Survival Analysis

Survival Analysis Overview

- Survival analysis examples
- Survival analysis set up and features
- Extensions of basic survival analysis
- Survival, hazard, and cumulative hazard functions
- Nonparametric analysis (Kaplan-Meier survival function)
- Parametric models (Exponential, Weibull, Gompertz, and Log-logistic)
- Semi-parametric models (Cox proportional hazard model)

Survival Analysis

Survival analysis is also called duration analysis, transition analysis, failure time analysis, and time-to-event analysis.

Survival analysis examples

- Finance: Loan performance (borrowers obtain loans and then they either default or continue to repay their loans)
- Economics: Firm survival and exit
- Economics: Time to retirement, finding a new job, etc.
- Economics: Adoption of new technology (firms either adopt the new technology or still haven't adopted it)

Survival analysis set up

- Subjects are tracked until an event happens (failure) or we lose them from the sample (censored observations).
- We are interested in how long they stay in the sample (survival).
- We are also interested in their risk of failure (hazard rates).

Survival analysis features

- The dependent variable is duration (time to event or time to being censored) so it is a combination of time and event/censoring.
    - time variable = length of time until the event happened or as long as they are in the study
    - the event variable = 1 if the event happened or 0 if the event has not yet happened
    - Instead of an event variable, a censor variable can be defined. The censored variable =1 if the event has not happened yet, and 0 if the event has happened.

Time	Event/ Failure	Censored	Explanation
15	0	1	Event hasn't happened yet (censored)
22	1	0	Event happened (not censored)
78	0	1	Event hasn't happened yet (censored)
34	1	0	Event happened (not censored)

- The hazard rate is the probability that the event will happen at time *t* given that the individual is at risk at time *t*.
- Hazard rates usually change over time.
    - The probability of defaulting on a loan may be low in the beginning but increases over the time of the loan.

Extensions of the basic survival analysis

- Multiple occurrences of events (multiple observations per individual)
  - borrower may have repeated restructuring of the loan
  - firm may adopt technology in some years but not others
- More than one type of event (include codes for events, e.g. 1, 2, 3, 4)
  - borrower may default (one type of event) or repay the loan earlier (a second type of event)
  - firms may adopt different types of technologies
- Two groups of participants
  - the effect of two types of educational programs on technology adoption rates
- Time-varying covariates
  - borrower's income may have changed during the study which caused the default.
- Discrete instead of continuous transition times
  - exits are measured in intervals (such as every month)
- There may different starting times – we need to measure time from the beginning time to the event.

Survival, hazard, and cumulative hazard functions

- The dependent variable duration is assumed to have a continuous probability distribution $f(t)$.
- The probability that the duration time will be *less than t* is:

$$F(t) = Prob(T \leq t) = \int_0^t f(s)ds$$

- *Survival function* is the probability that the duration will be *at least t*:

$$S(t) = 1 - F(t) = Prob(T \geq t)$$

- *Hazard rate* is the probability that the duration will end after time *t*, given that it has lasted until time *t*:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

- The hazard rate is the probability that an individual will experience the event at time *t* while that individual is at risk for experiencing the event.

Nonparametric models

- Nonparametric estimation is useful for descriptive purposes and to see the shape of the hazard or survival function before a parametric model with regressors is introduced.

Time t_j	Number at risk n_j	Number of events d_j	Number of censored observations	Hazard function $\lambda = d_j/n_j$	Cumulative hazard function $\Lambda(t_j)$	Survival function $S(t_j)$
3	100	10	3	10/100=0.1	0.1	1-0.1=0.9
4	100-10-3=87	3	2	3/87=0.034	0.1+0.034 =0.134	0.9*(1-0.034) =0.87
5	87-3-2=82	6	1	6/82=0.073	0.134+0.073 =0.207	0.87*(1-0.073)=0.81

- Think about the shapes of the hazard function and survival function plotted over time.

Survival analysis nonparametric procedure

- Sort the observations based on duration from smallest to largest $t_1 \leq t_2 \leq \cdots \leq t_n$
- For each duration, determine the number of observations at risk n_j (those still in the sample), the number of events d_j and the number of censored observations m_j.
- Calculate the hazard function as the number of events as a proportion of the number of observations at risk

$$\lambda(t_j) = \frac{d_j}{n_j}$$

- *Nelson-Aalen estimator of the cumulative hazard function* – calculated by summing up hazard functions over time:

$$\Lambda(t_j) = \sum \frac{d_j}{n_j}$$

- *The Kaplan-Meier estimator of the survival function* – take the ratios of those without events over those at risk and multiply that over time.

$$S(t_j) = \prod \frac{n_j - d_j}{n_j}$$

A few facts about the Kaplan-Meier survival function

- It is a decreasing step function with a jump at each discrete event time.
- Without censoring, the Kaplan-Meier estimator is just the empirical distribution of the data.

Parametric and semiparametric models

- Unlike the nonparametric estimation, the parametric models also allow the inclusion of independent variables.

Parametric models

- Parametric models can assume different parametric forms for the hazard function.

Parametric model	Hazard function λ	Survival function S
Exponential	γ	$\exp(-\gamma t)$
Weibull	$\gamma\alpha t^{\alpha-1}$	$\exp(-\gamma t^{\alpha})$
Gompertz	$\gamma\exp(\alpha t)$	$\exp(-(\gamma/\alpha)(e^{\alpha t}-1))$
Log-logistic	$\alpha\gamma^{\alpha}t^{\alpha-1}/(1+(\gamma t)^{\alpha})$	$1/(1+(\gamma t)^{\alpha})$

- The exponential model has a constant hazard rate over time.

Cox proportional hazard model

- The hazard rate in the Cox proportional hazard model is defined as:

$$\lambda(t|\mathbf{x},\beta) = \lambda_0(t)\exp(\mathbf{x}'\beta)$$

Estimation of the parametric models

- For the parametric and semiparametric models, report both the coefficients and hazard ratios.
- Interpretation of coefficients: a positive coefficient means that as the independent variable increases the time-to-event *decreases*, (lower duration or more likely for the event to happen).
- Interpretation of hazard rates: a hazard ratio of 2 (0.5) means that for a one unit increase in the x variable, the hazard rate (probability of event happening) increases by 100% (decreases by 50%). A hazard rate of greater than 1 means that it is more likely for the event to happen.

Coefficient	Hazard rate	Conclusion
Positive	>1	Lower duration, higher hazard rates (more likely for the event to happen).
Negative	(0,1)	Higher duration, lower hazard rates (less likely for the event to happen).

Survival Analysis Example

Survival Analysis Example

- Survival, hazard, and cumulative hazard functions
- Nonparametric analysis (Kaplan-Meier survival function)
- Parametric models (Exponential, Weibull, Gompertz, and Log-logistic)
- Semi-parametric models (Cox proportional hazard model)

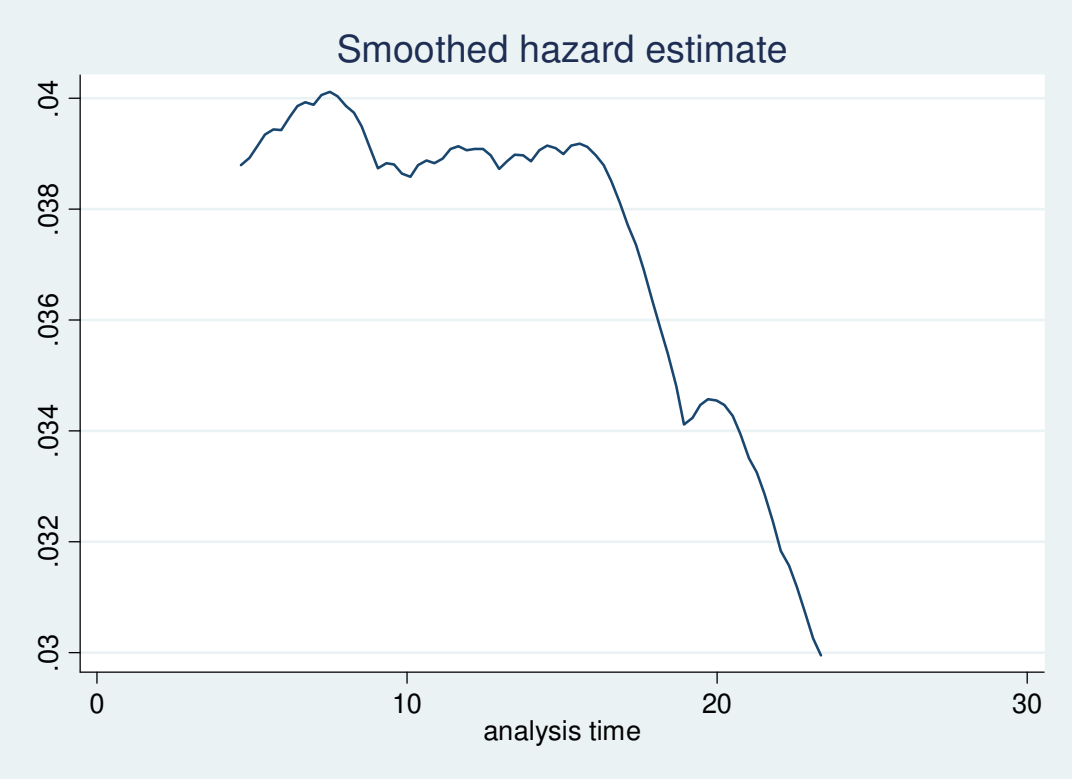
- We want to study the unemployment duration – the length of time it takes someone to find a full-time job.
- Data from the January Current Population Survey's Displaced Workers Supplements (DWS) for years 1986, 1988, 1990 and 1992.
- Dependent variable: duration (number of periods being unemployed), event (finding a job)
- Independent variables: log wage, claim unemployment insurance, and age.
- Summary statistics: Subjects tracked from 1 to 28 periods. They either find a job (event) or are still looking (censored). Number of subjects is 3,343; time at risk (periods summed over the subjects) is 20,887. Number of failures is 1,073 or 32% of sample has failed. Incidence rate is 5.13% which is the number of failures divided by the time at risk.

Survival function table

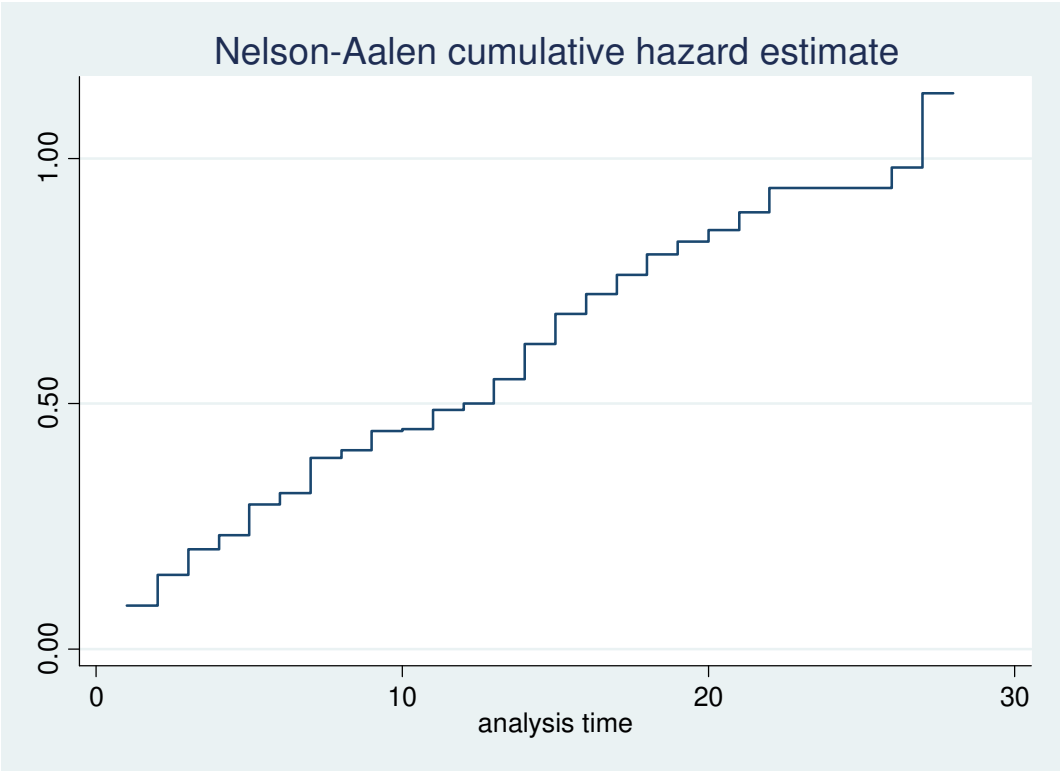
Time	Number of subjects	Failure/event	Net lost/censored	Survival Function
1	3343	294	246	0.91
2	2803	178	304	0.85
3	2321	119	305	0.81
4	1897	56	165	0.79
5	1676	104	233	0.74
6	1339	32	111	0.72
7	1196	85	178	0.67
….				
25	58	0	10	0.38
26	48	2	13	0.37
27	33	5	24	0.31
28	4	0	4	0.31

Nonparametric methods

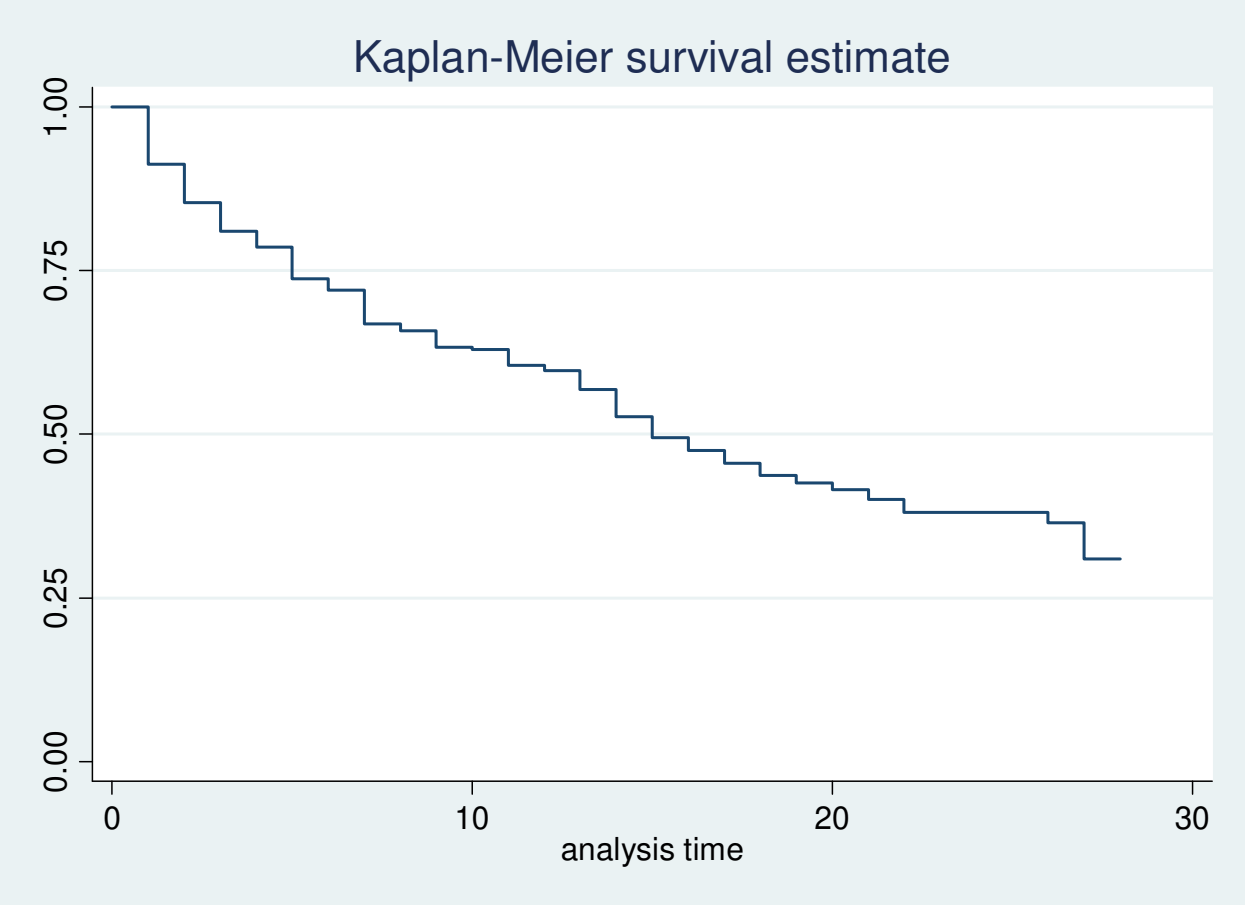
The hazard rate shows the probability of having the event (finding a job) going down from 4% to 3% over 25 time periods.



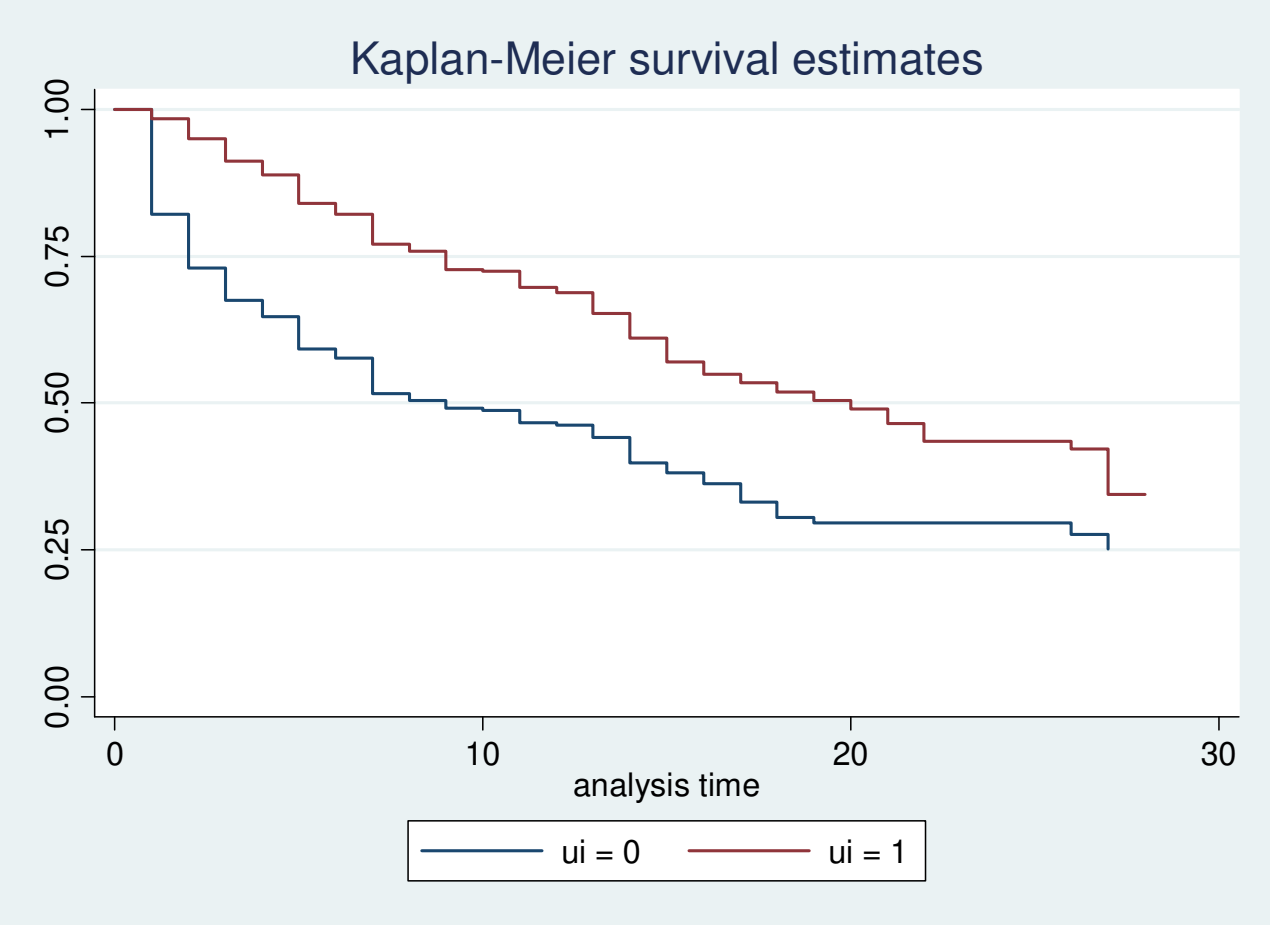
The Nelson-Aalen cumulative hazard estimate is non-decreasing.



The Kaplan-Meier survival function shows that survival probabilities go down to 31% over 28 time periods. This means that 31% of the individuals still have not found a job after 28 time periods.



Survival function for the group that has unemployment insurance (ui=1) and the group that does not have unemployment insurance (ui=0). The survival functions show at any point in time that claiming unemployment insurance is associated with higher survival rate. This means that if someone receives unemployment benefits he/she is more likely to still be unemployed.



Parametric regression model coefficients

Duration of unemployment	Exponential regression coefficients	Weibull regression coefficients	Gompertz regression coefficients	Cox proportional hazard model coefficients
Log wage	0.48*	0.49*	0.48*	0.46*
Claim unemployment insurance	-1.08*	-1.11*	-1.06*	-0.98*
Age	-0.01*	-0.01*	-0.01*	-0.01*

Results reported here are Stata results. SAS produces opposite signs for the exponential, Weibull, and Gompertz regression. R produces opposite signs for exponential and Weibull regression. Also, the Weibull regression in SAS and R give different estimates.

- Interpretation of the coefficients: Individuals with higher wages have *lower* unemployment duration, meaning will terminate unemployment faster. Individuals who claim unemployment insurances have *higher* unemployment durations, meaning they terminate unemployment slower.

Parametric regression model hazard rates

Duration of unemployment	Exponential regression hazard rates	Weibull regression hazard rates	Gompertz regression hazard rates	Cox proportional hazard model hazard rates
Log wage	1.62*	1.63*	1.61*	1.58*
Claim unemployment insurance	0.34*	0.33*	0.35*	0.37*
Age	0.99*	0.99*	0.99*	0.98*

- Interpretation of the hazard rates: A unit increase in the log wage is associated with 1.62-1 = 62% increase in the hazard rates. For individuals who claim unemployment insurance the hazard rates are 0.34-1 = 66% lower. In other words, individuals with higher wages are more likely to find a job and those that claim unemployment insurance are less likely to find a job.