Biost 517
Applied Biostatistics I

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

Lecture 18:
Extension to Other Simple Regression Models

December 6, 2010

1

Lecture Outline

- General Simple Regression Model
- Simple Logistic Regression
- Simple Proportional Hazards Regression

2

General Regression Model

3

Types of Variables

- Binary data
  - E.g., sex, death
- Nominal data: unordered, categorical data
  - E.g., race, marital status
- Ordinal categorical data
  - E.g., stage of disease
- Quantitative data
  - E.g., age, blood pressure
- Right censored data
  - E.g., time to death (when not everyone has died)

4

Summary Measures

- The measures commonly used to summarize and compare distributions vary according to the types of data
  - Means: binary; quantitative
  - Medians: ordered; quantitative; censored
  - Proportions: binary; nominal
  - Odds: binary; nominal
  - Hazards: censored
    - hazard = instantaneous rate of failure

5

Regression Models

- According to the parameter compared across groups
  - Means → Linear regression
  - Geom Means → Linear regression on logs
  - Odds → Logistic regression
  - Rates → Poisson regression
  - Hazards → Proportional Hazards regr
  - Quantiles → Parametric survival regr

6

General Regression

- General notation for variables and parameter

Y_i Response measured on the ith subject

X_i Value of the predictor 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.

7

Simple Regression

- General notation for simple regression model

$$g(\theta_i) = \beta_0 + \beta_1 \times X_i$$

$g(\;)$ "link" function used for modeling

β_0 "Intercept"

β_1 "Slope (for predictor X)"

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

8

Borrowing Information

- Use other groups to make estimates in groups with sparse data
- Intuitively: 67 and 69 year olds would provide some relevant information about 68 year olds
- Assuming straight line relationship tells us how to adjust data from other (even more distant) age groups
  - If we do not know about the exact functional relationship, we might want to borrow information only close to each group
    - (Next quarter: splines)

9

Defining "Contrasts"

- Define a comparison across groups to use when answering scientific question
  - If straight line relationship in parameter, slope is difference in parameter between groups differing by 1 year in X
  - If nonlinear relationship in parameter, slope is average difference in parameter between groups differing by 1 year in X
    - Statistical jargon: a "contrast" across the groups

10

Comparison of Methods

- The major difference between regression models is interpretation of the parameters
  - Summary: Mean, geometric mean, odds, hazards
  - Comparison of groups: Difference, ratio
- Issues related to inclusion of covariates remain the same
  - Address the scientific question
    - Predictor of interest; Effect modifiers
  - Address confounding
  - Increase precision

11

Simple Logistic Regression

Inference About the Odds

12

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"

13

Binary Response

- When using regression with binary response variables, we typically model the (log) odds using logistic regression
- Conceptually, there should be no problem modeling the proportion (which is the mean of the distribution)
- However, there are several technical reasons why we do not use linear regression very often with binary response

14

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)

15

Science: Case-Control Studies

- Scientific interest:
  - Distribution of "effect" across groups defined by "cause"
- Common sampling schemes
  - Cohort study: Sample by exposure
    - Estimate distribution of "effect" in exposure groups
  - Case-control study: Sample by outcomes
    - Estimate distribution of exposure in outcome groups
      - E.g., proportion (or odds) of smokers among people with or without cancer

16

Science: Case-Control Studies

- Estimable odds ratios for each sampling scheme
  - Cohort study
    - Odds of cancer among smokers : odds of cancer among nonsmokers
  - Case-control study
    - Odds of smoking among cancer : odds of smoking among noncancer
- Mathematically, the two odds ratios are the same

17

Science: Case-Control Studies

- The odds ratio is easily interpreted when trying to investigate rare events
  - Odds = prob / (1 – prob)
  - Rare event: (1 – prob) is approximately 1
    - Odds is approximately the probability
    - Odds ratio is approximately the risk ratio
      - Risk ratios are easily understood
- Case-control studies typically used when events are rare

18

Science: Linearity

- Proportions have to be between 0 and 1
- It is thus unlikely that a straight line relationship would exist between a proportion and any predictor
  - UNLESS the predictor itself is bounded
  - OTHERWISE there eventually must be a threshold above which the probability does not increase (or only increases a little)

19

Science: Effect Modification

- The restriction on ranges for probabilities also make it likely that effect modification will often be present with proportions
- Ex: 2 Yr Relapse rates by NadirPSA>4, BSS
  - If bone scan score < 3: A difference of 0.60
    - 40% of men with nadir PSA < 4 relapse in 24 months
    - 100% of men with nadir PSA > 4 relapse in 24 months
  - If bone scan score > 3:
    - 71% of men with nadir PSA < 4 relapse in 24 months
    - Thus impossible for men with nadir PSA > 4 to have an absolute difference of 0.60 higher

20

Why use the odds?

- The odds of an event are between 0 and infinity
  - Recall odds = prob / (1 – prob)
    - (Even better: log (odds) are between negative infinity and positive infinity)
  - Thus, there is a greater chance that linear relationships might hold without effect modification

21

Statistics: Mean-Variance

- Classical linear regression requires equal variances in each predictor group
  - With binary data, the variance within a group depends on the mean
    - For binary Y
      - E(Y) = p
      - Var (Y) = p(1 – p)
  - (With robust regression techniques, this problem not a limitation)

22

Simple Logistic Regression

- Modeling odds of binary response Y on predictor X

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

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

$X_i = 0$ $\quad$ $\text{log odds} = \beta_0$

$X_i = x$ $\quad$ $\text{log odds} = \beta_0 + \beta_1 \times x$

$X_i = x+1$ $\quad$ $\text{log odds} = \beta_0 + \beta_1 \times x + \beta_1$

23

Interpretation as Odds

- Exponentiation of regression parameters

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

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

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

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

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

24

Estimating Proportions

- Proportion = odds / (1 + odds)

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

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

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

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

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

25

Simple Logistic Regression

- Interpretation of the model
  - Odds when predictor is 0
    - Found by exponentiation of the intercept from the logistic regression: $\exp(\beta_0)$
  - Odds ratio between groups differing in the value of the predictor by 1 unit
    - Found by exponentiation of the slope from the logistic regression: $\exp(\beta_1)$

26

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

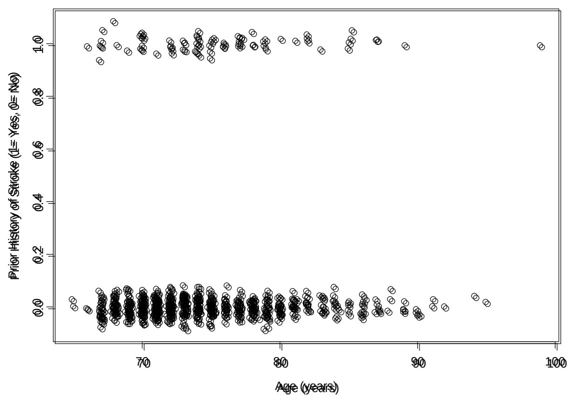
27

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

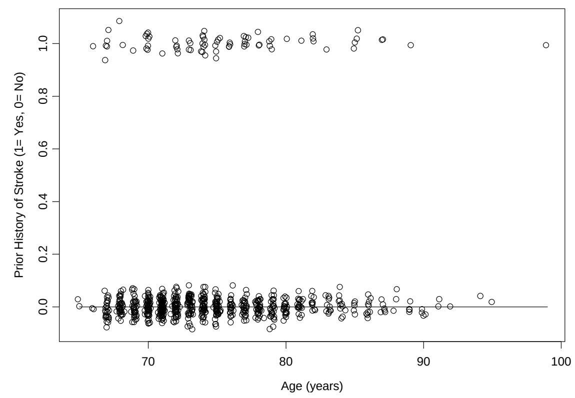
28

CVA (jittered) vs Age



29

Lowess Smooth of CVA vs Age



30

Characterization of Plot

- Clearly the scatterplot (even with superimposed smooth) is pretty useless with a binary response
  - (Note that we are estimating proportions– not odds– with this plot, so we can not even judge linearity for logistic regression)

31

Example: Regression Model

- Answer question by assessing linear trends in log odds of stroke by age
  - Estimate best fitting line to log odds of CVA within age groups

$$\log odds\left(CVA \middle| Age \right) = \beta_0 + \beta_1 \times Age$$

- An association will exist if the slope (β_1) is nonzero
  - In that case, the odds (and probability) of CVA will be different across different age groups

32

Parameter Estimates

```
. logit cva age
(iteration info deleted)
                        Number of obs   =        735
                        LR chi2(1)      =       2.45
                        Prob > chi2     =     0.1175
                        Log likelihood  = -240.98969
                        Pseudo R2       =     0.0051

  cva |  Coef StdErr     z    P>|z|   [95% Conf Int]
  age | .0336  .0210   1.59   0.111   -.0077  .0748
_cons | -4.69  1.591  -2.95   0.003   -7.810 -1.572
```

33

Interpretation of Stata Output

- Regression model for CVA on age
  - Intercept is labeled by "_cons"
    - Estimated intercept: -4.69
  - Slope is labeled by variable name: "age"
    - Estimated slope: 0.0336
  - Estimated linear relationship:
    - log odds relapse by nadir given by

$$\text{log odds } CVA = -4.69 + 0.0336 \times Age_i$$

34

Interpretation of Intercept

$$\text{log odds } CVA = -4.69 + 0.0336 \times Age_i$$

- Estimated log odds CVA for newborns is -4.69
  - Odds of CVA for newborns is $e^{-4.69}$ = 0.0092
  - Probability of CVA for newborns
    - Use prob = odds / (1+odds): .0092 / 1+.0092= .0091
- Pretty ridiculous to try to estimate
  - We never sampled anyone less than 67
  - In this problem, the intercept is just a tool in fitting the model

35

Interpretation of Slope

$$\text{log odds } CVA = -4.69 + 0.0336 \times Age_i$$

- Estimated difference in log odds CVA for two groups differing by one year in age is 0.0336, with older group tending to higher log odds
  - Odds Ratio: $e^{0.0336}$= 1.034
  - For 5 year age difference: $e^{5x0.0336}$= 1.034^5 = 1.183
- (If a straight line relationship is not true, we interpret the slope as an average difference in log odds CVA per one year difference in age)

36

Stata: “logit” versus “logistic”

- Given that we are rarely interested in the intercept, we might as well use the “`logistic`” command
  - It will provide inference for the odds ratio
    - We don’t have to exponentiate the slope estimate

37

Odds Ratios using “logistic”

```
.logistic cva age
Logistic regression     Number of obs   =        735
                        LR chi2(1)      =       2.45
                        Prob > chi2     =     0.1175
                        Log likelihood  = -240.98969
                        Pseudo R2       =     0.0051

cva  |Odds Ratio StdErr  z    P>|z|  [95% Conf Int]
age  |   1.034   .0218  1.59  0.111   .992    1.078
```

38

Comments on Interpretation

- I express this as a difference between group odds rather than a change with aging
  - We did not do a longitudinal study
- To the extent that the true group log odds have a linear relationship, this interpretation applies exactly
  - If the true relationship is nonlinear
    - The slope estimates the “first order trend” for the sampled age distribution
    - We should not regard the estimates of individual group probabilities / odds as accurate

39

Signal and Noise

- Note that the Signal and Noise idea does not apply so well here
  - We do not tend to quantify an “error distribution” with logistic regression

40

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

41

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 (odds)
    - Recall rule of thumb for chi-squared test based on expected number of events

42

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

43

Within Group Variance

- Assumptions about variance of response within groups for logistic regression
- Classically:
  - Mean variance relationship for binary data
    - Classical logistic regression estimates SE using model based estimates
    - Hence in order to satisfy this requirement, linearity of log odds across groups must hold
- Robust standard error estimates:
  - Allow unequal variances across groups
  - (Do not need the linearity of log odds)

44

Statistical Validity of Inference

- Inference (CI, P values) about odds of response in specific groups requires a further assumption
  - Assumption about adequacy of linear model

45

Linearity of Model

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

46

Statistical Validity of Inference

- Inference (prediction intervals, P values) about individual observations requires no further assumptions because we have binary data
  - If we know the mean (proportion), we know everything

47

Implications for Inference

- Regression based inference about associations is far more robust than estimation of group odds of response
  - A hierarchy of null hypotheses
    - Strong (and intermediate) null: Total independence of *Y* and *X*
      - A binary distribution only depends on the mean (proportion, odds)
    - Weak null: No linear trend in mean of *Y* across *X* groups

48

Under Strong Null

- If the response and predictor of interest were totally independent:
  - Probability of response, and hence the odds and log odds, would be the same in all groups
    - A flat line would describe the log odds response across groups (and a linear model is correct)
      - Slope would be zero
    - Within group variance would be correctly estimated by the model
    - In large sample sizes, the regression parameters are normally distributed

49

Under Weak Null

- Linear trend in means across predictor groups would lie on a flat line
  - Slope of best fitting line would be zero
  - Within group variance could vary from that predicted by model
  - 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

50

Classical Logistic Regression

- Inference about slope <u>tests</u> strong null
  - Tests make inference assuming the null
    - The data can appear nonlinear in log odds
      - Merely evidence strong null is not true
  - Limitations
    - We cannot be confident that there is a trend in the log odds across groups
      - Valid inference about trend demands correct model
    - We cannot be confident of estimates of group probabilities (odds)
      - Valid estimates of group means demands correct model

51

Robust Standard Errors

- Inference about slope <u>tests</u> weak null
  - Data can appear nonlinear in log odds
    - Robust SE estimates true variability
      - Does not use model based estimates of SE
    - Nonlinearity decreases precision, but inference still valid about first order (linear) trends
  - Only if linear relationship holds can we
    - Estimate group response probabilities (odds)

52

Choice of Inference

- Which inference is correct?
- Classical logistic regression and robust standard error estimates differ in the strength of necessary assumptions
  - As a rule, if all the assumptions of classical logistic regression hold, it will be more precise
    - (Hence, we will have greatest precision to detect associations if the linear model is correct)
  - The robust standard error estimates are, however, valid for detection of associations even in those instances

53

Implications for Inference

- Inference about associations is far more trustworthy than estimation of group means or individual predictions
  - Nonzero slope suggests an association between response and predictor
    - Inference about linear trends in log odds if use robust SE

54

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 odds of response across groups
  - Data suggests evidence of a trend toward larger (smaller) odds in groups having larger values of the predictor
  - (To the extent the data appears linear, estimates of the group odds will be reliable)

55

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, the odds of response)]*
    - There may be an association, but the best fitting line has a zero slope (a curvilinear association in the parameter)
    - There may be a first order trend in the log odds, but we lacked statistical precision to be confident that it truly exists (type II error)

56

Logistic Regression Inference

- The regression output provides
  - Estimates
    - Intercept: estimated log odds CVA when age = 0
    - Slope: estimated difference in log odds CVA for two groups differing by one year in age
  - Standard errors
  - Confidence intervals
  - P values testing for
    - Intercept= zero (odds= 1; prob= 0.5) (who cares?)
    - Slope= zero (test for linear trend in log odds)

57

Odds Ratios using "logistic"

```
.logistic cva age
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
```

58

Standard Error of Odds Ratio

- Logistic regression uses the log odds scale
  - Exponentiate estimates and CI to get inference on odds ratio
- Stata "`logistic`" provides estimates on odds ratio scale
  - Standard error is from "delta method"
  - CI is from exponentiating log odds CI

59

Delta Method Based SE

- In regression models encountered in this class, we can find SE of exponentiated slope parameters

$$\hat{\beta}_1 \stackrel{\sim}{\sim} N\left(\beta_1, se^2\left(\hat{\beta}_1\right)\right)$$

$$\Downarrow \quad \text{(delta method)}$$

$$e^{\hat{\beta}_1} \stackrel{\sim}{\sim} N\left(e^{\beta_1}, \left[e^{\beta_1} se\left(\hat{\beta}_1\right)\right]^2\right)$$

60

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."

61

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)

62

Simple Proportional Hazards Regression

Inference About Hazards

63

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

64

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

65

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

66

Survival Regression

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

67

Accelerated Failure Time Model

- Assume that a factor causes some 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 quantiles of survival distributions are constant across two group
  - E.g., report median ratios
- AFT models include the parametric exponential, Weibull, and lognormal models

68

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

69

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

70

Proportional Hazards Model

- Ignores the time that events occur
- 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

71

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

72

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

73

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

74

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)

75

Relationship to Survival

- Hazard function determines survival function

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

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

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

76

Stata

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

77

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

78

Scatterplots

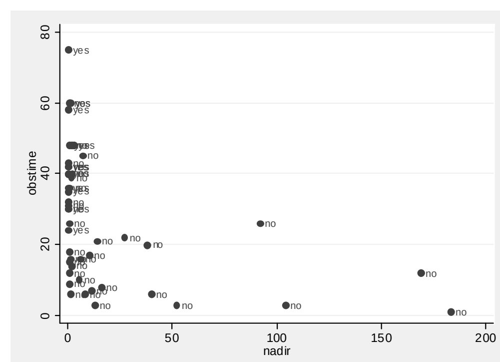
- Scatterplots of censored data are not scientifically meaningful

- 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

79

Obstime vs Nadir (by inrem)

- `scatter obstime nadir, mlabel(inrem)`



80

Characterization of Scatterplot

- Outliers
  - ??
- First order trends
  - Certainly downward slope: No censoring at high nadirs
- Second order trends
  - Must be curvilinear (but how much)
- Variability within groups
  - Highest with greater length of observation

81

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
                            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
```

82

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)

83

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

84

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

85

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

86

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)

87

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)

88

Prediction

- We rarely make inference about within group survival probabilities using the proportional hazards model
- We sometimes use estimated survival curves descriptively
  - Use estimates of baseline survival function
  - Exponentiate the baseline survival to find survival curve for specific covariates

89

Relationship to Survival

- Hazard function determines survival function

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

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

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

90

Implications for Inference

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

91

Classical PH Regression

- Inference about slope tests strong null
  - Tests make inference assuming the null
    - The data can appear nonproportional hazards or nonlinear in log hazard ratio
      - Merely evidence strong null is not true
  - Limitations
    - We cannot be confident that there is a trend in the hazard ratio across groups
      - Valid inference about trend demands correct model

92

Robust Standard Errors

- Inference about slope tests weak null
  - Data can appear nonproportional hazards or nonlinear in hazard ratio across groups
    - Robust SE estimates true variability
      - Does not use model based estimates of SE
    - Nonlinearity decreases precision, but inference still valid about first order (linear) trends

93

Choice of Inference

- Which inference is correct?
  - Classical PH regression and robust standard error estimates differ in the strength of necessary assumptions
    - As a rule, if all the assumptions of classical PH regression hold, it will be more precise
      - (Hence, we will have greatest precision to detect associations if the linear model is correct)
    - The robust standard error estimates are, however, valid for detection of associations even in those instances

94

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

95

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, the odds of response)
  - There may be an association, but the best fitting line has a zero slope (a curvilinear association in the parameter)
  - There may be a first order trend in the log hazard ratio, but we lacked statistical precision to be confident that it truly exists (type II error)

96

Estimation of Regression Model

```
. stset obstime relapse, robust
. 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)  =     16.79
Log lklhood  = -113.3       Prob > chi2 =    0.0000
_    t | HzRat StdErr    z    P>|z|    [95% Conf Int]
nadir | 1.016  .0038  4.10   0.000   1.008    1.023
```

97

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."

98

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

99

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
```

100

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

101

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)

102