

Biost 518

Applied Biostatistics II

.....

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

Lecture 4: Review of Simple Regression I

January 10, 2007

1

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

Lecture Outline

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

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

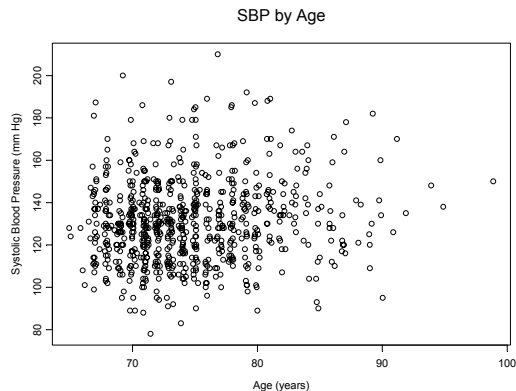
.....

- 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

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

11

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)

12

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

13

Simple Linear Regression

.....

14

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

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

15

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

$$\text{Model} \quad E[Y_i | X_i] = \beta_0 + \beta_1 \times X_i$$

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

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

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

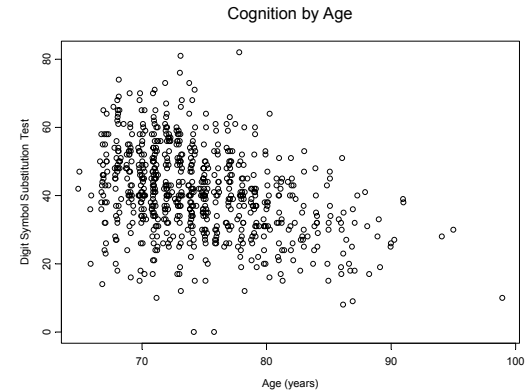
16

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 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)

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

21

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

22

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

23

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

24

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) ²⁵

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?)

26

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

27

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) ²⁸

28

Inference for the Geometric Mean

.....

Simple Linear Regression on Log Transformed Data

29

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”

30

Interpretation of Parameters

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

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

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

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

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

31

Interpretation of Parameters

-
- Restated model as log link for geometric mean

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

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

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

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

32

Interpretation of Parameters

- Interpretation of regression parameters by back-transforming model

– Exponentiation is inverse of log

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

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

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

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

33

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

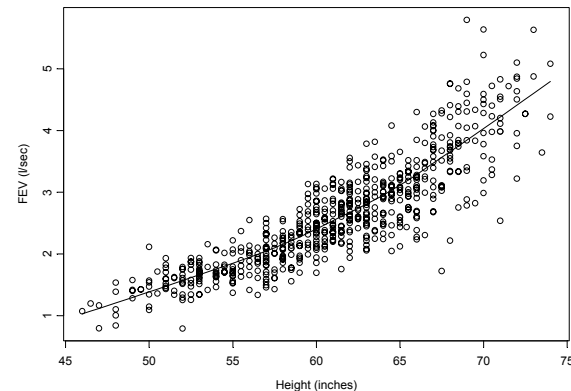
34

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

35

FEV versus Height



36

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

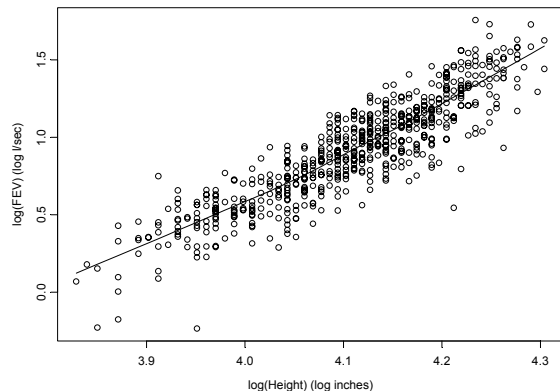
37

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

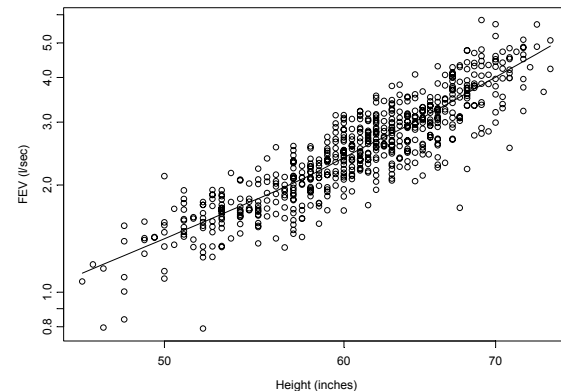
38

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

41

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

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

42

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)

43

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)

44

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

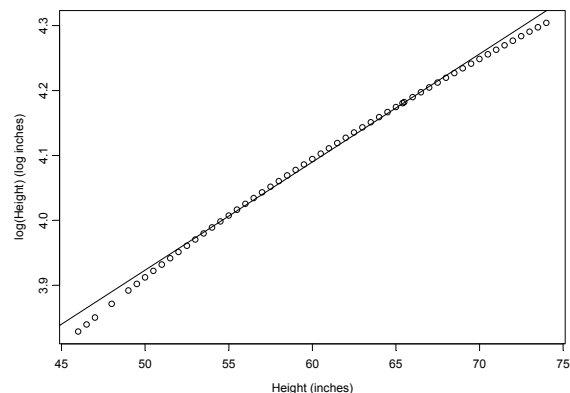
.....

- 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

49

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

50

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

51

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

52

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