

Applied Regression Analysis

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

Session 5

1

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

Relationship Between
Linear Regression
and t Tests
.....

2

Regression and t Tests



- Linear regression with a binary predictor (two groups) corresponds to the 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)

Relationship Between Linear Regression and Correlation



Regression and Correlation

- Pearson's correlation coefficient is intimately related to linear regression

–Correlation treats Y and X symmetrically, but we can relate it to the model of $E(Y | X)$ as a function of X

$$E(Y | X) = \beta_0 + \beta_1 \times X \qquad \beta_1 = \rho \frac{\sigma_Y}{\sigma_X}$$

$E(Y | X)$ mean Y within group having equal X

β_1 diff in mean Y per 1 unit diff in X

ρ true correlation between Y and X

σ_Y standard deviation of Y

σ_X standard deviation of X

5

Regression and Correlation

- More interpretable formulation of r :

$$r \approx \beta \sqrt{\frac{\text{Var}(X)}{\beta^2 \text{Var}(X) + \text{Var}(Y | X = x)}}$$

β = slope between Y and X

$\text{Var}(X)$ = variance of X in sample

$\text{Var}(Y | X = x)$ = variance of Y in groups that
have same value of X

(Vertical spread of data)

6

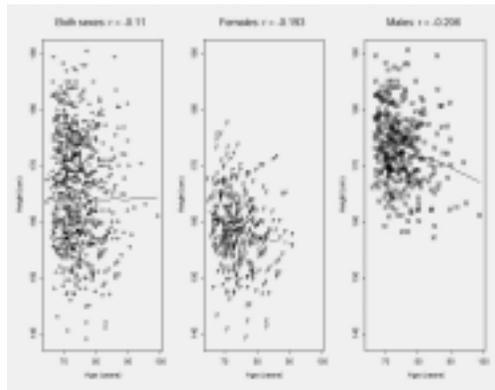
Regression and Correlation

- Correlation tends to increase in absolute value as
 - The absolute value of the slope of the line increases
 - The variance of data decreases within groups that share a common value of X
 - The variance of X increases

7

Example: Regression and Correlation

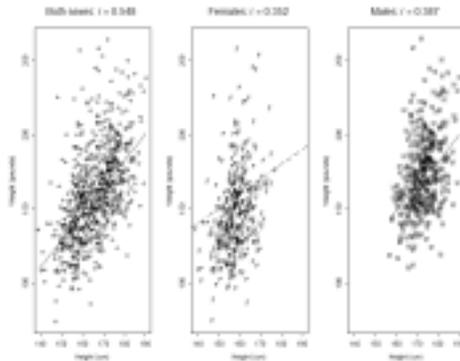
- Correlation between height and age in elderly
 - More extreme within each sex: lower $\text{Var}(Y | X)$



8

Example: Regression and Correlation.....

- Correlation between weight and height in elderly
 - More extreme in combined sexes: higher Var (X)



9

Correlation: Science vs Statistics.....

- Scientific use of correlation
 - It should be noted that
 - the slope between X and Y is of scientific interest
 - the variance of $Y|X=x$ is partly of scientific interest, but it can be affected by restricting sampling to certain values of another variable
 - E.g., var (Height | Age) is less in males than when both sexes are included
 - the variance of X is often set by study design
 - This is often not of scientific interest

10

Correlation: Science vs Statistics.....

- Ramifications for use in scientific literature
 - Two independent studies of the same phenomenon might estimate very similar slopes, but different correlations solely due to study design
 - Height vs Age
 - Males: Slope= -0.23 Corr= -0.21
 - Both: Slope= -0.20 Corr= -0.11

11

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

12

Simple Linear Regression on Log Transformed Data: Modeling theGeometric Mean.....

13

Regression on Geometric Means.....

- Geometric means of distributions are typically modeled using linear regression on log transformed data
 - The geometric mean is a common choice of population parameters 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”

14

Regression on Geometric Means.....

- Modeling of geometric mean of response Y on predictor X
 - Linear regression on log transformed Y
 - (I am using natural log)

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

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

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

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

15

Regression on Geometric Means.....

- Restated model as log link for geometric mean

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

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

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

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

16

Regression on Geometric Means.....

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

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

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

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

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

17

Regression on Geometric Means.....

- Interpretation of the model
 - 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

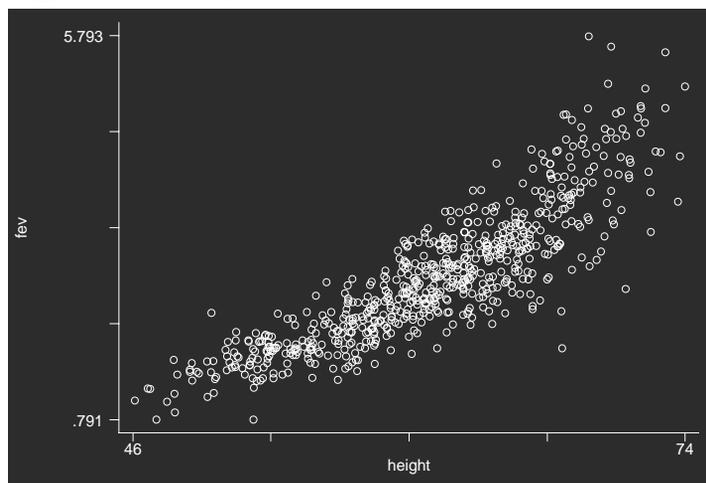
18

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

19

Scatterplot of FEV versus Height



20

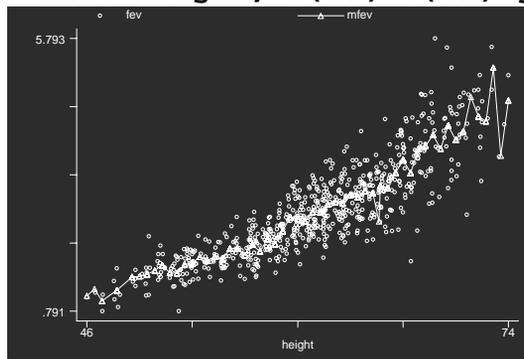
Example

- 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

21

Plot of Mean FEV versus Height

```
sort height
by height: egen mfev = mean (fev)
graph fev mfev height, s(oT) c(.1) j(1)
```



22

Example

- Choice of geometric mean for basis of model
 - Prior to looking at the data, we have good scientific justification for using geometric mean
 - FEV is a volume
 - Height is a linear dimension
 - Each dimension of lung size is likely proportional to height
 - Standard deviation likely proportional to height

Science $FEV \propto Height^3$

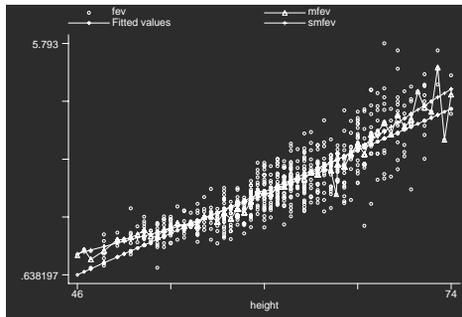
$$\sqrt[3]{FEV} \propto Height$$

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

23

Plot of Mean FEV versus Height

```
regress fev height
predict ffev
ksm fev height, lowess gen(smfev)
graph fev mfev ffev smfev height, s(oTdp)
c(.111)
```



24

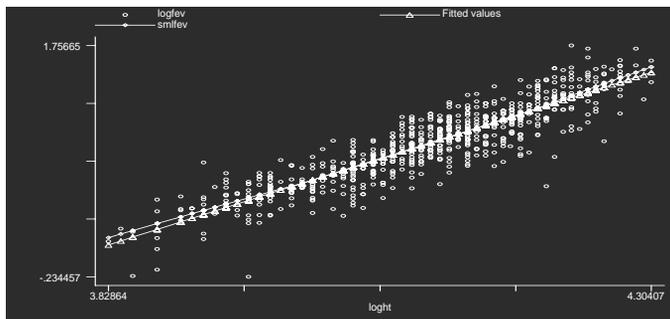
Example

- Modeling of log transformed FEV
 - 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

25

Plot of log (FEV) versus log (Height)

```
g logfev= log(fev)
g loght= log(height)
regress logfev loght
predict flfev
ksm logfev loght, lowess gen(smlfev)
graph fev flfev smlfev loght, s(oTd) c(.11)
```



26

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				[95% CI]	
logfev	Coef.	StErr	t	P> t		
loght	3.12	.068	46.15	0.000	2.99	3.26
_cons	-11.92	.278	-42.90	0.000	-12.47	-11.38

27

Log Transformed Predictors

- Interpretation of log transformed predictors with log link function
 - Log link function 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 with respect to geometric mean
 - Estimated ratio of geometric means

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

28

Interpretation of Stata Output

.....

- Scientific interpretation of the slope

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

- 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)

29

Transformation of the Predictor

.....

- Transformations of the predictor are typically chosen according to whether the model likely follows a straight line relationship
 - Linearity (“model fit”) is 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)

30

Transformation of the Predictor

.....

- It is rare that we truly know which transformation of the predictor would provide the 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

31

Transformation of the Predictor

.....

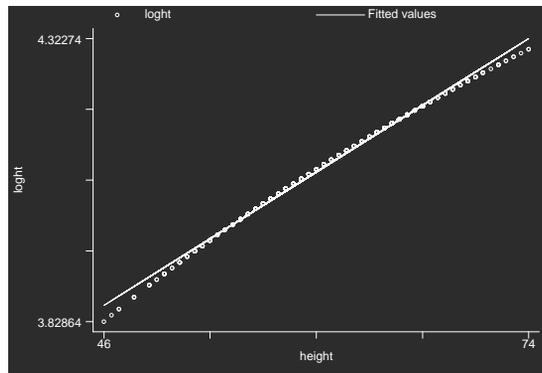
- It is best to choose the transformation of the predictor on scientific grounds
 - However, it is often the case that many functions are well approximated by a straight line over a small range of the data
 - Example: In the modeling of FEV as a function of height, the logarithm of height is approximately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

32

Transformation of the Predictor

.....

- Plot of log (Height) versus Height for FEV data with superimposed best fitting line



33

Transformation of the Predictor

.....

- 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

34

Transformation of the Predictor

.....

- 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

35

Transformation of the Predictor

.....

- Transformations of the predictor of interest should be dictated by the scientific question, which in turn depends on your level of previous knowledge about any association between response and POI
 - 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

36

Transformation of the Predictor

.....

- 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

37

Transformation of the Predictor

.....

- When modeling confounding variables, 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

38

Transformation of the Predictor

.....

- 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