Relationships between two quantitative variables

Overview

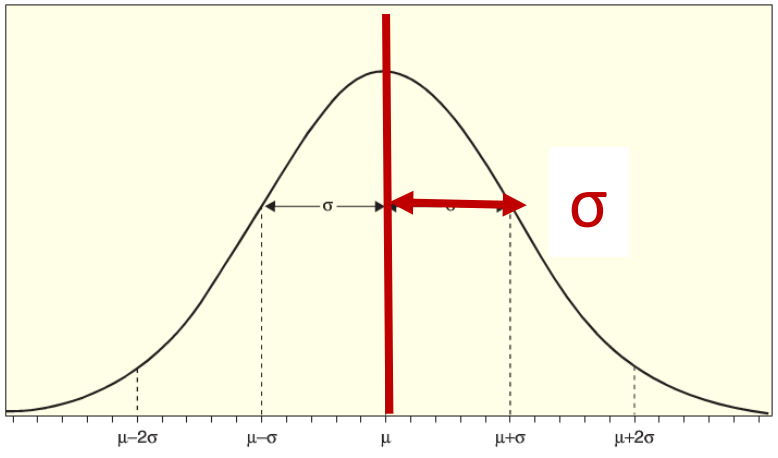
Quick review of a few concepts

Scatterplots

Correlation

Simple linear regression

μ

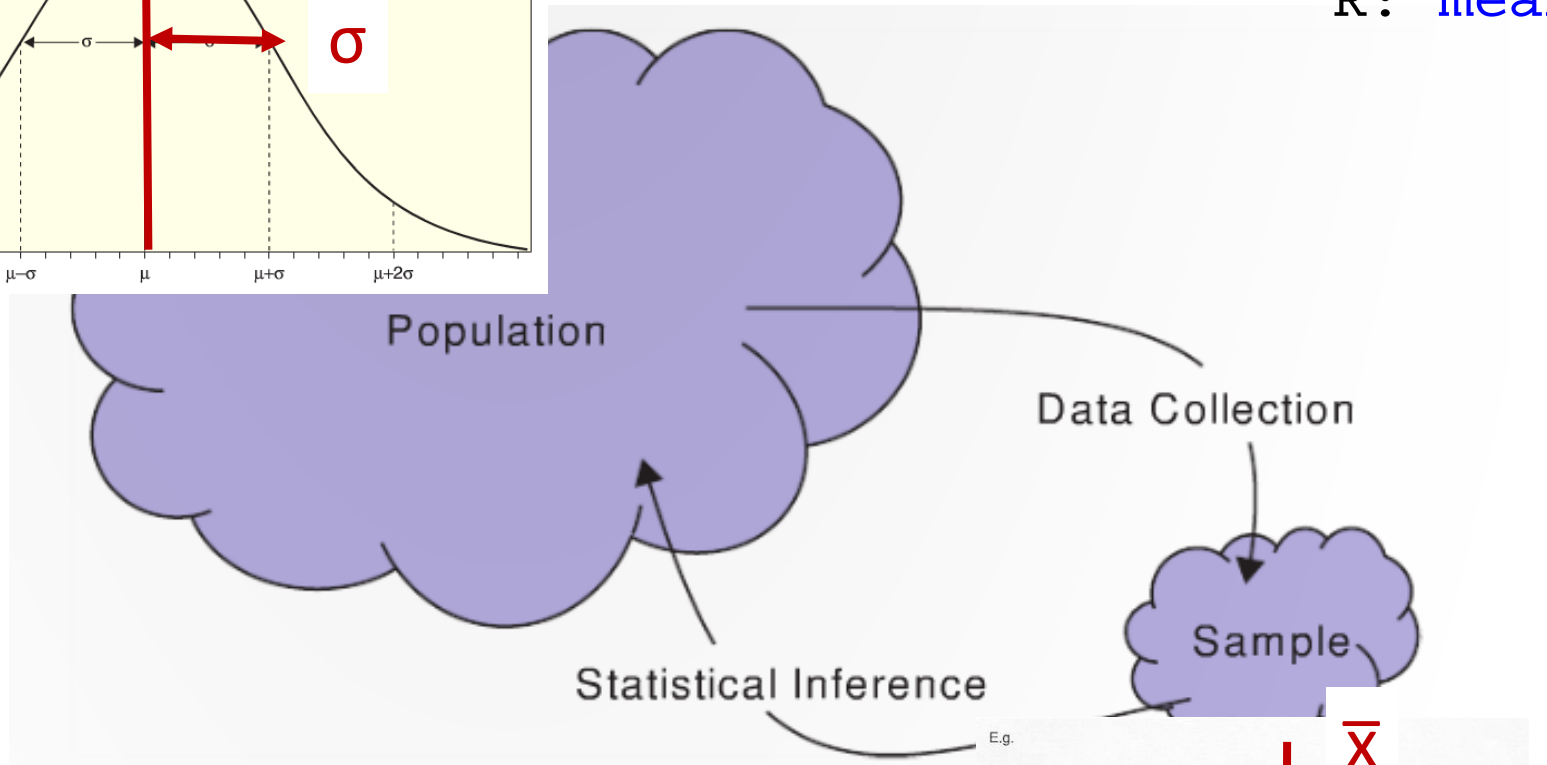


σ

Parameters

$$\bar{x} = \frac{\sum_i^n x_i}{n}$$

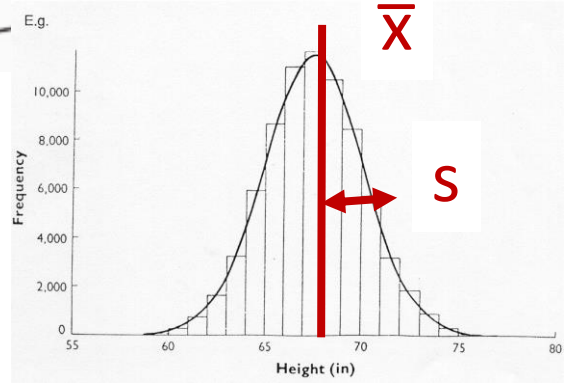
R: `mean(x)`



$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

R: `sd(x)`

Statistics



Review: z-scores

The z-scores tells how many standard deviations a value is from the mean

$$\text{z-score}(x_i) = \frac{x_i - \bar{x}}{s}$$

Which statistic is most impressive?

Z-score FGPct = 0.868

Z- score Points = 2.698

Z-score Assists = 1.965

Z-score Steals = 1.771

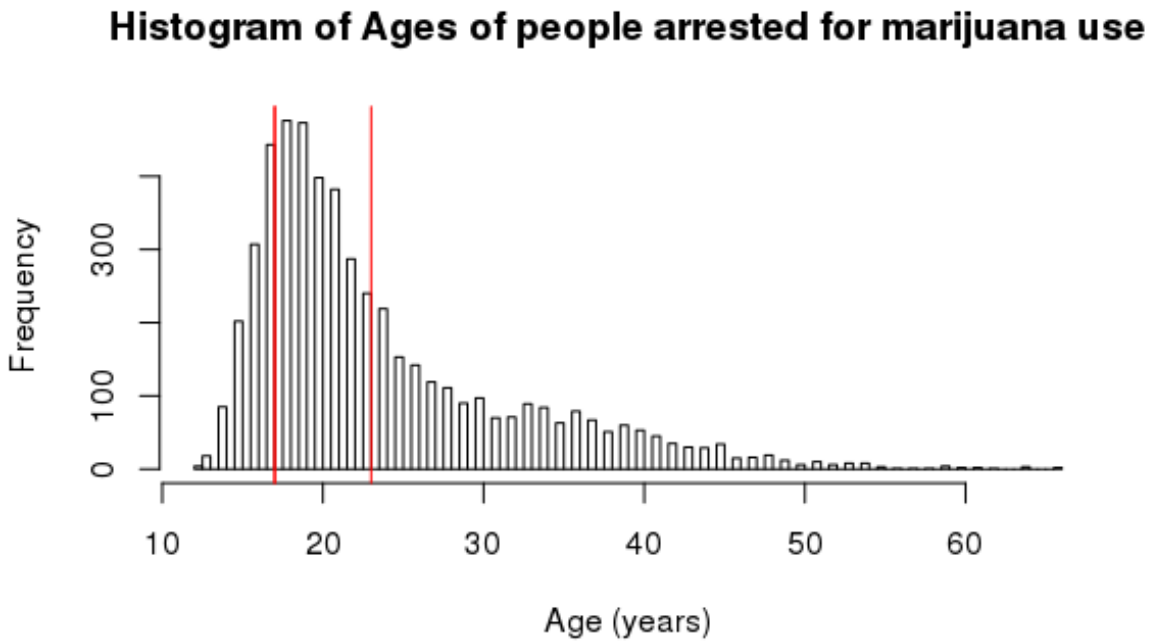
The normal pillow

Question: What percent of the pillow's mass is ± 2 standard deviations from the mean?

Answer: 95%

Review: quantiles (percentiles)

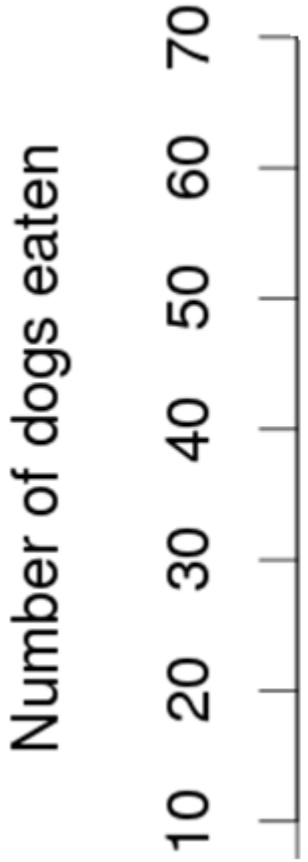
The **p^{th} percentile** is a quantitative value ***x*** which is greater than p percent of the data



60th percentile value is 23

i.e., 60% of the arrests were of ages 23 or less

Review: boxplot (5 number summary)



Maximum value that is not an outlier

Q_3 (75^{th} percentile)

Median

Q_1 (25^{th} percentile)

Minimum value that is not an outlier

Relationships between two quantitative variables

Two quantitative variables

In 1968, Joseph Fraumeni published a paper published in the Journal of the National Cancer Institute that examined the relationship between smoking and different types of cancer

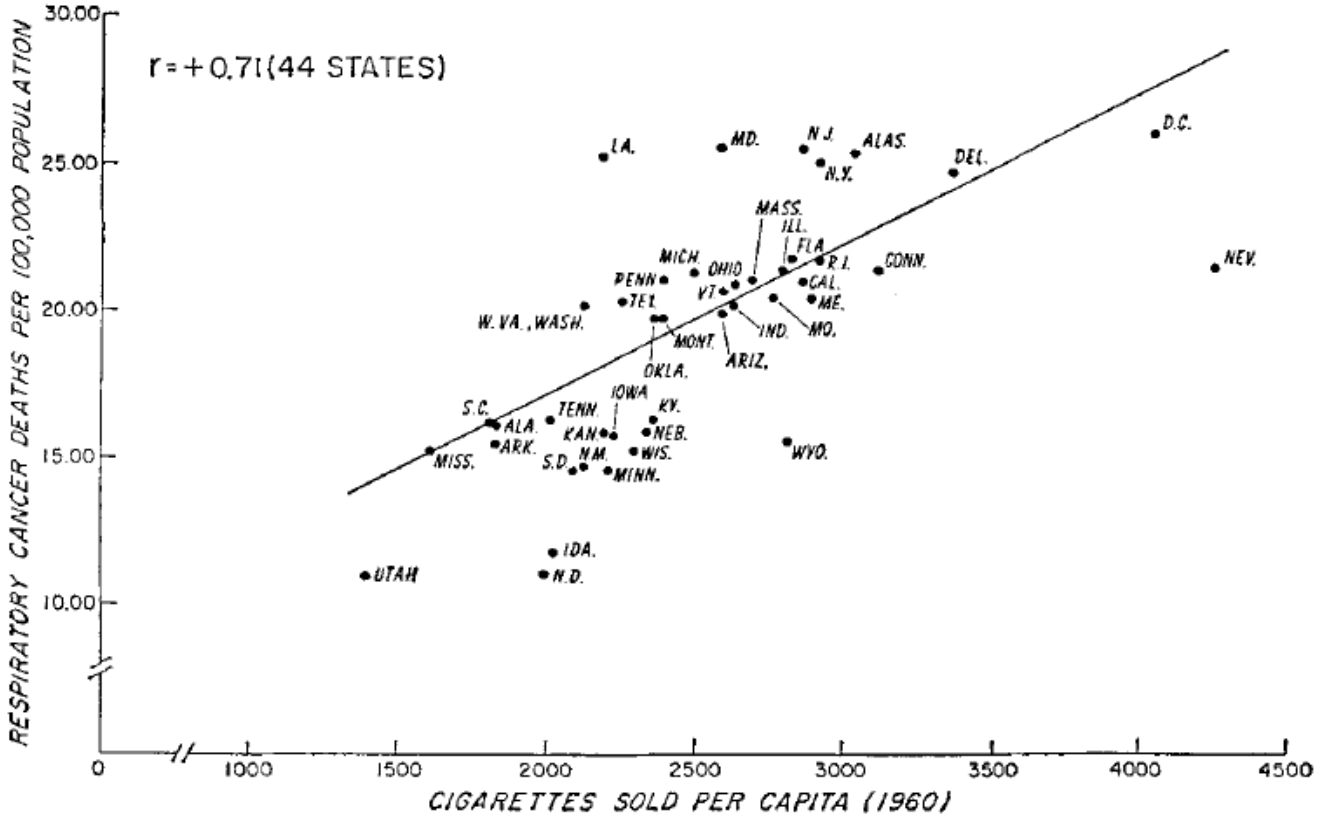
What are the…

- Cases?
- Variables?

State	Cig per capita	Bladder	Lung	Kidney	Leukemia
AL	18.2	2.9	17.05	1.59	6.15
AZ	25.82	3.52	19.8	2.75	6.61
AR	18.24	2.99	15.98	2.02	6.94
CA	28.6	4.46	22.07	2.66	7.06
CT	31.1	5.11	22.83	3.35	7.2
DE	33.6	4.78	24.55	3.36	6.45
DC	40.46	5.6	27.27	3.13	7.08

Relationship between smoking and lung cancer

TEXT-FIGURE 2.—Correlation between average annual age-adjusted death rates for respiratory tract cancer (1956–61) and *per capita* cigarette sales (1960) in 44 States.



JOURNAL OF THE NATIONAL CANCER INSTITUTE

J Fraumeni 1968

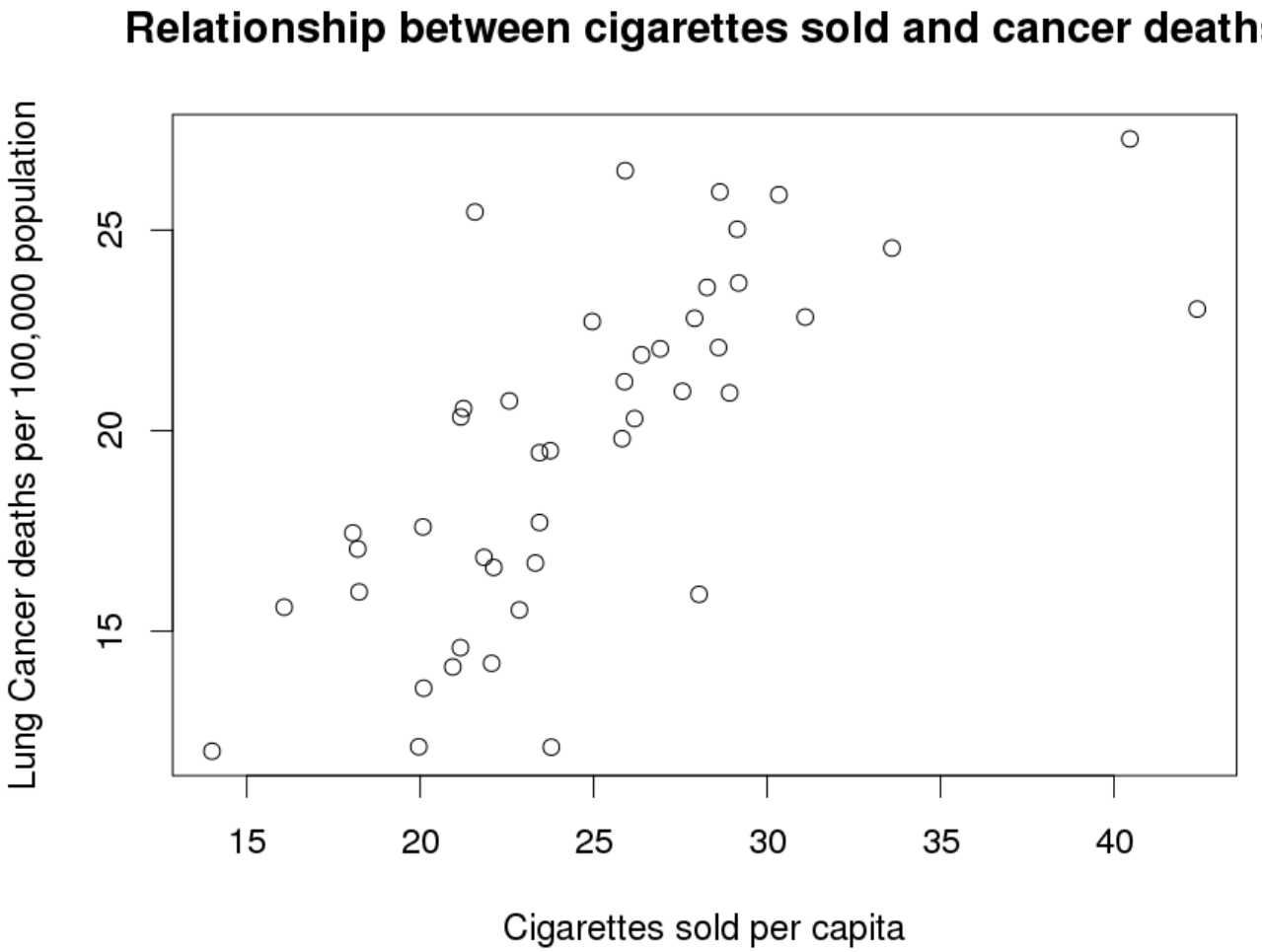
Scatterplot

A **scatterplot** graphs the relationship between two variables

- Each axis represents the value of one variables
- Each point the plot shows the value for the two variables for a single data case

If there is an explanatory and response variable, then the explanatory variable is put on the x-axis and the response variable is put on the y-axis

Relationship between smoking and lung cancer



R: `plot(x, y)`

Questions when looking at scatterplots

Do the points show a clear trend?

- Does it go upward or downward?
- How much scatter around the trend?

Does the trend seem be linear (follow a line) or is it curved?

Are there any outlier points?

Questions when looking at scatterplots

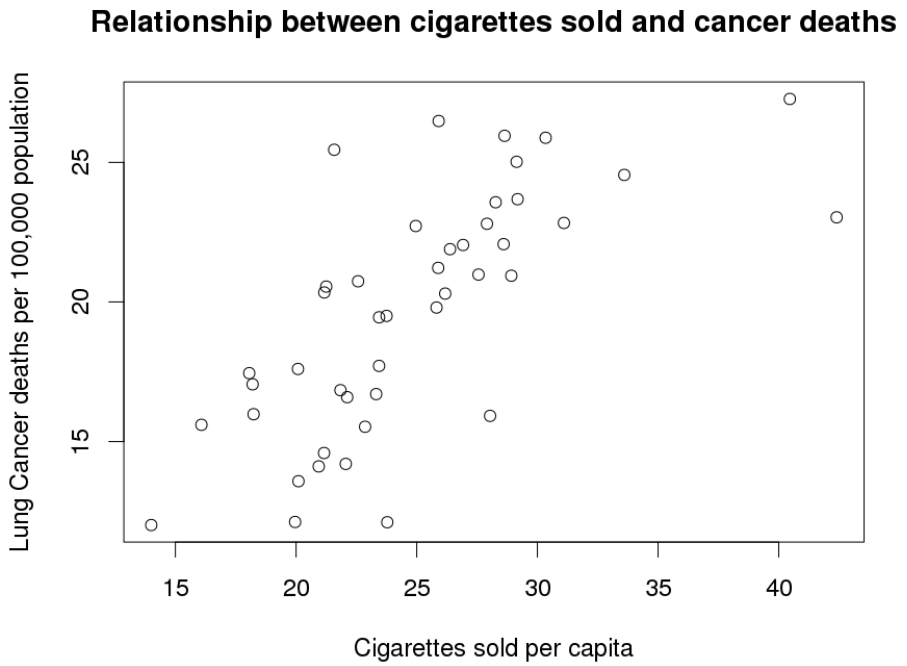
Do the points show a clear trend?

- Does it go upward or downward?
- How much scatter around the trend?

Does the trend seem be linear (follow a line) or is it curved?

Are there any outlier points?

Smoking and cancer



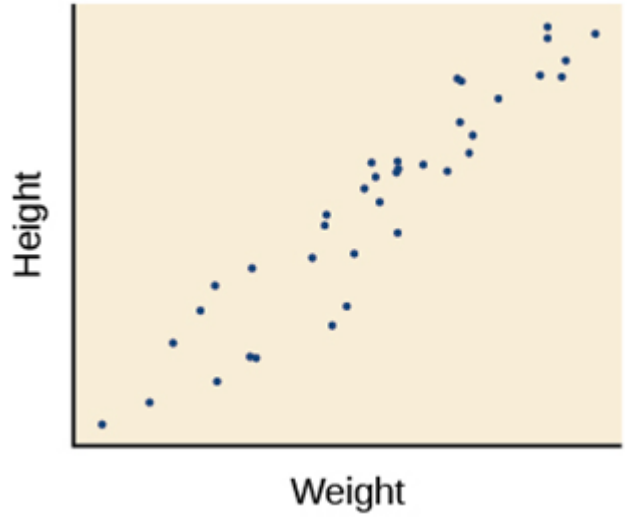
Positive, negative, no correlation

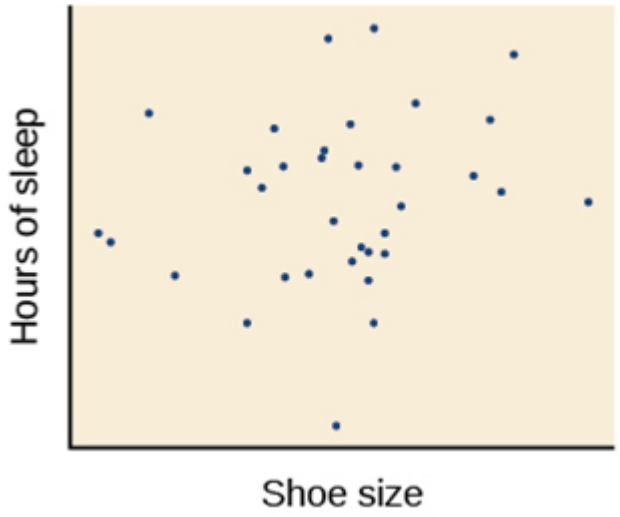
Do the points show a clear trend?

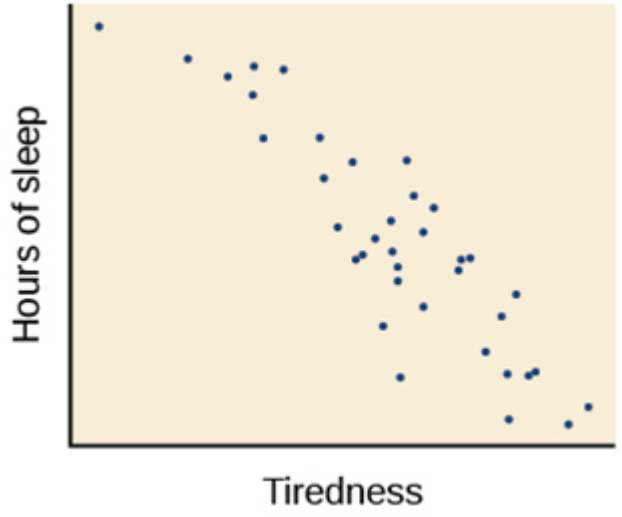
- Does it go upward or downward?
- How much scatter around the trend?

Does the trend seem be linear (follow a line) or is it curved?

Are there any outlier points?







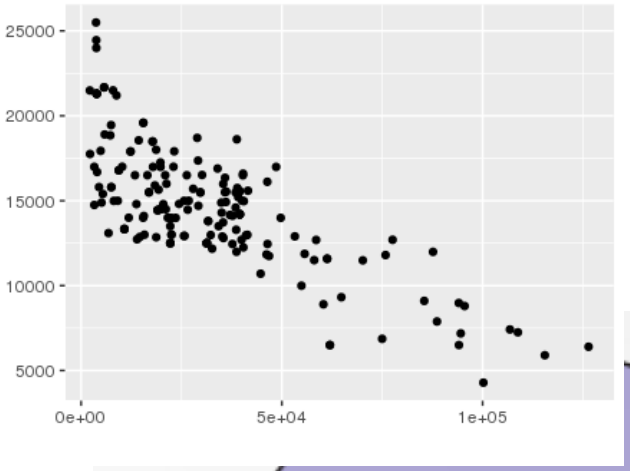
The correlation coefficient

The **correlation** is measure of the strength and direction of a linear association between two variables

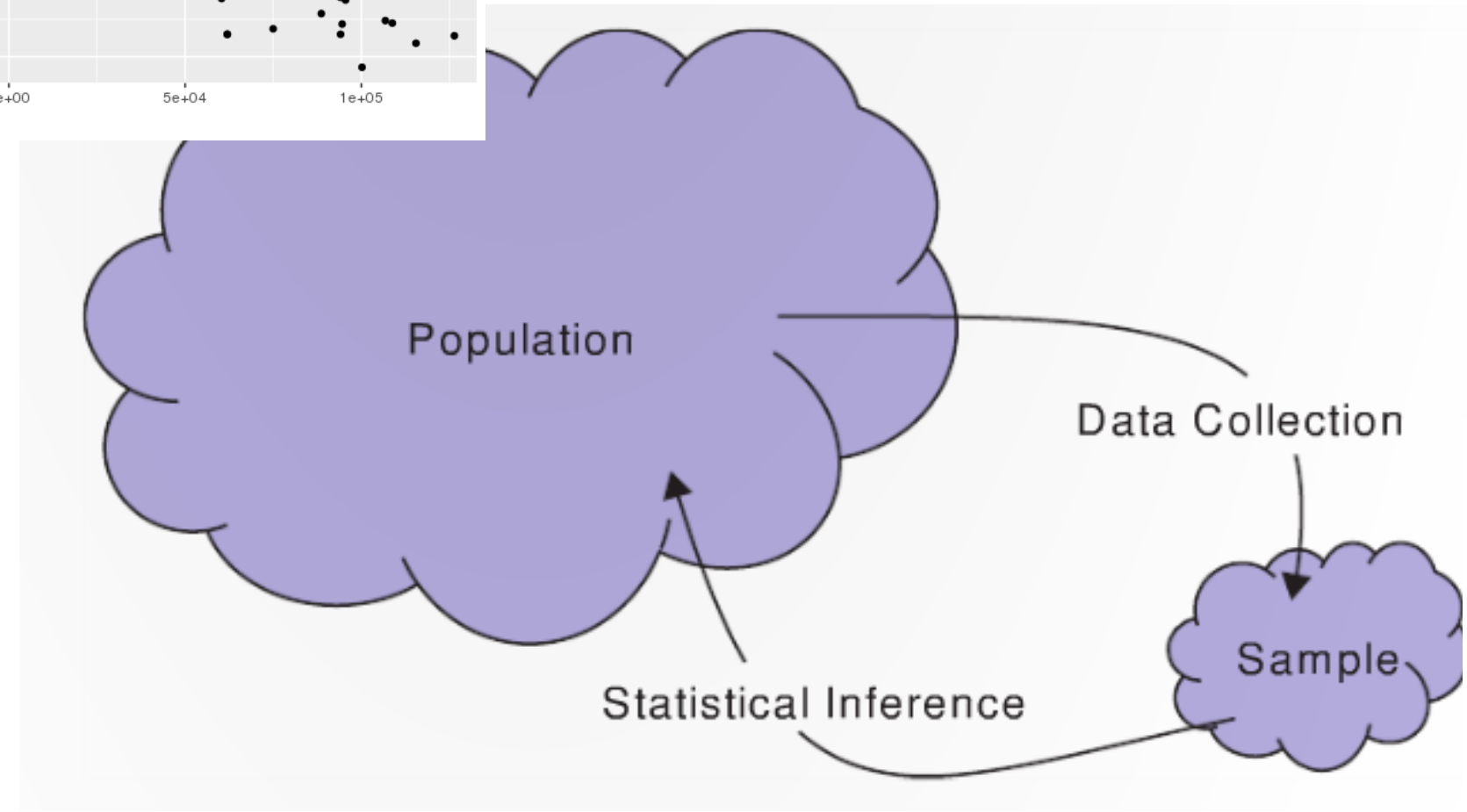
$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

- The correlation for a sample is denoted with **r**
- The correlation in the population is denoted with **ρ**
  (the Greek letter rho)

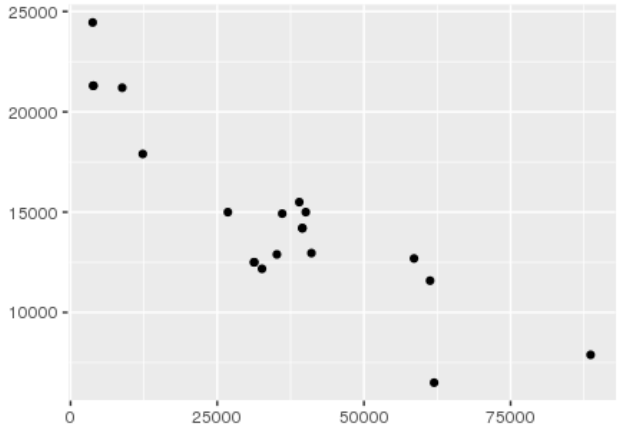
R: `cor(x, y)`



ρ parameter



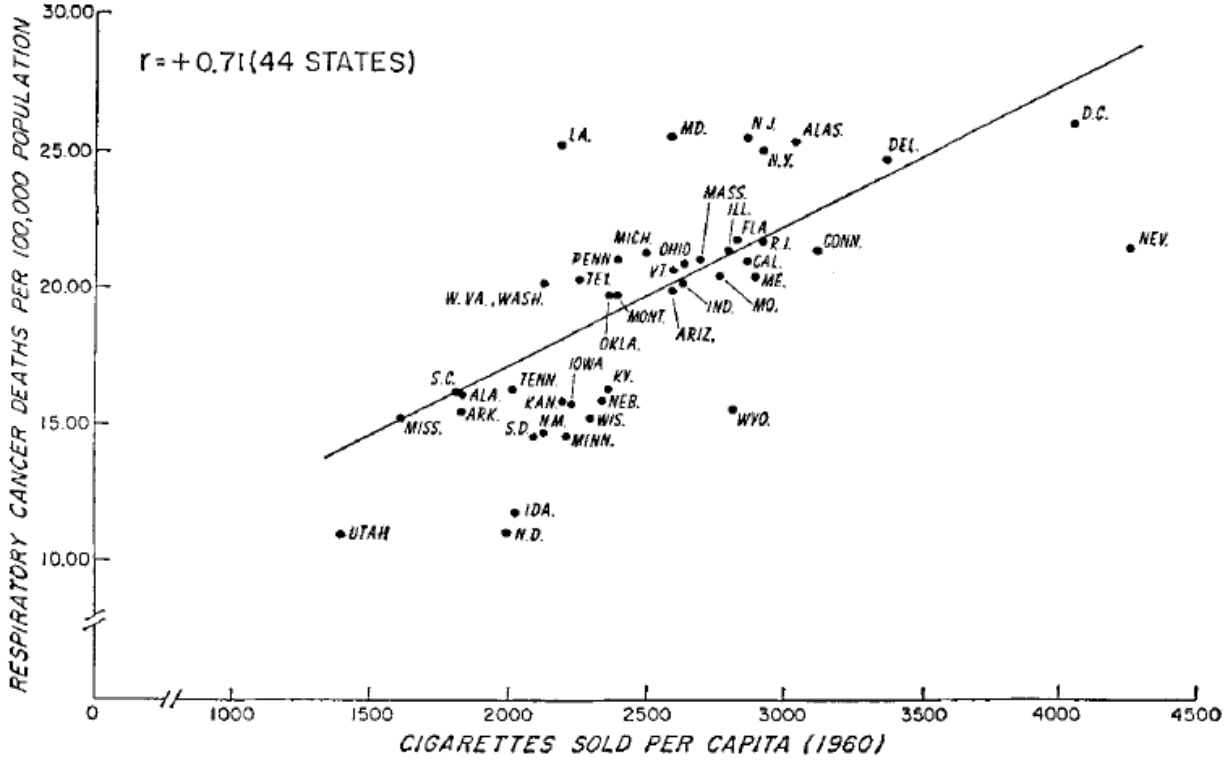
r statistic



Smoking and lung cancer correlation?

The **correlation** is measure of the strength and direction of a <u>linear association</u> between two variables

TEXT-FIGURE 2.—Correlation between average annual age-adjusted death rates for respiratory tract cancer (1956–61) and *per capita* cigarette sales (1960) in 44 States.



r = 0.71

Properties of the correlation

Correlation as always between -1 and 1: $-1 \leq r \leq 1$

The sign of r indicates the direction of the association

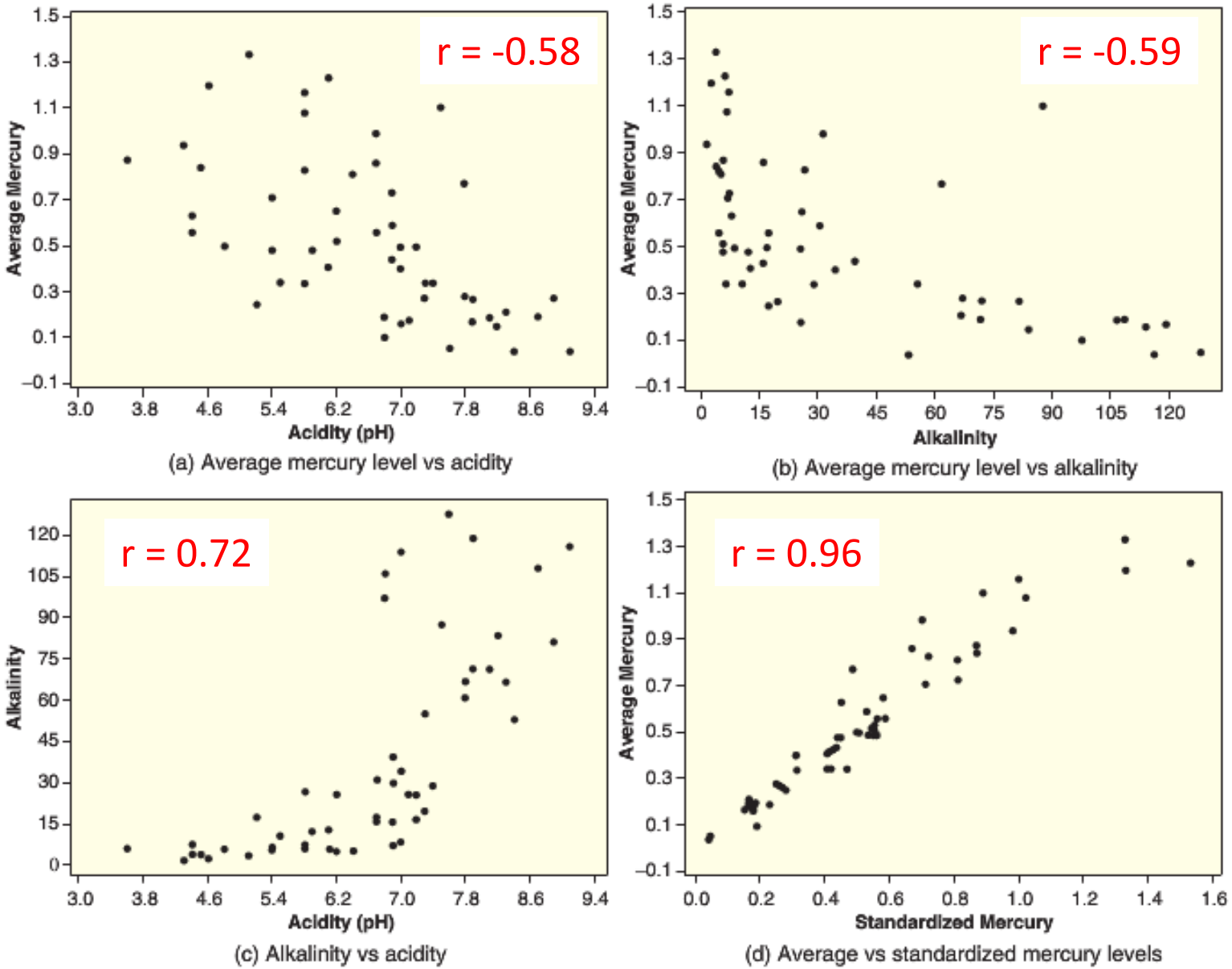
Values close to ± 1 show strong linear relationships, values close to 0 show no linear relationship

Correlation is symmetric: r = cor(x, y) = cor(y, x)

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

Florida lakes - guess the value of r

[Correlation game]



Let’s calculate some correlations

Is there an associate between cigarettes sold per capita and other types of cancer?

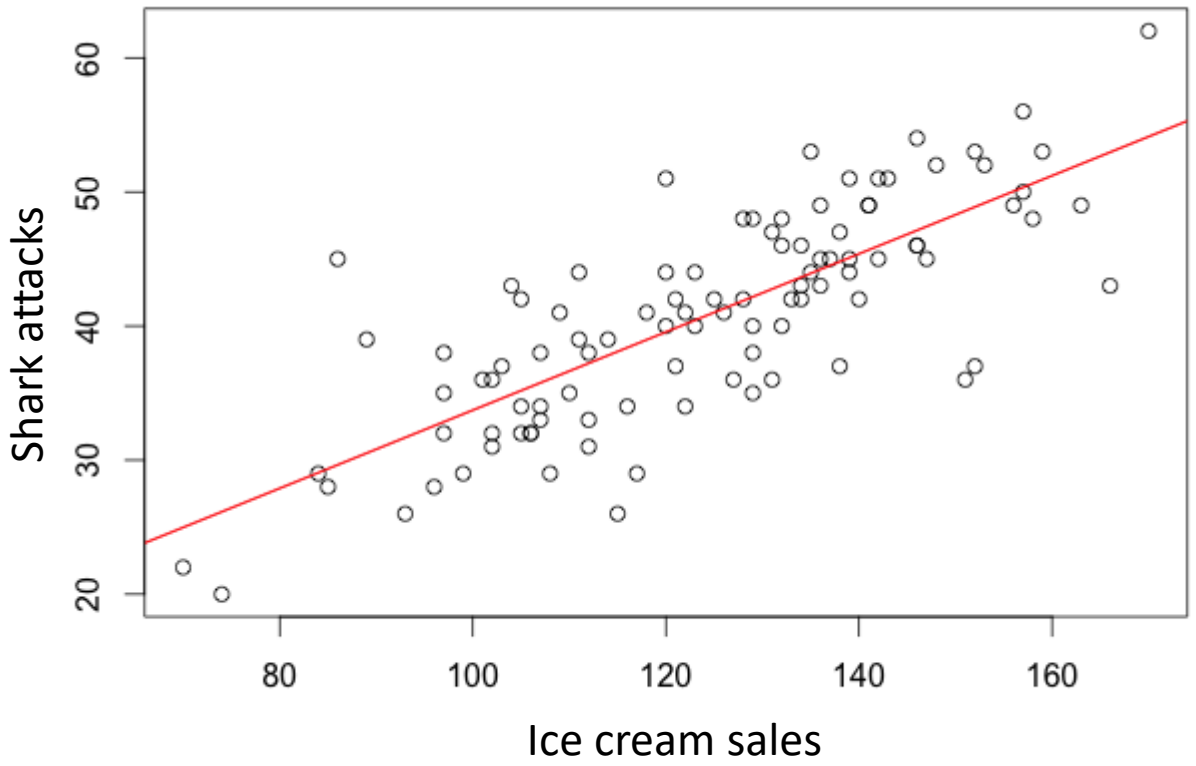
- Bladder cancer (BLAD)
- Kidney cancer (KID)
- Leukemia (LEUK)

```
# load the data
> download_class_data("smoking_cancer.Rda")
> load("smoking_cancer.Rda")

# create a scatter plot and calculate the correlation
> plot(smoking$CIG, smoking$LUNG)
> cor(smoking$CIG, smoking$LUNG)
```

Correlation caution #1

A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between two variables

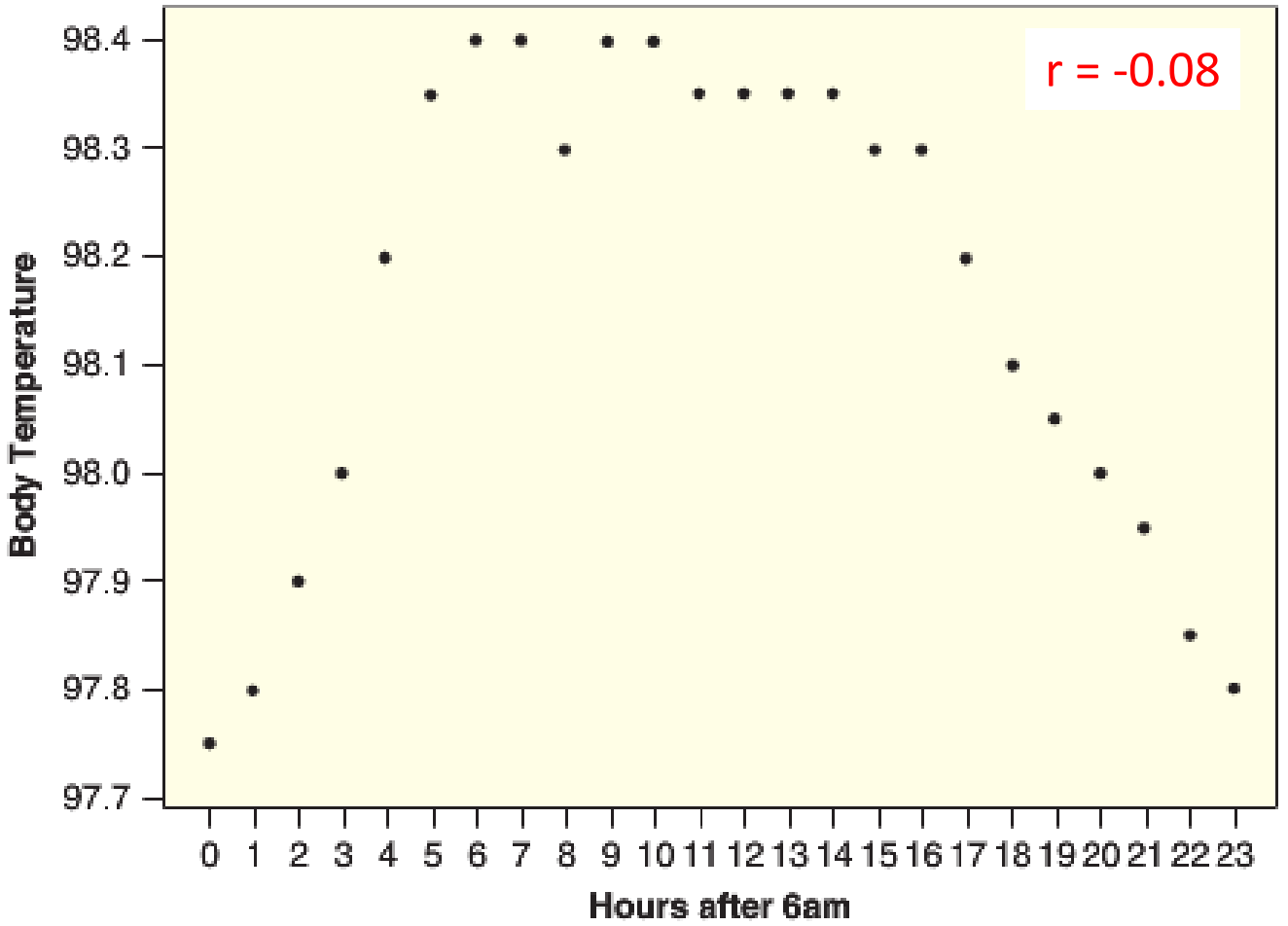



SOMEONE SENT ME ANOTHER ANONYMOUS EMAIL WITH A LINK TO AN ARTICLE ABOUT THE WORLD'S WORST BOSSES.
Dilbert.com DilbertCartoonist@gmail.com
I GET ONE OF THOSE EMAILS EVERY TIME I LEAVE YOUR CUBICLE. DID YOU THINK I WOULDN'T NOTICE THE CORRELATION?
11-28-11 © 2011 Scott Adams, Inc./Dist. by Universal Uclick
CORRELATION DOES NOT IMPLY CAUSATION.
SEND

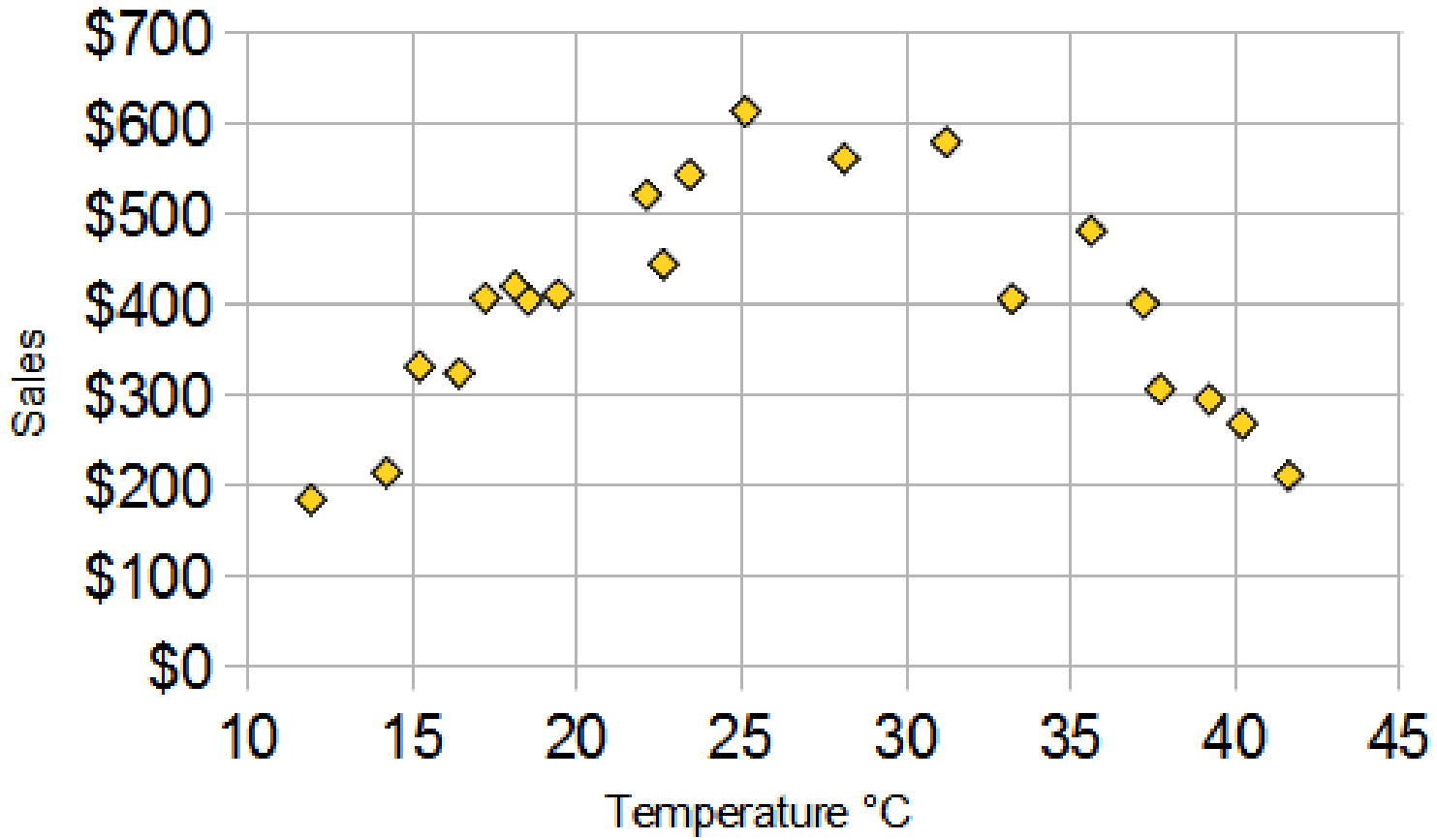
Correlation caution #2

A correlation near zero does not (necessarily) mean that two variables are not associated. Correlation only measures the strength of a linear relationship.

Body temperature as a function of time of the day

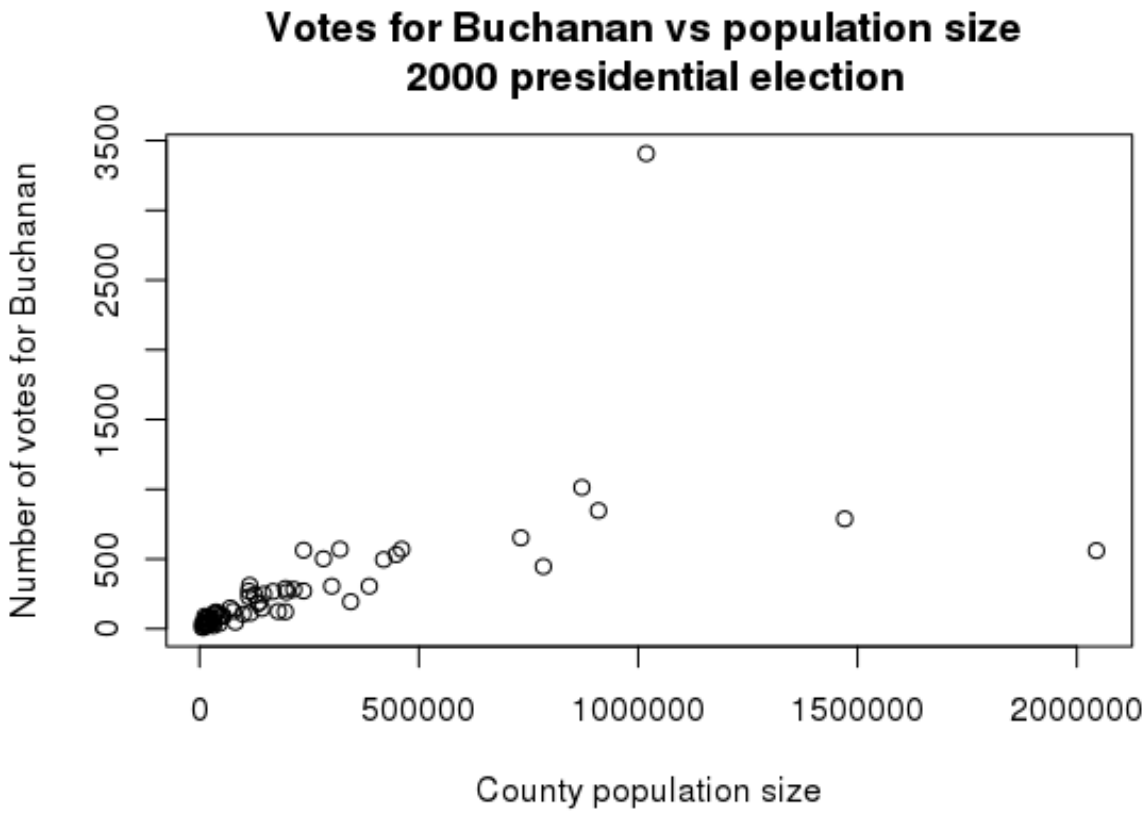


Ice cream sales and temperature



Correlation caution #3

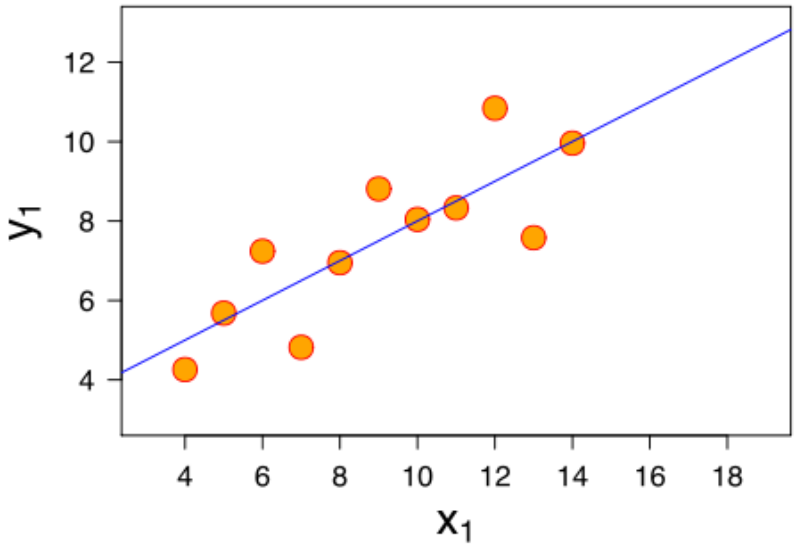
Correlation can be heavily influences by outliers. Always plot your data!

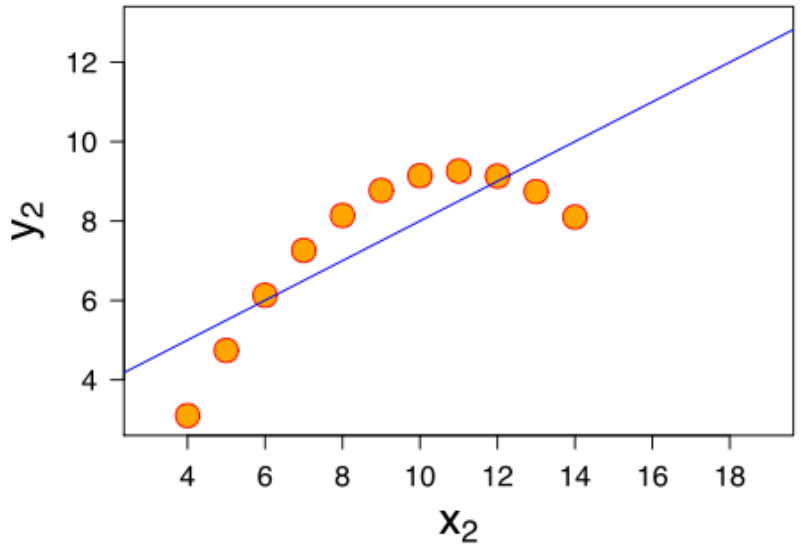


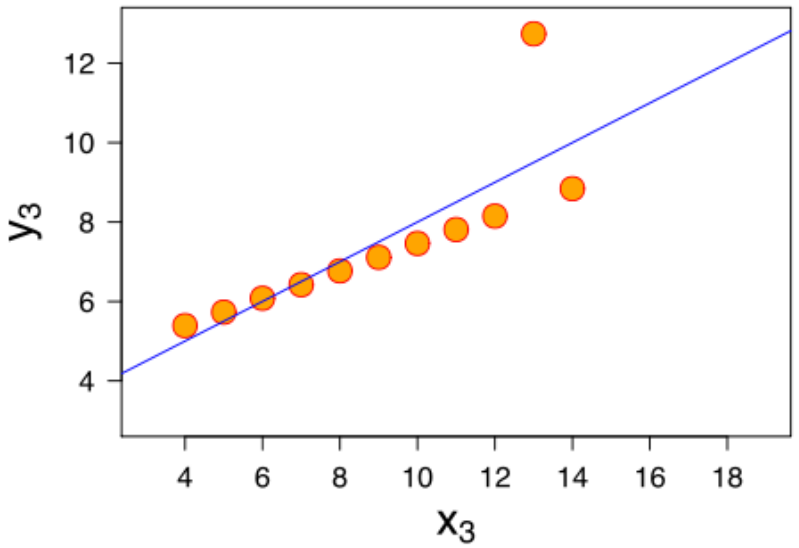
With Palm Beach
r = 0.61

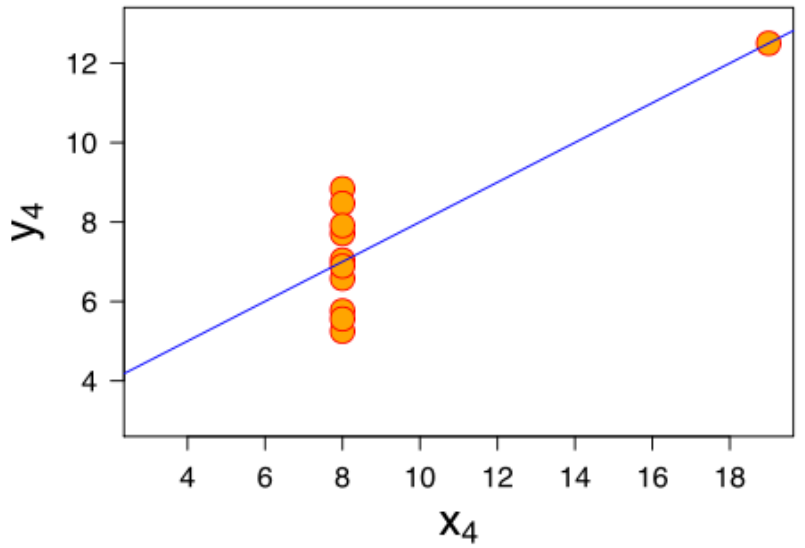
Without Palm Beach
r = .78

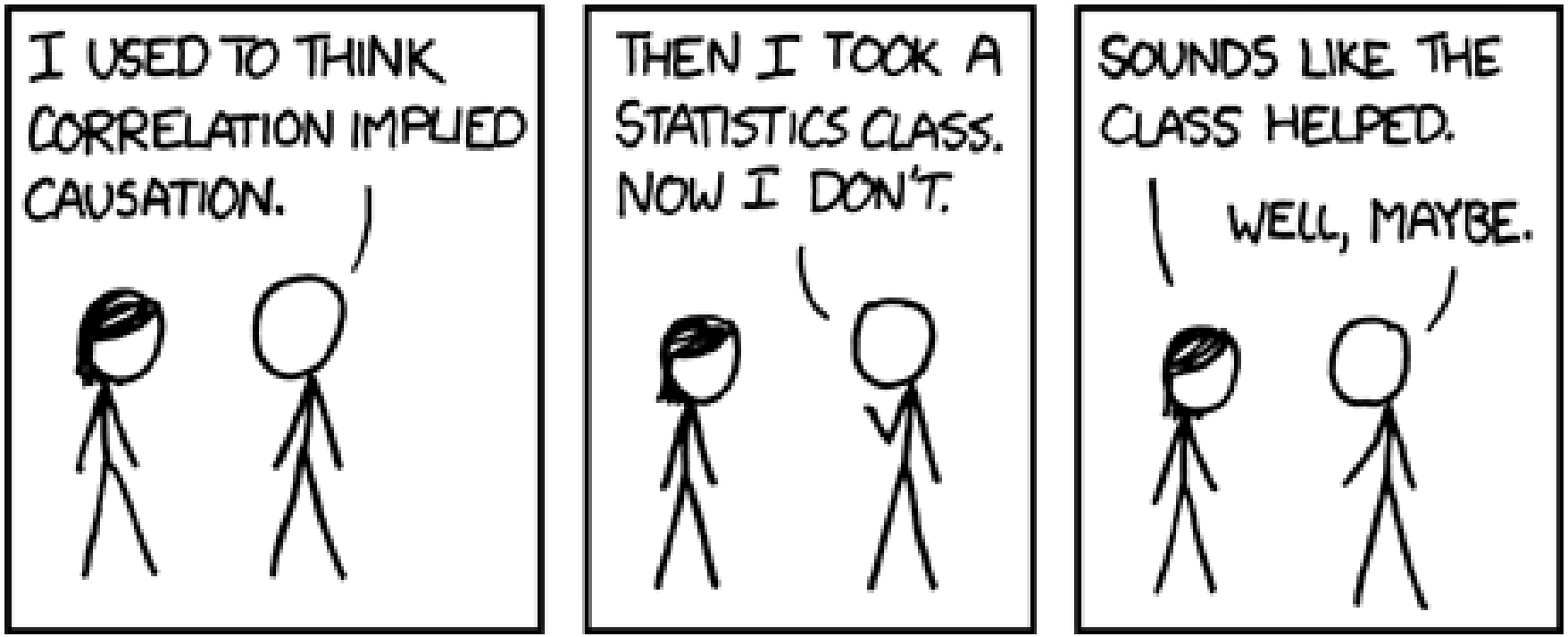
Anscombe's quartet (r = 0.81)










I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

More practice problems

Lock5 exercises first edition: 2.153, 2.155, 2.159, 2.177

Lock5 exercises second edition: 2.165, 2.167, 2.170, 2.191

Please get a copy of the textbook if you have not done so yet

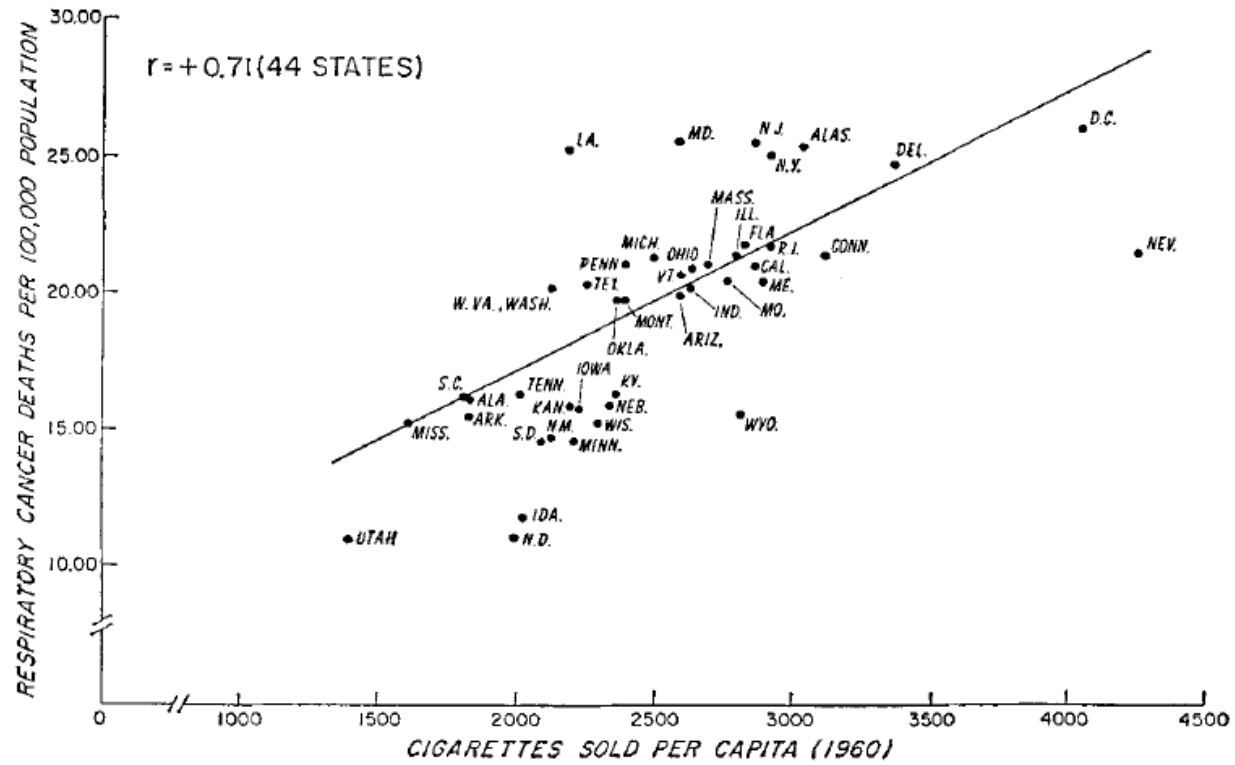
Regression

Regression is method of using one variable ***x*** *to predict* the value of a second variable ***y***

- i.e., $\hat{y} = f(x)$

In **linear regression** we fit a line to the data, called the **regression line**

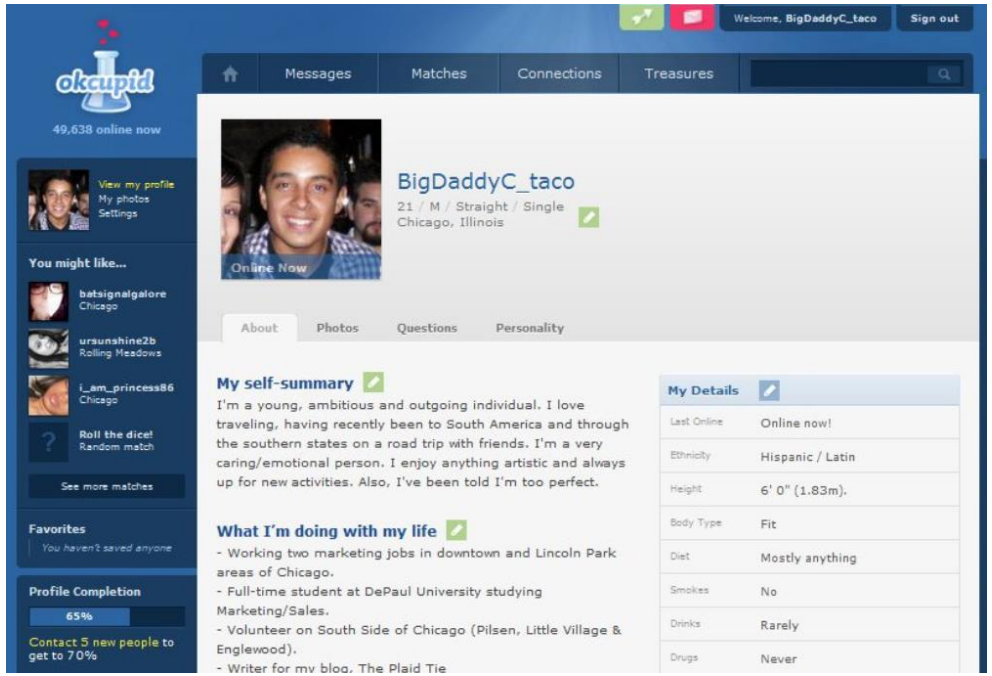
Cigarettes cancer regression line

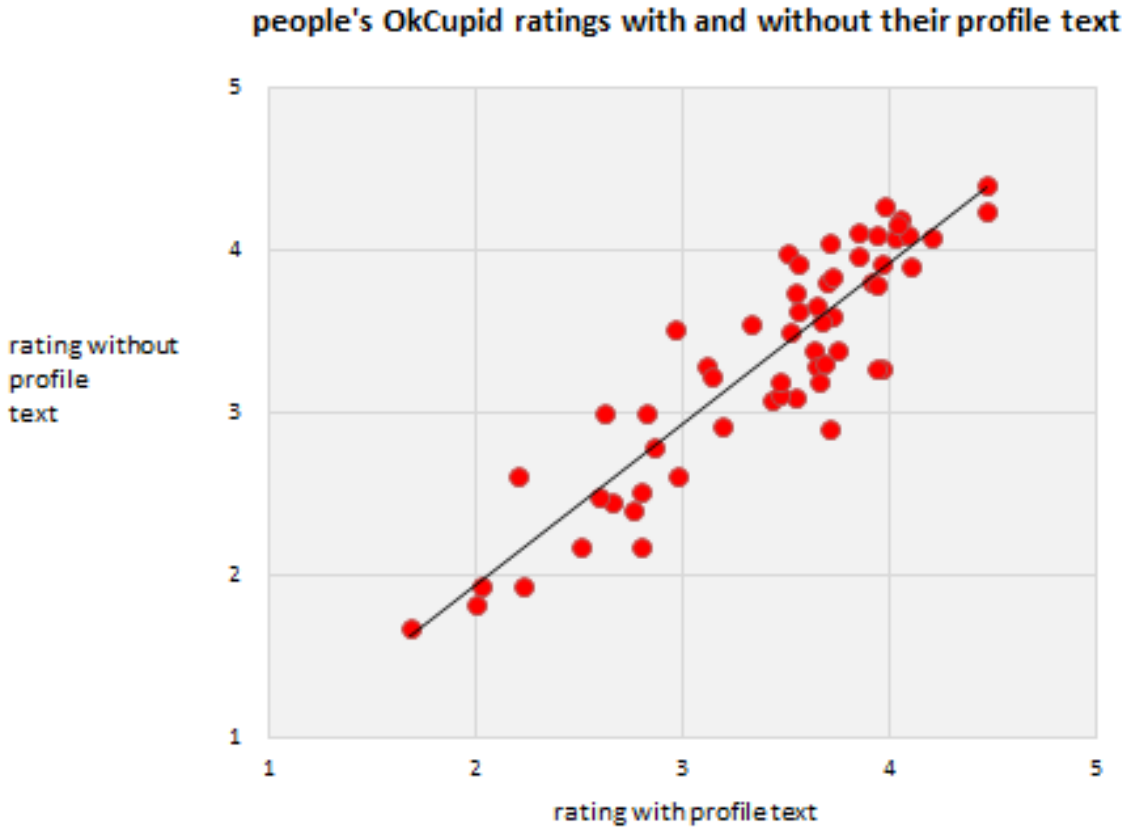


TEXT-FIGURE 2.—Correlation between average annual age-adjusted death rates for respiratory tract cancer (1956–61) and *per capita* cigarette sales (1960) in 44 States.

JOURNAL OF THE NATIONAL CANCER INSTITUTE

OkCupid text and images

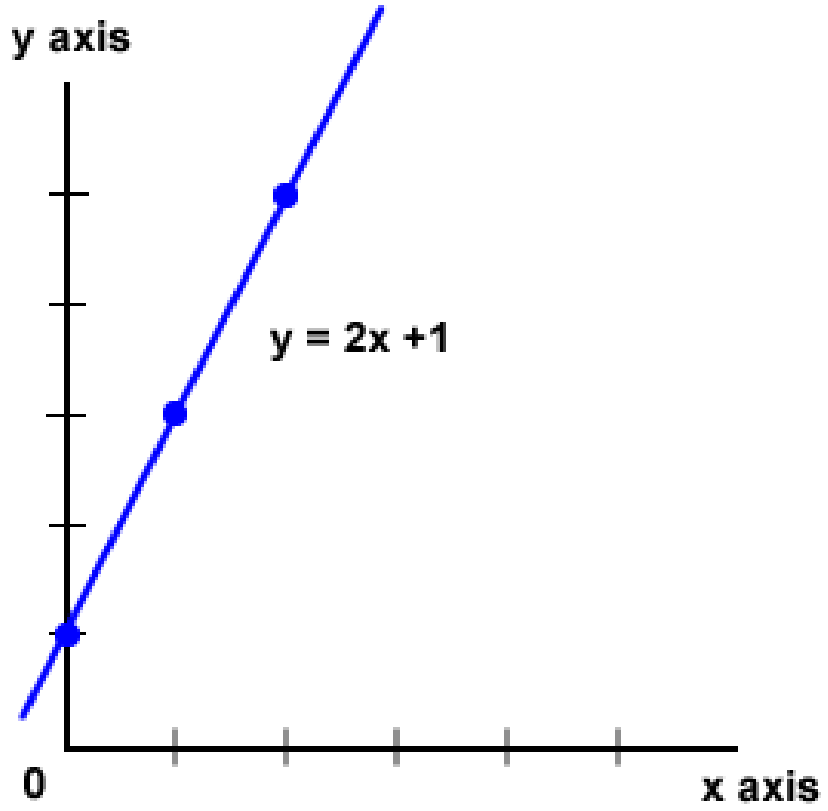




Equation for a line

What is the equation for a line?

$$\hat{y} = a + b \cdot x$$



Regression lines

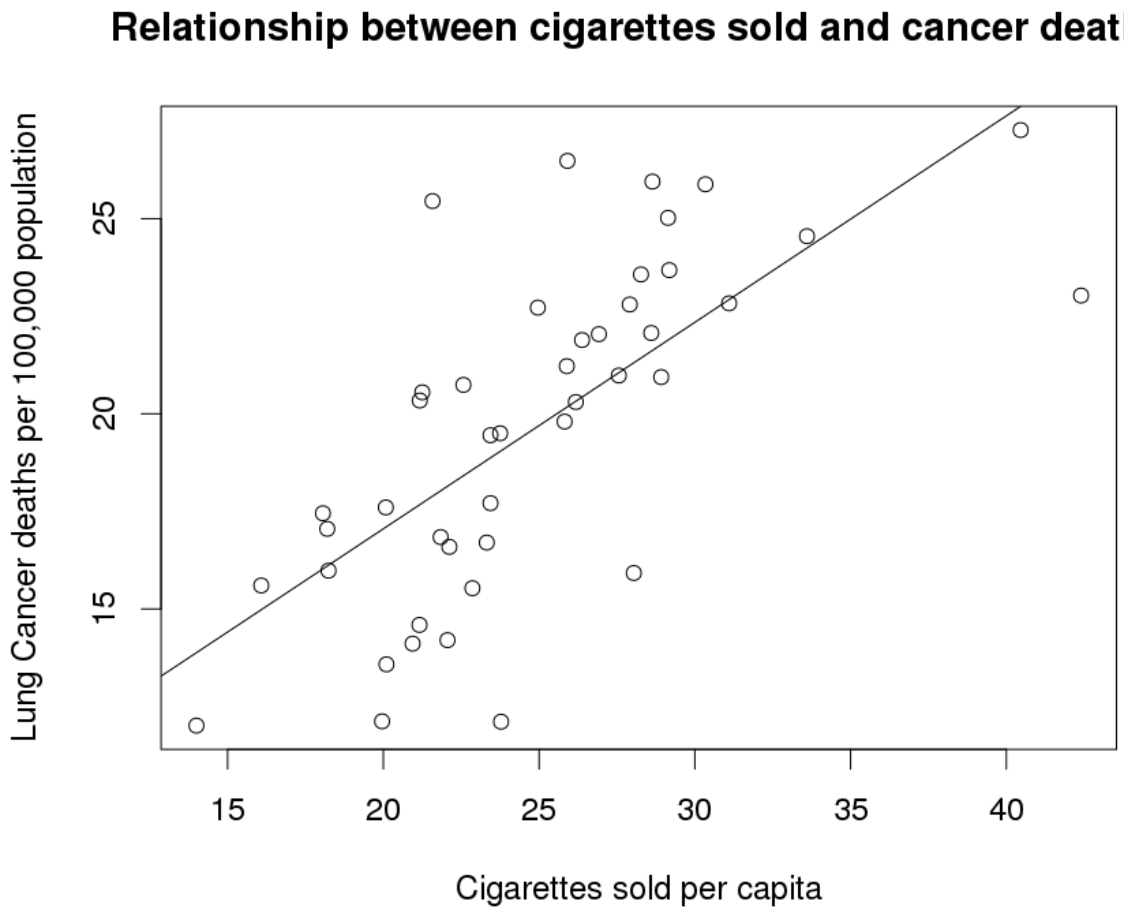
$\hat{y} = a + b \cdot x$

$Response = a + b \cdot Explanatory$

The slope ***b*** represents the predicted change in the response variable *y* given a one unit change in the explanatory variable *x*

The intercept ***a*** is the predicted value of the response variable y if the explanatory variable x were 0

Cancer smoking regression line



$\hat{y} = a + b \cdot x$

a = 6.47

b = 0.53

R: `lm(y ~ x)`

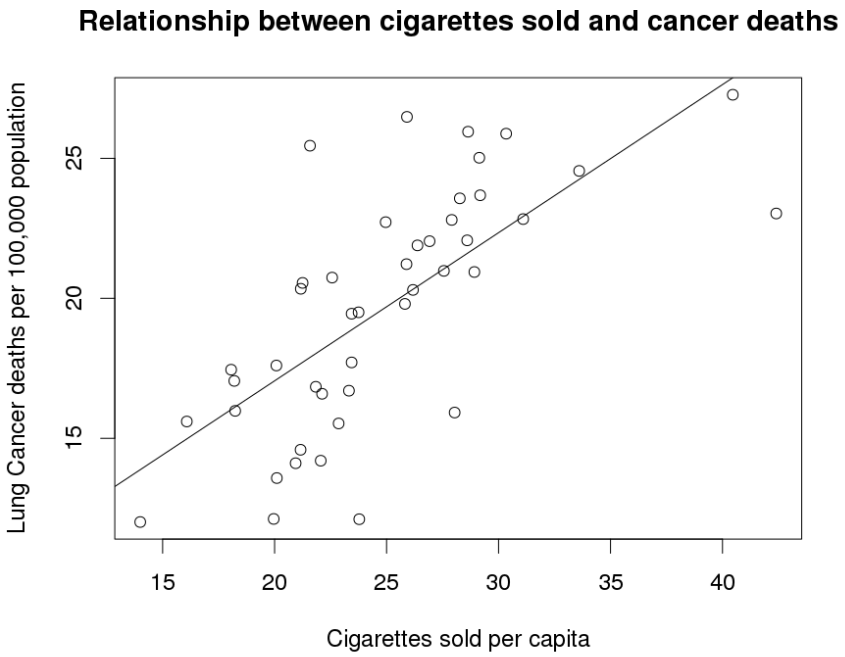
Using the regression line to make predictions

If a state sold 25 cigarettes per person

How many cancer deaths (per 100,000 people) would you expect?

$a = 6.47, \quad b = .53$

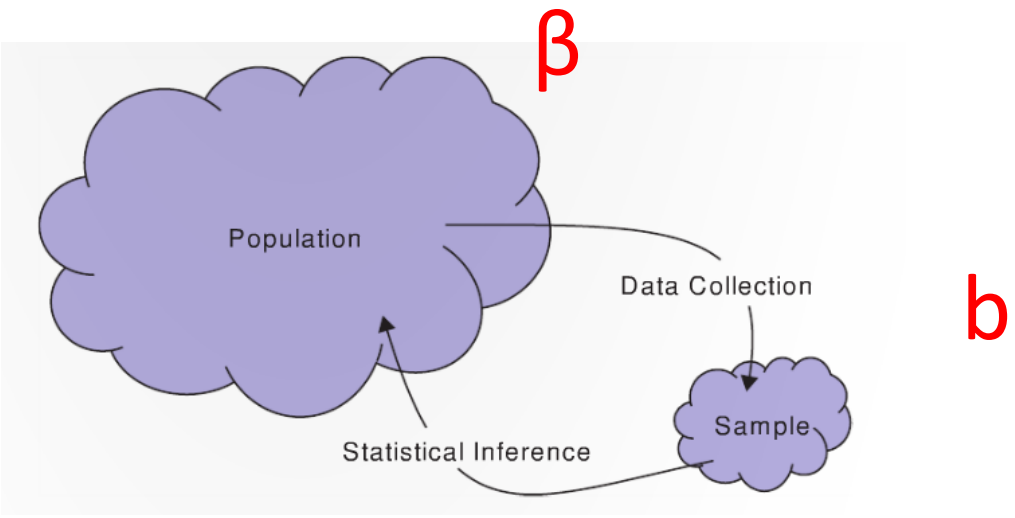
$\hat{y} = 6.47 + .53 \cdot x$



Notation

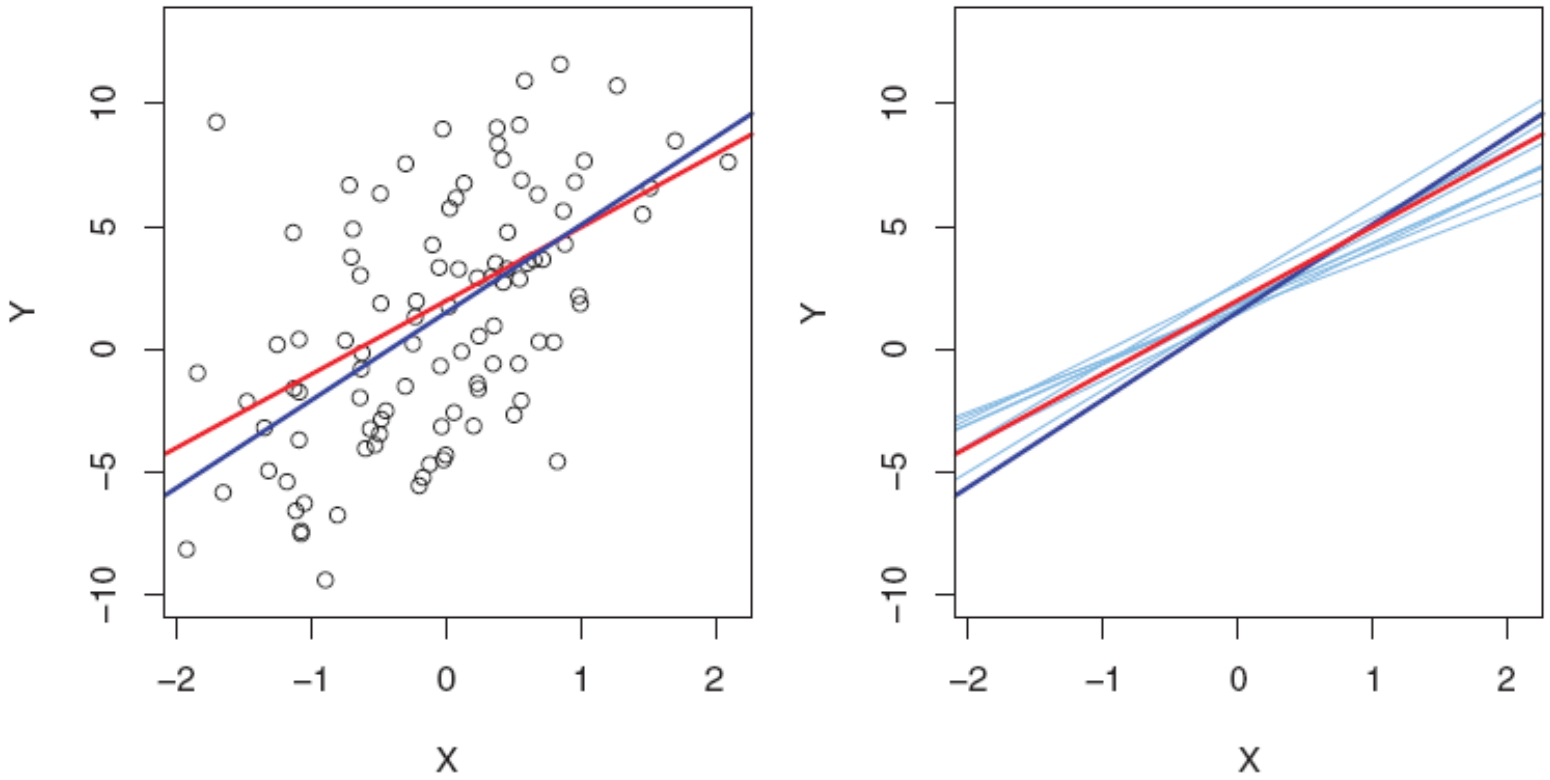
The letter **b** is typically used to denote the slope of the sample

The Greek letter **β** is used to denote the slope of the population



Population: β

Sample estimates: b

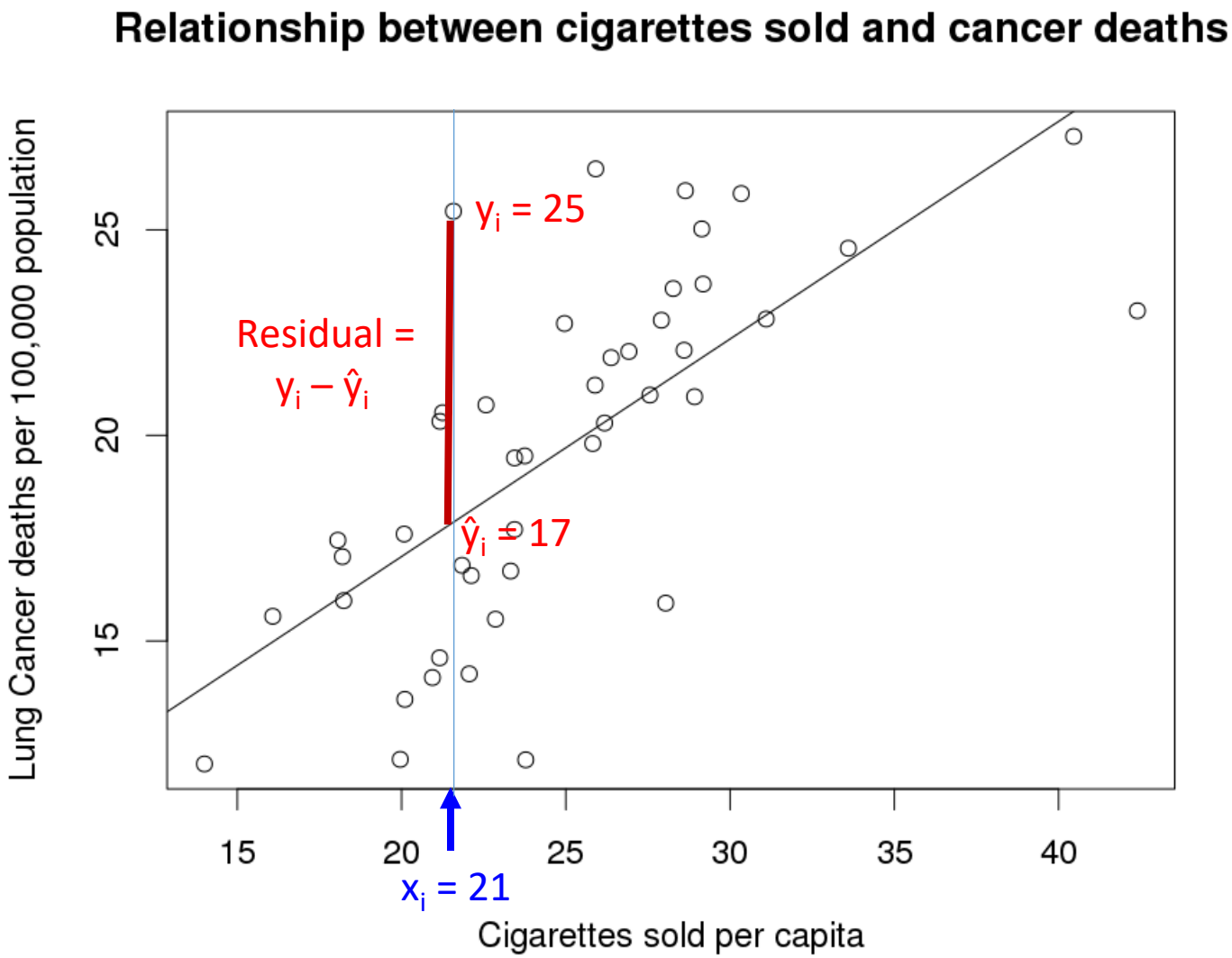


Residuals

The **residual** is the difference between an observed (y_i) and a predicted value ($\hat{y}_i$) of the response variable

$$Residual_i = Observed_i - Predicted_i = y_i - \hat{y}_i$$

Cancer smoking residuals

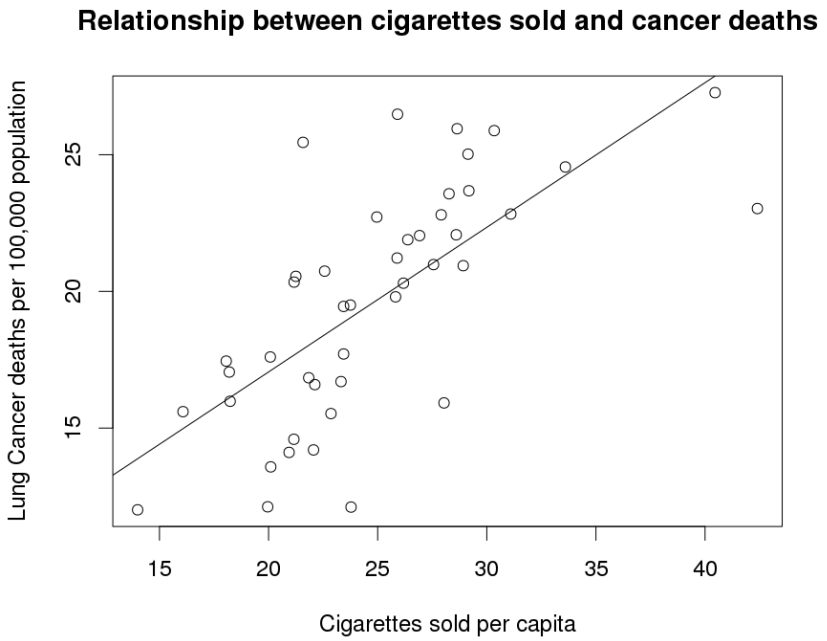


Cancer smoking residuals

Cancer obs (y)	Cancer pred (ŷ)	Residuals (y - ŷ)
17.05	16.10	0.95
19.80	20.13	-0.33
15.98	16.12	-0.14
22.07	21.60	0.47
22.83	22.93	-0.10
24.55	24.25	0.30
27.27	27.88	-0.61
23.57	21.24	2.14

Line of ‘best fit’

The **least squares line**, also called ‘**the line of best fit’**, is the line which minimizes the sum of squared residuals



Try to find the line of best fit

Cancer smoking residuals

Cancer obs (y)	Cancer pred (ŷ)	Residuals (y - ŷ)	Residuals² (y - ŷ)²
17.05	16.10	0.95	0.90
19.80	20.13	-0.33	0.11
15.98	16.12	-0.14	0.02
22.07	21.60	0.47	0.22
22.83	22.93	-0.10	0.01
24.55	24.25	0.30	0.09
27.27	27.88	-0.61	0.37
23.57	21.24	2.14	4.59

Let’s calculate regression lines in R

```
# download the smoking data
> download_class_data("smoking_cancer.Rda")

# create a scatter plot and calculate the correlation
> plot(smoking$CIG, smoking$LUNG)

# fit a regression model
> lm_fit <- lm(smoking$LUNG ~ smoking$CIG)

# examine the a and b coefficients
> coef(lm_fit)

# add the regression line to the plot
> abline(lm_fit)
```

Concepts for the relationship between two quantitative variables

A **scatterplot** graphs the relationship between two variables

The **correlation** is measure of the strength and direction of a linear association between two variables

- Value between -1 and 1

In **linear regression** we fit a line to the data, called the **regression line**

- We get coefficients for the slope (b) and the y-intercept (a)

The **residual** is the difference between an observed (y_i) and a predicted value ($\hat{y}_i$) of the response variable

- The regression line minimizes the sum of squared residuals