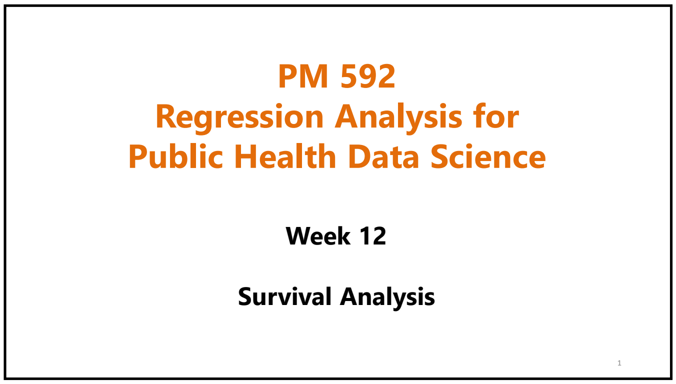
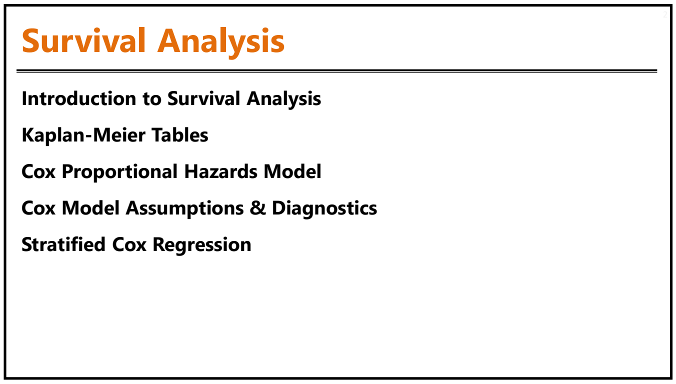


PM 592
Regression Analysis for Public Health Data Science

Week 12

Survival Analysis

1



Survival Analysis

Introduction to Survival Analysis

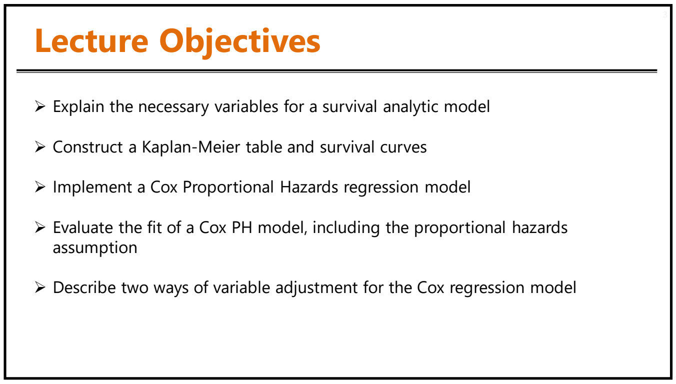
Kaplan-Meier Tables

Cox Proportional Hazards Model

Cox Model Assumptions & Diagnostics

Stratified Cox Regression

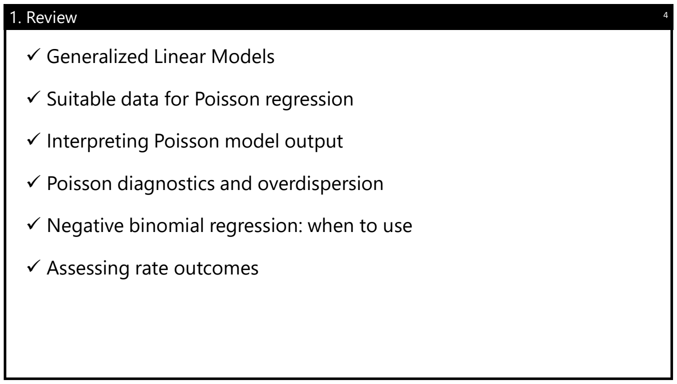
2



Lecture Objectives

- Explain the necessary variables for a survival analytic model
- Construct a Kaplan-Meier table and survival curves
- Implement a Cox Proportional Hazards regression model
- Evaluate the fit of a Cox PH model, including the proportional hazards assumption
- Describe two ways of variable adjustment for the Cox regression model

3

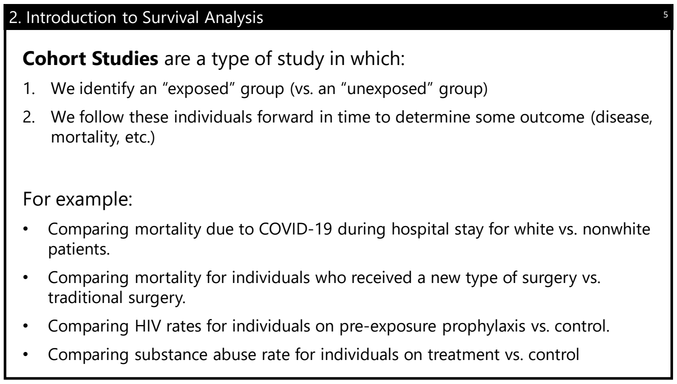


1. Review

4

- ✓ Generalized Linear Models
- ✓ Suitable data for Poisson regression
- ✓ Interpreting Poisson model output
- ✓ Poisson diagnostics and overdispersion
- ✓ Negative binomial regression: when to use
- ✓ Assessing rate outcomes

4



2. Introduction to Survival Analysis

5

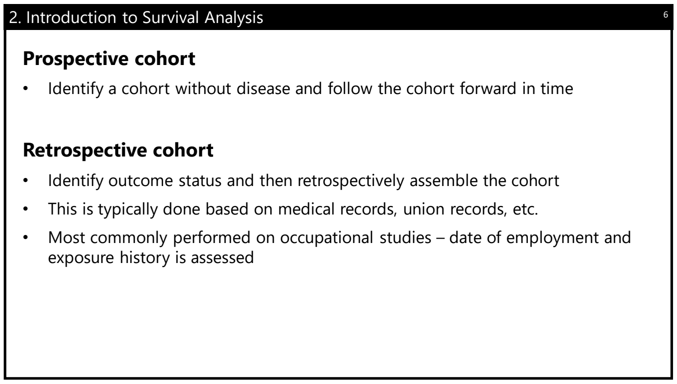
Cohort Studies are a type of study in which:

1. We identify an "exposed" group (vs. an "unexposed" group)
2. We follow these individuals forward in time to determine some outcome (disease, mortality, etc.)

For example:

- Comparing mortality due to COVID-19 during hospital stay for white vs. nonwhite patients.
- Comparing mortality for individuals who received a new type of surgery vs. traditional surgery.
- Comparing HIV rates for individuals on pre-exposure prophylaxis vs. control.
- Comparing substance abuse rate for individuals on treatment vs. control

5



2. Introduction to Survival Analysis

6

Prospective cohort

- Identify a cohort without disease and follow the cohort forward in time

Retrospective cohort

- Identify outcome status and then retrospectively assemble the cohort
- This is typically done based on medical records, union records, etc.
- Most commonly performed on occupational studies – date of employment and exposure history is assessed

6

2. Introduction to Survival Analysis

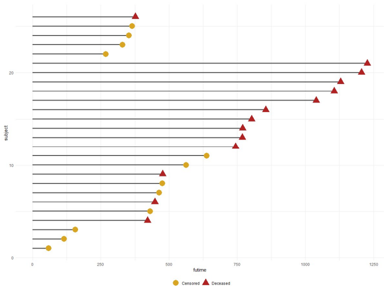
Example

Study	Inclusion	Entry Date	Outcome	Exposure	Unexposed
Japanese Atomic Bomb Survivors	All survivors within 2500 meters of epicenter and resident in city during 1950 census	Date of bomb	Cancer mortality, leukemia	Radiation level due to bomb	National mortality rates
Montana Smelter Workers (Historical)	Men employed >12 months prior to 12/31/56 in the smelter plant	Date at which 12 months of employment was completed, or 1/1/38	Cancer mortality		Montana male mortality rates
Framingham Heart Study	Men and women aged 30-62 in Framingham, MA, free of CVD	First recruitment	Heart disease, stroke, heart failure, and others	High blood pressure, high cholesterol, smoking, obesity, diabetes, physical activity, etc.	

2. Introduction to Survival Analysis

Example

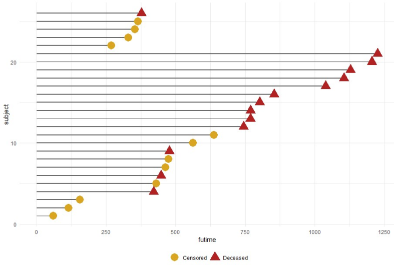
26 women were enrolled in a study comparing two different treatments for ovarian cancer.

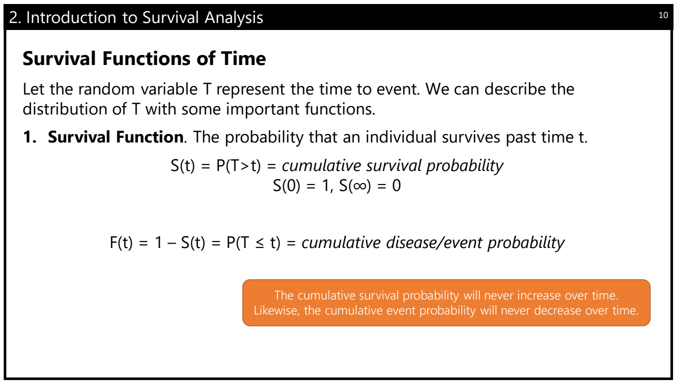


2. Introduction to Survival Analysis

Advantages

- Can estimate disease rates for exposed vs. unexposed.
- The observation time is taken into account (#diseased / #person-years at risk).
- Can define multiple outcomes of interest (e.g., mortality, other cancer type, other comorbidities).





2. Introduction to Survival Analysis 10

Survival Functions of Time

Let the random variable T represent the time to event. We can describe the distribution of T with some important functions.

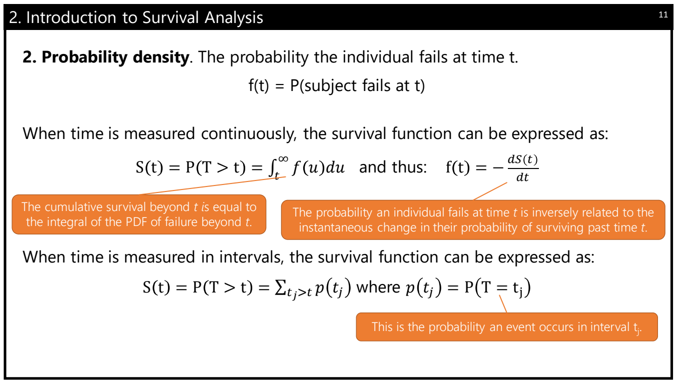
1. **Survival Function**. The probability that an individual survives past time t.

$$S(t) = P(T>t) = \textit{cumulative survival probability}$$
$$S(0) = 1, S(\infty) = 0$$

$$F(t) = 1 - S(t) = P(T \le t) = \textit{cumulative disease/event probability}$$

The cumulative survival probability will never increase over time. Likewise, the cumulative event probability will never decrease over time.

10



2. Introduction to Survival Analysis 11

2. Probability density. The probability the individual fails at time t.

$$f(t) = P(\text{subject fails at } t)$$

When time is measured continuously, the survival function can be expressed as:

$$S(t) = P(T > t) = \int_t^\infty f(u)du \quad \text{and thus:} \quad f(t) = -\frac{dS(t)}{dt}$$

The cumulative survival beyond *t is* equal to the integral of the PDF of failure beyond *t*.

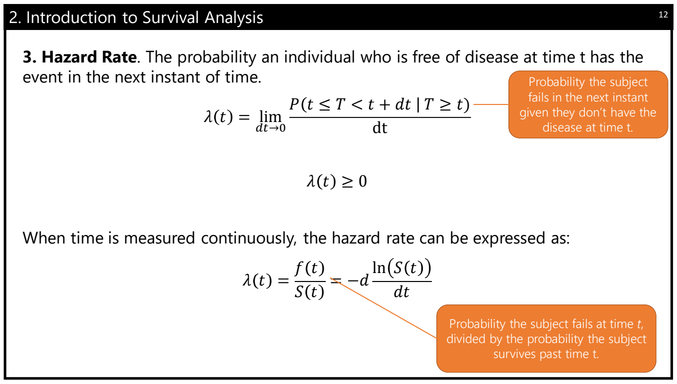
The probability an individual fails at time *t* is inversely related to the instantaneous change in their probability of surviving past time *t*.

When time is measured in intervals, the survival function can be expressed as:

$$S(t) = P(T > t) = \sum_{t_j > t} p(t_j) \text{ where } p(t_j) = P(T = t_j)$$

This is the probability an event occurs in interval t_j.

11



2. Introduction to Survival Analysis 12

3. Hazard Rate. The probability an individual who is free of disease at time t has the event in the next instant of time.

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \le T < t + dt \mid T \ge t)}{\text{dt}}$$

Probability the subject fails in the next instant given they don't have the disease at time t.

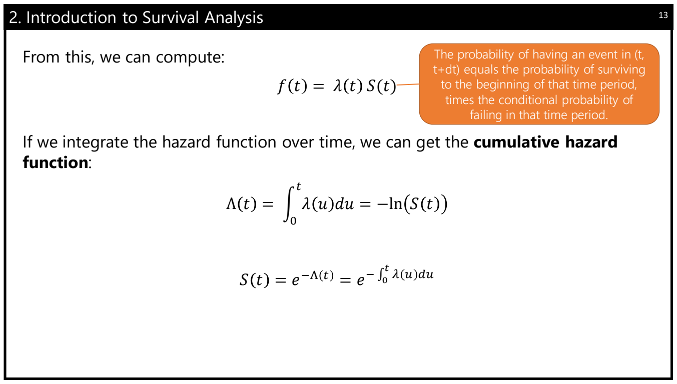
$$\lambda(t) \ge 0$$

When time is measured continuously, the hazard rate can be expressed as:

$$\lambda(t) = \frac{f(t)}{S(t)} = -d\frac{\ln(S(t))}{dt}$$

Probability the subject fails at time *t*, divided by the probability the subject survives past time t.

12



2. Introduction to Survival Analysis 13

From this, we can compute:

$$f(t) = \lambda(t)\, S(t)$$

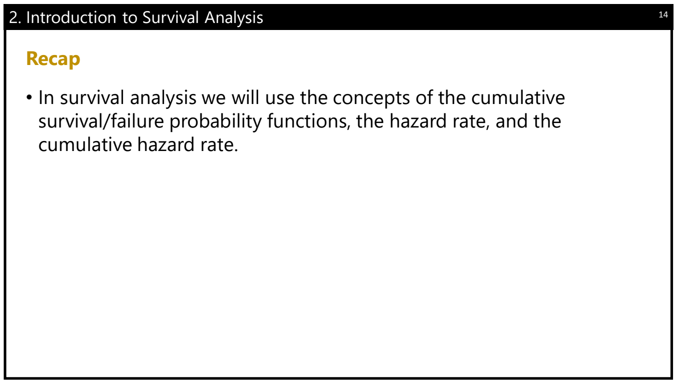
The probability of having an event in (t, t+dt) equals the probability of surviving to the beginning of that time period, times the conditional probability of failing in that time period.

If we integrate the hazard function over time, we can get the **cumulative hazard function**:

$$\Lambda(t) = \int_0^t \lambda(u)du = -\ln\big(S(t)\big)$$

$$S(t) = e^{-\Lambda(t)} = e^{-\int_0^t \lambda(u)du}$$

13

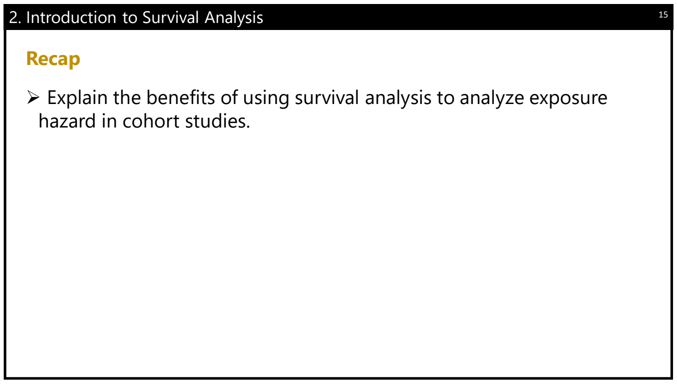


2. Introduction to Survival Analysis 14

Recap

- In survival analysis we will use the concepts of the cumulative survival/failure probability functions, the hazard rate, and the cumulative hazard rate.

14



2. Introduction to Survival Analysis 15

Recap

➢ Explain the benefits of using survival analysis to analyze exposure hazard in cohort studies.

15

3. Kaplan-Meier Tables

16

The simplest way of observing mortality over time is a **cohort** or **generation lifetable.**

- Uses mortality rates from a particular birth (or other) cohort.
- Observe the mortality of all persons from t_0 until all persons die (or are lost to follow-up)
- Answers epidemiologic questions regarding some outcome:
  - For acute disease, the case fatality rate can be a useful measure of survival.
  - For chronic disease, the case fatality rate is not a useful measure. A lifetable can provide specific information about the probability of surviving/dying within a specified time period after diagnosis.

16

3. Kaplan-Meier Tables

17

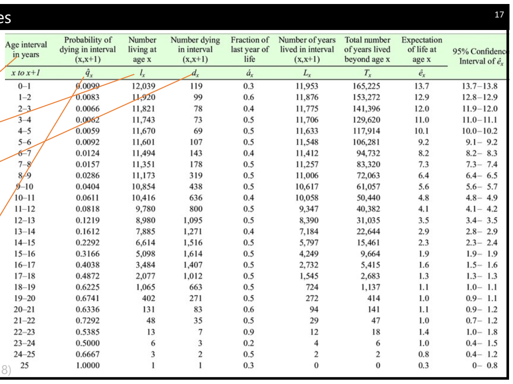
Partition time into a fixed sequence of intervals (usually 1 year).

l_i = # alive at beginning of interval *i*.

d_i = # died during interval *i*.

q_i = probability of dying in interval *i*, given subject is alive at the beginning of the interval.

u_i = # lost to follow-up during interval *i*
w_i = # withdrawn from study during interval *i*



Age interval in years	Probability of dying in interval (x,x+1)	Number living at age x	Number dying in interval (x,x+1)	Fraction of last year of life	Number of years lived in interval (x,x+1)	Total number of years lived beyond age x	Expectation of life at age x	95% Confidence Interval of $\hat{e}_x$
x to x+1	$\hat{q}_x$	l_x	d_x	$\acute{a}_x$	L_x	T_x	$\hat{e}_x$	
0–1	0.0099	12,039	119	0.3	11,953	165,225	13.7	13.7–13.8
1–2	0.0083	11,920	99	0.6	11,876	153,272	12.9	12.8–12.9
2–3	0.0066	11,821	78	0.4	11,775	141,396	12.0	11.9–12.0
3–4	0.0062	11,743	73	0.5	11,706	129,620	11.0	11.0–11.1
4–5	0.0059	11,670	69	0.5	11,633	117,914	10.1	10.0–10.2
5–6	0.0092	11,601	107	0.5	11,548	106,281	9.2	9.1– 9.2
6–7	0.0124	11,494	143	0.4	11,412	94,732	8.2	8.2– 8.3
7–8	0.0157	11,351	178	0.5	11,257	83,320	7.3	7.3– 7.4
8–9	0.0286	11,173	319	0.5	11,006	72,063	6.4	6.4– 6.5
9–10	0.0404	10,854	438	0.5	10,617	61,057	5.6	5.6– 5.7
10–11	0.0611	10,416	636	0.4	10,058	50,440	4.8	4.8– 4.9
11–12	0.0818	9,780	800	0.5	9,347	40,382	4.1	4.1– 4.2
12–13	0.1219	8,980	1,095	0.5	8,390	31,035	3.5	3.4– 3.5
13–14	0.1612	7,885	1,271	0.4	7,184	22,644	2.9	2.8– 2.9
14–15	0.2292	6,614	1,516	0.5	5,797	15,461	2.3	2.3– 2.4
15–16	0.3166	5,098	1,614	0.5	4,249	9,664	1.9	1.9– 1.9
16–17	0.4038	3,484	1,407	0.5	2,732	5,415	1.6	1.5– 1.6
17–18	0.4872	2,077	1,012	0.5	1,545	2,683	1.3	1.3– 1.3
18–19	0.6225	1,065	663	0.5	724	1,137	1.1	1.0– 1.1
19–20	0.6741	402	271	0.5	272	414	1.0	0.9– 1.1
20–21	0.6336	131	83	0.6	94	141	1.1	0.9– 1.2
21–22	0.7292	48	35	0.5	29	47	1.0	0.7– 1.2
22–23	0.5385	13	7	0.9	12	18	1.4	1.0– 1.8
23–24	0.5000	6	3	0.2	4	6	1.0	0.4– 1.5
24–25	0.6667	3	2	0.5	2	2	0.8	0.4– 1.2
25	1.0000	1	1	0.3	0	0	0.3	0– 0.8

Inoue, Kwan, & Sugiura (2018)

17

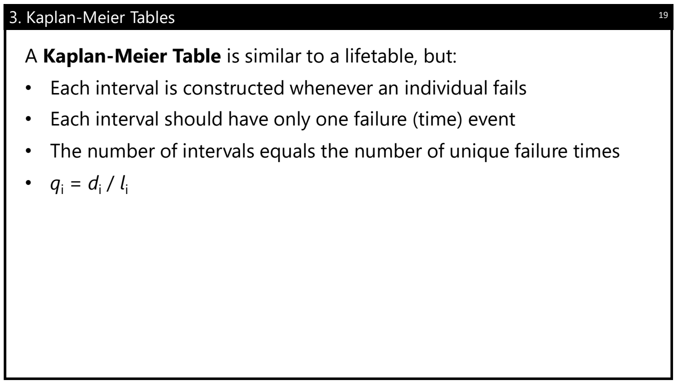
3. Kaplan-Meier Tables

18

The probability of surviving through each interval *i* is given as p_i = 1-q_i.

P_i is the cumulative probability of surviving to interval *i*, and is computed as $P_i = \prod_1^i p_i$

18

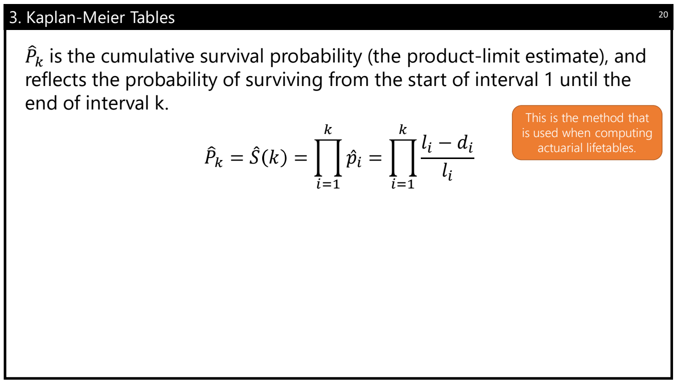


3. Kaplan-Meier Tables

A **Kaplan-Meier Table** is similar to a lifetable, but:

- Each interval is constructed whenever an individual fails
- Each interval should have only one failure (time) event
- The number of intervals equals the number of unique failure times
- $q_i = d_i / l_i$

19



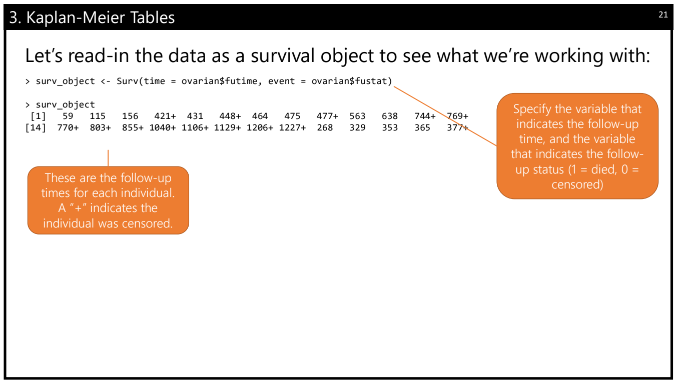
3. Kaplan-Meier Tables

$\hat{P}_k$ is the cumulative survival probability (the product-limit estimate), and reflects the probability of surviving from the start of interval 1 until the end of interval k.

$$\hat{P}_k = \hat{S}(k) = \prod_{i=1}^{k} \hat{p}_i = \prod_{i=1}^{k} \frac{l_i - d_i}{l_i}$$

This is the method that is used when computing actuarial lifetables.

20



3. Kaplan-Meier Tables

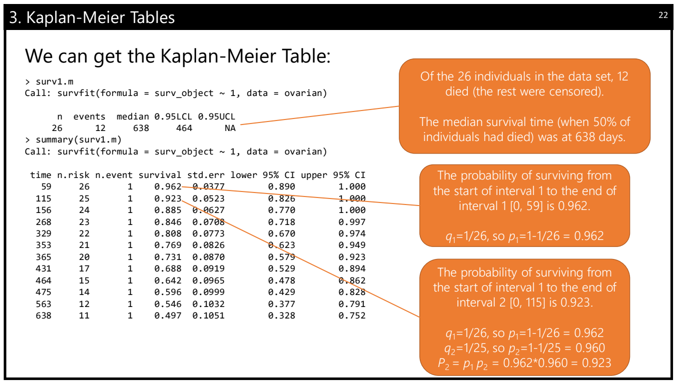
Let’s read-in the data as a survival object to see what we’re working with:

```
> surv_object <- Surv(time = ovarian$futime, event = ovarian$fustat)

> surv_object
 [1]   59   115   156   421+  431   448+  464   475   477+  563   638   744+  769+
[14]  770+  803+  855+ 1040+ 1106+ 1129+ 1206+ 1227+  268   329   353   365   377+
```

These are the follow-up times for each individual. A “+” indicates the individual was censored.

Specify the variable that indicates the follow-up time, and the variable that indicates the follow-up status (1 = died, 0 = censored)

21



3. Kaplan-Meier Tables

We can get the Kaplan-Meier Table:

```
> surv1.m
Call: survfit(formula = surv_object ~ 1, data = ovarian)

      n  events  median 0.95LCL 0.95UCL
     26      12     638     464      NA
> summary(surv1.m)
Call: survfit(formula = surv_object ~ 1, data = ovarian)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     26       1    0.962  0.0377        0.890        1.000
  115     25       1    0.923  0.0523        0.826        1.000
  156     24       1    0.885  0.0627        0.770        1.000
  268     23       1    0.846  0.0708        0.718        0.997
  329     22       1    0.808  0.0773        0.670        0.974
  353     21       1    0.769  0.0826        0.623        0.949
  365     20       1    0.731  0.0870        0.579        0.923
  431     17       1    0.688  0.0919        0.529        0.894
  464     15       1    0.642  0.0965        0.478        0.862
  475     14       1    0.596  0.0999        0.429        0.828
  563     12       1    0.546  0.1032        0.377        0.791
  638     11       1    0.497  0.1051        0.328        0.752
```

Of the 26 individuals in the data set, 12 died (the rest were censored).

The median survival time (when 50% of individuals had died) was at 638 days.

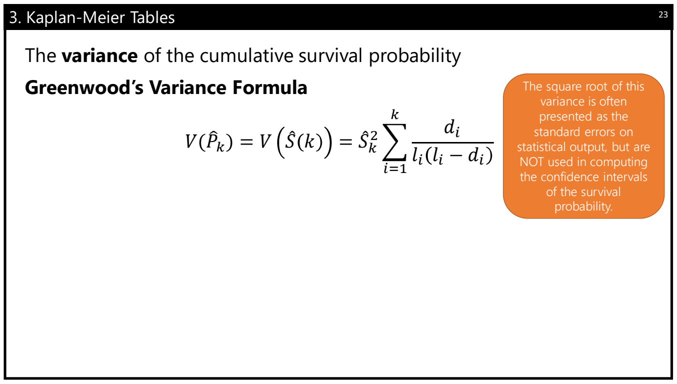
The probability of surviving from the start of interval 1 to the end of interval 1 [0, 59] is 0.962.

q_1=1/26, so p_1=1-1/26 = 0.962

The probability of surviving from the start of interval 1 to the end of interval 2 [0, 115] is 0.923.

q_1=1/26, so p_1=1-1/26 = 0.962
q_2=1/25, so p_2=1-1/25 = 0.960
$P_2 = p_1\, p_2$ = 0.962*0.960 = 0.923

22



3. Kaplan-Meier Tables

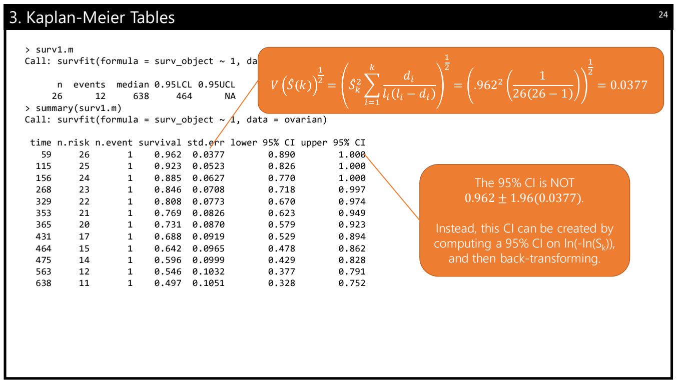
The **variance** of the cumulative survival probability

Greenwood's Variance Formula

$$V(\hat{P}_k) = V\left(\hat{S}(k)\right) = \hat{S}_k^2 \sum_{i=1}^{k} \frac{d_i}{l_i(l_i - d_i)}$$

The square root of this variance is often presented as the standard errors on statistical output, but are NOT used in computing the confidence intervals of the survival probability.

23



3. Kaplan-Meier Tables

```
> surv1.m
Call: survfit(formula = surv_object ~ 1, da

      n  events  median 0.95LCL 0.95UCL
     26      12     638     464      NA
> summary(surv1.m)
Call: survfit(formula = surv_object ~ 1, data = ovarian)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     26       1    0.962  0.0377        0.890        1.000
  115     25       1    0.923  0.0523        0.826        1.000
  156     24       1    0.885  0.0627        0.770        1.000
  268     23       1    0.846  0.0708        0.718        0.997
  329     22       1    0.808  0.0773        0.670        0.974
  353     21       1    0.769  0.0826        0.623        0.949
  365     20       1    0.731  0.0870        0.579        0.923
  431     17       1    0.688  0.0919        0.529        0.894
  464     15       1    0.642  0.0965        0.478        0.862
  475     14       1    0.596  0.0999        0.429        0.828
  563     12       1    0.546  0.1032        0.377        0.791
  638     11       1    0.497  0.1051        0.328        0.752
```

$$V\left(\hat{S}(k)\right)^{\frac{1}{2}} = \left(\hat{S}_k^2 \sum_{i=1}^{k} \frac{d_i}{l_i(l_i - d_i)}\right)^{\frac{1}{2}} = \left(.962^2\left(\frac{1}{26(26-1)}\right)\right)^{\frac{1}{2}} = 0.0377$$

The 95% CI is NOT
$0.962 \pm 1.96(0.0377)$.

Instead, this CI can be created by computing a 95% CI on ln(-ln(S_k)), and then back-transforming.

24

3. Kaplan-Meier Tables 25

Adding A Predictor

We can compute the life tables separately for individuals on treatment 1 (rx==1) vs. treatment 2 (rx==2)

```
> surv2.m <- survfit(surv_object ~ rx, data = ovarian)
> summary(surv2.m)
Call: survfit(formula = surv_object ~ rx, data = ovarian)

                rx=1 
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   59     13       1    0.923  0.0739        0.789        1.000
  115     12       1    0.846  0.1001        0.671        1.000
  156     11       1    0.769  0.1169        0.571        1.000
  268     10       1    0.692  0.1280        0.482        0.995
  329      9       1    0.615  0.1349        0.400        0.946
  431      8       1    0.538  0.1383        0.326        0.891
  638      5       1    0.431  0.1467        0.221        0.840

                rx=2 
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  353     13       1    0.923  0.0739        0.789        1.000
  365     12       1    0.846  0.1001        0.671        1.000
  464      9       1    0.752  0.1256        0.542        1.000
  475      8       1    0.658  0.1407        0.433        1.000
  563      7       1    0.564  0.1488        0.336        0.946
```

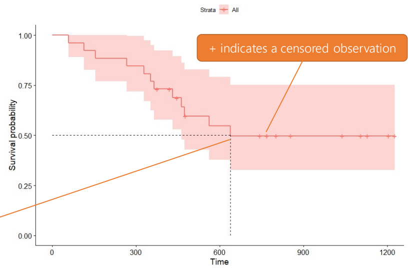
25

3. Kaplan-Meier Tables 26

Visualizing the Curves

It is easier to examine these curves visually

```
ggsurvplot(surv1.m, data = ovarian,
surv.median.line = "hv")
```



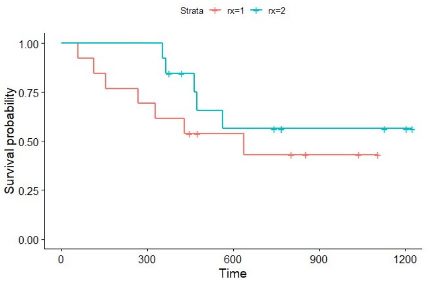
+ indicates a censored observation

Median survival is where the Kaplan-Meier curve crosses 50% survival probability.

26

3. Kaplan-Meier Tables 27

This figure shows us that the survival under Treatment 2 appears to be superior compared to survival under Treatment 1.



27

Comparing the Equality of Survival Curves

Method 1: Log-Rank Test

- $H_0: \lambda_1(t) = \lambda_2(t) = \lambda_3(t) =$ … for groups 1, 2, 3, etc.
- Each failure time contributes to the test statistic
- Treats the hazard function as proportional across groups over follow-up time

$$\chi^2_{J-1} = \sum_j \frac{\left(D_j - E_j\right)^2}{E_j}$$

J = # groups

D_j= Total observed failures in group *j* (summed over all failure times)

E_j = Expected failures in group *j* (summed over all failure times)

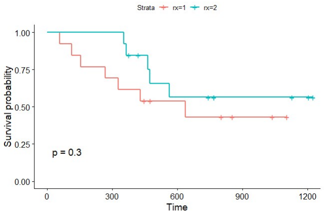
Under H_0, the expected failures in group *j* at time *t* = $e_{jt} = \frac{l_{jt} d_j}{l_j}$

28

Log-Rank Test

There is no statistically significant difference in survival curves between treatment groups (p=.30).

```
> survdiff(surv_object ~ rx, data = ovarian)
Call:
survdiff(formula = surv_object ~ rx, data = ovarian)

      N Observed Expected (O-E)^2/E (O-E)^2/V
rx=1 13        7     5.23     0.596      1.06
rx=2 13        5     6.77     0.461      1.06

 Chisq= 1.1  on 1 degrees of freedom, p= 0.3

> ggsurvplot(surv2.m, data = ovarian, pval = T)
```



29

Comparing the Equality of Survival Curves

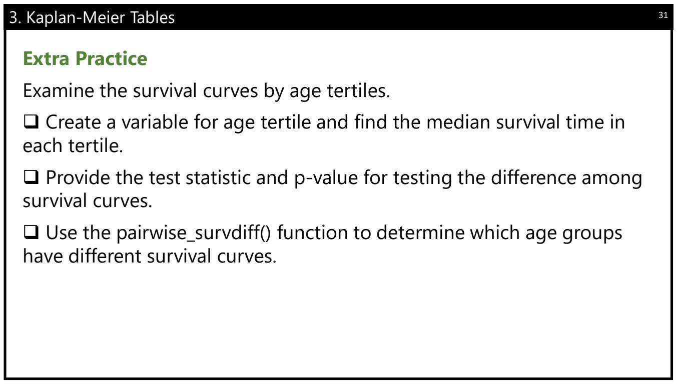
Method 2: Wilcoxon Test

- Use when hazards between groups may not differ proportionally across follow-up
- This method weights each failure time contribution by the number of subjects at-risk at that time, giving greater weight to earlier events when more subjects are at risk.
- This test is sensitive to different censoring patterns among groups.

Other tests can be implemented using differing values of rho in the survdiff() function as follows:

1	log-rank	
n[i]	Gehan-Breslow generalized Wilcoxon	
sqrt(n[i])	Tarone-Ware	
S1[i]	Peto-Peto's modified survival estimate	*S1(t) = cumprod(1 - e / (n + 1))*
S2[i]	modified Peto-Peto (by Andersen)	*S2(t) = S1[i] * n[i] / (n[i] + 1)*
FH[i]	Fleming-Harrington	The weight at *t_0 = 1* and thereafter is: *S(t[i - 1])^p * (1 - S(t)[i - 1]^q)*

30



3. Kaplan-Meier Tables 31

Extra Practice

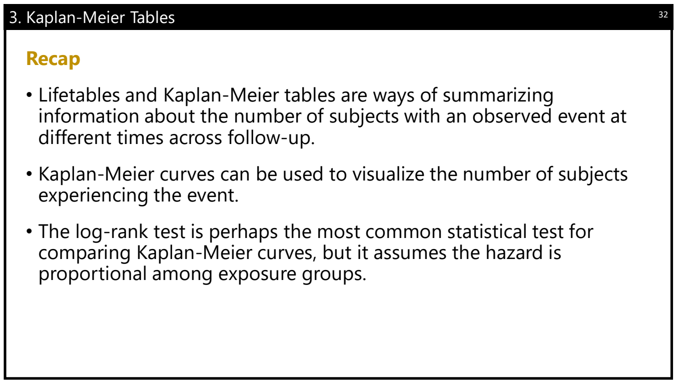
Examine the survival curves by age tertiles.

❑ Create a variable for age tertile and find the median survival time in each tertile.

❑ Provide the test statistic and p-value for testing the difference among survival curves.

❑ Use the pairwise_survdiff() function to determine which age groups have different survival curves.

31

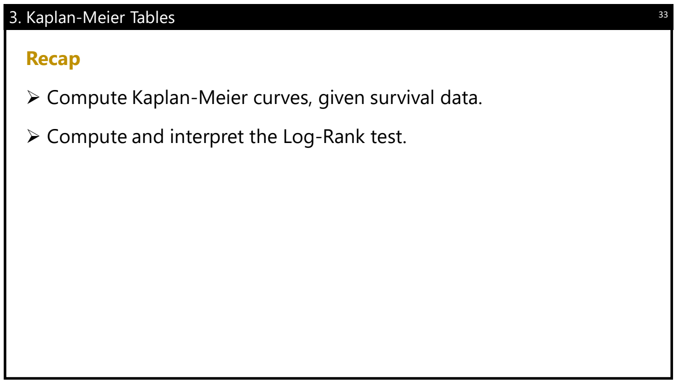


3. Kaplan-Meier Tables 32

Recap

- Lifetables and Kaplan-Meier tables are ways of summarizing information about the number of subjects with an observed event at different times across follow-up.
- Kaplan-Meier curves can be used to visualize the number of subjects experiencing the event.
- The log-rank test is perhaps the most common statistical test for comparing Kaplan-Meier curves, but it assumes the hazard is proportional among exposure groups.

32



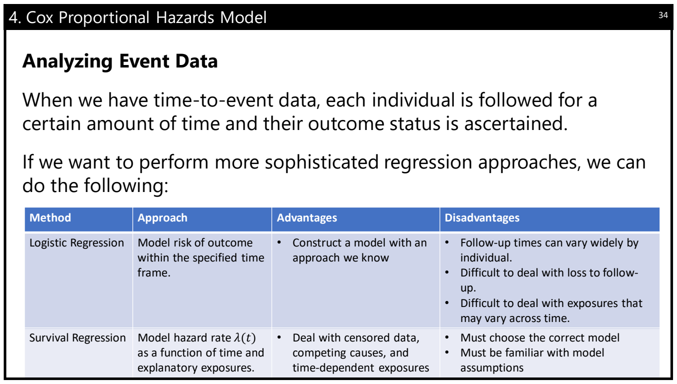
3. Kaplan-Meier Tables 33

Recap

➢ Compute Kaplan-Meier curves, given survival data.

➢ Compute and interpret the Log-Rank test.

33



4. Cox Proportional Hazards Model

34

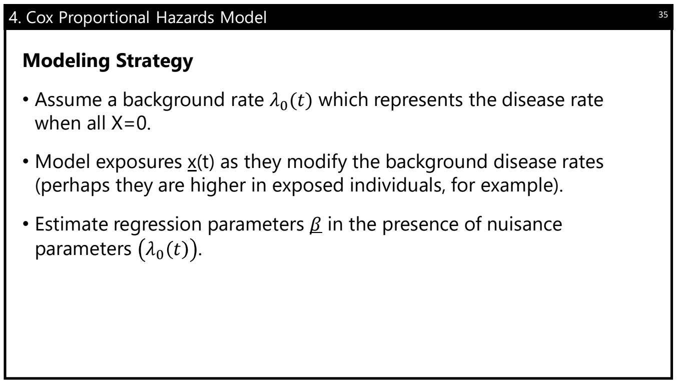
Analyzing Event Data

When we have time-to-event data, each individual is followed for a certain amount of time and their outcome status is ascertained.

If we want to perform more sophisticated regression approaches, we can do the following:

Method	Approach	Advantages	Disadvantages
Logistic Regression	Model risk of outcome within the specified time frame.	• Construct a model with an approach we know	• Follow-up times can vary widely by individual. • Difficult to deal with loss to follow-up. • Difficult to deal with exposures that may vary across time.
Survival Regression	Model hazard rate $\lambda(t)$ as a function of time and explanatory exposures.	• Deal with censored data, competing causes, and time-dependent exposures	• Must choose the correct model • Must be familiar with model assumptions

34



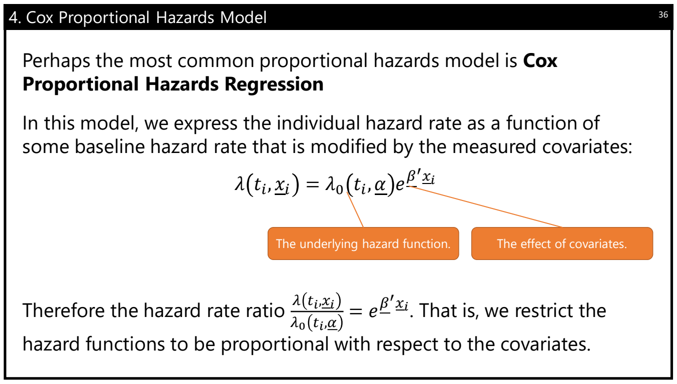
4. Cox Proportional Hazards Model

35

Modeling Strategy

- Assume a background rate $\lambda_0(t)$ which represents the disease rate when all X=0.
- Model exposures $\underline{x}$(t) as they modify the background disease rates (perhaps they are higher in exposed individuals, for example).
- Estimate regression parameters $\underline{\beta}$ in the presence of nuisance parameters $\left(\lambda_0(t)\right)$.

35



4. Cox Proportional Hazards Model

36

Perhaps the most common proportional hazards model is **Cox Proportional Hazards Regression**

In this model, we express the individual hazard rate as a function of some baseline hazard rate that is modified by the measured covariates:

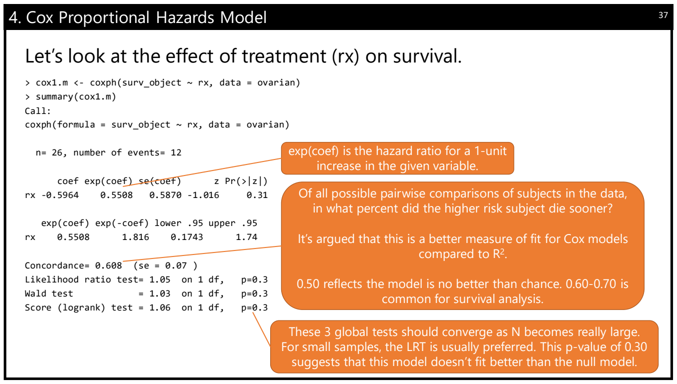
$$\lambda(t_i,\underline{x}_i) = \lambda_0\left(t_i,\underline{\alpha}\right)e^{\underline{\beta}'\underline{x}_i}$$

The underlying hazard function.

The effect of covariates.

Therefore the hazard rate ratio $\frac{\lambda(t_i,\underline{x}_i)}{\lambda_0(t_i,\underline{\alpha})} = e^{\underline{\beta}'\underline{x}_i}$. That is, we restrict the hazard functions to be proportional with respect to the covariates.

36



4. Cox Proportional Hazards Model

Let’s look at the effect of treatment (rx) on survival.

```
> cox1.m <- coxph(surv_object ~ rx, data = ovarian)
> summary(cox1.m)
Call:
coxph(formula = surv_object ~ rx, data = ovarian)

  n= 26, number of events= 12

      coef exp(coef) se(coef)      z Pr(>|z|)
rx -0.5964    0.5508   0.5870 -1.016     0.31

   exp(coef) exp(-coef) lower .95 upper .95
rx    0.5508      1.816    0.1743      1.74

Concordance= 0.608  (se = 0.07 )
Likelihood ratio test= 1.05  on 1 df,   p=0.3
Wald test            = 1.03  on 1 df,   p=0.3
Score (logrank) test = 1.06  on 1 df,   p=0.3
```

exp(coef) is the hazard ratio for a 1-unit increase in the given variable.

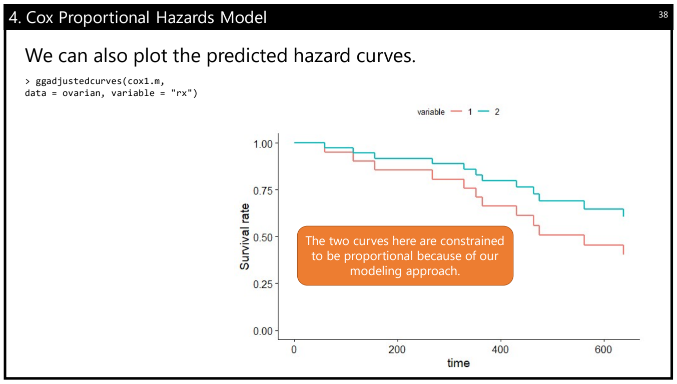
Of all possible pairwise comparisons of subjects in the data, in what percent did the higher risk subject die sooner?

It’s argued that this is a better measure of fit for Cox models compared to R^2.

0.50 reflects the model is no better than chance. 0.60-0.70 is common for survival analysis.

These 3 global tests should converge as N becomes really large. For small samples, the LRT is usually preferred. This p-value of 0.30 suggests that this model doesn’t fit better than the null model.

37



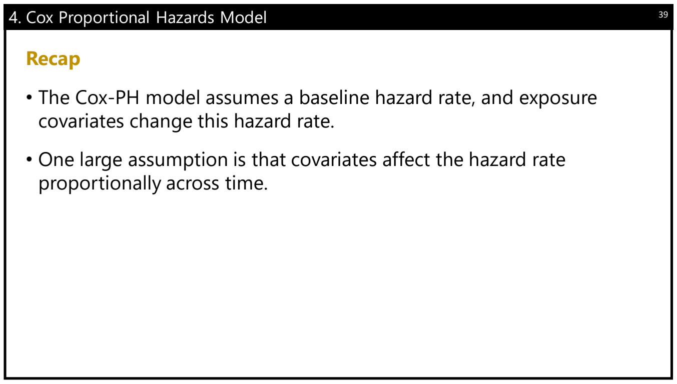
4. Cox Proportional Hazards Model

We can also plot the predicted hazard curves.

```
> ggadjustedcurves(cox1.m,
data = ovarian, variable = "rx")
```

The two curves here are constrained to be proportional because of our modeling approach.

38

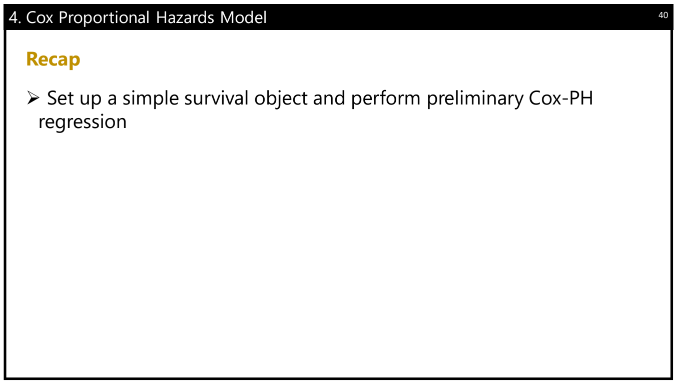


4. Cox Proportional Hazards Model

Recap

- The Cox-PH model assumes a baseline hazard rate, and exposure covariates change this hazard rate.
- One large assumption is that covariates affect the hazard rate proportionally across time.

39

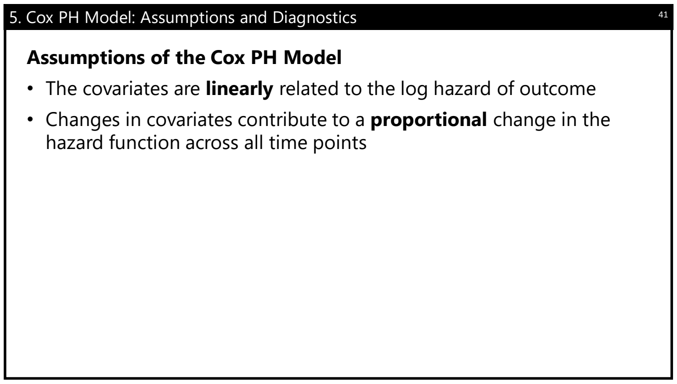


4. Cox Proportional Hazards Model 40

Recap

➢ Set up a simple survival object and perform preliminary Cox-PH regression

40

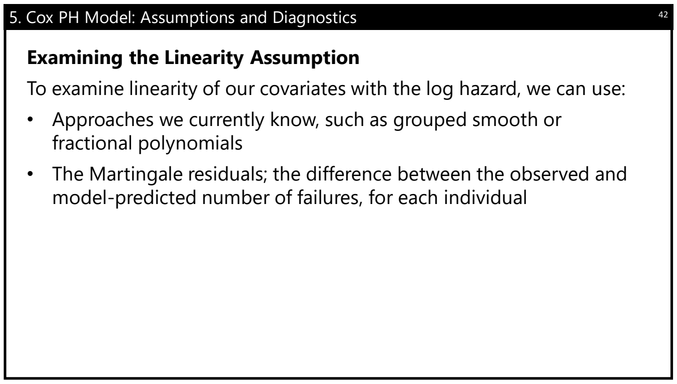


5. Cox PH Model: Assumptions and Diagnostics 41

Assumptions of the Cox PH Model

- The covariates are **linearly** related to the log hazard of outcome
- Changes in covariates contribute to a **proportional** change in the hazard function across all time points

41



5. Cox PH Model: Assumptions and Diagnostics 42

Examining the Linearity Assumption

To examine linearity of our covariates with the log hazard, we can use:

- Approaches we currently know, such as grouped smooth or fractional polynomials
- The Martingale residuals; the difference between the observed and model-predicted number of failures, for each individual

42

5. Cox PH Model: Assumptions and Diagnostics 43

Fractional Polynomials

Let's see if treatment is a significant predictor of hazard rate after adjusting for baseline age.

```
> mfp(Surv(futime, fustat) ~ fp(age) + rx, family = cox, data = ovarian)
Call: mfp(formula = Surv(futime, fustat) ~ fp(age) + rx, data = ovarian, family = cox)

Deviance table:
                       Resid. Dev
Null model  69.96988
Linear model           54.0838
Final model            54.0838

Fractional polynomials:
    df.initial select alpha df.final power1 power2
age          4      1  0.05        1      1      .
rx           1      1  0.05        1      1      .


Transformations of covariates:
           formula
age I((age/100)^1)
rx              rx

         coef exp(coef) se(coef)      z       p
age.1  0.1473    1.1587  0.04615  3.193 0.00141
rx.1  -0.8040    0.4475  0.63205 -1.272 0.20300

Likelihood ratio test=15.89  on 2 df, p=0.0003551 n= 26
```

Rx is a binary variable, so we don't need to consider its functional form.

When we examine age, MFP shows us the linear coding is sufficient.

43

5. Cox PH Model: Assumptions and Diagnostics 44

Martingale Residuals

The predicted number of failures for subject *i* at the end of follow-up time T_i is computed from the fitted model as:

$$\hat{\Lambda}(T_i) = \hat{\Lambda}_0(T_i)e^{\underline{\beta}'\underline{x}}$$

The Martingale residual should be linearly related to f(x), the optimal transformation of x.

The Martingale residual is calculated as:

$$r_{m_i} = \delta_i - \hat{\Lambda}(T_i)$$, where δ_i is the subject's failure status

Martingale residuals have a mean 0 with range $-\infty, 1$

44

5. Cox PH Model: Assumptions and Diagnostics 45

Martingale Residuals

- These aren't the usual type of residuals; there's no clear analogy between these residuals and those in linear regression.
- Think of these residuals generally as being some measure of difference in observed vs. predicted values.
- "Observed Y" doesn't make sense in survival time data, as survival is defined as presence of the event (Y=1 or Y=0) *and* the time at which the event occurred.
- A value of "1" indicates the person had the event but had a very small cumulative hazard.
- A large negative value indicates the person was censored (survived) but had a very large cumulative hazard.

45

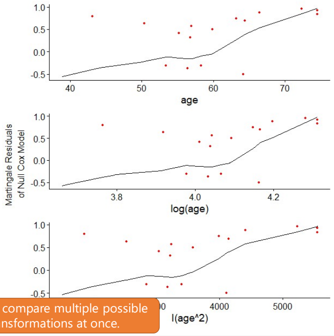
5. Cox PH Model: Assumptions and Diagnostics 46

Approach

- Run the Cox model with no covariates (null model)
- Generate Martingale residuals
- Plot the residuals against the values of the covariates under consideration
- Inspect the LOWESS line for a linear relationship

The following function does all this for us!

```
> ggcoxfunctional(surv_object ~ age + log(age) + I(age^2),
data = ovarian)
```



We can compare multiple possible transformations at once.

46

5. Cox PH Model: Assumptions and Diagnostics 47

Examining the Proportional Hazards Assumption

To test the PH assumption, we can use the scaled Schoenfeld residuals.

The Schoenfeld residual for variable *x* in subject *i* is calculated as:

$$r_s = D_i(x_i - \alpha_i),$$ where $D_i = 1$ when there is an event.

The Schoenfeld residual is the difference between the observed x_i for the subject who had the event and the weighted average of all x_i for all subjects in the risk set when the subject had the event.

$$\alpha_i = \frac{\sum_{l \epsilon R_i} x_l exp(\beta x_l)}{\sum_{l \epsilon R_i} exp(\beta x_l)}$$

The scaled Schoenfeld residual is calculated as:

$$\hat{\beta} + dVar(\hat{\beta}) r_s,$$ where d =total number of failure events

47

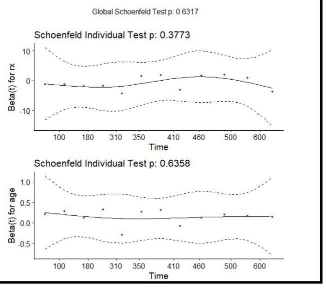
5. Cox PH Model: Assumptions and Diagnostics 48

Examining the Proportional Hazards Assumption

To test the PH assumption, we can use the scaled Schoenfeld residuals.

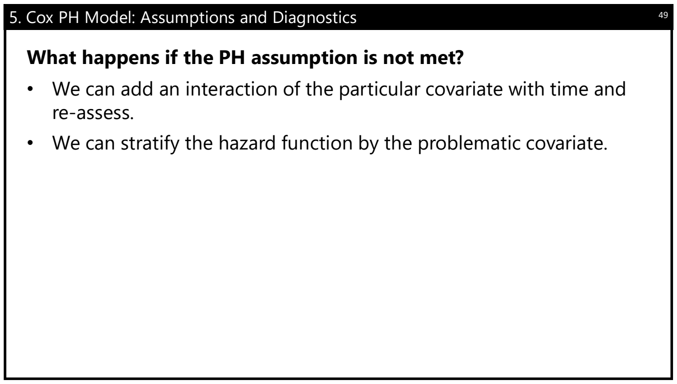
```
> cox.zph(cox2.m)
       chisq df    p
rx     0.780  1 0.38
age    0.224  1 0.64
GLOBAL 0.919  2 0.63

> ggcoxzph(cox.zph(cox2.m))
```

There is no violation of the proportional hazards assumption for treatment (p=.38).



These lines should be horizontal across time if the PH assumption is met.

48

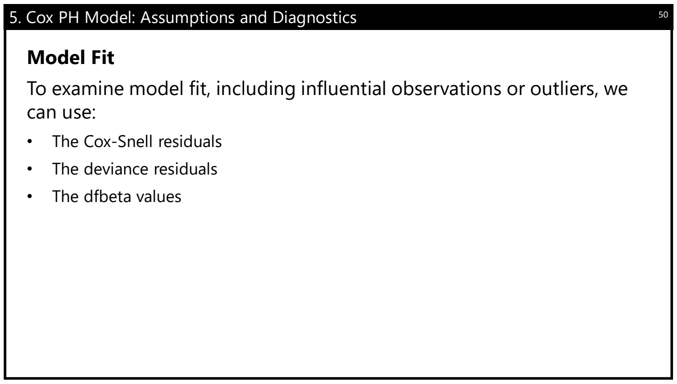


5. Cox PH Model: Assumptions and Diagnostics 49

What happens if the PH assumption is not met?

- We can add an interaction of the particular covariate with time and re-assess.
- We can stratify the hazard function by the problematic covariate.

49



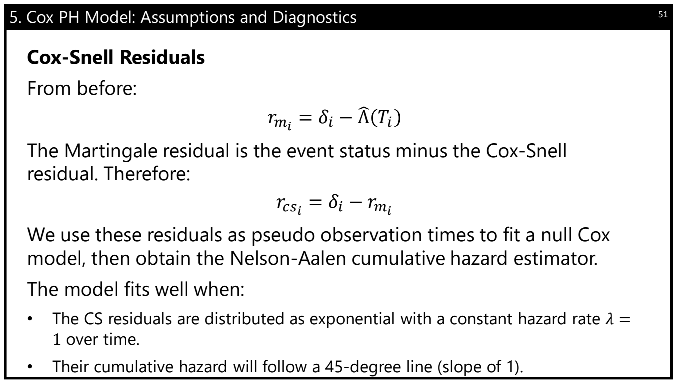
5. Cox PH Model: Assumptions and Diagnostics 50

Model Fit

To examine model fit, including influential observations or outliers, we can use:

- The Cox-Snell residuals
- The deviance residuals
- The dfbeta values

50



5. Cox PH Model: Assumptions and Diagnostics 51

Cox-Snell Residuals

From before:

$$r_{m_i} = \delta_i - \hat{\Lambda}(T_i)$$

The Martingale residual is the event status minus the Cox-Snell residual. Therefore:

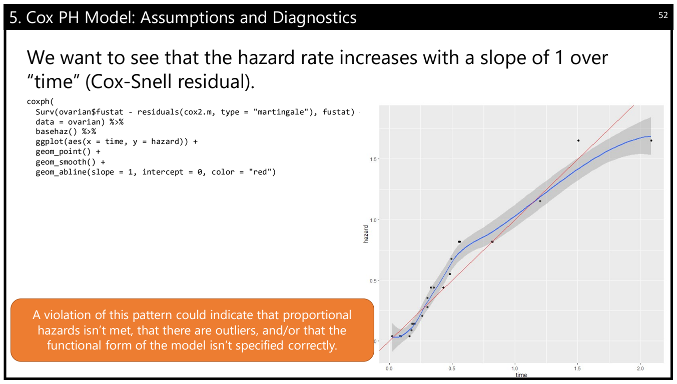
$$r_{cs_i} = \delta_i - r_{m_i}$$

We use these residuals as pseudo observation times to fit a null Cox model, then obtain the Nelson-Aalen cumulative hazard estimator.

The model fits well when:

- The CS residuals are distributed as exponential with a constant hazard rate $\lambda = 1$ over time.
- Their cumulative hazard will follow a 45-degree line (slope of 1).

51



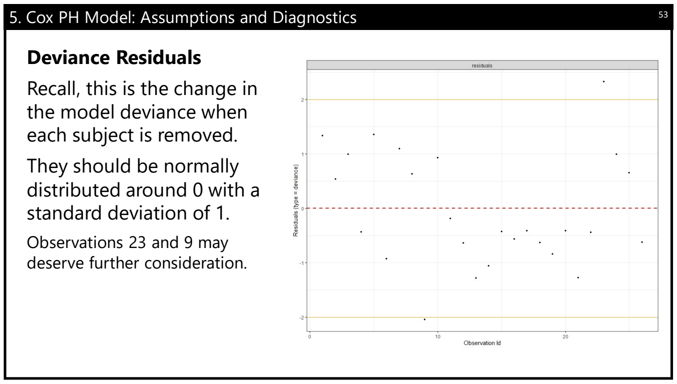
5. Cox PH Model: Assumptions and Diagnostics

We want to see that the hazard rate increases with a slope of 1 over "time" (Cox-Snell residual).

```
coxph(
  Surv(ovarian$fustat - residuals(cox2.m, type = "martingale"), fustat)
  data = ovarian) %>%
  basehaz() %>%
  ggplot(aes(x = time, y = hazard)) +
  geom_point() +
  geom_smooth() +
  geom_abline(slope = 1, intercept = 0, color = "red")
```

A violation of this pattern could indicate that proportional hazards isn't met, that there are outliers, and/or that the functional form of the model isn't specified correctly.

52



5. Cox PH Model: Assumptions and Diagnostics

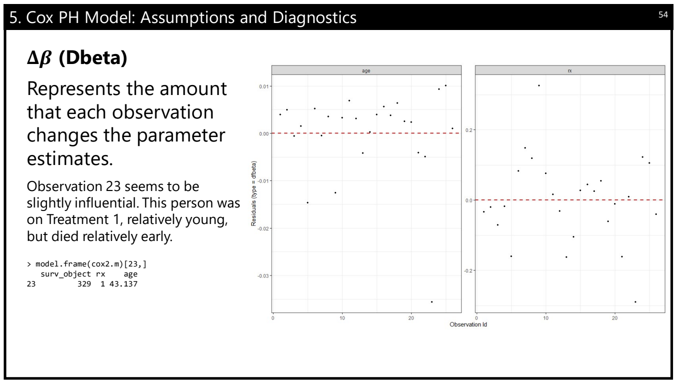
Deviance Residuals

Recall, this is the change in the model deviance when each subject is removed.

They should be normally distributed around 0 with a standard deviation of 1.

Observations 23 and 9 may deserve further consideration.

53



5. Cox PH Model: Assumptions and Diagnostics

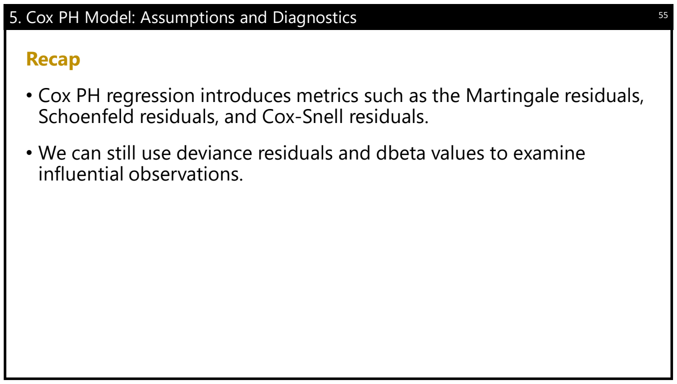
$\Delta\beta$ (Dbeta)

Represents the amount that each observation changes the parameter estimates.

Observation 23 seems to be slightly influential. This person was on Treatment 1, relatively young, but died relatively early.

```
> model.frame(cox2.m)[23,]
   surv_object rx    age
23         329  1 43.137
```

54

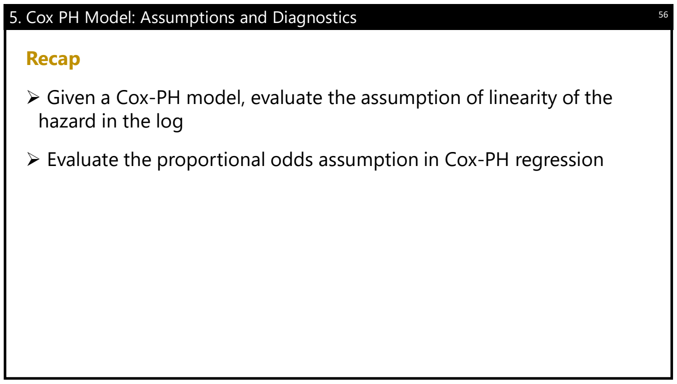


5. Cox PH Model: Assumptions and Diagnostics 55

Recap

- Cox PH regression introduces metrics such as the Martingale residuals, Schoenfeld residuals, and Cox-Snell residuals.
- We can still use deviance residuals and dbeta values to examine influential observations.

55

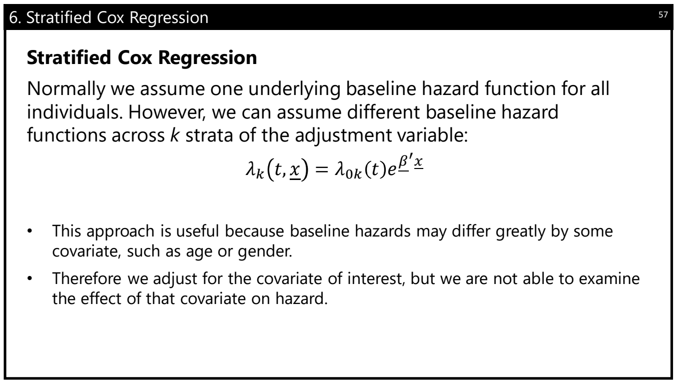


5. Cox PH Model: Assumptions and Diagnostics 56

Recap

- ➢ Given a Cox-PH model, evaluate the assumption of linearity of the hazard in the log
- ➢ Evaluate the proportional odds assumption in Cox-PH regression

56



6. Stratified Cox Regression 57

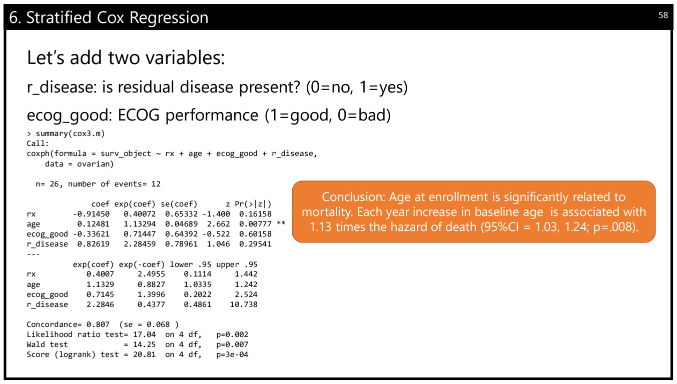
Stratified Cox Regression

Normally we assume one underlying baseline hazard function for all individuals. However, we can assume different baseline hazard functions across *k* strata of the adjustment variable:

$$\lambda_k(t,\underline{x}) = \lambda_{0k}(t)e^{\underline{\beta}'\underline{x}}$$

- This approach is useful because baseline hazards may differ greatly by some covariate, such as age or gender.
- Therefore we adjust for the covariate of interest, but we are not able to examine the effect of that covariate on hazard.

57



6. Stratified Cox Regression

Let’s add two variables:

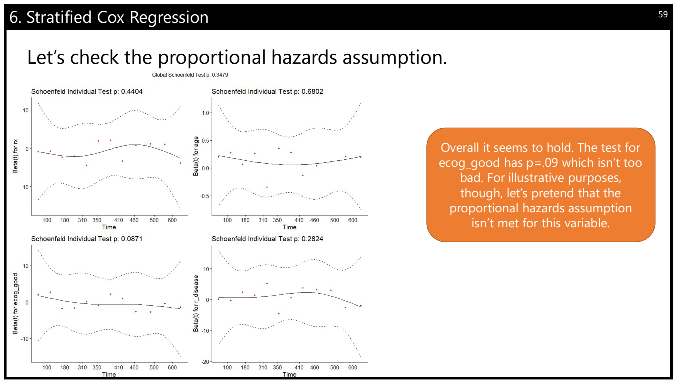
r_disease: is residual disease present? (0=no, 1=yes)

ecog_good: ECOG performance (1=good, 0=bad)

```
> summary(cox3.m)
Call:
coxph(formula = surv_object ~ rx + age + ecog_good + r_disease,
    data = ovarian)

  n= 26, number of events= 12

              coef exp(coef) se(coef)      z Pr(>|z|)
rx        -0.91450   0.40072  0.65332 -1.400  0.16158
age        0.12481   1.13294  0.04689  2.662  0.00777 **
ecog_good -0.33621   0.71447  0.64392 -0.522  0.60158
r_disease  0.82619   2.28459  0.78961  1.046  0.29541
---
          exp(coef) exp(-coef) lower .95 upper .95
rx           0.4007     2.4955    0.1114     1.442
age          1.1329     0.8827    1.0335     1.242
ecog_good    0.7145     1.3996    0.2022     2.524
r_disease    2.2846     0.4377    0.4861    10.738

Concordance= 0.807  (se = 0.068 )
Likelihood ratio test= 17.04  on 4 df,   p=0.002
Wald test            = 14.25  on 4 df,   p=0.007
Score (logrank) test = 20.81  on 4 df,   p=3e-04
```

Conclusion: Age at enrollment is significantly related to mortality. Each year increase in baseline age is associated with 1.13 times the hazard of death (95%CI = 1.03, 1.24; p=.008).

58

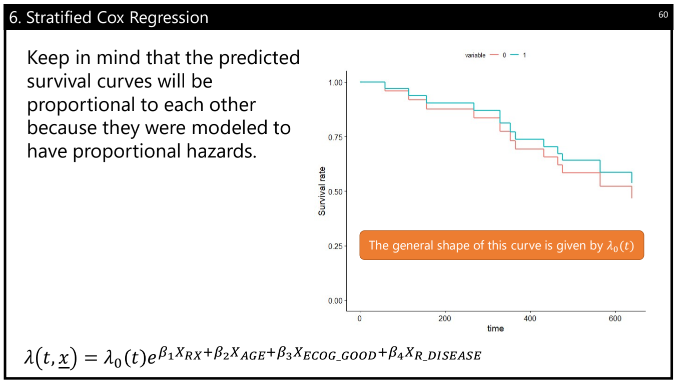


6. Stratified Cox Regression

Let’s check the proportional hazards assumption.

Overall it seems to hold. The test for ecog_good has p=.09 which isn’t too bad. For illustrative purposes, though, let’s pretend that the proportional hazards assumption isn’t met for this variable.

59



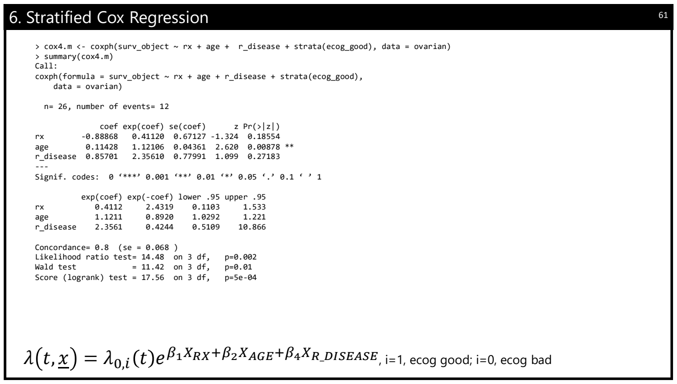
6. Stratified Cox Regression

Keep in mind that the predicted survival curves will be proportional to each other because they were modeled to have proportional hazards.

The general shape of this curve is given by $\lambda_0(t)$

$$\lambda(t,\underline{x}) = \lambda_0(t)e^{\beta_1 X_{RX}+\beta_2 X_{AGE}+\beta_3 X_{ECOG\_GOOD}+\beta_4 X_{R\_DISEASE}}$$

60

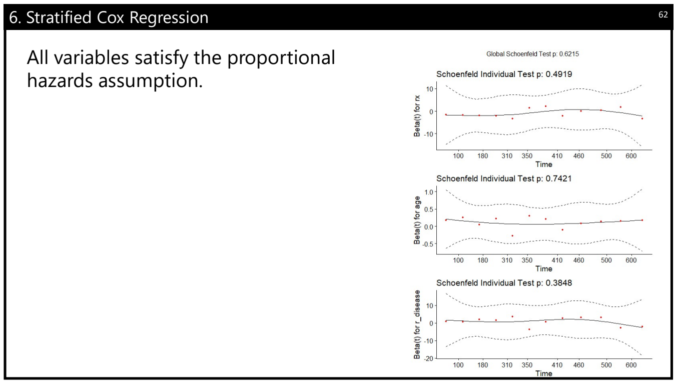


6. Stratified Cox Regression

```
> cox4.m <- coxph(surv_object ~ rx + age +  r_disease + strata(ecog_good), data = ovarian)
> summary(cox4.m)
Call:
coxph(formula = surv_object ~ rx + age + r_disease + strata(ecog_good),
    data = ovarian)

  n= 26, number of events= 12

              coef exp(coef) se(coef)      z Pr(>|z|)
rx        -0.88868   0.41120  0.67127 -1.324  0.18554
age        0.11428   1.12106  0.04361  2.620  0.00878 **
r_disease  0.85701   2.35610  0.77991  1.099  0.27183
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

          exp(coef) exp(-coef) lower .95 upper .95
rx           0.4112     2.4319    0.1103     1.533
age          1.1211     0.8920    1.0292     1.221
r_disease    2.3561     0.4244    0.5109    10.866

Concordance= 0.8  (se = 0.068 )
Likelihood ratio test= 14.48  on 3 df,   p=0.002
Wald test            = 11.42  on 3 df,   p=0.01
Score (logrank) test = 17.56  on 3 df,   p=5e-04
```

$\lambda(t,\underline{x}) = \lambda_{0,i}(t)e^{\beta_1 X_{RX}+\beta_2 X_{AGE}+\beta_4 X_{R\_DISEASE}}$, i=1, ecog good; i=0, ecog bad

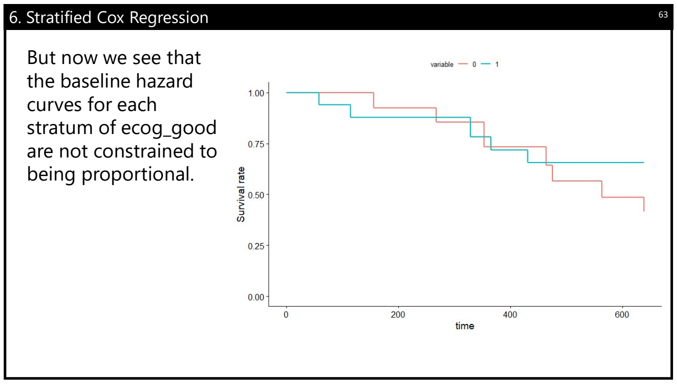
61



6. Stratified Cox Regression

All variables satisfy the proportional hazards assumption.

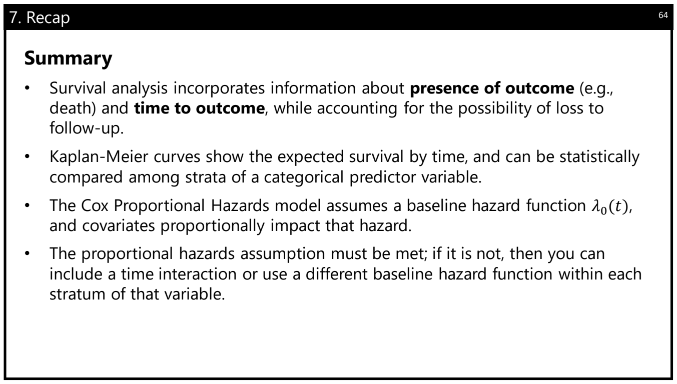
62



6. Stratified Cox Regression

But now we see that the baseline hazard curves for each stratum of ecog_good are not constrained to being proportional.

63

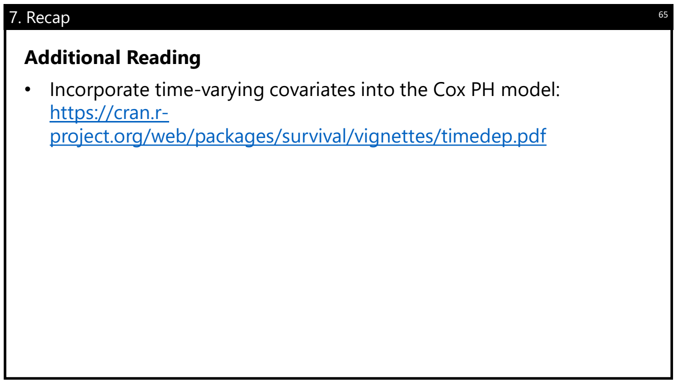


7. Recap 64

Summary

- Survival analysis incorporates information about **presence of outcome** (e.g., death) and **time to outcome**, while accounting for the possibility of loss to follow-up.
- Kaplan-Meier curves show the expected survival by time, and can be statistically compared among strata of a categorical predictor variable.
- The Cox Proportional Hazards model assumes a baseline hazard function $\lambda_0(t)$, and covariates proportionally impact that hazard.
- The proportional hazards assumption must be met; if it is not, then you can include a time interaction or use a different baseline hazard function within each stratum of that variable.

64

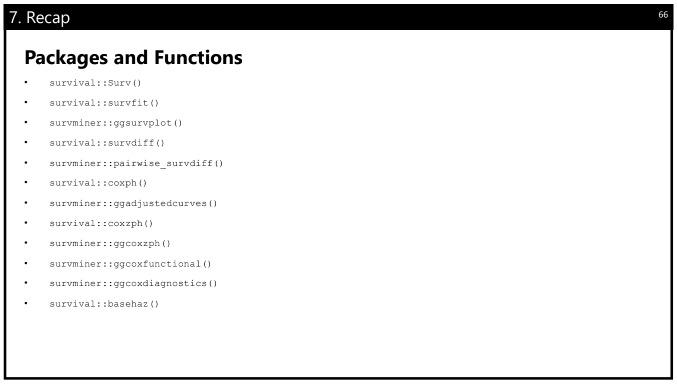


7. Recap 65

Additional Reading

- Incorporate time-varying covariates into the Cox PH model: https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf

65



7. Recap 66

Packages and Functions

- `survival::Surv()`
- `survival::survfit()`
- `survminer::ggsurvplot()`
- `survival::survdiff()`
- `survminer::pairwise_survdiff()`
- `survival::coxph()`
- `survminer::ggadjustedcurves()`
- `survival::coxzph()`
- `survminer::ggcoxzph()`
- `survminer::ggcoxfunctional()`
- `survminer::ggcoxdiagnostics()`
- `survival::basehaz()`

66