BIOS 312: MODERN REGRESSION ANALYSIS

James C (Chris) Slaughter
Department of Biostatistics
Vanderbilt University School of Medicine
james.c.slaughter@vanderbilt.edu

biostat.mc.vanderbilt.edu/CourseBios312

Copyright 2009-2012 JC Slaughter All Rights Reserved

Updated March 8, 2012

Contents

9 Simple Proportional Hazards Regression 5

9.1 Right Censored Data 5

9.2 Survival Regression 7

9.3 Proportional Hazards Regression 9

9.3.1 The simple PH regression model 10

9.3.2 Descriptive Plots 11

9.4 Software 16

9.5 Example Analysis 17

9.5.1 Descriptives 17

9.5.2 Regression model 19

9.6 Inference with PH Regression 21

9.6.1 Implications for Inference 23

9.7 Example Revisited 25

9.7.1 Robust standard errors 25

9.7.2 Log-transformed nadir PSA 26

9.8 Review: Interpretation of Slopes 27
9.8.1 Additive models 27
9.8.2 Multiplicative models 28

Chapter 9

Simple Proportional Hazards Regression

9.1 Right Censored Data

- When the response variable is time to event, and subjects are not followed long enough for the event to have occurred, the event times are *censored* at the last follow up
  - Definition: A survival time is said to be *right censored* at time t if it is only known to be greater that t
  - e.g. if measuring time to death over a 5 year study, all subjects may not have died after 5 years
- Survival data is special type of missing data: the exact value is not always known
  - Some outcomes are known exactly. This occurs when we observe the event at time t
  - Some outcomes are only known to exceed some value (right censored at time t)

- Thus, the response is typically represented by two variables
  - An observation time: Time to event or censoring, whichever came first
  - And indicator of event: Was the observation time an event or censored observation?
- Most statistical analyses assume that the reason for censoring is independent of the risk of the event
  - Example: Subject moves out of town, study runs out of money before followup is complete on everyone
  - We assume that censored subjects have the same event risk after they are no longer observed as subjects who remain in the study
  - Example where assumption is violated (informative censoring)
    - A subject is pulled off a drug for treatment failure (and censored at that time). Such a subject would likely be at increased risk for a negative outcome.
    - In a smoking cessation study, we are interested in measuring time until a subject starts smoking again. Subjects who drop out of the study (censored) often do so because they are likely to start smoking again.
- Statistical methods
  - Regression models we have studied to date are generally not valid for survival data
    - Because of right censoring, survival time cannot be analyzed as a continuous outcome (e.g. linear regression)
    - Because of unequal length of followup, survival (yes/no) cannot be analyzed using logistic regression

- In the presence of censored data, the usual descriptive statistics are not appropriate
  - Sample mean, sample median, simple proportions, sample standard deviation should not be used
  - Proper descriptives should be based on the Kaplan Meier estimates
- Specialized regression models are needed with censored data

9.2 Survival Regression

- Notation
  - Unobserved data
    - True times to events: $T_1^0, T_2^0, \ldots, T_n^0$
    - True censoring times: $C_1, C_2, \ldots, C_n$
  - Observed data
    - Observation times: $T_i = \min(T_i^0, C_i)$
    - Event indicators
      - $D_i = 1$ if $T_i = T_i^0$, or
      - $D_i = 0$ otherwise
- Two fundamental models used to describe the way that some factor might affect time to event
  - Accelerated failure time (less popular)
  - Proportional hazards (aka Cox model)

- Accelerated Failure Time model
  - Assume that a factor causes some subjects to spend their lifetime too fast
  - Basic idea: For every year in a reference group's lives, the other group ages "k" years
    - e.g. 1 human year is about 7 dog years
  - Assumes that ratios of quantiles of survival distribution are constant across groups
    - e.g. report median ratios: It takes 50% longer for half of the treated group to die than half of the control group
  - AFT models include the parametric exponential, Weibull, and lognormal models
  - Discussed in more detail in advanced survival analysis courses

- Proportional hazards model
  - Considers the instantaneous risk of failure at each time among those subjects who have not failed
  - The term "proportional hazards" assumes that the ratio of these instantaneous failure rates is constant in time between groups
  - Proportional hazards (Cox) regression treats the survival distribution within a group semi-parametrically
    - The baseline hazard is estimated without making any distributional assumptions
    - The hazard ratio is the parameter of interest and is used to compare groups

9.3 Proportional Hazards Regression

- Ignores the time that events occurred

- Looks at odds of choosing subjects relative to prevalence in the population
  - Can be derived as estimating the odds ratio of an event at each time that an event occurs

  - Proportional hazards model averages the odds ratio across all observed event times

  - If the odds ratio is constant over time between two groups, such an average results in a precise estimate of the hazard ratio

- Borrowing information
  - Uses other groups to make estimates in groups with sparse data

  - Borrows information across predictor groups
    * Intuitively, 67 and 69 year olds would provide some relevant information about 68 year olds

    * Assuming a straight line relationship tells us about other, even more distant, individuals

    * If we do not want to assume a straight line, we may only want to borrow information from nearby groups

  - Borrow information over time
    * Relative risk for an event at each time is presumed to be the same under proportional hazards

9.3.1 The simple PH regression model

- Modeling the log hazard over time as a function of covariates X

- "Baseline hazard" is unspecified. Baseline hazard is similar to an intercept

Model	$\log(\lambda(t\|X_i)) = \log(\lambda_0(t)) + \beta_1 \times X_i$
$X_i = 0$	log hazard at t is $\log(\lambda_0(t))$
$X_i = x$	log hazard at t is $\log(\lambda_0(t)) + \beta_1 \times x$
$X_i = x + 1$	log hazard at t is $\log(\lambda_0(t)) + \beta_1 \times x + \beta_1$

- Model on the hazard scale is found by exponentiating parameters

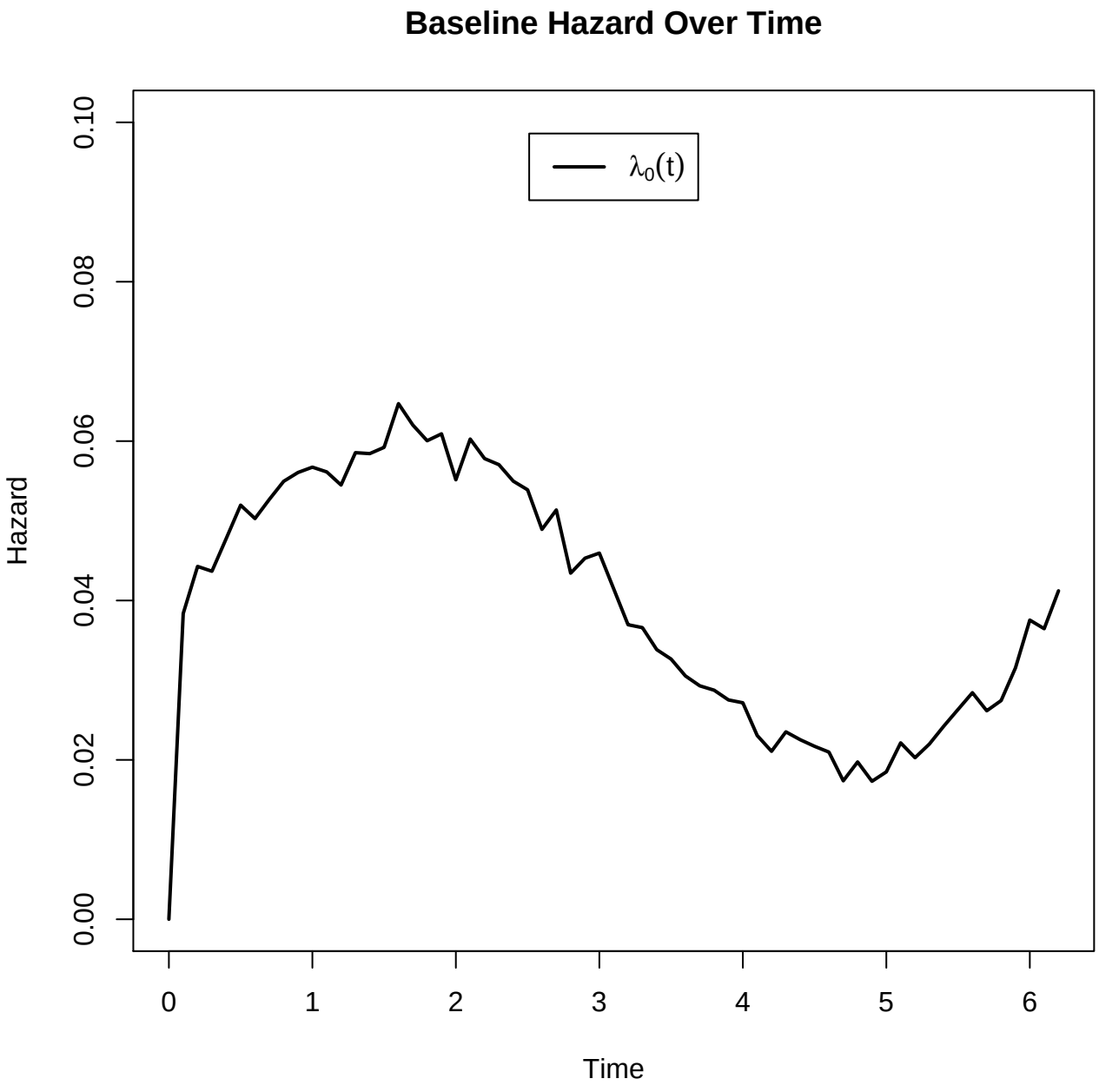
Model	$\lambda(t\|X_i) = \lambda_0(t) \times e^{\beta_1 \times X_i}$
$X_i = 0$	hazard at t is $\lambda_0(t)$
$X_i = x$	hazard at t is $\lambda_0(t) \times e^{\beta_1 \times x}$
$X_i = x + 1$	hazard at t is $\lambda_0(t) \times e^{\beta_1 \times x} \times e^{\beta_1}$

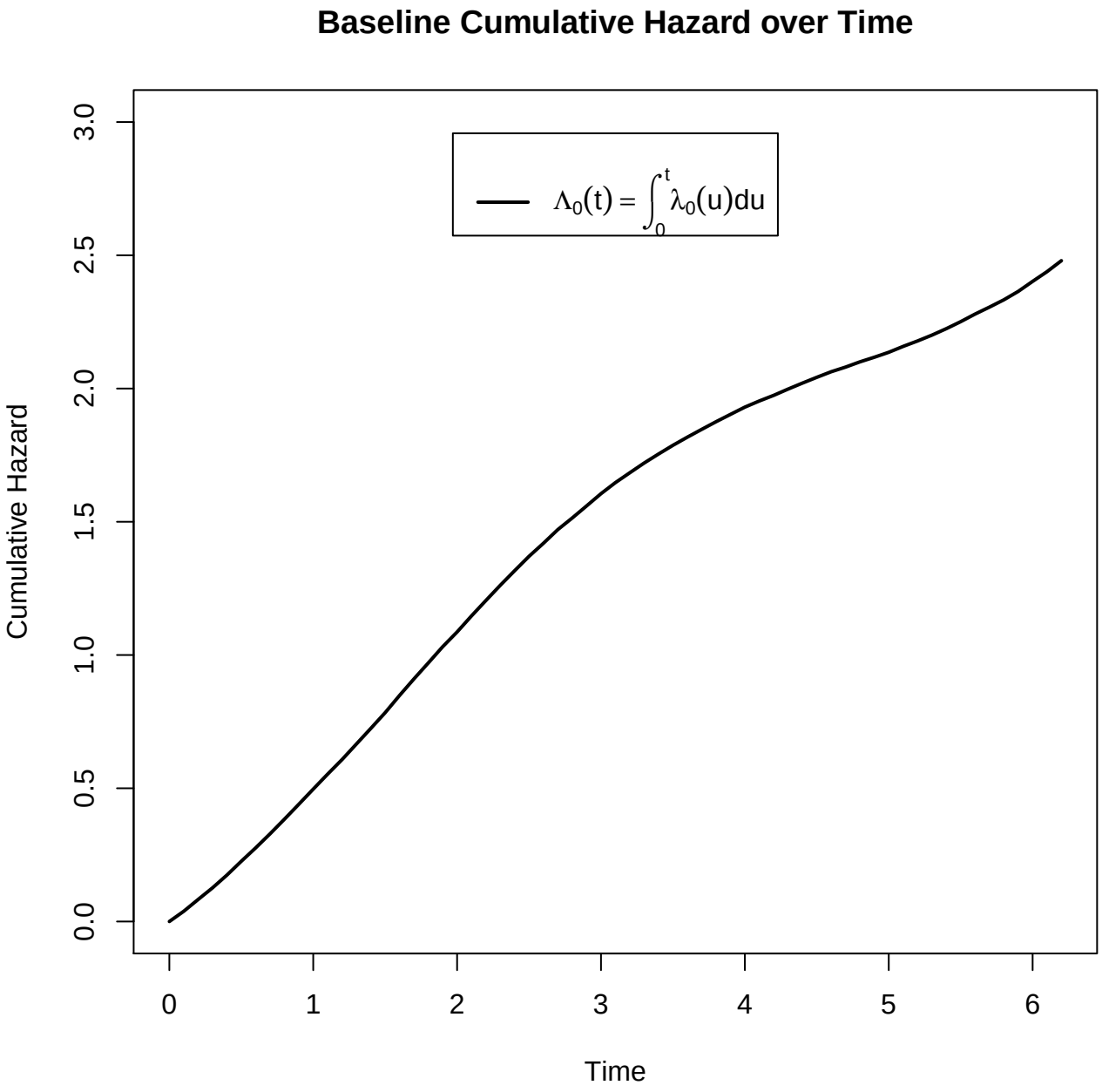
- Interpretation of the model
  - No intercept
    - Generally ignore the baseline hazard
  - Slope parameter
    - Hazard ratio found by exponentiating the slope from the PH regression: $\exp(\beta_1)$
    - Hazard ratio compared groups differing in the value of the predictor by 1 unit

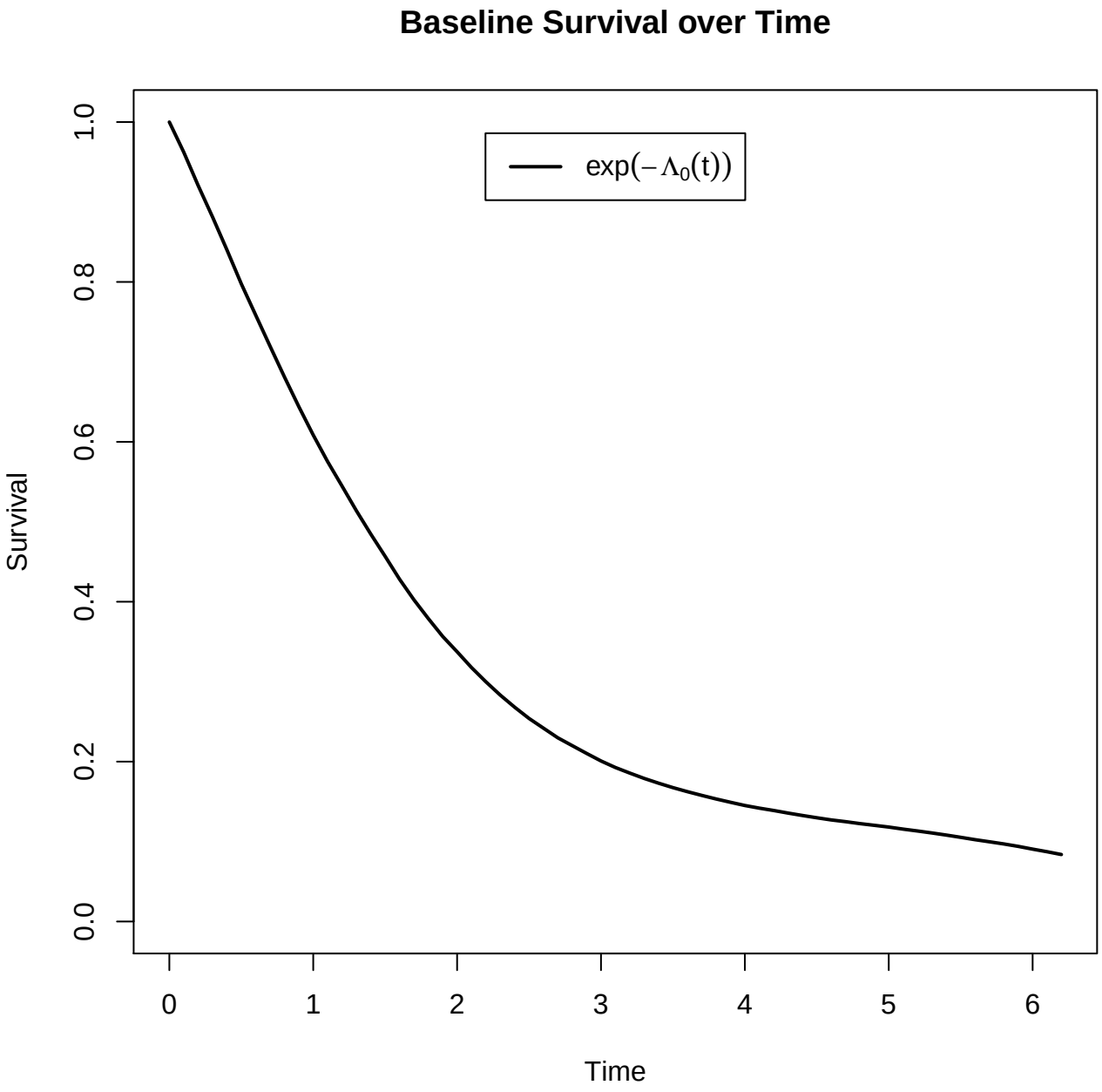
- Relationship to survival
  - Hazard function determines the survival function

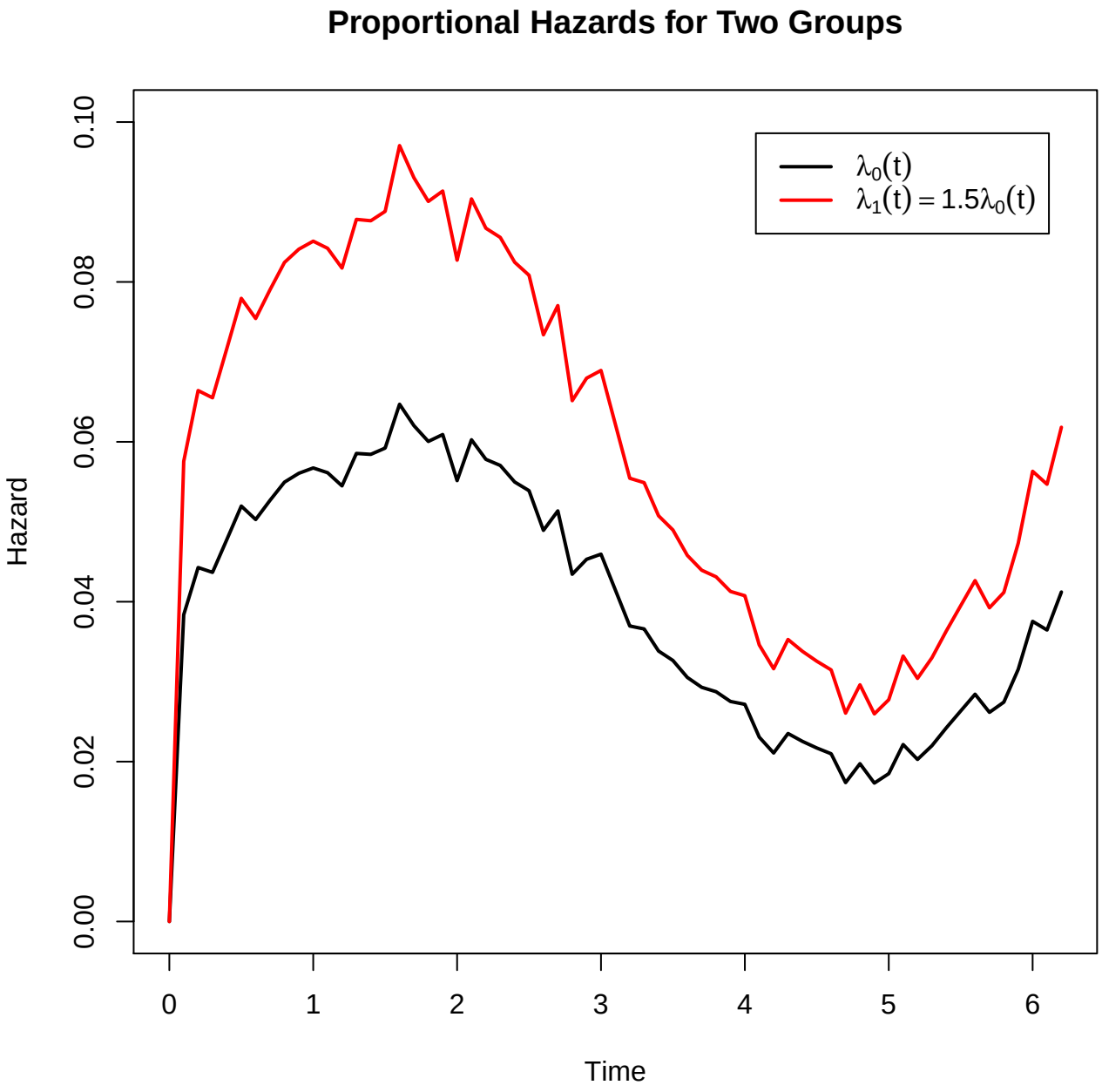
Hazard	$\lambda(t\|X_i) = \lambda_0(t) \times e^{\beta_1 \times X_i}$
Cumulative Hazard	$\Lambda(t\|X_i) = \int_0^t \lambda_0(u) \times e^{\beta_1 \times X_i}\, du$
Survival Function	$S(t\|X_i) = e^{-\Lambda(t\|X_i)} = [S_0(t)]^{e^{\beta_1 \times X_i}}$

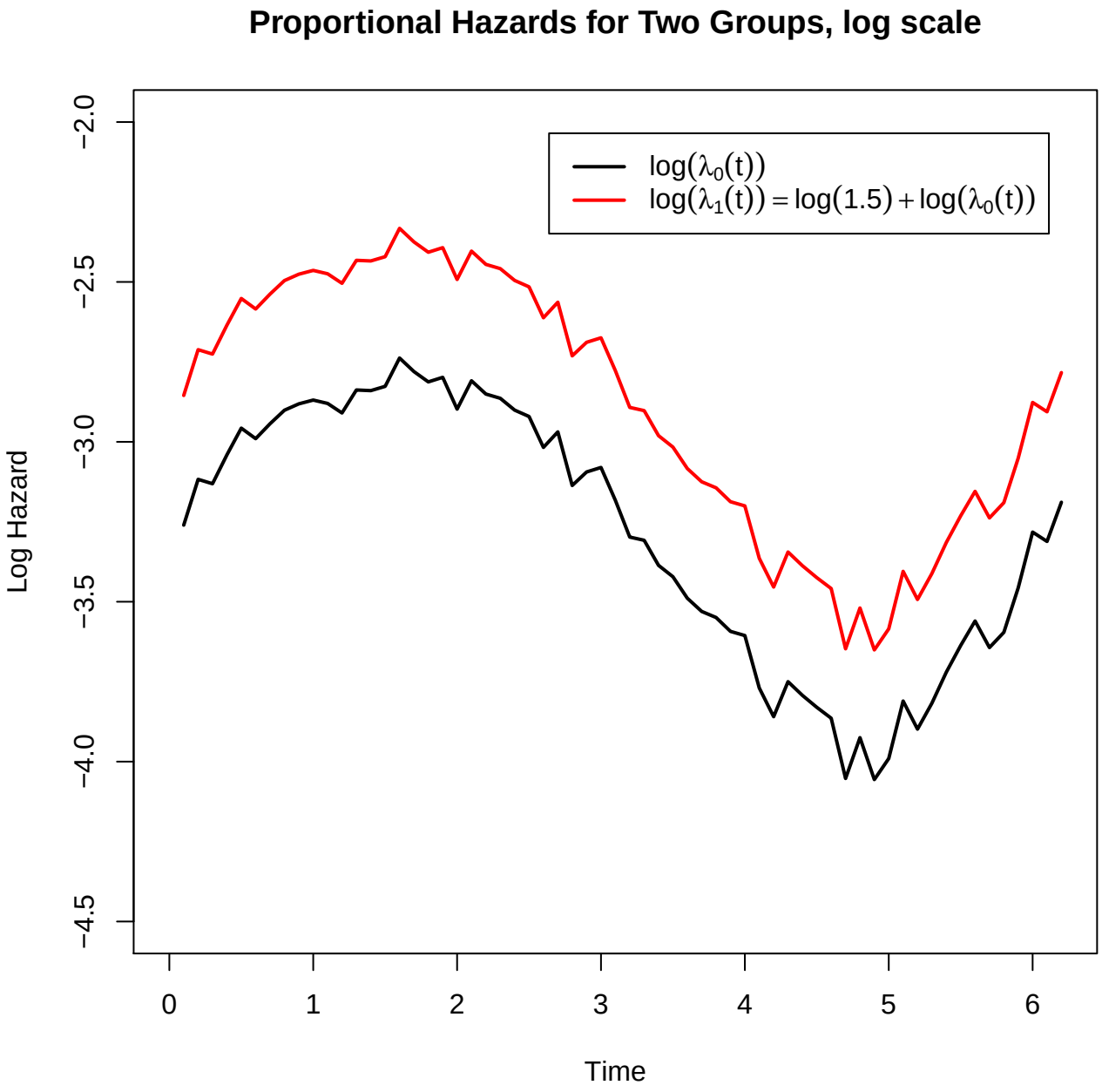
9.3.2 Descriptive Plots




Baseline Cumulative Hazard over Time
$\Lambda_0(t) = \int_0^t \lambda_0(u)du$
Cumulative Hazard
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
6
Time


Baseline Survival over Time
exp(−Λ0(t))
Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
Time


Proportional Hazards for Two Groups
Hazard
0.00
0.02
0.04
0.06
0.08
0.10
0
1
2
3
4
5
6
Time
λ0(t)
λ1(t) = 1.5λ0(t)


Proportional Hazards for Two Groups, log scale
log(λ0(t))
log(λ1(t)) = log(1.5) + log(λ0(t))
Log Hazard
−2.0
−2.5
−3.0
−3.5
−4.0
−4.5
0
1
2
3
4
5
6
Time

- Comments on plots
  - Baseline hazard can follow any functional form. This is the "non-parametric" part of the Cox proportional hazards model
  - The cumulative hazard is a non-decreasing function that starts at 0 at time 0. It is bounded by ∞.
  - Survival is a function of the cumulative hazard. Survival is 1 at time 0 and decreases over time. It is bounded by 0 and 1.
  - For a dichotomous predictor variable (two groups), the proportional hazards assumption is that $\lambda_0(t) = e^{\beta_1}\lambda_0(t)$
    * In the plots I illustrated $e^{\beta_1} = 1.5$ ($\beta_1 = 0.4054$).
    * On the hazard scale, this corresponds to $\lambda_1(t)$ always being 1.5 times larger than $\lambda_0(t)$.
    * On the log hazard scale, this corresponds to $log(\lambda_1(t))$ always being $log(1.5) = 0.4054$ units larger than $log(\lambda_0(t))$.
    * The proportional hazards assumption is the "parametric" part of the Cox proportional hazards model. Hence, the Cox proportional hazards model is referred to as being "semi-parametric".

9.4 Software

- Stata commands
  - "stset time event-indicator"
  - "stcox predictor, [robust]"
- R functions

  - "Surv(time, event-indicator)"

  - "coxph(Surv(time, event-indicator) ~ predictor)"

9.5 Example Analysis

9.5.1 Descriptives

- Prognostic values of nadir PSA relative to time in remission
  - PSA dataset: 50 men who received hormonal treatment for advanced prostate cancer

  - Followed at least 24 months for clinical progression, but exact followup varies from subject to subject

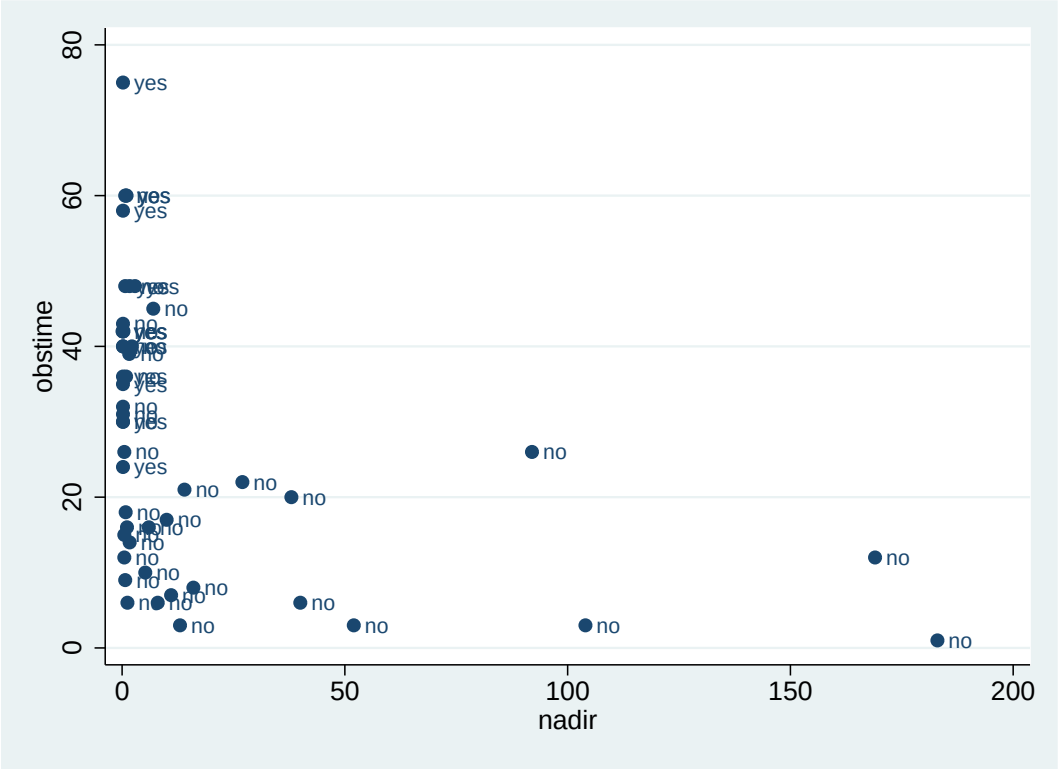
  - Nadir PSA: lowest level of serum prostate specific antigen achieved post treatment

- Scatterplots of censored data are not scientifically meaningful
  - It is best to not generate them unless you do something to indicate the censored dataset

  - We can label censored data, but have to remember that the true value may be anywhere larger than that

  - Following plot in Stata: "scatter obstime nadir, mlablel(inrem)"

- Characterization of plot
  - Outliers: Can't tell



- First order trends
  - Downward trending slop
  - No censoring at high nadir PSAs
- Second order trend: Must be curvilinear (but how much?)
- Variability within groups: Highest variability within lowest PSA groups

9.5.2 Regression model

```
. encode inrem, gen(inrem2)
. gen relapse = abs(inrem2 - 2)
. table relapse inrem

----------------------
          |   inrem
  relapse |   no   yes
----------+-----------
        0 |         14
        1 |   36
----------------------

. stset obstime relapse

     failure event:  relapse != 0 & relapse < .
obs. time interval:  (0, obstime]
 exit on or before:  failure

------------------------------------------------------------------------------
       50  total obs.
        0  exclusions
------------------------------------------------------------------------------
       50  obs. remaining, representing
       36  failures in single record/single failure data
     1423  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        75
```

- Create a relapse variable to indicate if a subject has relapsed or not
- Define the outcome using “stset”
  - Time to event (relapse or censored)
  - Failture indicator (relapse)

```
. stcox nadir

         failure _d:  relapse
   analysis time _t:  obstime

Iteration 0:   log likelihood = -118.96593
Iteration 1:   log likelihood = -113.88229
Iteration 2:   log likelihood = -113.29294
Iteration 3:   log likelihood = -113.28918
Iteration 4:   log likelihood = -113.28918
Refining estimates:
Iteration 0:   log likelihood = -113.28918

Cox regression -- Breslow method for ties

No. of subjects =           50                     Number of obs   =        50
No. of failures =           36
Time at risk    =         1423
                                                   LR chi2(1)      =     11.35
Log likelihood  =   -113.28918                     Prob > chi2     =    0.0008

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nadir |    1.01551   .0038122     4.10   0.000     1.008066     1.02301
------------------------------------------------------------------------------

. lincom 10*nadir, hr

 ( 1)  10 nadir = 0

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.166387   .0437864     4.10   0.000     1.083648    1.255443
------------------------------------------------------------------------------
```

- Interpretation of output
  - By default, stata gives the hazard ratio, not the coefficient
    - $\text{Hazard ratio} = e^{\text{coeff}}$
  - $\text{Hazard ratio} = 1.015^{\Delta nadir}$
  - Estimated hazard ratio for two groups differing by 1 in nadir PSA is found by exponentiating the slope (Stata only reports the hazard ratio)
    - Groups 1 unit higher nadir PSA have instantaneous event rate 1.015

fold higher (1.5% higher)

* Groups 10 units higher nadir PSA have instantaneous event rate $1.015^{10} = 1.166$ fold higher (16.6% higher)

9.6 Inference with PH Regression

- The ideas of Signal and Noise found in simple linear regression do not translate well to PH regression
  - We do not tend to quantify an error distribution with PH regression

- Valid statistical inference (CIs, p-values) about *associations* requires three general assumptions
  - Assumptions about approximately Normal distributions for the parameter estimates
    * Large N
      · Need for either robust standard errors or classical regression

      · Definition of large depends on the underlying probability distribution

  - Assumptions about independence of observations
    * Classical regression: Independence of all observation

    * Robust standard errors: Correlated observations within identified clusters

  - Assumptions about variance of observations within groups
    * Classical regression: Mean-variance relationship for binary data
      · Proportional hazards considers the hazard of event at every time

      · Hence in order to satisfy this requirement, need proportional hazards and linearity of predictor

    - Robust standard errors
        - Allows unequal variance across groups

        - Hence, do not need linearity of predictor or proportional hazards

- Valid statistical inference (CIs, p-values) about *hazard of response in specific groups* requires a further assumption
    - Assumption about the adequacy of the linear model
        - If we are trying to borrow information about the log hazards from neighboring groups, and we are assuming a straight line relationship, the straight line needs to be true

        - Needed for either classical or robust standard errors

        - Note that we can model transformations of the measured predictor

- We rarely make inference about within group survival probabilities using the proportional hazards model
    - Sometimes estimated survival curves are used descriptively
        - Use estimates of the baseline survival function

        - Exponentiate the baseline survival to find survival curve for specific covariate patterns

- Relationship to survival
    - Hazard function determines the survival function

Hazard	$\lambda(t\|X_i) = \lambda_0(t) \times e^{\beta_1 \times X_i}$
Cumulative Hazard	$\Lambda(t\|X_i) = \int_0^t \lambda_0(u) \times e^{\beta_1 \times X_i}\, du$
Survival Function	$S(t\|X_i) = e^{-\Lambda(t\|X_i)} = [S_0(t)]^{e^{\beta_1 \times X_i}}$

9.6.1 Implications for Inference

- The Moral: Regression based inference about *associations* is far more trustworthy than estimation of group odds of responses. Now, the story...

- We will now consider a hierarchy of null hypotheses
  - Strong (and intermediate) null: Total independence of time to event and X
    - The proportional hazards assumption holds because the same distribution in every X group
  - Weak null: No linear trend in hazard ratio across X groups

- Classical PH Regression
  - Inference about the slope tests the strong null
    - All tests make inference by assuming the strong null is true
    - If the data appear non-proportional hazards or nonlinear on the log hazard ratio scale, merely evidence the strong null is not true
  - Limitations
    - We cannot be confident that there is a trend in log HR across groups (valid inference about the trends demands a correct model)
    - We cannot be confident in estimate of group hazards (valid estimates of group probabilities require a correct model)

- Robust Standard Errors
  - Inference about the slope tests the weak null
    - All tests make inference by assuming the weak null is true

    - Data can appear non-proportional hazards or non-linear in hazard ratios across groups
        - Robust SE estimate true variability

        - Nonlinearity decreases precision, but inference about first-order (linear) trends still valid

- So, which inference is correct?
    - Classical PH regression and PH regression with robust standard errors differ in the strength of necessary assumptions

    - As a rule, if all of the assumptions of classical PH regression hold, it will be more precise
        - Hence, we will have the greatest precision to detect associations if the linear model is correct

    - The robust standard error methods are valid for detection of associations without relying on the classical assumptions

- Same moral as other regression methods: Regression based inference about *associations* is far more trustworthy than estimation of group means or individual predictions

- Interpreting "Positive" Results
    - Slope is statistically different from 0 using robust standard errors
        - Observed data is atypical of a setting with no linear trend in hazard ratio across groups

        - Data suggests evidence of a trend toward larger (or smaller) hazards in groups having larger values of the predictor

        - (To the extent the data appears linear, estimates of the group odds or probabilities will be reliable)

- Interpreting “Negative” Results
  - Many possible reasons why the slope is not statistically different from 0 using robust standard errors
    - There may be no association between the response and predictor
    - ~~There may be an association, but not in the parameter considered (the hazard of response)~~
    - There may be an association in the parameter considered, but the best fitting line has zero slope (e.g. a curvilinear association)
    - There may be a first order trend in the log odds, but we lacked the precision to be confident that it truly exists (a type II error)

9.7 Example Revisited

9.7.1 Robust standard errors

```
. stcox nadir, robust

Cox regression -- Breslow method for ties

No. of subjects      =           50                 Number of obs   =        50
No. of failures      =           36
Time at risk         =         1423
                                                    Wald chi2(1)    =     16.79
Log pseudolikelihood =   -113.28918                 Prob > chi2     =    0.0000

------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nadir |    1.01551   .0038139     4.10   0.000     1.008063    1.023013
------------------------------------------------------------------------------
```

“From proportional hazards regression analysis, we estimate that for each 1 ng/ml unit difference in nadir PSA, this risk of relapse is 1.6% higher in the group with the higher nadir PSA. This estimate is highly statistically significant

($p < 0.001$). A 95% CI suggests that this observation is not unusual if a group that has a 1 ng/ml higher nadir PSA might have risk of relapse that was anywhere from 0.8% higher to 2.3% higher than the group with lower nadir PSA."

9.7.2 Log-transformed nadir PSA

- Based on prior experience
  - A constant difference in PSA would not be expected to confer the same increased risk
    - Comparing 4 ng/ml to 10 ng/ml is not the same as comparing 104 ng/ml to 110 ng/ml
  - A multiplicative effect on risk might be better
    - Same increase in risk for each doubling of nadir PSA
    - Achieve this model by using log transformed nadir PSA

```
. generate lnnadir = log(nadir)
. stcox lnnadir, robust

Cox regression -- Breslow method for ties

No. of subjects      =           50                 Number of obs   =        50
No. of failures      =           36
Time at risk         =         1423
                                                    Wald chi2(1)    =     34.04
Log pseudolikelihood =   -107.31899                 Prob > chi2     =    0.0000

------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lnnadir |   1.535399   .1128444     5.83   0.000     1.329419    1.773293
------------------------------------------------------------------------------
```

- Hazard ratio is 1.54 for a e-fold difference in nadir PSA ($e = 2.7183$)
- It is more easy to understand doubling, tripling, 5-fold, 10-fold increases
  - For doubling: $\mathrm{HR} = 1.54^{log(2)} = 1.35$

  - For 5-fold: $\mathrm{HR} = 1.54^{log(5)} = 1.99$

  - Can similarly transform the upper and lower limits of the confidence interval

- The confidence interval and statistical test given in the Stata output is called a Wald test. Other tests (Score, Likelihood Ratio) are also possible.
  - All tests are asymptotically equivalent

  - The Wald test is easiest to obtain, but generally performs the poorest in small sample sizes. It is based on the coefficient estimate and standard error.

  - The Likelihood Ratio test performs the best in small samples. We will discuss it later, including how to obtain the test using post-estimation commands.

  - The Score test is not bad in small samples, but is often hard to obtain from software. It is exactly equal to the logrank test for binary outcomes and categorical predictors.

9.8 Review: Interpretation of Slopes

9.8.1 Additive models

- Identity link function

$$\text{Means} \quad \rightarrow \text{Linear regression}$$

- Interpretation of non-transformed slope: $\theta_X = \beta_0 + \beta_1 \times X$
  - β_1 : (Average) difference in summary measure between groups per 1 unit difference in X

  - $\Delta \times \beta_1$: (Average) difference in summary measure between groups per Δ unit difference in X

- Interpretation with log transformed slope: $\theta_X = \beta_0 + \beta_1 \times \log(X)$
  - $\log(k) \times \beta_1$: (Average) difference in summary measure between groups per k-fold difference in X

9.8.2 Multiplicative models

- Log link function

Geometric means	$\rightarrow$ Linear regression on log scale
Odds	$\rightarrow$ Logistic regression
Hazards	$\rightarrow$ Proportional Hazards (Cox) regression
Rates/Means	$\rightarrow$ Poisson regression
Medians	$\rightarrow$ Accelerated Failure Time regression

- Interpretation of non-transformed slope: $\log(\theta_X) = \beta_0 + \beta_1 \times X$
  - e^{β_1} : (Average) ratio of summary measure between groups per 1 unit difference in X
  - $e^{\Delta \times \beta_1}$: (Average) ratio of summary measure between groups per Δ unit difference in X

- Interpretation with log transformed slope: $\log(\theta_X) = \beta_0 + \beta_1 \times \log(X)$
  - $e^{\log(k) \times \beta_1} = k^{\beta_1}$: (Average) ratio of summary measure between groups per k-fold difference in X