Correlation

Lecture 45
Section 13.7

Robb T. Koether

Hampden-Sydney College

Mon, Apr 16, 2012

Outline

Outline

The Correlation Coefficient

- Recall

$$\begin{aligned} \text{SSX} &= \sum (x - \overline{x})^2, \\ \text{SSY} &= \sum (y - \overline{y})^2, \\ \text{SSXY} &= \sum (x - \overline{x})(y - \overline{y}). \end{aligned}$$

Then the correlation coefficient is

$$r = \frac{\text{SSXY}}{\sqrt{\text{SSX} \cdot \text{SSY}}}.$$

Some Correlations

- For the Free Lunch vs. Graduation Rate,

$$r = -0.8544.$$

Some Correlations

- For the Free Lunch vs. Graduation Rate,

$$r = -0.8544.$$

- For English SOL Passing Rate vs. Graduation Rate,

$$r = 0.7500.$$

Some Correlations

- For the Free Lunch vs. Graduation Rate,

$$r = -0.8544.$$

- For English SOL Passing Rate vs. Graduation Rate,

$$r = 0.7500.$$

- For the Teachers' Salary vs. Graduation Rate,

$$r = 0.0817.$$

Some Correlations

- For the Free Lunch vs. Graduation Rate,

$$r = -0.8544.$$

- For English SOL Passing Rate vs. Graduation Rate,

$$r = 0.7500.$$

- For the Teachers' Salary vs. Graduation Rate,

$$r = 0.0817.$$

- For the S/T Ratio vs. Graduation Rate,

$$r = 0.0022.$$

Another Formula for r

- Another formula for r is

$$r = \frac{1}{(n-1)} \sum \left(\frac{x - \overline{x}}{s_x} \right) \left(\frac{y - \overline{y}}{s_y} \right).$$

- This formula is based on the z-scores of x and y.

Another Formula for r

Example (Correlation Coefficient)

x	y	$\frac{x-\overline{x}}{s_x}$	$\frac{y-\overline{y}}{s_y}$	$\left(\frac{x-\overline{x}}{s_x}\right)\left(\frac{y-\overline{y}}{s_y}\right)$
1	8			
5	7			
9	7			
9	5			
11	3			

Another Formula for r

Example (Correlation Coefficient)

x	y	$\frac{x-\overline{x}}{s_x}$	$\frac{y-\overline{y}}{s_y}$	$\left(\frac{x-\overline{x}}{s_x}\right)\left(\frac{y-\overline{y}}{s_y}\right)$
1	8	-1.5		
5	7	-0.5		
9	7	0.5		
9	5	0.5		
11	3	1.0		

Another Formula for r

Example (Correlation Coefficient)

x	y	$\frac{x-\overline{x}}{s_x}$	$\frac{y-\overline{y}}{s_y}$	$\left(\frac{x-\overline{x}}{s_x}\right)\left(\frac{y-\overline{y}}{s_y}\right)$
1	8	-1.5	1.0	
5	7	-0.5	0.5	
9	7	0.5	0.5	
9	5	0.5	-0.5	
11	3	1.0	-1.5	

Another Formula for r

Example (Correlation Coefficient)

x	y	$\frac{x-\overline{x}}{s_x}$	$\frac{y-\overline{y}}{s_y}$	$\left(\frac{x-\overline{x}}{s_x}\right)\left(\frac{y-\overline{y}}{s_y}\right)$
1	8	-1.5	1.0	-1.50
5	7	-0.5	0.5	-0.25
9	7	0.5	0.5	0.25
9	5	0.5	-0.5	-0.25
11	3	1.0	-1.5	-1.50

Another Formula for r

Example (Correlation Coefficient)

x	y	$\frac{x-\overline{x}}{s_x}$	$\frac{y-\overline{y}}{s_y}$	$\left(\frac{x-\overline{x}}{s_x}\right)\left(\frac{y-\overline{y}}{s_y}\right)$
1	8	-1.5	1.0	-1.50
5	7	-0.5	0.5	-0.25
9	7	0.5	0.5	0.25
9	5	0.5	-0.5	-0.25
11	3	1.0	-1.5	-1.50
				-3.25

Another Formula for r

Example (Correlation Coefficient)

- Now compute

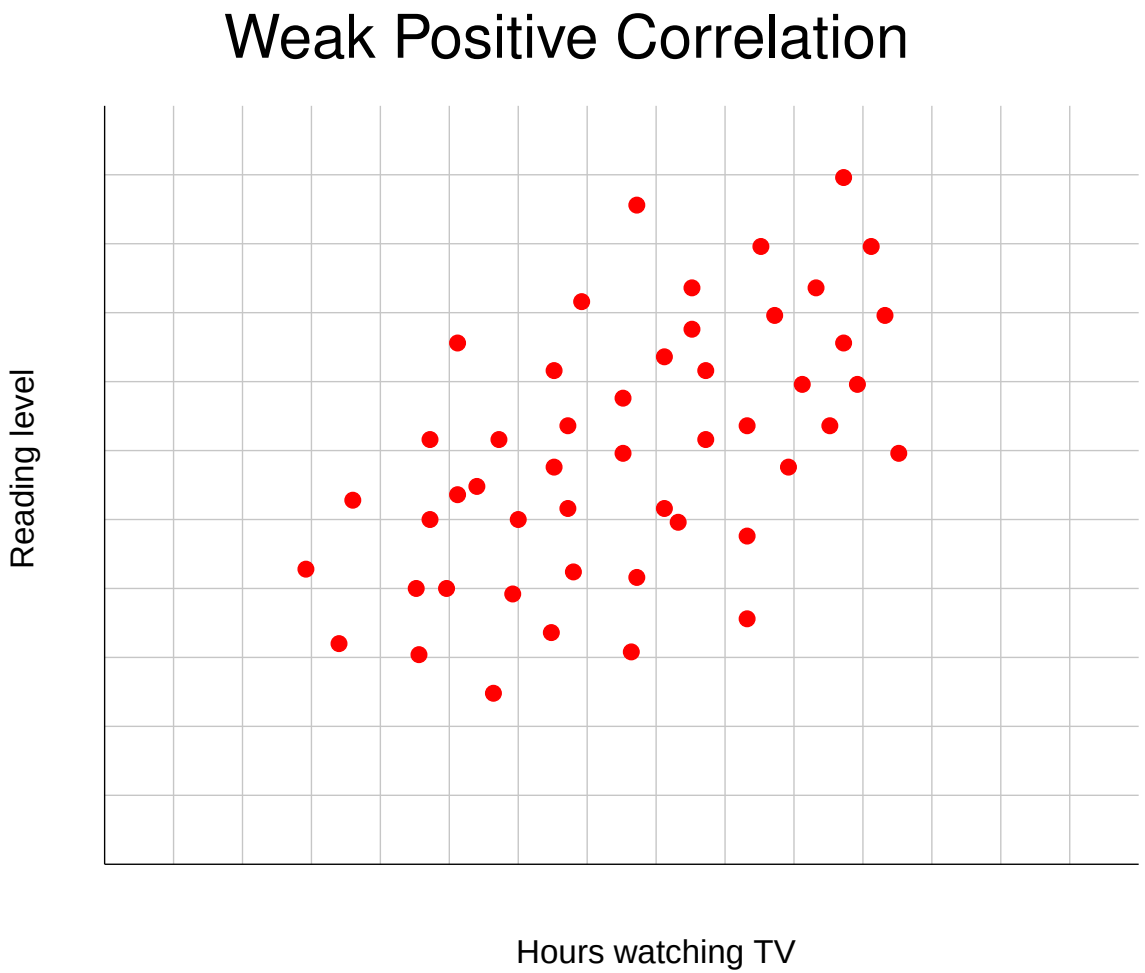
$$r = \frac{1}{4}(-3.25) = -0.8125.$$

Outline

The Danger of Hidden Variables

- Suppose we investigate the relationship in children (3rd - 5th grade) between number of hours spent watching TV and their reading level.
- One would expect a negative relationship.

The Danger of Hidden Variables



The Danger of Hidden Variables

- There appears to be a weak positive relationship between the two variables!

The Danger of Hidden Variables

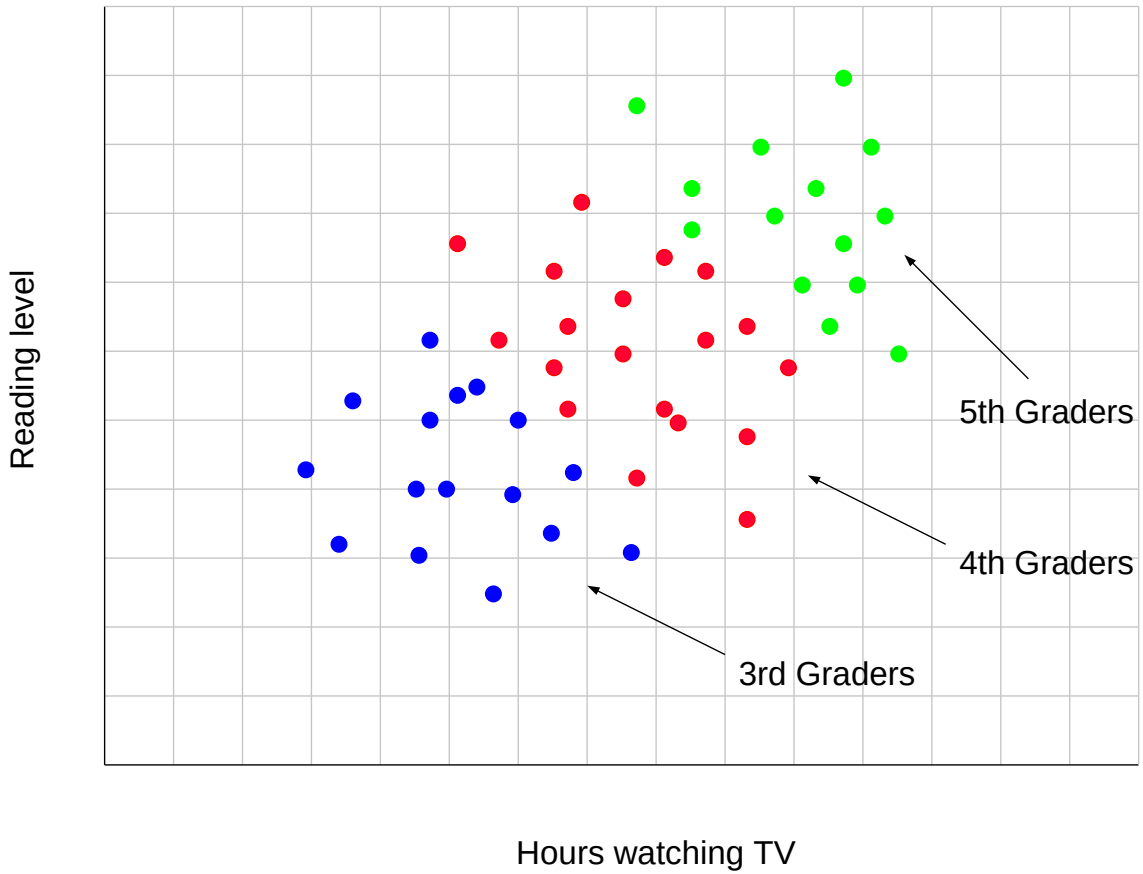
- There appears to be a weak positive relationship between the two variables!
- What an amazing discovery! Tell the kids to shut the books and watch more TV!

The Danger of Hidden Variables

- But now consider a possible third variable: grade level.
- Divide the sample into three groups according to grade level (3rd, 4th, and 5th).

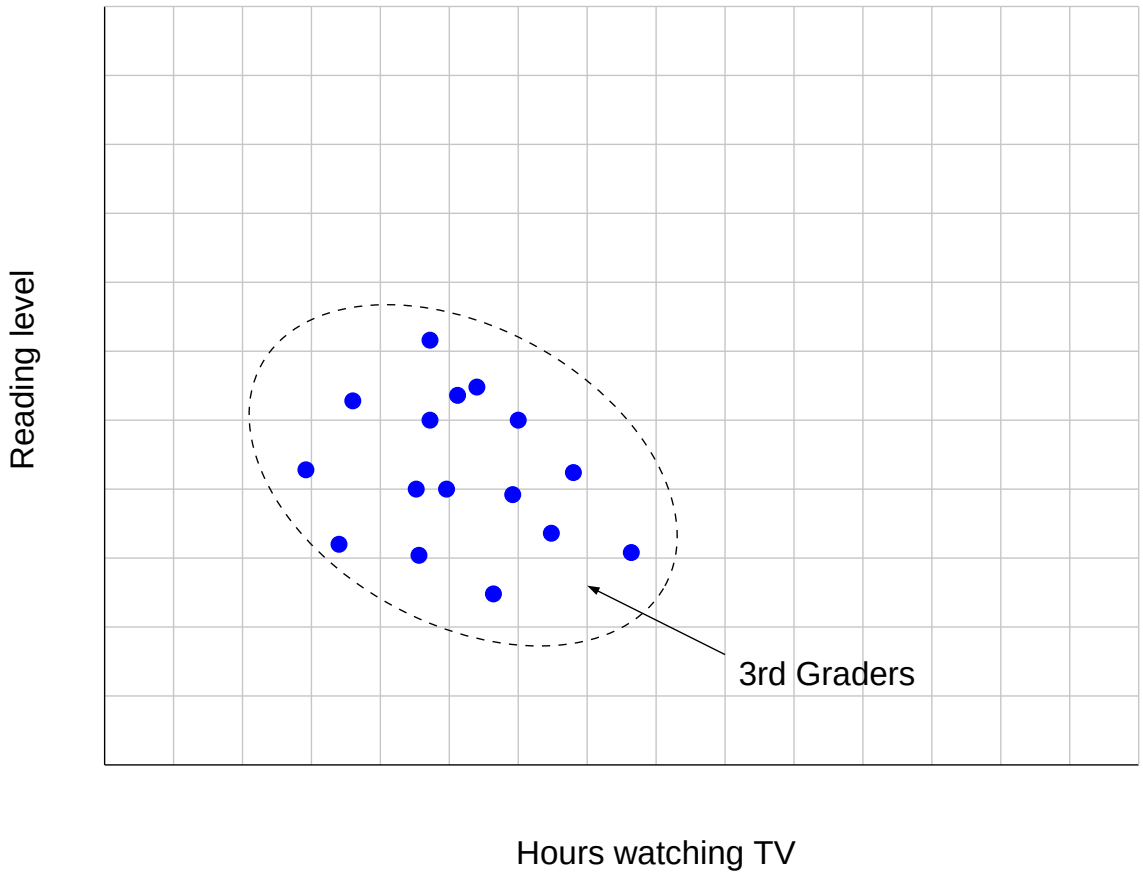
The Danger of Hidden Variables

Weak Positive Correlation?



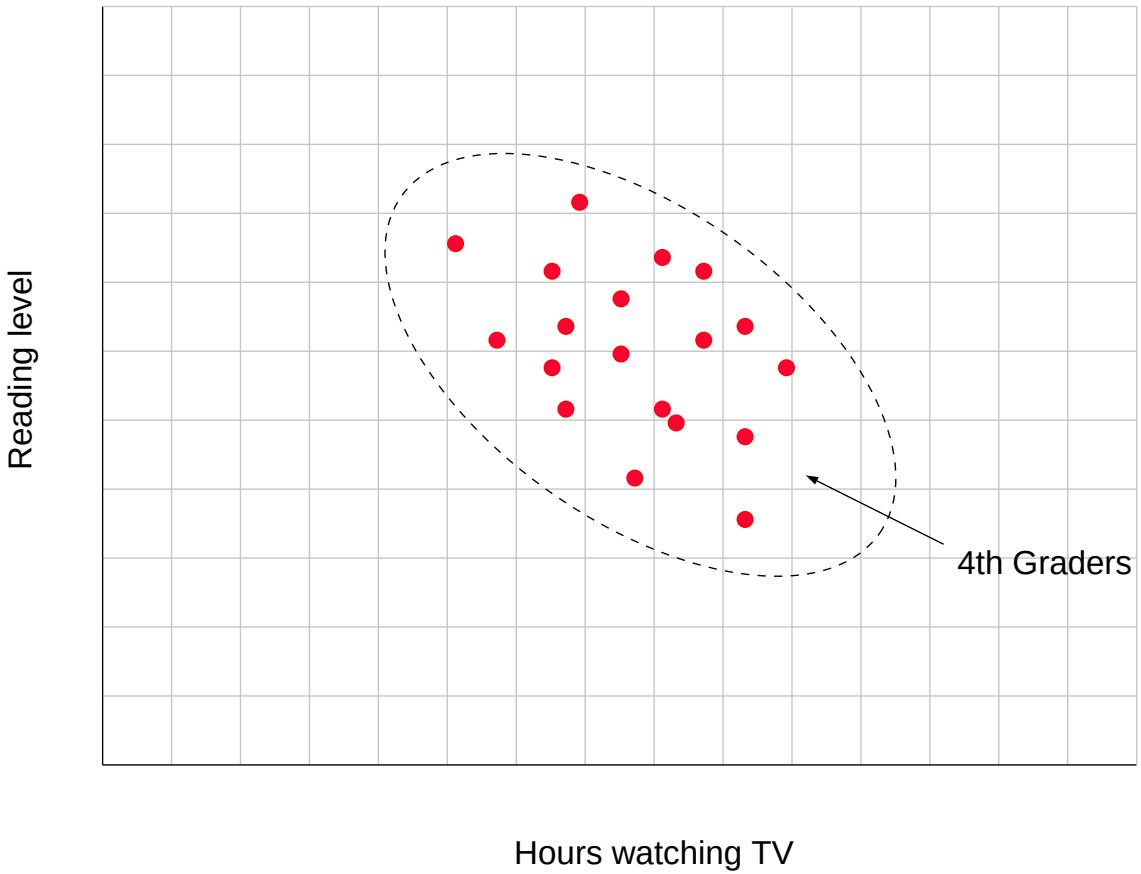
The Danger of Hidden Variables

Weak Negative Correlation Among 3rd Graders



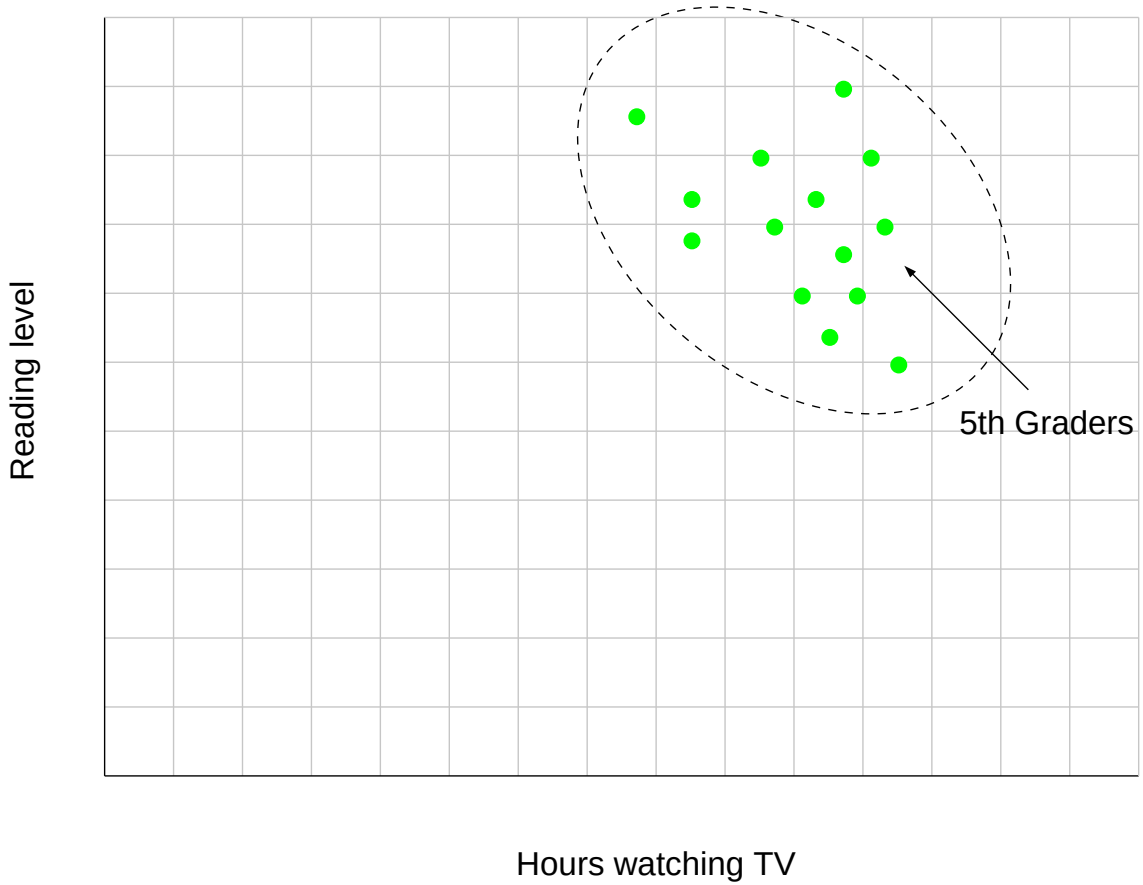
The Danger of Hidden Variables

Weak Negative Correlation Among 4th Graders



The Danger of Hidden Variables

Weak Negative Correlation Among 5th Graders

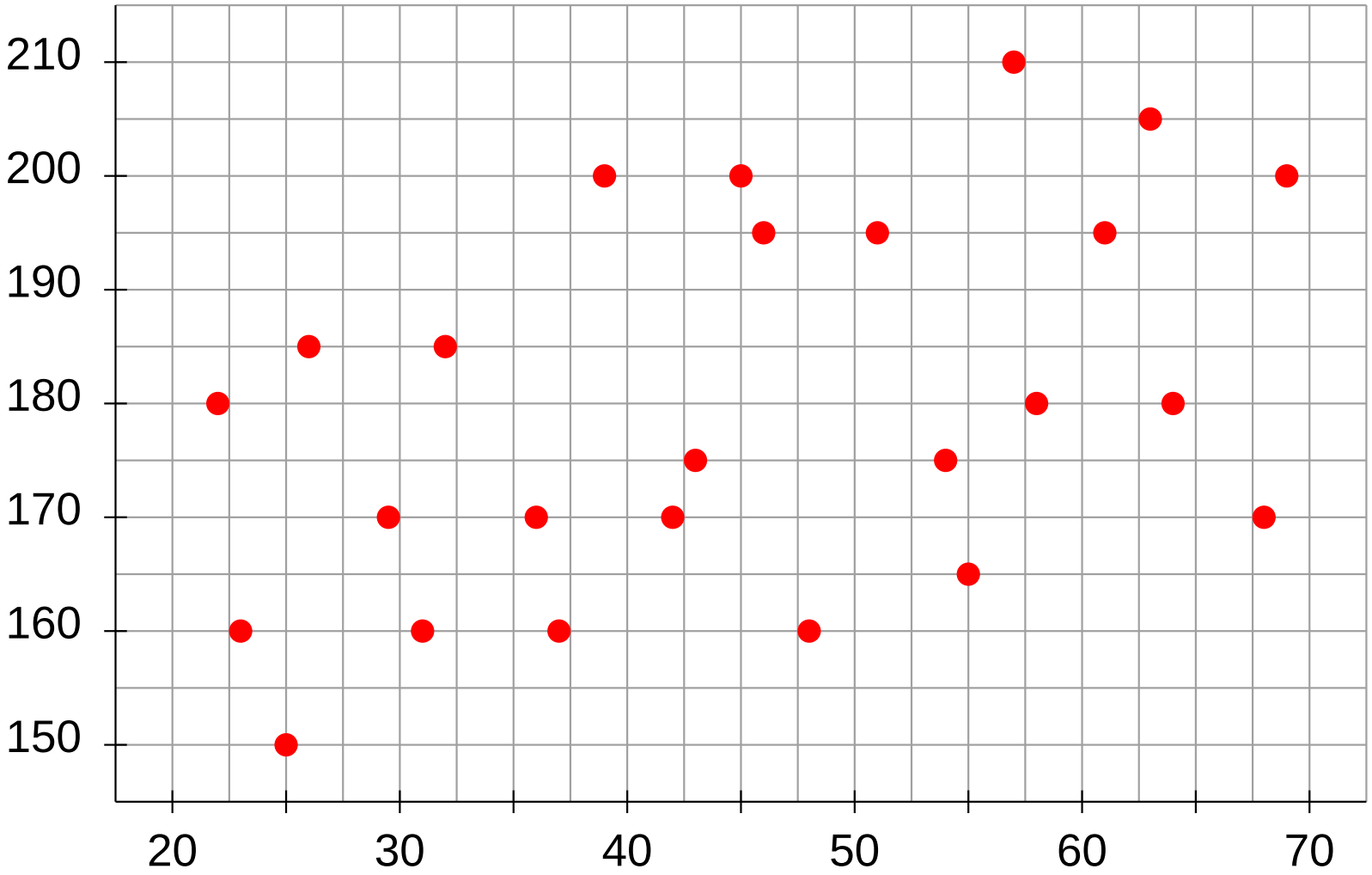


Correlating Averages

- Consider the following age (x) and weight (y) data.

x	y	x	y	x	y	x	y	x	y
22	180	31	160	42	170	51	195	61	195
23	160	32	185	44	175	54	175	63	205
25	155	36	170	45	200	55	165	64	180
26	185	37	160	46	195	57	210	68	170
29	170	39	200	48	160	58	180	69	200

Correlating Averages



Correlating Averages

- The regression line is

$$\hat{y} = 157.944 + 0.490x$$

and

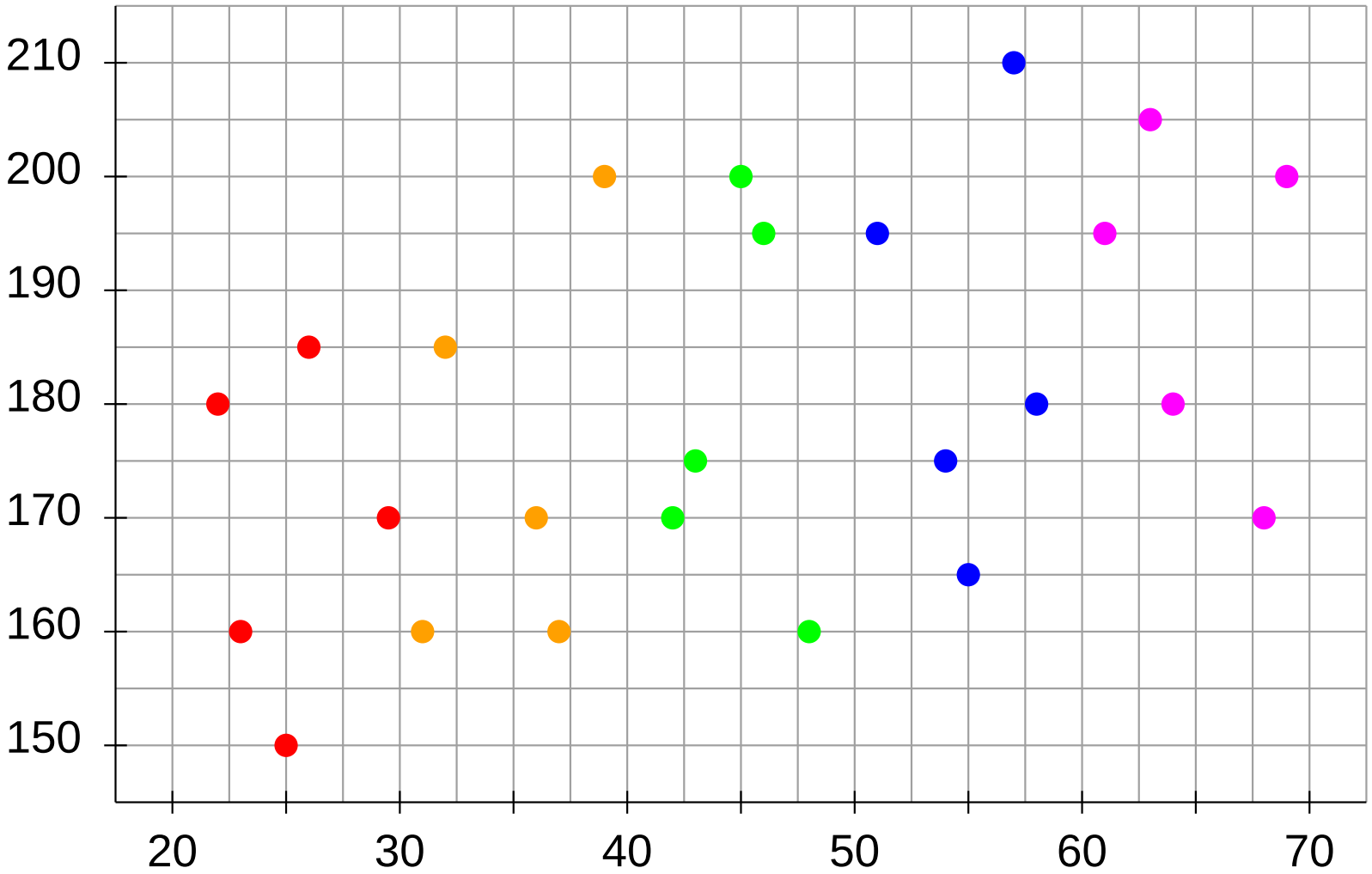
$$r = 0.4423.$$

- The association is positive and weak.

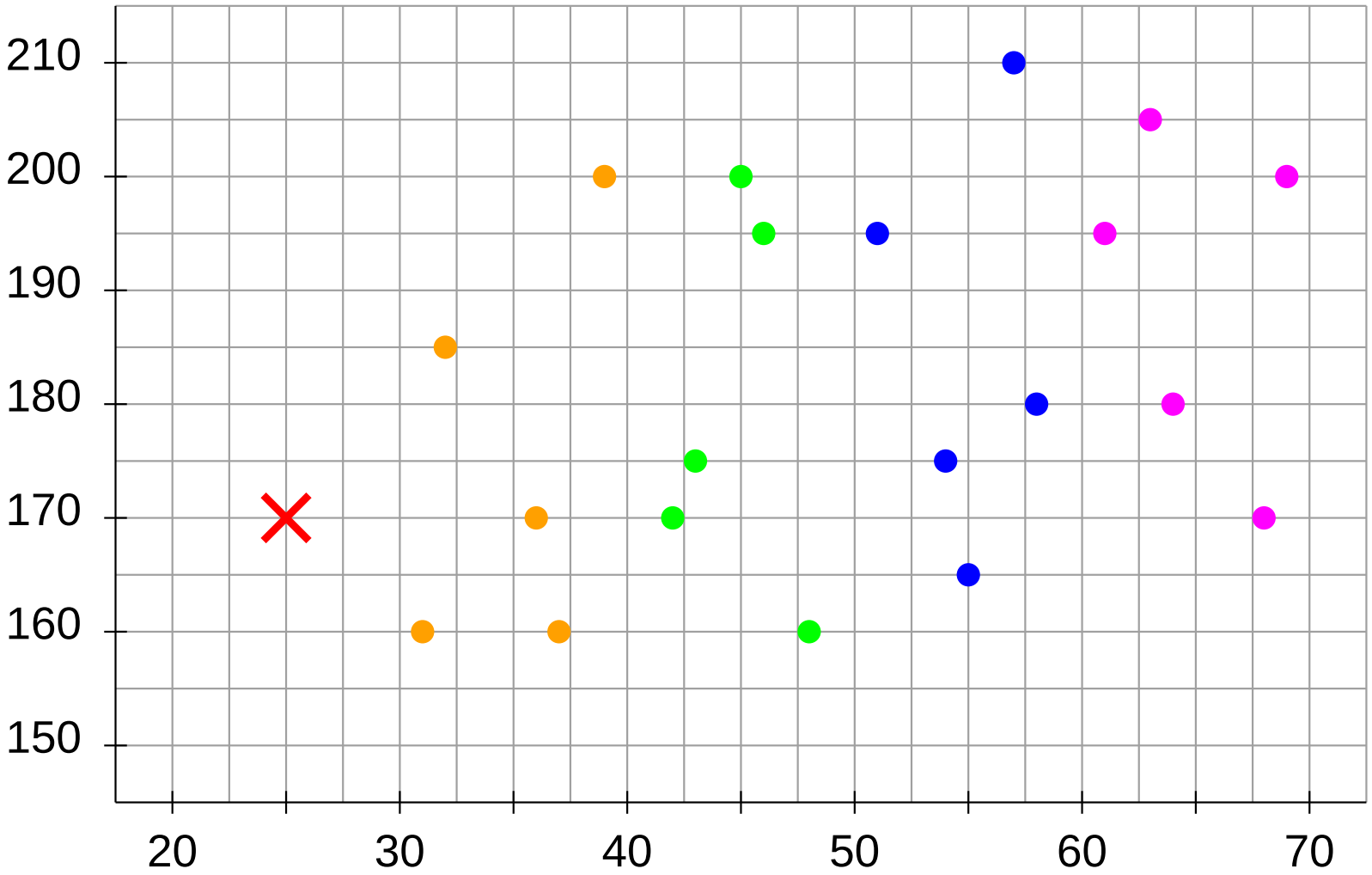
Correlating Averages

- Now group the data by age:
  - Group I: Age 20 - 29.
  - Group II: Age 30 - 39.
  - Group III: Age 40 - 49.
  - Group IV: Age 50 - 59.
  - Group V: Age 60 - 69.
- Represent each group as a single data point using the group's average age and average weight.

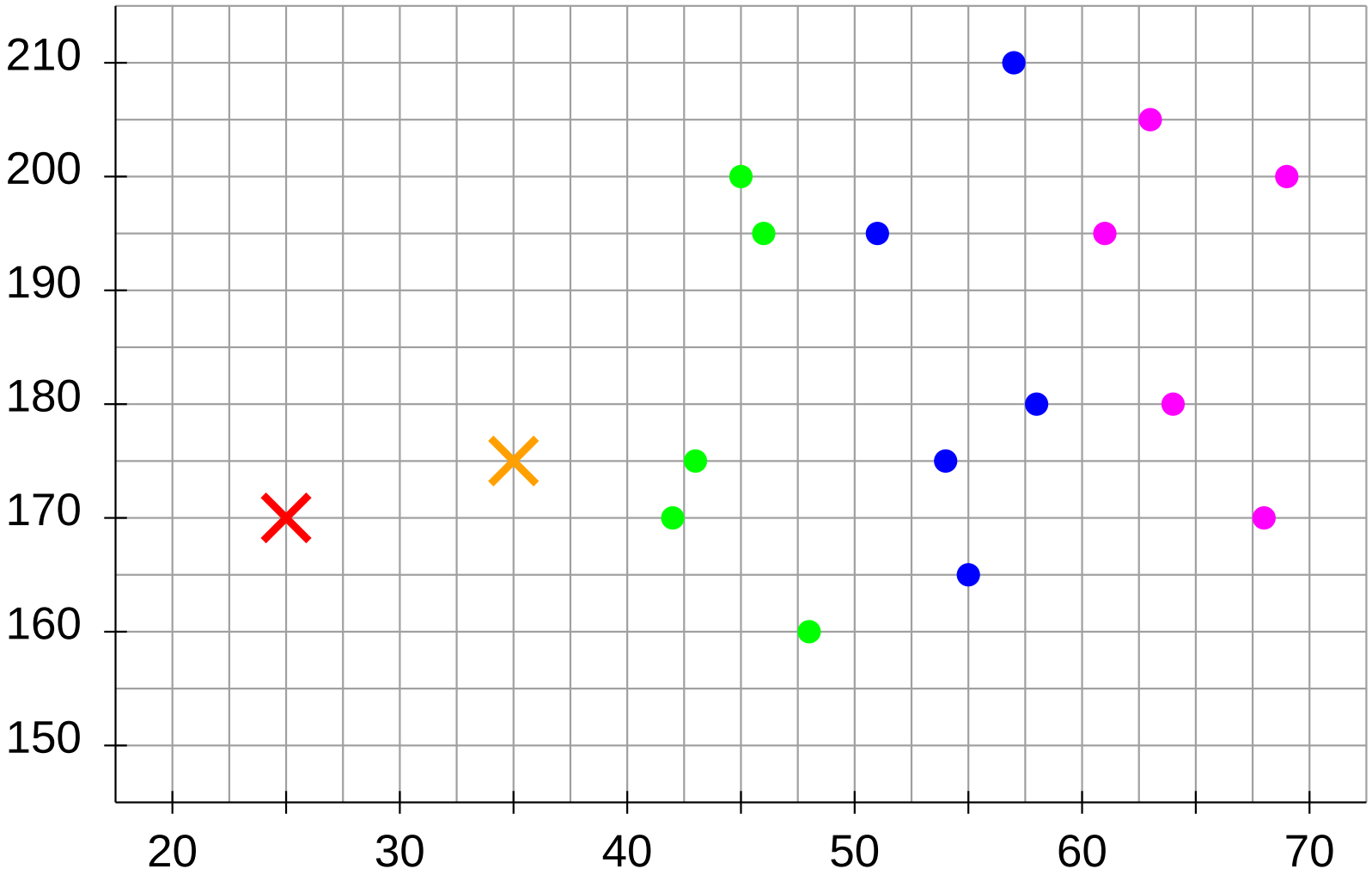
Correlating Averages



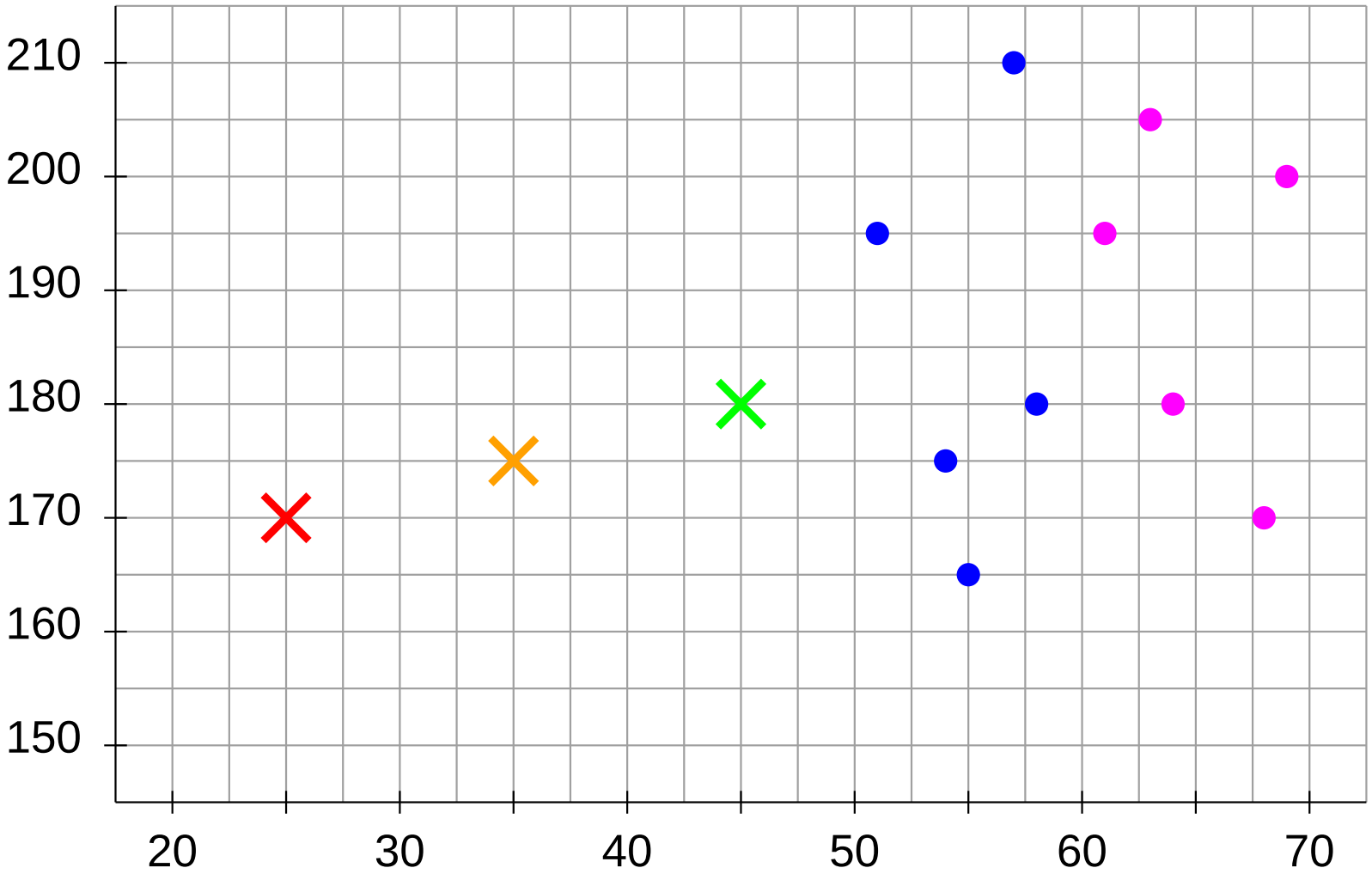
Correlating Averages



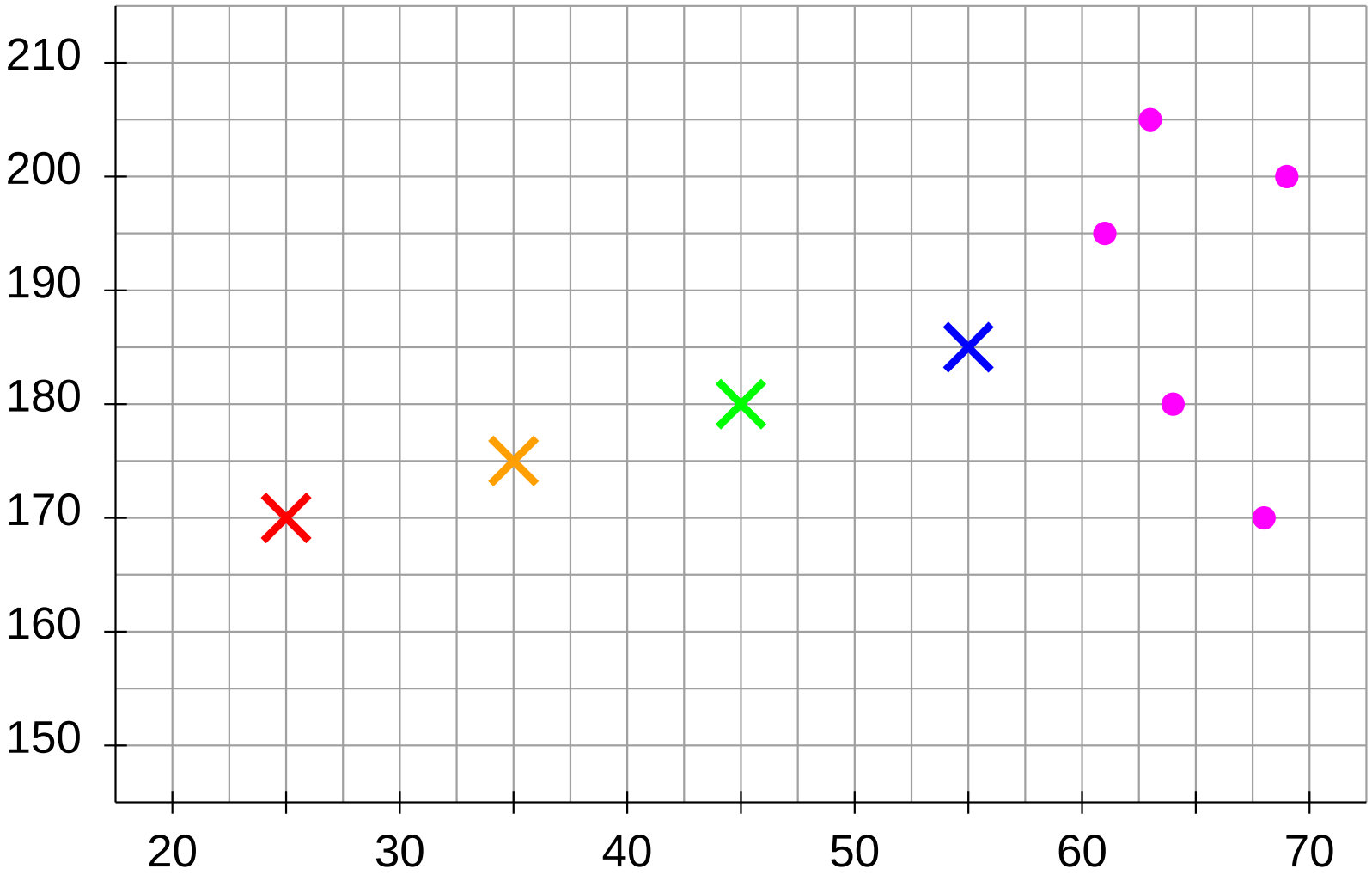
Correlating Averages



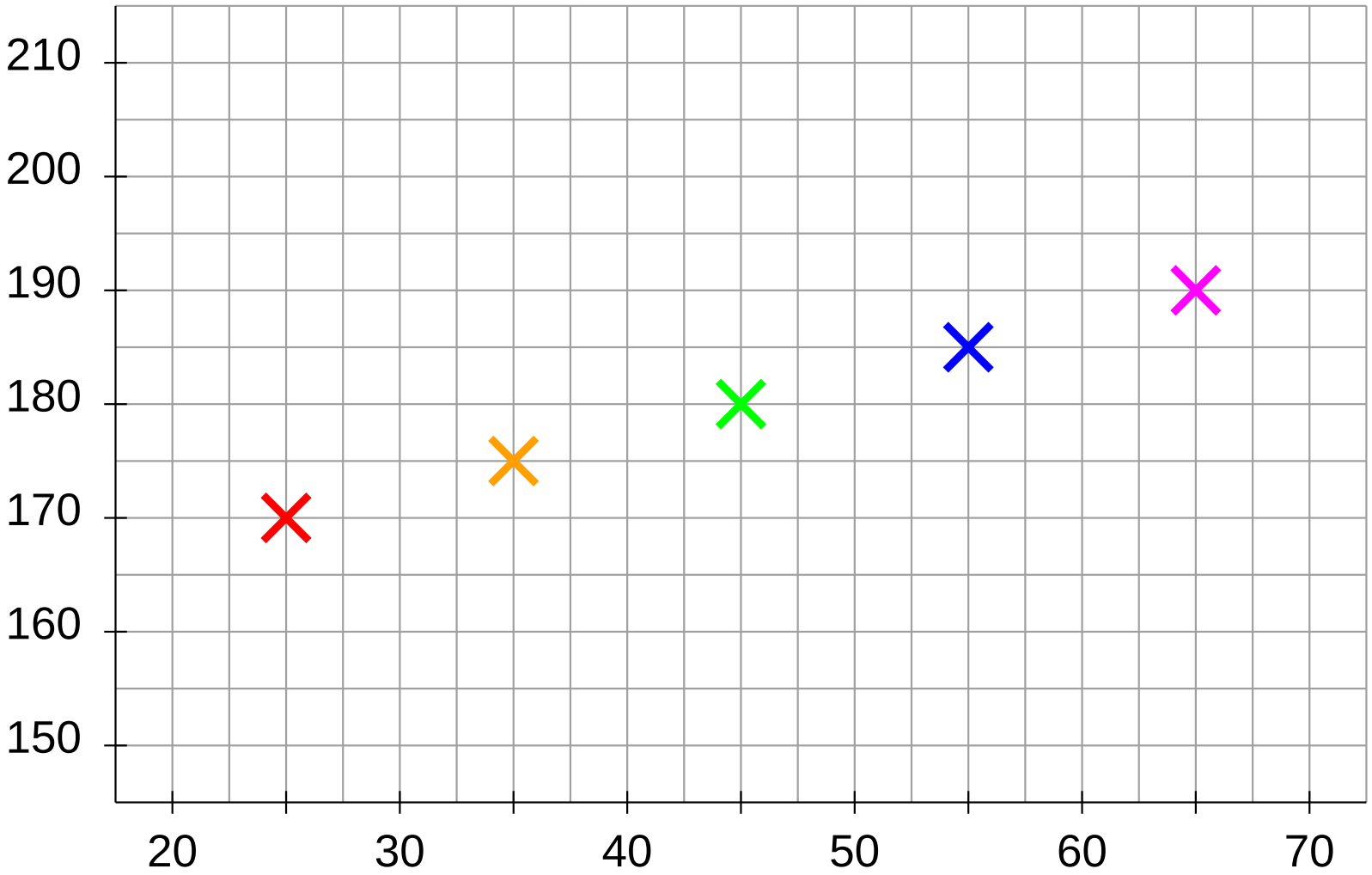
Correlating Averages



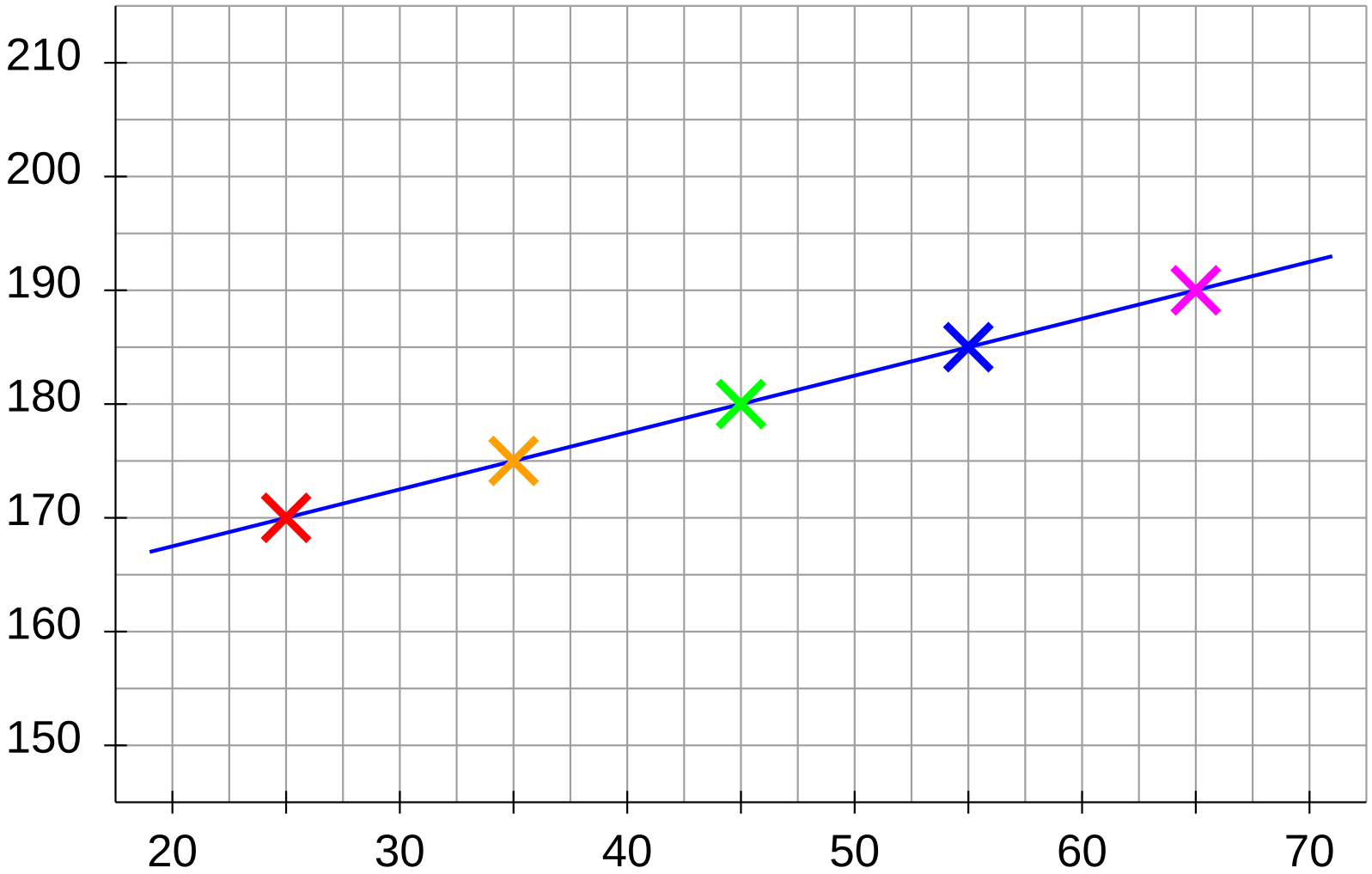
Correlating Averages



Correlating Averages



Correlating Averages



Correlating Averages

- The new data points are

$\overline{x}$	$\overline{y}$
25	170
35	175
45	180
55	185
65	190

Correlating Averages

- The regression line is

$$\hat{y} = 157.5 + 0.5x$$

and

$$r = 1.000.$$

- The association is positive and not just strong, but *perfect*!
- Does this mean that age is a *perfect* predictor of weight?

Outline

Assignment

Homework

- Read Section 13.7, pages 854 - 858.
- Let's Do It! 13.11, 13.14.
- Exercises 18, 19, 20, 32, page 858.