Survival Analysis and Homeownership - Determinants of the Transition Time into Homeownership

Alice Drube, Konstantin Goebler, Chris Kolb, Richard v. Maydell

August 12, 2018

Seminar Thesis

Statistical Programming Languages

Humboldt Universität zu Berlin

Ladislaus von Bortkiewicz Chair of Statistics

Student IDs: 592792/592286/592380/591378

Examiner: Petra Burdejova

Contents

List of Figures

List of Tables

1   Introduction

The present study applies survival analysis methods to German longitudinal data in order to examine the determinants of the transition time from renting to first-time homeownership.

The promotion of homeownership has long received a great deal of attention from policymakers because of its apparent social and financial benefits (Di and Liu 2007; Rohe and Lindblad 2013). Owning a home can generally be considered an indicator of long-term economic well-being and increased control over the living situation. It may also reflect status attainment and it is an important means of accumulating wealth and transferring it to the next generation. These pivotal roles justify homeownership as an important area of interdisciplinary study.

In the case of Germany, Grabka and Goebel (2018) recently found in a DIW study that the risk of poverty increases sharply for renters compared to homeowners. Over the last 25 years, the risk of poverty for homeowners in Germany did not change substantially. On the other hand, the risk of poverty for renters increased almost twofold from 16% in the early 1990's to roughly 30%, affecting mostly young adults in their early thirties. At the same time, Sagna and Voigtländer (2018) found that the number of first owners is declining since 2016, suggesting that younger individuals are particularly affected. This motivates us to investigate which factors facilitate or interfere with the transition to a socio-economically more secure demographic, namely homeowners. Self-evidently, those findings have important implications for both policymakers and researchers.

Transitioning from renting to owning appears to be a complex process, demanding for a deeper investigation of potential determinants. Although a large strand of literature is concerned with homeownership and its policy implications, only few studies use longitudinal data to observe and investigate the progression into homeownership.

Economics primarily focus on the importance of income, wealth, race, and education in the transition (Boehm and Schlottmann 2004; Di and Liu 2007; King 1980), whereas another, somewhat less prominent branch studies low-income individuals facing credit constraints (Linneman and Wachter 1989). Other social sciences emphasize the role of social dynamics in tenure choice. Grinstein-Weiss et al. (2011) and Fisher and Gervais (2011) for example find that marriage is a key trigger for transitioning into homeownership, while the results of Fischer and Khorunzhina (2015) suggest that divorce has a detrimental effect on homeownership rates. Lastly, Kauppinen, Skifter, and Hedman (2015) find that immigrants transition at a considerably lower rate than native inhabitants.

All these findings guide us in the selection of relevant covariates. In our analysis, we ex-

amine the impact of household income, educational attainment, migration background, marriage/divorce, and spatial factors on the transition time from renting to homeownership.

In our report, we first introduce the basic survival concepts needed to understand our empirical application in Section 2. Then, we describe which data we use and how we created our final data set from the raw files in Section 3. Descriptives, summary statistics and data visualizations further provide a better overview over the data. Section 4 comprises the main part of our report: Most comprehensively, we describe and interpret the results from our non-parametric, semi-parametric and fully parametric approaches, before going on to compare these models. To better illustrate the programming issue at hand, we include the most important lines of code in our text. Lastly, Section 5 provides a brief summary of the main results.

2 Review of Survival Analysis Concepts

2.1 Basic Concepts

Survival analysis is widely understood as a set of models where the outcome variable is the time that elapses until some event of interest occurs. Due to its application in a large number of disciplines, the response is also known as survival time (biostatistics), duration (econometrics), or failure time (industrial statistics). The event could be death, marriage, divorce, or the duration of a strike.

Why don't we just use linear regression and model the expected survival time as a linear function of the predictors? There are numerous reasons for preferring survival models over standard methods, such as ordinary least squares estimation (OLS) or logistic regression. First, the survival time is restricted to be non-negative and heavily right-skewed. In order to adequately apply OLS in this setting, the survival times would first need to be transformed to remove this restriction. Second, the probability to survive longer than some arbitrary point in time is often of more interest than merely the expected survival time. Third, the hazard function incorporated in survival models may provide deeper insight into the failure mechanism compared to OLS coefficients.

Lastly, and most importantly, standard models such as OLS and logistic regression can not account for **censoring**, thus rendering these procedures inconsistent. Censoring is present when some survival times are only incompletely observed. The most common type of censoring encountered in applications is right-censoring, i.e. observing survival from the start, but only for some time, without an event occurring. This could be because subjects did not

experience an event until the study was ended, or because they were lost to follow-up during the survey period. Hence, censoring is a specific kind of missing data problem that frequently occurs in time-to-event studies.

Survival models overcome these problems by explicitly incorporating information from both censored and uncensored observations in model parameter estimation. Typically, survival models still need to assume uninformative censoring[1], i.e. that the censoring mechanism is independent of the failure mechanism. The dependent variable in survival analysis is defined by the pair (T, δ), where $T \geq 0$ denotes observed survival time, and $\delta \in \{0, 1\}$ denotes the status/failure indicator, i.e. whether an event has occurred.

There are several equivalent ways to characterize the distribution of a survival random variable. Additionally to the density function, $f(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t+\Delta t)}{\Delta t}$, and the cumulative distribution function $F(t) = \mathbb{P}(T \leq t) = \int_0^t f(u)\,du$, the **survival function** plays a central role in survival analysis. It is defined as the probability of surviving past time t, i.e.

$$S(t) := \mathbb{P}(T > t) = 1 - F(t) = \int_t^\infty f(u)\,du \quad . \tag{1}$$

The **hazard function**, given by

$$\lambda(t) := \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \geq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} \quad , \tag{2}$$

is usually referred to as the instantaneous rate at which events arrive, conditional on no prior event having occurred. The hazard function can be obtained from the survival function and *vice versa* via:

$$S(t) = \exp\left(-\int_0^t \lambda(u)\,du\right) = \exp\left(-\Lambda(t)\right) \quad , \tag{3}$$

where $\Lambda(t) = \int_0^t \lambda(u)\,du = -\log(S(t))$ is called the cumulative hazard function.

Other quantities of interest are usually derived from those two functions, for example the expected survival time $\mu = \mathbb{E}[T]$, or quantiles such as median survival.

Generally, it is of interest to estimate the effect of one, or several covariates such as age, gender, or income on the time to event. A plethora of models are available to analyze the impact of a set of covariates on survival time. These can broadly be categorized as non-parametric, semi-parametric and fully parametric approaches.

---

[1]This assumption, though very often violated in medical applications does not impose any problems in the context of survey data, as panel mortality tends to be unrelated to the specific question at hand.

2.2 Parametric Methods

Parametric estimators assume that the distribution of survival times follows a certain, *a priori* known, probability distribution. Distributions popular in survival analysis include the Weibull, Exponential or Log-normal distribution. Parametric survival models can generally be considered as either belonging to the class of *Accelerated Failure Time* (AFT) models or *Proportional Hazard* (PH) models (Kleinbaum and Klein 2010). In PH models, the covariates' effect is to increase the hazard multiplicatively by some constant, whereas AFT models assume covariates to accelerate (or decelerate) the survival time by some constant. In this paragraph, only AFT models will be described, as the most important PH models (Weibull and Exponential) can also be re-parametrized as AFT models. Moreover, the PH assumption is delineated in more detail later in subsection 2.4. AFT models typically are of the form

$$\log(T_i) = \underbrace{x_i'\beta}_{=\mu} + \sigma\epsilon_i \quad , \tag{4}$$

where β is the coefficient vector, x_i a vector of covariates, ϵ_i the additive error and σ a parameter that scales the error. The survival time T has been logged because their distributions tend to be right-skewed. The model then corresponds to

$$T_i = \exp\left(x_i'\beta\right) \times u_i^{\sigma} \quad , \tag{5}$$

where $\epsilon_i = \log(u_i)$ defines the multiplicative error u_i. For example, in the generic case where $\sigma = 1$, specifying $\epsilon_i \sim \mathcal{N}(\cdot, \cdot)$ will result in survival times that are distributed as log-normal. Conversely, if u_i is Weibull distributed, ϵ_i follows an extreme value distribution.[2] The model coefficients β_k have a simple interpretation: if $\beta_k > 0$, exposure to the k-th covariate benefits survival (i.e. lowers the probability of transitioning into home ownership), whereas if $\beta_k < 0$, exposure harms survival.

Model parameters are usually estimated using Maximum-Likelihood, a procedure that easily accommodates the inclusion of exogenous variables in the model, as well as effectively accounts for censoring in the Likelihood function. The respective Likelihood for parametric survival models share a common functional form. Recall, that $\delta_i = 1$ means that an event occurred for individual i, while $\delta_i = 0$ means that no event occurred, i.e. the individual is censored. Then, given some data $(T_i, \delta_i)_{i=1,\dots,n}$, the adequate Likelihood is given by:

$$\mathcal{L}(\theta|T,\delta) = \prod_{i:\delta_i=1} f(T_i) \prod_{i:\delta_i=0} S(T_i) \tag{6}$$

However, parametric models operate under heavily restrictive assumptions, casting doubt on their consistency if the chosen probability distribution is not plausibly justified. Yet, if

---

[2]The same holds for the Exponential distribution, which is simply a special case of the Weibull distribution.

the precise survival mechanism is known to the researcher, parametric models allow for vastly more precise characterizations of the survival distribution. This is particularly important because parametric approaches permit the estimation of the baseline hazard, a term later introduced in subsection 2.4. Furthermore, there are also less restrictive parametric approaches, for example flexible parametric proportional hazard models (Royston and Parmar 2002), which utilize restricted cubic splines to provide a better approximation to the true survival and hazard functions, relative to classical parametric distributional choices.

2.3   Non-Parametric Methods

Non-parametric estimators avoid these distributional assumptions altogether and operate by exclusively using information obtained from the data itself. The most popular non-parametric estimator of survival probabilities is the Kaplan-Meier (product limit) estimator (Kaplan and Meier 1958), closely related to its "cousin", the Fleming-Harrington or Nelson-Aalen estimator (Fleming and Harrington 1979; Nelson 1969, 1972). Both estimators use observed ratios of number of events and number of people at risk to derive an estimate of the survival probability, while taking censoring into account[3]. The Kaplan-Meier estimator is given by

$$\hat{S}_{KM}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \ , \tag{7}$$

where t_i are the distinct event times, d_i denotes the number of events at time t_i, and n_i the number of individuals at risk at time t_i.

As introduced in Equation 3, the cumulative hazard $\Lambda(t) = \int_0^t \lambda(u)\,\mathrm{d}u = -\log(S(t))$ can also be directly estimated by an approach originally developed by Nelson (1969, 1972):

$$\hat{\Lambda}(t) = \sum_{i:t_i \leq t} \frac{d_i}{n_i} \tag{8}$$

where d_i and n_i are as above. Obviously, the Nelson-Aalen estimator is an increasing right-continuous step function with increments d_i/n_i at the observed failure times.
Due to the relation $\Lambda(t) = -\log(S(t))$, a natural estimator developed by Fleming and Harrington (1979) is given by:

$$\hat{S}_{FH}(t) = \exp\left(-\hat{\Lambda}(t)\right) \tag{9}$$

Even though the Kaplan-Meier (KM) and the Fleming-Harrington (FH) estimators are asymp-

---

[3]Without the presence of censoring, the Kaplan-Meier estimator is simply $\hat{S}(t) = 1 - \hat{F}(t)$, where $\hat{F}(t)$ is the empirical CDF.

totically equivalent, they may differ in small samples[4] and thus a comparison can later prove to be worthwhile.

The KM and FH estimators are often used to compute simple univariate statistics, such as median survival or quantiles. Nevertheless, they can also be used to compare the estimated survival curves of two or more different groups. Examining differences in survival curves between groups usually requires some kind of testing procedure. The most popular of which is the so-called log-rank or Mantel-Haenszel test (Kuritz, Landis, and Koch [1988]), although a myriad of modifications exist. In applications, this type of χ^2-test is used to determine whether there are significant differences in survival curves between two or more groups (e.g. treated vs. non-treated). Assuming the two groups to be compared are denoted as A and B, the (weighted) log-rank test statistic is given by

$$\text{LogRank} = \frac{U^2}{\mathbb{V}(U)} \sim \chi^2_{(1)} \quad , \tag{10}$$

where $U = \sum_{t_i} w_{t_i} \left(o^A_{t_i} - e^A_{t_i} \right)$ is the sum of the weighted differences between observed and expected (under the Null) event counts at time t_i, and $\mathbb{V}(U)$ its variance. Groups A and B of course are interchangeable and yield the same statistic.

While the standard log-rank test specifies $w_{t_i} = 1 \; \forall \, t_i$, it might be worthwhile to modify the weighting structure with respect to the application at hand. In our analysis, we use modified Peto-Peto weights that calibrate the log-rank test to be more sensitive regarding early differences in survival times, and improve robustness in the presence of high censoring rates (see Arboretti et al. [2017]), which is particularly indicated in our data.[5]

2.4 Semi-Parametric Methods

As the name suggests, semi-parametric estimators make less assumptions than fully parametric models, but more assumptions than the non-parametric methods discussed above. By far, the most widely disseminated model from this class is the Cox proportional hazard (PH) model (Cox [1972]). In contrast to the Kaplan-Meier estimator, its focus is not on the survival curve, but on the hazard function. Additionally, it allows for the incorporation of covariates that shift the hazard proportionally. In particular, it assumes that the hazard function can be apportioned into a part that does not depend on covariates (the so-called baseline hazard), and a part that does.

The baseline hazard is the hazard an individual faces if all covariates are set to zero (akin

---

[4]Colosimo et al. ([2002]) preform Monte Carlo simulations in order to establish which of the two performs better under various settings.

[5]The modified Peto-Peto weights are given by $w_{t_i} = \frac{n_{t_i}}{1+n_{t_i}} \prod_{t_i} (1 - e_{t_i})/(1 + n_{t_i})$.

to an intercept in linear regression). The covariates are assumed to have a multiplicative effect on the baseline hazard. The model's semi-parametric nature stems from the fact that it is possible to consistently estimate the effect of covariates on the hazard function, without ever needing to specify the baseline hazard, via partial Maximum-Likelihood. Consequently, the Cox PH model is particularly popular because it provides easy to interpret information without making many assumptions. The Cox model specifies the hazard function as follows:

$$\lambda(t|x_i) = \lambda_0(t) \exp\left(x_i'\beta\right) \quad , \tag{11}$$

where $\lambda_0(t)$ is the baseline hazard, and x_i a vector of time-invariant characteristics, such as age at study enrollment, or gender. It is immediately obvious that the Hazard Ratio (HR), i.e. the ratio of two hazards, is constant over time and given by $\exp(\beta_k)$ if the k-th predictor is increased by one unit. This assumption is called the **proportional hazards assumption**, and it is a crucial part of every Cox PH regression to check whether this assumption is met.

The methods discussed above cover the most common methods for the analysis of time-to-event data in the presence of right-censoring.

3   Data Cleaning and Descriptives

The empirical results in this paper are based on micro-level data from the German Socio-Economic Panel (SOEP), a panel study that follows over 20,000 individuals in 12,000 households and features a rich set of items on socio-economic and living conditions, as well as attitude studies. The survey was started in 1984 in West Germany and 1990 in East Germany, and is conducted annually by the German Institute for Economic Research (DIW Berlin). In particular, the SOEP contains comprehensive information on thematically relevant characteristics, such as income, education, households' tenure choice (renting vs. owning), and a wide choice of demographic indicators.

Reasonably enough, we only want to include individuals who are relevant to our research topic, namely those who are capable of acquiring a dwelling. Thus, we restrict our investigation to individuals aged between 25 and 65[6]. Unfortunately, information on tenure choice is only available at the household level, our analysis however is located at the individual level[7]. Thus, some choice of disaggregation is required. Consequently, we only consider household heads, and match the household's rental status to their individual characteristics. In addition, we uniquely want to observe one individual per household tenure

---

[6]It is of course perfectly possible for individuals to progress into homeownership past the age of 65. However, it is save to assume that the passive population is subject to deviating transitioning factors.

[7]It would be entirely fruitless to track households over time as it is impossible to decide whose characteristics to consider.

choice, as otherwise it is not immediately clear whose characteristics we should match to the household-level variables. Therefore, we remove households that have dissolved, i.e. where household heads have changed over the observation period.

Since we are interested in the determinants of the transition time from renting to first-time homeownership, we further exclude individuals who acquired a residence through inheritance or endowment. In essence, our final data set includes all individuals surveyed by the SOEP who start out as renters at age 25 and are observed until they either acquire a dwelling or drop out of the study.

3.1   Implementation in R

After some crucial descriptives we present our data-cleaning strategy. This is then followed by a sufficiently detailed account on how we generate our final survival analysis data set from the raw `.csv` files provided by the SOEP.

Generally, we use `.RDA` files that allow us to save and reload our R data structures. Mainly, we process our datasets using `data.frame()` and `data.table()` objects. We choose this approach due to the objects' easy handling and to efficiently subset the data, select and compute on rows and columns, and perform aggregation by group. Most basic data manipulations are done with packages from the `tidyverse` collection, especially `dplyr` and `tidyr`.

Since the data cleaning process - to arrive from the raw `.csv` files at our final `datfinal.RDA` survival data set (that includes a data frame called `dat`) - was tedious and cumbersome, we choose to spare the reader of the repetitive and hard-to-follow lines of code. Rather, we confine to illustrate the general procedure by which we created our data set and highlight some stylized programming routines, such that the main part of the report is left to the actual analysis.

The SOEP provides the user with a number of raw data files, from which we have assembled our final data set. The raw files we use are `pequiv.csv`, `pgen.csv` and `pffadl.csv` for individual-level characteristics, and `hgen.csv` and `hbrutto.csv` for household-level information. We then subset the relevant variables in each data set and later merge them into one file containing all relevant information.

In the next step, we perform basic data cleaning on the long-format merged data set, which includes, *inter alia*:

- excluding dissolved households, i.e. households where the household head did not remain the same throughout the observation period

- excluding households who acquired homeownership through inheritance or endowment

- excluding individuals who are not observed from age 25 onward (e.g. 30 to 46)

- excluding individuals who already started out as homeowners at age 25

In essence, we are now left with 25 year old renters, who we observe until they transition or drop out of the study (right-censoring). The next paragraphs outline the steps necessary to extract the information on survival times and censor status that we seek.

Most crucially, we begin with the integer variable `hgowner` from `hgen.csv`. The variable indicates the ownership status of a household, i.e. whether the current habitation is owned or rented.[8]. Notably, the variable `hgowner` is essential to our investigation, as any transition over time is constructed from it. Practically speaking, we use the function `shift()` to compare the housing status with previous years, and infer from that whether a transition into homeownership has taken place.[9] From the specification of `hgwoner`, we first construct simplified dichotomous variables indicating rent or ownership. Then, we compare the lagged value for the dummy variable `rent` with current `owner` status. Subsequently, we construct a `change` variable by looking at any alterations between the current status of housing and its lagged value. Specifying our survival object, we have to define a failure (or event) variable. The indicator variable `event` flags all years where a transition from rent into ownership has taken place. In the spirit of survival analysis terminology, transitioning to homeownership is considered a failure, while remaining in rent means survival. Furthermore, we create a variable `firstfailyear` that indicates the year of transition, given that an event has occurred for the individual in question.

To further determine the first and last year of observation, we create the variables `firstyear` and `lastyear`. In doing so, we employ the generic function `aggregate`[10], which splits data frames into subsets and computes functions aggregated on group levels. In our case, we aggregate our data on the household identifier variable `hid`[11]. In order to find the first year of observation, we specify the function `min(x)` to be computed over all years. As a result, we receive a two-column `list` that combines the household identifier `hid` with the minimum `syear` per household. The new variable generated by `aggregate` is then merged to our original data frame. Analogously, we do the same for last year of observation with `max(x)`:

```
# create first year of observation variable
```

---

[8]Description of variable `hgowner`:
(-1) No Answer, (1) Owner, (2) Main Tenant, (3) Sub-Tenant, (4) Tenant, (5) Living In A Home

[9]We consequently `group by` the variable `hid`.

[10]The function is included in the R `stats` package.

[11]We could theoretically also use the person identifier `pid`, since we matched one household to exactly one household head.

```r
fyears <- aggregate(syear ~ hid, data, function(x) min(x))
names(fyears)[names(fyears) == "syear"] <- "firstyear"
data = left_join(data, fyears, by = "hid")
rm(fyears)

# create last year of observation variable
lyears <- aggregate(syear ~ hid, data, function(x) max(x))
names(lyears)[names(lyears) == "syear"] <- "lastyear"
data = left_join(data, lyears, by = "hid")
rm(lyears)
```

After creating these variables, we can now distinguish whether a transition from rent to homeownership has taken place. In case of multiple events, we choose not to consider all potential subsequent ones after the first as we are primarily interested in the time interval up to first event and repeat buyers are likely subject to different relocation restrictions (Guiso and Jappelli 2002).

In order to achieve this, we apply the `data.table` command .SD[1] to generate the variable `firstfailyear`. The .SD[1] option flags all rows where the first transition from rent into ownership[12] takes place.

```r
# subset observations to include only those before and up to first
    event per hid
# example: 0 0 0 0 1 1 1 => 0 0 0 0 1 and 0 0 1 0 1 => 0 0 1
# flag syear at first event for each household where failure occurs

first.fail <- data[failure == 1, .SD[1], by = hid]
first.fail <- first.fail[, c("hid", "syear")]
names(first.fail)[names(first.fail) == "syear"] <- "firstfailyear"
data = left_join(data, first.fail, by = "hid")
rm(first.fail)
```

Now that we have created the `firstfailyear` variable, we can conveniently exclude all unwanted observations after the first event.

However, simply using `subset(data, syear<=firstfailyear)` does not yield the desired results, since all censored observations take on `NA` values in the variable `firstfailyear` and would consequently be excluded. Hence, the exclusion is implemented by the following somewhat awkward lines of code:

```r
# subset: only include syear<=firstfailyear and censored units
data1 <- subset(data, syear <= firstfailyear)
data2 <- subset(data, is.na(firstfailyear))
data <- rbind(data1, data2)
rm(data1, data2)
```

---

[12]After the first time where the variable event takes on the value 1.

At this point, we have delineated the most important parts of the data cleaning process. Next to some final remarks on our data-cleaning procedure the consecutive subsections further describe how the survival object (T, δ) can be generated from the variables above.

3.2 Variables and Imputation

In the previous subsection, we describe how we manipulate and merge the raw .csv files provided by the SOEP in order to assemble our own data set. This subsection briefly explains how we create the time (i.e. transition time to homeownership) and event (i.e. whether or not an event actually occurred in the observation period) variables. Moreover, besides the survival pair, we use a number of covariates in our analysis, part of which required some transformations from their original specification in the SOEP. Lastly, we describe two approaches we take to impute missing data.

As explained in Section 2, the survival object consists of a pair (T, δ), where T is the survival time, and δ indicates whether an event (i.e. a transition) has taken place or the individual is being right-censored. Let us first create δ, the event variable. To do this, recall the change variable previously introduced, which indicates whether or not an individual changed its rental status, relative to the previous year. This variable only takes the value 1 if a change from renting to owning has occurred (and zero or some negative value otherwise). To create the event variable, we compute the maximum of the change variable for each respondent (our identifying variable here is hid, that is, household ID). Then, if the maximum per individual is 1, we conclude that an event has occurred. Further, we assign zero to the event variable if the computed maximum is less than 1.

```
# mark all units where failure occurs at some point
failure.dat <- aggregate(change ~ hid, data, function(x) max(x))
names(failure.dat)[names(failure.dat) == "change"] <- "event"
data = left_join(data, failure.dat, by = "hid")
data$event[data$event != 1] <- 0
rm(failure.dat)
```

Next, we create the time (Time to Event) variable, i.e. the time that elapses starting from the first year of observation (at age 25) until either an event has occurred or the individual is right-censored. We implement this using an if-else statement on the previously introduced firstyear, lastyear, firstfailyear and event variables.

```
# create time to event variable
data <- mutate(data,
                time = ifelse(data$event==1,
                          data$firstfailyear-data$firstyear+1,
                          data$lastyear - data$firstyear +1))
```

11

If an event has occurred, `time` takes the value `firstfailyear-firstyear+1`, and if not `lastyear-firstyear+1`, which is simply the period of observation.

However, we do not simply want to work with (T, δ), but we also want to incorporate covariates in our models to investigate their effect on survival time. Keep in mind that most survival models in their standard format only take time-invariant covariates as input. If the variable we want to include is already constant, then no further steps are needed. However, for time-dependent covariates we choose to take their value at age 25. Utilizing only one, namely the first realization per individual and covariate leaves us with some missing values – additionally to the values that were already missing beforehand.

To overcome this obstacle, we take two different approaches to impute missing values. Figure 1 depicts the missing value pattern by covariate and survey year, demonstrating that most covariates do not have missings in the first place. The variable `firstfailyear` has a high percentage of missings by design, as many subjects never experience "failure", i.e. transitioning into homeownership. Naturally, the fraction of missing values in this variable increases in time, as the proportion of incomplete survival times increases, and hence `firstfailyear` is assigned `NA` by the generating code.

Figure 1: Pattern of missing data before random forest imputation
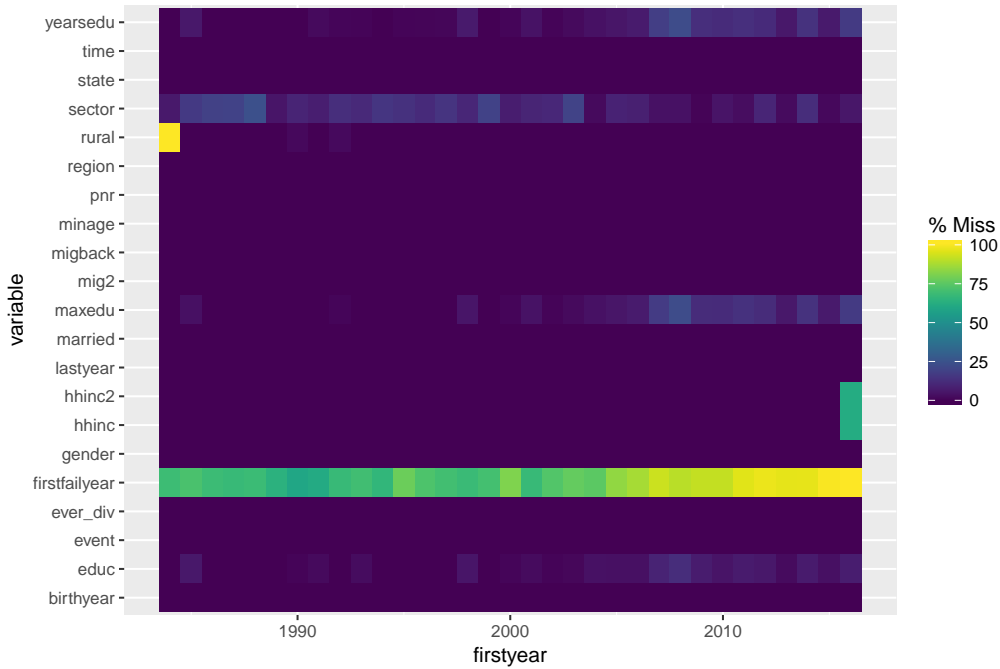
Our first approach is to logically impute the missing values from the following years. Hence, we check if the covariate is also missing at age 26 and 27, which provides us with a much higher imputation quality as compared to statistically more sophisticated methods. For better understanding, we present that procedure for the variable `i11101`, later renamed `hhinc`:

```
#### Impute pre-government HH income #####
# If income value is NA take value next year otherwise of year
    after

setDT(data)[, shiftincome:= lead(i11101), hid]
setDT(data)[, shift2income:= lead(shiftincome), hid]
data <- mutate(data, hhincimp =
        ifelse(is.na(data$i11101),
        ifelse(is.na(data$shiftincome),
        ifelse(is.na(data$shift2income), NA ,
              data$shift2income), data$shiftincome),data$i11101))
```

In the instance of `yearsedu`, i.e. years of education, we impute the variable with the maximum years of education over all observation periods, which is often deemed as a sufficiently adequate time-invariant measure of educational attainment.

Although this imputation approach is highly successful, it still leaves us with some missing observations where none of the subsequent observations appeared to valid. Handling the remaining ones, we choose to implement a modern, non-parametric Random Forest imputation method, outlined in Tang and Ishwaran (2017). This innovative procedure is implemented in R's `randomForestSRC` package, providing a convenient and flexible framework for fast imputation applications. This approach has the advantage that it can handle mixed types of data with satisfactory performance, even when data is missing not at random (MNAR). We specified on-the-fly imputation using unsupervised multivariate node splitting, however, as the procedure requires some explaining, we decide to omit the details for reasons of clarity.

3.3 Summary Statistics and Visualization

After having illustrated how we create our final data set, generate the relevant covariates and impute missing data, we present some summary statistics and visualizations to get a better understanding of the statistical issue at hand.

First, we generate a table that gives us a synoptic overview over our covariates. Implementing that, we make use of the `stargazer` package, a powerful tool for producing publication-quality LaTeX tables from R output.

```
sum.dat <- dat
sum.dat <- select(sum.dat,
                 one_of(c("time", "event", "hhinc", "maxedu",
                         "birthyear", "gender", "region", "rural",
                          "married", "ever_div", "migback")) )

stargazer(sum.dat, type="latex",
        summary.stat=c("mean", "sd", "min",
                         "median", "max", "n"),
```

```
        title = "Summary Statistics",
        covariate.labels = c("Time to Event","Event","HH Income",
                             "Educ. (years)", "Birthyear",
                             "Gender","EastGermany","Rural",
                             "Married", "Ever Divorced",
                             "Migr. Background"),
        digits=2,
        summary.logical = TRUE)
```

The result is Table 1. In total, we observe 3,928 individuals that fit our previously described requirements for study enrollment. It can be seen that the average transition time is 6 years, while 18% of individuals experience an event (progressing from renting to home-ownership). However, due to the high amount of right-censoring our data exhibits, these values can not be interpreted as expected survival time and event probability, respectively. They should only be understood as a descriptive statistic of the variable. What is more, we see that the longest observed time to event is 33 years, which stems from the fact that we observe individuals aged 25 or older from 1984 onwards.

Table 1: Summary Statistics

Statistic	Variable	Mean	St. Dev.	Min	Median	Max	N
Time to Event	time	6.13	5.93	1	4	33	3,928
Event	event	0.18	0.39	0	0	1	3,928
HH Income	hhinc	22.94	18.61	0	20.76	248.21	3,928
Educ. (years)	maxedu	12.27	2.70	0	11.5	18	3,928
Birthyear	birthyear	1978.5	9.79	1959	1980	1992	3,928
Gender	gender	0.54	0.50	0	1	1	3,928
East Germany	region	0.21	0.41	0	0	1	3,928
Rural	rural	0.27	0.45	0	0	1	3,928
Married	married	0.22	0.41	0	0	1	3,928
Ever Divorced	ever_div	0.05	0.21	0	0	1	3,928
Migr. Background	migback	0.31	0.46	0	0	1	3,928

The other variables included behave broadly as expected, with the exception of ever_div, whose mean appears to be oddly low. However, the portion of divorced individuals who did not previously live in a self-owned dwelling might actually be that small. Thus, inspection of the summary statistics do not indicate that our sample selection process has rendered the SOEP data unrepresentative.

While data can typically be well described by summary tables, survival data should also be visualized to get an intuitive understanding of its temporal nature. To do this, we create a plot with time on the x-axis, and subject IDs on the y-axis. Each horizontal line then represents an individual, from age 25 until they stop being observed. A cross at the end of a

line indicates that an event has occurred, while a circle indicates that the individual is censored, i.e. dropped out of the study. Since plotting all 3,928 observations would overcrowd the plot with lines, circles and crosses, we opt to visualize 500 randomly drawn individuals with the following lines of `ggplot` code:

```
dat.str <- dat
dat.str$pnr <- sample(1:nrow(dat), nrow(dat), replace=F)
dat.str1 <- subset(dat.str, pnr<500)

ggplot(dat.str1, aes(x = pnr, y = time)) +
  geom_linerange(aes(ymin = 0, ymax = time), size=0.5) +
  geom_point(aes(shape=as.factor(event),color=as.factor(event)),
  stroke = 1.3, cex = 2) +
  scale_shape_manual(values = c(1,4)) +
  labs(color="Event") +
  guides(shape=F) +
  labs(y = "Time (years)", x = "Subject ID") + coord_flip() +
  theme_classic()
```

The visualization of survival times for 500 randomly drawn individuals is depicted in Figure 2. Since we observe only discrete time values (years), the circles and crosses are aligned vertically at the corresponding integers. Looking at the plot, the large amount of censoring becomes apparent, particularly during the first few years. This graphical representation of our data underscores the absolutely vital need for statistical methods that account for censoring regarding our research topic.
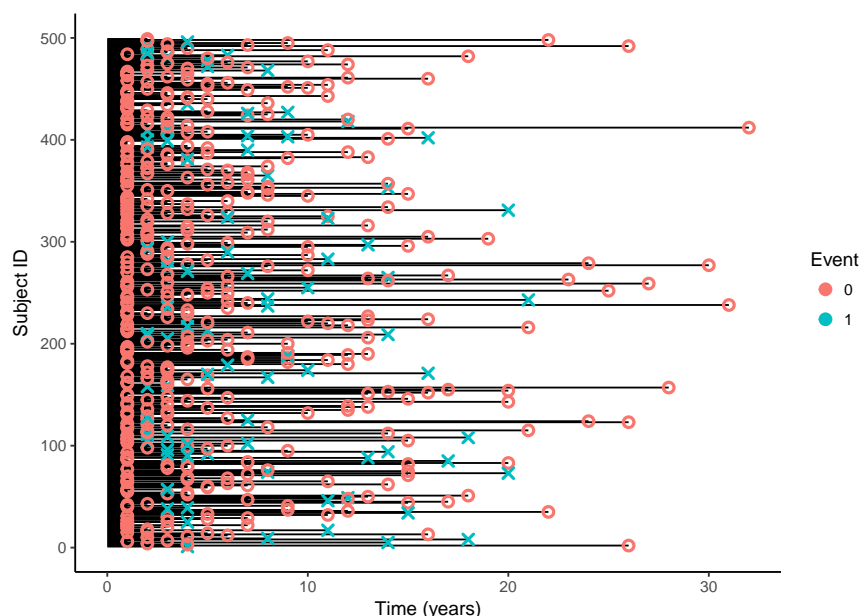
Figure 2: Visualization of SOEP survival data. For better visibility, only 500 randomly drawn individuals are visualized ($n = 3,928$). DataVisualization

Next, we inspect the distribution of survival times more closely. We do this by creating a

histogram of completed survival times, with the corresponding estimated density on top:

```
dens.dat <- subset(dat, event==1)


hist(dens.dat$time, # histogram
     col="Dodger Blue", # column color
     border="black",
     prob = TRUE, # show densities instead of frequencies
     xlab = "Time",
     ylab = "Relative Frequency",
     xlim = c(0, 30),
     ylim = c(0, 0.28),
     breaks=30,
     main = "Survival Time Density")
lines(density(dens.dat$time, from = 0, to = 30), # density plot
      lwd = 2, # thickness of line
      col = "red")
```

The resulting plot can be examined in Figure 3. As expected, the distribution seems heavily right-skewed, with its modus already obtained after some 5 years. Yet, we must interpret this plot with caution, due to the previously mentioned censoring bias. Thus, we can hypothesize that the true distribution of survival times is shifted to the right, as sample attrition becomes more likely the longer we observe individuals.

Figure 3: Density of completed survival times

With regard to the spatial dimension we are interested whether the distribution of survival times differs among Federal States in Germany. For example, we expect the distribution to be right-shifted for densely populated states, as price levels are higher and anecdotal

evidence tells us that individuals in urban areas typically acquire homeownership later in life. To investigate state differences, we produce as similar plot as Figure 3, only that this time, we stratify by state:
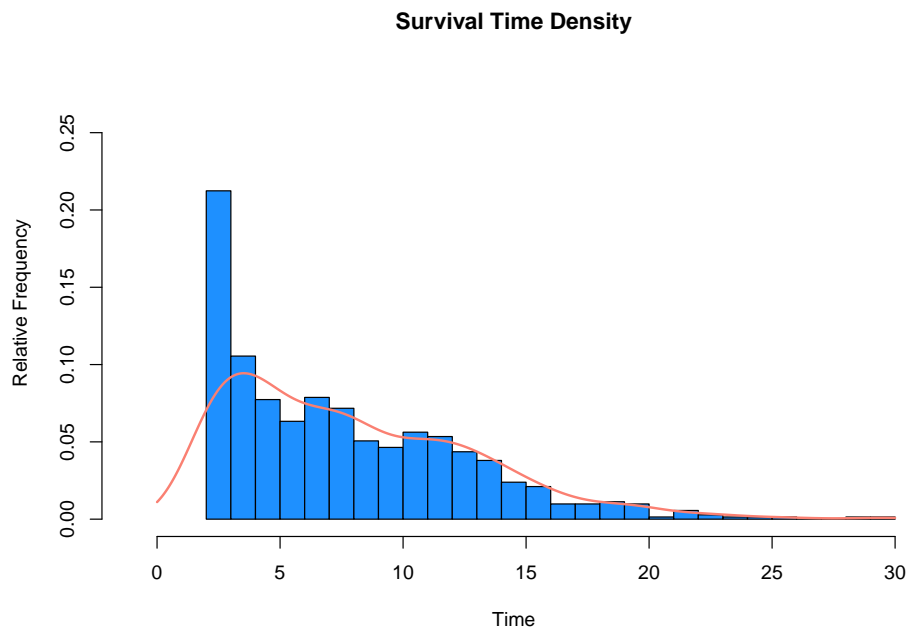
```
#### distribution of time to event by state ###
dist <- subset(dat, event==1)
levels(dist$state)

ggplot(
  dist,
  aes(x = dist$time, y = dist$state, fill=dist$state)) +
  geom_density_ridges_gradient(
    aes(fill = ..x..), scale = 2.5, size = 0.3) +
  scale_fill_gradientn(
    colours = c("#0D0887FF", "#CC4678FF", "#F0F921FF"),
    name = "Duration")+
  theme_ridges() +
  labs(title = 'Density of Time to Event',
       x = "Time", y = "States")
```

Figure 4 shows the produced graph. The general appearance of the state densities corresponds to the shape of the overall density seen in Figure 3. The vast majority of densities is heavily right-skewed, with their modus being achieved around 5-7 years. Interestingly, the distributions of Brandenburg and Bremen exhibit multiple pronounced peaks, and Saarland's distribution even seems left-skewed. Some of these differences can be ascribed to structural differences in transition times from renting to homeownership, while others can be attributed to data issues and few observations per state.

Next, after having described how we create our final data set datfinal.RDA with its associated data frame dat, impute missing values and present visualizations of the data, we move on to the results section. Here, we apply various inferential non-parametric, semiparametric and fully parametric methods that can account for right-censoring. Using these techniques, we hope to validly uncover the determinants of the transition time from renting to homeownership, without falling prey to the numerous pitfalls the data offers.
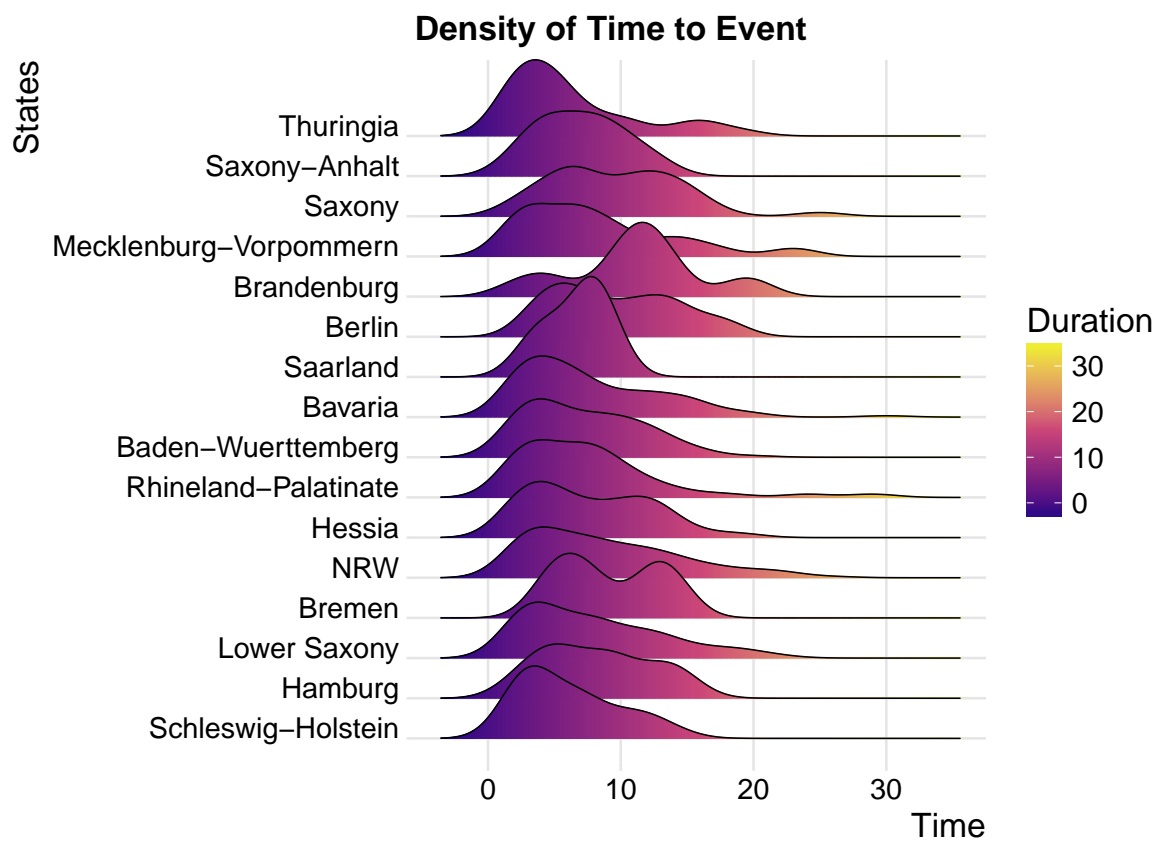
17

Figure 4: Density of Survival States by Federal State

SurvivalDens

4 Survival Analysis Results

After having presented the necessary descriptives, we engage into a comprehensive investigation of important factors regarding the transition from renting to homeownership.
As outlined in Section 1, surprisingly few attempts have been made to examine the longitudinal transition behavior from renting into homeownership. Therefore, our seminar paper neatly fits into that gap for the case of Germany.
The analysis follows a thoroughly structured path, initiating a non-parametric investigation of the transitioning times from renting into homeownership. Then we proceed to a semi-parametric design brushing upon covariate effects and somehow involved testing methods, until we lastly turn to fully parametrized survival models.

4.1 Application in R

Due to the wide range of scientific fields making use of survival analysis, its documentation in R is adequately comprehensive. Hence, the following analysis will predominantly make use of three packages: First and most importantly, the `survival` package (Therneau 2015), which comprises all core survival analysis routines carried out in the subsequent investigation.
Second, the `flexsurv` package (Jackson 2017), which allows us to fit flexible parametric survival models, including Royston and Parmar (2002)'s flexible spline model described in Section 2.
And third, the `survminer` package (Kassambara, Kosinski, and Biecek 2017), which permits us to generate flexible and visually appealing survival graphs.
In general, the dependent variable in survival model analyses is a so-called *survival object*, created by the `Surv()` function of the `survival` package. A survival object can be thought of as a matrix containing the (potentially right-censored) time-to-event variable in the first column and a binary variable indicating if an event has occurred in the second column. Thus, it is R's implementation of the survival pair (T, δ) described in subsection 3.2.
In practice, we define the survival object via the `Surv()` function as follows:

```
surv.obj <- Surv(time, event, type="right")
```

where time and event variables are as described in Section 3.
Broadly speaking, survival analysis is concerned with the distribution of survival times. With the methodology introduced in Section 2, we are able to estimate non-parametric, as well as semi-parametric, and fully parametric survival models.

Why would we be concerned with parameterizing our survival distribution when we have unconditional methods at hand? As clarified above, the goal is to determine transition-mechanism from renting into homeownership. Thus, we are interested in the relationship

between the *survival* in rent and several covariates introduced in Section 3. For this reason, we wish to estimate parameters with respect to the covariates of interest. Fully nonparametric approaches such as the Kaplan-Meier or the Nelson-Aalen estimator allow for a comparison of different subgroups, but not for a quantification of the respective effect sizes. Moreover, given that the survival distribution is defined correctly, parametric methods are more efficient compared to distribution-free methods.

4.2 Non-Parametric Estimators

We begin however, with a first glance at our survival distribution by fitting the Kaplan-Meier and Fleming-Harrington estimators. In R, this is easily done by using the function `survfit()`:

```
# Kaplan -Meier estimator
km.fit <- survfit(Surv(time,event, type="right") ~ 1, data=dat,
    type="kaplan -meier")
# Fleming -Harrington estimator
fh.fit <- survfit(Surv(time,event, type="right") ~ 1, data=dat,
    type="fleming -harrington")
```

Here, both estimators are chosen at their simplest, fitting as input our survival object against a symbolic 1, indicating that the whole sample is used without stratification. In addition, we refrain from making use of R's basic plotting function, but rather invest into an appealing visualization relying on the `survminer` package.

In order to customize the graphical tools of `survminer` to our needs, we write a function that fits the otherwise rather cumbersome syntax neatly in. Our function relies heavily on `ggsurvplot_combine()` which simply combines multiple `survfit` objects on the same plot:

```
#### Function for ggsurvplot_combine #####
# choose functional argument (default is Survival Function)

nonparametricKurves <- function(fun=NULL) {
   ggsurvplot_combine(kmfh.all, data=dat, conf.int=T,
                      legend.labs=c("KM", "Fleming -Harrington")
   , legend.title="Model",
                      fun=fun,
                      risk.table=F,
                      cumcensor=FALSE,
                      censor=FALSE,
                      linetype=c(1,1),
                      size = 0.3)
                                    }
```

When leaving all arguments as default, `nonparametricKurves()` produces both the Kaplan-Meier and the Fleming-Harrington estimators with their respective 95% confidence inter-

val. The result can be viewed in Figure 5's left-hand graph.
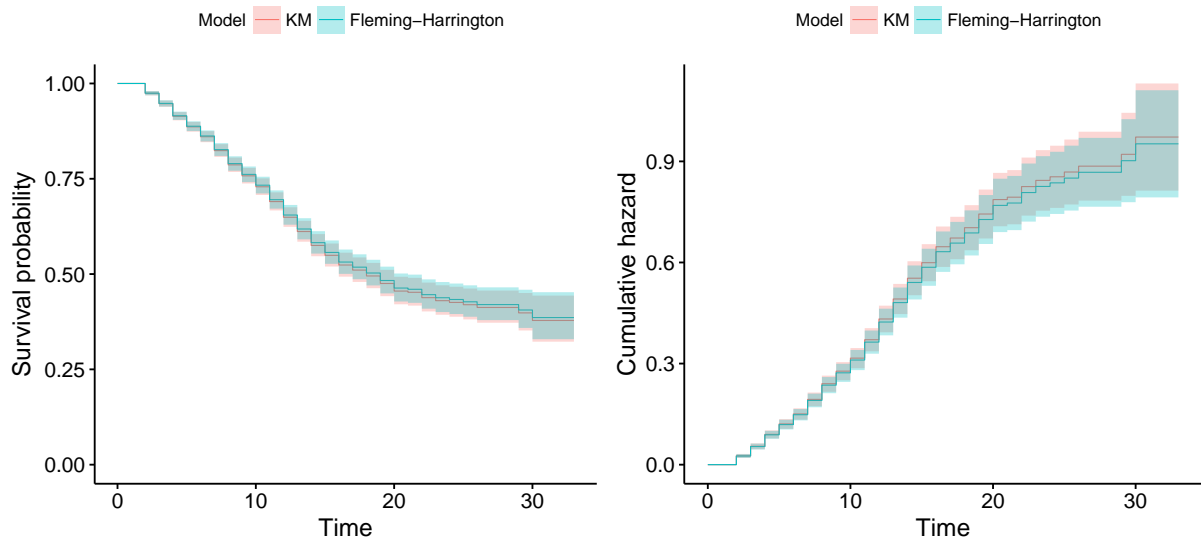
Figure 5: Comparison of KM and FH estimators of the survival function. Nonpara_Curves

As expected, the KM and FH estimators do not differ substantially since our sample size is relatively large. Looking at the survival function, one can clearly see that there is a steady transition into homeownership in the first 15 or so years. Recalling that we restricted our sample to individuals who were enrolled into the study at age 25, both KM and FH imply that roughly at age 40 (after 15 years) half of our population has transitioned into homeownership. After the full observational period there still remain roughly 37% of the population who have not (yet) progressed into homeownership.

Next, let us consider the cumulative hazard curve estimated by both KM and FH. By specifying `nonparametricKurves("cumhaz")`, we are able to plot the cumulative hazard function, which can be examined as the right-hand graph of Figure 5. It is a convenient feature of our function that we could use any functional argument in order to transform the survival function. Put differently, `nonparametricKurves(f(y)=-log(y))` would yield exactly the same result as our `"cumhaz"` specification above, due to the fact that we can derive the hazard function[13] from the survival function and vice versa, as already indicated in Equation 3. Finally, using the following code we can produce the graphs in Figure 5:

```
#Survival Function
surv.all <- nonparametricKurves()
#Cumulative Hazard Function
cumhaz.all <- nonparametricKurves("cumhaz")

# put all plots in one graph
kmfh.glist <- list(surv.all, cumhaz.all)
```

---

[13]The most common transformations of the survival functions are given intuitive names (see for more insight Kassambara, Kosinski, and Biecek 2017).

```
    arrange_ggsurvplots(kmfh.glist, print = TRUE, ncol = 2, nrow = 1)
```

As delineated above, we are not able to determine any effect sizes when fitting our non-parametric estimators. However, it is perfectly possible to compare time-invariant characteristics by comparing various strata of interest.[14]

Let us begin simply by investigating any differences in the survival function when separating our population in East and West Germans[15]

In R, we can now simply fit our survival object against the stratification variable of interest, in this case `region`:

```
### KM by region
wide.fit <- survfit(Surv(time, event, type="right") ~ region, data=
    dat)
```

Again we have written a function in order to customize producing our stratified survival graphs. For this, we once more make use of the `ggsurvplot()` function, specify the desired options, and apply it to the grouping variable of interest.

Implemented in code, the function takes the following form:

```
#### Function for Kaplan Meier Curves by Strata ####
###  wide.fit = survobject, labs= category description, Legend
    Title, umber of functions, confidence intervals
kmGroupKurves <- function(labs,title,line=c(1,1),conf=T){
  ggsurvplot(wide.fit, conf.int=conf,
                legend.labs=labs, legend.title=title,
                censor=F,
                palette = "strata",
                risk.table = T,
                pval=TRUE,
                risk.table.height=.25,
                ylim=c(0,1),
                xlim=c(0,30),
                surv.median.line="hv",
                linetype=line,
                size = 0.5)
                                                          }
```

Default settings include confidence intervals, the display of the risk table[16], a line depicting the median survival of the specified groups, and lastly the p-value of the modified Peto-Peto weighted log-rank test (Peto and Peto 1972) for differences in survival curves.

---

[14]One remark shall be made regarding gender differences. Despite the fact that theoretically there are large gender effects to be expected regarding *who* acquires the dwelling, we have household information at hand regarding the homeownership status. Since the male breadwinner model is still predominant in Germany (see Siermniska, Frick, and Grabka 2010), female and male individuals do not differ in regard to when they transition into homeownership.

[15]As is well researched, there are still large differences especially in past and present wealth accumulation between East and West Germans (see for e.g. Grabka 2014).

[16]The risk table provides an easy-to-follow synopsis of the number of individuals at risk over time, grouped by the specified strata.

Consequently, we are able to construct the stratified KM estimator for East and West Germans, which can be viewed in the left-hand graph in Figure 6:

```
#### KM by region (East/West)#####
wide.fit <- survfit(
                Surv(time, event, type="right") ~ region, data=dat)

km.reg <- kmGroupKurves(c("West", "East"), "Region")
```
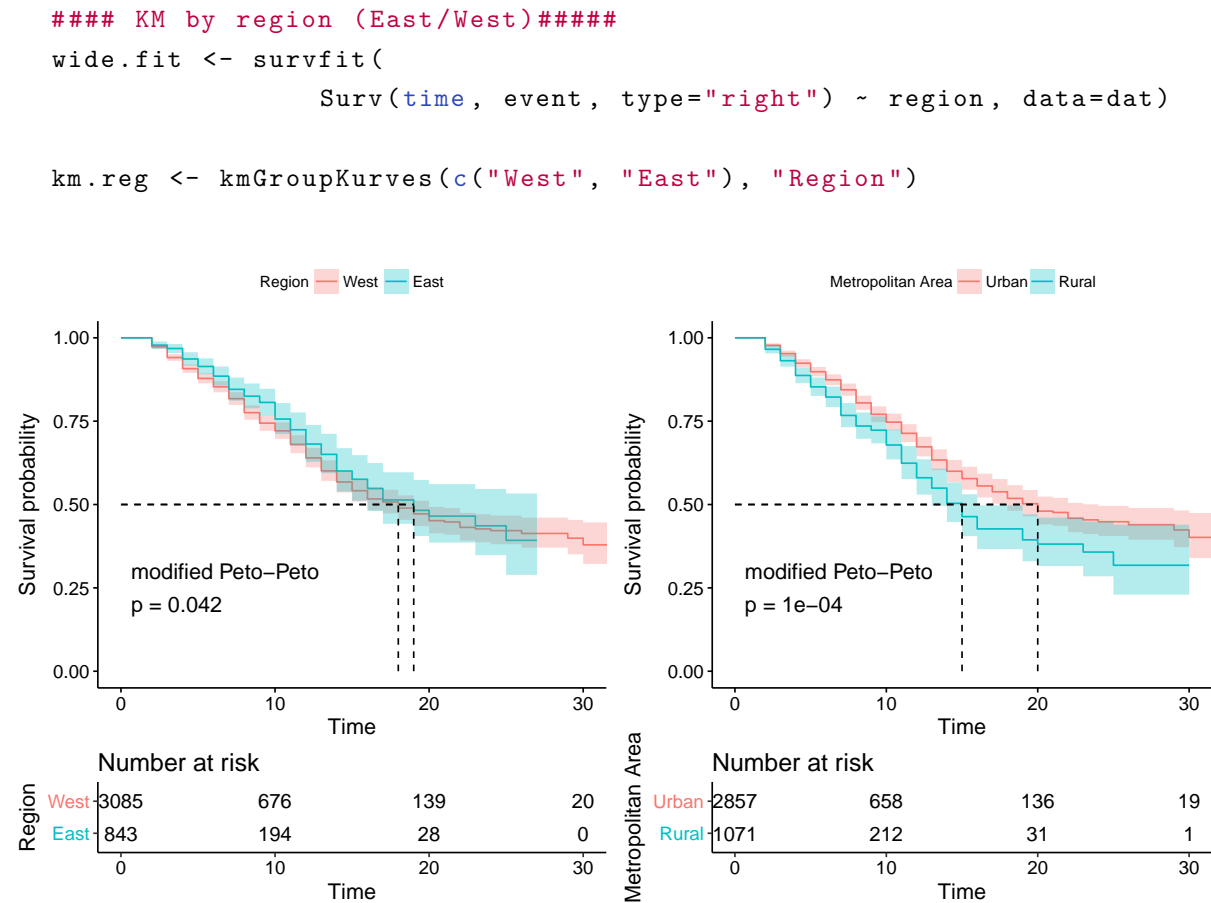

Figure 6: Stratified Kaplan-Meier estimates for survival in rent.

It appears that the West German survival function almost consistently lies below that of East Germans, suggesting that West Germans tend to transition into homeownership earlier than East Germans. The difference if the stratified KM estimator is just significant at the 5% level.

When repeating the same procedure in order to determine whether there exist substantial differences between living in rural or urban areas in Germany, a much clearer distinction is found. Respondents living in urban areas consistently transition slower throughout the whole observation period. The median survival time is roughly five years longer for urban homeowners, and a larger share of them has not transitioned at the end of the observation period of 30 years.

Yet, even more striking regarding survival times is the discrepancy between high and low income individuals. Figure 7 displays the stratified KM estimator.

Individuals belonging to the high income group have a pre-government household income strictly greater than the median income in our sample. We produce the graph by the following lines of code:
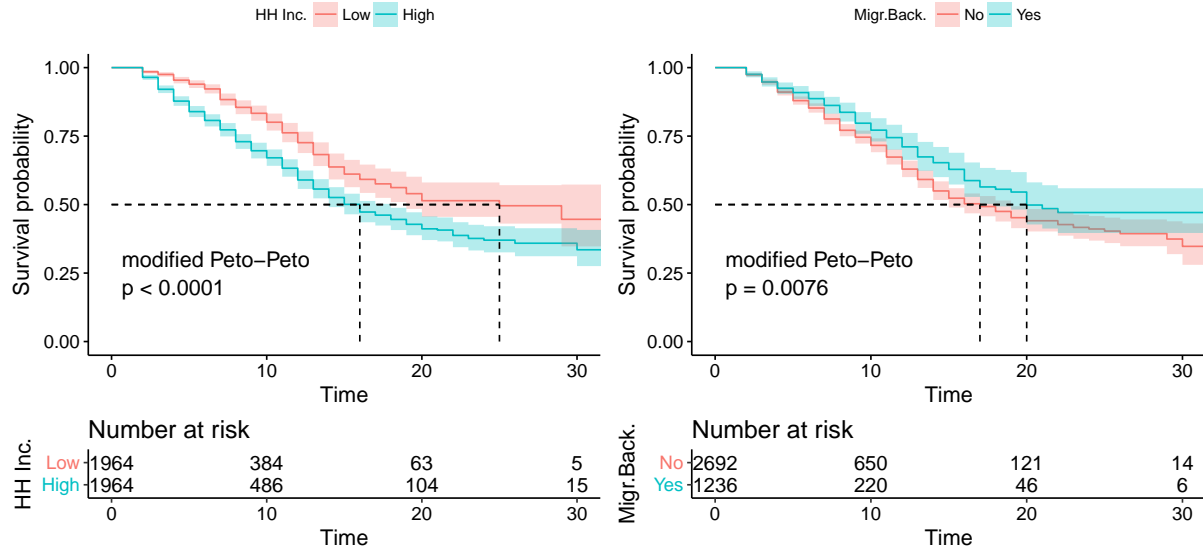
Figure 7: Stratified Kaplan-Meier estimates for survival in rent

```r
medinc <- median(as.numeric(dat$hhinc), na.rm=TRUE)
dat.inc <- mutate(dat, highinc = ifelse(dat$hhinc > medinc, 1, 0))
summary(dat.inc$highinc)
#define survival object and fit KM estimator
wide.fit <- survfit(Surv(time, event, type="right") ~ highinc, data
    =dat.inc)

km.inc <- kmGroupKurves(c("Low", "High"),"HH Inc.")
```

Evidently, and not surprising, high income individuals consistently transition into home-ownership faster than low income individuals. The median survival of the high income group appears to be almost ten years shorter than that of the low income group. What is more, almost 50% of low income individuals do not experience the event at all.

Similarly, yet not as pronounced, the stratification on migration background portrays significant differences between the KM curves. As expected, individuals with migration background transition to homeownership both at a later stage and many, in fact, never do. This can largely be explained by the high correlation between migration background and the tendency to have lower household incomes (see also Grabka and Goebel 2018).

As hypothesized in Section 1, marital status is expected to greatly influence the decision and not least the financial means of acquiring a dwelling. Figure 8 unfolds the transition effects of both whether the respondent is married at the beginning of the observation period and whether or not she ever divorced.

The divorce effect depicted in the right-hand graph appears to be considerably and surprisingly strong. However, the presence of selection bias can be conjectured due to the comparatively few individuals at risk. However, intuitively, going through a divorce puts a large financial burden on both parties involved, resulting in lower transition probabilities
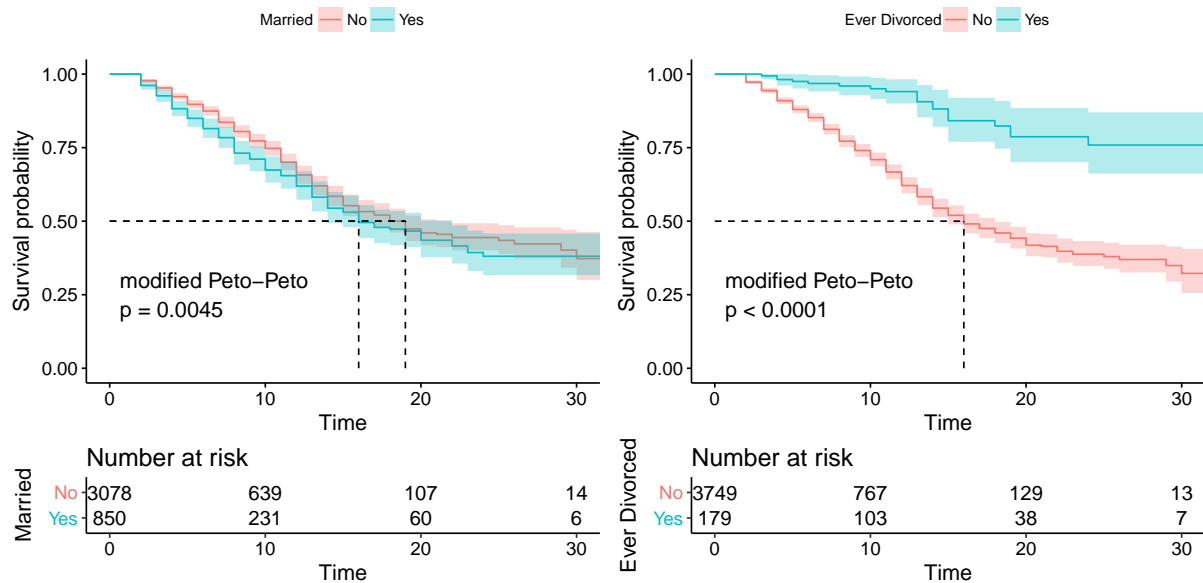
Figure 8: Stratified Kaplan-Meier estimates for survival in rent

into homeownership.

Similarly, when married, financial stability increases and long-term planning horizons expand (see the discussion in Sierminska, Frick, and Grabka 2010). This together with the increased likelihood of producing offspring leads to earlier transitioning into homeownership when compared to non-married individuals. This can clearly be observed in the left-hand graph of Figure 8. In the end, both groups end up with a similar share of respondents having experienced the event, which aptly fits into our explanation.

Lastly, let us inspect different levels of educational attainment and their impact on survival probabilities. The following code produces Figure 9:

```
#define survival object and fit KM estimator
wide.fit <- survfit(Surv(time, event, type="right") ~ educ, data=
    dat)

km.edu <- kmGroupKurves(c("Elementary", "Medium", "Higher voc.", "
    High"),"Education", line = c(1,1,1,1), conf=F)
```

In favor of improved clarity, we opt not to display confidence intervals.

Intuitively, highly educated individuals in the first five or so years transition more slowly into homeownership due to prolonged time of schooling. However, they quickly catch up when they are at the age of 30, ultimately ending up with the lowest rate of survival in rent. On the other hand, individuals with elementary educational attainment experience by far the lowest transition probabilities. On top of that, slightly more than 50% of elementarily educated individuals remain in rent at the end of our observation period. Medium and higher vocational education levels neatly sandwich into these two strata.

Here again, the Peto-Peto transformed log-rank test reports highly significant differences
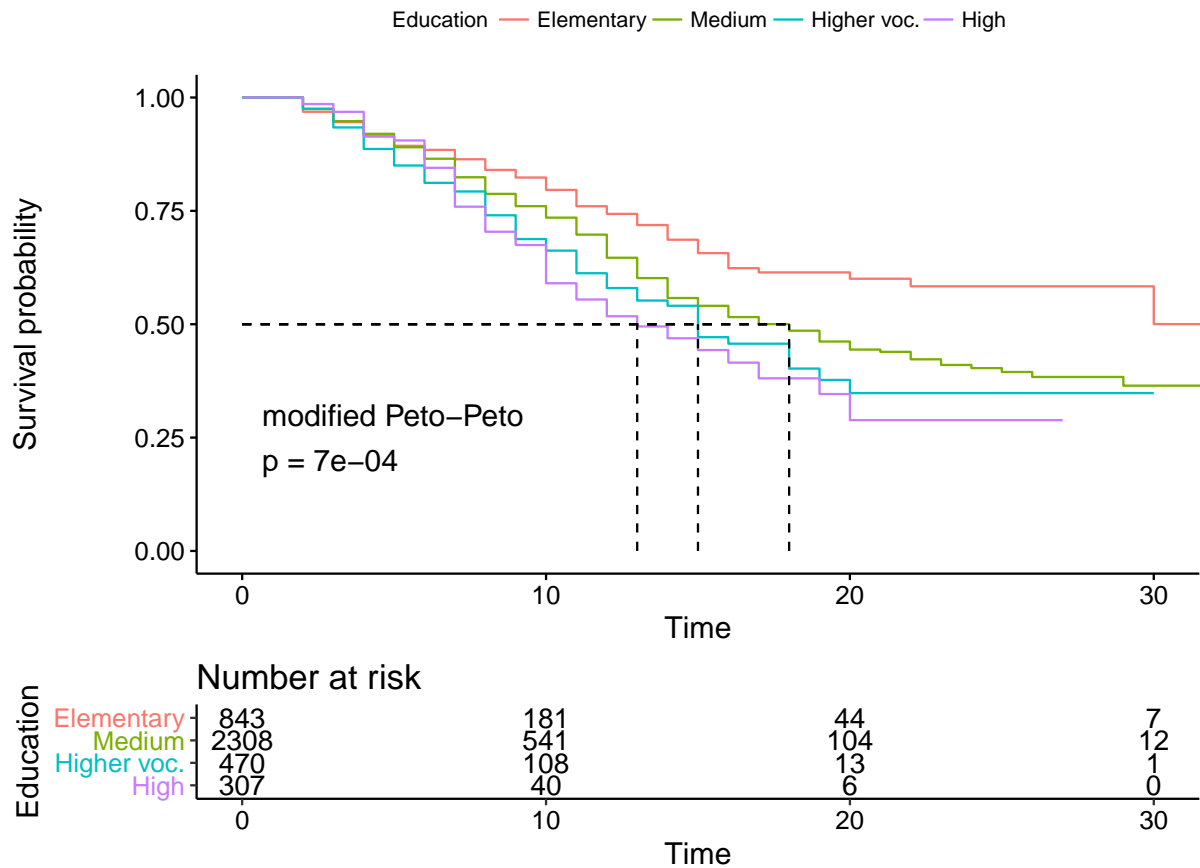
Figure 9: Kaplan-Meier estimates for different education levels

between the stratified survival curves.

In conclusion, factors determining the transitioning from rent into homeownership can easily be exposed when utilizing non-parametric techniques with respect to the survival function. Thus, we are able to confirm the existence of directional effects of different stratification variables, as presented above. Additionally, we can refrain from invoking any distributional assumptions. On the other hand, we cannot make inference regarding effect sizes, or control for the presence of confounding factors.

However, Cox regression provides an elegant method that overcomes these shortcomings, which will be applied and discussed in the next subsection.

4.3   Semi-Parametric Estimators

The Cox Proportional Hazard Model has already been introduced in Section 2. Fortunately, our way forward is rather clear due to the structure of the investigation conducted in our non-parametric analysis. On account of the limited inferential insight the KM estimator provides on covariates' effect sizes, Cox PH regression is employed to do exactly that.

26

Consequently, we fit a linear predictor function to the logarithm of the relative hazard $\lambda(t)/\lambda_0(t)$, i.e. the log ratio of the hazard function to the baseline hazard:

$$\log\left(\frac{\lambda(t|x_i)}{\lambda_0(t)}\right) = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \ ,$$

which is just a different way of writing Equation 11. The x_{ik} denote the time-invariant covariates considered in Section 4.2.

In R, the Cox PH regression model is estimated using the `coxph()` function, again located in the `survival` package. Its general appearance is very similar to the `lm()` function of the `stats` package. Hence, the right-hand side of the formula of `coxph()` corresponds to `lm()` while the left-hand side is a survival object identical to the one defined for the KM Estimator.

```
coxsurv <- Surv(dat$time, dat$event, type="right")
```

Now that we have specified `coxsurv` as our survival object, we can stipulate our Cox PH model[17] as follows:

```
# define formula
coxform <- as.formula("coxsurv ~ hhinc + rural + maxedu + region +
    migback + married + ever_div")

# estimate Cox regression
cox.ph <- coxph(coxform, data=dat)
```

The regression results are displayed in Table 4.3. In favor of readability, we confine to reporting only the most relevant output.

The first and second columns display the time-independent hazard ratios (HR), obtained by exponentiating the estimated coefficients. The third column displays 95% confidence intervals, and the last column the respective *p-values*[18] In general, both the coefficient, and the exponentiated coefficient are interpretable as the effect on the hazard. While a HR greater than one indicates a positive effect on the hazard, the same holds for coefficients greater than zero. A HR smaller than one translates to a negative regression coefficient. For zero coefficients, we obtain a HR of 1, indicating no effect on the hazard function.

For example, living in a rural, as opposed to an urban area, increases the hazard of transitioning into homeownership by 54.8%, *ceteris paribus*. Of course, the coefficient's numerical value of 0.437 has no direct interpretation as an effect size. Rather, it merely indicates

---

[17]We have examined closely whether robust covariance matrix estimators are appropriate in our setting. `coxph()` allows us to fit the so-called Wei, Lin, Weissfeld (WLW) (Lin and Wei 1989) robust variance estimates. However, we do not find convincing evidence for heteroskedasticity and thus refrain from specifying the WLW correction due to efficiency improvement.

[18]The *p-value* is retrieved from standard normal distribution at its z value. In this context, the z value is just the ratio of each regression coefficient to its standard error, which is given in parenthesis. The resulting Wald statistic is asymptotically standard normal distributed (see Lin and Wei 1989).

	Coefficient	HR = exp(β)	95% CI	p-value
HH Income	0.018	1.018	[1.015, 1.022]	< 0.0001
	(0.002)			
Rural	0.437	1.548	[1.301, 1.841]	< 0.0001
	(0.089)			
Years of Education	0.034	1.035	[1.002, 1.068]	0.03
	(0.016)			
East Germany	-0.242	0.785	[0.639, 0.966]	0.02
	(0.105)			
Migration Background	-0.284	0.753	[0.617, 0.919]	0.0052
	(0.102)			
Married	0.275	1.317	[1.103, 1.572]	0.0023
	(0.090)			
Ever Divorced	-1.482	0.227	[0.145, 0.356]	< 0.0001
	(0.229)			
Observations	3,928	Log Likelihood		−4,896.1
Wald Test	218.6	LR Test		214.4

Table 2: Output from Cox PH regression. Standard errors are in parentheses. The hazard ratios depict the (time-constant) ratio of hazard functions for a one-unit increase in the predictor.

the effect's direction, which, however, suffices for most purposes. Likewise, having had a divorce decreases the hazard of transitioning, which is revealed by the variable's negative coefficient. Looking at the HR, we can quantify the effect as a 77% decrease in the transition hazard.

Regarding the coefficient for `HH income`, one might conclude that its impact on the hazard rate is negligibly small. However, recalling that the coefficients display the effect of a one-unit increase in the predictor, and `HH income` is given as annual income in 1000 Euro (in 2010 prices), its effect can be tremendous. For instance, an increase in annual income by 10.000 EUR corresponds to an increase in the annual transition probability from rent to homeownership by $(\exp(0.018)^{10} - 1) \times 100 = 19.72\%$, holding all other variables constant. Thus, a high household income is in principle able to compensate for the negative effects on the transition hazard when choosing to reside in urban areas.

In summary, the findings from our Cox regression are consistent with the non-parametric findings above. Increased pre-government household income, living in rural areas, higher educational attainment[19], and being married have positive effects on the transition probability to homeownership, while living in East Germany, having a migration background, and having had a divorce negatively impact the transition. Both the global Wald Test and the Likelihood-Ratio Test are highly significant at the 0.1% significance level, suggesting the

[19]We decided to use the quasi-continuous years of education variable, as opposed to the categorical ISCED-97 classification in the non-parametric setting. This is largely because the Cox PH model can deal much more easily with non-ordinal variables.

whole model indeed has some merit and is not a product of chance.

4.3.1 Model Diagnostics

Even though our methods are semi-parametric, that does not mean that there is no need to determine whether our Cox regression adequately describes the underlying data. Before we address the essential model diagnostics however, let us first test for the presence of influential observations via *Delta-Beta plots* in our model. These graphs contain so-called Delta-Beta quantities, that are obtained as $\Delta\beta_{ki} = \hat{\beta}_k - \hat{\beta}_{k(i)}$, where $\hat{\beta}_{k(i)}$ is the estimate of β_k with the i-th subject removed from the data set.

Figure 10 visualizes any overly influential observations, utilizing the function ggcoxdiagnostics in the survminer package:

```
ggcoxdiagnostics(cox.ph, type = "dfbetas",
        ox.scale= "observation.id", hline.col = "darkgreen", hline.
    alpha = 0.5, point.alpha = 0.4, point.col = "lightcoral", sline.
    alpha = 0.4, sline.col = "dodgerblue")
```

The function calculates the score residuals, namely each individual's contribution to the score vector[20]. These are then transformed such the approximate change in the coefficient vector $\hat{\beta}$ is retrieved if an individual observation were to be dropped. In addition, we rescale the change in coefficients by that coefficients' standard error to facilitate comparability. Inspecting Figure 10, there is little evidence for influential observations, except for a few observations that carry considerably more influence than others. Ever Divorced seems to be highly unbalanced, owing to the fact that we have very few observations of individuals who have ever had a divorce.

The most crucial assumption invoked by Cox PH models is the proportional hazards assumption, which needs to be statistically examined in order to draw valid inference. Luckily, there is a straightforward approach to check whether the PH assumption holds for each variable, as well as globally.

The general idea, developed by Schoenfeld (1982), is to test the null hypothesis of proportional hazards by considering the so called *scaled Schoenfeld residuals*[21] and plot them against time. Under proportional hazards, the residuals are independent of time.

In R, we can access the Schoenfeld residuals either by adopting the residuals(*model*, "scaledsch") function of the stats package, where model is a coxph model object. How-

---

[20]The score vector is the first derivative of the log-likelihood function.

[21]Recall Equation 11 where we estimate our model with partial maximum likelihood. Provided that our PH model holds, $\hat{\beta}$ is the true regression coefficient and the Schoenfeld residuals just correspond to the difference between the observed covariates and the weighted mean of the covariates of the population at risk. Scaling these residuals by the inverse of the covariance matrix of $\hat{\beta}$ and fitting them against some transformation of time $g(t)$ such that $\beta(t) = \beta + \rho g(t)$, we can test $H_0 : \rho = 0$ (see Grambsch and Therneau 1994, for a more detailed discussion).
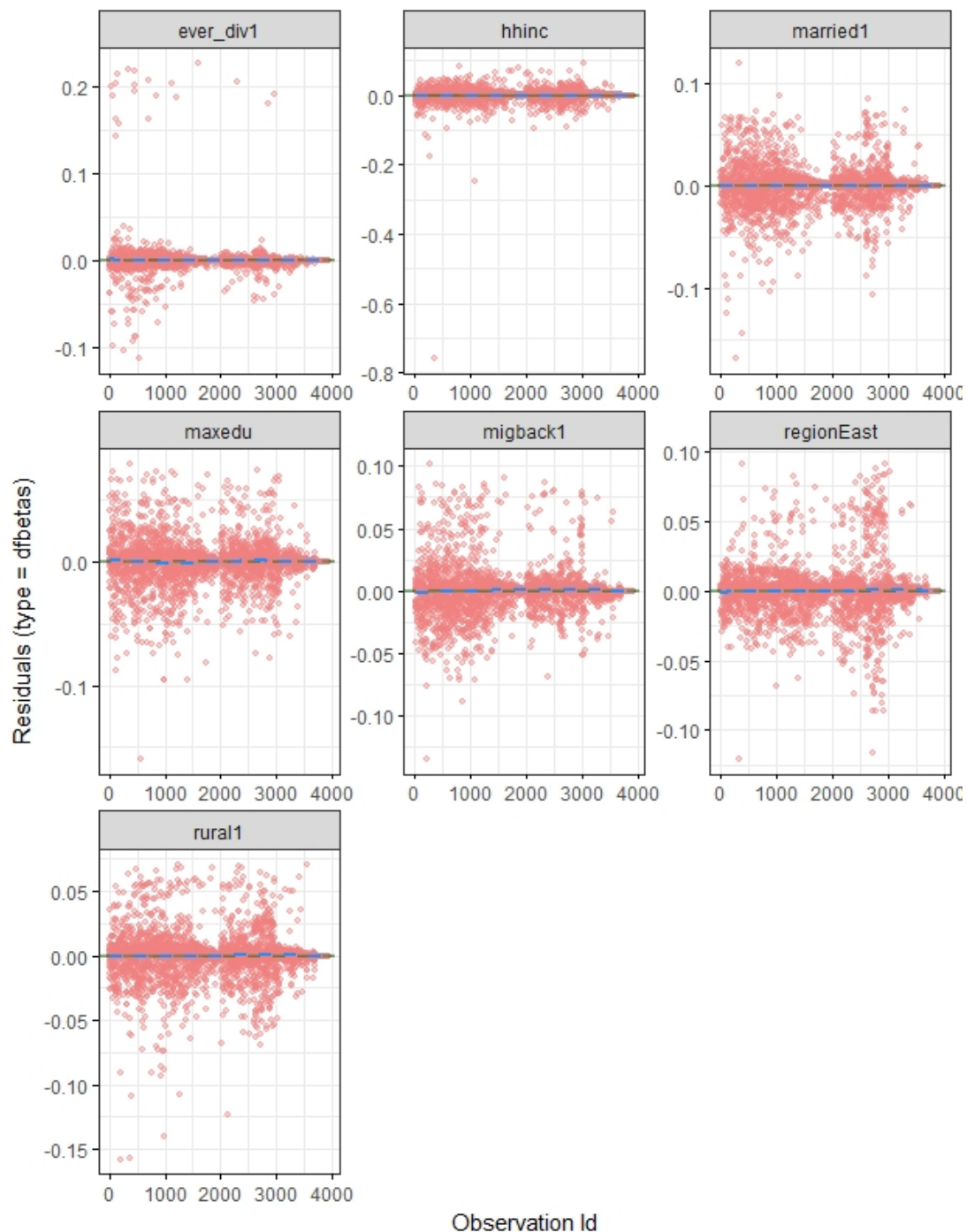
Figure 10: Identification of influential observations through Delta-Beta residuals. [InflObs_DeltaBeta]

ever, the `survival` package again provides us with a more handy solution by employing the `cox.zph()` function. It calculates the above introduced test of the proportional hazards assumption for each of the covariates distinctly and combined. Concerning the required transformation of time $g(t)$, we adhere to the literature and choose the *Kaplan Meier* estimate of the survival function, $\hat{S}_{KM}(t)$:

```
# Schoenfeld test of cox ph assumption
coxtest <- cox.zph(cox.ph, transform = "km")
```

```
coxtest
stargazer(coxtest$table, out = "schoenfeld.tex")
```

	rho	chisq	p
HH Income	-0.060	1.800	0.180
Rural	-0.017	0.190	0.660
Years of Education	0.120	10.000	0.001
East Germany	0.046	1.500	0.230
Migration Background	0.047	1.600	0.200
Married	-0.039	1.100	0.290
Ever Divorced	0.080	4.500	0.034
GLOBAL		26.000	0.001

Table 3: Schoenfeld test for proportional-Hazard assumpution of Cox PH regression. The null hypothesis states that the covariates do no depend on time.

As Table 3 exposes, there is strong evidence for non-proportional hazards for `Years of Education` and `Ever Divorced`. Furthermore, the `Global` Schoenfeld test suggests the null hypothesis of all covariates together being independent of time should be rejected. The latter finding is not particularly problematic, since we do not aim to fit an overall model that correctly captures the hazard of transitioning into homeownership. Much more, our interest lies in the importance of individual covariates.

It is also possible to do a graphical diagnostic check of the PH assumption.. The function `ggcoxzph()` located in the `survminer` package lets us visualize our findings from Table 3 in Figure 11, producing for each covariate a plot of the scaled Schoenfeld residuals against time:

```
# scaled Schoenfeld plots for HH income and years of education
ggcoxzph(coxtest, resid = T, point.alpha = 0.4, var=c("hhinc","
    maxedu"))
```

Figure 11's upper graph displays the Schoenfeld residuals of household pre-government income as red scatter points, together with a fitted natural spline and confidence bands at two standard errors. Schoenfeld residuals, should the PH assumption be valid, have the sample path of a random walk and the fitted smoothed regression line is then expected to have a zero slope. The violation of this assumption is particularly obvious in the lower graph, depicting the Schoenfeld individual test for `Years of Education`. There is a clear inverse U-shaped pattern in the data points, prompting the Schoenfeld test to reject the null. This is not entirely surprising, because as we have seen in the non-parametric analysis, educational attainment does not portray a strictly linear relationship to the survival probability. Individuals with high educational attainment tend to enter the job market at a later stage, reducing the early transition probability into homeownership. Figure 14 in Appendix A displays scaled Schoenfeld residuals for the full set of covariates. It turns out that

31
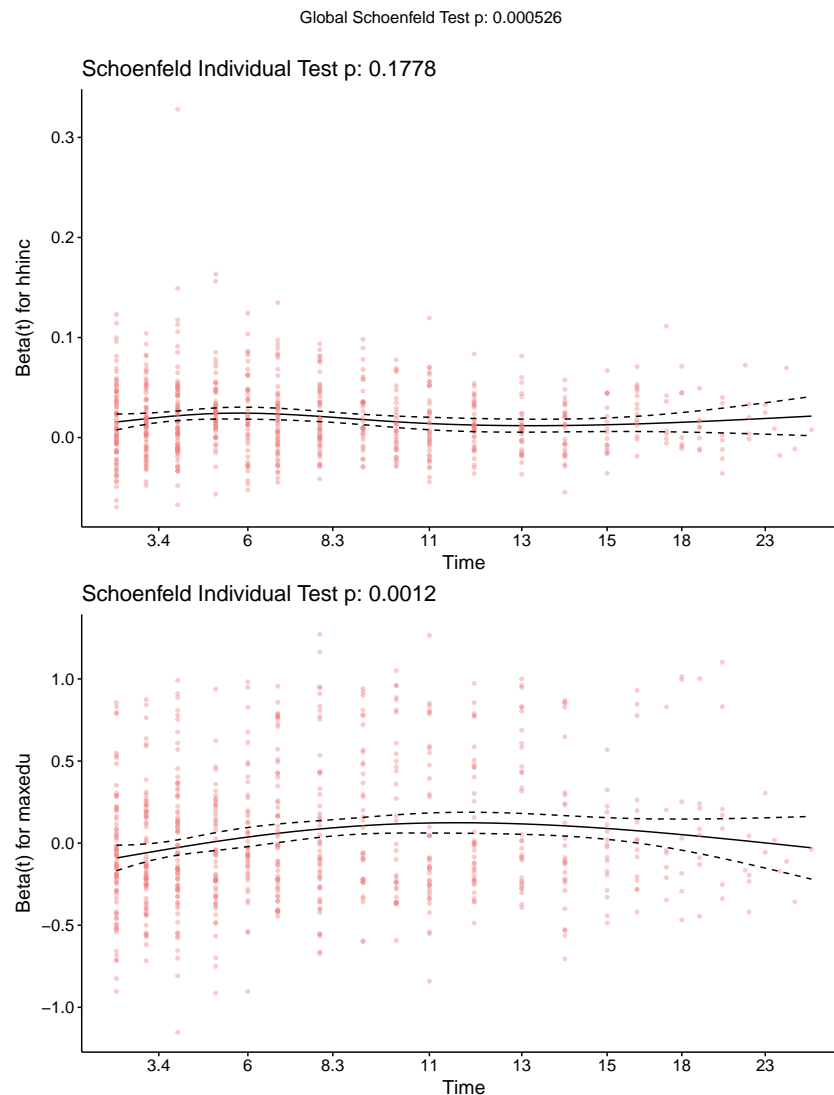
Figure 11: Scaled Schoenfeld residuals for HH income and years of education. Schoenf_IndivTest

for the variable `Ever Divorced`, the number of observations in the divorced category is just too limited, which does not permit us to draw valid conclusions about non-proportional hazards.

Dealing with violations of the proportional-hazard assumption is not always easy and straight forward. One way of tackling time-dependencies is to introduce interactions between the covariates in question and time (see Huang and Liu 2006, for greater detail). However, doing so requires us to assume a particular form of interaction which may lead to serious misspecification. In case of `Years of education`, the nonlinearity with respect to time is not immediately clear, especially when considering our non-parametric findings.

Another method is to stratify the Cox PH regression, such that we allow for distinct baseline hazards in each of the strata. This is the more appealing approach, as it avoids making additional assumptions. Yet, in the presence of large amounts of censoring, stratification might create more problems than it solves, as proportionality of hazards is hard to examine

when observations become scarce, as is the case for `Ever Divorced`.

Next to the aforementioned difficulties of coping with violations of the proportional hazard assumption, Cox PH models intently treat the baseline hazard as a high-dimensional nuisance parameter bringing forth the semi-parametric nature of the Cox Model (see Royston and Parmar 2002).

However, the form of the hazard function might be of direct interest since it can be highly instructional with regard to the time-course of the progression into homeownership. With that in mind, some kind of parametrization can help both to overcome non-proportional hazards and to provide us with precise information about the (baseline) hazard function.

4.4   Comparison to Fully Parametric Approaches

Generally, in parametric survival models, both the hazard function and the functional form of the covariates are specified. Doing so grants us the opportunity to predict survival times and hazard rates, as well as mean survival times and survival quantiles. Furthermore, in the case of correct model specification, parametric models have more power and are more efficient compared to semi-parametric models. In that spirit, note that these parametric models rely on full maximum likelihood.

All praise apart however, as the name suggests, fully parametric models are only efficient and generate correct results when the parametric form is correctly specified. And not surprisingly, choosing the appropriate hazard function together with the covariate form proves to be the crux. With the help of data driven methods to assess whether the specified form appears to fit the data, and prior knowledge from our non- and semi-parametric results, effective model selection can be performed.

As already introduced in section 2.2, we subsequently examine the most common Accelerated Failure Time (AFT) and Proportional Hazards (PH) models. Royston and Parmar (2002)'s flexible parametric proportional hazard approach serves as a crude benchmark for the true shape of the hazard and survival curves. This is because although it lacks the interpretability other parametric methods provide, its flexible nature typically implies a good fit, regardless of whether the researcher is agnostic about the true survival distribution. In other words, flexible models are less likely to fall prey to model misspecification. Recall, as in Equation 4, that the family of AFT (with the Weibull and Exponential also exhibiting PH) survival distributions can be parametrized in terms of location and scale parameters for the natural logarithm of the survival time:

$$\log(T_i) = x_i' \beta + \sigma \epsilon_i \quad,$$

where the parametric assumption concerns only ϵ_i (see Haile 2015).

The `flexsurvreg()` function in the `flexsurv` package deals with all common AFT models

including Weibull[22], Exponential, Log-Logistic, and Log-Normal distributions.

```
# define survival object
coxparm <- Surv(dat$time, dat$event, type="right")
# define model formula
parmform <- as.formula("coxparm ~ hhinc + rural + maxedu + region +
    migback + married + ever_div")
# Flexible splines (Royston and Parmar 2002)
flex.spline <- flexsurvspline(coxparm ~ 1, data = dat, k = 2, scale
    = "odds")
# Weibull distribution
weibull <- flexsurvreg(formula = parmform, data = dat, dist = "
    weibull")
# Exponential distribution
expo <- flexsurvreg(formula = parmform, data = dat, dist = "exp")
# Log-Logistic distribution
loglog <- flexsurvreg(formula = parmform, data = dat, dist = "
    llogis")
# Log-normal distribution
lnormal <- flexsurvreg(formula = parmform, data = dat, dist = "
    lnorm")
```

Naturally, we preserve the survival object and the model formula used ealier in order to uncover the best model fit. In order to get a feel for the properties of each distribution, let us first plot the estimated hazard functions for each of the five parametric models we consider:

```
# plot hazard functions
ggplot(data.frame(summary(expo, type = "hazard")), aes(x = time)) +
  geom_line(aes(y = est, col = "Exponential")) +
  geom_line(data = data.frame(summary(weibull, type = "hazard")),
    aes(y = est, col = "Weibull")) +
  geom_line(data = data.frame(summary(loglog, type = "hazard")),
    aes(y = est, col = "Log-Logistic")) +
  geom_line(data = data.frame(summary(lnormal, type = "hazard")),
    aes(y = est, col = "Log-Normal")) +
  geom_line(data = data.frame(summary(flex.spline, type = "hazard")
    )), aes(y = est, col = "Flexible Splines")) +
  labs(x = "Time (years)", y = "Hazard Function", col = "Models") +
    theme_classic()
```

Figure 12 depicts the resulting plot produced by ggplot. Looking at the the hazard function estimated by flexible parametric splines, the survival time seems to exhibit positive *duration dependence* during the first 10 or so years, before the relationship reverses to negative dependence.

---

[22]Weibull AFT regression used to be implemented in the WeibullReg() function located in the survreg package. However, the flexsurv package now combines all distributional options, including more flexible arguments, in one package.
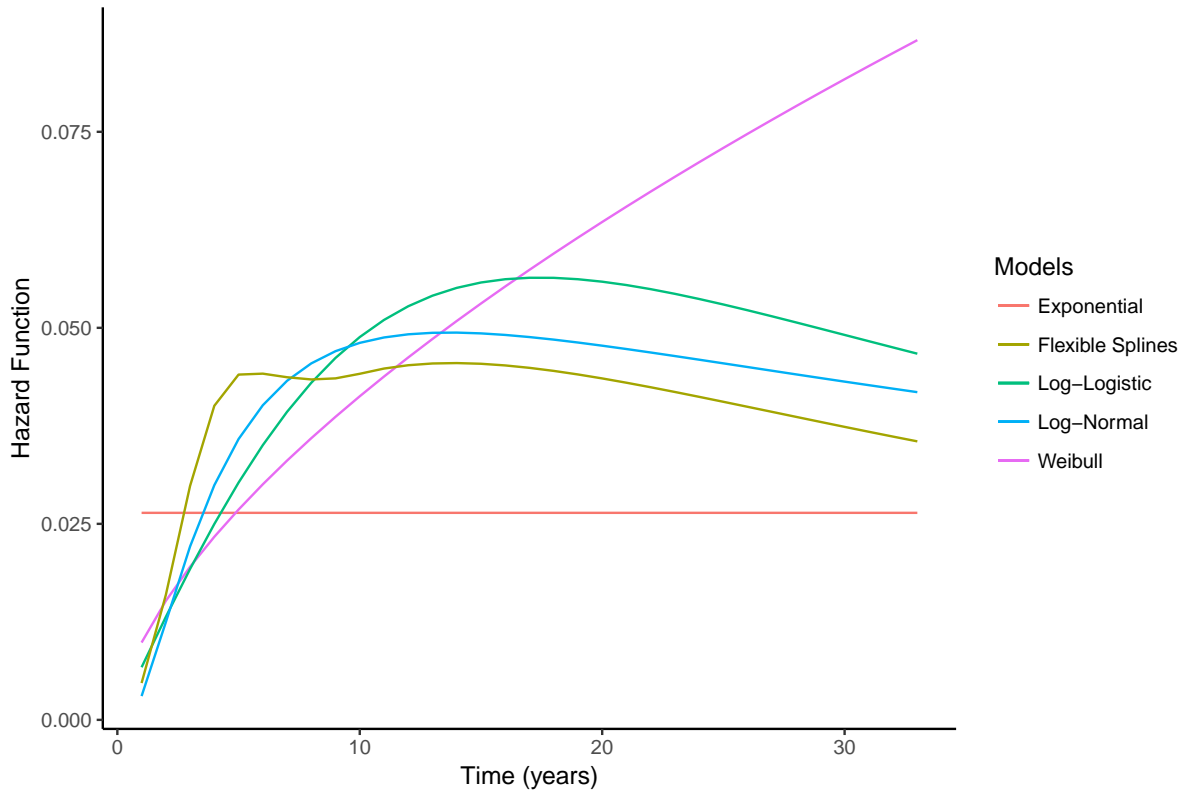
Figure 12: Estimated hazard function for different models. ComparHaz

As the Weibull and the Exponential[23] distribution can not accommodate non-monotonic behavior of the hazard function, they likely make bad model candidates. The log-logistic and log-normal models look more promising, as both can handle switching signs of duration dependence, with the log-normal model seeming more close to the true shape's approximation. The graph in whole suggests the hazard of transitioning into homeownership reaches its maximum at around age 35, with monotonically decaying tails on both sides. Moreover, the hazard function appears to be right-skewed, implying a sharp increase in transition probability before peak hazard, as well as a slow decline afterwards. From a graphical inspection alone, the log-normal model seems to fit our data best.

Next, we contrast the different models by plotting their estimated survival curves. Now, not only parametric flexible splines can serve as a benchmark, but furthermore we also graph the results from our non-parametric and semi-parametric models. The graph is obtained with the same lines of `ggplot` code as Figure 12, only without specifying the `summary` option `type = "hazard"`, which then automatically defaults to the survival function as the graphical output. To further add the non-parametric Kaplan-Meier and semi-parametric Cox estimates, we first have to estimate the models and save their output as a data frame

---

[23]The hazard function of the Exponential model is simply a constant, meaning that the hazard rate does not depend on time.

using the `ggfortify` package:

```
# Kaplan -Meier estimator
kapm <- survfit(coxparm ~ 1, data=dat)
library(ggfortify)
# puts survival table from kapm object into a data frame
kap.dat <- fortify(kapm)
# Cox PH model
cox.ph <- coxph(formula=parmform , data=dat)
# fortify cox model output
cox.dat <- fortify(survfit(cox.ph, conf.int = F))
```

before adding the following two lines to the above described `ggplot` code:

```
+ geom_step(data = kap.dat, aes(x=time , y=surv , colour = "Kaplan -
    Meier"), size = 0.37)
+ geom_step(data = cox.dat, aes(x=time , y=surv , colour = "Cox PH"),
    size = 0.37)
```

Figure 13 shows the estimated survival curves for our four classical parametric distributions, flexible parametric splines, the Kaplan-Meier estimator, and Cox regression. For the latter, benchmark survival curves, but not hazard functions, can be computed with one caveat: An unconditional and general survival curve can not be obtained for semi-parametric Cox PH models, because all results are conditional on covariates. Rather, it is standard in the literature (and in R packages) to consider the simulated survival curve at the mean of covariates (see Jackson 2016, for greater detail).
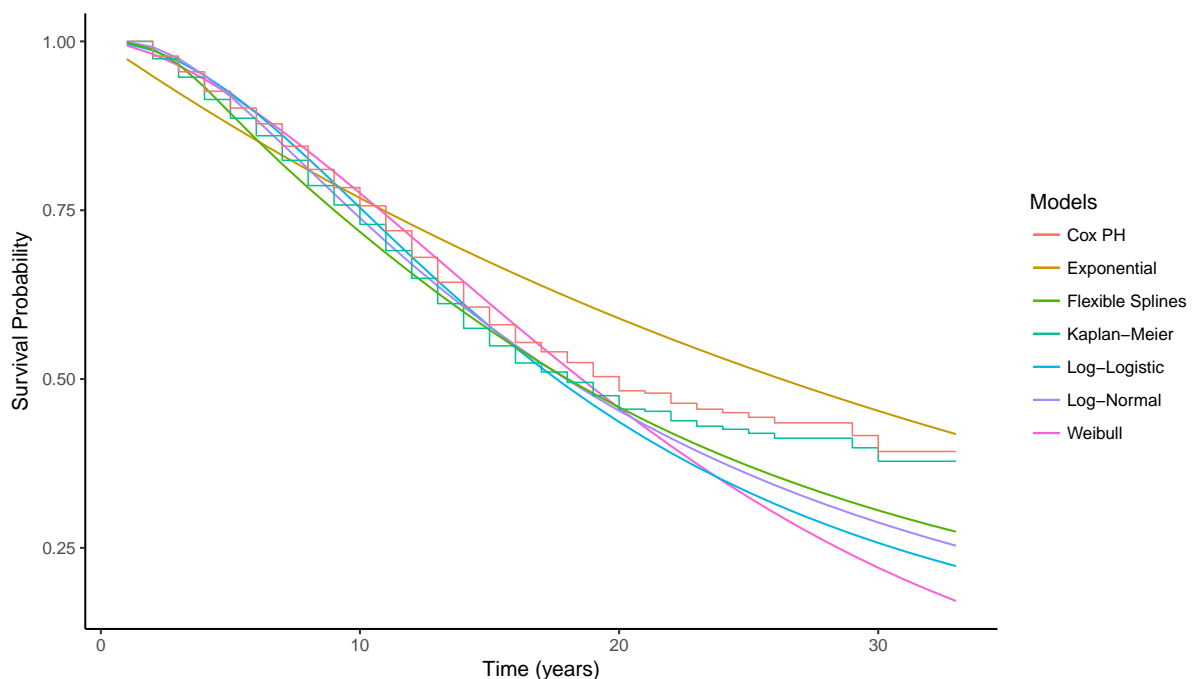

Figure 13: Estimated survival function for different models.

ComparSurv

36

At first glance, it becomes obvious that the two step functions - the non-parametric and semi-parametric curves - are closely aligned. Additionally, also the survival curve obtained from Royston and Parmar's model closely imitates their path and curvature. Together, the three curves almost coinciding provides strong evidence for their validity. Unsurprisingly, the functional form of the Exponential model is too restrictive for it to have a good fit. The estimate for the Weibull model also diverges toward the later years, although model uncertainty at this stage is high due to sample attrition. The log-normal and log-logistic models again exhibit relatively good fit, only showing minor discrepancies. This is plausible, as their distributions look very much alike, relative to the others. This second graphical examination also suggests the log-normal or the log-logistic approach as the best parametric model candidate.

As mentioned before, the most fundamental aspect of assessing parametric model fit is to inspect whether the chosen parametric form suits the data. Until now, this has been done graphically. Naturally, however, we prefer a numerical approach that provides us with a mathematically more rigorous solution. Demler, Paynter, and Cook (2015) assess the calibration of several goodness-of-fit tests to survival data. Unfortunately, they are predominantly designed to check the proportional hazard assumption, leaving the distributional specifications assessed before in question.

Fortunately enough, there is a simple way out. The Akaike Information Criterion (AIC), introduced by Akaike (1974), is a well-known method, capable of performing appropriate model selection when the model is estimated via the maximum-likelihood principle.

	Weibull	Exponential	Log-Logistic	Log-Normal
AIC	5,985	6,255	5,915	5,881
Log-Likelihood	-2,983	-3,119	-2,949	-2,932

Table 4: AIC for model selection of the fully parametric AFT survival models. The lowest AIC corresponds to the best model choice. InfCrit_Surv

As this is the case for our fully parametric survival models, Table 4 reports the AIC values for each of the AFT models, along with their respective Log-Likelihood.

The AIC confirms our graphical evaluations, where the log-normal model exhibits the best fit of all parametric models considered. Unfortunately, we cannot compare this with the semi-parametric Cox PH regression[24], because there is no full likelihood obtained in that model.

---

[24]Generally, parametric models rely on observed event times while semi-parametric models such as the Cox PH model are based on the order of event times

Now that we have determined the best parametric fit, we can at last take a look at the parameter estimates. Table 5 reports the results from four different models: The Cox PH model, which corresponds to the results outlined in Table 4.3 and three of the AFT models. The exponential model has been dropped from the table for clarity, as its fit was already determined to be inferior.

	Survival models:			
	Cox PH	Weibull	Log-Logistic	Log-Normal
	(1)	(2)	(3)	(4)
HH Income	0.018***	−0.011***	−0.012***	−0.012***
	(0.002)	(0.001)	(0.001)	(0.001)
Rural	0.440***	−0.300***	−0.290***	−0.280***
	(0.089)	(0.055)	(0.058)	(0.058)
Years of Educ.	0.034**	−0.027***	−0.018*	−0.008
	(0.016)	(0.010)	(0.010)	(0.010)
East	−0.240**	0.130**	0.160**	0.170**
	(0.100)	(0.065)	(0.067)	(0.068)
Migr. Backgr.	−0.280***	0.160**	0.190***	0.170***
	(0.100)	(0.063)	(0.065)	(0.064)
Married	0.280***	−0.150***	−0.170***	−0.150**
	(0.090)	(0.056)	(0.060)	(0.060)
Ever Divorced	−1.500***	0.990***	1.000***	1.000***
	(0.230)	(0.140)	(0.130)	(0.120)
Constant		3.700***	3.300***	3.200***
		(0.150)	(0.150)	(0.150)
Observations	3,928	3,928	3,928	3,928
Log Likelihood	−4,896	−2,983	−2,949	−2,932
χ^2 (df = 7)		232***	241***	226***
Score (Logrank) Test	211.000*** (df = 7)			

*p<0.1; **p<0.05; ***p<0.01

Table 5: Comparing semi- and fully parametric models. Note that fully parametric models should report coefficient estimates corresponding largely to the Cox PH coefficients but with opposite sign.

As we have already concluded, the Log-Normal parametrization fits our data best, hence we will inspect its properties more closely.

Withal, additional care has to be taken with respect to parameter interpretation. The Weibull, as well as the Log-Logistic and the Log-Normal model all belong to the family of AFT models[25], which means that the reported coefficients are associated with so-called *acceleration factors*.

Let us consider a simple example: As we have already discussed earlier, living in rural Germany as opposed to an urban area increases the hazard of transitioning from rent into homeownership by 54.8%, *ceteris paribus*. Now, considering our Log-Normal model, the coefficient is estimated to be -0.280. Solving the estimated survival function for our parameter, we can conclude that our acceleration factor is given by $\hat{\gamma}_k = exp(\hat{\beta}_k)$. This means that a one-unit change in x_k corresponds to a $100 \times [1 - \hat{\gamma}_k]$ percent change in the expected survival time.

With $100 \times [1 - exp(-0.280)]$ being equal to 24.42, we observe that, holding all other factors constant, by living in rural Germany the estimated survival time in rent decreases by 24.42%.

Furthermore, in the Log-Normal model the variable for educational attainment is no longer significant, which is not so surprising since in its effect is likely to be mediated through higher income, which we have already controlled for. Satisfactorily enough, we find that the coefficient range is less extreme as compared to the Cox PH model. Living in East Germany increases expected survival in rent by 18.5%, while being married decreases it by 13.9%.

In essence, after deciding on the functional form, the Log-Normal model provides an easily interpretable tool that is capable of predicting expected survival times in a straightforward manner. Even though there is a thin line between model misspecification and convenience, we are confident that following the steps from investigating non-parametric methods to a full parametrization of the survival function and its covariate structure, provides us with sufficient *a priori* knowledge to robustly make that decision.

---

[25]The Weibull model also belongs to family of proportional hazard models. However, there is a transformation required when rewriting the AFT into PH.

5 Summary and Conclusion

In our empirical analysis, we investigate the determinants of the transition time from rent into first-time homeownership using longitudinal data from Germany's largest annual survey on socio-economic factors and living conditions. To do this, we employ a number of non-parametric, semi-parametric and parametric methods from the field of survival analysis.

In our non-parametric analysis, we first characterize the shape of the survival curve. We find that median survival in rent is about 15 years (after age 25), whereas roughly 30% of subjects stay in rent throughout the whole observation period of 30 years. Then, we proceed by comparing stratified Kaplan-Meier survival curves and assess whether they are significantly different from each other through Log-Rank testing. Regarding the spatial factors we incorporate in our models, we find that living in East Germany, as well as living in an urban area, is associated with a significantly lower probability of transitioning into homeownership. The divide is especially pronounced for urban vs. rural areas. Unsurprisingly, high income is also associated with less time spent in rent, while having a migration background predicts a lower transition probability. Likewise, being married is negatively related to survival time, meaning that married individuals transition faster into homeownership. Having had a divorce, however, puts a considerable financial burden on both parties, and hence is positively associated with remaining in rent. Lastly, we investigate the way educational attainment is connected to homeownership. We find that survival in rent is inversely related to education, however, this effect could merely be mediated by the higher income that comes with better education.

To account for confounding variables and quantify effect sizes, we employ a semi-parametric Cox PH model. Via the Cox model, we can consistently estimate covariate effects via partial Maximum-Likelihood without having to specify the baseline hazard first. The results from this regression are in line with our findings from non-parametric methods. However, now we can quantify effect sizes: For example, increaseing household income by 1,000 Euros is estimated to increase the hazard function by 1.8%. Likewise, living in East Germany is associated with a 21.5% reduction in the hazard function. Even though the Cox PH model is semi-parametric in nature, it still invokes some assumptions that are not always satisfied, above all the Proportional Hazards assumption. Therefore, we run a number of diagnostic checks, including a Schoenfeld residual test, finding that the PH assumption holds for all but two variables - years of education and having had a divorce. For education, this can be traced back to the fact that higher education first increases the survival function (because individuals stay in school for longer), and then drastically reduces survival once they are in the work force due to higher income. Therefore, educational attainment "rotates" the

survival curve clock-wise, violating the PH assumption.

In a last step, we estimate a number of fully parametric survival models and perform model selection in order to uncover the best fit. As a benchmark, we use a flexible splines model that approximates the true shape of the survival and hazard functions, and compare that to the results from our parametric models. Furthermore, the results from our parametric models are broadly consistent with our non-parametric and semi-parametric approaches. Upon graphical inspection, we find that the Log-Normal distribution most accurately fits our observed survival data. Akaike's Information Criterion confirms our impression.

The consistency of results across the wide variety of methods lends credibility to our findings. In conclusion, our findings robustly shed light on some of the most relevant determinants of the transition time from renting to first-time homeownership, describing their significance and effect sizes.

References

Akaike, H. 1974. "A new look at the statistical model identification". *IEEE transactions on automatic control* 19 (6): 716–723.

Arboretti, R., et al. 2017. "Nonparametric combination tests for comparing two survival curves with informative and non-informative censoring". *Statistical methods in medical research*: 0962280217710836.

Boehm, T. P., and A. M. Schlottmann. 2004. "The dynamics of race, income, and homeownership". *Journal of Urban Economics* 55 (1): 113–130.

Colosimo, E., et al. 2002. "Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators". *Journal of Statistical Computation and Simulation* 72 (4): 299–308.

Cox, D. R. 1972. "Regression Models and Life-Tables". *Journal of the Royal Statistical Society. Series B (Methodological)* 34 (2): 87–22.

Demler, O. V., N. P. Paynter, and N. R. Cook. 2015. "Tests of calibration and goodness-of-fit in the survival setting". *Statistics in medicine* 34 (10): 1659–1680.

Di, Z. X., and X. Liu. 2007. "The importance of wealth and income in the transition to homeownership". *Cityscape*: 137–151.

Fischer, M., and N. Khorunzhina. 2015. "Housing Decisions under Divorce Risk".

Fisher, J. D., and M. Gervais. 2011. "Why has home ownership fallen among the young?" *International Economic Review* 52 (3): 883–912.

Fleming, T. R., and D. P. Harrington. 1979. "Nonparametric estimation of the survival distribution in censored data". *Unpublished manuscript.*

Grabka, M. M. 2014. "Private Vermögen in Ost-und Westdeutschland gleichen sich nur langsam an". *DIW-Wochenbericht* 81 (40): 959–966.

Grabka, M. M., and J. Goebel. 2018. "Einkommensverteilung in Deutschland: Realeinkommen sind seit 1991 gestiegen, aber mehr Menschen beziehen Niedrigeinkommen". *DIW-Wochenbericht* 85 (21): 449–459.

Grambsch, P. M., and T. M. Therneau. 1994. "Proportional hazards tests and diagnostics based on weighted residuals". *Biometrika* 81 (3): 515–526.

Grinstein-Weiss, M., et al. 2011. "The effect of marital status on home ownership among low-income households". *Social Service Review* 85 (3): 475–503.

Guiso, L., and T. Jappelli. 2002. "Private transfers, borrowing constraints, and timing of homeownership". *Journal of money, credit, and banking* 34 (2): 315–339.

Haile, S. R. 2015. "Weibull AFT Regression Functions in R". *Cran-R.*

Huang, J. Z., and L. Liu. 2006. "Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form". *Biometrics* 62 (3): 793–802.

Jackson, C. 2017. "flexsurv: Flexible parametric survival and multi-state models, 2014". *URL http://CRAN. R-project. org/package= flexsurv. R package version 0.5.[p 114]*.

Jackson, C. H. 2016. "flexsurv: a platform for parametric survival modeling in R". *Journal of Statistical Software* 70.

Kaplan, E. L., and P. Meier. 1958. "Nonparametric estimation from incomplete observations". *Journal of the American statistical association* 53 (282): 457–481.

Kassambara, A., M. Kosinski, and P. Biecek. 2017. "survminer: Drawing Survival Curves usinggplot2". *R package version 0.3* 1.

Kauppinen, T. M., H. Skifter, and L. Hedman. 2015. "DETERMINANTS OF IMMIGRANTSENTRY TO HOMEOWNERSHIP IN THREE NORDIC CAPITAL CITY REGIONS". *Geografiska Annaler: Series B, Human Geography* 97 (4): 343–362.

King, M. A. 1980. "An econometric model of tenure choice and demand for housing as a joint decision". *Journal of Public Economics* 14 (2): 137–159.

Kleinbaum, D. G., and M. Klein. 2010. *Survival analysis*. Vol. 3. Springer.

Kuritz, S. J., J. R. Landis, and G. G. Koch. 1988. "A general overview of Mantel-Haenszel methods: applications and recent developments". *Annual review of public health* 9 (1): 123–160.

Lin, D. Y., and L.-J. Wei. 1989. "The robust inference for the Cox proportional hazards model". *Journal of the American statistical Association* 84 (408): 1074–1078.

Linneman, P., and S. Wachter. 1989. "The impacts of borrowing constraints on homeownership". *Real Estate Economics* 17 (4): 389–402.

Nelson, W. 1969. "Hazard plotting for incomplete failure data". *Journal of Quality Technology* 1 (1): 27–52.

— . 1972. "Theory and applications of hazard plotting for censored failure data". *Technometrics* 14 (4): 945–966.

Peto, R., and J. Peto. 1972. "Asymptotically efficient rank invariant test procedures". *Journal of the Royal Statistical Society. Series A (General)*: 185–207.

Rohe, W. M., and M. Lindblad. 2013. "Reexamining the social benefits of homeownership after the housing crisis".

Royston, P., and M. K. Parmar. 2002. "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects". *Statistics in medicine* 21 (15): 2175–2197.

Sagna, P., and M. Voigtländer. 2018. "Die Zahl der Ersterwerber sinkt weiter". *IW-Kurzbericht*.

Schoenfeld, D. 1982. "Partial residuals for the proportional hazards regression model". *Biometrika* 69 (1): 239–241.

Sierminska, E. M., J. R. Frick, and M. M. Grabka. 2010. "Examining the gender wealth gap". *Oxford Economic Papers* 62 (4): 669–690.

Tang, F., and H. Ishwaran. 2017. "Random forest missing data algorithms". *Statistical Analysis and Data Mining: The ASA Data Science Journal* 10 (6): 363–377.

Therneau, T. M. 2015. *A Package for Survival Analysis in S.* Version 2.38.

Appendices

A   Analysis
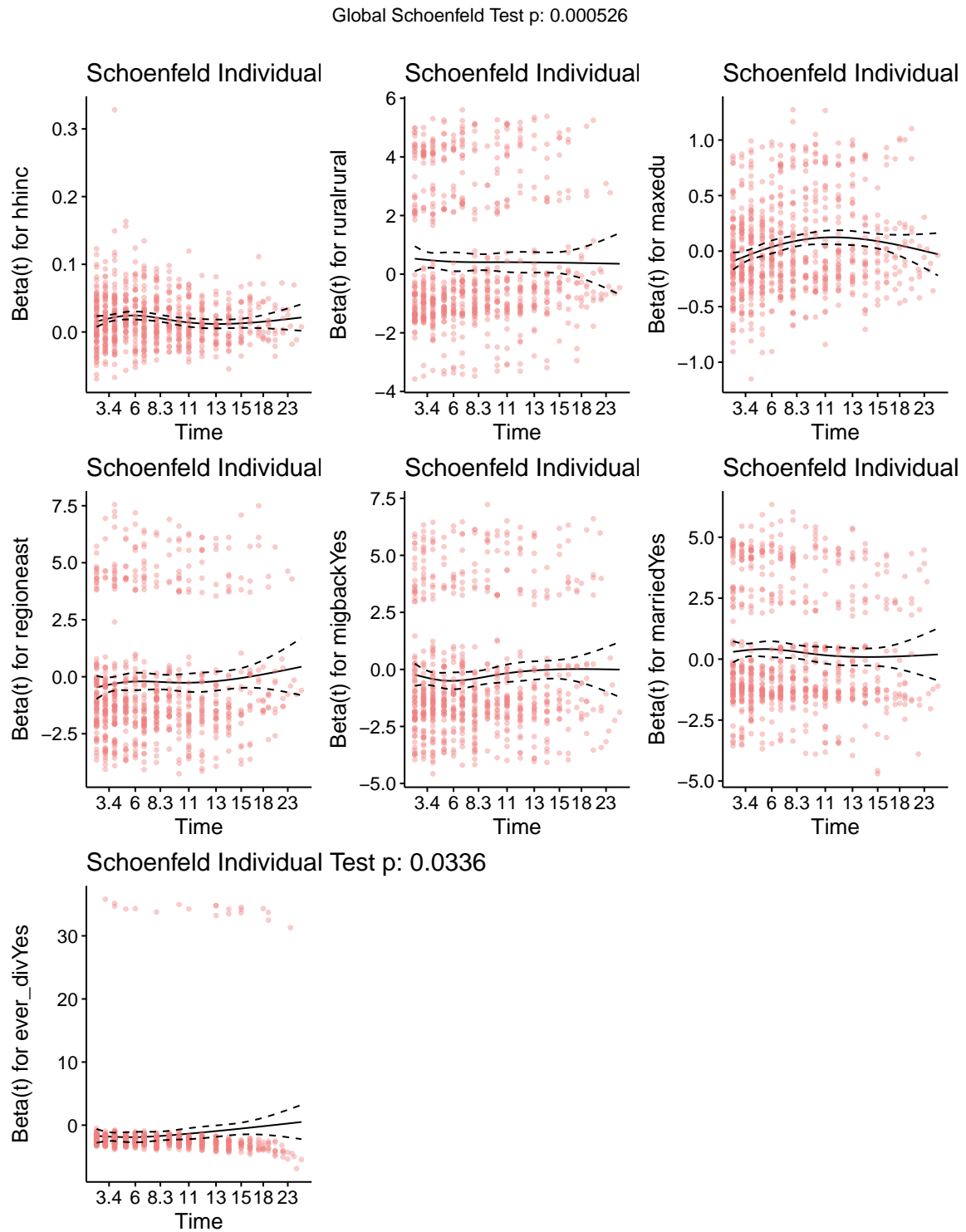
Figure 14: Scaled Schoenfeld residuals for all covariates.

Schoenf_IndivTest

Appendices

B Source Code

```r
###############################################################
### Packages ##################################################
###############################################################

if (!require("pacman")) install.packages("pacman")
pacman::p_load(
  bigmemory,
  biganalytics,
  plm,
  car,
  varhandle,
  foreign,
  survival,
  rms,
  survminer,
  randomForestSRC,
  ggRandomForests,
  pec,
  flexsurv,
  ggfortify,
  broom,
  grid,
  gridExtra,
  stargazer,
  reporttools,
  forestplot,
  Hmisc,
  pastecs,
  rio,
  haven,
  VIM,
  naniar,
  ggridges,
  GGally,
  # always load these last
  dplyr,
  readr,
  ggplot2,
  tidyverse)
```

```r
##############################################################
##############################################################
####Descriptives##############################################
##############################################################
##############################################################

#Note1: All scripts and SOEP data files need to be in the same
    directory
#Note2: .path.R file needs to be specified by the user

#Structure:
#0: Set-up
#1: Summary table
#2: Data structure visualization
#3: Density plot of survival times
#4: Density plots by federal states


##############################################################
#### Set Up ##################################################
##############################################################


# clear workspace
rm(list=ls())

# setwd(path) in path.R
source(".path.R")

# install and load packages
source("packages.R")

# load data
load("datfinal.RDA")


##############################################################
#### Summary Table ###########################################
##############################################################

# recode factors as numeric for them to be included in table

sum.dat <- dat
sum.dat$gender <- as.integer(as.character(sum.dat$gender))
sum.dat$married <- as.integer(as.character(sum.dat$married))
sum.dat$ever_div <- as.integer(as.character(sum.dat$ever_div))
sum.dat$region <- as.integer(sum.dat$region)-1
sum.dat$rural <- as.integer(as.character(sum.dat$rural))
```

```r
sum.dat$migback <- as.integer(as.character(sum.dat$migback))

sum.dat <- select(sum.dat,
                  one_of(c("time", "event", "hhinc",
                           "maxedu","birthyear","gender",
                           "region", "rural", "married",
                           "ever_div", "migback")) )

stargazer(sum.dat, type="latex",
          summary.stat=c("mean", "sd", "min",
                         "median", "max", "n"),
          title = "Summary Statistics",
          covariate.labels = c("Time to Event", "Event",
                               "HH Income", "Educ. (years)",
                               "Birthyear", "Gender", "East Germany",
                               "Rural", "Married",
                               "Ever Divorced", "Migr. Background"),
          digits=2,
          summary.logical = TRUE)

rm(sum.dat)


###############################################################
#### Data Structure Visualization ############################
###############################################################

# permutate persID to get rid of pnr (=syear) dependencies in
   survival time

dat.str <- dat
dat.str$pnr <- sample(1:nrow(dat), nrow(dat), replace=F)

# plot data set
ggplot(dat.str, aes(x = pnr, y = time)) +
  geom_linerange(aes(ymin = 0, ymax = time), size=0.3) +
  geom_point(aes(shape = as.factor(event),
  color = as.factor(event)), stroke = 0.5, cex = 1) +
  scale_shape_manual(values = c(3,4)) +
  guides(shape = F, color = F) +
  labs(y = "Time (years)", x = "Subject ID") + coord_flip() +
  theme_classic()

# plot too crowded, look only at subsample of 500
dat.str1 <- subset(dat.str, pnr<500)

ggplot(dat.str1, aes(x = pnr, y = time)) +
  geom_linerange(aes(ymin = 0, ymax = time), size=0.5) +
```

```r
  geom_point(aes(shape = as.factor(event),
  color = as.factor(event)), stroke = 1.3, cex = 2) +
  scale_shape_manual(values = c(1,4))  +
  labs(color="Event") + guides(shape=F) +
  labs(y = "Time (years)", x = "Subject ID") + coord_flip() +
  theme_classic()


################################################################
#### Density plot of survival time ###########################
################################################################

# subset individuals who  transitioned into ownership
dens.dat <- subset(dat, event==1)

# frequency plot
ggplot(dens.dat, aes(x=time)) +
  geom_histogram(binwidth = 0.2, aes(fill = ..count..) ) + theme_
    classic() +
  labs(y = "Count", x = "Time (years)")

# histogram 1
hist(dens.dat$time, breaks = 15, freq = F, xlab = 'Time', ylim = c(
    0, 0.2), ylab = 'Relative Frequency', main = 'Histogram of
    Survival Times')
lines(density(dens.dat$time, na.rm = T, from = 0, to = 30))

#histogram 2
hist(dens.dat$time, # histogram
     col="Dodger Blue", # column color
     border="black",
     prob = TRUE, # show densities instead of frequencies
     xlab = "Time",
     ylab = "Relative Frequency",
     xlim = c(0, 30),
     ylim = c(0, 0.28),
     breaks=30,
     main = "Survival Time Density")
lines(density(dens.dat$time, from = 0, to = 30), # density plot
      lwd = 2, # thickness of line
      col = "red")



################################################################
#### Density plots by federal states #########################
################################################################

#### distribution of time to event by state ###
```

```
dist <- subset(dat, event==1)
dist$yeargroup <- as.factor(dist$firstyear)

ggplot(
  dist,
  aes(x = dist$time, y = dist$state, fill=dist$state)) +
  geom_density_ridges_gradient(
    aes(fill = ..x..), scale = 2.5, size = 0.3) +
  scale_fill_gradientn(
    colours = c("#0D0887FF", "#CC4678FF", "#F0F921FF"),
    name = "Duration") +
  theme_ridges() +
  labs(title = 'Density of Time to Event', x = "Time", y = "States"
    )

rm(dist)


#### distribution of hhinc by State ###
dist <- subset(dat, hhinc<90000)
dist$yeargroup <- as.factor(dist$firstyear)

# gradient color style
ggplot(
  dist,
  aes(x = dist$hhinc, y = dist$state, fill=dist$region)) +
  geom_density_ridges_gradient(
    aes(fill = ..x..), scale = 2.5, size = 0.3) +
  scale_fill_gradientn(
    colours = c("#0D0887FF", "#CC4678FF", "#F0F921FF"),
    name = "Duration") +
  theme_ridges() +
  labs(title = 'Density of Household Income ',
    x = "Household Income", y = "States")

# alternating color style
ggplot(
  dist,
  aes(x = hhinc, y = state, fill = state)) +
  geom_density_ridges(scale = 2.5) + theme_minimal() +
  scale_fill_cyclical(values = c("blue", "green")) +
  labs(title = 'Density of Household Income',
    x = "Household Income", y = "States")

rm(dist)
```

```
###############################################################
###############################################################
### Functions #################################################
###############################################################
###############################################################


### Function for ggsurvplot_combine ###



# store survplot object and choose functional argument (default is
   Survival Function)

nonparametricKurves <- function(fun=NULL) {
   ggsurvplot_combine(kmfh.all, data=dat, conf.int=T,
                      legend.labs=c("KM", "Fleming-Harrington")
   , legend.title="Model",
                      fun=fun,
                      risk.table=F,
                      cumcensor=FALSE,
                      censor=FALSE,
                      linetype=c(1,1),
                      size = 0.3)}



### Function for Kaplan Meier Curves by Strata ###



# Storing survObject, labs= category description, Legend Title
kmGroupKurves <- function(labs,title,line=c(1,1),conf=T){
  ggsurvplot(wide.fit, conf.int=conf,
               legend.labs=labs, legend.title=title,
               censor=F,
               palette = "strata",
               risk.table = T,
               pval=TRUE,
               pval.method = TRUE,
               log.rank.weights = "S2",
               risk.table.height=.25,
               ylim=c(0,1),
               xlim=c(0,30),
               surv.median.line="hv",
               linetype=line,
               size = 0.5)}
```

```
################################################################
################################################################
#### ANALYSIS #################################################
################################################################
################################################################

#Note1: All scripts and SOEP data files need to be in the same
    directory
#Note2: .path.R file needs to be specified by the user
#Note3: This script needs to load functions.R script first

#Structure:
#0: Set-up
#1: Comparison of KM/FH estimators
#2: KM by strata
#3: Cox PH model
#4: Cox PH model diagnostics
#5: Comparison plots (survival curve, hazard function)
#6: Comparison tables (estimated coefficients)

################################################################
#### Set Up ###################################################
################################################################

# clear workspace
rm(list=ls())

# setwd(path) in path.R
source(".path.R")
source("functions.R")

# install and load packages
source("packages.R")

# load data
load("datfinal.RDA")

################################################################
#### Comparison Kaplan-Meier&Nelson-Aalen/Fleming-Harrington##
################################################################

# Kaplan-Meier estimator
km.fit <- survfit(Surv(time,event, type="right") ~ 1, data=dat,
    type="kaplan-meier")

# Fleming-Harrington estimator
fh.fit <- survfit(Surv(time,event, type="right") ~ 1, data=dat,
```

```r
    type="fleming-harrington")

kmfh.all <- list(km.fit, fh.fit)

#Survival Function
surv.all <- nonparametricKurves()

#Cumulative Event Function: f(y)=1-y
#cumprop.all <- nonparametricKurves("event")

#Cumulative Hazard Function
cumhaz.all <- nonparametricKurves("cumhaz")

# put all plots in one graph
kmfh.glist <- list(surv.all, cumhaz.all)
arrange_ggsurvplots(kmfh.glist, print = TRUE, ncol = 2, nrow = 1)


############################################################
#### KM by strata ##########################################
############################################################


#### KM by gender ###
# 0 = male, 1 = female
wide.fit <- survfit(Surv(time, event, type="right") ~ gender, data=
    dat)
km.sex <- kmGroupKurves(c("Male", "Female"), "Gender")

rm(wide.fit)


### KM by Metropolitan Area ###
# 0 = Urban, 1 = Rural
wide.fit <- survfit(Surv(time, event, type="right") ~ rural, data=
    dat)
km.urban <- kmGroupKurves(c("Urban", "Rural"), "Metropolitan Area")

rm(wide.fit)


#### KM by married ###
wide.fit <- survfit(Surv(time, event, type="right") ~ married, data
    =dat)
km.marr <- kmGroupKurves(c("No", "Yes"), "Married")

rm(wide.fit)


#### KM by ever_divorced ###
wide.fit <- survfit(Surv(time, event, type="right") ~ ever_div,
    data=dat)
```

```r
km.div <- kmGroupKurves(c("No", "Yes"), "Ever Divorced")

rm(wide.fit)

#### KM by region ####
wide.fit <- survfit(Surv(time, event, type="right") ~ region, data=
    dat)
km.reg <- kmGroupKurves(c("West", "East"), "Region")

rm(wide.fit)

#### KM by migback ####
wide.fit <- survfit(Surv(time, event, type="right") ~ migback, data
    =dat)
km.mig <- kmGroupKurves(c("No", "Yes"), "Migr.Back.")

rm(wide.fit)

#### KM by highinc/lowinc ###

#define highinc variable as above median household income
medinc <- median(as.numeric(dat$hhinc), na.rm=TRUE)
dat.inc <- mutate(dat, highinc = ifelse(dat$hhinc > medinc, 1, 0))
summary(dat.inc$highinc)

#define survival object and fit KM estimator
wide.fit <- survfit(Surv(time, event, type="right") ~ highinc, data
    =dat.inc)
km.inc <- kmGroupKurves(c("Low", "High"),"HH Inc.")

rm(wide.fit, medinc, dat.inc)

#### KM by educ (ISCED 97) ###
wide.fit <- survfit(Surv(time, event, type="right") ~ educ, data=
    dat)
km.edu <- kmGroupKurves(c("Elementary", "Medium", "Higher voc.", "
    High"),"Education", line = c(1,1,1,1), conf=F)

rm(wide.fit)

### KM by cohorts 84-87 and 94-97 ###

dat <- mutate(dat, cohort8494 = ifelse (dat$firstyear<=1987, 1,
    ifelse(dat$firstyear>=1994 & dat$firstyear<=1997, 2, NA)))
summary(dat$cohort8494)
table(dat$cohort8494)
```

```
#define survival object and fit KM estimator
wide.fit <- survfit(Surv(time, event, type="right") ~ cohort8494,
    data=dat)
km.coh <- kmGroupKurves(c("84-87", "94-97"),"Cohorts")

rm(wide.fit)


### generate arranged plots ###

km.glist1 <- list(km.inc, km.mig)
km.plot1 <- arrange_ggsurvplots(km.glist1, ncol = 2, nrow = 1,
            print = FALSE, risk.table.height = 0.25,
            surv.plot.height = 1)


km.glist2 <- list(km.reg, km.urban)
km.plot2 <- arrange_ggsurvplots(km.glist2, ncol = 2, nrow = 1,
            print = FALSE, risk.table.height = 0.25,
            surv.plot.height = 1)


km.glist3 <- list(km.marr,km.div)
km.plot3 <- arrange_ggsurvplots(km.glist3, ncol = 2, nrow = 1,
            print = FALSE, risk.table.height = 0.25,
            surv.plot.height = 1)

#### print 'KM by strata' plots ###

print(km.plot1)
print(km.plot2)
print(km.plot3)
print(km.edu)



################################################################
### Cox Proportional Hazards Regression #######################
################################################################

# survival package to estimate models, survminer package for plots
    and diagnostics

# define survival object
coxsurv <- Surv(dat$time, dat$event, type="right")

# define formula
coxform <- as.formula("coxsurv ~ hhinc + rural + maxedu + region +
    migback + married + ever_div")

# estimate Cox regression
```

```r
cox.ph <- coxph(coxform, data=dat)
summary(cox.ph)

# Cox PH model table
stargazer(cox.ph)
displayCoxPH(cox.ph, cap = "", dig.coef = 3, dig.p = 2)

# Forest plot of results
dat <- within(dat,{
  rural <- factor(rural, labels = c("urban", "rural"))
  region <- factor(region, labels = c("west", "east"))
  migback <- factor(migback, labels = c("No", "Yes"))
  married <- factor(married, labels = c("No", "Yes"))
  ever_div <- factor(ever_div, labels = c("No", "Yes"))
})
cox.ph <- coxph(coxform, data=dat)

ggforest(cox.ph)



################################################################
#### Cox Model Diagnostics #####################################
################################################################


### testing proportional hazards assumption (Schoenfeld) ###

# Schoenfeld test
coxtest <- cox.zph(cox.ph, transform = "km")
coxtest
stargazer(coxtest$table, out = "schoenfeld.tex")

# graphical Schoenfeld test of cox ph assumption #

# scaled Schoenfeld plots for two selected variables
ggcoxzph(coxtest, resid = T, point.col="lightcoral", point.alpha =
    0.4, var=c("hhinc","maxedu"))

# scaled Schoenfeld plots for all variables
ggcoxzph(coxtest, point.alpha = 0.4, point.col="lightcoral")

### testing for influential Observations ####

# use Delta-Beta residuals to detect influential observations
    scaled by standard errors of coefficients
ggcoxdiagnostics(cox.ph, type = "dfbetas", ox.scale= "observation.
    id", hline.col = "darkgreen", hline.alpha = 0.5, point.alpha = 0
    .4, point.col = "lightcoral", sline.alpha = 0.4, sline.col = "
```

```
    dodgerblue")

rm(cox.ph, coxtest, coxsurv, coxform)


############################################################
#### KM, Cox PH and Parametric distributions plot ###########
############################################################

# define survival object
coxparm <- Surv(dat$time, dat$event, type="right")


# define model formula
parmform <- as.formula("coxparm ~ hhinc + rural + maxedu + region +
    migback + married + ever_div")


# Kaplan-Meier estimator
kapm <- survfit(coxparm ~ 1, data=dat)


# fortify puts survival table from kapm object into a data frame
kap.dat <- fortify(kapm)


# Cox PH model
cox.ph <- coxph(formula=parmform, data=dat)
summary(cox.ph)


# fortify cox model output
cox.dat <- fortify(survfit(cox.ph, conf.int = F))


# Flexible splines (Royston and Parmar 2002)
flex.spline <- flexsurvspline(coxparm ~ 1, data = dat, k = 2,
  scale = "odds")


# Weibull distribution
weibull <- flexsurvreg(formula = parmform, data = dat, dist = "
   weibull")


# Exponential distribution
expo <- flexsurvreg(formula = parmform, data = dat, dist = "exp")


# Log-Logistic distribution
loglog <- flexsurvreg(formula = parmform, data = dat, dist = "
   llogis")


# Log-normal distribution
lnormal <- flexsurvreg(formula = parmform, data = dat, dist = "
   lnorm")
```

```r
### plot all curves together ###

# Note: plot for Cox PH model is (by default) at average of
    covariates
grid.arrange(
  ggplot(data.frame(summary(expo)), aes(x = time)) +
    geom_line(aes(y = est, col = "Exponential")) +
    geom_line(data = data.frame(summary(weibull)), aes(y = est, col
    = "Weibull")) +
    geom_line(data = data.frame(summary(loglog)), aes(y = est, col
    = "Log-Logistic")) +
    geom_line(data = data.frame(summary(lnormal)), aes(y = est, col
    = "Log-Normal")) +
    geom_line(data = data.frame(summary(flex.spline)), aes(y = est,
    col = "Flexible Splines")) +
    geom_step(data = kap.dat, aes(x=time, y=surv, colour = "Kaplan-
    Meier"), size = 0.37)+
    geom_step(data = cox.dat, aes(x=time, y=surv, colour = "Cox PH"
    ), size = 0.37)+
    labs(x = "Time (years)", y = "Survival Probability", col = "
    Models") + theme_classic(),
  ggplot(data.frame(summary(expo, type = "hazard")), aes(x = time))
     +
    geom_line(aes(y = est, col = "Exponential")) +
    geom_line(data = data.frame(summary(weibull, type = "hazard")),
     aes(y = est, col = "Weibull")) +
    geom_line(data = data.frame(summary(loglog, type = "hazard")),
   aes(y = est, col = "Log-Logistic")) +
    geom_line(data = data.frame(summary(lnormal, type = "hazard")),
     aes(y = est, col = "Log-Normal")) +
    geom_line(data = data.frame(summary(flex.spline, type = "hazard
    ")), aes(y = est, col = "Flexible Splines")) +
    labs(x = "Time (years)", y = "Hazard Function", col = "Models")
     + theme_classic(),
  ncol = 2
)


# only survival curves as single plot
ggplot(data.frame(summary(expo)), aes(x = time)) +
  geom_line(aes(y = est, col = "Exponential")) +
  geom_line(data = data.frame(summary(weibull)), aes(y = est, col =
    "Weibull")) +
  geom_line(data = data.frame(summary(loglog)), aes(y = est, col =
    "Log-Logistic")) +
  geom_line(data = data.frame(summary(lnormal)), aes(y = est, col =
    "Log-Normal")) +
  geom_line(data = data.frame(summary(flex.spline)), aes(y = est,
```

```r
    col = "Flexible Splines")) +
  geom_step(data = kap.dat, aes(x=time, y=surv, colour = "Kaplan-
    Meier"), size = 0.37)+
  geom_step(data = cox.dat, aes(x=time, y=surv, colour = "Cox PH"),
     size = 0.37)+
  labs(x = "Time (years)", y = "Survival Probability", col = "
    Models") + theme_classic()

# only hazard functions as single plot
ggplot(data.frame(summary(expo, type = "hazard")), aes(x = time)) +
  geom_line(aes(y = est, col = "Exponential")) +
  geom_line(data = data.frame(summary(weibull, type = "hazard")),
    aes(y = est, col = "Weibull")) +
  geom_line(data = data.frame(summary(loglog, type = "hazard")),
    aes(y = est, col = "Log-Logistic")) +
  geom_line(data = data.frame(summary(lnormal, type = "hazard")),
    aes(y = est, col = "Log-Normal")) +
  geom_line(data = data.frame(summary(flex.spline, type = "hazard")
    ), aes(y = est, col = "Flexible Splines")) +
  labs(x = "Time (years)", y = "Hazard Function", col = "Models") +
     theme_classic()




###############################################################
#### Cox PH and Parametric distributions tables #############
###############################################################

# stargazer only compatible with survreg, not flexsurvreg

# define survival object
coxparm <- Surv(dat$time, dat$event, type="right")

# define model formula
parmform <- as.formula("coxparm ~ hhinc + rural + maxedu + region +
    migback + married + ever_div")

# Cox PH model
cox.ph.tab <- coxph(formula=parmform, data=dat)
summary(cox.ph.tab)

# Weibull distribution
weibull.tab <- survreg(formula = parmform, data = dat, dist = "
    weibull")

# Exponential distribution
expo.tab <- survreg(formula = parmform, data = dat, dist = "exp")
```

```r
# Log-Logistic distribution
loglog.tab <- survreg(formula = parmform, data = dat, dist = "
    loglogistic")

# Log-normal distribution
lnormal.tab <- survreg(formula = parmform, data = dat, dist = "
    lognormal")

# Results table
stargazer(cox.ph.tab, weibull.tab, loglog.tab, lnormal.tab, align=F
    , out = "comparison.tex")

# AIC for parametric models (model selection)

AICs <- matrix(data = NA, nrow = 4, ncol = 2)
AICs[1, 1] <- weibull$AIC
AICs[1, 2] <- weibull$loglik

AICs[2, 1] <- expo$AIC
AICs[2, 2] <- expo$loglik

AICs[3, 1] <- loglog$AIC
AICs[3, 2] <- loglog$loglik

AICs[4, 1] <- lnormal$AIC
AICs[4, 2] <- lnormal$loglik

rownames(AICs) <- c("Weibull", "Exponential", "Log-Logistic", "Log-
    Normal")
colnames(AICs) <- c("AIC", "Log-Likelihood")
t(AICs)

#Log-normal has best fit
stargazer(t(AICs), out="aic.tex")
```

Statutory Declaration

We hereby confirm that we have authored this Seminar paper independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Berlin, August 12th, Alice Drube & Konstantin Göbler & Chris Kolb & Richard v. Maydell