

Relationships between two
quantitative variables

Overview

Quick review

- Standard deviations, z-scores, percentiles, boxplots

Relationships between two quantitative variables

- Scatter plots
- Correlation

If there is time

- Simple linear regression

Announcement

Homework 3 has been posted!

It is due on Gradescope on **Sunday September 21st at 11pm**

- **Be sure to mark each question on Gradescope!**

Also, be sure to use the “education” partition on the YCRC server

Note: The homework involves taking a “Quiz” on Canvas to test your knowledge

- You should study for the Quiz before you take it, although you will only be graded on completion and not on how many questions you get right

Memory per CPU core in GiB

5

Partitions

education

Reservation (optional)

☐ I would like to receive an email when the session starts

Additional modules (optional)

provide additional modules. Module names are separated by a space.

☐ Check the box to view more options

Launch

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

This is LeBron's best statistic (relative to his peers)
But should we be impressed with this?

- i.e., maybe z-scores of 2.7 are common?

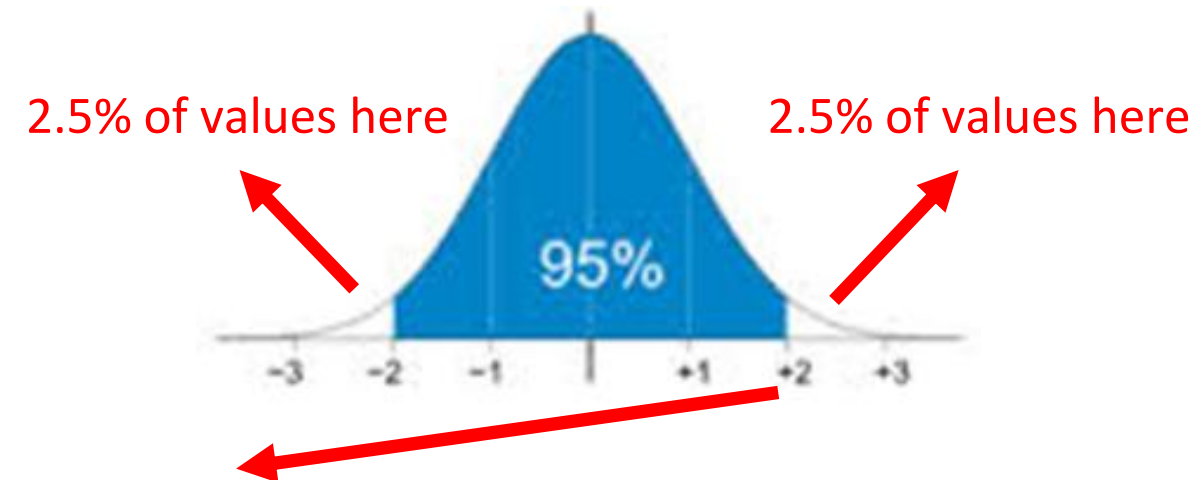
The normal pillow

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

- **Answer:** 95%

Question: If the values are normally distributed, how frequently should we expect a randomly selected z-score to be more than 2?

- **Answer:** 2.5% of the time



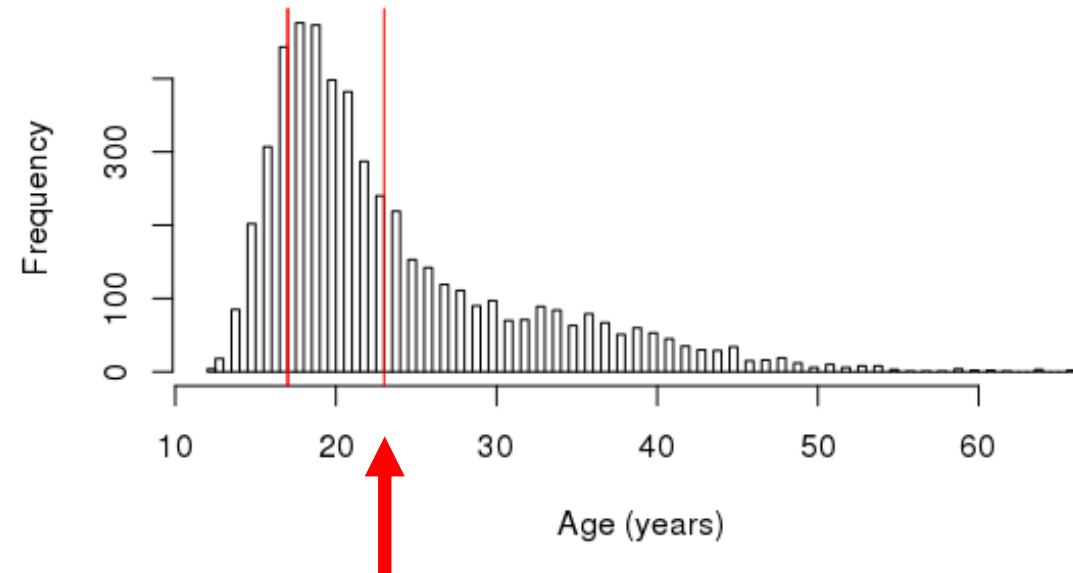
97.5% of values are less than 2

Review: quantiles (percentiles)

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

Let's look at the age people were arrested for using marijuana in Toronto

Histogram of Ages of people arrested for marijuana use



60th percentile value is 23

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

In R: `quantile(Arrests$age, .6)`

New: Quantiles of the normal distribution

We can also get quantiles of a normal distribution

Question: what is the 97.5% quantile of a standard normal distribution?

- i.e., What z-score value is greater than 97.5% of the data in a standard normal distribution?



97.5% of values are less than 2

In R: `qnorm(.975)`

The quantile universe

Five-Number Summary = (minimum, Q_1 , median, Q_3 , maximum)

Q_1 = 25th percentile, Q_3 = 75th percentile

Range = maximum – minimum

Interquartile range (IQR) = $Q_3 - Q_1$

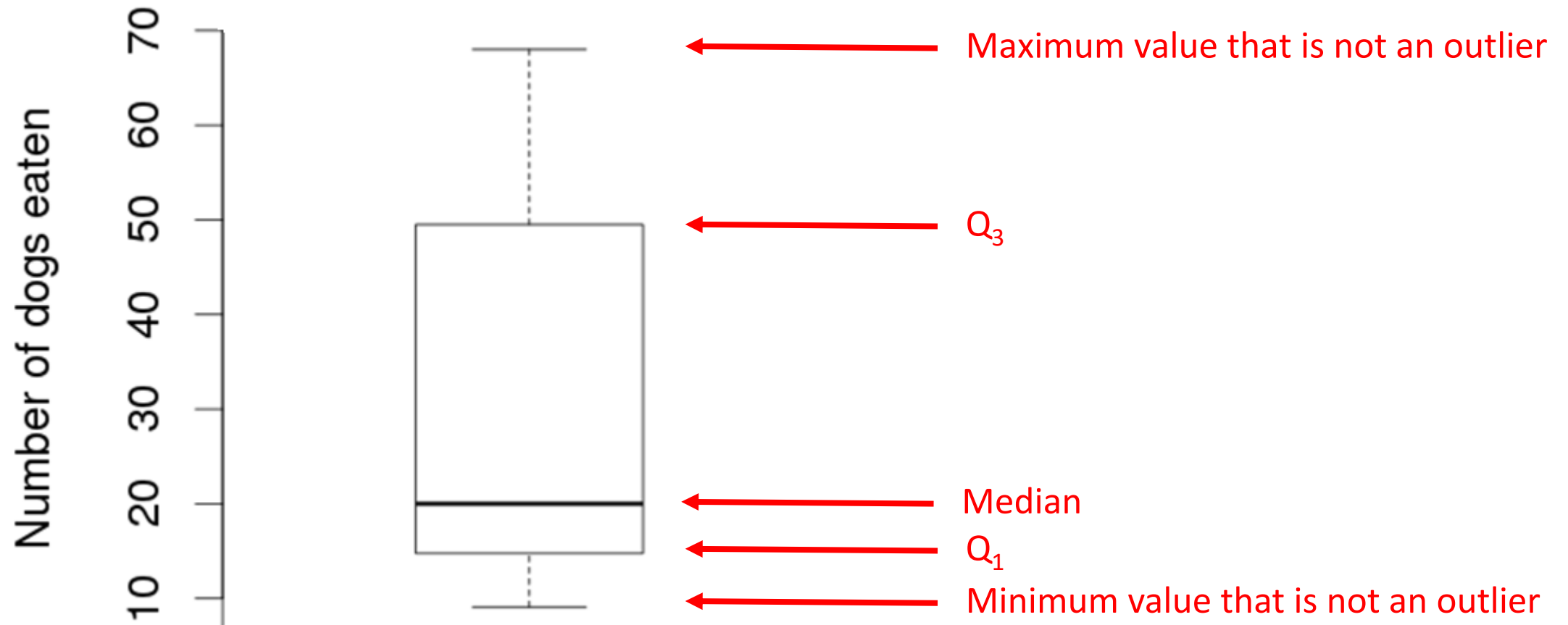
As a rule of thumb, we call a data value an **outlier** if it is:

Smaller than: $Q_1 - 1.5 * IQR$

Larger than: $Q_3 + 1.5 * IQR$

In R: `fivenum(v)`

Box plot of the number of hot dogs eaten by the men's contest winners 1980 to 2010

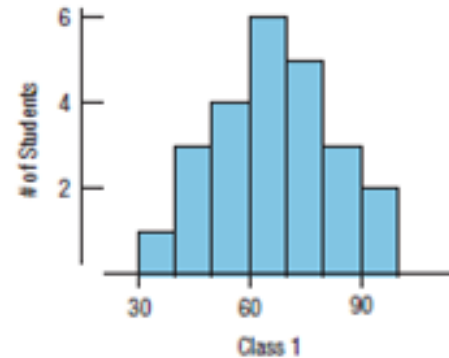


R: `boxplot(v)`

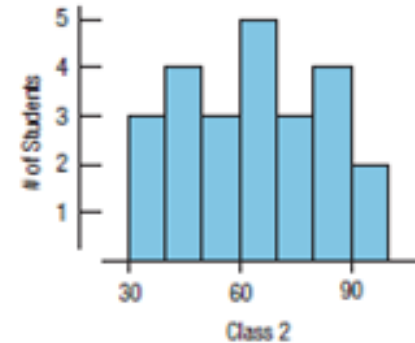
Box plots extract key statistics from histograms

Question: which Box plot goes with which histogram?

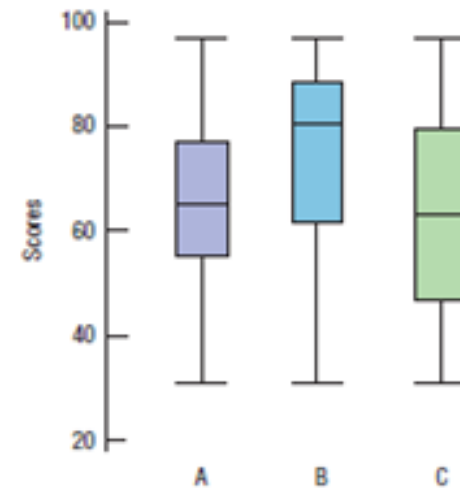
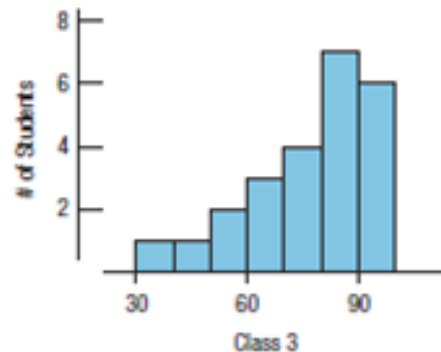
Histogram 1



Histogram 2



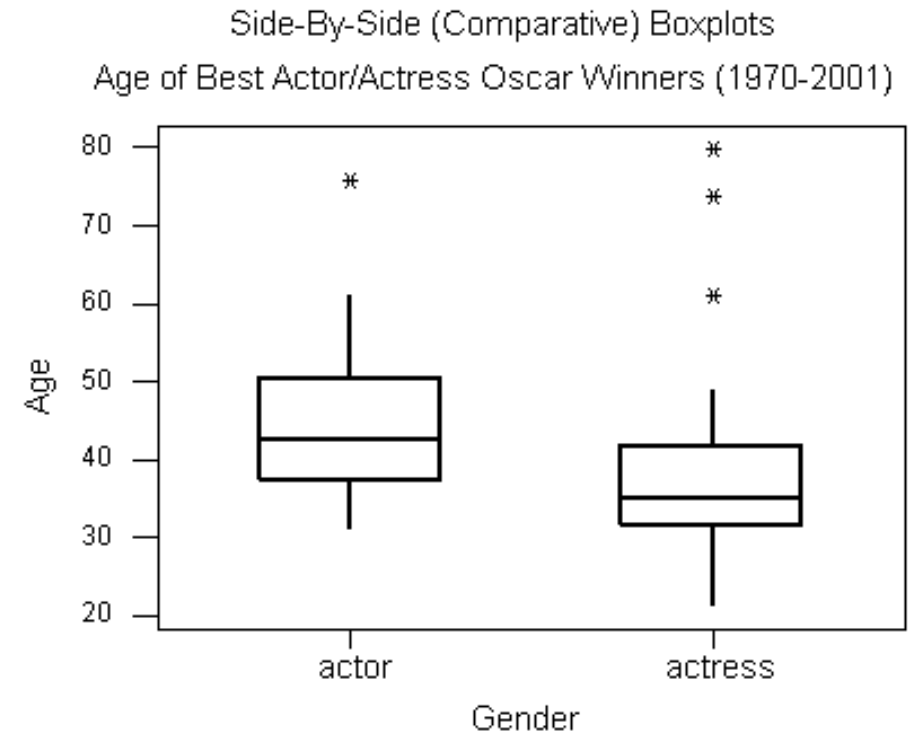
Histogram 3



Comparing quantitative variables across categories

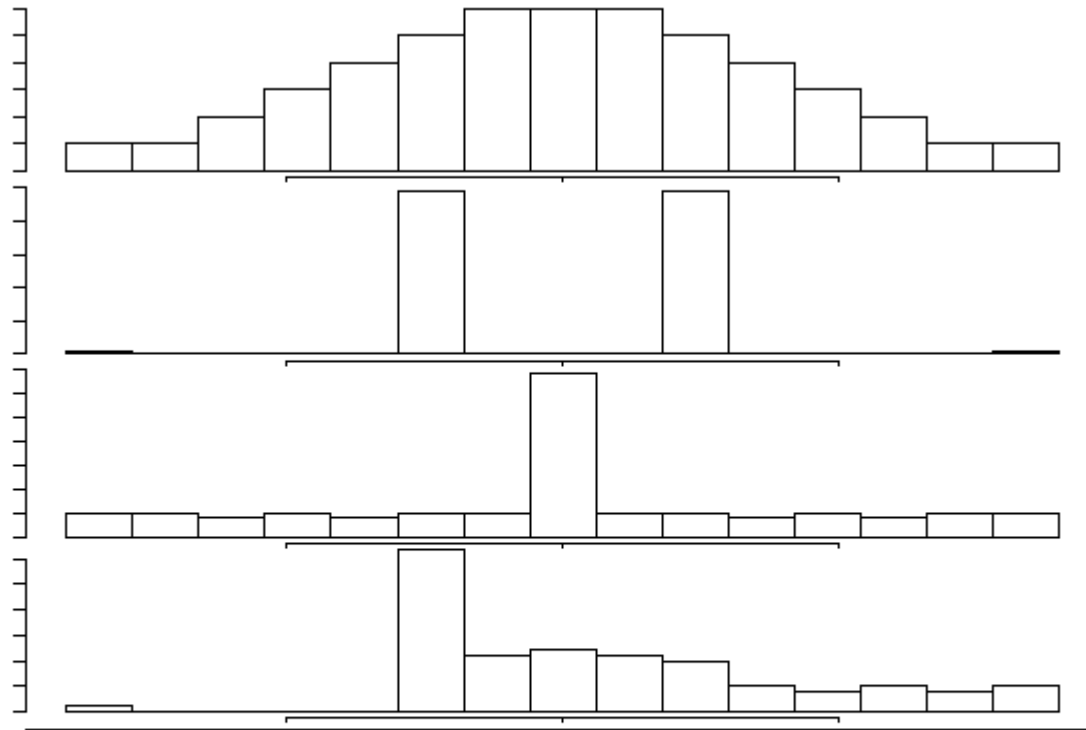
Often one wants to compare quantitative variables across categories

Side-by-Side graphs are a way to visually compare quantitative variables across different categories



```
boxplot(v1, v2, names = c("name 1", "name 2"), ylab = "y-axis name")
```

Box plots don't capture everything



Do you think the box plots for these distributions look similar?

Relationships between two
quantitative variables

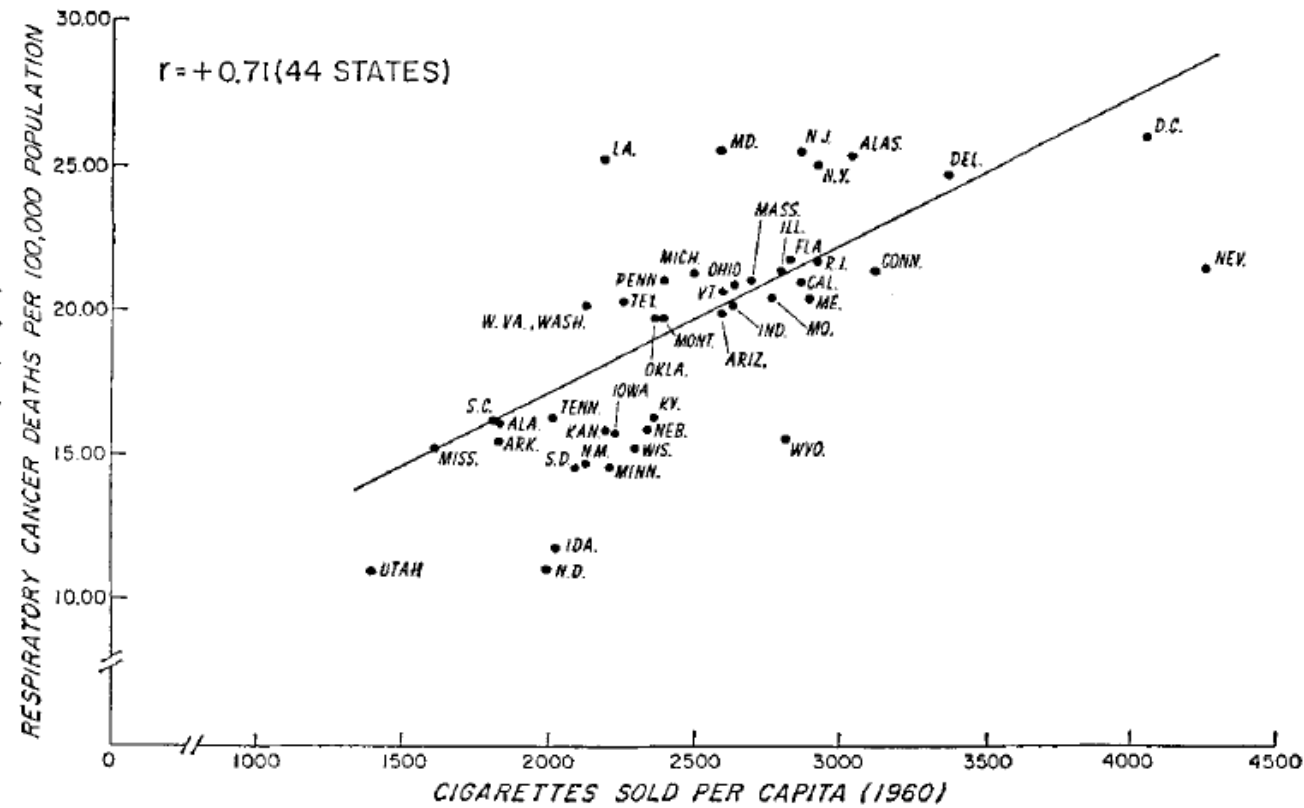
Two quantitative variables

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

State	Cig per capita	Bladder	Lung	Kidney	Leukemia
AL	1,820	2.9	17.05	1.59	6.15
AZ	2,582	3.52	19.8	2.75	6.61
AR	1,824	2.99	15.98	2.02	6.94
CA	2,860	4.46	22.07	2.66	7.06
CT	3,110	5.11	22.83	3.35	7.2
DE	3,360	4.78	24.55	3.36	6.45
DC	40,460	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.



Scatterplot

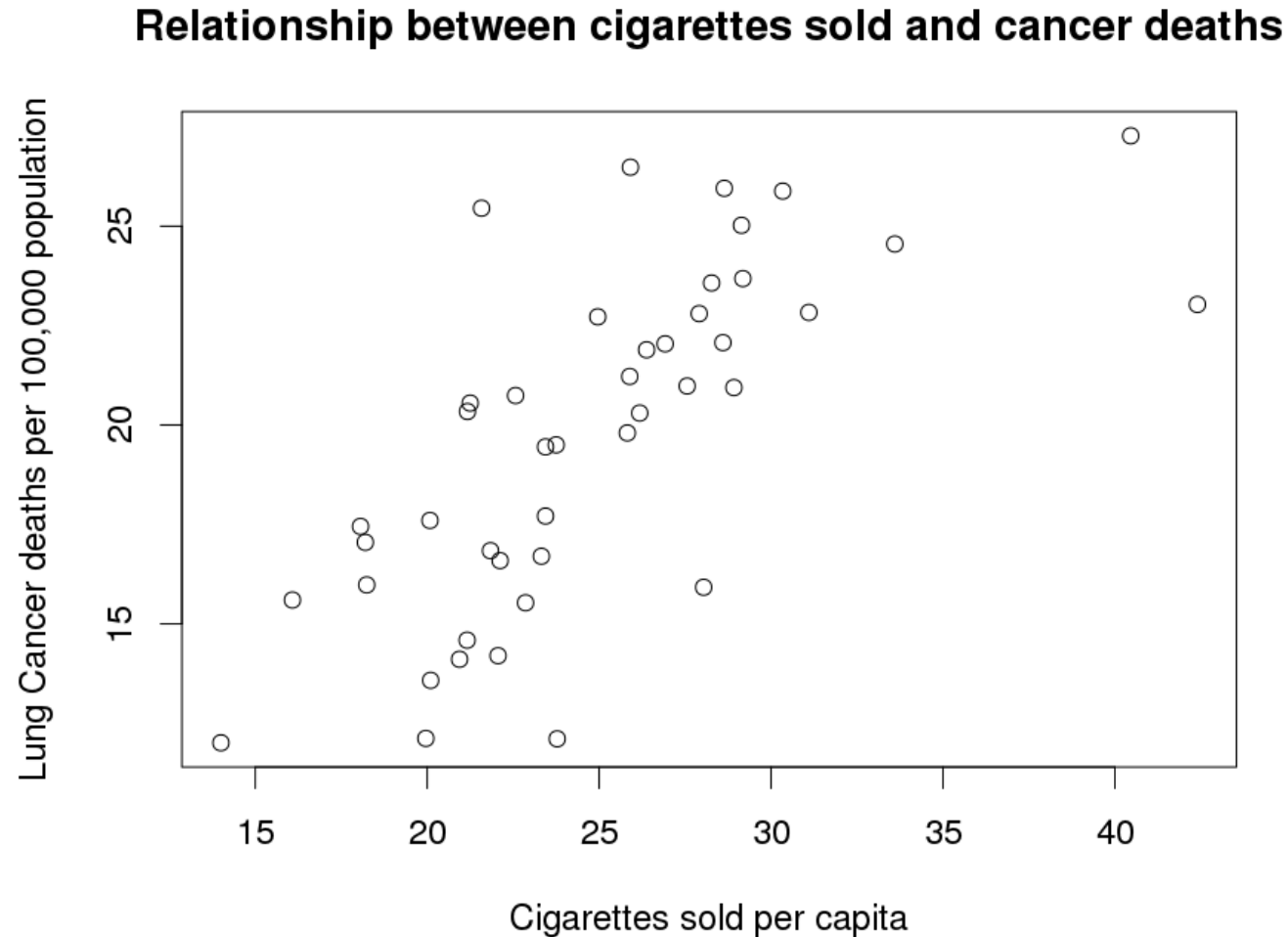
A **scatterplot** graphs the relationship between two variables

- 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?

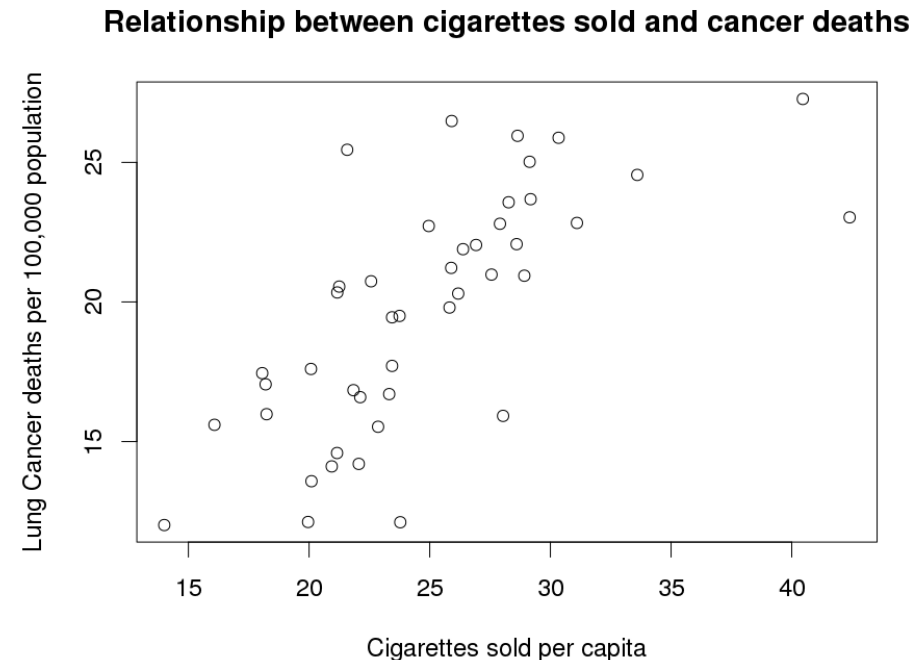
Does it go upward or downward?

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

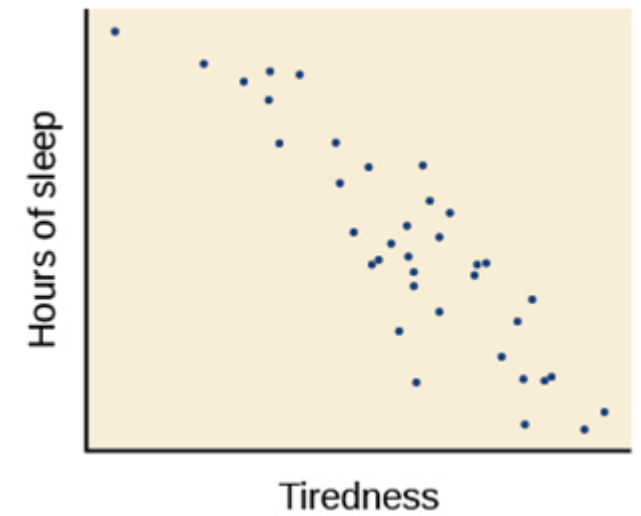
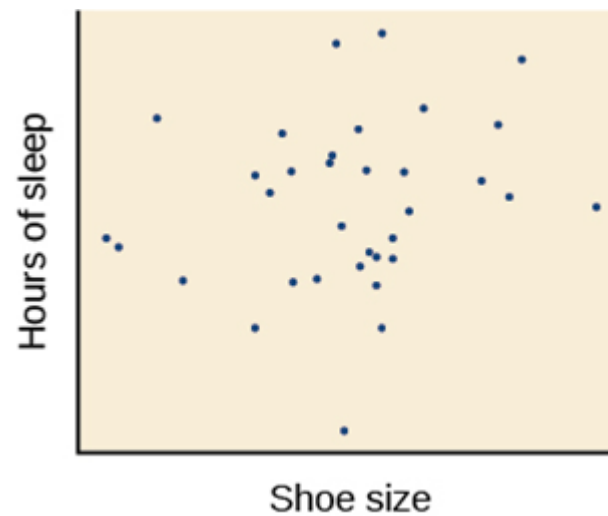
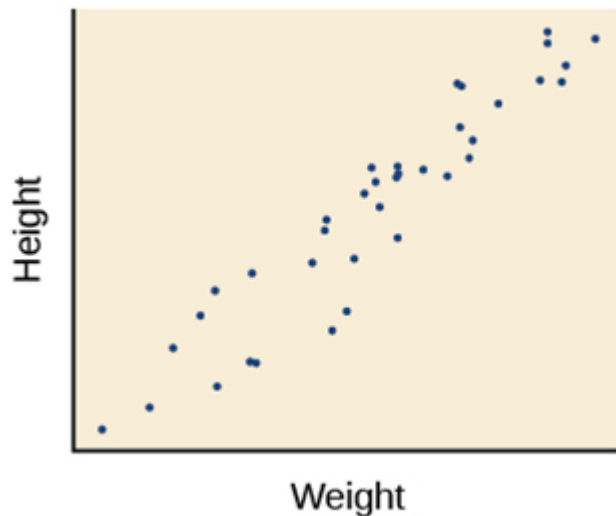
Do the points show a clear trend?

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?



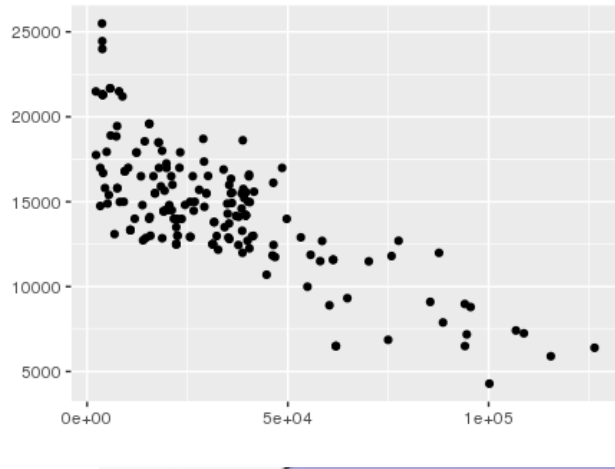
The correlation coefficient

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

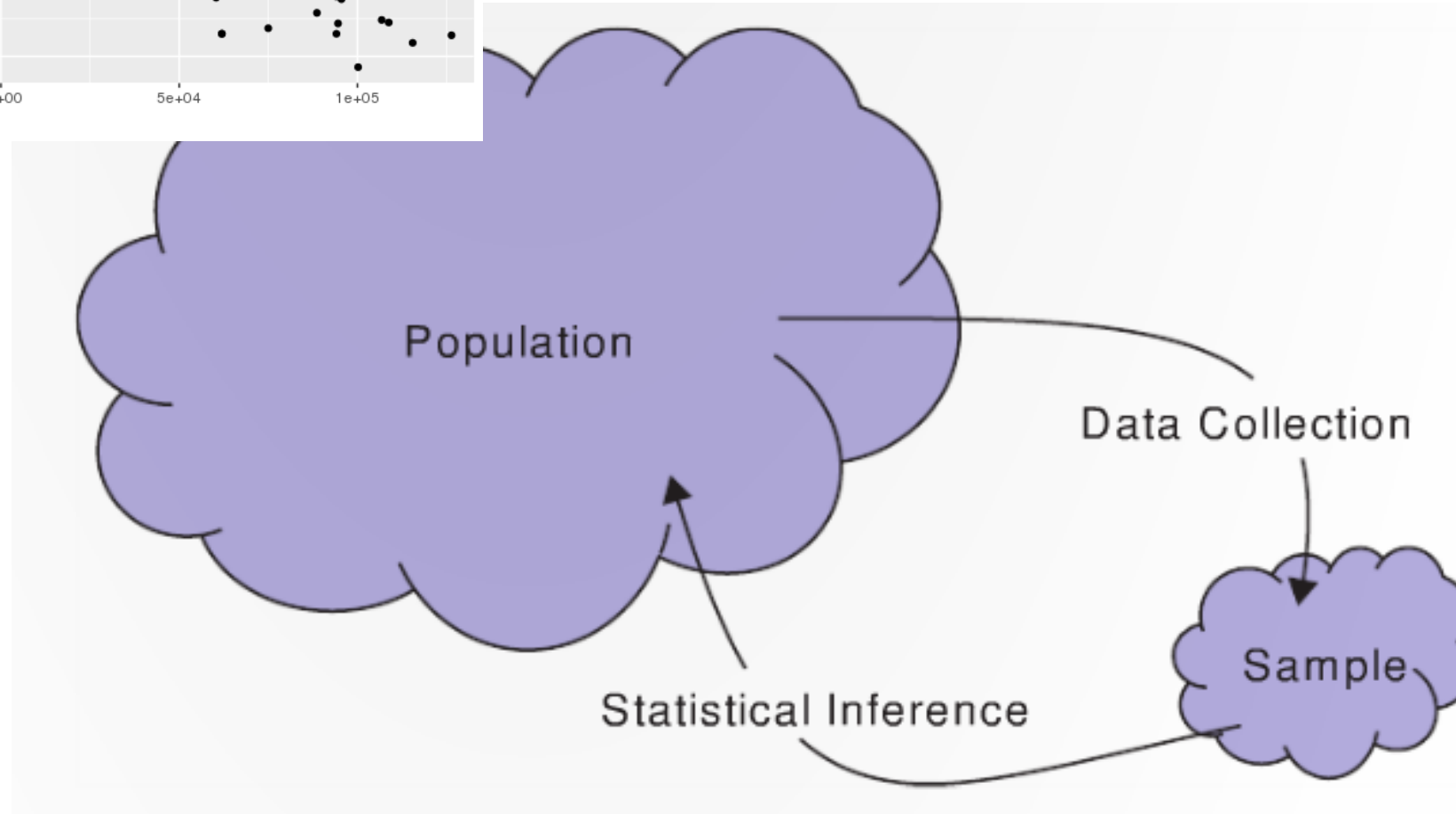
$$r = \frac{1}{(n - 1)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

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

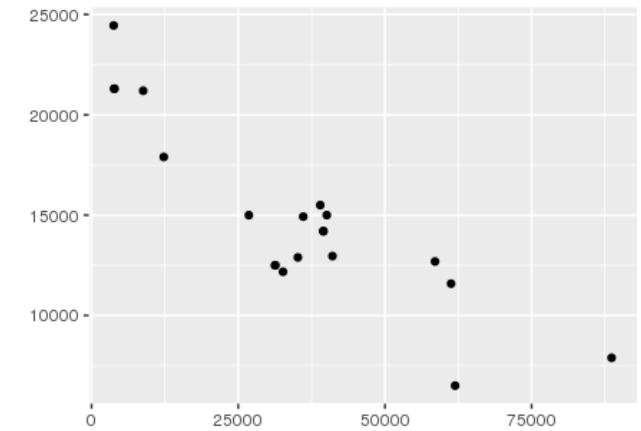
R: `cor(x, y)`



ρ parameter



r statistic

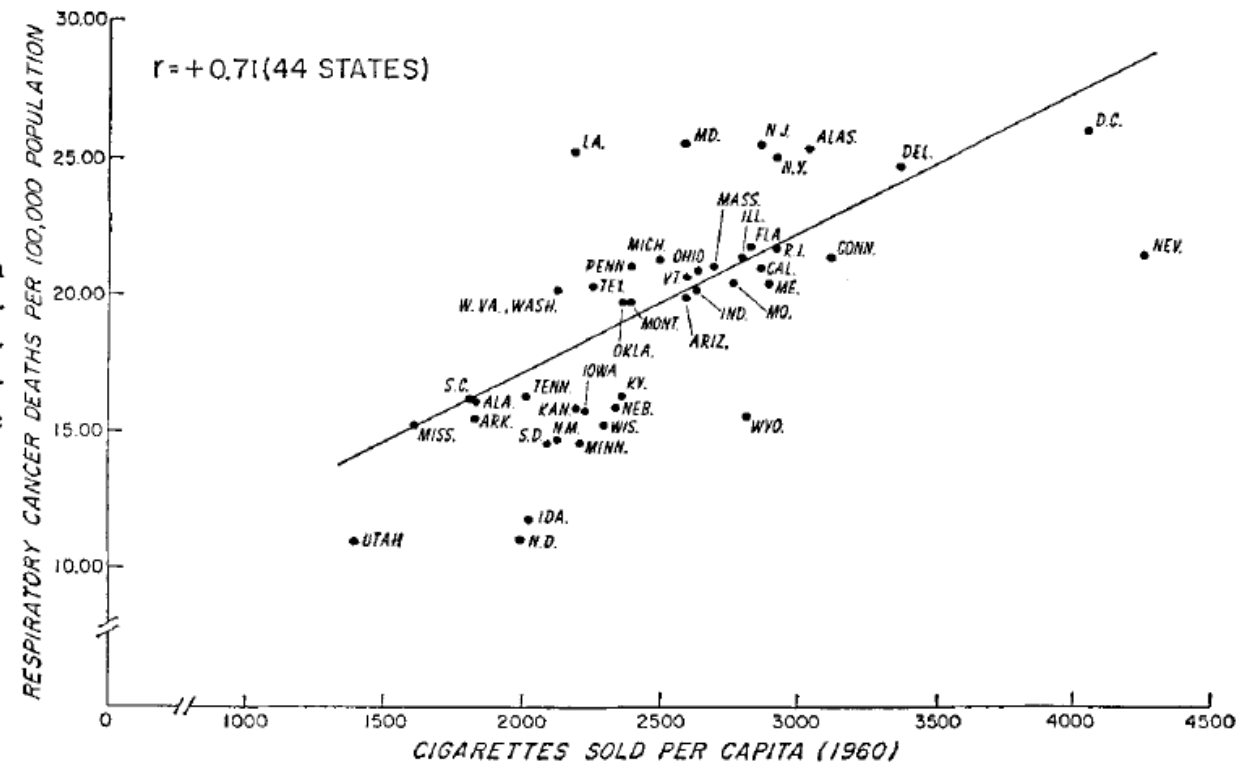


Smoking and lung cancer correlation?

The **correlation** is measure of the strength and direction of a linear association 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 is always between -1 and 1: $-1 \leq r \leq 1$

The sign of r indicates the direction of the association

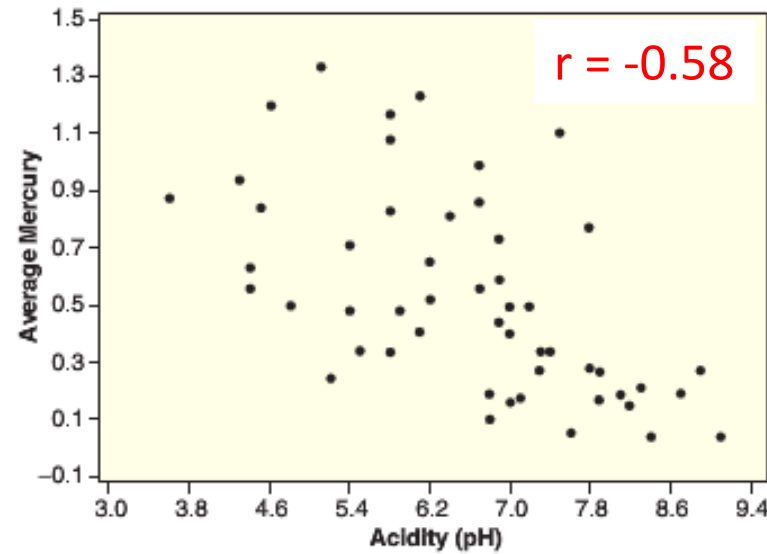
Values close to ± 1 show strong linear relationships, values close to 0 show no linear relationship

Correlation is symmetric: $r = \text{cor}(x, y) = \text{cor}(y, x)$

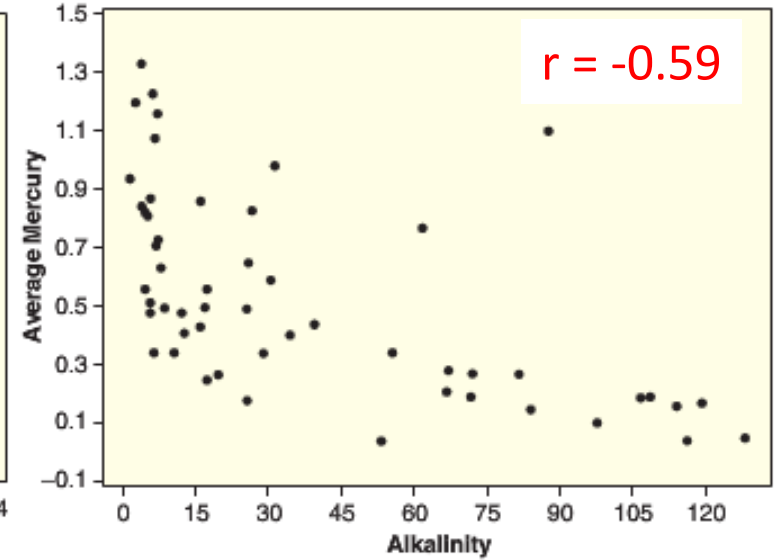
$$r = \frac{1}{(n-1)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Florida lakes

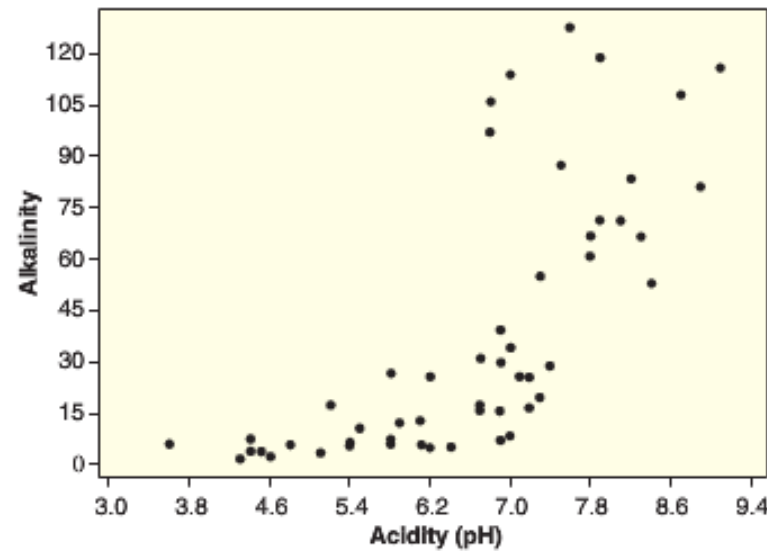
Correlation game



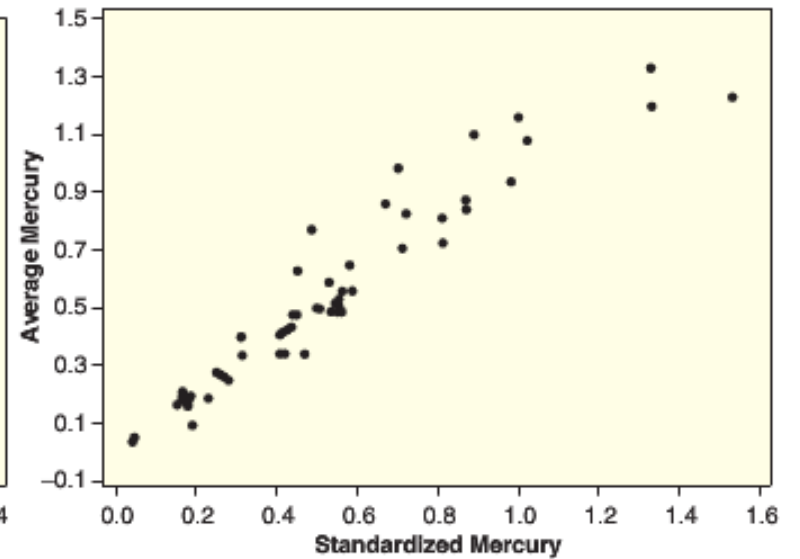
(a) Average mercury level vs acidity



(b) Average mercury level vs alkalinity



(c) Alkalinity vs acidity



(d) Average vs standardized mercury levels

Let's calculate some correlations in R

load the data

```
load("smoking_cancer.Rda")
```

create a scatter plot and calculate the correlation


```
plot(smoking$CIG, smoking$LUNG)
```

```
cor(smoking$CIG, smoking$LUNG)
```

	STATE	CIG	BLAD	LUNG	KID	LEUK
1	AL	1820	2.90	17.05	1.59	6.15
2	AZ	2582	3.52	19.80	2.75	6.61
3	AR	1824	2.99	15.98	2.02	6.94
4	CA	2860	4.46	22.07	2.66	7.06
5	CT	3110	5.11	22.83	3.35	7.20



Number of cigarette's sold per capita



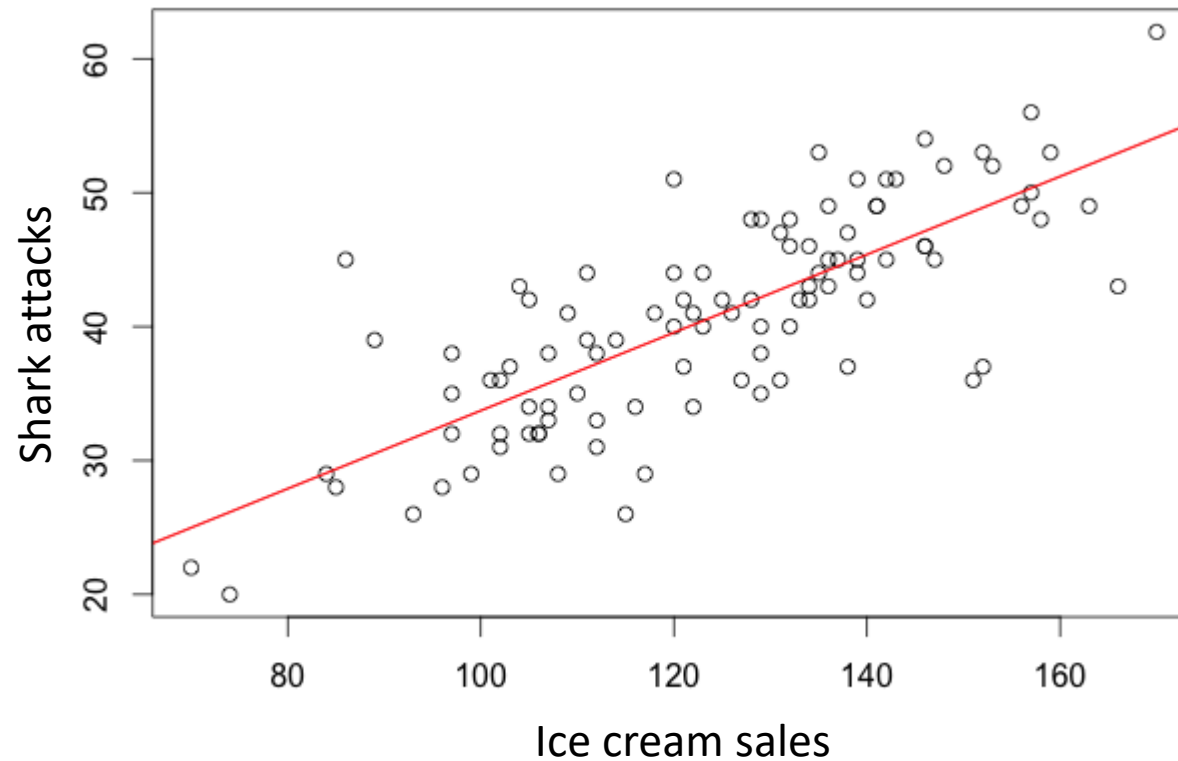
Cancer rates per 10k for bladder, lung, kidney and leukemia

Try it in R

Correlation cautions

Correlation caution #1

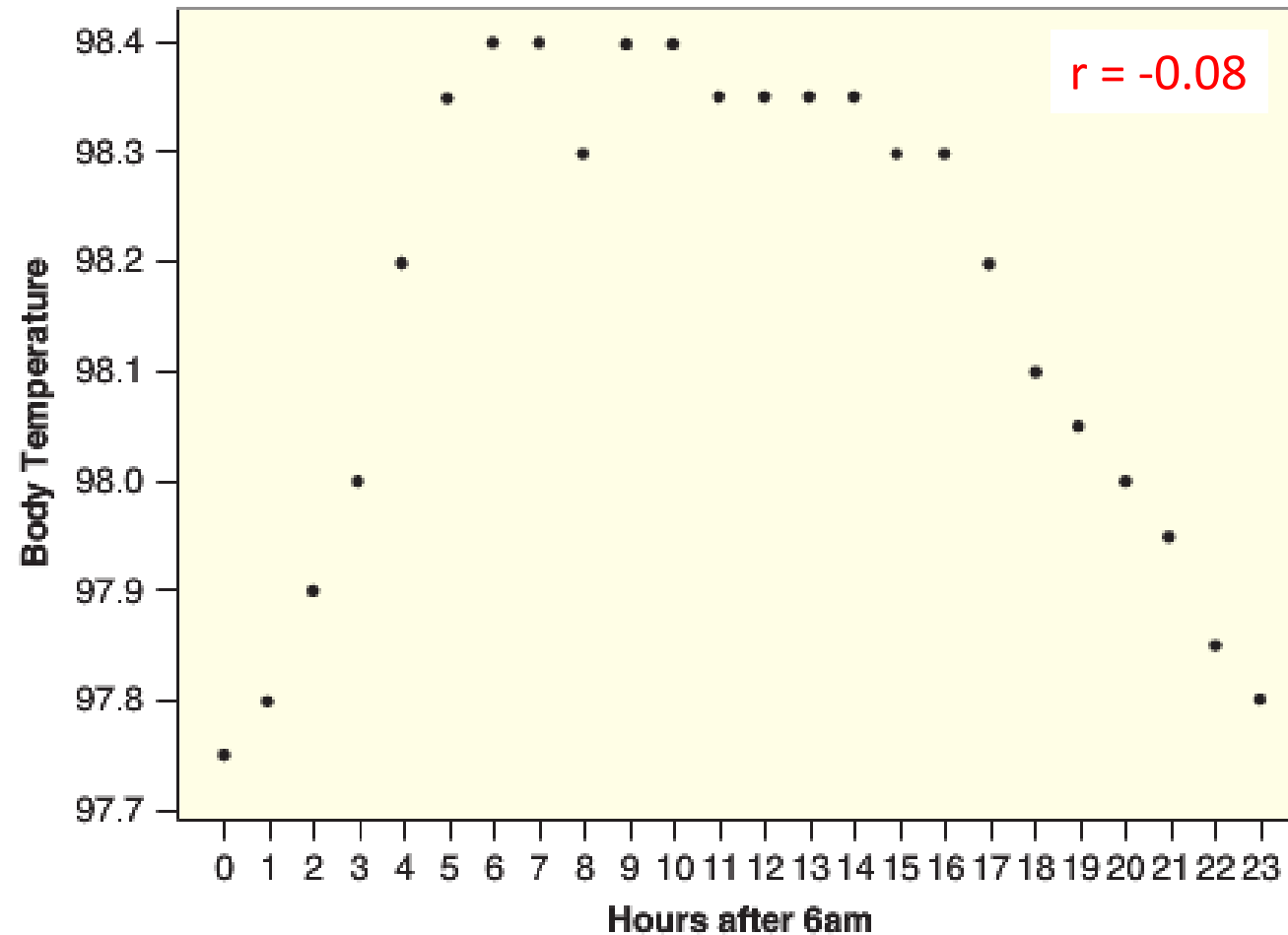
A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between two variables



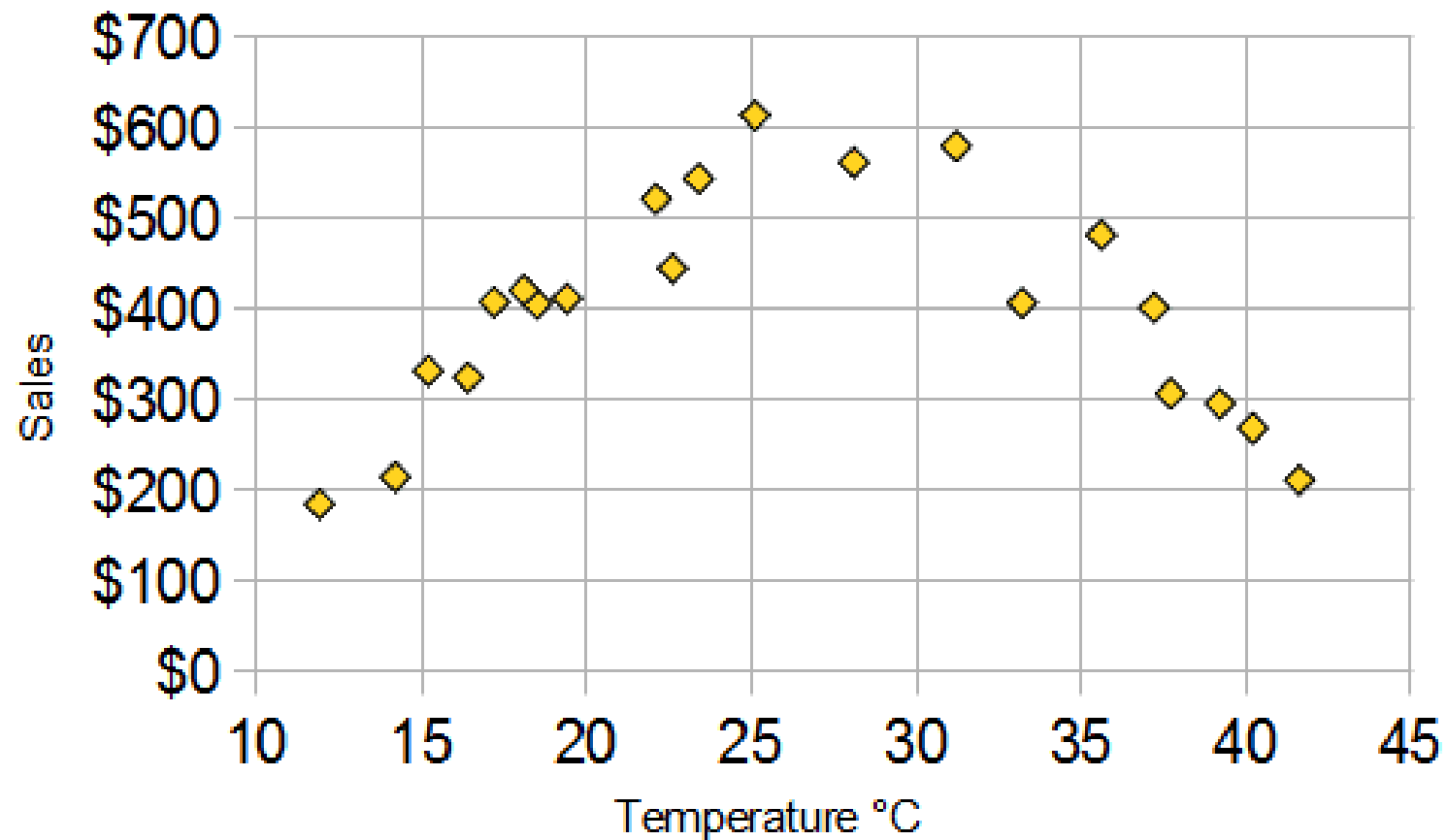
Correlation caution #2

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

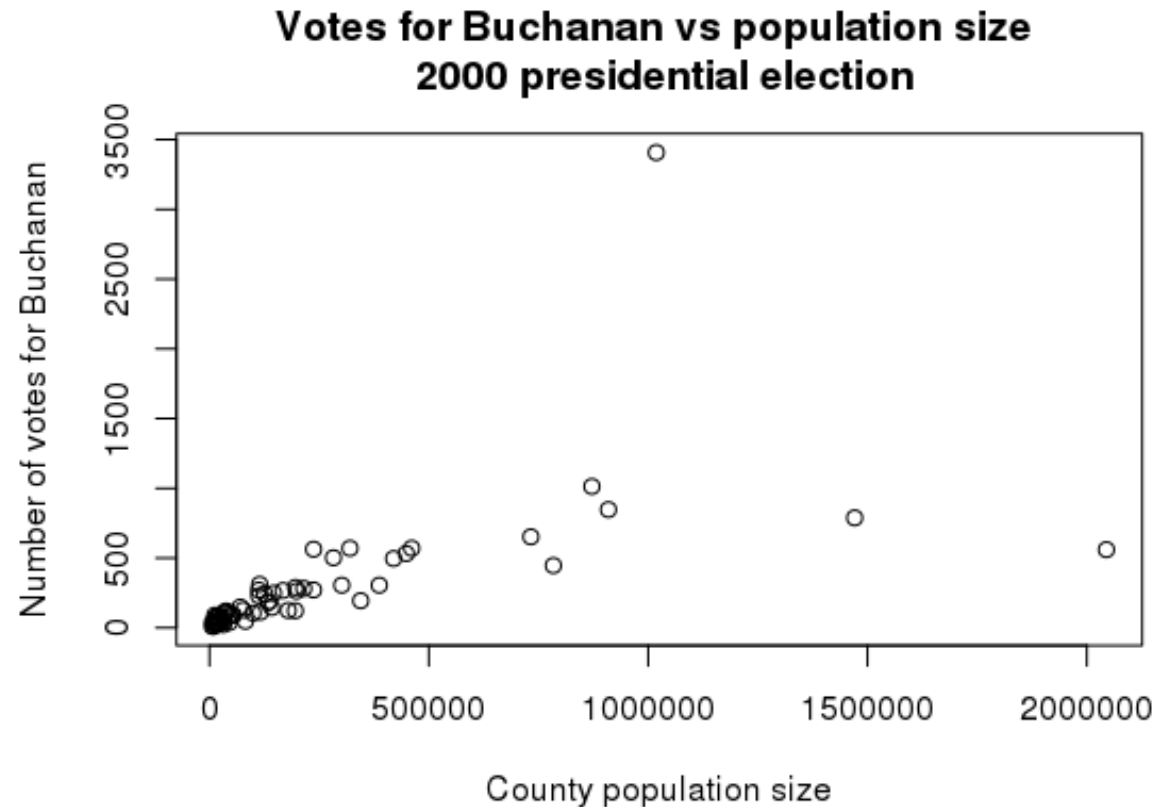


Ice cream sales and temperature



Correlation caution #3

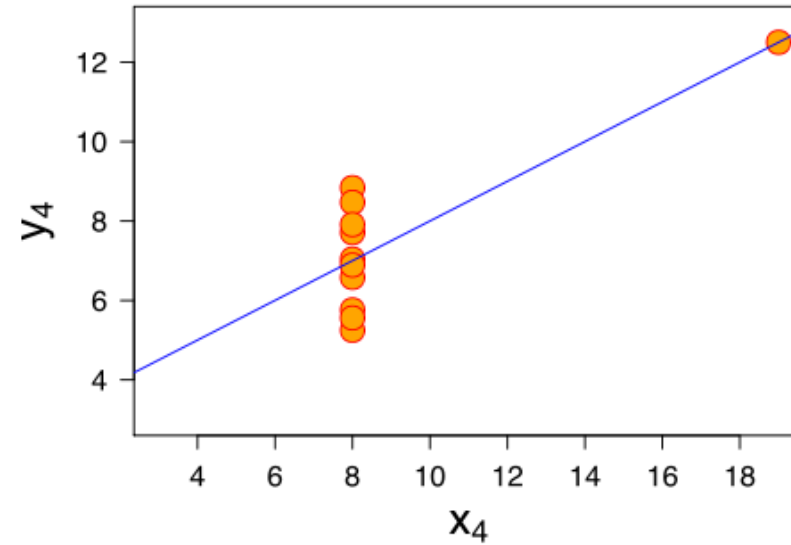
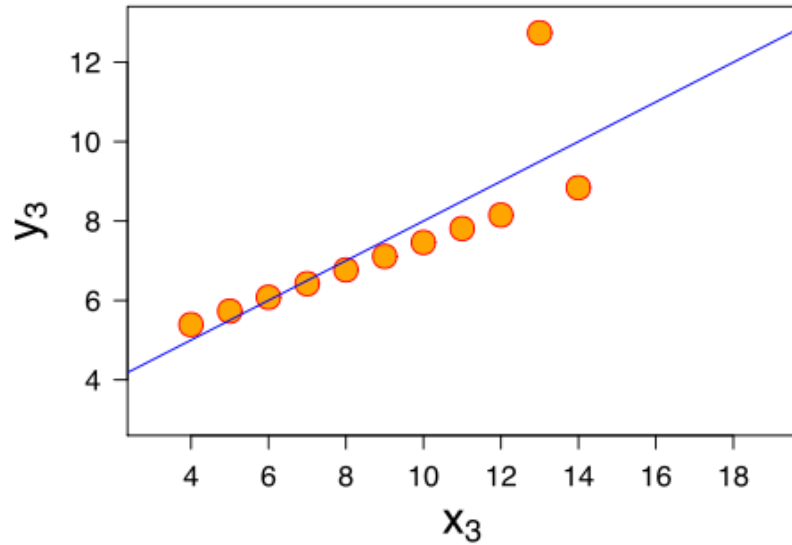
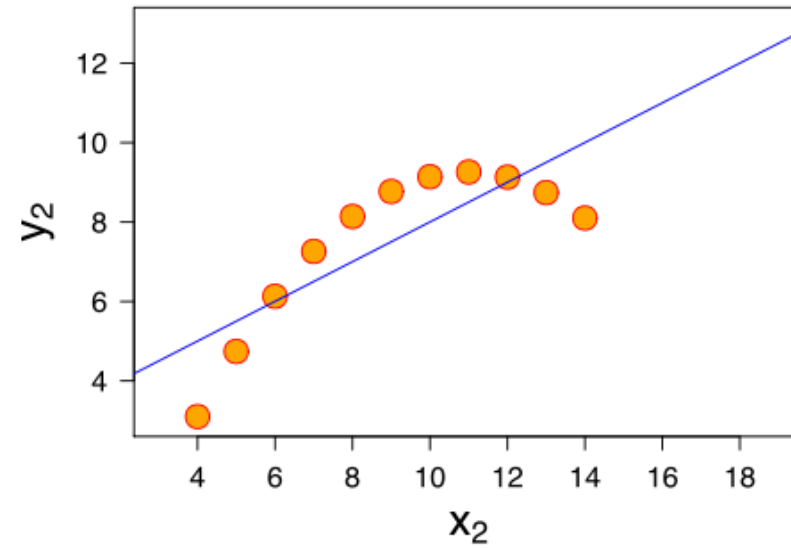
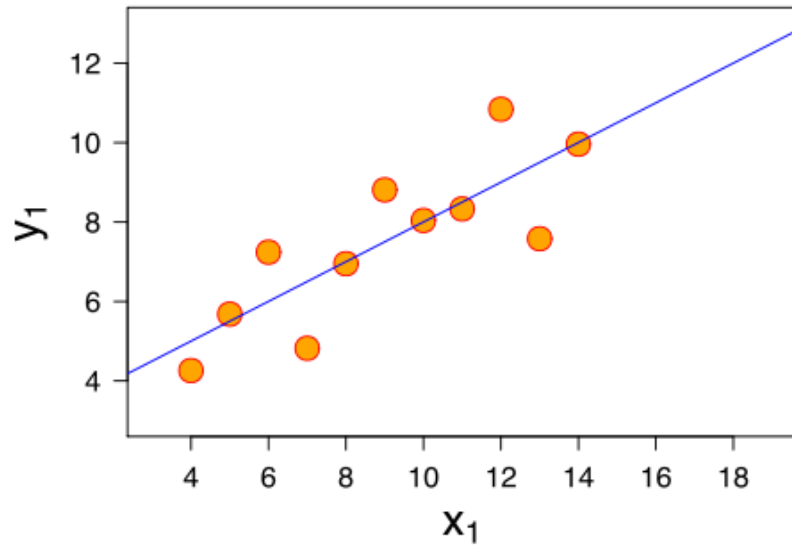
Correlation can be heavily influenced by outliers. Always plot your data!



With Palm Beach
 $r = 0.61$

Without Palm Beach
 $r = .78$

Anscombe's quartet ($r = 0.81$)



Linear regression

Regression

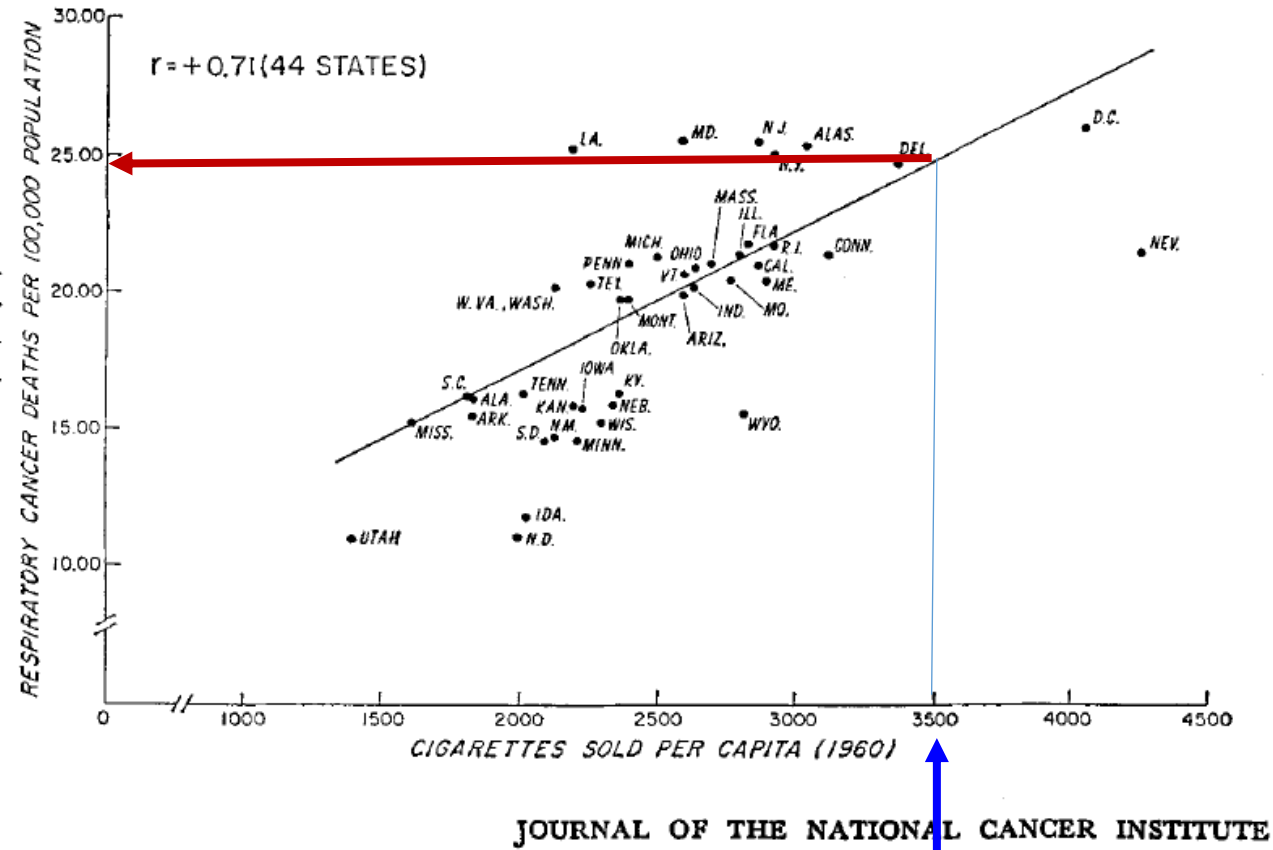
Regression is method of using one variable x to predict the value of a second variable y

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

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

Cigarette 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.

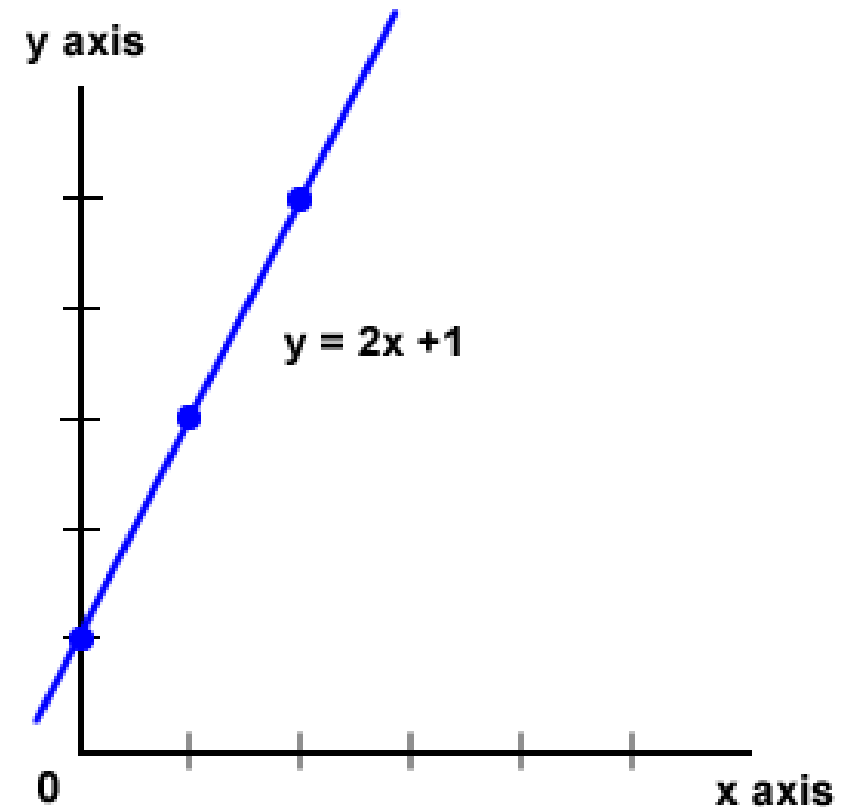


$$x_i = 3500$$

Equation for a line

What is the equation for a line?

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



Regression lines

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

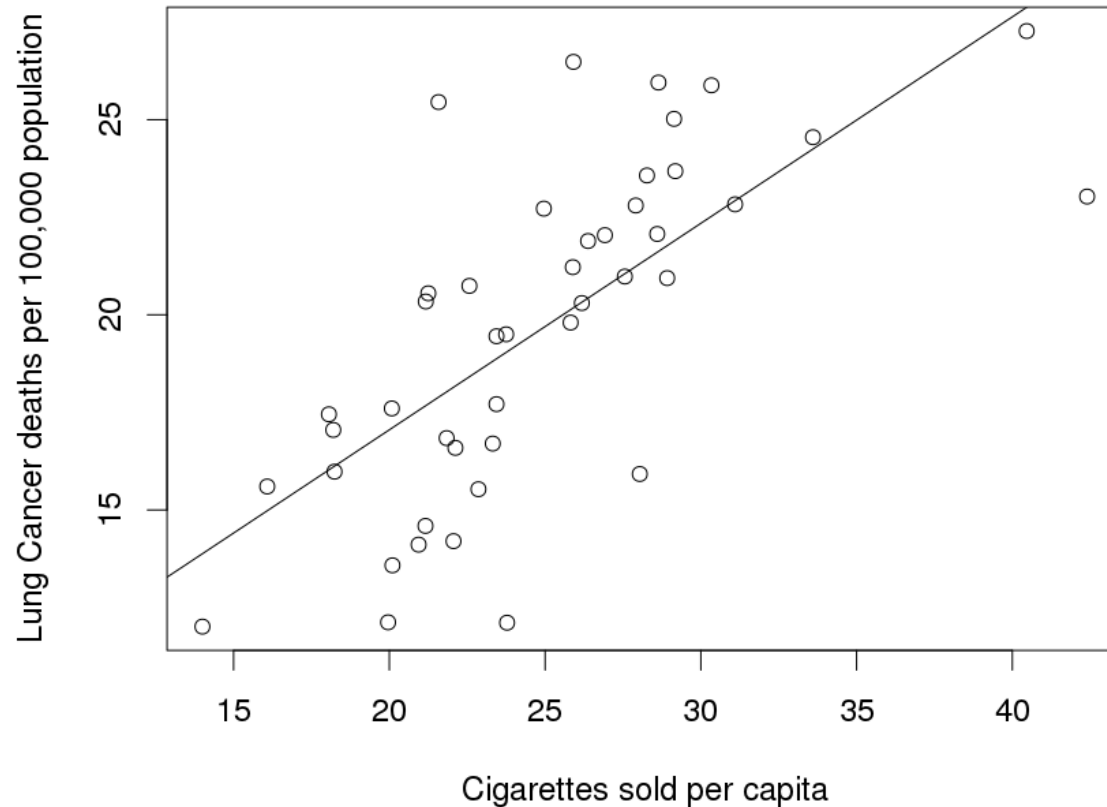
$$\textit{Response} = a + b \cdot \textit{Explanatory}$$

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

Relationship between cigarettes sold and cancer deaths



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

$$a = 6.47$$

$$b = 0.0053$$

$$R: \text{lm}(y \sim x)$$

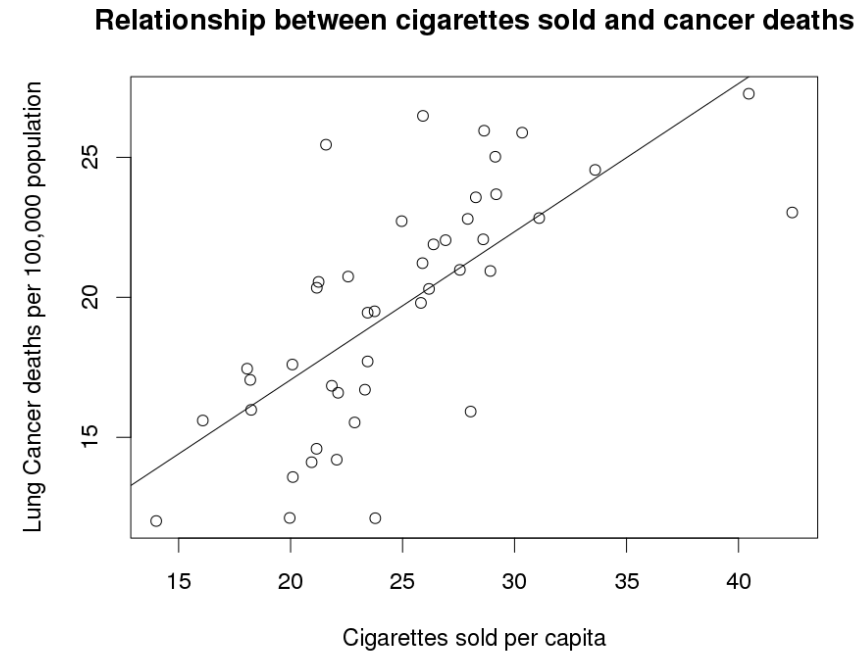
Using the regression line to make predictions

If a state sold 25 (hundred) cigarettes per person

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

$$a = 6.47, \quad b = .0053$$

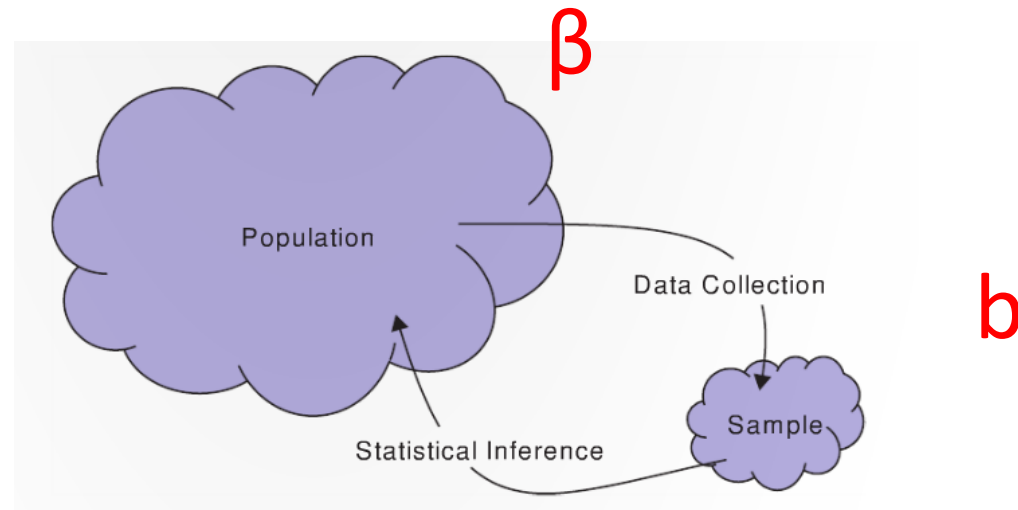
$$\hat{y} = 6.47 + .0053 \cdot x$$



Notation

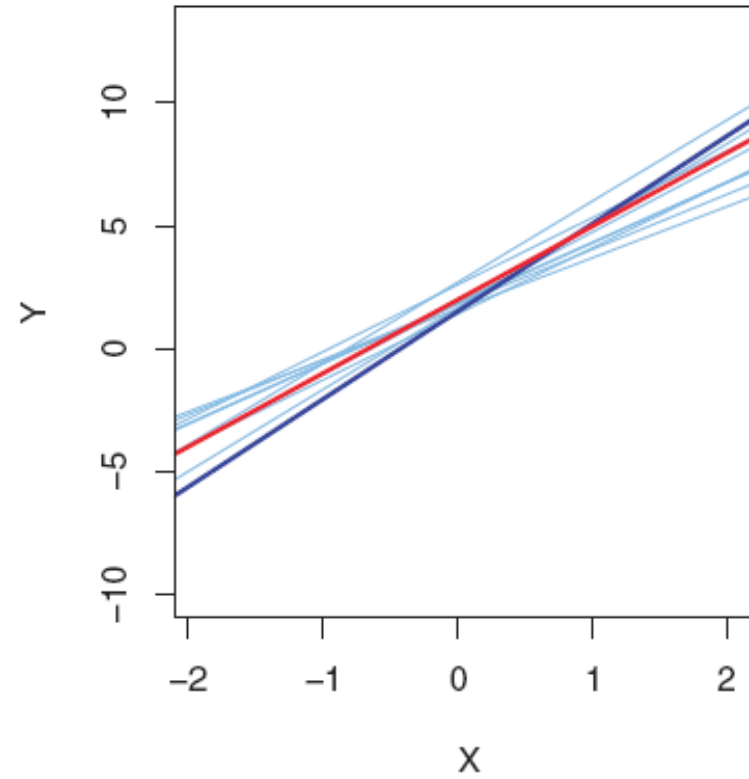
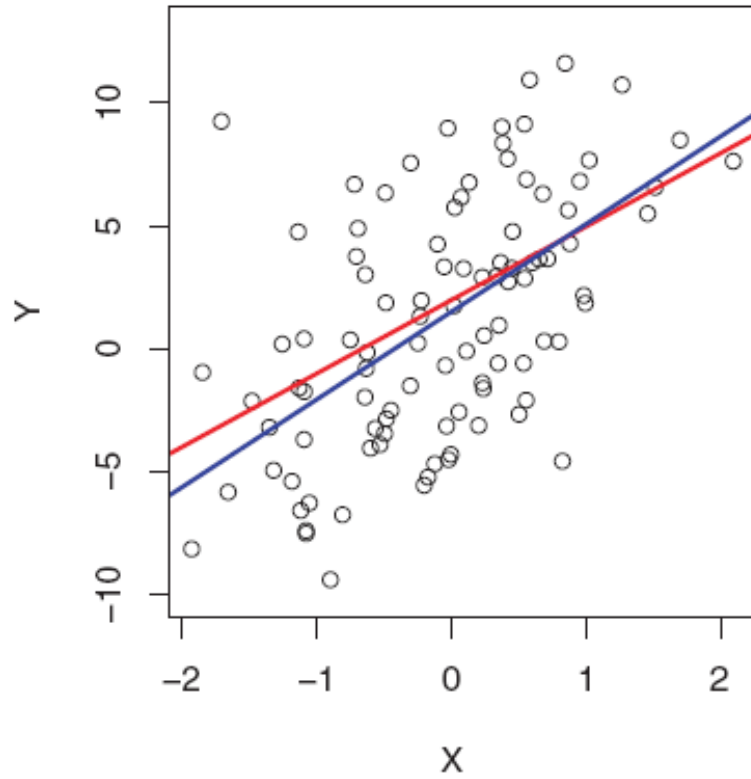
The letter **b** is typically used to denote the slope of the sample

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



Population: β

Sample estimates: b



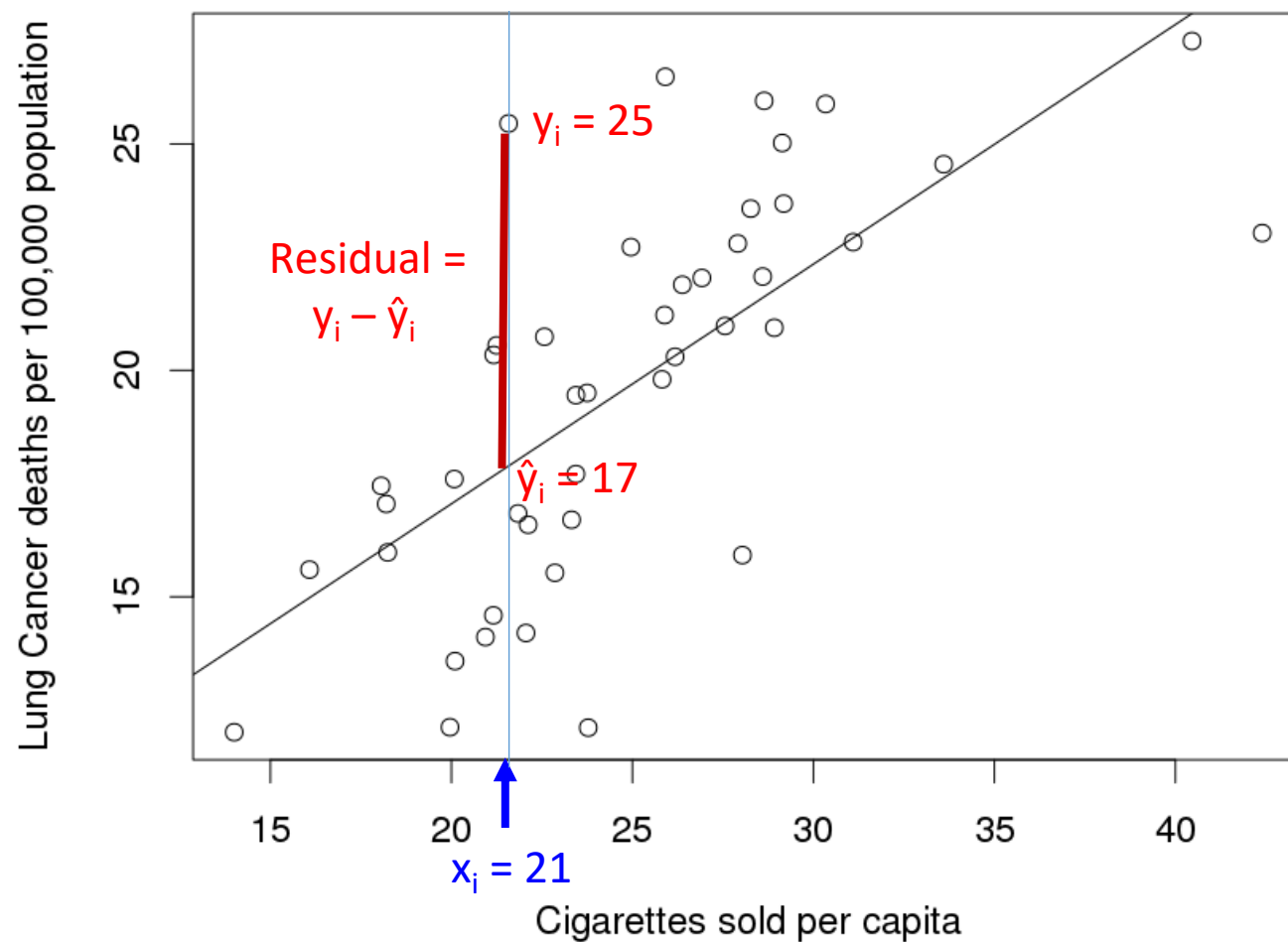
Residuals

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

Relationship between cigarettes sold and cancer deaths

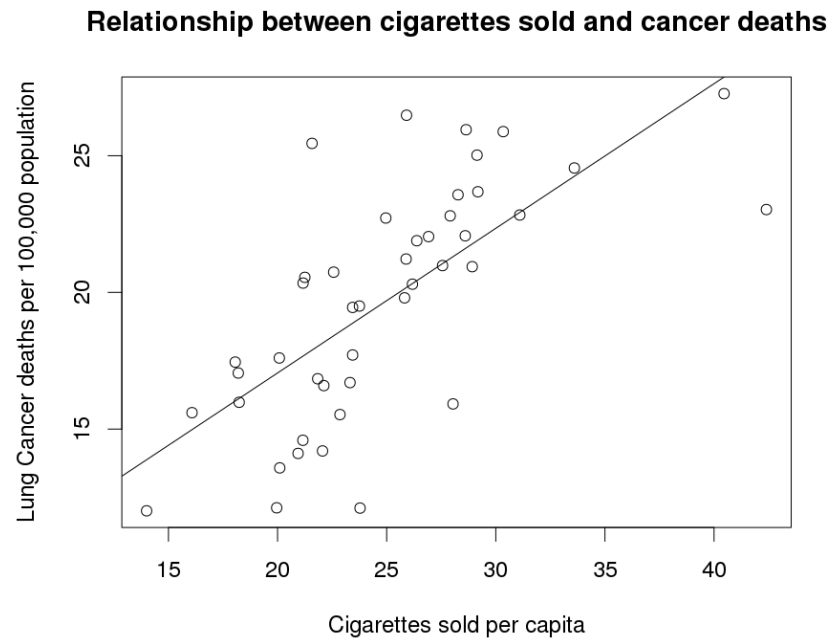


Cancer smoking residuals

Cancer obs (y)	Cancer pred (\hat{y})	Residuals ($y - \hat{y}$)
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

Least squares line

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

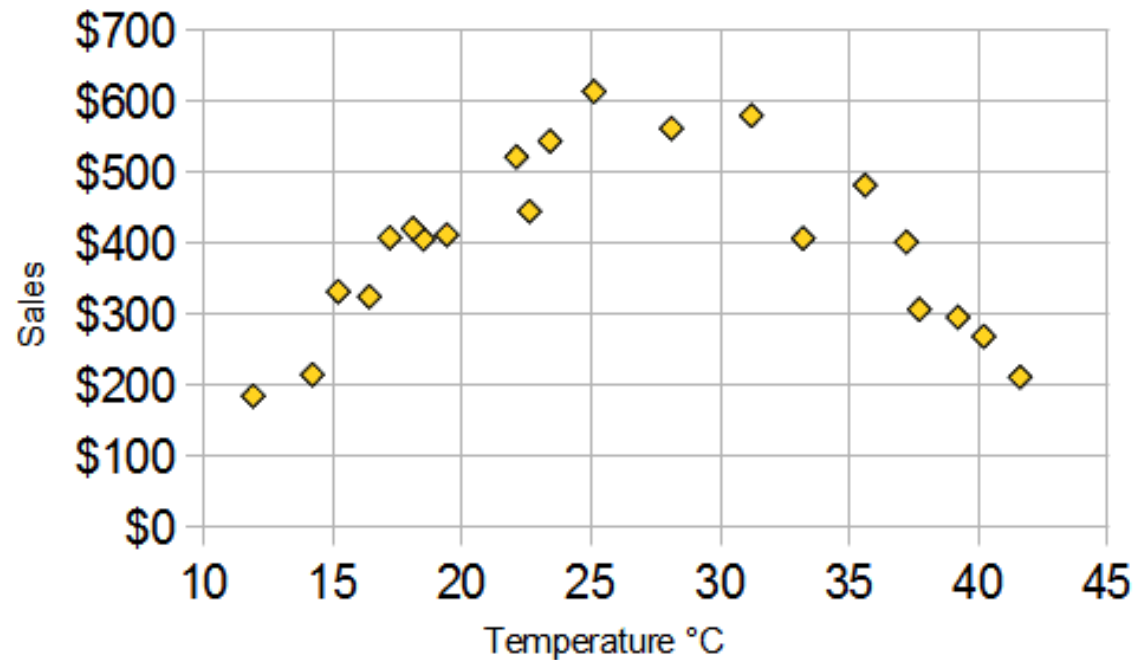
Cancer obs (y)	Cancer pred (\hat{y})	Residuals ($y - \hat{y}$)	Residuals² ($y - \hat{y}$)²
17.05	16.10	0.95	0.90
19.80	20.13	-0.33	0.11
15.98	16.12	-0.14	0.02
22.07	21.60	0.47	0.22
22.83	22.93	-0.10	0.01
24.55	24.25	0.30	0.09
27.27	27.88	-0.61	0.37
23.57	21.24	2.14	4.59

Regression caution # 1

Avoid trying to apply the regression line to predict values far from those that were used to create the line. i.e., do not extrapolate too far

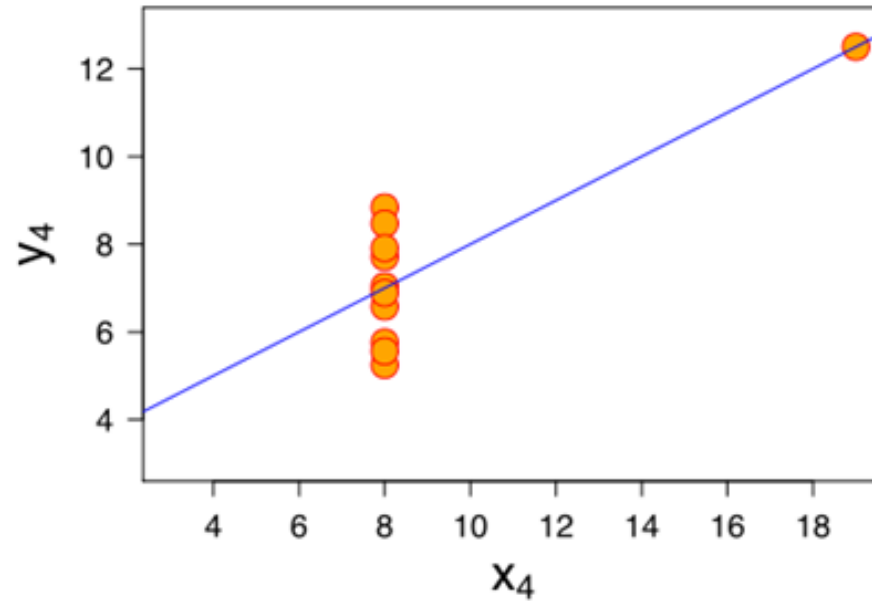
Regression caution # 2

Plot the data! Regression lines are only appropriate when there is a linear trend in the data.



Regression caution #3

Be aware of outliers – they can have an huge effect on the regression line.



Linear regression in R

Regression lines in R

load the data

```
load("states_smoking.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)
```

get the a and b coefficients

```
coef(lm_fit)
```

add a 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