

Biost 517
Applied Biostatistics I
.....
Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 17:
Simple Regression

December 5, 2011

1

© 2002, 2003, 2005 Scott S. Emerson, M.D., Ph.D.

Lecture Outline
.....

- General Regression Setting
- Simple Regression Models
 - Linear Regression
 - Inference About Geometric Means
 - Logistic Regression
 - Proportional Hazards Regression
- Additional Comments About Inference

2

General Regression Setting
.....

3

Two Variable Setting
.....

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

4

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

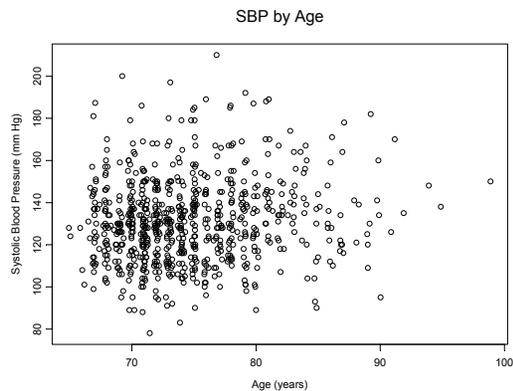
5

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

9

Guiding Principle

“Everything is regression.”

- Scott Emerson

10

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)

11

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

12

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

13

General Regression

- General notation for variables and parameter

Y_i Response measured on the i th subject

X_i Value of the predictor for the i th 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.

14

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" of "linear predictor"

β_1 "Slope (for predictor X)" of "linear predictor"

- The link function is usually either
 - None (also called identity) for an "additive model"
 - Most common when analyzing means of continuous Y
 - Log for a "multiplicative model"
 - Analyzing geometric means, odds, rates, hazards

15

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)

16

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

17

Next Quarter: Multiple Regression

- General notation for simple regression model

$$g(\theta_i) = \beta_0 + \beta_1 \times X_{1i} + \beta_2 \times X_{2i} + \dots + \beta_p \times X_{pi}$$

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

β_0 "Intercept" of "linear predictor"

β_j "Slope (for predictor X_j)" of "linear predictor"

- The link function is usually either
 - None (also called identity) for an "additive model"
 - Most common when analyzing means of continuous Y
 - Log for a "multiplicative model"
 - Analyzing geometric means, odds, rates, hazards

18

Uses of Multiple Regression

- Modeling complex scientific factors
 - Smoking
 - Indicator of ever smoked, pack years, years since quit...
 - U-shaped trends
 - Dummy variables modeling each group independently
- Adjusting for covariates
 - Confounding
 - Precision
- Modeling effect modification
 - Include interaction terms

19

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

20

Usual Regression Output

- Estimates
 - Intercept: estimated $g(\theta)$ when all predictors are 0
 - Slope for each predictor: estimated difference in $g(\theta)$ for two groups that
 - differ by one unit in corresponding predictor,
 - but agreeing in all other predictors
- Standard errors
- Confidence intervals
- P values testing for
 - Intercept of zero (who cares?)
 - Slope of zero (test for association in θ)

21

Interpretation: Identity Link

$$\theta(Y | X) = \beta_0 + \beta_1 \times X$$

- Intercept: β_0 - value of θ for a group with $X=0$
 - Quite often not of scientific interest because out of range of data or even impossible
- Slope: β_1 - (avg) diff in θ across groups differing in X by 1 unit
 - Usually measures association between Y and X
- Most common examples:
 - Linear regression when θ is mean

22

Interpretation: Log Link

$$\log[\theta(Y | X)] = \beta_0 + \beta_1 \times X$$

- Intercept: $\exp(\beta_0)$ - value of θ for a group with $X=0$
 - Quite often not of scientific interest because out of range of data or even impossible
- Slope: $\exp(\beta_1)$ - (avg) ratio of θ across groups differ in X by 1
 - Usually measures association between Y and X
- Most common examples:
 - Linear regression on $\log(Y)$: $\exp(\beta_1)$ is geom mean ratio
 - Logistic regression: $\exp(\beta_1)$ is odds ratio
 - Proportional hazard regression: $\exp(\beta_1)$ is hazard ratio

23

Inference with Regression

- Regression analysis is commonly used to answer statistical questions of type
 - 3. Estimating distribution parameters in population
 - 4. Comparing distributions across groups
 - 5. Predicting future individual observations
- Assumptions needed for valid inference depend on question
 - To detect associations (comparing distributions) we need
 - Approximate normality of estimated slopes
 - Correct modeling of dependence among observations
 - Correct modeling of variance within groups
 - To use linear predictor to estimate θ we also need (in addition)
 - Correct linear relationship
 - To predict future values of Y we also need (in addition)
 - Correct assumptions about distribution of Y

24

Motivating Example

.....

25

Example: Questions

.....

- Association between blood pressure and age
- Scientific question:
 - Does aging affect blood pressure?
- Statistical question:
 - Does the distribution of systolic blood pressure differ across age groups?
 - Acknowledges variability of response
 - Acknowledges uncertainty of cause and effect
 - Differences could be related to calendar time of birth instead of age

26

Example: Definition of Variables

.....

- Response: Systolic blood pressure
 - continuous
- Predictor of interest (grouping): Age
 - continuous
 - an infinite number of ages are possible
 - we probably will not sample every one of them
- (Linear regression is most often used with a continuous response variable and a continuous POI or any POI adjusted for other variables
 - BUT: It makes perfect sense with binary POI
 - Arguments could even be made for the case of binary response, though this is nonstandard)

27

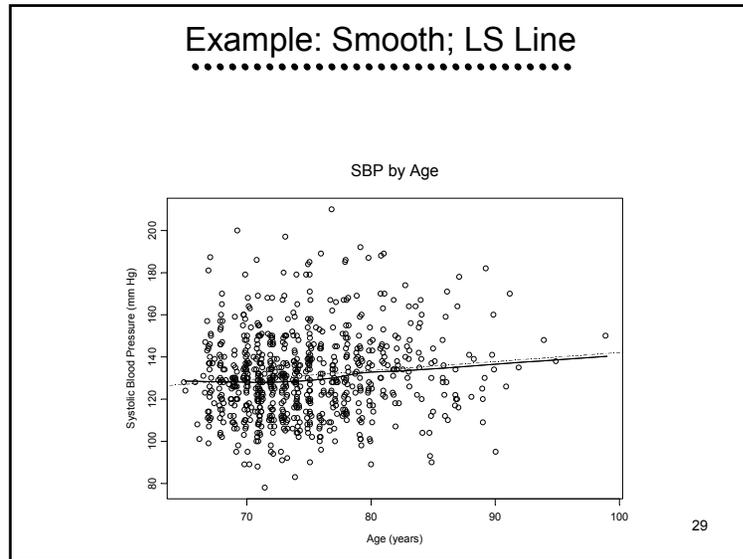
Example: Regression Model

.....

- Answer question by assessing linear trends in, say, average SBP by age
 - Estimate best fitting line to average SBP within age groups
- An association will exist if the slope (β_1) is nonzero
 - In that case, the average SBP will be different across different age groups

$$E(SBP | Age) = \beta_0 + \beta_1 \times Age$$

28



“Rule of Thumb”

.....

- The regression model thus produces something similar to “a rule of thumb”
 - E.g., “Normal SBP is 100 plus half your age”

$$E(SBP | Age) = 100 + 0.5 \times Age$$

30

Example: Estimates, Inference

.....

```
. regress sbp age
```

	Number of obs =	735			
<u>Source</u>	<u>SS</u>	<u>df</u>	<u>MS</u>	F(1, 733) =	10.63
Model	4056	1	4056.4	Prob > F =	0.0012
<u>Residual</u>	<u>279740</u>	<u>733</u>	<u>381.6</u>	R-squared =	0.0143
Total	283796	734	386.6	Adj R-squared =	0.0129
				Root MSE =	19.536

sbp	Coef.	St.Err.	t	P> t	[95% Conf Int]
age	.431	.132	3.26	0.001	.172 .691
_cons	98.9	9.89	10.01	0.000	79.5 118.4

$$E(SBP | Age) = 98.9 + 0.431 \times Age$$

31

Use of Regression

.....

- The regression “model” serves to
 - Make estimates in groups with sparse data by “borrowing information” from other groups
 - Define a comparison across groups to use when answering scientific question

32

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)

33

Defining “Contrasts”

- Define a comparison across groups to use when answering scientific question
- If straight line relationship in means, slope is difference in mean SBP between groups differing by 1 year in age
- If nonlinear relationship in means, slope is average difference in mean SBP between groups differing by 1 year in age
 - Statistical jargon: a “contrast” across the means

34

Linear Regression Inference

- The regression output provides
 - Estimates
 - Intercept: estimated mean when age = 0
 - Slope: estimated difference in average SBP for two groups differing by one year in age
 - Standard errors
 - Confidence intervals
 - P values testing for
 - Intercept of zero (who cares?)
 - Slope of zero (test for linear trend in means)

35

Example: Interpretation

“From linear regression analysis, we estimate that for each year difference in age, the difference in mean SBP is 0.43 mmHg. A 95% CI suggests that this observation is not unusual if the true difference in mean SBP per year difference in age were between 0.17 and 0.69 mmHg. Because the P value is $P < .0005$, we reject the null hypothesis that there is no linear trend in the average SBP across age groups.”

36

Simple Linear Regression

.....

37

Ingredients: Regression Model

.....

- Response: Mean of this variable compared across groups
 - Typically an uncensored continuous random variable
 - But truly can sometimes be used with discrete variables
- Predictor: Indicates the groups to be compared
 - Can be continuous or discrete (including binary)
- Model: We typically consider a “linear predictor function” that is linear in the modeled predictors
 - Expected value (mean) of Y for a particular value of X

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

38

Use of Straight Line Relationship

.....

- Algebra: A line is of form $y = mx + b$
 - With no variation in the data, each value of y would lie exactly on a straight line
 - Intercept b is value of y when $x=0$
 - Slope m is difference in y per unit difference in x
- In the real world
 - Response within groups is variable
 - “Hidden variables”
 - Inherent randomness
 - The line describes the central tendency of the data in a scatterplot of the response versus the predictor

39

Ingredients: 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

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

40

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 $E[Y_i | X_i] = \beta_0 + \beta_1 \times X_i$

$X_i = 0$ $E[Y_i | X_i = 0] = \beta_0$

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

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

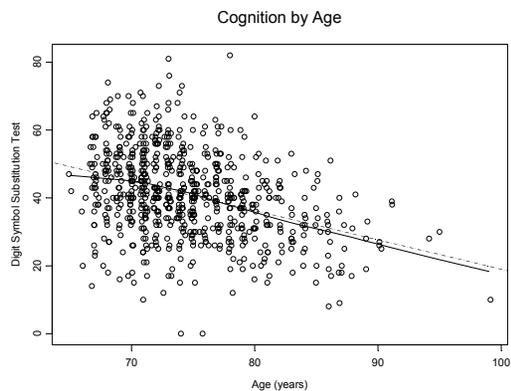
41

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

42

Example: Lowess, LS Line



43

Least Squares Estimation

```
. regress dsst age
```

Source	SS	df	MS	Nbr of obs =	723
Model	15377	1	15377	F(1, 721) =	109.57
Residual	101191	721	140.3	Prob > F =	0.0000
Total	116569	722	161.4	R-squared =	0.1319
				Adj R-sqr =	0.1307
				Root MSE =	11.847

dsst	Coef.	StdErr	t	P> t	[95% C I]
age	-.863	.0825	-10.47	0.000	-1.03 - .701
cons	105	6.16	17.11	0.000	93.3 117

44

Useful Output

.....

```

. regress dsst age

                                Nbr of obs =      723

                                Prob > F    =  0.0000
                                R-squared    =  0.1319
                                Adj R-sqr   =  0.1307
                                Root MSE  = 11.847

-----+-----
 dsst | Coef.  StdErr   P>|t|   [95% C I]
-----+-----
   age |  -.863   .0825    0.000  -1.03  -.701
   cons |   105    6.16    0.000   93.3   117
-----+-----
    
```

45

Deciphering Stata Output: Means

.....

- Estimates of within group means
 - Intercept is labeled “_cons”
 - Estimated intercept: 105.
 - Slope is labeled by variable name: “age”
 - Estimated slope: -.863
 - Estimated linear relationship:
 - Average DSST by age given by

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

46

Deciphering Stata Output: SD

.....

- Estimates of within group standard deviation
 - Within group SD is labeled “Root MSE”
 - Estimated within group SD: 11.85
 - This presumes constant variance in age groups
 - If not, this is in based on average within group variance

47

Interpretation of Intercept

.....

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

- 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

48

Interpretation of Slope

.....

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

- 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: $5 \times -0.863 = -4.32$
 - 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)

49

Comments on Interpretation

.....

- I express this as a difference between group means rather than a change with aging
 - We did not do a longitudinal study
- To the extent that the true group means 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 means as accurate

50

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

51

Ex: Classical Linear Regression

.....

```
. regress dsst age
```

Source		SS	df	MS	Nbr of obs =	723
					F(1, 721) =	109.57
Model		15377	1	15377	Prob > F =	0.0000
Residual		101191	721	140.3	R-squared =	0.1319
					Adj R-sqr =	0.1307
Total		116569	722	161.4	Root MSE =	11.847

dsst		Coef.	StdErr	t	P> t	[95% C I]
age		-.863	.0825	-10.47	0.000	-1.03 - .701
_cons		105	6.16	17.11	0.000	93.3 117

52

Classical Linear Regression

- Inference for association based on slope
 - Strong null based inference
 - P value < .0001 suggests distribution of DSST differs across age groups
 - T statistic: -10.47 (Who cares?)
 - Under assumptions of homoscedasticity
 - 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: -1.03, -0.701

53

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

54

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?)

55

Choice of Inference

- Which inference is correct?
- Classical linear regression and robust standard error estimates differ in the strength of necessary assumptions
- As a rule, if all the assumptions of classical linear 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

56

Choosing the Correct Model

“All models are false, some models are useful.”

- George Box

57

Choosing the Correct Model

“In statistics, as in art, never fall in love with your model.”

- Unknown

58

Alternative Representation

- Sometimes linear regression models are expressed in terms of the response instead of the mean response
 - Includes an “error” modeling difference between observed value and expectation

Model $Y_i = \beta_0 + \beta_1 \times X_i + \varepsilon_i$

59

Signal and Noise

Model $Y_i = \beta_0 + \beta_1 \times X_i + \varepsilon_i$

- The response is divided into two parts
 - The mean (systematic part or “signal”)
 - The “error” (random part or “noise”)
 - difference between the observed value and the corresponding group mean
 - ε_i is called the error
- The error distribution describes the within-group distribution of response

60

Estimates of Error Distribution

.....

- The error distribution is estimated from the residuals

Residual $\hat{\epsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 \times X_i)$

- The mean of the errors is assumed to be 0
- The sample standard deviation of the residuals is reported as the “Root Mean Squared Error”

61

Example

.....

- Thus we estimate within group SD of 11.85 in the DSST vs age example
 - Classical linear regression:
 - SD for each age group
 - Robust standard error estimates:
 - Square root of average variances across groups

62

Relationships to Previous Methods

.....

- Linear regression on a binary predictor
 - Classical LR: exactly the t test that presumes equal variances
 - Robust SE: approximates t test that allows unequal variances
 - “Huber-White sandwich estimator”
 - Stata: `“regress dsst age, robust”`
- Classical simple linear regression
 - Test for slope is exactly the test for significant correlation

63

Inference for the Geometric Mean

.....

Simple Linear Regression on Log Transformed Data

64

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”

65

Interpretation of Parameters

.....

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

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

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

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

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

66

Interpretation of Parameters

.....

- Restated model as log link for geometric mean

Model $\log GM[Y_i | X_i] = \beta_0 + \beta_1 \times X_i$

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

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

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

67

Interpretation of Parameters

.....

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

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

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

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

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

68

Interpretation of Parameters

- Geometric mean when predictor is 0
 - Found by exponentiation of the intercept from the linear regression on log transformed data: $\exp(\beta_0)$
- 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: $\exp(\beta_1)$
- Confidence intervals for geometric mean and ratios found by exponentiating the CI for regression parameters

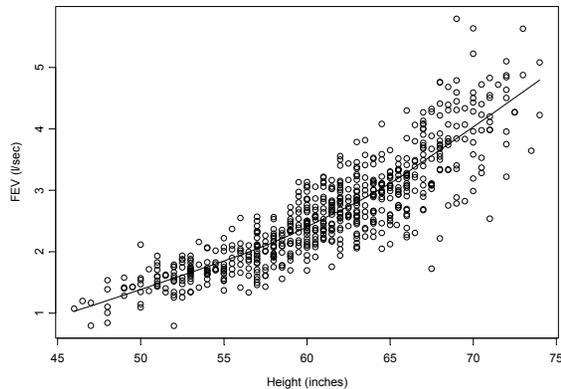
69

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

70

FEV versus Height



71

Characterization of Scatterplot

- Detection of outliers
 - None obvious
- Trends in FEV across groups
 - FEV tends to be larger for taller children
- Second order trends
 - Curvilinear increase in FEV with height
- Variation within height groups
 - “heteroscedastic”: unequal variance across groups
 - mean-variance relationship: higher variation in groups with higher FEV

72

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

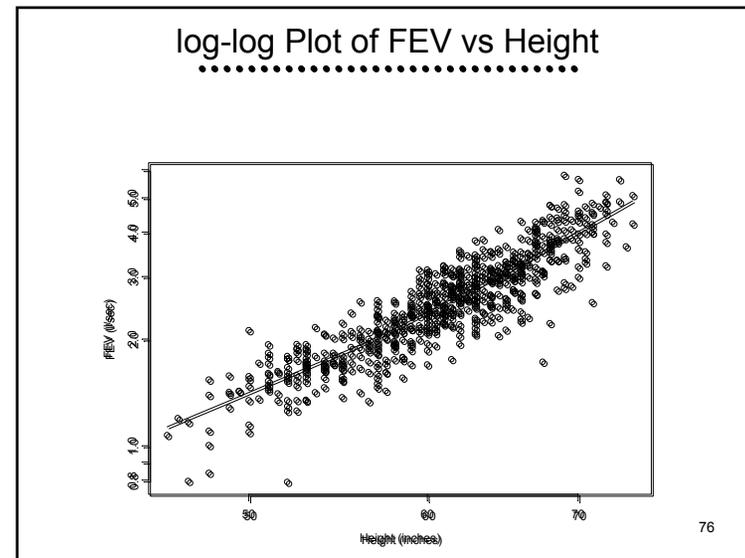
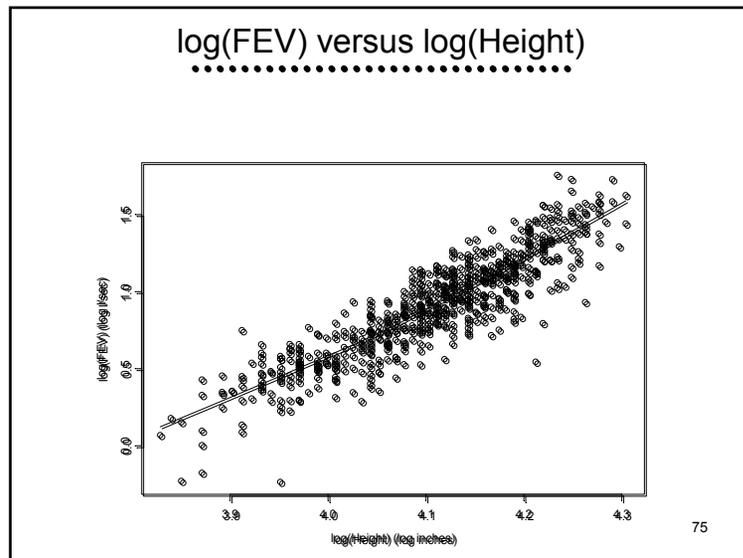
73

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

74



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

77

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

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

78

Interpretation of Stata Output

.....

- Scientific interpretation of the slope

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

- 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)

79

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)

80

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

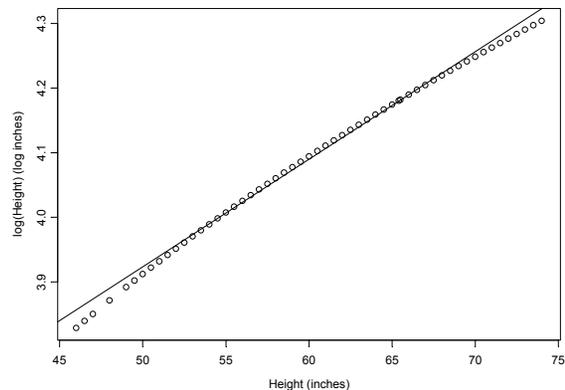
81

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

82

log(Height) versus Height



83

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

84

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

85

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

86

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

87

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

88

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

89

Simple Logistic Regression

Inference About the Odds

90

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”

91

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

92

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)

93

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

94

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

95

Science: Case-Control Studies

- The odds ratio is easily interpreted when trying to investigate rare events
 - Odds = $\text{prob} / (1 - \text{prob})$
 - Rare event: $(1 - \text{prob})$ 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

96

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)

97

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

98

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

99

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
 - $E(Y) = p$
 - $\text{Var}(Y) = p(1 - p)$
 - (With robust regression techniques, this problem not a limitation)

100

Simple Logistic Regression

.....

- Modeling odds of binary response Y on predictor X

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

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

$X_i = 0$ $\log \text{odds} = \beta_0$
 $X_i = x$ $\log \text{odds} = \beta_0 + \beta_1 \times x$
 $X_i = x+1$ $\log \text{odds} = \beta_0 + \beta_1 \times x + \beta_1$

101

Interpretation as Odds

.....

- Exponentiation of regression parameters

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

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

$X_i = 0$ $\text{odds} = e^{\beta_0}$
 $X_i = x$ $\text{odds} = e^{\beta_0} \times e^{\beta_1 \times x}$
 $X_i = x+1$ $\text{odds} = e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}$

102

Estimating Proportions

.....

- Proportion = odds / (1 + odds)

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

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

$X_i = 0$ $p_i = e^{\beta_0} / (1 + e^{\beta_0})$
 $X_i = x$ $p_i = \frac{e^{\beta_0} \times e^{\beta_1 \times x}}{1 + e^{\beta_0} \times e^{\beta_1 \times x}}$
 $X_i = x+1$ $p_i = \frac{e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}}{1 + e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}}$

103

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: $\exp(\beta_1)$

104

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

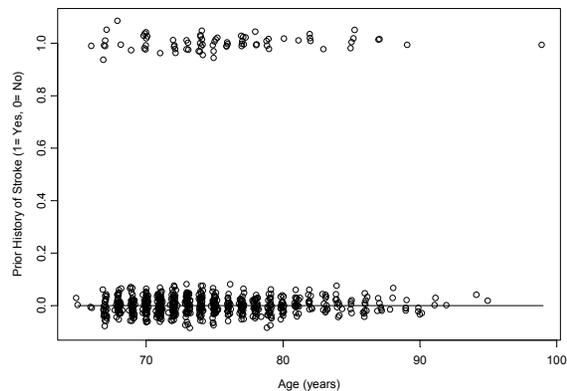
105

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

106

Lowess Smooth of CVA vs Age



107

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)

108

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

$$\text{logodds}(CVA | Age) = \beta_0 + \beta_1 \times Age$$

- 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

109

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

110

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 CVA by age group given by

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

111

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

112

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: $e^{5 \times 0.0336} = 1.034^5 = 1.183$
- (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)

113

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

114

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

115

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

116

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

117

Simple Proportional Hazards Regression

.....

Inference About Hazards

118

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

119

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

120

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

121

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

122

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

123

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

124

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

125

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

126

Simple PH Regression Model

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

Model $\log(\lambda(t | X_i)) = \log(\lambda_0(t)) + \beta_1 \times X_i$

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

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

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

127

Model on Hazard scale

- Exponentiating parameters

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

$X_i = 0$ 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 $t = \lambda_0(t) \times e^{\beta_1 \times x} \times e^{\beta_1}$

128

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: $\exp(\beta_1)$

129

Relationship to Survival

.....

- Hazard function determines survival function

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

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

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

130

Stata

.....

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

131

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

132

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

133

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 likelihood = -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

134

Interpretation of Stata Output

.....

- Scientific interpretation of the slope

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

- 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)

135

Additional Comments Regarding Validity of Inference

.....

136

Inference with Regression

- Most commonly encountered questions
 - Quantifying distributions
 - Describing the distribution of response Y within groups by estimating the mean $E(Y | X)$
 - 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
 - Often we use the mean or geometric mean

137

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

138

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
 - 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

139

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

140

Within Group Variance

- Assumptions about variance of response within groups for linear regression
- Classically:
 - Equal variances across groups
- Robust standard error estimates:
 - Allow unequal variances across groups

141

Statistical Validity of Inference

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

142

Linearity of Model

- Assumption 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)

143

Statistical Validity of Inference

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

144

Distribution of Errors

- Assumption about distribution of errors within each group for prediction intervals with linear regression
- Classically:
 - Errors have the same normal distribution within each group
- Possible extension:
 - Errors have the same distribution within each group, though it need not be normal
 - Not implemented in any software that I know of

145

Prediction and Robust SE

- If you are using robust standard error estimates, prediction intervals based on linear regression models is inappropriate
 - Prediction intervals based on linear regression assume common error distribution across groups

146

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

147

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 the mean response 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

148

Under Intermediate Null

- Means 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

149

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 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

150

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 the means
 - Valid inference about means demands homoscedasticity
 - We cannot be confident of estimates of group means
 - Valid estimates of group means demands linearity

151

Robust Standard Errors

- Inference about slope tests 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
 - Only if linear relationship holds can we
 - Test intermediate null
 - Estimate group means

152

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 means if use robust SE

153

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

154

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 mean response)
 - There may be an association in the parameter considered, but the best fitting line has a zero slope (a curvilinear association in the parameter)
 - 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)

155

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

156

Model Checking
.....

“Blood suckers hide ‘neath my bed”

“Eyepennies”, Mark Linkous (Sparklehorse)

157

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)

158

Choice of Inference
.....

- My general recommendation:
 - 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?

159

Inference on Group Means
.....

- Inference about estimation of group means 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

160