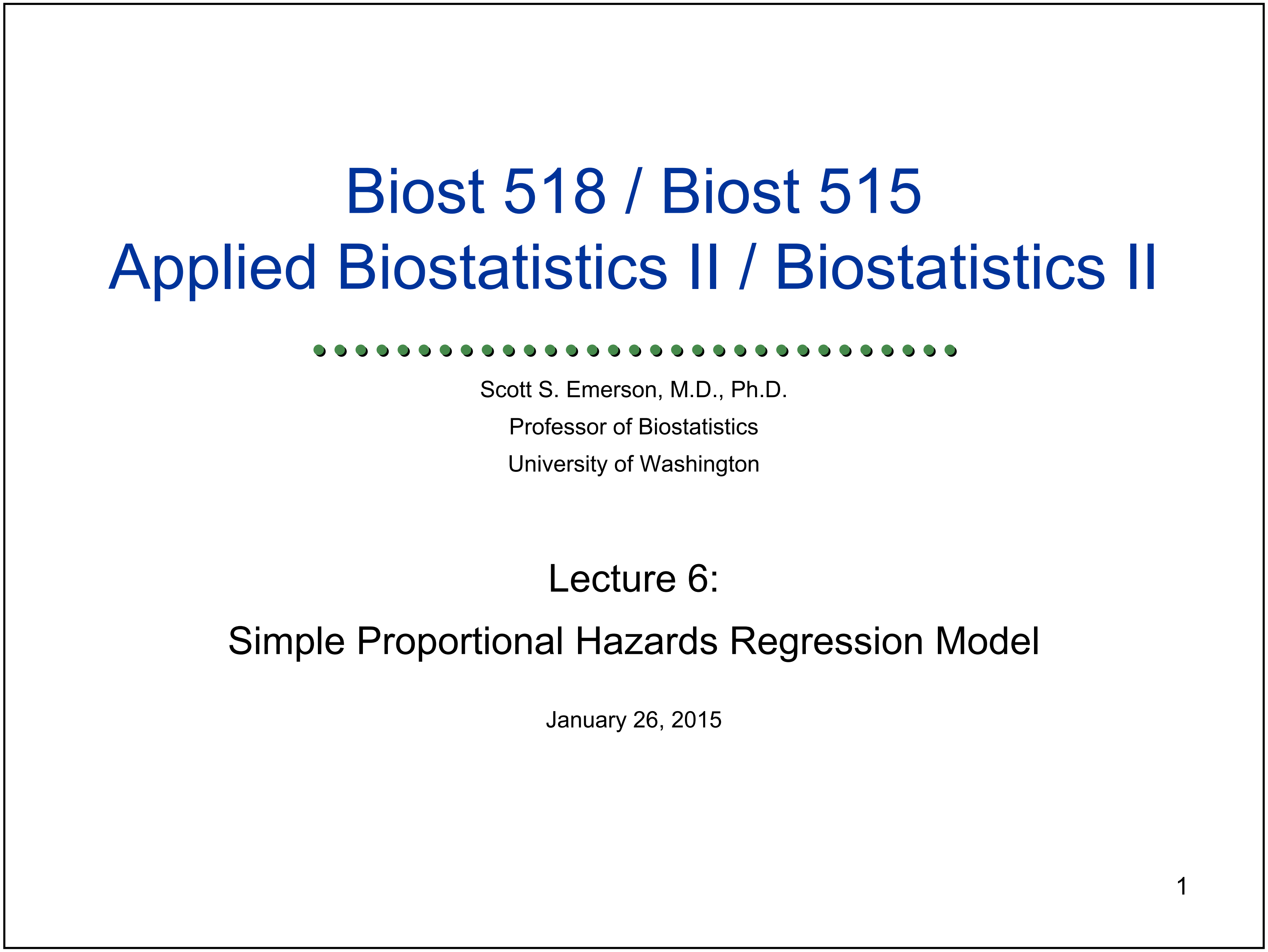


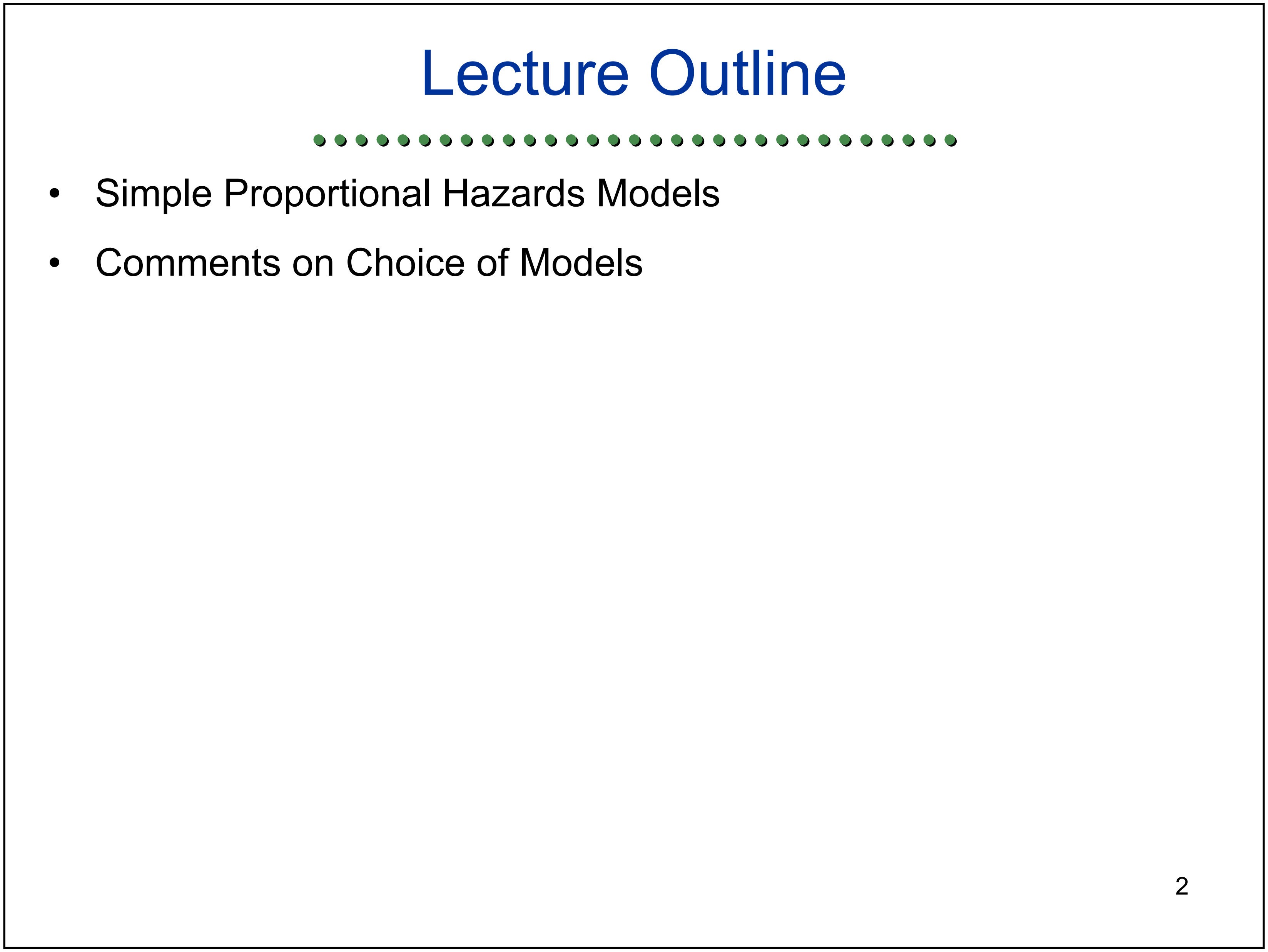
Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 6:
Simple Proportional Hazards Regression Model

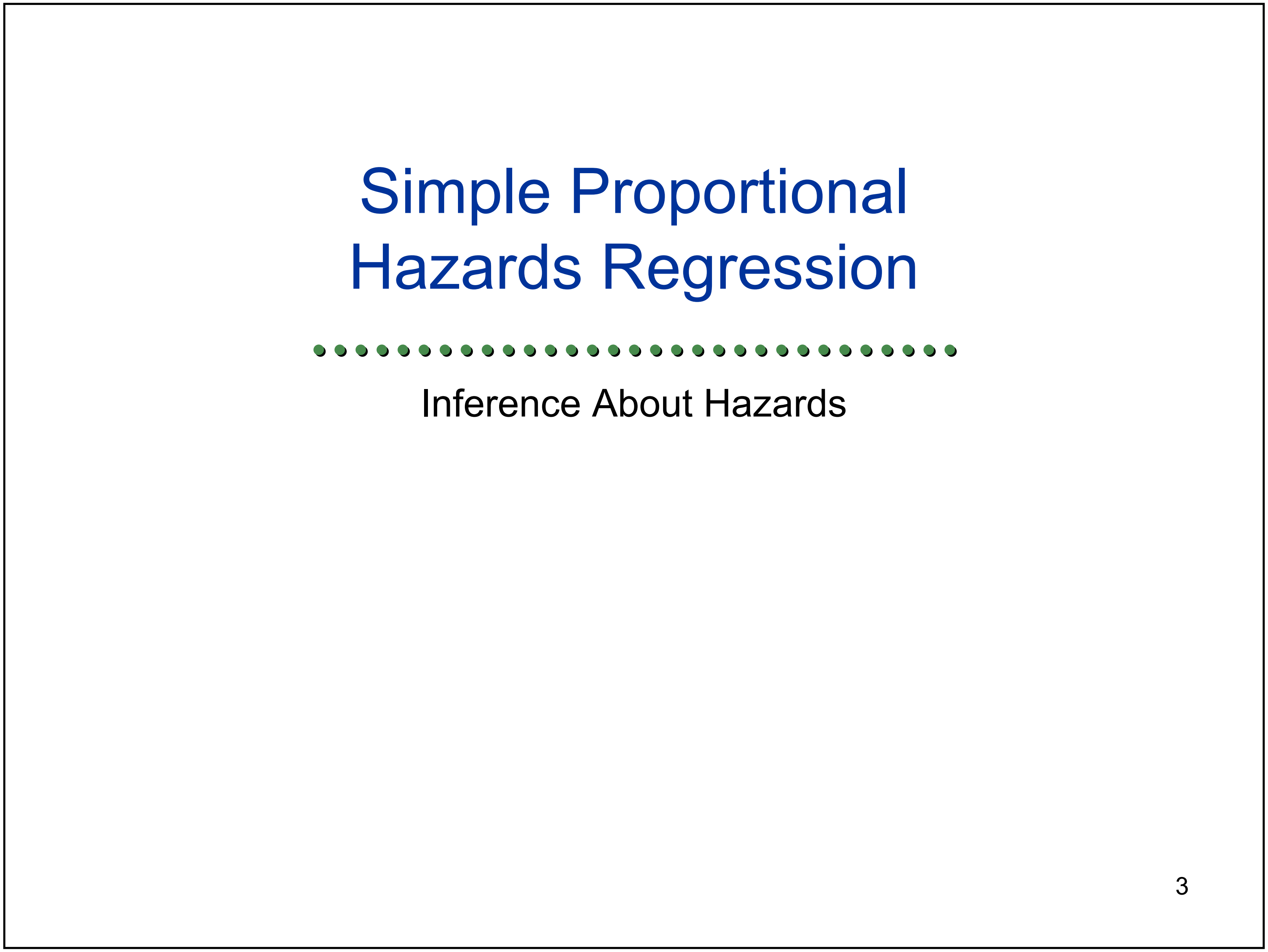
January 26, 2015

1



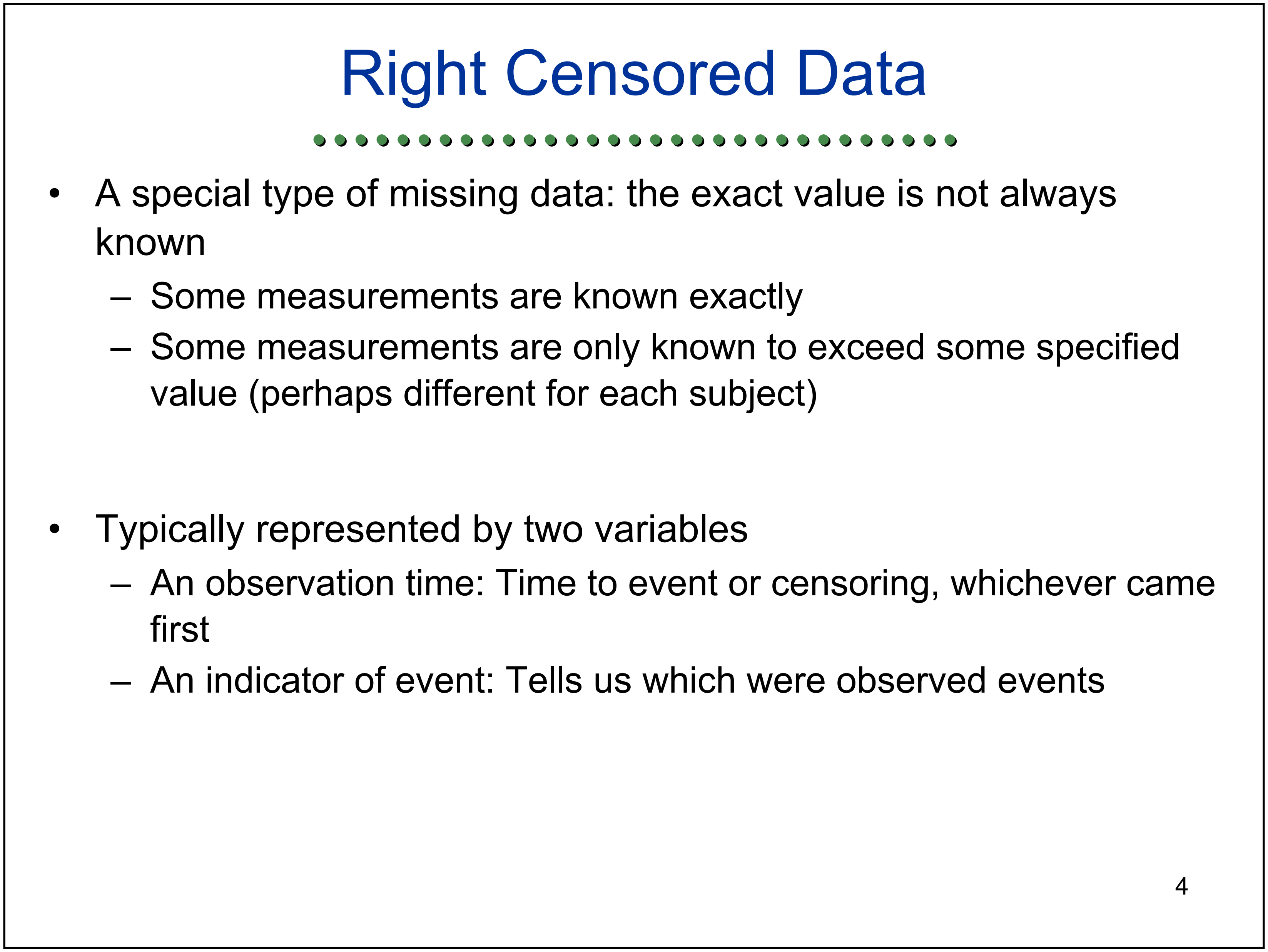
Lecture Outline

- Simple Proportional Hazards Models
- Comments on Choice of Models



Simple Proportional Hazards Regression

Inference About Hazards



Right Censored Data

- A special type of missing data: the exact value is not always known
  - Some measurements are known exactly
  - Some measurements are only known to exceed some specified value (perhaps different for each subject)

- Typically represented by two variables
  - An observation time: Time to event or censoring, whichever came first
  - An indicator of event: Tells us which were observed events

Notation

Unobserved :

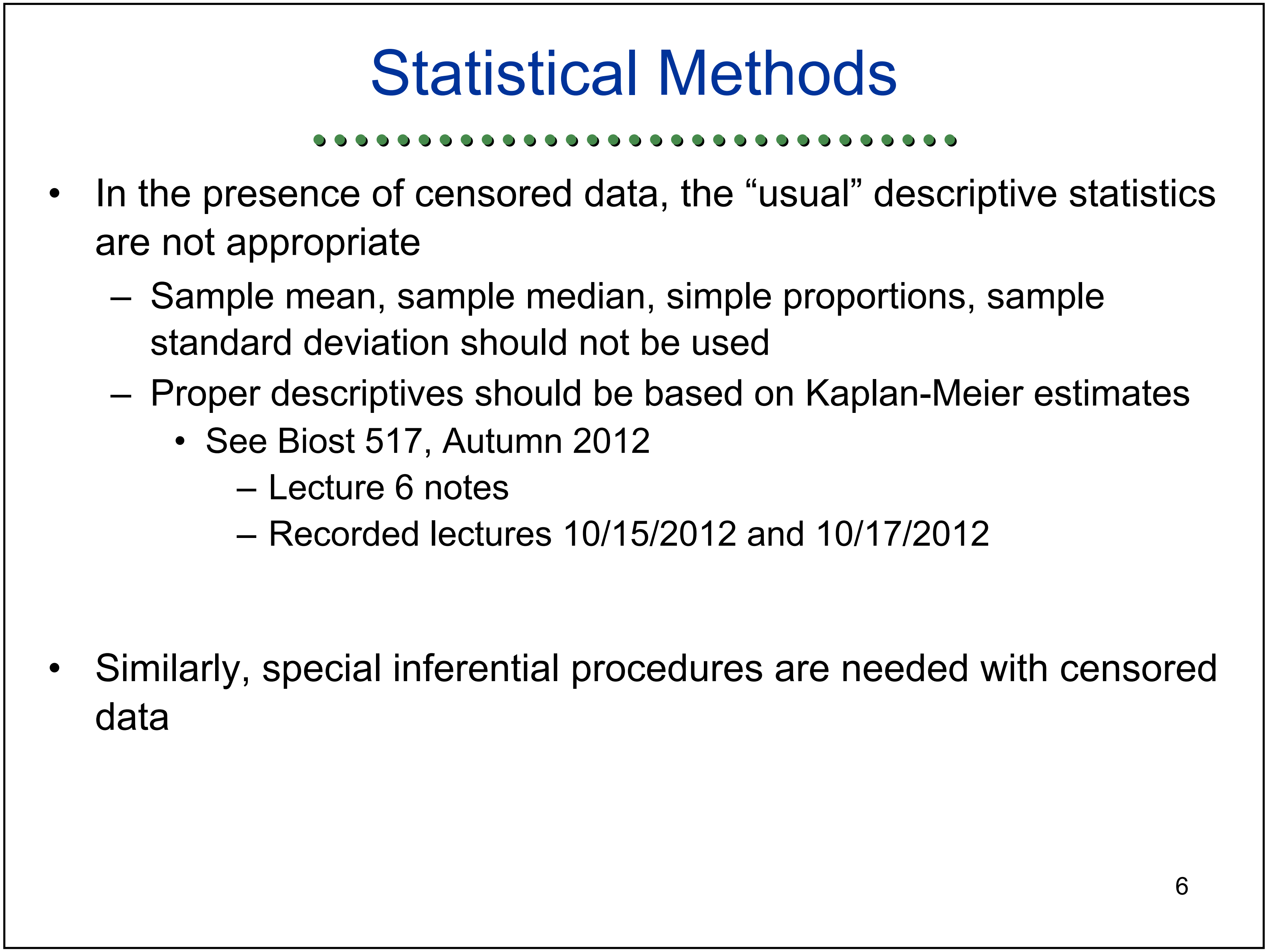
True times to event : $\left\{T_1^0, T_2^0, \ldots, T_n^0\right\}$

Censoring Times : $\left\{C_1, C_2, \ldots, \mathrm{C}_n\right\}$

Observed data :

Observation Times : $T_i = \min\left(T_i^0, C_i\right)$

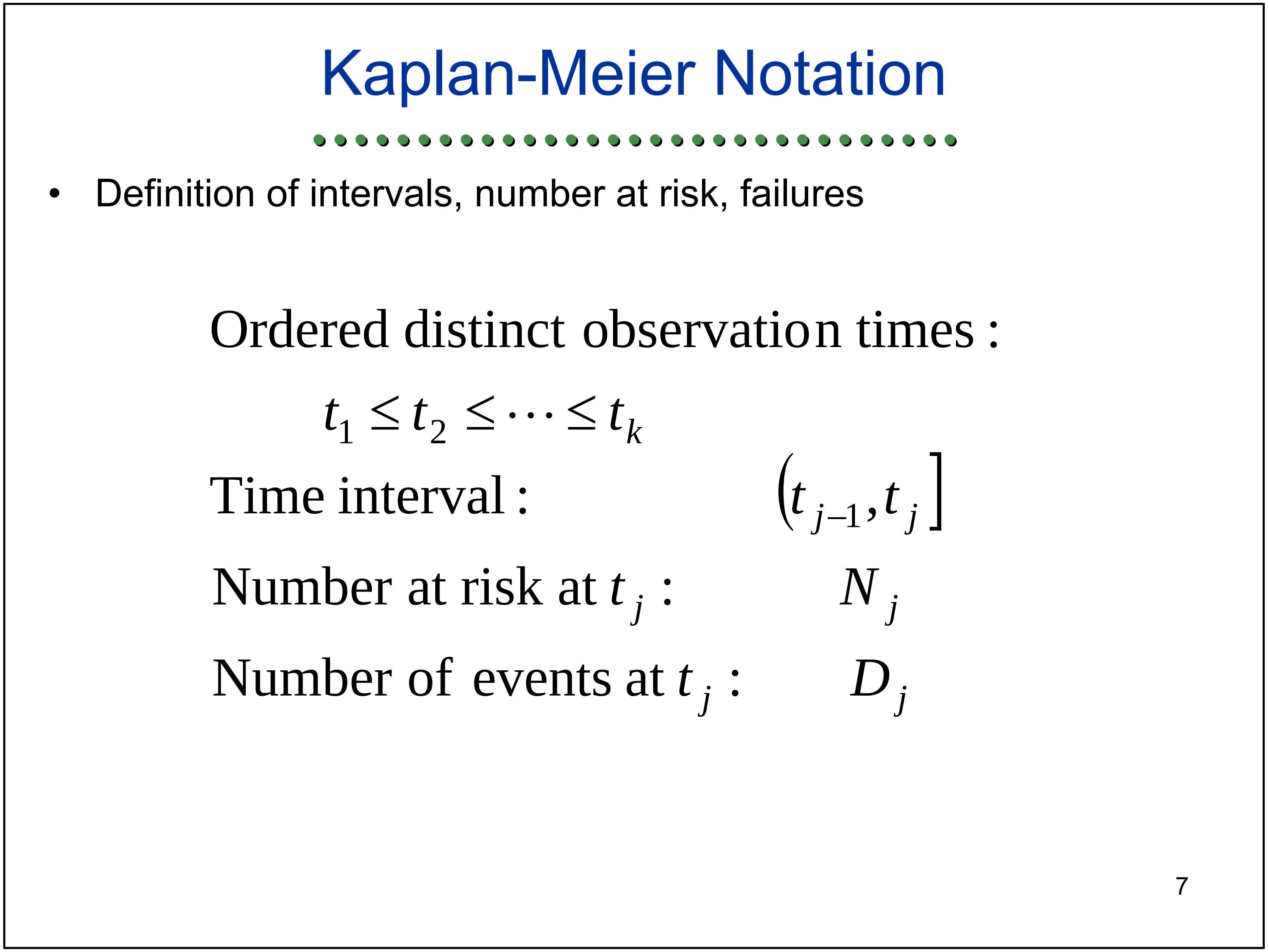
Event indicators : $\mathrm{D}_i = \begin{cases} 1 & \text{if } T_i = T_i^0 \\ 0 & \text{otherwise} \end{cases}$



Statistical Methods

- 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 Kaplan-Meier estimates
    - See Biost 517, Autumn 2012
      - Lecture 6 notes
      - Recorded lectures 10/15/2012 and 10/17/2012

- Similarly, special inferential procedures are needed with censored data



Kaplan-Meier Notation

- Definition of intervals, number at risk, failures

Ordered distinct observation times :

$$t_1 \leq t_2 \leq \cdots \leq t_k$$

Time interval : $\left(t_{j-1}, t_j\right]$

Number at risk at t_j : N_j

Number of events at t_j : D_j

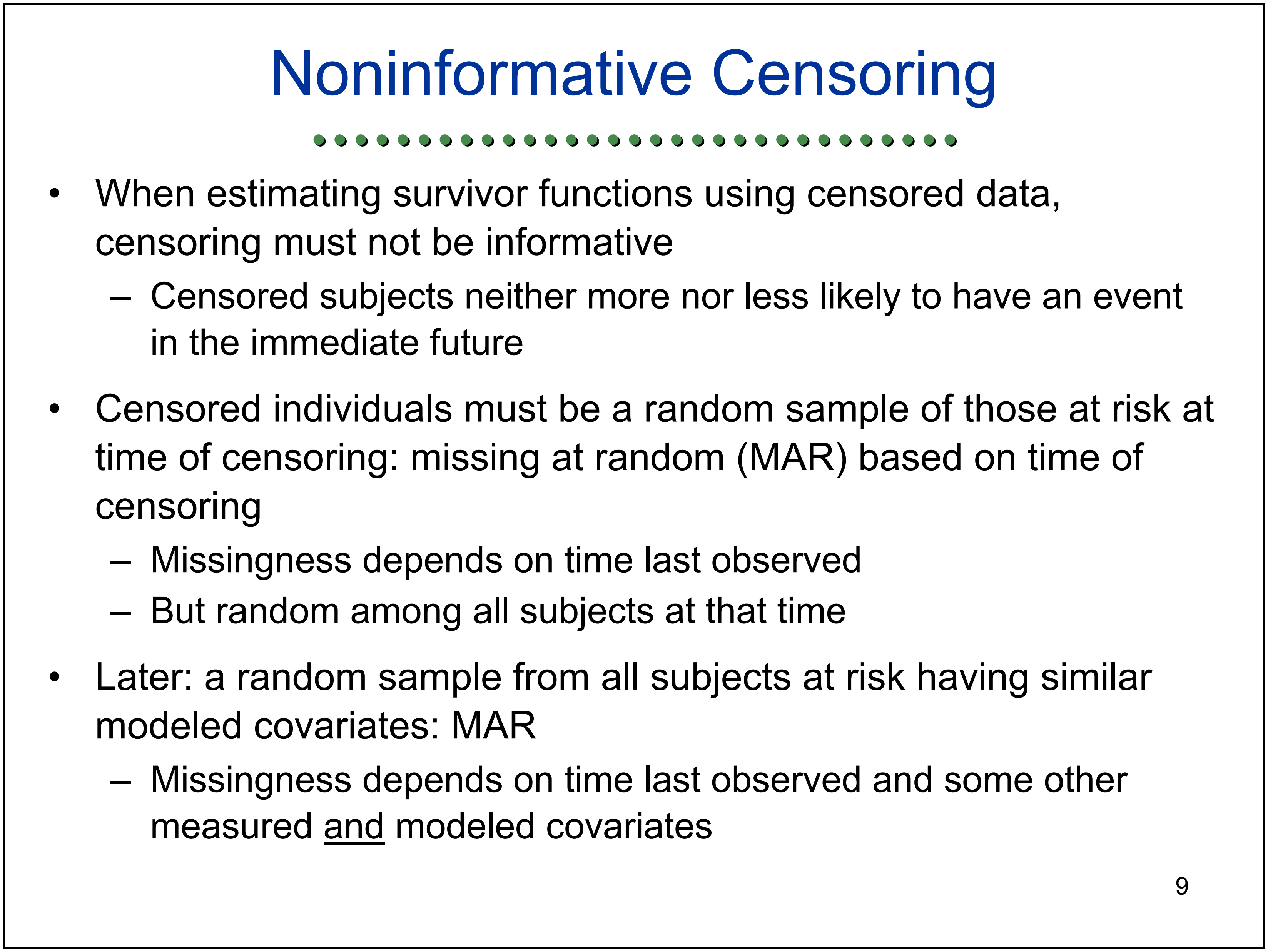
Kaplan-Meier Hazard Estimates

- Computation of hazard and conditional probability of survival in interval based on
  - number at risk at the beginning of the interval, and
  - number having an event during the interval

$$\text{Hazard for event in interval:} \qquad \frac{D_j}{N_j}$$

$$\text{Conditional probability of survival in interval:}$$

$$\Pr\left(T^0 \geq t_j \mid T^0 \geq t_{j-1}\right) = 1 - \frac{D_j}{N_j}$$

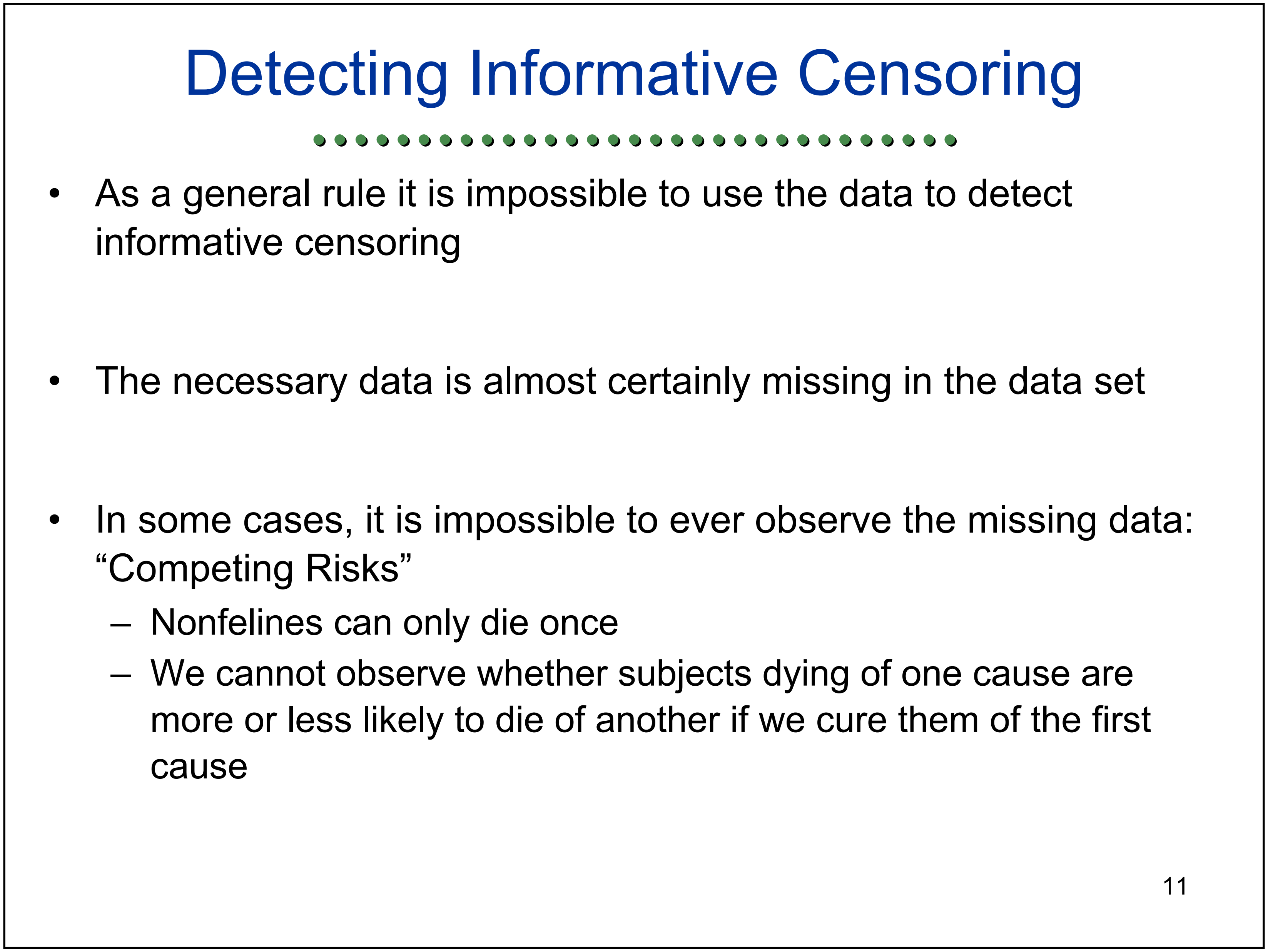


Noninformative Censoring

- When estimating survivor functions using censored data, censoring must not be informative
  - Censored subjects neither more nor less likely to have an event in the immediate future
- Censored individuals must be a random sample of those at risk at time of censoring: missing at random (MAR) based on time of censoring
  - Missingness depends on time last observed
  - But random among all subjects at that time
- Later: a random sample from all subjects at risk having similar modeled covariates: MAR
  - Missingness depends on time last observed and some other measured and modeled covariates

Informative Censoring Examples

- Subjects in a RCT are withdrawn due to treatment failure
  - (likely they would die sooner than those remaining)
- Subjects in a RCT in a fatal condition are lost to follow up when they go on vacation
  - (likely they are healthier than those remaining)
- Leukemia patients in a RCT of bone marrow transplantation are censored if they die of infections rather than dying of cancer
  - (they might have had a more effective regimen to wipe out existing cancer)



Detecting Informative Censoring

- As a general rule it is impossible to use the data to detect informative censoring

- The necessary data is almost certainly missing in the data set

- In some cases, it is impossible to ever observe the missing data: “Competing Risks”
  - Nonfelines can only die once
  - We cannot observe whether subjects dying of one cause are more or less likely to die of another if we cure them of the first cause

Kaplan-Meier Survival Estimate

- Estimating survival probability with noninformative censoring

$$S(t) = Pr\ (T^0 > t)$$

$$\text{Cumulative probability of survival:}$$

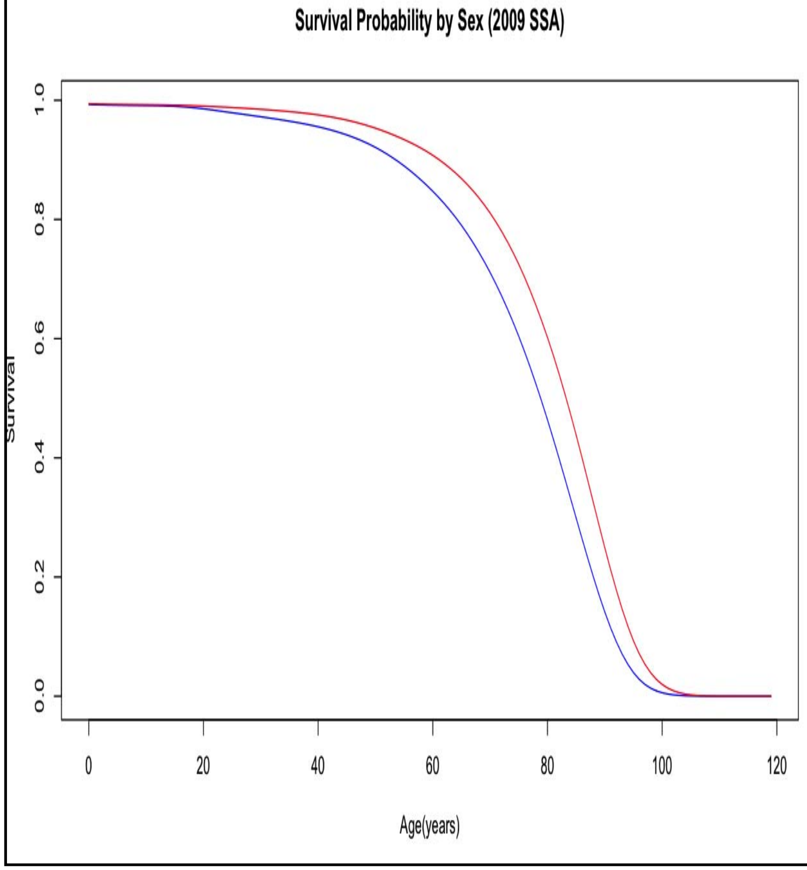
$$\Pr\left(T^0 > t_j\right) = \Pr\left(T^0 > t_j \mid T^0 > t_{j-1}\right)\Pr\left(T^0 > t_{j-1}\right)$$

$$\hat{S}\left(t_j\right) = \left(1 - \frac{D_j}{N_j}\right) \times \left(1 - \frac{D_{j-1}}{N_{j-1}}\right) \times \cdots \times \left(1 - \frac{D_1}{N_1}\right)$$

$$= \prod_{i=1}^{j}\left(1 - \frac{D_i}{N_i}\right)$$

Comparing Survival Curves: S(t) vs t

- Commonly used summary measures can be seen in a plot of survival curves



- Difference in survival at t_0
  - Vertical separation at t_0
- Difference in quantiles
  - Horizontal separation at p
- Difference in means
  - Area between curves
- Hazard
  - Slope divided by height of the curve
  - (not exactly clear)

13

Survival Regression

- There are two fundamental models used to describe the way that some factor might affect time to event

- Accelerated failure time
  - Models quantiles of a distribution on a multiplicative scale
    - Usually parametric models: exponential, Weibull, log logistic…
  - Exponentiated slopes are median ratios

- Proportional Hazards
  - Models hazard function of a distribution on a multiplicative scale
    - Usually semi-parametric model (but also parametric Weibull)
  - Exponentiated slopes are hazard ratios

Accelerated Failure Time Model

- Assume factors causes subjects to spend their lifetime too fast

- The basic idea: For every year in a reference group's lives, the other group "ages" k years
  - E.g.: 1 human year = 7 dog years

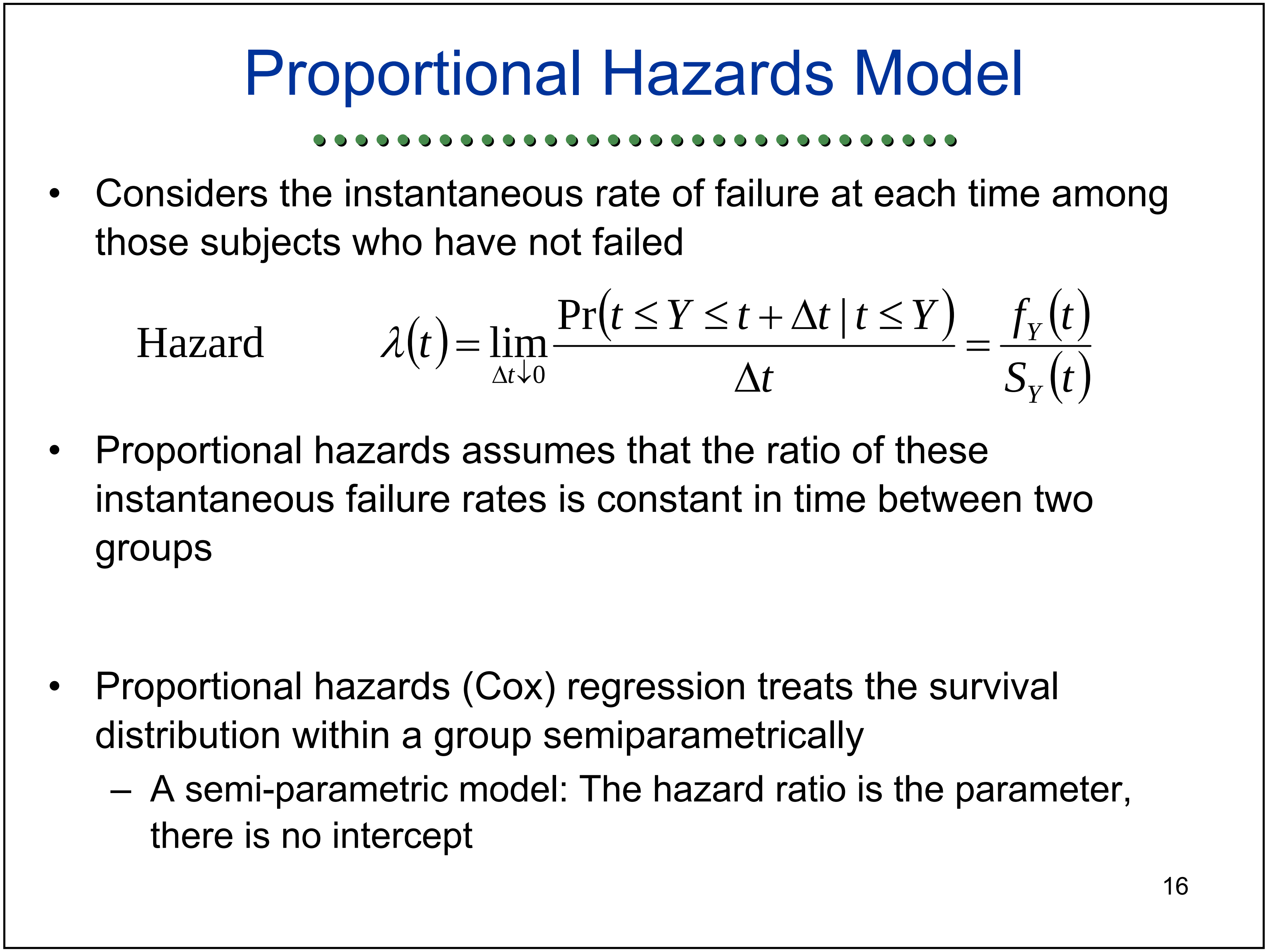
- Ratios of distribution quantiles are constant across two group
  - E.g., report median ratios

- AFT models include parametric exponential, Weibull, lognormal
  - "Error" distribution often does not have mean 0

$$\log(Y_i) = \beta_0 + \beta_1 \times X_i + \cdots \beta_p \times W_{(p-1)i} + \sigma\varepsilon_i$$

$$\varepsilon_i \sim \text{some parametric distribution (usually)}$$

15



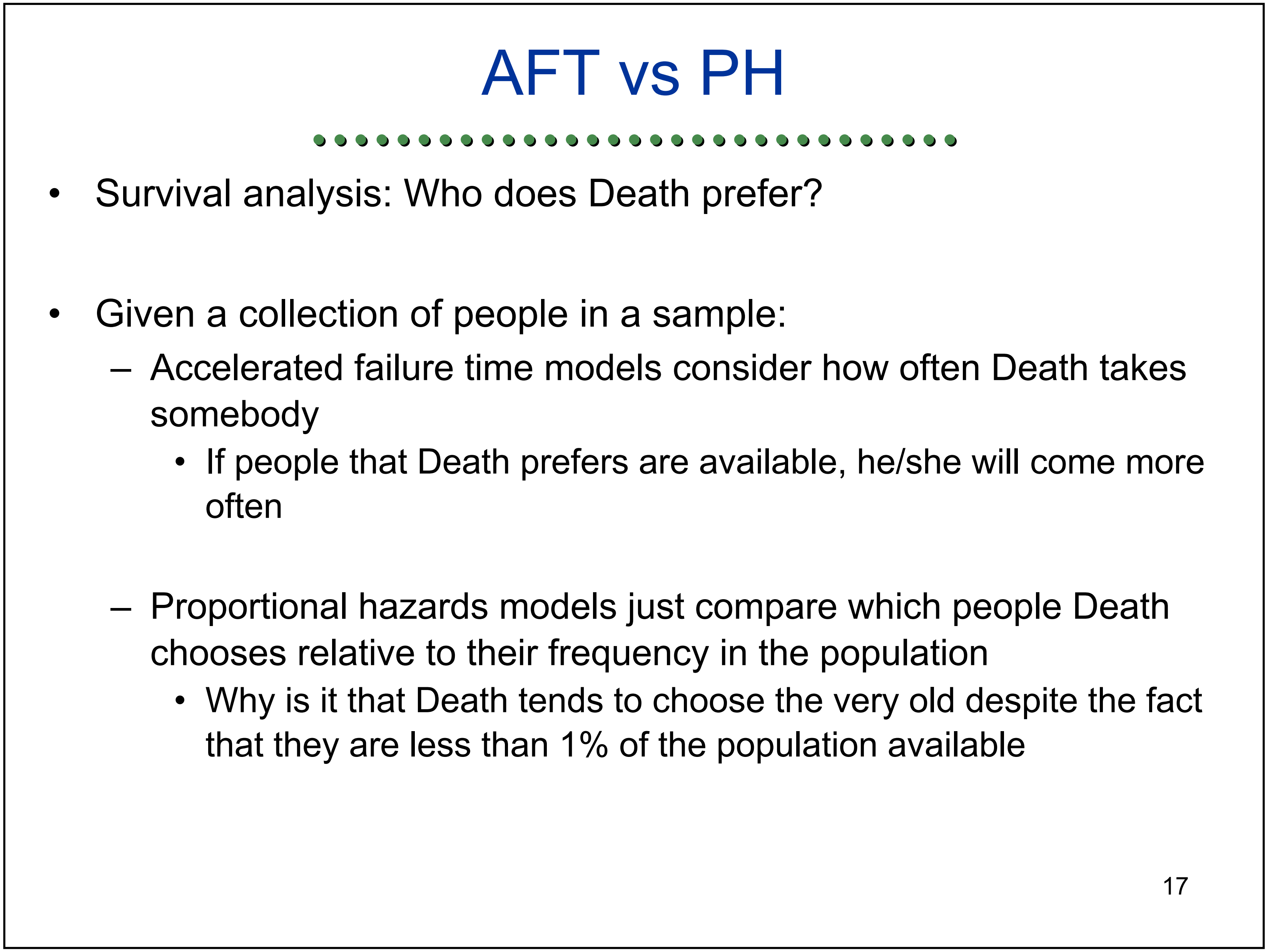
Proportional Hazards Model

- Considers the instantaneous rate of failure at each time among those subjects who have not failed

Hazard $$\lambda(t)=\lim_{\Delta t\downarrow 0}\frac{\Pr\left(t\le Y\le t+\Delta t\,|\,t\le Y\right)}{\Delta t}=\frac{f_Y(t)}{S_Y(t)}$$

- Proportional hazards assumes that the ratio of these instantaneous failure rates is constant in time between two groups

- Proportional hazards (Cox) regression treats the survival distribution within a group semiparametrically
  - A semi-parametric model: The hazard ratio is the parameter, there is no intercept

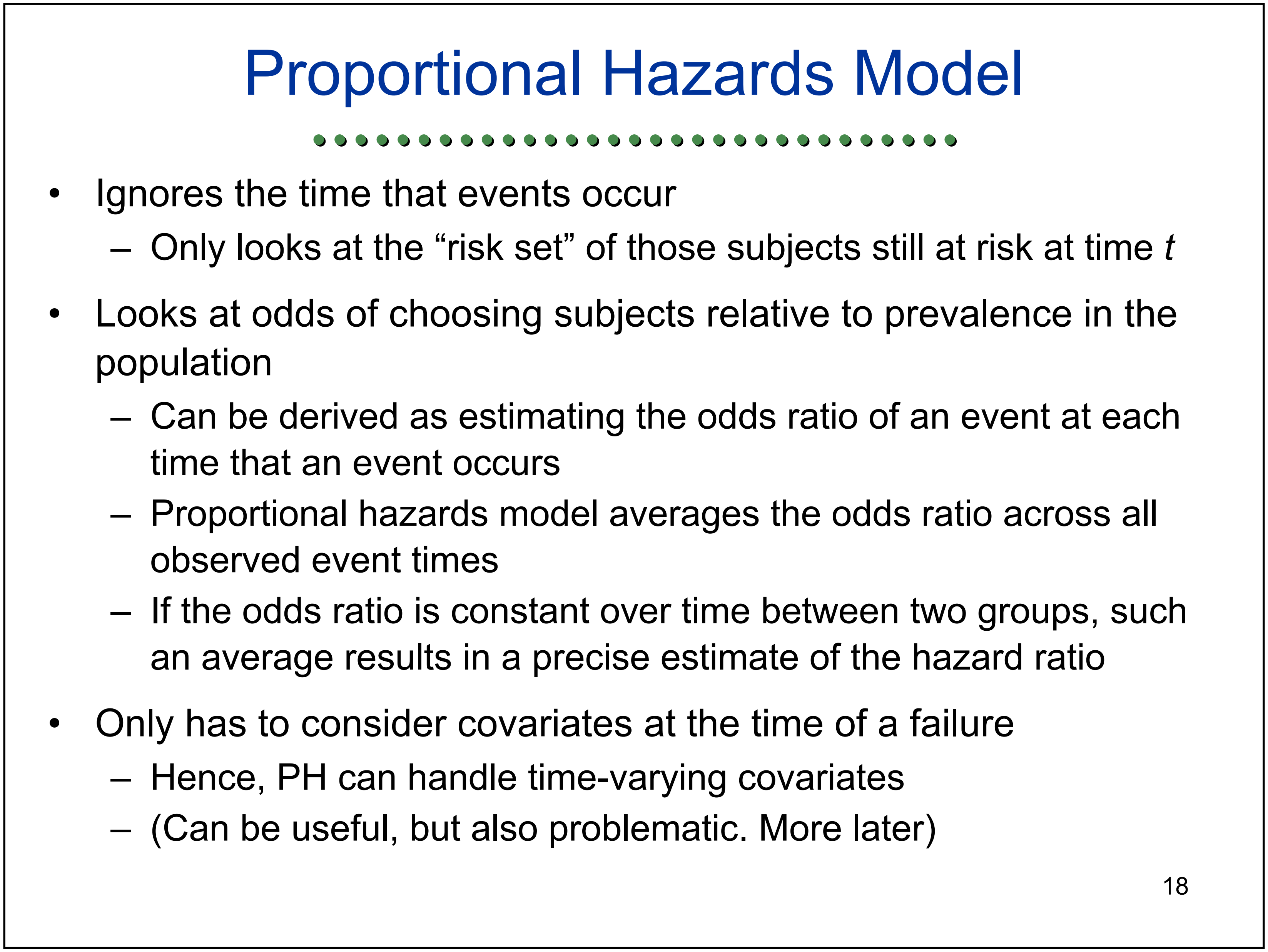


AFT vs PH

- Survival analysis: Who does Death prefer?

- Given a collection of people in a sample:
  - Accelerated failure time models consider how often Death takes somebody
    - If people that Death prefers are available, he/she will come more often

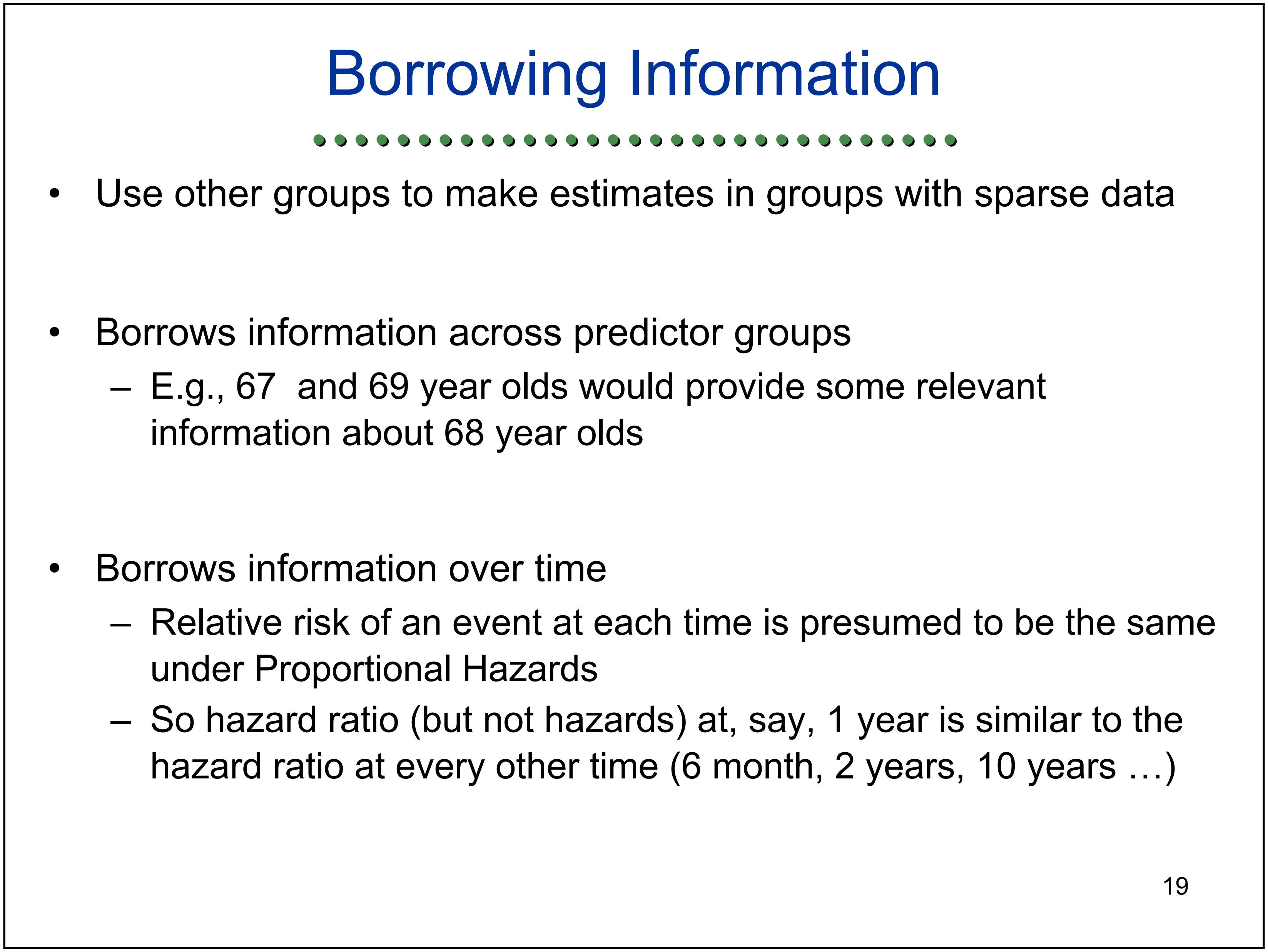
  - Proportional hazards models just compare which people Death chooses relative to their frequency in the population
    - Why is it that Death tends to choose the very old despite the fact that they are less than 1% of the population available



Proportional Hazards Model

- Ignores the time that events occur
  - Only looks at the “risk set” of those subjects still at risk at time *t*
- 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
- Only has to consider covariates at the time of a failure
  - Hence, PH can handle time-varying covariates
  - (Can be useful, but also problematic. More later)

18



Borrowing Information

- Use other groups to make estimates in groups with sparse data

- Borrows information across predictor groups
  - E.g., 67 and 69 year olds would provide some relevant information about 68 year olds

- Borrows information over time
  - Relative risk of an event at each time is presumed to be the same under Proportional Hazards
  - So hazard ratio (but not hazards) at, say, 1 year is similar to the hazard ratio at every other time (6 month, 2 years, 10 years …)

Simple PH Regression Model

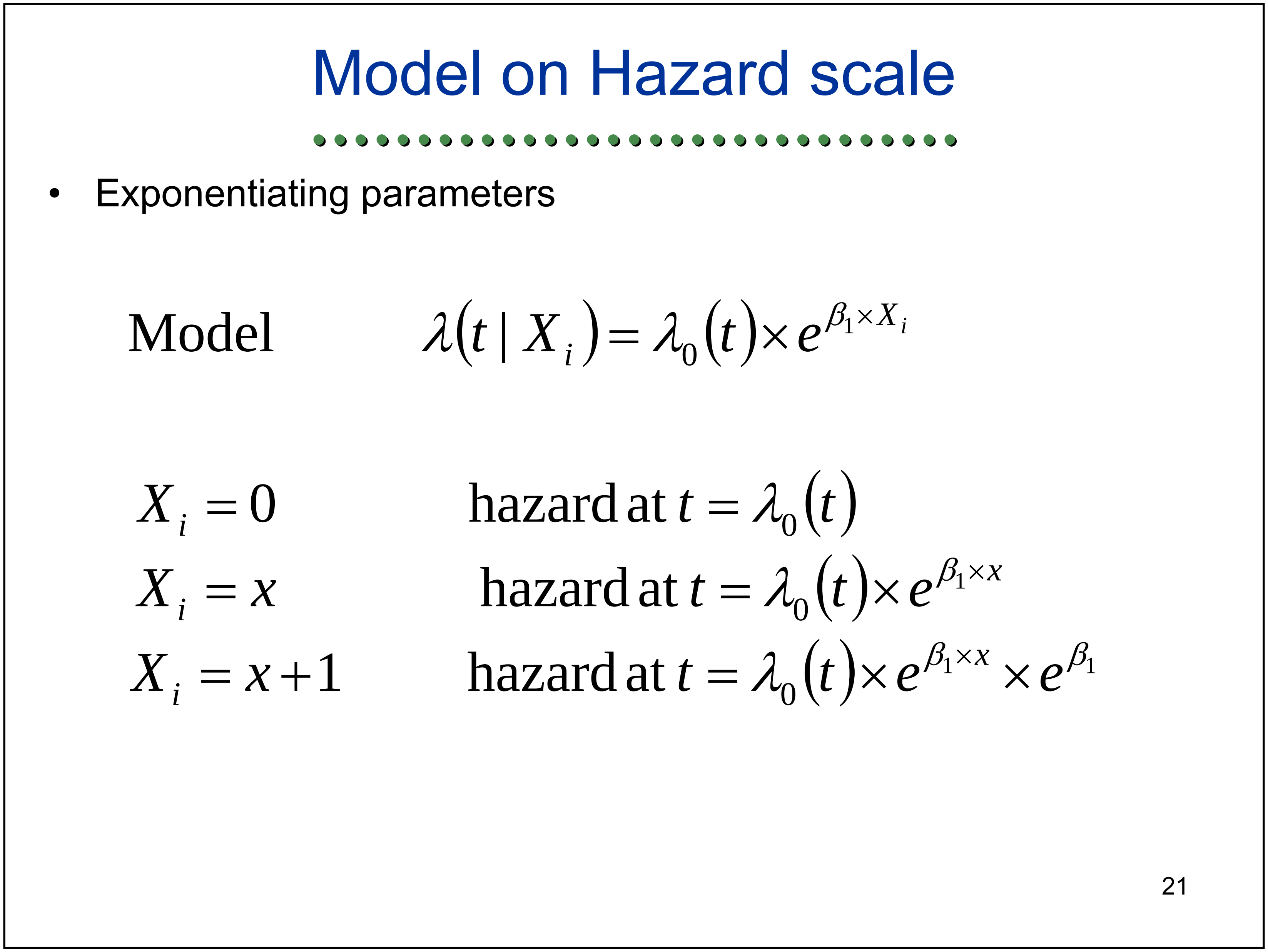
- “Baseline” hazard function is unspecified
  - Similar to an intercept

Model $\log\left(\lambda(t \mid X_i)\right) = \log\left(\lambda_{i0}(t)\right) + \beta_1 \times X_i$

$X_i = 0$ $\log \text{ hazard at } t = \log\left(\lambda_0(t)\right)$

$X_i = x$ $\log \text{ hazard at } t = \log\left(\lambda_0(t)\right) + \beta_1 \times x$

$X_i = x + 1$ $\log \text{ hazard at } t = \log\left(\lambda_0(t)\right) + \beta_1 \times x + \beta_1$



Model on Hazard scale

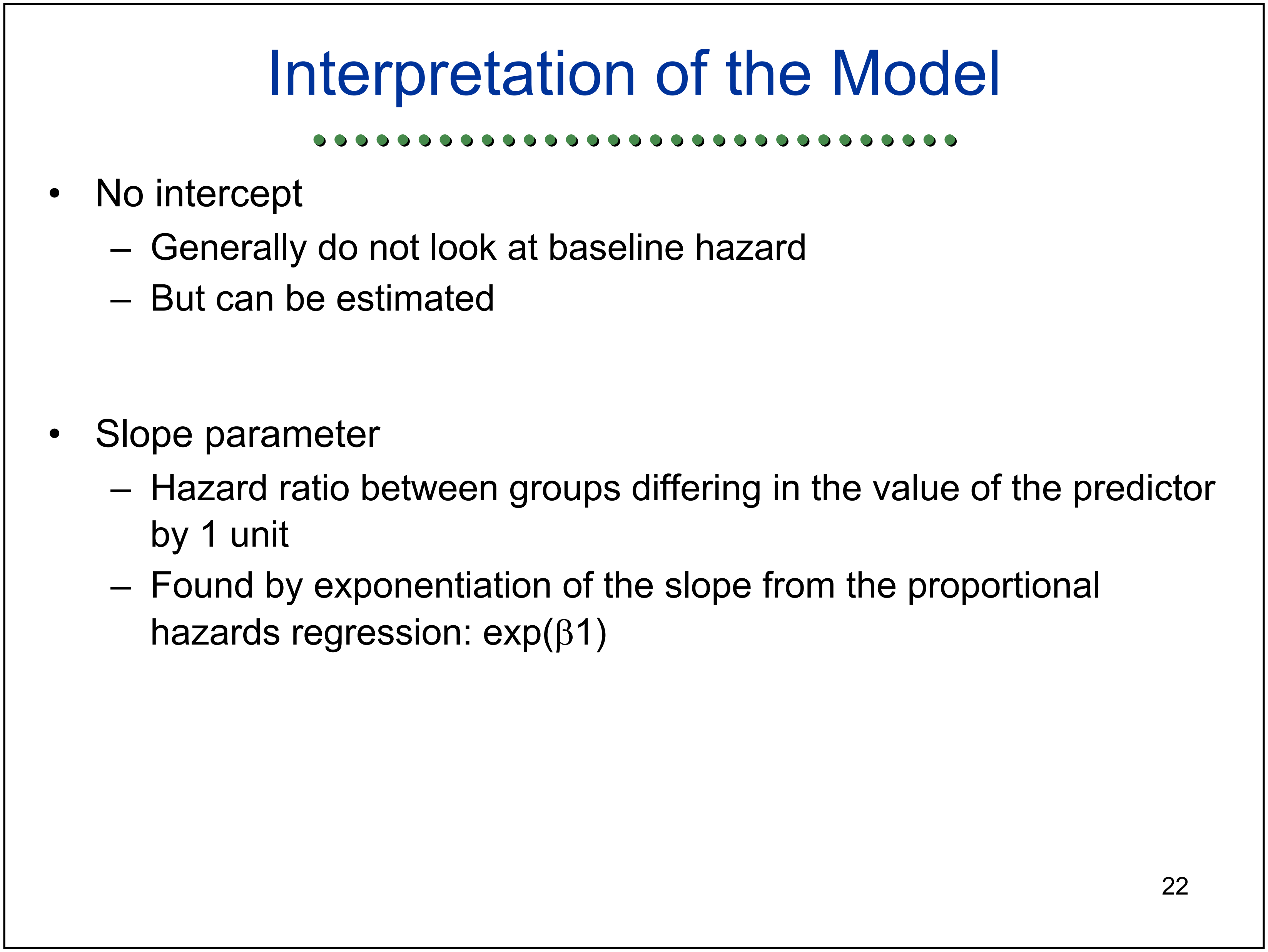
- Exponentiating parameters

Model $$\lambda(t \mid X_i) = \lambda_0(t) \times e^{\beta_1 \times X_i}$$

$$X_i = 0 \qquad \text{hazard at } t = \lambda_0(t)$$

$$X_i = x \qquad \text{hazard at } t = \lambda_0(t) \times e^{\beta_1 \times x}$$

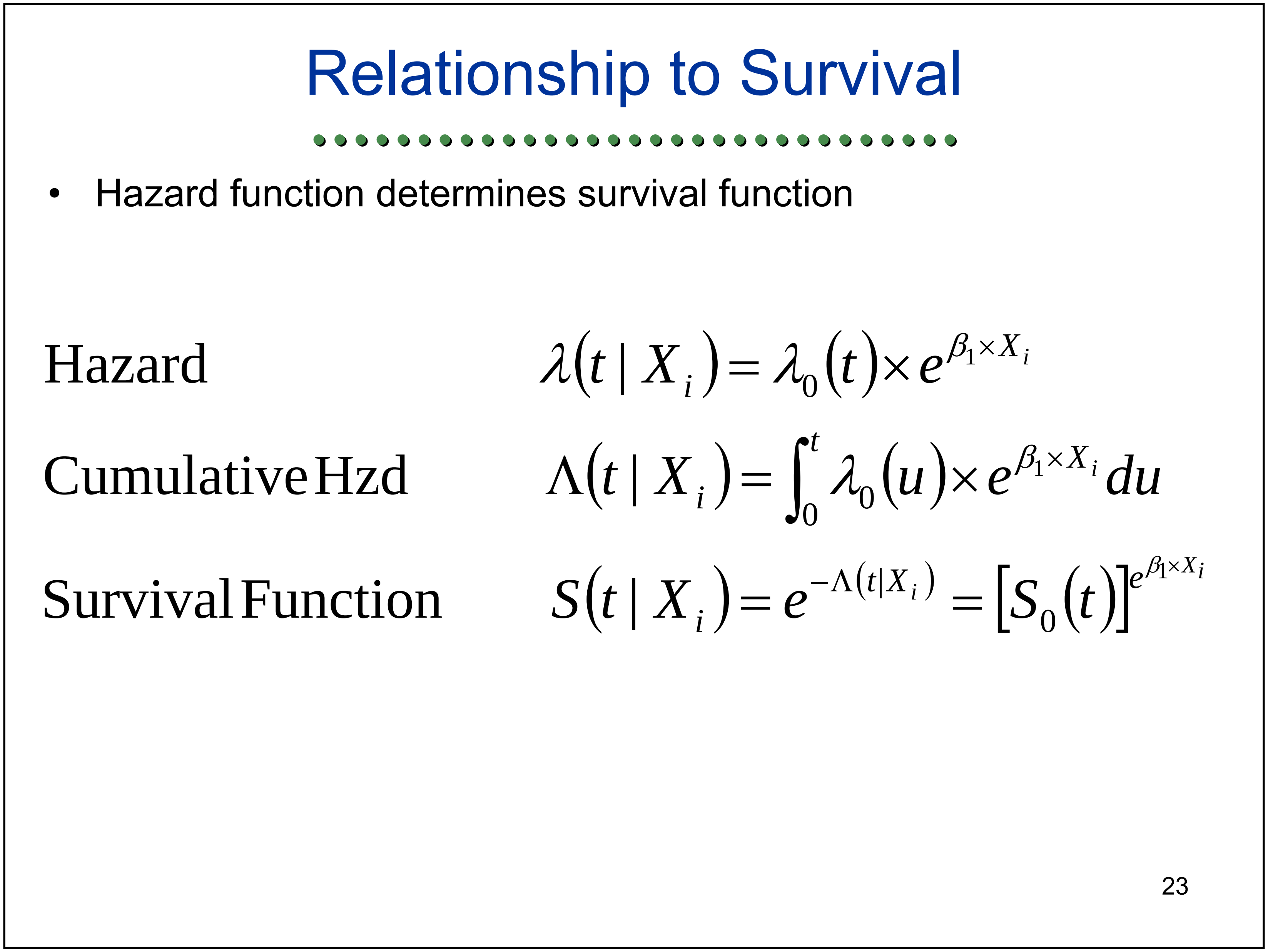
$$X_i = x + 1 \qquad \text{hazard at } t = \lambda_0(t) \times e^{\beta_1 \times x} \times e^{\beta_1}$$



Interpretation of the Model

- No intercept
  - Generally do not look at baseline hazard
  - But can be estimated

- Slope parameter
  - Hazard ratio between groups differing in the value of the predictor by 1 unit
  - Found by exponentiation of the slope from the proportional hazards regression: exp(β1)



Relationship to Survival

- Hazard function determines survival function

Hazard $\lambda\left(t \mid X_i\right) = \lambda_0\left(t\right) \times e^{\beta_1 \times X_i}$

Cumulative Hzd $\Lambda\left(t \mid X_i\right) = \int_0^t \lambda_0\left(u\right) \times e^{\beta_1 \times X_i} du$

Survival Function $S\left(t \mid X_i\right) = e^{-\Lambda\left(t \mid X_i\right)} = \left[S_0\left(t\right)\right]^{e^{\beta_1 \times X_i}}$

Stata Commands

- Same general idea as for other regression models, but:
  - Because it takes two variables to specify a censored observation, Stata makes you first declare those two variables
    - **`stset`** ***`obstime eventind`***
  - Then you do not have to specify the response variable when you execute the regression command
    - `stcox predvar,[robust] [nohr]`
  - (Note that R has you specify a “survival object” that has both variables in it)
- By default, Stata reports estimates on the hazard ratio scale
  - Specifying the option `nohr` will cause Stata to print log HR
  - (Intercept is the “baseline hazard function” and never printed with the standard regression output, though it can be estimated)

Similarity to Other Regressions

- Proportional hazards regression uses *maximum partial likelihood estimation* to find parameter estimates
- There is no real concept of a “saturated model”, because we are always borrowing information across time
- In large samples, the regression parameter estimates are approximately normally distributed
  - P values and CI that are displayed for each parameter estimate are Wald- based estimates

$$95\%\,\mathrm{CI}:\ (estimate) \pm (crit\ value) \times (std\ err) \qquad \hat{\beta} \quad \pm \quad z_{1-\alpha/2} \times s\hat{e}\left(\hat{\beta}\right)$$

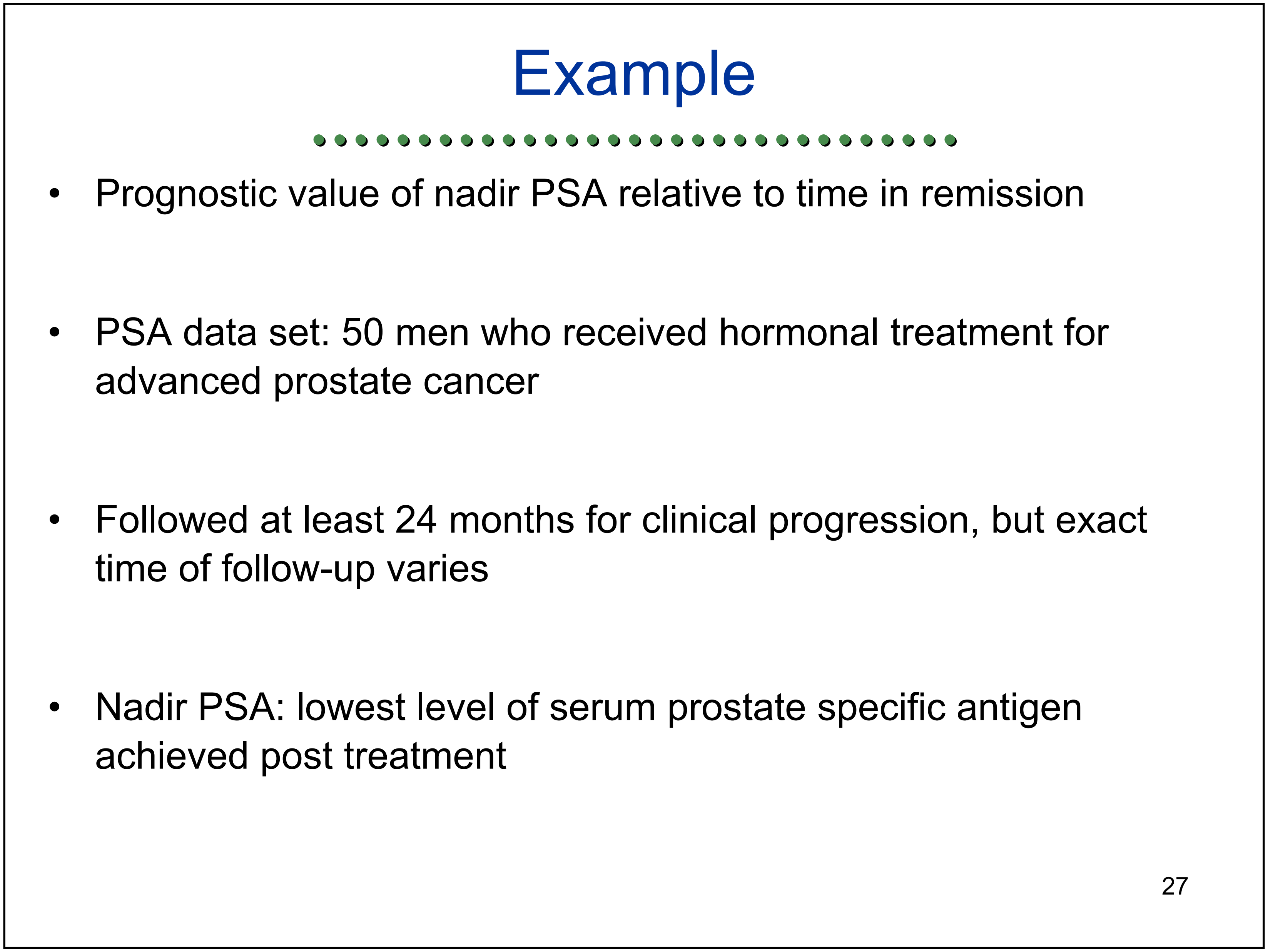
$$\text{Test stat}: \qquad Z = \frac{(estimate)-(null)}{(std\ err)} \qquad\qquad Z = \frac{\hat{\beta}-\beta_0}{s\hat{e}\left(\hat{\beta}\right)}$$

Technical Details

- Unlike linear regression, there is no closed form expression to find the PH regression parameter estimates

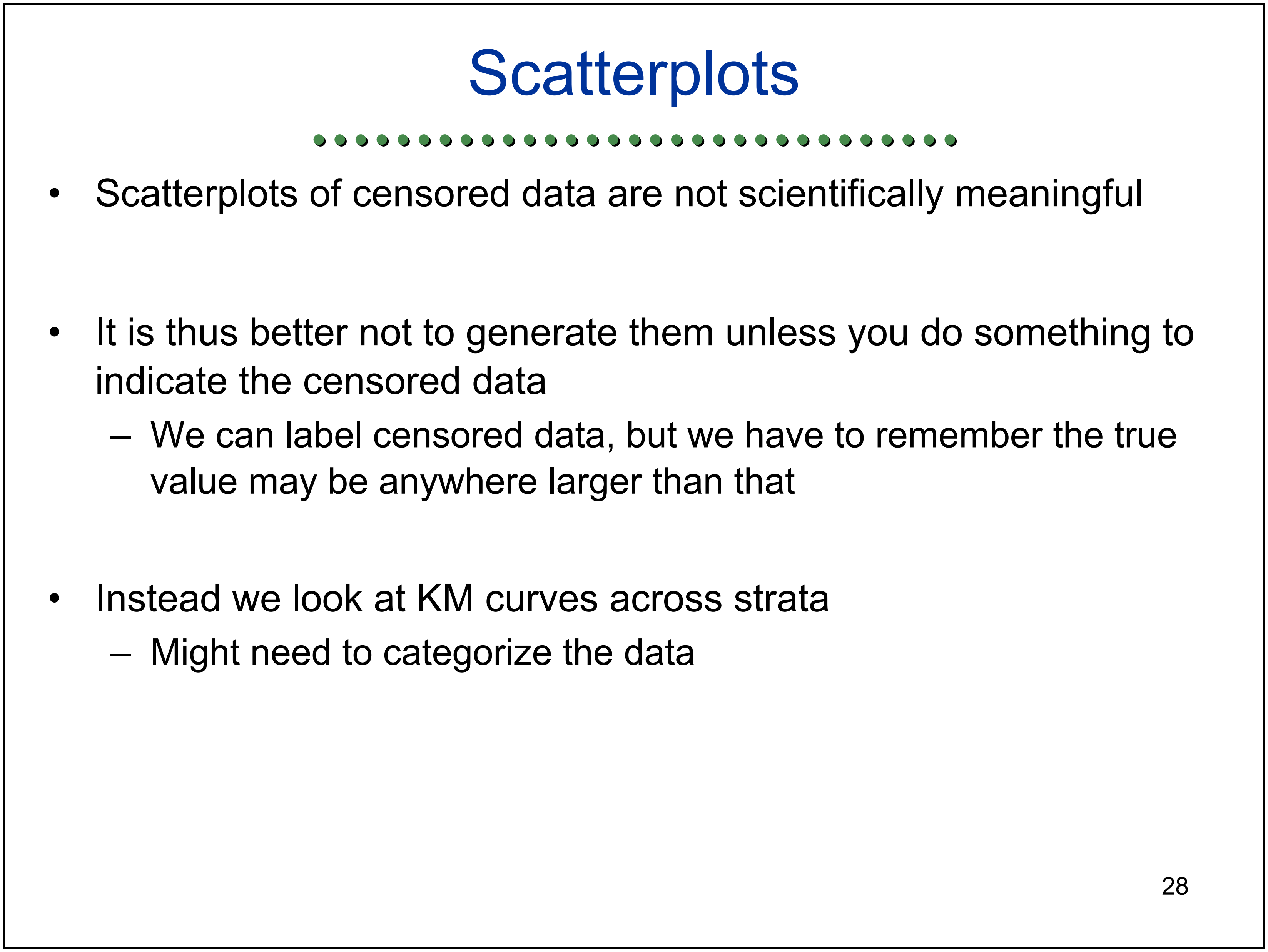
- Instead, computer programs use an iterative search

- This search can fail in settings if some parameter corresponds to comparisons in which one group all has events prior to the other group having any events
  - In this setting, PH regression parameters modeling the log HR are trying to estimate positive or negative infinity
  - The sample size is too small for the model



Example

- Prognostic value of nadir PSA relative to time in remission
- PSA data set: 50 men who received hormonal treatment for advanced prostate cancer
- Followed at least 24 months for clinical progression, but exact time of follow-up varies
- Nadir PSA: lowest level of serum prostate specific antigen achieved post treatment



Scatterplots

- Scatterplots of censored data are not scientifically meaningful

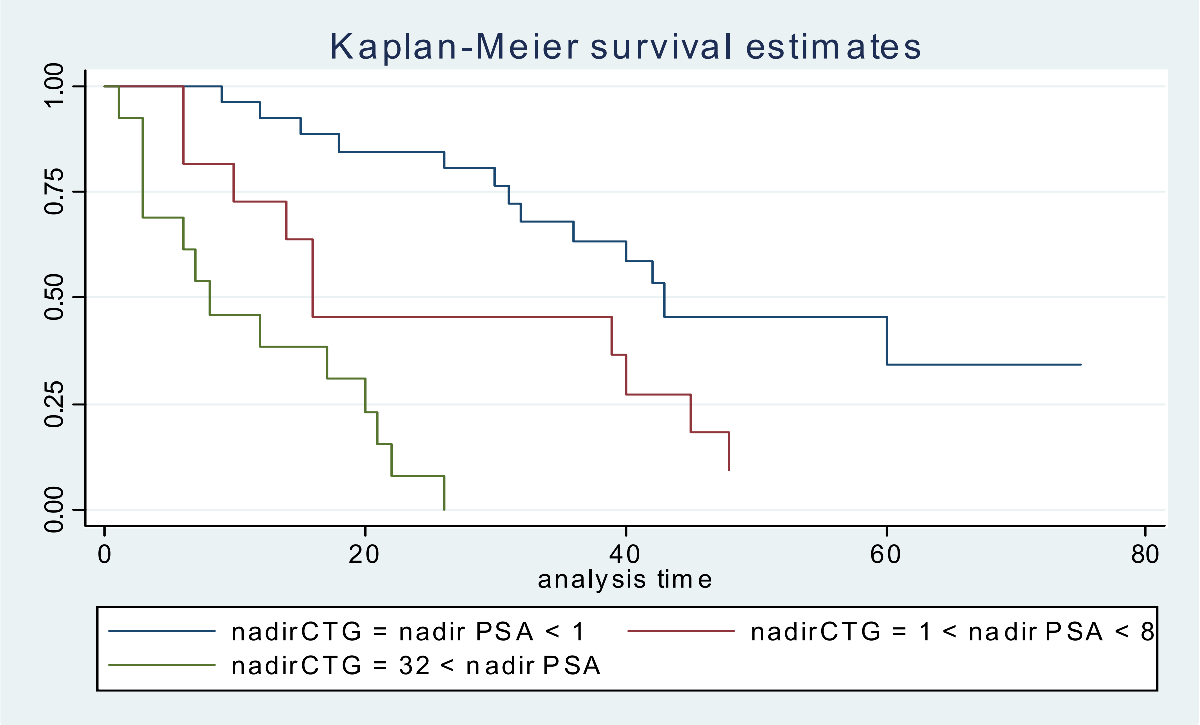
- It is thus better not to generate them unless you do something to indicate the censored data
  - We can label censored data, but we have to remember the true value may be anywhere larger than that

- Instead we look at KM curves across strata
  - Might need to categorize the data

Stratified Kaplan-Meier Plots

- Kaplan-Meier estimates of the distribution of time to relapse according to nadir PSA

```
. stset obstime relapse
. recode nadir 0/1=1 1/8=4 8/max=32, gen(nadirCTG)
. sts graph, by(nadirCTG)
```



29

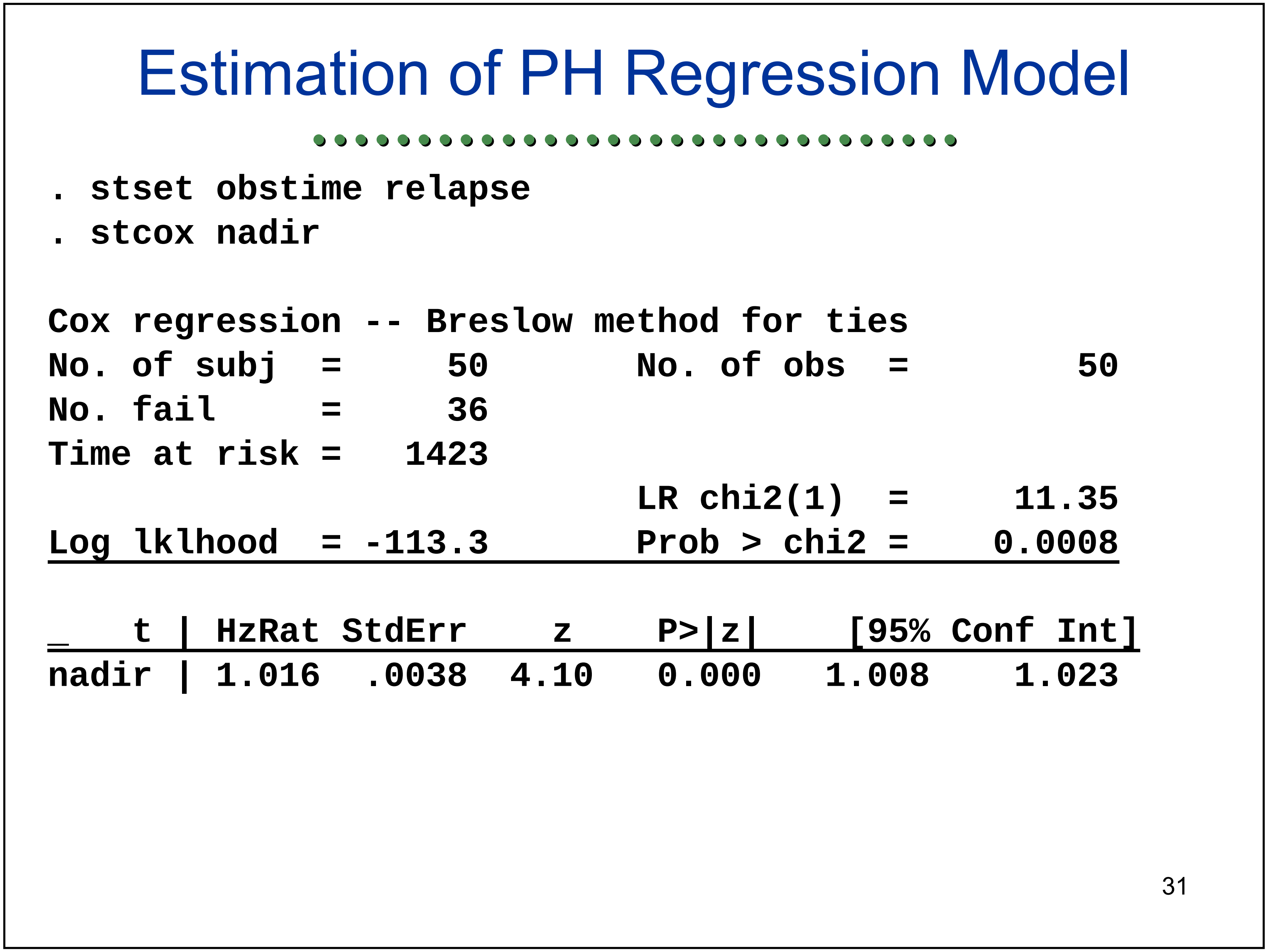
Tabulated Survival Estimates

- We can also get Stata to give us tables of estimates at specific times (here I chose 12, 24, and 36 months)

```
. stset obstime relapse
. sts list, by(nadirCTG) at(12 24 36)

        failure _d:  relapse
  analysis time _t:  obstime

              Beg.                          Survivor      Std.
    Time     Total      Fail                Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
nadir PSA < 1
      12        25         2                  0.9231    0.0523     0.7260    0.9802
      24        22         2                  0.8462    0.0708     0.6404    0.9393
      36        15         5                  0.6351    0.0978     0.4137    0.7918
1 < nadir PSA < 8
      12         9         3                  0.7273    0.1343     0.3708    0.9028
      24         7         3                  0.4545    0.1501     0.1666    0.7069
      36         7         0                  0.4545    0.1501     0.1666    0.7069
32 < nadir PSA
      12         6         8                  0.3846    0.1349     0.1405    0.6280
      24         2         4                  0.0769    0.0739     0.0048    0.2920
      36         1         1                       .         .          .         .
```

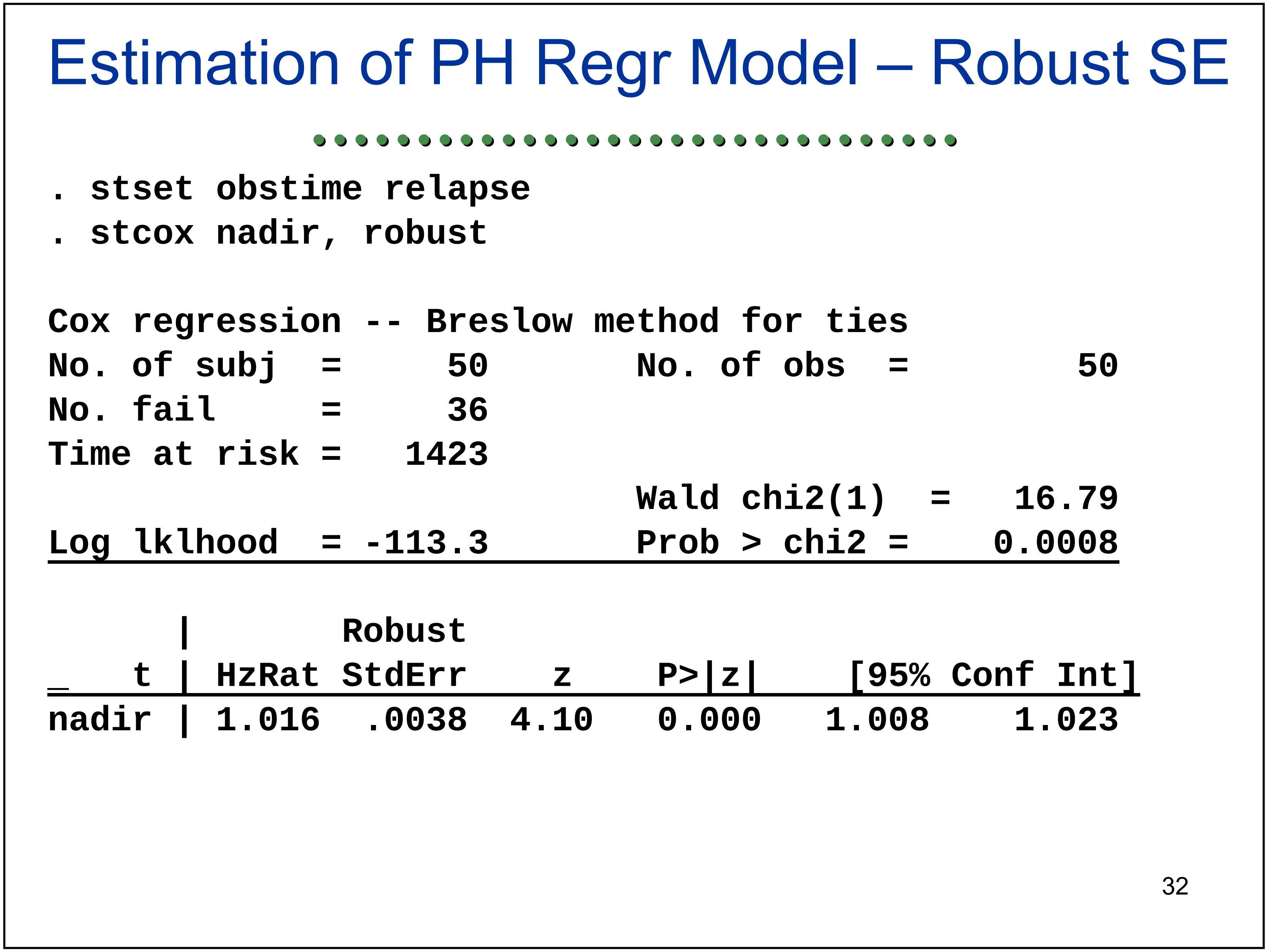


Estimation of PH Regression Model

```
. stset obstime relapse
. stcox nadir

Cox regression -- Breslow method for ties
No. of subj   =     50        No. of obs   =         50
No. fail      =     36
Time at risk =   1423
                              LR chi2(1)   =      11.35
Log lklhood   = -113.3        Prob > chi2 =     0.0008

_    t | HzRat  StdErr    z     P>|z|     [95% Conf Int]
nadir | 1.016   .0038   4.10   0.000   1.008     1.023
```

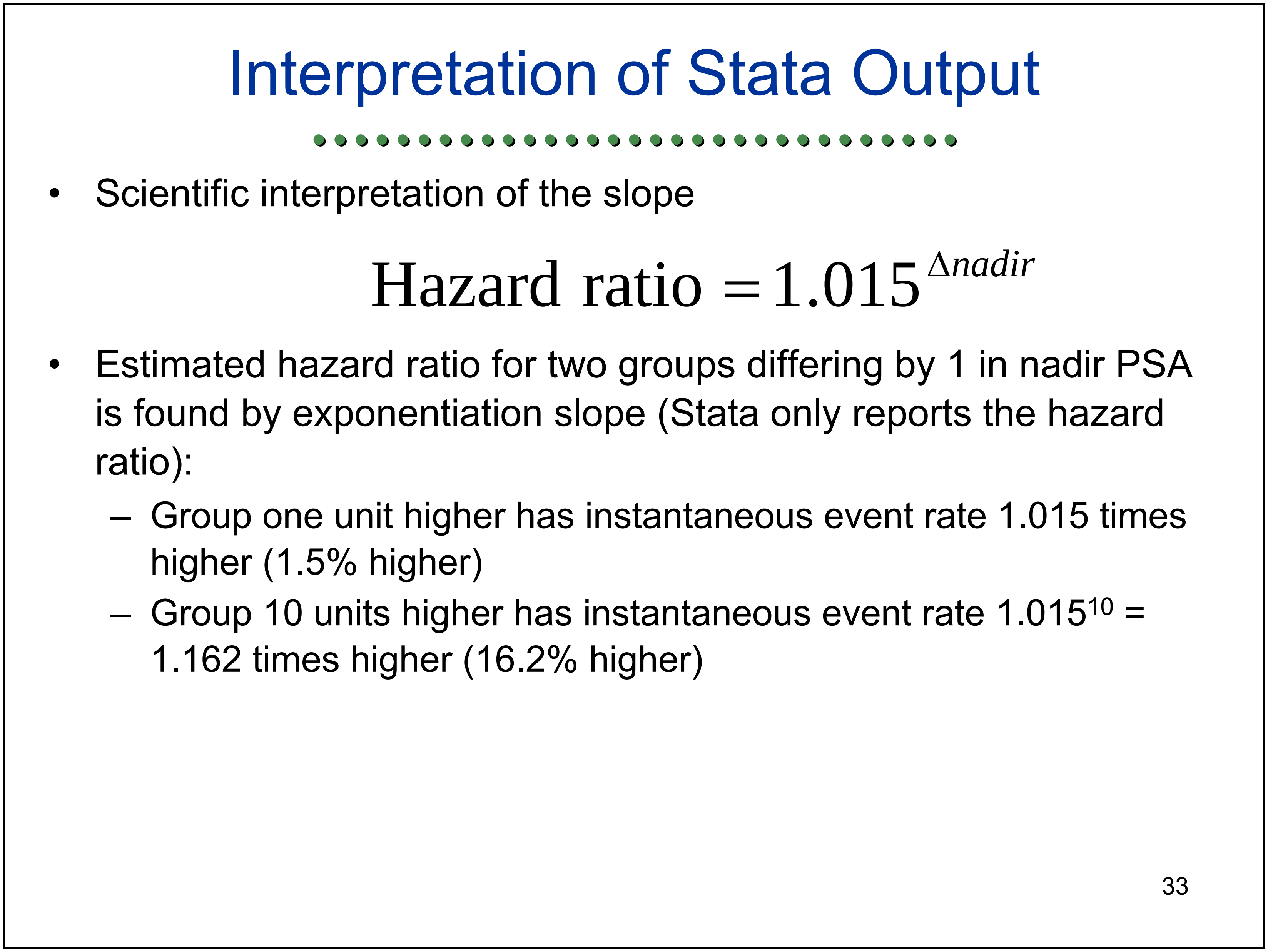


Estimation of PH Regr Model – Robust SE

```
. stset obstime relapse
. stcox nadir, robust

Cox regression -- Breslow method for ties
No. of subj   =      50       No. of obs   =        50
No. fail      =      36
Time at risk =    1423
                              Wald chi2(1)   =   16.79
Log lklhood   = -113.3        Prob > chi2 =    0.0008

      |       Robust
_    t | HzRat StdErr     z     P>|z|     [95% Conf Int]
nadir | 1.016   .0038  4.10    0.000    1.008     1.023
```



Interpretation of Stata Output

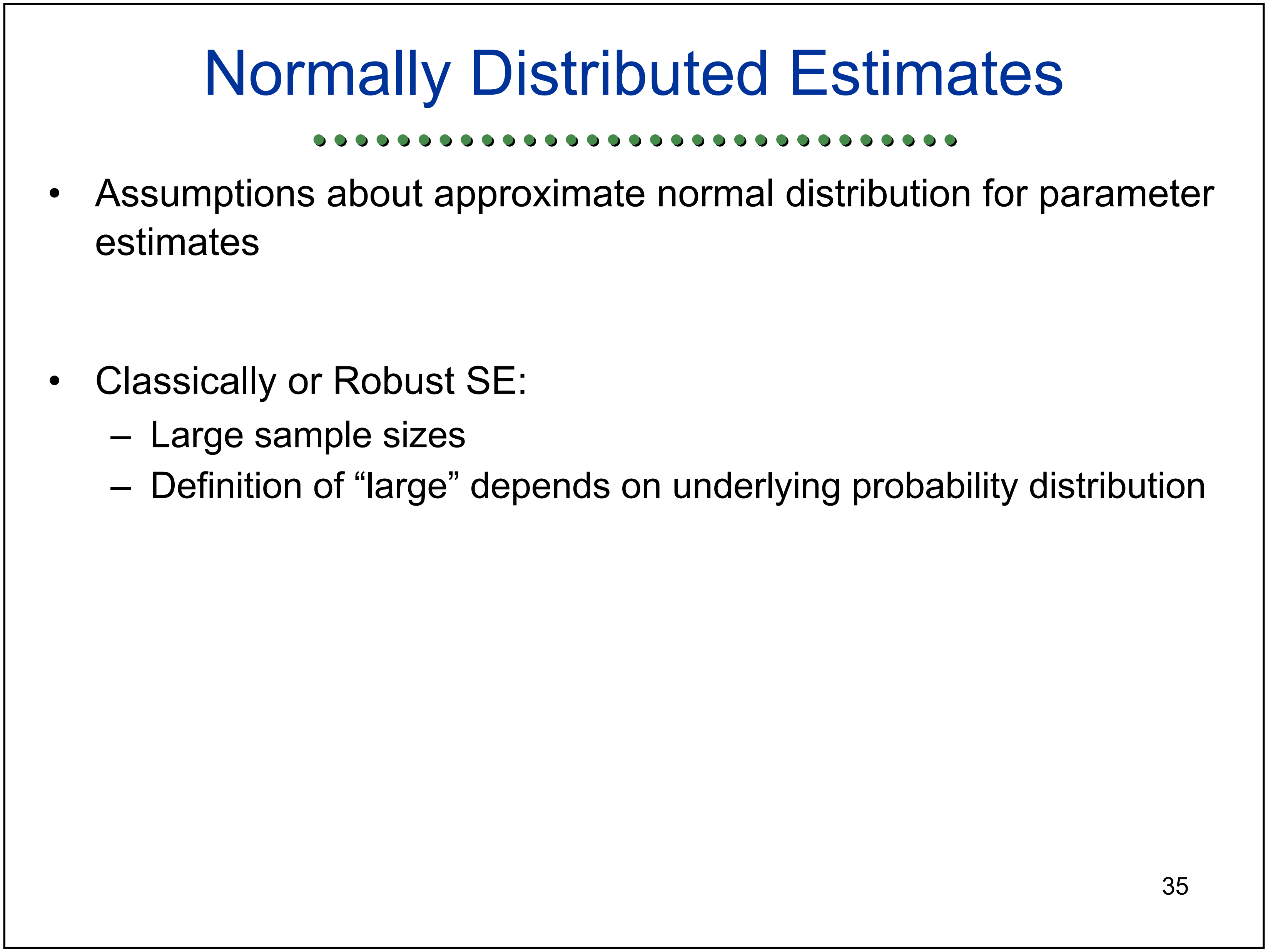
- Scientific interpretation of the slope

$$\text{Hazard ratio} = 1.015^{\Delta nadir}$$

- Estimated hazard ratio for two groups differing by 1 in nadir PSA is found by exponentiation slope (Stata only reports the hazard ratio):
  - Group one unit higher has instantaneous event rate 1.015 times higher (1.5% higher)
  - Group 10 units higher has instantaneous event rate 1.015^{10} = 1.162 times higher (16.2% higher)

Statistical Validity of Inference

- Inference (CI, P vals) about associations requires three general assumptions
  - Assumptions about approximate normal distribution for parameter estimates
  - Assumptions about independence of observations
  - Assumptions about variance of observations within groups



Normally Distributed Estimates

- Assumptions about approximate normal distribution for parameter estimates

- Classically or Robust SE:
  - Large sample sizes
  - Definition of “large” depends on underlying probability distribution

Independence / Dependence

- Assumptions about independence of observations for PH regression

- Classically:
  - All observations are independent

- Robust standard error estimates:
  - Allow correlated observations within identified clusters

Within Group Variance

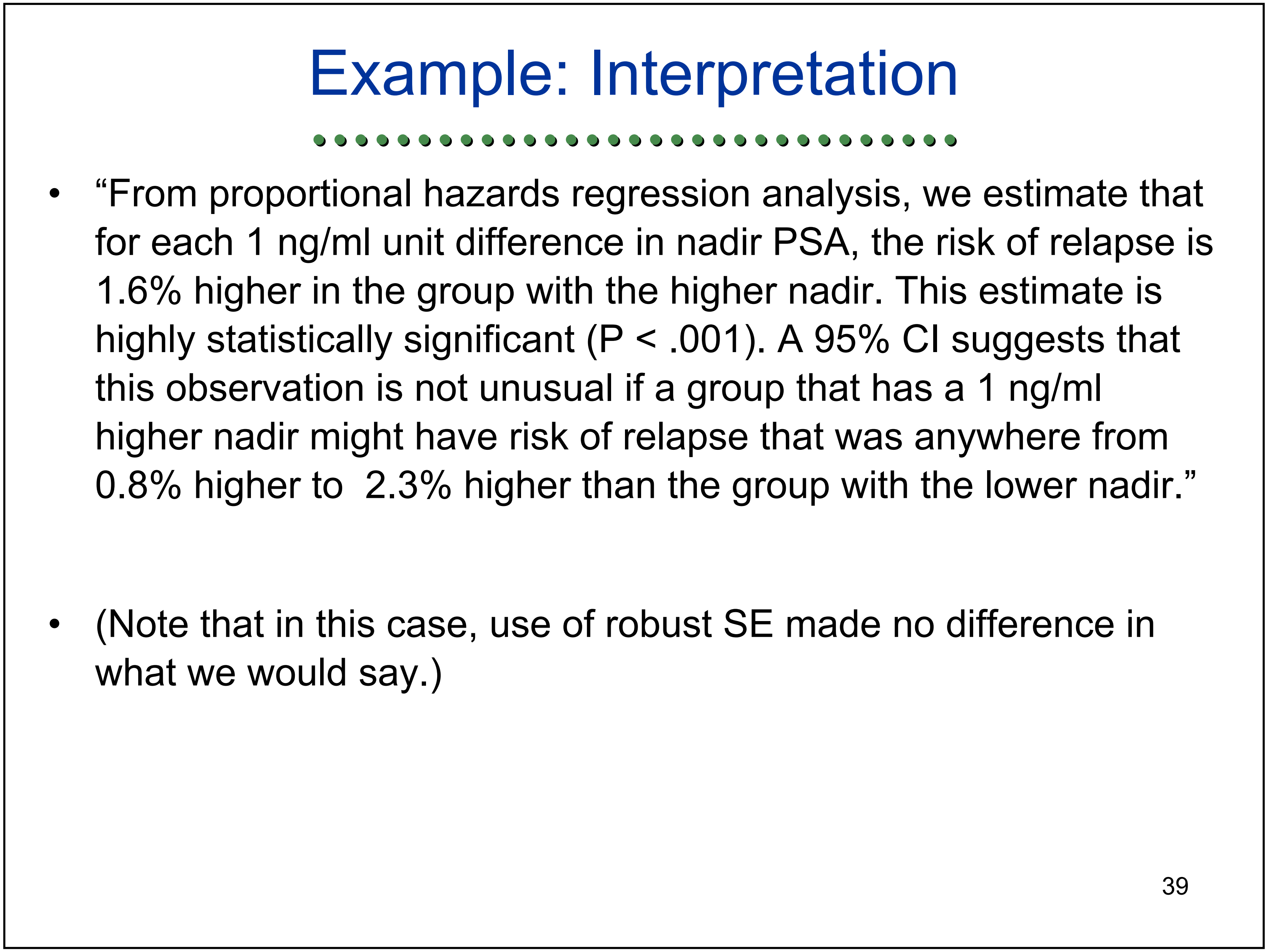
- Assumptions about variance of response within groups for proportional hazards regression

- Classically:
  - Mean variance relationship for binary data
    - Proportional hazards considers odds of event at every time
    - Need proportional hazards and linearity of predictor

- Robust standard error estimates:
  - Allow unequal variances across groups
  - (Do not need proportional hazards or linearity if sample size sufficiently large)

Linearity of Model

- Assumption about adequacy of linear model for prediction of group survival curves
- In PH regression, the baseline hazard function is the “intercept”
  - It does not need to be estimated in order to estimate regression parameters
  - It can be estimated afterward and used to estimate a survival curve for each covariate group
- In order for this to be valid, we would need both
  - The log hazard ratio across groups is linear in the modeled predictor (we can model transformations of the measured variable)
  - The proportionals hazards assumption has to hold



Example: Interpretation

- “From proportional hazards regression analysis, we estimate that for each 1 ng/ml unit difference in nadir PSA, the risk of relapse is 1.6% higher in the group with the higher nadir. This estimate is highly statistically significant (P < .001). A 95% CI suggests that this observation is not unusual if a group that has a 1 ng/ml higher nadir might have risk of relapse that was anywhere from 0.8% higher to 2.3% higher than the group with the lower nadir.”

- (Note that in this case, use of robust SE made no difference in what we would say.)

Log Transformed NadirPSA

- Based on prior experience a log transformation of PSA was what I would have considered *a priori*

- A constant difference in PSA would not be expected to confer same increase in 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
  - Use log transformed nadir PSA

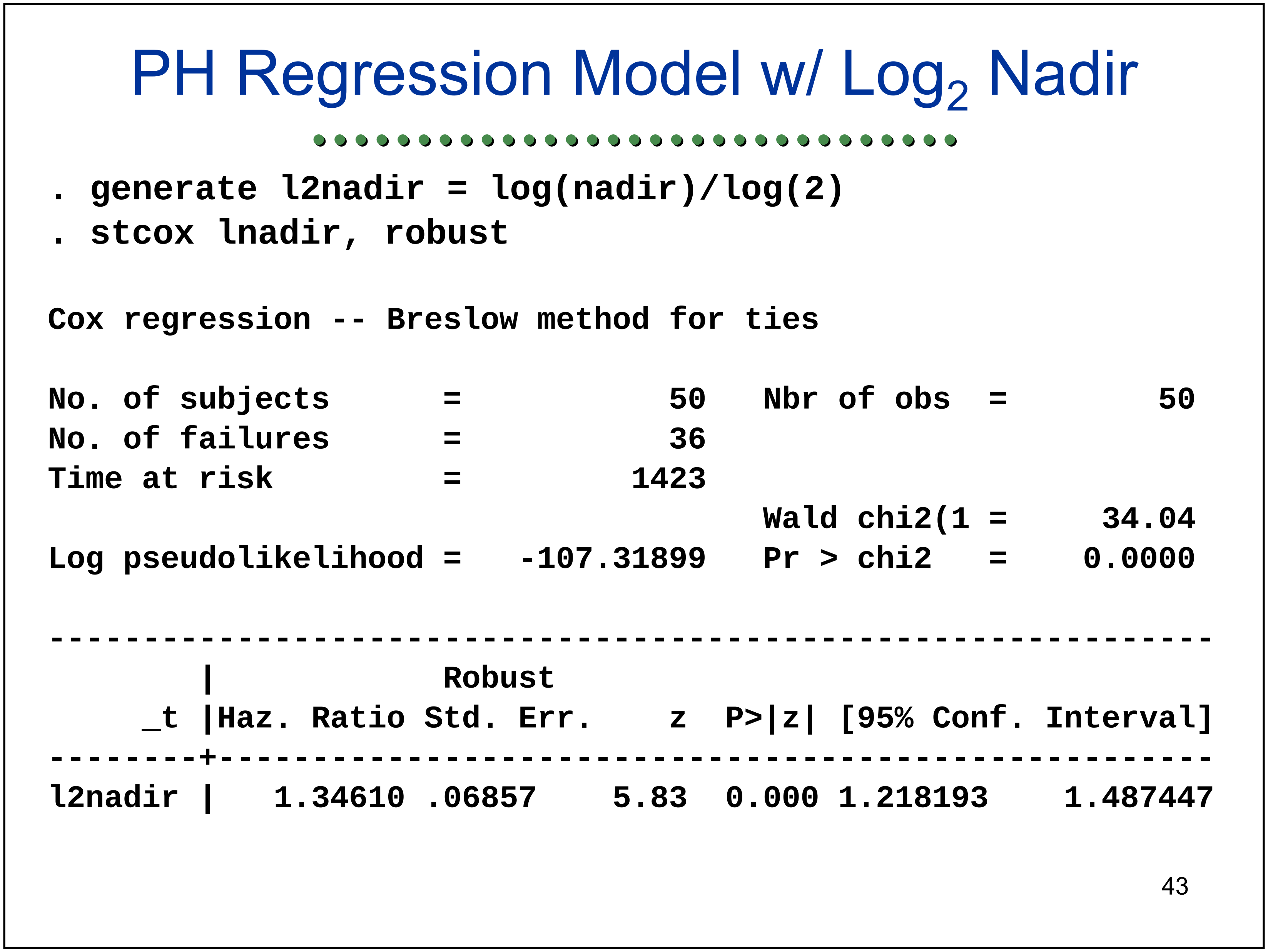
PH Regression Model w/ Log Nadir

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

Cox regression -- Breslow method for ties
No. of subj   =     50        No. of obs   =        50
No. fail      =     36
Time at risk  =   1423
                              LR chi2(1)   =     34.04
Log lklhood   = -107.3        Prob > chi2 =    0.0000
_     t | HzRat StdErr     z     P>|z|    [95% Conf Int]
lnadir | 1.54   .113   5.83   0.000     1.33     1.77
```

Interpretation of Parameters

- Hazard ratio is 1.54 for an e-fold difference in nadir PSA
  - e = 2.7183
- I can more easily understand doubling, tripling, 5-fold, 10-fold increases
  - For doubling: HR : $1.54^{\log(2)} = 1.35$
- If I were going to reference a doubling, it might be easier just to compute the base 2 logarithm of nadir PSA, and then Stata would have done all the calculations I needed



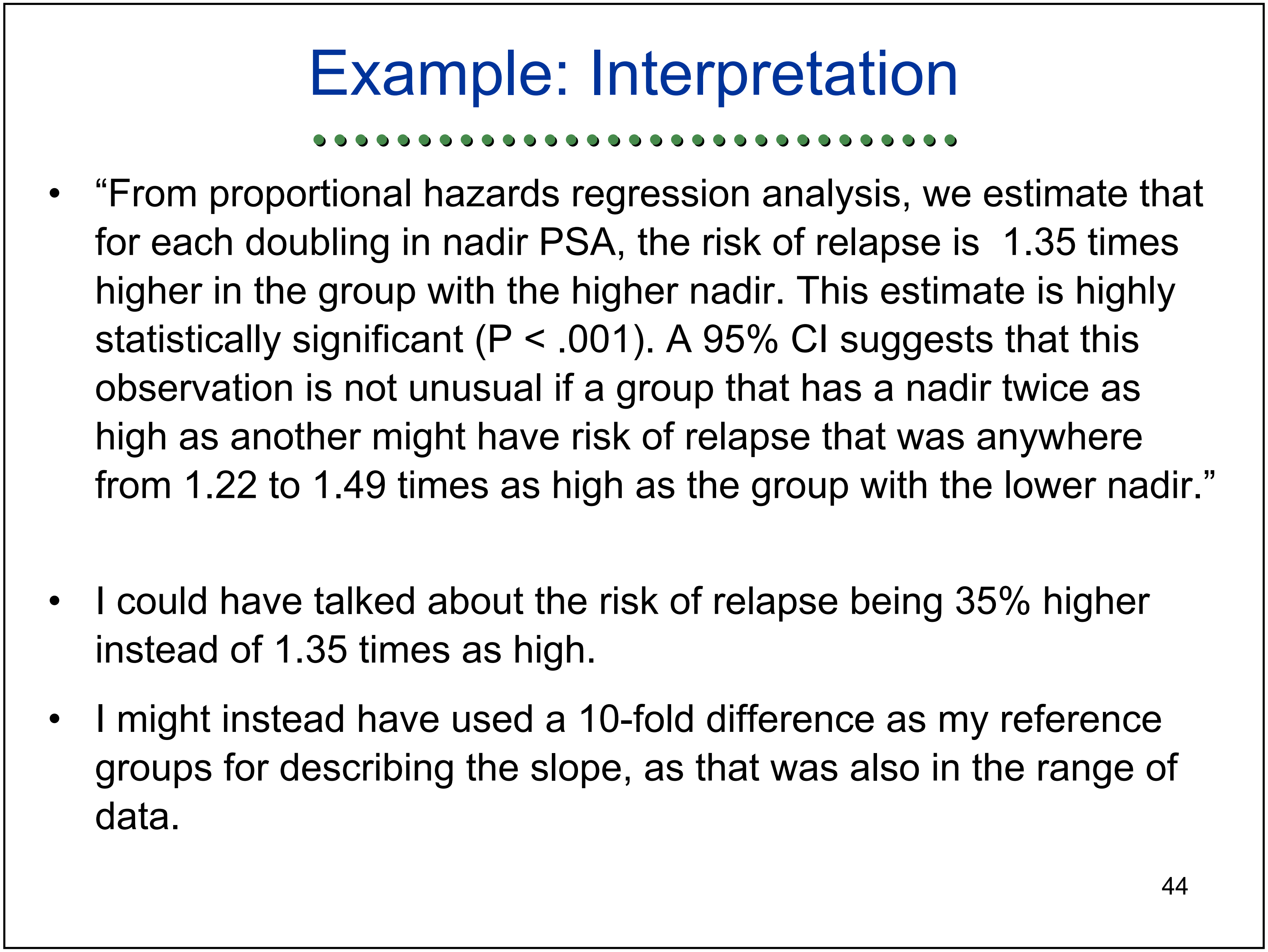
PH Regression Model w/ Log$_2$ Nadir

```
. generate l2nadir = log(nadir)/log(2)
. stcox lnadir, robust

Cox regression -- Breslow method for ties

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

--------------------------------------------------------------
        |            Robust
     _t |Haz. Ratio Std. Err.    z  P>|z| [95% Conf. Interval]
--------+-----------------------------------------------------
l2nadir |   1.34610 .06857    5.83  0.000 1.218193    1.487447
```



Example: Interpretation

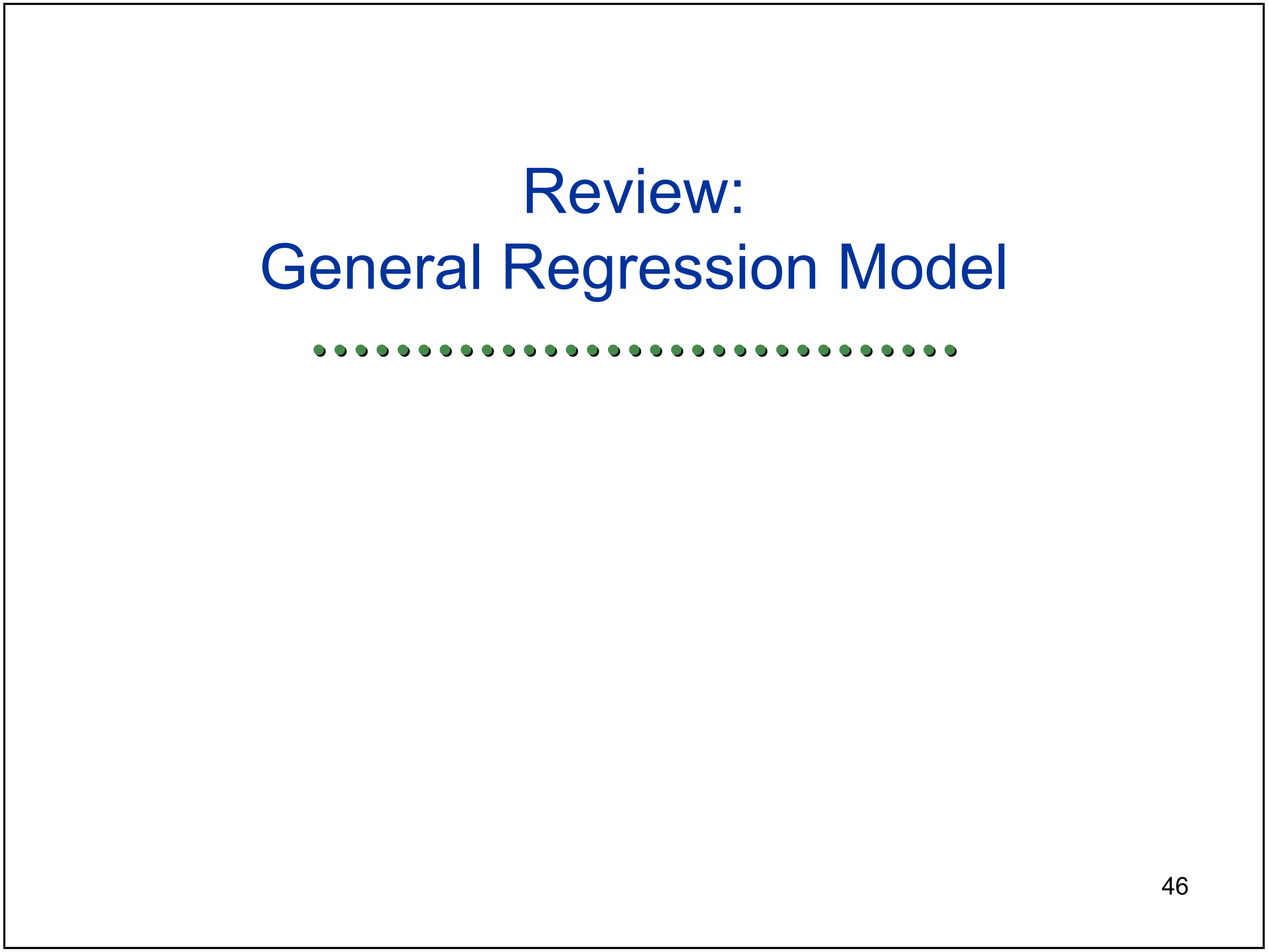
- “From proportional hazards regression analysis, we estimate that for each doubling in nadir PSA, the risk of relapse is 1.35 times higher in the group with the higher nadir. This estimate is highly statistically significant (P < .001). A 95% CI suggests that this observation is not unusual if a group that has a nadir twice as high as another might have risk of relapse that was anywhere from 1.22 to 1.49 times as high as the group with the lower nadir.”

- I could have talked about the risk of relapse being 35% higher instead of 1.35 times as high.
- I might instead have used a 10-fold difference as my reference groups for describing the slope, as that was also in the range of data.

PH Regression and Logrank Test

- Proportional hazards regression with a binary predictor (two groups) corresponds to the logrank test
  - For this reason, when only doing a two sample test, I tend to suggest using the hazard ratio from PH regression as your quantification of the difference in survival

- Three possible statistics from proportional hazards regression
  - Wald: The test based on the estimate and SE
  - Score: Corresponds to logrank test, but not given in Stata output
  - Likelihood ratio test: Can be obtained using post-regression commands in Stata (covered with adjustments for covariates)



Review:
General Regression Model

46

General Regression

- General notation for variables and parameter

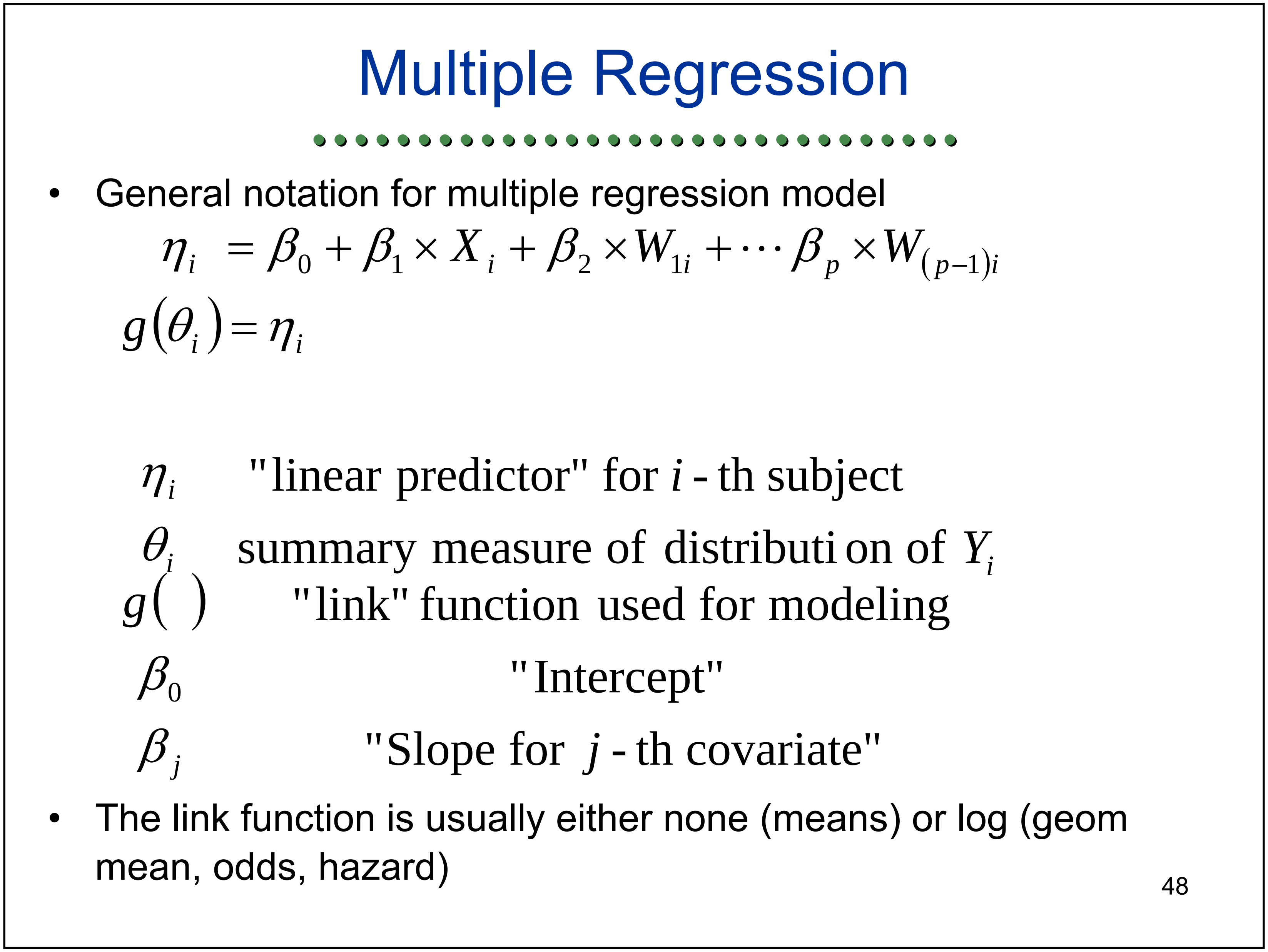
Y_i Response measured on the ith subject

X_i Value of the POI for the ith subject

$W_{1i},\ldots W_{(p-1)i}$ Value of additional covariates for the ith subject

θ_i Parameter of distribution of Y_i

- The parameter might be the mean, geometric mean, odds, rate, instantaneous risk of an event (hazard), etc.



Multiple Regression

- General notation for multiple regression model

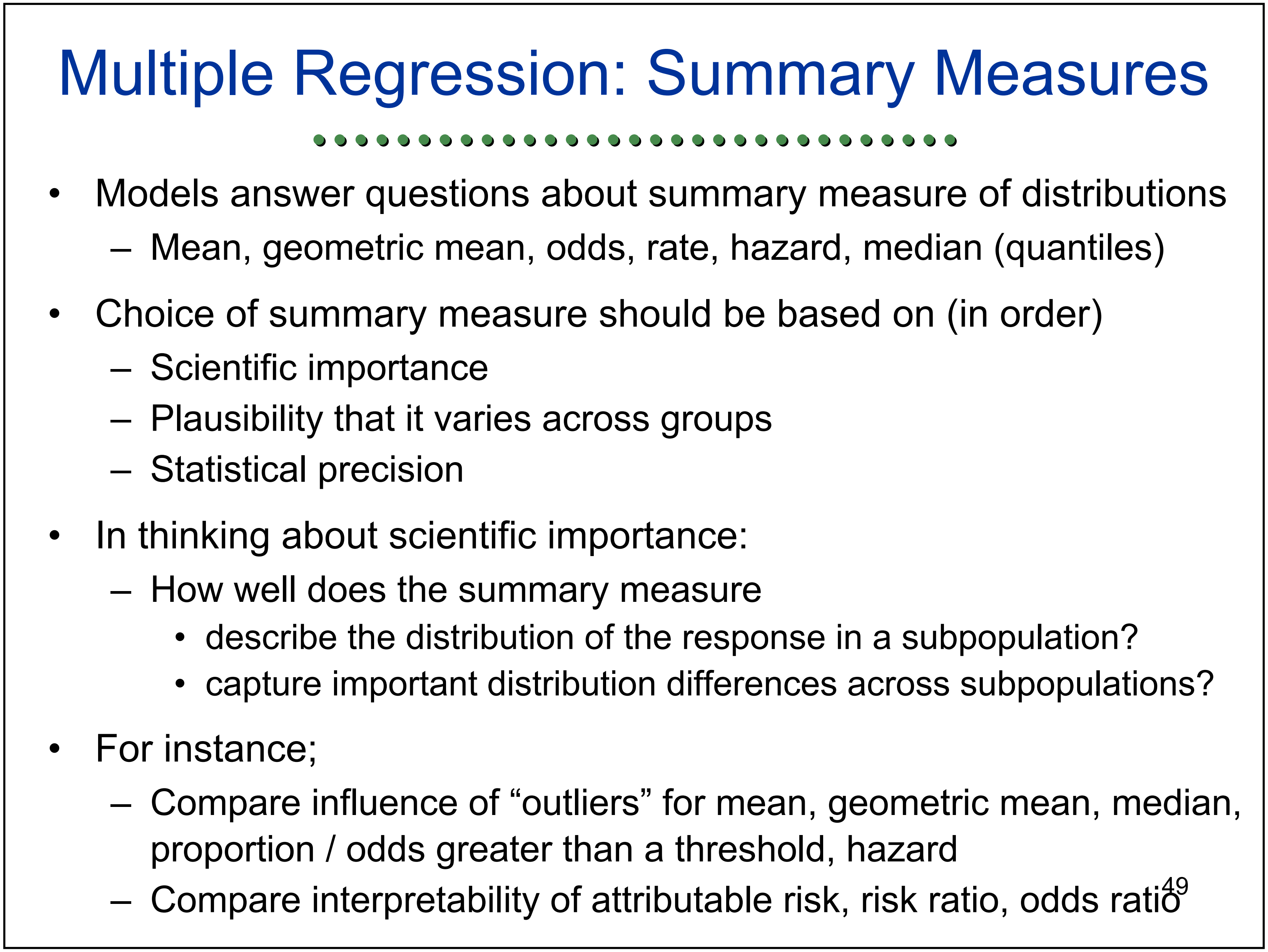
$$\eta_i = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_{1i} + \cdots \beta_p \times W_{(p-1)i}$$

$$g(\theta_i) = \eta_i$$

η_i	"linear predictor" for i - th subject
θ_i	summary measure of distributi on of Y_i
$g(\ )$	"link" function used for modeling
β_0	"Intercept"
β_j	"Slope for j - th covariate"

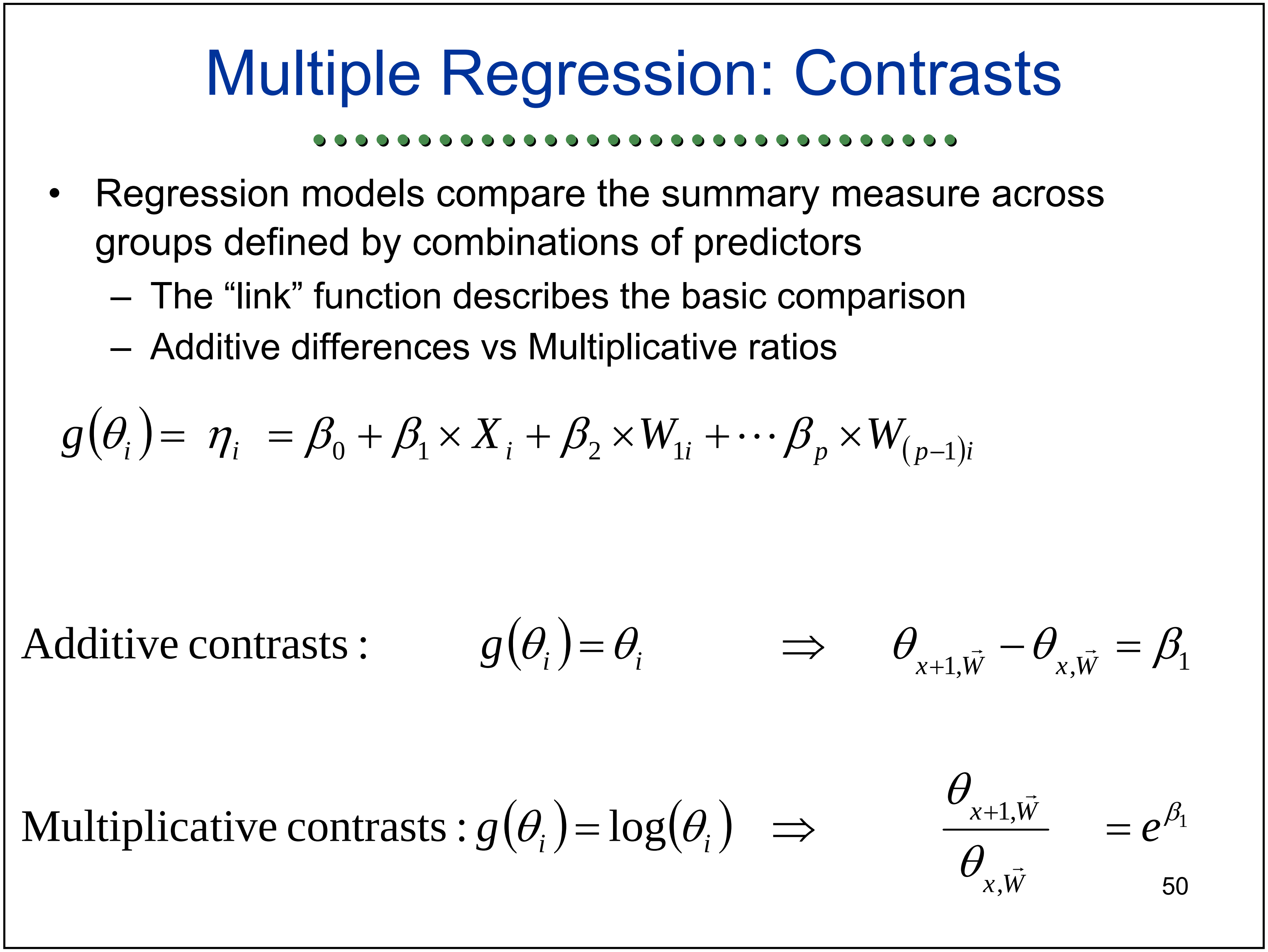
- The link function is usually either none (means) or log (geom mean, odds, hazard)

48



Multiple Regression: Summary Measures

- Models answer questions about summary measure of distributions
  - Mean, geometric mean, odds, rate, hazard, median (quantiles)
- Choice of summary measure should be based on (in order)
  - Scientific importance
  - Plausibility that it varies across groups
  - Statistical precision
- In thinking about scientific importance:
  - How well does the summary measure
    - describe the distribution of the response in a subpopulation?
    - capture important distribution differences across subpopulations?
- For instance;
  - Compare influence of “outliers” for mean, geometric mean, median, proportion / odds greater than a threshold, hazard
  - Compare interpretability of attributable risk, risk ratio, odds ratio



Multiple Regression: Contrasts

- Regression models compare the summary measure across groups defined by combinations of predictors
  - The “link” function describes the basic comparison
  - Additive differences vs Multiplicative ratios

$$g(\theta_i) = \eta_i = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_{1i} + \cdots \beta_p \times W_{(p-1)i}$$

$$\text{Additive contrasts:} \quad g(\theta_i) = \theta_i \quad \Rightarrow \quad \theta_{x+1,\vec{W}} - \theta_{x,\vec{W}} = \beta_1$$

$$\text{Multiplicative contrasts:} \; g(\theta_i) = \log(\theta_i) \quad \Rightarrow \quad \frac{\theta_{x+1,\vec{W}}}{\theta_{x,\vec{W}}} = e^{\beta_1}$$

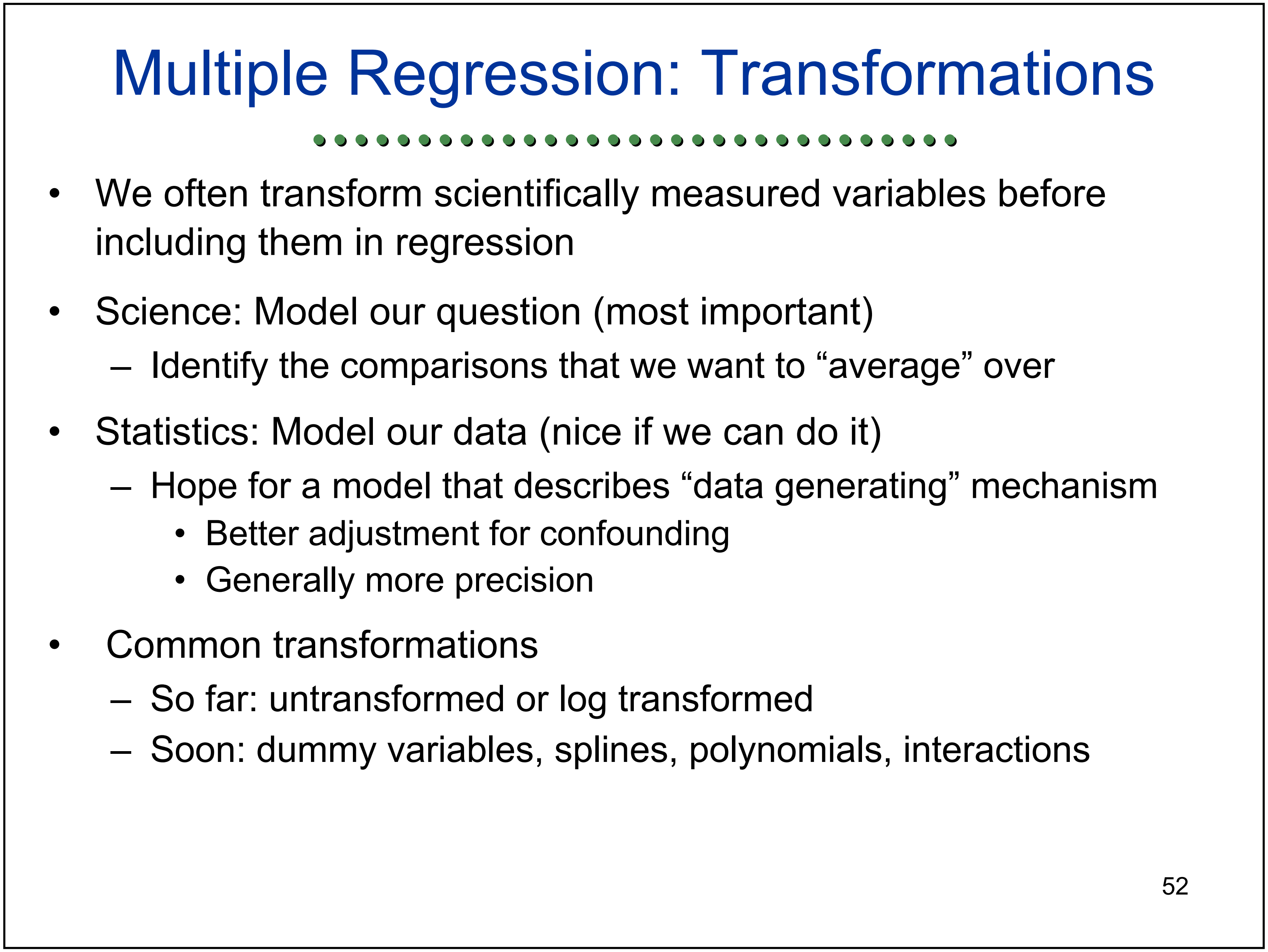
Multiple Regression: Linear Predictor

- Key point: The “linear predictor” is linear in the parameters
  - Modeled covariates *X, W_1, W_2, …* are often transformations of scientifically measured variables, but still called “linear predictor”
- Regression models compare the summary measure across groups defined by combinations of predictors
  - By specifying the form of variables in the “linear predictor” we are describing which comparisons give similar answers
  - The slope parameters measure “average” comparisons across 1 unit difference in the ***modeled*** covariates

$$g(\theta_i) = \eta_i = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_{1i} + \cdots \beta_p \times W_{(p-1)i}$$

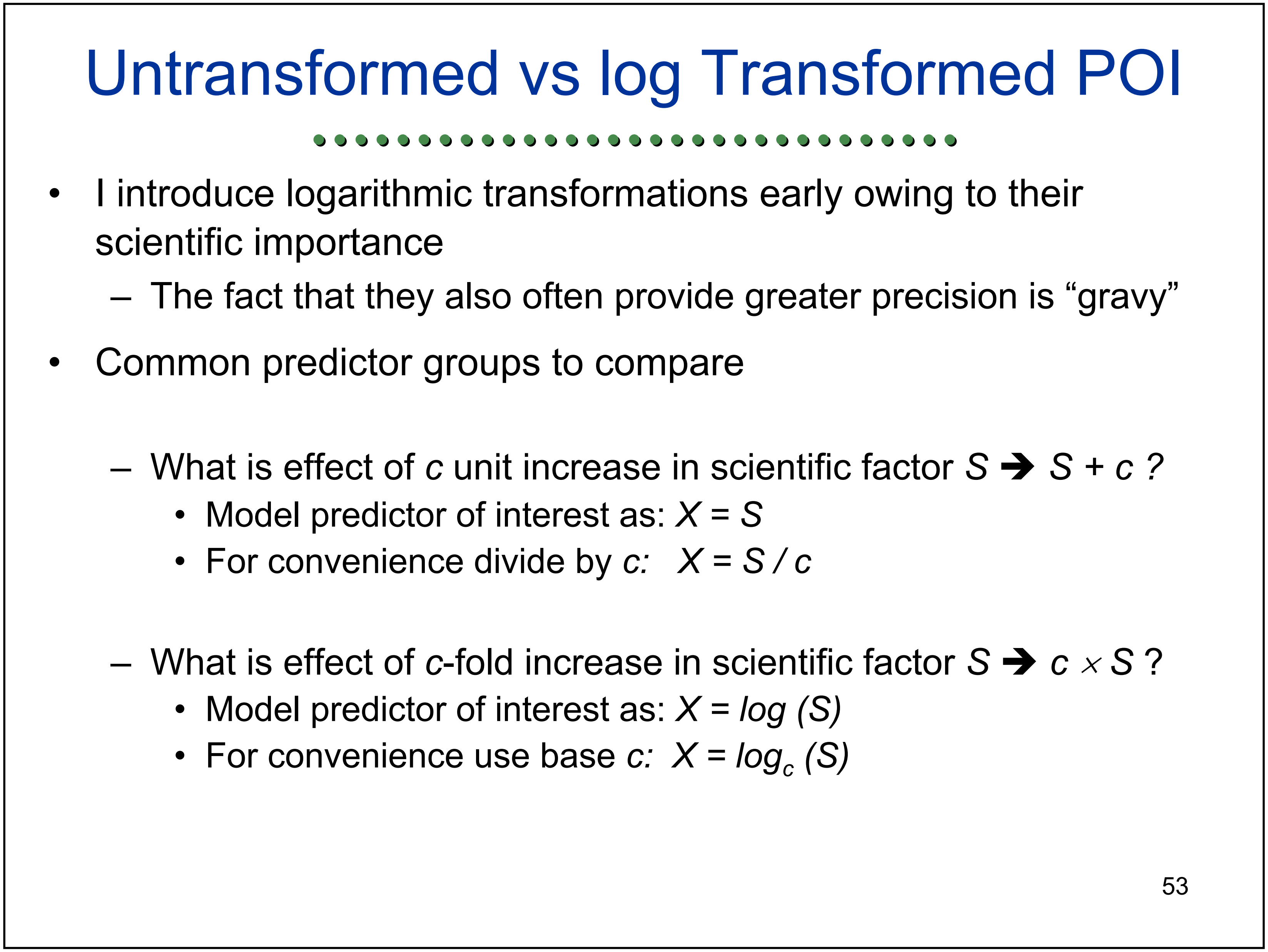
$$\text{Additive contrasts: } \quad g(\theta_i) = \theta_i \quad \Rightarrow \quad \theta_{x+1,\vec{W}} - \theta_{x,\vec{W}} = \beta_1$$

$$\text{Multiplicative contrasts: } g(\theta_i) = \log(\theta_i) \quad \Rightarrow \quad \frac{\theta_{x+1,\vec{W}}}{\theta_{x,\vec{W}}} = e^{\beta_1}$$



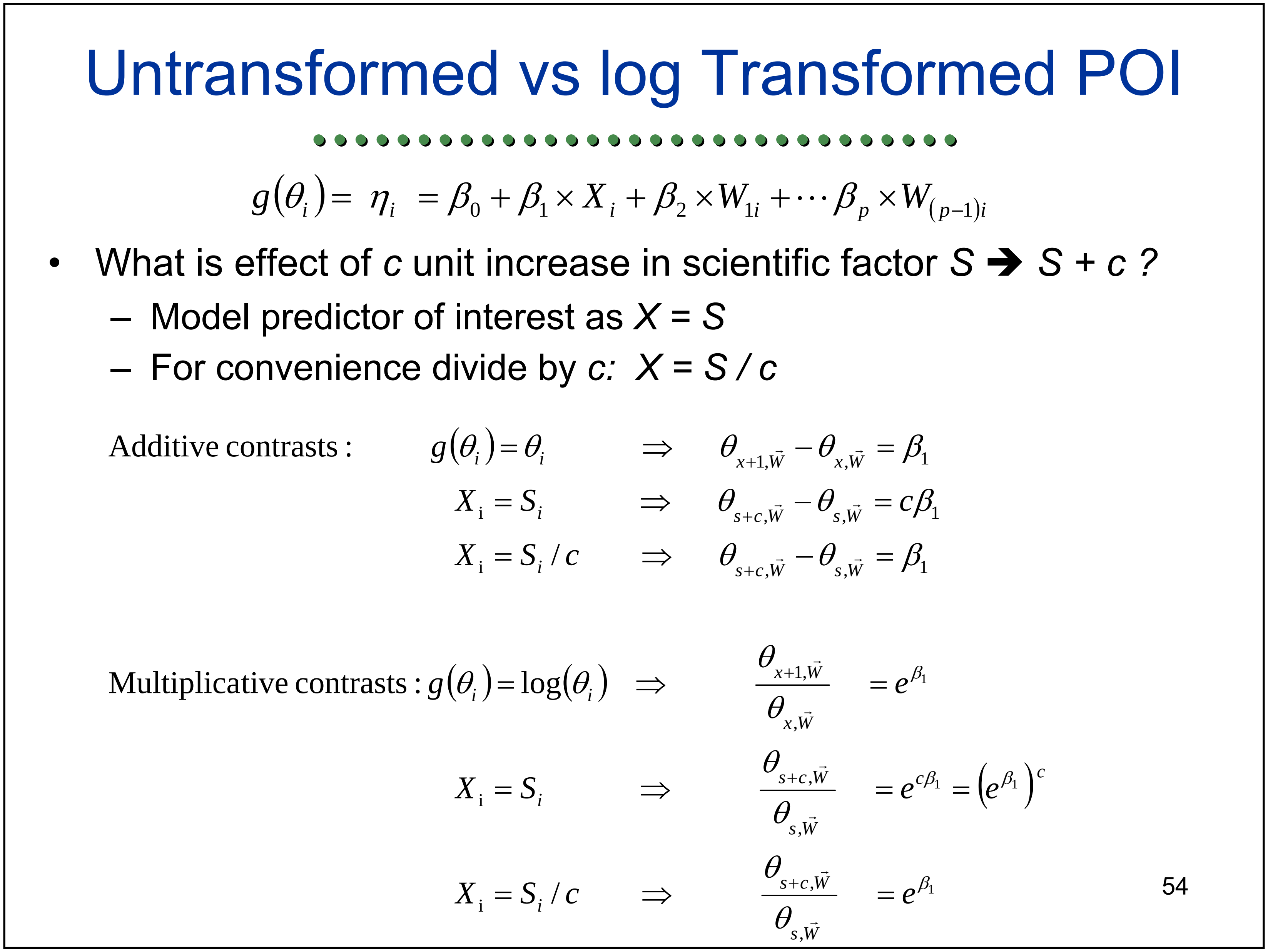
Multiple Regression: Transformations

- We often transform scientifically measured variables before including them in regression
- Science: Model our question (most important)
  - Identify the comparisons that we want to “average” over
- Statistics: Model our data (nice if we can do it)
  - Hope for a model that describes “data generating” mechanism
    - Better adjustment for confounding
    - Generally more precision
- Common transformations
  - So far: untransformed or log transformed
  - Soon: dummy variables, splines, polynomials, interactions



Untransformed vs log Transformed POI

- I introduce logarithmic transformations early owing to their scientific importance
  - The fact that they also often provide greater precision is “gravy”
- Common predictor groups to compare
  - What is effect of *c* unit increase in scientific factor $S \rightarrow S + c$?
    - Model predictor of interest as: $X = S$
    - For convenience divide by *c*: $X = S / c$
  - What is effect of *c*-fold increase in scientific factor $S \rightarrow c \times S$?
    - Model predictor of interest as: $X = log\ (S)$
    - For convenience use base *c*: $X = log_c\ (S)$



Untransformed vs log Transformed POI

$$g(\theta_i) = \eta_i = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_{1i} + \cdots \beta_p \times W_{(p-1)i}$$

- What is effect of *c* unit increase in scientific factor *S* ➔ *S* + *c* ?
  - Model predictor of interest as *X* = *S*
  - For convenience divide by *c:* *X* = *S* / *c*

$$\begin{aligned}
&\text{Additive contrasts}: && g(\theta_i) = \theta_i && \Rightarrow && \theta_{x+1,\vec{W}} - \theta_{x,\vec{W}} = \beta_1 \\
& && X_{\mathrm{i}} = S_i && \Rightarrow && \theta_{s+c,\vec{W}} - \theta_{s,\vec{W}} = c\beta_1 \\
& && X_{\mathrm{i}} = S_i / c && \Rightarrow && \theta_{s+c,\vec{W}} - \theta_{s,\vec{W}} = \beta_1
\end{aligned}$$

$$\begin{aligned}
&\text{Multiplicative contrasts}: && g(\theta_i) = \log(\theta_i) && \Rightarrow && \frac{\theta_{x+1,\vec{W}}}{\theta_{x,\vec{W}}} = e^{\beta_1} \\
& && X_{\mathrm{i}} = S_i && \Rightarrow && \frac{\theta_{s+c,\vec{W}}}{\theta_{s,\vec{W}}} = e^{c\beta_1} = \left(e^{\beta_1}\right)^c \\
& && X_{\mathrm{i}} = S_i / c && \Rightarrow && \frac{\theta_{s+c,\vec{W}}}{\theta_{s,\vec{W}}} = e^{\beta_1}
\end{aligned}$$

Untransformed vs log Transformed POI

$$g(\theta_i) = \eta_i = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_{1i} + \cdots \beta_p \times W_{(p-1)i}$$

- What is effect of *c*-fold increase in scientific factor *S* ➔ *c* × *S* ?
  - Model predictor of interest as *X* = *log (S)*
  - For convenience use base *c:* *X* = *log*$_c$ *(S)*

$$\text{Additive contrasts}: \quad g(\theta_i) = \theta_i \quad \Rightarrow \quad \theta_{x+1,\vec{W}} - \theta_{x,\vec{W}} = \beta_1$$

$$X_{\text{i}} = \log(S_i) \quad \Rightarrow \quad \theta_{s\times c,\vec{W}} - \theta_{s,\vec{W}} = \log(c) \times \beta_1$$

$$X_{\text{i}} = \log_c(S_i) \Rightarrow \quad \theta_{s\times c,\vec{W}} - \theta_{s,\vec{W}} = \beta_1$$

$$\text{Multiplicative contrasts}: g(\theta_i) = \log(\theta_i) \quad \Rightarrow \quad \frac{\theta_{x+1,\vec{W}}}{\theta_{x,\vec{W}}} = e^{\beta_1}$$

$$X_{\text{i}} = \log(S_i) \quad \Rightarrow \quad \frac{\theta_{s\times c,\vec{W}}}{\theta_{s,\vec{W}}} = e^{\log(c)\times\beta_1} = c^{\beta_1} = \left(e^{\beta_1}\right)^{\log(c)}$$

$$X_{\text{i}} = \log_c(S_i) \Rightarrow \quad \frac{\theta_{s\times c,\vec{W}}}{\theta_{s,\vec{W}}} = e^{\beta_1}$$

55

Stata: Post Estimation Commands

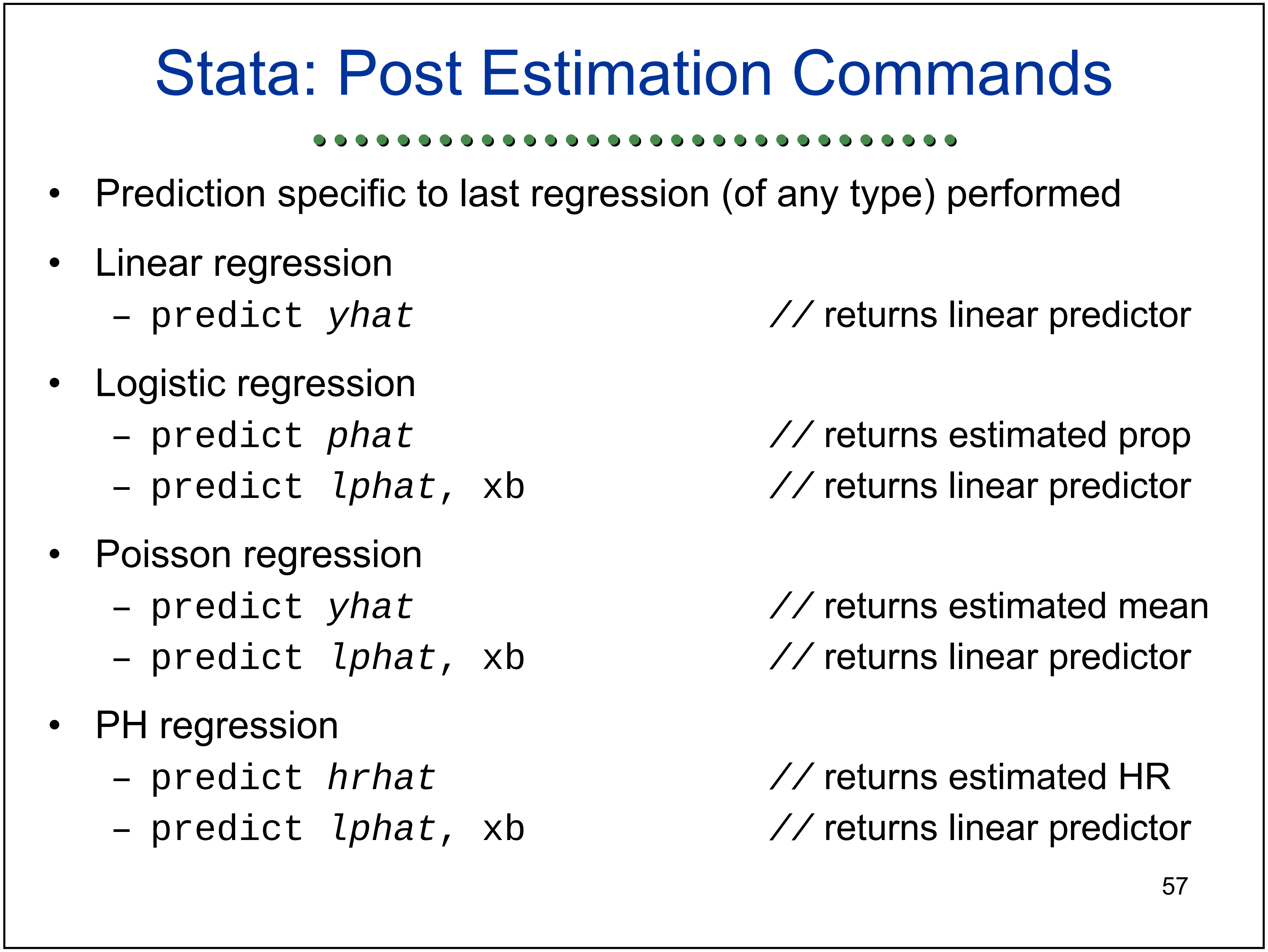
- Stata has a suite of command that can be used for estimation and testing after any regression
  - `predict` can be used to obtain estimates of the linear predictor and / or the summary measure θ

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 \times X_i + \hat{\beta}_2 \times W_{1i} + \cdots \hat{\beta}_p \times W_{(p-1)i}$$

$$\text{Additive models: } \quad g(\theta_i) = \theta_i \quad \Rightarrow \quad \hat{\theta}_i = \hat{\eta}_i$$

$$\text{Multiplicative models: } g(\theta_i) = \log(\theta_i) \quad \Rightarrow \quad \hat{\theta}_i = e^{\hat{\eta}_i}$$

56



Stata: Post Estimation Commands

- Prediction specific to last regression (of any type) performed
- Linear regression
  - `predict yhat` // returns linear predictor
- Logistic regression
  - `predict phat` // returns estimated prop
  - `predict lphat, xb` // returns linear predictor
- Poisson regression
  - `predict yhat` // returns estimated mean
  - `predict lphat, xb` // returns linear predictor
- PH regression
  - `predict hrhat` // returns estimated HR
  - `predict lphat, xb` // returns linear predictor

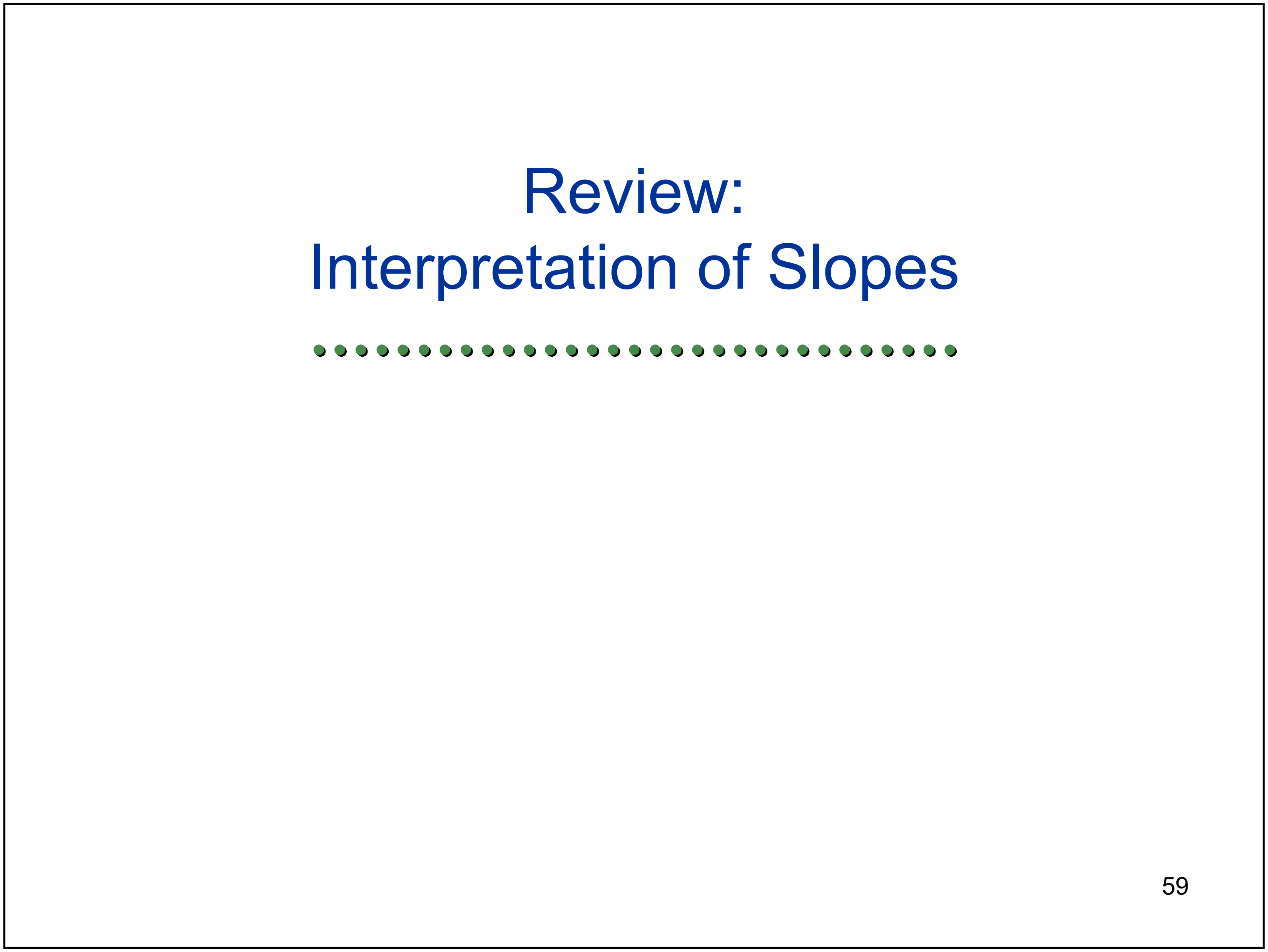
57

Stata: Post Testing Commands

- Stata also allows testing and estimation of combinations of covariates

```
– test v1 v2 …                  //Wald test that coeff all 0
– testparm v*                   //Wald test allows wildcards
– lrtest model1                 //LR test comparing models
– lincom 3 * v1 + 2 * v2        //Wald based estimation
```

- We will discuss these further as we get into multiple regression



Review:
Interpretation of Slopes

“Additive Models”

- Identity link function
  - Means: linear regression

$$\theta_X = \beta_0 + \beta_1 \times X$$

“Additive Models”: Slope

- Interpretation of slope:

  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

$$\theta_X = \beta_0 + \beta_1 \times X$$

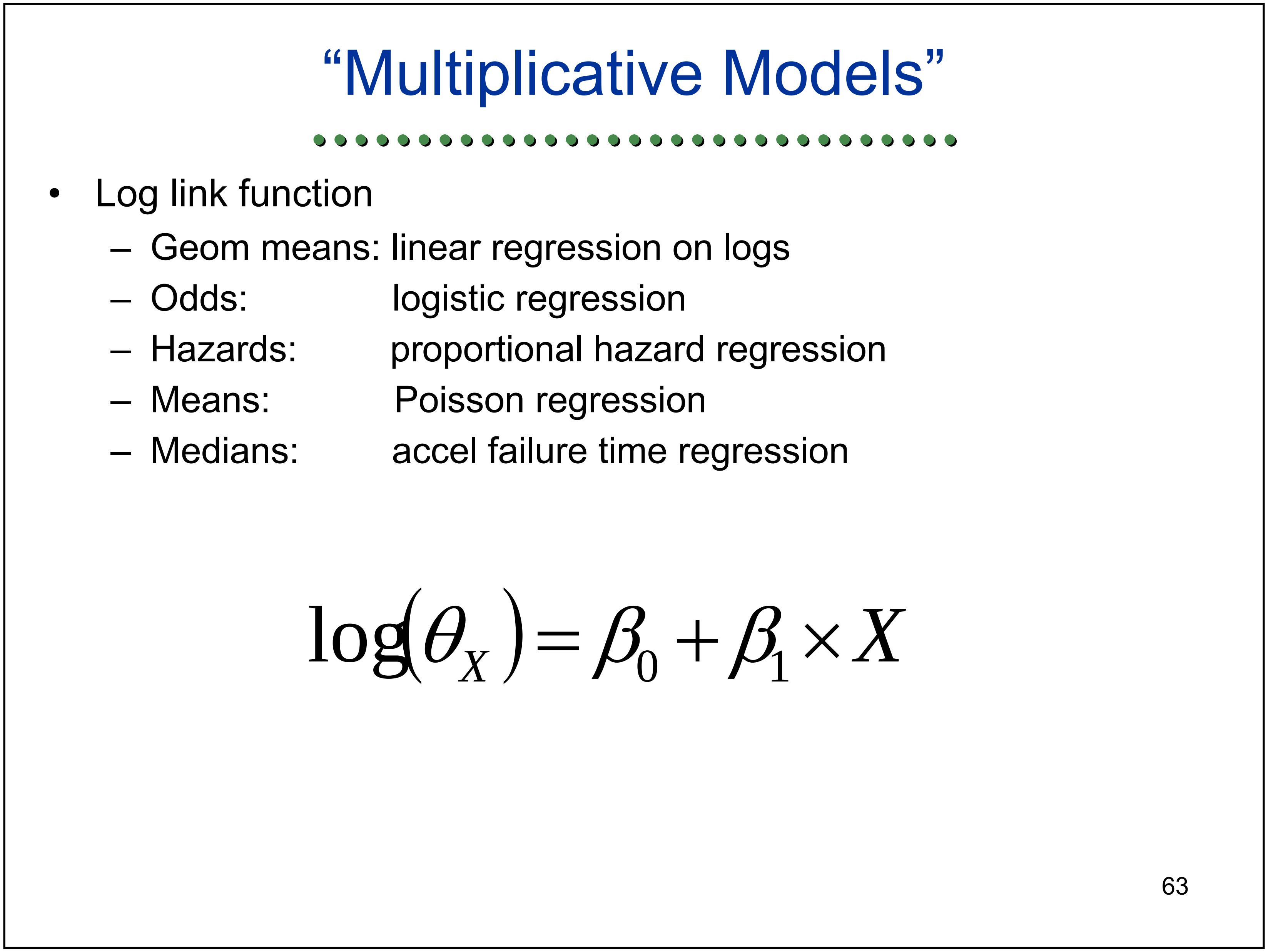
"Additive Models": log(X)

- Slope with log transformed predictor

  log(k) $\times \beta_1$: (Average) Difference in summary measure between groups per *k*-fold difference in *X*

$$\theta_X = \beta_0 + \beta_1 \times \log(X)$$

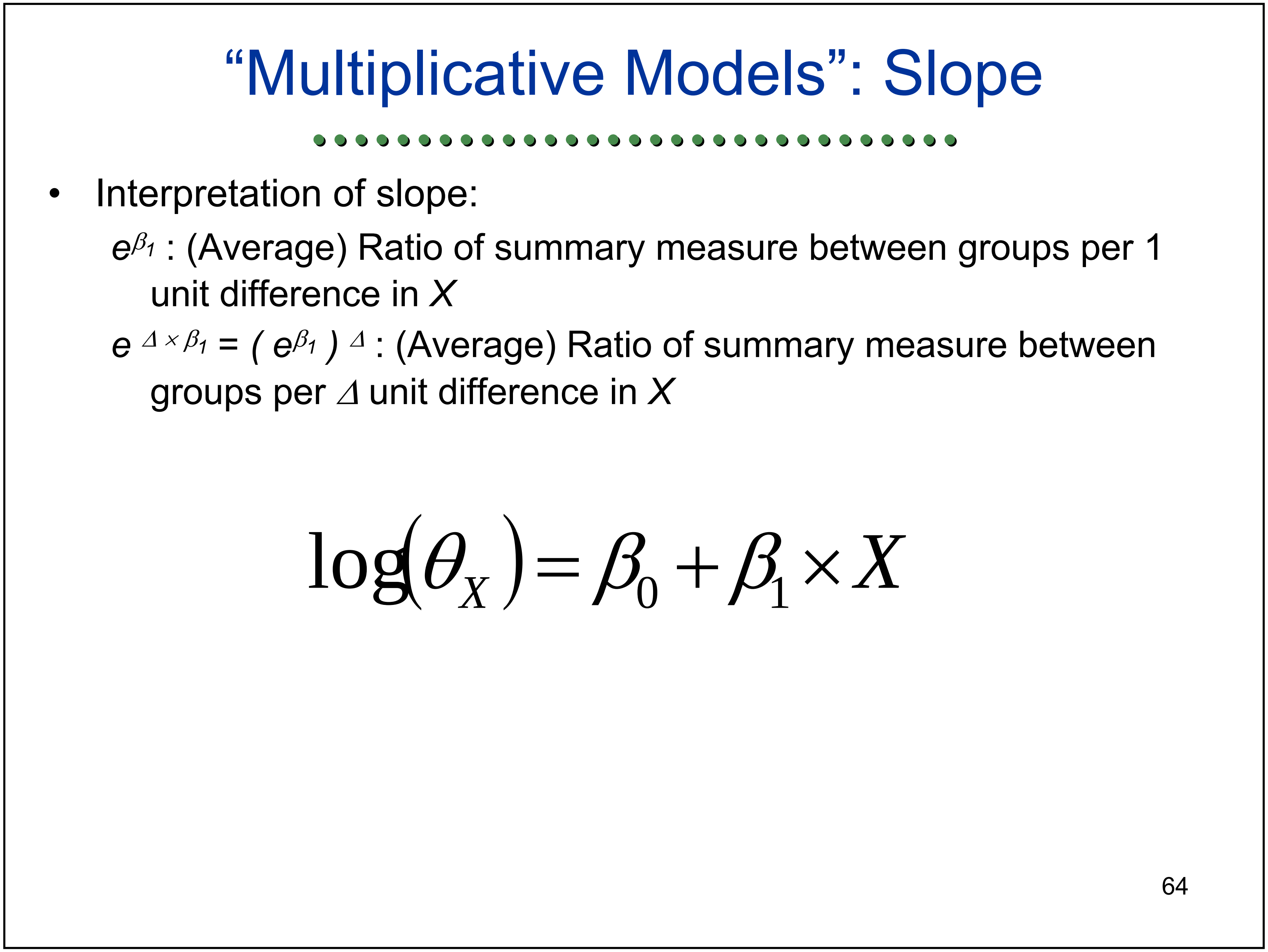
62



“Multiplicative Models”

- Log link function
  - – Geom means: linear regression on logs
  - – Odds: logistic regression
  - – Hazards: proportional hazard regression
  - – Means: Poisson regression
  - – Medians: accel failure time regression

$$\log(\theta_X) = \beta_0 + \beta_1 \times X$$



“Multiplicative Models”: Slope

- Interpretation of slope:

  e^{β_1} : (Average) Ratio of summary measure between groups per 1 unit difference in X

  $e^{\Delta \times \beta_1} = (e^{\beta_1})^{\Delta}$: (Average) Ratio of summary measure between groups per Δ unit difference in X

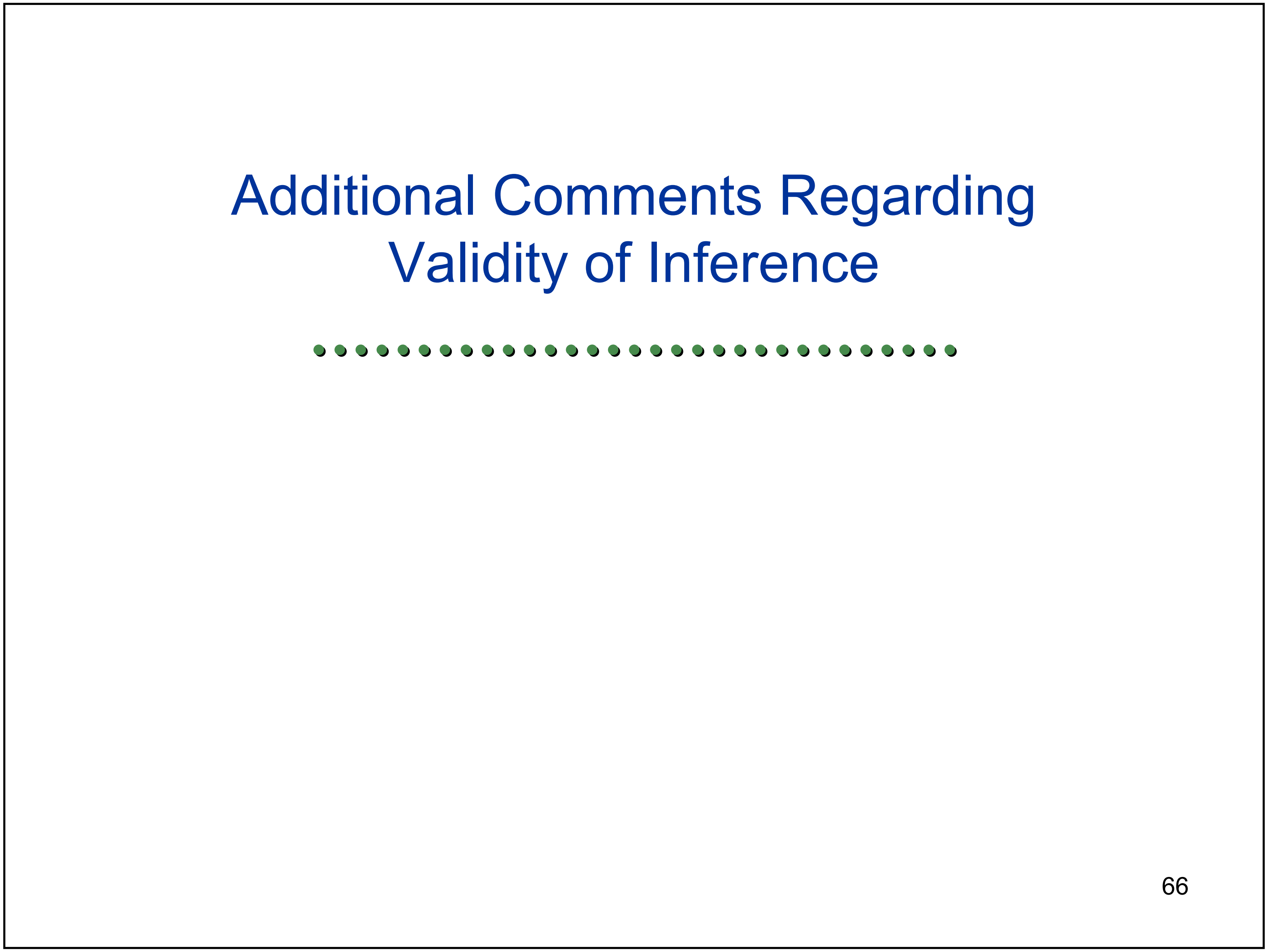
$$\log(\theta_X) = \beta_0 + \beta_1 \times X$$

“Multiplicative Models”: log(X)

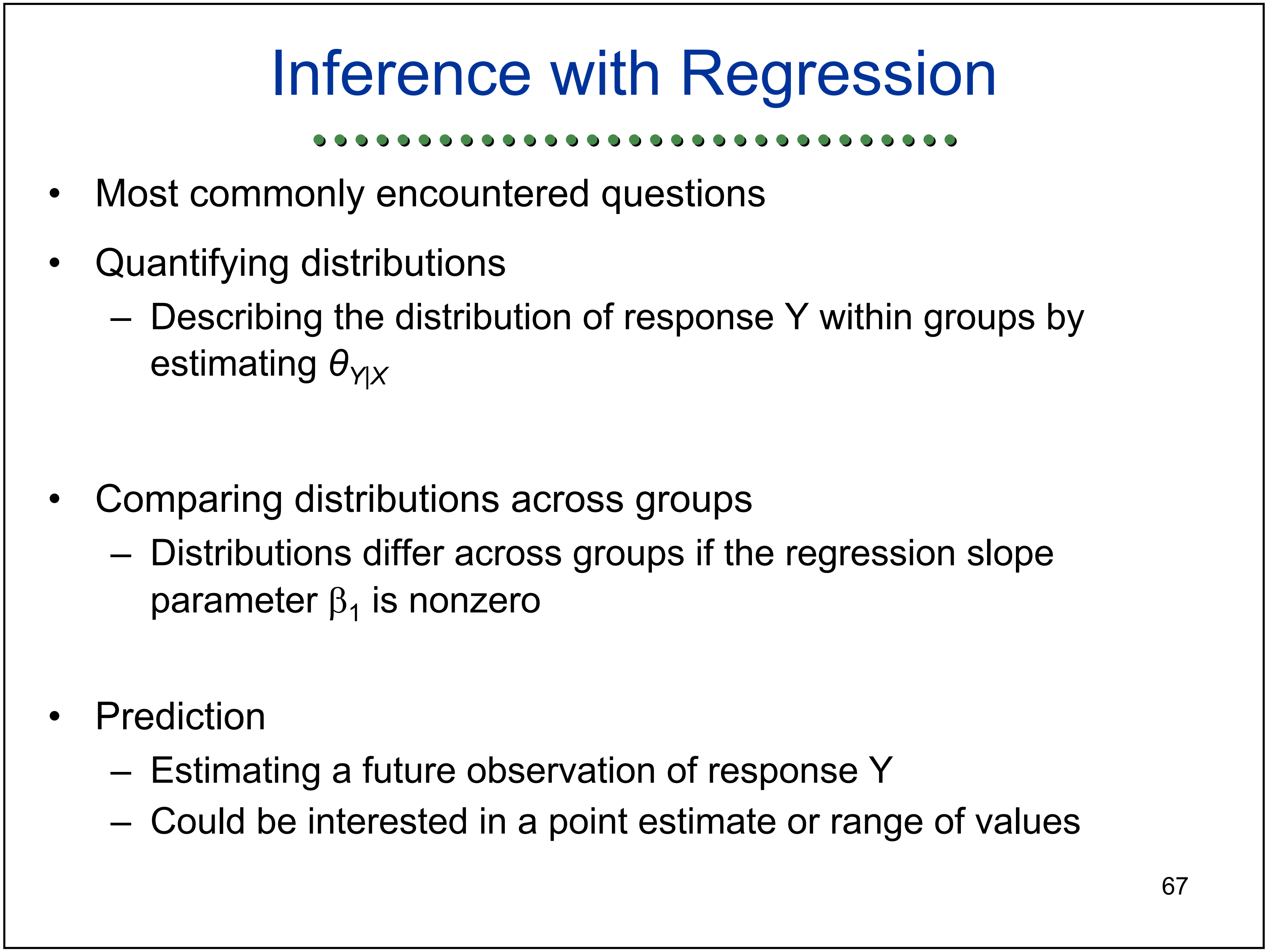
- Slope with log transformed predictor

  $e^{\log(k) \times \beta_1} = k^{\beta_1} = (e^{\beta_1})^{\log(k)}$: (Average) Ratio of summary measure between groups per *k*-fold difference in *X*

$$\log(\theta_X) = \beta_0 + \beta_1 \times \log(X)$$

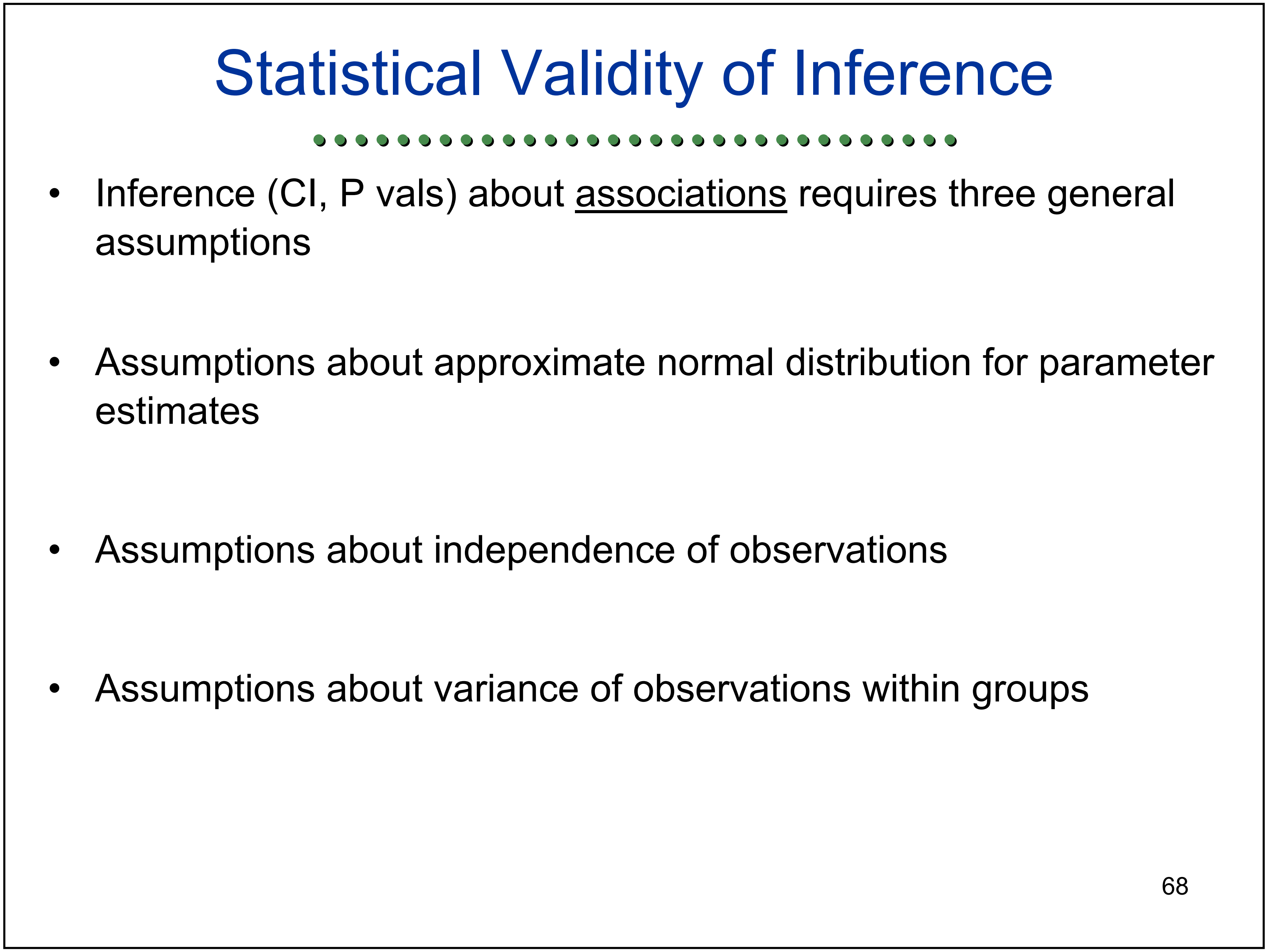


Additional Comments Regarding Validity of Inference



Inference with Regression

- Most commonly encountered questions
- Quantifying distributions
  - Describing the distribution of response Y within groups by estimating $\theta_{Y|X}$
- Comparing distributions across groups
  - Distributions differ across groups if the regression slope parameter β_1 is nonzero
- Prediction
  - Estimating a future observation of response Y
  - Could be interested in a point estimate or range of values



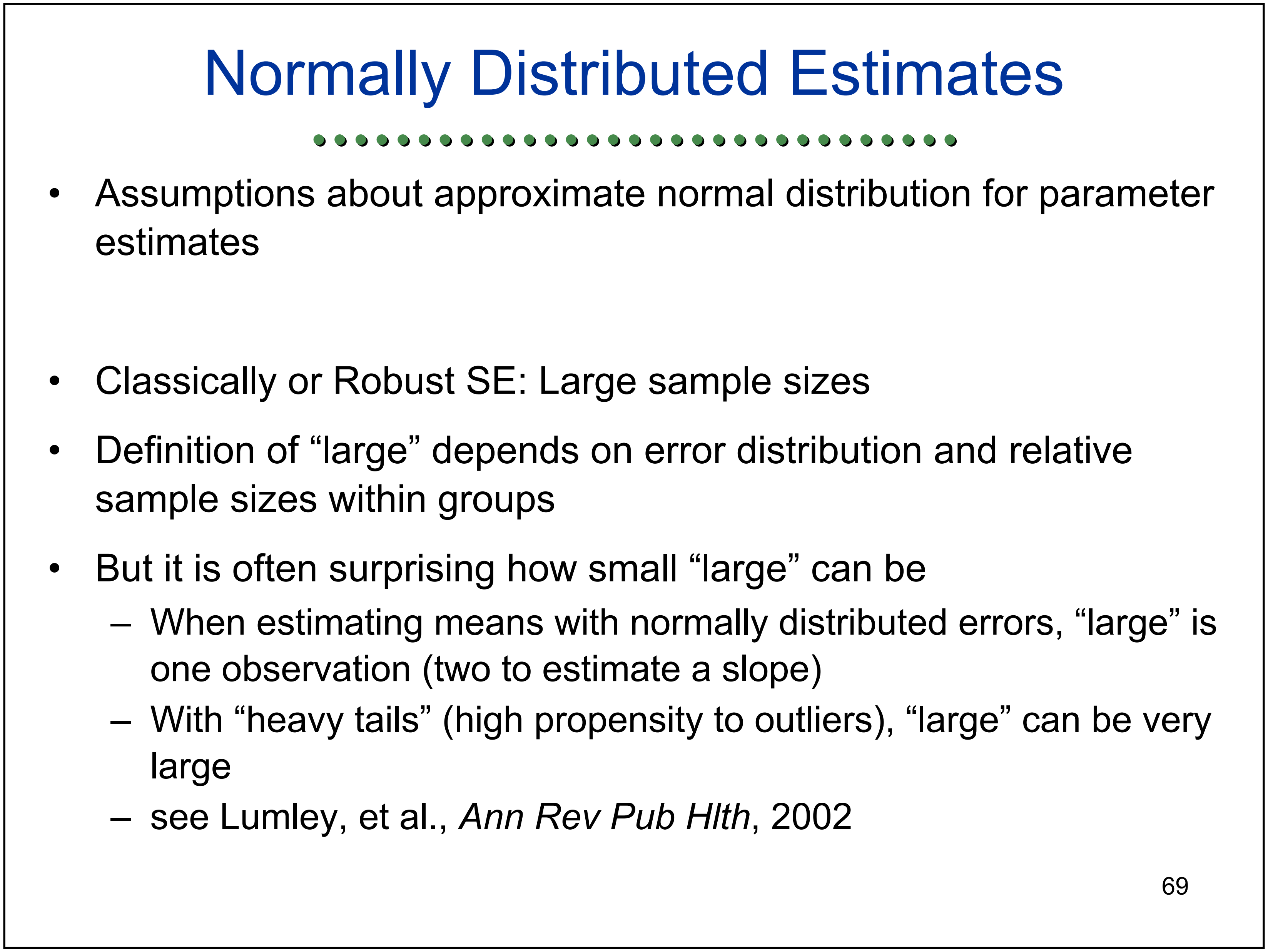
Statistical Validity of Inference

- Inference (CI, P vals) about associations requires three general assumptions

- Assumptions about approximate normal distribution for parameter estimates

- Assumptions about independence of observations

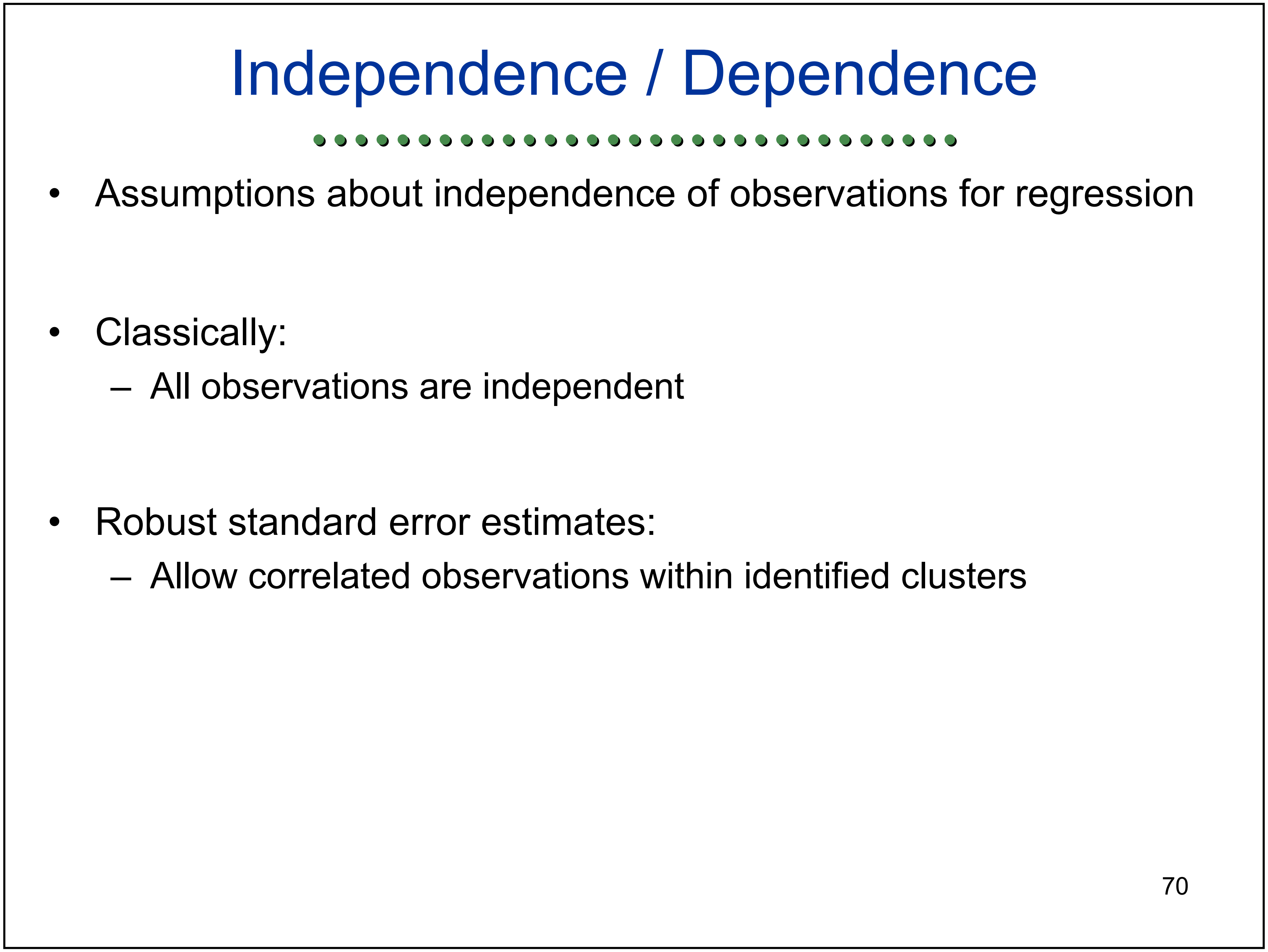
- Assumptions about variance of observations within groups



Normally Distributed Estimates

- Assumptions about approximate normal distribution for parameter estimates

- Classically or Robust SE: Large sample sizes
- Definition of “large” depends on error distribution and relative sample sizes within groups
- But it is often surprising how small “large” can be
  - When estimating means with normally distributed errors, “large” is one observation (two to estimate a slope)
  - With “heavy tails” (high propensity to outliers), “large” can be very large
  - see Lumley, et al., *Ann Rev Pub Hlth*, 2002

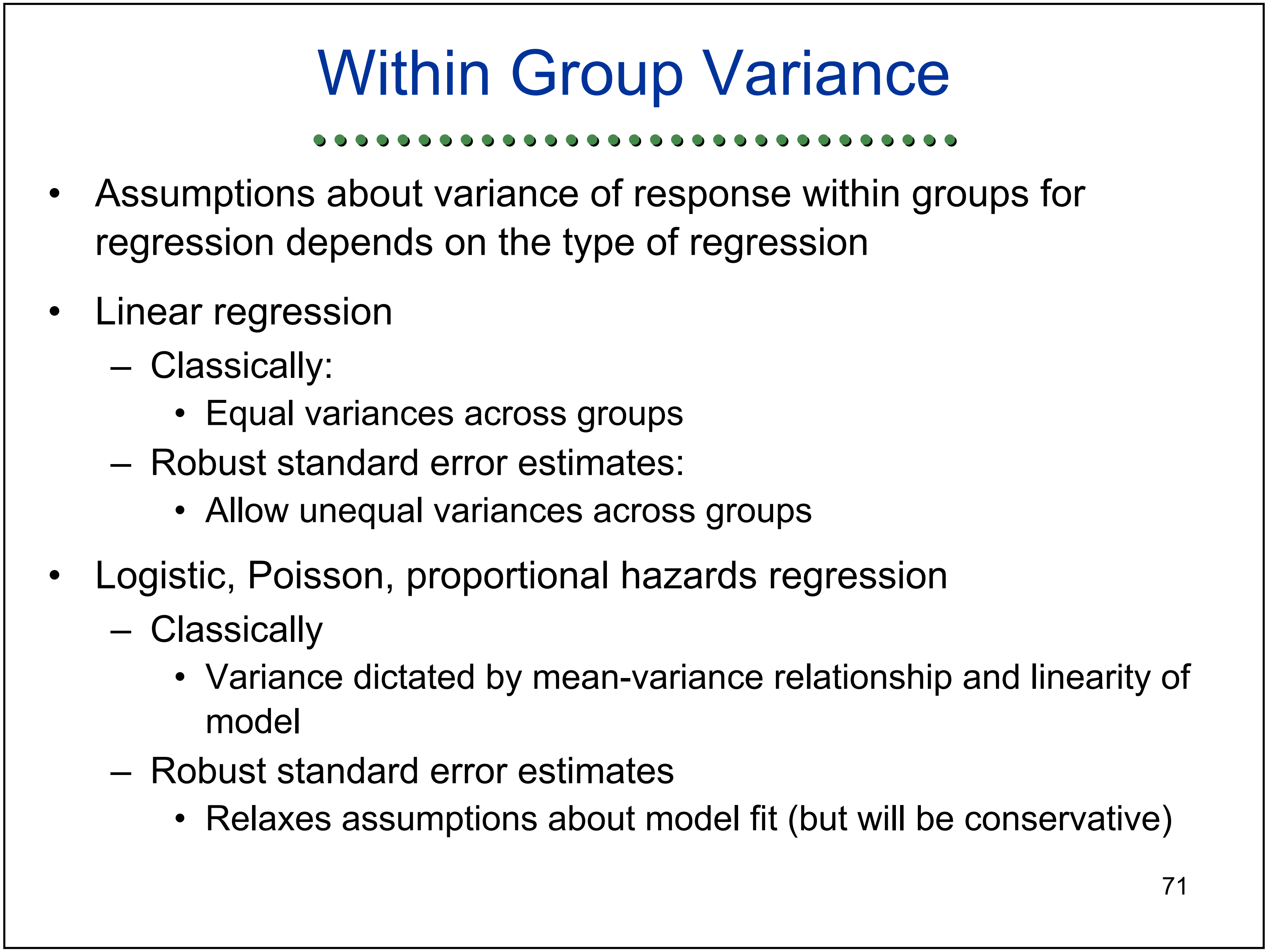


Independence / Dependence

- Assumptions about independence of observations for regression

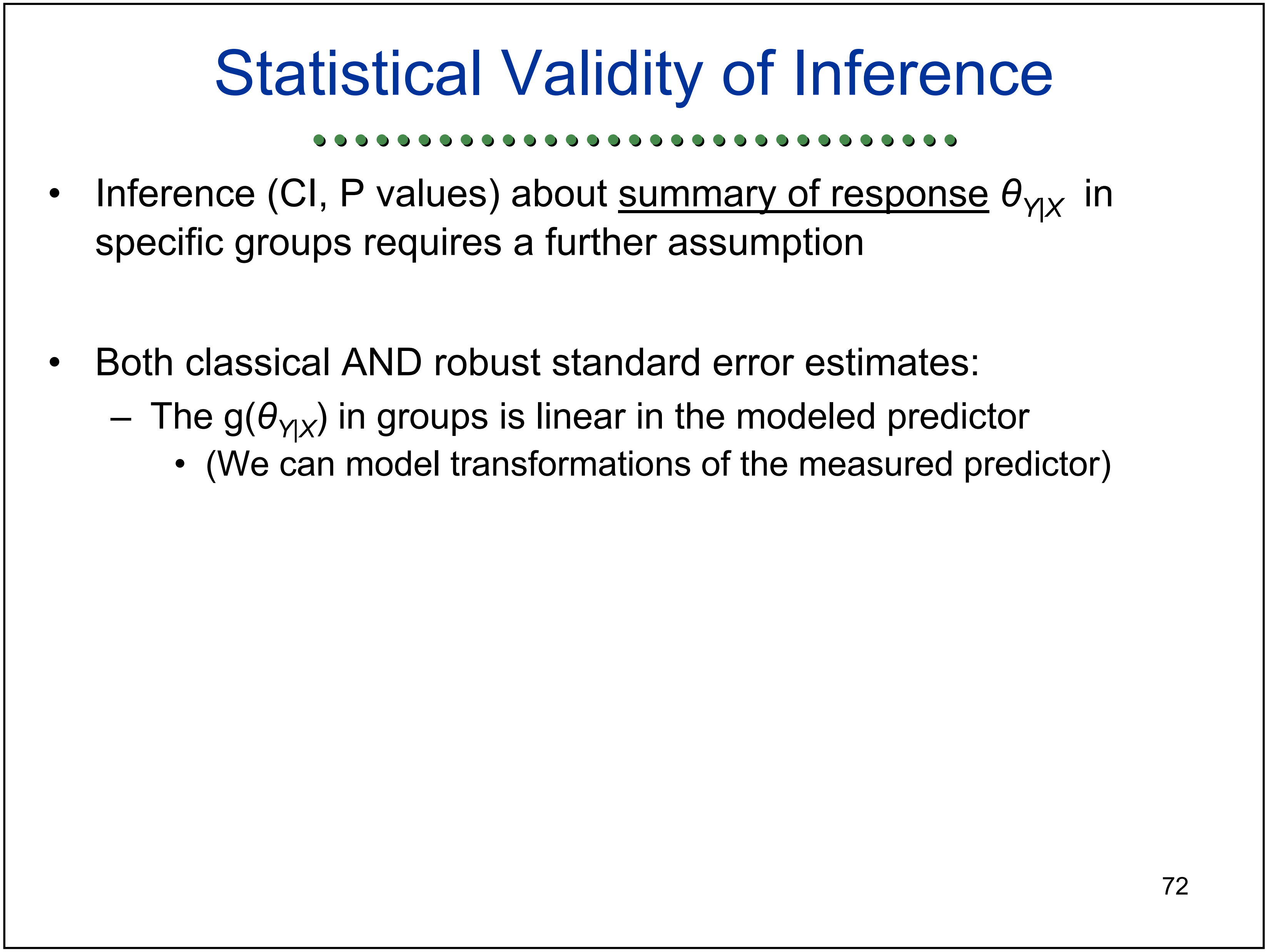
- Classically:
  - All observations are independent

- Robust standard error estimates:
  - Allow correlated observations within identified clusters



Within Group Variance

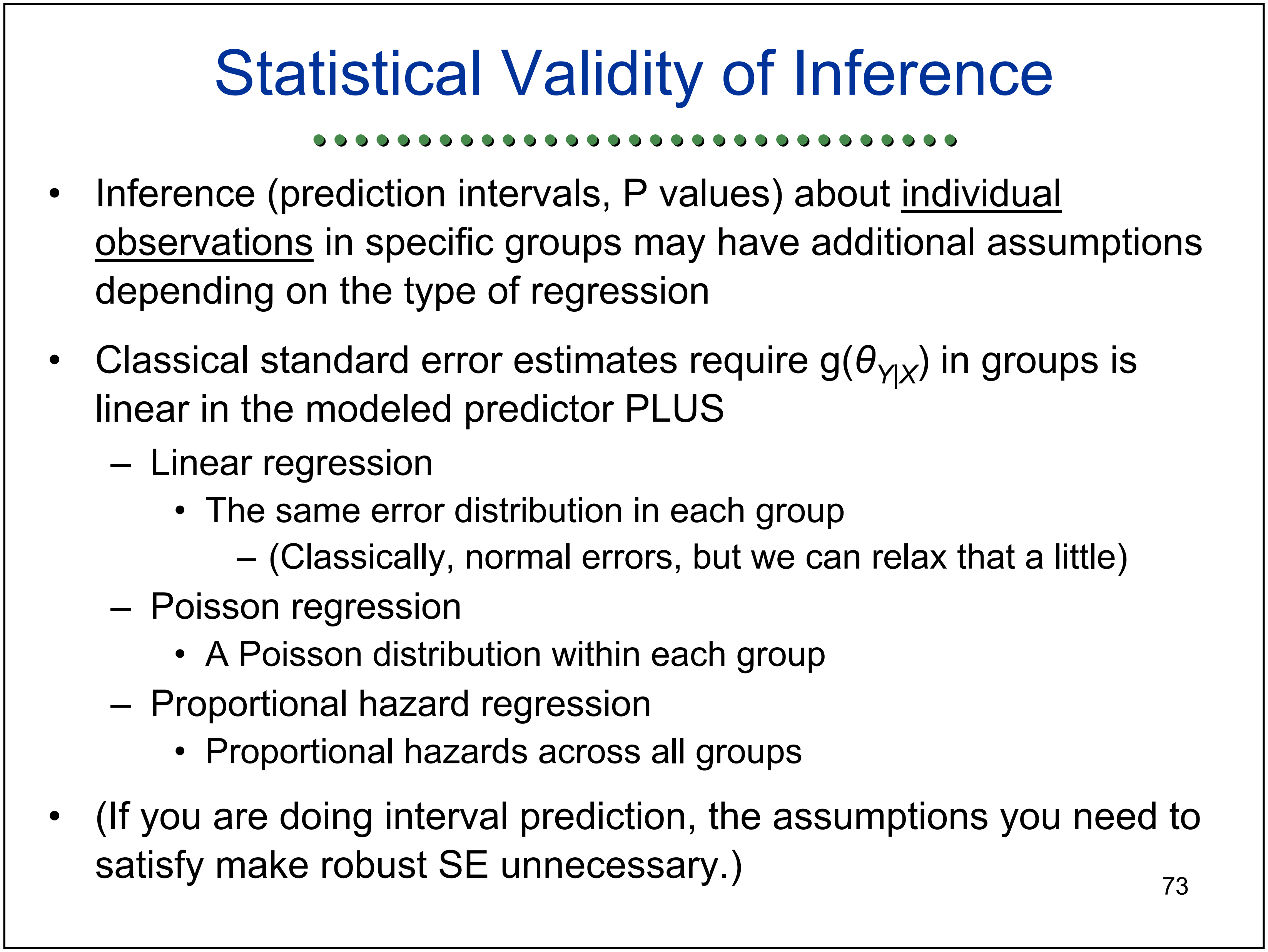
- Assumptions about variance of response within groups for regression depends on the type of regression
- Linear regression
  - Classically:
    - Equal variances across groups
  - Robust standard error estimates:
    - Allow unequal variances across groups
- Logistic, Poisson, proportional hazards regression
  - Classically
    - Variance dictated by mean-variance relationship and linearity of model
  - Robust standard error estimates
    - Relaxes assumptions about model fit (but will be conservative)



Statistical Validity of Inference

- Inference (CI, P values) about summary of response $\theta_{Y|X}$ in specific groups requires a further assumption

- Both classical AND robust standard error estimates:
  - The $g(\theta_{Y|X})$ in groups is linear in the modeled predictor
    - (We can model transformations of the measured predictor)



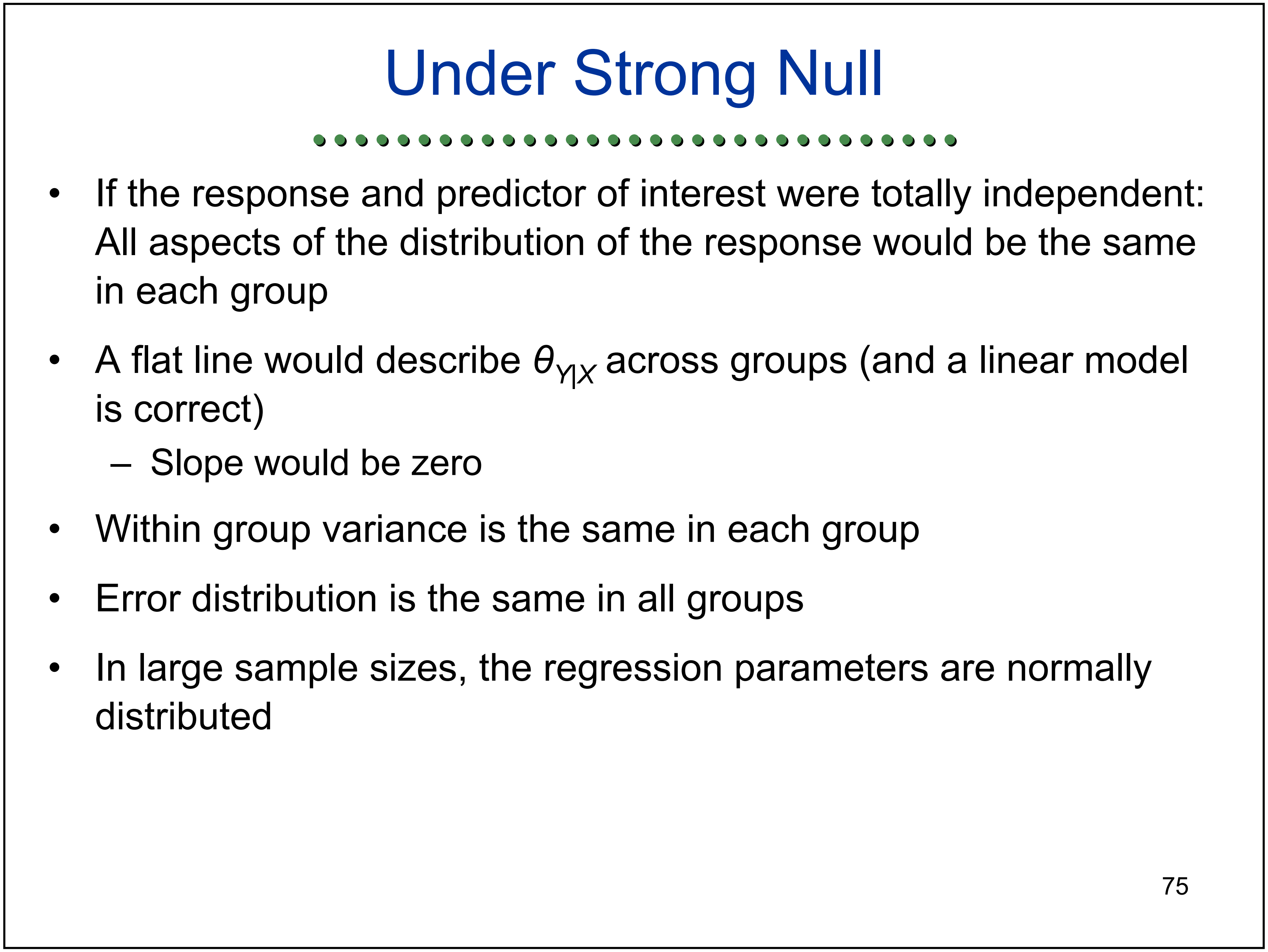
Statistical Validity of Inference

- Inference (prediction intervals, P values) about individual observations in specific groups may have additional assumptions depending on the type of regression
- Classical standard error estimates require g($\theta_{Y|X}$) in groups is linear in the modeled predictor PLUS
  - Linear regression
    - The same error distribution in each group
      - (Classically, normal errors, but we can relax that a little)
  - Poisson regression
    - A Poisson distribution within each group
  - Proportional hazard regression
    - Proportional hazards across all groups
- (If you are doing interval prediction, the assumptions you need to satisfy make robust SE unnecessary.)

Implications for Inference

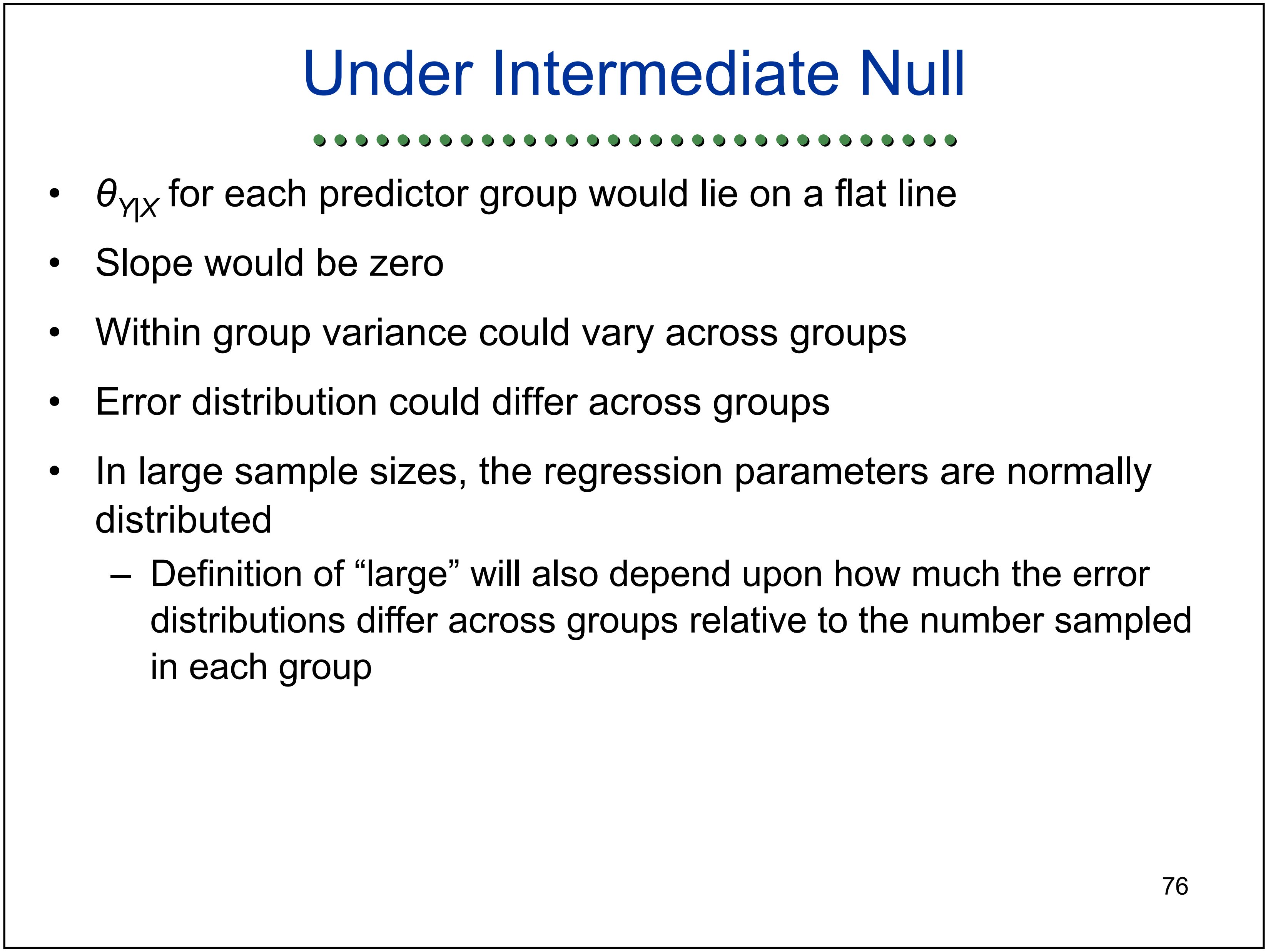
- Regression based inference about associations is far more robust than estimation of group means or individual predictions

- A hierarchy of null hypotheses
  - Strong null: Total independence of *Y* and *X*
  - Intermediate null: Mean of *Y* the same for all *X* groups
  - Weak null: No linear trend in mean of *Y* across *X* groups



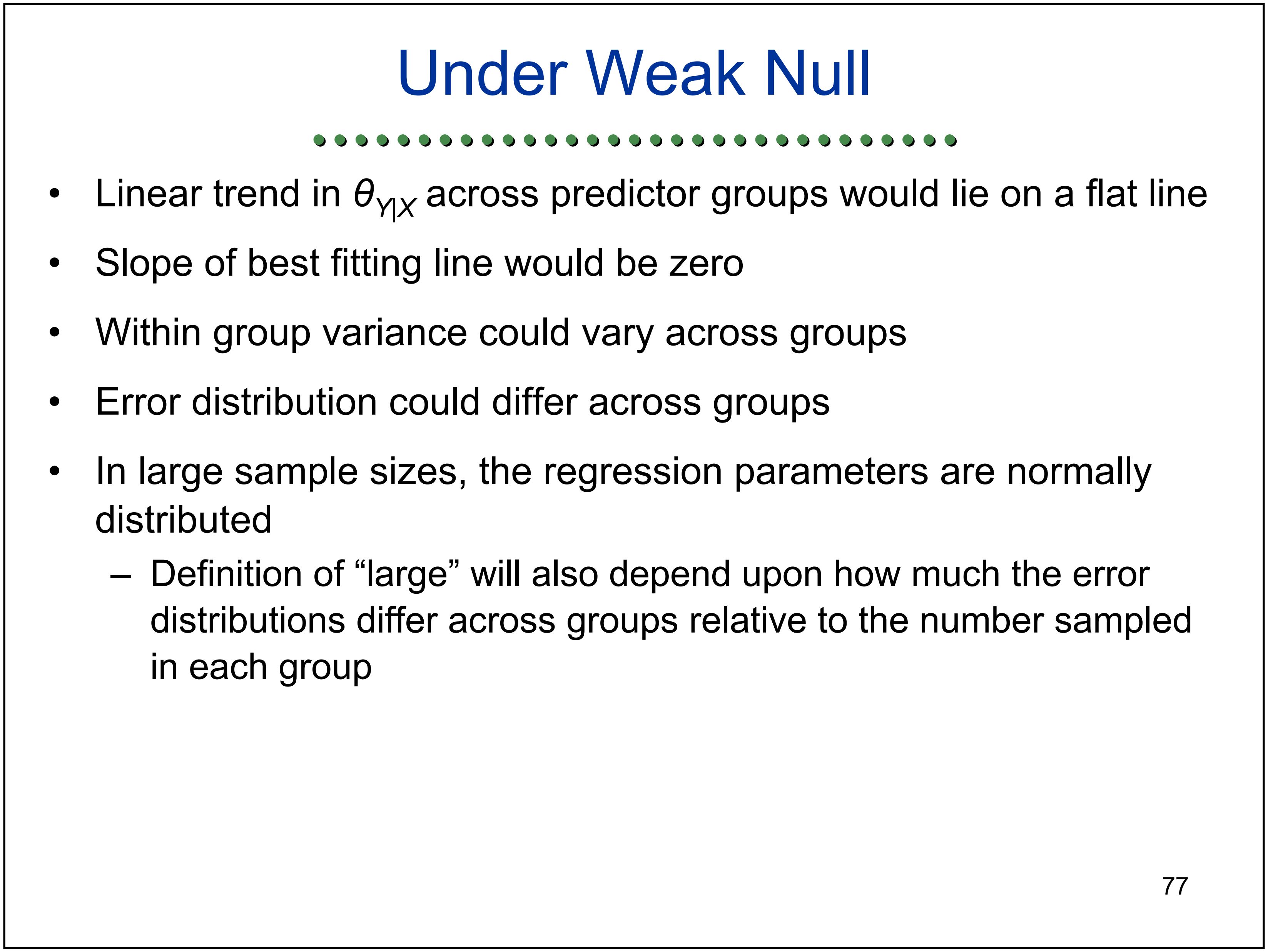
Under Strong Null

- If the response and predictor of interest were totally independent: All aspects of the distribution of the response would be the same in each group
- A flat line would describe $\theta_{Y|X}$ across groups (and a linear model is correct)
  - Slope would be zero
- Within group variance is the same in each group
- Error distribution is the same in all groups
- In large sample sizes, the regression parameters are normally distributed



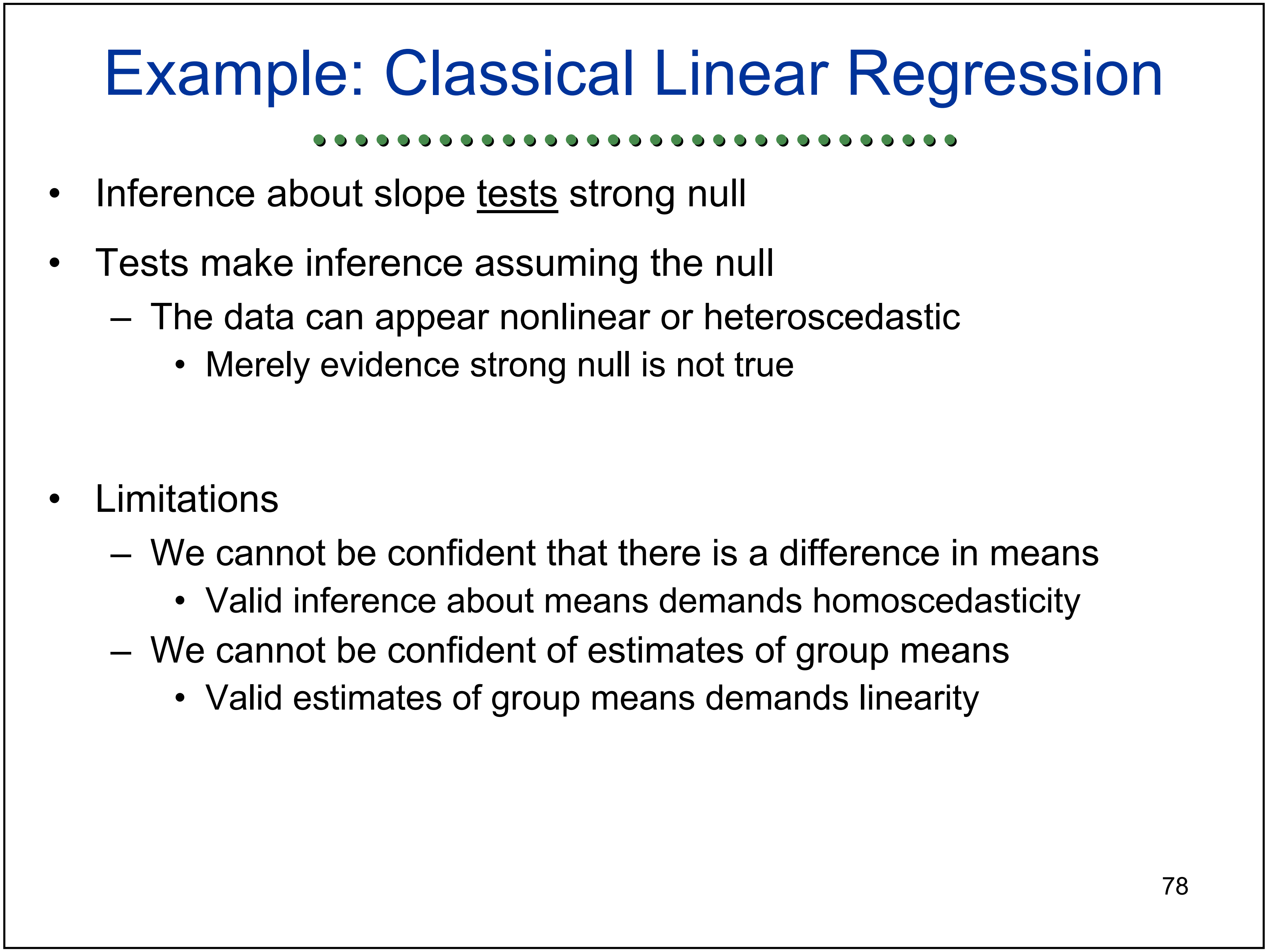
Under Intermediate Null

- $\theta_{Y|X}$ for each predictor group would lie on a flat line
- Slope would be zero
- Within group variance could vary across groups
- Error distribution could differ across groups
- In large sample sizes, the regression parameters are normally distributed
  - Definition of “large” will also depend upon how much the error distributions differ across groups relative to the number sampled in each group



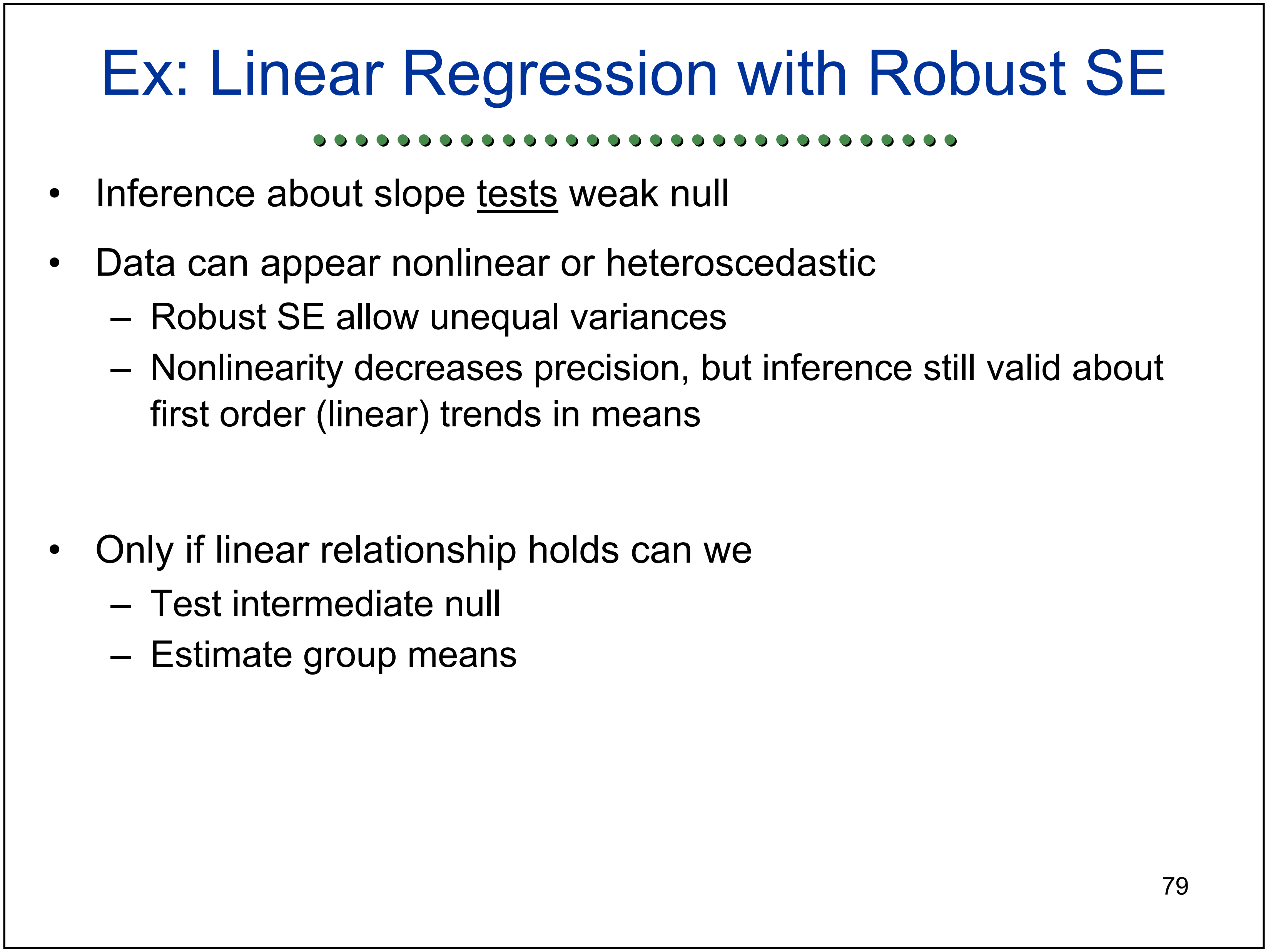
Under Weak Null

- Linear trend in $\theta_{Y|X}$ across predictor groups would lie on a flat line
- Slope of best fitting line would be zero
- Within group variance could vary across groups
- Error distribution could differ across groups
- In large sample sizes, the regression parameters are normally distributed
  - Definition of “large” will also depend upon how much the error distributions differ across groups relative to the number sampled in each group



Example: Classical Linear Regression

- Inference about slope tests strong null
- Tests make inference assuming the null
  - The data can appear nonlinear or heteroscedastic
    - Merely evidence strong null is not true
- Limitations
  - We cannot be confident that there is a difference in means
    - Valid inference about means demands homoscedasticity
  - We cannot be confident of estimates of group means
    - Valid estimates of group means demands linearity



Ex: Linear Regression with Robust SE

- Inference about slope <u>tests</u> weak null
- Data can appear nonlinear or heteroscedastic
  - Robust SE allow unequal variances
  - Nonlinearity decreases precision, but inference still valid about first order (linear) trends in means
- Only if linear relationship holds can we
  - Test intermediate null
  - Estimate group means

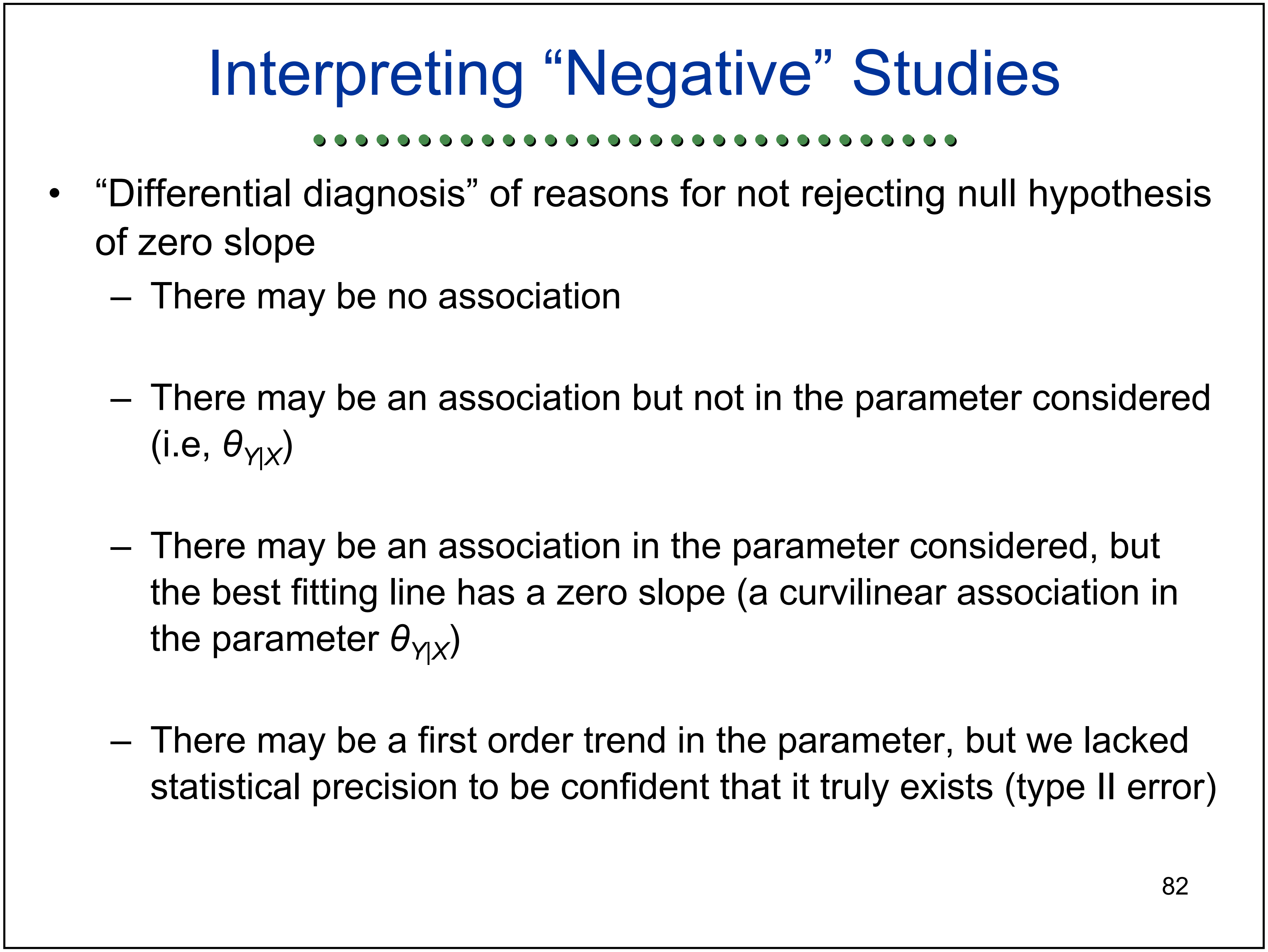
Implications for Inference

- Inference about associations is far more trustworthy than estimation of group $\theta_{Y|X}$ or individual predictions

- Nonzero slope suggests an association between response and predictor
  - Inference about linear trends in $\theta_{Y|X}$ if use robust SE

Interpreting “Positive” Results

- If slope is statistically significant different from 0 using robust SE
  - Observed data is atypical of a setting with no linear trend in $\theta_{Y|X}$ response across groups
    - Data suggests evidence of a trend toward larger (smaller) $\theta_{Y|X}$ in groups having larger values of the predictor
    - (To the extent the data appears linear, estimates of the group $\theta_{Y|X}$ will be reliable)
  - Data may be typical of a setting with no linear trend in $\theta_{Y|X}$ response across groups, but we were unlucky in our sampling



Interpreting “Negative” Studies

- “Differential diagnosis” of reasons for not rejecting null hypothesis of zero slope
  - There may be no association
  - There may be an association but not in the parameter considered (i.e, $\theta_{Y|X}$)
  - There may be an association in the parameter considered, but the best fitting line has a zero slope (a curvilinear association in the parameter $\theta_{Y|X}$)
  - There may be a first order trend in the parameter, but we lacked statistical precision to be confident that it truly exists (type II error)

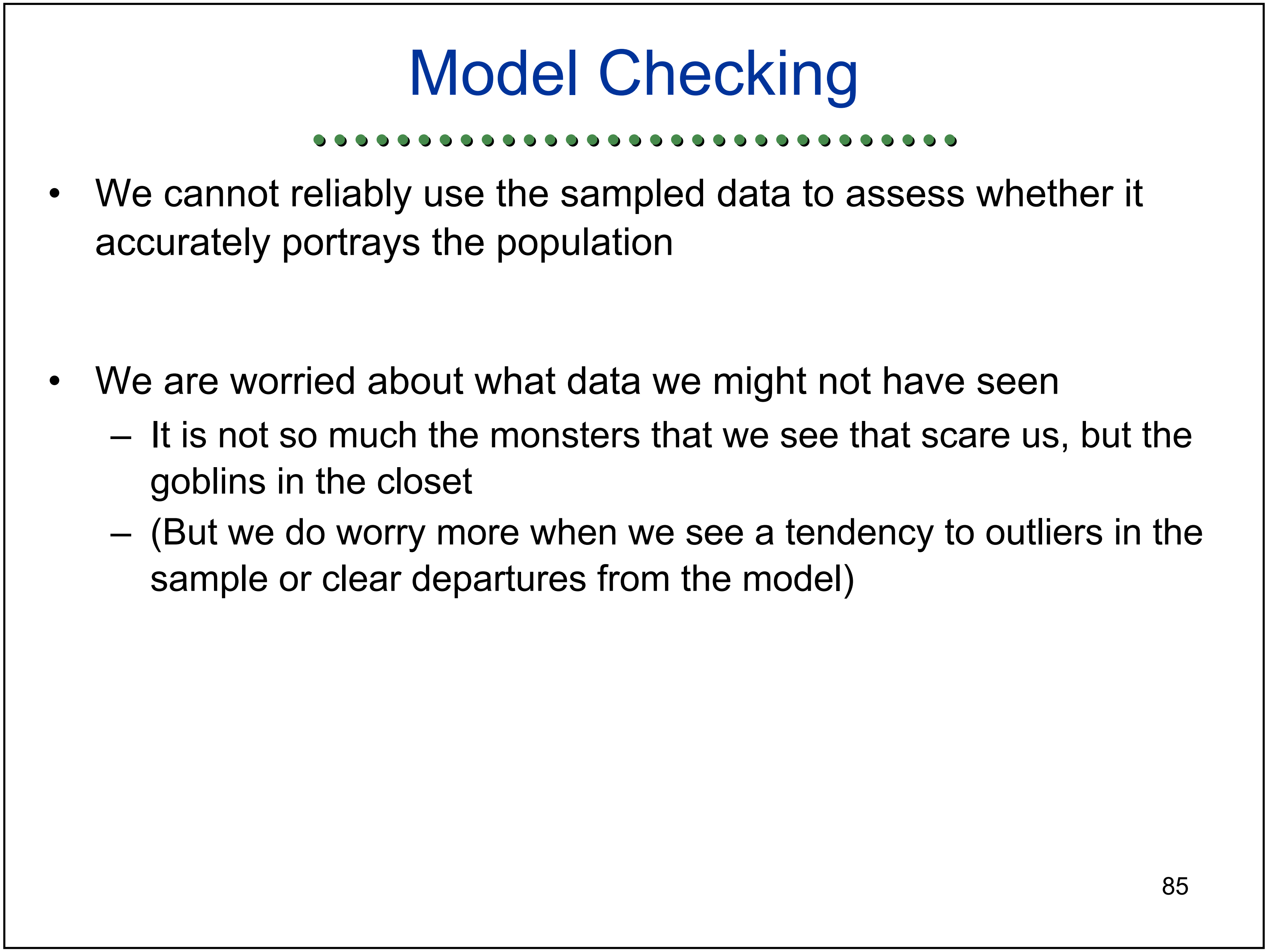
Model Checking

- Much statistical literature has been devoted to means of checking the assumptions for regression models
- I believe model checking is generally fraught with peril, as it necessarily involves multiple comparisons

Model Checking

“Blood suckers hide ‘neath my bed”

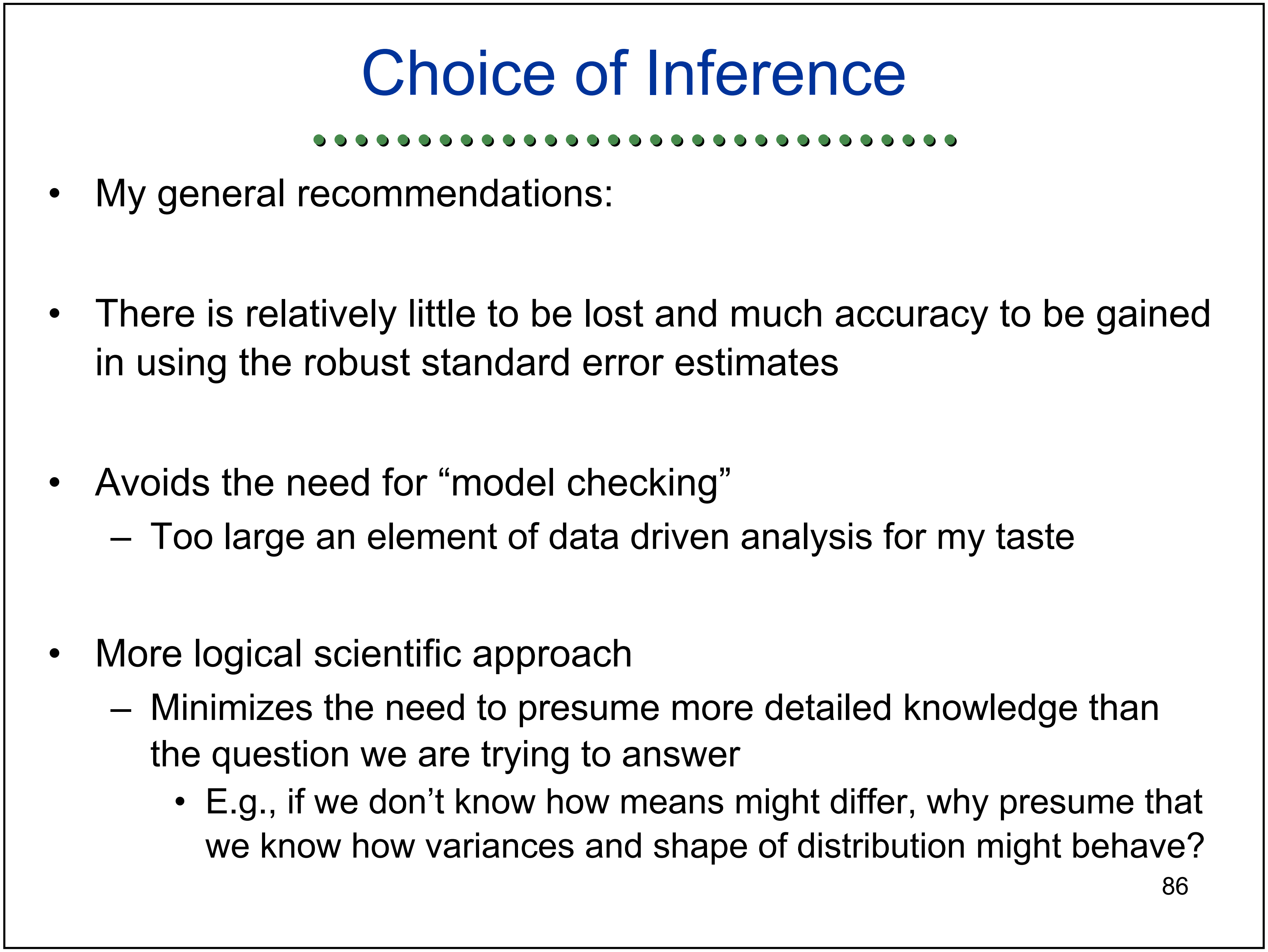
“Eyepennies”, Mark Linkous (Sparklehorse)



Model Checking

- We cannot reliably use the sampled data to assess whether it accurately portrays the population

- We are worried about what data we might not have seen
  - It is not so much the monsters that we see that scare us, but the goblins in the closet
  - (But we do worry more when we see a tendency to outliers in the sample or clear departures from the model)



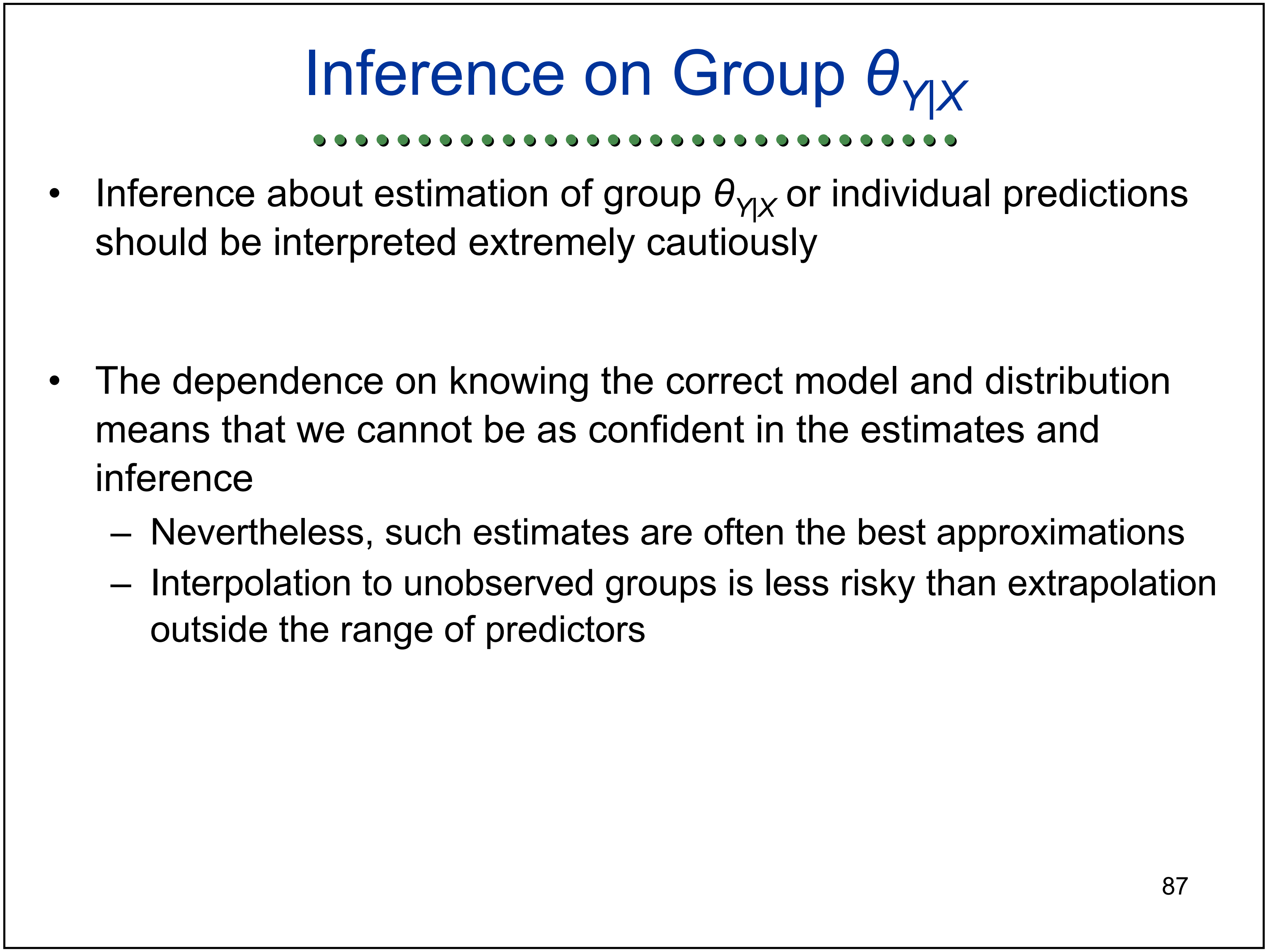
Choice of Inference

- My general recommendations:

- There is relatively little to be lost and much accuracy to be gained in using the robust standard error estimates

- Avoids the need for “model checking”
  - Too large an element of data driven analysis for my taste

- More logical scientific approach
  - Minimizes the need to presume more detailed knowledge than the question we are trying to answer
    - E.g., if we don’t know how means might differ, why presume that we know how variances and shape of distribution might behave?



Inference on Group $\theta_{Y|X}$

- Inference about estimation of group $\theta_{Y|X}$ or individual predictions should be interpreted extremely cautiously

- The dependence on knowing the correct model and distribution means that we cannot be as confident in the estimates and inference
  - Nevertheless, such estimates are often the best approximations
  - Interpolation to unobserved groups is less risky than extrapolation outside the range of predictors