Correlation and Regression

MOS5e chapter 3

In a experiment, 16 student volunteers at the Ohio State University drank a randomly assigned number of cans of beer. Thirty minutes later, a police officer measured their blood alcohol content (BAC, g/dl).

Here we have two quantitative variables for each of 16 subjects.

1. How many beers they drank, and
2. Their blood alcohol level (BAC)

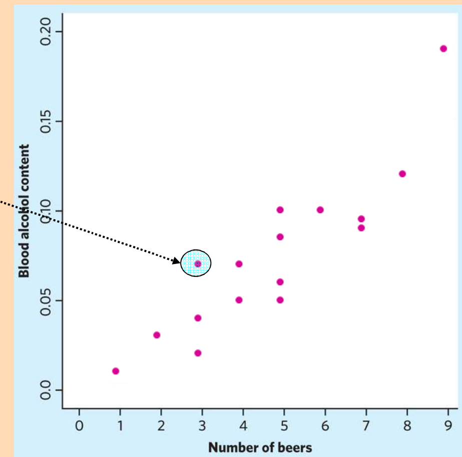
We are interested in **the relationship between the two variables**: How is one affected by changes in the other one?

Subject	**Number of Beers**	**BAC**
1	**5**	**0.1**
2	**2**	**0.03**
3	**9**	**0.19**
4	**7**	**0.095**
5	**3**	**0.07**
6	**3**	**0.02**
7	**4**	**0.07**
8	**5**	**0.085**
9	**8**	**0.12**
10	**3**	**0.04**
11	**5**	**0.06**
12	**5**	**0.05**
13	**6**	**0.1**
14	**7**	**0.09**
15	**1**	**0.01**
16	**4**	**0.05**

Scatterplots

In a **scatterplot** one axis is used to represent each of the variables, and the data are plotted as points on the graph. A scatterplot should be roughly square and not leave a lot of extra space on the edges.

Subject	Beers	BAC
1	5	0.1
2	2	0.03
3	9	0.19
4	7	0.095
5	3	0.07
6	3	0.02
7	4	0.07
8	5	0.085
9	8	0.12
10	3	0.04
11	5	0.06
12	5	0.05
13	6	0.1
14	7	0.09
15	1	0.01
16	4	0.05



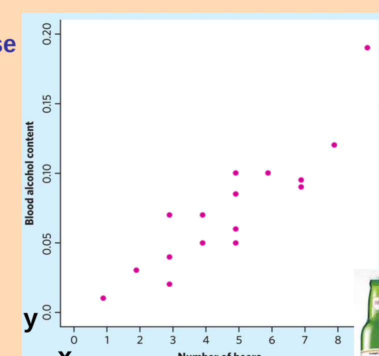
Explanatory and response variables

A **response (or dependent) variable:**

An **explanatory (or independent) variable**:

When there is an obvious explanatory variable, it is plotted on the *x* (horizontal) axis.

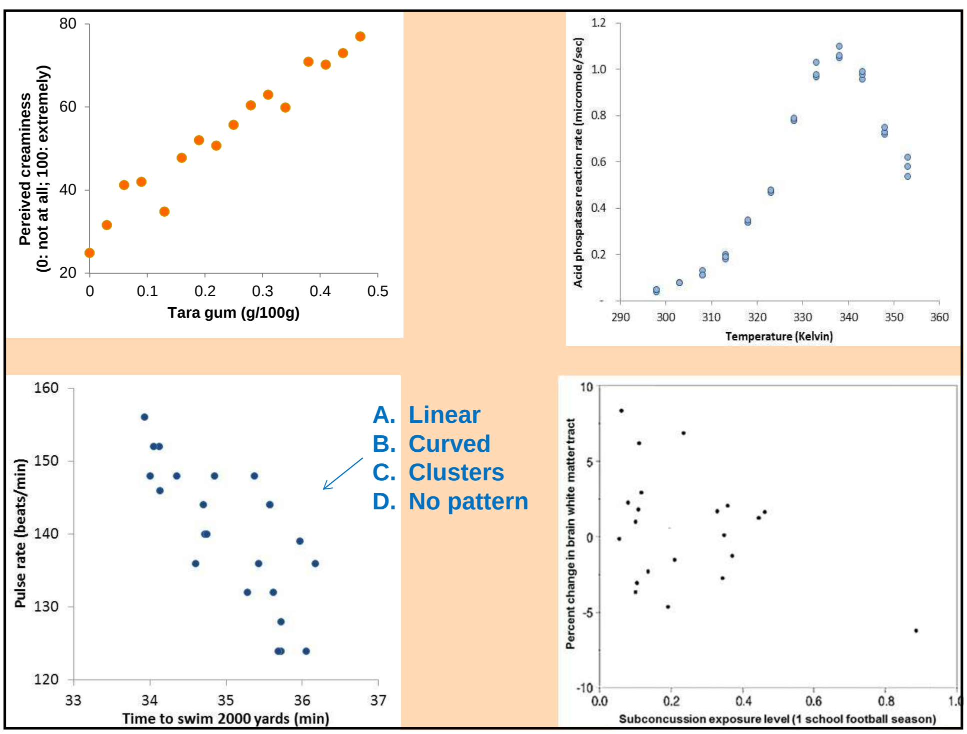
Response *BAC*



Explanatory *number of beers*

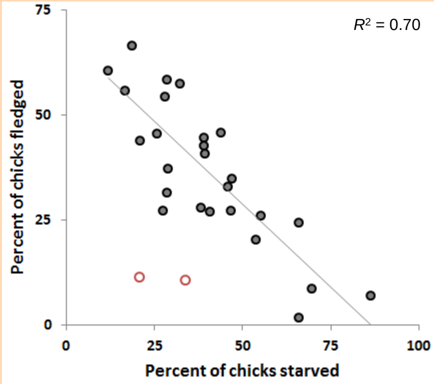
Interpreting scatterplots

- We look for an overall pattern in the points to describe the association between the two quantitative variables:
  - **Form**: linear, curved, clusters, no pattern
  - **Direction**: for linear patterns, positive (upward) or negative (downward)
  - **Strength**: weak (lots of scatter) to strong (points closely fit the form)
- … and deviations from that pattern:
  - **Outliers:** points that fall outside of the overall pattern of the relationship



Long-term study of Magellanic penguins at Punta Tombo, Argentina (1983-2010).

Climate Change Increases Reproductive Failure in Magellanic Penguins (2014)
doi:10.1371/journal.pone.0085602.g003



Punta Tombo is arid with low annual precipitation. The 2 open circles represent 1991 and 1999, when rain killed over 40% of chicks each year, and were not included in the regression.

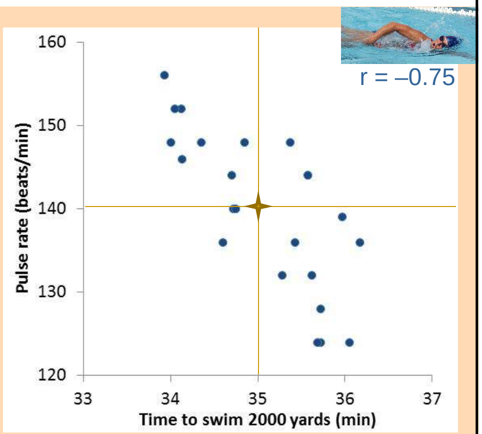
The linear correlation coefficient, r

The **linear correlation coefficient** is a measure of the **direction and strength** of a relationship.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Time to swim: $\bar{x} = 35$, $s_x = 0.7$

Pulse rate: $\bar{y} = 140$ $s_y = 9.5$



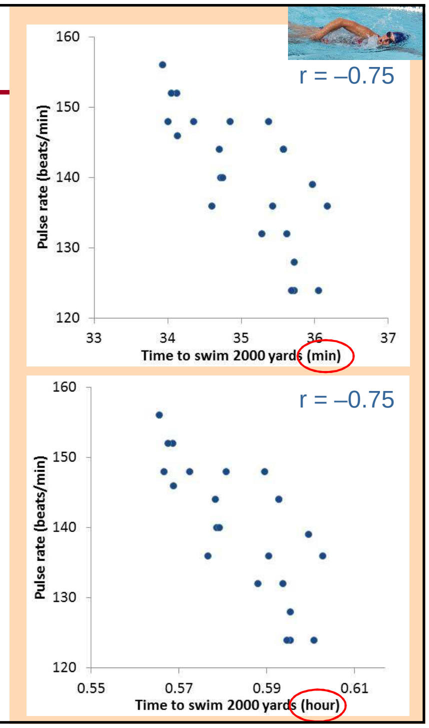
r has no unit

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

standardized value of x (unitless)

standardized value of y (unitless)

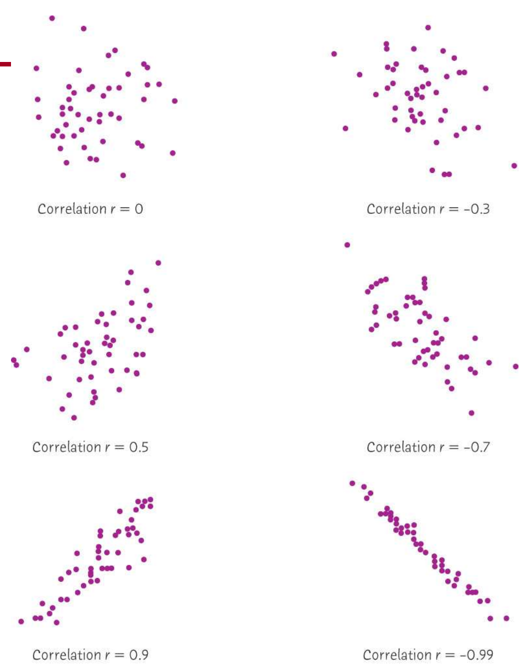
Note that this calculation isn't affected by which of the two variables in used for the x axis.



r ranges from −1 to +1

Strength is indicated by the absolute value of r

Direction is indicated by the sign of r (+ or –)

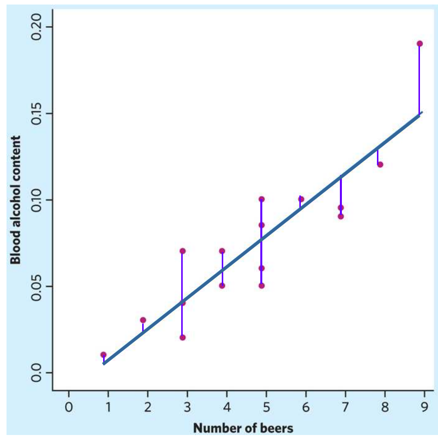


The least-squares regression line

The **least-squares regression line** is the unique line such that the sum of the squared **vertical distances** (**residuals**) between the data points and the line is the smallest possible.

sample data = model + <u>residuals</u>

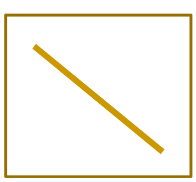
residual = actual – predicted
= $y - \hat{y}$



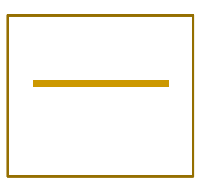
Notation

$\hat{y}$ is the predicted *y* value on the regression line

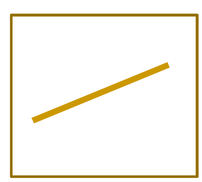
$$\boxed{\hat{y} = \text{intercept } + \text{slope } x} \qquad \boxed{\hat{y} = b_0 + b_1 x}$$



slope < 0



slope = 0



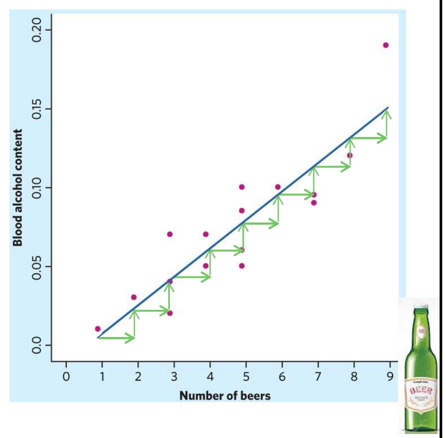
slope > 0

Not all calculators/software use this convention. Other notations include:

$$\hat{y} = ax + b$$
$$\hat{y} = a + bx$$
$$\hat{y} = \text{variable\_name } x + \text{constant}$$

Interpretation

The **slope** of the regression line describes how much we expect *y* to change, on average, for every unit change in *x*.



The **intercept** is a necessary mathematical descriptor of the regression line. It does not describe a specific property of the data.

Mathematical properties

The **slope of the regression line, *b*,** equals: $$b_1 = r\frac{s_y}{s_x}$$

r is the correlation coefficient between x and y
s_y is the standard deviation of the response variable y
s_x is the standard deviation of the explanatory variable x

The **intercept, *a*,** equals: $$b_0 = \bar{y} - b_1\bar{x}$$

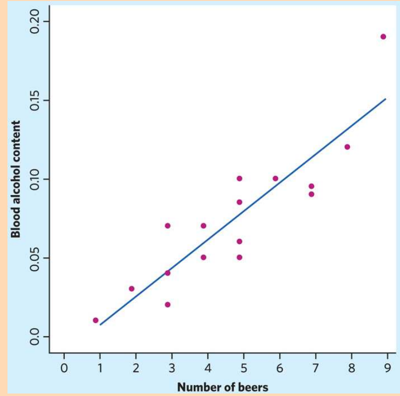
$\bar{x}$ and $\bar{y}$ are the respective means of the x and y variables

y or $\hat{y} = 0.01796x - 0.013$

$\text{BAC} = 0.01796 \text{ Beers} - 0.013$

```
Predictor       Coef    SE Coef       T       P
Constant    -0.01270    0.01264   -1.00   0.332
Beers       0.017964   0.002402    7.48   0.000

S = 0.0204410   R-Sq = 80.0%   R-Sq(adj) = 78.6%
```



SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.894338148 ← *r*			
R Square	0.799840723			
Adjusted R Square	0.785543632			
Standard Error	0.020440951			
Observations	16			
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-0.012700604	0.012637502	-1.004993204	0.331955132
Number of Beers	0.017963762	0.002401703	7.479592058	2.96948E-06

The slope is:

A) 0.01796
B) –0.0127
C) 0.8943
D) 0.0204

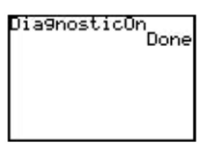
TI calculator : linear regression / correlation

First, you need to set up the regression function in your calculator.

This must be done ONCE only. No need to repeat next time.

In order to compute the correlation coefficient r between paired data of quantitative variables, we first must make sure that the calculator's diagnostics are turned on. To turn on the setting, press [CATALOG] (i.e., [2nd] **0**) and scroll down to the **DiagnosticOn** command. Press [ENTER] to bring the command to the Home screen, then press [ENTER] again.

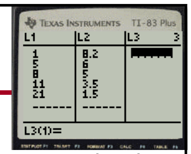
[CATALOG] ([2nd] **0**).



Press [ENTER].

Now if paired data is entered into lists, then we can find the correlation with the **LinReg(ax+b)** or **LinReg(a+bx)** commands from the [STAT] CALC screen.

TI calculator : linear regression / correlation

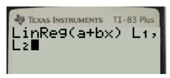


Enter the data into 2 lists

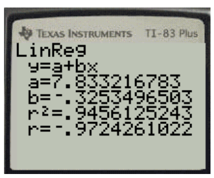
STAT **/ CALC** then

Use either **LinReg(ax+b)** or **LinReg(a+bx)**

then enter: **L1, L2**



Select the list for x values first, then the list for y values

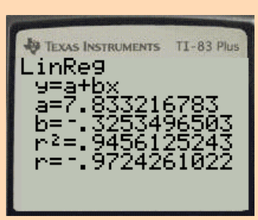


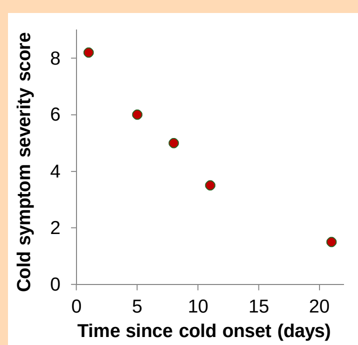
→ in this case, a is the intercept and b is the slope
→ r is the linear correlation coefficient

Researchers examined the relationship between the time since the onset of a cold and the severity of cold symptoms, based on a review of medical literature. Here are the findings:

days	severity
1	8.2
5	6
8	5
11	3.5
21	1.5

Describe the relationship.
Give an appropriate mathematical model.





Least-squares regression is only for linear associations

Don’t compute the regression line until you have confirmed that there is a linear relationship between x and y.

ALWAYS PLOT THE RAW DATA

These data sets all give a linear regression equation of about $\hat{y} = 3 + 0.5x$.

But don’t report that until you have plotted the data.

Data Set A

x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68

Data Set B

x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

Data Set C

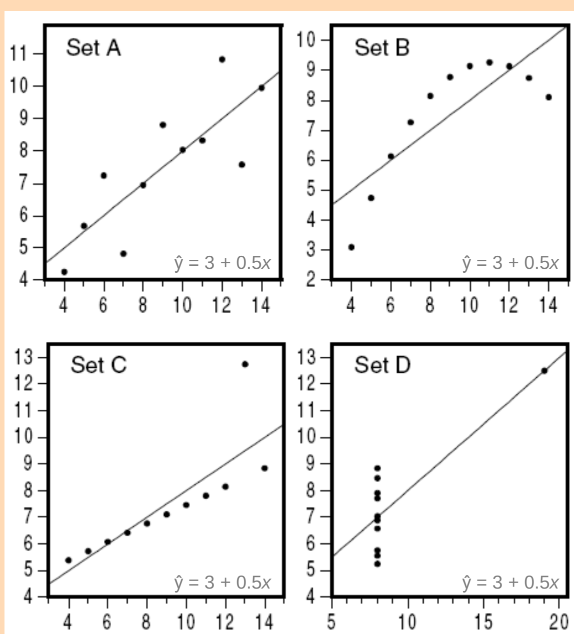
x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Data Set D

x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

Source: Frank J. Anscombe, "Graphs in statistical analysis," The American Statistician, 27 (1973), pp. 17–21.

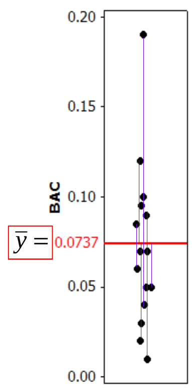
Which data set is best suited for a regression analysis? **A**, **B**, **C**, or **D**?



```
Variable  Count     Mean   Variance
BAC (y)      16  0.07375     0.0019
```

$$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

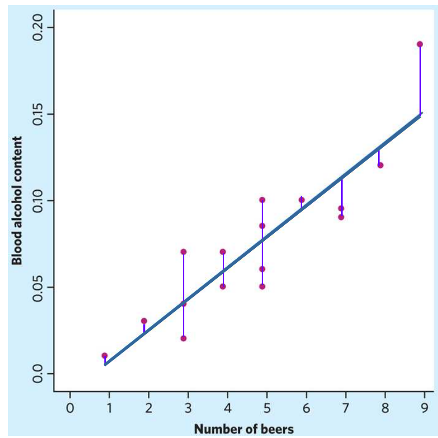
$$s_y^{\,2} = \frac{SSTO}{n-1}$$



residuals = $y_i - \hat{y}$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

The least-squares regression line is by definition the line with the smallest sum of squared residuals.



The coefficient of determination, r^2

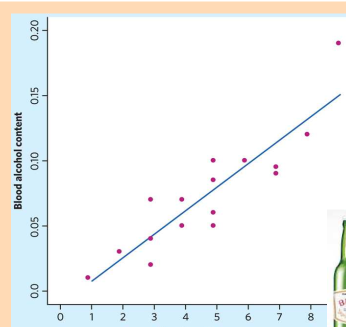
r^2, the **coefficient of determination**, is the squared value of the correlation coefficient.

$$r^2 = \frac{SSTO - SSE}{SSTO}$$

r^2 represents **the fraction of the variance in y that can be explained by the regression model.**

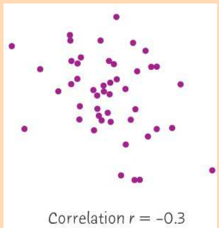
Regression Statistics	
Multiple R	0.894338148
R Square	0.799840723

This linear regression model explains 80% of individual variations in BAC.

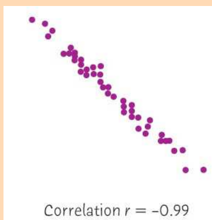


r represents the direction and strength a of a linear relationship

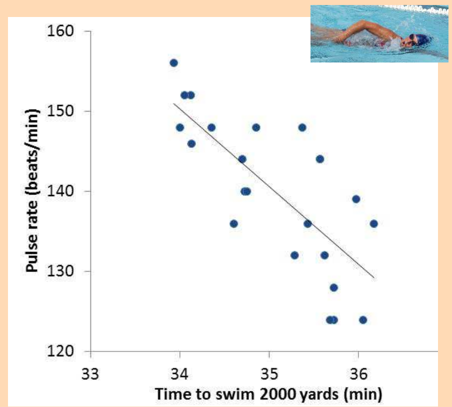
r^2 indicates what fraction of the variation in *y* can be explained by the linear regression model



$r^2 = 0.09$



$r^2 = 0.9801$



$r = -0.746$

$r^2 = 0.556$

Which is true of the slope of the least-squares regression line?

A) It has the same sign as the correlation.

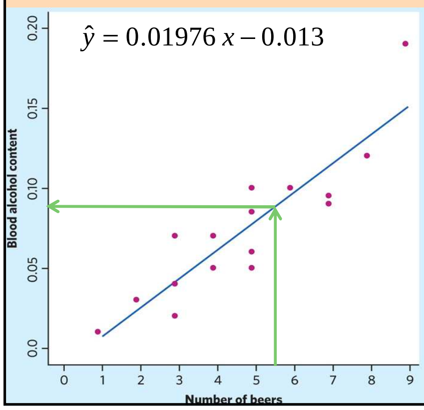
B) The square of the slope equals the fraction of the variation in the response variable that is explained by the explanatory variable.

C) It is unitless.

Making predictions

Use the equation of the least-squares regression to **predict** *y* for any value of *x* **within the range studied**.

Predication outside the range is extrapolation. ***Avoid extrapolation***.



What would we expect for the BAC after drinking 5.5 beers?

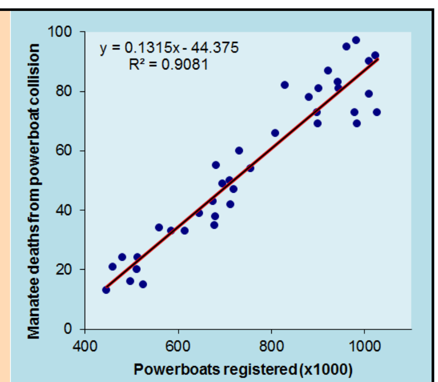
$$\hat{y} = 0.01796x - 0.013$$

$$\hat{y} = 0.01796(5.5) - 0.013 \approx 0.086\,\text{mg/ml}$$

Data for 1977-2014

Positive linear relationship

$\hat{y} = 0.1315x - 44.4$



If Florida were to limit the number of powerboats to 500,000, what could we expect the number of manatee deaths to be in that year?

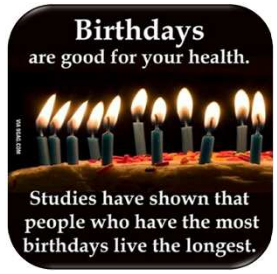
A) ~21 B) ~ 65 C) ~109 D) ~65,006

What if Florida were to limit the number of powerboats to 200,000?

Association does not imply causation

Association, however strong, does NOT necessarily imply causation.

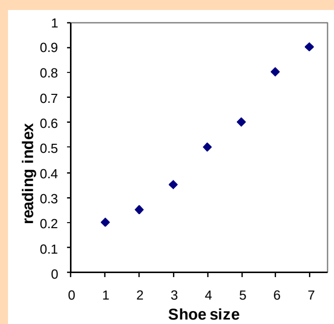
The observed association could have an external cause or be a coincidence.



- A **lurking variable** is a variable that is not among the explanatory or response variables in a study, and yet may influence the relationship between the variables studied.
- We say that two variables are **confounded** when their effects on a response variable cannot be distinguished from each other.

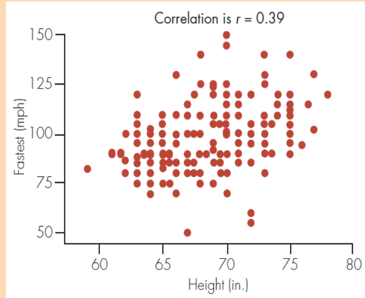
What is most likely the lurking variable, if any, in each case?

Strong positive association between the shoe size and reading skills in young children.

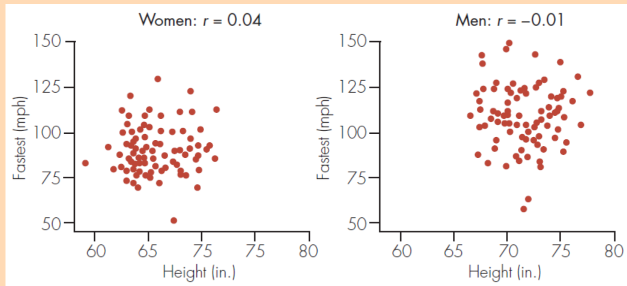


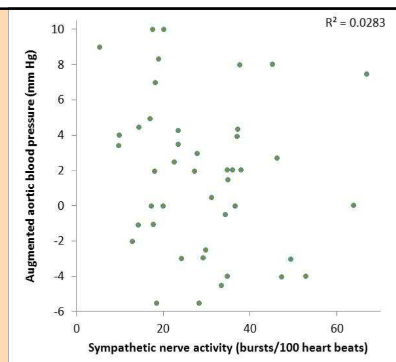
Negative association between moderate amounts of wine-drinking and death rates from heart disease in developed nations.

College student heights and responses to the question "What is the fastest you have ever driven a car?" Looks like a mild positive linear relationship.

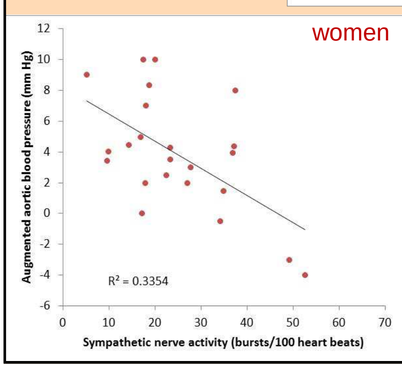


Same data broken down by sex.

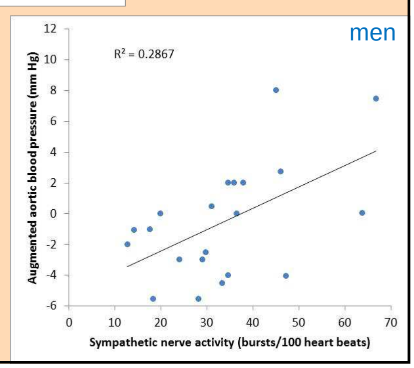




Relationship between muscle sympathetic nerve activity and a measure of arterial stiffness in young adults.



Same data broken down by sex.



Establishing causation

Establishing causation from an observed association can be done if:

1) The association is strong.
2) The association is consistent.
3) Higher doses are associated with stronger responses.
4) The alleged cause precedes the effect.
5) The alleged cause is plausible.

Lung cancer is clearly associated with smoking.

What if a genetic mutation (lurking variable) caused people to both get lung cancer <u>and</u> become addicted to smoking?

It took years of research and accumulated indirect evidence to reach the conclusion that smoking causes lung cancer.

<u>Other observed associations with an established conclusion of causality</u>

- Second-hand smoking cause of lung cancer, heart disease, etc.
- Man-made activity source of increased lead pollution and cause of neurodevelopmental damage
- Zika virus infection during pregnancy and microcephaly in newborn (WHO declaration 2016)
  who.int/emergencies/zika-virus/situation-report/31-march-2016/en/

<u>Observed associations with a causal component still hotly argued</u>

- Consumption of added sugar and obesity / metabolic syndrome
- Man-made activity and global climate change
- Concussions and depression / CTE (chronic traumatic encephalopathy)
  www.nytimes.com/2016/03/25/sports/football/nfl-concussion-research-tobacco.html

<u>Completely debunked causal association</u>

- Vaccines do NOT cause autism – fraudulent study www.bmj.com/content/342/bmj.c5347.full