

# Duration Models

Paul Goldsmith-Pinkham

February 21, 2023

# Today's topic: duration models

- First question is: what's a duration model?
- Second question: why do we care?
- Third question: what are ways to estimate them? What are estimands that are identified in these settings?

# What's a duration model?

- A duration model is just what it sounds like – a model relating to duration of an event
- Why would we need a special model for this?
  1. Data measurement: measuring durations accurately is challenging!
  2. Estimation: the mapping between theory and estimation can require more sophisticated models
- We'll start by just discussing what's special about durations

## First, some examples

- Lancaster (1979) - unemployment spells
  - $t = 0$ : unemployment begins
  - Spell ends: employment
- Galiani, Gertler and Schargrotsky (2005) - privatizing water service
  - $t = 0$ : the year 1990
  - Spell ends: water service privatized
- Palmer (2015) - mortgage default
  - $t = 0$ : Mortgage origination
  - Spell ends: mortgage default
- Rose (2020) - supervised release
  - $t = 0$ : Release
  - Spell ends: New arrest or revocation

## More examples

- Engle and Russell (1998) – Irregularly spaced transaction data
  - $\{t_0, t_1, \dots, t_n, \dots\}$  denote arrival times of transactions
  - Condition on previous events – “autoregressive conditional duration” models
- Carlson et al. (2015) - bankruptcy
  - $t = 0$  retirement from NFL
  - Spell ends: bankruptcy filing
- Goldsmith-Pinkham & Gilbukh (2021) - moving houses
  - $t = 0$  buy a home
  - Spell ends: buy a new home

## Running example

- One very common duration is the duration of tenure in housing
- Let  $Y_i$  denote the length of time that individual  $i$  lives in their home.
  - If we have better data, we can even consider  $Y_{is}$  to be the length of time that individual  $i$  lives in their home in spell  $s$
  - Multiple spells, e.g. panel data
- A number of notable things could affect this tenure:
  - Age of the individuals
  - The housing cycle
  - The business cycle
  - Whether or not they are homeowners

## Types of Duration Data we observe

- The “truth”: A single duration observed for person  $i$ ,  $Y_i \in [0, T]$ 
  - In theory, the duration could be unbounded  $T = \infty$ , but could be maximally bounded (e.g. max lifetime of human)
- What kind of data do we observe?
  - First example is best case but given unbounded nature of  $T$ , unrealistic for everything
  - For a given observation, we see spell start and spell end








## Types of Duration Data we observe

- Sometimes it's even less information in our sampling scheme
- We know when the spells began, but can only identify whether or not exits occurred as of period  $c_i$ 
  - Easy to envision data sampling like this – checking in on a pool of individuals, and identifying whether they stuck around
- This leaves us with less information, but so long as the censoring (e.g. the time of check-in) is sufficiently random, also addressable

Case	Start	End
Full	$t_0$	$t_1$
Right-Censor	$t_0$	$\min\{t_1, c_i\}$
Indicator	$t_0$	$1(Y_i < c)$



The diagram illustrates a timeline from  $t=0$  to  $t=T$ . A blue horizontal line segment starts at  $t=0$  and ends at  $t=c_i$ , representing the observed duration. An orange horizontal line segment starts at  $t=c_i$  and ends at  $t=T$ , representing the censored period. Vertical tick marks are placed at  $t=0$ ,  $t=c_i$ , and  $t=T$ .

## Stock Sampling vs. Flow Sampling

- In these cases so far, we observe when the start of the duration begins
  - This is called flow sampling
- An alternative sampling procedure samples from the stock of existing individuals
  - E.g. Stock sampling
- What issues does this create? Two scenarios:
  - If we observe how long the duration lasted at time of sampling (e.g. the start time), we still need to account for the sample selection from stock sampling
  - If we don't observe the start time, creates a version of left-censoring
- **Left-censoring creates serious problems** – need to make stronger assumptions

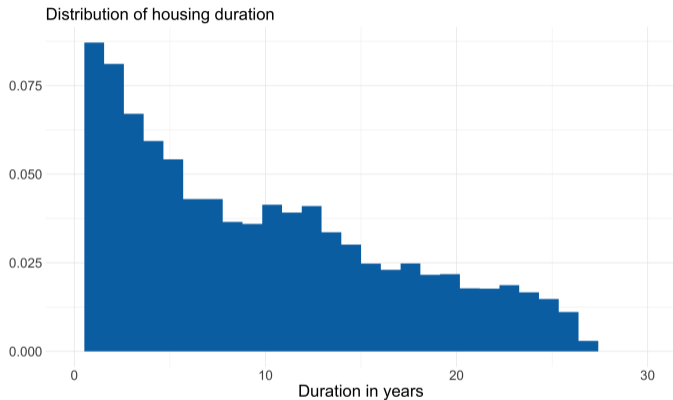
Sampling	Case	Start	End	Adjustment
Flow	Full	$t_0$	$t_1$	No
Flow	Right-Censor	$t_0$	$\min\{t_1, c_j\}$	Yes
Flow	Indicator	$t_0$	$1(Y_j < c)$	Yes
Stock	Full	$t_0$	$t_1$	Yes
Stock	Right-Censor	$t_0$	$\min\{t_1, c_j\}$	Yes
Stock	Indicator	$t_0$	$\min\{t_1, c_j\}$	Yes

## Key takeaway

- Understanding the sampling structure of your data is always important
  - Particularly important with duration data
- However, censoring problems are not unique to duration data
  - E.g., wage data can be censored/truncated due to reservation wages or survey measurement
  - However, in duration data, right-censoring is quite common
- These are important features to consider for understanding the data generating process for your sample (and the population)
- However, a more important question is what are you interested in?
  - E.g. what is your estimand?
  - Effect of a treatment on average length of duration? Median duration?
  - Consider  $Y_i = \alpha + T_i\beta + \epsilon_i$  – this is well-defined when  $T_i$  is randomly assigned, but censoring still causes issues

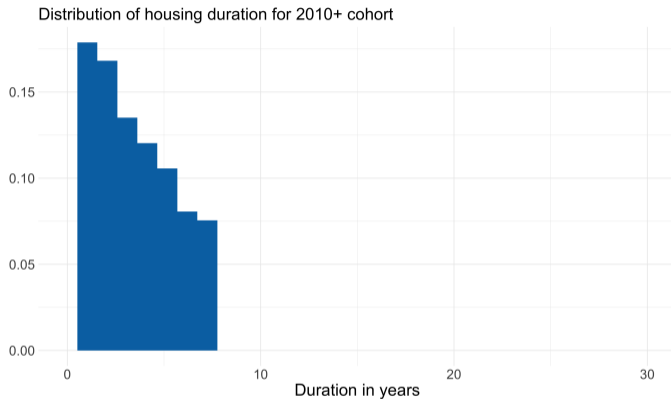
# Consider the example of housing

- Length of time between housing transactions
- Sample is drawn in 2017m8, but we see every transaction
  - Implication: data is censored at 2017m8, which creates different censoring horizons depending on when the home was last bought
- We can first examine full distribution of durations



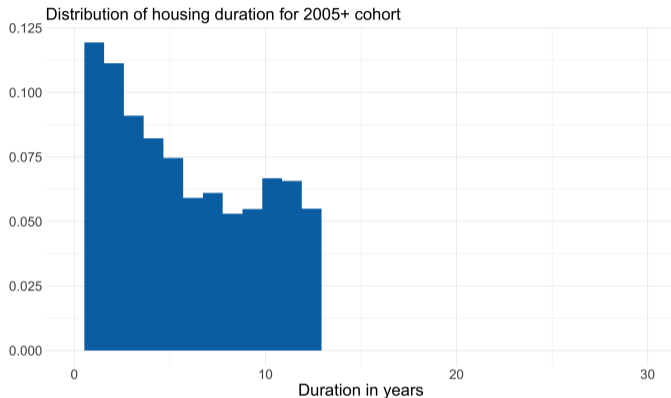
## Consider the example of housing

- Length of time between housing transactions
- Sample is drawn in 2017m8, but we see every transaction
  - Implication: data is censored at 2017m8, which creates different censoring horizons depending on when the home was last bought
- If we focus on 2010+ cohorts, truncation problem is obvious, but shape is not



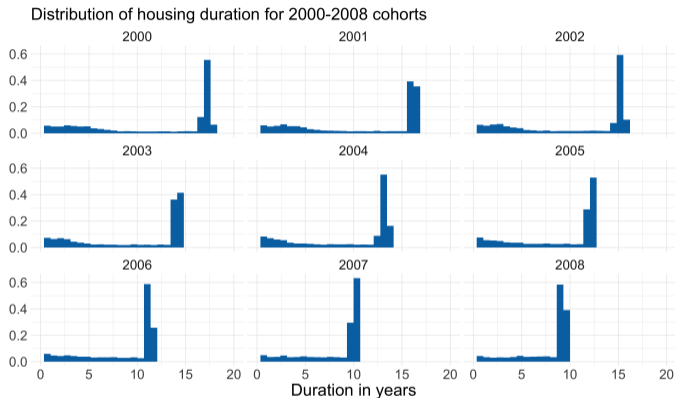
# Consider the example of housing

- Length of time between housing transactions
- Sample is drawn in 2017m8, but we see every transaction
  - Implication: data is censored at 2017m8, which creates different censoring horizons depending on when the home was last bought
- If we focus on 2005+ cohorts, it's clear there's heterogeneity



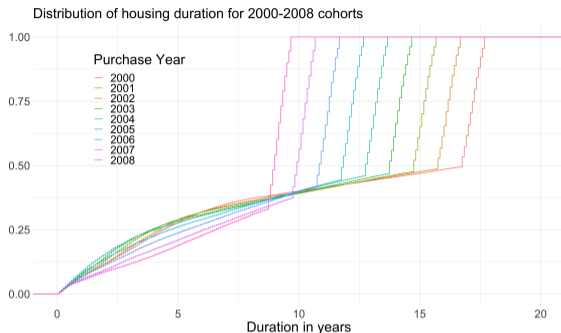
# Consider the example of housing

- Length of time between housing transactions
- Sample is drawn in 2017m8, but we see every transaction
  - Implication: data is censored at 2017m8, which creates different censoring horizons depending on when the home was last bought
- Censoring issue is very apparently within a given year



# Consider the example of housing

- Can we calculate the average duration, without further assumptions?
  - Purely non-parametrically? No. Even with random sampling, we don't know the DGP and the average could be completely unbounded
  - Can explore this further in partial identification
- If we are willing to make more assumptions (next), then yes!
- Before we do that – worth recognizing that there are other estimands that we can identify
  - E.g., the quantiles – for 2000 cohort, the median





## Pivoting to hazard models

- We'll now discuss some parametric ways that papers address these problems
- Duration modeling is, in many cases, focused on *hazard* modeling. Why?
  - Hazard has natural economic theory tie-ins
  - Adjusts appropriately for the “survival” of individuals
- There is nothing more powerful in these settings than anything else we've studied – by using parametric models, you are able to account for data issues, but require additional assumptions

## Quick aside: some formal definitions

- Let  $F(y) = \Pr(Y \leq y)$  be the probability of a duration no longer than  $y$ , and  $f(y)$  the corresponding density
- Then,  $S(y) = 1 - F(y)$  is known as the survival function (the probability you'll survive until  $y$ )
- This lets us define the hazard function  $h(y) = \frac{f(y)}{S(y)}$ , which is the probability of an event occurring, condition on surviving until  $y$ .
- Key features of the hazard:
  - Conditions on the population surviving until  $y$  (rather than everyone)
  - Can be time varying
  - Summarizes all characteristics of  $F$
- Effectively think of it as a transformation of the distribution

# Why the hazard function? (Van Den Berg (2001))

- So why use a hazard model?  
E.g. extra structure
- Van Den Berg lays out some reasons in his Handbook chapter on duration modeling
  - The hazard is a concise way to summarize the state of the remaining sample

The hazard function is the focal point of econometric duration models. That is, properties of the distribution of  $T$  are generally discussed in terms of properties of  $\theta$ . There are two major reasons for this. First, and most importantly, this approach is dictated by economic theory. In general, theories that aim at explaining durations focus on the rate at which the subject leaves the state at duration  $t$  given that he has not done so yet. In particular, they explain the hazard at  $t$  in terms of external conditions at  $t$  as well as the underlying economic behavior of the subjects that are still in the state at  $t$ . Theoretical predictions about a duration distribution thus run by way of the hazard of that distribution. It is obvious that if the completion of a spell is at least partly affected by external conditions that change over time (e.g., due to external shocks), and if one attempts to describe behavior of the subject over time in a changing environment, then it is easier to think about the rate of leaving at  $t$  given that one has not done so than to focus on the unconditional rate of leaving at  $t$ . In the next section we provide some examples of such theories.

# Why the hazard function? (Van Den Berg (2001))

- So why use a hazard model?  
E.g. extra structure
- Van Den Berg lays out some reasons in his Handbook chapter on duration modeling
  - The hazard is a concise way to summarize the state of the remaining sample
  - More effectively captures time-varying characteristics
  - Deals well with right-censoring

It is often stated that a major advantage of using the hazard function as a basic building block of the model is that it facilitates the inclusion of time-varying covariates. This is, of course, part of the argument of the previous paragraph; it reformulates the issue from the point of view of a builder of reduced-form models.

The second major advantage of using the hazard function as the basic building block of the model is entirely practical. Real-life duration data are often subject to censoring of high durations. In that case it does not make sense to model the duration distribution for those high durations.

## Simple hazard example

- Imagine that people move houses because of life events, and they arrive randomly with a random rate  $\theta(t)$ ,
  - the expected number of life events in a short time period is  $\theta(t)dt$
  - For now, assume it's constant – e.g.  $\theta = \theta(t)$

- This implies a distribution of life events that is exponential with mean  $1/\theta$ :

$$f(y) = \theta \exp(-y\theta) \text{ for } y > 0, F(y) = 1 - e^{-\theta y}$$

- This distribution is extremely nice, since it has the lack of memory property, e.g.

$$E(Y - c | Y > c) = 1/\theta,$$

irrespective of  $c$

- Hence, the hazard rate is exactly  $\theta$ !

## Hazard modeling data sampling

- We can use this setup to study our duration cases previously
- Consider our data sampling and the likelihood:
  - With fully observed flow sampling, the likelihood is

$$\begin{aligned}L(\theta) &= \prod_{i=1}^n f(y_i|\theta) = \prod_{i=1}^n h(y_i|\theta)S(y_i|\theta) \\ &= \prod_{i=1}^n \theta e^{-\theta y} \text{ under exponential}\end{aligned}$$

- With right censoring (not censored  $\rightarrow d_i = 1$ ) and flow sampling, the likelihood is

$$L(\theta) = \prod_{i=1}^n f(y_i|\theta)^{d_i} S(c_i|\theta)^{1-d_i} = \prod_{i=1}^n h(y_i|\theta)^{d_i} S(y_i|\theta)^{d_i} S(c_i|\theta)^{1-d_i}$$

## Hazard modeling data sampling

- With stock sampling, where you sample from the *stock* of individuals (rather than the flow)
  - A given draw is sampled to have lived for  $s_i$  periods
- Then, the likelihood is

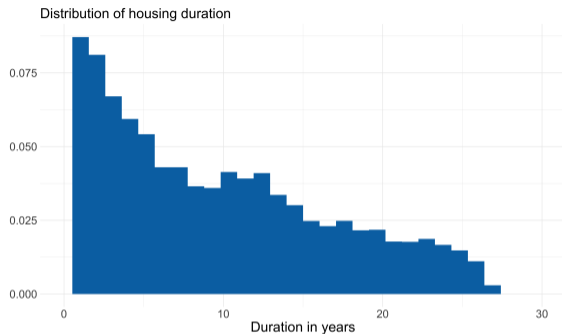
$$L(\theta) = \prod_{i=1}^n \frac{f(y_i|\theta)}{S(s_i|\theta)} = \prod_{i=1}^n h(y_i|\theta) \frac{S(y_i|\theta)}{S(s_i|\theta)}$$

- With right censoring (not censored  $\rightarrow d_i = 1$ ) such that we do not track the observations, the likelihood is

$$L(\theta) = \prod_{i=1}^n \left( \frac{f(y_i|\theta)}{S(s_i|\theta)} \right)^{d_i} (S(s_i|\theta))^{(1-d_i)}$$

# Non-parametric estimations of survival with censoring

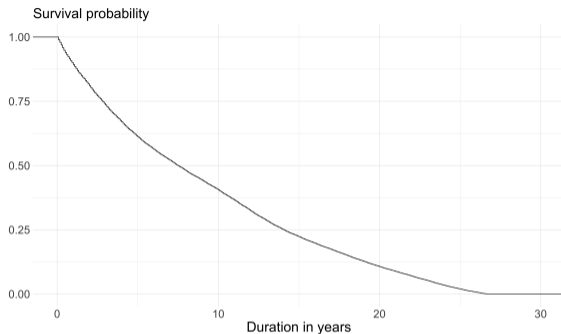
- In the full sample of the housing example, we were plotting the density of the variable  $t_i = \min\{y_i, c_i\}$ 
  - e.g., the true failure time or when it is censored





# Non-parametric estimations of survival with censoring

- In the full sample of the housing example, we were plotting the density of the variable  $t_i = \min\{y_i, c_i\}$ 
  - e.g., the true failure time or when it is censored
- However there's a lot of censoring in this full sample. How accurately does this map to the probability of someone staying in a home?
  - E.g. can I use this to estimate  $S(t) = Pr(Y_i > t)$ ?
- Short answer: not directly. Need to adjust for censoring and can do so non-parametrically
  - We'll use the Kaplan-Meier estimator



## Kaplan-Meier: non-parametric estimations of survival

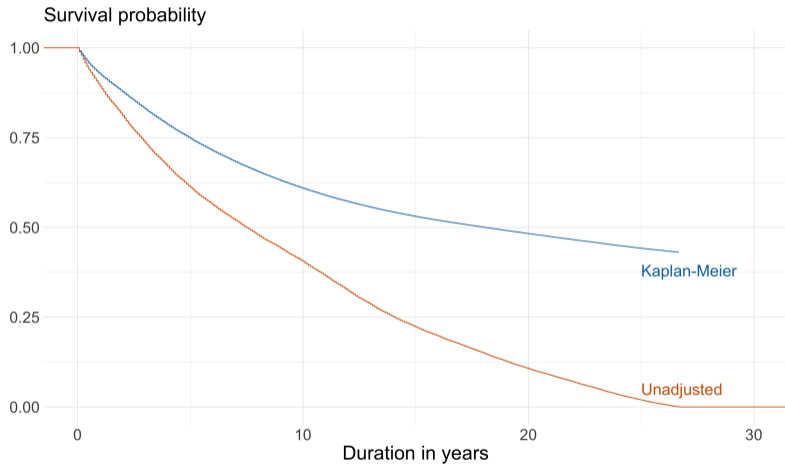
- Kaplan-Meier estimator exploits the fact that the survival up to period  $t$ ,  $S(t)$ , can be thought of as the joint probability of  $t$  non-exits in a row:  $S(t) = \prod_{j=1}^t (1 - h(j))$
- Then, this implies that  $f(t) = h_t \prod_{j=1}^{t-1} (1 - h(j))$  and  $f(1) = h(1)$ .
- We need to just estimate  $h(\cdot)$  for every time period. Let  $a_j = \sum_i 1(Y_i \geq t)$ ,  $e_j = \sum_i 1(Y_i = j \cap Y_i \leq c_i)$

$$\begin{aligned} L(h) &= \prod_{i=1}^n f(y_i)^{d_i} S(c_i)^{1-d_i} \\ &= \prod_j h_j^{e_j} (1 - h_j)^{a_j - e_j} \end{aligned}$$

- Our MLE is  $\hat{h}_j = e_j / a_j$ 
  - Can consider standard errors and testing around these estimates

# Kaplan-Meier: non-parametric estimations of survival

- Ignoring the censoring makes a big difference



## Hazard modeling data sampling

- Clearly, ignoring the censoring will bias your estimates - either of  $\theta$  in a parameterized model, or of  $h(t)$  in non-parametric estimates
  - There are two ways one could naively ignore it - toss any data that's censored, or treat the censoring as real data
  - Both will give over-estimates of  $\theta$  in the exponential case
- To see the intuition, let  $t_i = \min(y_i, c_i)$ , and note the likelihood.

$$L(\theta) = \prod_{i=1}^n h(t_i|\theta)^{d_i} S(t_i|\theta)$$

- Now assume the exponential and solve for the MLE  $\hat{\theta} = \bar{d}/\bar{t}$ . Consider how the estimates change if you either throw out data, or mislabel the censoring
  - If you ignore the censoring,  $\bar{d} \rightarrow 1$ , and  $\bar{t}$  doesn't change. Upward bias in  $\hat{\theta}$
  - If you drop the censored obs, then the MLE is  $\bar{\theta} = \sum d_i / \sum d_i t_i$ . Note that the numerator remains unchanged, but the denominator decreases. Upward bias in  $\hat{\theta}$

## Value of hazard modeling

- So far, the main value of the modeling is to add parametric structure to capture the univariate features of duration
  - E.g., we want to know the properties of a censored random variable
- However, in many cases we want to know the effect of some variable on the duration
  - E.g.  $Y = D\beta + \epsilon$
- What is the downside of simply running a regression like above? As Van Den Berg discusses above:
  - Hazard rate will more concisely capture a meaningful characteristic
  - Time-varying characteristics are more easily accomodated (e.g., how does the above linear regression incorporate a changing minimum wage schedule?)
  - The simple linear regression approach doesn't deal with censoring well

## A defense of regression

- An aside in defense of a simple linear regression approach
- While hazard modeling is tightly connected to economic models, it can feel non-transparent (as many non-linear models do)
- Simple linear regression models *can* address censoring in two ways:
  - Indicators of “survived to year  $K$ ”, so long as year  $K$  is not censored
  - quantile regression
- It is possible to use these combinations to do a number of robust analyses
  - In fact, I would highly recommend that any hazard modeling done is also supported by simple linear regression as well
- However any linear regression needs to be *extremely* aware of the data sampling process

## The workhorse of hazard modeling - Proportional Hazard

- In some settings, there is theory driving the hazard modeling, and that should determine your approach
  - in more reduced form settings, you need a workhorse model that is flexible
- In hazard models, this model is the Cox proportional hazards model:

$$\theta(t|x) = \phi(t)\theta_0(x), \quad (1)$$

where  $\theta_0 = \exp(x\beta)$  usually.  $\phi(t)$  is the *baseline* hazard, and gives the underlying shape of the hazard function.

- The characteristics of individuals  $x$ , move around the level of the hazard curve, but do not change the overall shape (e.g.  $\theta_0$  is not directly a function of time, other than through  $x$ )
- In more complicated settings, we can allow for unobserved heterogeneity  $v$  in what is known as the *mixed* proportional hazards model:

$$\theta(x|t, v) = \phi(t)\theta_0(x)v \quad (2)$$

## Unobserved heterogeneity

$$\theta(x|t, v) = \phi(t)\theta_o(x)v \quad (3)$$

- Key result from Lancaster that is easy to understand - hazard rate could be time varying (which is interesting from a theoretical point) or it could be unobserved heterogeneity
- Consider the following example. Suppose there are two types of people in the population, “movers” and “stayers”, movers are share  $p$  and stayers  $1 - p$ . Movers have a of  $\theta_m = 2$ . Stayers have a lower lower rate of  $\lambda_s = 1$ .
- As time goes on, the share of each in the remaining population changes, shifting the overall hazard rate
  - However, it's not that hazard rates are changing, but instead a compositional impact of unobserved heterogeneity
- Using multi-spell data is a way to address this issue, similar to panel data with unobserved heterogeneity
  - See Van Den Berg (2001) and Rose (2020) for details on cites



## Estimation

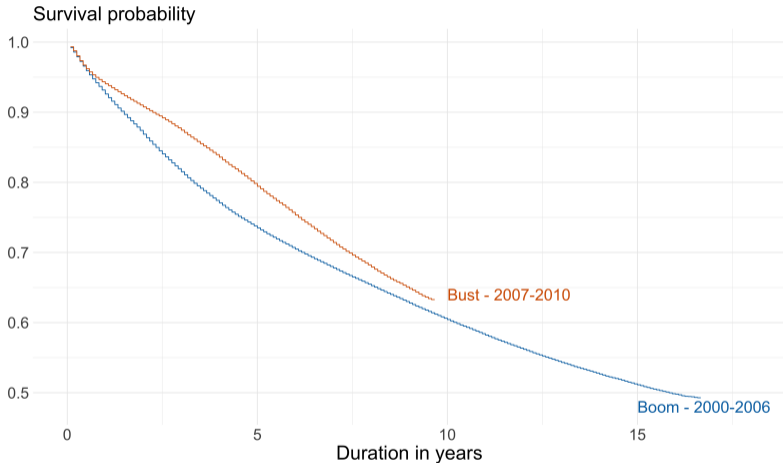
- This approach, as with the simple hazard model, uses likelihood methods
- The key complications are:
  - The baseline hazard model
  - The heterogeneity
- How to deal with the baseline hazard?  $\lambda(t)$  is a nuisance parameter for the estimation of the  $\theta_0$ .
- The Cox approach ( $v = 1$ ) for the nuisance parameter exploits the proportionality: at any given event, the *partial likelihood* for unit  $i$  that fails at period  $t$  is

$$\begin{aligned}L_i(\beta) &= \frac{\phi(t)\theta_0(X_i\beta)}{\sum_{j:Y_j>Y_i} \phi(t)\theta_0(X_j\beta)} \\ &= \frac{\theta_0(X_i\beta)}{\sum_{j:Y_j>Y_i} \theta_0(X_j\beta)}\end{aligned}$$

(this is analogous to the solution in conditional logit)

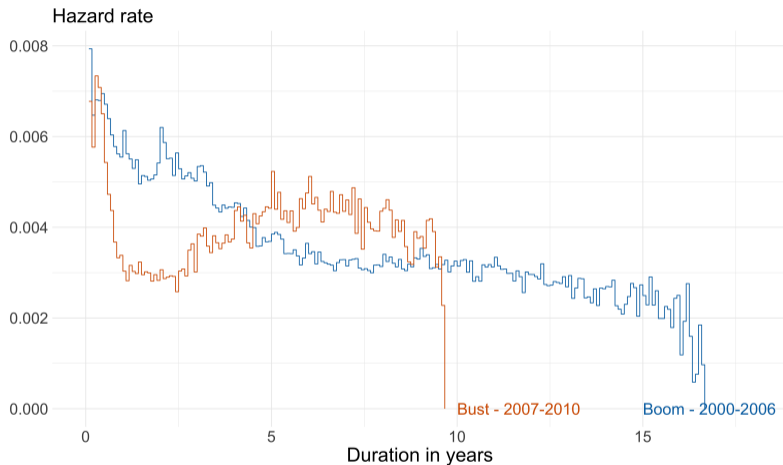
# Simplifying the intuition

- In the case on a non-time-varying treatment (e.g. a baseline covariate) that is binary (or discrete), we don't need to get quite so complicated
- What if we just compared Kaplan-Meier survival functions?
  - Let's compare boom and bust houses



# Simplifying the intuition

- Very intuitive and straightforward to compare hazards
- Doesn't look like proportional hazards is a reasonable assumption
  - This is a standard model fit check
- Key downside: doesn't accommodate unobserved heterogeneity, nor time-varying characteristics



## Sometimes you have to get complicated

- Sometimes a more complicated model is worthwhile, and you can't do a simple comparison
  - Time-varying treatments being an obvious case
- Key tension (discussed in Abbring and Van Den Berg (2003)) – when is the *timing* of the time-varying treatments?
  - Anticipation of treatments will confound your estimates
- Additional complicating factor: competing risks
  - What if there are multiple simultaneous states one could transition to?
  - e.g. Unemployment could go to employment, or leaving the labor force
  - e.g. Mortgage could go to default or prepayment
  - Our strategies above ignore this and assume the risks are independent
    - This is clearly a strong assumption
- See Honore and Lleras-Muney (2006) for a discussion on the fundamental identification issue in competing risks