A time-dynamic duration model for driver's license holding in discrete-time *

Siamak Baradaran

Christer Persson

Muriel Beser Hugosson

Anders Karlström

sia@kth.se

KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden

December 22, 2016

Abstract

Significance of information on individual driver's license holding has mainly been discussed in context of its influence on vehicle ownership and traffic safety analysis in the literature. Despite its importance and influence on other predictive models, for instant on vehicle ownership models,the acquirement mechanism of driver's license is rarely modeled. The aim of this research has been to develop a descriptive parametric model that could help us to understand individuals propensity to acquire driver's license from a behavioral perspective. We also sought to incorporate longitudinal trends and to identify potential time-dependent factors/variables, in order to detect and understand their time-dynamic nature. In this research, we managed to construct a so called *complementary log-log model*, employing the discrete time survival modeling framework. We have been able to to show that, despite the fact that discrete time modeling domain is far more

*This research has been carried out at the Department for System Analysis and Economics and Center for Transportation Studies at Royal Institute of Technology, KTH, in Stockholm - Sweden and been sponsored by Swedish Transport Administration. We would like to thank Mr. Lars Johansson at Swedish Transportation Administration for his support of this project.

restricted than the one for continues time analysis, it is possible to compose attractive dynamic models that can thoroughly emulate the underlying dynamic processes of interest. We could show that unlike earlier hypothesis, growth rate of female license holders is in proportion with male population, yet lower. We could also verify earlier hypothesis about the broad dependencies between young adults propensity ot acquire a license and their parents income and vehicle ownership. Furthermore our model shows that individuals who live in large cities acquire their license later than those living in towns and villages.

Introduction

Significance of information on individual driver's license holding has mainly been discussed in context of its influence on vehicle ownership and traffic safety analysis in the literature. In models that aim to replicate trip behavior, driver's license holding variable is predominantly utilized for adjustment of individuals choice sets[1]. For that reason, propensity of individuals towards acquirement of driver's license and their dynamics over time, exhibits crucial information for prediction of future shares of license holders.

Despite their importance and influence on other predictive models, driver's license holding models are quite rare, in particular when compared to related models such as vehicle ownership or mode choice models. Reviewing litterateur in search of such models also appears to be a discouraging pursuit. driver's license holding models are rarely mentioned and then rather briefly.

There are however few causal models one can refer to. The Italian System of Interurban Passenger Trip demand model system, SIMPT, has an integrated driver's license holding model in it's mobility choice module; Cascetta (2001). The model has a binomial logit structure with license possession or non-possession as dependent alternatives and socioeconomic and spatial variables as independents.

The ANTONIN[2] model for Paris is a disaggregate multinomial logit model and includes a driver's license holding module. The driver's license module calculates the probability of having a driver's license whereas the total number of licenses in Ile de France is generated by a cohort model that also assimilates the generations effect into account; Tuinenga & Pieters (2006).

[1] In a mode choice model for instant, non-possession of a driver's license will exclude car trip alternative from the the individual's choice set.

[2] ANTONIN: ANalyse des Transports et de lOrganisation de Nouvelles INfrastructures

There is also one independent Swedish driver's license model developed by Cedersund & Henriksson (2006). The authors construct seperate independent linear regression models, categorized by age groups for individuals between eighteen and twenty four years old. The individual models merely incorporate a variable reflecting cost of acquiring a license and another predictor representing share of individuals with post high school education. As a result of the chosen simplicity and the implicit categorization, the model is not suitable for mimicking potential trends and is consequently not appropriate for prognosis purposes.

The Swedish national transport model system, SAMPERS, doesn't include a nominal model for driver's license holding. Number of individuals possessing a driver's license is therefore computed using actual shares of individuals holding a license in the base year[3]. This share is then adjusted to the anticipated population growth level for the targeted prognosis year. Since driver's license holding data is required at zonal level by the vehicle ownership and mode choice models, the total number of license holders is finally distributed to the zones by utilizing population and age information.

Despite the rareness of driver's license holding models, there exists numerous descriptive analysis where trends in development of driver's license acquirement are discussed, essentially with regards to socioeconomic and spatial aspects of the studied population. These analysis predominantly aim to identify ongoing dynamic trends in driver's license holding.

For instant, several related analysis shows that growth in number of driver's license holders is mainly induced by increasing number of female drivers. Some of these studies also show that young people in much of Europe, Australia and North America are acquiring their licenses at older ages, compared to earlier generations and those who do, drive less, see for instant Delbosc & Graham (2014) or McDonald & Trowbridge (2009).

Salonen (2003) studied young individuals in Sweden and showed that share of nineteen to twenty one year's individuals who live with their parents has increased from fifty, to more than seventy percent between 1978 and 2001. He also concludes that young adults dwell longer than earlier generations in building their own households. He singles out the altered job market as the prevailing reason behind the change. This also seems to be the main explanatory factor behind young peoples delayed introduction in job market, higher education levels, extended dependencies to parents and of course, postponement of driver's license acquirement.

[3]Shares of individual, holding a driver's license is gathered from travel survey data for the base year.

The aim of this research is to develop a descriptive model that could help us to understand the propensity of individuals in acquiring driver's license from a behavioral perspective. The candidate model should also permit for convenient analysis of changing policies and be sensitive towards potential lagged behaviors such as growth saturation or potential declining trends.

Moreover we would like to examine the accuracy of some of the claimed statements by other studies regarding gender, job market dynamics and dependencies to parents.

This paper is outlined as follows. Several feasible models and their attributes and limitations are discussed in the following two section, followed by a short section on employed data. Results from the estimated models are presented thereafter. In the last section, major findings are described and few practical findings are concluded.

Review of potential modeling approaches

Since our aim is to construct a causal model that also is responsive towards potential temporal trends in acquirement of driver's license, we are restricted to build a model that incorporate temporal aspects of the behavior, which is the reason we have to utilize a dynamic model framework.

To our knowledge, no dynamic model has yet been developed for acquirement of driver's license. Literature on transportation related dynamic models on the other hand is dominated by two classes of models, *Dynamic Discrete Choice Models*, DDCM, and *event history/duration models*.

In a DDCM model setup, choices are assumed to be independent (mutually exclusive) and all potential choices are considered (collectively exhaustive). In case of driver's license acquirement, the model estimates the probability that the individual acquires a license in each single point in time. The continues time space will therefore need to be discretized into short enough time intervals during which, the observable part of a individual's utility can be assumed to be stationary and therefore, computable.

The subsequent presumption that also needs to be made is that individuals are assumed to be able to appreciate, and most certainly, compare the utility of owning versus not owning a driver's license on any given time in time-space. Since utility of alternatives change with time[4], the probability of

[4]For example the individual may build a family, move to a bigger city, etc, which may influence the individuals choice of acquiring a driver's license.

acquiring a license becomes a temporal variable, hence the model becomes dynamic, see for instant Train (1986) or de Jong & Kitamura (2009).

The main problem with DDCM models rises from the required discretization of the continues time space. A coarse time resolution may violate the stationary state of utilities, which is required for the sake of computability. Conversely, too densely defined time intervals would explode the number of choices in the choice-set and would consequently make the computation process, exhaustive.

It can also be determinedly questionable to assume that the an individual can appreciate the utility of acquiring a license for all discrete time intervals, from the time s/he become eligible for acquiring a license.

Event history models, even called *duration models* or *hazard models* allow for examination of the longitudinal progression of the probability that an event occurs. These models estimate the hazard or probability of occurrence of the event in focus, given that it has not taken place until a specific point in time.

Unlike dynamic discrete choice models, duration models estimate the length of elapsed time to the event rather than probabilities of the event happening within different discrete time interval. Therefore there is no need for discrimination of continues time. This enticing feature makes duration models more suitable for studies such as ours.

There are several transportation related duration models that one can refer to such as Gilbert (1992), Hensher & Mannering (1994), de Jong (1996), Yamamoto et al. (1999) or Rashidi et al. (2011), all developed for modeling vehicle ownership. It is however possible and, as it will be demonstrated in this research, plausible to conduct a duration model for our purpose.

Yamaguchi (1990) advocate that the point of interest with duration models is to identify patterns and causes of the change over time. Bennett (1999) also argue that the predominant purpose of duration analysis is to allow for possibility of duration dependencies in the model. These aspects of duration models qualifies them for our purpose, which is to be able to reproduce potential behavioral trends. From this perspective, the purpose of our model is to describe why and for how long an individual remain in same state, in this case as non-owner of a driver's license.

Events such as acquirement of drive's license may occur at any instant in the time continuum. The way the longitudinal data is registered however is the aspect that categorizes these models into two distinct classes of models, *continuous time* and *discrete time* models.

In case of continuous time survival models, we either need to know the exact time of occurrence of the event or the observation intervals need to be sufficiently small in order to make it reasonable to assume continuity. The length of such events are therefore non-negative real numbers. Continuous time survival models are often utilized in the field of bio-medicine.

Discrete time survival models are primarily employed in the field of social science and are branched in two different groups of their own. For the first group, the time scale of occurrence of event is essentially discrete. This for instant is the case for elections or school gradings, which occur in certain discrete dates.

The second group of discrete time duration models employ so called *grouped*, *banded* or *interval censured* data and refer to events that occur in continuous time, while their observations are made within discrete time intervals e.g. weekly, monthly or annually, see Singer & Willett (1993) for instant.

The data utilized in this research represent events that arrive in continues time, while they are assembled from databases that are updated annually. Furthermore, no information is provided on exact time of occurrence of the event within each specific year. Potential development of events are literally registered through comparison of data for individual driver's license holding states, between two subsequent years. Therefore the data can not qualify for continues modeling approach and has to be recognized as interval censured.

Let us start with few definitions. The individual state may potentially change at any time, from non-owner to owner[5] in the time space. This change in individual state is demarcated by an event, in this case acquirement of a license. The time interval that the individual remains within the same state is called a *spell* and it's length is denoted as *duration*, T. The starting point of the spell is assumed to be the point in time that the individual become eligible for acquiring a driver's license [6] and the spell is concluded if:

- the individual has acquired a license,
- data on that individual has been discontinued,
- or the individual has not yet acquired a license by the end of the spell.

[5]Here, the process of acquiring driver's license is assumed to be irreversible and we ignore the fact that the license can be revoked.

[6]Individuals become eligible to acquire a driver's license once they turned 18 in Sweden.

Following same mathematical notation as Rodriguez [7], we denote the length of the spell in the discrete time space by the random variable T so that $t_0 < t_1 < ..., t_k$. Since events in this case are designated to discrete time, they are sequenced to finite, non-overlapping and contiguous time periods where dates representing points in time, $t_0 = 0, t_1, t_2, ..., t_k$ are the discrete time interval boundaries. The discrete time intervals are consequently defined as

$$[0 = t_0, t_1], (t_1, t_2], (t_2, t_3], ..., (t_{k-1}, t_k = \infty] \tag{1}$$

$f(t_j)$ is then defined as cumulative distribution function of realization of T (e.g. acquiring driver's license), also known as *failure function*in survival analysis.

$$f(t_j) = f_j = Pr\{T = t_j\} \quad . \tag{2}$$

The so called *survival function*, $S(t_j)$ is then defined as

$$S(t_j) = S_j = 1 - f(t_j) = Pr\{T \geq= t_j\} = \sum_{k=j}^{\infty} f_j \quad . \tag{3}$$

Observe that $S(t)$ and $f(t)$ are both probabilities with values between 0 and 1. Furthermore the survival probability, $S(t)$, is strictly decreasing with time. The value of survival function at the start of discrete time interval j is given by

$$Pr(T > t_{j-1}) = 1 - f(t_{j-1}) \quad , \tag{4}$$

and at the end of interval j is given by

$$Pr(T > t_j) = 1 - f(t_j) \quad . \tag{5}$$

The conditional probability of occurrence of the event in interval, (t_j), also called *discrete hazard* or *interval hazard* is defined as

$$\lambda(t_j) = \lambda_j = Pr\{T = t_j | T \geq t_j\} = \frac{f_j}{S_j} \quad . \tag{6}$$

The conditionality has the implication that in order to survive to discrete time interval t_j, the individual has to survive all discrete time intervals prior to t_j so the Survival function from equation 3 can be rewritten in terms of hazards in preceding intervals.

$$S(t_j) = S_j = (1 - \lambda_{t_1})(1 - \lambda_{t_2})...(1 - \lambda_{t_{j-1}}) \quad . \tag{7}$$

[7] Rodrguez, G. (2007). Lecture Notes on Generalized Linear Models. URL: http://data.princeton.edu/wws509/notes/

Hazard rate obviously is a function of survival time $\lambda(t)$. In order to allow it to represent variations between individuals, depending on their characteristics, they need to be specified. so let $\boldsymbol{X}$ be a set of $X \times 1$ vector of characteristics of individuals, $X_1, X_1, ..., X_k$. These characteristics are generally incorporated in the model through introduction of a linear combination of the characteristics

$$\boldsymbol{\beta X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \quad , \tag{8}$$

where $\boldsymbol{\beta} = \beta_0, \beta_1, ..., \beta_k$ are to be estimated.

Depending on the type of data, truly discrete or continues and grouped, the functional form of hazard and survival model can be constructed in different ways, see for instant Beck et al. (1998) and Carter & Signorino (2010) . The choice of functional form should obviously be based on our understanding of the underlying process. Three different functional forms are commonly adopted, *logistic*, *piece-wise exponential* and so called *complementary log-log* or *c-log-log* specification. The logistic form is favored most frequently in literature[8] and is useful in situations in which, time to event is intrinsically discrete[9]. However, Kalbfleisch & Prentice (1982) show that the c-log-log specification is uniquely appropriate for grouped data in continuous time and under proportional hazards framework, which also is the case with our data.

We start with the survival function in proportional hazards framework Cox (1972), which can be written as

$$S(t_j|\boldsymbol{x}_i) = S_0(t_j)^{\, e^{(\boldsymbol{x}_i'\boldsymbol{\beta})}} \quad , \tag{9}$$

where $S(t_j|\boldsymbol{x}_i)$ is the probability that individual i with covariate values $\boldsymbol{x}_i$ will survive up to discrete time interval t_j. $S_0(t_j)$ is the so called *baseline survival function* and describes the relative risk for individuals with $\boldsymbol{x} = 0$ and $e^{(\boldsymbol{x}_i'\boldsymbol{\beta})}$ is a proportionate relative increase or reduction in risk, associated with the characteristics of $\boldsymbol{x}$ in individual i. Equation 7 can then be rewritten as

$$1 - \lambda(t_j|\boldsymbol{x}_i) = [1 - \lambda_0(t_j)]^{\, e^{(\boldsymbol{x}_i'\boldsymbol{\beta})}} \quad , \tag{10}$$

and the hazard function for individual i can be written as

$$\lambda(t_j|\boldsymbol{x}_i) = 1 - [1 - \lambda_0(t_j)]^{\, e^{(\boldsymbol{x}_i'\boldsymbol{\beta})}} \quad . \tag{11}$$

[8] Because of their interpretation capabilities and due to the fact that software for the former is more available than for piece-wise and log-log model

[9] A intrinsically discrete-time process is a process that the event occurs in more or less discrete time, like grading of student that happens at the end of semester, where it is more convenient to measure time in umber of semesters rather than in months.

$\lambda_0(t_j)$ in equations 10 and 11 represent the baseline hazard, also called the shape function, and summarize the pattern of duration dependence. It however can not be observed and assumptions needs to be made about it's distribution such as linear assumption, j, or some transformation of it such as $ln(j)$, j^2 or j^3 or replaced by a cubic spline function, which have a smoothing effect and is generally efficient, see Beck et al. (1998). Linearization of the right hand side of equation 11 results in

$$log\Big(- log\big(1 - \lambda(t_j|\boldsymbol{x}_i)\big)\Big) \; = \; \alpha_j \; + \; \boldsymbol{x}_i'\boldsymbol{\beta} \quad , \tag{12}$$

where

$$\alpha_j \; = \; log\Big(- log\big(1 - \lambda_0(t_j)\big)\Big) \quad , \tag{13}$$

hence the notion complementary log-log function.

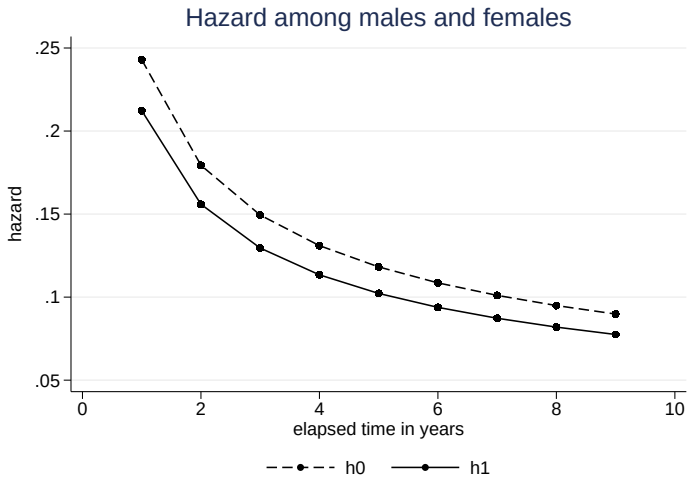

Figure 1: Distribution of hazard (y-axis) over time (x-axis) among male and female individuals.

The proportional hazard structure is attractive as it satisfies the separability assumption about effect of baseline hazard from contribution of predictors. This property implies that difference in hazards between two individuals i and m, in any given time interval is proportional with their observable heterogeneity, values of their independent variables $\boldsymbol{x}_i$ and $\boldsymbol{x}_m$ and moreover is independent of time. In other words, it explains variations in hazard rate's for different groups of population, for instant among females and males. In figure 1 for instant we can see that the shape of the hazard function is similar among females and males while they deviate with regards to their absolute value, visible along y-axis.

Censuring and truncation

Time is the most characterizing essence in survival analysis and in order to understand censuring and truncation, we need to comprehend the relationship between survival data and time. The manner individual observations are recorded has great consequence on the outcome of survival models. We have already mentioned the limitations we face, using discrete time data, which necessitates certain assumptions and reduces modeling capability. In this section we discuss further concerns related to incompleteness of the data.

In most studies, data is restricted to shorter observation periods and are to be considered as incomplete. Incompleteness of data leads to censuring and truncation and is a prevalent problem, see for instant Carter & Signorino (2013) or Cain et al. (2011). Censuring and truncation is best explained studying figure 2. Time is measured on the x-axis, while the lines, numbered

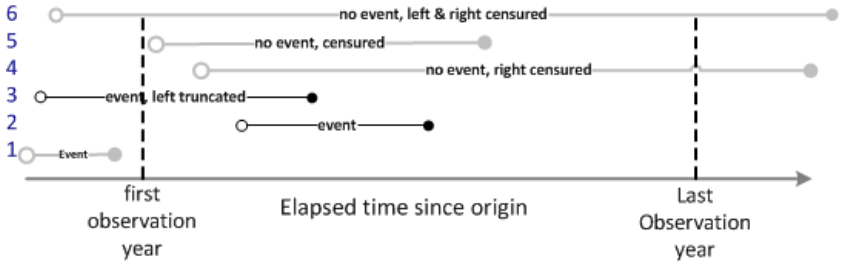

Figure 2: Different censuring and truncation scenarios.

one to six, represent different spells, yielding different individuals. The empty rings on the left of each line represent the start of each spell and the filled rings at the right of each line represent the exit time. We may have spells of different starting times and lengths. The two vertical lines encase the time interval, during which, data is available to us, denoted first and last observation years.

1. Spell one starts and ends before the first observation year and is therefore left-censured.

2. Spell two starts and ends within our observation years and is therefore not censured and is fully modeled.

3. Spell three started before the observations started, while it ended within the period and is left-truncated.

4. Spell four starts within the observation years and ends after. Since no event is registered within the observation years, the spell is right censured but has been accounted for in the model within the period from start of spell until the last observation year.

5. Spell five is entirely within the observation years and is accounted for in the model fully, however the individual exits (we lost track of her/him) and since no event is registered, she or he will be censured when the spell ends (last observation year).

6. Spell six has both start and exit outside the observation period and is only accounted for by the model while s/he was within the observation period. S/he is both left truncated and right censured since no events were observed.

Left censured individuals, those who acquired license before the start of observation period, create a bigger problem since starting time of their spell is unknown and the only way of dealing with them is to assume constant hazard through time which would violate most hazard models. These observations would be non-informative, if we could assume that they are independent of the survival distribution with no pattern and are left out randomly. If censuring is not independent, then censuring is said to be informative and would introduce bias in the models. In our case, there is not much we can do about the left censured data as we are completely uninformed about them.

We are modeling a non-repetitive event, which means that all individuals who were observed by the last observation year, 2011, and who either had or had not acquired a license, are part of the population we study. Right censuring would be a problem, in case we would have left out individuals who had not acquire a license by the last observation year. This however is not the case and all individuals are accounted for during the spell, even though they had not acquired a license by the end of hte spell.

Left-truncation present unique problems through introduction of bias in the model, explained for instant by Guo (1993) and Cain et al. (2011). However, there are techniques that can be used to account for Left truncation. As been showed in simulations by Cain et al. (2011), these techniques greatly reduces that bias. Nevertheless, as it also been showed by Cain et al. (2011), the estimates become increasingly unstable as the amount of truncation approaches or exceeds 50% of all observations.

In our case, the process of interest develops over many years. This implies that individual's age, at the onset of the process, varies considerably across the population. The earliest individual in our study turned eighteen years old, back in 1921 while s/he was observed in our data, first in 2003. As the first observation year is 2003, all individuals who turned eighteen years old before 1985, $(= 2003 - 18)$ are either left-censured[10] or left-truncated. The bias introduced by left truncation would increase with time as expected

[10]Since there are no information on the starting time of their event.

likelihood of individuals acquiring driver's license decreases with age. In other words, inclusion of left-truncated individuals will lower the hazard and increase the survival probability and spell length as result.

Reviewing our data, we found that number of truncated observations were more than 50% of all observations and as been mentioned earlier we have left censuring problem as well. This leaves us with no choice but to exclude all censured and truncated individuals, who are individuals that turned eighteen before 1985.

A comforting fact from our analysis shows that majority of individuals in our data acquire a license within first three to four years of their spell, which reduces the importance of inclusion of older data and justifies our decision to only use fully observed data.

1 Data

Two unique data sets are employed in this project. Both data sets include observations for years 2003 through 2011. The first data set is extracted from annual tax declarations, gathered for all adult individuals in Sweden and includes individual's socio-economic attributes. Heads of households and number of potential adult children are traceable through specific key-tables. This future has been used to determine total household income[11] and number of vehicles owned by the household as well as number of potentially accessible vehicles owned by one of the households head's parent/s, if the individual is an adult and living with her/his parent/s.

The other data set consists of information from annual vehicle inspections. This data set has been utilized to identify all vehicles owned by each individual and was later matched with the socio-economic data from the first data set.

The final data we compiled for modeling consist of 10% random sample of all unique adult individuals in Sweden, who are observed between 2003 to 2011. After exclusion of censured and truncated data, almost 340.000 single-year observations were left representing more than 100.000 unique individuals. Around 11.000 new individuals are observed entering the spell annually. Figure 3 shows that almost 20% of individuals who entered the spell between 2003 to 2011 got a license within the first year of spell. Around 20.000 individuals received their license in the second year of spell, which

[11]Household income is the individual income in case of single individuals and sum of individual incomes in case of married couples.

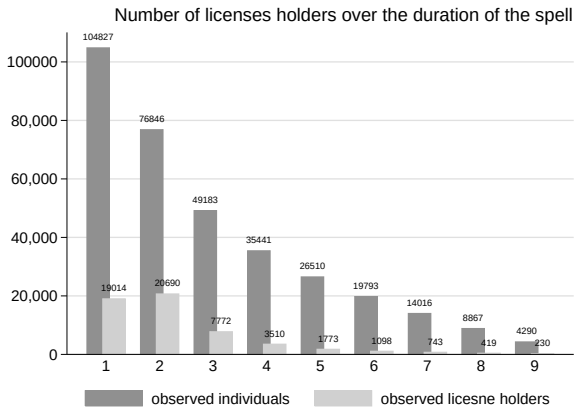

Figure 3: Total number of individuals and those who acquire their license over the duration of our data.

is almost 27% of all individuals who made it to second year. Same share was 16% in third, 10% in fourth, 7% in fifth, 5.5% in sixth, 5.3% in seventh, 4.7% in eighth and 5.3% in ninth year of spell.

The left graph in figure 4 shows that around 20% of individuals who turn eighteen, acquire their license within the same year. corresponding share of license holders among nineteen years old individuals is few percents more than first group, while it decreases to barely 7% among twenty years old individuals and continues to decline, the older the individuals become. share of males are slightly higher among eighteen and nineteen years old individuals while share of females become at least as large as males, moving into the third spell year.

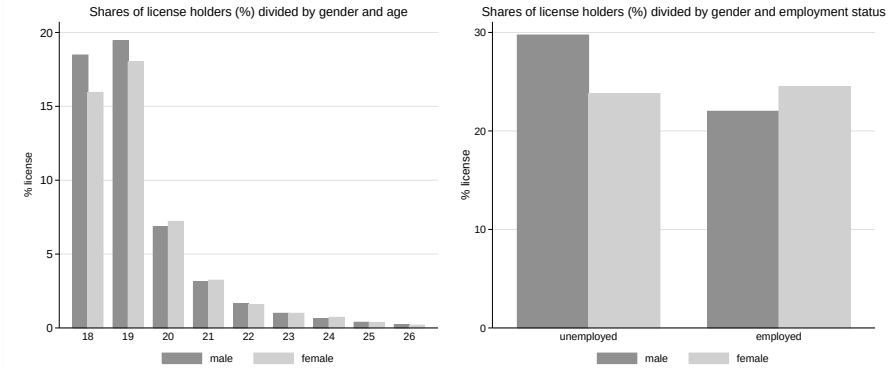

Figure 4: Share of new license holders by age and gender, left figure, and by employment status and gender, right figure.

The right graph in figure 4 shows that share of unemployed driver's license holders are slightly larger than employed individuals, which perhaps shouldn't come as surprising since almost 50% of individuals get their licenses by the time they've turned 20 years old and it is well known that

unemployment rates among young adults are much higher than general population.

The left graph in figure 5 shows distribution of observed individuals and observed license holders over population of urbanized areas in Sweden, categorized by their population in thousands, on x-axis. The right graph shows same data but this time both individuals and license holders are shown as percent of each urban population category. It can bee clearly seen that there are more license holders living in smaller villages and cities and number of license holders decrease as population grows.

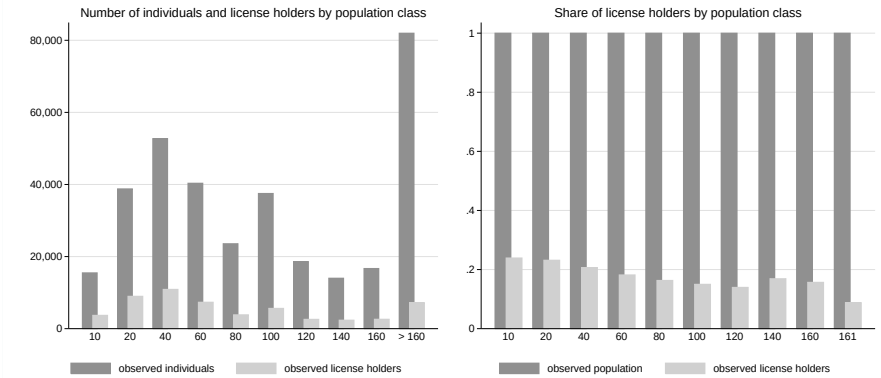

Figure 5: Number (left figure) and share (right figure) of new license holders by population class.

Models and results

As been explained, we were constrained to employ discrete-time model framework in consideration of the fact that our data is interval truncated. Following previous discussions , we have therefore formulated a *cloglog* proportional hazard model that also satisfies the separability assumption discussed earlier.

Choice of functional form for baseline hazard has been discussed briefly in earlier sections. It has been argued, by among others Bennett (1999), that functional form of the candidate model should be determined in conjunction with hypothesis made, potential dynamics in the data and applicable predictors. For sake of comparability, we have constructed three different models. All models share same set of covariates and only differ with regard to shape of their baseline hazard, λ_j. For the first model we utilized a log-transformed time function, $ln(j)$ which can be thought of as discrete-time analogue to the continuous-time Weibull model. The other two models are

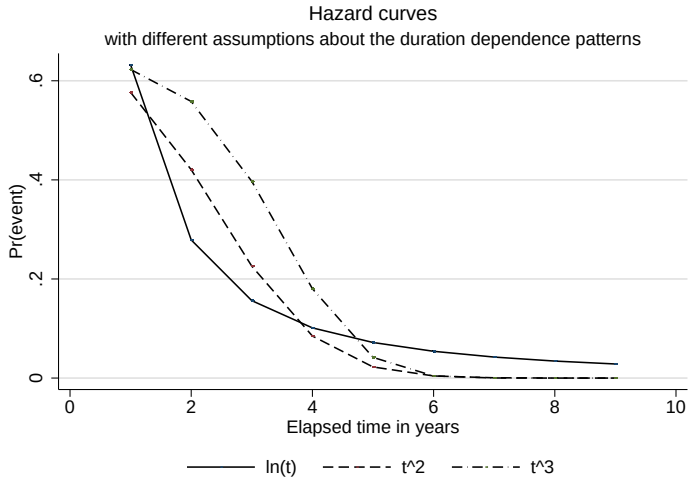

Figure 6: Probability of event, here as acquiring a license, on y-axis over elapsed time on x-axis for models with log-transformed, quadratic and cubic time functions.

using a quadratic and cubic transformation of time. Figure 6 shows the distribution of hazard over time in years, excluding effects from covariates[12]. It would have been interesting to test a model with cubic spline function as well. This has however not been possible with the current software version we have access to through Statistics Sweden.

Variables called *gender*, *employment status*, *student* and *vehicle access* have binary form[13]. Remainder covariates in table 1 are continuous variables. Parents income has been chosen due to individual's young ages and could indicate potential dependencies to parents, which was one of the hypothesis we would like to investigate. We have already mentioned the variable *vehicle access*, which shows if the individual have access to vehicles owned by parents[14]. This variable can also indicate potential dependencies of young individuals to their parents.

One of more important independent variables to account for would be level of accessibility by different transportation modes. It would for instant be a good idea to include a ratio between accessibility by public transportation-mode and by car-mode. Such data was however not available and we were constrained to use population data as a proxy, instead. Our hypothesis is that accessibility by public transportation-mode increases with size of cities (or in this case population). This implies that the bigger the population is

[12]Value of all x's are sett to zero.

[13]0 = male, unemployed, not student and no access to parents vehicles, while 1 = female, employed, student and have access to at least one vehicle.

[14]Vehicles owned by parents are assumed to be accessible to their children if they share address.

(the larger the city is), the higher would accessibility be with public transportation .This could potentially decrease individuals propensity towards acquiring driver's license at young ages.

	logisitc	quadratic	cubic
age	-0.363***	-0.107***	-0.0118***
gender	-0.200***	-0.205***	-0.205***
employment status	0.676***	0.700***	0.696***
student	-0.086***	-0.225***	-0.203***
num. of children	0.278***	0.152***	0.160***
ln(parenets income)	0.539***	0.218***	0.230***
vehicles access	0.467***	0.450***	0.451***
ln(population)	-0.176***	-0.189***	-0.189***
ln(time)	0.535***		
$time^2$		-0.023***	
$time^3$			-0.004***
num. of observations	117755	117755	117755
$log-likeliehood$	-62075.5	-62216.3	-62175.4
degrees of freedom	9	9	9
AIC	124169.1	124450.6	124368.8
BIC	124256.2	124537.7	124455.9

Table 1: Estimated model parameters for logistic, quadratic and cubic models.

Table 1 shows the estimated results. It is possible to compare the three different models using model statistics, gray area in the table, since all three models share same additive independent variables other than their baseline hazard. We can see that the logistic model supports our data best, comparing *log-likelihood*, *Akaike information criterion* (AIC) and *Bayesian information criterion* (BIC) values.

Age seems to have a negative effect on occurrence of event since it's coefficient has negative sign, which seems reasonable as number of events decrease with duration that is equal with $age + 18$ in our case[15]. This is also illustrated by figure 7, which moreover reveals that the hazard rate is higher among male individuals than females.

Coefficient of the population variable has, as been expected, negative sign. This confirms our hypothesis that it is less likely for individuals living in

[15]Duration of the spell is measure form starting state that is the same year as the individual turn 18 and ends at the event or if the individual is censured.

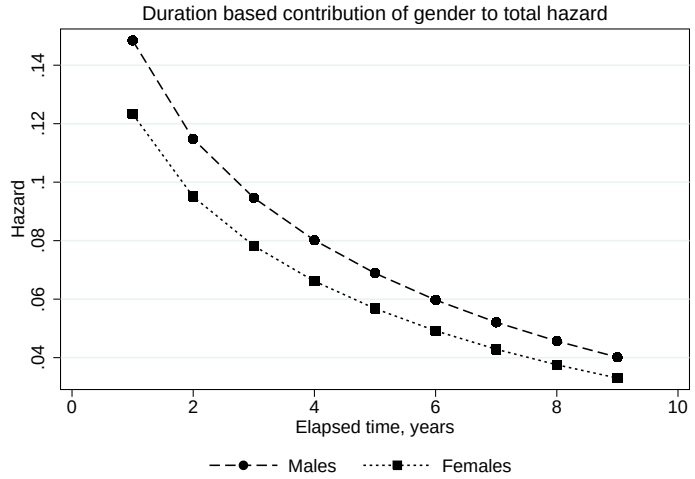

Figure 7: Comparison of distribution of hazard (probability to event) over the duration of study between females and male individuals.

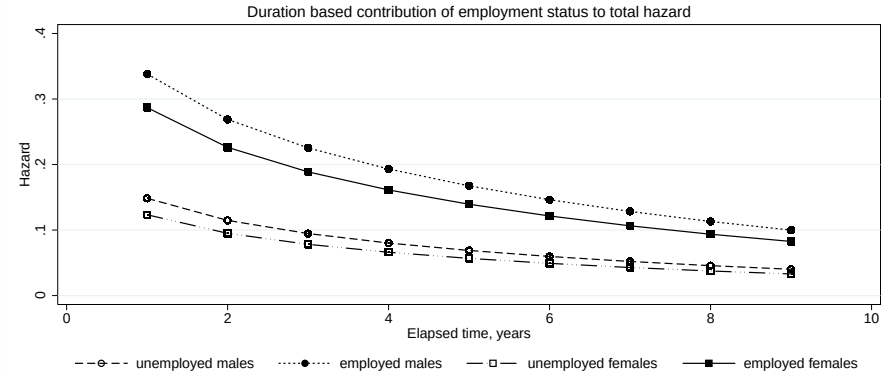

Figure 8: Comparison of distribution of hazard (probability to event) over the duration of study between employed and unemployed individuals and divided among male and female individuals.

large cities to acquire driver's license, compared to those living in towns or villages.

It is also worth noticing that parents income and access to own parents vehicles seems to have a great positive effect on probability of acquiring driver's license. Remainder of dependent variables have positive effect on the hazard (and corollary negative effect on survival) and shortens the duration of spells.

Conclusions

The aim of this research has been to develop a descriptive parametric model that could help us to understand individuals propensity to acquire driver's

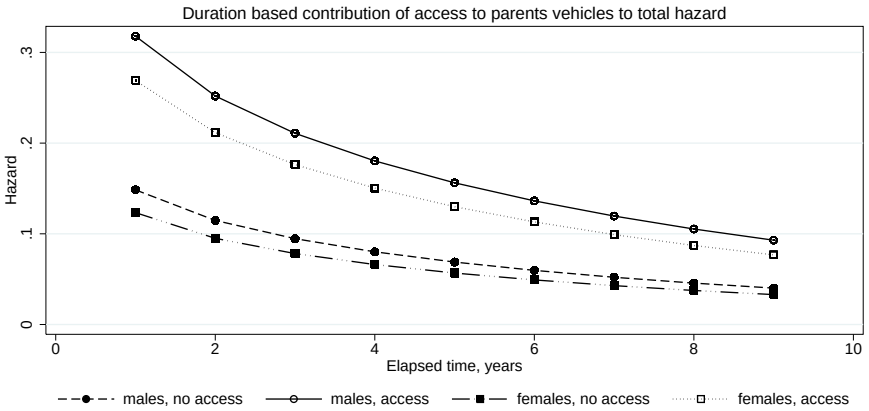

Figure 9: Comparison of distribution of hazard (probability to event) over the duration of study between individuals with and without access to parents vehicles divided by gender.

license from a behavioral perspective. We also sought to incorporate longitudinal trends and to identify potential time-dependent factors/variables, in order to detect and understand their time-dynamic nature.

In this research, we managed to construct a so called *complementary log-log model*, employing the discrete time survival modeling framework. We have been able to to show that, despite the fact that discrete time modeling domain is far more restricted than the one for continues time analysis, it is possible to compose attractive dynamic models that can thoroughly emulate the underlying dynamic processes of interest.

Comparison of estimated *log-likelihood* values as well as their corresponding Akaike and Bayesian information criterions suggests that the model with it's baseline hazard described as log-transformed time, performs best in case of our data.

Not surprisingly, variables age and being student decreases the probability of acquiring license but what which is a surprise is that male individuals are assigned higher hazard than females during the length of the study and seemingly with steady rate. This is contradicting the earlier hypothesis by Delbosc & Graham (2014) or McDonald & Trowbridge (2009) that growth in number of new licenses is induced by growing number of female drivers or we can that our study among Swedish individuals can not confirm their hypothesis.

Our research, however verifies Salonen's hypothesis,Salonen (2003) that many young individuals are highly dependent on their parents which was illustrated through independent variables *parents income* and *vehicle access*.

Estimated coefficient of the population variable is negative, which means that it is less likely for individuals who live in large cities to acquire driver's license, compared to those living in towns and villages, which was expected.

References

Beck, N., Katz, J. N. & Tucker, R. (1998), 'Taking time seriously: Time-series-cross-section analysis with a binary dependent variable', *American Journal of Political Science* **42**(4), pp. 1260–1288.

Bennett, D. S. (1999), 'Parametric models, duration dependence, and time-varying data revisited', *American Journal of Political Science* **43**(1), pp. 256–270.

Cain, K. C., Harlow, S. D., Little, R. J., Nan, B., Yosef, M., Taffe, J. R. & Elliott, M. R. (2011), 'Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes', *American Journal of Epidemiology* **173**(9), 1078–1084.

Carter, D. B. & Signorino, C. S. (2010), 'Back to the future: Modeling time dependence in binary data', *Political Analysis* **18**(3), 271–292.

Carter, D. B. & Signorino, C. S. (2013), 'Good times, bad times: Left censoring in grouped binary duration data'.

Cascetta, E. (2001), *Transportation Supply Models*, Vol. 49 of *Applied Optimization*, Springer US, pp. 23–94.

Cedersund, H. Á. & Henriksson, P. (2006), 'En modell för att prognostiser ungdomar körkortstagande', *VTI-Publications, VTI-Report* (511).

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society. Series B (Methodological)* **34**(2), 187–220.

de Jong, G. (1996), 'A disaggregate model system of vehicle holding duration, type choice and use', *Transportation Research Part B: Methodological* **30**(4), 263–276.

de Jong, G. & Kitamura, R. (2009), 'A review of household dynamic vehicle ownership models: holdings models versus transactions models', *Transportation* **36**(6), 733–743.

Delbosc, A. & Graham, C. (2014), 'Changing demographics and young adult driver license decline in melbourne, australia (19942009)', *Transportation* **41**(3), 529–542.

Gilbert, C. C. (1992), 'A duration model of automobile ownership', *Transportation Research Part B: Methodological* **26**(2), 97 – 114.

Guo, G. (1993), 'Event-history analysis for left-truncated data', *Sociological Methodology* **23**, pp. 217–243.

Hensher, D. A. & Mannering, F. L. (1994), 'Hazardbased duration models and their application to transport analysis', *Transport Reviews* **14**(1), 63–82.

Kalbfleisch, J. & Prentice, R. (1982), 'The statistical analysis of failure time data', *Canadian Journal of Statistics* **10**(1), 64–66.

McDonald, N. & Trowbridge, M. (2009), 'Does the built environment affect when american teens become drivers? evidence from the 2001 national household travel survey', *Journal of Safety Research* **40**(3), 177 – 183.

Rashidi, T., Mohammadian, A. & Koppelman, F. (2011), 'Modeling interdependencies between vehicle transaction, residential relocation and job change', *Transportation* **38**(6), 909–932.

Salonen, T. (2003), 'Ungas ekonomi och etablering; en studie om förändrade vilkor från 1970-talet till 2000-talet inledning', *Ungdomsstyrelsens skrifter* (9).

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

Train, K. (1986), MIT Press, Series in Transportation Studies, Cambridge, Massachusetts, pp. 733–743.

Tuinenga, J. G. & Pieters, M. (2006), Antonin: Updating and comparing a transport modelfor the paris region, presented at the European Transport Conference, Strasbourg.

Yamaguchi, K. (1990), 'Logit and multinomial logit models for discrete-time event-history analysis: a causal analysis of interdependent discretestate processes', *Quality and Quantity* **24**(3), 323–341.

Yamamoto, T., Kitamura, R. & Kimura, S. (1999), 'TRB paper number: 990810 Competing Risks Duration Model of Household Vehicle Transactions with Indicators of Changes in Explanatory Variables'.