Biost 518
Applied Biostatistics II

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

Lecture 4:
Review of Simple Regression

January 13, 2006

Lecture Outline

- General Regression Setting
- Inference on Means
- Inference about Geometric Means
- Inference about Odds
- Inference about Hazards

General Regression Setting

Two Variable Setting

- Many statistical problems consider the association between two variables
  - Response variable
    - (outcome, dependent variable)
  - Grouping variable
    - (predictor, independent variable)

Addressing Scientific Question

- Compare the distribution of the response variable across groups that are defined by the grouping variable
  - Within each group, the value of the grouping variable is constant

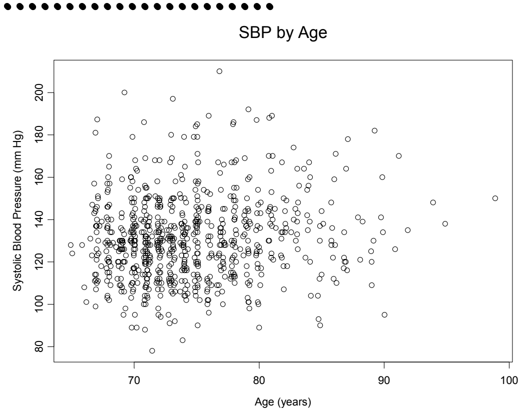
5

Intro Course Classification

- Characterize statistical analyses by
  - Number of samples (groups), and
  - Whether subjects in groups are independent
- Correspondence with two variable setting
  - By characterization of grouping variable
    - Constant: One sample problem
    - Binary: Two sample problem
    - Categorical: k sample problem (e.g., ANOVA)
    - Continuous: Infinite sample problem
      - Regression

6

Example: SBP and Age



7

Regression Methods

- Regression extends one and two sample statistics (e.g., the t test) to the infinite sample problem
  - While we don't really ever have (or care) about an infinite number of samples, it is easiest to use models that would allow that in order to handle
    - Continuous predictors of interest
    - Adjustment for other variables

8

Regression vs Two Samples

- When used with a binary grouping variable common regression models reduce to the corresponding two variable methods
  - Linear regression with a binary predictor
    - Classical: t test with equal variance
    - Robust SE: t test with unequal variance (approx)
  - Logistic regression with a binary predictor
    - Score test: Chi squared test for association
  - Cox regression with a binary predictor
    - Score test: Logrank test

Guiding Principle

"Everything is regression."
- Scott Emerson

Uses of Regression

- Two major uses of regression
  - Borrow information to address "sparse data" in some groups
    - E.g., 68 and 70 year olds provide information about 69 year olds
    - Question: How far away do you want to go?
  - Provide a statistical "contrast" to compare distribution of response across groups
    - Think of a "slope" as an average comparison of summary measures per unit difference in the grouping variable

Regression Inference

- Estimates
  - Slope: (average) contrasts across groups
  - Fitted values: estimated summary measure in a group
- Standard errors
- Confidence intervals
- P values testing for
  - Intercept of zero (who cares?)
  - Slope of zero (test for linear trend in summary measures)

Robust Standard Errors

- I have recommended the use of robust standard errors
  - Relaxes assumptions about variance of data within groups
  - Allows tests of weak null hypotheses
    - Statements about equality of summary measures rather than equality of entire distributions
  - Soon: Allows regression with correlated data

Simple Linear Regression

Interpretation

- Interpretation of "regression parameters"
  - Intercept β_0: Mean Y for a group with X=0
    - Quite often not of scientific interest
      - Often outside range of data, sometimes impossible
  - Slope β_1: Difference in mean Y across groups differing in X by 1 unit
    - Usually measures association between Y and X

$$E(Y \mid X) = \beta_0 + \beta_1 \times X$$

Derivation of Interpretation

- Simple linear regression of response Y on predictor X
  - Mean for an arbitrary group derived from model
  - Interpretation of parameters by considering special cases

Model	$E[Y_i \mid X_i] = \beta_0 + \beta_1 \times X_i$
$X_i = 0$	$E[Y_i \mid X_i = 0] = \beta_0$
$X_i = x$	$E[Y_i \mid X_i = x] = \beta_0 + \beta_1 \times x$
$X_i = x+1$	$E[Y_i \mid X_i = x+1] = \beta_0 + \beta_1 \times x + \beta_1$

Example: Mental Function by Age

- Cardiovascular Health Study
  - A cohort of ~5,000 elderly subjects in four communities followed with annual visits
    - A subset of 735 subjects
  - Mental function measured at baseline by Digit Symbol Substitution Test (DSST)
  - Question: How does performance on DSST differ across age groups

17

Example: Scatterplot

18

Statistical Validity of Inference

- Inference (CI, P vals) about associations requires three general assumptions
  - Approximate normal distribution for estimates
    - Normal data or large N
  - Assumptions about independence of observations
    - Independence or identified clusters
  - Assumptions about variance of observations within groups
    - Robust SE: relaxes requirement for equal variance

19

Prediction of Group Means

- Additional ssumption about adequacy of linear model for prediction of group means with linear regression
  - Classically OR robust standard error estimates:
    - The mean response in groups is linear in the modeled predictor
      - (We can model transformations of the measured predictor)

20

Prediction Intervals

- Inference (prediction intervals) about individual observations in specific groups has still another assumption
  - Assumption about distribution of errors within each group
    - Normally distributed errors

Regression in Stata

- Inference based on either classical linear regression or robust standard errors
  - Classical linear regression
    - "`regress respvar predictor`"
      - E.g., `regress dsst age`
  - Robust standard error estimates
    - "`regress respvar predictor, robust`"
      - E.g., `regress dsst age, robust`
  - The two approaches differ in CI and P values, not estimates

Ex: Robust Standard Errors

```
. regress dsst age, robust
Linear regression
                                              Number of obs =      723
                                              F(  1,    721) =  130.72
                                              Prob > F      =  0.0000
                                              R-squared     =  0.1319
                                              Root MSE      =  11.847

       |        Robust
  dsst |   Coef StdErr       t    P>|t|   [95% Conf  Int]
   age | -.863   .0755  -11.43   0.000    -1.01   -.715
 _cons |   105    5.71   18.45   0.000     94.1     117
```


Interpretation of Intercept

$$E\left[DSST_i \,\middle|\, Age_i\right] = 105 - 0.863 \times Age_i$$

- Estimated mean DSST for newborns is 105
  - Pretty ridiculous estimate
    - We never sampled anyone less than 67
    - Maximum value for DSST is 100
    - Newborns would in fact (rather deterministically) score 0
- In this problem, the intercept is just a mathematical construct to fit a line over the range of our data

Interpretation of Slope

$$E[DSST_i \mid Age_i] = 105 - 0.863 \times Age_i$$

- Estimated difference in mean DSST for two groups differing by one year in age is -0.863, with older group averaging a lower score
  - For 5 year age difference: 5 x -0.863 = - 4.32
  - For 10 year age difference: - 8.63
- (If a straight line relationship is not true, we interpret the slope as an average difference in mean DSST per one year difference in age)

Robust Standard Errors

- Inference for association based on slope
  - Weak null based inference
    - Estimated trend in mean DSST by age is an average difference of -.863 per one year differences in age (DSST lower in older)
    - CI for trend: -1.01, -0.715
    - P value < .0001 suggests mean DSST differs across age groups
      - T statistic: -11.43 (Who cares?)

Inference for Correlation

- Hypothesis tests for a nonzero correlation are EXACTLY the same as a test for a nonzero slope in classical linear regression
  - Interestingly:
    - The statistical significance of a given value of r depends only on the sample size
      - Correlation is far more of a statistical than a scientific measure

Regression and t Tests

- Linear regression with a binary predictor (two groups) corresponds to familiar t tests
  - Classical linear regression: Two sample t test which presumes equal variances (exactly the same)
  - Robust standard error estimates: Two sample t test which allows unequal variances (nearly the same)
  - Identified clusters with robust standard error estimates: Paired t test (nearly the same)

Inference for the Geometric Mean

Simple Linear Regression on Log Transformed Data

Regression on Geometric Means

- Geometric means of distributions are typically analyzed by using linear regression on log transformed data
  - Common choice for inference when a positive response variable is continuous, and
    - we are interested in multiplicative models,
    - we desire to downweight outliers, and/or
    - the standard deviation of response in a group is proportional to the mean
      - "Error is +/- 10%" instead of "Error is +/- 10"

Interpretation of Parameters

- Linear regression on log transformed Y
  - (I am using natural log)

Model $$E[\log Y_i \mid X_i] = \beta_0 + \beta_1 \times X_i$$

$X_i = 0$ $$E[\log Y_i \mid X_i = 0] = \beta_0$$

$X_i = x$ $$E[\log Y_i \mid X_i = x] = \beta_0 + \beta_1 \times x$$

$X_i = x+1$ $$E[\log Y_i \mid X_i = x+1] = \beta_0 + \beta_1 \times x + \beta_1$$

Interpretation of Parameters

- Restated model as log link for geometric mean

Model $$\log \mathrm{GM}[Y_i \mid X_i] = \beta_0 + \beta_1 \times X_i$$

$X_i = 0$ $$\log GM[Y_i \mid X_i = 0] = \beta_0$$

$X_i = x$ $$\log GM[Y_i \mid X_i = x] = \beta_0 + \beta_1 \times x$$

$X_i = x+1$ $$\log GM[Y_i \mid X_i = x+1] = \beta_0 + \beta_1 \times x + \beta_1$$

Interpretation of Parameters

- Interpretation of regression parameters by back-transforming model
  - Exponentiation is inverse of log

Model $\quad GM[Y_i \mid X_i] = e^{\beta_0} \times e^{\beta_1 \times X_i}$

$X_i = 0 \quad GM[Y_i \mid X_i = 0] = e^{\beta_0}$

$X_i = x \quad GM[Y_i \mid X_i = x] = e^{\beta_0} \times e^{\beta_1 \times x}$

$X_i = x + 1 \quad GM[Y_i \mid X_i = x+1] = e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}$

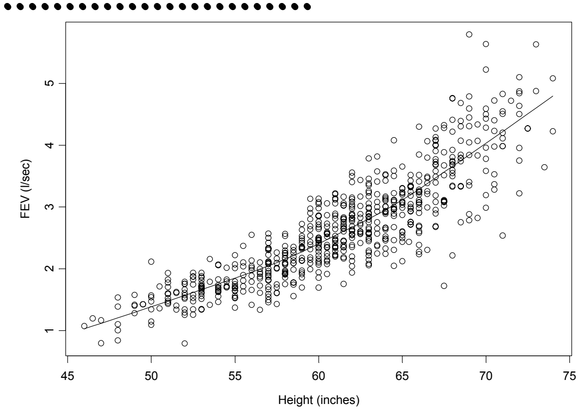
Interpretation of Parameters

- Geometric mean when predictor is 0
  - Found by exponentiation of the intercept from the linear regression on log transformed data: exp(β0)
- Ratio of geometric means between groups differing in the value of the predictor by 1 unit
  - Found by exponentiation of the slope from the linear regression on log transformed data: exp(β1)
- Confidence intervals for geometric mean and ratios found by exponentiating the CI for regression parameters

Example

- Trends in FEV with height
  - FEV data set
    - A sample of 654 healthy children
    - Lung function measured by forced expiratory volume (FEV)
      - maximal amount of air expired in 1 second
    - Question: How does FEV differ across height groups

FEV versus Height



Choice of Summary Measure

- Scientific justification for geometric mean
  - FEV is a volume
  - Height is a linear dimension
    - – Each dimension of lung size is proportional to height
  - Standard deviation likely proportional to height

Science $FEV \propto Height^3$

$\sqrt[3]{FEV} \propto Height$

Statistics $\log(FEV) \propto 3\log(Height)$

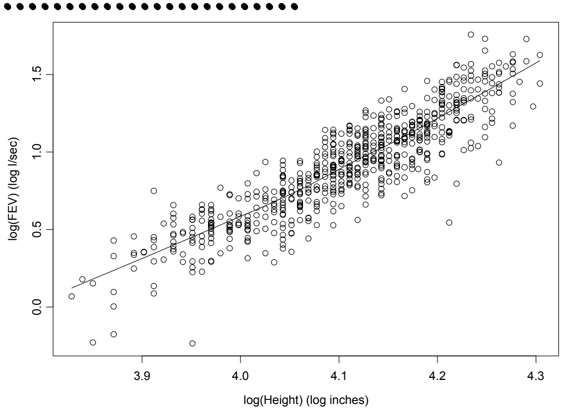
37

Model Geometric Mean

- Science dictates any of the models
  - – Statistical preference for transformation of response
    - May transform to equal variance across groups
    - “Homoscedasticity” allows easier inference
  - – Statistical preference for log transformation
    - Easier interpretation: multiplicative model
    - Compare groups using ratios

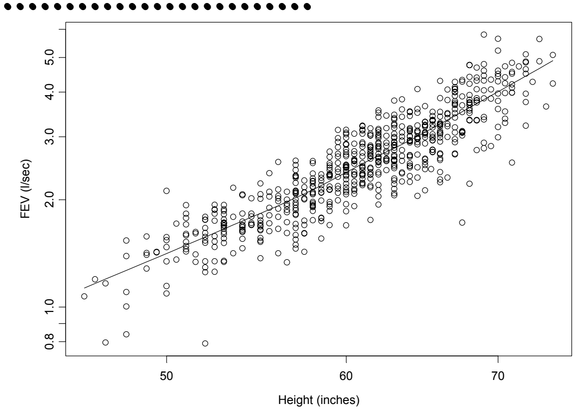
38

log(FEV) versus log(Height)



39

log-log Plot of FEV vs Height



40

Estimation of Regression Model

```
. regress logfev loght, robust
Regression with robust standard errors

                                      Number of obs =     654
                                      F(  1,    652) = 2130.18
                                      Prob > F      =  0.0000
                                      R-squared     =  0.7945
                                      Root MSE      =   .1512
------------------------------------------------------------------
       |         Robust
logfev |  Coef. StErr      t   P>|t|       [95% CI]
------------------------------------------------------------------
 loght |   3.12  .068  46.15  0.000    2.99      3.26
 _cons | -11.92  .278 -42.90  0.000  -12.47    -11.38
```


Log Transformed Predictors

- Interpretation of log transformed predictors with log link function
  - Log link used to model the geometric mean
    - Exponentiated slope estimates ratio of geometric means across groups
  - Compare groups with a k-fold difference in their measured predictors
    - Estimated ratio of geometric means

$$\exp\left(\log(k)\times \beta_1\right) = k^{\beta_1}$$

Interpretation of Stata Output

- Scientific interpretation of the slope

$$\log \text{GM}\left[FEV_i \,\middle|\, loght_i\right] = -11.9 + 3.12 \times loght_i$$

  - Estimated ratio of geometric mean FEV for two groups differing by 10% in height (1.1-fold difference in height)
    - Exponentiate 1.1 to the slope: $1.1^{3.12} = 1.35$
      - Group that is 10% taller is estimated to have a geometric mean FEV that is 1.35 times higher (35% higher)

Why Transform Predictor?

- Typically chosen according to whether the data likely follow a straight line relationship
  - Linearity (“model fit”) necessary to predict the value of the parameter in individual groups
    - Linearity is not necessary to estimate existence of association
    - Linearity is not necessary to estimate a “first order trend” in the parameter across groups having the sampled distribution of the predictor
    - (Inference about these two questions will tend to be conservative if linearity does not hold)

Choice of Transformation

- Rarely do we know which transformation of the predictor provides best “linear” fit
  - As always, there is a danger in using the data to estimate the best transformation to use
    - If there is no association of any kind between the response and the predictor, a “linear” fit (with a zero slope) is the correct one
    - Trying to detect a transformation is thus an informal test for an association
      - Multiple testing procedures inflate the type I error

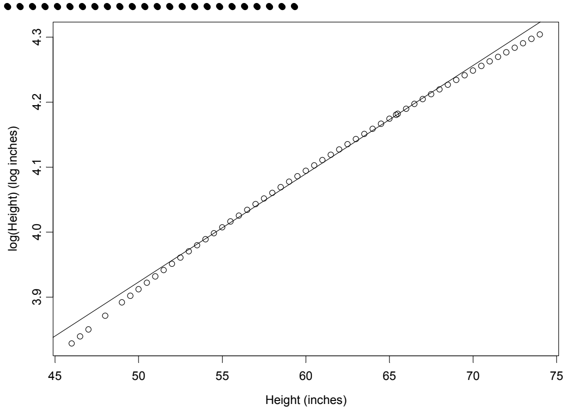
45

Sometimes Does Not Matter

- It is best to choose the transformation of the predictor on scientific grounds
  - However, it is often the case that many functions are well approximated by a straight line over a small range of the data
    - Example: In the modeling of FEV as a function of height, the logarithm of height is approximately linear over the range of heights sampled

46

log(Height) versus Height



47

Untransformed Predictors

- It is thus often the case that we can choose to use an untransformed predictor even when science would suggest a nonlinear association
  - This can have advantages when interpreting the results of the analysis
    - E.g., it is far more natural to compare heights by differences than by ratios
      - Chances are we would characterize two children as differing by 4 inches in height rather than as the 44 inch child as being 10% taller than the 40 inch child

48

Statistical Role of Variables

- Looking ahead to multiple regression: The relative importance of having the "true" transformation for a predictor depends on the statistical role
  - Predictor of Interest
  - Effect Modifiers
  - Confounders
  - Precision variables

Predictor of Interest

- In general, don't worry about modeling the exact relationship before you have even established that there is an association (binary search)
  - Searching for the best fit can inflate the type I error
  - Make most accurate, precise inference about the presence of an association first
    - Exploratory analyses can suggest models for future analyses

Effect Modifiers

- Modeling of effect modifiers is invariably just to test for existence of the interaction
  - We rarely have a lot of precision to answer questions in subgroups of the data
  - Patterns of interaction can be so complex that it is unlikely that we will really capture the interactions across all subgroups in a single model
    - Typically we restrict future studies to analyses treating subgroups separately

Confounders

- It is important to have an appropriate model of the association between the confounder and the response
  - Failure to accurately model the confounder means that some residual confounding will exist
  - However, searching for the best model may inflate the type I error for inference about the predictor of interest by overstating the precision of the study
    - Luckily, we rarely care about inference for the confounder, so we are free to use inefficient means of adjustment, e.g., stratified analyses

Precision Variables

- When modeling precision variables, it is rarely worth the effort to use the “best” transformation
  - We usually capture the largest part of the added precision with crude models
  - We generally do not care about estimating associations between the response and the precision variable
    - Most often, precision variables represent known effects on response

Simple Logistic Regression

Inference About the Odds

Logistic Regression

- Binary response variable
- Allows continuous (or multiple) grouping variables
  - But is OK with binary grouping variable also
- Compares odds of response across groups
  - “Odds ratio”

Why not Linear Regression?

- Many misconceptions about the advantages and disadvantages of analyzing the odds
- Reasons that I consider valid
  - Scientific basis
    - Use of odds ratios in case-control studies
    - Plausibility of linear trends and no effect modifiers
  - Statistical basis
    - Mean variance relationship (if not using robust SE)

Simple Logistic Regression

- Modeling odds of binary response Y on predictor X

Distribution $\Pr(Y_i = 1) = p_i$

Model $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \times X_i$

$X_i = 0$ log odds $= \beta_0$

$X_i = x$ log odds $= \beta_0 + \beta_1 \times x$

$X_i = x+1$ log odds $= \beta_0 + \beta_1 \times x + \beta_1$

57

Interpretation as Odds

- Exponentiation of regression parameters

Distribution $\Pr(Y_i = 1) = p_i$

Model $\left(\frac{p_i}{1-p_i}\right) = e^{\beta_0} \times e^{\beta_1 \times X_i}$

$X_i = 0$ odds $= e^{\beta_0}$

$X_i = x$ odds $= e^{\beta_0} \times e^{\beta_1 \times x}$

$X_i = x+1$ odds $= e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}$

58

Estimating Proportions

- Proportion = odds / (1 + odds)

Distribution $\Pr(Y_i = 1) = p_i$

Model $p_i = \frac{e^{\beta_0} \times e^{\beta_1 \times X_i}}{1 + e^{\beta_0} \times e^{\beta_1 \times X_i}}$

$X_i = 0$ $p_i = e^{\beta_0} / \left(1 + e^{\beta_0}\right)$

$X_i = x$ $p_i = \frac{e^{\beta_0} \times e^{\beta_1 \times x}}{1 + e^{\beta_0} \times e^{\beta_1 \times x}}$

$X_i = x+1$ $p_i = \frac{e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}}{1 + e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}}$

59

Simple Logistic Regression

- Interpretation of the model
  - Odds when predictor is 0
    - Found by exponentiation of the intercept from the logistic regression: exp(β0)
  - Odds ratio between groups differing in the value of the predictor by 1 unit
    - Found by exponentiation of the slope from the logistic regression: exp(β1)

60

Stata

- “`logit respvar predvar, [robust]`
  - Provides regression parameter estimates and inference on the log odds scale
    - Intercept, slope with SE, CI, P values
- “`logistic respvar predvar, [robust]`
  - Provides regression parameter estimates and inference on the odds ratio scale
    - Only slope with SE, CI, P values

Example

- Prevalence of stroke (cerebrovascular accident- CVA) by age in subset of Cardiovascular Health Study
  - Response variable is CVA
    - Binary variable: 0= no history of prior stroke, 1= prior history of stroke
  - Predictor variable is Age
    - Continuous predictor

Odds Ratios using “logistic”

```
.logistic cva age, robust
Logistic regression    Number of obs    =        735
                       LR chi2(1)       =       2.52
                       Prob > chi2      =     0.1127
                       Log likelihood   = -240.98969
                       Pseudo R2        =     0.0051

cva |Odds Ratio StdErr  z    P>|z|  [95% Conf Int]
age |   1.034   .0219  1.59  0.113   .992    1.078
```


Example: Interpretation

“From logistic regression analysis, we estimate that for each year difference in age, the odds of stroke is 3.4% higher in the older group, though this estimate is not statistically significant (P = .113). A 95% CI suggests that this observation is not unusual if a group that is one year older might have odds of stroke that was anywhere from 0.8% lower or 7.8% higher than the younger group.”

Logistic Regression and χ^2 Test

- Logistic regression with a binary predictor (two groups) corresponds to familiar chi squared test
  - Three possible statistics from logistic regression
    - Wald: The test based on the estimate and SE
    - Score: Corresponds to chi squared test, but not given in Stata output
    - Likelihood ratio test: Can be obtained using post-regression commands in Stata (next quarter)

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

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
- Similarly, special inferential procedures are needed with censored data

Notation

Unobserved:

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

Censoring Times: $\{C_1, C_2, \ldots, C_n\}$

Observed data:

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

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

Proportional Hazards Model

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

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

Simple PH Regression Model

- "Baseline" hazard function is unspecified
  - Similar to an intercept

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

$X_i = 0$ log hazard at $t = \log(\lambda_0(t))$

$X_i = x$ log hazard at $t = \log(\lambda_0(t)) + \beta_1 \times x$

$X_i = x+1$ log hazard at $t = \log(\lambda_0(t)) + \beta_1 \times x + \beta_1$

Model on Hazard scale

- Exponentiating parameters

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

$X_i = 0$ hazard at $t = \lambda_0(t)$

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

$X_i = x + 1$ 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)

Stata

- "`stcox obsvar eventvar, [robust]`"
  - Provides regression parameter estimates and inference on the hazard ratio scale
    - Only slope with SE, CI, P values

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

Estimation of 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
                             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 <u>associations</u> 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 linear 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)

Linearity of Model

- Assumption about adequacy of linear model for prediction of group odds of response with logistic regression
  - The log hazard ratio across groups is linear in the modeled predictor
    - (We can model transformations of the measured predictor)

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."

Log Transformed NadirPSA

- Based on prior experience
  - 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

Estimation of Regression Model

```
. 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 : $1.54^{\log(2)} = 1.35$

PH Regression and Logrank Test

- Proportional hazards regression with a binary predictor (two groups) corresponds to the logrank test
  - 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 (next quarter)