# Class 6 Outline

1. Survival analysis models
2. Weibull complementary log-log transformation, Standard errors, 95% CI for *S(t)*
3. Log-rank test for comparing survival curves
4. Cox proportional hazards regression model
5. Back to the AML Example
6. *Optional Example- FYI*: CABG
7. Summary

# 0. Learning Objectives

- Describe and use a log-rank test to compare two survival curves
- Describe and use the Cox proportional hazards regression model to compare survival experience.

Key words – survival function, Cox regression, partial likelihood, log-rank test, relative hazard, hazard ratio, proportional hazards, baseline hazard

# 1. Survival Analysis Models

- Survival analysis models relate to data in which the response variable is the time until an event occurs

- Regression models determine how times to an event depend on predictors; usually focus on the hazard (incidence) rates of events

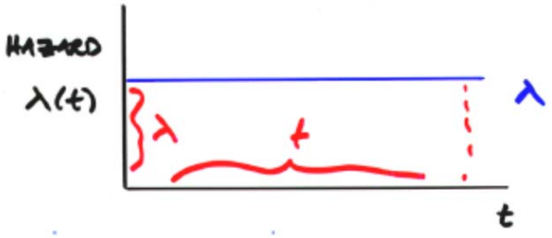
# 1.1 Parametric Models for Survival

- Exponential survival distribution (special case of Weibull distribution)

- Weibull survival distribution

# 1.2a Exponential Survival Distribution

- If the survival times follow an exponential distribution, the hazard function will look like:

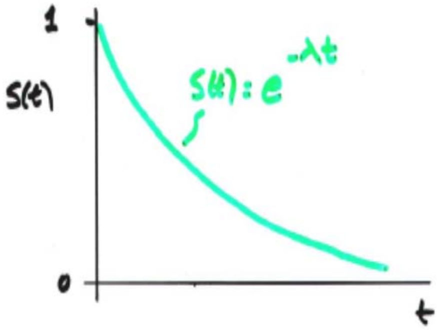


- The above graph shows a constant hazard - risk of event in small interval is always the same (e.g., light bulbs failing)

5

# 1.2b Exponential Survival Distribution

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



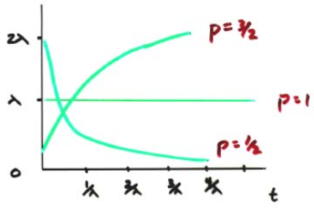
6

# 1.3a Weibull Survival Distribution

- “Simple” generalization of exponential distribution of survival times

$$S(t) = e^{-(\lambda t)^p}$$

- The Weibull distribution allows:
  - Hazard **increasing** with time ( **p>1** )
  - Hazard is **constant** with time ( **p=1** ) corresponds to an exponential distribution
  - Hazard **decreasing** with time ( **p<1** )



# 2. Complementary Log-Log Function

- Given a Weibull distribution: $S(t) = e^{-(\lambda t)^p}$
- One can derive the complementary *log-log (CLL)* transformation of *S(t)* for a Weibull distribution of survival times

$$
\begin{aligned}
\log[-\log S(t)] &= \log[-\log e^{-(\lambda t)^p}] \\
&= \log(\lambda t)^p \\
&= p \log \lambda + p \log t \\
&= \beta_0 + \beta_1 \log t
\end{aligned}
$$

- CLL(S(t)) is a linear function of log t

# 2.1 Use of CLL for Checking the Fit of a Weibull Survival Distribution

- Estimate S(t) using the Kaplan-Meier method
- Estimate CLL, the complementary log-log transformation of S(t):

$$\hat{v}(t) = \log(-\log \hat{S}(t))$$

- Plot $\hat{v}(t)$ vs. $\log t$
- If the plot *approximates a straight line*, then the Weibull distribution for survival times is a reasonable choice

# 2.2a Use of CLL for SE of S(t)

- The formula for the variance of the CLL, $\hat{v}(t) = \log(-\log \hat{S}(t))$ is

$$\mathrm{V\hat{a}r}(\hat{v}(t)) = \frac{\displaystyle\sum_{j:\, tj \le t} \frac{y_j}{n_j(n_j - y_j)}}{\left[\displaystyle\sum_{j:\, tj \le t} \log\left(\frac{n_j - y_j}{n_j}\right)\right]^2}$$

- Using the formula for the variance for the CLL, the SE of the estimated *v(t)* is

$$SE_{CLL}(t) = \sqrt{\mathrm{V\hat{a}r}((\hat{v}(t))}$$

## 2.2b Use CLL to Obtain 95% CI for S(t)

1. Get 95% CI for *v(t)*:

$$\hat{v}(t) \ \pm \ 1.96 \cdot SE_{CLL}(t)$$

2. Transform back to get 95% CI for *S(t)*:
   Use the inverse transformation

$$S(t)=e^{(-e^{v(t)})}$$

to get the 95% CI for *S(t)*:

$$\left(e^{(-e^{\hat{v}(t)-1.96SE_{CLL}(t)})}, e^{(-e^{\hat{v}(t)+1.96SE_{CLL}(t)})}\right)=[\hat{S}(t)]^{e^{\pm 1.96SE_{CLL}(t)}}$$

## 2.3 AML Data: Stata’s 95% CI for S(t)

(NOTE: **Stata** uses the *CLL* transformation for 95% CI on *S(t)*

Example: Back to the AML data

```
Beg.          Net            Survivor      Std.
  Time    Total   Fail   Lost           Function     Error     [95% Conf. Int.]
-----------------------------------------------------------------------------------
   --------------------------------------------------------------
     9       11      1      0             0.9091    0.0867     0.5081    0.9867
    13       10      1      1             0.8182    0.1163     0.4474    0.9512
    18        8      1      0             0.7159    0.1397     0.3502    0.8990
    23        7      1      0             0.6136    0.1526     0.2658    0.8353
    28        6      0      1             0.6136    0.1526     0.2658    0.8353
    31        5      1      0             0.4909    0.1642     0.1673    0.7534
    34        4      1      0             0.3682    0.1627     0.0928    0.6570
    45        3      0      1             0.3682    0.1627     0.0928    0.6570
    48        2      1      0             0.1841    0.1535     0.0117    0.5250
   161        1      0      1             0.1841    0.1535     0.0117    0.5250
```

## 2.4a AML Data: 95% CI Using CLL

- Using the formula for the estimated variance for the CLL at time 13:

$$\hat{\text{Var}}(\hat{v}(13)) = \frac{\sum_{j:\, tj \le 13} \frac{y_j}{n_j(n_j\text{-}y_j)}}{\left[\sum_{j:\, tj \le 13} \log\left(\frac{n_j\text{-}y_j}{n_j}\right)\right]^2}$$

$$= \frac{\left(\frac{1}{11(11\text{-}1)} + \frac{1}{10(10\text{-}1)}\right)}{\left[\log\frac{10}{11} + \log\frac{9}{10}\right]^2} = 0.502$$

## 2.4b AML Data: 95% CI Using CLL

- The 95% confidence interval for S(13) using the SE for the CLL is

$$\hat{S}(13)^{e^{\pm 1.96 SE_{CLL}(13)}} = (0.818^{e^{-1.96(0.709)}}, 0.818^{e^{+1.96(0.709)}})$$

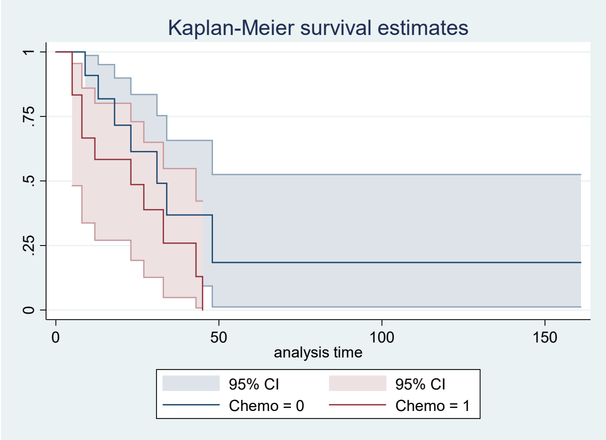
$$= \boxed{(0.447, 0.951)}$$

- Using the SE for the CLL provides a good option for calculating the 95% CI for S(t) and agrees with Stata

## 2.5 Graph of Estimated Survivor Functions with 95% CIs

```
.sts graph, by(Chemo) ci
```

$\widehat{S(t)}$



**t (weeks)**

## 3. Log-rank Test for Comparing Survivor Curves

- Are two survivor curves the same?
- Use the times of events: *t1, t2, ...* (do not include censoring times)
- Treat each event and its "set of persons still at risk" (i.e., risk set) at each time *tj* as an independent table
- Make a 2 × 2 table ***at each*** $\boldsymbol{t_i}$

| | Event | No Event | Total |
|---|---|---|---|
| Group A | $a_j$ | $n_{jA}$- $a_j$ | $n_{jA}$ |
| Group B | $c_j$ | $n_{jB}$-$c_j$ | $n_{jB}$ |
| Total | $d_j$ | $n_j$-$d_j$ | $n_j$ |

# 3.1 Calculating Expected Number of Events for the Log-rank Test

- At each event time $t_j$, under assumption of equal survival (i.e., SA(t) = SB(t) ), the expected number of events in Group A out of the total events ($d_j$=$a_j$ +$c_j$) is in proportion to the numbers at risk in group A to the total at risk at time $t_j$:

$$Ea_j = d_j \cdot \frac{n_{jA}}{n_j}$$

- Differences between $a_j$ and $Ea_j$ represent evidence against the null hypothesis of equal survival in the two groups

# 3.2 Log-rank Test as a Chi-Squared Statistic

- Use the Cochran Mantel-Haenszel idea of pooling over events j to obtain the log-rank chi-squared statistic with one degree of freedom

$$X_{LR}^2 = \frac{\left[\sum_j (a_j - Ea_j)\right]^2}{\sum_j V\hat{a}r\ a_j} \sim X_1^2$$

where $V\hat{a}r\ (a_j) = \frac{d_j(n_j - d_j)n_{jA}n_{jB}}{n_j^2(n_j - 1)}$

# 3.3 Log-rank Test in Stata

```
.sts test Chemo

         failure _d:  failed == 1
   analysis time _t:  t
                 id:  id


Log-rank test for equality of survivor functions

      |   Events         Events
Chemo |  observed       expected
------+------------------------
0     |         7          10.13
1     |        10           6.87
------+------------------------
Total |        17          17.00

            chi2(1) =       2.61
            Pr>chi2 =     0.1061
```

# 3.4 What Does the Log-rank Test Compare?

- It measures distance between curves by the summation over event times of the difference in hazards:

$$\sum_{\text{event times}:t_j} w(t_j)[h_1(t_j) - h_0(t_j)], \; w(t_j) \equiv 1$$

- If the hazards cross, the test loses power; early positive differences are partially offset by later negative differences
  - It is possible for the hazards to cross and the survival curves not to cross

- There are alternative tests
  - Different weight w( ) functions
    - Scale change and the generalized Wilcoxon

## 4.1a Cox Regression Model

- With a single covariate, Cox model provides the same inference as the log-rank statistic

- With multiple covariates, for example:
  - $X_1$ = treatment indicator
  - $X_2$ = gender
  - $X_3$ = CD4 cell count

- The model assumes proportional hazards

  h(t |treatment) = h(t |control)· constant($X_1$,$X_2$,$X_3$)

## 4.1b Cox Regression Model

- The hazards and survival curves are related by:

$$h(t\,|X_1,X_2,X_3) = h_0(t)\cdot e^{\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}$$

$$S(t\,|X_1,X_2,X_3) \;=\; [S_0(t)]^{e^{\beta_1 X_1+\beta_2 X_2+\beta_3 X_3}}$$

- One does not need to know $h_0(t)$, the baseline hazard, in order to estimate the coefficients

- Ease of implementation has made the Cox model the “t-test of survival analysis”

## 4.1c Cox Regression Model

- The regression model for the hazard function (the instantaneous incidence rate) as a function of p explanatory (X) variables is specified as follows:

  log hazard:

  $log\ h(t{:}\ X) = log\ h_0(t) + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_p X_p$

  hazard:

  $$h(t;X) = h_0(\mathrm{t}) \cdot e^{\beta_1 X_1} \cdot e^{\beta_2 X_2} \cdots \cdot e^{\beta_p X_p}$$

  $$h(t;X) = h_0(\mathrm{t}) \cdot e^{X\beta}$$

  Interpretation of $h_0(t)$:

  Hazard (incidence) rate as a function of time when all X's are zero; often must center X's to make $h_0(t)$ interpretable

## 4.2 Interpretation of Cox Regression Coefficients

- Interpretation of $e^{\beta 1}$

  $e^{\beta 1}$ is the relative hazard associated with a one unit change in $X_1$ (i.e., $X_1$+1 vs. $X_1$), holding other X's constant, at every time

- Synonymous terms: relative hazard, hazard ratio, "relative rate", "relative risk"

- Other β's have similar interpretations

  Note: $e^{X\beta}$ "multiplies" the baseline hazard $h_0(t)$ by the same amount regardless of the time t. This is therefore a "proportional hazards" model – the effect of any (fixed) X is the same at any time during follow-up

# 4.3 Cox Model as a Semi-Parametric Model

- David Cox (1972) showed how to estimate β without having to assume a model for $h_0(t)$
- β is the focus whereas $h_0(t)$ is a nuisance variable
- "Semi-parametric"
  - – $h_0(t)$ is the baseline hazard – "non-parametric" part of the model
  - – Xβ are the regression coefficients – "parametric" part of the model

# 4.4 Hazards and Risk Sets

- Let the survival times (times to failure) be:

  $t_1 < t_2 < \ldots. < t_k$
- And let the "risk sets" corresponding to these times be:

  $R_1, R_2, \ldots.., R_k$ where

  $R_i$ = the set of persons at risk of the event just before time $t_i$
- Then we can write:

$$= \frac{\text{hazard of failed person}}{\text{hazards of individuals who could have failed at } t_i}$$

- Choose β so that the individual who failed at each time was most likely, relative to the others who might have failed
- Connection to conditional logistic regression

# 5. Back to the AML Example

- Consider a clinical trial in patients with acute myelogenous leukemia (AML) comparing two groups of patients: no maintenance treatment with chemotherapy (*X=0*) -vs- maintenance chemotherapy treatment (*X=1*)

| Group | Weeks in remission -- ie, time to relapse |
|---|---|
| Maintenance chemo (*X=1*) | *9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+* |
| No maintenance chemo (*X=0*) | *5, 5, 8, 8, 12, 16+, 23, 27, 30+, 33, 43, 45* |

- + indicates a censored time to relapse; e.g., 13+ = more than 13 weeks to relapse

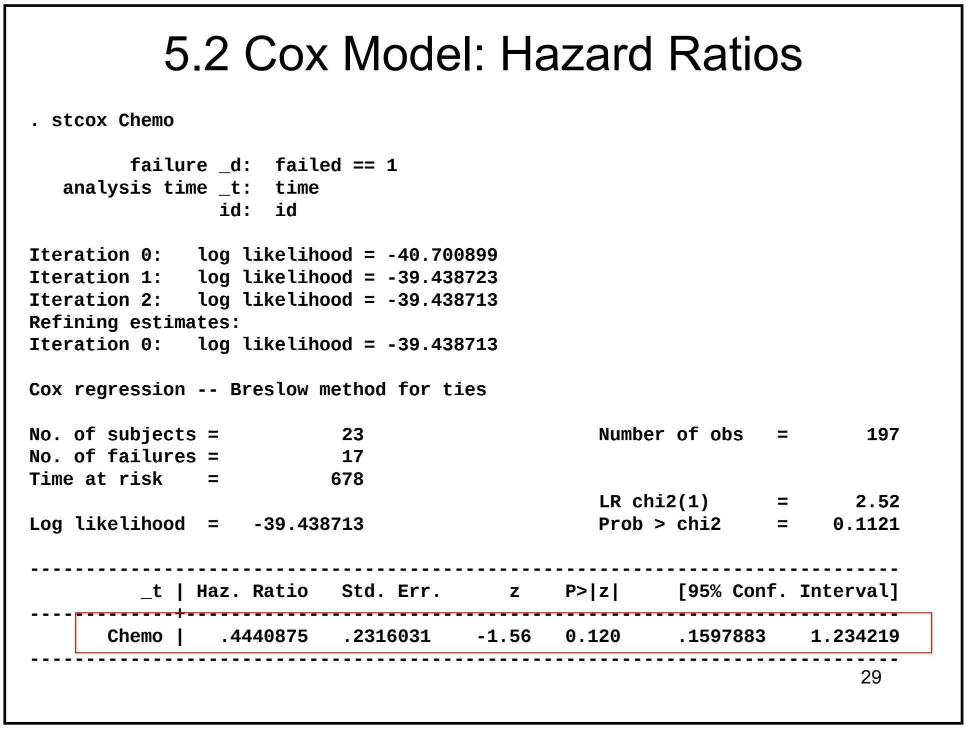
# 5.1 Cox PH Model for the AML Data

- Semi-parametric model for the hazard (incidence) rate for relapse in the AML data set

$$h_j(t) = h_0(\mathrm{t}) \cdot e^{X_j \beta}$$

- where $h_j(t)$ is the hazard for person j at week t, $h_0(t)$ is the hazard if $X_j$=0 (not maintained group) and $e^{Xj\beta}$ is the multiplicative effect of $X_j$=1 (maintained group)

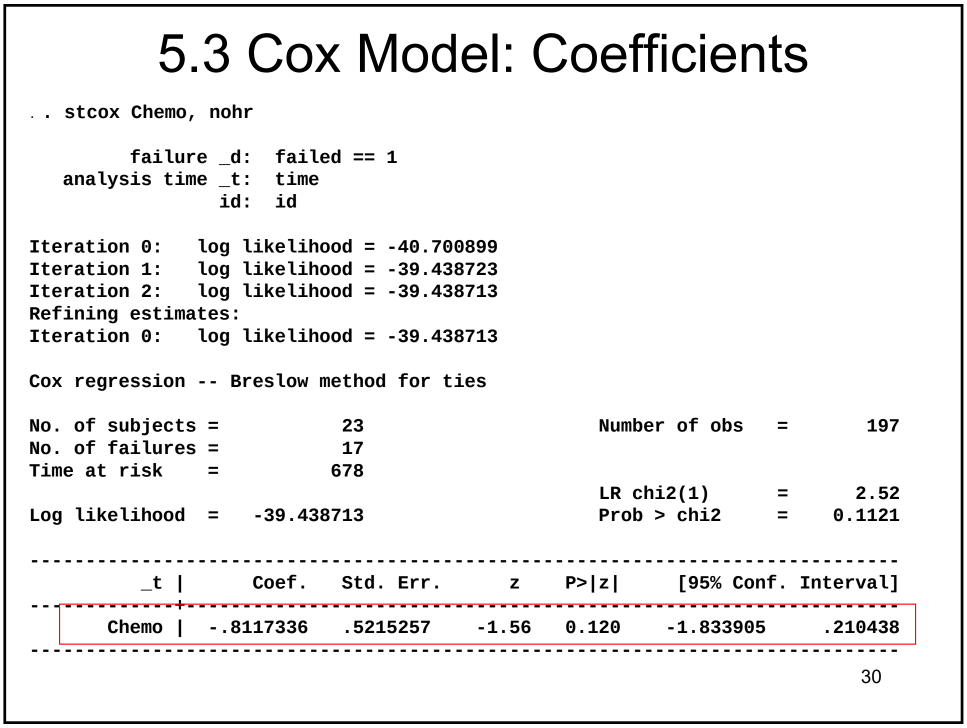
- Hazard ratio = $e^{\beta}$



# 5.2 Cox Model: Hazard Ratios

```
. stcox Chemo

         failure _d:  failed == 1
   analysis time _t:  time
                 id:  id

Iteration 0:   log likelihood = -40.700899
Iteration 1:   log likelihood = -39.438723
Iteration 2:   log likelihood = -39.438713
Refining estimates:
Iteration 0:   log likelihood = -39.438713

Cox regression -- Breslow method for ties

No. of subjects =           23                     Number of obs   =        197
No. of failures =           17
Time at risk    =          678
                                                   LR chi2(1)      =       2.52
Log likelihood  =   -39.438713                     Prob > chi2     =     0.1121

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Chemo |   .4440875   .2316031    -1.56   0.120     .1597883    1.234219
------------------------------------------------------------------------------
```

29



# 5.3 Cox Model: Coefficients

```
. . stcox Chemo, nohr

         failure _d:  failed == 1
   analysis time _t:  time
                 id:  id

Iteration 0:   log likelihood = -40.700899
Iteration 1:   log likelihood = -39.438723
Iteration 2:   log likelihood = -39.438713
Refining estimates:
Iteration 0:   log likelihood = -39.438713

Cox regression -- Breslow method for ties

No. of subjects =           23                     Number of obs   =        197
No. of failures =           17
Time at risk    =          678
                                                   LR chi2(1)      =       2.52
Log likelihood  =   -39.438713                     Prob > chi2     =     0.1121

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Chemo |  -.8117336   .5215257    -1.56   0.120    -1.833905     .210438
------------------------------------------------------------------------------
```

30

# 5.4 Cox Model: Interpretation of Coefficients

- $b_1$= -0.812 = difference in the log hazard rate of AML relapse in the maintained group (X=1) versus the not maintained group (X=0)

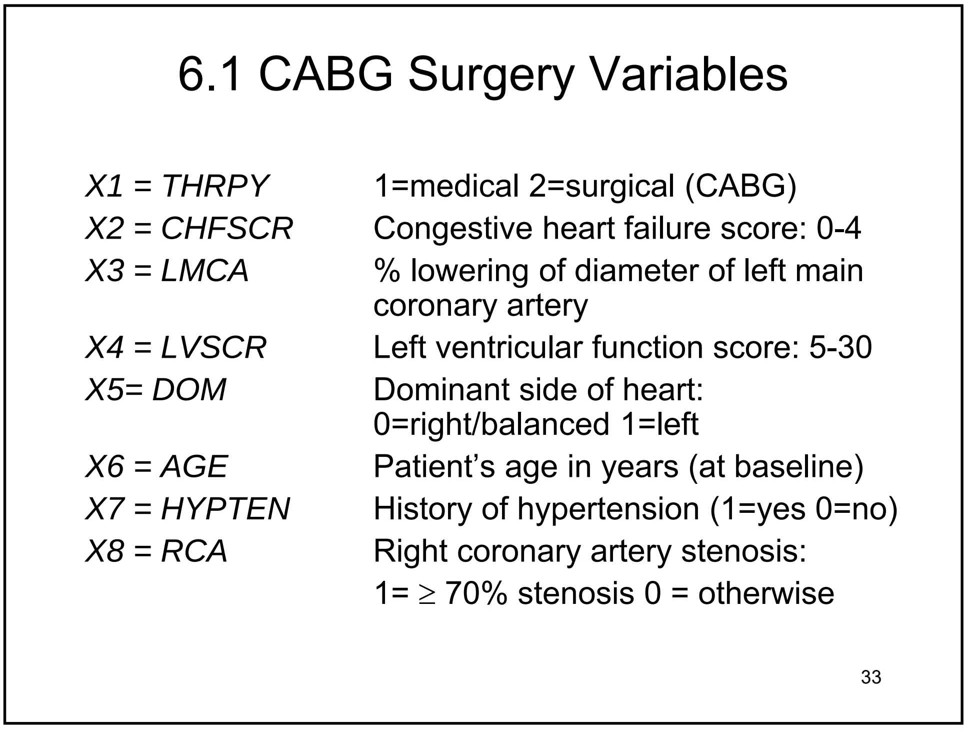
  $e^{b1}$ = $\boxed{0.44}$ = the hazard ratio of AML relapse in the maintained group versus the not maintained group

- 95% CI for $\beta_1$ : $b_1 \pm 1.96\ se(b_1)$ = (-1.83, 0.210)

  95% CI for $e^{\beta 1}$ : ($e.^{-1.83}$, $e.^{210}$) = (0.16, 1.23)

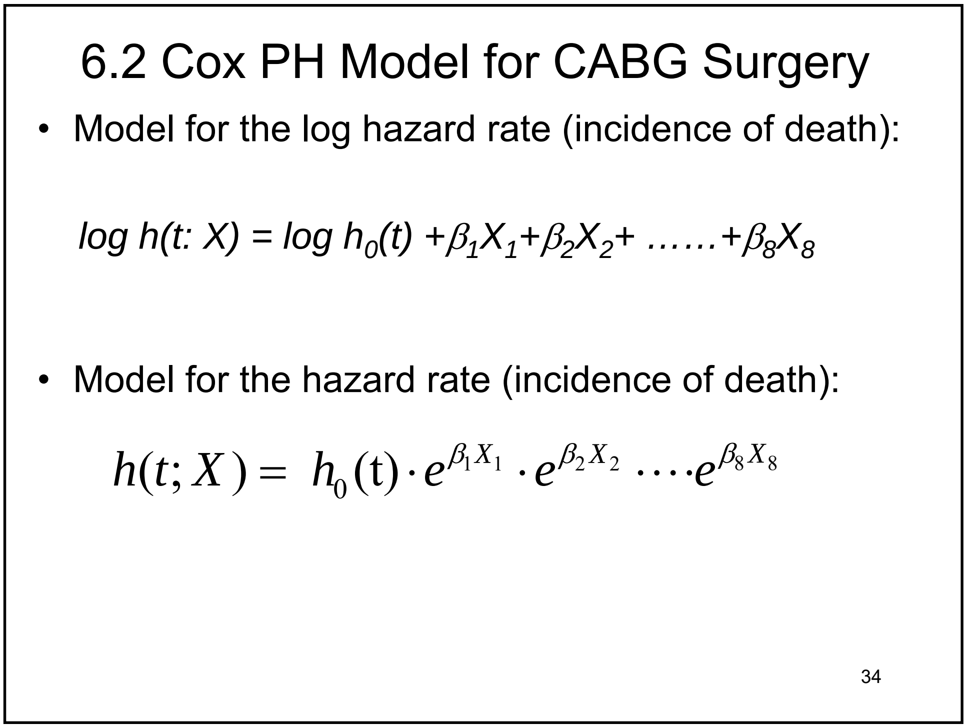
# 6. *Optional Example- FYI:* CABG Surgery Data Set

- Cox model to compare two treatments, controlling for several predictors (Fisher and Van Belle, 1993)
  - Compare surgical (CABG) with medical treatment for left main coronary heart disease
  - Use mortality (time to death) as the response variable
  - Control for 7 risk factors (age at baseline and 6 coronary status measures) in making the comparison
  - Time variable is time from treatment initiation to death or censoring due to the end of the study or lost to follow-up



## 6.1 CABG Surgery Variables

| | |
|---|---|
| *X1 = THRPY* | 1=medical 2=surgical (CABG) |
| *X2 = CHFSCR* | Congestive heart failure score: 0-4 |
| *X3 = LMCA* | % lowering of diameter of left main coronary artery |
| *X4 = LVSCR* | Left ventricular function score: 5-30 |
| *X5= DOM* | Dominant side of heart:<br>0=right/balanced 1=left |
| *X6 = AGE* | Patient’s age in years (at baseline) |
| *X7 = HYPTEN* | History of hypertension (1=yes 0=no) |
| *X8 = RCA* | Right coronary artery stenosis:<br>1= $\geq$ 70% stenosis 0 = otherwise |



## 6.2 Cox PH Model for CABG Surgery

- Model for the log hazard rate (incidence of death):

$$\log h(t: X) = \log h_0(t) + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_8 X_8$$

- Model for the hazard rate (incidence of death):

$$h(t;X) = h_0(\mathrm{t}) \cdot e^{\beta_1 X_1} \cdot e^{\beta_2 X_2} \cdots \cdot e^{\beta_8 X_8}$$

## 6.3 Cox PH Model Results for CABG Surgery

| Variable | Estimated $\beta$ = b | SE(b) | Z=b/SE(b) |
|---|---|---|---|
| *THRPY* | *-1.0777* | *.1668* | *-6.46* |
| *CHFSCR* | *.2985* | *.0667* | *4.48* |
| *LMCA* | *.0178* | *.0049* | *3.63* |
| *LVSCR* | *.1126* | *.0182* | *6.19* |
| *DOM* | *1.2331* | *.3564* | *3.46* |
| *AGE* | *.0423* | *.0098* | *4.32* |
| *HYPTEN* | *-.5428* | *.1547* | *-3.51* |
| *RCA* | *.5285* | *.2923* | *1.81* |

## 6.4a Interpretation of THRPY Coefficient

- What is the relative hazard of death for the CABG group compared to the medical group, adjusted for age and other risk factors?
  - $b_1$= -1.0777 = difference in the log hazard rate in the CABG group (X=2) and the medical group (X=1) = log (hazard rate ratio)
  - $e^{b1} = e^{-1.0777}$ = 0.34 = hazard ratio comparing the CABG group (X=2) and the medical group (X=1)

Note:

Coding 2 = CABG, 1 = Medical gives the same results as coding 1
1 = CABG, 0 = Medical

## 6.4b Interpretation of THRPY Coefficient

- 95% CI for $\beta_1$ : $b_1 \pm 1.96\ se(b_1)$
  $= -1.0777 \pm 1.96(0.1668)$
  $= (-1.406, -0.750)$

- 95% CI for $e^{\beta 1}$ : $(e^{-1.406}, e^{-0.750}) = (0.25, 0.47)$

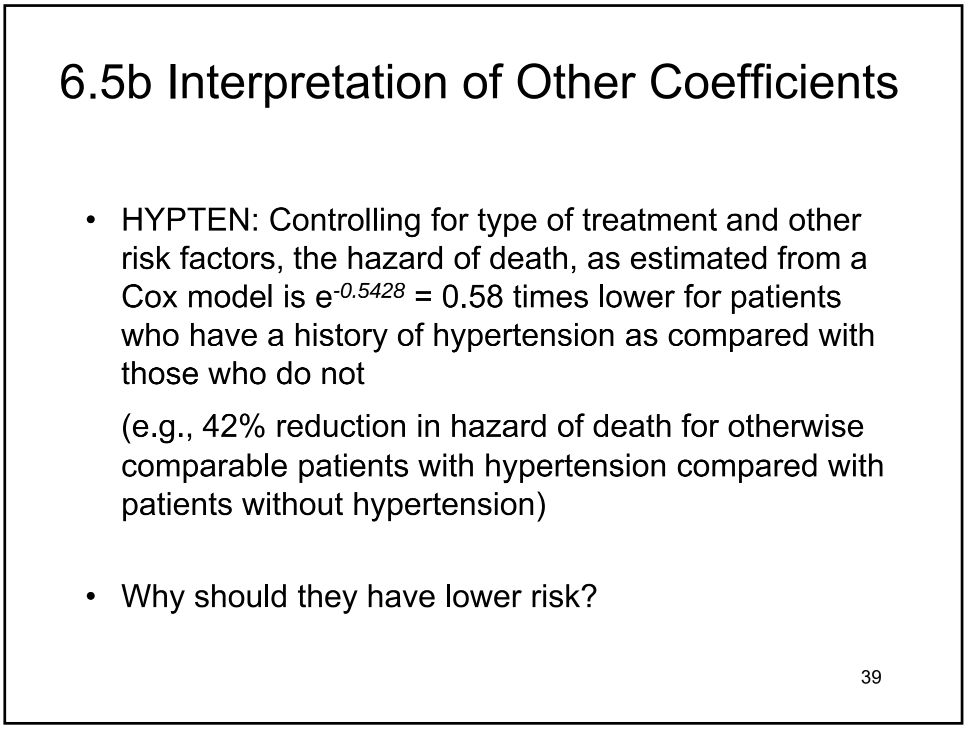
- Adjusted HR = 0.34,     95% CI (0.25, 0.47)

- Thus, there is an estimated 66% reduction in the hazard (“risk”) of death for otherwise comparable patients treated with CABG as compared with patients treated medically

## 6.5a Interpretation of Other Coefficients

- CHFSCR: Controlling for type of treatment and other risk factors, the hazard of death, as estimated from a Cox model is $e^{0.2985}$ = 1.35 times higher per unit increase in CHF score

- AGE: Controlling for type of treatment and other risk factors, the hazard of death, as estimated from a Cox model is $e^{0.0423}$ = 1.04 times higher per year of age



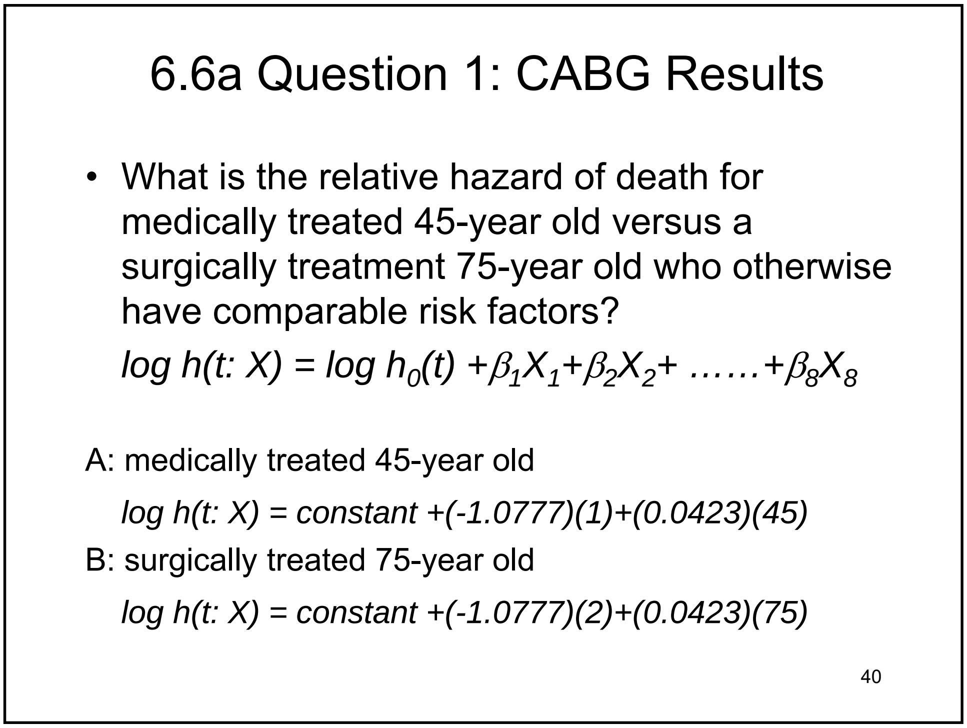
## 6.5b Interpretation of Other Coefficients

- HYPTEN: Controlling for type of treatment and other risk factors, the hazard of death, as estimated from a Cox model is $e^{-0.5428}$ = 0.58 times lower for patients who have a history of hypertension as compared with those who do not

  (e.g., 42% reduction in hazard of death for otherwise comparable patients with hypertension compared with patients without hypertension)

- Why should they have lower risk?

39



## 6.6a Question 1: CABG Results

- What is the relative hazard of death for medically treated 45-year old versus a surgically treatment 75-year old who otherwise have comparable risk factors?

  $\log h(t{:}\ X) = \log h_0(t) + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_8 X_8$

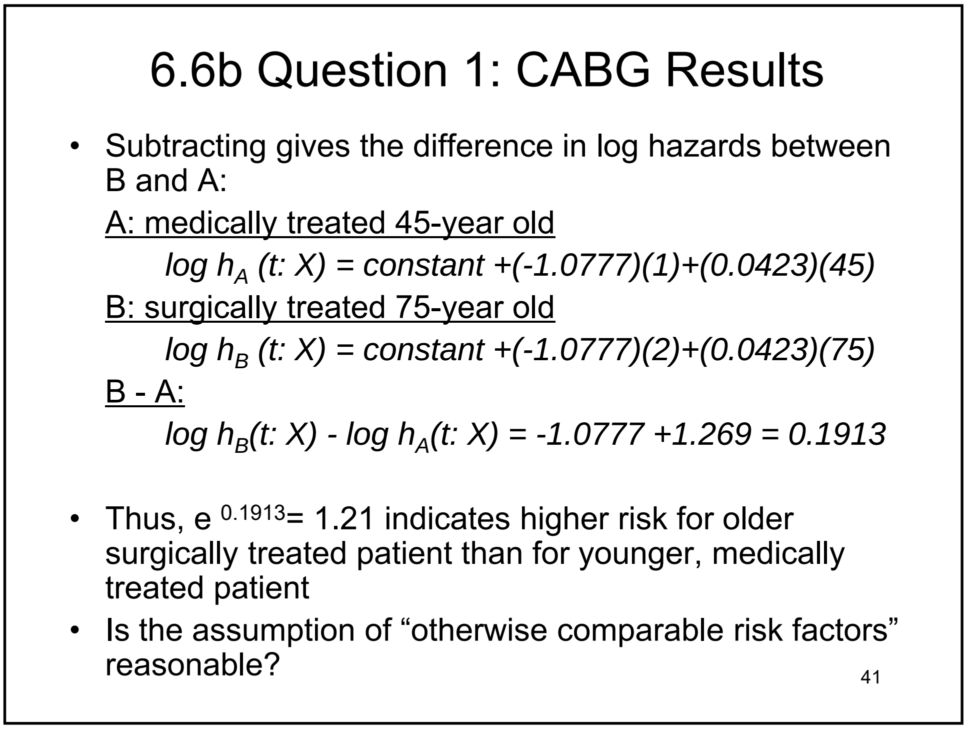
A: medically treated 45-year old

$\log h(t{:}\ X) = \text{constant} + (-1.0777)(1) + (0.0423)(45)$

B: surgically treated 75-year old

$\log h(t{:}\ X) = \text{constant} + (-1.0777)(2) + (0.0423)(75)$

40



## 6.6b Question 1: CABG Results

- Subtracting gives the difference in log hazards between B and A:

  <u>A: medically treated 45-year old</u>

  $\log h_A\,(t{:}\ X) = \text{constant} + (-1.0777)(1) + (0.0423)(45)$

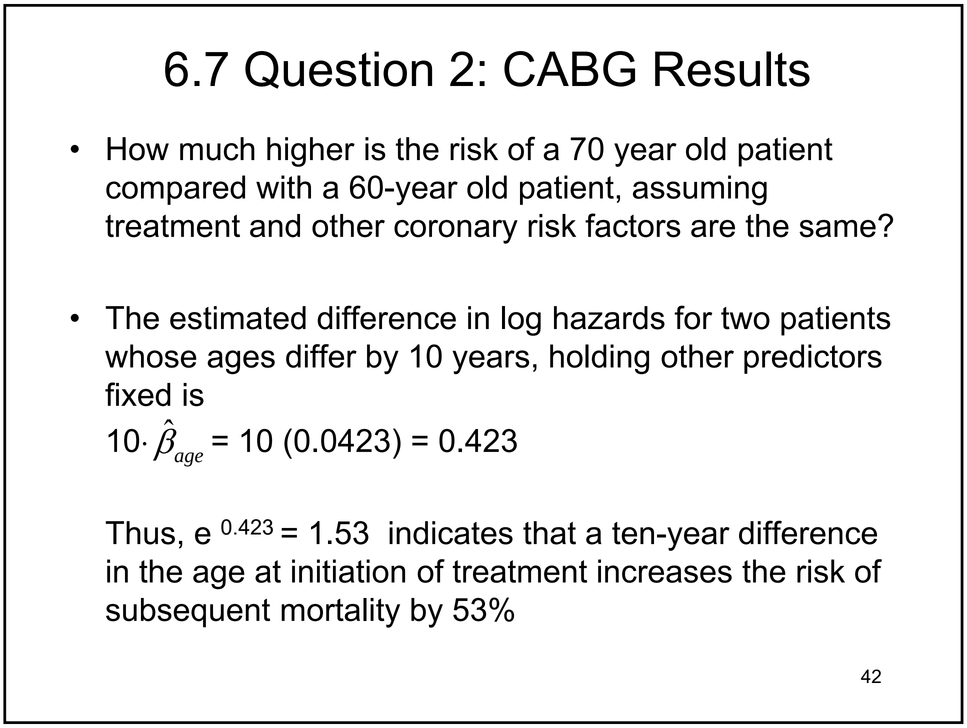
  <u>B: surgically treated 75-year old</u>

  $\log h_B\,(t{:}\ X) = \text{constant} + (-1.0777)(2) + (0.0423)(75)$

  <u>B - A:</u>

  $\log h_B(t{:}\ X) - \log h_A(t{:}\ X) = -1.0777 + 1.269 = 0.1913$

- Thus, $e^{0.1913} = 1.21$ indicates higher risk for older surgically treated patient than for younger, medically treated patient
- Is the assumption of “otherwise comparable risk factors” reasonable?



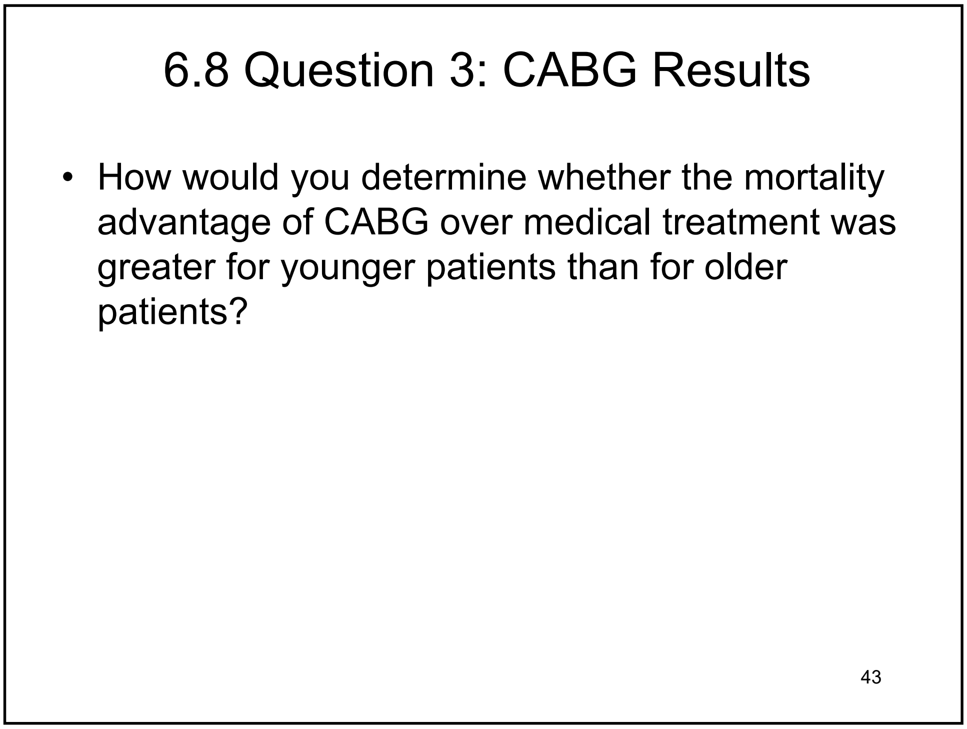
## 6.7 Question 2: CABG Results

- How much higher is the risk of a 70 year old patient compared with a 60-year old patient, assuming treatment and other coronary risk factors are the same?

- The estimated difference in log hazards for two patients whose ages differ by 10 years, holding other predictors fixed is

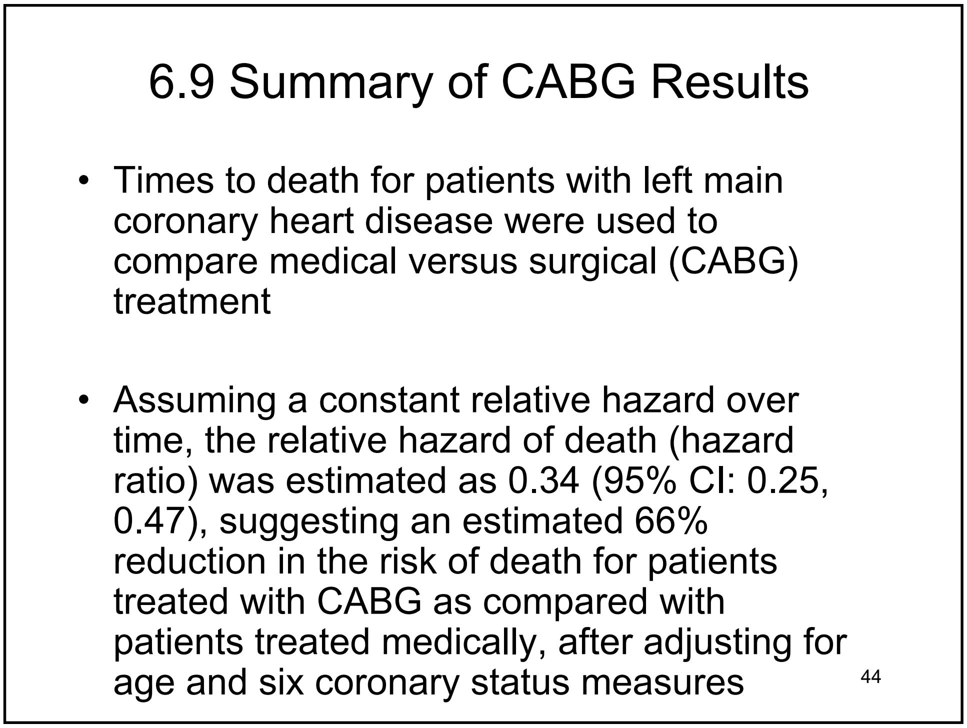
  $10 \cdot \hat{\beta}_{age} = 10\ (0.0423) = 0.423$

  Thus, $e^{0.423} = 1.53$ indicates that a ten-year difference in the age at initiation of treatment increases the risk of subsequent mortality by 53%



## 6.8 Question 3: CABG Results

- How would you determine whether the mortality advantage of CABG over medical treatment was greater for younger patients than for older patients?



## 6.9 Summary of CABG Results

- Times to death for patients with left main coronary heart disease were used to compare medical versus surgical (CABG) treatment
- Assuming a constant relative hazard over time, the relative hazard of death (hazard ratio) was estimated as 0.34 (95% CI: 0.25, 0.47), suggesting an estimated 66% reduction in the risk of death for patients treated with CABG as compared with patients treated medically, after adjusting for age and six coronary status measures

# 7a. Summary

- The Weibull probability distribution can be used to describe survival times (with the exponential distribution as a special case); the complementary log-log (CLL) function can be derived
- The 95% confidence interval for the survivor function can be best estimated using the standard error for the CLL function
- The log-rank statistic can be used to compare two survival curves; equal weight is given to each event time

# 7b. Summary

- The Cox proportional hazards regression model for the log hazard rate $\lambda_j$ as a function of p explanatory (X) variables is specified as follows:

$$log\ h(t: X) = log\ h_0(t) + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_p X_p$$

- Interpretation of $h_0(t)$: hazard (incidence) rate as a function of time when all X's are zero; often must center X's to make $h_0(t)$ interpretable

# 7c. Summary

- The model for the expected hazard rate is:

$$h(t;X) = h_0(\mathrm{t}) \cdot e^{\beta_1 X_1} \cdot e^{\beta_2 X_2} \cdot \cdots \cdot e^{\beta_p X_p}$$

$$h(t;X) = h_0(\mathrm{t}) \cdot e^{X\beta}$$

- And, $e^{\beta}$ is the relative hazard (hazard ratio) associated with a one unit change in $X_1$ (i.e., $X_1$+1 vs. $X_1$), holding other X's constant, independent of time

<u>Note</u>: $e^{X\beta}$ "multiplies" the baseline hazard $h_0(t)$ by the same amount regardless of the time t. This is therefore a "proportional hazards" model – the effect of any (fixed) X is the same at any time during follow-up