Biost 517
Applied Biostatistics I

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

Lecture 17:
Simple Linear Regression

December 2, 2009

1

Lecture Outline

- General Regression Setting
- Motivating Example
- Simple Linear Regression
- Relationship to Correlation
- Relationship to t Tests
- Inference about Geometric Means

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

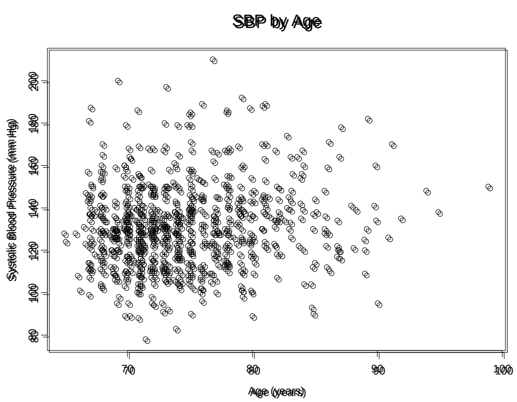
5

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

Motivating Example

11

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

12

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

13

Example: Regression Model

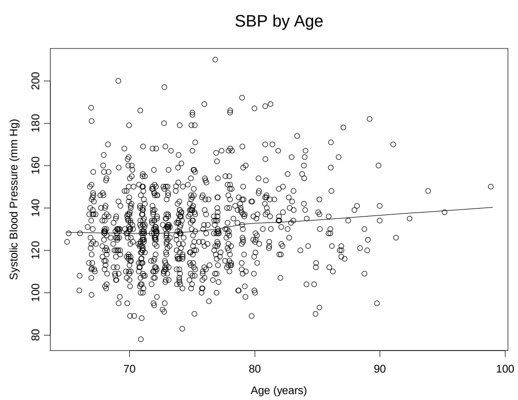
- Answer question by assessing linear trends in, say, average SBP by age
  - Estimate best fitting line to average SBP within age groups

$$E\left(SBP \mid Age \right) = \beta_0 + \beta_1 \times Age$$

- – An association will exist if the slope (β_1) is nonzero
  - In that case, the average SBP will be different across different age groups

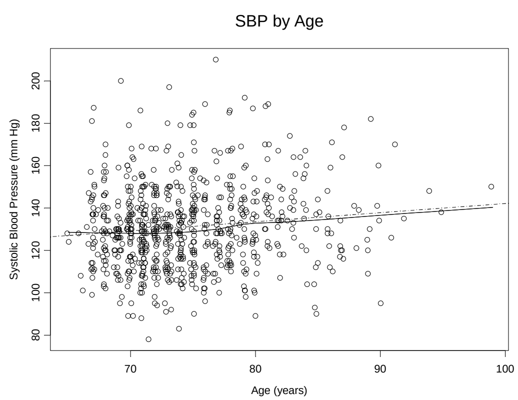
14

Example: Scatterplot



15

Example: Smooth; LS Line



16

"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"

$$E\left(SBP \mid Age \right) = 100 + 0.5 \times Age$$

17

Example: Estimates, Inference

```
. regress sbp age
                                        Number of obs =     735
   Source |      SS   df      MS        F(  1,    733) =   10.63
    Model |    4056    1 4056.4         Prob > F      =  0.0012
 Residual | 279740 733   381.6          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
```

$$E\left(SBP \mid Age \right) = 98.9 + 0.431 \times Age$$

18

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

19

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)

20

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 I average difference in mean SBP between groups differing by 1 year in age
    - Statistical jargon: a "contrast" across the means

21

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)

22

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."

23

Simple Linear Regression

24

Ingredients: Response

- The distribution of this variable will be compared across the groups
    - Linear regression models the mean of this variable
    - Log transformation of the response corresponds to modeling the geometric mean
  - Notation:
    - It is extremely common (99 of 100 statisticians agree) to use Y to denote the response variable when discussing general methods

25

Ingredients: Predictor

- Predictor (grouping) variables
  - Group membership is measured by a variable
  - Notation
    - When not using mnemonics, I will tend to use X to denote a predictor variable
    - (When we proceed to multiple regression, I will use subscripts to denote different predictors)

26

Ingredients: Regression 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

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

27

Deterministic World: 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*

28

With Variability: Statistics

- 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

29

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

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

30

Derivation of Interpretation

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

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

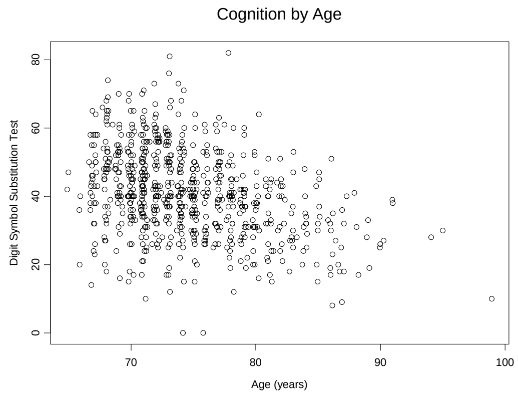
31

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

32

Example: Scatterplot



33

Example: Stratified Descriptives

Age	N	Nonmsgn	Mean	Std Dev
67	4	4	39.25	11.03
68	22	21	44.05	12.50
69	79	79	46.62	12.40
70	72	71	44.85	12.63
71	69	68	47.09	10.85
72	75	75	42.19	12.86
73	64	64	43.22	10.06
74	39	39	41.15	12.21
75	44	44	40.84	15.76
76	32	32	39.03	11.41
77	39	37	40.11	12.69

34

Example: Stratified Descriptives

Age	N	Nonmsgn	Mean	Std Dev
78	36	36	38.56	11.11
79	33	33	36.61	9.78
80	28	28	36.21	8.90
81	19	19	32.95	11.84
82	15	14	30.93	8.94
83	12	12	35.08	9.06
84	14	12	29.92	12.18
85	9	9	35.56	9.37
86	7	7	18.43	5.71

35

Example: Stratified Descriptives

Age	N	Nonmsgn	Mean	Std Dev
87	5	4	31.50	8.50
88	5	5	33.60	12.72
89	5	4	26.25	6.70
90	3	1	26.00	
91	1	1	38.00	
92	2	2	33.50	7.78
93	1	1	30.00	
97	1	1	10.00	

36

Stata: Plot of Stratified Means

- Using “egen” to get group specific statistics

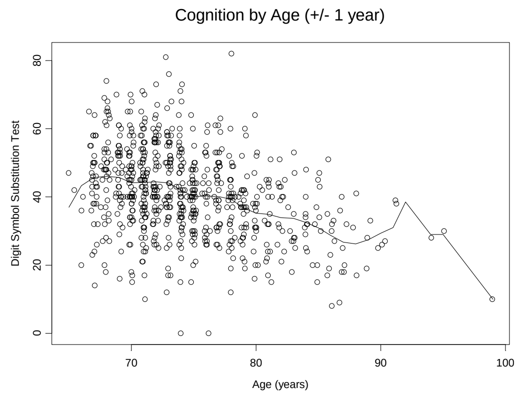
```
.sort age
.by age: egen mdsst = mean (dsst)
.twoway (scatter dsst age,
  jitter(3) (line mdsst age)
```

37

Example: Stratified Means

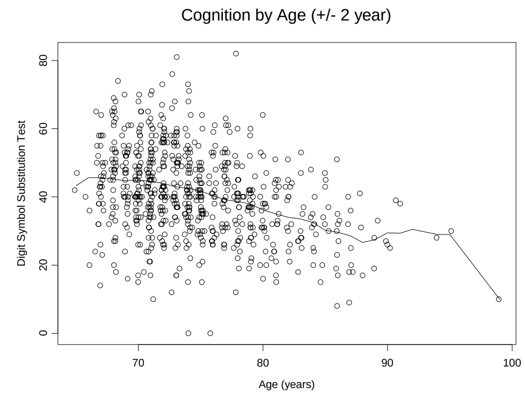
38

Example: Moving Average (± 1y)



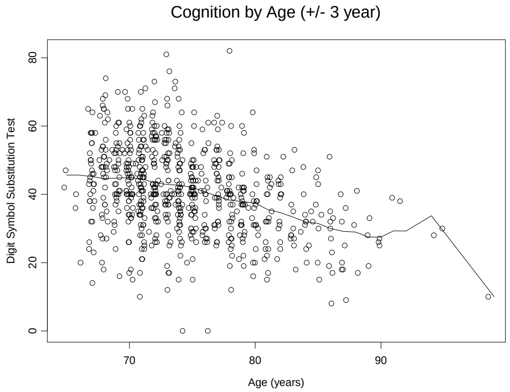
39

Example: Moving Average (± 2y)



40

Example: Moving Average (± 3y)



41

Least Squares Estimation

```
. 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
```

42

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
```

43

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

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

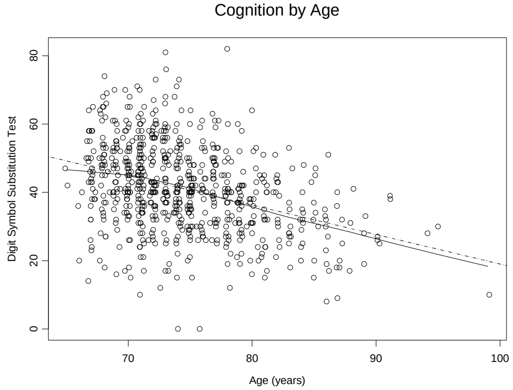
44

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

45

Example: Lowess, LS Line



46

Interpretation of Intercept

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

- 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

47

Interpretation of Slope

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

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

48

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

49

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

50

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

51

Estimates of Error Distribution

- The error distribution is estimated from the residuals

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

- 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"

52

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

53

Inference with Regression

- Most commonly encountered questions
  - Prediction
    - Estimating a future observation of response Y
    - Often we use the mean or geometric mean
  - 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

54

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

55

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

56

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

57

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

58

Statistical Validity of Inference

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

59

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)

60

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

61

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

62

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

63

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

64

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

65

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

66

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

67

Classical Linear Regression

- Inference about slope <u>tests</u> 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

68

Robust Standard Errors

- Inference about slope <u>tests</u> 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

69

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

70

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)

71

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)

72

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

73

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
```

74

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

75

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
```

76

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?)

77

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

78

Choosing the Correct Model

"All models are false, some models are useful."

- George Box

79

Choosing the Correct Model

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

- Unknown

80

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

81

Model Checking

"Blood suckers hide 'neath my bed"

"Eyepennies", Mark Linkous (Sparklehorse)

82

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)

83

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?

84

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

85

Relationship Between Linear Regression and Correlation

86

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 E(Y | X) as a function of X

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

$E(Y \mid X)$ 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

87

Regression and Correlation

- More interpretable formulation of r :

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

β = slope between Y and X

$Var(X)$ = variance of X in sample

$Var(Y \mid X = x)$ = variance of Y in groups that have same value of X (Vertical spread of data)

88

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

89

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

90

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

91

Relationship Between Linear Regression and t Tests

92

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)

93

Example: DSST and Stroke

- Association between DSST and stroke (cerebrovascular accident- CVA)
  - CVA is a binary predictor
  - Compare
    - t test with equal variances and classical linear regression
      - Estimates, standard errors, CI, P values exactly equal
    - t test with unequal variances and robust SE
      - Estimates exactly equal; standard errors, CI, P values approximately equal

94

Classical LS vs Equal Var t Test

```
. ttest dsst, by(cva)
Two-sample t test with equal variances
Grp |      Mean Std. Err. Std. Dev.   [95% Conf Interval]
  0 | 41.70507 .4847756  12.3689     40.75315    42.65698
  1 | 35.19444 1.677038  14.23014    31.85053    38.53836
diff| 6.510625 1.56047               3.447018    9.574232
       t =   4.1722           Pr(|T|>|t|)= 0.0000

. regress dsst cva
dsst |     Coef.  StdErr    t   P>|t| [95% Conf. Interval]
 cva | -6.510625 1.56047 -4.17 0.000 -9.574232   -3.447018
_cons| 41.70507  .492439 84.69 0.000 40.73828    42.67185
```

95

Classical LS vs Equal Var t Test

- Note correspondences
  - Group 0
    - Sample mean reported in t test is exactly the same as intercept reported in classical regression
      - Standard error, CI differ because regression uses a pooled standard deviation
  - Difference between group means
    - Estimate, standard error, CI, P values from t test are exactly the same as slope, SE, CI, P values from classical least squares regression

96

Robust SE vs Uneq Var t Test

```
. ttest dsst, by(cva) unequal
Two-sample t test with unequal variances
Grp |      Mean  Std. Err. Std. Dev.  [95% Conf Interval]
  0 | 41.70507  .4847756  12.3689   40.75315    42.65698
  1 | 35.19444  1.677038  14.23014  31.85053    38.53836
diff| 6.510625  1.745699            3.038684    9.982566
       t =   3.7295           Pr(|T| > |t|) = 0.0003

. regress dsst cva, robust
     |            Robust
dsst |      Coef. Std Err.    t   P>|t|  [95% Conf Intval]
 cva | -6.510625 1.736774 -3.75 0.000 -9.92036 -3.10089
_cons|  41.70507 .4850745 85.98 0.000 40.75274  42.6574
```

97

Classical LS vs Equal Var t Test

- Note correspondences
  - Group 0
    - Sample mean reported in t test is exactly the same as intercept reported in regression
      - Standard error, CI differ because regression uses a pooled standard deviation
  - Difference between group means
    - Estimate from t test is exactly the same as slope
    - Standard error, CI, P values from t test differ only slightly from regression with robust SE
      - Has to do with using n versus n-2 in variance estimates

98

Inference for the Geometric Mean

Simple Linear Regression on Log Transformed Data

99

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”

100

Interpretation of Parameters

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

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

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

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

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

101

Interpretation of Parameters

- Restated model as log link for geometric mean

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

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

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

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

102

Interpretation of Parameters

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

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

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

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

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

103

Interpretation of Parameters

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

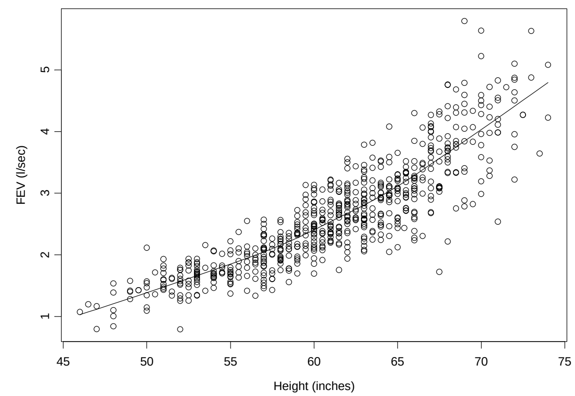
104

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

105

FEV versus Height



106

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

107

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

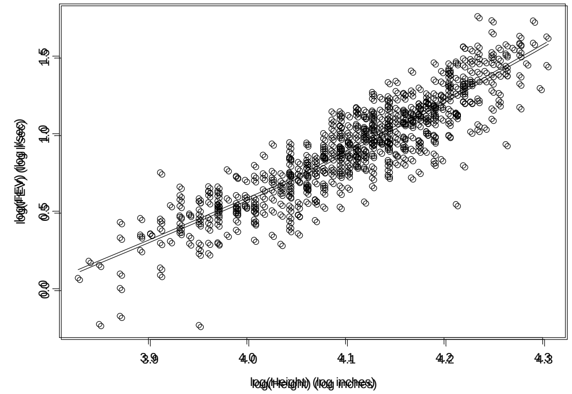
108

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

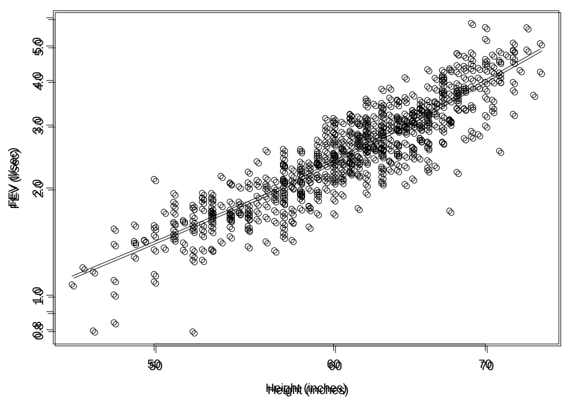
109

log(FEV) versus log(Height)



110

log-log Plot of FEV vs Height



111

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
```

112

Log Transformed Predictors

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

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

113

Interpretation of Stata Output

- Scientific interpretation of the slope

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

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

114

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)

115

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

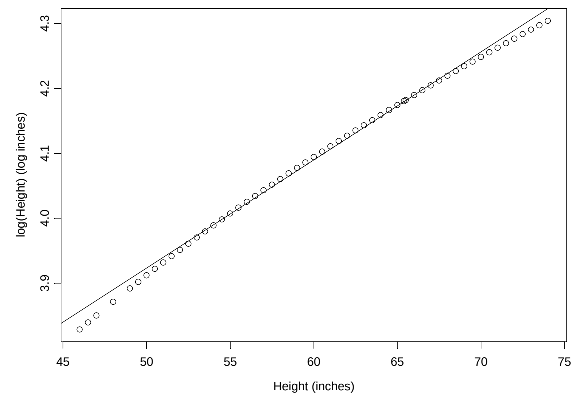
116

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

117

log(Height) versus Height



118

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

119

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

120

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

121

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

122

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

123

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

124