is only one year after first possibly layoff, this process looks like the most suitable one how to catch most of first spells (those who are unemployed this time they are probably unemployed for first time and so we caught first spell; it is relatively short period of time to already lose another job).

We want to find length of unemployment (that lasted during week 57). We know starting date from record. We need to find end date of this unemployment.
Few scenarios can happen:

- Current record ended (one was stated as unemployed) and later there exist another record about this individual that states him/ her as employed (we look at first one that follows after current record) then we take start date from record that states him/ her as employed and difference between this start date (employed) and start date (unemployed) is final lenght of unemployment (that lasted during week 57). As far as record of unemployment is followed by record of employment we can observe true length of unemployment and so data is uncensored.

- Current record ended (one was stated as unemployed) and later only records that state this person as unemployed follow. In this case (no record about employment follows), we take this individual as he/ she is unemployed whole time from start date (from unemployed record that lasted during week 57) until date December 15^{th} 2012 (end of experiment) and for length we set the difference between these

two dates. The individual did not find job until experiment ended and so we did not observe real length of unemployment, we just know it was at least of this length. Such data are censored.

- Current record did not end (end date was set to December 15^{th} 2012). To get length we just take difference between end date and start date of this record and as far as this individual did not find job until experiment ended, data is stated as censored (similarly as above).

We got dataset with IDs, length of unemployment (that lasted during 57 week) and indicator (that indicates if data are censored or not(0/1)). We made one correction with data. If length of unemployment was shorter than 6 days (suspisious) we looked at data and they went through closer examination (6 records about participants and 24 about non-participants). Such short length of unemployment was probably result of contradictory definition of employments (between Inskrivna and Sokandekategoribyten) or by mistakes.

By using list that contains IDs of participants and non-participants we are able seperate records into these two groups. Now, we have two datasets that are containing all important information (ID, length of unemployment and indicator) about participants and non-participants (we denote Training and NonTraining).

In terms of being able to compare results from Training and NonTraining, balanced groups are needed. We balanced these two groups by using propensity scores. Propensity scores selects individuals from Training and NonTraining by similarities of individuals' characteristics. By balancing we want to eliminate factor that had impact on one's decision if participate in training or not. Matching was done by using 'optimal matching' algorithm applied on propensity scores (calculated via logistic regression model), following variables were included: gender, country of birth, duration of previous unemployment, education, region, dummy for registration before the latest unemployment, registration month and dummy of previous vocational training within the field of expertise. We used function matchit() in R (package: MatchIt). By matching we got groups (matchedTraining and matchedNonTraining) with 'similar' individuals. Balancing is necessary in terms of comparing comparable.

Finally, we got all important data and we can optimize parameters from hazard func-

tion. We were optimizing log-likelihood for all four datasets (Training, NonTraining, matchedTraining and matchedNonTraining) and we found optimal values for parameters a, b and c (see results in Section 4). We used function optim() in R. To find suitable initialize values for parameters we plotted minus log-likelihood in Matlab and made video for values: $a \in (0; 5000]$, $b \in (0; 2000]$, $c \in [-5000; 0) \cup (0; 5000]$ and $a \in [-5000; 0)$, $b \in [-5000; 0)$, $c \in [-5000; 0) \cup (0; 5000]$ (see one frame from video Figure 10).

3.2.1 Estimated hazard and estimated integrated hazard function

According to Kiefer (1988), we can get estimated hazard function as:

$$\hat{\lambda}(t_j) = \frac{h_j}{n_j}$$

that corresponds to the number of "failures" at duration t_j divided by the number "at risk" at duration t_j.

and estimated integrated hazard function as:

$$\hat{\Lambda}(t_j) = \sum_{i \leq j} \hat{\lambda}(t_i)$$

After we modeled estimated hazard function (see Figure 6) and estimated integrated hazard function (see Figure 7) for our data, we can see convex shape of estimated integrated hazard what implies that our hazard is increasing. This fact supports the idea of S shape hazard function. However, we need to mention that estimates for longer durations are less accurate (for any inference) than for shorter ones (Kiefer, 1988).

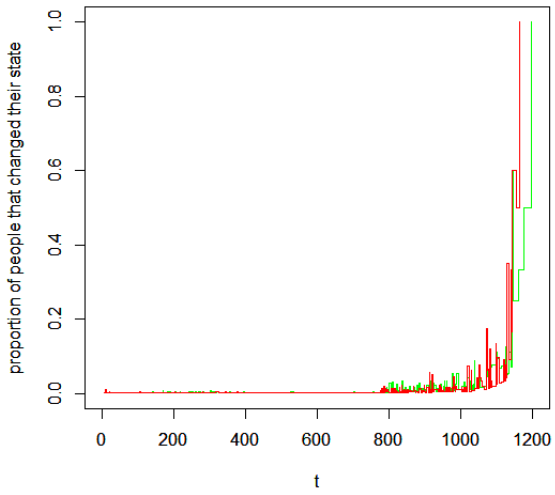

Figure 6: Comparison of two estimated hazard functions (from Training and NonTraining)

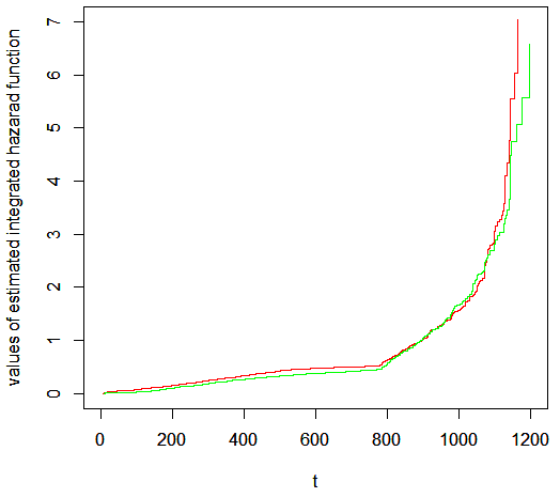

Figure 7: Comparison of two estimated integrated hazard functions (from Training and NonTraining)

4 Results

One aim of this study is to apply our S-shaped hazard function to real dataset. We introduce here results from data about Volvo Cars training project. From Section 3.2, we got two kinds of datasets: before balancing and after balancing. And then we have two datasets for each kind: people who participated in training and people who did not participate (from now we denote people who participated on training as 'Training', people who did not participate on training as 'NonTraining', Training after balancing as 'matchedTraining' and NonTraining after balancing as 'matchedNonTraining'). So, in total we got four datasets. All four datasets contain only necessary information (ID, length of duration and indicator) for getting optimal paramaters a, b and c from our $log - likelihood$ function (see Equation 12).

We found global maximum of $log - likelihood$ by using function $optim()$ in R (because of function optim() is set as it is looking for minimum from now we denote: behind log-likelihood we mean minus log-likelihood and therefore from now we are looking for minimum instead of maximum). "optim()" function requires initial values for parameters to be optimized. As far as our log-likelihood contains three parameters and because of our log-likelihood function is very sensitive to small changes in parameters it is essential to put reasonably good guess for initial values into $optim()$ function. The visualization of $log - likelihood$ helped us for initial guess. We allocated two local minimums for each dataset to which used to converge any initial values. Convergency to 1^{st} local minimum (see Table 1 and Table 2 below) took longer process than convergency to 2^{nd} local minimum. By comparing two values at these local minimums we get global minimum and so final optimal parameters for certain dataset.

Firstly, we allocated two local minimums for Training and NonTraining datasets.

Table 1: Local-optimal parameters (a;b;c) of log-likelihood for Training and Non-Training

Before balancing	Training	Non-Training
1^{st} local minimum	$(8.81e-10; 5260.27; 6.999)$	**(1.9786e-9;4197.029;5.999)**
2^{nd} local minimum	**(13.888;1746.8544;-6.8422)**	$(0.3526; 1880.844; -0.061)$

Table 2: Values of log-likelihood for local-optimal parameters (a;b;c) for Training and Non-Training

Values of log-likelihood	Training	Non-Training
1^{st} local minimum	3293.685	**6320.292**
2^{nd} local minimum	**3293.182**	6327.298

Those parameters are the closest ones to true parameters of 'all' trained people and 'all' non-trained people by assuming that hazard function is defined by Equation 10. These parameters give us the most accurate information about 'all' people who participated in training and separately, about 'all' people who did not participate in training. Interpretation of these two sets of parameters need to be isolated from each other. Illustration of hazard function for these two datasets looks like:

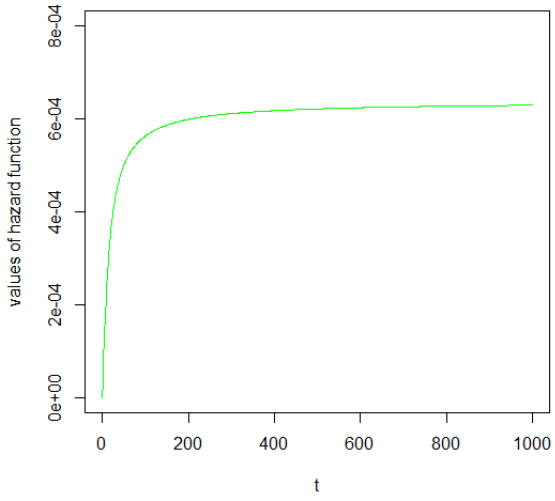

Figure 8: Shape of hazard function with optimal parameters for Training

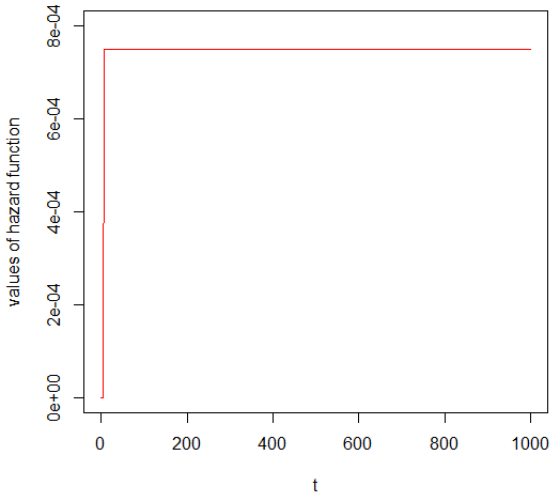

Figure 9: Shape of hazard function with optimal parameters for NonTraining

Secondly, we allocated two local minimums of log-likelihood for Training and Non-Training after balancing.

Table 3: Local-optimal parameters (a;b;c) of log-likelihood for Training and Non-Training after balancing

After balancing	Matched Training	Matched Non-Training
1^{st} local minimum	**(4.296754e-7;5546.592;7.820406)**	**(1.3435288e-10;4795.294;6.999)**
2^{nd} local minimum	$(2.441257; 399.440022; -10.072105)$	$(0.1592279; 997.742; -0.2100552)$

Table 4: Values of log-likelihood for local-optimal parameters (a;b;c) for Training and Non-Training after balancing

Values of log-likelihood	Matched Training	Matched Non-Training
1^{st} local minimum	**1898.59**	**2066.04**
2^{nd} local minimum	1899.212	2068.49

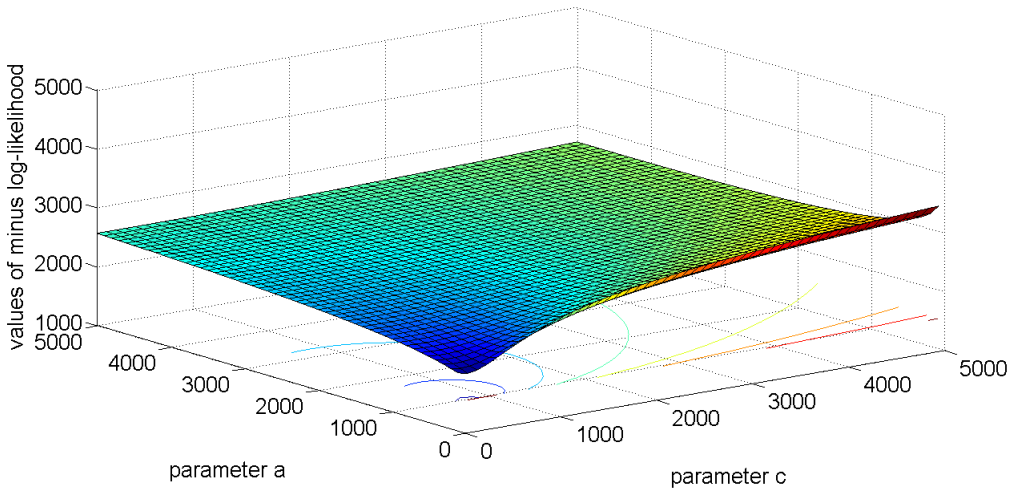

Figure 10: Illustration of minus log-likelihood function for fixed $b = 5546.592$ for matchedTraining

After zoom in Figure 10, we can see that peek that points to point $(0, 0, 0)$ does not go directly to $(0, 0, 0)$ but in c-direction it goes little bit away from 0 and that is how the optimal value for c parameter converges to 7.820406.

Finally, by balancing groups we got sets of parameters that are comparable and therefore also program can be evaluated from these results. Illustration of comparison

of two hazard functions follows:

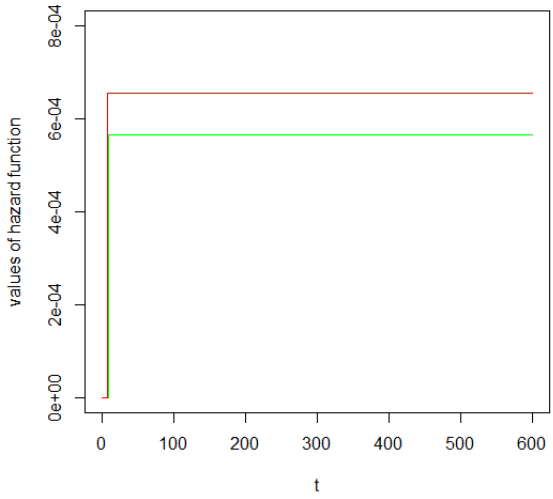

Figure 11: Comparison of hazard function for matchedTraining(green) and matchedNonTraining(red)

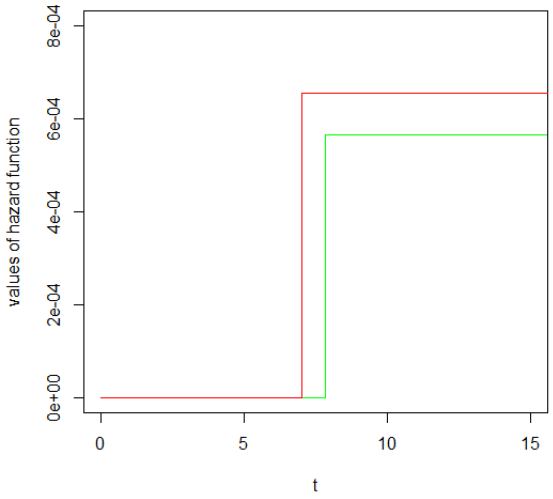

Figure 12: Zoom in of Figure 11, hazard function for matchedTraining(green) and matchedNonTraining(red)

By comparing parameter c for matchedTraining($c = 7.820406$) and matchedNonTraining($c = 6.999$) we can see there is no big difference (duration is measured in days) but still according to our results it takes longer time to find a job for one from matchedTraining than one from matchedNonTraining at the beginning (from stated time, here week 57). In this case, for such small difference and small values of parameter c, it does not seem to be reasonable to relate our results with term "lock-in effect". It can be tested through Wald test (process is described in Subsection 2.2, we did not make test here because of computational difficulty with 2^{nd} derivative of log-likelihood and following inverse Fisher information matrix) if lock-in effect is significant. Rest two parameters a and b are more crucial in our case. Specific values for these two parameters decide about height of hazard function. According to Figure 11, we can see that pair of parameters $(a; b) = (4.296754e - 7; 5546.592)$ for matchedTraining and pair of parameters $(a; b) = (1.3435288e - 10; 4795.294)$ for matchedNonTraining result in fact that hazard function for matchedNonTraininng is above the one for mathcedNonTraining at any time t. As far as matchedNonTraining have bigger chance to change their

state of unemployed at any time t we announce Volvo Cars training project as unsuccessful. However, if we look at velocities (Figure 13) and accelerations (Figure 14) we can see that both velocity and acceleration for matchedTraining reach enormously greater values than for matchedNonTraining but radical increasment starts sooner for matchedTraining (hazard for matchedTraining increases sooner than for matchedNonTraining what results in lock-in effect). From Figure 13 and Figure 14, we can conclude training had positive effect on participants (their velocity is much higher than one for non-participants) but positive effect was not big enough to say participants got general competitive advantage to non-participants (even there is huge difference in velocities (positively for matchedTraining) it was not enough compares to existing horizontal movement, lock-in effect). Hazard for participants never catched hazard for non-participants (see Figure 11). The answer why hazard for matchedTraining did not reach sufficient level of height to catch hazard for matchedNonTraining is probably hidden in Figure 15 and Figure 16. We can see from these figures that both accelerations are assymetric but they have different tendency. Negative peek in acceleration for matchedTraining is twice greater than positive peek which means velocity goes fast up but then it goes down even twice faster (it is equivalent to say that change in convex part of hazard is smaller than change in concave part of hazard). In acceleration for matchedNonTraining, negative peek is much smaller than positive peek which means increasment (change in convex part of hazard) of velocity of hazard is steeper (bigger) than decreasment (change in concave part of hazard). Acceleration of hazard gives us information about changes in concavity/ convexity of certain hazard.

If we define successful program as one in which participants get ahead of non-participants (in terms of hazard function), which means intersection of their hazards appears at any time t_0, e.g. Figure 3, then existence of intersection is our interest. Existence of intersection depends on two factors: velocity of hazards (how fast they raise) and acceleration of hazards (which is telling us tendency of changes in convex and concave parts in hazard). Specific situations result in intersection or no intersection.

Examination of explicit relation between existence of intersection, velocity of hazards and acceleration of hazards stays for future work.

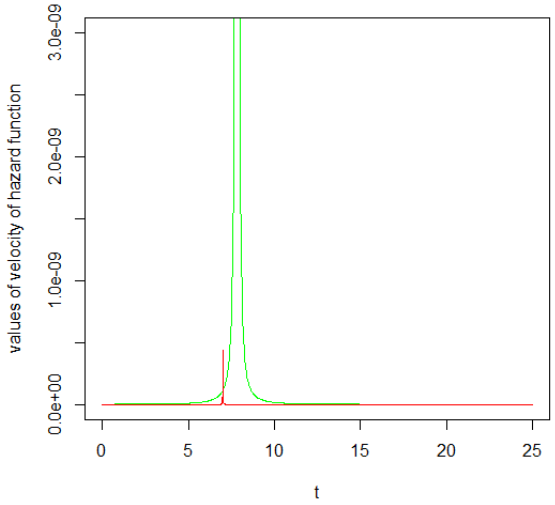

Figure 13: Comparison of velocities of hazard function for matchedTraining(green) and matchedNonTraining(red)

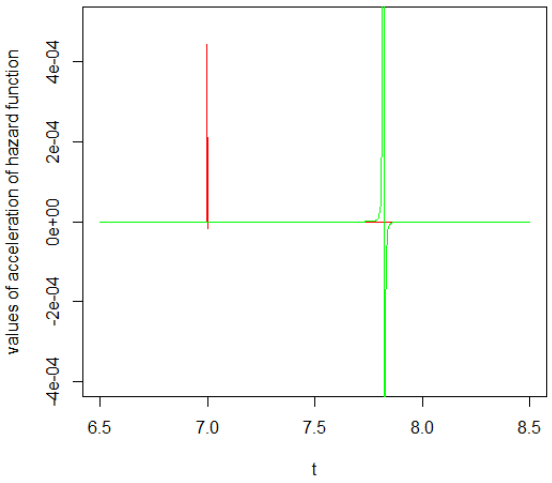

Figure 14: Comparison of accelerations of hazard function for matchedTraining(green) and matchedNonTraining(red)

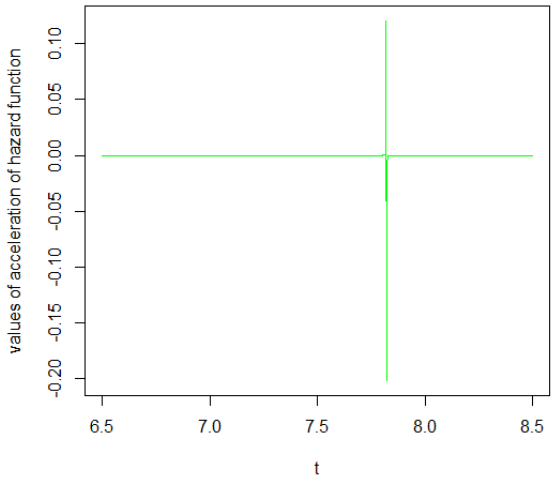

Figure 15: Acceleration of hazard function for matchedTraining

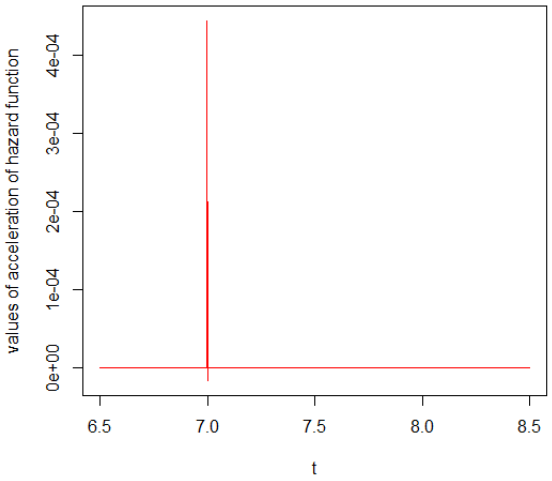

Figure 16: Acceleration of hazard function for matchedNonTraining

Concluding discussion

In our study, we have introduced our S-shaped hazard function based on the arctangent function with three additional parameters. The S-shaped hazard function was designed in such way that the parameter c provides information about lock-in effect and even more we are able to test whether this lock-in effect is significant or not. Until now, we know about no other approach that would be able to measure the lock-in effect by a single parameter. Included parameters enable the function to move in different directions and to approach data as closely as possible.

We have also derived the likelihood function and we have applied our S-shaped hazard function on data from Volvo Cars Project. Results were quite surprising for this particular dataset. In final version of the hazard functions, we did not expect such steep increase in the beginning phase followed by flat shape later (see Figure 11). Few reasons follow that could impact final results.

Contradictory data. Sometimes data were collapsing, for instance: according to records from Inskrivna it could very well happen that an individual was unemployed but according Sokandekategoribyten the same person was also employed at the same time. In these cases, we took the shortest possible unemployment time. Contradictory states at certain time probably exist because of the obvious lack of unemployment definition [2] or simply because of falsely stated dates (for example if an officer did not close record at the true date and so the record has lasted longer than it should have). As Heckman et al. (1999) mentions in his study, improvement in data significantly affects results. Therefore we advice to put emphasize on data quality in the future.

Finding global minimum. Because surface of our minus log-likelihood function significantly varies with changing parameters, it is quite challenging to find global minimum (three optimal parameters) of this function. Even the function "optim()" in R had trouble to converge. The problem is there are more local minima available and they are located 'far enough' from each other to not converge to global minimum from any initial values but only converge to the closest local minimum (sometimes the limited number of iterations is also problematic). Because of this problem findings of global

[2] original states of (un)employment in two datasets differ in such way that it is not possible to match them properly

minima are becoming more of a manual process than an automatic one which is certainly a big limitation. We assume global minimum as minimum of known minima which does not have to be necessarily right. It is possible that our parameters do not correspond to a true global minimum. Closer examination would be more time consuming and more sophisticated computational and automatic process would be needed to produce reliable analysis.

Violation of independency assumption. In duration data, we assume probability that individual's state can be changed is independent of time. It is likely that this assumption is not fulfiled in our case. As far as there were almost 6000 workers redundant in short period of time (and mostly from the same region) and we have only one job market it could happen that people who did not participate in the program found jobs (occupied places that were free until then) and filled the job market. And after training finished, participants do not have the same chance to find a job as non-participants had during the period when training was held. This hypothesis is difficult to test but if it is indeed true then the probability to change one's state depends on time and we cannot consider this data as duration data.

Also limitations for theoretical part of our approach are present. There are no covariates included. Developing this part is quite challenging and stays for future work as well as examination of more spells. We only looked at first spell of unemployment. We could go further and investigate also next spells and look at long-term effect. Very useful feature of approach would be explicit relation how existence of intersection of two hazards depends on velocity of hazards and acceleration of hazards. Motivation for future work can be also to derive this relation.

References

Md. M. Alam, Kenneth Carling, and Ola Naas. Utvrdering av det arbetsmarknadspolitiska projektet Volvo Cars och dess underleverantrer. *Working papers in transport, tourism, information technology and microdata analysis*, 2012.

J. de Koning, M. Koss, and A. Verkaik. A quasi-experimental evaluation of the evaluation of the vocational training centre for adults. *Environment and Planning C: Government and Policy*, 9, 1991.

A. Harkman, F. Jansson, and A. Tamas. Effects, defects, and prospects- an evaluation of labor market training in sweden, 1996.

J. J. Heckman, R. J. Lalonde, and J. A. Smith. *The Economics and Econometrics of Active Labor Market Programs*, volume 3A. Elsevier Science B. V., 1999.

N. M. Kiefer. Economic duration data and hazard functions. *Journal of Economic Literature*, 26, 1988.

T. Lancaster. *The Econometric Analysis of Transition Data*. Cambridge University Press, 1990.

P. Thierry and M. Sollogoub. French employment policies for youth. An econometric evaluation. *Revue-Economique*, 46, 1995.