Example: Can an increase in non-exercise activity (e.g. fidgeting) help people gain less weight?

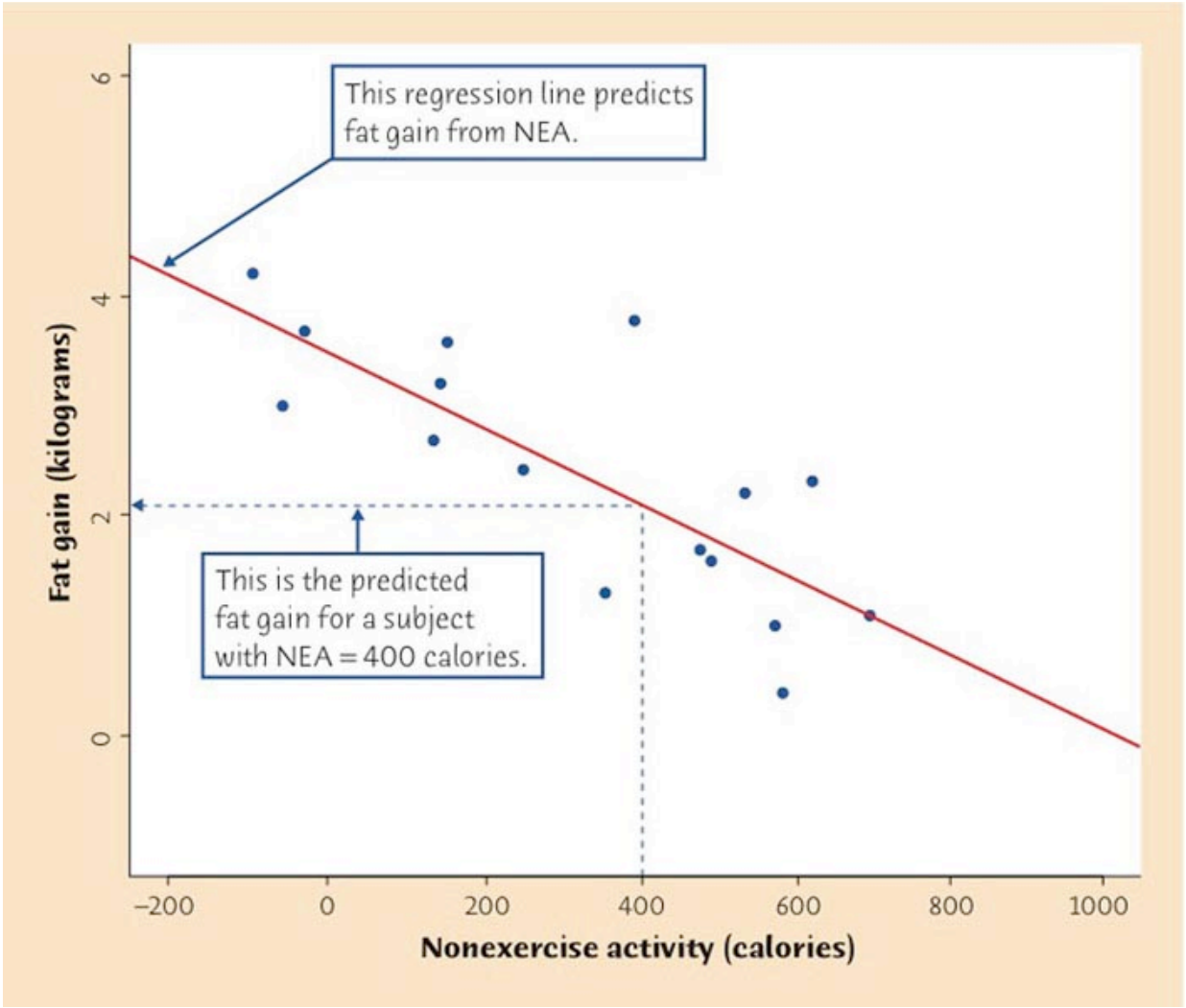
16 subjects overfed for 8 weeks

Explanatory: change in energy use from non-exercise activity (calories)

Response: fat gain (kilograms)

Data see p. 92

Correlation coefficient: r = –0.7786


This regression line predicts fat gain from NEA.
This is the predicted fat gain for a subject with NEA = 400 calories.
Fat gain (kilograms)
0
2
4
6
–200
0
200
400
600
800
1000
Nonexercise activity (calories)

Review of linear equations:

x = explanatory variable
y = response variable

$$y = a + bx$$

b = slope = (change in response)/(change in explanatory)
a = y-intercept

Texas Instruments Graphing Calculator

```
LinReg
 y=a+bx
 a=3.505122916
 b=-.003441487
 r²=.6061492049
 r=-.7785558457
```

fat gain = 3.505 – (0.00344 × NEA change)

Minitab

Session

Regression Analysis: fat versus nea

The regression equation is
fat = 3.51 - 0.00344 nea

Predictor	Coef	SE Coef	T	P
Constant	3.5051	0.3036	11.54	0.000
nea	-0.0034415	0.0007414	-4.64	0.000

S = 0.739853 R-Sq = 60.6% R-Sq (adj) = 57.8%

Microsoft Excel

Book1

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression statistics*					
4	Multiple R	0.778555846				
5	R Square	0.606149205				
6	Adjusted R Square	0.578017005				
7	Standard Error	0.739852874				
8	Observations	16				
9						
10		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	
11	Intercept	3.505122916	0.303616403	11.54458	1.53E-08	
12	nea	-0.003441487	0.00074141	-4.64182	0.000381	
13						

Output | nea data

Example: change in non-exercise activity (NEA) and fat gain

fat gain = 3.505 – (0.00344 × NEA change)

Questions:
(a) What does the slope tell us?

Solution:
(a) $\text{slope} = \dfrac{-0.00344}{1}$

For each additional calorie of NEA, fat gain decreases by 0.00344 kg.
Equivalently, for each additional 100 calories, fat gain decreases by 0.344 kg (≈ 0.75 pound)

fat gain = 3.505 – (0.00344 × NEA change)

Questions:
(b) Predict the fat gain for a person whose NEA increases by 400 calories.

Answer:
Substitute:
fat gain = 3.505 – (0.00344 × 400) = 2.13 kg

Example: 5th grade students
Explanatory: IQ test scores
Response: reading test scores

Reading score = –33.4 + (0.882 × IQ score)

(a) Interpret slope: slope = 0.882
For each 1 point IQ increases, reading test scores increase by 0.882 points.

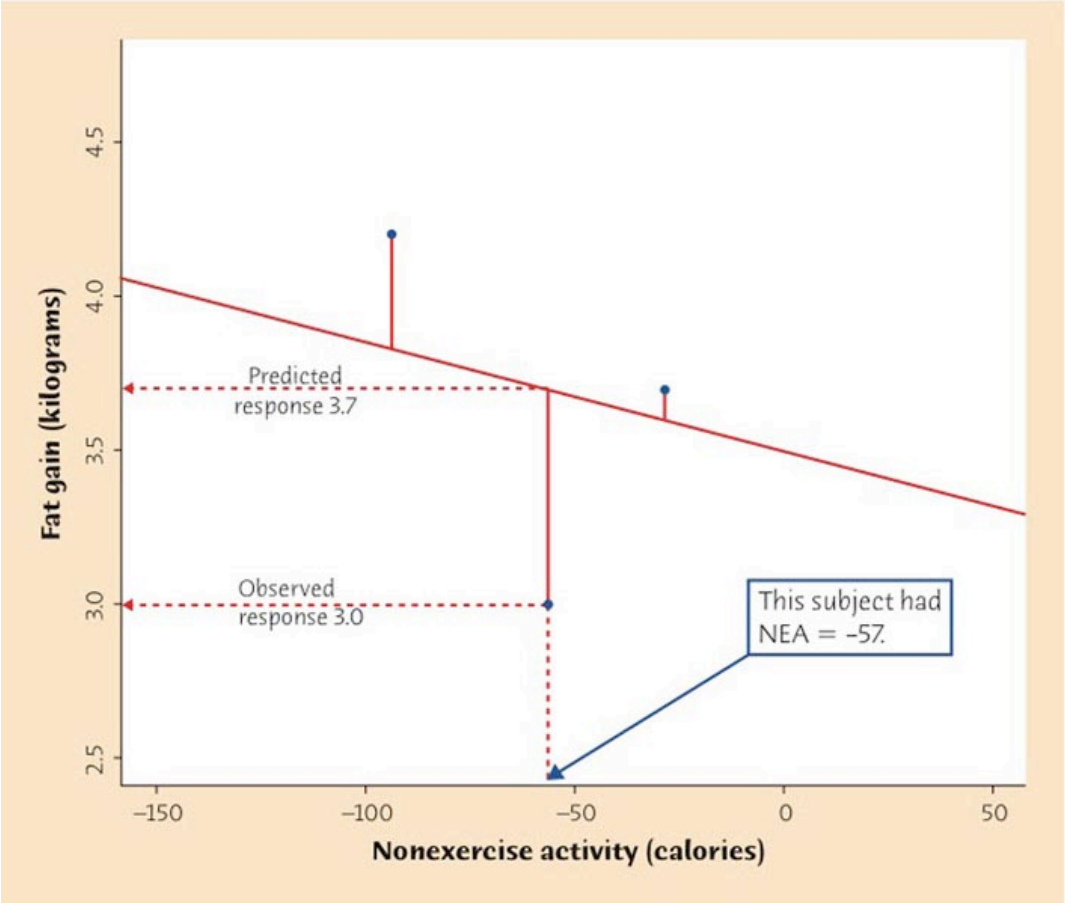
(b) Predict reading scores for children with IQs of 90 and 140:
Substitute:

–33.4 + (0.882 × 90) = 45.98
–33.4 + (0.882 × 140) = 90.08

(c) r = 0.557 for this data. Do we expect the predictions of reading test scores from IQ in part (b) to be accurate?

No, because r = 0.557 means that the correlation between reading test scores and IQ is fairly weak.

Starting from the data, how do we find the best line?
Answer: Minimize the vertical distance of the data to the line.



Explanatory:	x_1	x_2	…	x_n	$\longrightarrow \bar{x}, s_x$
Response:	y_1	y_2	…	y_n	$\longrightarrow \bar{y}, s_y$

The equation of the **Least-Squares Regression Line** is

$$\hat{y} = a + bx$$

with **slope**

$$b = r\frac{s_y}{s_x}$$

and **intercept**

$$a = \bar{y} - b\bar{x}$$

TI-83 or 84: Correlation and linear regression

0. One time preliminary step:
   CATALOG (2nd+'0') → **DiagnosticOn** (Enter twice)

1. **STAT → EDIT → 1:Edit**,
   enter explanatory in L1, response in L2
2. **STAT → CALC → 8:LinReg(a+bx)** (Enter)

Quick practice Example:

L_1:	12	16	20
L_2:	1.60	1.85	1.95

Answer: $a = 1.1$, $b = 0.04375$, $r = 0.970725$,
so the regression line is $y = 1.1 + 0.04375\,x$

Example: Starbucks Coffee

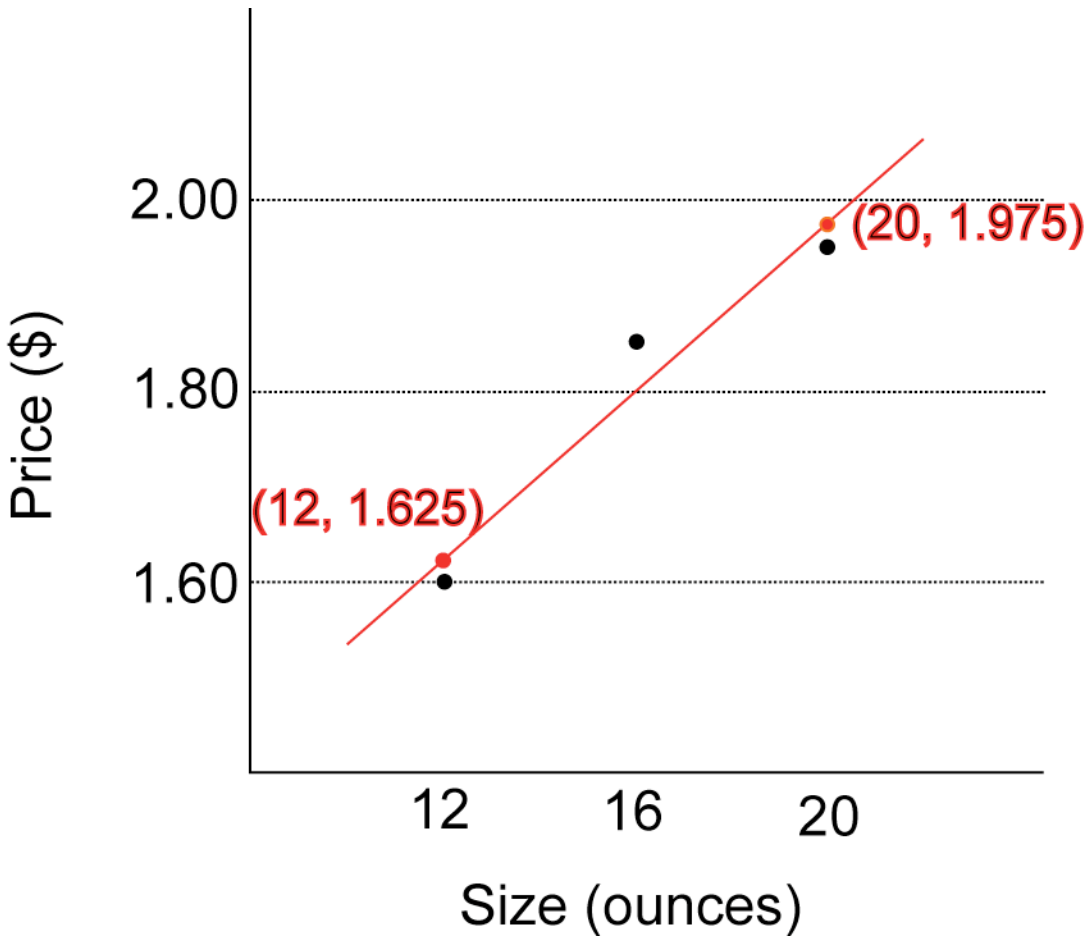
	Size (ounces):	Price ($):
Tall	12	1.60
Grande	16	1.85
Venti	20	1.95

We calculate r = 0.970725, a = 1.1, b = 0.04375

Regression equation: price = 1.1 + (0.04375 × size)

Predicted prices for ...

- Tall: price = 1.1 + (0.04375 × 12) = 1.625
- Grande: price = 1.1 + (0.04375 × 16) = 1.80
- Venti: price = 1.1 + (0.04375 × 20) = 1.975


2.00
1.80
1.60
Price ($)
(20, 1.975)
(12, 1.625)
12
16
20
Size (ounces)

The discrepancy between actual and predicted is called a residual.

Residual = actual y – predicted y

- Tall: residual = 1.60 – 1.625 = –0.025
- Grande: residual = 1.85 – 1.80 = 0.05
- Venti: residual = 1.95 – 1.975 = –0.025

It's not on the menu, but "Short" = 8 ounces.

predicted price = 1.1 + (0.04375 × 8) = 1.45
actual price (Sherman Oaks) = 1.50
residual = 0.05

1. The distinction between explanatory and response variables is essential. Don't get them backwards!

2. The regression line always passes through the point $(\bar{x},\bar{y})$

3. The correlation *r* describes the strength of the linear relationship. More specifically,

the value of r^2 is the percentage of the variation in y that is explained by the regression line

Examples:

1. Explanatory: non-exercise activity “NEA” (calories)
   Response: fat gain (kilograms)
   $r = -0.7786$
   $r^2 = 0.6062$

   So 60.62% of the variation in fat gain is explained by the linear relationship with NEA.

2. Explanatory: Size of a cup of Starbucks coffee (oz)
   Response: Price ($)
   $r = 0.970725$
   $r^2 = 0.942308$

   So 94% of the variation in Starbucks prices is explained by the linear relationship with size.

Example:
Explanatory: Vehicle weight (pounds)
Response: Gas Mileage (mpg)

Toyota models: $r = -0.639$
Honda models: $r = -0.795$

For which company are predictions of mileage based on weight more accurate?

Answer: Honda. For Honda, $r^2 = (-0.795)^2 = 63\%$ of variation in mileage is due to its relationship with weight, while for Toyota only $(-0.639)^2 = 41\%$ of variation in mileage is due to weight.

INFLUENTIAL OBSERVATIONS:

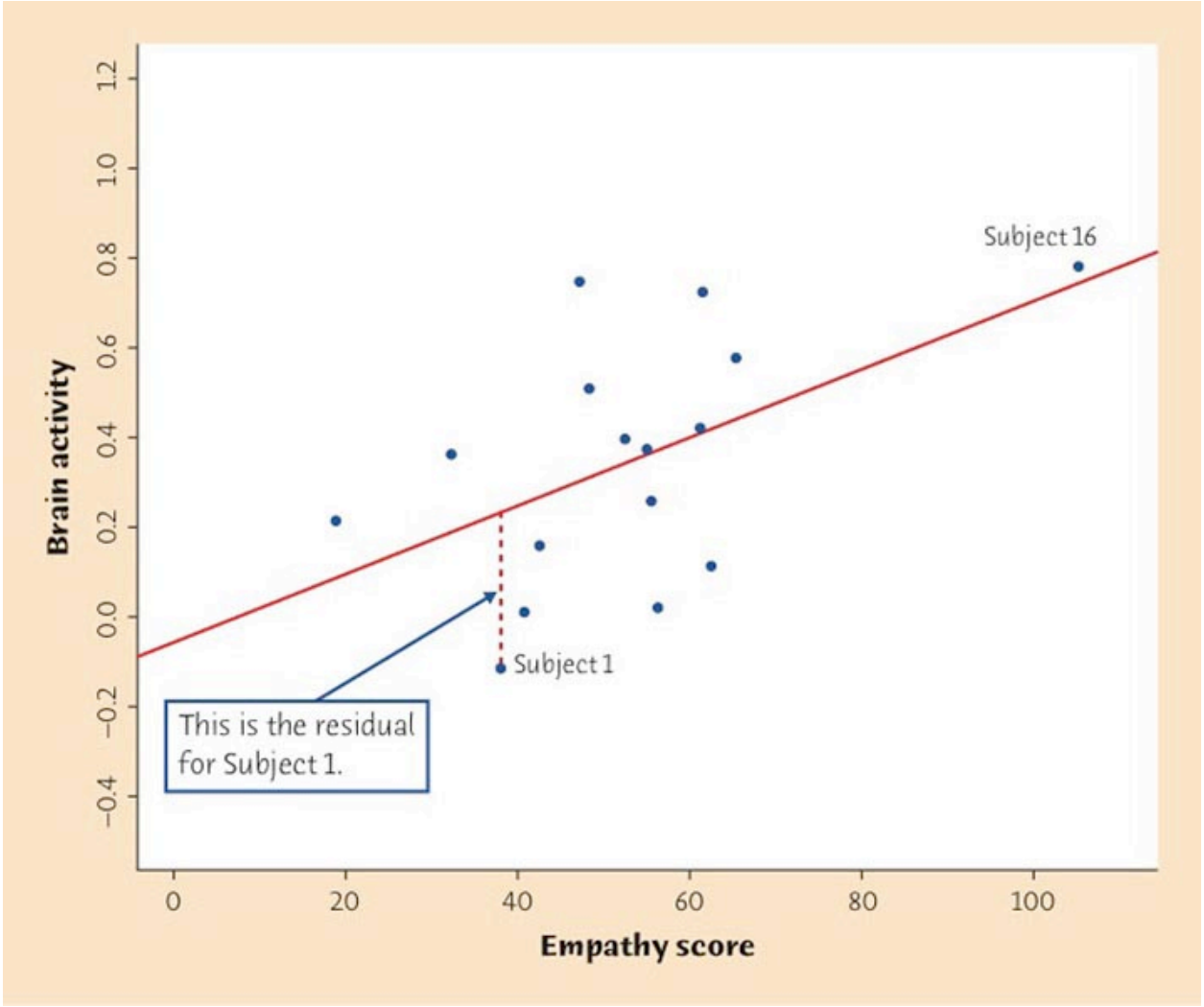
An observation is influential for a statistical quantity if removing it markedly changes that quantity.

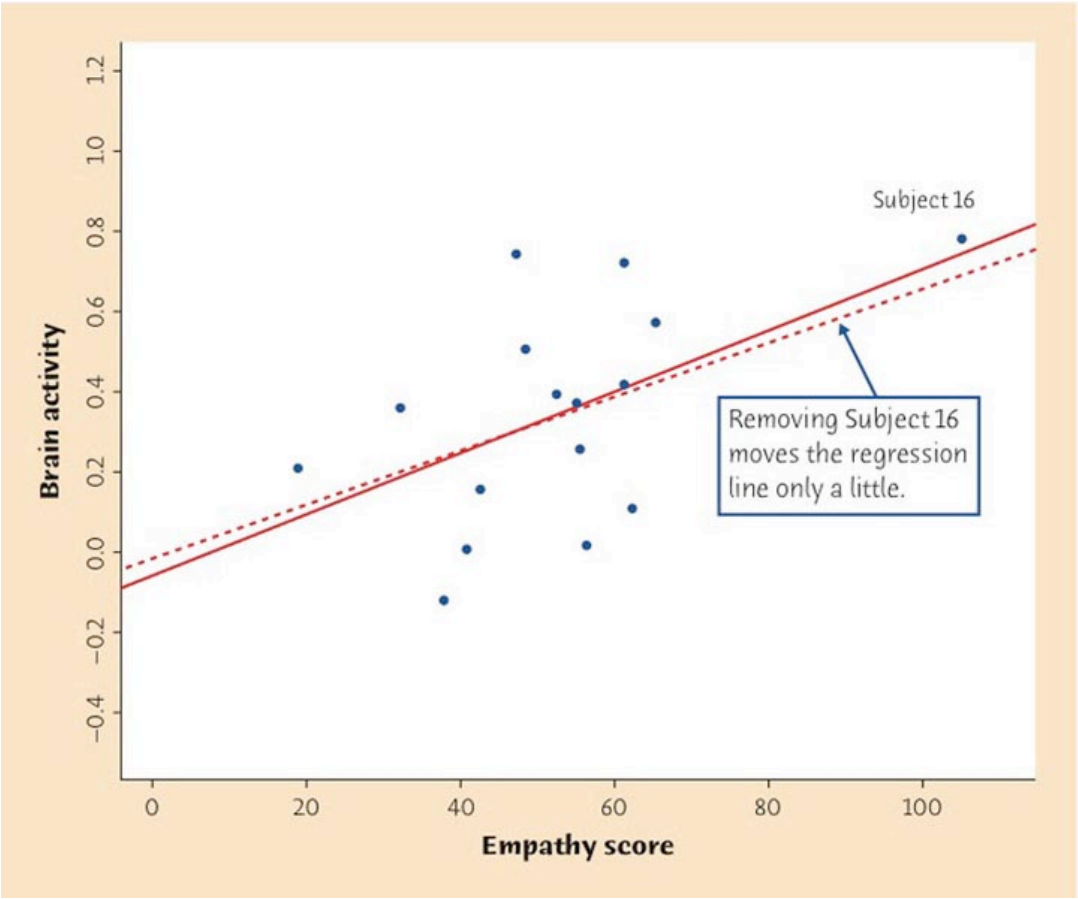
Outliers are often (but not always) influential.

Example: Psych experiment to measure if empathy is physiological.

Explanatory: Empathy score for woman
Response: Brain activity when husband shocked with electrode

(data page 99)


Brain activity
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
20
40
60
80
100
Empathy score
Subject 16
Subject 1
This is the residual for Subject 1.

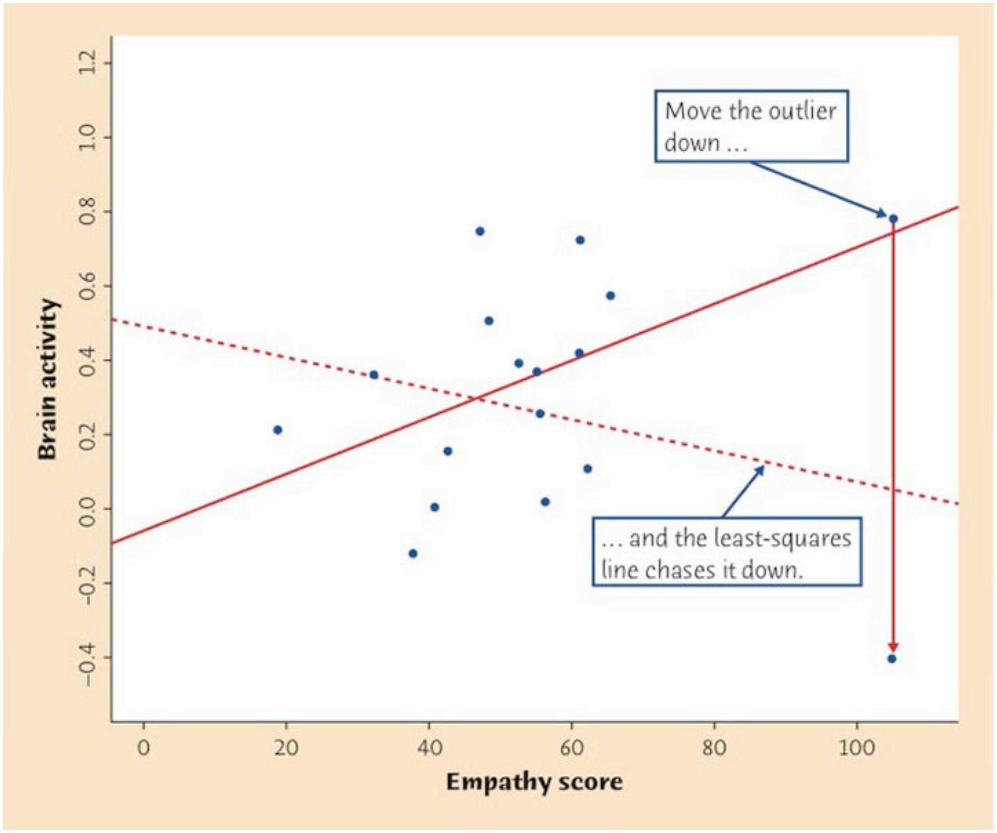


- Subject 16 is not influential for the regression line.
- Subject 16 is influential for r

  r all points = 0.515

  r without Subject 16 = 0.331

If the outlier is moved down ...



... then Subject 16 is also influential for the regression line.

Cautions about correlation/linear regression:

- Correlation and regression lines only describe linear relationships, not curved ones.

- Influential observations may lead to misleading conclusions.

- Beware extrapolation: using the regression line for predictions outside the range where the answers are reasonable.

  Example: for a child, between ages 3-10:
  explanatory: age
  response: height
  strong linear correlation, slope b ≈ 2.5 inches per year

  "Prediction" x = 30 years old → height over 75 feet tall!

- Beware the lurking variable:

Examples of correlations:

a) for countries: t.v. sets per household / life expectancy
b) for students: musical study / grades in math class
c) for countries: wine consumption / heart disease rates
d) for cars: time since last oil change / gas mileage

In all these cases, other variables help explain the correlation:

a) wealth of the country
b) family background
c) diet; wealth of the country
d) ?

Because of lurking variables ...

Correlation does not imply causation.

Examples of correlations:

a) for countries: t.v. sets per household / life expectancy
b) for students: musical study / grades in math class
c) for countries: wine consumption / heart disease rates
d) for cars: time since last oil change / gas mileage
e) for people: cigarette smoking / lung cancer
f) for the planet: atmospheric CO_2 levels / Earth's temperature

How can you figure out when there is causation?
Answer: experiments!