

4. ASSOCIATION AND CORRELATION: FPP Ch 7,8,9  
4.1 THE SCATTER PLOT (FPP: SCATTER DIAGRAM)

Blank page for your notes

- Often we are interested in relationships among quantitative variables:

Weight and cholesterol level of individuals

Gas economy (mpg) and capacity (cu. feet) in cars

Heights of fathers and sons

Boat registrations and manatee deaths in Florida

Per capita income and education levels across countries.

- If we have pairs of quantitative observations on two variables on the same subject, we can draw a scatter plot.

- One value goes on the "x-axis" (horizontal axis) and the other on the "y-axis" (vertical axis).

- If we are considering cause-and-effect, we put the explanatory variable (independent variable) on the horizontal axis, and the response variable (dependent variable) on the vertical axis.

- But ASSOCIATION IS NOT CAUSE, and there may be no good reason to choose which variable goes on which axis.

Example: Heights of fathers and sons

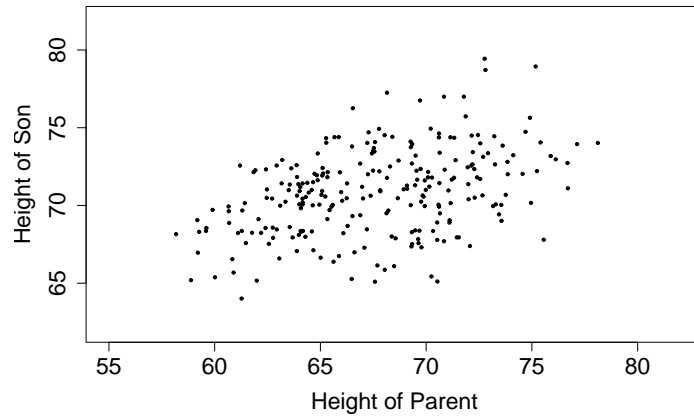
Example: Heights of husbands and wives

#### 4.2 ADDING SYMBOLS FOR ADDITIONAL INFORMATION

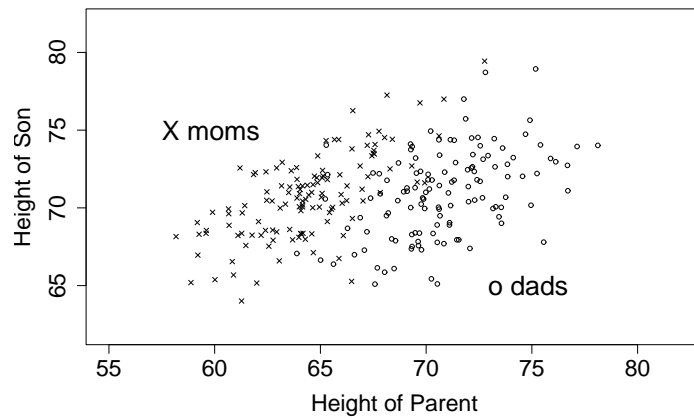
Blank page for your notes

- 250 recent retirees of a company were asked about their height, and that of their eldest (adult) son.

A very weak association .... (??)



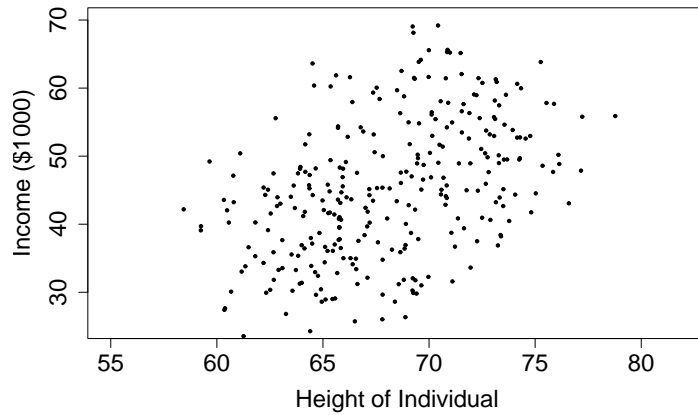
- .. becomes a strong association (as expected)



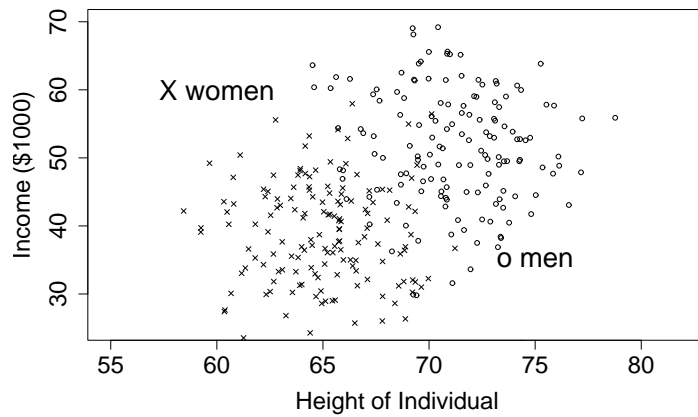
## THE REVERSE EFFECT

Blank page for your notes

- The same retirees were asked about their income.  
an apparent strong association:

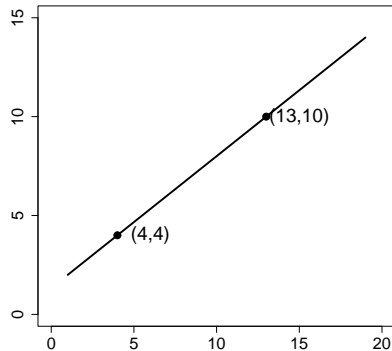


- ... becomes no association at all !!

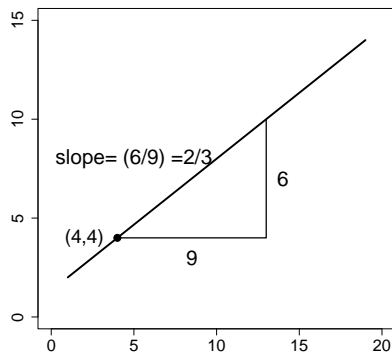


### 4.3 PLOTTING LINES (FPP Ch 7)

- Two points on the (x,y)-plot determine a line (joining them).



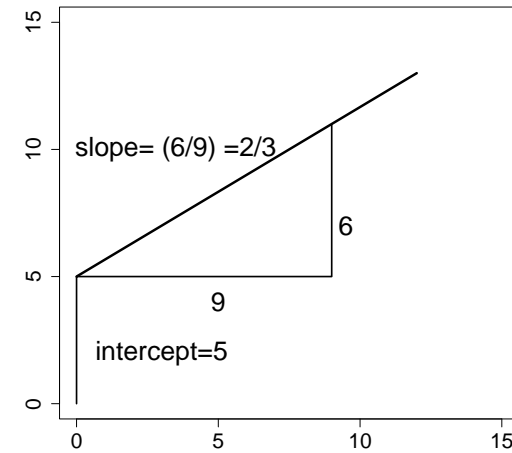
- One point and the slope determine a line.



- Slope and intercept determine a line:

$$y = \text{intercept} + \text{slope} \times x$$

Example:  $y = 5 + (2/3) \times x$



- Points above or below a line?

If  $y$  is bigger than  $(\text{intercept} + \text{slope} \times x)$  then point is above.

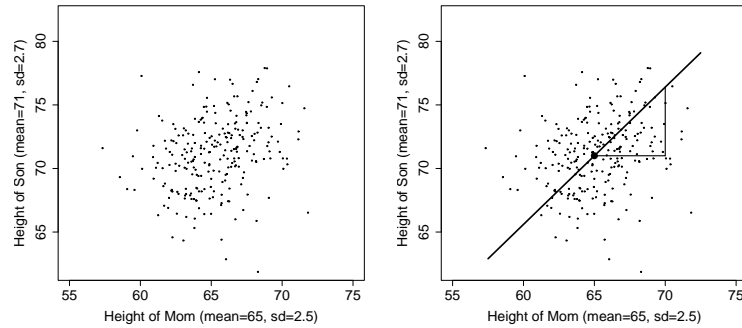
If  $y$  is smaller than  $(\text{intercept} + \text{slope} \times x)$  then point is below.

- The point  $(x=6, y=11)$  is above.  
The point  $(x=9, y=7)$  is below.

#### 4.4 THE SD LINE

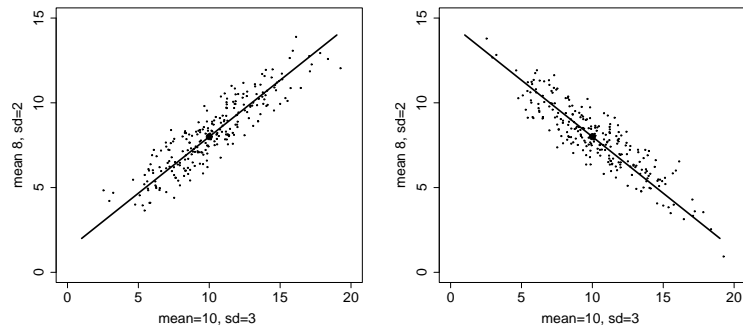
Blank page for your notes

- The SD line goes through the point of averages.
- The slope of the SD line is  
(SD of y) divided by (SD of x).  
For a negative association, it goes the other way.
- Sometimes it is easier to see the pattern if we draw the SD line.

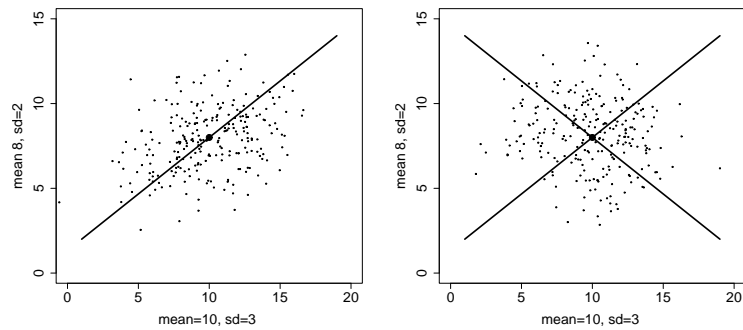


- For a strong association, points are clustered round the SD line.

Blank page for your notes



- For a weak association, or none, they are not.



#### 4.5 STANDARDIZED VALUES

Blank page for your notes

- Recall, we are interested in how many SD an observation is from the average.
- The standardized value of an observation is (value - mean) divided by SD.
- Standardized values have mean 0 and SD equal to 1.
- Example:

The data			y-value deviation from mean	y-value squared dev	Standardized	
x-value	y-value				x-value	y-value
3.0	3.0		0.0	0.0	-0.92	0.00
4.2	2.0		-1.0	1.0	-0.07	-0.71
5.8	4.0		1.0	1.0	1.06	0.71
6.0	5.0		2.0	4.0	1.20	1.41
2.5	1.0		-2.0	4.0	-1.27	-1.41
sum	21.5	15.0	0.0	10.0	0.0	0.0
avg	4.3	3.0		2.0		

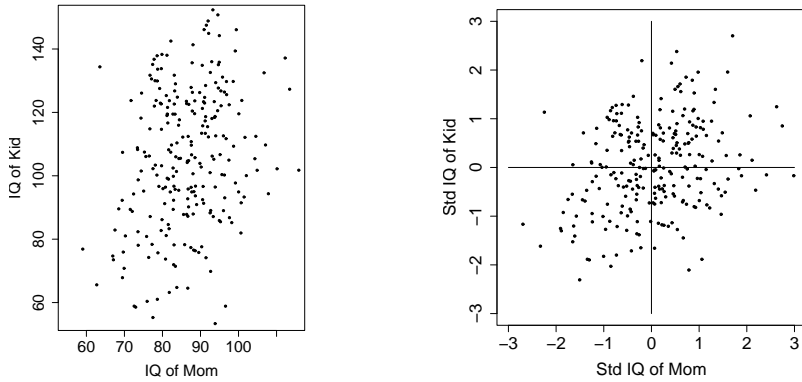
For the x-values, the mean is 4.3 and the SD is 1.42 from page 3.4. The standardized values are then as shown.

For the y-values, the mean is 3.0 and the SD is  $\sqrt{2.0}$  or 1.41. The standardized values are then as shown.

#### 4.6 USING SD VALUES IN SCATTERPLOTS

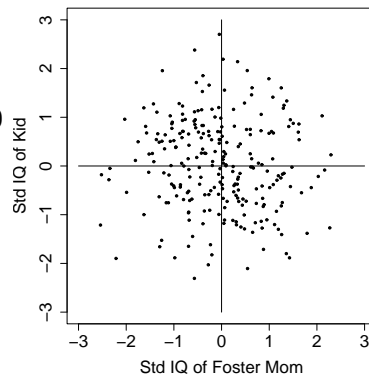
Blank page for your notes

- Can be useful to see strength of association more easily  
IQ test results of kids 4-7 and (biological) mom IQ.



- Can be misleading !!  
IQ test results of same kids 4-7 and foster mom IQ.

- $\text{corr}(\text{kid}, \text{bio-mom}) = 0.3$   
(as for non-fostered)
- $\text{corr}(\text{kid}, \text{foster-mom}) = 0.0$   
But  $\text{mean}(\text{bio-mom}) = 86$ ,  
 $\text{mean}(\text{kids}) = 106$ ,  
 $\text{mean}(\text{foster-mom}) = 110$





#### 4.7 THE CORRELATION COEFFICIENT

Blank page for your notes

- Computing the correlation coefficient:

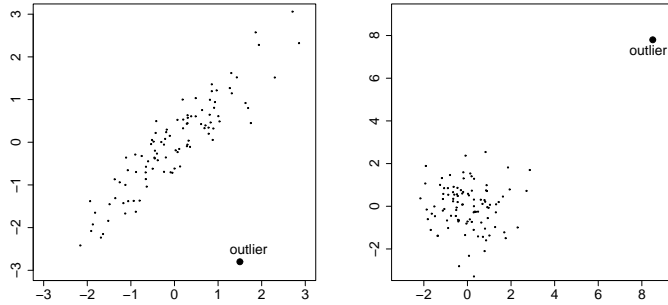
original values		standardized values		product	
x-value	y-value	x-value	y-value		
3.0	3.0	-0.92	0.00	0.00	
4.2	2.0	-0.07	-0.71	0.49	
5.8	4.0	1.06	0.71	0.75	
6.0	5.0	1.20	1.41	1.70	
2.5	1.0	-1.27	-1.41	1.79	
sum	21.5	15.0	0.0	0.0	4.29
avg	4.3	3.0	0.0	0.0	0.86

- Compute the standardized values.
- Multiply the corresponding standardized x- and y-values.
- Take the average of these products.
- FINALLY, the answer (if I did it right) is 0.86
- The correlation has no units. It is a number between -1 and +1.
- The correlation is unchanged by
  - interchanging the variables
  - adding the same number to all values of 1 variable
  - multiplying all the values of one variable by the same positive number.

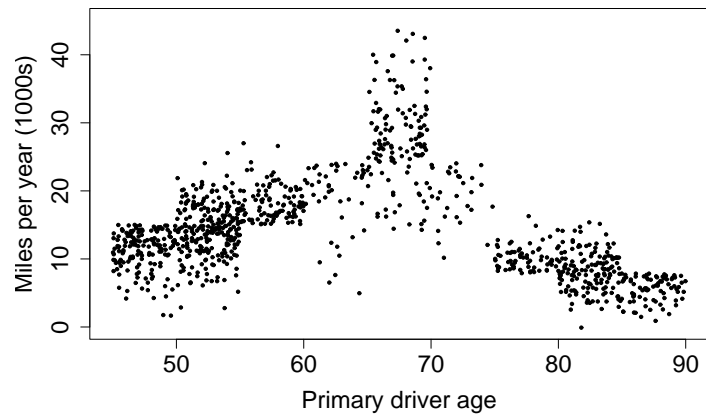
## 4.8 OUTLIERS AND NON-LINEAR ASSOCIATION

Blank page for your notes

- Examples:



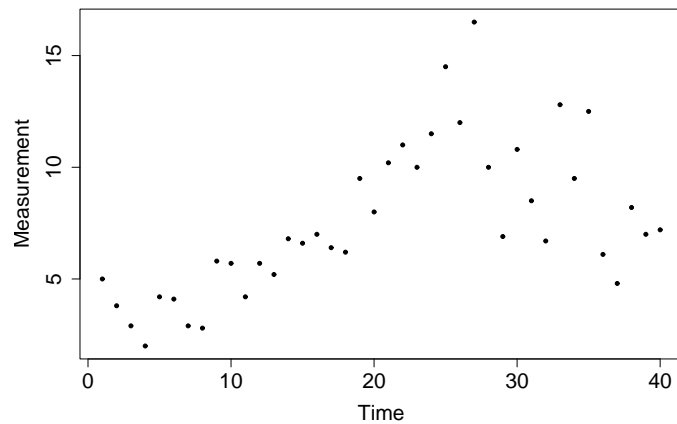
- Again the cars (1000 mpy) and primary driver age:



#### 4.9 HOMEOPATHY STUDY 7 THE ASSOCIATION WITH TIME

Blank page for your notes

- Recall there was apparently a big differences between the first half of the Lab1 values and the second half.
- In detail here are the 40 values, plotted against the order:



- Message of the last sets of scatterplots:  
Always look at the scatterplot: do not rely on the numerical correlations.

#### 4.10 ASSOCIATION IS NOT CAUSE

- In an observational study, an association, even a consistent one, is not evidence of cause.
- There may be a confounding factor:
  - Ice-cream sales and water-sports injuries.
  - Manatee deaths and boat registrations in Florida (??).
- There may be a time trend or effect:
  - Email use and dental cavities.
  - Monthly rainfall and price of tomatoes.
- In a small sample:
  - results may be due to chance.
- In a controlled experiment:
  - we can control one variable, and observe the other.

Blank page for your notes