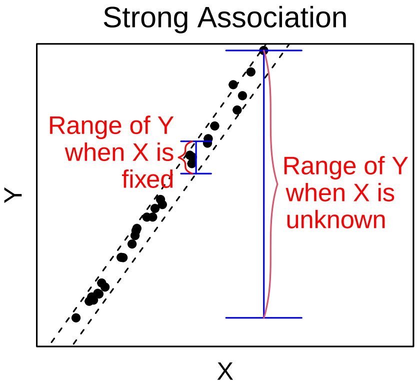
STAT 220 Lecture 22
Correlation

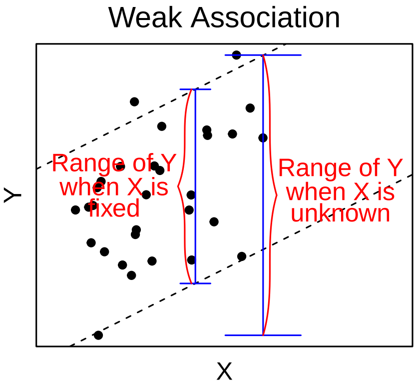
Section 8.1.4 in OpenIntro Statistics (4ed)

Yibi Huang
Department of Statistics
University of Chicago

Recall in Chapter 1, we assessed the **strength** of the association between two variables by eyeballs.



Small spread of Y
when X is known

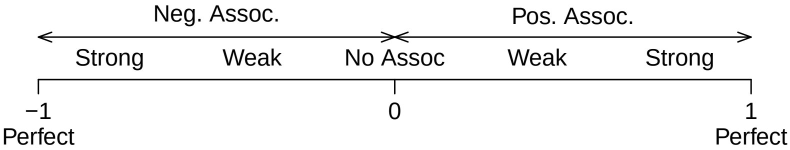


Large spread of Y
when X is known

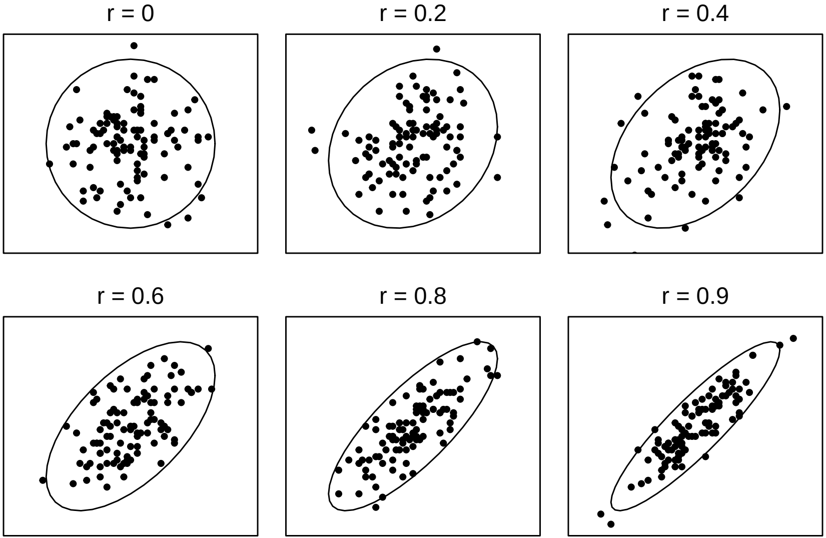
Correlation = Correlation Coefficient, r

Correlation r is a numerical measure of the **direction** and **strength** of the **linear** relationship between two numerical variables.

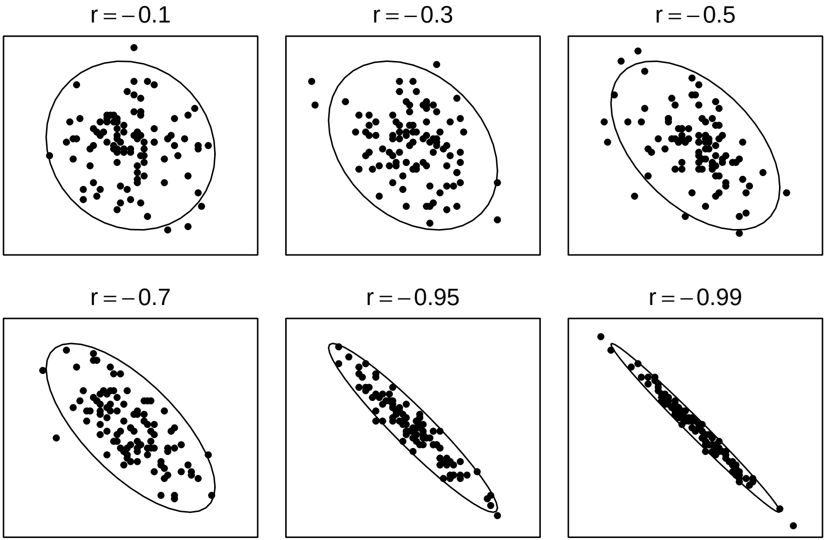
- $-1 \leq r \leq 1$ always
- Sign of r — direction of association
  - $r > 0$: positive association
  - $r < 0$: negative association
- Magnitude of r — strength of association
  - $r \approx 0$: very weak linear relation
  - large $|r|$: strong linear relationship
  - $r = -1$ or $r = 1$: *only* when all the data points on the scatterplot lie exactly along a *straight line*



Positive Correlations



Negative Correlations



Formula for Computing the Correlation Coefficient "r"

$$(x_1, y_1)$$
$$(x_2, y_2)$$
$$(x_3, y_3)$$
$$\vdots$$
$$(x_n, y_n)$$

The **correlation coefficient** r
(or simply, **correlation**) is defined as:

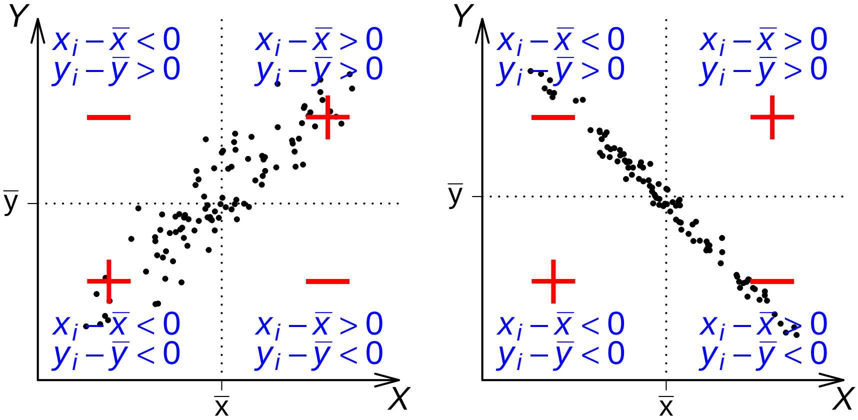
$$r = \frac{1}{n-1} \sum_{i=1}^{n} \underbrace{\left(\frac{x_i - \bar{x}}{s_x}\right)}_{z\text{-score of } x_i} \times \underbrace{\left(\frac{y_i - \bar{y}}{s_y}\right)}_{z\text{-score of } y_i}.$$

where s_x and s_y are respectively the sample SD of X and of Y.

Usually, we find the correlation using software rather than by manual computation.

Why r Measures the Direction of a Linear Relation?

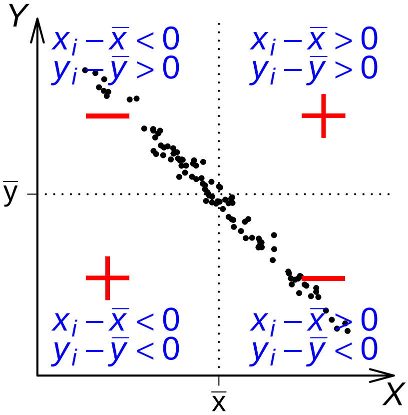
What is the sign of $(x_i - \bar{x})(y_i - \bar{y})$?



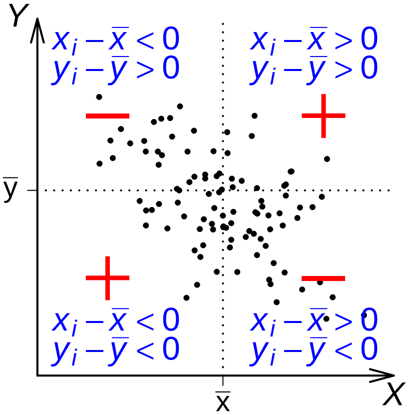
r > 0 as most points have $(x_i - \bar{x})(y_i - \bar{y}) > 0$

r < 0 as most points have $(x_i - \bar{x})(y_i - \bar{y}) < 0$

Why r Measures the Strength of a Linear Relation?



r Has a Larger Magnitude



r Has a Smaller Magnitude

Correlation r Has No Unit

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \underbrace{\left(\frac{x_i - \bar{x}}{s_x} \right)}_{z\text{-score of } x_i} \times \underbrace{\left(\frac{y_i - \bar{y}}{s_y} \right)}_{z\text{-score of } y_i}.$$

After standardization, the z-score of neither x_i nor y_i has a unit.

- So r is unit-free.
- So we can compare r between data sets, where variables are measured in different units or when variables are different. E.g. we may compare the

  $$r \text{ between [swim time and pulse]},$$

  with the

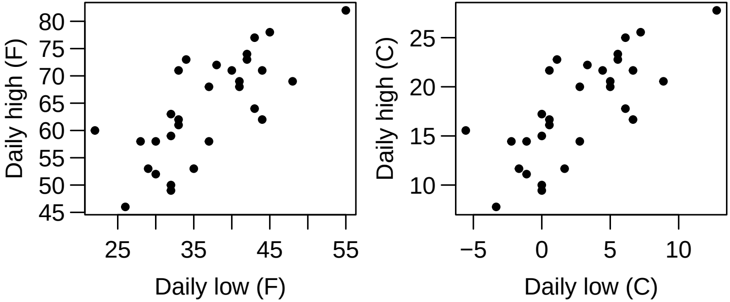
  $$r \text{ between [swim time and breathing rate]}.$$

Correlation r Has No Unit (2)

Changing the units of variables does not change the correlation coefficient r, because we get rid of all the units when we standardize them (get z-scores).

E.g., correlations are not affected by whether the temperatures are recorded in $^\circ F$ or $^\circ C$ since

$$C = \frac{5}{9}(F - 32).$$

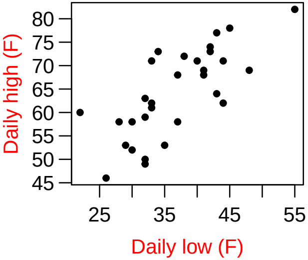


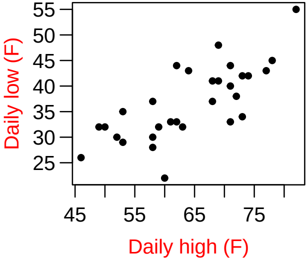
Correlation Does Not Distinguish X & Y

Sometimes one uses the X variable to predict the Y variable. In this case, X is called the *explanatory variable*, and Y the *response*. The correlation coefficient r does not distinguish between the two. It treats x and y symmetrically.

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

Swapping the x-, y-axes doesn't change r (both $r = 0.74$.)

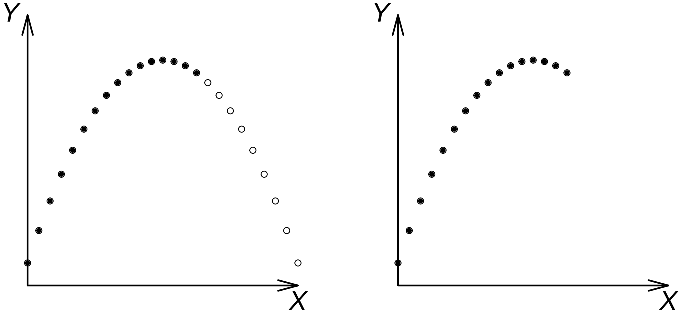




Correlation Doesn't Reflect Strength of Nonlinear Relations

Both scatter plots below show perfect nonlinear relations.

All points fall on the curve $y = 2 - x^2/2$.



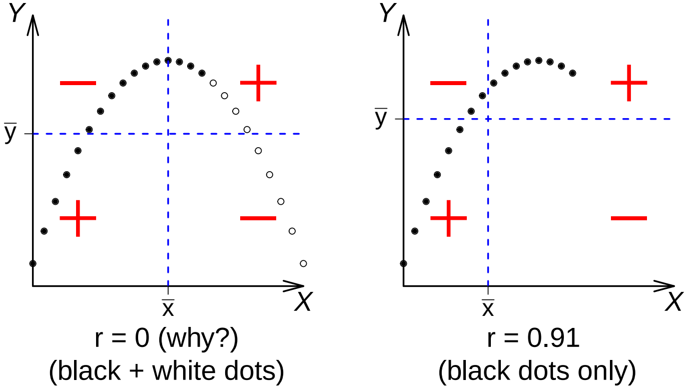
r = 0 (why?)
(black + white dots)

r = 0.91
(black dots only)

Correlation Doesn't Reflect Strength of Nonlinear Relations

Both scatter plots below show perfect nonlinear relations.

All points fall on the curve $y = 2 - x^2/2$.

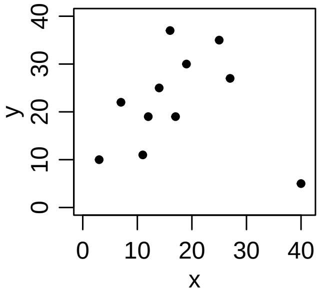


No matter how strong the relation, r is never 1 or -1 if it's nonlinear.

r can even be 0, like the plot above.

Correlation Is VERY Sensitive to Outliers

A single outlier can change r drastically.

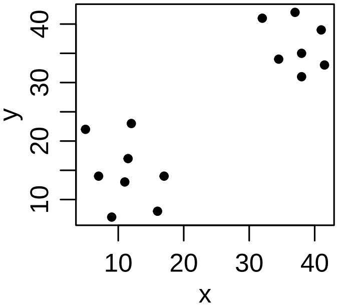


R command to calculate correlation: `cor()`.

For the left plot, the correlation is ...

```
x = c(3 ,11,40,17,12,7,14,27,19,25,16)
y = c(10,11,5,19,19,22,25,27,30,35,37)
cor(x,y)            # including the outlier
[1] 0.004447
cor(x[-3], y[-3])  # excluding the outlier
[1] 0.6912
```

When Data Points Are Clustered ...



Each of the two clusters in the left plot shows a weak negative association ($r = -0.336$ and -0.323).
But the whole diagram shows a moderately strong positive association ($r = 0.849$).

- This is an example of the **Simpson's paradox**.
- An overall r can be misleading when data points are clustered.
- Cluster-wise r's should be reported as well.

Example: Auto Data

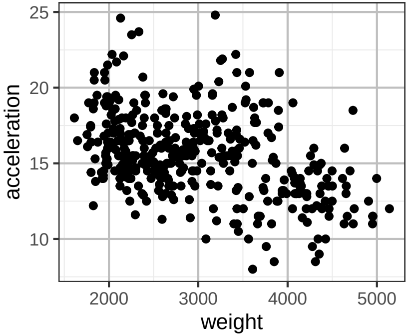
Auto is a data set containing information about 392 car models in the 1980s. The variables include

- `acceleration`: Time to accelerate from 0 to 60 mph (in seconds)
- `horsepower`: Engine horsepower
- `weight`: Vehicle weight (lbs.)

You can load the data by running the command below.

```
Auto = read.table("https://www.stat.uchicago.edu/~yibi/s224/data/Auto.txt", h=T)
```

Example: Auto Data (2)

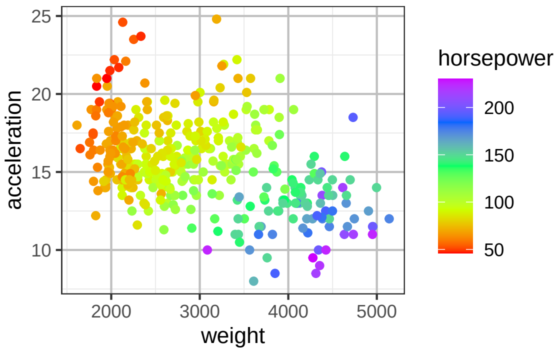



```
cor(Auto$weight, Auto$acceleration)
[1] -0.4168
```

From the scatter plot and the correlation, `weight` and `acceleration` are negatively correlated. Is that reasonable?

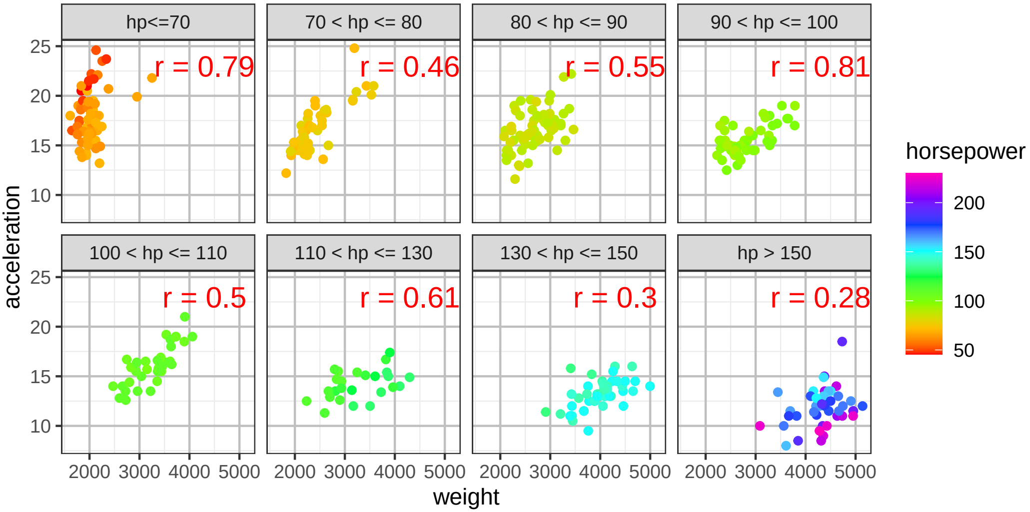
Example: Auto Data (3)

```
ggplot(Auto, aes(x=weight, y=acceleration, col=horsepower)) +
  geom_point() + scale_color_gradientn(colours = rainbow(5))
```



Consider car models of similar `horsepower`, are `weight` and `acceleration` positively or negatively correlated?

Example: Auto Data (4)

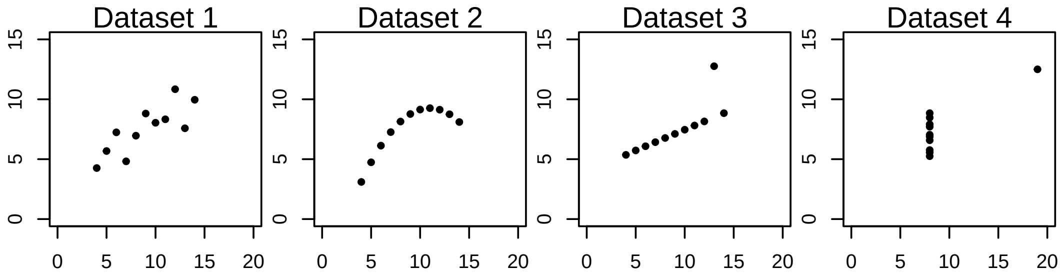


Always Check the Scatter Plots (1)

The 4 data sets below have identical $\overline{x}$, $\overline{y}$, s_x, s_y, and r.

	Dataset 1		Dataset 2		Dataset 3		Dataset 4	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.96	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.75	13	12.76	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.10	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.10	4	5.36	19	12.50
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Ave	9	7.5	9	7.5	9	7.5	9	7.5
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

Always Check the Scatter Plots (2)



- For Dataset 2, y can be predicted exactly from x. But $r < 1$, because r only measures **linear** association.
- For Dataset 3, r would be 1 instead of 0.82 if the outlier were actually on the line.

The correlation coefficient can be misleading in the presence of outliers, multiple clusters, or nonlinear association.

Correlation Indicates Association, *Not* Causation

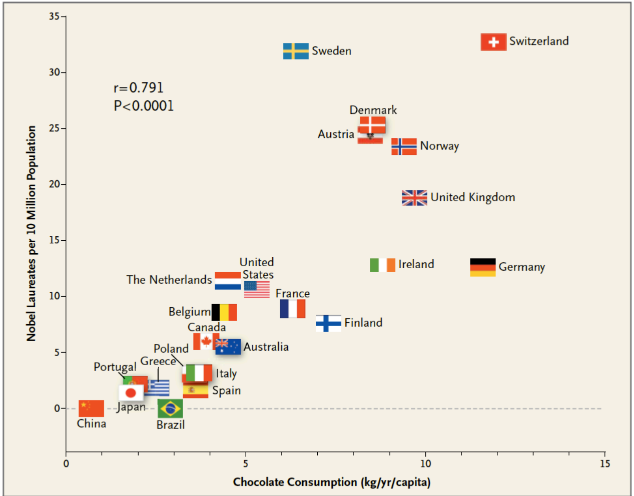


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Source: http://www.nejm.org/doi/full/10.1056/NEJMon1211064

Questions

- Why do both variables have to be numerical when computing their correlation coefficient?
- If the law requires women to marry only men 2 years older than themselves, what is the correlation of the ages between husbands and wives?