Agenda

Survival analysis

1. Administrative
2. Modeling “survival”
3. Time-to-event models & censored data
4. Hazard/survival models

ADMINISTRATIVE

Presentation order:

- Mahjoube
- Jordan
- Zacharie
- Ella
- Kaitlin
- Zekai
- Terhas
- Trent

Time-to-event models

For many real-world problems, a model needs to account for *uncertainty in event timing*

- Recidivism
- Career length
- Life expectancy
- Time to residency
- Regime length
- Protest event length
- ...

Rossi (1980) recidivism data

Observational data for one year after release for 432 convicted people in Maryland in the late 1970s.

Treatment: receiving “financial aid”

Rossi, Peter Henry, Richard A. Berk, and Kenneth J. Lenihan. 1980. *Money, Work, and Crime: Experimental Evidence*. Academic Press.

ID	week_arrested	arrested	financial_aid	black	…
1	20	1	0	1	
2	52	0	1	1	
3	52	0	0	0	
4	17	1	0	1	
…	…	…	…	…	

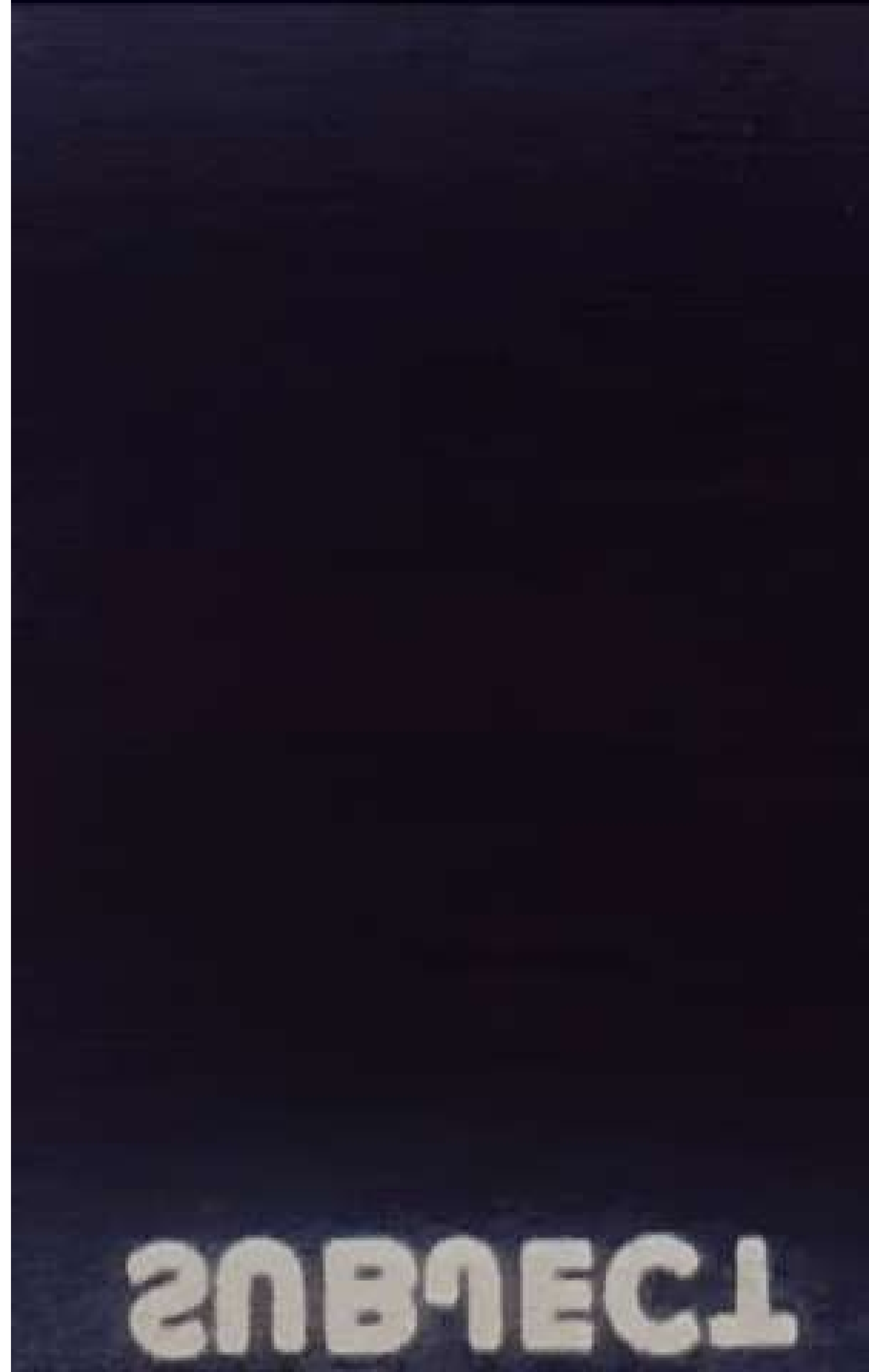
DURATION VS. HAZARD

Two common approaches to modeling *sources* of differences in timing:

Duration:
Model the duration of the event as a random variable with expectation determined by individual characteristics

Hazard:
Model the event as a (conditional) Bernoulli random variable with the probability determined by individual characteristics

Modeling duration



MODELING DURATION

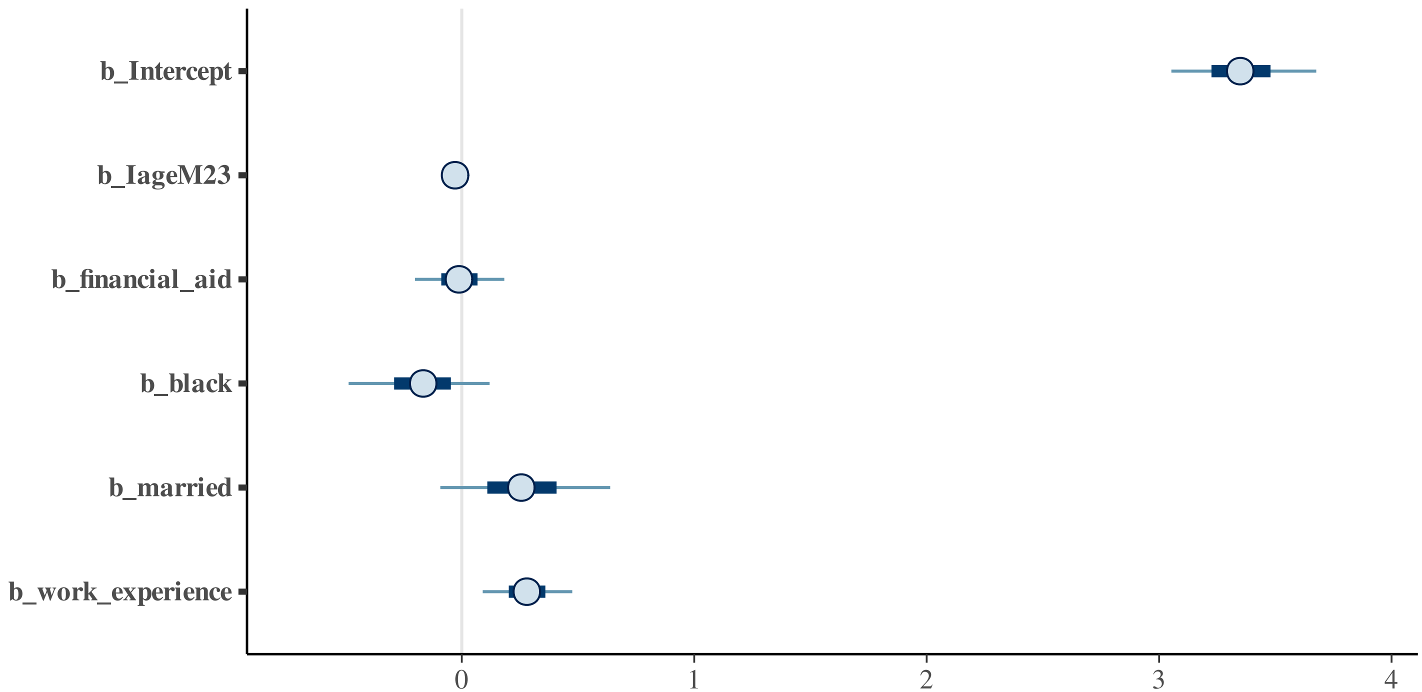
Choose an outcome distribution (e.g. Gamma, Weibull, exponential, log-normal, ...) and follow the same strategy as we have for other GLM models:

$$W_i \sim \text{Gamma}\left(\lambda, \frac{\lambda}{\mu_i}\right)$$
$$\log(\mu_i) = \beta_0 + \beta_1 T_i + \beta_2 B_i + \ldots$$

Problem:
*What do we do with people that were not arrested during the year? (**censored data**)*

Option 1 (bad)

We *could* drop observations for whom we do not know whether or when they were arrested. **This is a bad idea** and will almost certainly lead to biased results (see lecture on missing data)



N = 114 (out of 432)

Option 2 (better)

A better approach is to treat the arrest timing for those who were not arrested during the 52 week sample period as *missing data* for which we have a definite lower limit.

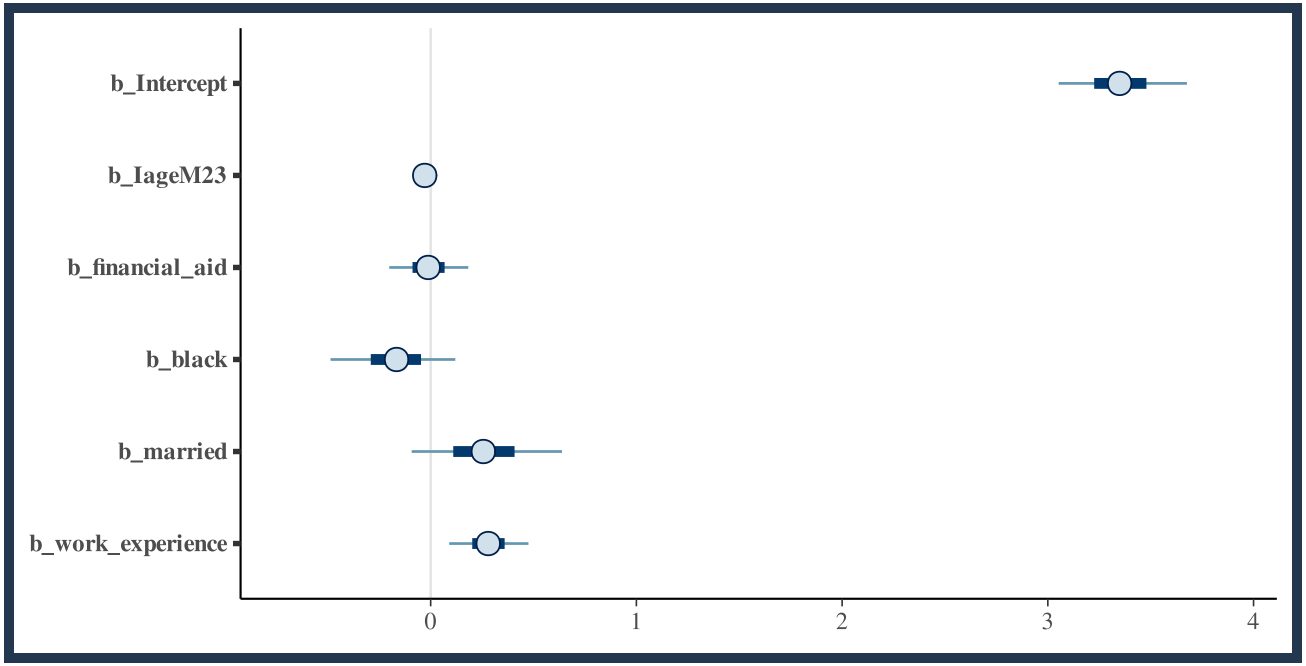
There are many robust ways to deal with this sort of censored data. In a Bayesian context the most common approach is treat the missing values as parameters with strong priors

Types of censoring:

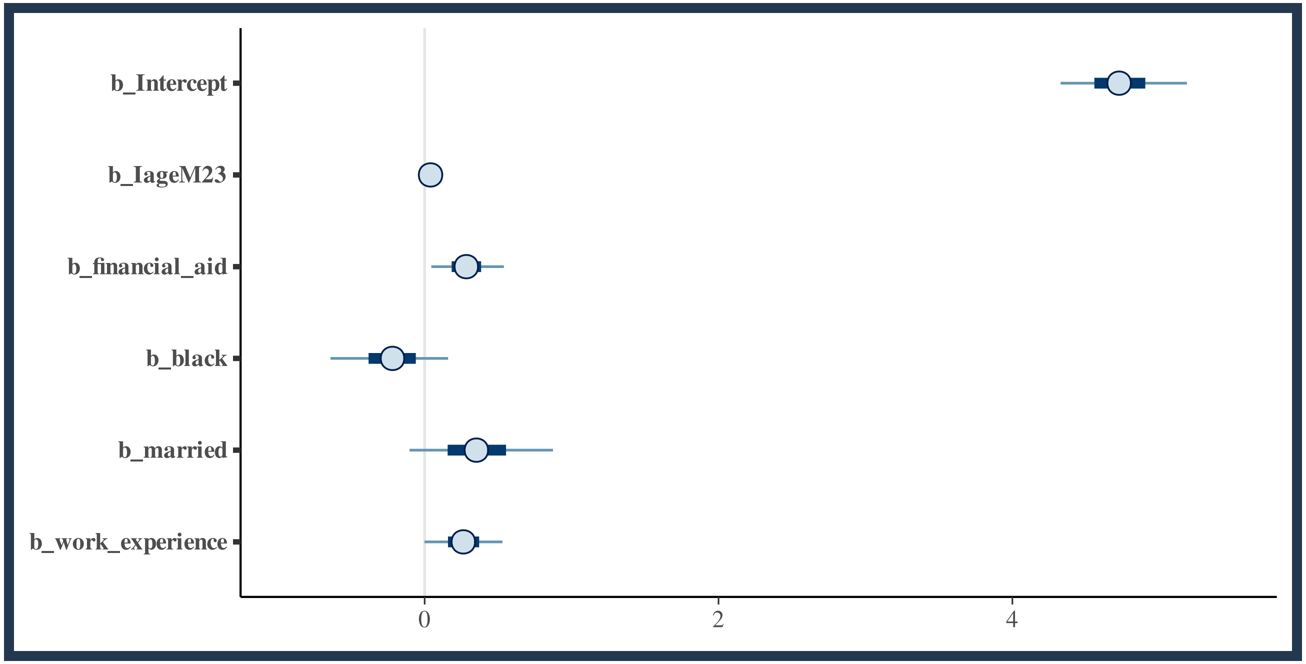
> - *Right-censored* — We only observe the data *before* a specific time
> - *Left-censored* — We only observe the data *after* a specific time
> - *Interval-censored* — We only observe the data within a specfic time interavl

CENSORED DATA

Dropping censored observations (bad)



Modeling censoring process (better)



Modeling hazard

HAZARD & SURVIVAL FUNCTIONS

Survival function

- $S(t) = \text{Prob}(Y > t)$
- Probability that the event will happen after a given time *t*.

Hazard function

- $\lambda(t) = \lim_{\delta \to 0} \frac{\text{Prob}(t \leq Y \leq t + \delta | Y > t)}{\delta}$
- "Instantaneous" probability of the even happening, conditional on it *not* having happened already

Neither of these is observed

Discrete version of Cox Proportional Hazard Model:

$$\mathrm{Prob}(W_i = t | W_i \geq t) = \lambda_0(t)\mu_i$$

$$\log(\mu_i) = \beta_0 + \beta_1 T_i + \beta_2 B_i + \ldots$$