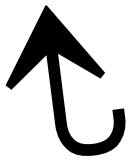
BST 210
Applied Regression Analysis



Lecture 22
Plan for Today

Introduction to Survival Analysis:

- A new outcome variable in town: Time!
- Where have we seen a similar model?
- Survival data: definition, notation and censoring
- Kaplan-Meier estimate of survivor curve

Motivation

Methods discussed previously have facilitated assessing *events* and *time* in the following ways:

- (Linear)
  - how *events* relate to potential *time* factor (i.e. age, minutes)
- (Logistic)
  - whether *event* occurred or not (yes/no)
  - potentially adjusting for *time* factor (i.e. age, minutes, year)
- (Poisson)
  - how many *event*s occurred (count)
  - over specified interval of *time* (i.e. 1 min, 24 hrs, 1 year)
- Next…Survival
  - a slightly different take on *events* and *time*!

Motivation

Most recently:

- Recall Poisson random variable $Y \sim \text{Poisson}(\lambda t)$, defined by

$$P(Y = y) = \frac{e^{-\lambda t}(\lambda t)^y}{y!}$$

where

Y counts number of events occurring in time interval t

y = 0, 1, 2, … counts are independent

λ = (constant) number of cases per unit time (incidence rate)

and

$$\log(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \log(t)$$

(or)

$$y = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \log(t))$$

Motivation

Now what if we had for example:

- T_i = *time* to some *event* for patient i (critical distinction)
- $X_{i1}, X_{i2}, \ldots X_{ip}$ covariates for patient i

then $T_i \mid X_{i1}, X_{i2}, \ldots X_{ip}$ follows what we call an exponential distribution

$$f(t) \sim \lambda e^{-\lambda t}$$

and $h(t) = \lambda$ can be thought of as the 'hazard' of event occurring at t, where

$$h(t|x_1, x_2, \ldots x_p) = h_0 \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)$$

with h_0 = baseline 'hazard' (when all covariates = 0)

- Does this form of a model look familiar?

Motivation

Where have we seen this before?

- In Poisson regression (has exponential form, recall) we model the number of *events* Y occurring with rate λ over a fixed amount of *time* t
- In the Exponential model, we also have λ and t, but we are actually estimating the amount of *time until an event* occurs, where that event occurs with hazard λ (instantaneous rate of occurrence)
- Similar: Both are based on understanding the *rate at which events occur* (and thus likelihood based estimation of β's for covariates associated with these rates are essentially the same across the 2 models!)
- Different: Each model has different consideration of how *t (time)* is factored in (this will be the critical difference between survival data vs non)
- Let's now introduce survival data and the analysis of such →

Survival Data: It's just a matter of 'time'

- In some studies, the response variable of interest is the ***length of time between an initial observation and the occurrence of a subsequent event***
- The event is often called a *failure*
- The time from the initial observation until failure is called the ***survival time***
- Examples: Time from birth until death, time from start of treatment until serious adverse event, time from randomization to relapse or death, time from entry in a cohort study until myocardial infarction, time between live births, time until marriage, duration of geographic stay, etc.
- Do you think distributions of survival times would tend to be skewed or not?

Survival Data

- **The analysis of incidence rates** (e.g., Poisson regression) does not allow event rates to vary over time, after controlling for covariates (though breaking up of time periods for a subject can allow this)

- **The analysis of proportions** (e.g., logistic regression) only considers whether an event occurs or does not occur

- **Survival analysis**

  - uses the time from the starting point to the occurrence of the event

  - can allow the incidence rate to vary over time

Goals of Survival Analysis

Same themes as previous linear modeling…

- To estimate the distribution of survival times for a population
- To test the equality of survival distributions (e.g., treated vs. control group, smokers vs. nonsmokers)
- To estimate and control for the effects of other covariates when investigating the relationship between a predictor variable and survival time
- To assess model fit

Functions defining time T

- There are essentially 4 related functions, all which help to define our survival (or failure) times T.

- Let's let random variable *T* represent the time from a start point to an event of interest (e.g., time from start of treatment to serious adverse event, time from disease remission to recurrence)

- By definition, *T* must be ≥ 0

- The *survival function* S(*t*) is then defined as:

$$S(t) = P(T > t)$$

- S(t) is the proportion of individuals who are event-free at time t, or the probability of not having the event until time t.

Functions defining time T

- It can also be shown that

$$S(t) = P(T > t)$$
$$= 1 - P(t \leq T)$$
$$= 1 - F(t)$$
$$= \int_t^{\infty} f(x)\mathrm{d}x$$

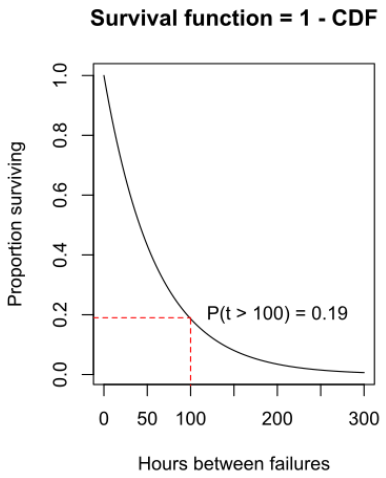
- where

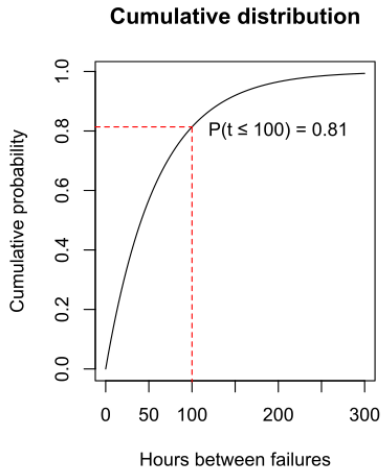
$$f(t) = F'(t) = -S'(t)$$

is the probability density function, and

$$F(t) = P(t \leq T) = 1 - S(t)$$

is the cumulative distribution function.





Survival Function

Note that by definition,

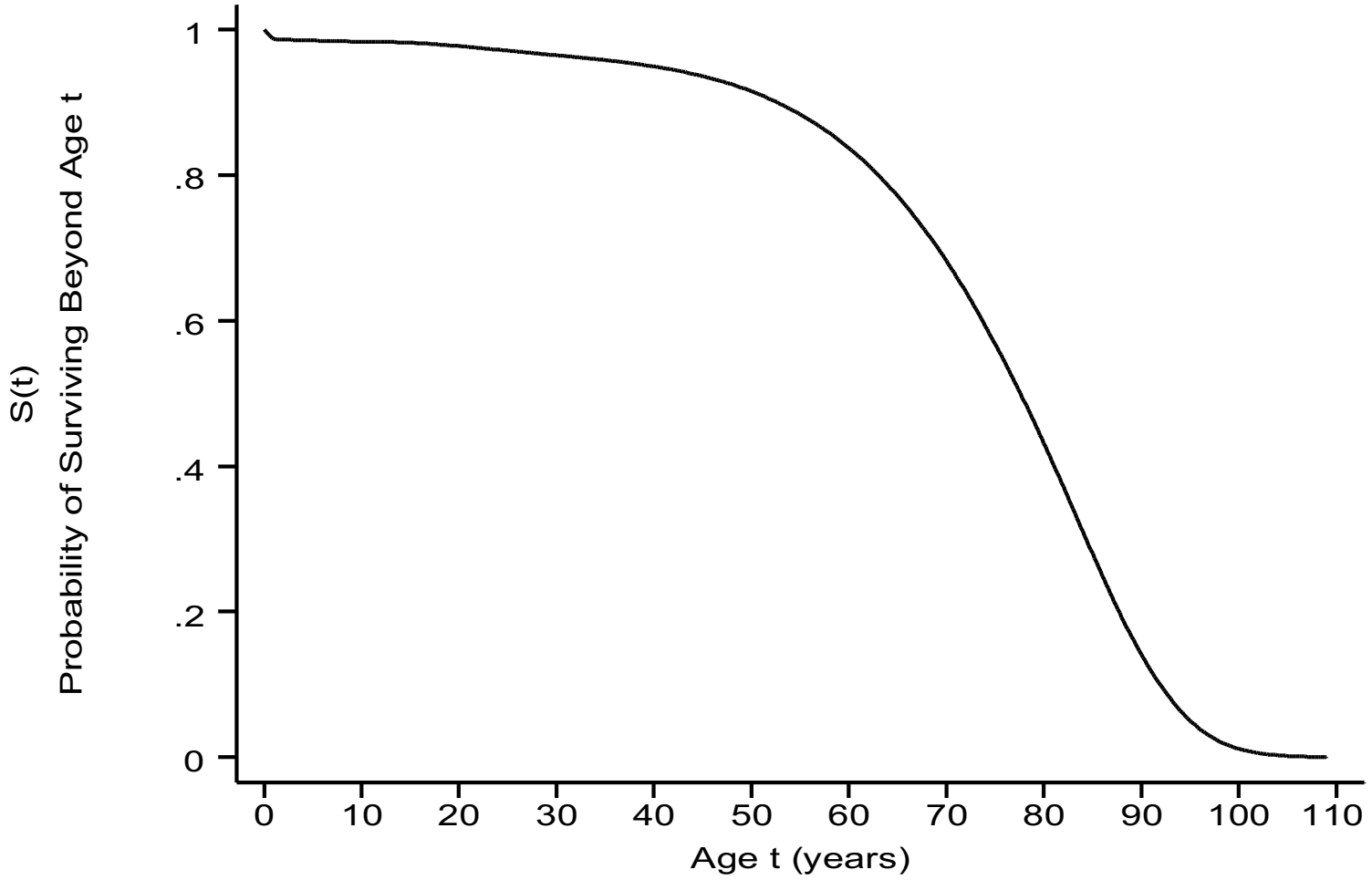
(a) $0 \leq S(t) \leq 1$ for each $t \geq 0$

(b) $S(0) = 1$

(c) If $t_2 \geq t_1$, then $S(t_2) \leq S(t_1)$, i.e., survival functions are non-increasing over time

- The graph of a survival function $S(t)$ versus time t is called a *survival curve*

Survival Curve



Survival Data

- As we previously noted, distributions of survival times are frequently skewed to the right

- Thus, mean survival time is usually not a good summary measure; the mean is pulled up toward the outlying values (and there may be censored values, making estimation of the mean difficult)

- Instead, estimation of percentiles of the survival distribution is often preferable

Example: Chemotherapy for Leukemia

Let's get a feel for what survival data actually looks like…

- A pilot study was conducted among leukemia patients currently in remission
- Leukemia patients were randomized to two groups:

  (a) maintenance chemotherapy group

  (b) control group (no chemotherapy)

- Day of randomization was the start point of the study

Example: Chemotherapy for Leukemia

- What do you think would be the goal of such a study?
- The goal of the study was to compare the survival experience of the two treatment groups
- Survival time is defined as time of randomization to time to relapse of leukemia
- The endpoint is not just whether or not remission was maintained (e.g., binary), but for how long it was maintained (e.g., time to event)

Example: Chemotherapy for Leukemia

- There were 11 patients randomized to the maintenance chemotherapy group, and 12 patients randomized to the control group
- Time to relapse was measured for each subject
- These times can also be thought of as length of complete remission (in weeks)

Example: Chemotherapy for Leukemia

'Times to relapse' in weeks for this study are:

- Maintenance chemotherapy group (*n* = 11)

9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+

- Control group (*n* = 12)

5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

- *This is all cool…but what do the + signs represent?*

Caveat with Survival Data: *Censoring*

- This brings us to perhaps the most defining characteristic of survival data…we call it *censoring* (basically a missing data problem)

- Since most studies occur over a finite time period, the event of interest may not have occurred for some subjects during the study period

- All that is known is that the time to an event *T* is greater than the period of follow-up *C*, where *C* is called the *censoring time*

- For subjects who have an event during the study period, we have the actual event time *T*

Example: Chemotherapy for Leukemia

Recall the times in weeks for this example study are:

- Maintenance chemotherapy group (n = 11)

9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+

- Control group (n = 12)

5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

Example: Chemotherapy for Leukemia

What do we see here…

- Patient # 2 in the maintenance chemotherapy group had a remission time of 13 weeks, which indicates that he/she relapsed during the study at 13 weeks

- Patient # 3 in the maintenance chemotherapy group has a remission time of 13+, which means that the patient remained in remission for 13 weeks, did not have an event (i.e., relapse of leukemia), and was then *censored* at that time point

- What does that mean, and how does censoring work?

Causes of Censoring

- We could see no event by the end of the study (study closed)
- Withdrawn from the study prior to its end (i.e., loss to follow-up), due to either

  - lack of interest of the participant, or

  - an adverse event which presented a medical contraindication for continuing with the study

- Death from another cause (e.g., death due to cancer in a study where the endpoint is myocardial infarction)

Effects of Censoring

- Standard statistical methods could be used if all event times *T* were observed (continuous measurements)
- …although the distribution of event times would most likely be skewed (meaning we might apply nonparametric methods or need transformations of the data, e.g., log(*T*), sqrt(*T*))
- However, the presence of censoring makes this impossible, since some event times are unknown

Types of Censoring

So how does censoring tend to present itself?

(a) Right censoring

(b) Left censoring

(c) Interval censoring

Right Censoring

- We know that the event time *T* is greater than the censoring time *C*
- This is the most common form of censoring
- For example, a subject enrolled in a clinical trial might be followed until the end of the trial without having an event
- Loss to follow-up before an event occurs is another example of right censoring

Left Censoring

- We know that the event time T is smaller than a specific time C (i.e., $T < C$)
- For example, for the condition retinitis pigmentosa (RP), an important endpoint is time to legal blindness, defined as a visual acuity of 20/200 or worse, or a visual field of $< 20^{o}$ in both eyes
- A patient seen at an RP clinic and enrolled in a study of adults at age 18 years may already have reached the endpoint of legal blindness
- In this case, we would only know that $T < 18$

Interval Censoring

- This type of censoring occurs when we know that the event occurred within an interval, but do not know the exact time of the event

- For example, a subject comes to an RP clinic at age 20 and is not legally blind, but comes back at age 30 and is legally blind

- In this case, $20 < T < 30$

Effects of Censoring on Inference

- To make valid comparisons of distributions of survival times between groups, we need to assume that the censoring time (C) is independent of the survival time (T)
- (should ring familiar with topic of missing data)

- This assumption implies that:

$$P(C \mid T = t) = P(C) \text{ for all } t$$

- This is called *noninformative censoring*

Effects of Censoring on Inference

- An example of when the noninformative censoring assumption is valid is in the case of administrative censoring
- A study is terminated at a fixed date before an event has occurred
- T is unknown, but C is equal to the study termination date minus the enrollment date

Effects of Censoring on Inference

- The assumption of noninformative censoring is not valid if subjects who are at higher risk for an endpoint (i.e., *T* is short) tend to also be at higher risk for an adverse event and subsequent loss-to-follow-up (i.e., *C* is also short)

- Another example would be a smoking cessation study, where the endpoint is time to relapse

- Subjects are randomized to receive either nicotine gum or placebo gum

Effects of Censoring on Inference

- Subjects who stop responding to phone calls or emails at 2 months (*C* is short) might be more likely to relapse subsequently (*T* is short) than subjects who do respond and remain in the study (*C* is long)

- The assumption of independence of *C* and *T* is usually untestable, since we do not completely observe the (*T*, *C*) pair for all subjects

Effects of Censoring on Inference

- *Main point!* The estimation of survival probabilities is complicated due to the presence of censored data

- For example, in the leukemia example, there are 7 out of 11 patients (64%) in the maintenance group who survived for at least 20 weeks

- Does this mean that the estimated survival probability at 20 weeks for the maintenance group should be

$$\hat{S}(20) = 0.64?$$

Example: Chemotherapy for Leukemia

Recall the times in weeks for this example study are:

- Maintenance chemotherapy group (n = 11)

9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+

- Control group (n = 12)

5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

Effects of Censoring on Inference

- The answer to whether $\hat{S}(20) = 0.64$ is: probably not
- This estimate will be *biased downwards* because it does not take censoring into account
- Why would this be?
- The third subject in the maintenance group survived for 13 weeks and then withdrew
- The subject might have survived for 20+ weeks if he or she had been followed longer – we just don't know, and instead our 'censoring-naive' calculation counts this subject as a failure, contributing to a decrease in S(t)

Options for Estimation and Inference

- What options then, do we have for estimating survival curves and thus probabilities, in the presence of censored data?
- A nonparametric method for estimating survival curves that takes censoring into account is called the *Kaplan-Meier estimator* or the *product-limit estimator*
- No assumptions are made about the underlying distribution of survival times (ie nonparametric)
- Drawback: cannot account for any covariate data
- Let's examine this approach →

Example: Chemotherapy for Leukemia

- Let us first consider the maintenance group only, and *(incorrectly) assume there is no censoring* (remove the + signs)

- Times to event in weeks are:

  Maintenance chemotherapy group (*n* = 11)

  9, 13, 13, 18, 23, 28, 31, 34, 45, 48, 161

- Now we just need to define who is *at risk* and when -->

Define a Risk Set

- We define a patient to be at risk for an event at time *t* if they have not experienced an event before time *t* and are not yet censored at time *t*

- Thus, for the maintenance chemotherapy group (n=11), the risk set consists of

  11 patients at 0 weeks, 11 patients at 1 week, ..., 11 patients at 9 weeks

  10 patients at 10 weeks, ..., 10 patients at 13 weeks

  8 patients at 14 weeks, ..., 8 patients at 18 weeks, etc.

  (where recall that weeks = 9, 13, 13, 18, 23, 28, 31, 34, 45, 48, 161)

Product Limit Method

How then does the Kaplan-Meier estimation method work?
Let…

- t_i = distinct observed failure times (uncensored) in increasing order so that

$$t_1 < t_2 < t_3 < \ldots < t_{m-1} < t_m$$

- m = number of distinct failure times
- n_i = number of subjects in the risk set at time t_i
- d_i = number of failures (events) at time t_i

Product Limit Method

And let:

$\hat{p}_i = d_i / n_i$

= probability of failure at time t_i given that a subject is in the risk set at time t_i.

Let $\hat{q}_i = 1 - \hat{p}_i$

= probability of surviving beyond time t_i given that a subject is in the risk set at time t_i.

Kaplan-Meier Estimate

Then we have

- S(0) = 1
- Estimated survival decreases after each of the failure times
- In order for subject *i* to have $T_i > t_k$, subject *i* needs to

  (1) be at risk at time t_1 and have $T_i > t_1$,
  (2) be at risk at time t_2 and have $T_i > t_2$,
  …
  (k) be at risk at time t_k and have $T_i > t_k$

Kaplan-Meier Estimate

- The Kaplan-Meier Estimate S(t) simply multiplies these 'running' conditional probabilities together up through each event time

- Thus, using the multiplication rule of probability,

$$\Pr(T_i > t_k) = \Pr(T_i > t_1 \,|\, \textit{in risk set at time } t_1)$$
$$\text{x } \Pr(T_i > t_2 \,|\, \text{in risk set at time } t_2)$$
$$\text{x } \ldots$$
$$\text{x } \Pr(T_i > t_k \,|\, \text{in risk set at time } t_k).$$

Kaplan-Meier Estimate

- Note that this can also be stated in terms of the survival function, as:

$S(t_i) = P(T > t_i)$

$= P(T > t_1) \times P(T > t_2 \mid \text{survived to } t_1)$

$\times P(T > t_3 \mid \text{survived to } t_2) \times \ldots$

$\times P(T > t_i \mid \text{survived to } t_{i-1})$

- In general, the K-M survival function is estimated as

$$\hat{S}(t) = \prod_{j=1}^{k} (1 - d_j / n_j) = \prod_{j=1}^{k} \hat{q}_j \ \ for\ t_k \le t < t_{k+1}.$$

Kaplan-Meier Estimate

- Where specifically for a certain risk set we have:

$$\hat{S}(t) = 1 \; for \; 0 \le t < t_1,$$

$$\hat{S}(t) = \hat{\Pr}(T > t_1 \,|\, in \; risk \; set \; at \; time \; t_1)$$

$$= 1 - d_1 / n_1 = \hat{q}_1 \; for \; t_1 \le t < t_2,$$

$$\hat{S}(t) = \hat{\Pr}(T > t_1 \,|\, in \; risk \; set \; at \; time \; t_1) \; \text{x} \; \hat{\Pr}(T > t_2 \,|\, in \; risk \; set \; at \; time \; t_2)$$

$$= (1 - d_1 / n_1)(1 - d_2 / n_2) = \prod_{j=1}^{2} (1 - d_j / n_j) \; for \; t_2 \le t < t_3.$$

Example: Chemotherapy and Leukemia (no censoring)

n_j	t_j	$S(t_j)$
11	0	1
11	9	1 – (1/11) = 10/11 = 0.909
10	13	(10/11) x (1 – 2/10) = 8/11 = 0.727
8	18	(8/11) x (1 – 1/8) = 7/11 = 0.636
7	23	(7/11) x (1 – 1/7) = 6/11 = 0.545
6	28	(6/11) x (1 – 1/6) = 5/11 = 0.455
5	31	(5/11) x (1 – 1/5) = 4/11 = 0.364
4	34	(4/11) x (1 – 1/4) = 3/11 = 0.273
	etc.	

Example: Chemotherapy for Leukemia

- Now return the *censored* observations
- Times to event in weeks are:
- Maintenance chemotherapy group (n = 11)

9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+

- Control group (n = 12)

5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45

Risk Set (censoring)

- We still define a patient to be at risk for an event at time *t* if they have not experienced an event before time *t* and are not yet censored just before time *t*

- A subject who is censored at time *t* is assumed to have had no event up to time *t*, and then gets censored (taken out of the risk set) just after time *t*

Risk Set

- Therefore, the third patient in the maintenance group (coded as 13+) is in the risk set at 13 weeks, but is not followed beyond 13 weeks

- Thus, for the maintenance chemotherapy group, the risk set consists of

  11 patients at 0 weeks, 11 patients at 1 week, ..., 11 patients at 9 weeks

  10 patients at 10 weeks, ..., 10 patients at 13 weeks

  8 patients at 14 weeks, ..., 8 patients at 18 weeks, etc.

Example: Chemotherapy and Leukemia

- For the maintenance group, $t_1 = 9$, $t_2 = 13$, $t_3 = 18$, etc.
- Note that the subject who is censored at 13 weeks is assumed to survive longer than the subject who fails at 13 weeks

Example: Chemotherapy and Leukemia (with censoring)

n_j	t_j	$S(t_j)$	
11	0	1	
11	9	1 – (1/11) = 10/11 = 0.909	
10	13	(10/11) x (1 – 1/10) = 9/11 = 0.818	
8	18	(9/11) x (1 – 1/8) = (9/11) (7/8) = 0.716	
7	23	0.716 x (1 – 1/7) = 0.614	
5	31	0.614 x (1 – 1/5) = 0.491	
4	34	0.491 x (1 – 1/4) = 0.368	
2	48	0.368 x (1 – 1/2) = 0.184	etc.

Estimating Survival Curves

- The data set must have separate columns for:

  (a) the survival time (possibly censored)

  (b) the failure indicator

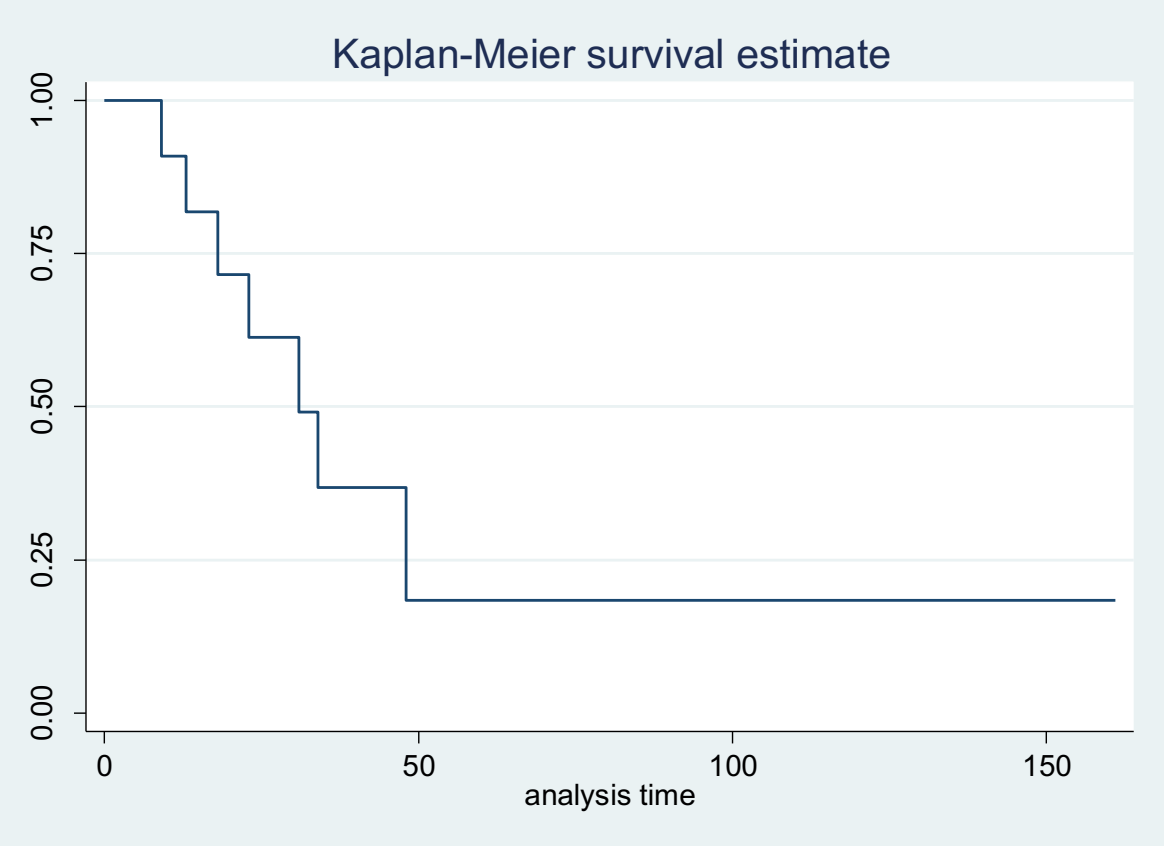
  = 1 if a subject fails

  = 0 if a subject is censored

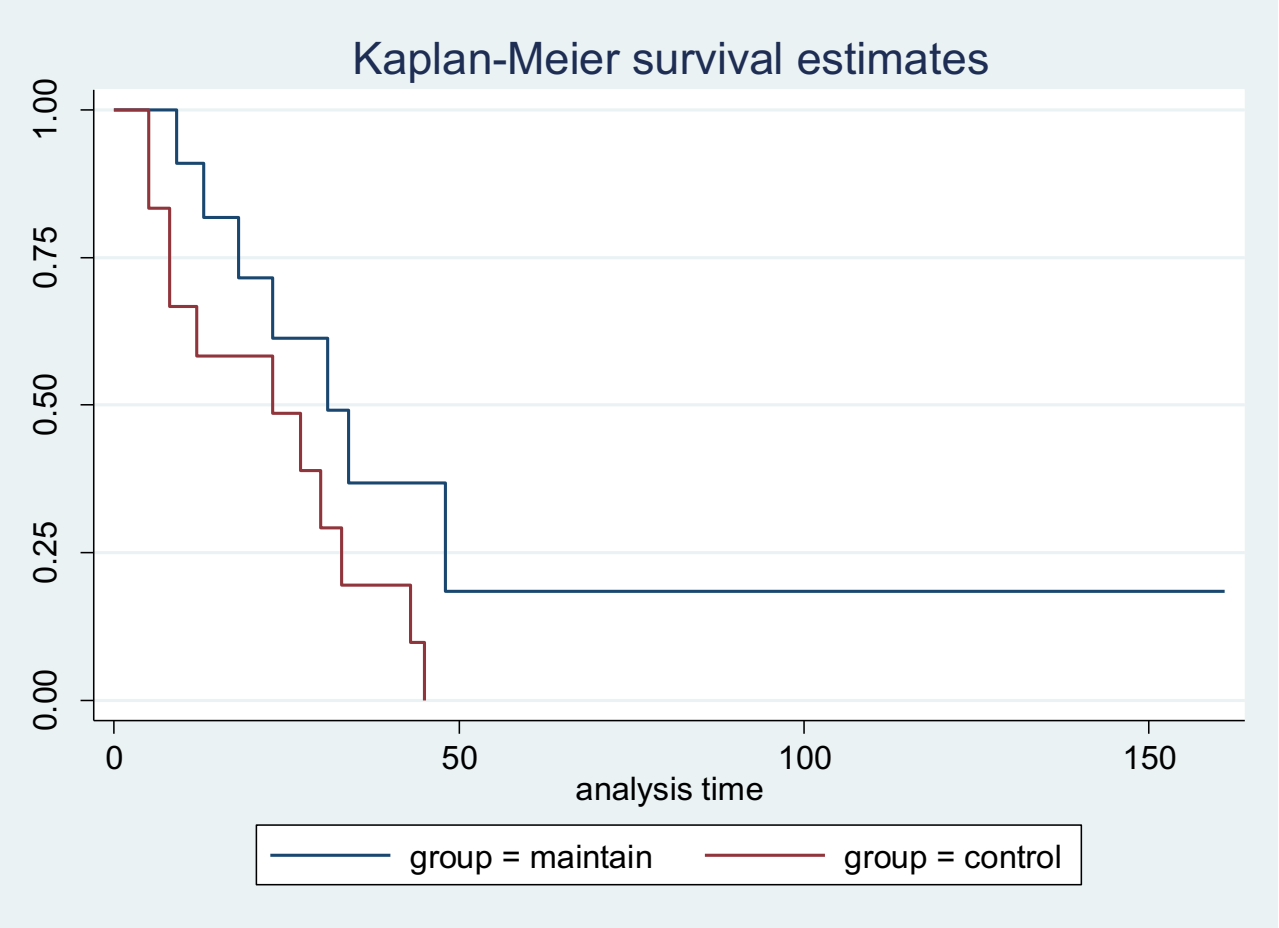
  (c) a group indicator variable (if there is one)

- Or (b') the censoring indicator (= 1 if a subject is censored, = 0 if a subject fails)

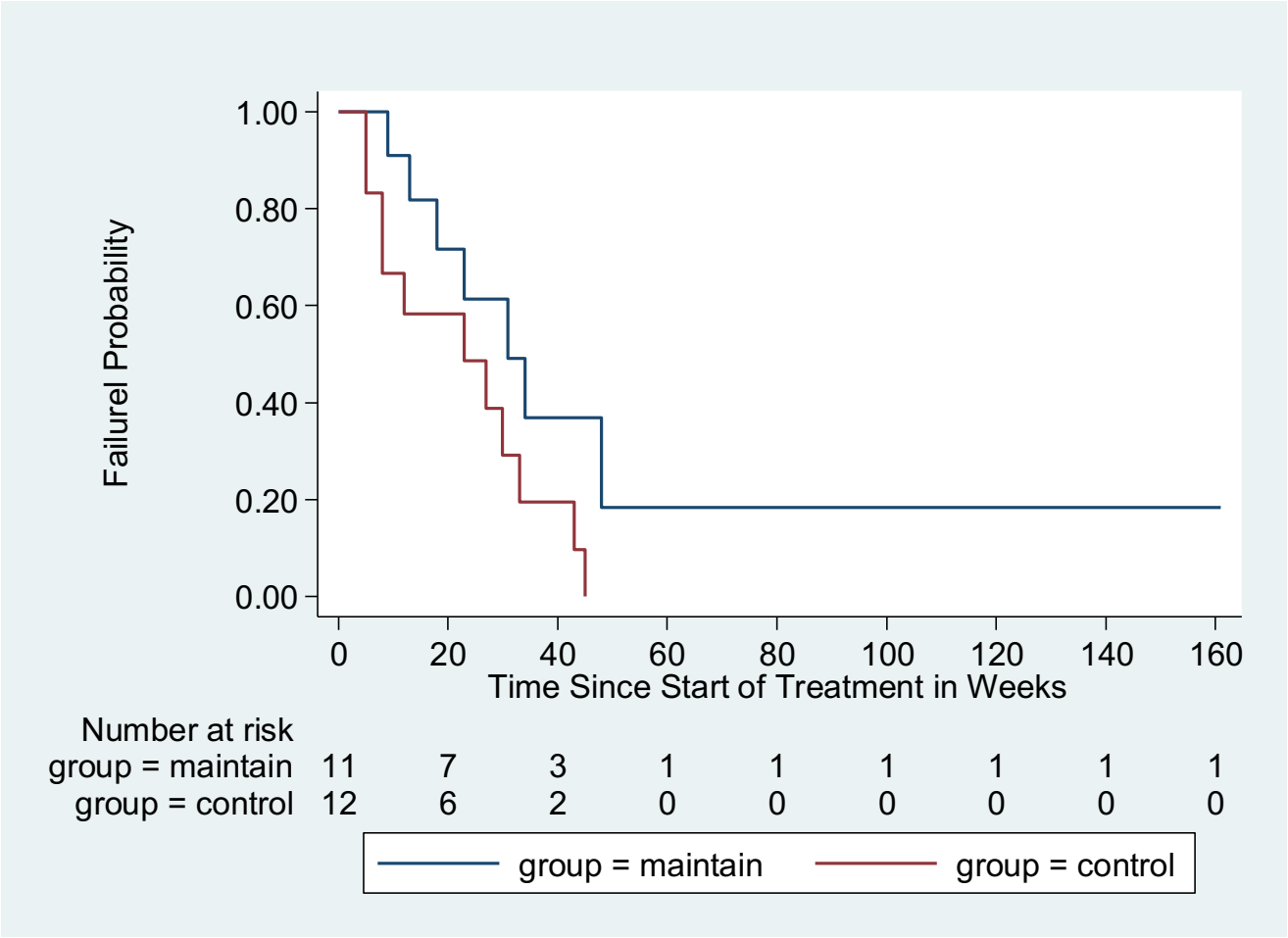
Estimated Survival Curve



Estimated Survival Curves (two groups)



Estimated Survival Curves (two groups)



Survival Function

- The survival function allows us to estimate the probability of survival beyond a specified time point *t*
- In the chemotherapy maintenance group, the estimated probability of surviving at least 23 weeks is 0.614
- The estimated probability of surviving at least 20 weeks is 0.716
- *But, what is a confidence interval or standard error estimate for these quantities?*

Interval Estimation for Survival Probabilities

- The most popular method for obtaining interval estimates for survival probabilities is to use *Greenwood's formula*

- Although survival probabilities must lie between 0 and 1, this is different than the calculation of confidence intervals for binomial proportions -- the denominator changes over time as subjects drop out of the risk set

- Greenwood's method is based on calculating a confidence interval for $\log[S(t)]$ rather than $S(t)$ itself, since the natural log of the survival function is more closely normally distributed than the survival function itself

- We then exponentiate the lower and upper bounds to get a confidence interval for $S(t)$

Greenwood's Formula

Using the delta method, it can be shown that:

$$\text{var}\{\ln[\hat{S}(t)]\} = \sum_{\{j:t_j \le t\}} \frac{d_j}{n_j(n_j - d_j)}$$

Also, $\ln[\hat{S}(t)]$ tends to be more normally distributed than $\hat{S}(t)$.

Hence, a 100% x (1-α) CI for ln[S(t)] is given by:

$$(c_1, c_2) = \ln[\hat{S}(t)] \pm z_{1-\alpha/2}\sqrt{\text{var}\{\ln[\hat{S}(t)]\}}.$$

The corresponding 100% x (1-α) CI for S(t) is $[\exp(c_1), \exp(c_2)]$.

This is known as Greenwood's formula.

Confidence Interval for $S(t_k)$

- In the example, there are 3 survival times in the maintenance group that are ≤ 20 weeks: 9, 13, and 18 weeks

- The variance of log[$S(t)$] will be summed over these 3 time points

Confidence Interval for $S(t_k)$

- $\hat{S}(20) = \hat{S}(18) = 0.716$ and $\ln[\hat{S}(20)] = -0.3342$.

  Also, from Greenwood's formula, since there were $n_i = 11$ patients in the risk set at 9 weeks, 10 patients at 13 weeks and 8 patients at 18 weeks, and $d_i = 1$ patient failed(relapsed) at each of these time points, we have:

  $$\text{var}[\ln(\hat{S}(20)] = \frac{1}{11(10)} + \frac{1}{10(9)} + \frac{1}{8(7)} = 0.0381.$$

  $$se[\ln(\hat{S}(20)] = \sqrt{0.0381} = 0.1951.$$

Confidence Interval for $S(t_k)$

- Thus, a 95% CI for ln[S(20)] is given by :

  $-0.3342 \pm 1.96(0.1951) = (-0.7166, 0.0482) = (c_1, c_2)$.

  The corresponding 95% CI for S(20) is :

  $[\exp(-0.7166), \exp(0.0482)] = (0.49, 1.05)$.

  Since survival estimates cannot be greater than 1,

  we truncate the upper limit and obtain a 95% CI for S(20) of $(0.49, 1.00)$.

Confidence Interval for $S(t_k)$

- Since exp(x) can be > 1, it is possible that confidence intervals based on the natural log transformation will yield upper confidence limits for $S(t)$ that are > 1, as seen in the previous example

- An alternative approach is to use the *complementary log-log transformation* instead of the log transformation when constructing confidence intervals

Confidence Interval for $S(t_k)$

- The complementary log-log transformation is given by:

$$y(t) = \log\{-\log[S(t)]\}$$

- The advantage of this transformation is that if we solve for $S(t)$ as a function of $y(t)$, we obtain:

$$S(t) = e^{-e^{y(t)}}, \text{where} -\infty < y(t) < \infty.$$

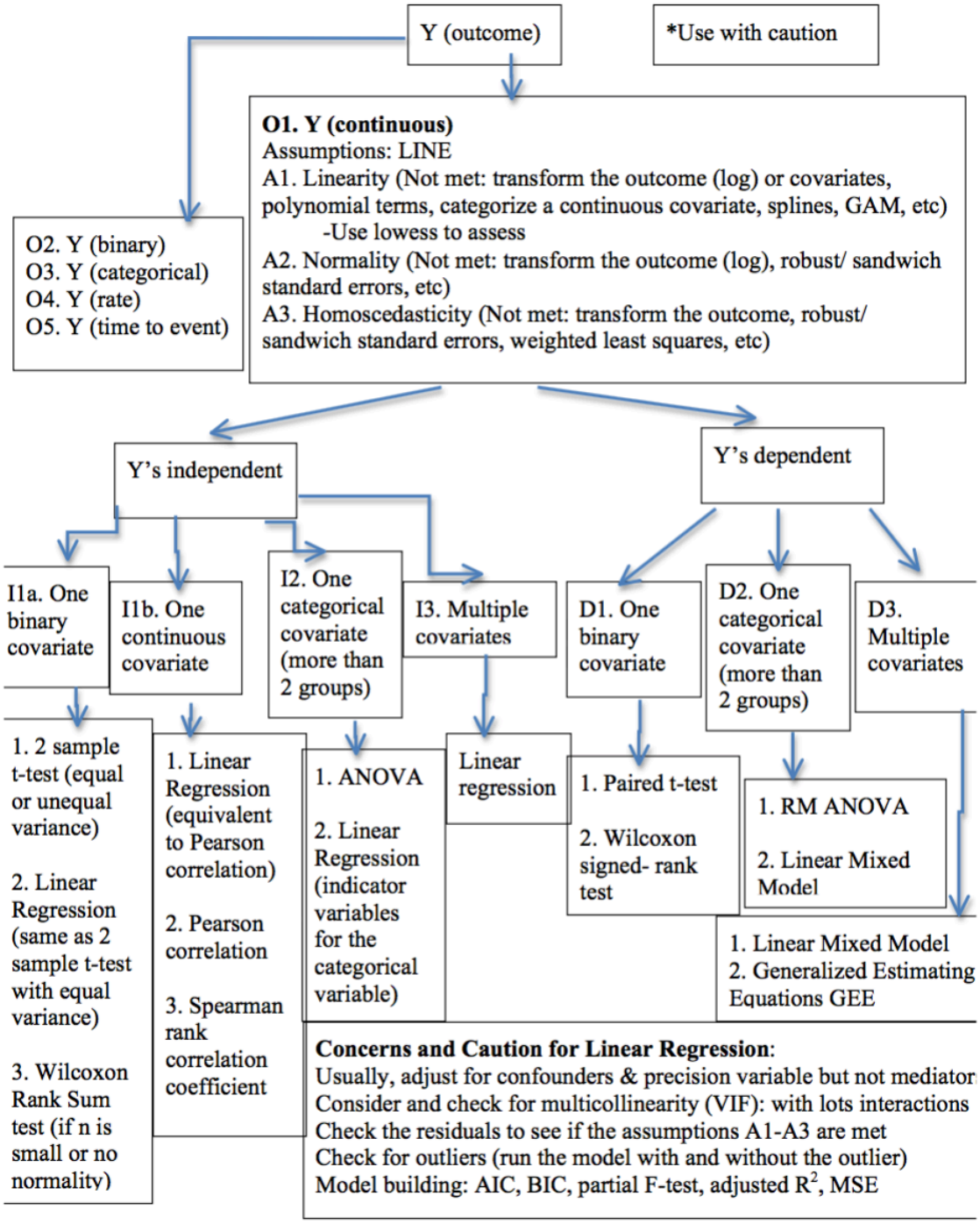
Confidence Interval for $S(t_k)$

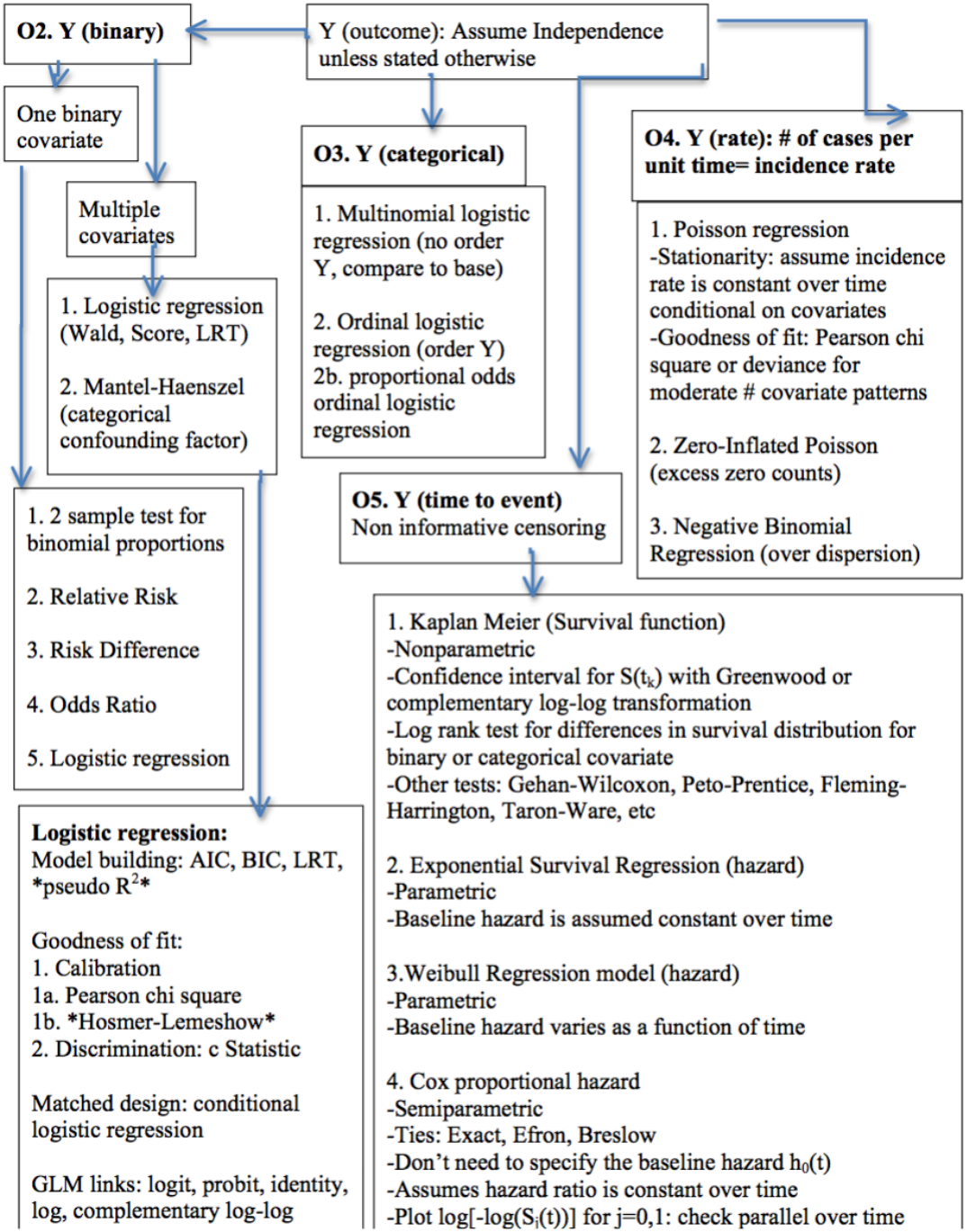
- Since $S(t) = 1$ when $y(t) = -\infty$, and $S(t) = 0$, when $y(t) = \infty$, it implies that if we obtain confidence limits for $y(t)$ and transform back to the $S(t)$ scale, the corresponding confidence limits for $S(t)$ will always be between 0 and 1.

- The variance formula is more complicated than for the log transformation

Coming Up

- Log Rank test to compare survival curves
- Much more on survival analysis! – including the Cox proportional hazards model


Y (outcome)
*Use with caution
O1. Y (continuous)
Assumptions: LINE
A1. Linearity (Not met: transform the outcome (log) or covariates, polynomial terms, categorize a continuous covariate, splines, GAM, etc)
-Use lowess to assess
A2. Normality (Not met: transform the outcome (log), robust/ sandwich standard errors, etc)
A3. Homoscedasticity (Not met: transform the outcome, robust/ sandwich standard errors, weighted least squares, etc)
O2. Y (binary)
O3. Y (categorical)
O4. Y (rate)
O5. Y (time to event)
Y's independent
Y's dependent
I1a. One binary covariate
I1b. One continuous covariate
I2. One categorical covariate (more than 2 groups)
I3. Multiple covariates
D1. One binary covariate
D2. One categorical covariate (more than 2 groups)
D3. Multiple covariates
1. 2 sample t-test (equal or unequal variance)
2. Linear Regression (same as 2 sample t-test with equal variance)
3. Wilcoxon Rank Sum test (if n is small or no normality)
1. Linear Regression (equivalent to Pearson correlation)
2. Pearson correlation
3. Spearman rank correlation coefficient
1. ANOVA
2. Linear Regression (indicator variables for the categorical variable)
Linear regression
1. Paired t-test
2. Wilcoxon signed- rank test
1. RM ANOVA
2. Linear Mixed Model
1. Linear Mixed Model
2. Generalized Estimating Equations GEE
Concerns and Caution for Linear Regression:
Usually, adjust for confounders & precision variable but not mediator
Consider and check for multicollinearity (VIF): with lots interactions
Check the residuals to see if the assumptions A1-A3 are met
Check for outliers (run the model with and without the outlier)
Model building: AIC, BIC, partial F-test, adjusted R^2, MSE


Y (outcome): Assume Independence unless stated otherwise
O2. Y (binary)
One binary covariate
Multiple covariates
1. Logistic regression (Wald, Score, LRT)
2. Mantel-Haenszel (categorical confounding factor)
1. 2 sample test for binomial proportions
2. Relative Risk
3. Risk Difference
4. Odds Ratio
5. Logistic regression
Logistic regression:
Model building: AIC, BIC, LRT, *pseudo R^2*
Goodness of fit:
1. Calibration
1a. Pearson chi square
1b. *Hosmer-Lemeshow*
2. Discrimination: c Statistic
Matched design: conditional logistic regression
GLM links: logit, probit, identity, log, complementary log-log
O3. Y (categorical)
1. Multinomial logistic regression (no order Y, compare to base)
2. Ordinal logistic regression (order Y)
2b. proportional odds ordinal logistic regression
O4. Y (rate): # of cases per unit time= incidence rate
1. Poisson regression
-Stationarity: assume incidence rate is constant over time conditional on covariates
-Goodness of fit: Pearson chi square or deviance for moderate # covariate patterns
2. Zero-Inflated Poisson (excess zero counts)
3. Negative Binomial Regression (over dispersion)
O5. Y (time to event)
Non informative censoring
1. Kaplan Meier (Survival function)
-Nonparametric
-Confidence interval for $S(t_k)$ with Greenwood or complementary log-log transformation
-Log rank test for differences in survival distribution for binary or categorical covariate
-Other tests: Gehan-Wilcoxon, Peto-Prentice, Fleming-Harrington, Taron-Ware, etc
2. Exponential Survival Regression (hazard)
-Parametric
-Baseline hazard is assumed constant over time
3.Weibull Regression model (hazard)
-Parametric
-Baseline hazard varies as a function of time
4. Cox proportional hazard
-Semiparametric
-Ties: Exact, Efron, Breslow
-Don't need to specify the baseline hazard $h_0(t)$
-Assumes hazard ratio is constant over time
-Plot $\log[-\log(S_i(t))]$ for j=0,1: check parallel over time