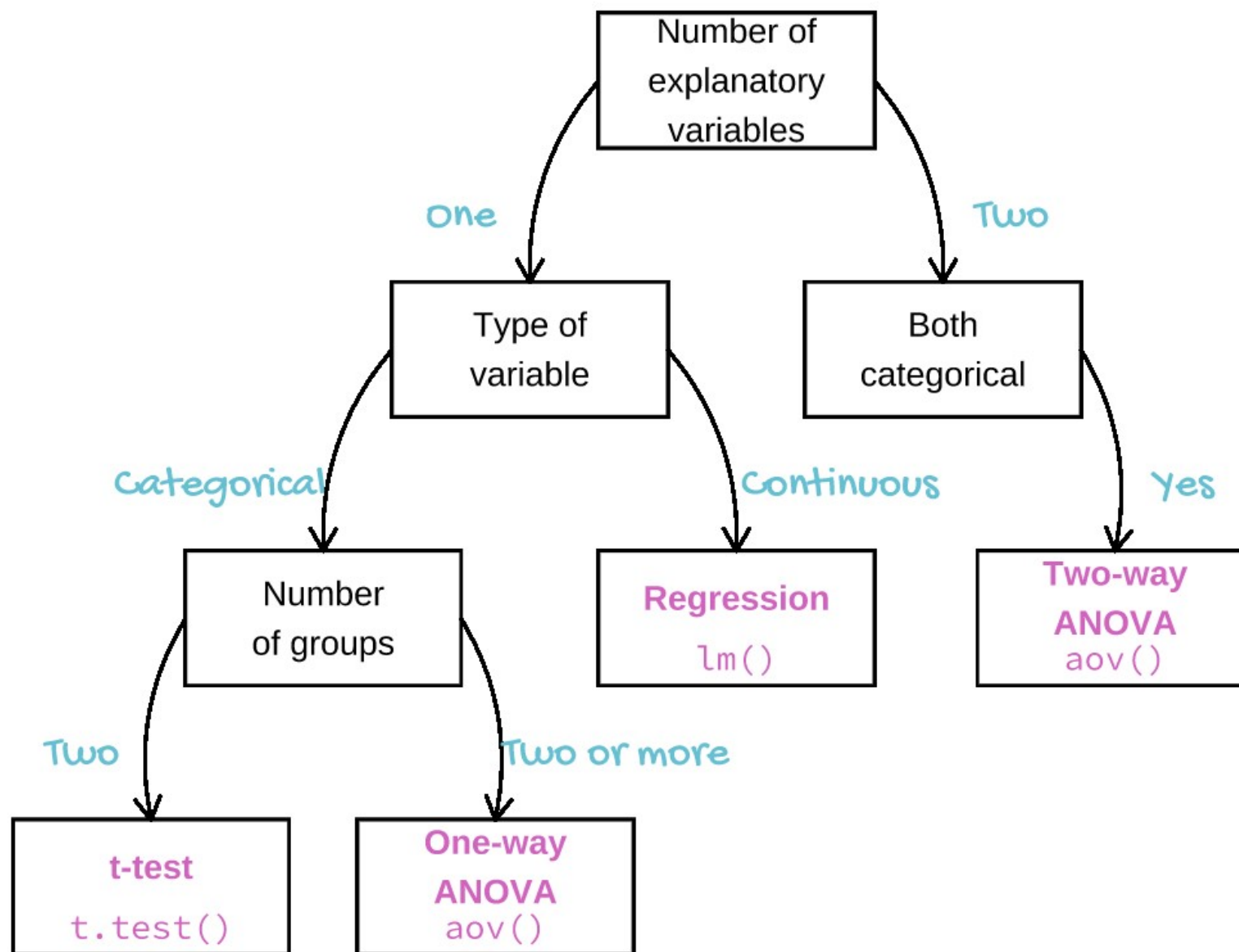# Emma Rand
Data Analysis in R

Correlation and Regression

Previous weeks

- We have considered data with categorical explanatory variables

Number of
explanatory
variables

One → Type of
variable

Two → Both
categorical

Categorical → Number
of groups

Continuous → **Regression**
`lm()`

Yes → **Two-way
ANOVA**
`aov()`

Two → **t-test**
`t.test()`

Two or more → **One-way
ANOVA**
`aov()`

Summary of this week

- Situations where our explanatory variable is 'continuous' rather than categorical.
- Parametric and non-parametric correlation
  - Meaning
  - Assumptions
  - Carrying out, interpreting and Reporting
  - Tests of correlation coefficients
- Regression
  - Meaning and terminology
  - Carrying out, interpreting and Reporting
  - Assumptions
  - Assessment of fit (explanatory power)

Learning objectives for the week

By the end of this week the successful student should be able to :

- Explain the principles of correlation and of regression and know when each can be applied (MLO 1)

- Select, appropriately correlation and regression (MLO 2)

- Apply and interpret a correlation in R (MLO 3 and 4)

- Appreciate the difference between statistical significance and biological significance (MLO 1 and 4)

- Apply and interpret a simple linear regression in R (MLO 3 and 4)

- Evaluate whether the assumptions of regression are met (MLO 2)

- Summarise and illustrate with appropriate R figures test results scientifically (MLO 3 and 4)

Introduction

Similar but different

Correlation

- Linear
- Association
- Axes can be switched
- two randomly sampled continuous or ordered discrete variables
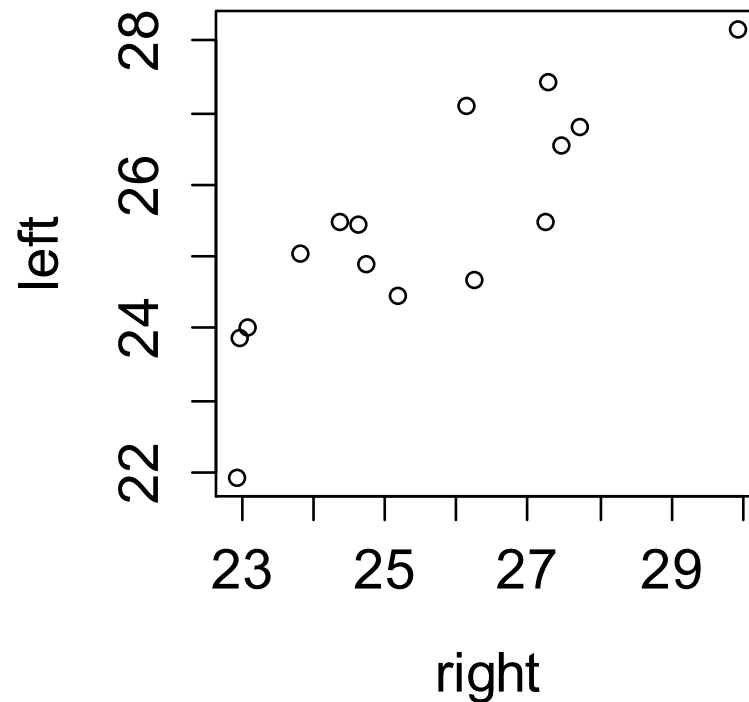- Scatter plot, no line

Regression

- Linear
- Prediction
- Axes cannot be switched
- X is "sampled without error"; y randomly sampled for each x
- Scatter plot, must have the regression line

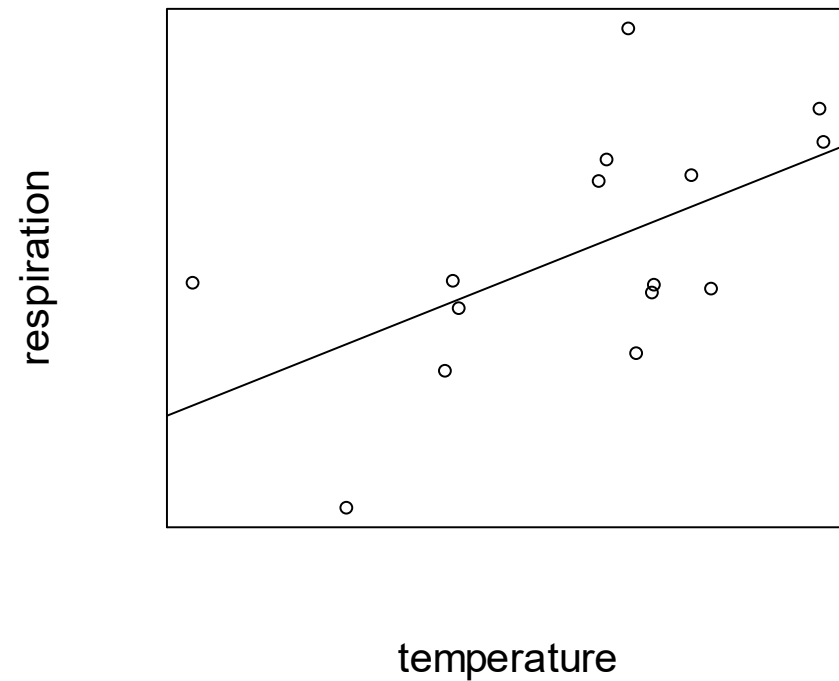# Similar but different

Correlation

Length of Ulna (cm)

Regression

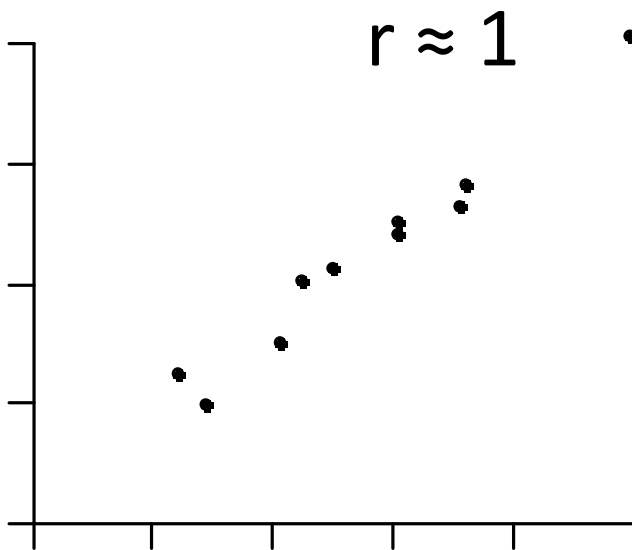**Manipulate/choose x, measure y**

Correlation

Basics

- Sample correlation: *r*
- Reflects degree of linear association between two sampled variables: -1 to +1
- Several types but same principles
- Here: Pearson's (Pearson's Product Moment Correlation Coefficient)
- Parametric

Example of correlations

r ≈ 1

r ≈ -1

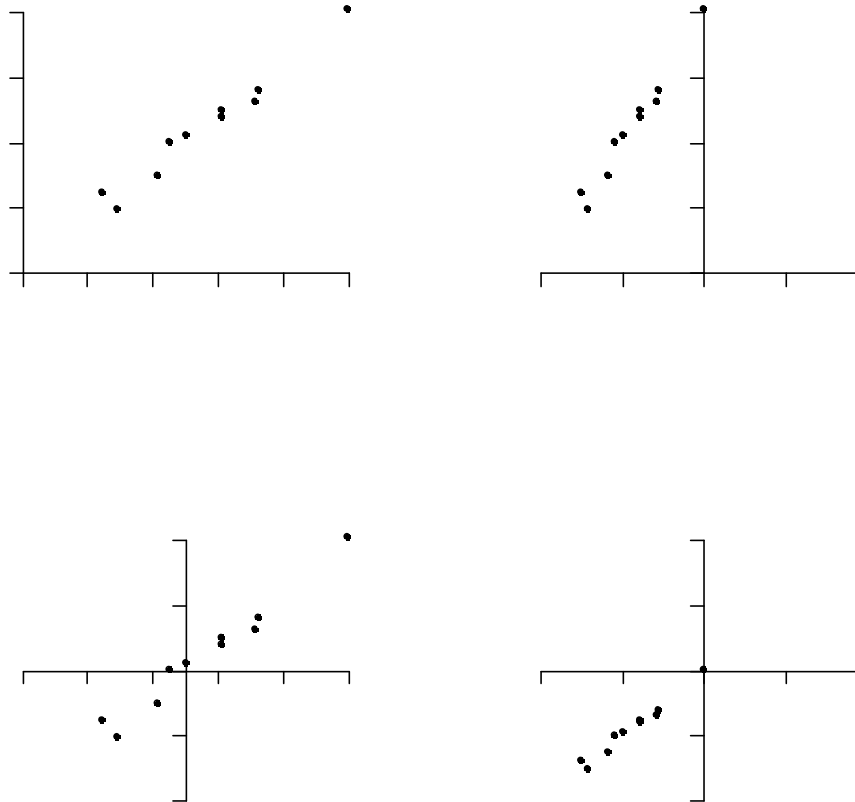Positive: Highest scores on one axis associated with highest scores on other
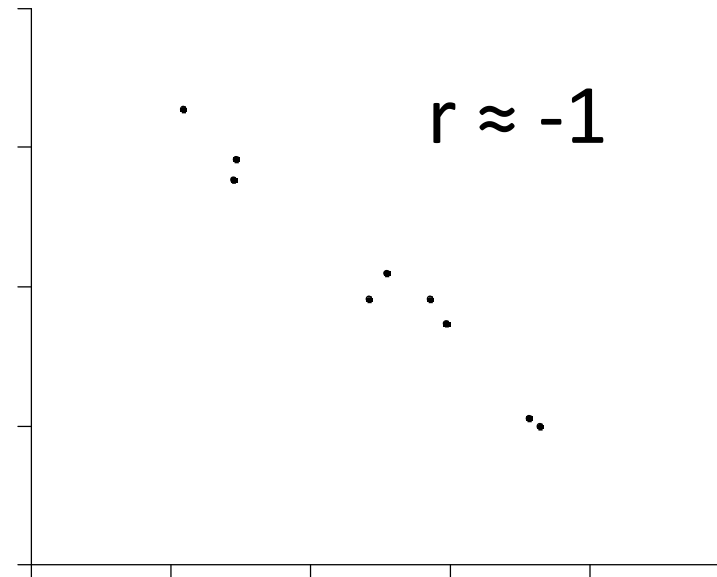
Negative: Highest scores on one axis associated with lowest scores on other

Example of positive correlations

r ≈ 1

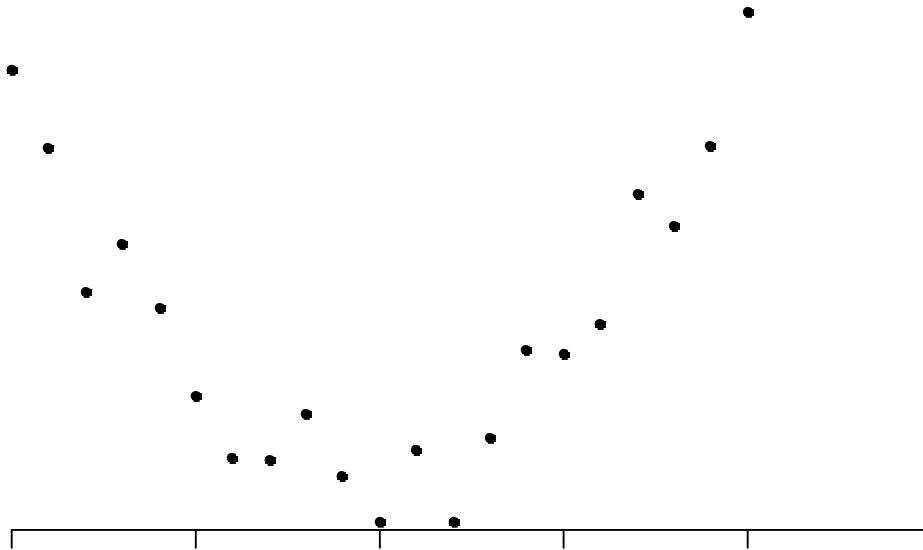Highest scores on one axis associated with highest scores on other

Correlation but not linear

r ≈ 0

Cannot use Pearson's PMMC

Correlation
Example

Wheat seeds: High quality visualization of the internal kernel structure by a soft X-ray technique and 7 measurements taken:

- Area

- Perimeter

- Compactness

- Kernel length

- Kernel width

- Asymmetry coefficient

- Length of kernel groove

We will examine the correlation between compactness and kernel width

# Reading in and examining the structure of the data

```
file <- "data-raw/seeds_dataset.xlsx"
seeds <- read_excel(file, sheet = "seeds_dataset")
glimpse(seeds)
Observations: 70
Variables: 7
$ area          <dbl> 15.26, 14.88, 14.29, 13.84, 16.14, 14.38, 14.69, 14.11, 1...
$ perimeter     <dbl> 14.84, 14.57, 14.09, 13.94, 14.99, 14.21, 14.49, 14.10, 1...
$ compactness   <dbl> 0.8710, 0.8811, 0.9050, 0.8955, 0.9034, 0.8951, 0.8799, 0...
$ kernal_length <dbl> 5.763, 5.554, 5.291, 5.324, 5.658, 5.386, 5.563, 5.420, 6...
$ kernel_width  <dbl> 3.312, 3.333, 3.337, 3.379, 3.562, 3.312, 3.259, 3.302, 3...
$ asymmetry_coef <dbl> 2.2210, 1.0180, 2.6990, 2.2590, 1.3550, 2.4620, 3.5860, 2...
$ groove_length <dbl> 5.220, 4.956, 4.825, 4.805, 5.175, 4.956, 5.219, 5.000, 5...
```

Plot your data

Plot your data: roughly

```
ggplot(data = seeds, aes(x = compactness, y = kernel_width)) +
    geom_point()
```
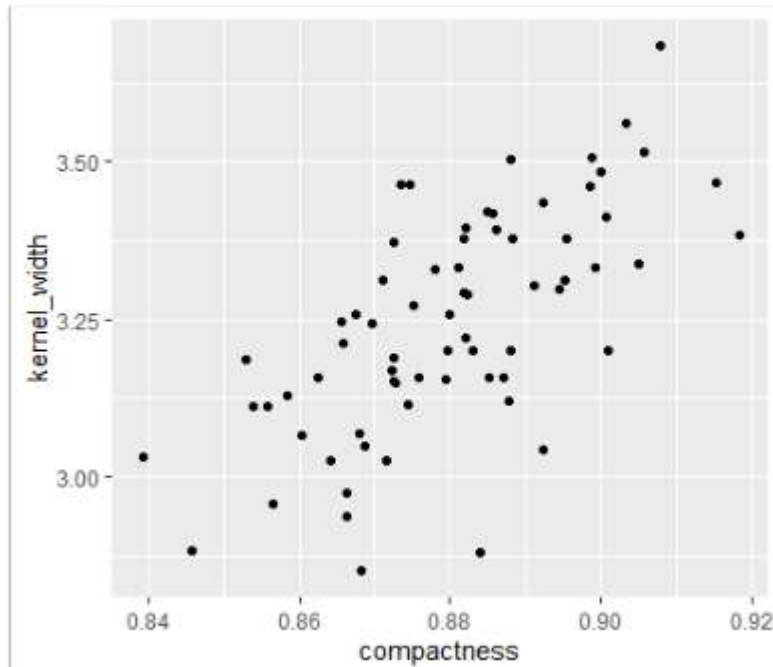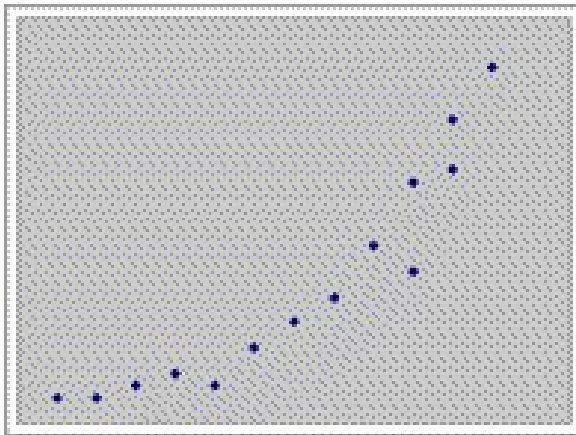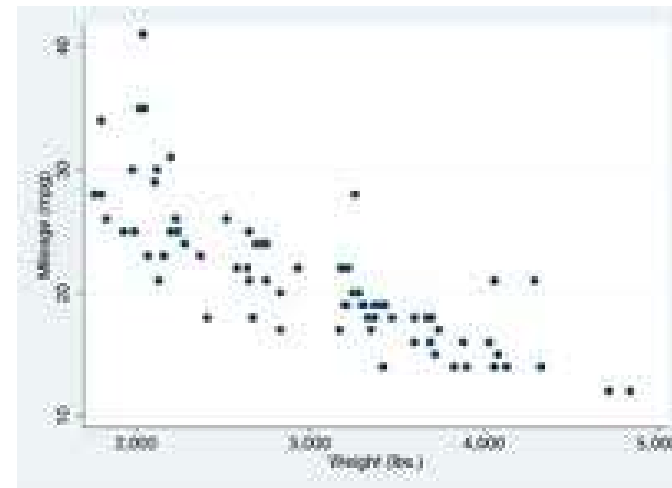

Check roughly linear

This looks ok

Plot your data

Not suitable for linear correlation
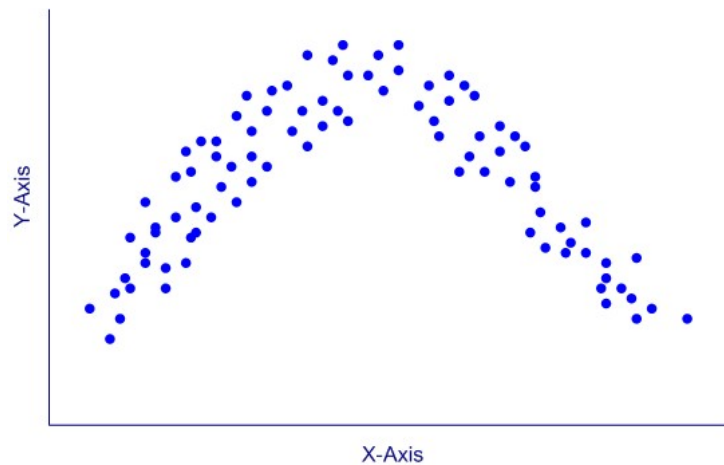
Running the test

```
cor.test(data = seeds, ~ compactness + kernel_width)

        Pearson's product-moment correlation

data:  compactness and kernel_width
t = 7.3738, df = 68, p-value = 2.998e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5117537 0.7794620
sample estimates:
      cor
0.6665731
```

A variable is not being explained:
~ compactness + kernel_width

Running the test

```
cor.test(data = seeds, ~ compactness + kernel_width)

        Pearson's product-moment correlation


data:  compactness and kernel_width
t = 7.3738, df = 68, p-value = 2.998e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5117537 0.7794620
sample estimates:
      cor
0.6665731
```

Gives type of correlation done. Pearson's in the default

Running the test

```
cor.test(data = seeds, ~ compactness + kernel_width)

        Pearson's product-moment correlation

data:  compactness and kernel_width
t = 7.3738, df = 68, p-value = 2.998e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5117537 0.7794620
sample estimates:
      cor
0.6665731
```

Gives type of correlation done.
Pearson's in the default

Running the test

```
cor.test(data = seeds, ~ compactness + kernel_width)

        Pearson's product-moment correlation

data:  compactness and kernel_width
t = 7.3738, df = 68, p-value = 2.998e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5117537 0.7794620
sample estimates:
      cor
0.6665731
```

The correlation coefficient, *r*

Correlation
Running the test

```
cor.test(data = seeds, ~ compactness + kernel_width)

        Pearson's product-moment correlation

data:  compactness and kernel_width
t = 7.3738, df = 68, p-value = 2.998e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5117537 0.7794620
sample estimates:
      cor
0.6665731
```

Confidence interval on, *r*

Running the test

```
cor.test(data = seeds, ~ compactness + kernel_width)

        Pearson's product-moment correlation

data:  compactness and kernel_width
t = 7.3738, df = 68, p-value = 2.998e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5117537 0.7794620
sample estimates:
      cor
0.6665731
```

Test of whether r is significantly from zero

Reporting the result

```
data:  compactness and kernel_width
t = 7.3738, df = 68, p-value = 2.998e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5117537 0.7794620
sample estimates:
      cor
0.6665731
```

- There is a significant positive correlation (r = 0.67) between compactness and kernel width (t = 7.37; $d.f.$ = 68, p < 0.001).

Understanding the test of significance

- The R output contains a test of whether r = 0

- uses t
$$t = \frac{\text{statistic - hypothesised value}}{\text{estimated SE of the statistic}}$$

- For correlation: $t_{[d.f.]} = \frac{r}{s.e.}$

- Where standard error of r is $\sqrt{\frac{1-r^2}{N-2}}$

  - d.f. are N-2
- Sensitivity to sample size

Statistical significance ≠Biological significance

```
data:  x1 and x2
t = -2.2154, df = 9998, p-value = 0.02675
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.041733016 -0.002552103
sample estimates:
        cor
-0.02215107
```
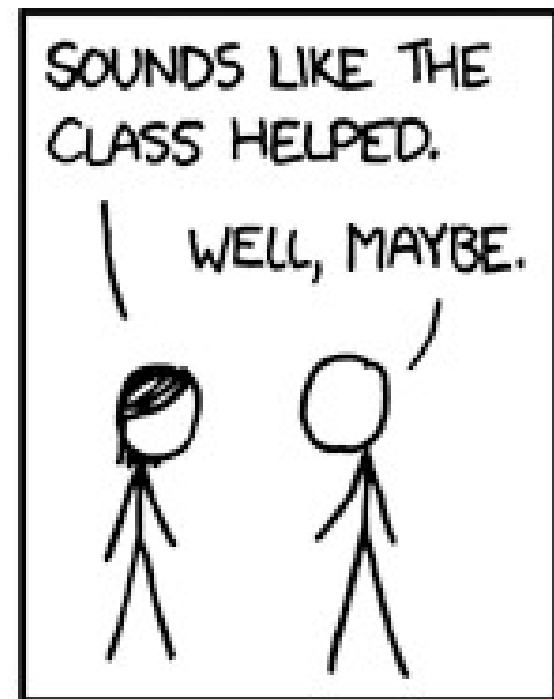
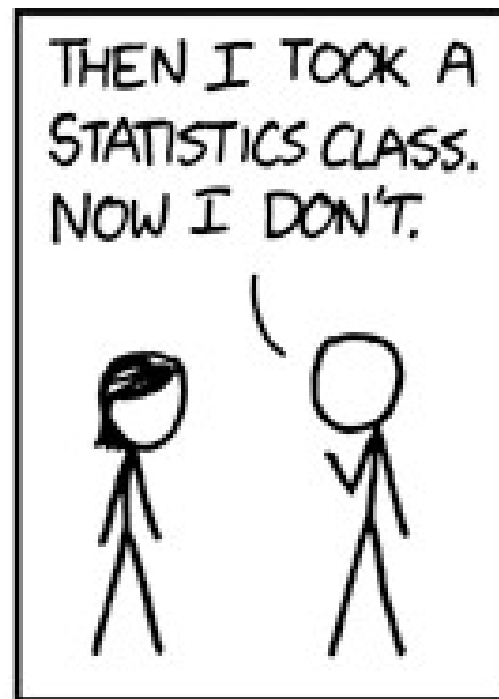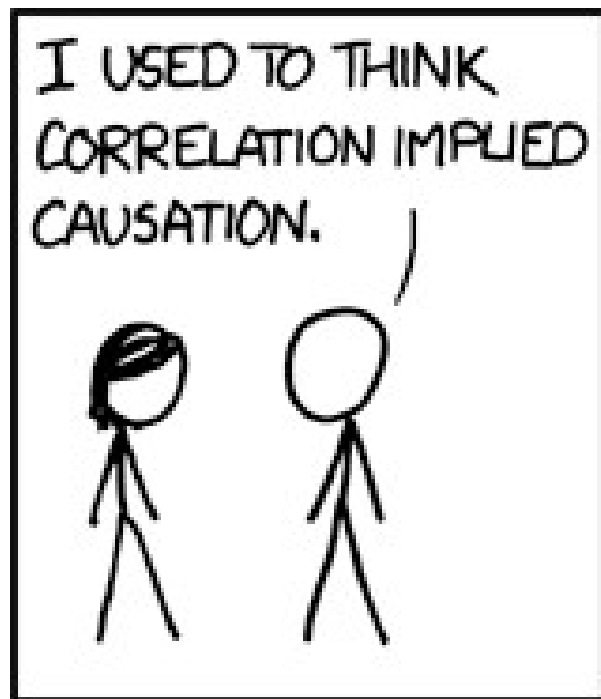Large samples mean small probability

But *r* is tiny!

Our knowledge determines biological significance
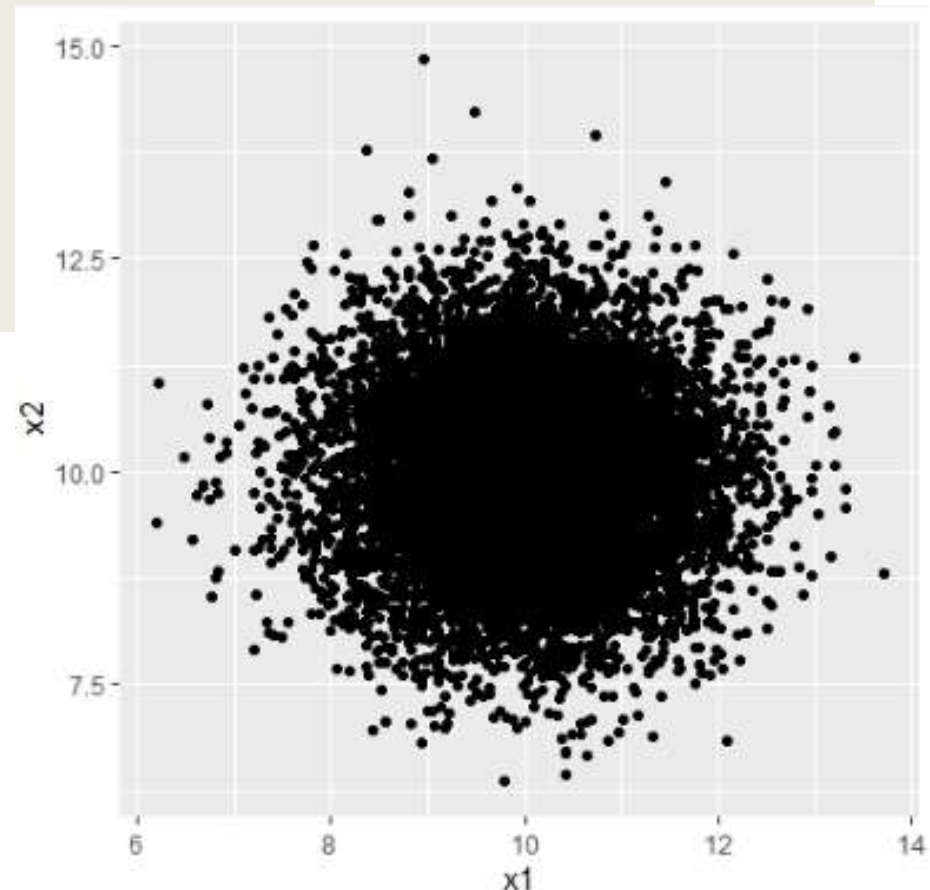
Correlation summary

- Association

- Pearson's is Parametric; Spearman's is non-parametric

- Two randomly sampled continuous/ordered variables

- Function in R:
  `cor.test(data = df, ~ x1 + x2)`

- quote r, its significance and n

- if scatterplot included do NOT show a fitted line

Regression

Regression

- Linear
- Prediction: One variable causes the other
- Axes cannot be switched
- X is "sampled without error"; y randomly sampled for each x
- We will consider linear regression only best fitting straight line:
$$y = b_1 x + b_0$$

The terminology

Response = b_1 * explanatory + b_0

$y = b_1x - b_0$

b_1 is the slope
b_0 is the intercept

Residual
observed y - predicted y
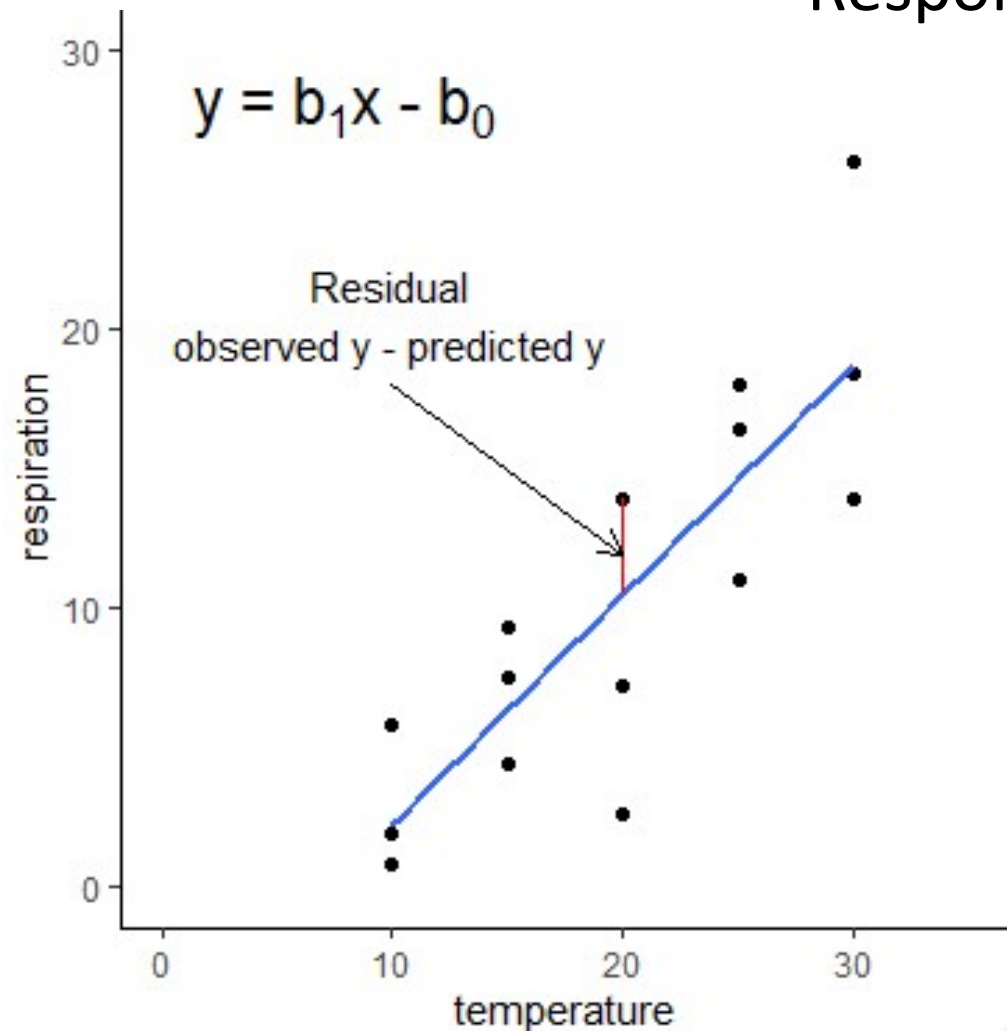
Null hypothesis

Can be expressed as:
- $b_1 = 0$
- x cannot predict *y*
- Regression line doesn't explain variance in *y*

Assumptions
- Normality and homoscedascity of residuals
- X is "sampled without error";
- y randomly sampled for each x and normally distributed

Example

Brine Shrimp (*Artemia salina*) were put in water baths at 10C, 15C, 20C, 25C, 30C and their respiration rate measured (units)

	temperature	respiration
1	10	0.785
2	10	5.784
3	10	1.879
4	15	9.331
5	15	4.412
6	15	7.515
7	20	13.852
8	20	2.633
9	20	7.157
10	25	17.983
11	25	16.426
12	25	11.029
13	30	18.353
14	30	13.934
15	30	25.965

Plot your data

Plot your data: roughly

```
ggplot(data = shrimp, aes(x = temperature, y = respiration)) +
  geom_point()
```
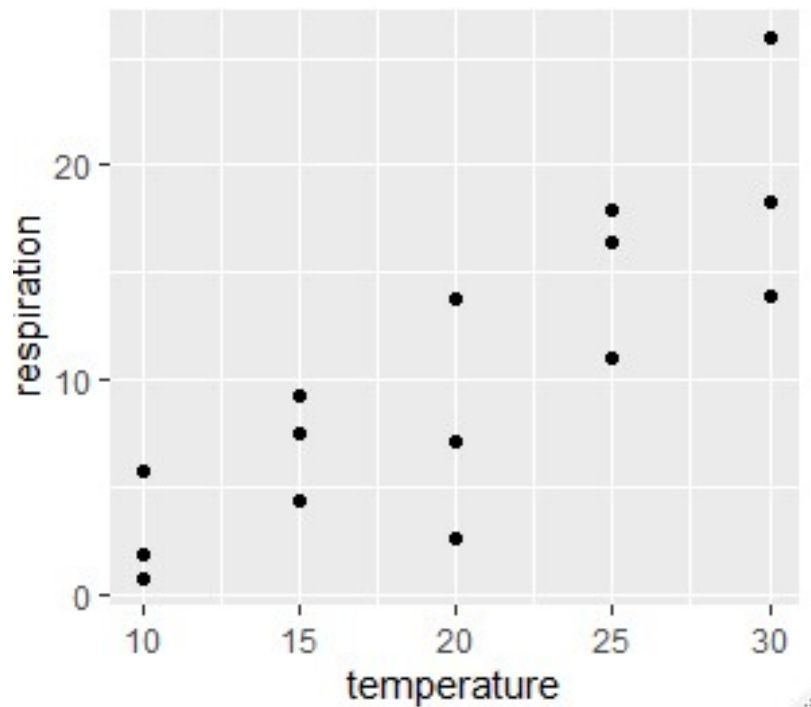

Check roughly
linear

This looks ok

Running the test

```
mod <- lm(data = shrimp,
          respiration ~ temperature)
summary(mod)
```

Understanding the output

Core statistical ideas – very extendable. You will see again next year

```
Call:
lm(formula = respiration ~ temperature, data = shrimp)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8362 -2.6216 -0.3377  3.1854  7.2433

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.0359     3.1560  -1.912   0.0781 .
temperature   0.8253     0.1488   5.547 9.43e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.074 on 13 degrees of freedom
Multiple R-squared:  0.703,      Adjusted R-squared:  0.6801
F-statistic: 30.77 on 1 and 13 DF,  p-value: 9.433e-05
```

Regression
Understanding the output

Core statistical ideas – very extendable. You will see again next year

```
Call:
lm(formula = respiration ~ temperature, data = shrimp)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8362 -2.6216 -0.3377  3.1854  7.2433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.0359     3.1560  -1.912   0.0781 .
temperature   0.8253     0.1488   5.547 9.43e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.074 on 13 degrees of freedom
Multiple R-squared:  0.703,      Adjusted R-squared:  0.6801
F-statistic: 30.77 on 1 and 13 DF,  p-value: 9.433e-05
```

b_0 and b_1

$$y = 0.83x - 6.03$$

Regression
Understanding the output

```
Call:
lm(formula = respiration ~ temperature, data =

Residuals:
    Min      1Q  Median      3Q     Max
-7.8362 -2.6216 -0.3377  3.1854  7.2433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.0359     3.1560  -1.912   0.0781 .
temperature  0.8253     0.1488   5.547 9.43e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.074 on 13 degrees of freedom
Multiple R-squared:  0.703,      Adjusted R-squared:  0.680
F-statistic: 30.77      1 and 13 DF,  p-value: 9.433e-05
```

Test: $b_0 = 0$
Often not impt

Test: $b_1 = 0$
Always of interest

Test of 'model'
Same as $b_1 = 0$
in single
regression

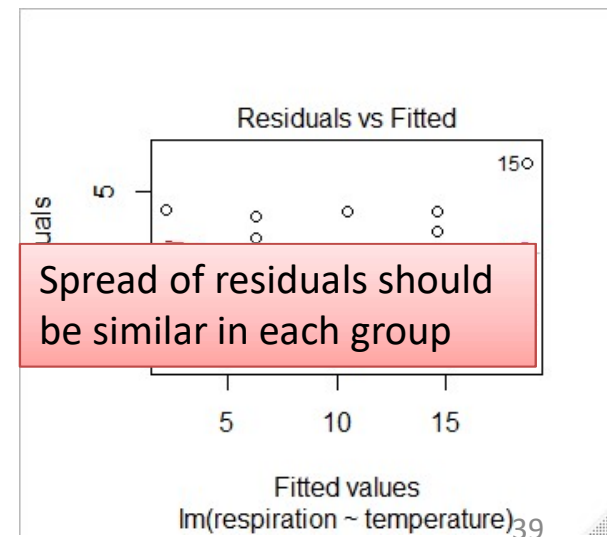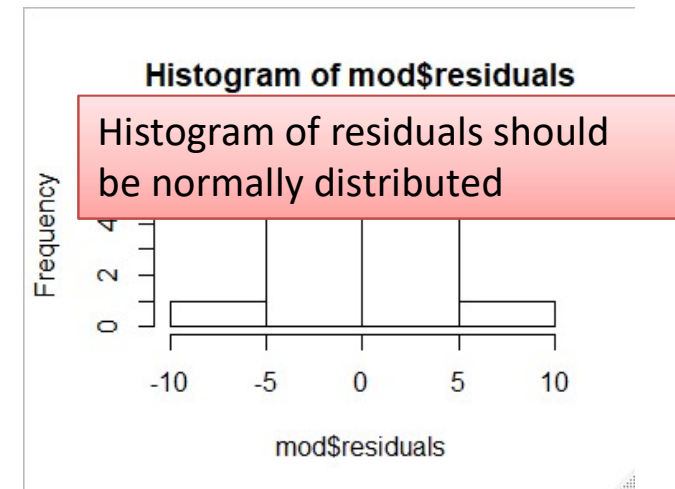Multiple R-squared: Proportion of y explained by x

Checking Assumptions

Residuals are calculated for you already!

```
hist(mod$residuals)
shapiro.test(mod$residuals)

        Shapiro-Wilk normality test

data:  (mod$residuals)
W = 0.97969, p-value = 0.9673
plot(mod, which = 1)
```


Histogram of mod$residuals

Histogram of residuals should be normally distributed

Residuals vs Fitted

Spread of residuals should be similar in each group

Fitted values
lm(respiration ~ temperature)

Reporting the results

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.0359      3.1560  -1.912   0.0781 .
temperature   0.8253      0.1488   5.547 9.43e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
   1

Residual standard error: 4.074 on 13 degrees of freedom
Multiple R-squared:  0.703,   Adjusted R-squared:  0.6801
F-statistic: 30.77 on 1 and 13 DF,  p-value: 9.433e-05
```
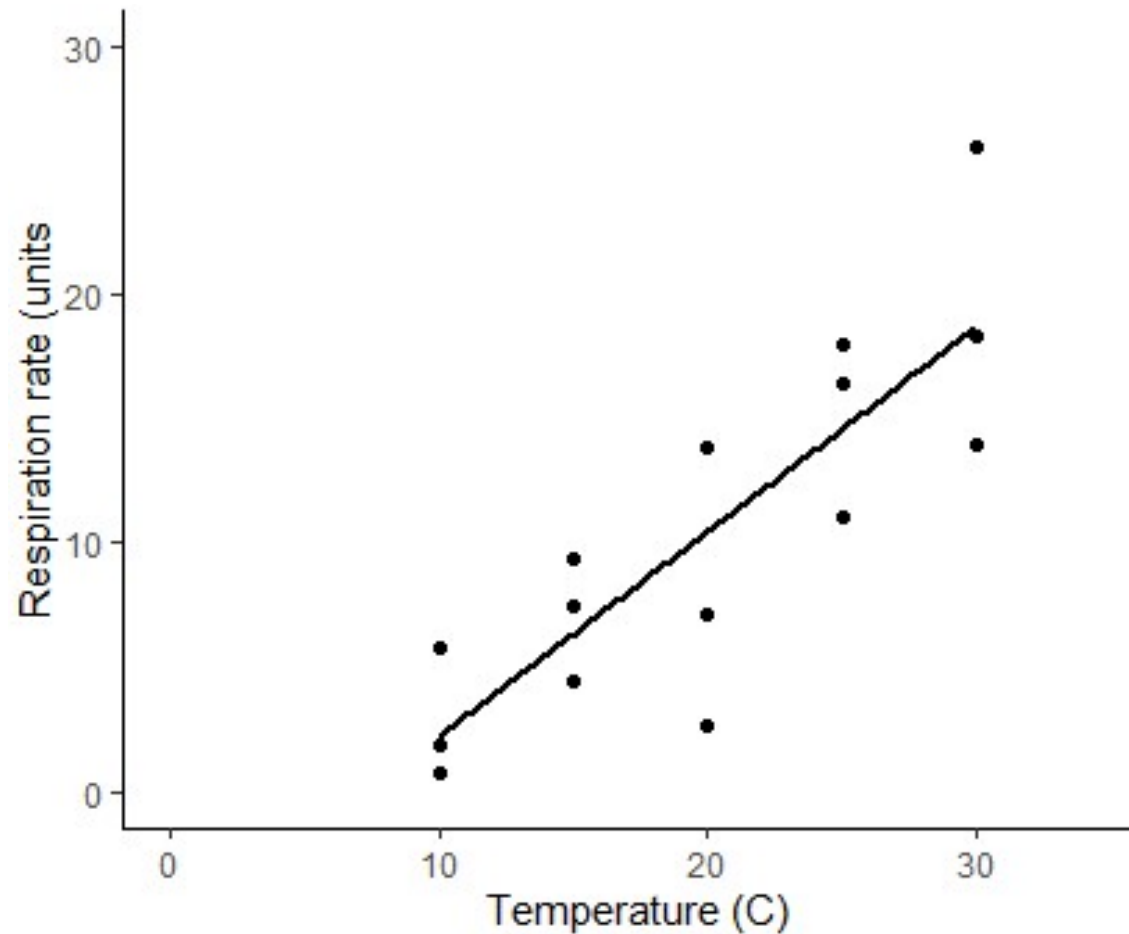
Reporting the result: "significance, direction, magnitude"

The temperature explained a significant amount of the variation in respiration rate (ANOVA: F = 30.8; $d.f.$ = 1, 13; p < 0.001). The regression line is: Respiration rate= 0.83 * temperature - 6.04

Reporting the results: figure

Regression summary

- Linear
- Prediction: One variable causes the other
- Axes cannot be switched
- X is "sampled without error"; y randomly sampled for each x
- linear regression: $y = b_1 x + b_0$
  Function in R:
  ```
  mod <- lm(data = df, y ~ x)
  Summary(mod)
  ```
- quote regression equation and test result (either ANOVA or t
- Scatterplot with fitted line (data and the model)