Modeling Duration Outcomes for Social Science Researchers

Frederick J. Boehmke

University of Iowa

May 7, 2019

Prepared for presentation at The University of Kentucky.

Duration Models

- Also known as survival models or time-to-event models.
- A broad class of estimation strategies for modeling the timing of events.
- Arose initially in medical research, but now ubiquitous in social science research.
- A wide variety of considerations to evaluate when specifying them.

Political Science Examples of Duration Outcomes

1. How long until a state/city/country adopts a policy?
2. How long until treaties are signed?
3. How long do wars or civil conflicts last?
4. How long do governments last?
5. How long until Presidential nominees are confirmed?
6. When do candidates go negative?

What are the Main Features of Duration Data?

1. Start with a collection of subjects.
2. Identify the duration event of interest.
3. Collect data on their spells: when the event of interest starts and when it ends.
4. Collect information on variables that may influence the duration of spells.

Questions to Ask Before Starting Your Analysis

1. Do I have a discrete or continuous measure of failure time?
2. Are all spells fully observed?
3. Does the baseline hazard change over time?
4. Do any independent variables change over the course of a spell?
5. Do I have multiple failures per subject?
6. Are all subjects at risk of failure?
7. Are there conceptually different ways in which subjects can fail?

Outline of Workshop

1. Discrete versus continuous outcomes:
    1. Discrete for failures in distinct periods: logit/probit;
    2. Continuous for failure times in exact or small units.

Outline of Workshop

1. Discrete versus continuous outcomes:
    1. Discrete for failures in distinct periods: logit/probit;
    2. Continuous for failure times in exact or small units.
2. Questions about censoring:
    1. Do we fully observe beginning, middle, and end of all durations?
    2. Most commonly: do all units fail in the study period?
    3. If not, need to account for censoring in our analysis.

Outline of Workshop

1. Discrete versus continuous outcomes:
    1. Discrete for failures in distinct periods: logit/probit;
    2. Continuous for failure times in exact or small units.
2. Questions about censoring:
    1. Do we fully observe beginning, middle, and end of all durations?
    2. Most commonly: do all units fail in the study period?
    3. If not, need to account for censoring in our analysis.
3. Accounting for duration dependence:
    1. Does the chance of the event increase or decrease over time?
    2. In discrete version, include variables for time trend;
    3. In continuous version, need to pick the right distribution.

Outline of Workshop

1. Time-varying versus time-invariant covariates
    1. Do independent variables change over time?
    2. If yes, data structure is somewhat different.

Outline of Workshop

1. Time-varying versus time-invariant covariates
    1. Do independent variables change over time?
    2. If yes, data structure is somewhat different.
2. Multiple events:
    1. Repeated failures of the same type suggest controlling for event number;
    2. Competing failure types allow modeling each distinctly.

Outline of Workshop

1. Time-varying versus time-invariant covariates
    1. Do independent variables change over time?
    2. If yes, data structure is somewhat different.
2. Multiple events:
    1. Repeated failures of the same type suggest controlling for event number;
    2. Competing failure types allow modeling each distinctly.
3. Miscellaneous extensions:
    1. Split population models;
    2. Non-random sample selection;
    3. Pooled event history analysis.

Goals

Gain a general understanding of ...

- The general structure of duration outcomes;
- Statistical details of basic estimators for discrete and continuous duration data;
- The most common set of considerations for basic duration data;
- Resources and concepts related to more advanced issues;
- How to work with duration data in Stata;
- Practical advice based on my experiences.

Goals of the Computer Exercises

- Examples to learn basic commands;
- Guided learning using variations on basic examples;
- Start simple, then get more complicated;
- Play with you data — what matters and how?;
- Use good programming practices (you'll pick up some tricks).

Presentation Conventions

1. Stata commands are indicated in `typewriter font`.
2. Stata commands you can run are preceded by a "." and in `typewrite font`.
3. Commands may break across two lines of slides, but they must be one line in the Stata command prompt (they can wrap, but no carriage returns).
4. Some special characters may not copy properly from pdf, e.g., _, ''.
5. I'll share batch files for all the computer exercises we do.

General Concepts

- Identify the start of the duration process.
- Measure the time of failure for each unit.
- Determine the appropriate unit of time:
  1. How is the outcome variable measured?
  2. Do events occur in discrete periods?
  3. Do covariates change over time?

Key Concepts

Risk Set: the set of observations that are still at risk of experiencing the event at time t, $R(t) = \{i : Y_i \geq t\}$.

Survival function: the proportion of observations at risk at time t.

Hazard rate: the chance of failure at time t conditional on surviving to time t.

Spell: The period of time during which a subject is at risk of failure.

The Funny Terminology of Duration Analysis

Given its origins in medical research, we have terms like ...

- Survival: how long does a patient with a disease live?
- Failures: experiencing the event was usually a bad thing!
- Cure models: can some people never get the disease?
- Frailties: Is there unmeasured heterogeneity in the population?

Discrete Duration Models

- Also known as Event History Analysis (EHA);
- Outcome is measured in regular intervals (e.g., annually) and covariates typically change over those periods as well;
- Data are structured in a cross-sectional time-series manner by period and unit;
- Typically estimators include logit, probit, or cloglog;
- Censoring is handled in a straightforward manner.

Discrete vs. Continuous

- Discrete duration data code failure *within* a time period rather than *at* an exact time.
- Think of data failing in a given year without regard to exactly when.
- Or a study with regular, set followup times.
- The outcome is no longer a continuous outcome, but a binary indicator for failure at each discrete time period.

Figure: Continuous Duration Data

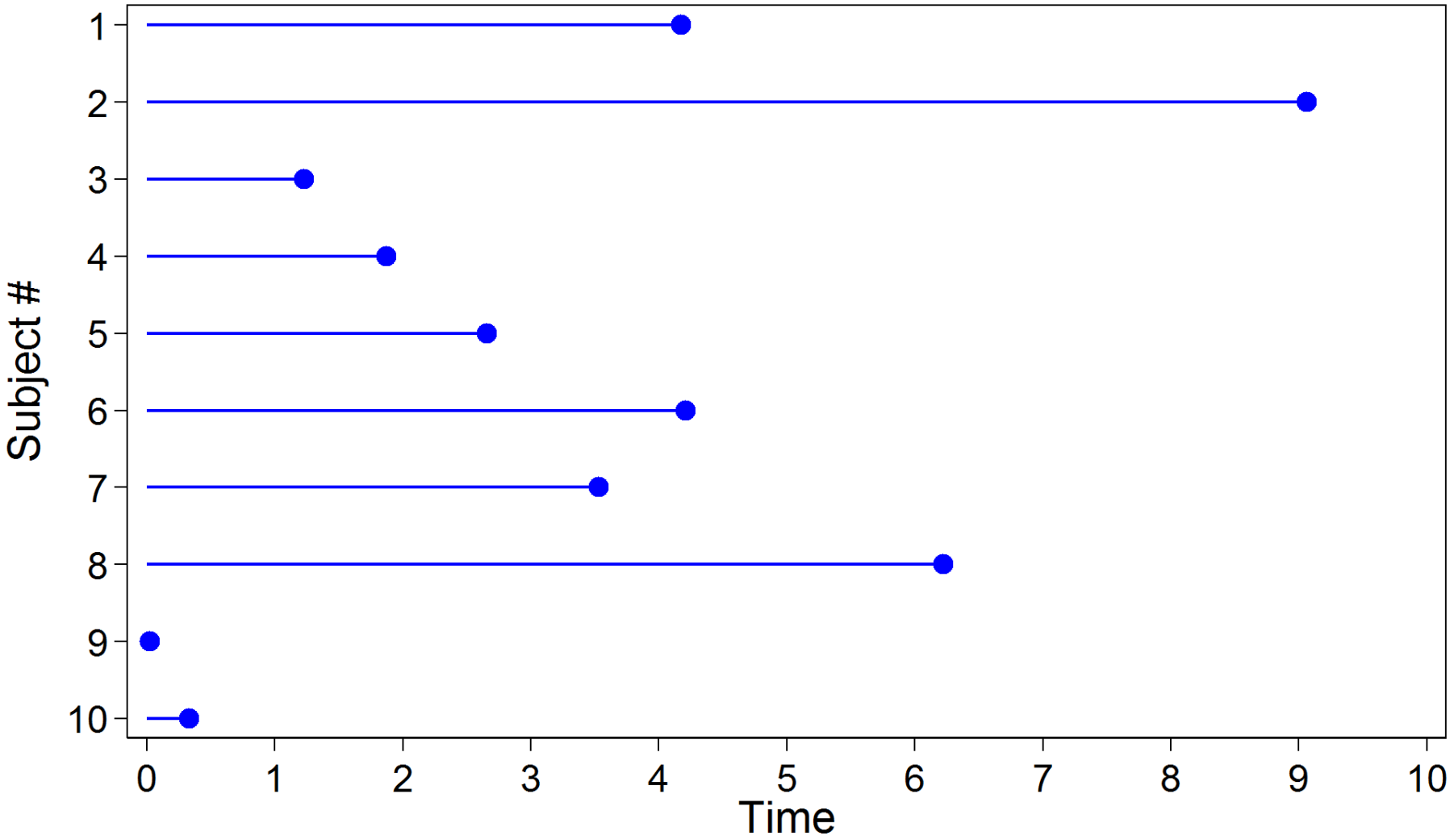


Figure: Discrete Duration Data Converted from Continuous Time

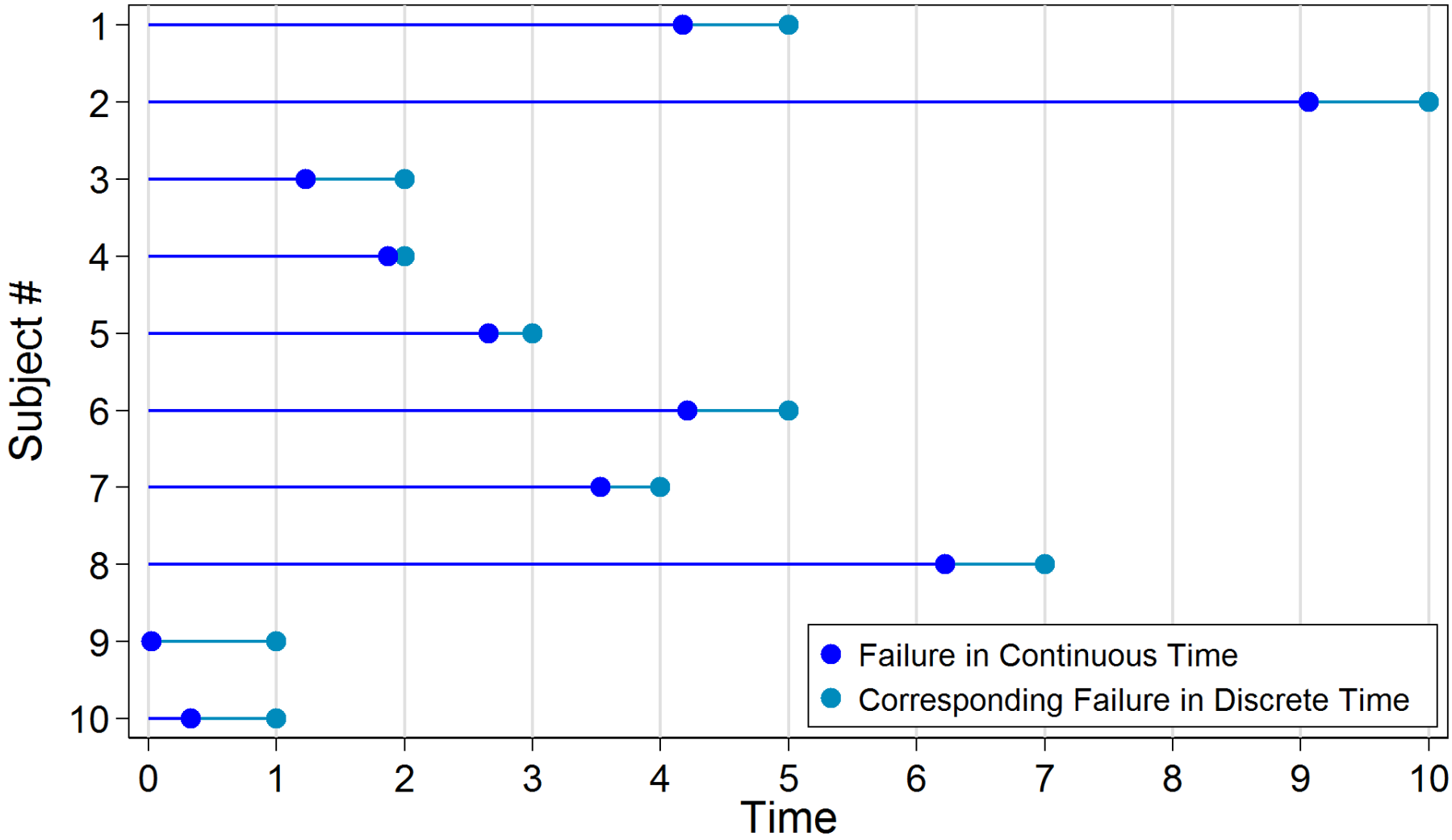
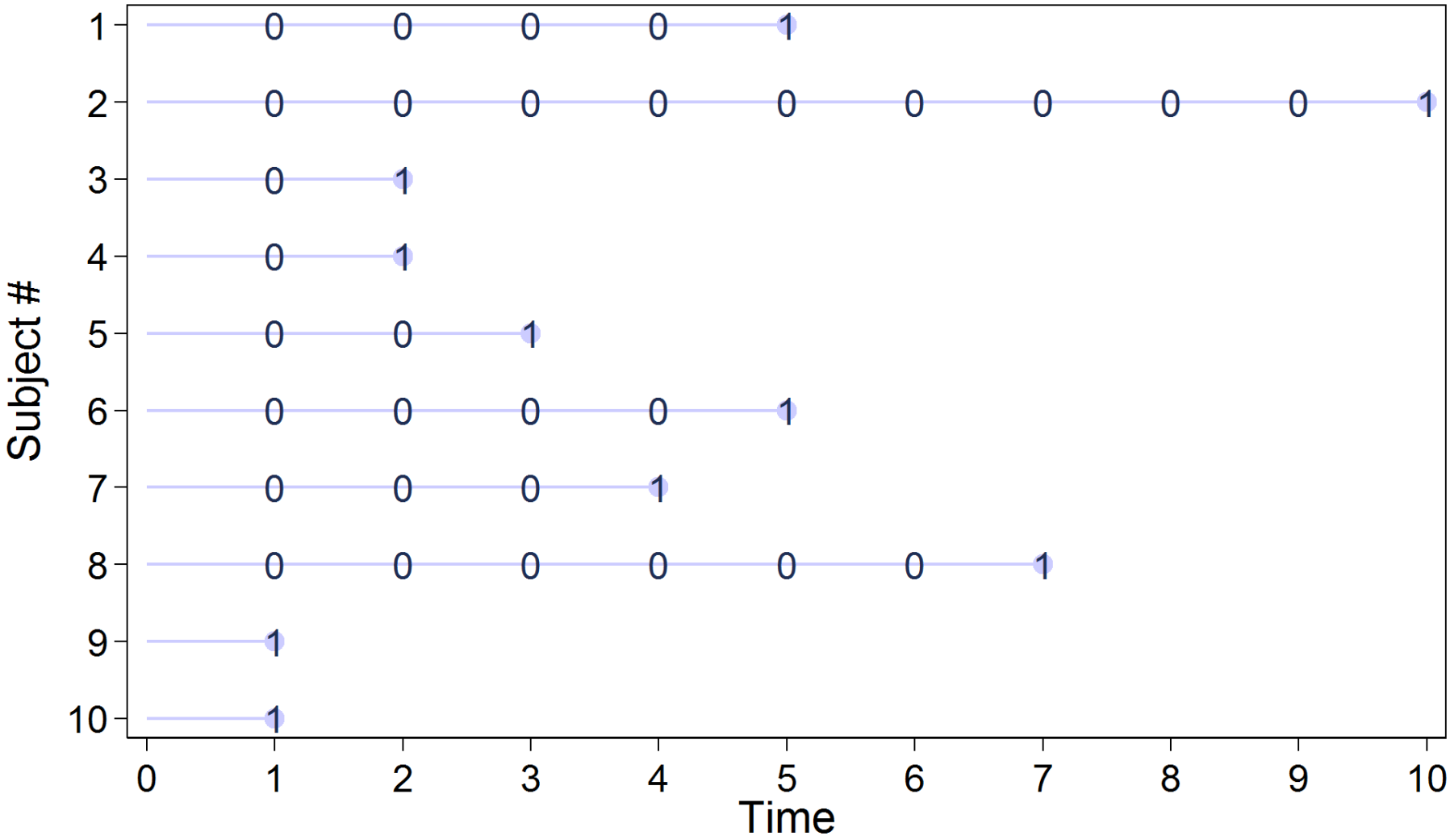


Figure: Discrete Duration Data: How Outcome is Coded



What do Discrete Duration Data Look Like?

ID	Time	Fails At	Y	X
1	1	5	0	-0.68
1	2	5	0	-0.68
1	3	5	0	-0.68
1	4	5	0	-0.68
1	5	5	1	-0.68
1	6	5	.	-0.68
1	7	5	.	-0.68
1	8	5	.	-0.68
1	9	5	.	-0.68
1	10	5	.	-0.68
2	1	10	0	2.44
2	2	10	0	2.44
2	3	10	0	2.44
2	4	10	0	2.44
2	5	10	0	2.44
2	6	10	0	2.44
2	7	10	0	2.44
2	8	10	0	2.44
2	9	10	0	2.44
2	10	10	1	2.44

Estimation

- Run a logit, probit, or cloglog model with discrete Y as the dependent variable.
- Risk set is naturally accounted for by how we code the dependent variable:
  1. $Y_{it} = 0$ means at risk at t and does not fail at t;
  2. $Y_i = 1$ means at risk and fails at t;
  3. $Y_i = .$ means not at risk at t.
- Results can be interpreted in the usual manner as the probability of the event occurring in a given time period.

Interpretation

- Results can be interpreted in the usual manner for a discrete choice model, e.g., using `predict` or `margins` in Stata, as the probability of the event occurring in a given time period (the hazard):

$$\Pr(Y_{it} = 1|X_{it}) = \frac{exp(X_{it}\beta)}{1 + exp(X_{it}\beta)}.$$

- Since the single-period effects may seem small, you may also want to generate the survival function to convey the cumulative effect of a covariate:

$$S(T|X_i) = \Pr(Y_{it} = 0 \forall t < T|X_{it}) = \prod_{t=1}^{T} \Pr(Y_{it} = 0|X_{it}).$$

Accounting for Duration Dependence

Does the baseline hazard rate change over time independently of changes in covariates? If so, we can include:

- A linear time trend;
- A quadratic, cubic, or higher-order polynomial of time;
- Splines of time;
- Time fixed effects.

Accounting for Duration Dependence: Recommendations

- Linear time trend is usually not enough;
- A quadratic or cubic is usually pretty good – just test for adding additional terms;
- Splines are very flexible, but make sure to look into picking the number and location of knots;
- Time fixed effects tend to be overkill and often lead to data separation issues.

Computer Exercise for Discrete EHA

Commands for this are in `exercise01discrete.do`.

- Open the file `exercise01discrete.dta`;
- Create the dependent variable from adoption years;
- Create some basic graphs of the duration process;
- Run a discrete EHA model;
- Control for duration dependence in various ways;
- Generate predicted values to interpret the model.

Continuous Time Duration Models

- Outcome is measured very precisely (e.g., days or months);
- Covariates might change or may be constant;
- Data could be structured as cross-sectional or cross-section time-series, depending on covariate structure.
- Here we have to pick a distribution to capture duration dependence;
- Stata has a suite of commands for survival data which will handle censoring easily once the data are declared appropriately.

What do Duration Data Look Like?

Subject	Start Time	End Time	Duration	X
1	2.21	6.38	4.18	−0.68
2	4.89	13.95	9.06	2.44
3	4.27	5.50	1.23	0.02
4	5.73	7.61	1.87	−0.39
5	0.26	2.92	2.65	0.80
6	2.39	6.60	4.21	0.55
7	6.71	10.25	3.53	1.83
8	4.59	10.82	6.22	−0.26
9	2.07	2.10	0.03	−1.41
10	6.02	6.35	0.33	−0.12

Figure: Example of Duration Data: Start and End Times

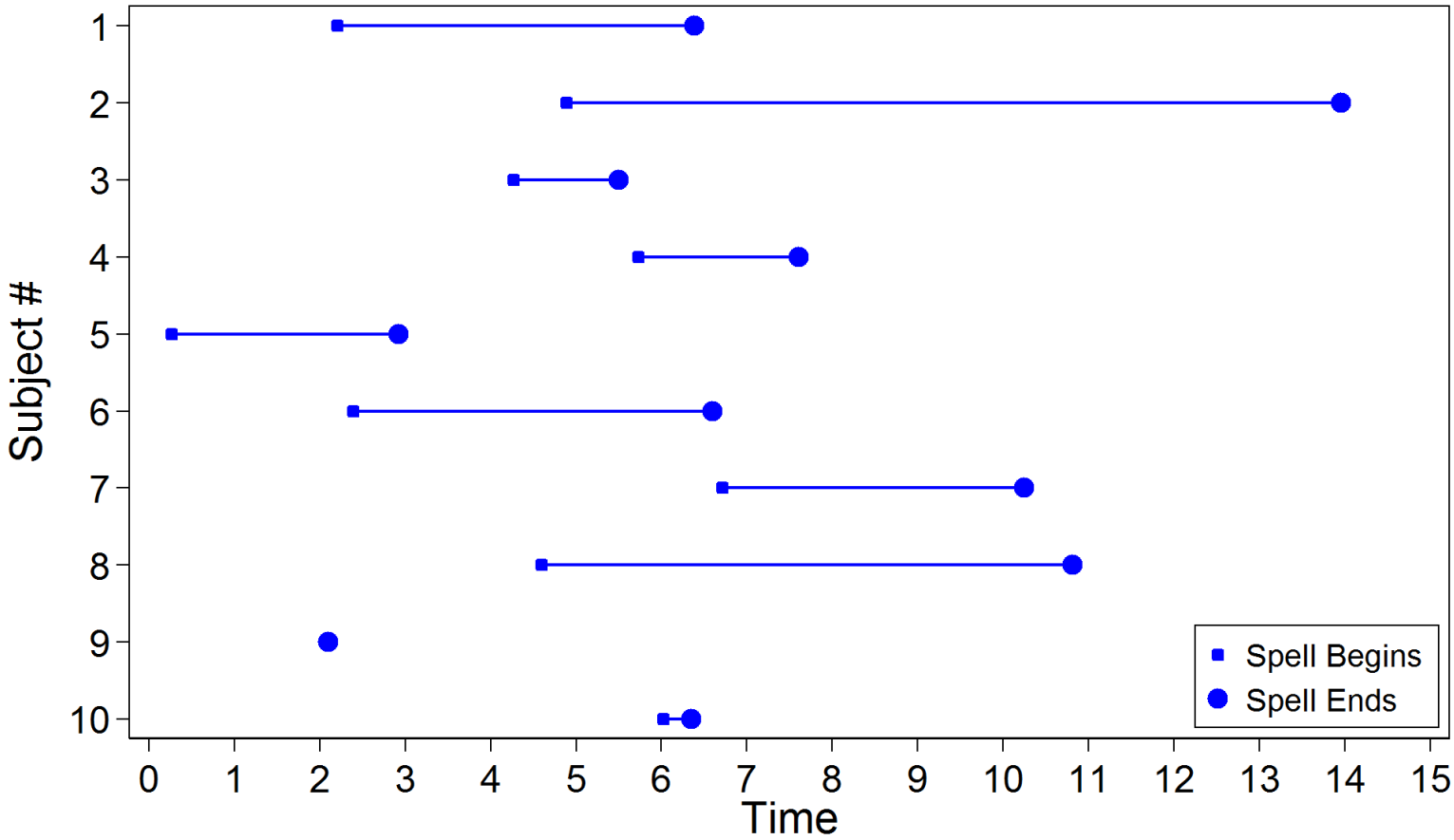
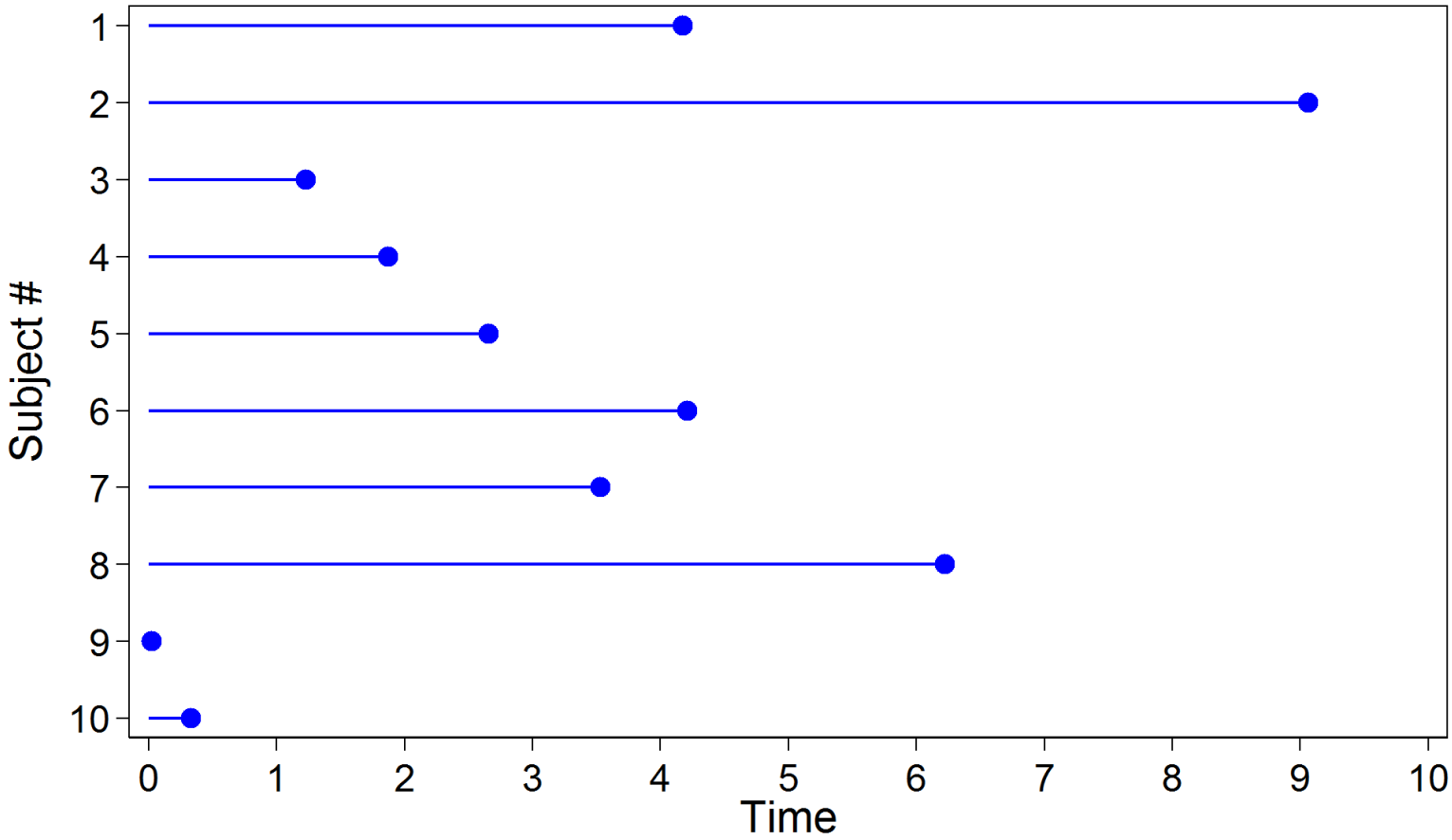


Figure: Example of Duration Data: Time to Failure



Relating Continuous Time Durations to OLS

- Helpful to think of as similar to OLS data-generating process:

$$Y_i = X_i\beta + u_i.$$

Relating Continuous Time Durations to OLS

- Helpful to think of as similar to OLS data-generating process:

$$Y_i = X_i\beta + u_i.$$

- But that would be problematic since durations must be non-negative, so we modify a bit:

$$\ln(Y_i) = X_i\beta + \ln(u_i).$$

Relating Continuous Time Durations to OLS

- Helpful to think of as similar to OLS data-generating process:

$$Y_i = X_i\beta + u_i.$$

- But that would be problematic since durations must be non-negative, so we modify a bit:

$$\ln(Y_i) = X_i\beta + \ln(u_i).$$

- Then rewrite by exponentiating both sides:

$$\begin{aligned} \exp(\ln(Y_i)) &= \exp(X_i\beta + \ln(u_i)), \\ Y_i &= \exp(X_i\beta)\exp(\ln(u_i)), \\ Y_i &= \exp(X_i\beta)u_i. \end{aligned}$$

Relating Continuous Time Durations to OLS

- Helpful to think of as similar to OLS data-generating process:

$$Y_i = X_i\beta + u_i.$$

- But that would be problematic since durations must be non-negative, so we modify a bit:

$$\ln(Y_i) = X_i\beta + \ln(u_i).$$

- Then rewrite by exponentiating both sides:

$$\begin{aligned} \exp(\ln(Y_i)) &= \exp(X_i\beta + \ln(u_i)), \\ Y_i &= \exp(X_i\beta)\exp(\ln(u_i)), \\ Y_i &= \exp(X_i\beta)u_i. \end{aligned}$$

- This keeps the duration outcome, Y_i, positive and the error scales the outcome up or down.

Choosing a Distribution for u_i

- With continuous time durations we have many possible distributions for u_i from which to select.
- Choice is important since it implies features of the duration, most notably the shape of the baseline hazard over time. Common examples include:
  1. Exponential: hazard rate stays constant;
  2. Weibull: hazard can increase or decrease;
  3. Log-normal: Hazard can increase then decrease;
  4. Gamma: Flexible and allows previous as special cases.
- Covariates then model changes in the expected duration.
- These estimators make the proportional hazards assumption: the effect of an independent variable on the hazard stays constant over time.
- Semi-parametric Cox approach allow for arbitrary hazard rates.

Moving from Errors to Outcomes

Once we assume a distribution for the error term we can characterize the distribution of Y.

- We typically assume the data is generated as follows:

$$Y_i = \exp(X_i\beta)u_i.$$

- To obtain the distribution of Y_i we solve for u_i:

$$u_i = Y_i \exp(-X_i\beta).$$

- To simplify the expression write $\lambda_i = \exp(-X_i\beta)$. Thus

$$u_i = Y_i\lambda_i.$$

- The distribution of Y is then used to:
  - Derive the likelihood and estimate the parameters;
  - Interpret the results.

Additional Interpretations of Durations

Given the density, $f(Y)$, and its associated cdf, $F(Y)$, we often focus on two additional representations of the distribution.

- The survival function, which captures the probability a subject has not experienced the event by time t:

$$S(Y) = 1 - F(Y).$$

Additional Interpretations of Durations

Given the density, $f(Y)$, and its associated cdf, $F(Y)$, we often focus on two additional representations of the distribution.

- The survival function, which captures the probability a subject has not experienced the event by time t:

$$S(Y) = 1 - F(Y).$$

- The hazard rate, which captures the instantaneous chance of failure at time t given that a subject has survived to t:

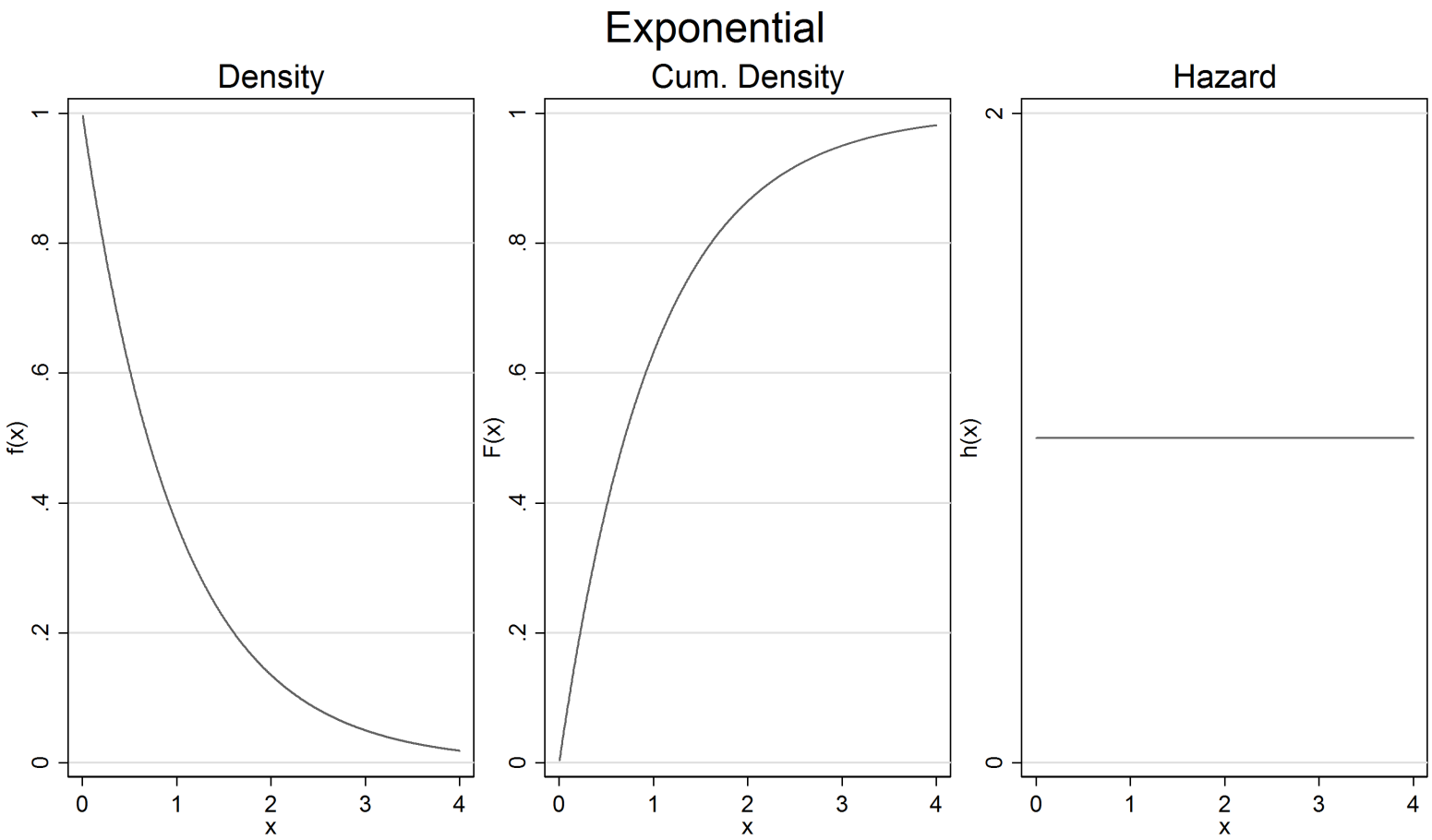
$$h(Y) = \frac{f(Y)}{1 - F(t)} = \frac{f(Y)}{S(Y)}.$$

The Exponential

$$
\begin{aligned}
F(u_i) &= 1 - \exp(-u_i), \\
F(Y_i|X_i) &= 1 - \exp(-Y_i\lambda_i); \\
S(Y_i|X_i) &= 1 - F(Y_i|X_i), \\
&= \exp(-Y_i\lambda_i); \\
f(Y_i|X_i) &= \lambda_i \exp(-Y_i\lambda_i), \\
h(Y_i|X_i) &= \frac{f(Y_i|X_i)}{1 - F(Y_i|X_i)}, \\
&= \frac{\lambda_i \exp(-Y_i\lambda_i)}{\exp(-Y_i\lambda_i)}, \\
&= \lambda_i.
\end{aligned}
$$

Note that the hazard, λ_i, does not depend on time.

Figure: Shapes of Exponential Durations



The Weibull

$$
\begin{aligned}
F(u_i) &= 1 - \exp(-u_i^p), \\
F(Y_i|X_i) &= 1 - \exp(-(Y_i\lambda_i)^p); \\
S(Y_i|X_i) &= 1 - F(Y_i|X_i), \\
&= \exp(-(Y_i\lambda_i)^p); \\
f(Y_i|X_i) &= p\lambda_i^p Y_i^{p-1}\lambda_i \exp(-(Y_i\lambda_i)^p), \\
h(Y_i|X_i) &= \frac{f(Y_i|X_i)}{1 - F(Y_i|X_i)}, \\
&= \frac{p\lambda_i^p Y_i^{p-1}\lambda_i \exp(-(Y_i\lambda_i)^p)}{\exp(-(Y_i\lambda_i)^p)}, \\
&= p\lambda_i^p Y_i^{p-1}.
\end{aligned}
$$

Note that the hazard *does* depend on time (via Y_i).

Figure: Shapes of Weibull Durations

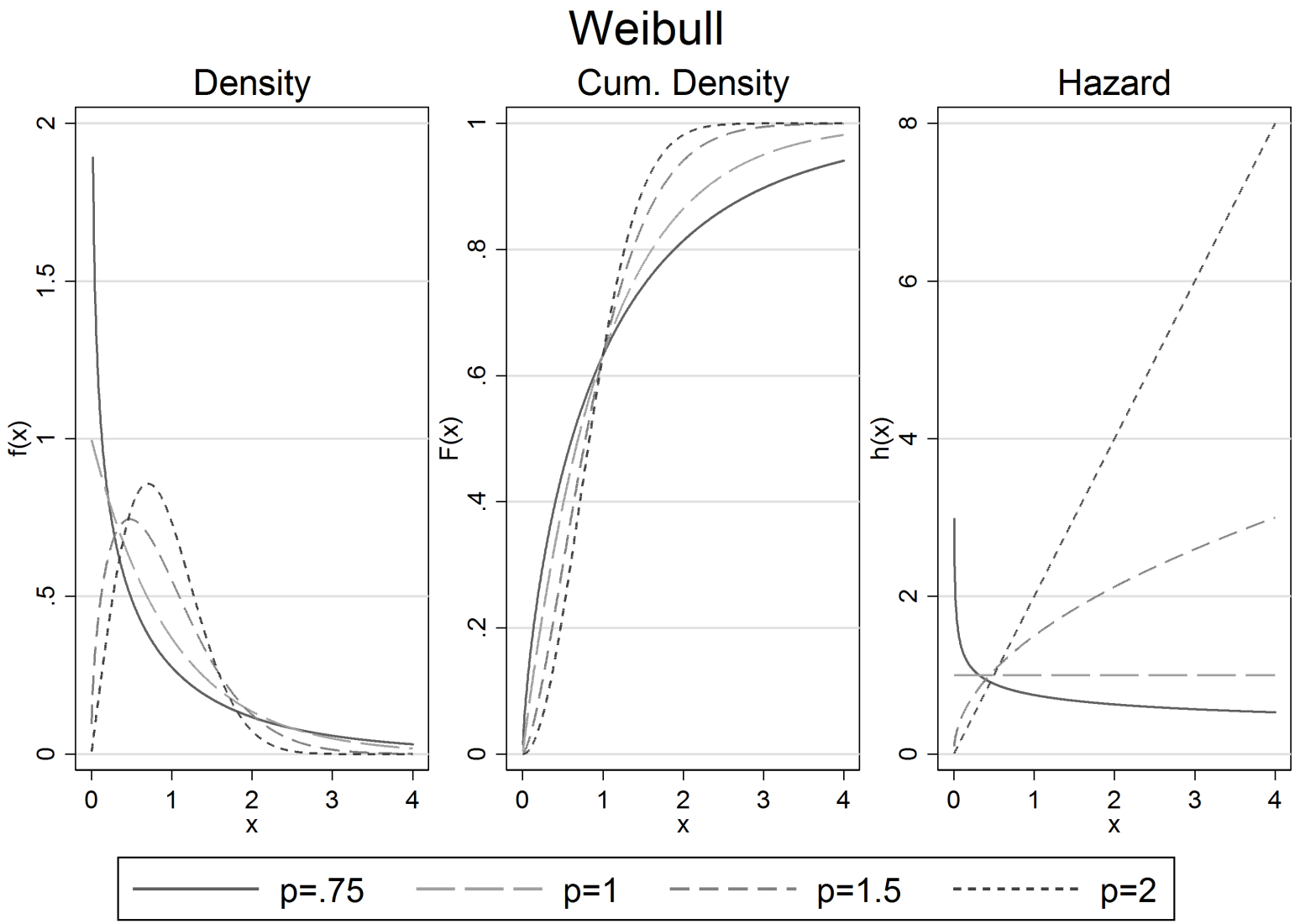
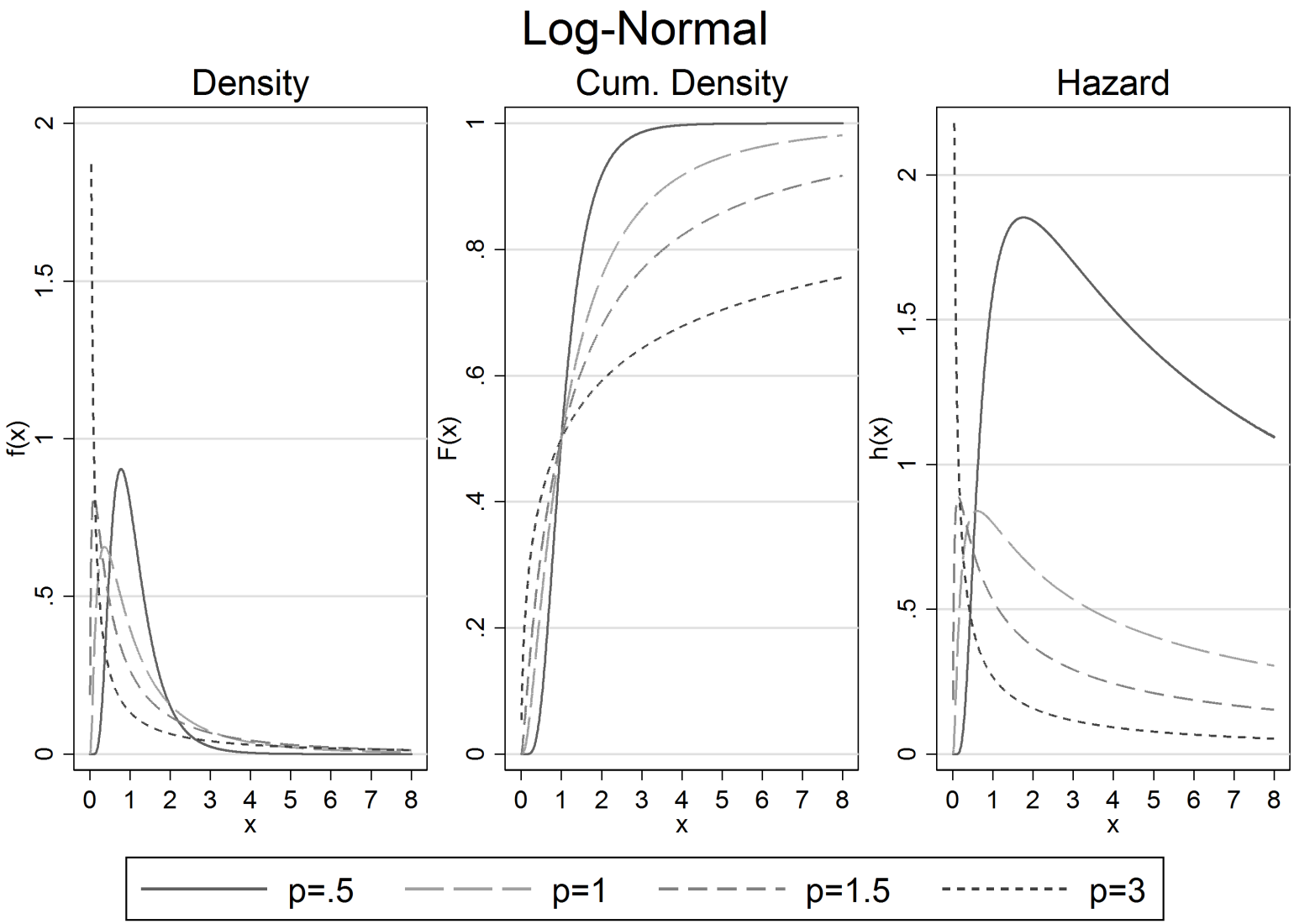


Figure: Shapes of Log-Normal Durations



The Generalized Gamma

The generalized gamma is especially helpful since it allows for increasing and decreasing hazards.

$$f(y_i|\lambda_i) = \frac{p\lambda_i(\lambda_i y_i)^{p\kappa-1}\exp(-(\lambda_i y_i)^p)}{\Gamma(\kappa)}.$$

Table: Special Cases of the Generalized Gamma

κ	p	**Distribution**
1	1	Exponential
1	p	Weibull
0	p	log-normal
κ	1	one parameter gamma

Estimation

Whichever distribution we select, we use the density to estimate the parameters via maximum likelihood.

$$L(\theta|Y,X) = \prod_{i=1}^{n} f(-\lambda_i y_i).$$

Interpretation

- We can calculate $E[Y_i|X_i]$ just like in regression and then do marginal effects or first differences. Expected values for selected distributions are as follows:

$$\begin{aligned} \text{exponential} \quad & \exp(X\widehat{\beta}); \\ \text{Weibull} \quad & \Gamma\left(\frac{1+\widehat{p}}{\widehat{p}}\right)\exp(X\widehat{\beta}); \\ \text{log-normal} \quad & \exp(\widehat{\sigma}^2/2)\exp(X\widehat{\beta}). \end{aligned}$$

- We can also plot the hazard or survival function over time (and do predictions at different values or first differences, etc.).
- Stata's `st` suite of duration commands includes options for interpretation as does its `predict` command.

The Cox Model

- This takes a semi-parametric approach to avoid the problem of assuming a parametric hazard function.
- It is very flexible.
- Intuitively, it evaluates the conditional probability that an observation fails when it does given all the remaining observations that could fail at that time.
- Thus it ignores the exact failure times and considers only the order.
- This leads to a ratio of densities through which the baseline hazard cancels out.
- Tied failure times complicate it a little, but solutions exist.

Intuition Behind the Cox Model

Write the hazard generally as:

$$h(y_i|X_i) = \exp(X_i\beta)h_0(y_i).$$

The conditional probability that observation i fails at time y_i, given that one of the surviving observations fails at y_i, is:

$$\begin{aligned}
\Pr(i \text{ fails at } y_i | \text{ someone fails at } y_i) &= \frac{h(y_i|X_i)}{\sum_{j \in R(y_i)} h(y_i|X_j)}, \\
&= \frac{\exp(X_i\beta)h_0(y_i)}{\sum_{j \in R(y_i)} \exp(X_j\beta)h_0(y_i)}, \\
&= \frac{\exp(X_i\beta)}{\sum_{j \in R(y_i)} \exp(X_j\beta)}.
\end{aligned}$$

Estimation for the Cox Model

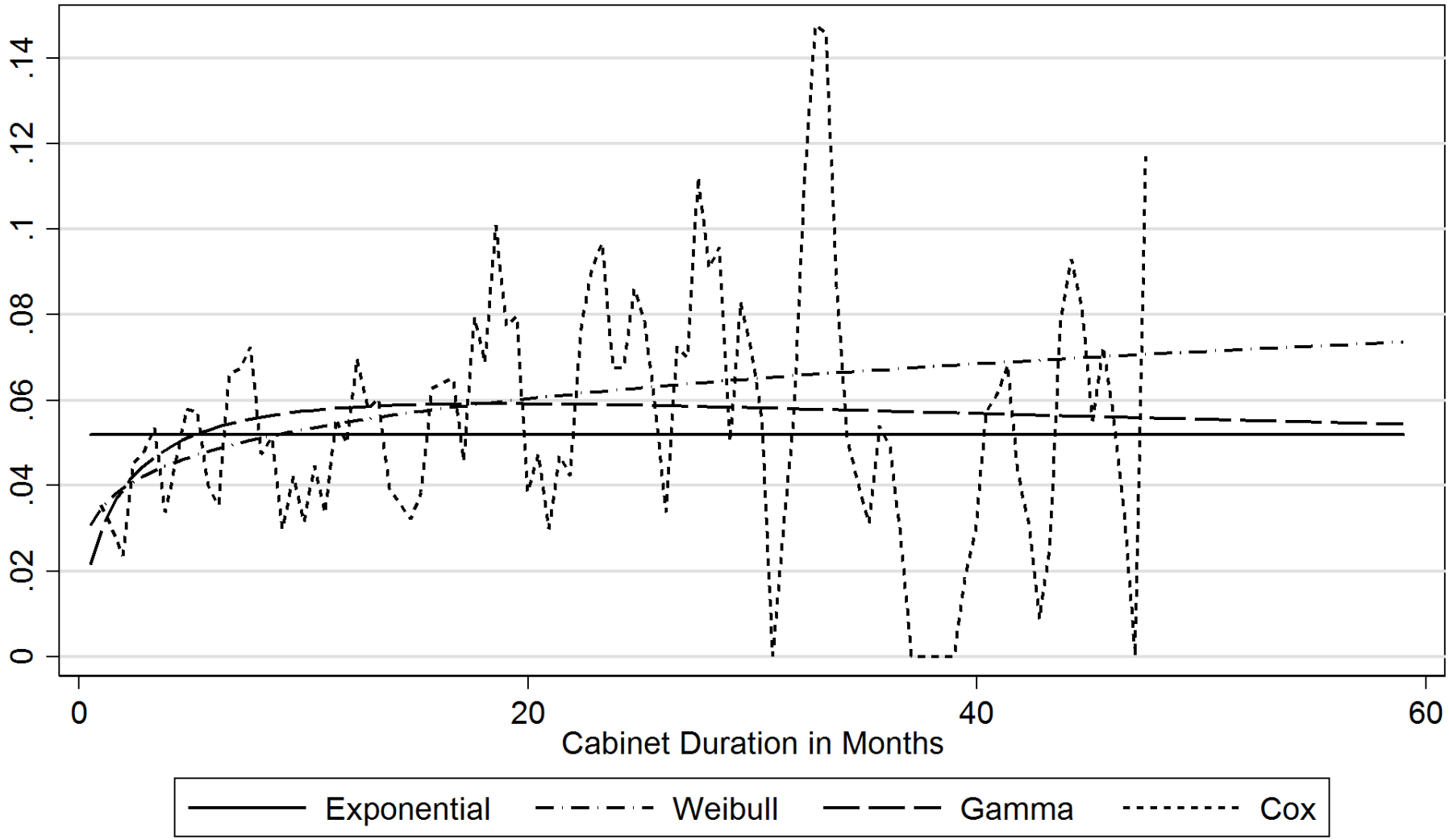
Order the observations by their observed failure time such that:

$$y_1 < y_2 < y_3 < \ldots < y_n.$$

We can define the partial likelihood function as follows:

$$\begin{aligned}
\mathcal{L}_p(\beta|Y,X) &= \prod_{i=1}^{n} \frac{h_i(y_i)}{\sum_{j \in R(y_i)} h_j(y_i)}, \\
&= \prod_{i=1}^{n} \frac{\exp(X_i\beta)}{\sum_{j \in R(y_i)} \exp(X_j\beta)}.
\end{aligned}$$

Figure: Comparison of Hazard Rates for Cabinet Durations



Different Ways of Reporting Parameters

There are multiple ways to report the parameters depending on the "interpretation" of the model, i.e., hazard versus time to failure.

Table: Reported Estimates for Weibull and exponential models in Stata

Interpretation	**Reported**	
	Weibull	*Exponential*
Hazard Ratio	$\exp(-\beta \times p)$	$\exp(-\beta)$
Prop. Hazard	$-\beta \times p$	$-\beta$
A.F.T.	β	β

Different Ways of Reporting Parameters

Here's why this makes sense (using the Weibull to illustrate):

$$\begin{aligned} h(t) &= p\lambda_i^p y_i^{p-1}, \\ &= p\exp(-X_i\beta)^p y_i^{p-1}, \\ &= p\exp(-p(X_i\beta))y_i^{p-1}, \\ &= p\exp(-X_i\beta p)y_i^{p-1}, \\ &= p\exp(-X_i\beta')y_i^{p-1}, \end{aligned}$$

where $\beta' = \beta p$.

Right Censoring

- Right censoring occurs when we do not observe a failure time.
- This could happen for many reasons:
  1. The study ends before all units fail;
  2. Units exit the study for unrelated reasons;
  3. The methods of failure becomes impossible (e.g., war ends, a cure is found).
- We need to account for this in our estimation.
- The Cox handles this easily.
- Parametric models require a bit more work.

Figure: Illustration of Right Censoring: Data Before Censoring

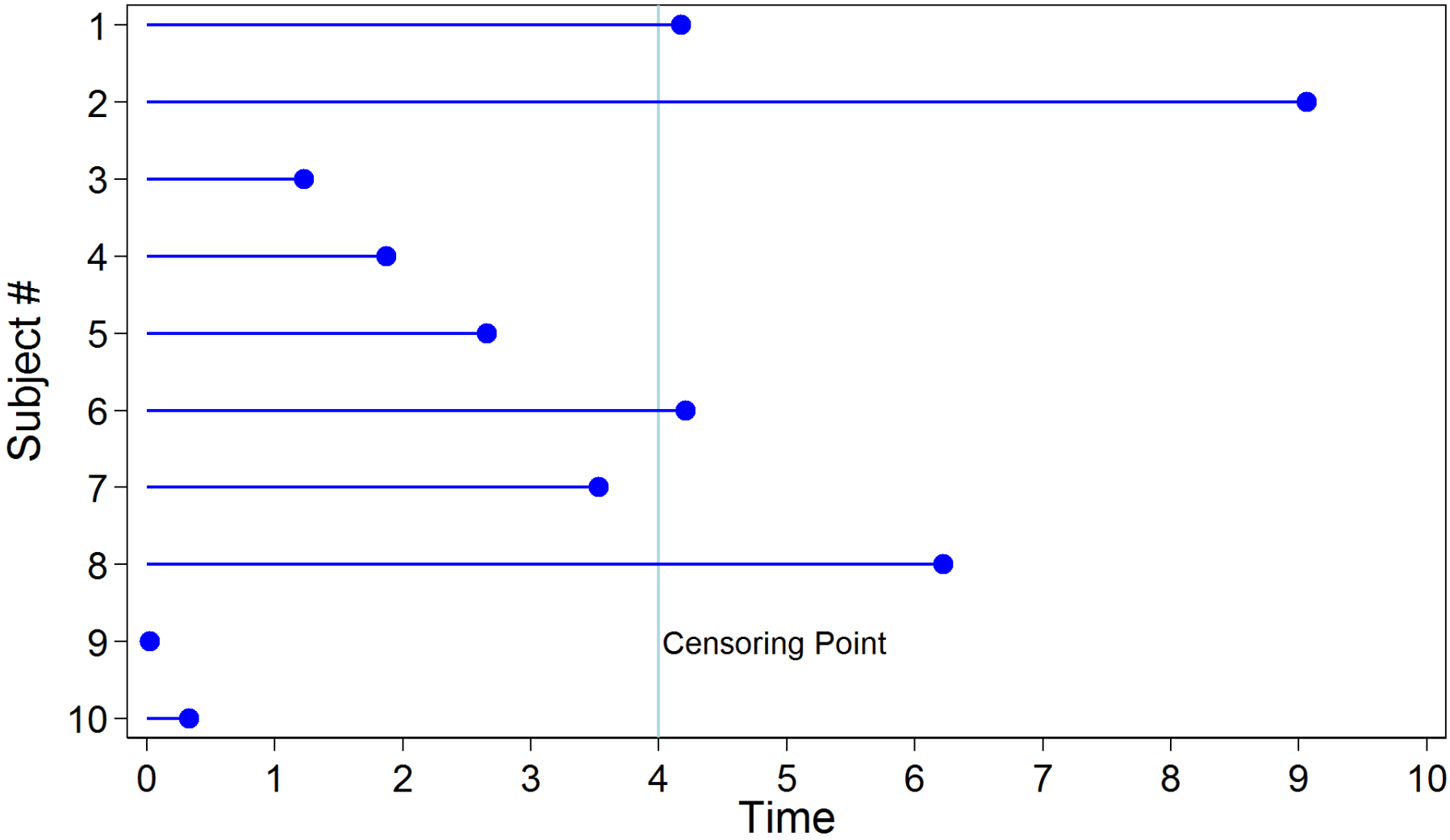


Figure: Illustration of Right Censoring: The Underlying Data

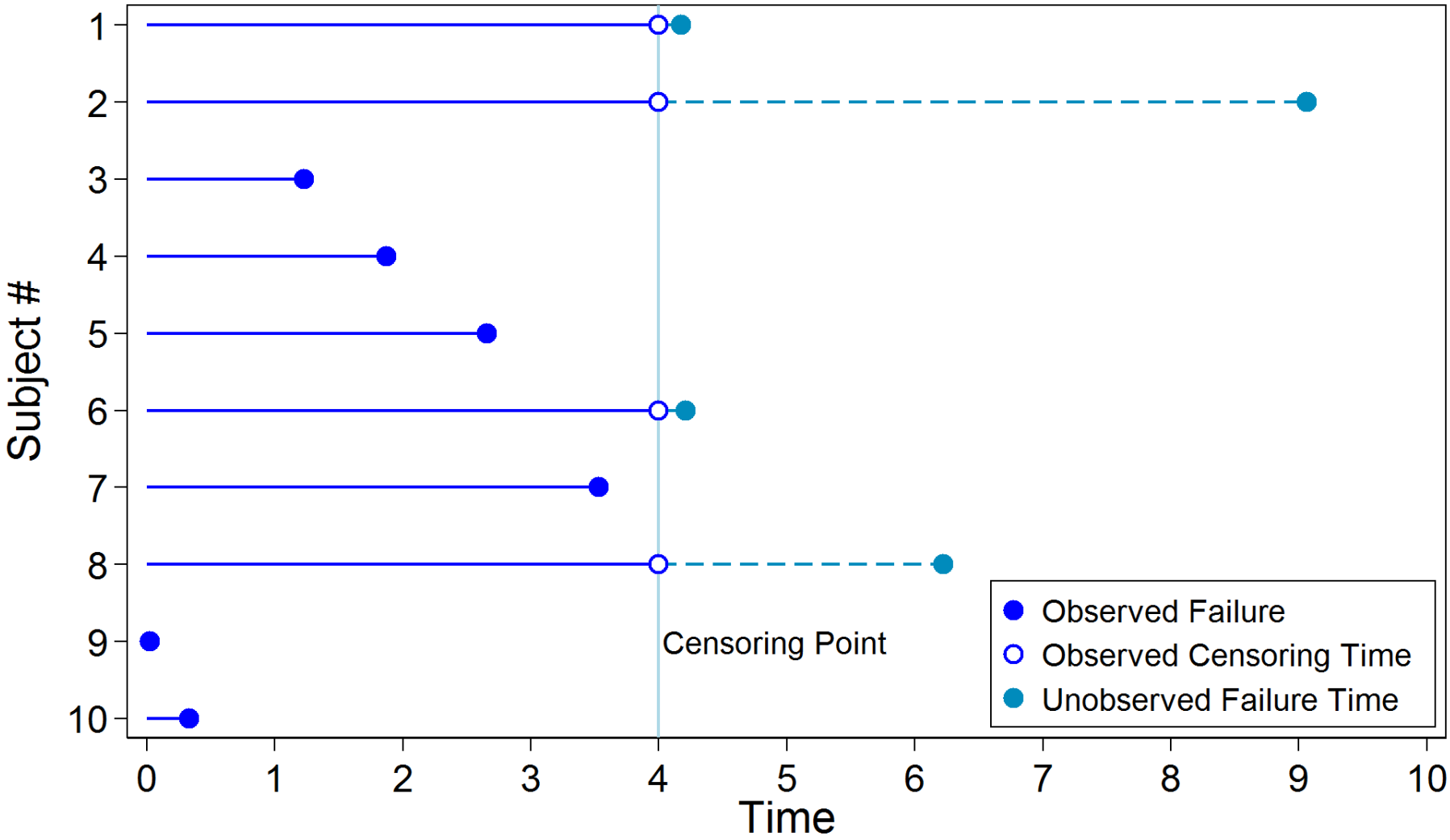
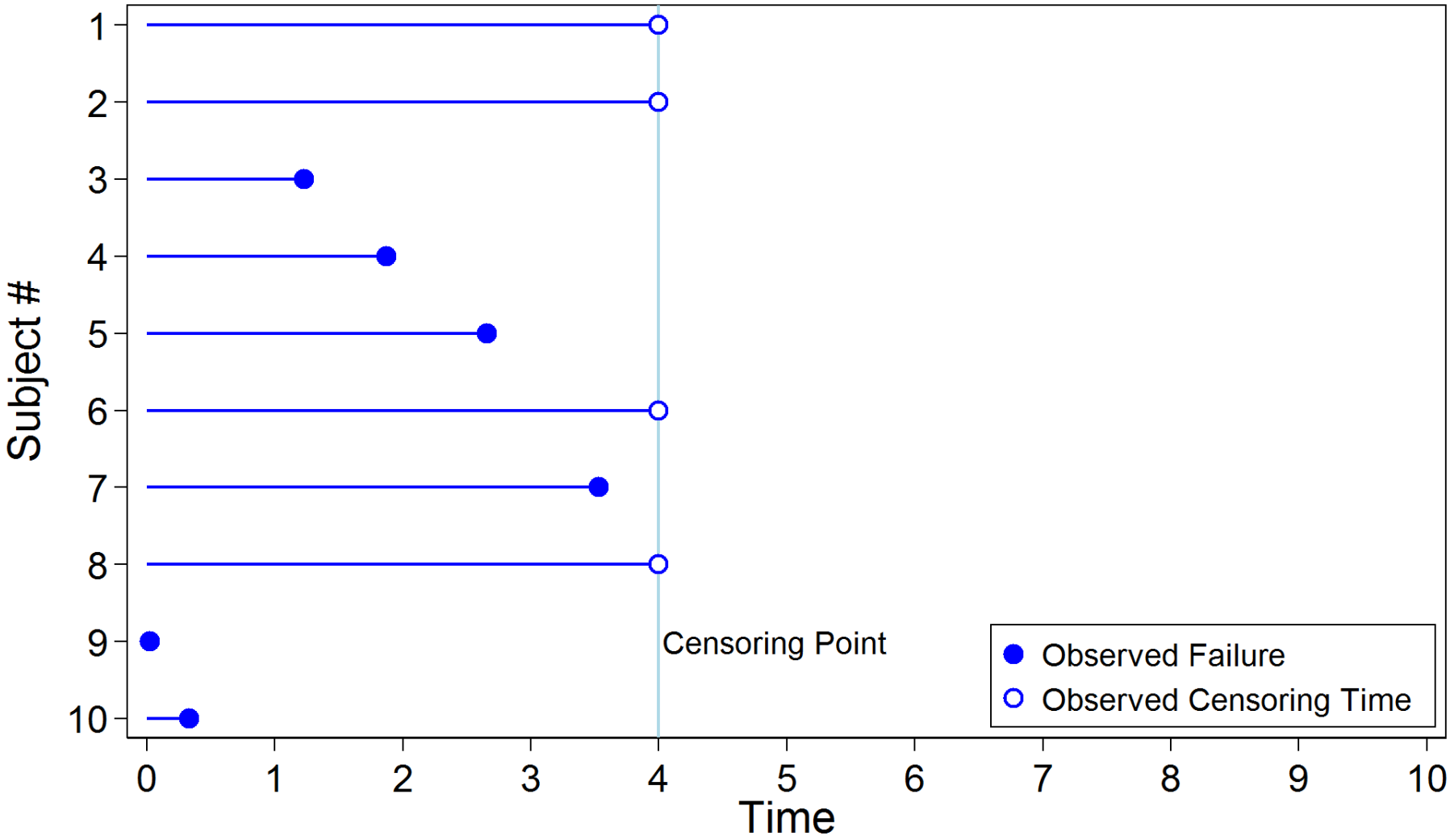


Figure: Illustration of Right Censoring: What we Observe



Accounting for Right Censoring

- Parametric models account for this by constructing the likelihood out of the density for observed failures and the survival function for right censored cases.
  - For observed failures we use their density (here the exponential):

$$f(Y_i|X_i) = \lambda_i \exp(-Y_i\lambda_i).$$

  - For censored cases, Y_i^c, we use the survival function:

$$S(Y_i^c|X_i) = \exp(-Y_i^c\lambda_i).$$

- The Cox model accounts for it via its construction:
  - Observations appear in the numerator of the partial likelihood only when they fail;
  - All observations still at risk at the time of failure appear in the denominator.

The Likelihood with Right Censoring

Estimation for parametric models works as before except we now separate observed failures from censored cases. Let $c_i = 1$ denote that an observation is right censored at y_i^c and $c_i = 0$ indicate that its failure time, y_i, is observed.

$$L(\theta|Y, X, C) = \prod_{c_i=0} f(-\lambda_i y_i) \prod_{c_i=1} S(-\lambda_i y_i^c).$$

Survival Analysis in Stata

For continuous-time models, Stata has an excellent set of routines under the `st` suite of commands to handle survival time data.

stset	declare the data to be survival time and indicate censoring, repeated failures, etc.;
stdescribe	learn about the features of the data;
stsum	learn descriptive statistics about the duration outcome;
sts	Plot the hazard and survival function;
streg	Run a parametric duration model;
stcox	Run a Cox model;
stcurve	Plot predicted hazard and survival functions based on model results.

The Multiple Interpretations of Variables' Effects

- Different distributions allows different interpretations:
  1. Hazard ratio;
  2. Proportional hazard;
  3. Accelerated failure time.
- If the data generating process is $Y_i = \exp(X_i\beta)u_i$, then

Stata Option	**Interpretation**	**Reported**	
		Weibull	*Exponential*
(none)	Hazard Ratio	$\exp(-\beta \times p)$	$\exp(-\beta)$
`nohr`	Prop. Hazard	$-\beta \times p$	$-\beta$
`time`	A.F.T.	β	β

- Not all distributions share the same reporting options.
- Key take away: make sure to know what form you are getting and to be clear in your tables!

Example: Position Taking on NAFTA

Assume we have single-failure data with time-invariant covariates (one record per subject) and possible right-censoring. Here we examine the number of months a cabinet lasts.

```
use duration02continuous, clear
histogram durat
stset durat
stdescribe
stsum
sts graph, survival
sts graph, survival by(caretk2)
```

Computer Exercise for Continuous-Time Duration Models

Commands for this are in `exercise02continuous.do`.

- Open the file `exercise02continuous.dta`;
- Explore the data;
- Use the `stset` command to declare survival time data;
- Run a continuous-time survival model;
- Explore different parametric specifications and the Cox model;
- Account for right censoring.

Time-Varying Covariates

- Many analysis have independent variables that change at somewhat regular intervals.
- This is easy in discrete EHA since the data are already arranged over time.
- For continuous models we have to set up the data with a new observation for each period in which a variable changes – they don't have to be constant units, though.
- Ignoring this would only consider the value at the time of failure rather than the chance of failure at the different values of the TVCs.
- We then identify units when we `stset` the data.

TVCs in NAFTA

I have modified the data to include the (lagged) daily count of the number of supporters and opponents of NAFTA as well as the net difference. This changes day by day.

- Now we need multiple records per observation – one for every day it is at risk.
- We then create a variable indicating the day of failure (announcement).
- We have to `stset` the data to account for this new structure.

Nonproportional Hazards

As noted, standard models assume that the proportionate effect of a variable on the hazard does not change over time, here illustrated with the Weibull:

$$\begin{aligned} h(Y_i|X_i) &= p\lambda_i^p Y_i^{p-1}, \\ &= p\exp(-X_i\beta)^p Y_i^{p-1}; \end{aligned}$$

Nonproportional Hazards

As noted, standard models assume that the proportionate effect of a variable on the hazard does not change over time, here illustrated with the Weibull:

$$\begin{aligned} h(Y_i|X_i) &= p\lambda_i^p Y_i^{p-1}, \\ &= p\exp(-X_i\beta)^p Y_i^{p-1}; \\ h(Y_i|X_i+1) &= p\exp(-(X_i+1)\beta)^p Y_i^{p-1}; \end{aligned}$$

Nonproportional Hazards

As noted, standard models assume that the proportionate effect of a variable on the hazard does not change over time, here illustrated with the Weibull:

$$
\begin{aligned}
h(Y_i|X_i) &= p\lambda_i^p Y_i^{p-1}, \\
&= p\exp(-X_i\beta)^p Y_i^{p-1}; \\
h(Y_i|X_i+1) &= p\exp(-(X_i+1)\beta)^p Y_i^{p-1}; \\
\frac{h(Y_i|X_i+1)}{h(Y_i|X_i)} &= \frac{p\exp(-(X_i+1)\beta)^p Y_i^{p-1}}{p\exp(-X_i\beta)^p Y_i^{p-1}}, \\
&= \frac{\exp(-(X_i+1)\beta)^p}{\exp(-X_i\beta)^p}, \\
&= \exp(-\beta)^p.
\end{aligned}
$$

(Note that $\exp(-p\beta)$ is the hazard ratio interpretation.)

Nonproportional Hazards

As noted, standard models assume that the proportionate effect of a variable on the hazard does not change over time, here illustrated with the Weibull:

$$
\begin{aligned}
h(Y_i|X_i) &= p\lambda_i^p Y_i^{p-1}, \\
&= p\exp(-X_i\beta)^p Y_i^{p-1}; \\
h(Y_i|X_i+1) &= p\exp(-(X_i+1)\beta)^p Y_i^{p-1}; \\
\frac{h(Y_i|X_i+1)}{h(Y_i|X_i)} &= \frac{p\exp(-(X_i+1)\beta)^p Y_i^{p-1}}{p\exp(-X_i\beta)^p Y_i^{p-1}}, \\
&= \frac{\exp(-(X_i+1)\beta)^p}{\exp(-X_i\beta)^p}, \\
&= \exp(-\beta)^p.
\end{aligned}
$$

(Note that $\exp(-p\beta)$ is the hazard ratio interpretation.)

But this assumption may not hold.

Testing the PH Assumption

- If this assumption is violated, we get biased estimates.
- To test the PH assumption in a parametric model, use piecewise regression or explicit interactions with time.
- To test the PH assumption in a Cox model, we calculate the model's residuals.
  - A global test plots these against time;
  - A single variable test plots them against the variable.
- Then do explicit statistical tests.

Testing the PH Assumption

- This is often tested using the Schoenfeld residuals, which are related to the score function.
- These calculate the average difference between the value of a covariate for all units that fail at t_k to all units at risk at t_k.
- The test evaluates whether this average difference correlates with time.
- Under Grambsch and Therneau's scaled adjustment and if the time scale is linear it simplifies to a regression of the scaled residuals on time.

Correcting for NPH

- If there is evidence of NPH, add an interaction of the offending variable with time (usually $\ln(t)$), re-run the model, and repeat.
- Note that these tests can detect other model specification issues (e.g., omitted variables).
- Also note that once you add an interaction with time, you have TVCs.

Computer Exercise for Nonproportional Hazards

Commands for this are in `exercise03nph.do`.

- Open the file `exercise03nph.dta`;
- Explore the data;
- Use the `stset` command to declare survival time data;
- Run a Cox continuous-time survival model;
- Test for NPH;
- Correct for NPH using Stata's `tvc()` option.;
- Correct for NPH "by hand" by converting the data to TVC using `stsplit` and adding the interaction with time to correct NPH.

Repeated Events

- Repeated events are common in Political Science application (e.g., war, government duration).
- If we treat them independently, then nothing needs to be done – just treat them as completely distinct units.
- But this may often be incorrect.
- We must consider when observations enter and leave the risk set over time;
- And whether and how to account for heterogeneity across events and units.

Figure: Single-Failure Continuous Duration Data

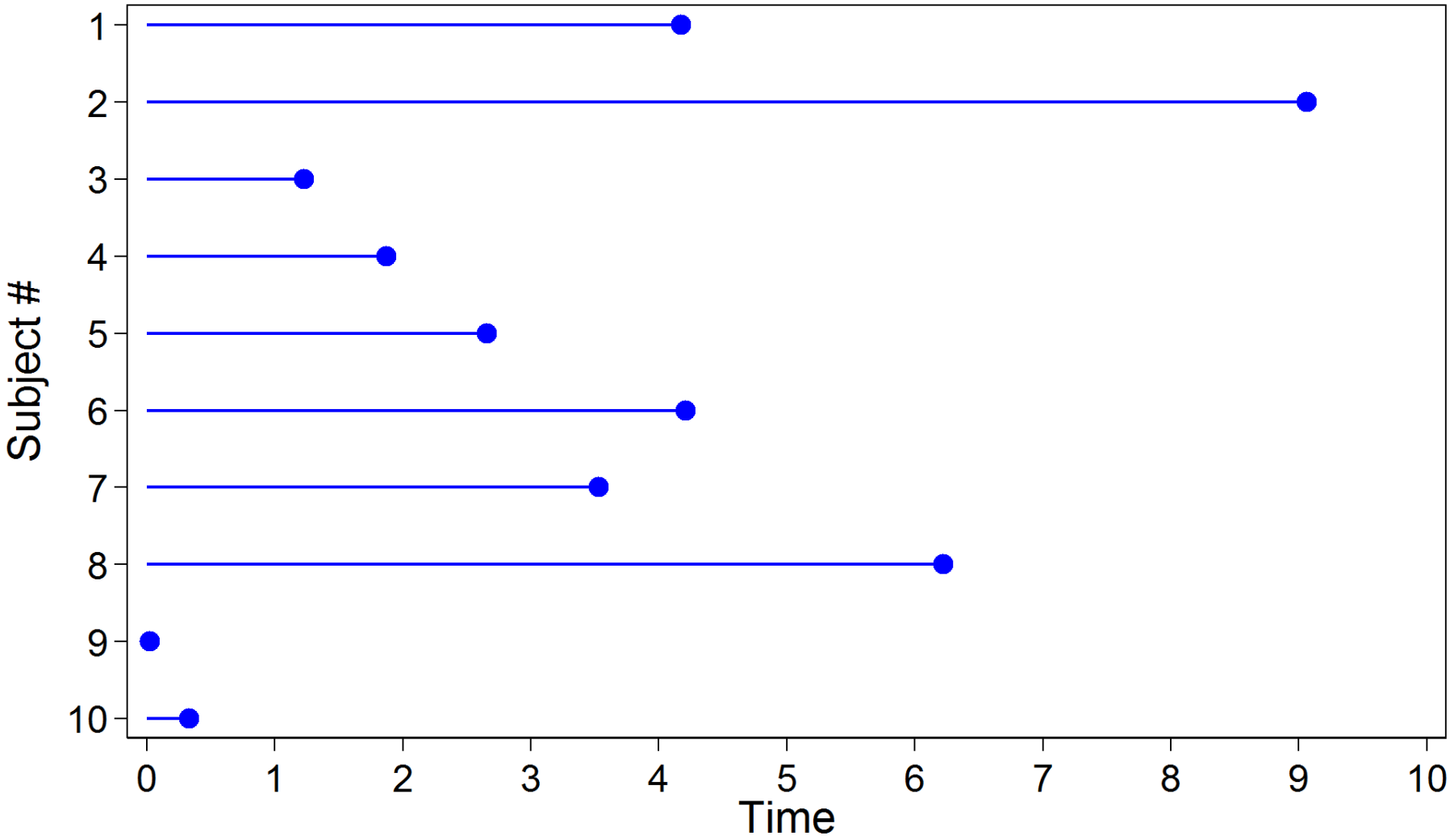


Figure: Repeated Events: What we Observe

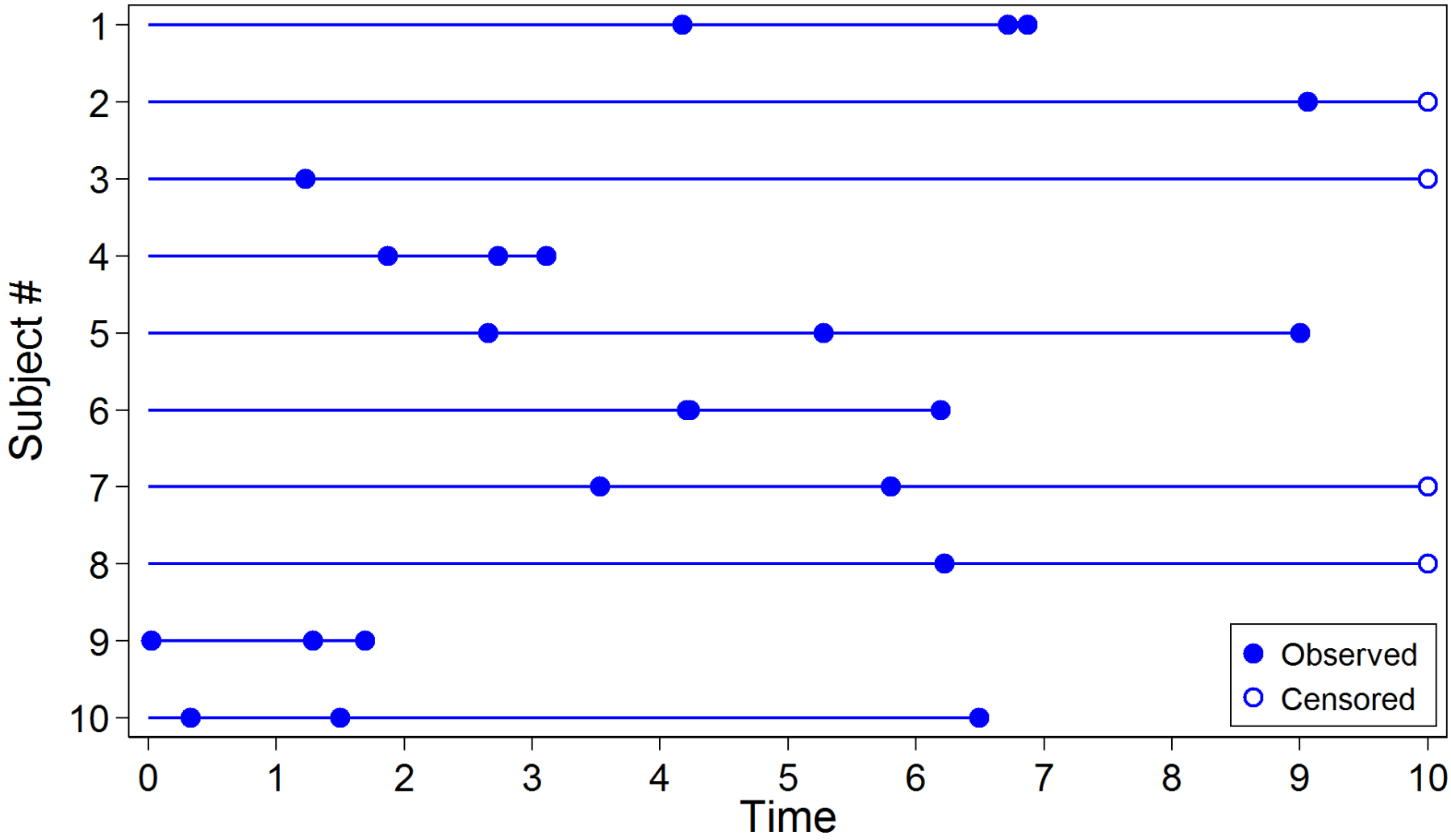


Figure: Repeated Events: Gap Time Setup

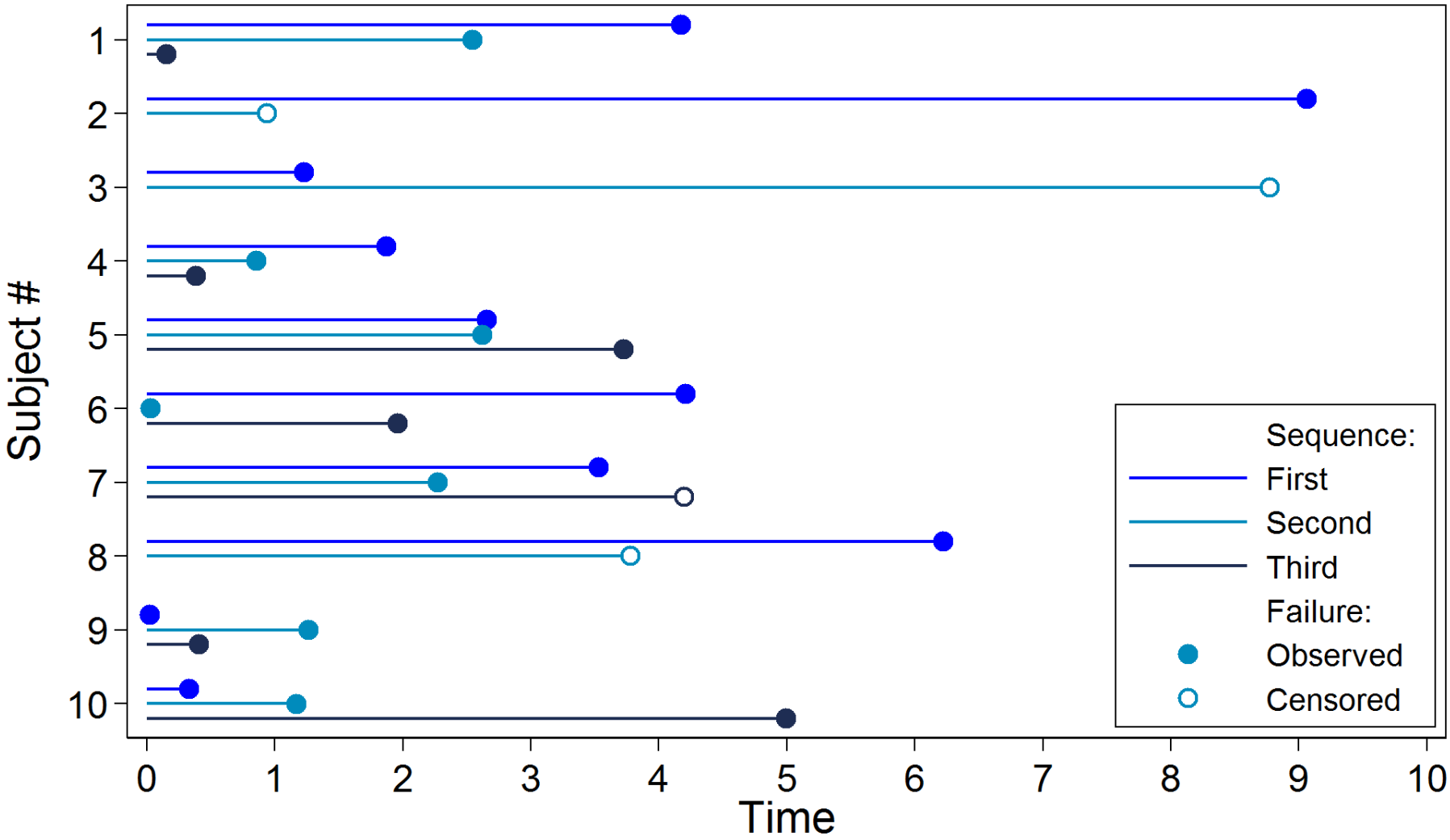


Figure: Repeated Events: Total Time Setup

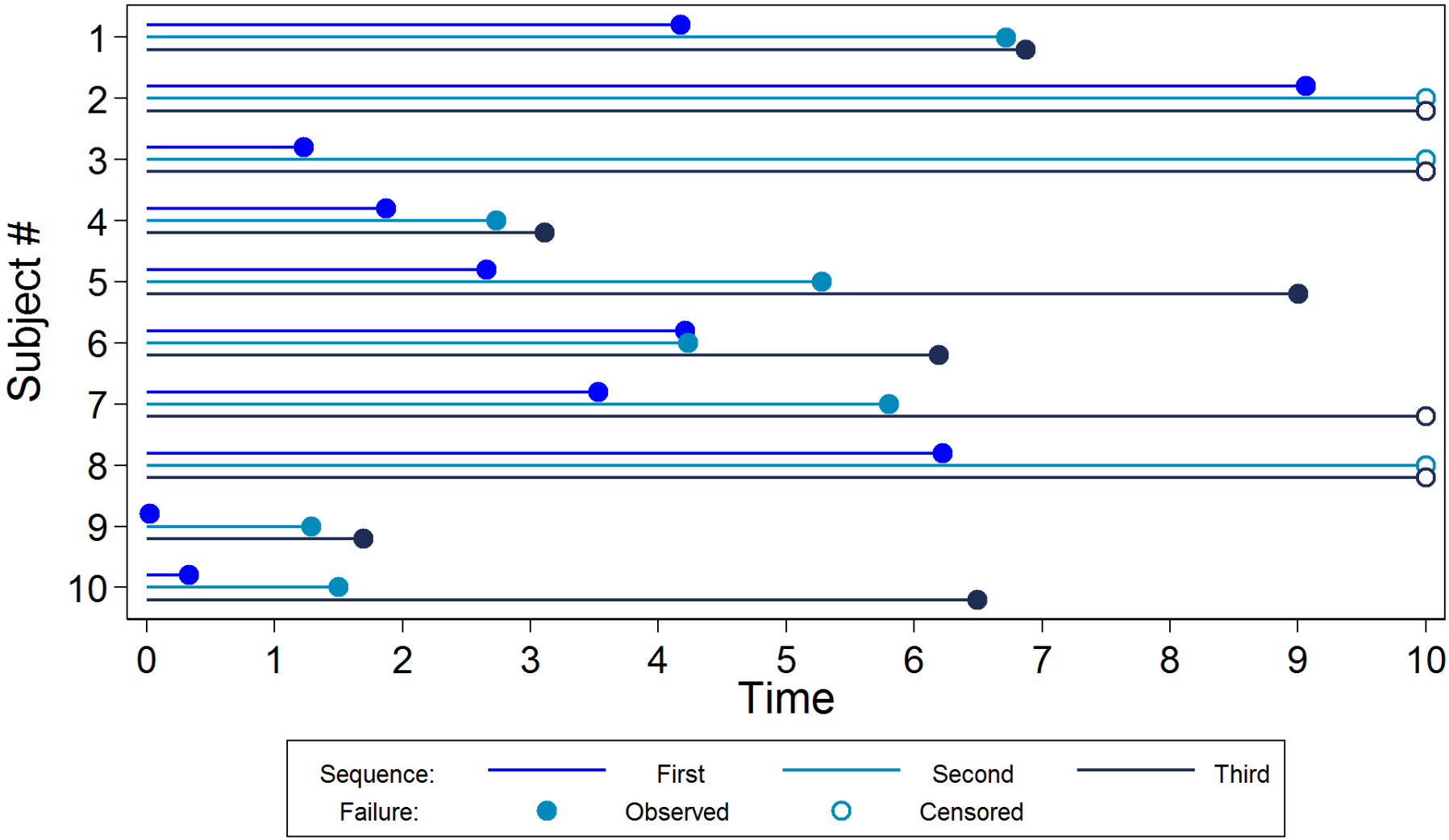
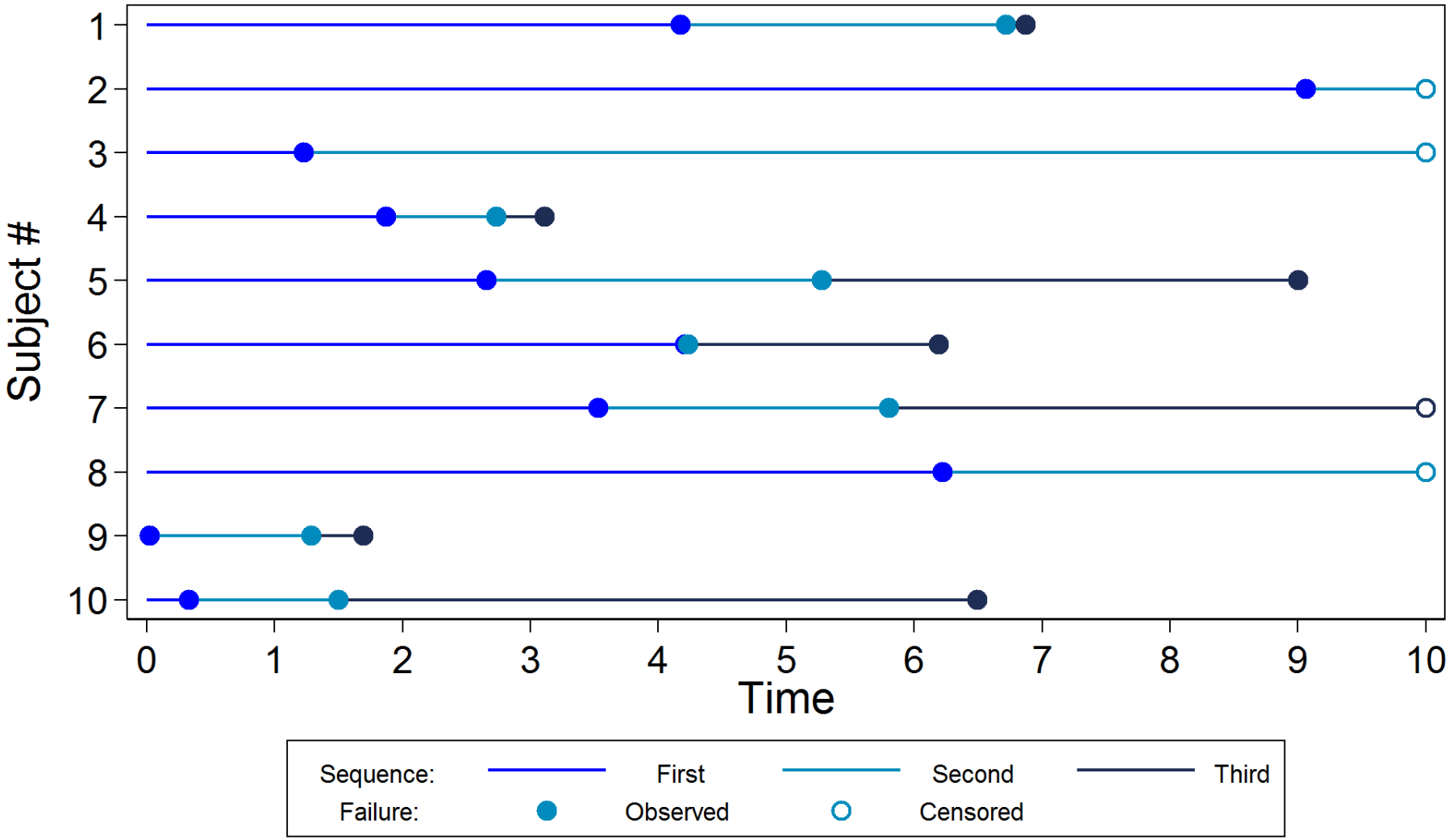


Figure: Repeated Events: Counting Process Time Setup



Key Choices Repeated Events

- Our first choice is to treat them as:
  1. Gap time: time since last failure;
  2. Total time: all begin at the origin (t=0);
  3. Counting process: subsequent events begin at failure time rather than resetting to zero.
- Our second choice covers heterogeneity:
  1. Does the baseline hazard increase by event number?
  2. Do variables have different effects by event number?
  3. Do we want to account for frailties across units (i.e., random effects)?

A typical Political application will evaluate gap time with restricted hazards, stratifying by failure via fixed effects.

Competing Risks

- Competing risks occur when there is more than one way to fail:
  1. Legislator's career could end in retirement or election loss;
  2. War could end with a negotiated settlement or a clear victor.
- It's likely that there will be different factors that determine each type of failure.
- In a data set we only observe one type of failure (the first one).
- So we have to treat the other(s) as censored at the point of observed failure.
- Easy to estimate with separate continuous duration models for each failure type by redeclaring the data with the appropriate failure indicator. Or with a multinomial logit model for discrete EHA.

Figure: Single-Failure Continuous Duration Data

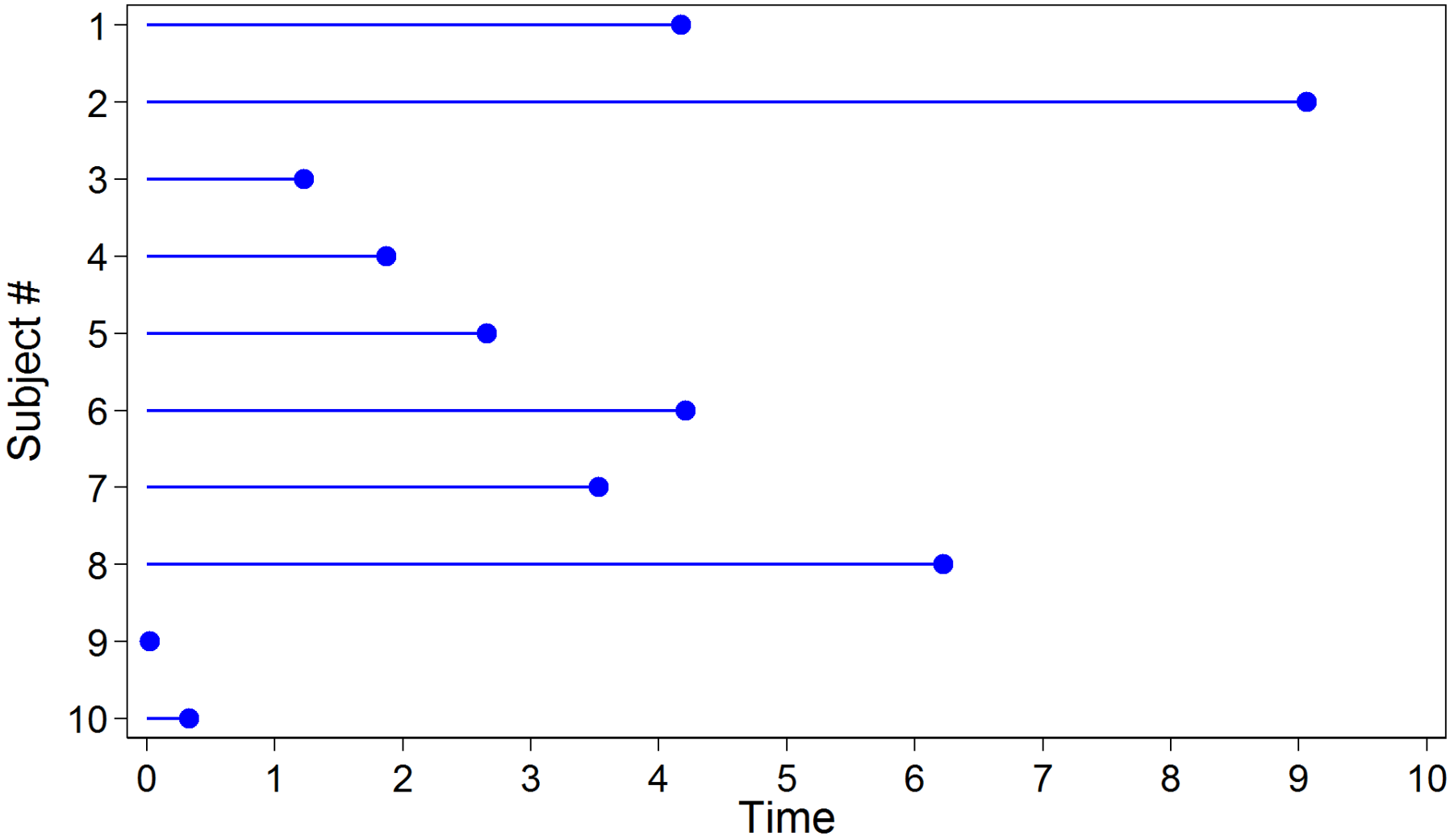


Figure: Competing Risks: The Underlying Data

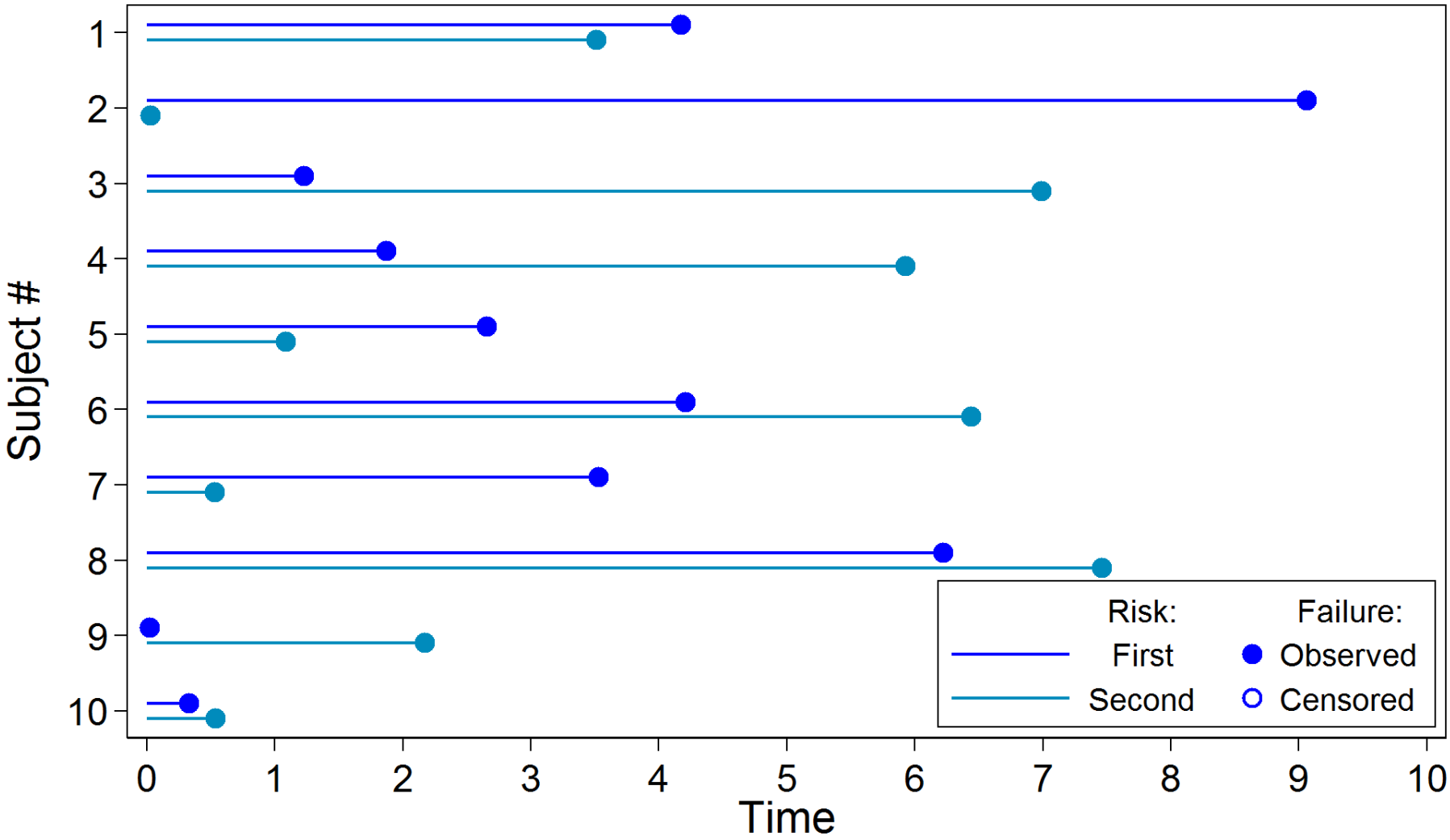
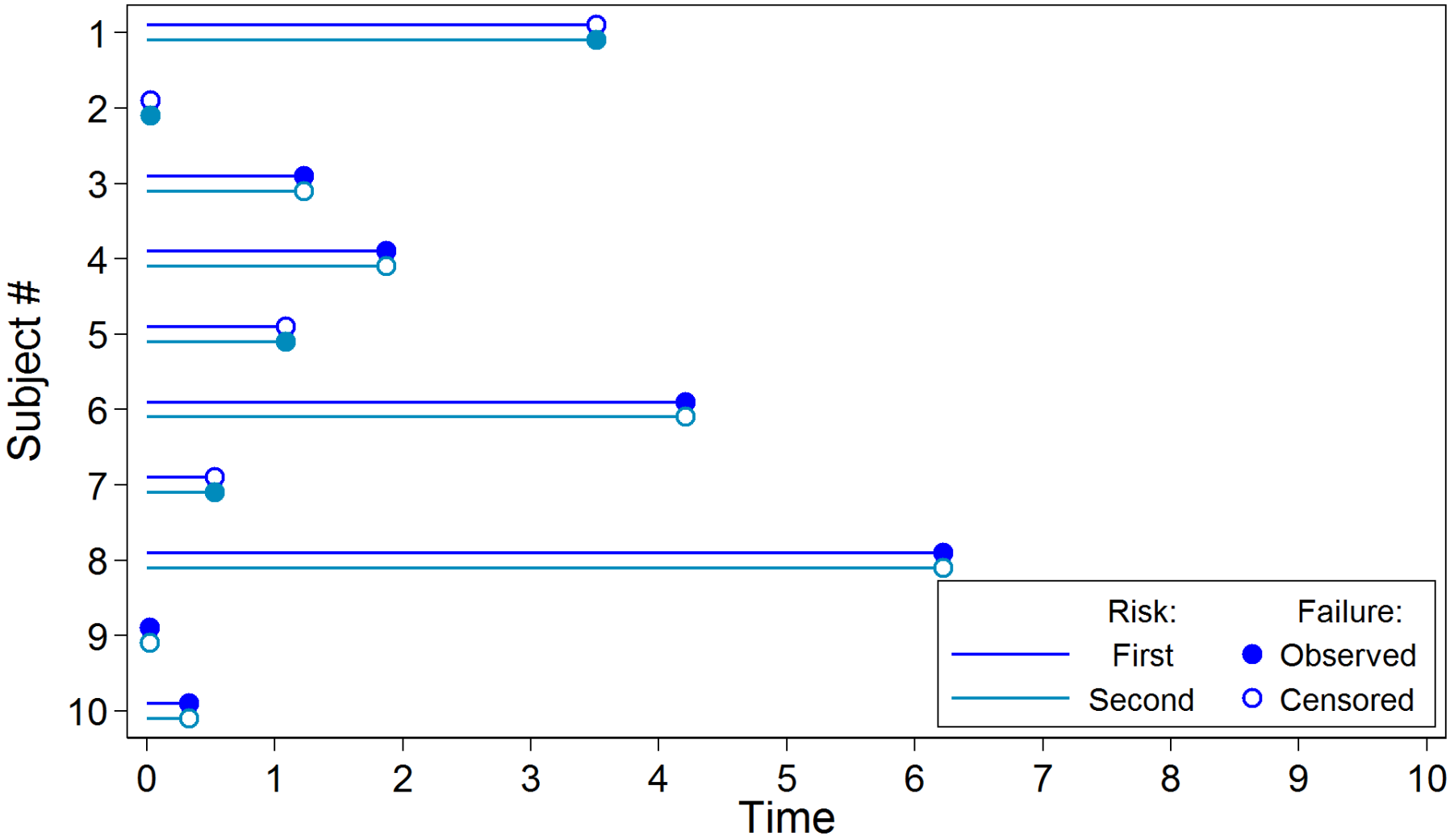


Figure: Competing Risks: What we Observe



Competing Risks: Estimation

The time of failure is the first event to occur:

$$Y_i = \min\{Y_{i1}, Y_{i2}, \ldots Y_{iK}\}$$

We build the likelihood by combining this failure time, k, with right censoring for the other failure types, $j \neq k$:

$$\begin{aligned}
\Pr(Y_i = y_{ik}|X_i) &= \Pr(Y_{ik} = y_{ik}|X_{ik}) \prod_{j \neq k} \Pr(Y_{ij} > y_{ik}|X_{ij}), \\
&= f(y_{ik}|X_{ik}) \prod_{j \neq k} S_j(y_{ik}|X_{ik}); \\
\mathcal{L}(\theta|Y, X) &= \prod_{i=1}^{n} \left[f(y_{ik}|X_{ik}) \prod_{j \neq k} S_j(y_{ik}|X_{ik}) \right].
\end{aligned}$$

Computer Exercise for Competing Risks

Commands for this are in `exercise04cr.do`.

- Open the file `exercise04cr.dta`;
- Explore the data;
- Use the `stset` command to declare survival time data;
- Run a Cox continuous-time survival model;
- Now allow for competing risks and compare the results.

Split Population Models

- Our data might include units that will never experience the event – they are "cured".
- But we can't distinguish this from right censoring in most cases.
- So we assume we have two types of observations in the data: those that will fail (but might not during the study period) and those that will never fail.
- We account for the probability an observation is "cured" in estimation and can even include covariates to model this probability as a logit.
- In Stata, these models can be estimated if one installs the `spsurv` or `lncure` packages.

Split Population Models: Estimation

Start with a binary choice model for population assignment:

$$Z_i = \begin{cases} 0 \text{ if } W_i\gamma + \epsilon_i \leq 0 \\ 1 \text{ if } W_i\gamma + \epsilon_i > 0. \end{cases}$$

There is one way an observation fails:

$$\begin{aligned} \Pr(Y_i = y_i|X_i, W_i) &= \Pr(Y_i = y_i|X_i, W_i, Z_i = 0)\Pr(Z_i = 1|W_i), \\ &= f(y_i|X_i, Z_i = 0)\Pr(Z_i = 0|W_i). \end{aligned}$$

While there are two ways that an observation might not fail:

$$\begin{aligned} \Pr(Y_i > T_i|X_i, W_i) &= \Pr(Y_i > T_i|X_i, W_i, Z_i = 1)\Pr(Z_i = 1|W_i) + \Pr(Z_i = 0|W_i) \\ &= S(T_i|X_i, Z_i = 1)\Pr(Z_i = 1|W_i) + \Pr(Z_i = 0|W_i), \\ &= \exp(-T_i\lambda_i)\frac{1}{1+\exp(W_i\gamma)} + \frac{\exp(W_i\gamma)}{1+\exp(W_i\gamma)}. \end{aligned}$$

Nonrandom Sample Selection

- Units may self-select into the duration process and the ones that do may be different than the ones that don't.
- This is the continuous time duration analog to Heckman's selection model for linear regression.
- Bias occurs when unobserved factors that influence selection also influence duration.
- Ignoring selection leads to biased estimates.
- The solution parallels Heckman's: model both processes simultaneously and allow for correlation in the error terms.

Pooled EHA

- Used for simultaneously estimating clusters of similar events, e.g.:
  1. Pro- and anti-abortion policies;
  2. Position timing on lots of bills.
- Essentially stacks separate EHA data and runs one big model to examine common influences.
- Need to think about accounting for heterogeneity across events: baseline hazards, duration dependence, coefficients.
- The goal is to balance between full parsimony and completely separate estimating across events.
- Multilevel modeling can provide some nice leverage on this tradeoff.

Computer Exercise for Time-Varying Covariates

Commands for this are in `exercise05tvc.do`.

- Open the file `exercise05tvc.dta`;
- Run a Weibull continuous-time survival model;
- Now let's create multiple record per subject data;
- Re-`stset` the data and compare the results;
- Then we'll create time-vary variables and include them in our analysis.