Lecture 13: Introduction to Correlation

• 13. 1 Overview

- Fundamental definitions of covariance, correlation, and regression

- Importance as measures of association

- Mathematical and statistical models

- Algebraic notation

- Questions

    - What is the strength of the relationship between two variables?

    - What is the shape of the relationship?

    - How can one make predictions from one variable to another?

• 13. 2 Correlation -- Introduction

- Until now our independent variables have been <u>discrete</u> (nominal)

    e.g. *Experimental* vs. *Control*

- But the IV can also be continuous (interval or ratio)

    e.g., *Height* could be related to *weight,* but we don't have "height groups"

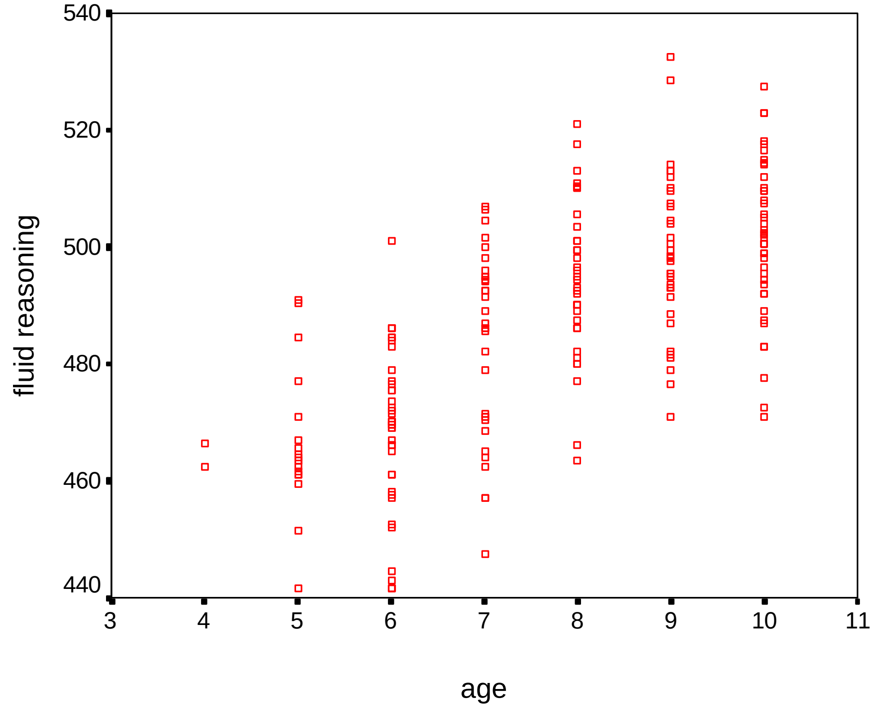
- We want to know the degree (direction) and extent (magnitude) of linear relations – We use the Pearson product-moment correlation coefficient, denoted by "*r*"

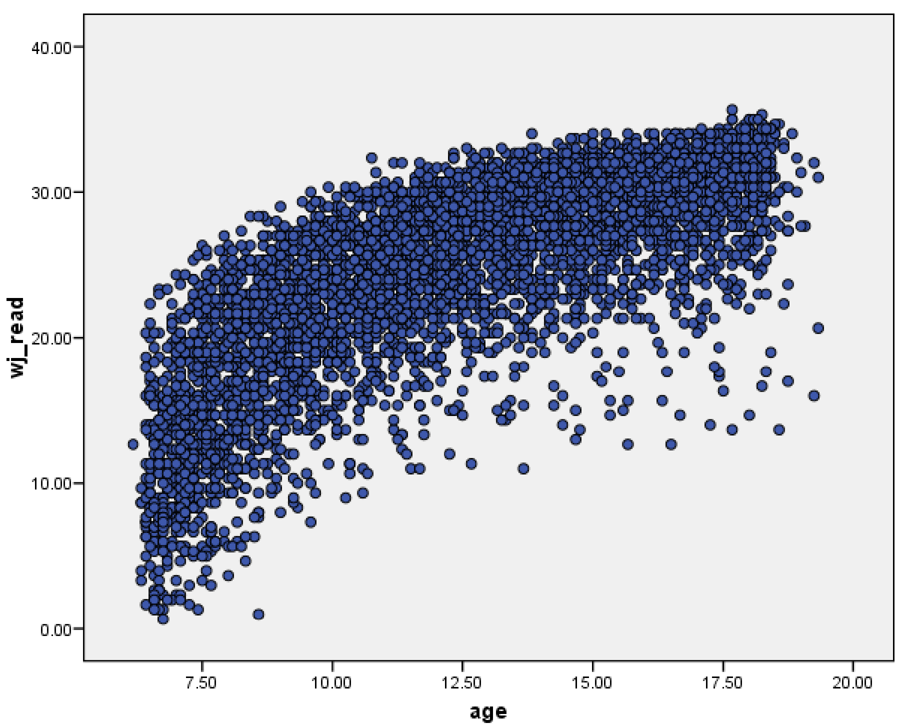
$$r = \frac{\sum (Z_x)(Z_y)}{n-1}$$

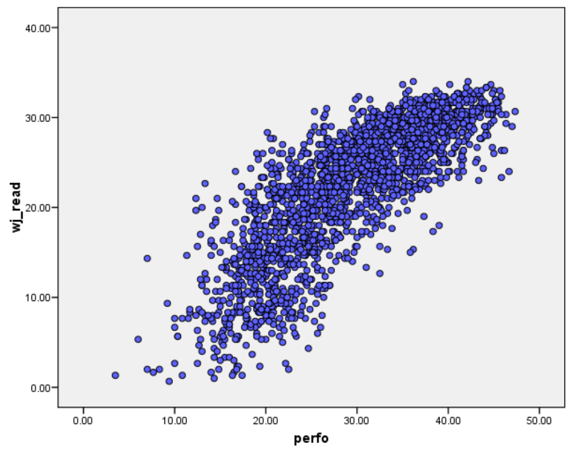
- A correlation coefficient describes the degree of linear relationship between two variables

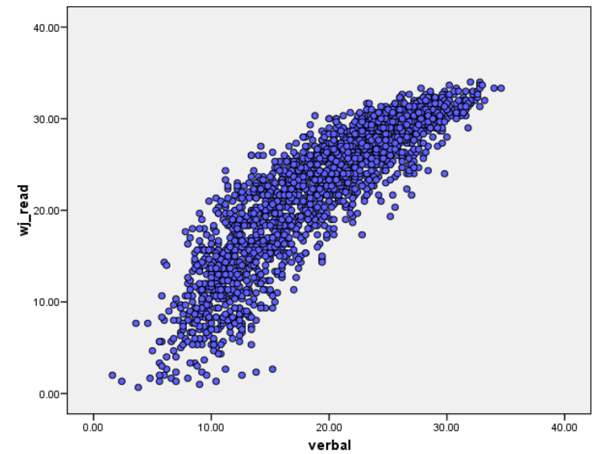
- **13. 3 Correlation -- Illustration**

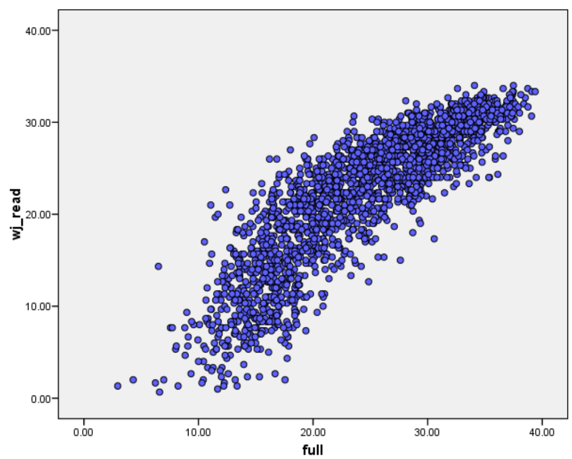
- Some examples of linear relationships (via scatterplot)

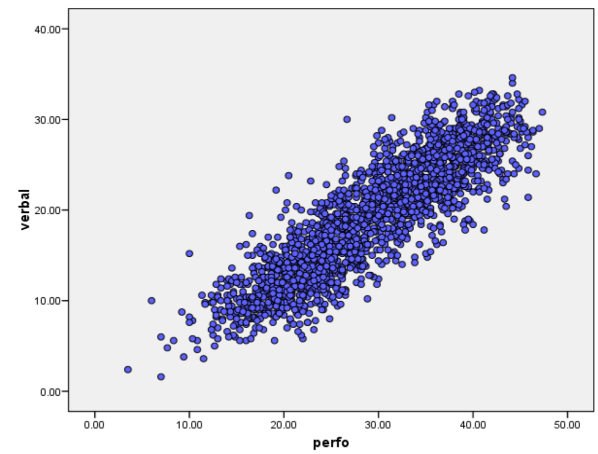


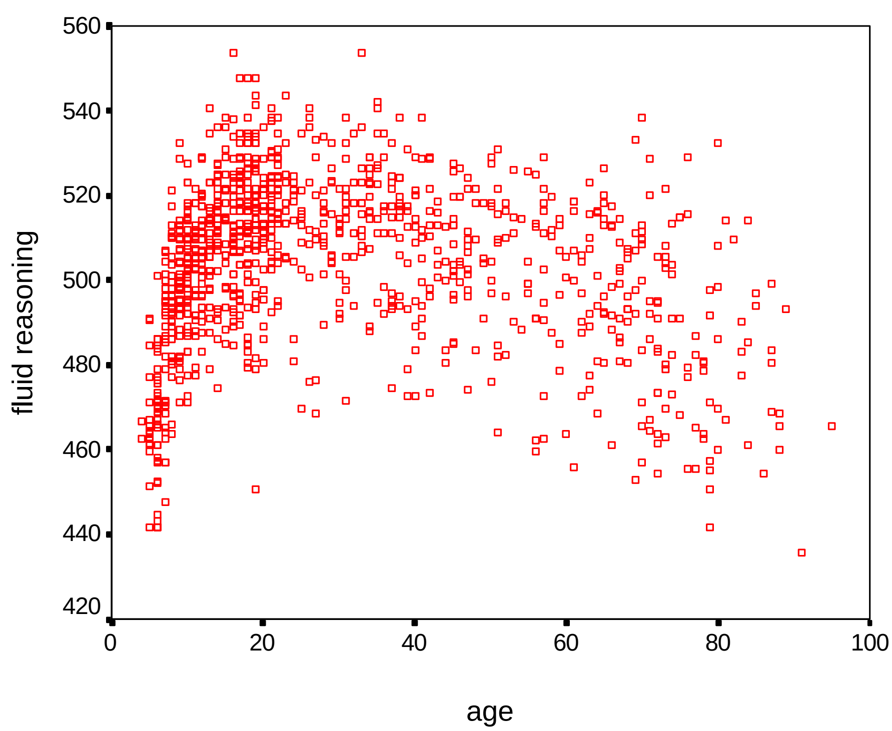


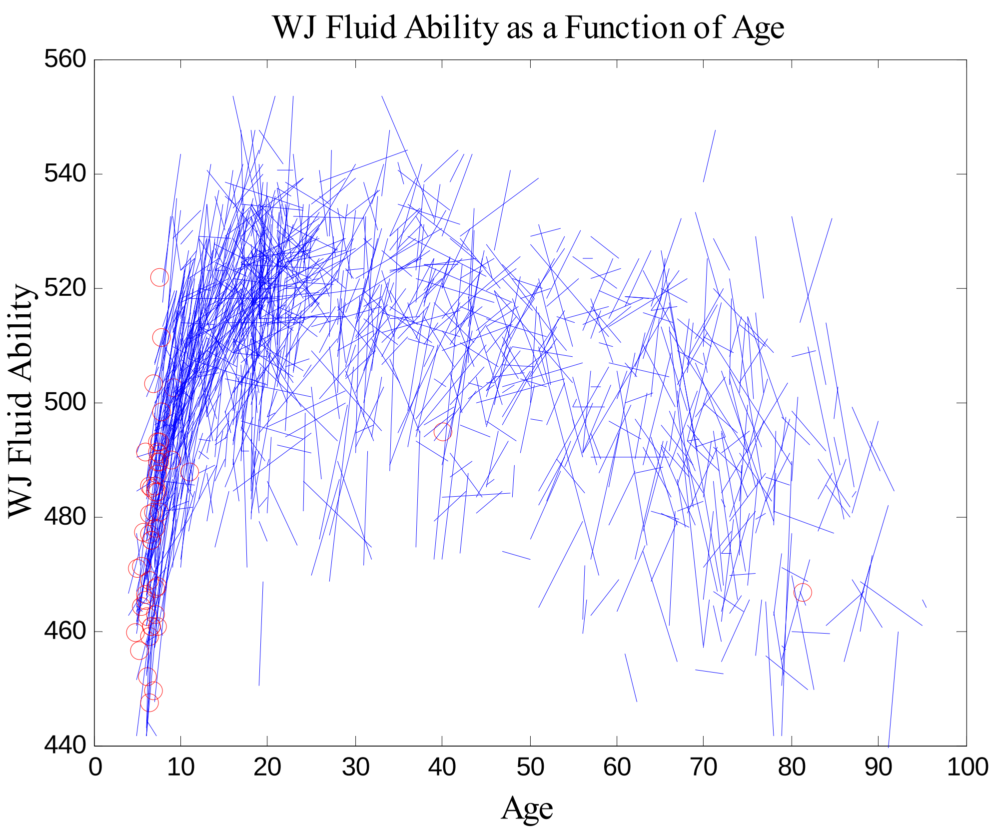

wj_read
perfo

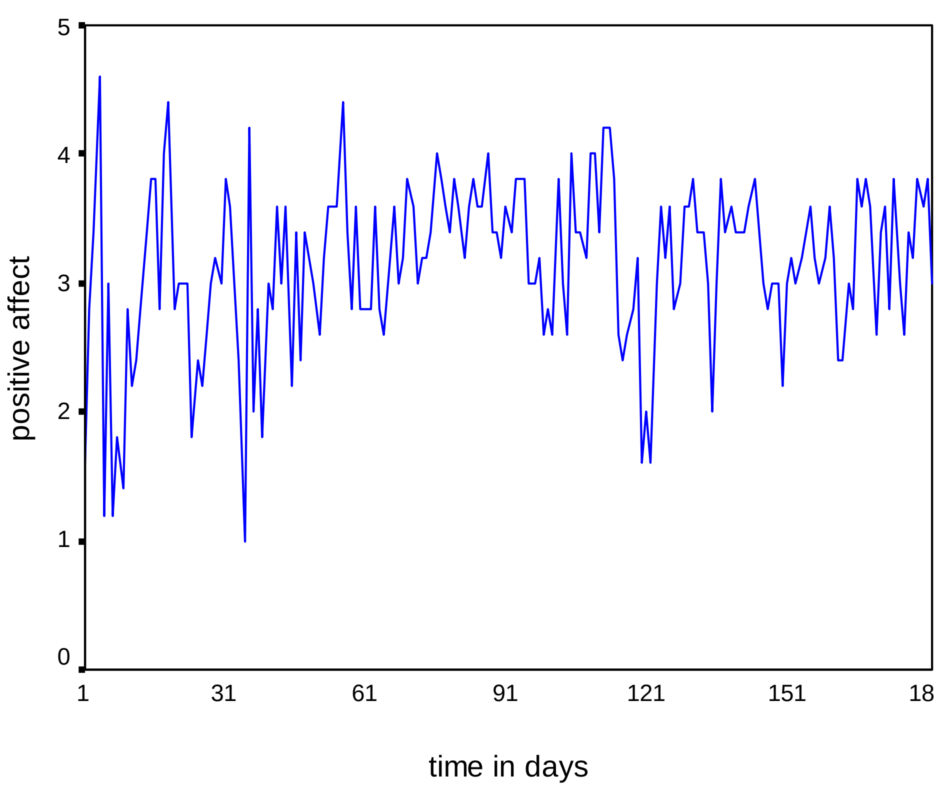

wj_read
verbal


wj_read
full


verbal
perfo


fluid reasoning
age


WJ Fluid Ability as a Function of Age
WJ Fluid Ability
560
540
520
500
480
460
440
0
10
20
30
40
50
60
70
80
90
100
Age


positive affect
5
4
3
2
1
0
1
31
61
91
121
151
18
time in days

• 13. 4 Correlation -- Interpretation

- A correlation is the degree of linear association between two variables

- A correlation is the degree to which the data points cluster around a regression line, or line of best fit

- A correlation is the regression slope if both x and y are rescaled to have variances equal to 1.0

If we rescale x to z-scores with: $$Z_x = \frac{x - \bar{x}}{s_x}$$

and rescale y to z-scores with: $$Z_y = \frac{y - \bar{y}}{s_y}$$

then regress z_y onto z_x, the slope will be r.

- A correlation is the square root of the proportion of variance in y that is "explained" by x, and vice versa

  - Equivalently, r^2 is the proportion of variance in y explained by x, and vice versa

- A correlation r is the sample estimate of the population correlation, ρ (rho)

- Correlations can be positive or negative

- Correlations range between –1.0 and +1.0, inclusive

  $r = 0$ means x and y are not linearly related

- Correlation does not imply causation (unless x is something we manipulate experimentally)

• 13. 5 Covariance

- <u>Covariance</u> – an unstandardized measure of the relationship between two variables

- A correlation is a “standardized” covariance – the covariance between two variables whose scales have been altered so that their variances are 1.0

- Definitional formula (the average of the cross-products of the data)

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

- Example

$$\mathbf{A} = \begin{bmatrix} x & y \\ 1 & 2 \\ 2 & 8 \\ 3 & 6 \\ 4 & 4 \\ 5 & 10 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x-\bar{x} & y-\bar{y} \\ -2 & -4 \\ -1 & 2 \\ 0 & 0 \\ 1 & -2 \\ 2 & 4 \end{bmatrix}$$

$$\mathbf{C_x} = \frac{1}{N-1}\mathbf{X'X} = \frac{1}{4}\begin{bmatrix} -2 & -1 & 0 & 1 & 2 \\ -4 & 2 & 0 & -2 & 4 \end{bmatrix}\begin{bmatrix} -2 & -4 \\ -1 & 2 \\ 0 & 0 \\ 1 & -2 \\ 2 & 4 \end{bmatrix}$$

$$= \frac{1}{4}\begin{bmatrix} 10 & 12 \\ 12 & 40 \end{bmatrix} = \begin{bmatrix} 2.5 & 3 \\ 3 & 10 \end{bmatrix} = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix}$$

- No index of strength of association – not a descriptive idea of the size of the association

- The covariance depends on the scale of the variables

- The solution is to put both variables in the same metric…standardized scores

• 13. 6 From Covariance to Correlation

- Standardize *X* and *Y* first and then get their covariance

$$\mathrm{cov}(Z_x, Z_y) = \frac{\sum (Z_x - \overline{Z}_x)(Z_y - \overline{Z}_y)}{n-1}; \quad r = \frac{\sum (Z_x)(Z_y)}{n-1}$$

$$Z_x = \frac{\sum (x - \bar{x})}{s_x}; \; Z_y = \frac{\sum (y - \bar{y})}{s_y};$$

$$r = \frac{C_{xy}}{S_x S_y}$$

r = Pearson product moment correlation coefficient

- **13. 7 Example**

- What is the relationship between self-esteem and number of friends?

X (SE)	Y (Friends)	Z_x	Z_y	$(Z_x)(Z_y)$
1	2	-1.44	-1.16	1.64
3	4	0	-0.39	0
4	6	0.71	0.39	0.27
4	8	0.71	1.16	0.82
$\overline{X}$ = 3	$\overline{Y}$ = 5			Σ = 2.73
S_x = 2.58	S_y = 1.41			

$$r = \frac{\sum (Z_x)(Z_y)}{n-1}; \quad r = \frac{2.73}{3} = .91$$

- There is a high association between the two variables (but no directionality)

- Knowing a person's SE tells us a lot about the number of friends they are likely to have

- r^2 = .84; about 84% of the variance in number of friends is explained by SE alone

- **13. 8 From Covariance to Correlation (In Matrix Form)**

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array}\left[\begin{array}{ccc} x_1 & x_2 & x_3 \\ 9 & 9 & 12 \\ 9 & 16 & 10 \\ 12 & 10 & 25 \end{array}\right] =$$

$$\begin{bmatrix} \frac{1}{s_{x1}} & 0 & 0 \\ 0 & \frac{1}{s_{x2}} & 0 \\ 0 & 0 & \frac{1}{s_{x3}} \end{bmatrix} \begin{bmatrix} 9 & 9 & 12 \\ 9 & 16 & 10 \\ 12 & 10 & 25 \end{bmatrix} \begin{bmatrix} \frac{1}{s_{x1}} & 0 & 0 \\ 0 & \frac{1}{s_{x2}} & 0 \\ 0 & 0 & \frac{1}{s_{x3}} \end{bmatrix} =$$

$$\begin{bmatrix} \frac{9}{s_{x1}s_{x1}} & \frac{9}{s_{x1}s_{x2}} & \frac{12}{s_{x1}s_{x3}} \\ \frac{9}{s_{x2}s_{x1}} & \frac{16}{s_{x2}s_{x2}} & \frac{10}{s_{x2}s_{x3}} \\ \frac{12}{s_{x3}s_{x1}} & \frac{10}{s_{x3}s_{x2}} & \frac{10}{s_{x3}s_{x3}} \end{bmatrix} = \begin{bmatrix} 1 & sym & sym \\ \frac{9}{12} & 1 & sym \\ \frac{12}{15} & \frac{10}{20} & 1 \end{bmatrix} = \begin{bmatrix} 1 & sym & sym \\ .75 & 1 & sym \\ .80 & .50 & 1 \end{bmatrix}$$

- **13. 9 Significance Testing**

- But maybe there is really no relationship between X and Y in the population

- In other words, perhaps $\rho = 0$, even though $r = .91$

- Is $r = .91$ significantly different from zero?

H_0: $\rho = 0$
H_1: $\rho \neq 0$

$df = N - 2$

- We can use a t test.
    - Testing H_0: $\rho = 0$ is similar to performing a one-sample t-test, where we compare the observed correlation to a fixed value of 0

$$t = \frac{r}{SE_r} \qquad SE_r = \sqrt{\frac{1 - r^2}{N - 2}}$$

$$t = \frac{.91}{\sqrt{\frac{1-(.91)^2}{4-2}}} = \frac{.91}{\sqrt{\frac{.1719}{2}}} = \frac{.91}{\sqrt{.086}} = \frac{.91}{.293} = 3.11$$

$t_{(.05,\,2)} = 4.303$

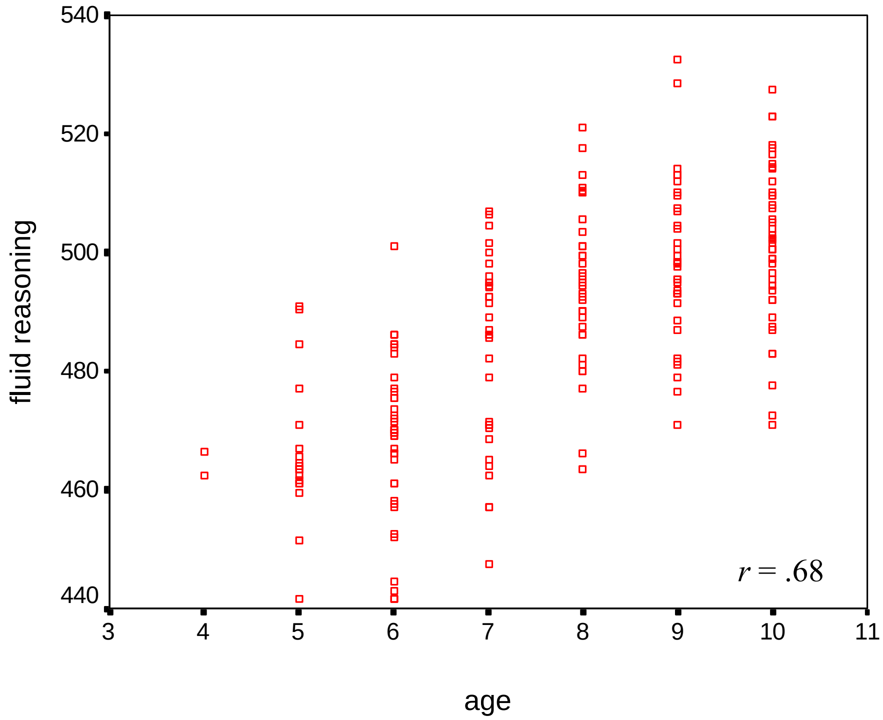
observed t < critical t, thus we retain H_0: $\rho = 0$

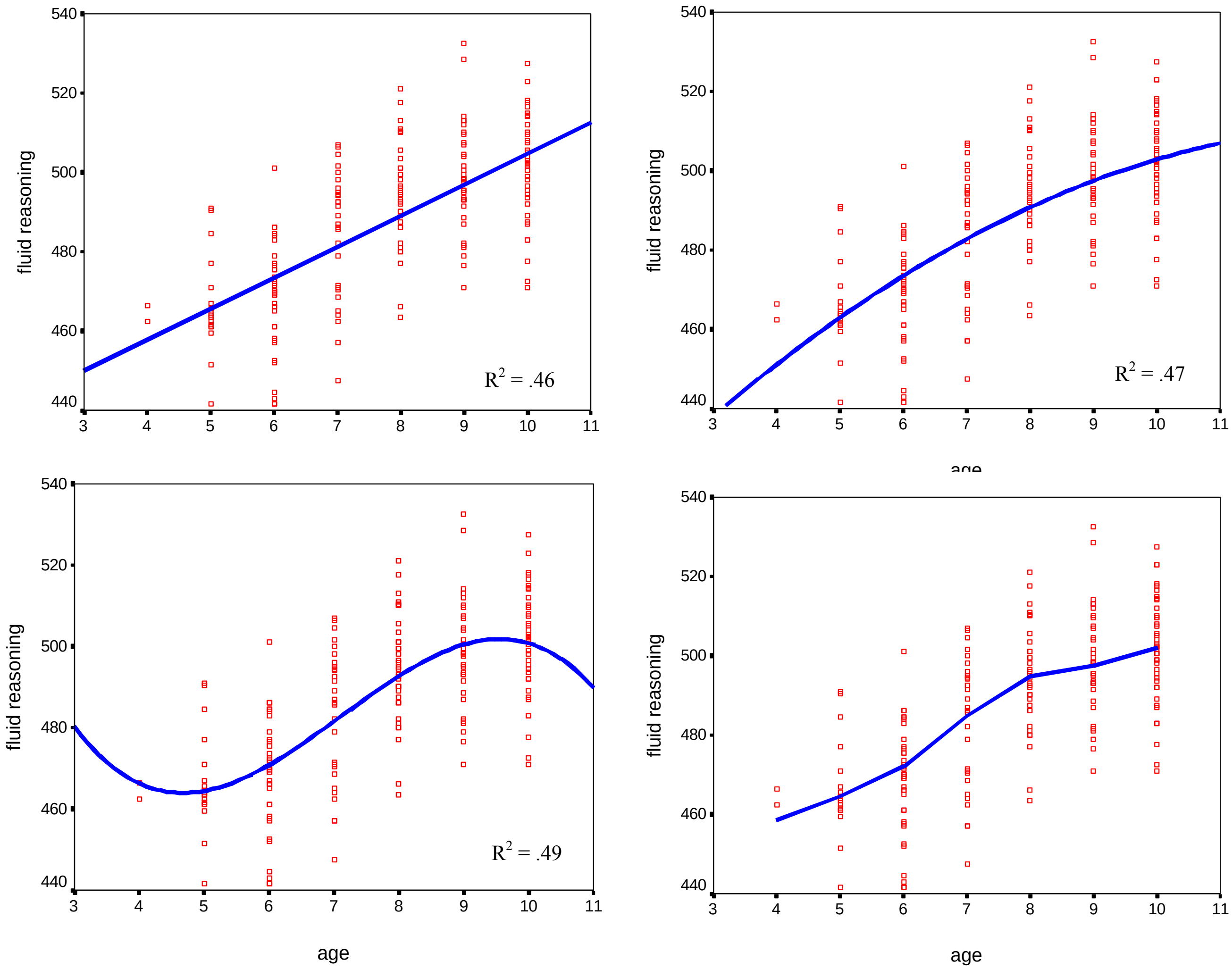
- Other tests are also possible

H_0: ρ = fixed value
H_0: $\rho_1 = \rho_2$
H_0: $\rho_1 = \rho_2 = \rho_3 = \ldots = \rho_k$
H_0: $\rho_1 = \rho_2 = \rho_3 = \ldots = \rho_k$

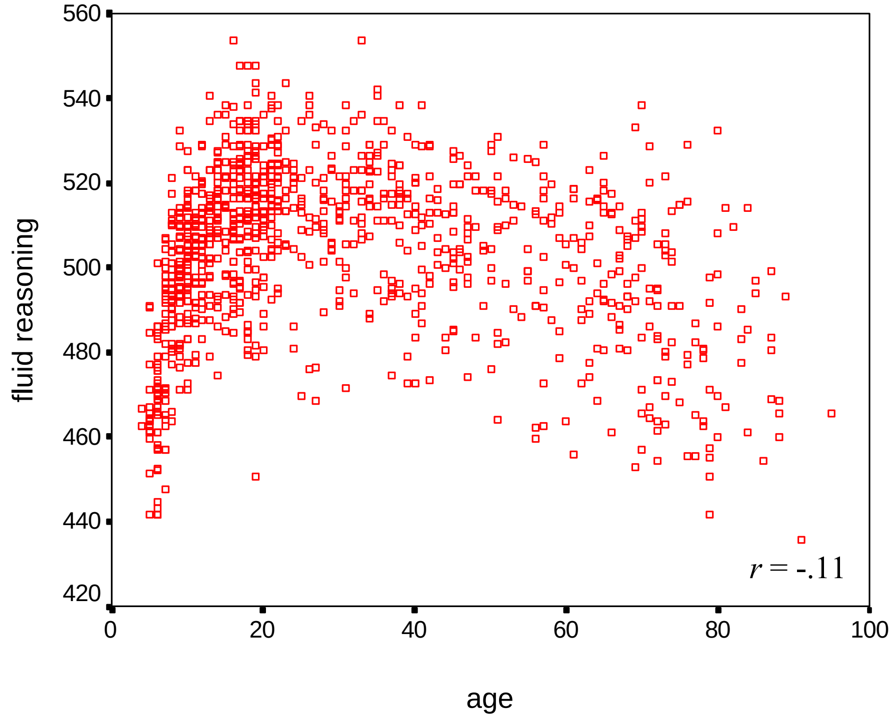
- **13. 10 Factors Affecting *r***

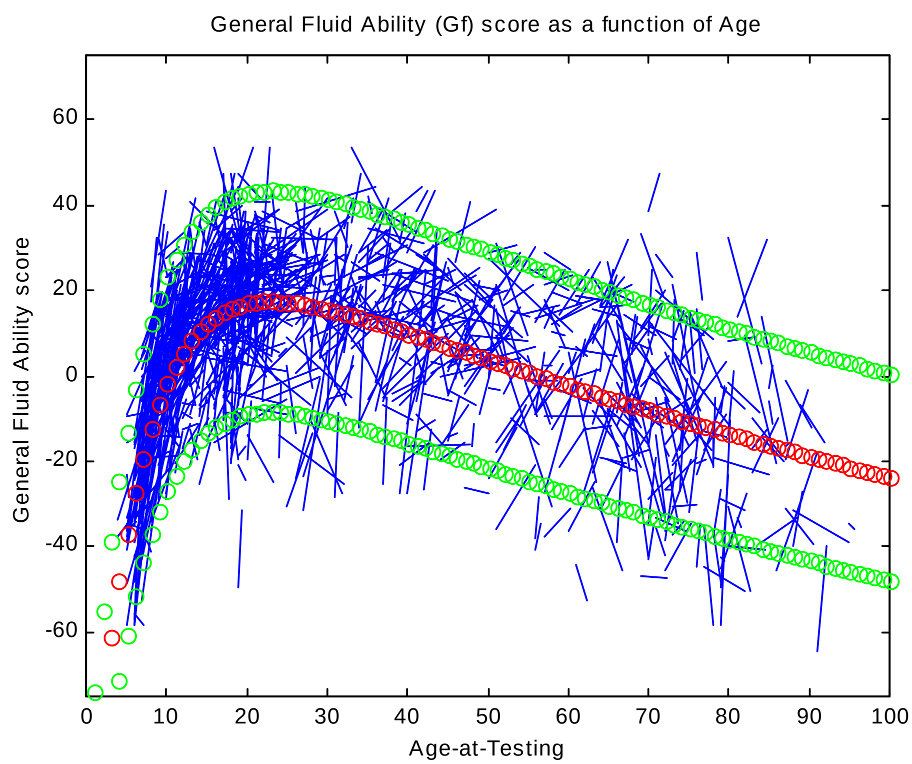
- Linear relationships

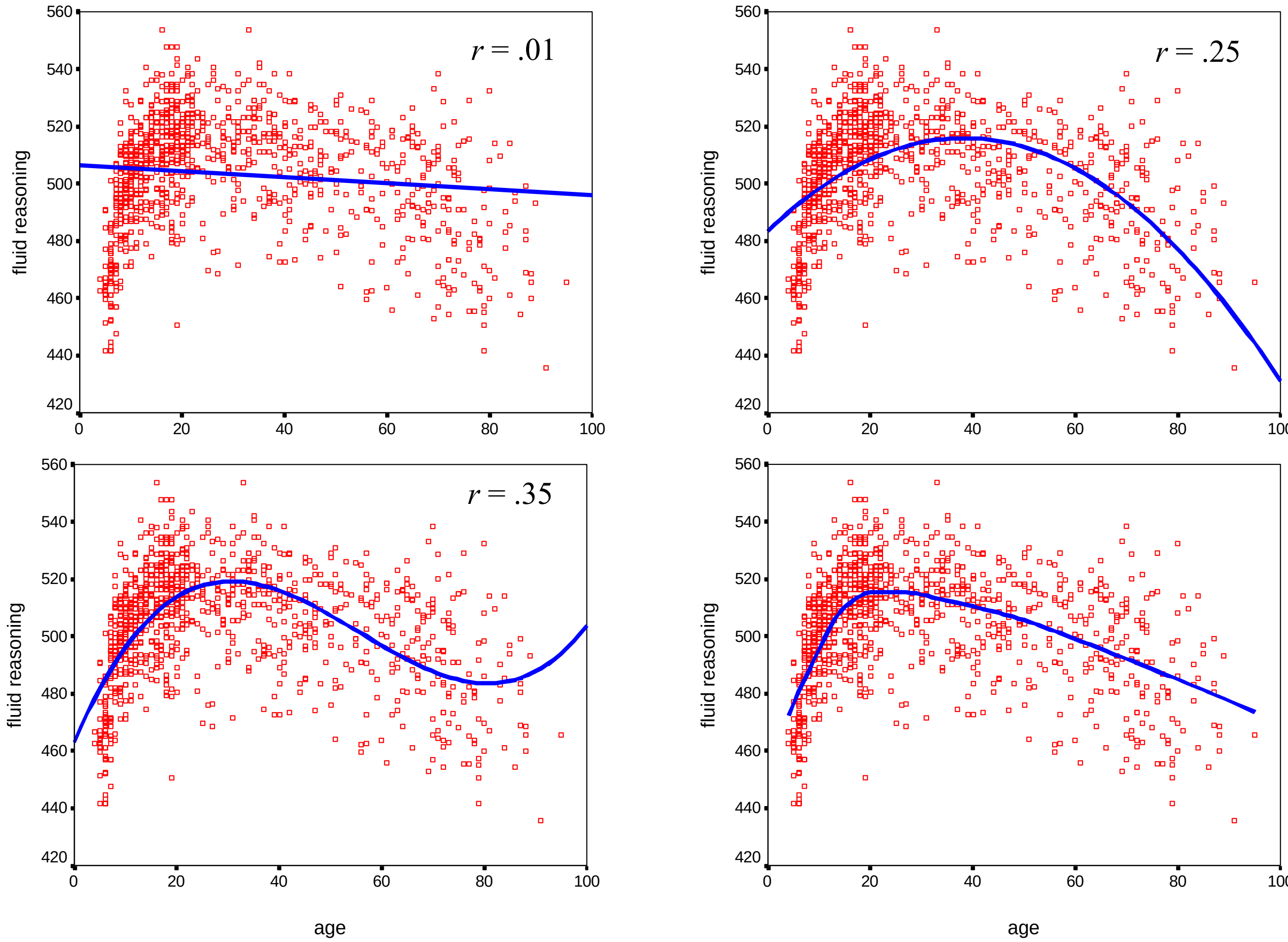



fluid reasoning
age
$R^2 = .46$
$R^2 = .47$
$R^2 = .49$
540
520
500
480
460
440
3
4
5
6
7
8
9
10
11

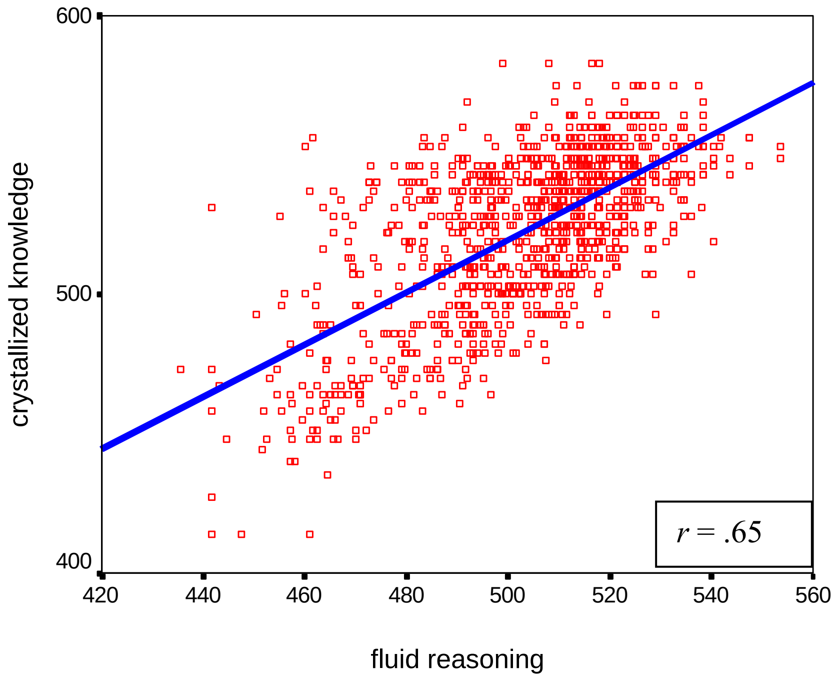
- Nonlinear relationships

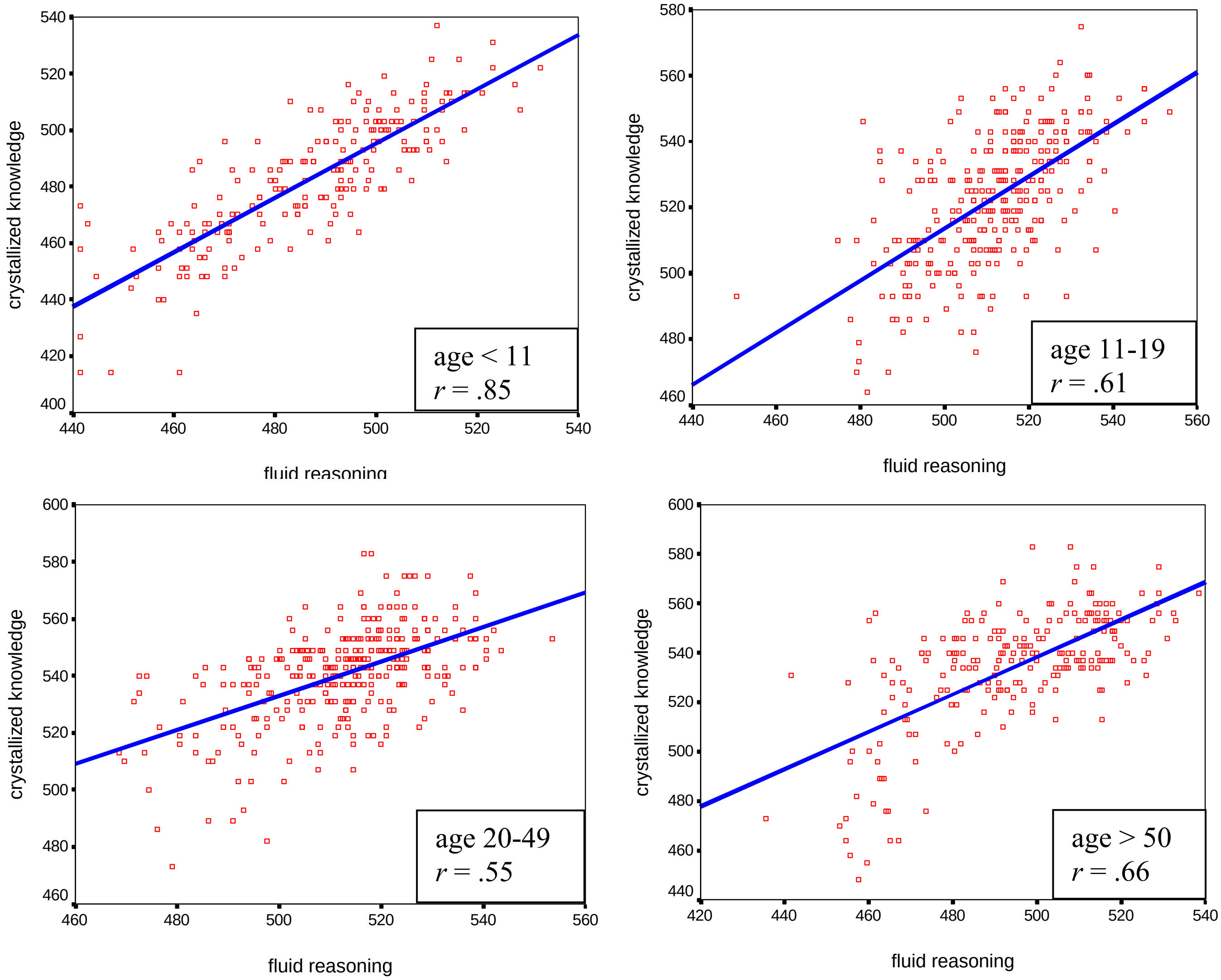





r = .01
r = .25
r = .35
fluid reasoning
age
560
540
520
500
480
460
440
420
0
20
40
60
80
100

- Third variable (also restriction in range)





- *Other Factors*
    - Outliers
    - Heteroskedasticity
    - Curvilinearity
    - Selection (e.g., restricted range)
    - Mismatched distributions
    - Group membership

• 4. 11 Correlation and Causality

- If X and Y are correlated, there are several directions of causality, including:
  - X could be causing Y $\quad X \rightarrow Y$
  - Y could be causing X $\quad Y \rightarrow X$
  - Some other factor Z could be causing both X and Y $\quad Z \rightarrow X, Y$
  - Both X and Y are causing each other $\quad X \leftrightarrow Y$

- Ruling out directions of causality
  - Additional knowledge
    - correlation between sleep one night and mood next day cannot be due to mood next day causing better sleep the night before (but we can't say that sleep → mood either)
  - Randomization
    - if individuals are randomly assigned to two groups, and a manipulation (sleep deprivation) is performed on one group (experimental) but not to the other (control), any relationship between X and Y (sleep and mood) that is different between groups should be due to the manipulation

• 4. 12 Correlation Coefficient vs. Correlational Methods

- Correlation coefficient (r or ρ) is a statistical procedure
- Correlational method is a type of research design that does not involve true experiments (with randomization)
- Correlational methods do not necessarily use the correlation coefficient as the statistical procedure for analyzing the data

- ## 4. 13 Restriction in Range

- It is important to examine the correlation between X and Y when considering the entire range of the variables

    - Association between SAT scores and performance in college (or GRE and performance in graduate school)

    - Association between age and memory across the life span