

Applied Regression Analysis

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

Session 3

1

Applied Regression Analysis

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

Part 2: Simple Linear Regression

2

Lecture Outline

- Topics:
 - Simple linear regression
 - Descriptive statistics
 - Interpretation of model
 - Inference
 - Relationship to t tests
 - Relationship to correlation
 - Transformation of variables

3

Homework #1

4

Sample A

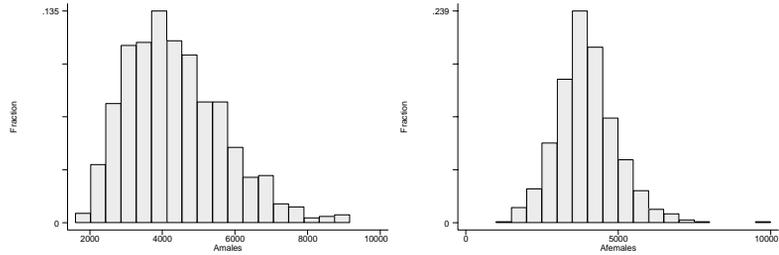


Figure 1: Histograms of monthly salaries for males (left) and females (right) at University A.

	Mean	Std. Dev.	Min	25th %ile	Median	75th %ile	Max	% > 50k	% > 75k
SAMPLE A									
Males	4366	1332	1600	3361	4174	5160	9154	50.6	9.1
Females	3968	987	1418	3313	3933	4535	9556	37.7	2.4

Sample B

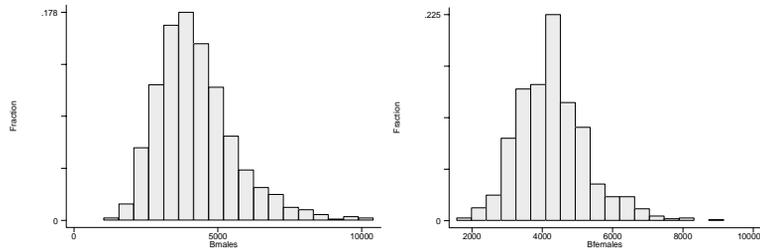


Figure 2: Histograms of monthly salaries for males (left) and females (right) at University B.

	Mean	Std. Dev.	Min	25th %ile	Median	75th %ile	Max	% > 50k	% > 75k
SAMPLE B									
Males	4247	1348	1392	3296	4048	4951	10411	45.6	8.1
Females	4271	993	1563	3607	4220	4812	9025	52.1	4.7

Sample C

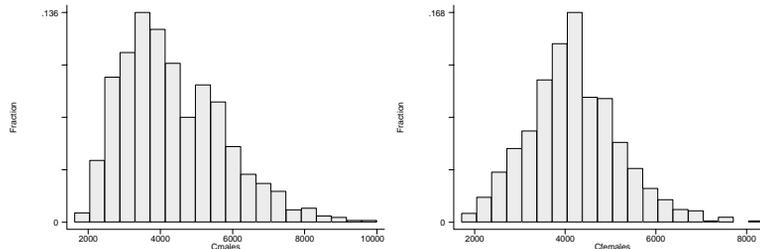


Figure 3: Histograms of monthly salaries for males (left) and females (right) at University C.

	Mean	Std. Dev.	Min	25th %ile	Median	75th %ile	Max	% > 50k	% > 75k
SAMPLE C									
Males	4362	1399	1623	3298	4105	5242	9833	47.8	10.2
Females	4183	1004	1711	3559	4112	4806	8381	48.1	3.1

- *ptidno*= patient identification number uniquely identifying each patient
- *date*= date of visit in MMDDYY format
- *age*= age of patient in years
- *male*= sex of patient in coded format (0= female, 1= male)
- *clinic*= code for clinic visited (0= family practice, 1= pediatrics, 2= medicine, 3= surgery, 4= obstetrics, 5= emergency room)
- *temp*= patient's temperature in degrees Celsius at that visit
- *sbp*= patient's systolic blood pressure in mm Hg at that visit
- *radimag*= indicator that radiologic imaging was used at that visit (0= no, 1= yes)

.....

The following table presents descriptive statistics for the dataset.

	n	msg	mean	std dev	min	25%-ile	median	75%-ile	maximum
ptid	600	0	995	573	1	507	998	1482	2000
age	600	0	60.27	6.79	38.59	55.61	60.03	64.99	80.76
male	600	0	0.68	0.47	0.00	0.00	1.00	1.00	1.00
chol	600	0	222.56	33.76	159.00	208.00	216.00	236.50	395.00
sbp	600	0	134.58	30.92	90.04	107.56	129.12	160.47	199.37
mi	600	0	0.17	0.37	0.00	0.00	0.00	0.00	1.00
marij	600	0	0.23	0.42	0.00	0.00	0.00	0.00	1.00

9

Descriptive Statistics Useful with Linear Regression

.....

10

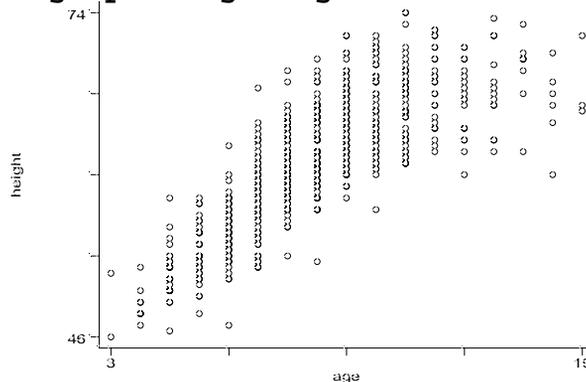
Scatterplots

- A graph of Y versus X
 - Most useful for two continuous variables
 - If variables are discretely measured, jittering can be helpful
 - “jittering”: adding a little noise to the data to break ties
 - I tend to try to jitter to allow visualization of all points, but still try to keep discrete levels separate: use a spread of about 40% the separation between categories

11

Example: Scatterplot (Unjittered)

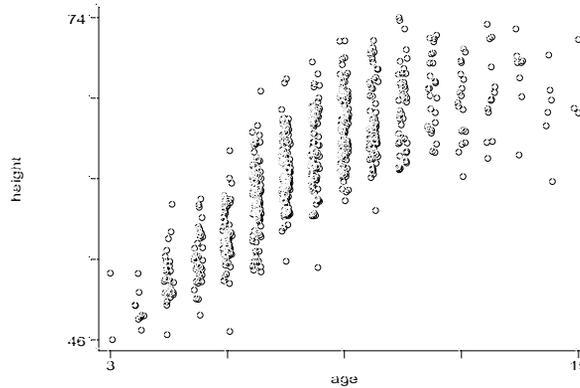
- FEV Data: Scatterplot of Height versus Age
 - Stata: `graph height age`



12

Example: Scatterplot (Jittered)

- FEV Data: Jittered scatterplot of Height versus Age
 - Stata: `graph height age, jitter(1)`



13

Example: Scatterplot (Jittered)

- Observations:
 - Tendency toward increased height for older ages
 - First order trend is upward
 - Hint of curvilinear relationship
 - Suggestion of increasing variability with increased height
 - Must be careful when judging variability from range
 - Need to compare range of equal numbers of data in area with equal slopes

14

Superimposing Curves on Scatterplots.....

- It is often helpful to place curves over a scatterplot to help see trends in the data
 - Theoretical relationship
 - If theory prescribes a supposed relationship
 - Least squares line
 - Best fitting line (but it forces it to be a line)
 - Smooths (e.g., lowess)
 - Curve that represents smooth approximation to the data

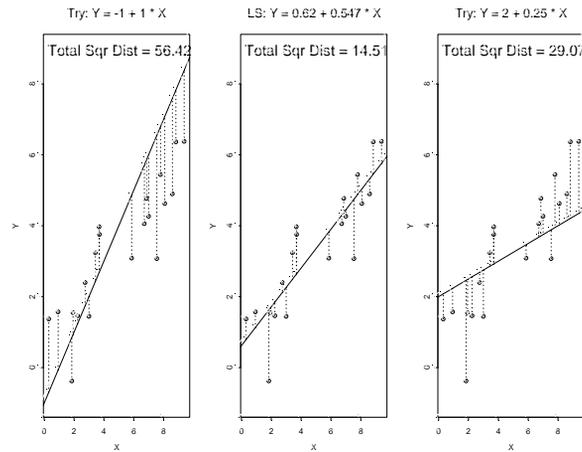
15

Least Squares Estimation of a 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
 - Computers find such estimates easily

16

Least Squares Estimation of a Line



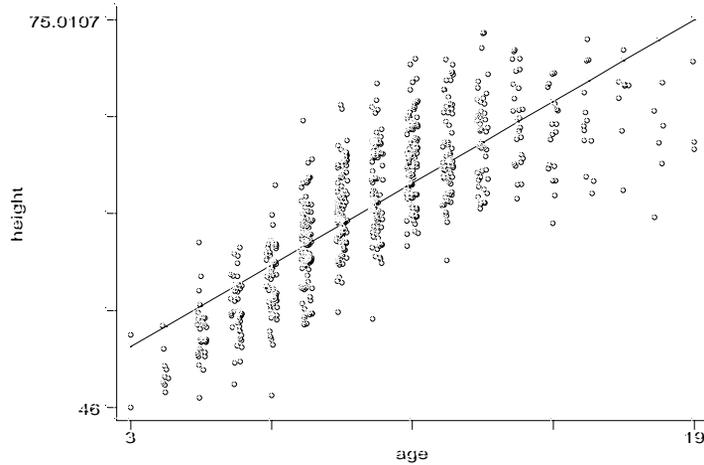
17

Example: Height versus Age

- Stata commands to superimpose a least squares line on a scatter plot
 - "regress height age"
 - Computes Least Squares line
 - "predict fit"
 - Stores estimated values in variable *fit*
 - "sort age"
 - Line is best drawn with sorted data
 - "graph height fit age, s(oi) c(.1)"
 - Scatterplot of height vs age: symbol o; no line
 - Scatterplot of fit vs age: no symbol; connected

18

Example: Height versus Age



19

Example: Height versus Age

- Observations
 - Clearly increasing trend in data
 - Our eye tends to like to detect lines, so it takes careful inspection to decide a line is not the best fit
 - Note that at lowest ages and highest ages most data tend to be on one side of line rather than symmetric about line

20

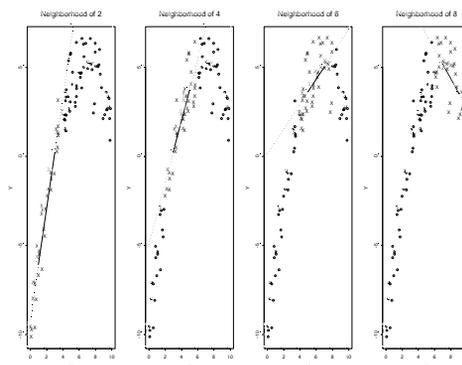
Locally Weighted Scatterplot Smoother.....

- Lowess: A smoother to find a smooth curve approximating relationship in the data
 - For every value of X, fits straight lines in a neighborhood of that value
 - “Bandwidth” is width of window defining neighborhood
 - Weights closer data more heavily when finding the line
 - Combines the estimates from different regions to form a smooth curve

21

Locally Weighted Scatterplot Smoother.....

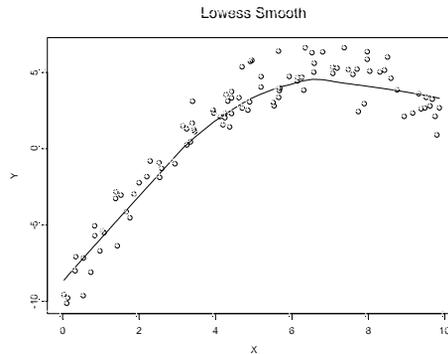
- Conceptual algorithm
 - Least squares lines fit in neighborhoods



22

Locally Weighted Scatterplot Smoother.....

- Conceptual algorithm
 - Lowess combines locally fit least squares lines



23

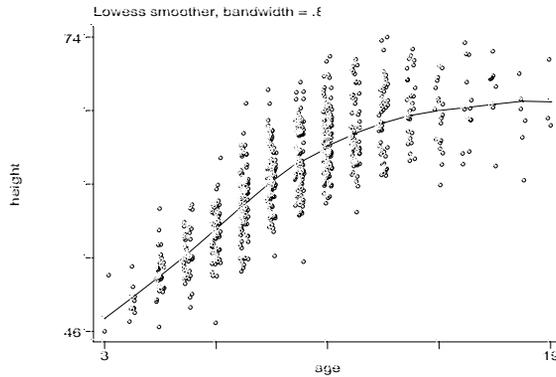
Lowess Smooths: Stata Commands.....

- “Kernel smoothing”
 - “`ksm yvar xvar, lowess`”
 - Options:
 - Plot is generated unless option `nograph` is used
 - Control degree of smoothing with `bwidth(#)`
 - # (between 0 and 1) is width of window used in locally estimated lines
 - Save smoothed estimates with `gen(smvar)`
 - `smvar` is a variable name of your choice
 - Useful when plotting stratified smooths

24

Example: Height versus Age

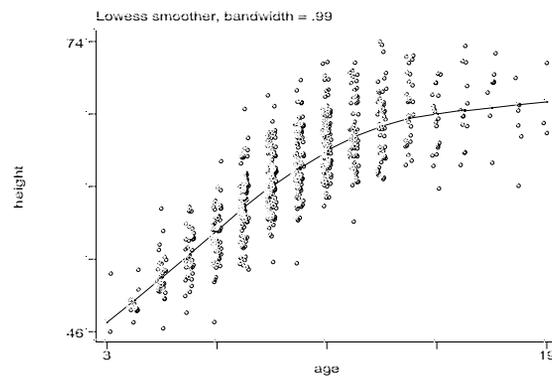
- ksm height age, lowess jitter(1)
 - Default bandwidth is 0.8



25

Example: Height versus Age

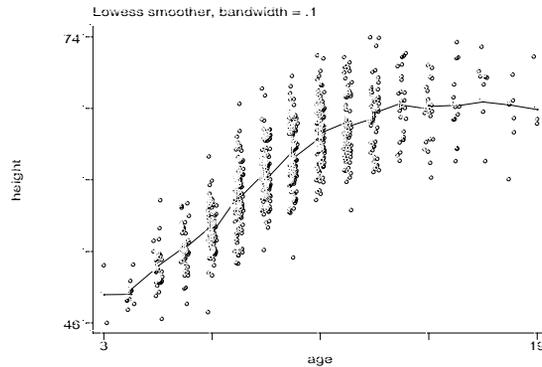
- ksm height age, lowess jitter(1) bwidth(0.99)
 - Wider bandwidth produces more smoothing



26

Example: Height versus Age

- ksm height age, lowess jitter(1) bwidth(0.1)
 - Narrower bandwidth produces less smoothing



27

Example: Height versus Age

- Observations:
 - Lowess smooth shows that height tends to increase pretty linearly with age up until about age 11 or 12
 - Height levels off in late teens with little change in mean height

28

Other Smoothers

.....

- Many different methods of smoothing data have been proposed
 - Lowess is often criticized due to the way it can accentuate data near the end of its range
 - One should not make too much of the way the estimate curve wiggles at the extremes of the data
 - For my purposes, almost any smoother will do
 - I just want to have something that is not forced to be a line, and something that I did not draw
 - I can be just as biased as anyone