Biost 518 / Biost 515
Applied Biostatistics II / Biostatistics II

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

Lecture 2:
Simple Linear Regression Model

January 7, 2015

1

Lecture Outline

- Motivating Example
- Simple Linear Regression Models
- Inference About Geometric Means

2

Motivating Example

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

4

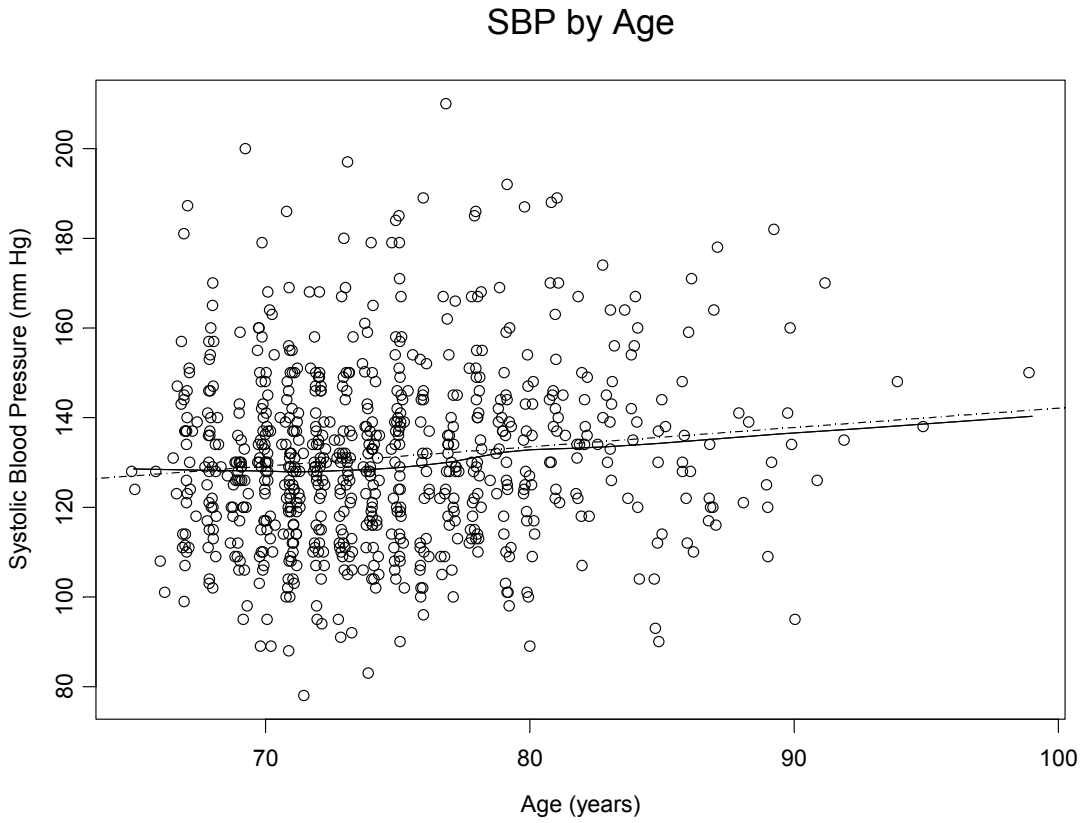
Example: Definition of Variables

- 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

5

Example: Descriptive Statistics

- Graphical: Jittered scatterplot with superimposed smooth, ?LS fit
  - Response on y-axis, predictor on x-axis



6

Example: Descriptive Statistics (R)

- Tabular: Stratified descriptive statistics
  - Strata by scientifically relevant categories (not quintiles)

```
age5 <- 5*trunc((age - 65)/5)+65

descrip(sbp,strata=age5)
                N Msng  Mean   Std Dev  Min   25%   Mdn   75%   Max
sbp:All       735   0   131.1   19.66    78   118   130   142   210
sbp:  Str 65  117   0   129.3   18.24    95   118   128   137   200
sbp:  Str 70  305   0   128.8   18.38    78   116   129   138   197
sbp:  Str 75  187   0   132.8   21.28    90   117   130   144   210
sbp:  Str 80   81   0   138.0   19.81    89   124   137   148   189
sbp:  Str 85   35   0   129.8   21.60    90   116   128   138   182
sbp:  Str 90    8   0   138.6   22.78    95   132   138   151   170
sbp:  Str 95    2   0   144.0   8.485   138   141   144   147   150
```

7

Example: Descriptive Statistics (Stata)

- Tabular: Stratified descriptive statistics
  - Strata by scientifically relevant categories (not quintiles)

```
. g age5 = 5 * int((age - 65) / 5)
. egen age5 = cut(age) at(65(5)100)  // alternative approach

. tabstat sbp, by(age5) col(stat) stat(n mean sd min q max) format

Summary for variables: sbp
     by categories of: age5
```

age5	N	mean	sd	min	p25	p50	p75	max
65	117	129	18	95	118	128	137	200
70	305	129	18	78	116	129	138	197
75	187	133	21	90	117	130	144	210
80	81	138	20	89	124	137	148	189
85	35	130	22	90	116	128	139	182
90	8	139	23	95	130	138	154	170
95	2	144	8	138	138	144	150	150
Total	735	131	20	78	118	130	142	210

8

General Regression

- General notation for variables and parameter

Y_i Response measured on the ith subject

X_i Value of the POI for the ith subject

θ_i Parameter of distribution of Y_i

- The parameter might be the mean, geometric mean, odds, rate, instantaneous risk of an event (hazard), etc.

9

Simple Regression

- General notation for multiple regression model

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

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

β_0 "Intercept"

β_1 "Slope for Pred of Interest X"

- The link function is usually either none (means) or log (geom mean, odds, hazard)
- (With binary data we sometimes also consider
  - logit link: log [p / (1-p)]
  - complementary log log link: log (- log p) = log λ + log Δ
    - for proportions measuring cumulative incidence, which relate to log link on exponential hazards)

10

Example: Linear 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

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

11

"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"

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

- Linear regression estimates parameters using "least squares"
  - Most efficient average-based estimates for homoscedastic data
    - Asymptotically normal distribution for estimates
  - (Most efficient estimation when data is normal within groups
    - Normal distribution for estimates even in small sample sizes)

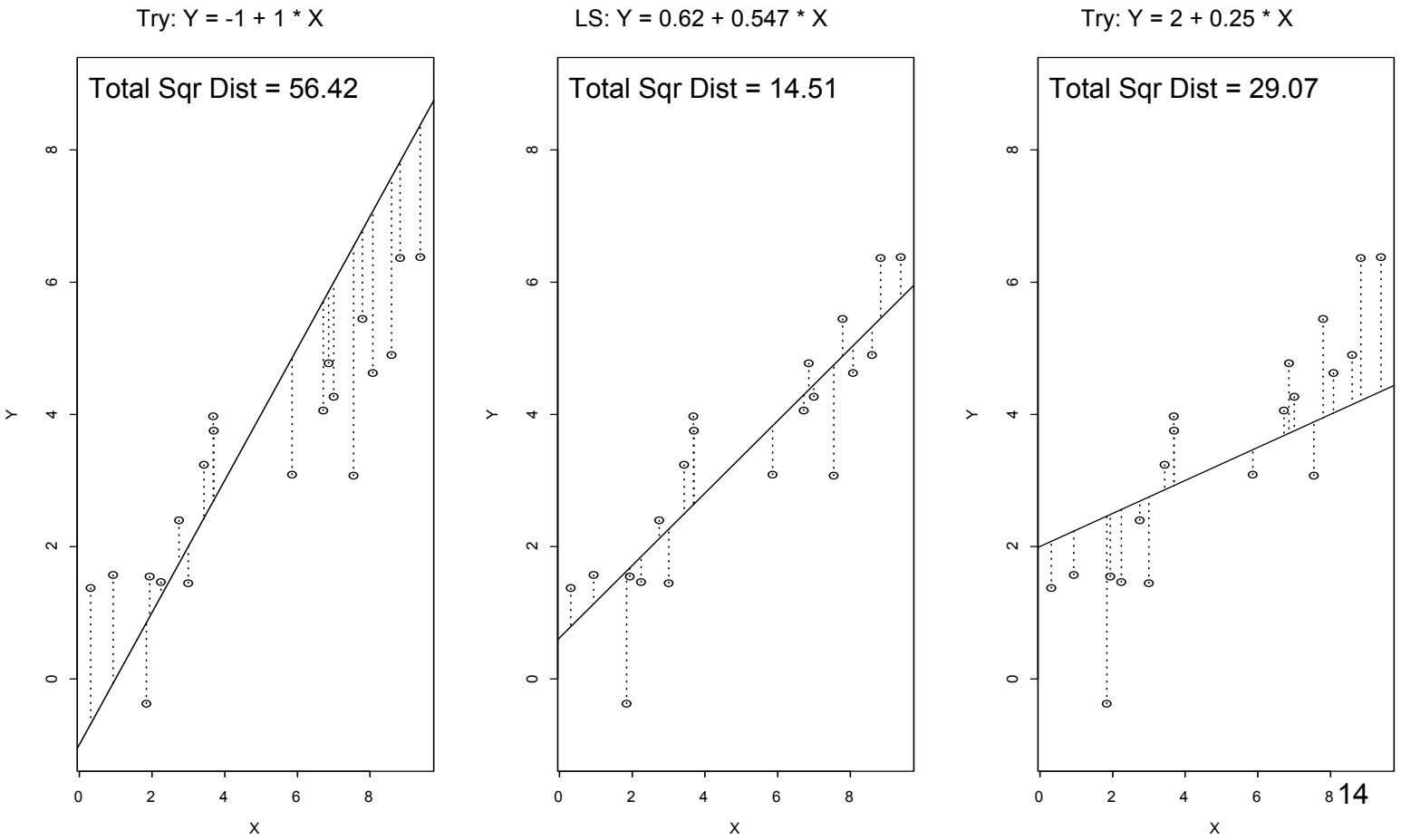
12

Least Squares Line

- Find the straight line that minimizes total squared vertical distance from data to line
  - Conceptually: Trial and error search
    - Guess a formula for a line
    - Compute total squared distance from data to line
    - Iterate until smallest number found
  - Calculus:
    - Find a formula based on derivatives
  - Real life:
    - Computers find such estimates easily

13

Conceptual Example



Example: Estimates, Inference in R

```
regress("mean", sbp, age)
reg(y ~ sbp, data=mri)

Residuals:
    Min      1Q  Median      3Q     Max
-51.568 -13.843  -0.568  10.432  77.845

Coefficients:
          Estimate NaiveSE RobustSE 95%L  95%H  F stat df
   Pr(>F)
Intercept   98.9    9.89   9.82    79.7  118    101.6  1  <
   0.00005
age        0.431   0.132  0.132   0.172  0.691  10.65  1
   0.0012

Residual standard error: 19.54 on 733 degrees of freedom
Multiple R-squared: 0.01429,  Adjusted R-squared: 0.01295
F-statistic: 10.63 on 1 and 733 DF,  p-value: 0.001165
```

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

Example: Estimates, Inference in Stata

```
. 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
```

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

16

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

17

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
    - (Later: splines)

18

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
  - Regression in some sense considers all possible pairwise contrasts, and then averages them in a special way

- 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

19

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)

20

Example: Interpretation

"From linear regression analysis, we estimate that for each year difference in age between two populations, 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 two sided P value is P < .0005, we reject the null hypothesis that there is no linear trend in the average SBP across age groups."

21

Example: Interpretation

- Note specification of point estimate, CI, and p value
  - Response: SBP (measured in mmHg)
  - Summary measure: mean
  - Contrast of summary measure across groups: difference
  - Predictor of interest: age
  - Difference in POI across groups being compared: 1 year

"From linear regression analysis, we estimate that for **each year difference in age** between two populations, 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 two sided P value is P < .0005, we reject the null hypothesis that there is no linear trend in the average SBP across age groups." 22

Simple Linear Regression

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

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

24

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

25

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

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

26

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

27

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

28

Example: Lowess, LS Line

29

Least Squares Estimation

```
. regress dsst age

  Source |       SS       df       MS   Nbr of obs =     723
---------+------------------------------   F(1, 721)  =  109.57
   Model |  15377.0   1  15377.0          Prob > F   =  0.0000
Residual |  101191.0 721  140.3           R-squared  =  0.1319
---------+------------------------------   Adj R-sqr  =  0.1307
   Total |  116569.0 722  161.4           Root MSE   =  11.847

------------------------------------------------------------------------------
        dsst |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    -.8630    .0825    -10.47   0.000    -1.025    -.7010
       _cons |     105.0     6.16     17.11   0.000     92.93     117.1
------------------------------------------------------------------------------
```

30

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.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.8631   .0825    -10.46   0.000    -1.025    -.7011
       _cons |   105     6.16      17.04   0.000     93.3      117
------------------------------------------------------------------------------
```

31

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

```
  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
```

32

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

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

  - Example: Fitted value for 70 year olds;

$$E\left[DSST_i \,\middle|\, Age_i = 70\right] = 105 - 0.863 \times 70 = 44.59$$

33

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

```
 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
```

34

Interpretation of Intercept

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

- 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

35

Reparameterization: Location

- It is possible to reparameterize our model in order to make the intercept more interpretable
  - Two models are the same if they have the same fitted values

$$\text{Original model:} \quad E(Y \mid X) = \beta_0 + \beta_1 \times X$$

$$\text{Recenter } X: \quad X^* = X - 65$$

$$\begin{aligned}\text{Reparameterization:} \quad E(Y \mid X) &= \beta_0^* + \beta_1^* \times X^* \\ &= \beta_0^* + \beta_1^* \times (X - 65) \\ &= (\beta_0^* - \beta_1^* \times 65) + \beta_1^* \times X \\ &= \beta_0 + \beta_1 \times X\end{aligned}$$

36

Reparameterization: Intercept Changes

```
. g yrabove65= age-65
. regress dsst yrabove65
      Source |       SS       df       MS              Number of obs =     723
-------------+------------------------------           F(  1,   721) =  109.57
       Model |  15377.4797     1  15377.4797            Prob > F      =  0.0000
    Residual |  101191.195   721  140.348398            R-squared     =  0.1319
-------------+------------------------------           Adj R-squared =  0.1307
       Total |  116568.675   722  161.452458            Root MSE      =  11.847

------------------------------------------------------------------------------
        dsst |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   yrabove65 |  -.8633297   .0824779   -10.47   0.000    -1.025255   -.7014041
       _cons |   49.22311   .8959871    54.94   0.000     47.46406    50.98217
------------------------------------------------------------------------------

. regress dsst age
      Source |       SS       df       MS              Number of obs =     723
-------------+------------------------------           F(  1,   721) =  109.57
       Model |  15377.4797     1  15377.4797            Prob > F      =  0.0000
    Residual |  101191.195   721  140.348398            R-squared     =  0.1319
-------------+------------------------------           Adj R-squared =  0.1307
       Total |  116568.675   722  161.452458            Root MSE      =  11.847

------------------------------------------------------------------------------
        dsst |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.8633297   .0824779   -10.47   0.000    -1.025255   -.7014041
       _cons |   105.3395   6.157026    17.11   0.000     93.25171    117.4274
------------------------------------------------------------------------------
```

Interpretation of New Intercept

$$E\left[DSST_i \mid YrAbove65_i\right] = 49.2 - 0.863 \times YrAbove65_i$$

- Estimated mean DSST for 65 year olds is 49.2

- In this parameterization, the intercept has more relevance to our sampling scheme
  - But it is still not all that relevant to our question about associations between DSST and age

38

Interpretation of Slope

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

- 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)

39

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

40

Reparameterization: Scale

- It is possible to reparameterize our model in order to make the slope more interpretable
  - Two models are the same if they have the same fitted values

$$\text{Original model:} \quad E(Y \mid X) = \beta_0 + \beta_1 \times X$$

$$\text{Rescale } X \text{ to decades:} \quad X^* = X/10$$

$$\begin{aligned}\text{Reparameterization: } E(Y \mid X) &= \beta_0^* + \beta_1^* \times X^* \\ &= \beta_0^* + \beta_1^* \times (X/10) \\ &= \beta_0^* + (\beta_1^*/10) \times X \\ &= \beta_0 + \beta_1 \times X\end{aligned}$$

41

Reparameterization: Rescale Slope by 10

```
. g ageD= age/10
. regress dsst ageD
      Source |       SS       df       MS              Number of obs =     723
-------------+------------------------------           F(  1,   721) =  109.57
       Model |  15377.4797     1  15377.4797            Prob > F      =  0.0000
    Residual |  101191.195   721  140.348398            R-squared     =  0.1319
-------------+------------------------------           Adj R-squared =  0.1307
       Total |  116568.675   722  161.452458            Root MSE      =  11.847

        dsst |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
        ageD |  -8.633297   .8247794   -10.47   0.000    -10.25255   -7.014041
       _cons |   105.3395   6.157026    17.11   0.000     93.2517    117.4274


. regress dsst age
      Source |       SS       df       MS              Number of obs =     723
-------------+------------------------------           F(  1,   721) =  109.57
       Model |  15377.4797     1  15377.4797            Prob > F      =  0.0000
    Residual |  101191.195   721  140.348398            R-squared     =  0.1319
-------------+------------------------------           Adj R-squared =  0.1307
       Total |  116568.675   722  161.452458            Root MSE      =  11.847

        dsst |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
         age |  -.8633297   .0824779   -10.47   0.000    -1.025255   -.7014041
       _cons |   105.3395   6.157026    17.11   0.000     93.25171    117.4274
```

Regression in R

- Inference based on either classical linear regression or robust standard errors
  - Classical linear regression (assume homoscedasticity)
    - **"regress("mean",respvar,predictor,robustSE=F)"**
    - **"reg(respvar ~ predictor, robustSE=F)**
      - E.g., **reg(dsst ~ age, robustSE=F)**
  - Robust standard error estimates (allows heteroscedasticity)
    - **"regress("mean",respvar,predictor)"**
    - **"reg(respvar ~ predictor)**
      - E.g., **reg(dsst ~ age)**
  - The two approaches differ in CI and P values, not estimates

43

Regression in Stata

- Inference based on either classical linear regression or robust standard errors
  - Classical linear regression (assume homoscedasticity)
    - **"regress respvar predictor"**
      - E.g., **regress dsst age**
  - Robust standard error estimates (allows heteroscedasticity)
    - **"regress respvar predictor, robust"**
      - E.g., **regress dsst age, robust**
  - The two approaches differ in CI and P values, not estimates

44

Ex: Classical Linear Regression

```
. regress dsst age

  Source |       SS       df       MS   Nbr of obs =     723
---------+------------------------------   F(1, 721)  =  109.57
   Model |  15377.0   1  15377.0          Prob > F   =  0.0000
Residual |  101191.0 721  140.3           R-squared  =  0.1319
---------+------------------------------   Adj R-sqr  =  0.1307
   Total |  116569.0 722  161.4           Root MSE   =  11.847

------------------------------------------------------------------------------
        dsst |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    -.8631    .0825    -10.47   0.000    -1.025    -.7012
       _cons |     105.0    6.161     17.11   0.000     92.93     117.1
------------------------------------------------------------------------------
```

45

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?)
  - The "overall F test" tests that some variable in the model matters
    - In simple linear regression, there is only one variable
    - Equivalent to the t test for the slope: p values will agree exactly
    - $F = 109.57 = (-10.4676)^2$

- 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

46

What if Heteroscedastic?

- What if the variances within each group are not equal?
- With t test we knew
  - Group with small sample size and higher variance ➔
    - t test that presumes equal variance is anti-conservative inference
      - Reported p values are too small
      - Reported CI is too narrow
  - Group with small sample size and lower variance ➔
    - t test that presumes equal variance is conservative inference
      - Reported p values are too high
      - Reported CI is too wide
- With linear regression similar findings for skewness of X
  - Anti-conservative inference if higher within group variance (Var[Y|X]) in outlying values of X
  - Conservative inference if lower within group variance (Var[Y|X]) in outlying values of X

47

Example: Lowess, LS Line

48

Example: Stratified Descriptives

- By scientifically relevant intervals

```
. tabstat dsst, by(age5) col(stat) stat(n mean sd min q max) format

Summary for variables: dsst
     by categories of: age5
```

age5	N	mean	sd	min	p25	p50	p75	max
65	117.0	45.6	12.4	14.0	38.0	47.0	53.0	74.0
70	302.0	43.9	12.7	0.0	36.0	43.0	53.0	81.0
75	185.0	39.2	11.0	0.0	31.0	40.0	45.0	82.0
80	78.0	34.1	10.0	15.0	27.0	32.0	40.0	64.0
85	33.0	28.6	10.8	8.0	19.0	30.0	35.0	51.0
90	6.0	30.5	6.3	25.0	26.0	27.5	38.0	39.0
95	2.0	20.0	14.1	10.0	10.0	20.0	30.0	30.0
Total	723.0	41.1	12.7	0.0	32.0	40.0	50.0	82.0

49

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   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.8631   .0755   -11.43   0.000    -1.01   -.715
       _cons |     105    5.71    18.45   0.000     94.1     117
------------------------------------------------------------------------------
```

50

Estimates the Same

```
. 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% Conf Int]
  age | -.863   .0825 -10.47  0.000  -1.03  -.701
_cons |   105    6.16  17.11  0.000   93.3    117


. regress dsst age, robust
                               Nbr of obs   =     723
                               F(1, 721)    =  130.72
                               Prob > F     =  0.0000
                               R-squared    =  0.1319
                               Root MSE     =  11.847      (but not as relevant)

      |        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
```

51

Inference is Different

```
. 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% Conf Int]
  age | -.863   .0825 -10.47  0.000  -1.03  -.701
_cons |   105    6.16  17.11  0.000   93.3    117


. regress dsst age, robust
                              Nbr 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
```

52

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?)
  - Again, F = 130.72 = $(-11.43)^2$

53

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

54

Choosing the Correct Model

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

- George Box

55

Choosing the Correct Model

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

- Unknown

56

Example: Interpretation

"From linear regression analysis using Huber-White estimates of the standard error, we estimate that for each year difference in age between two populations, the difference in mean DSST is 0.863 points lower in the older population. A 95% CI suggests that this observation is not unusual if the true difference in mean DSST were between .715 and 1.01 points lower per year difference in age. Because the two sided P value is P < .0005, we reject the null hypothesis that there is no linear trend in the average DSST across age groups."

57

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

58

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

59

Estimates of Error Distribution

- The error distribution is estimated from the residuals

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

  - 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"

60

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
    - (Gives a very rough idea of the magnitude of variances if heteroscedastic)

61

Relationships to Previous Methods: Corr

- Classical simple linear regression
  - Test for slope is exactly the test for significant correlation
  - R^2 in simple LR is squared correlation: .0033 = .0573^2

```
. pwcorr dsst weight, sig
             |     dsst   weight
-------------+------------------
        dsst |   1.0000
             |
      weight |   0.0573   1.0000
             |   0.1239

. regress dsst weight
      Source |       SS       df       MS              Number of obs =     723
-------------+------------------------------           F(  1,   721) =    2.37
       Model |  382.385284     1  382.385284           Prob > F      =  0.1239
    Residual |   116186.29   721  161.146033           R-squared     =  0.0033
-------------+------------------------------           Adj R-squared =  0.0019
       Total |  116568.675   722  161.452458           Root MSE      =  12.694

------------------------------------------------------------------------------
        dsst |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   .0236155   .0153305     1.54   0.124    -.0064822    .0537131
       _cons |   37.27787   2.498129    14.92   0.000     32.37339    42.18235
------------------------------------------------------------------------------
```

Relationships to Previous Methods: T test

- 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 male, robust`"

- Classical simple linear regression
  - Test for slope is exactly the test for significant correlation

63

Binary Predictor Example: Estimates

- A "saturated model": Number of groups = number of parameters
  - The predictor variable used in the analysis only had two values
  - The regression model has two parameters
  - We are not borrowing information across the groups for the mean
  - Each group mean can be fit exactly
    - Intercept is the sample mean for females
    - Intercept plus slope is the sample mean for males

- We could of course reparameterize our model
  - **`female = 1 - male`**
  - **`regress dsst female`**
    - Then intercept would be the sample mean for males
    - Intercept plus slope would be sample mean for females

64

Example: DSST by Sex (female reference)

```
. ttest dsst, by(male)
Two-sample t test with equal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     367    42.42779    .6680565    12.79812    41.11408    43.7415
       1 |     356    39.64326    .6610097    12.47191    38.34327   40.94325
---------+--------------------------------------------------------------------
combined |     723    41.05671    .4725559    12.70639    40.12896   41.98446
---------+--------------------------------------------------------------------
    diff |            2.784534    .9401746                .9387276   4.630341
------------------------------------------------------------------------------
diff = mean(0) - mean(1)                                      t =   2.9617
Ho: diff = 0                     degrees of freedom =      721

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9984         Pr(|T| > |t|) = 0.0032          Pr(T > t) = 0.0016

. regress dsst male
      Source |       SS       df       MS              Number of obs =     723
-------------+------------------------------           F(  1,   721) =    8.77
       Model |  1401.14463     1  1401.14463           Prob > F      =  0.0032
    Residual |   115167.53   721  159.733052           R-squared     =  0.0120
-------------+------------------------------           Adj R-squared =  0.0106
       Total |  116568.675   722  161.452458           Root MSE      =  12.639

------------------------------------------------------------------------------
        dsst |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   -2.784534   .9401746    -2.96   0.003    -4.630341   -.9387276
       _cons |    42.42779   .6597272    64.31   0.000     41.13258    43.72301
------------------------------------------------------------------------------
```

Example: DSST by Sex (male reference)

```
. ttest dsst, by(male)
Two-sample t test with equal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     367    42.42779    .6680565    12.79812    41.11408     43.7415
       1 |     356    39.64326    .6610097    12.47191    38.34327    40.94325
---------+--------------------------------------------------------------------
combined |     723    41.05671    .4725559    12.70639    40.12896    41.98446
---------+--------------------------------------------------------------------
    diff |            2.784534    .9401746                .9387276    4.630341
------------------------------------------------------------------------------
diff = mean(0) - mean(1)                                      t =   2.9617
Ho: diff = 0                     degrees of freedom =      721

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9984         Pr(|T| > |t|) = 0.0032          Pr(T > t) = 0.0016

. regress dsst female
      Source |       SS       df       MS              Number of obs =     723
-------------+------------------------------           F(  1,   721) =    8.77
       Model |  1401.14463     1  1401.14463           Prob > F      =  0.0032
    Residual |   115167.53   721  159.733052           R-squared     =  0.0120
-------------+------------------------------           Adj R-squared =  0.0106
       Total |  116568.675   722  161.452458           Root MSE      =  12.639

------------------------------------------------------------------------------
        dsst |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |    2.784534   .9401746     2.96   0.003     .9387276    4.630341
       _cons |    39.64326    .669842    59.18   0.000     38.32818    40.95833
------------------------------------------------------------------------------
```

66

Binary Predictor Example: Classical LR

- Inference from classical linear regression corresponds to t test that presumes equal variances
- t test for equal variances p value is exactly the test for nonzero slope
- CI for slope is exactly CI for difference in means

67

Example: DSST by Sex

```
. ttest dsst, by(male)
Two-sample t test with equal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     367    42.42779    .6680565    12.79812    41.11408     43.7415
       1 |     356    39.64326    .6610097    12.47191    38.34327    40.94325
---------+--------------------------------------------------------------------
combined |     723    41.05671    .4725559    12.70639    40.12896    41.98446
---------+--------------------------------------------------------------------
    diff |            2.784534    .9401746                .9387276    4.630341
------------------------------------------------------------------------------
diff = mean(0) - mean(1)                                      t =   2.9617
Ho: diff = 0                     degrees of freedom =      721

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9984         Pr(|T| > |t|) = 0.0032          Pr(T > t) = 0.0016

. regress dsst male
      Source |       SS       df       MS              Number of obs =     723
-------------+------------------------------           F(  1,   721) =    8.77
       Model |  1401.14463     1  1401.14463           Prob > F      =  0.0032
    Residual |   115167.53   721  159.733052           R-squared     =  0.0120
-------------+------------------------------           Adj R-squared =  0.0106
       Total |  116568.675   722  161.452458           Root MSE      =  12.639

------------------------------------------------------------------------------
        dsst |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   -2.784534   .9401746    -2.96   0.003    -4.630341   -.9387276
       _cons |    42.42779   .6597272    64.31   0.000     41.13258    43.72301
------------------------------------------------------------------------------
```

Binary Predictor Example: Classical LR

- However, the CI for the intercept is not the CI for the females printed with the t test output, because in regression we use the pooled SD

One sample $$\bar{Y}_F \pm t_{.025,n_F-1} \times \frac{s_F}{\sqrt{n_F}}$$

Regression $$\hat{\beta}_0 \pm t_{.025,n_M+n_F-1} \times \frac{RMSE}{\sqrt{n_F}}$$

$$RMSE = \sqrt{s^2_{pool}} = \sqrt{\frac{(n_M-1)s^2_M + (n_F-1)s^2_F}{n_M+n_F-2}}$$

69

Example: DSST by Sex

```
. ttest dsst, by(male)
Two-sample t test with equal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     367    42.42779    .6680565    12.79812    41.11408    43.7415
       1 |     356    39.64326    .6610097    12.47191    38.34327    40.94325
---------+--------------------------------------------------------------------
combined |     723    41.05671    .4725559    12.70639    40.12896    41.98446
---------+--------------------------------------------------------------------
    diff |            2.784534    .9401746                .9387276    4.630341
------------------------------------------------------------------------------
diff = mean(0) - mean(1)                                      t =   2.9617
Ho: diff = 0                     degrees of freedom =      721

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9984         Pr(|T| > |t|) = 0.0032          Pr(T > t) = 0.0016

. regress dsst male
      Source |       SS       df       MS              Number of obs =     723
-------------+------------------------------           F(  1,   721) =    8.77
       Model |  1401.14463     1  1401.14463           Prob > F      =  0.0032
    Residual |   115167.53   721  159.733052           R-squared     =  0.0120
-------------+------------------------------           Adj R-squared =  0.0106
       Total |  116568.675   722  161.452458           Root MSE      =  12.639

------------------------------------------------------------------------------
        dsst |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   -2.784534   .9401746    -2.96   0.003    -4.630341   -.9387276
       _cons |    42.42779   .6597272    64.31   0.000     41.13258    43.72301
------------------------------------------------------------------------------
```

Inference for the Geometric Mean

Simple Linear Regression on Log Transformed Data

71

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"

72

Interpretation of Parameters

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

Model $$E\left[\log Y_i \,\middle|\, X_i\right] = \beta_0 + \beta_1 \times X_i$$

$X_i = 0$	$E\left[\log Y_i \,\middle|\, X_i = 0\right] = \beta_0$
$X_i = x$	$E\left[\log Y_i \,\middle|\, X_i = x\right] = \beta_0 + \beta_1 \times x$
$X_i = x+1$	$E\left[\log Y_i \,\middle|\, X_i = x+1\right] = \beta_0 + \beta_1 \times x + \beta_1$

73

Interpretation of Parameters

- Restated model as log link for geometric mean

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

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

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

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

74

Interpretation of Parameters

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

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

$X_i = 0$	$GM\left[Y_i \,|X_i = 0\right] = e^{\beta_0}$
$X_i = x$	$GM\left[Y_i \,|X_i = x\right] = e^{\beta_0} \times e^{\beta_1 \times x}$
$X_i = x+1$	$GM\left[Y_i \,|X_i = x+1\right] = e^{\beta_0} \times e^{\beta_1 \times x} \times e^{\beta_1}$

75

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

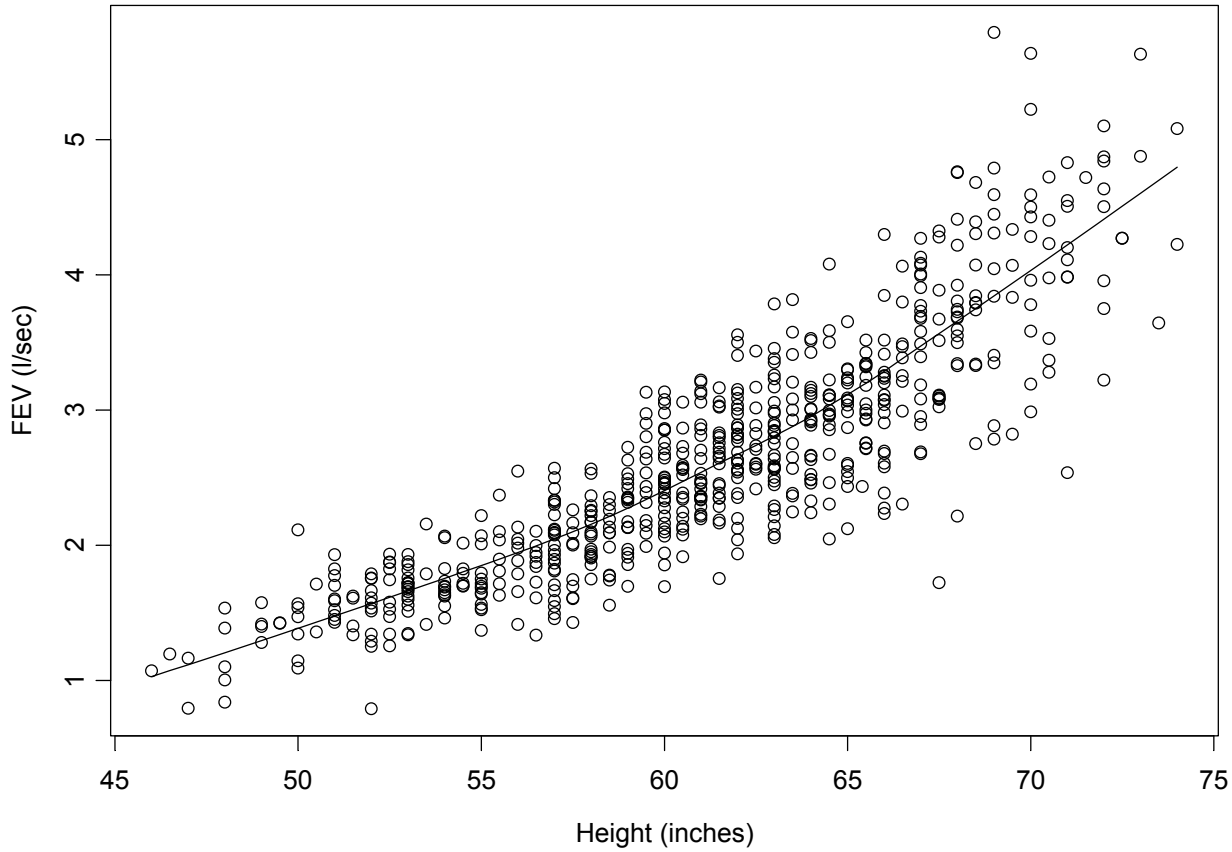
76

Example: Trends in FEV by 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 (L/sec)
- Question: How does FEV differ across height groups

77

FEV versus Height



78

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

79

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

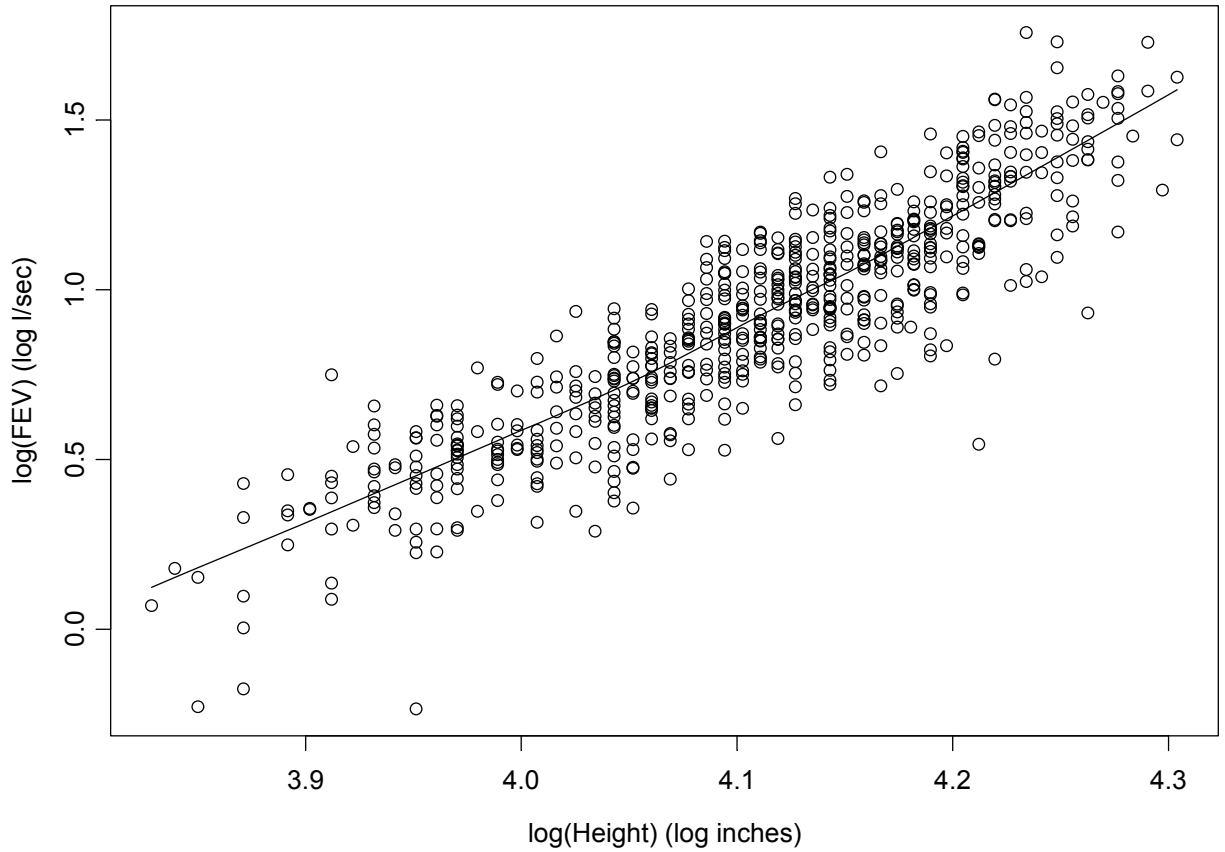
80

Model Geometric Mean

- Science dictates any of the models
- Statistical preference for transformation of response
  - In presence of heteroscedasticity "best linear unbiased estimator" requires weighting observations unequally
    - Okay if linear model truly holds
    - Scientifically unpleasing if linear model does not hold
  - Instead, we may be able to transform to equal variance across groups
    - "Homoscedasticity" tends toward easier and more precise inference when weighting all individuals equally
- Statistical preference for log transformation
  - Easier interpretation: multiplicative model
  - Compare groups using ratios

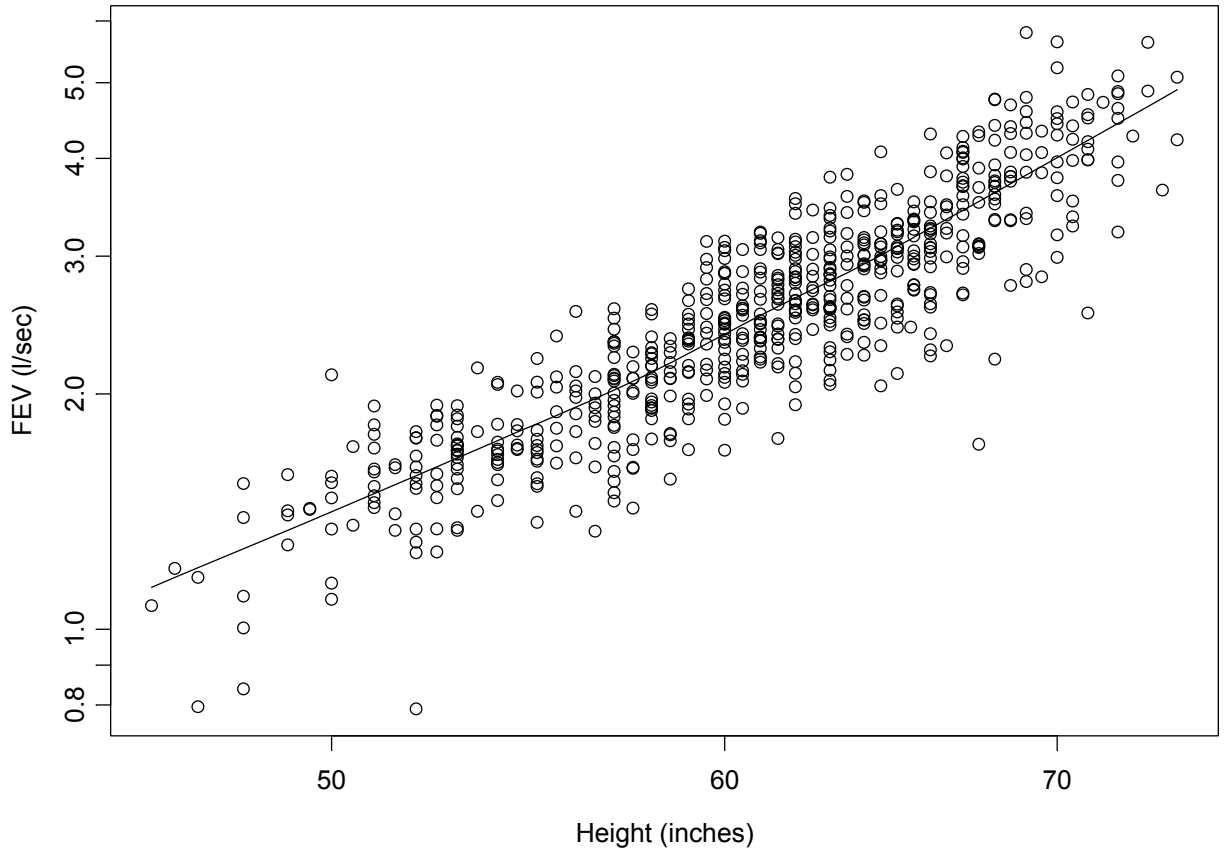
81

log(FEV) versus log(Height)



82

log-log Plot of FEV vs Height



83

Estimation of Regression Model (Stata)

```
. 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
```

84

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

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

85

Interpretation of Stata Output

- Scientific interpretation of the slope

$$\log \mathrm{GM}\left[FEV_i \,\middle|\, loght_i\right] = -11.9 + 3.12 \times loght_i$$

- 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)

86

More Interpretable Estimates (Stata)

- Have Stata
  - exponentiate coefficients and
  - rescale log height to base 1.1
  - (note the Root MSE still refers to the SD of log transformed data)
  - (note that Robust SE have been transformed in complicated way, but CI computed on the log scale so are just transformations)

```
. g loght = log(height) / log(1.1)
. regress logfev loght, robust eform(Geom Mn)

Linear regression                               Number of obs =     654
                                                F(  1,   652) = 2130.18
                                                Prob > F      =  0.0000
                                                R-squared     =  0.7945
                                                Root MSE      =   .1512
------------------------------------------------------------------------------
             |               Robust
      logfev |   Geom Mn   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       loght |   1.346846   .0086893    46.15   0.000     1.329892    1.364017
       _cons |   6.65e-06   1.85e-06   -42.90   0.000     3.85e-06    .0000115
------------------------------------------------------------------------------
```

More Interpretable Estimates (R)

- R regression on geometric mean returns exponentiated coefficients by default
  - rescale log height to base 1.1

```
loght= log(height) / log(1.1)
regress("geom",fev,loght)
regress(fnctl = "geom", y = fev, model = loght)

Coefficients:
             Est NaivSE  RbstSE   e(Est)   e(95%L)  e(95%H)  Fstat df Pr(>F)
Intercept  -11.9  0.256   0.278  6.65e-6  3.853e-6  1.147e-5  1840 1  < 0.00005
loght      0.298 5.93e-3 6.45e-3  1.347     1.330      1.3    2130 1  < 0.00005

Residual standard error: 0.1512 on 652 degrees of freedom
Multiple R-squared:  0.7945,      Adjusted R-squared:  0.7941
F-statistic:  2520 on 1 and 652 DF,  p-value: < 2.2e-16
```

88

Example: Interpretation

"From linear regression analysis on log transformed FEV using Huber-White estimates of the standard error, we estimate that when comparing two groups of children differing in height by 10%, the geometric mean FEV is 34.7% higher in the taller population. A 95% CI suggests that this observation is not unusual if the true relationship between geometric means were such that the taller group's geometric mean FEV were between 33.0% and 36.4% higher than that in the shorter group. Because the two sided P value is P < .0005, we reject the null hypothesis that there is no linear trend in the average DSST across height groups."

89

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)

90

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

91

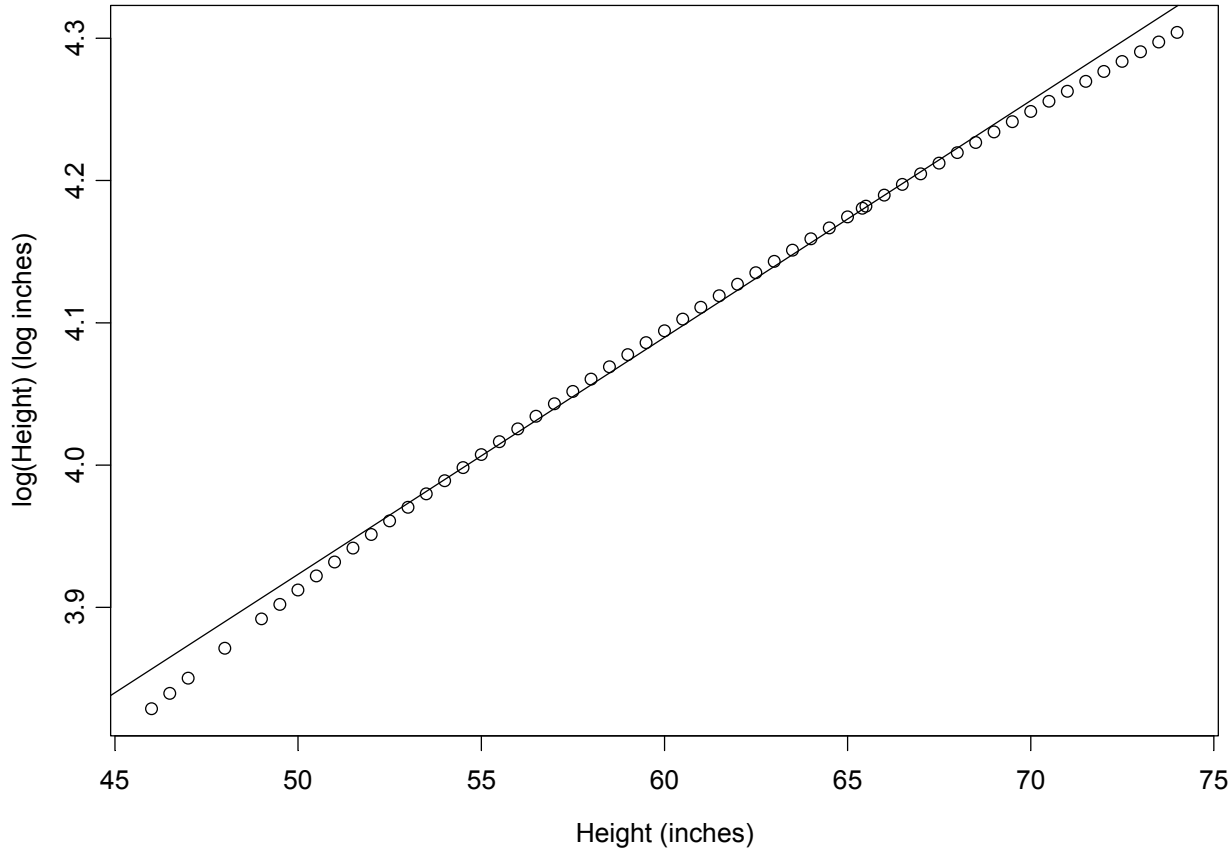
Sometimes Does Not Matter

- 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

92

log(Height) versus Height



93

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

94

Estimation of Regression Model

- In Stata, the regress command will not backtransform for you by default

```
. regress logfev height, robust eform(Geom Mn)

Linear regression                               Number of obs =     654
                                                F(  1,   652) = 2155.08
                                                Prob > F      =  0.0000
                                                R-squared     =  0.7956
                                                Root MSE      =  .15078
------------------------------------------------------------------------------
             |               Robust
       logfev |    Geom Mn   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   1.053501   .0011828    46.42   0.000     1.051181    1.055826
       _cons |   .1031767    .007073   -33.13   0.000     .0901824    .1180434
------------------------------------------------------------------------------
```

95

Example: Interpretation

"From linear regression analysis on log transformed FEV using Huber-White estimates of the standard error, we estimate that for every 1 inch difference in height between two groups of children, the geometric mean FEV is 5.35% higher in the taller population. A 95% CI suggests that this observation is not unusual if the true relationship between geometric means were such that the taller group's geometric mean FEV were between 5.12% and 5.58% higher for each 1 inch difference in height. Because the two sided P value is P < .0005, we reject the null hypothesis that there is no linear trend in the average DSST across height groups."

96

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

97

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

98

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

99

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

100

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

101