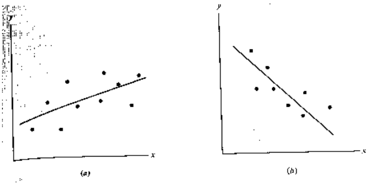
Correlation

Goal: Find cause and effect links between variables.

What can we conclude when two variables are highly **correlated**?



Positive Correlation	**Negative Correlation**
High values of x are associated with high values of y.	High values of x are associated with low values of y.

The **correlation coefficient,** R^2 is a number between 0 and 1. Values near 1 show strong correlation; values near 0 show weak correlation.

Calculating the R^2 Statistic

To calculate R^2, you need data points **AND** a best fit linear regression. Calculate:

- The **error sum of squares**: $SSE = \sum_i \left[y_i - f(x_i)\right]^2$.

⋆ SSE is the variation between the data and the function. ⋆

- The **total corrected sum of squares**: $SST = \sum_i \left[y_i - \bar{y}\right]^2$, where $\bar{y}$ is the average y_i value.

⋆ SST is the variation solely due to the data. ⋆

- Now calculate $R^2 = 1 - \frac{SSE}{SST}$.

⋆ R^2 is the proportion of variation explained by the function. ⋆

Calculating the R^2 Statistic

Example. (cont. from notes p. 29) What is R^2 for the data set:
$\{(1.0, 3.6), (2.1, 2.9), (3.5, 2.2), (4.0, 1.7)\}$?

Recall that the regression line is $f(x) = -0.605027x + 4.20332$.

- The **error sum of squares**: $SSE = \sum_i \left[y_i - f(x_i)\right]^2$.

$$\begin{aligned} SSE &= (3.6 - f(1.0))^2 + (2.9 - f(2.1))^2 + (2.2 - f(3.5))^2 + (1.7 - f(4.0))^2 \\ &= (.0017)^2 + (-0.033)^2 + (0.114)^2 + (-0.083)^2 = 0.0210 \end{aligned}$$

- The **total corrected sum of squares**: $SST = \sum_i \left[y_i - \bar{y}\right]^2$.

First, calculate $\bar{y} = (3.6 + 2.9 + 2.2 + 1.7)/4 = 2.6$

$$\begin{aligned} SST &= (3.6 - 2.6)^2 + (2.9 - 2.6)^2 + (2.2 - 2.6)^2 + (1.7 - 2.6)^2 \\ &= (1)^2 + (0.3)^2 + (-0.4)^2 + (-0.9)^2 = 2.06 \end{aligned}$$

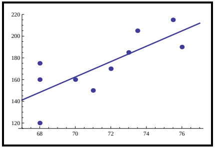
- Now calculate $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{0.0210}{2.06} = 1 - .01 = 0.99$.

Another R^2 Calculation

Example. Estimating weight from height.

To the right is a list of heights and weights for ten students. We can calculate the line of best fit:

$$(\text{weight}) = 7.07(\text{height}) - 333.$$



ht.	wt.
68	160
70	160
71	150
68	120
68	175
76	190
73.5	205
75.5	215
73	185
72	170

Now find the correlation coefficient: $(\overline{w} = 173)$

$SSE = \sum_{i=1}^{10}(w_i - [(7.07)h_i - 333])^2 \approx 2808$

$SST = \sum_{i=1}^{10}(w_i - 173)^2 = 6910$

So $R^2 = 1 - (2808/6910) = 0.59$, a good correlation.

We can do better by introducing another variable:

Multiple Linear Regression

Add waist measurements to the list:

We wish to calculate a relationship such as:

$$(\text{weight}) = a(\text{height}) + b(\text{waist}) + c.$$

Do a linear regression to find the *best-fit plane.*
Apply again the least-squares criterion. Minimize:

$$SSE = \sum_{(h_i, ws_i, wt_i)} \left[wt_i - (a \cdot h_i + b \cdot ws_i + c) \right]^2,$$

which finds that the best fit plane is (coeff sign)

$$(\text{weight}) = 4.59(\text{height}) + 6.35(\text{waist}) - 368.$$

ht.	wst.	wt.
68	34	160
70	32	160
71	31	150
68	29	120
68	34	175
76	34	190
73.5	38	205
75.5	34	215
73	36	185
72	32	170

Multiple Linear Regression

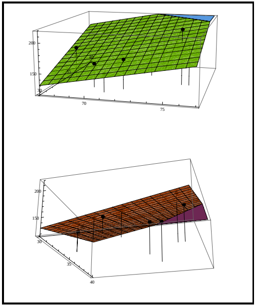
Visually, we can see that we might expect a plane to do a better job fitting the points than the line.

- Now calculate R^2.

Calculate $SSE = \sum_{i=1}^{10}(w_i - f(h_i, ws_i))^2 \approx 955$

SST does not change: (why not?)

$\sum_{i=1}^{10}(w_i - 173)^2 = 6910$



ht.	wst.	wt.
68	34	160
70	32	160
71	31	150
68	29	120
68	34	175
76	34	190
73.5	38	205
75.5	34	215
73	36	185
72	32	170

So $R^2 = 1 - (955/6910) = 0.86$, an excellent correlation.

Notes about the Correlation Coefficient

Example. Cancer and Fluoridation. (pp. 188–189)
Does fluoride in the water cause cancer?

Variables:
L = log of years of fluoridation
A = % of population over 65.
C = cancer mortality rate

Use a linear regression to find that
$C = \mathbf{27.1}L + 181$, with an $R^2 = 0.047$.

Compare to a multiple linear regression of
$C = \mathbf{0.566}L + 10.6A + 85.8$, with an $R^2 = 0.493$.

- Be suspicious of a low R^2.
- Signs of coefficients tell positive/negative correlation.
- Cannot determine relative influence of one variable in one model without some gauge on the magnitude of the data.
- Can determine relative influence of one variable in two models.

Notes about the Correlation Coefficient

Example. Time and Distance (pp. 190)
Data collected to predict driving time from home to school.

Variables:
T = driving time
M = miles driven
S = Last two digits of SSN.

Use a linear regression to find that
$T = 1.89M + 8.05$, with an $R^2 = 0.867$.

Compare to a multiple linear regression of
$T = 1.7M + 0.0872S + 13.2$, with an $R^2 = 0.883$!

- R^2 increases as the number of variables increase.
- This doesn't mean that the fit is better!

Causation

If we have high correlation, we'd like to determine causation.

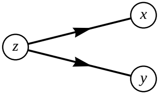
To visually represent the direction of causality between variables, use arrows. For example, if x causes y, we draw an arrow from x to y.

The ways in which two variables may have strong correlation are:

I. Simple Causality 

II. Reverse Causality 

III. Mutual Causality 

IV. Hidden/Confounding Variable 

V. Complete Accident/Coincidence 

Simple Causality

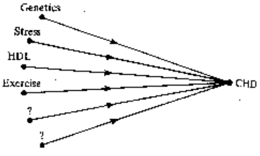
I. Simple Causality $x \longrightarrow y$

We say that variables x and y are related by **simple causality** if the level of x *determines* the level of y.

Example 2 (pp. 171–173) deals with high blood pressure. After plotting blood pressure (x) with deaths from heart disease (y), there is high correlation.

A chain of causation can be deduced that makes the argument for simple causality:

high blood pressure → arteries clog → lack of oxygen in heart → heart disease



(b)

Many factors have been determined that increase the chance for heart disease.

Reverse Causality

II. Reverse Causality $x \longleftarrow y$

We say that variables x and y are related by **reverse causality** if the level of x *is determined by* the level of y.

Example. Islanders in South Pacific determined that healthy people had body lice and sick people didn't. The islanders concluded that more body lice means better health. However, everyone had lice and lice prefer healthy hosts.

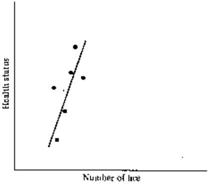


Figure 7

Example. Human birth rate and stork population: "storks bring babies".

Mutual Causality / Feedback

III. Mutual Causality $x \rightleftarrows y$

We say that variables x and y are related by **mutual causality** if changes in x produce changes in y and vice versa.

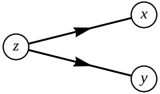
Example. Car dealers:

If you plot car sales and advertising budget for a large set of car dealers, you will likely find a strong correlation.

Do car sales pay for advertising
or does advertising drive sales?

They are mutually reinforcing,
so this is an example of mutual causality.

Hidden Variable Causes Both

IV. Hidden/Confounding Variable 

We say that x and y are in a **spurious relationship** if the levels of both x and y are determined by the level of a **confounding variable** z.

Example. In a city, the number of churches there are is highly correlated with the number of liquor stores.

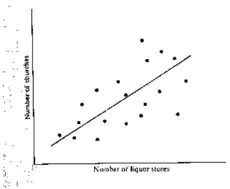


Figure 10

- Simple causation would imply:
- Reverse causation would imply:

In this instance, there is a confounding variable: ____________________________.

Complete Accident

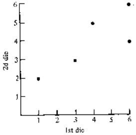
V. Complete Accident/Coincidence x y

If none of the above four cases apply, x and y are unrelated.

Take two dice. Roll each five times. Plot the value of one die versus the value of the other die for the five rolls. Often there will be no correlation.

One instance of correlation occurred, with an R^2 of 0.672 (relatively high!)

Roll	Number showing: 1st die	2d die
#1	1	2
#2	3	3
#3	4	5
#4	5	4
#5	6	6



An example of a correlation by coincidence.

Example. Perhaps with students and SSN's?

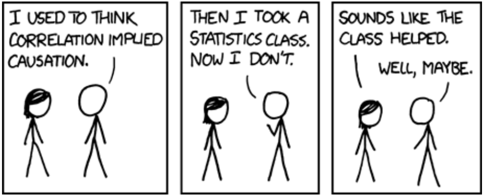
- The chance of this occurring decreases as more observations are taken.

Correlation does not imply causation!

Groupwork: Justify the correlations between the following variables:

- As ice cream sales increase, the rate of drowning deaths increase.
- The more firemen fighting the fire, the larger the fire grows.
- With fewer pirates on the open seas, global warming has increased.
- The more people in my Facebook group, the faster it grows.

What is the joke below?



Source: `http://xkcd.com/552/`