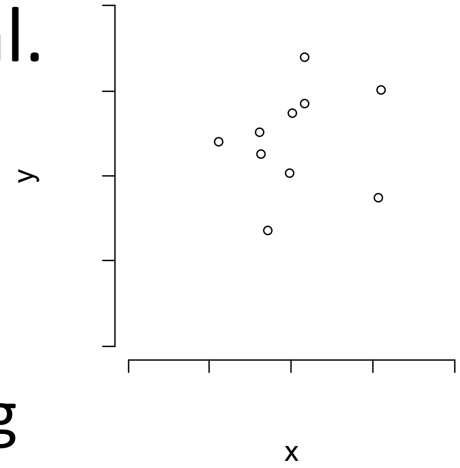
17C Laboratory & Professional Skills: Data Analysis

Laboratory & Professional skills for Bioscientists
Term 2: Data Analysis in R

Correlation and Regression

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 actively following the lecture and practical and carrying out the independent study the successful student will be able to:

- Explain the principles of correlation and of regression (MLO 1)
- Apply (appropriately), interpret and evaluate the legitimacy of, both in R (MLO 2, 3 and 4)
- Summarise and illustrate with appropriate R figures test results scientifically (MLO 3 and 4)

Correlation and Regression

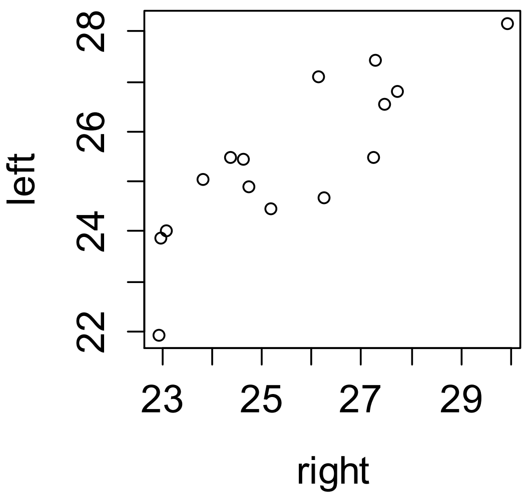
Similar but different

- Similar
  - Linear
  - Two continuous/ordered variables
  - Illustrated with a scatter plot
- Different
  - Correlation is association; regression is prediction
  - In correlation axes can be switched; in regression axis cannot be switched
  - Do not put a line of best fit on a correlation graph; regression graph must have the regression line

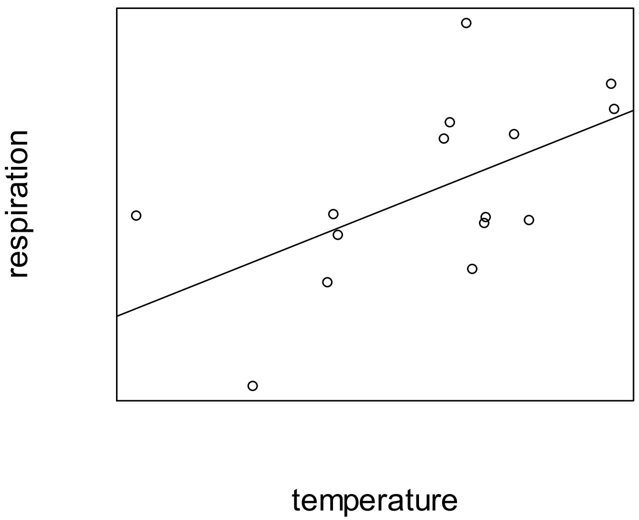
Correlation and Regression

Similar but different

Length of Ulna (cm)



Manipulate/choose x, measure y



Correlation

- Linear association
- No cause and effect
- Axes could be swapped

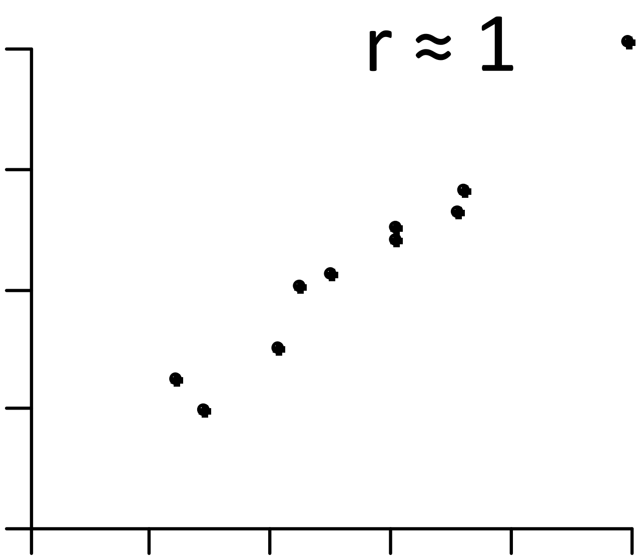
Regression

- Linear relationship
- Cause and effect
- Axes cannot be swapped

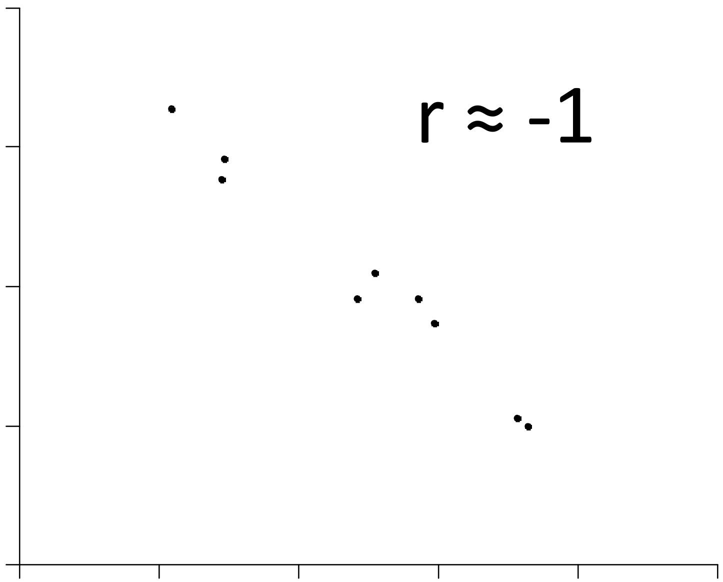
Basics

- Pearson’s (Pearson’s Product Moment Correlation Coefficient)
- Parametric
- Sample correlation: r
- Reflects degree of linear association between two sampled variables: -1 to +1

Example of correlations



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

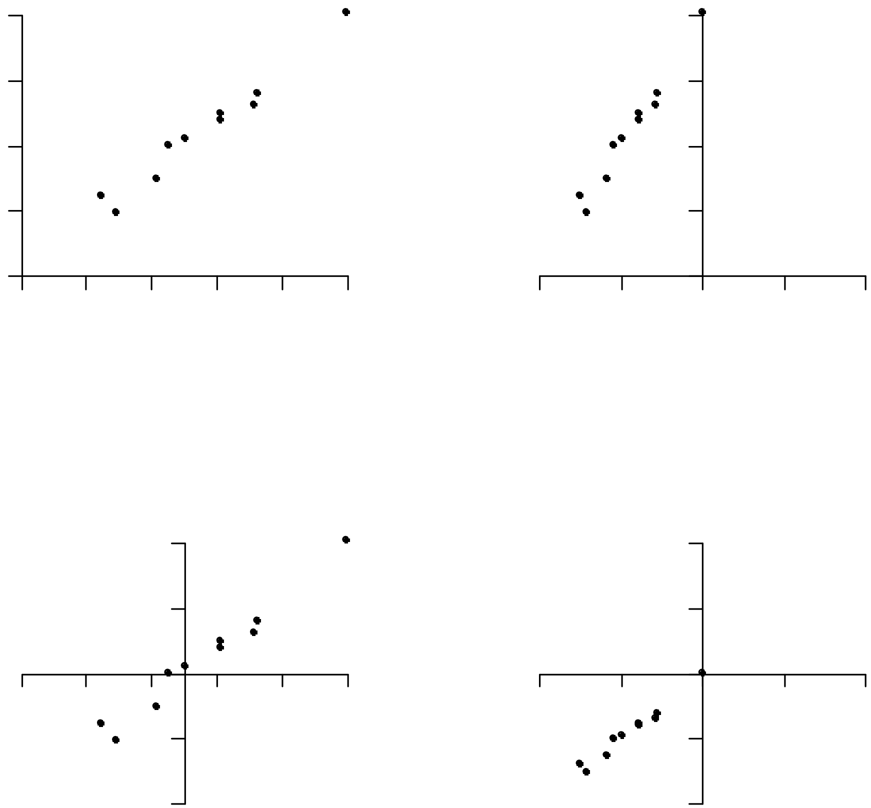


Negative: Highest scores on one axis associated with lowest scores on other

Correlation

Example of positive correlations

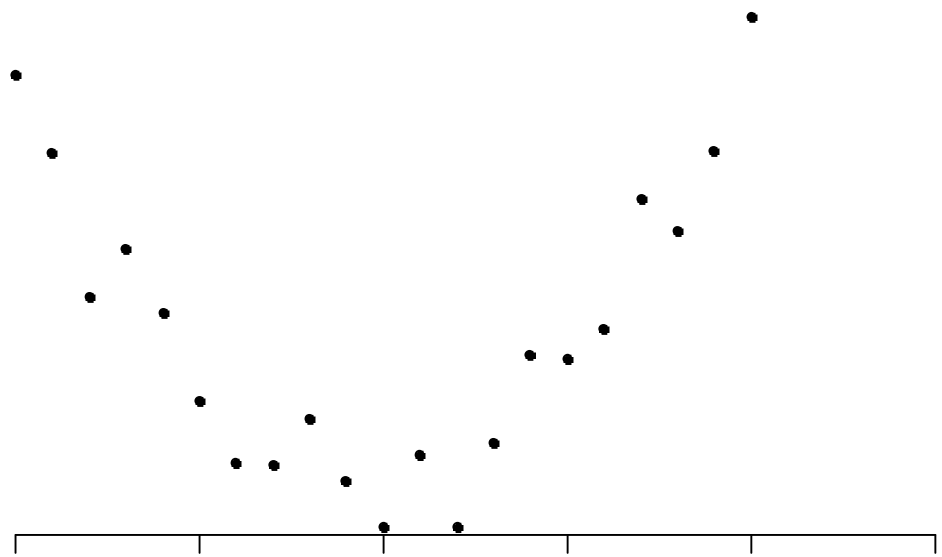
r ≈ 1



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
Length of kernel.
Width of kernel.
Asymmetry coefficient.
Length of kernel groove.

Correlation

Example

lect 09 correlation and regression.R* × | seeds ×

Filter

	area	perimeter	compactness	kernal_length	kernel_width	asymmetry_coef	groove_length
1	15.26	14.84	0.8710	5.763	3.312	2.2210	5.220
2	14.88	14.57	0.8811	5.554	3.333	1.0180	4.956
3	14.29	14.09	0.9050	5.291	3.337	2.6990	4.825
4	13.84	13.94	0.8955	5.324	3.379	2.2590	4.805
5	16.14	14.99	0.9034	5.658	3.562	1.3550	5.175
6	14.38	14.21	0.8951	5.386	3.312	2.4620	4.956
7	14.69	14.49	0.8799	5.563	3.259	3.5860	5.219
8	14.11	14.10	0.8911	5.420	3.302	2.7000	5.000
9	16.63	15.46	0.8747	6.053	3.465	2.0400	5.877
10	16.44	15.25	0.8880	5.884	3.505	1.9690	5.533
11	15.26	14.85	0.8696	5.714	3.242	4.5430	5.314
12	14.03	14.16	0.8796	5.438	3.201	1.7170	5.001
13	13.89	14.02	0.8880	5.439	3.199	3.9860	4.738
14	13.78	14.06	0.8759	5.479	3.156	3.1360	4.872
15	13.74	14.05	0.8744	5.482	3.114	2.9320	4.825
16	14.59	14.28	0.8993	5.351	3.333	4.1850	4.781
17	13.99	13.83	0.9183	5.119	3.383	5.2340	4.781
18	15.69	14.75	0.9058	5.527	3.514	1.5990	5.046

Two-way ANOVA example

Reading in and examining the structure of the data

```
library(readxl)
file <- "../data/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...
```

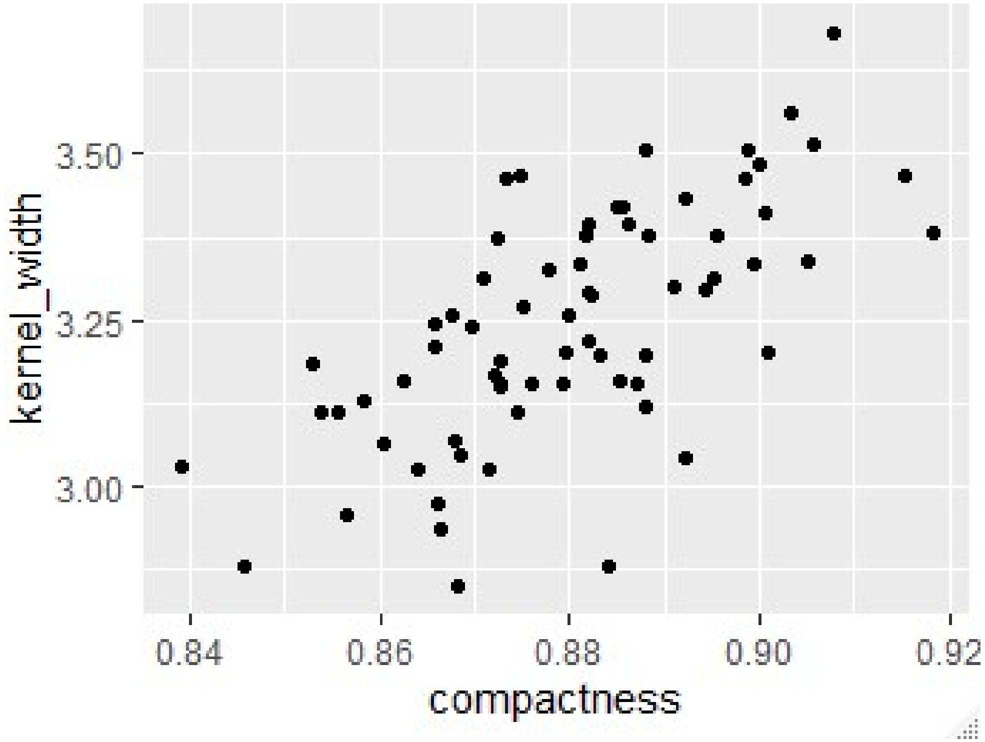
Assumptions: “bivariate normal”

Common sense

Plot your data

Plot your data: roughly

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

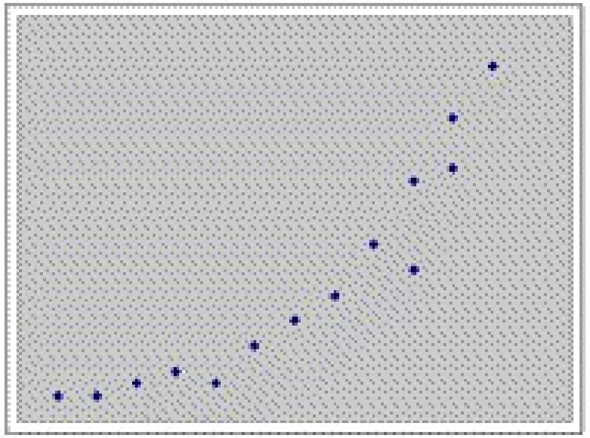


Check roughly linear

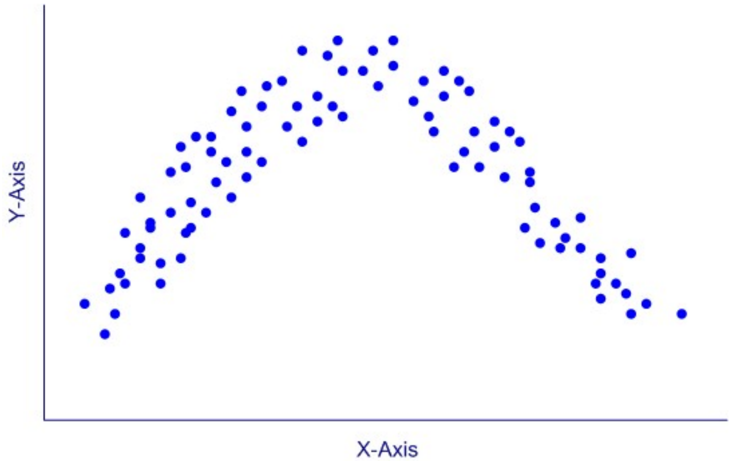
This looks ok

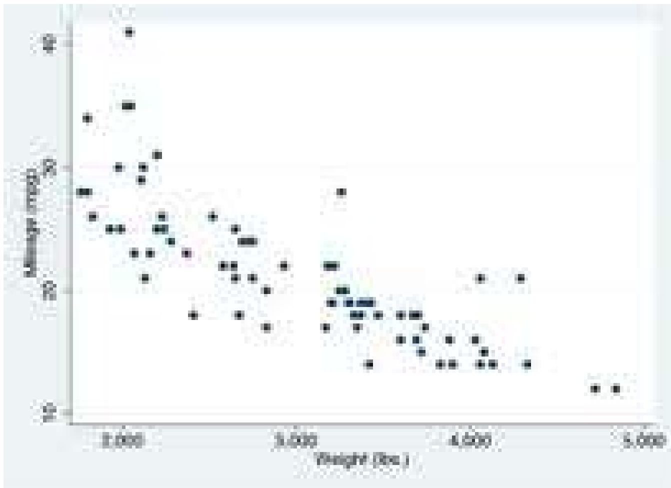
Correlation

Plot your data



Not suitable for linear correlation





Running the test

```
cor.test(seeds$compactness, seeds$kernel_width)
   Pearson's product-moment correlation

data:  seeds$compactness and seeds$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

t-test of whether *r* is different from zero

Correlation coefficient, *r*

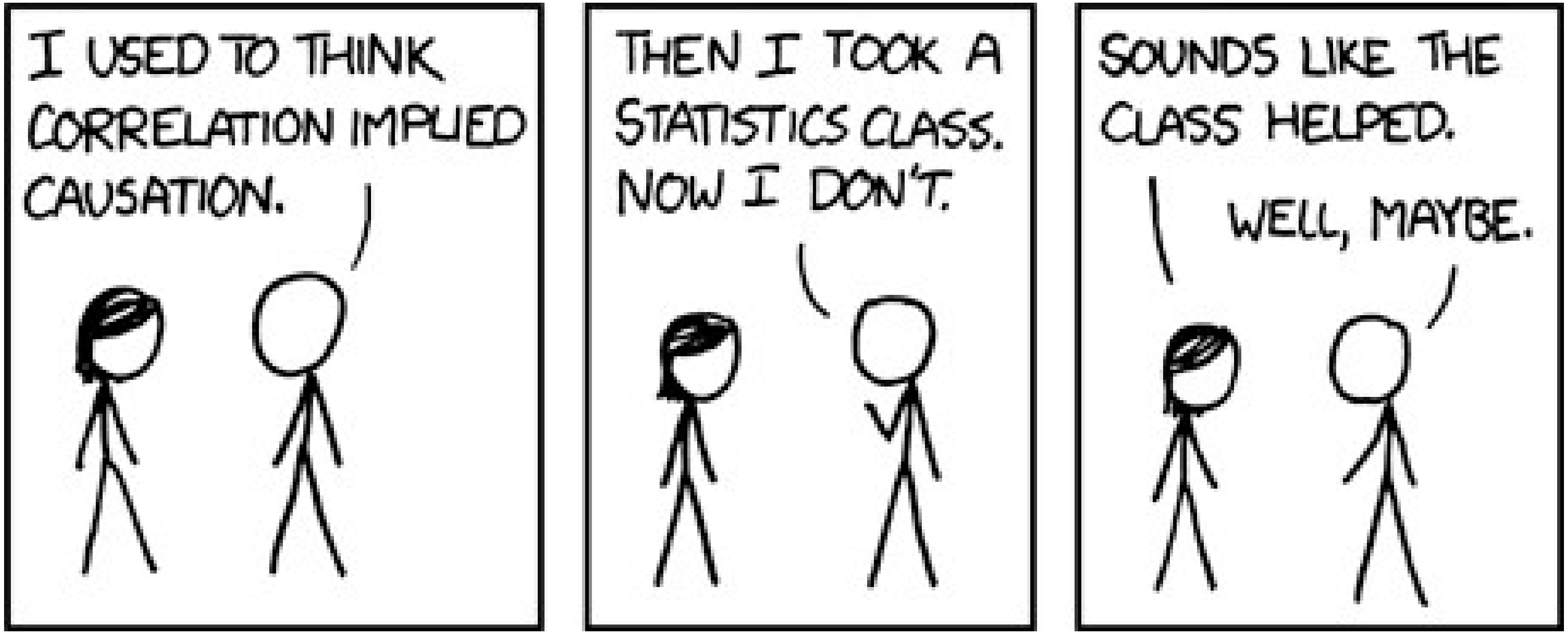
Reporting the result

```
data:  seeds$compactness and seeds$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


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

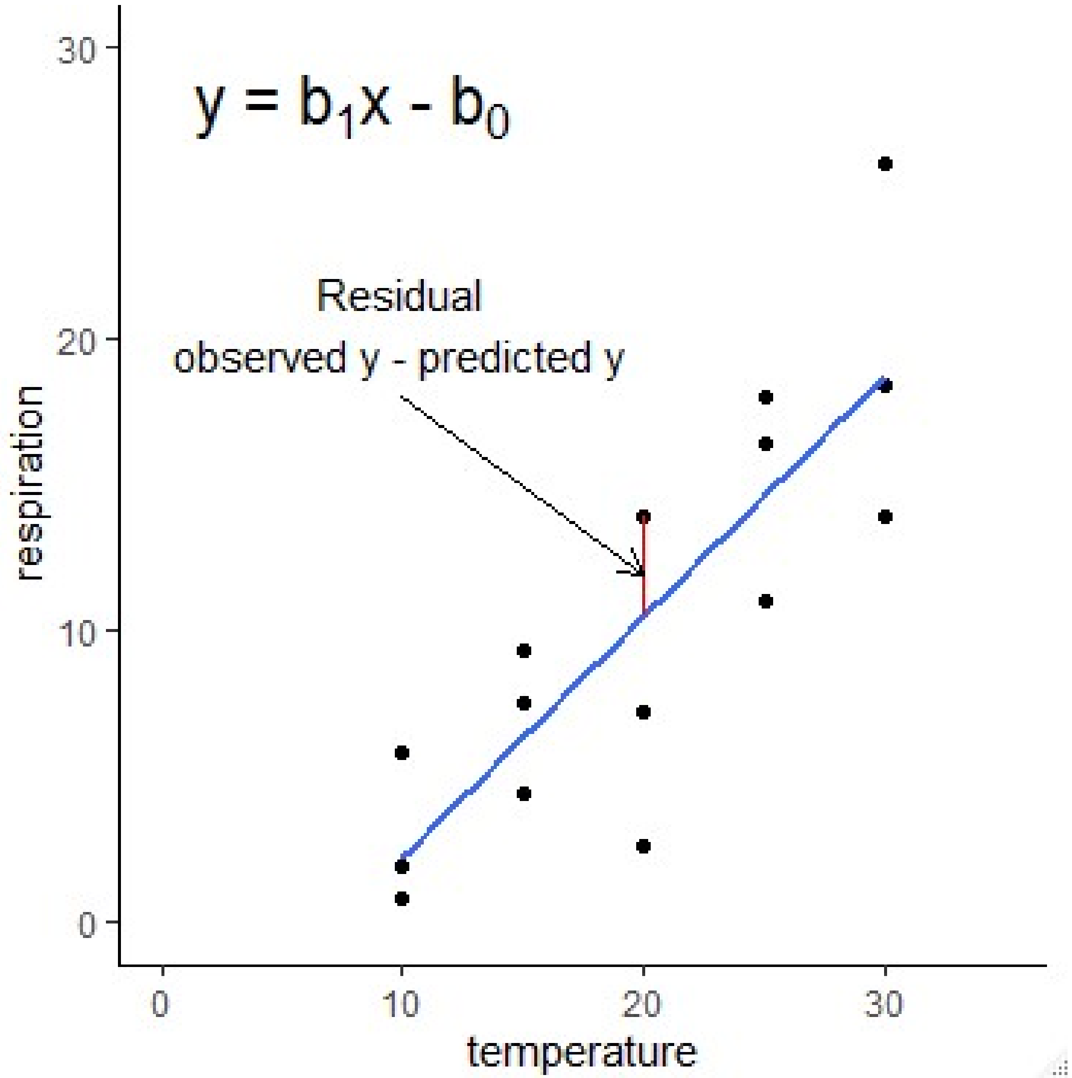
Regression

- Prediction
- One variable causes the other
- Axes matter
- We will consider linear regression only
  best fitting straight line:

$$y = b_1 x + b_0$$

Regression

The terminology



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
- *y* values are independent
- *x* is measured is chosen/set

Example

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

Assumptions

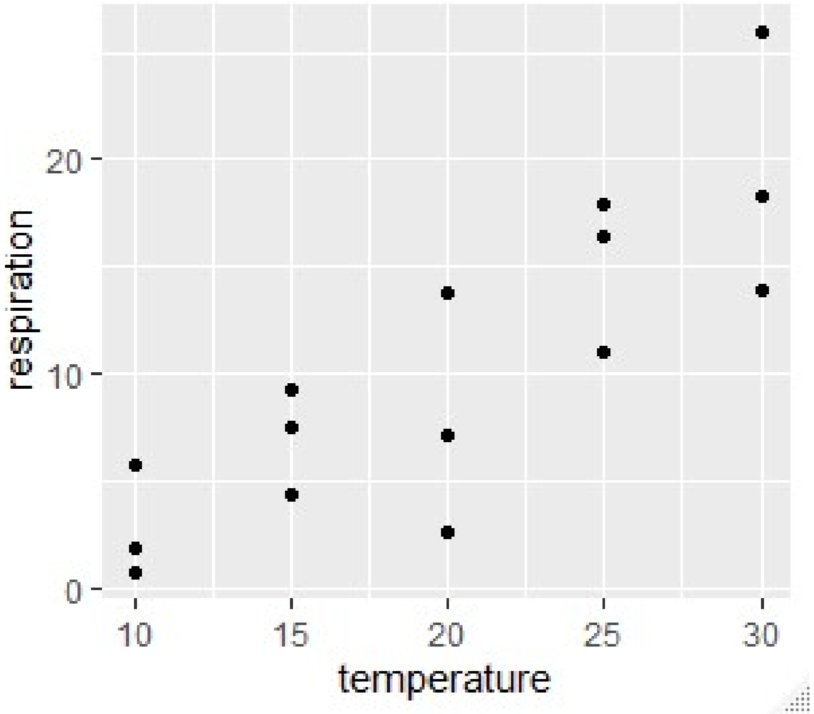
- Normality and homoscedascity of residuals
- *y* values are independent
- *x* is measured is chosen/set

	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
```

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

Reporting the results

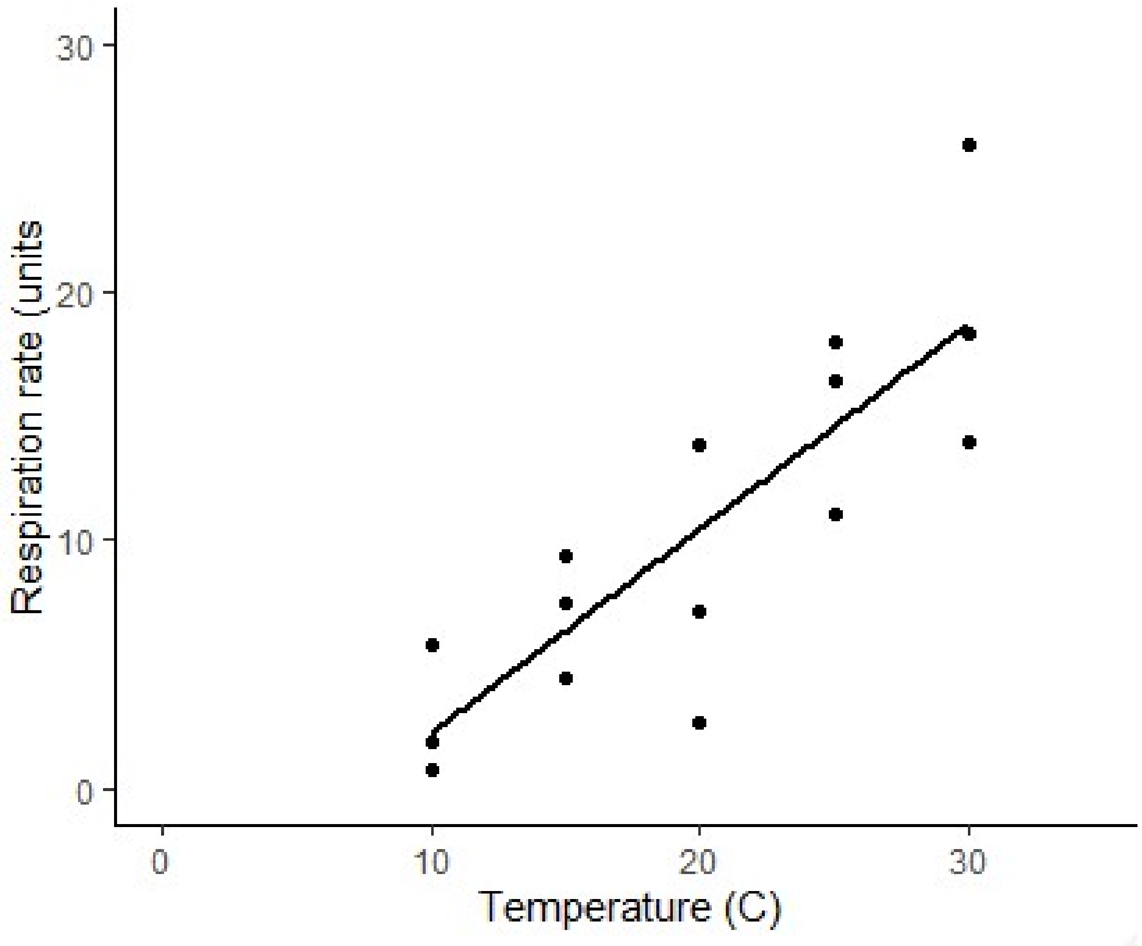
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.0359     3.1560  -1.912   0.0781 .
temperature   0.8253     0.1488   5.547 9.43e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
    1

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

Regression

Reporting the results: figure

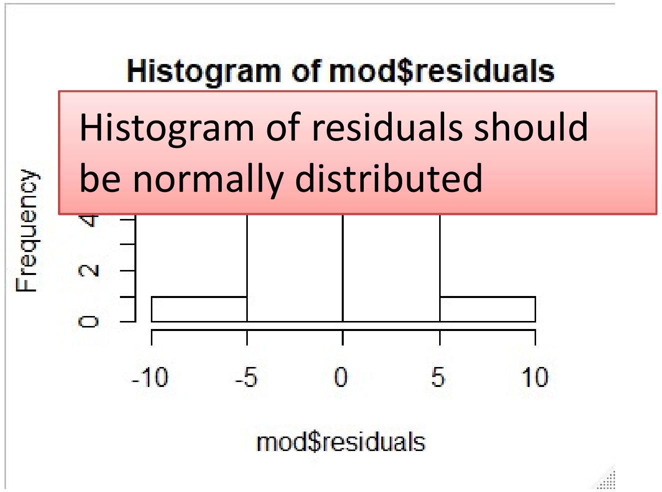


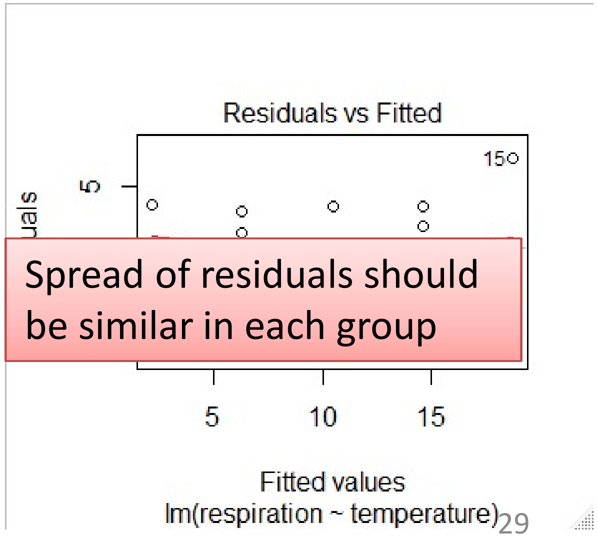
Regression

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)
```



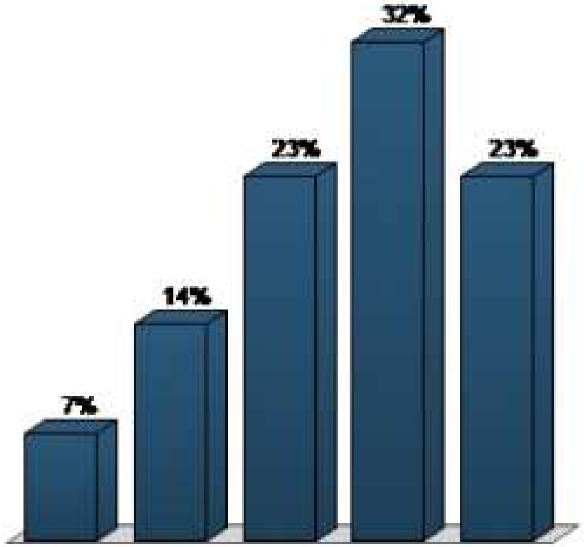


Summary of reporting

- Correlation - association
  - quote r, its significance (*p*) and *n*
  - if scatterplot included do NOT show a fitted line
- Regression - relationship
  - quote regression equation and test result (either ANOVA or *t*)
  - may also quote r^2 but not *r*
  - if scatterplot included do show a fitted line

2. I will enjoy the data analysis part of the 17C module? (Multiple Choice)

	Responses	
	Percent	Count
Definitely agree	6.96%	8
Probably agree	13.91%	16
Neutral	23.48%	27
Probably disagree	32.17%	37
Definitely disagree	23.48%	27
Totals	**100%**	**115**



Learning objectives for the week

By actively following the lecture and practical and carrying out the independent study the successful student will be able to:

- Explain the principles of correlation and of regression (MLO 1)
- Apply (appropriately), interpret and evaluate the legitimacy of, both in R (MLO 2, 3 and 4)
- Summarise and illustrate with appropriate R figures test results scientifically (MLO 3 and 4)