SOEE1475 Statistics and Data Analysis
Lecture 3: Bivariate statistics

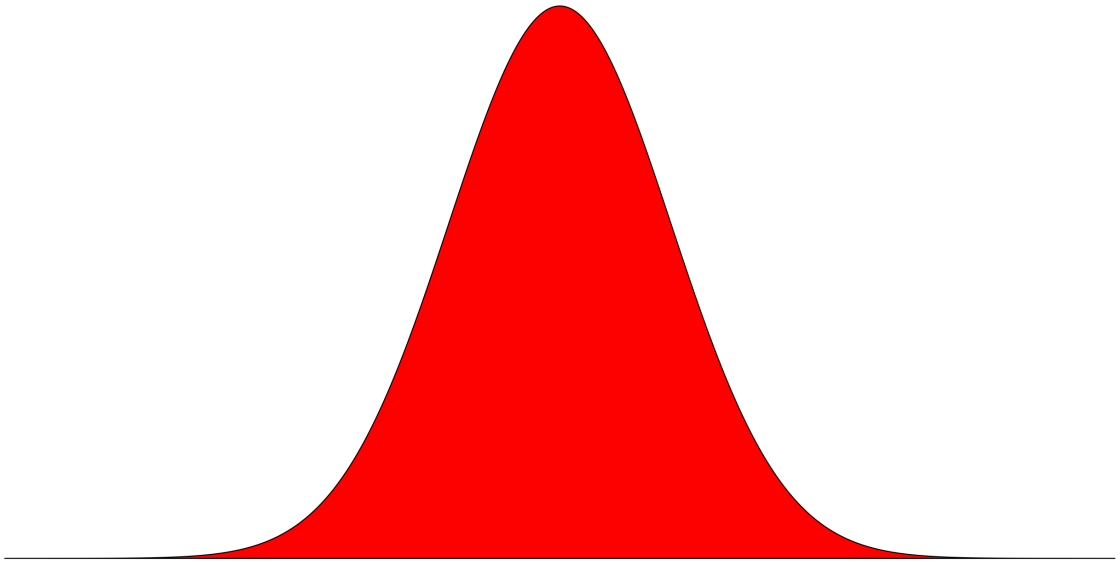
Graeme T. Lloyd

Today

- Introduction to bivariate data
- Regression
- Correlation

Univariate data

Population



Sample

Bivariate data

Dependent / Response variable

The variable we wish to explain; by convention *y*

Independent / Explanatory variable

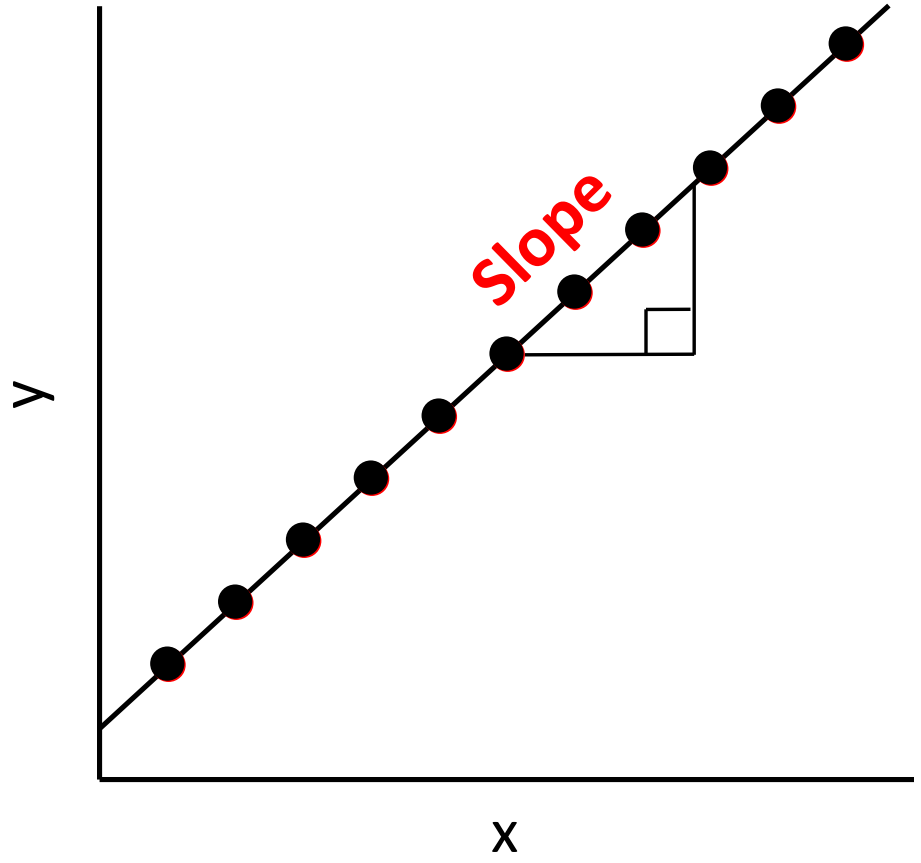
The variable we believe explains y; by convention *x*

Bivariate data

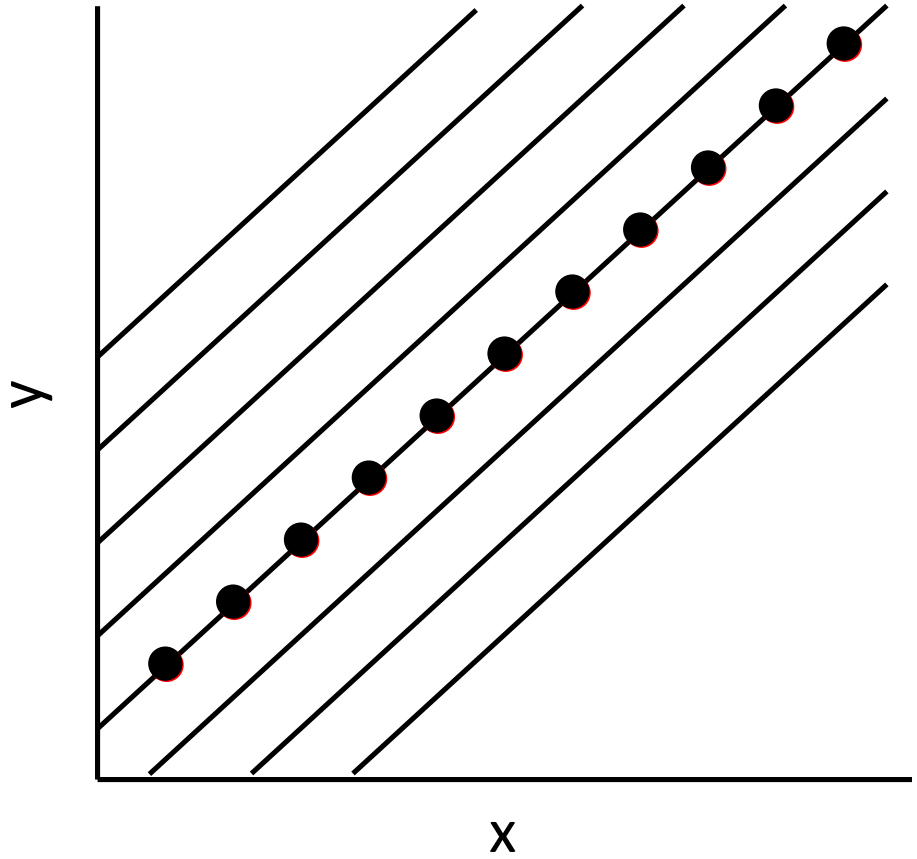
Atmospheric CO_2

Global temperature

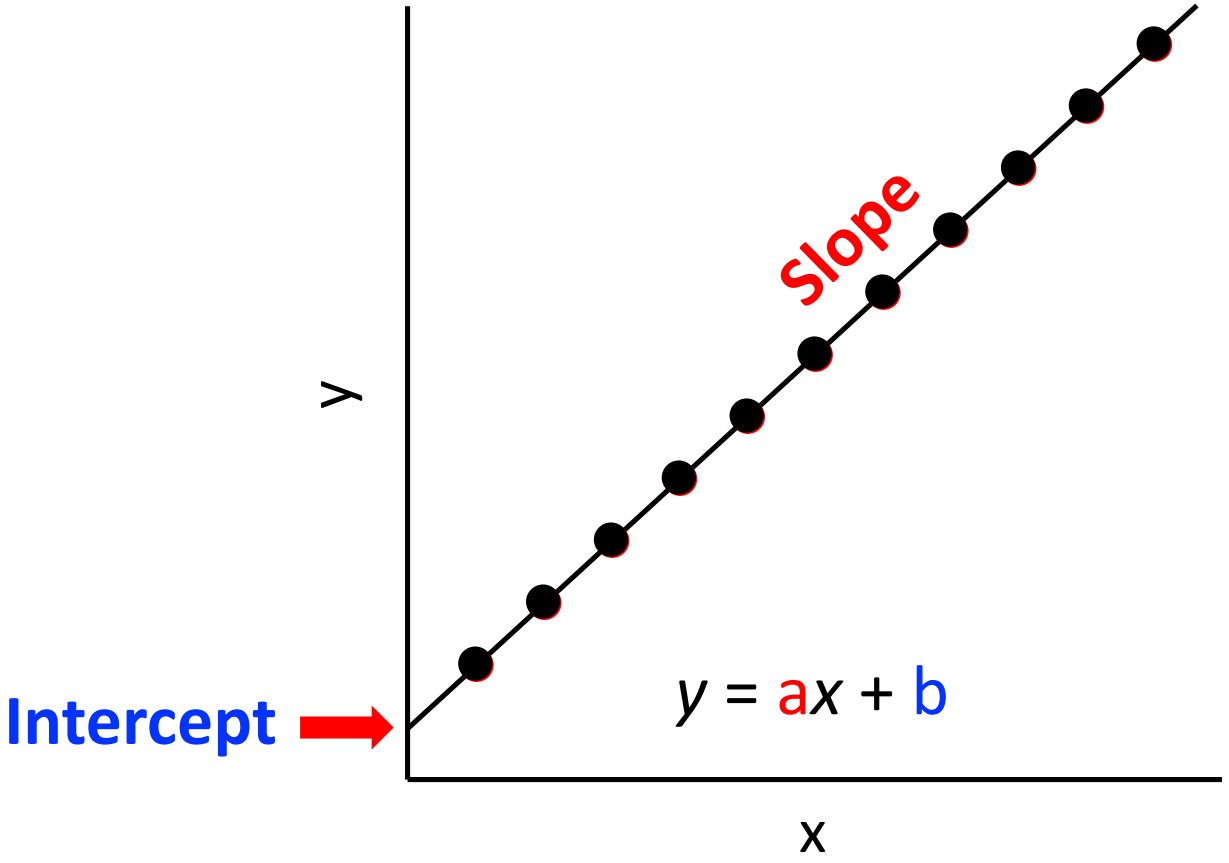
Linear regression



Linear regression

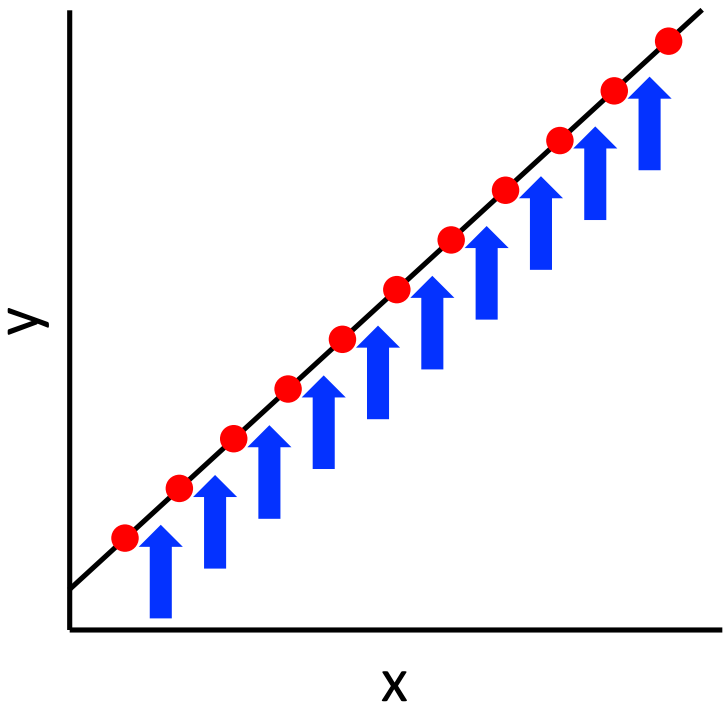


Linear regression

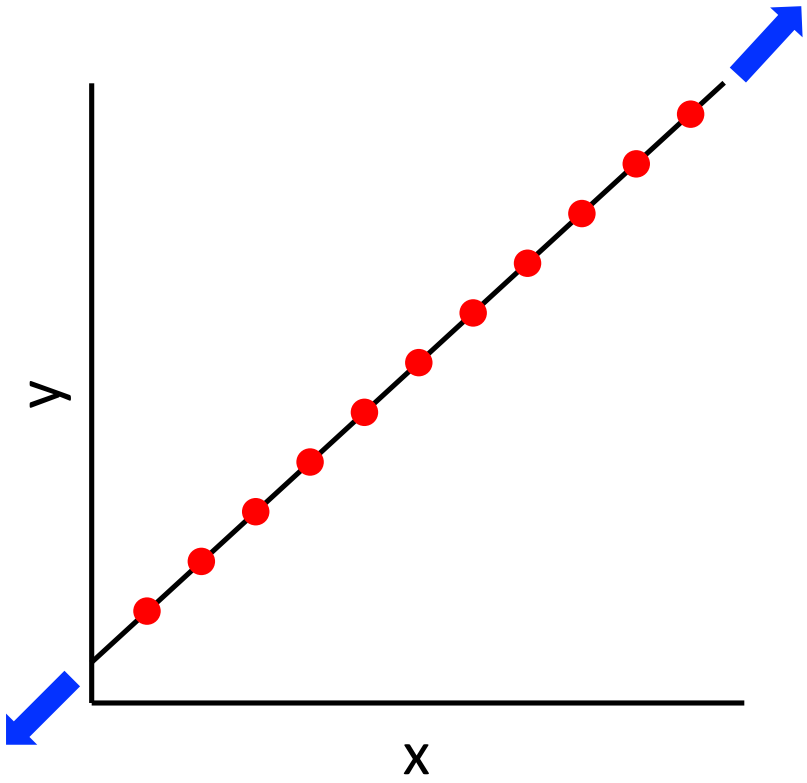


Linear regression

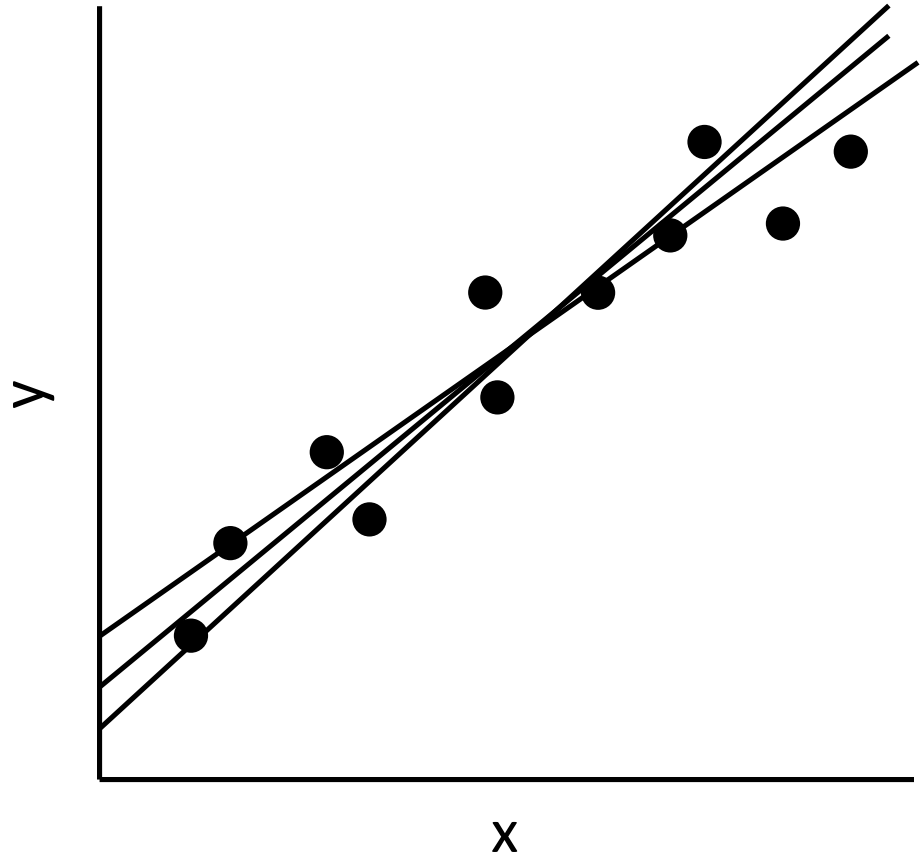
Interpolation



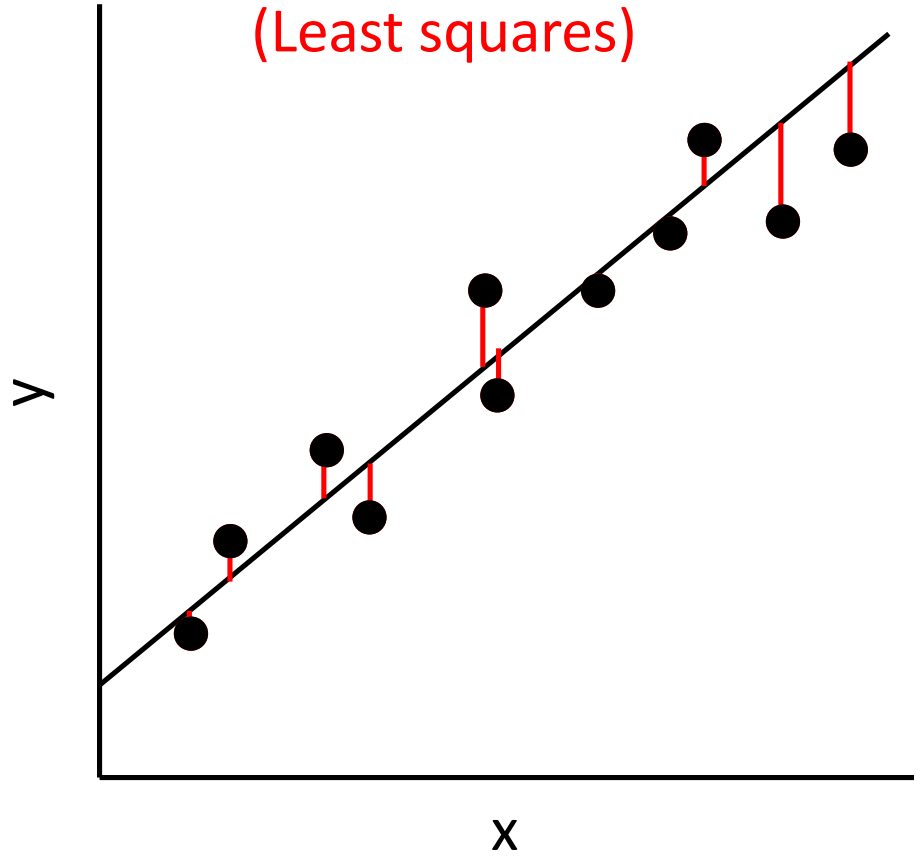
Extrapolation



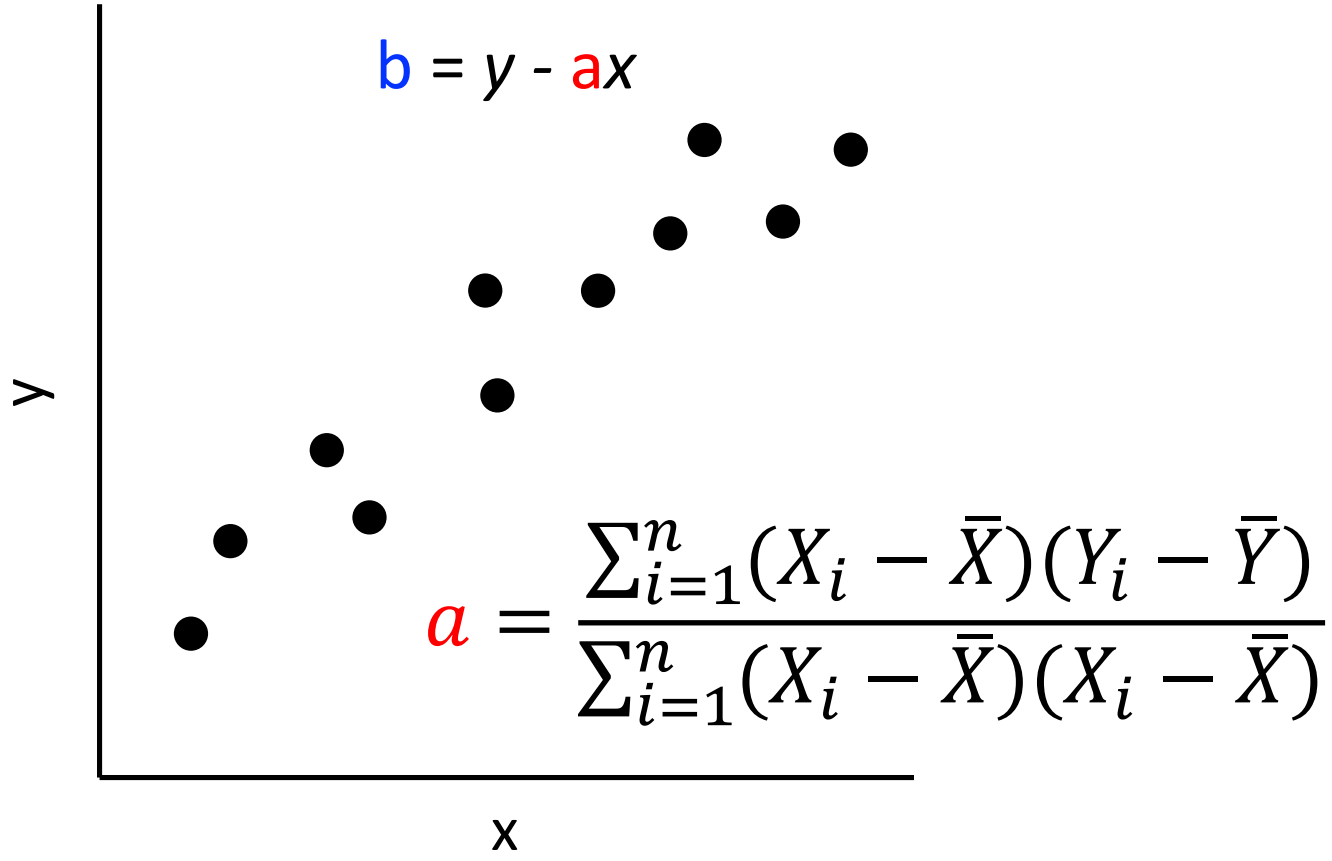
Linear regression



Linear regression



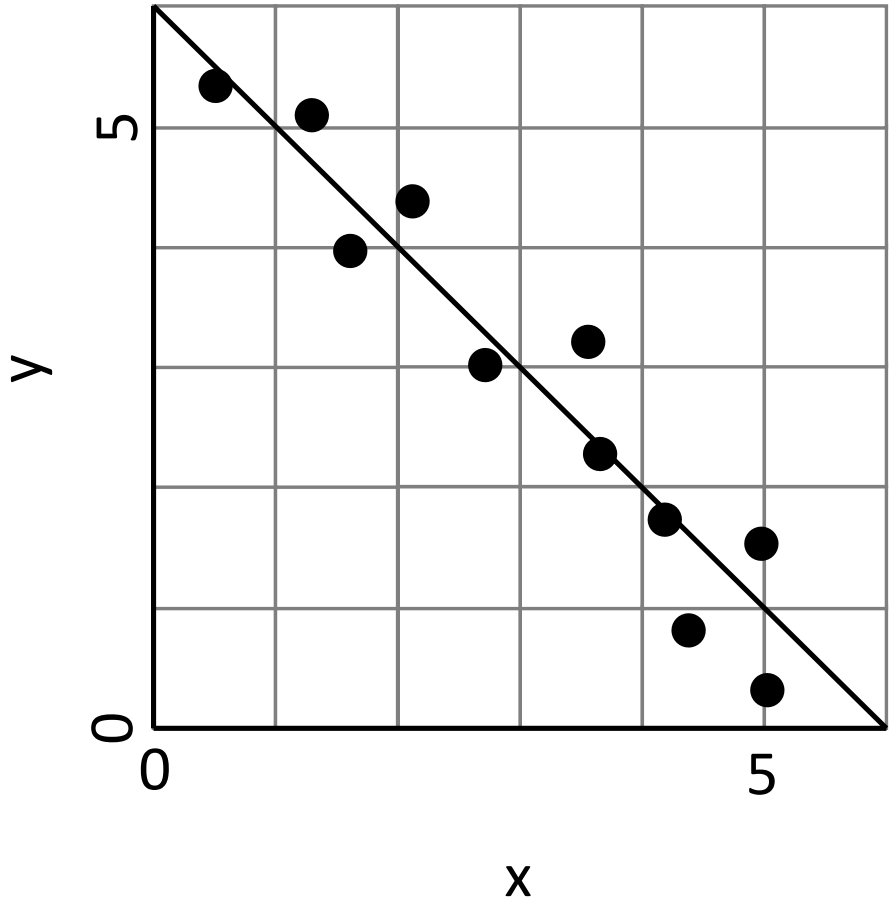
Linear regression



Linear regression

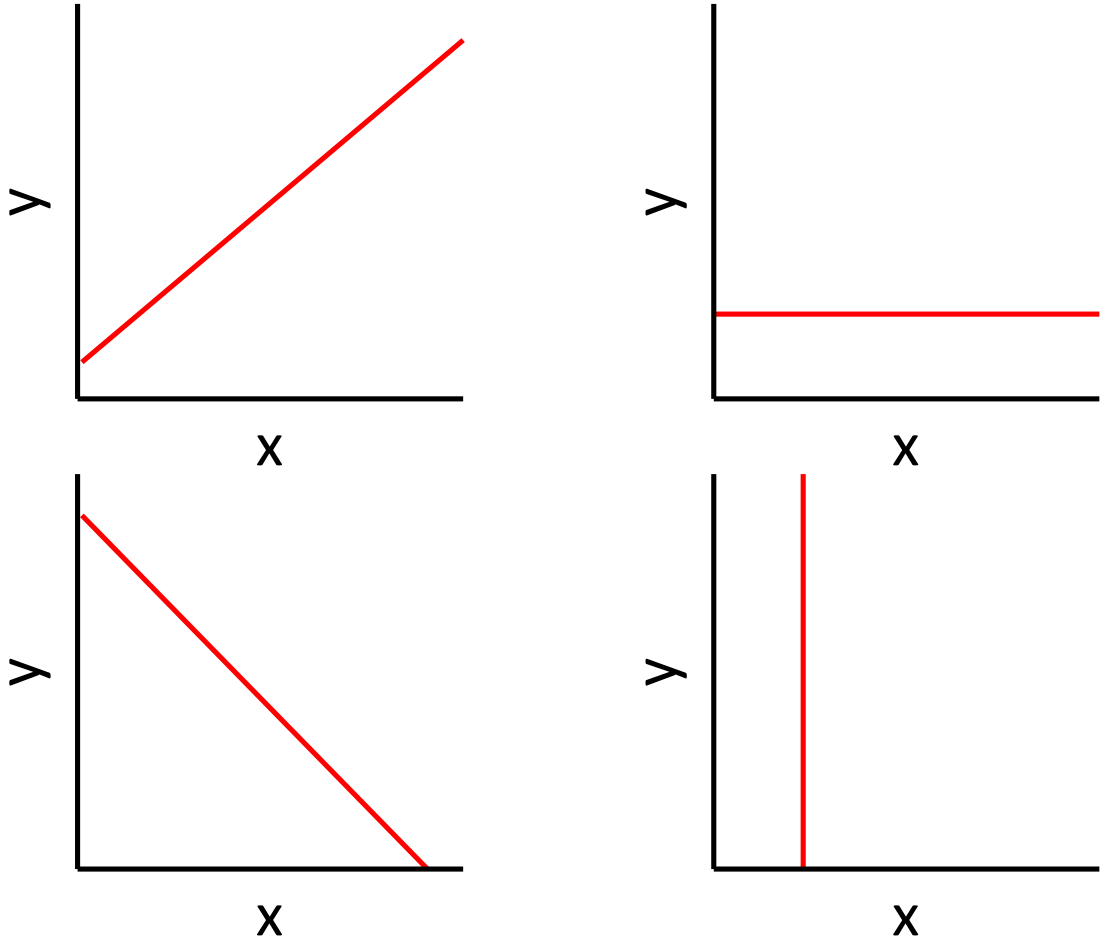
```
slope <- lm(y ~ x)$coefficients[2]
intercept <- lm(y ~ x)$coefficients[1]
```

Linear regression

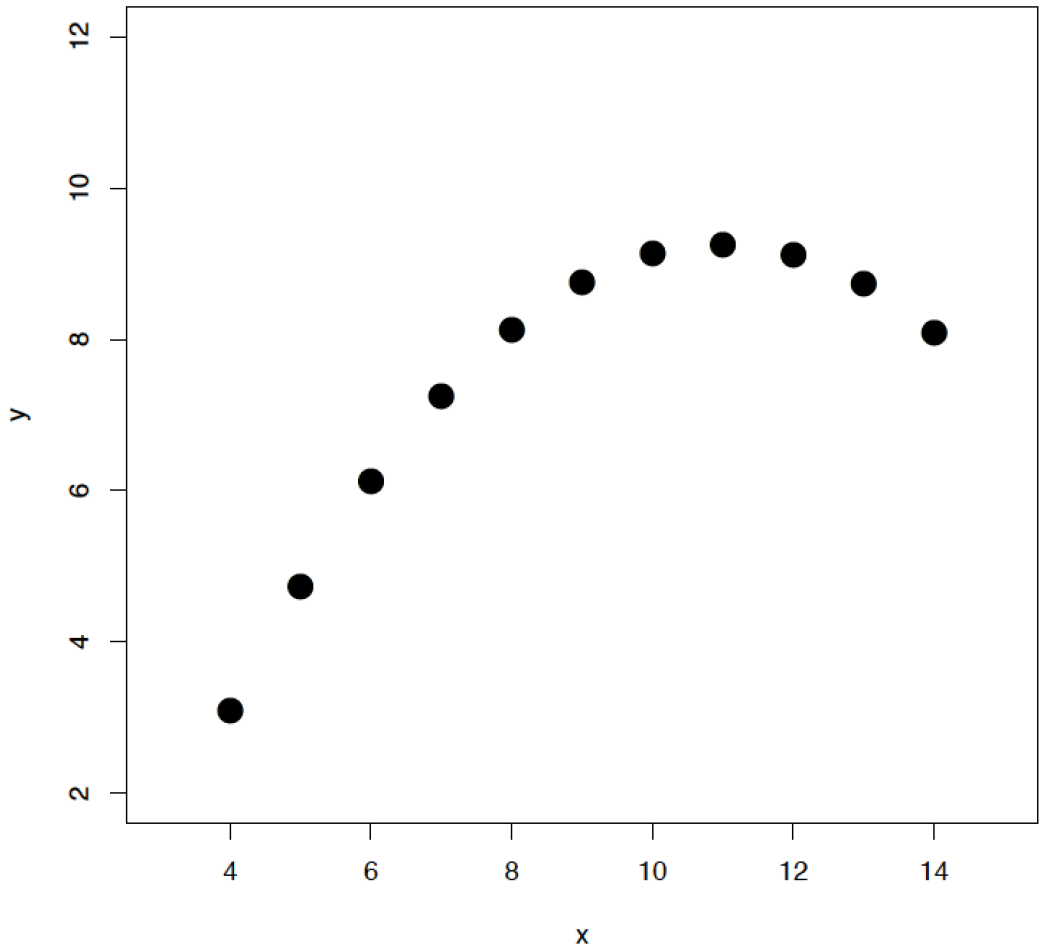


$y = -1x + 6$

Interpreting slopes

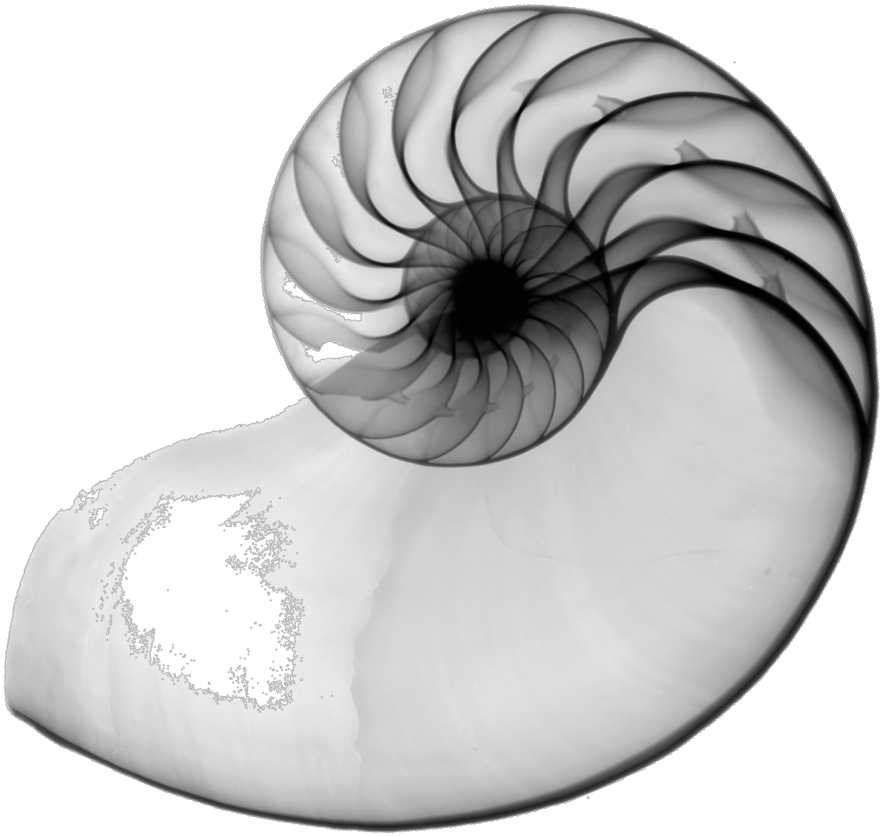


More complex regressions

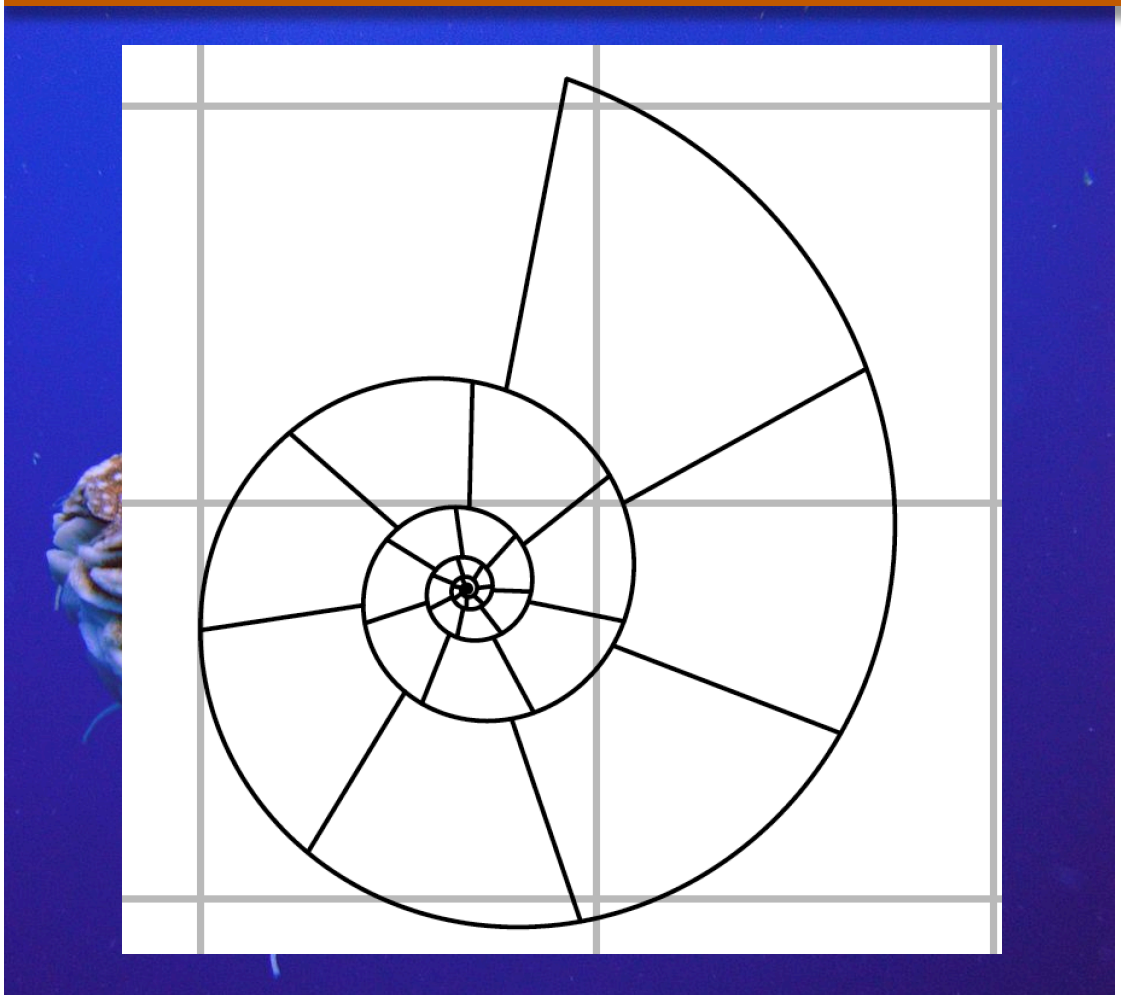


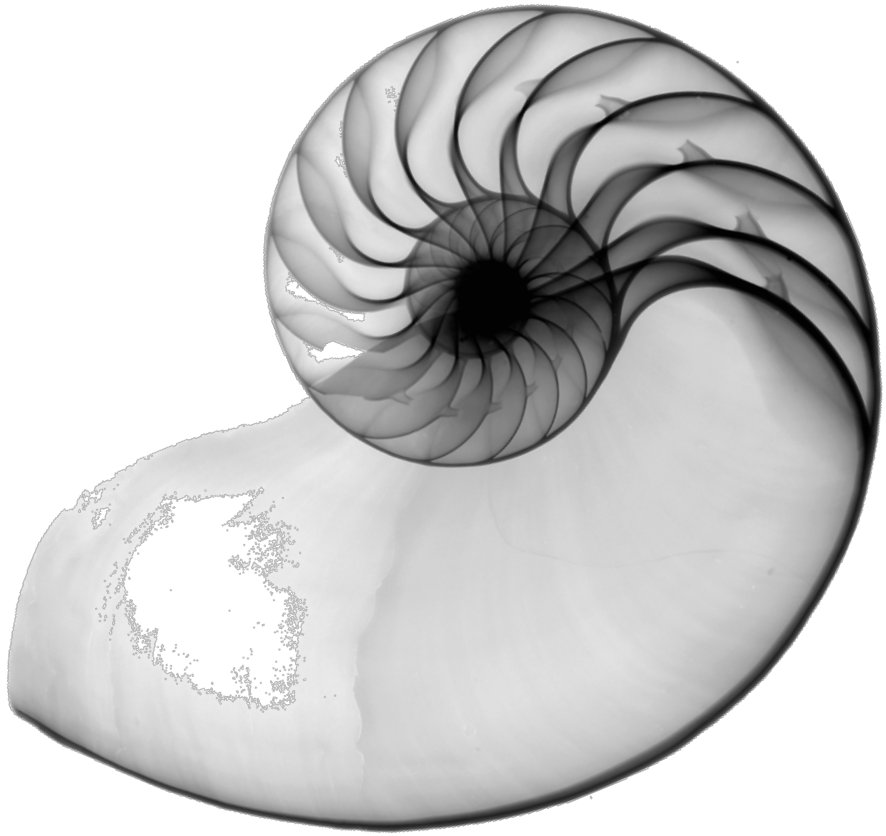

```
lm(y ~ poly(x, 2))
lm(y ~ poly(x, 3))
lm(y ~ poly(x, 4))
```

Exercise



Exercise



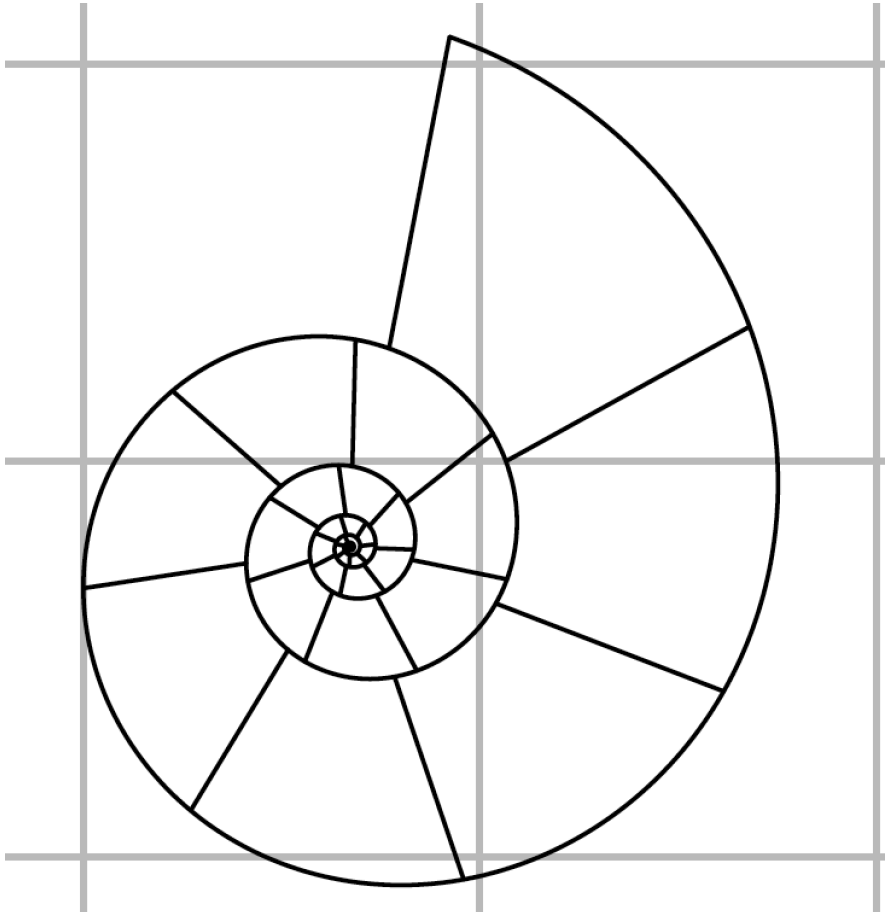


Exercise

N Chambers

Diameter (mm)

Exercise

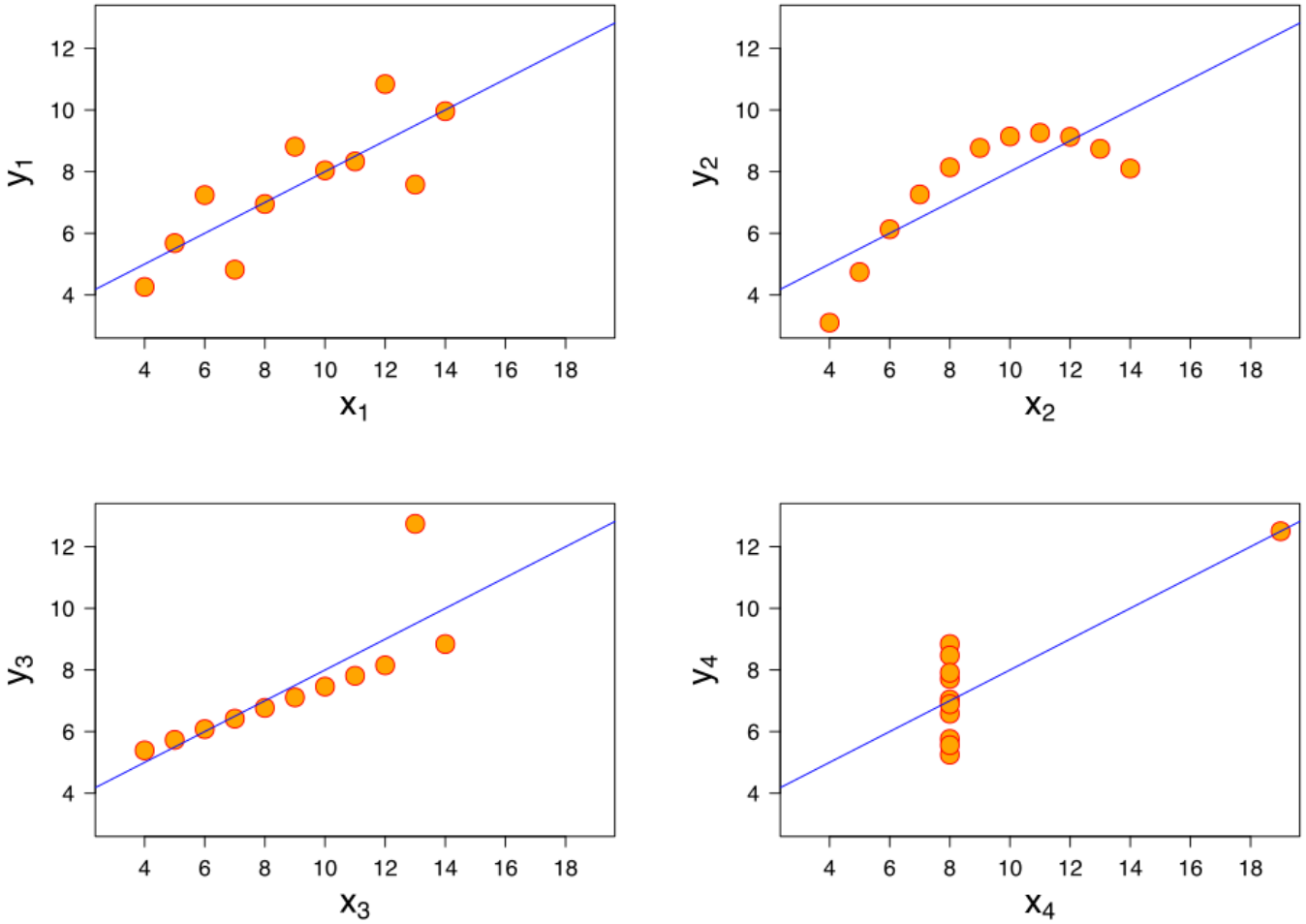


```
x <- c(20, 22, 25, 28, 30)
y <- c(70, 90, 125, 190, 280)
```

$$a = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}$$

$$b = \bar{y} - a\bar{x}$$

Anscombe's quartet



Measuring relationships between variables

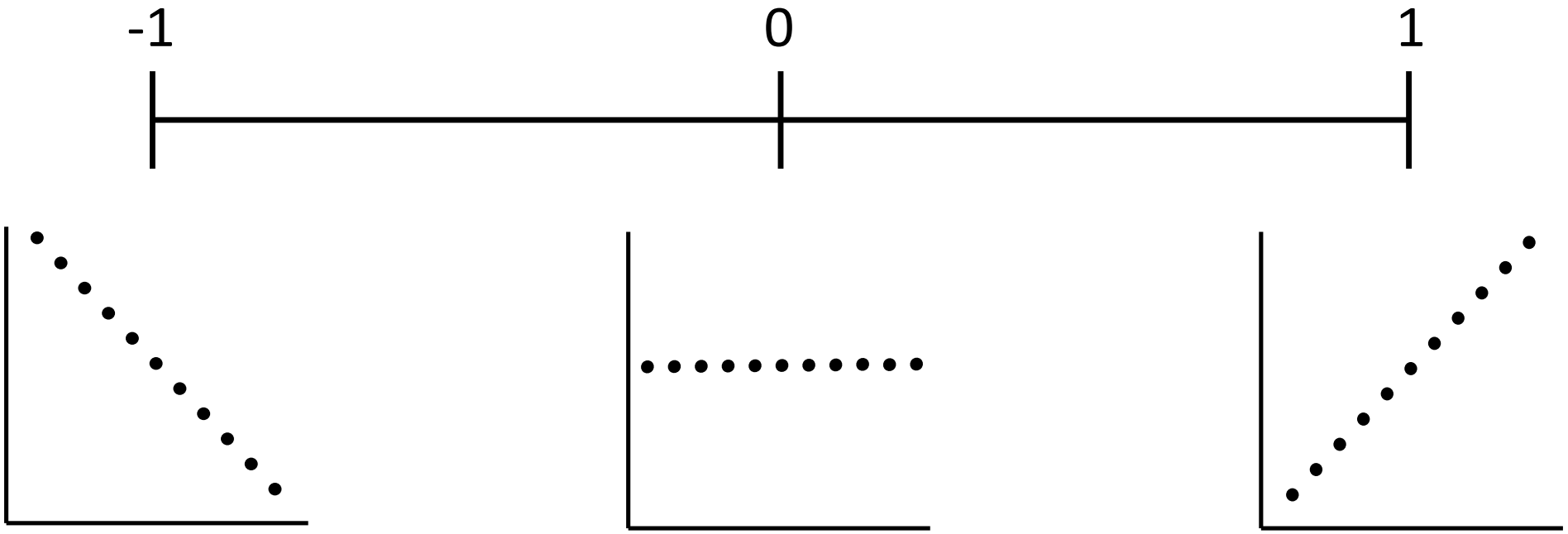
Strength of relationship

How strong of an effect does x have on y?

Sign of relationship

Does y increase (+ve) or decrease (-ve) as x increases?

Pearson correlation coefficient, *r*



Pearson correlation coefficient, *r*

Step one

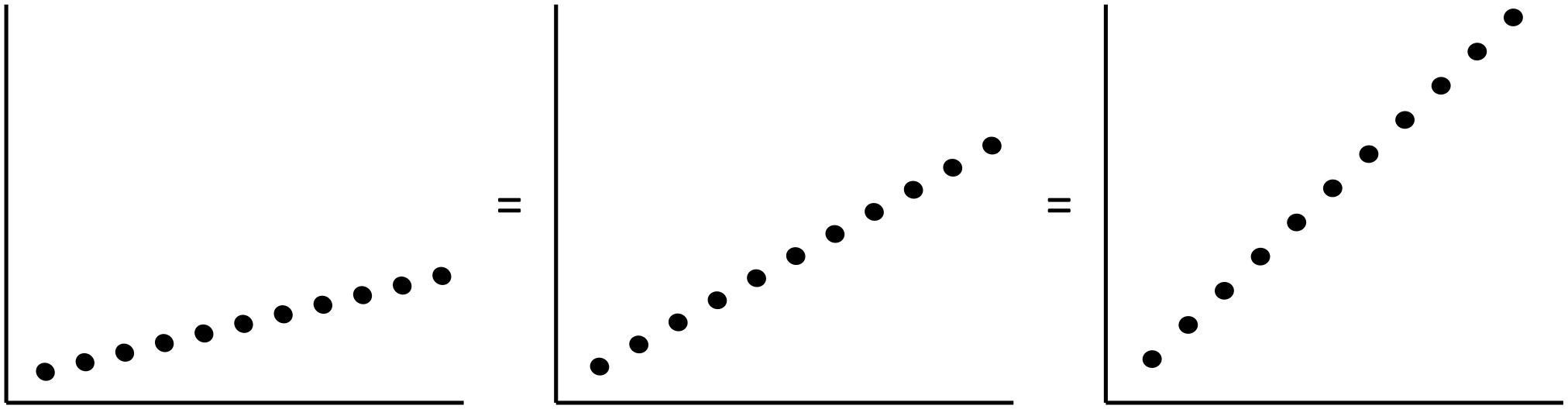
Normalise data (Z-statistic)

Step two

Place in expression for r that yields required values on -1 to 1 scale

Pearson correlation coefficient, *r*

Normalise



Pearson correlation coefficient, *r*

Normalise

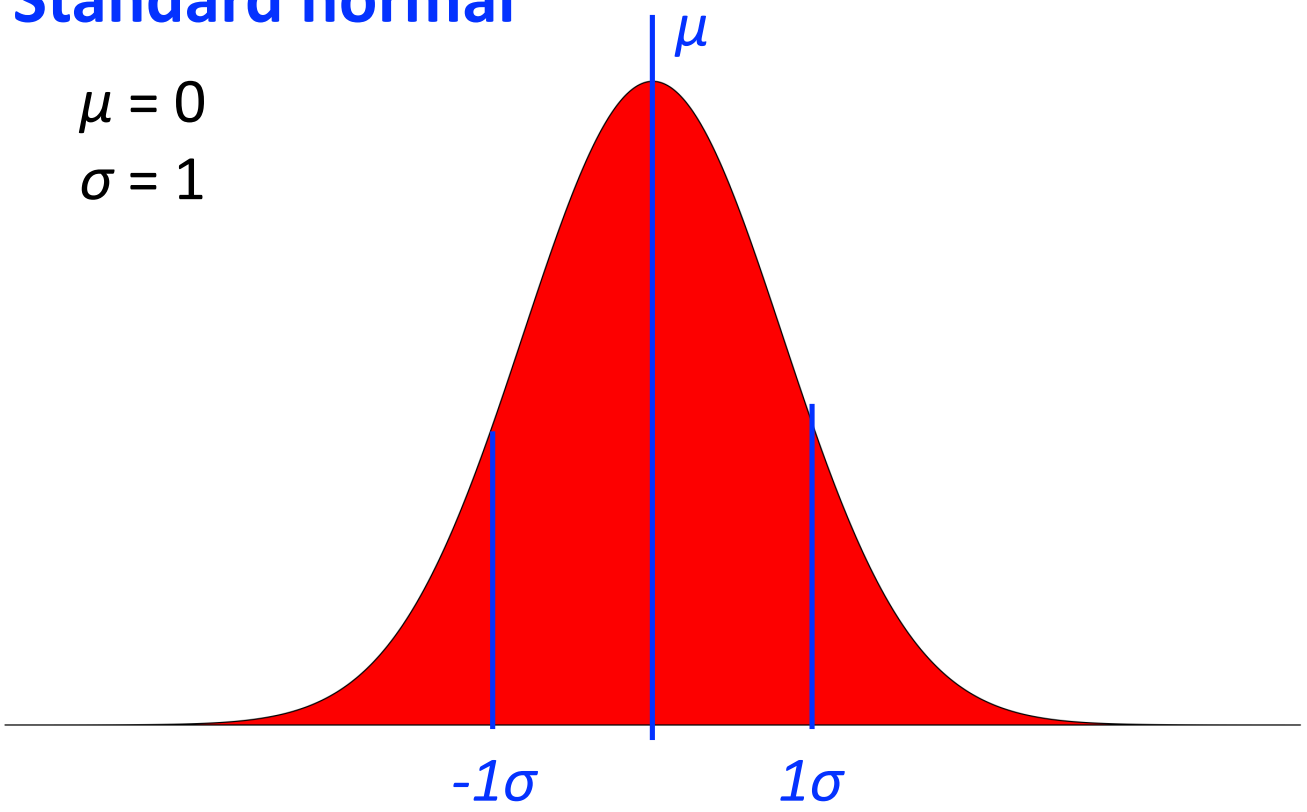
$$Z_{xi} = \frac{X_i - \bar{X}}{s_X}$$

$$Z_{yi} = \frac{Y_i - \bar{Y}}{s_Y}$$

Standard normal

$\mu = 0$

$\sigma = 1$



Pearson correlation coefficient, *r*

Expression for r

$$r = \frac{\sum_{i=1}^{n}(Z_{xi} \times Z_{yi})}{n-1}$$

Pearson correlation coefficient, *r*

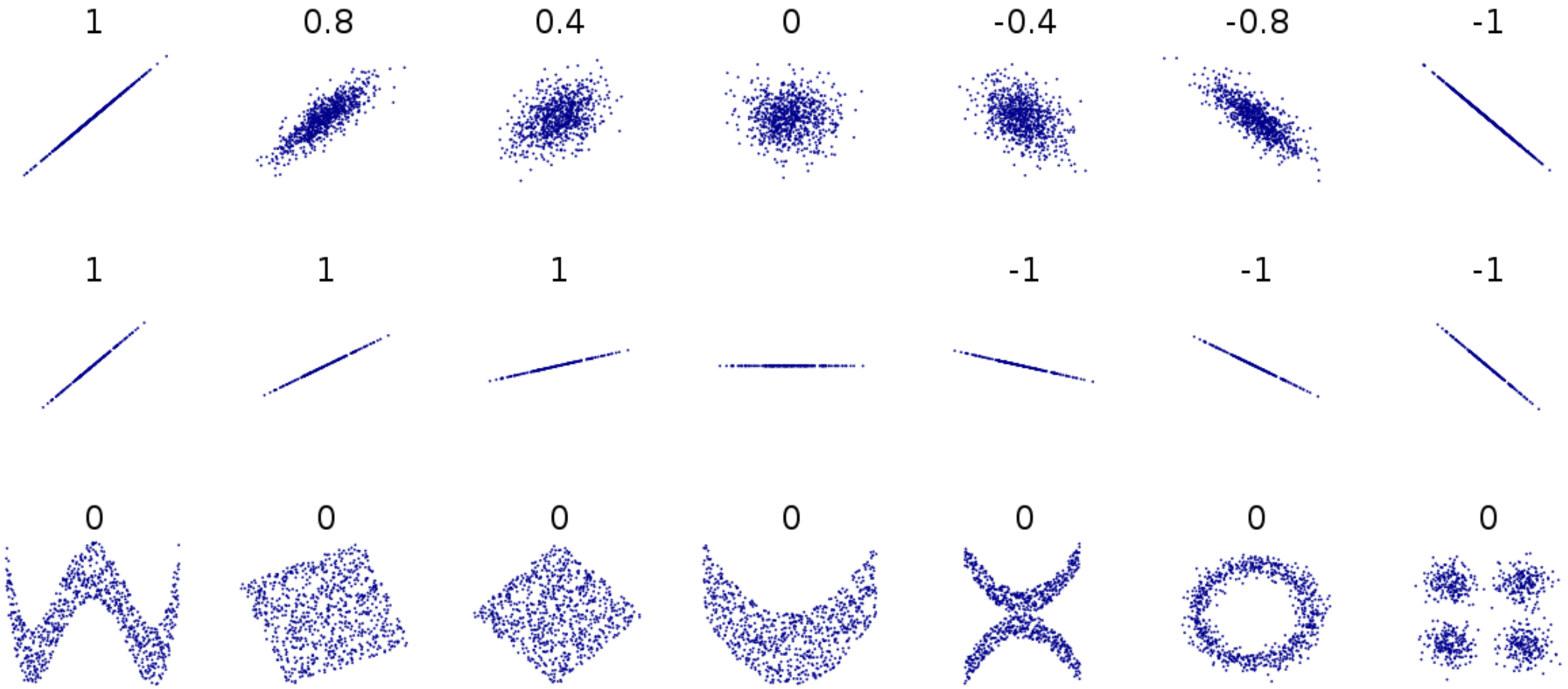
```
cor.test(x = x, y = y)$estimate
```

```
0.9832309
```

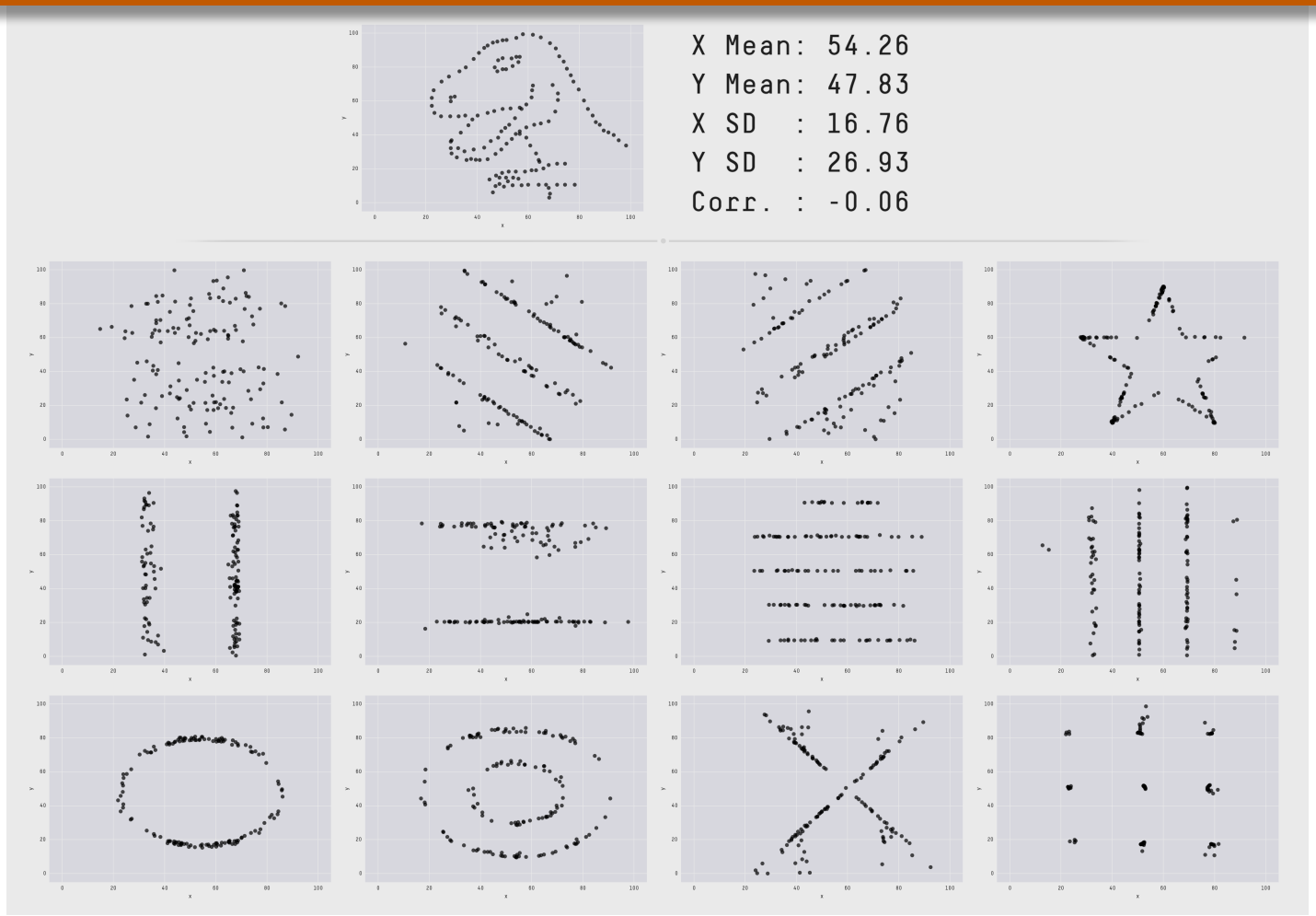
```
cor.test(x = x, y = y)$p.value
```

```
0.009772999
```

Pearson correlation coefficient, *r*

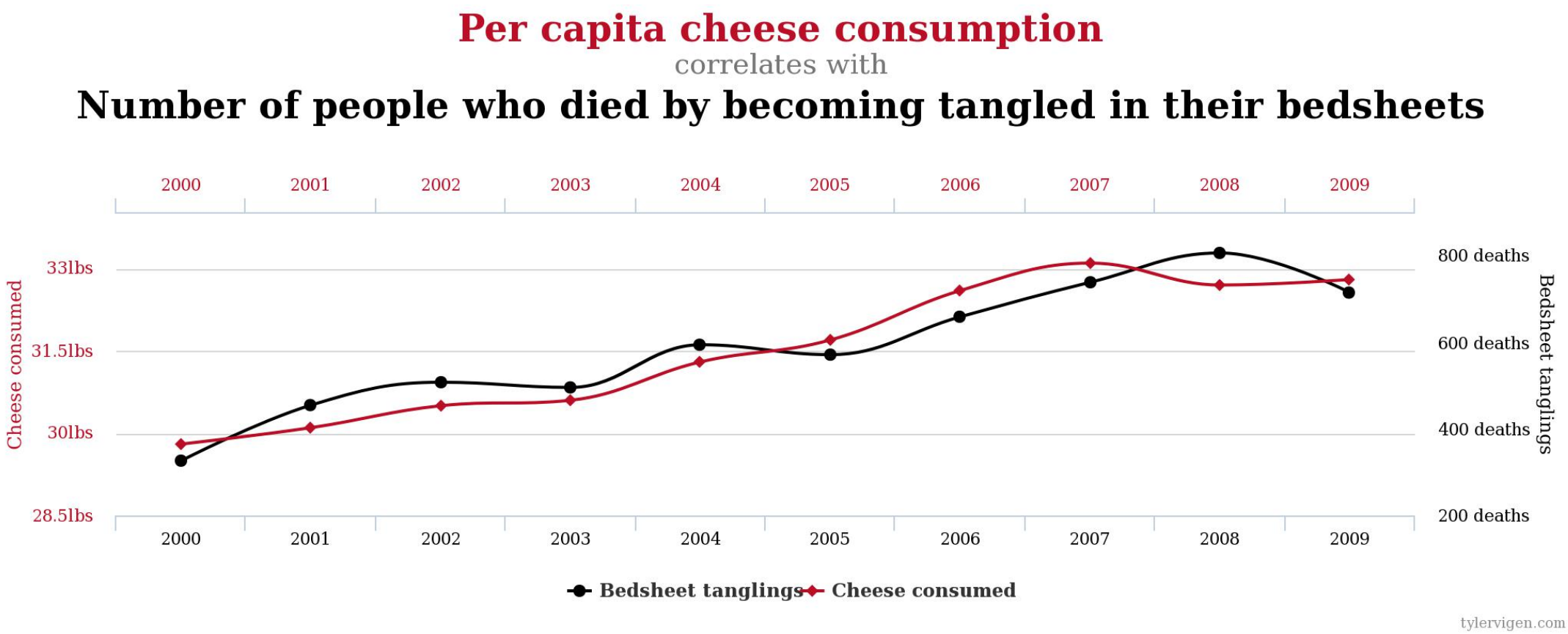


Pearson correlation coefficient, *r*



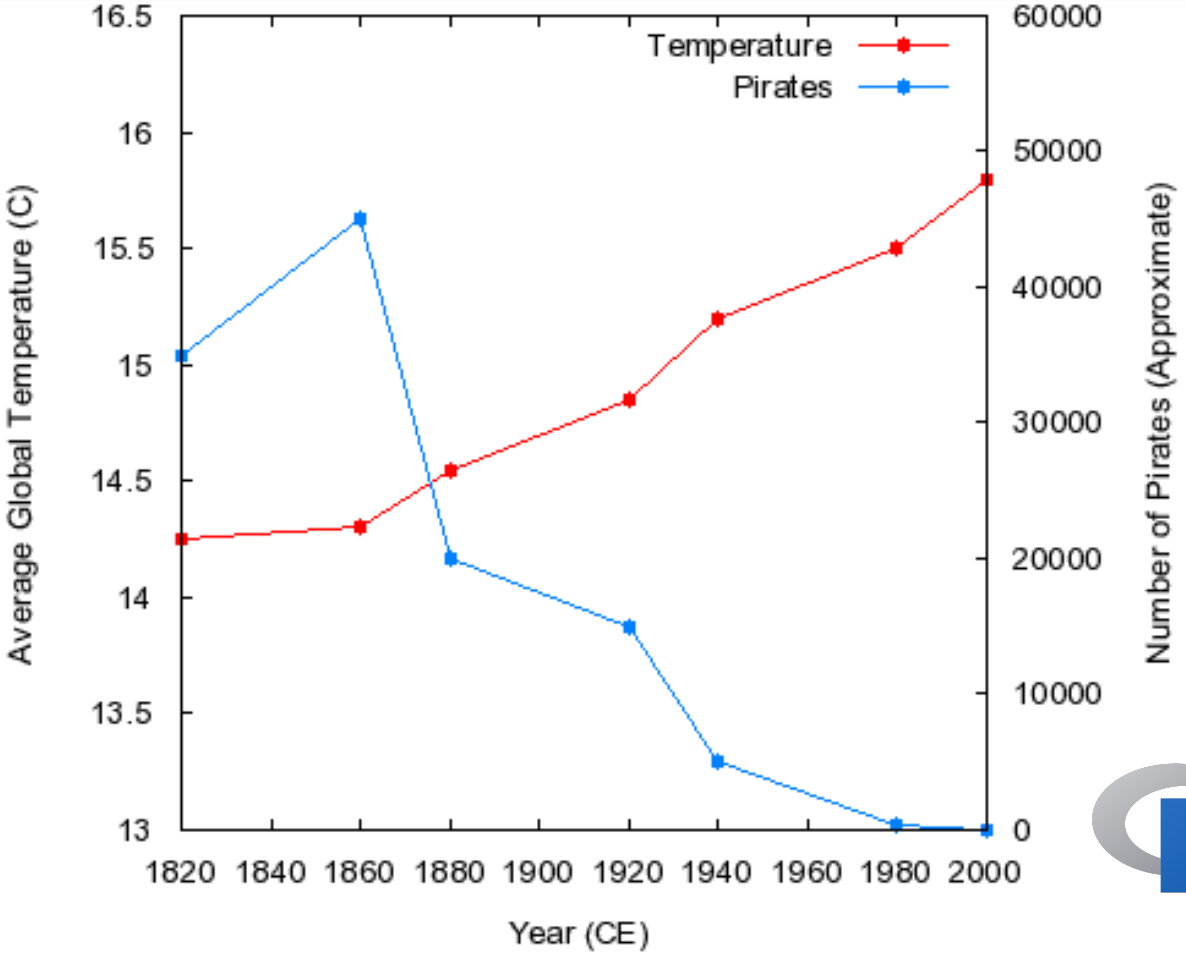
Source: autodeskresearch.com/publications/samestats

Correlation and causation



Source: tylervigen.com/spurious-correlations

Autocorrelation




```
cor.test(x = pirates,
  y = temp)$estimate
      -0.9242241
cor.test(x = pirates,
  y = temp)$p.value
      0.00291353
```