

Determinants of the Transition Time into Homeownership Using Survival Analysis

Alice Drube
Konstantin Göbler
Chris Kolb
Richard v. Maydell

Statistical Programming Languages
Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin
<http://lvb.wiwi.hu-berlin.de>



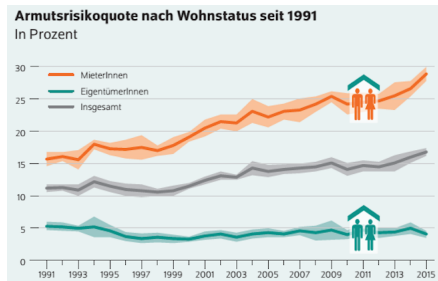
Outline

1. Introduction
2. Survival Analysis Basics
3. Data and Descriptives
4. Application in R
5. Discussion
6. Conclusion



Motivation

- Grabka and Goebel (2018) recently found in a DIW study: **risk of poverty** increasing sharply for renters compared to homeowners



- The share of homeowners increased to almost 50% in last 25 years
- Risk of poverty for homeowners did not change substantially since 1991
- Risk of poverty for renters increases almost twofold since 1991 to roughly 30%
- Particularly, this concerns young adults in their mid 30's



- At the same time, Sagna and Voigtländer (2018) found that the number of first owners is declining since 2016
- Again providing evidence for younger individuals being primarily affected
⇒ *Transitioning from renting to owning appears to be a complex process, demanding for deeper investigation*

Research Question

Which factors determine duration of transition from renting to owning?



Survival Analysis Basics

- Class of statistical models analyzing the **time to event**, *i.e.* the time that elapses between occurrence of two events
- Need to account for **right-censoring** of data (at time of observation survival may not have ended)
 - ⇒ standard models (e.g. OLS) are inconsistent
 - ⇒ Survival Analysis methods
- Main interest is in risk of 'failure' event (hazard) rather than expected survival time (OLS)
- **Central question:** What is the probability that survival ends given that it has not ended yet?



Important concepts in Survival Analysis:

1. **Survival Function:**

$$S(t) := \mathbb{P}(T > t) = \int_t^{\infty} f(u)du$$

i.e. the probability of 'surviving' longer than t .

2. **Hazard Function:**

$$\lambda(t) := \lim_{\Delta t \rightarrow 0} \frac{\mathbb{P}(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{1 - F(t)}$$

i.e. the instantaneous failure rate.

3. The survival function can be recovered from the hazard rate (sufficient to estimate one and get other 'for free'):

$$S(t) = \exp \left[- \int_0^t \lambda(u) du \right]$$



Survival Data Structure

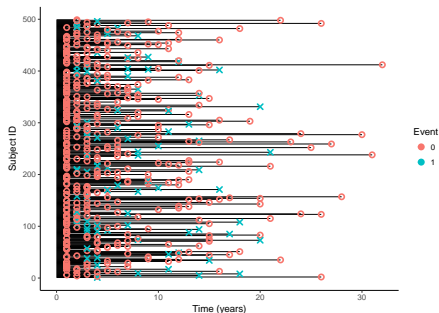


Figure 1: Visualization of SOEP survival data. For better visibility, only 500 randomly drawn individuals are visualized ($n = 3,928$).

 DataVisualization

Estimating and Testing Survival Models

- **Non-parametric methods** estimate the survival curve and hazard function without any strict distributional assumptions
 - ▶ Kaplan-Meier estimator (workhorse model)
 - ▶ Nelson-Aalen estimator
- Stratify analysis by subgroups (e.g. by household income levels, geographical region, educational attainment etc.)
- Test for group differences in survival curves using **non-parametric Log-rank (Mantel-Cox) test**:

$$H_0 : S_1(t) = S_2(t) \quad \text{vs.} \quad H_1 : S_1(t) \neq S_2(t)$$



Estimating and Testing Survival Models

- **Cox Proportional Hazard Model:** popular semi-parametric model, specifying the hazard function as

$$\lambda(t|x_i) = \lambda_0(t) \exp(x_i' \beta)$$

- ▶ estimation of $\lambda_0(t)$ not needed
 - ▶ consistent partial MLE for β obtainable (Cox 1972) without specification of $\lambda_0(t)$
 - ▶ proportional hazards assumption: ratio of hazards of any two individuals is constant over time (**Schoenfeld residual test**)
 - ▶ test significance of covariates using likelihood-based testing
- Another alternative are fully **parametric models**, assuming a specific distribution of the survival times (e.g. Weibull)



The GSOEP

- The German Socio-Economic Panel (Wagner, Frick, and Schupp 2007) is a representative longitudinal survey.
 - ▶ started in 1984 (1991 for East Germany)
 - ▶ follows over 20,000 individuals in over 11,000 households
 - ▶ published annually by the DIW
- 16,099 individuals & 5,863 households in 1984
- 24,649 individuals & 11,155 households in 2013
- Contains large array of items on socio-economic indicators, living conditions, labor market status, and attitudes.



Data Cleaning

We construct our data set in the following way using the tidyverse package collection, esp. dplyr:

1. Match information on homeownership status of household with data on household head
2. Restrict sample to individuals
 - ▶ who are observed from age 24/25 onward
 - ▶ who start out as renters, not owners
 - ▶ who did not inherit a dwelling
3. Identify 'failure events' by homeownership status changing from renting to owning
4. Identify 'time to event' by 'year of event' minus 'year of first observation'



Data Cleaning

- Finally, a set of (time-invariant) covariates is included to investigate the determinants of individual survival times in rent:
 - ▶ Gender
 - ▶ Marital Status
 - ▶ Having had a divorce
 - ▶ Migration background
 - ▶ Pre-Government Household Income
 - ▶ Years of Education
 - ▶ State of Residence
 - ▶ East/West Germany
 - ▶ Spatial Type (city/rural)
 - ▶ ISCED97 Education Variable



Missing Data Pattern

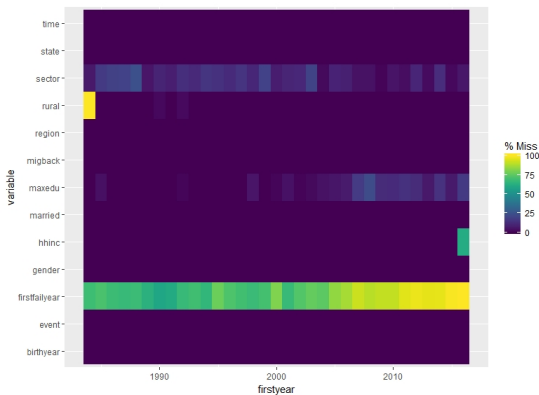


Figure 2: Pattern of missing data before random forest imputation

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	0.27	0.45	0	0	1	3,928
Rural	rural	0.22	0.41	0	0	1	3,928
Married	married	0.05	0.21	0	0	1	3,928
Ever Divorced	ever_div	0.31	0.46	0	0	1	3,928



Density of Survival Times by Federal State

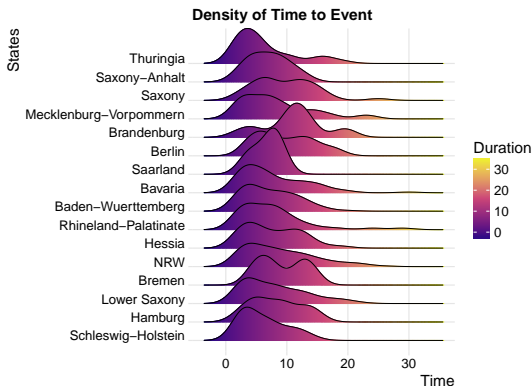


Figure 3: Survival times of non-censored cases by state  SurvivalDens

Survival Analysis in R

Exhaustive overview over relevant R routines for survival analysis provided by CRAN Task Review:

<https://cran.r-project.org/web/views/Survival.html>

- ▣ `survival` package: core package for non-parametric and parametric survival analysis
- ▣ `survminer` package: includes functions for 'nicer' plots
- ▣ `survreg` package: estimation of parametric proportional hazards and accelerated failure time models (e.g. Weibull, Log-normal, Exponential, ...)



Small code snippet

```
#installing packages, loading omitted
install.packages(c("survival", "survminer"))

#define survival object and fit KM estimator
km.fit<-survfit(Surv(time, event) ~ educ, data=
  dat)

#plot survival curves
surv<-ggsurvplot(km.fit, conf.int=T, pval=T)
print(surv)

#Cox PH model
cox.ph<-coxph(Surv(time, event) ~ hhinc+rural+
  maxedu+region+migback , data=dat)
print(cox.ph)
```



K-M stratified by Education Levels

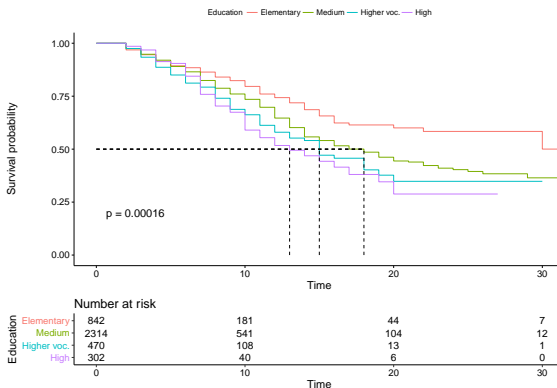



Figure 4: Non-parametric estimation of survival function  GroupKM

K-M stratified by HH Inc and Migr. Back.

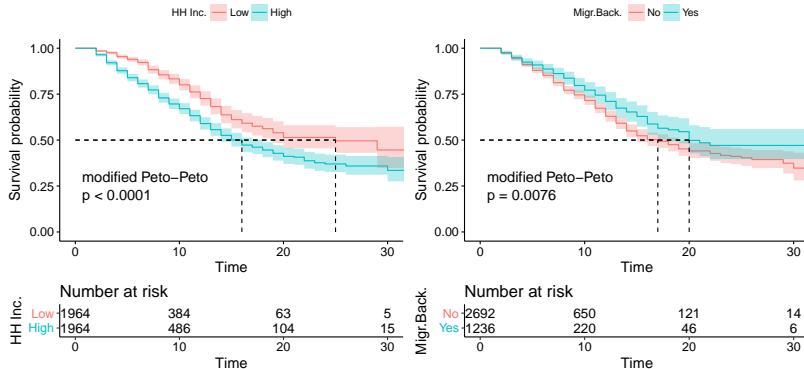


Figure 5: Non-parametric estimation of survival function  GroupKM

Cox PH Regression

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

Table 1: Cox PH regression. Standard errors are in parentheses. The hazard ratios depict the (time-const.) ratio of hazard functions for a one-unit increase in the predictor.



Comparison of Parametric Approaches

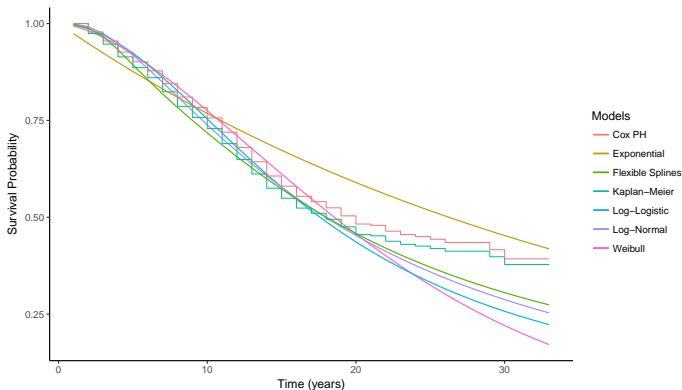



Figure 6: Comparison of estimated survival functions  ComparSurv

Comparison of Parametric Approaches

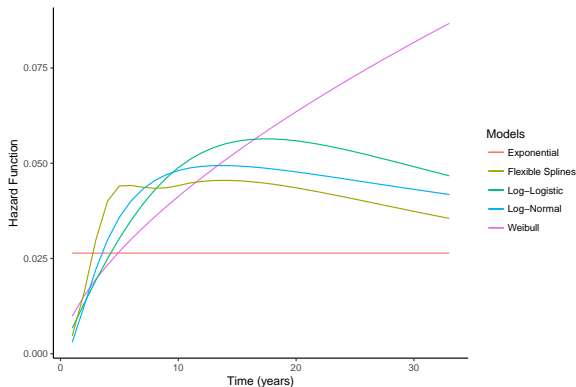


Figure 7: Comparison of estimated hazard functions  ComparHazard

Comparison of Parametric Approaches

- Flexible splines and non-/semi-parametric approaches serve as benchmark for parametric models
- Clearly, the Exponential distribution is a bad fit due to its restrictive parametrization (constant hazard)
- Graphical inspection concludes best fit for either Log-logistic or Log-normal distribution
- Model selection via Akaike's Information Criterion confirms Log-normal as best parametric fit



Discussion

- Unsurprisingly, socio-economic variables such as household income are a strong predictor of transitioning into homeownership (increases hazard by 1.8% per 1,000 Euro yearly household income)
- Likewise, having a migration background is estimated to be about as detrimental as living in East Germany, regarding the transition time.
- Lastly, marriage and divorce also greatly influence housing tenure choice







Conclusion

- Different approaches from the statistical field of survival analysis have proven to be powerful tools to understand the determinants and mechanisms of the transition time from rent at age 25 into first-time homeownership
- Analyzing data from Germany's largest annual survey, the GSOEP, we find a range of socio-economic, as well as spatial, variables significantly determine the transition time
- Consistency of results across models strengthen credibility of estimates



Bibliography

-  Cox, D. R. 1972. "Regression Models and Life-Tables". *Journal of the Royal Statistical Society. Series B (Methodological)* 34 (2): 87–22.
-  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.
-  Sagna, P., and M. Voigtländer. 2018. "Die Zahl der Ersterwerber sinkt weiter". *IW-Kurzbericht*.
-  Wagner, G., J. Frick, and J. Schupp. 2007. "The German Socio-Economic Panel study (SOEP)-evolution, scope and enhancements".