Introduction to Q&C and linear models revisited

58I Lab and Prof Skills II Quantitative and Computational skills

Lecture Overview

Introduction to Q&C skills strand

- Q&C skills strand in 58I
- Data Skills in degree program - roadmap

Linear models revisited

- Stage 1 - revision, brief!
- Linear models - what are they?
- Revisiting regression, t-tests and ANOVA as linear models

Learning Objectives for 58I

1. To be able to generate a testable hypothesis.
2. To design and conduct experiments to test this hypothesis, with appropriate controls.
3. To have practical experience of a range of techniques relevant to the discipline.
4. To work effectively within a team.
5. To be able to write a scientific report based on practical work.
6. To communicate scientific information and ideas in the form of a variety of media to a variety of audiences.
7. To use appropriate graphical methods to produce data figures with appropriately detailed legends.
8. To use relevant statistical or other analytical methods to analyse data.
9. To research scientific literature in a given area, and write an extended and well-structured account.

Assessment of Q&C: Express competency in Experimental Design and Bioscience Techniques (and elsewhere).

Topics covered in 58I Q&C

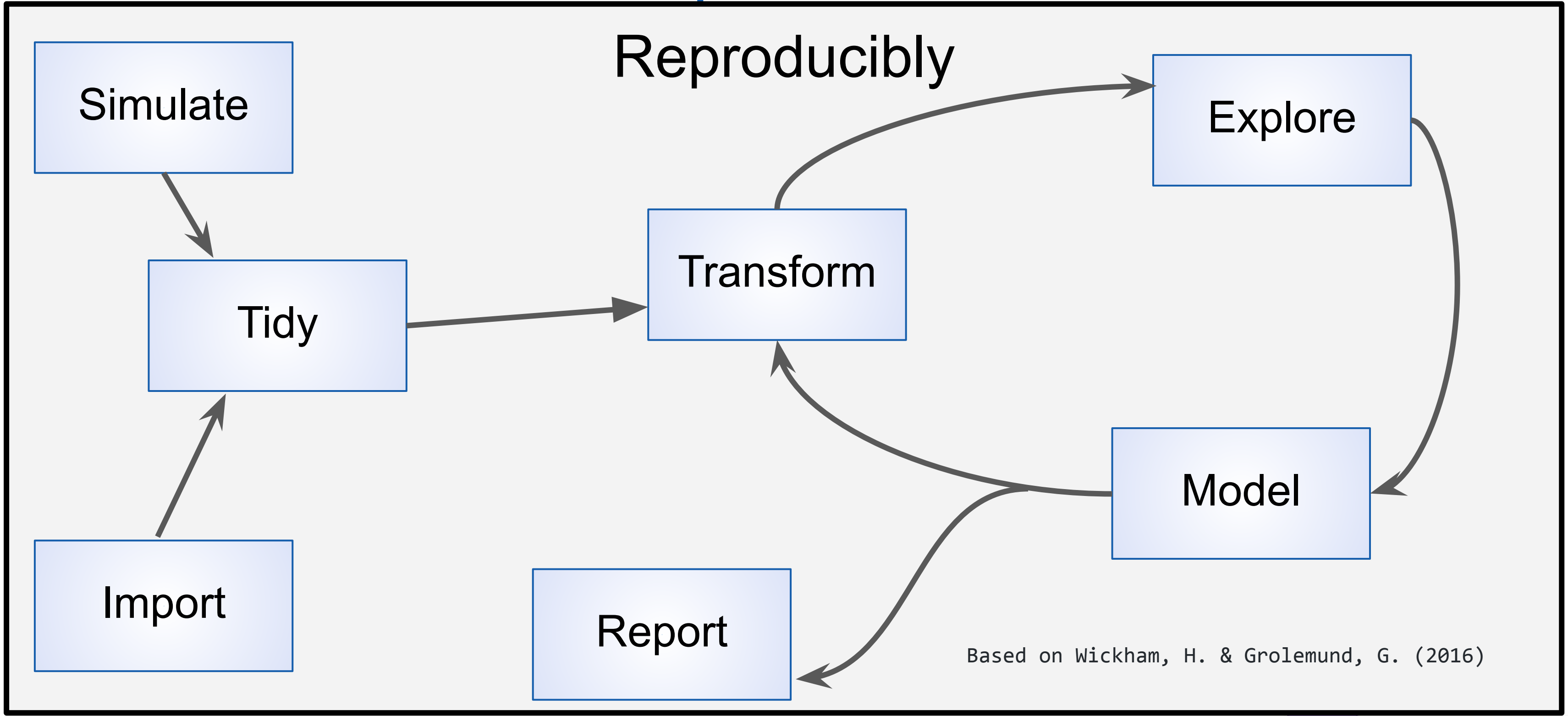
Impossible to cover everything to you might ever need!

Chosen topics are: foundational, follow stage 1 well, widely applicable (in this module and beyond), transferable conceptually:

- Generalised Linear Models:
- Non-linear Models (non-linear regression)

Methods which are very specific to the Experimental Design / Bioscience Technique taken are covered in that option. Talk to your project leader.

Data Skills are reproducible actions with data



Reproducibly

Simulate

Tidy

Import

Transform

Explore

Model

Report

Based on Wickham, H. & Grolemund, G. (2016)

ROADMAP: Stage 1

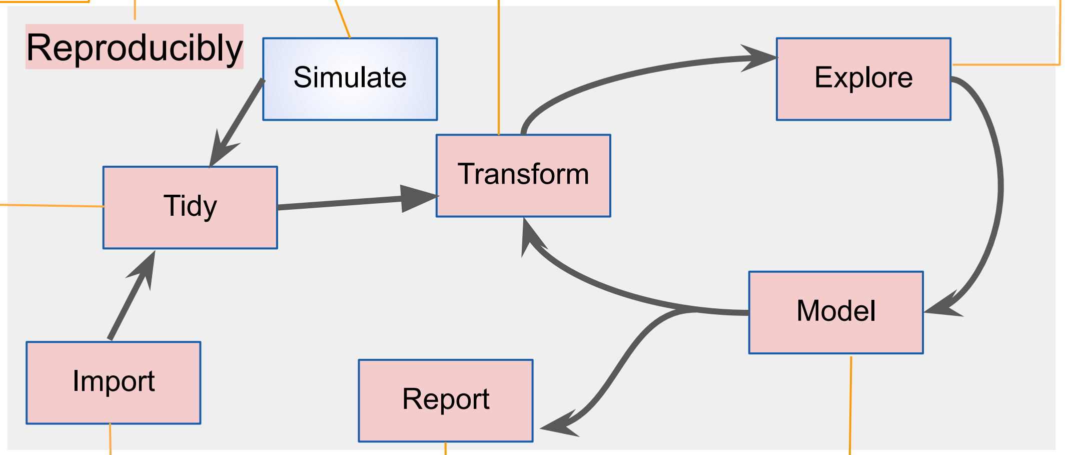
Introductory

Everything scripted
Code commenting
Organisation of analysis

Abstraction

ranking,
logging

Simple plots:
histograms
Normality testing
Summary stats



What ‘tidy’ data are but little tidying.

Changing variable names and types
Factor levels
Wide to long reshaping

Fundamental concepts in hypothesis testing
CI, Linear models (*t*-tests, ANOVA, regression), correlation

Multiple comparison

Selection:
Assumptions
Model fit: not really

From files - all but unusually complex
.txt, .xlsx, .csv, .sav, .dta

Relative paths
Separators
…..and more

“significance, direction, magnitude”
Figures: legends, saving
Not fully reproducibly

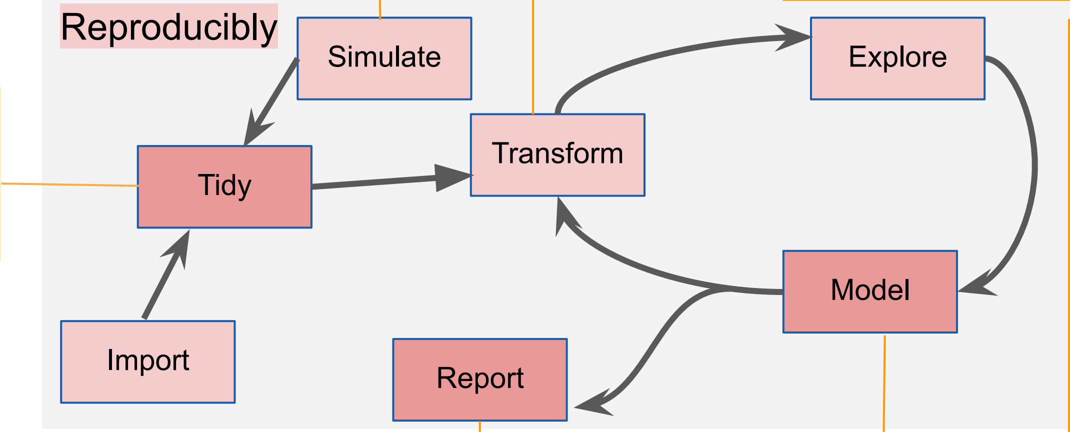
Introductory

Intermediate

Stage 2

Depending on options:
Abstraction
Running and interpreting particular models

Inevitably



Reproducibly

Simulate

Tidy

Import

Transform

Explore

Model

Report

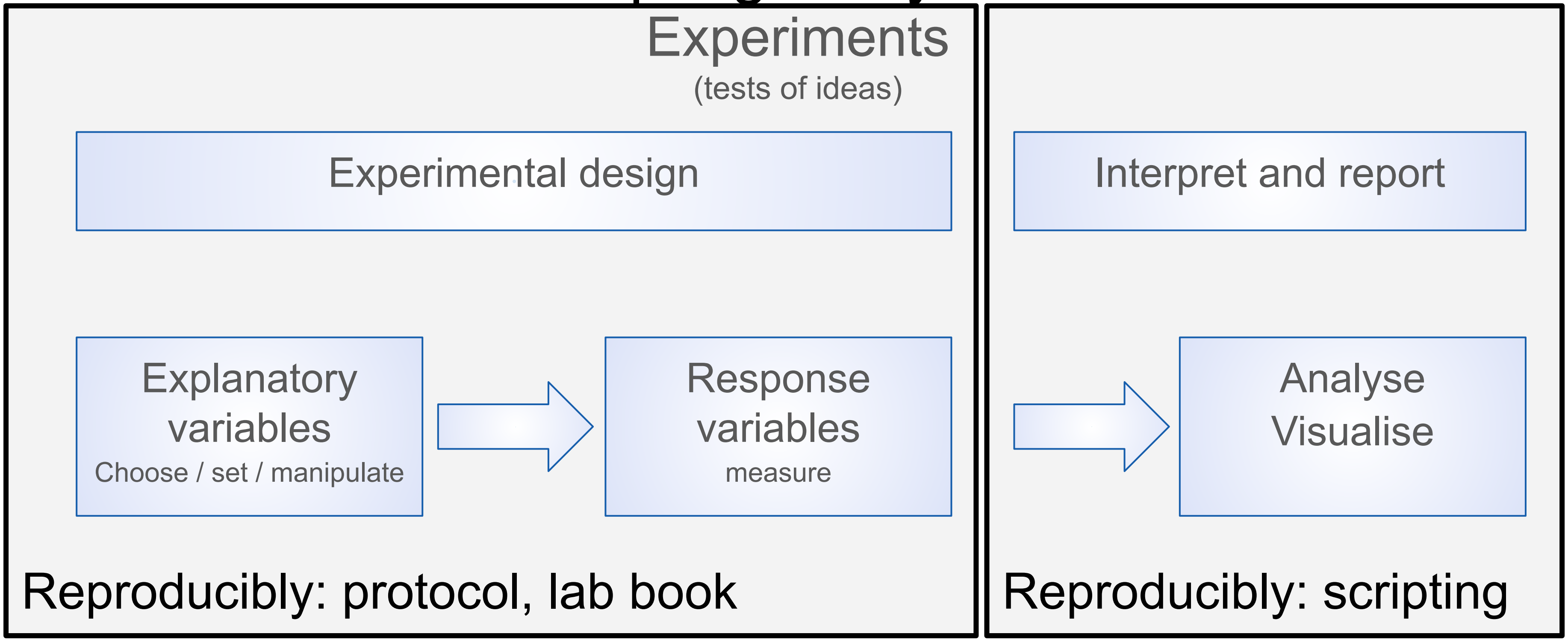
Depending on options:
Proportions
Z score standardisation
Coefficient of variation
Log to base 2
Subtraction of noise/background
Scaling/reversing experimental steps
PCR Relative quantification
RPKM quantification

Explicitly:
Stage 1 tests in LM framework (increased conceptual complexity)
More LM
GLM - Binomial and Poisson
Odds ratios
Deviance measures of fit
More on Multiple comparisons
Non-linear regression

Depending on options:
Mixed models
FDR
GWAS
bootstrapping

Multi panel figures
Complex domain specific figures

The rationale for scripting analysis



Experiments

(tests of ideas)

Experimental design

Explanatory variables

Choose / set / manipulate

Response variables

measure

Reproducibly: protocol, lab book

Interpret and report

Analyse

Visualise

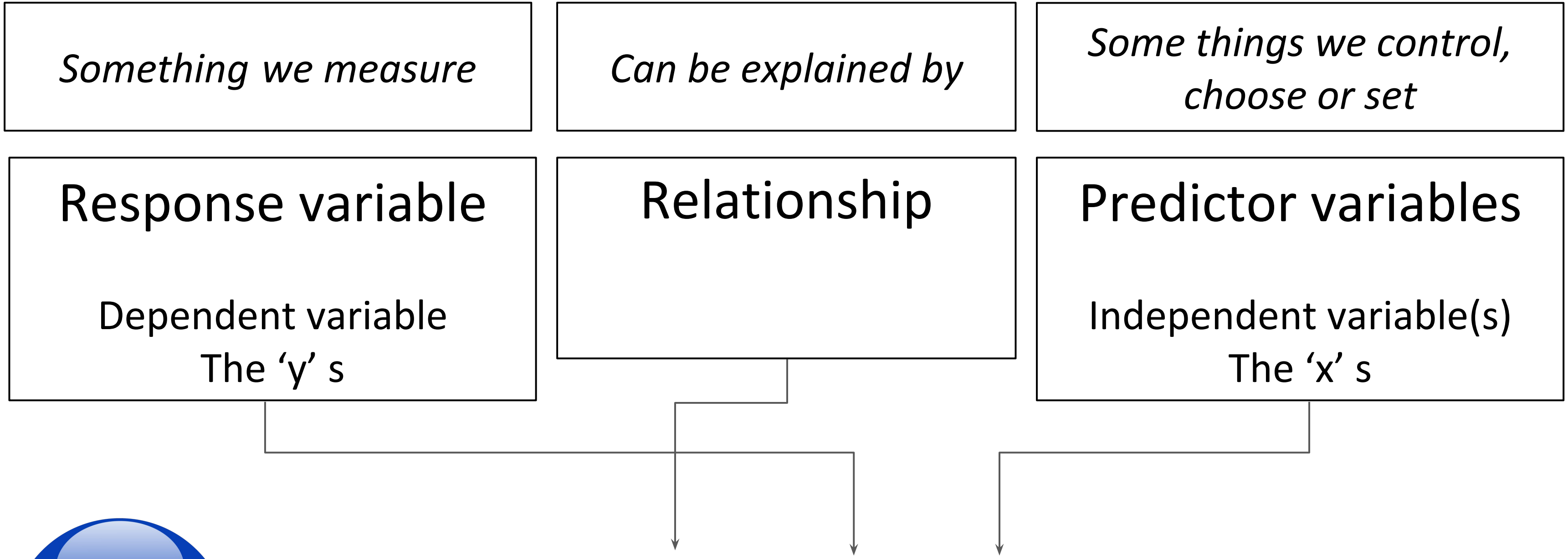
Reproducibly: scripting

Why R?

It’s a good choice but not the only option.

- R caters to “users who do not see themselves as programmers, but then allows them to slide gradually into programming”
- Community, active, relatively diverse
- Language designed for data analysis and visualisation so makes those easy
- Open source, Free,
- Reproducibility - R markdown, R’s “killer feature”

Stage 1 Revision: experiments and analysis



Something we measure

Can be explained by

Some things we control, choose or set

Response variable

Dependent variable
The 'y' s

Relationship

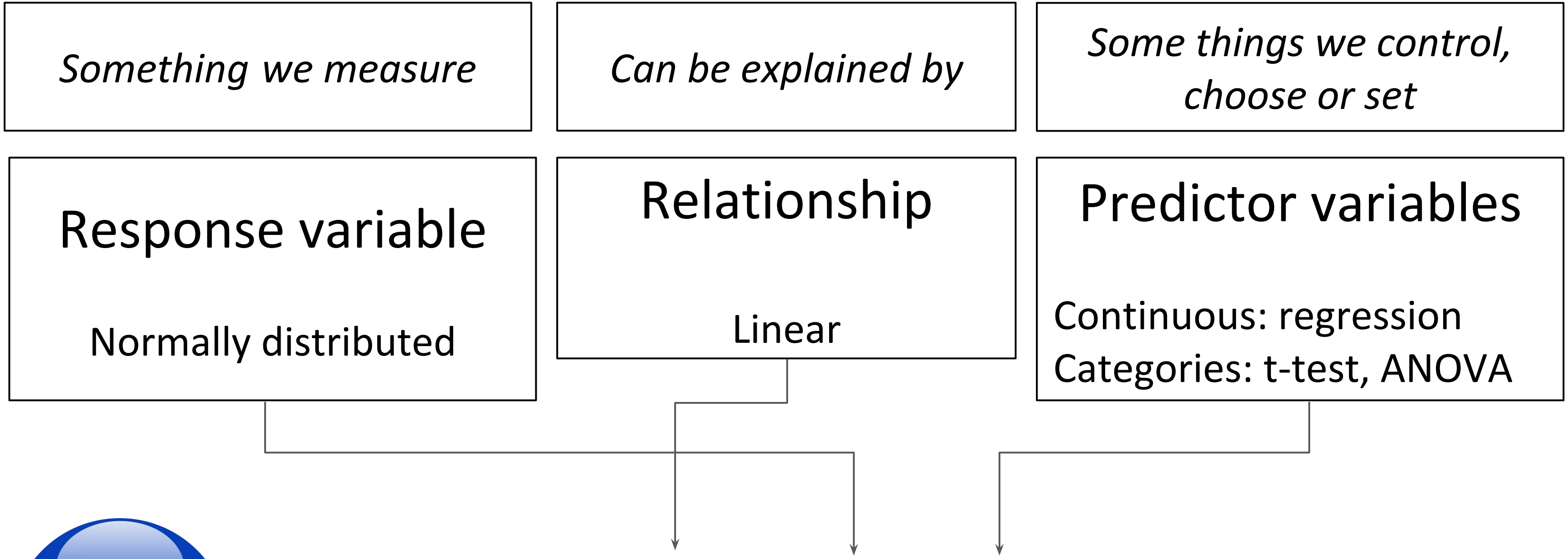
Predictor variables

Independent variable(s)
The 'x' s

$$\text{function}(y \sim x)$$

$$\text{function}(y \sim x_1 * x_2)$$

Stage 1 Revision: experiments and analysis



Something we measure

Can be explained by

Some things we control, choose or set

Response variable

Normally distributed

Relationship

Linear

Predictor variables

Continuous: regression
Categories: t-test, ANOVA

$$\texttt{function}(y \sim x)$$

$$\texttt{function}(y \sim x_1 * x_2)$$

Contact time: 1 lecture + 4 workshops

Lecture 1 : Introduction to Generalised Linear Models (ER)

Workshop 1: Linear Models (ER)

T-tests, ANOVA and regression are used when we have a continuous response variable. We revisit these using a linear modelling framework. This means using a single function `lm()` rather than three different ones and enhancing our understanding of the concepts underlying the tests.

Workshop 2: Generalised Linear Models for Poisson distributed data (ER)

Workshop 3: Generalised Linear Models for Binomially distributed data (ER)

Workshop 4: Non-linear regression and dynamics (JWP)

Lecture Overview

Introduction to Q&C skills strand

- Q&C skills strand in 58I ✓
- Data Skills in degree program - roadmap ✓

Linear models revisited

- Stage 1 - revision, brief! ✓
- Linear models - what are they? ←
- Revisiting regression, t-tests and ANOVA as linear models

Learning objectives

By actively following this lecture and undertaking the exercises in workshop 1 the successful student will be able to:

- Explain the the link between t-tests, ANOVA and regression
- Appropriately apply linear models using lm()
- Interpret the results using summary() and anova() and relate them to the outputs of t.test() and aov()

What are linear models?

Something you have already met!

Equation to explain, with a linear relationship, one response variable with one or more explanatory variables: $y = ax_1 + bx_2 + ...$

Procedure	Response	Explanatory	R	Stage 1 examples
Single linear regression	Continuous	1 Continuous	y ~ x	mand ~ jh mass ~ day
Two-sample t-test	Continuous	1 categorical (2 levels)	y ~ x	adiponectin ~ treatment time ~ status
One-way ANOVA	Continuous	1 categorical (2 or more levels)	y ~ x	myoglobin ~ species
Two-way ANOVA	Continuous	2 categorical (2 or more levels each)	y ~ x1*x2	para ~ season * species diameter ~ agent * species

Key points

T-tests, ANOVA and regression are fundamentally the same, collectively called ‘general linear models’. They can be carried out in R with `lm()`

There are other linear models too

The concept can be extended to ‘generalised linear models’ for different types of response. Generalised linear models are carried out in R with `glm()`

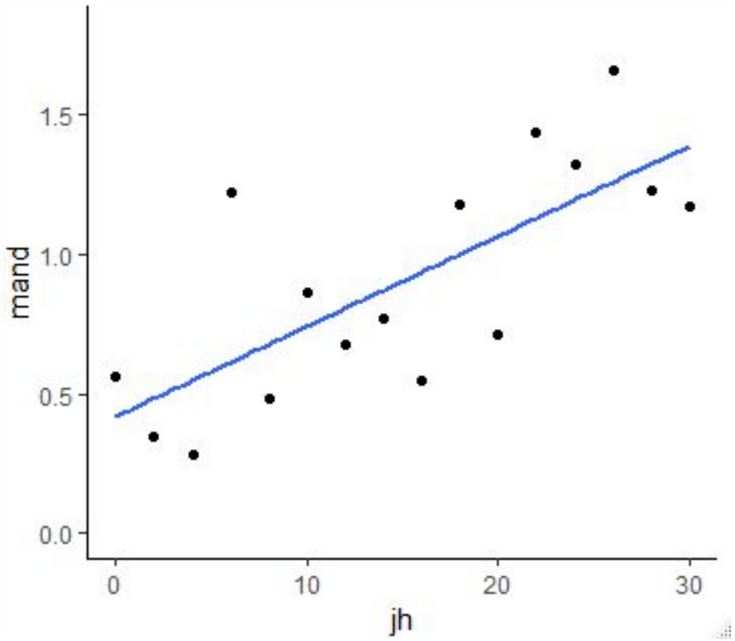
The output of `lm()` looks more complex, at first, than the outputs of `t.test()` and `aov()`

The output of `glm()` is like that for `lm()`. So we will revisit regression, t-tests and ANOVA using `lm()` to help you understand the output.

Revisiting: Regression - this is exactly as last year!

Concentration of juvenile hormone (JH) and mandible length in stag beetles

```
mod <- lm(data = stag, mand ~ jh)
```



Revisiting: Regression - this is exactly as last year!

```
mod <- lm(data = stag, mand ~ jh)
```

```
summary(mod)
Call:
lm(formula = mand ~ jh, data = stag)

Residuals:
     Min        1Q    Median        3Q       Max 
-0.38604 -0.20281 -0.09751  0.15034  0.60690 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.419338   0.139429   3.008  0.00941 **
jh          0.032294   0.007919   4.078  0.00113 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.292 on 14 degrees of freedom
Multiple R-squared:  0.5429,    Adjusted R-squared:  0.5103 
F-statistic: 16.63 on 1 and 14 DF,  p-value: 0.00113
```

Revisiting: Regression - this is exactly as last year!

mand = 0.42 + 0.03*jh

```
mod <- lm(data = stag, mand ~ jh)
```

```
summary(mod)
Call:
lm(formula = mand ~ jh, data = stag)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.38604 -0.20281 -0.09751  0.15034  0.60690 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.419338   0.139429   3.008  0.00941 **
jh          0.032294   0.007919   4.078  0.00113 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.292 on 14 degrees of freedom
Multiple R-squared:  0.5429,	Adjusted R-squared:  0.5103 
F-statistic: 16.63 on 1 and 14 DF,  p-value: 0.00113
```

Intercept

Slope

Test of intercept

Test of slope

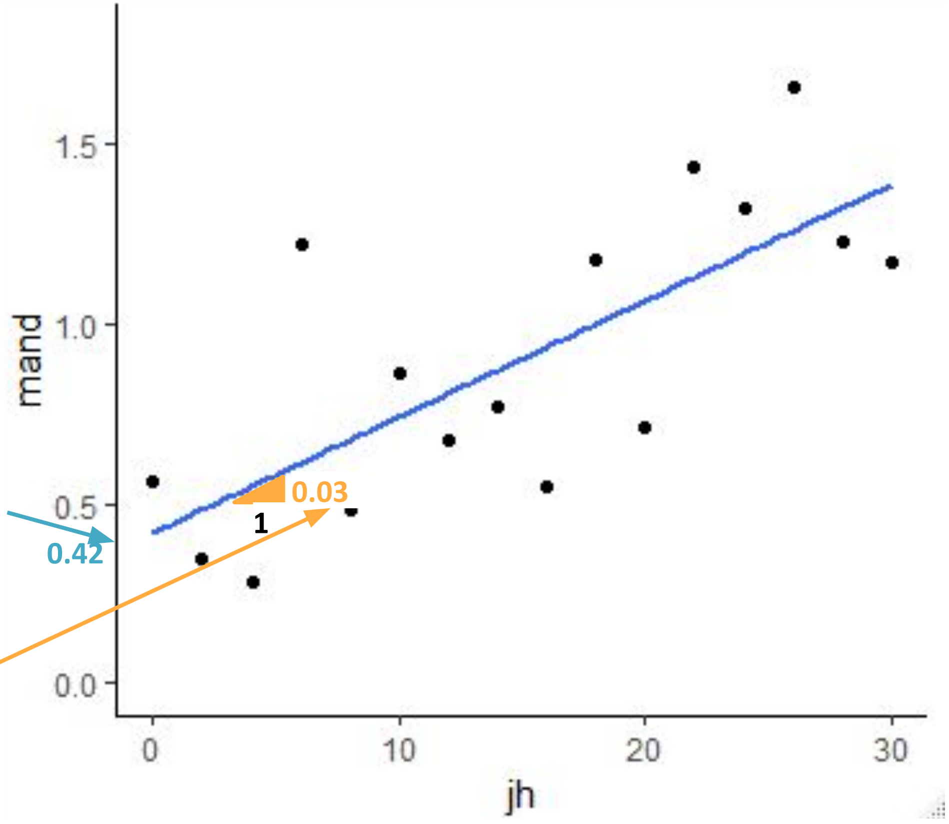
% of variation in y explained by x
“model fit”

Test of model

Revisiting: Regression - thi

```
mod <- lm(data = stag, mand ~ jh)
```

```
summary(mod)
Call:
lm(formula = mand ~ jh, data = stag)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.38604 -0.20281 -0.09751  0.15034  0.60690 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.419338   0.139429   3.008  0.00941
jh          0.032294   0.007919   4.078  0.00113
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 0.292 on 14 degrees of freedom
Multiple R-squared:  0.5429,  Adjusted R-squared:  0.5103 
F-statistic: 16.63 on 1 and 14 DF,  p-value: 0.00113
```



Revisiting: Regression - this is exactly as last year!

```
mod <- lm(data = stag, mand ~ jh)

   summary(mod)
   Call:
   lm(formula = mand ~ jh, data = stag)

   Residuals:
        Min       1Q   Median       3Q      Max
   -0.38604 -0.20281 -0.09751  0.15034  0.60690

   Coefficients:
               Estimate Std. Error t value Pr(>|t|)
   (Intercept) 0.419338   0.139429   3.008  0.00941 **
   jh          0.032294   0.007919   4.078  0.00113 **
   ---
   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

   Residual standard error: 0.292 on 14 degrees of freedom
   Multiple R-squared:  0.5429,    Adjusted R-squared:  0.5103
   F-statistic: 16.63 on 1 and 14 DF,  p-value: 0.00113
```

When only one continuous variable after the ~

….

P value for slope of single variable
=
P value of whole model

This will not be true for more for
i) one-way anova with more than 2 gps
ii) two-way anova
iii) other linear models

Revisiting: two-sample t-test using t.test()

t.test(y ~ x, data = mydata, var.equal = T)

Example 1 from 17C.
Is there a significant difference between the masses of male and female chaffinches?

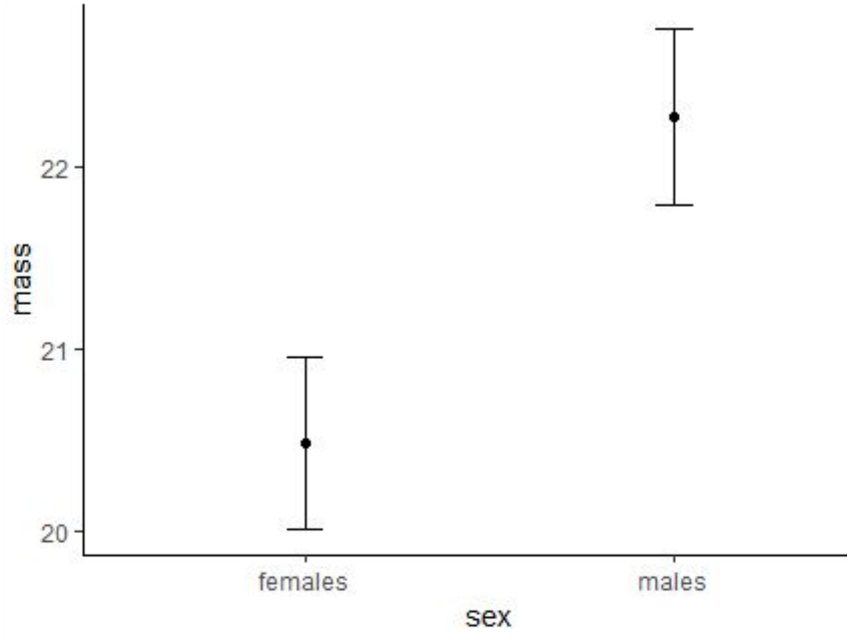
```
t.test(mass ~ sex, data = chaff, var.equal = T)

        Two Sample t-test
data:  mass by sex
t = -2.6471, df = 38, p-value = 0.01175
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.167734 -0.422266
sample estimates:
mean in group females   mean in group males 
               20.480                22.275 
```

Example 2 from 08C.
Does treatment with Nicotinic acid affect adiponectin secretion compared to control treatment?

```
t.test(adiponectin ~ treatment, data = adip, var.equal = T)

       Two Sample t-test
data:  adiponectin by treatment
t = -3.2728, df = 28, p-value = 0.00283
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.1910762 -0.7342571
sample estimates:
  mean in group control mean in group nicotinic 
               5.546000                7.508667 
```

Revisiting: two-sample t-test using t.test()

t.test(y ~ x, data = mydata, var.equal = T)



```
t.test(mass ~ sex, data = chaff, paired = F, var.equal = T)

	Two Sample t-test
data:  mass by sex
t = -2.6471, df = 38, p-value = 0.01175
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.167734 -0.422266
sample estimates:
mean in group females   mean in group males
               20.480                22.275
```

The means are significantly different

Alternative way to state:

- Sex has a significant effect on mass

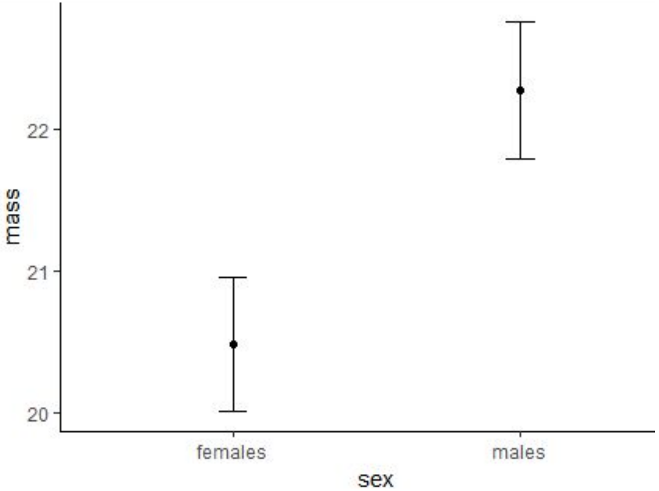
Revisiting: Comparing t.test() with lm()

Using t.test

```
t.test(mass ~ sex, data = chaff, paired = F, var.equal = T)

        Two Sample t-test
data:  mass by sex
t = -2.6471, df = 38, p-value = 0.01175
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.167734 -0.422266
sample estimates:
mean in group females   mean in group males
               20.480                22.275
```

Output of lm() to do a t-test looks the same as the output of lm() to do a regression.

Mathematically the same thing!



Using lm()

```
mod <- lm(mass ~ sex, data = chaff)
summary(mod)
Call:
lm(formula = mass ~ sex, data = chaff)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.2750 -1.7000 -0.3775  1.6200  4.1250 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  20.4800     0.4795  42.712   <2e-16 ***
sexmales      1.7950     0.6781   2.647   0.0118 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.144 on 38 degrees of freedom
Multiple R-squared:  0.1557,	Adjusted R-squared:  0.1335 
F-statistic: 7.007 on 1 and 38 DF,  p-value: 0.01175
```

Difference is significant

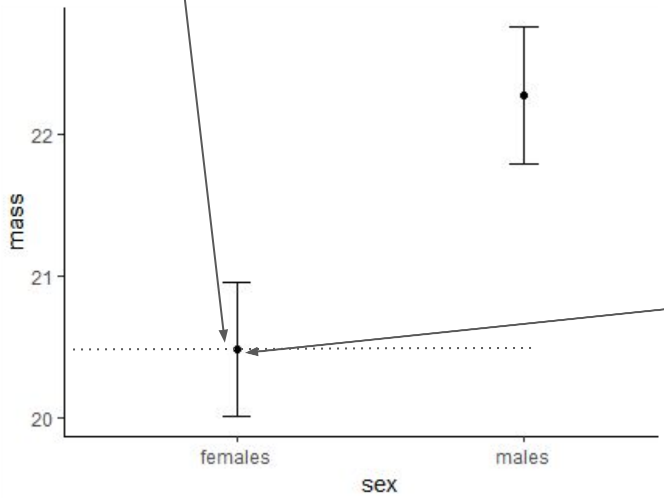
Revisiting: Comparing t.test() with lm()

Using t.test

```
t.test(mass ~ sex, data = chaff, paired = F, var.equal = T)

       Two Sample t-test
data:  mass by sex
t = -2.6471, df = 38, p-value = 0.01175
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.167734 -0.422266
sample estimates:
mean in group females   mean in group males
               20.480                22.275
```

Intercept is mean of 'lowest' level of factor

I.e., equivalent to x = 0 in regression



Using lm()

```
mod <- lm(mass ~ sex, data = chaff)
summary(mod)
Call:
lm(formula = mass ~ sex, data = chaff)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.2750 -1.7000 -0.3775  1.6200  4.1250 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  20.4800     0.4795  42.712   <2e-16 ***
sexmales      1.7950     0.6781   2.647   0.0118 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.144 on 38 degrees of freedom
Multiple R-squared:  0.1557,	Adjusted R-squared:  0.1335 
F-statistic: 7.007 on 1 and 38 DF,  p-value: 0.01175
```

Female mean sig diff from 0. Not important

Revisiting: Comparing t.test() with lm()

Using t.test

```
t.test(mass ~ sex, data = chaff, paired = F, var.equal = T)

        Two Sample t-test
data:  mass by sex
t = -2.6471, df = 38, p-value = 0.01175
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.167734 -0.422266
sample estimates:
mean in group females   mean in group males
               20.480                22.275
```

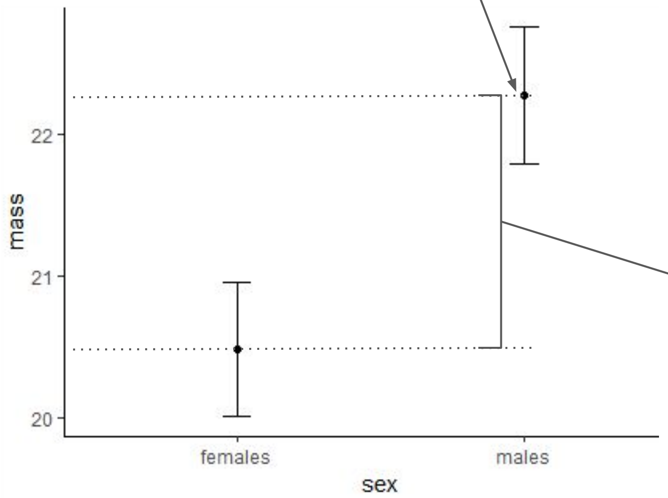
Difference between intercept and next level (i.e., the slope)

I.e., Changing x by 1 unit makes y go up by the value of slope

Using lm()

```
mod <- lm(mass ~ sex, data = chaff)
summary(mod)
Call:
lm(formula = mass ~ sex, data = chaff)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.2750 -1.7000 -0.3775  1.6200  4.1250 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  20.4800     0.4795  42.712   <2e-16 ***
sexmales      1.7950     0.6781   2.647   0.0118 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.144 on 38 degrees of freedom
Multiple R-squared:  0.1557,	Adjusted R-squared:  0.1335 
F-statistic: 7.007 on 1 and 38 DF,  p-value: 0.01175
```

Difference is significant



Why use lm()?

Extendable! These are particular cases but a linear models include any number of continuous and categorical explanatory variables.

Procedure	Response	Explanatory	R	Stage 1 examples
Single linear regression	Continuous	1 Continuous	y ~ x	mand ~ jh mass ~ day
Two-sample t-test	Continuous	1 categorical (2 levels)	y ~ x	adiponectin ~ treatment time ~ status
One-way ANOVA	Continuous	1 categorical (2 or more levels)	y ~ x	myoglobin ~ species
Two-way ANOVA	Continuous	2 categorical (2 or more levels each)	y ~ x1*x2	para ~ season * species diameter ~ agent * species

Why use lm()?

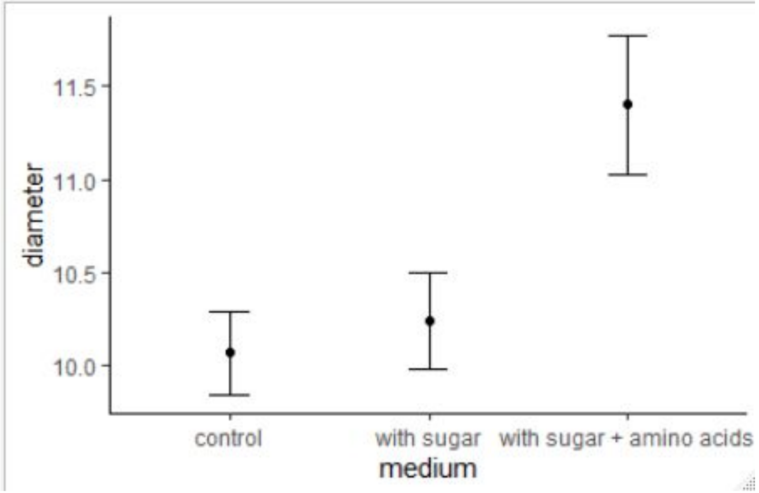
For example...

Procedure	Response	Explanatory	R	Stage 1 examples
Single linear regression	Continuous	1 Continuous	y ~ x	mand ~ jh mass ~ day
Two-sample t-test	Continuous	1 categorical (2 levels)	y ~ x	adiponectin ~ treatment time ~ status
One-way ANOVA	Continuous	1 categorical (2 or more levels)	y ~ x	myoglobin ~ species
Two-way ANOVA	Continuous	2 categorical (2 or more levels each)	y ~ x1*x2	para ~ season * species diameter ~ agent * species
	Continuous	1 categorical and 1 continuous	y ~ x1*x2	

Revisiting: One-way ANOVA

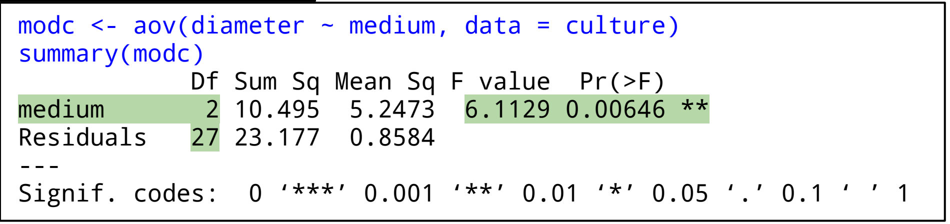
```
mod <- aov(y ~ x, data = mydata)
summary(mod)
```

```
modc <- aov(diameter ~ medium, data = culture)
summary(modc)
            Df Sum Sq Mean Sq F value  Pr(>F)   
medium       2 10.495  5.2473  6.1129 0.00646 **
Residuals   27 23.177  0.8584                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



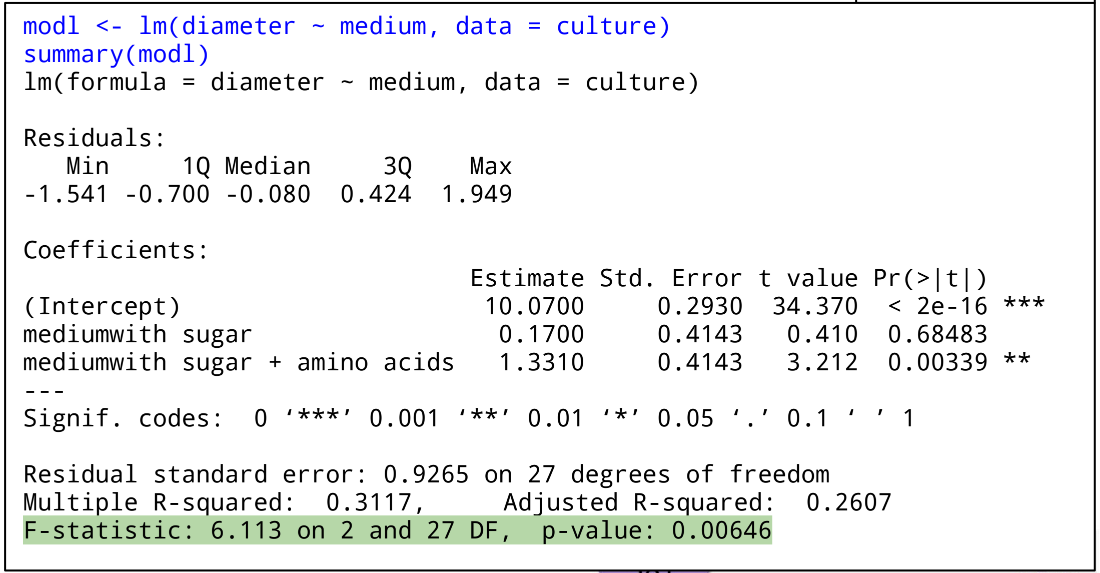
Revisiting: One-way ANOVA

Using aov()



```
modc <- aov(diameter ~ medium, data = culture)
summary(modc)
            Df Sum Sq Mean Sq F value  Pr(>F)
medium       2 10.495  5.2473  6.1129 0.00646 **
Residuals   27 23.177  0.8584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Using lm()



```
modl <- lm(diameter ~ medium, data = culture)
summary(modl)
lm(formula = diameter ~ medium, data = culture)

Residuals:
   Min      1Q Median      3Q     Max
-1.541 -0.700 -0.080  0.424  1.949

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     10.0700     0.2930  34.370  < 2e-16 ***
mediumwith sugar                 0.1700     0.4143   0.410  0.68483
mediumwith sugar + amino acids   1.3310     0.4143   3.212  0.00339 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9265 on 27 degrees of freedom
Multiple R-squared:  0.3117,     Adjusted R-squared:  0.2607
F-statistic: 6.113 on 2 and 27 DF,  p-value: 0.00646
```

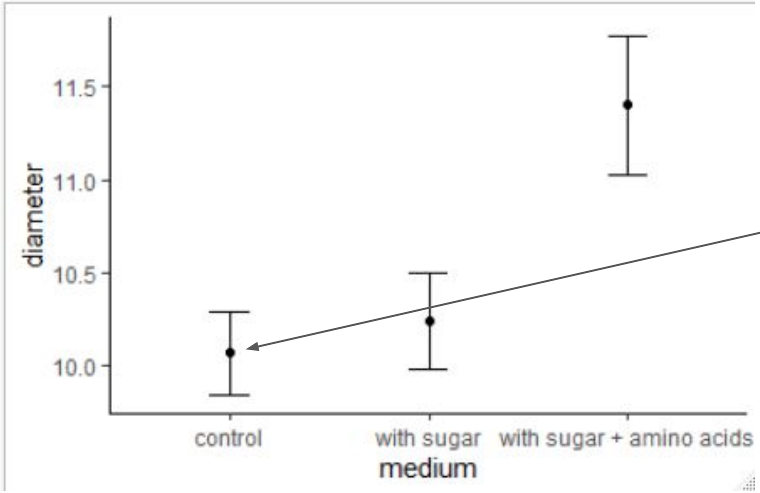
Revisiting: One-way ANOVA

Using aov()

```
modc <- aov(diameter ~ medium, data = culture)
summary(modc)
            Df Sum Sq Mean Sq F value  Pr(>F)
medium       2 10.495  5.2473  6.1129 0.00646 **
Residuals   27 23.177  0.8584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Intercept is mean of ‘lowest’ level of factor

I.e., equivalent to x = 0 in regression



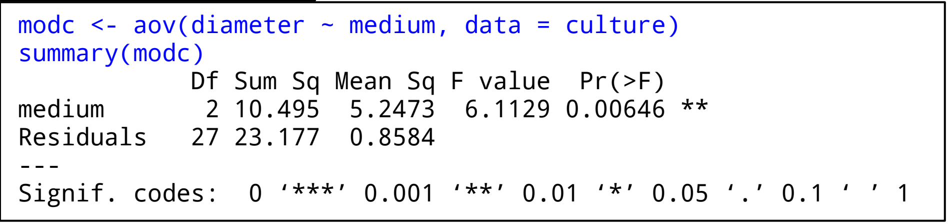
Using lm()

```
modl <- lm(diameter ~ medium, data = culture)
summary(modl)
lm(formula = diameter ~ medium, data = culture)

Residuals:
   Min     1Q Median     3Q    Max
-1.541 -0.700 -0.080  0.424  1.949

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    10.0700     0.2930  34.370  < 2e-16 ***
mediumwith sugar                0.1700     0.4143   0.410  0.68483
mediumwith sugar + amino acids  1.3310     0.4143   3.212  0.00339 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9265 on 27 degrees of freedom
Multiple R-squared:  0.3117,    Adjusted R-squared:  0.2607
F-statistic: 6.113 on 2 and 27 DF,  p-value: 0.00646
```

Control mean sig diff from 0. Not important

Revisiting: One-way ANOVA

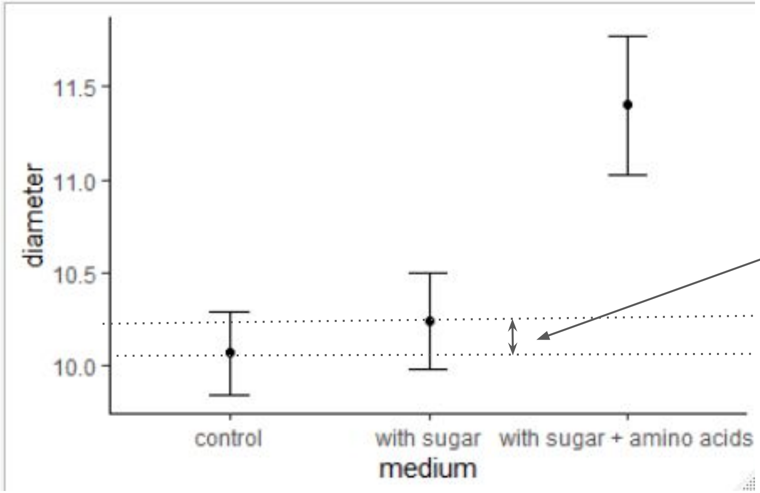
Using aov()

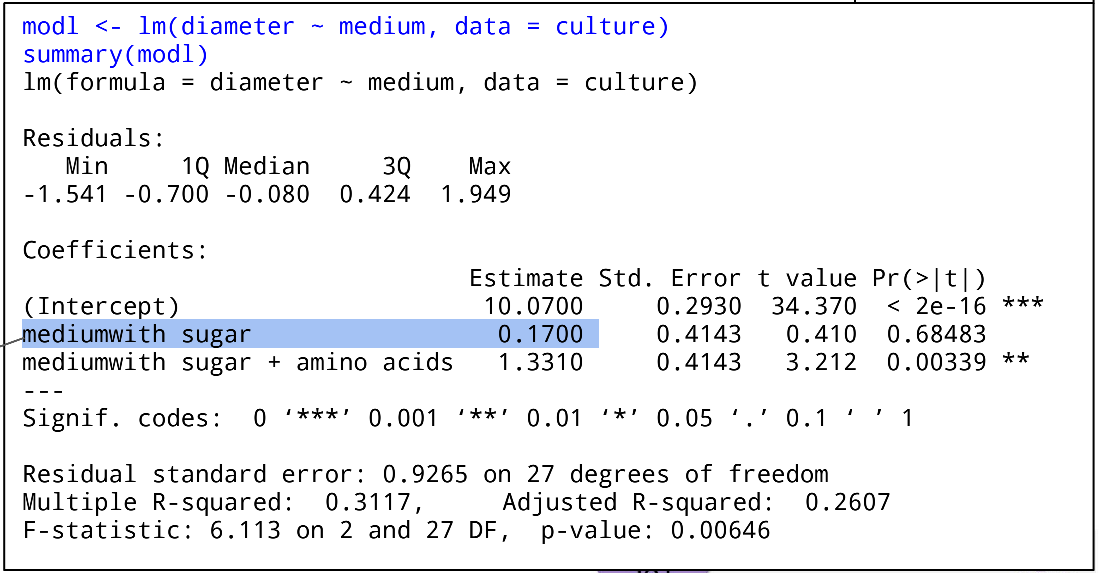


```
modc <- aov(diameter ~ medium, data = culture)
summary(modc)
            Df Sum Sq Mean Sq F value  Pr(>F)
medium       2 10.495  5.2473  6.1129 0.00646 **
Residuals   27 23.177  0.8584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Difference between intercept and next level

Using lm()



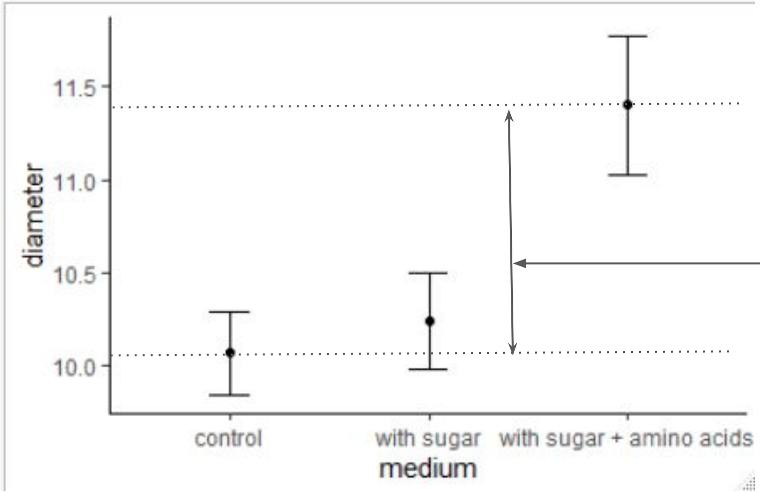


```
modl <- lm(diameter ~ medium, data = culture)
summary(modl)
lm(formula = diameter ~ medium, data = culture)

Residuals:
   Min     1Q Median     3Q    Max
-1.541 -0.700 -0.080  0.424  1.949

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     10.0700     0.2930  34.370  < 2e-16 ***
mediumwith sugar                 0.1700     0.4143   0.410  0.68483
mediumwith sugar + amino acids   1.3310     0.4143   3.212  0.00339 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9265 on 27 degrees of freedom
Multiple R-squared:  0.3117,     Adjusted R-squared:  0.2607
F-statistic: 6.113 on 2 and 27 DF,  p-value: 0.00646
```

Revisiting: One-way ANOVA

Using aov()

```
modc <- aov(diameter ~ medium, data = culture)
summary(modc)
            Df Sum Sq Mean Sq F value  Pr(>F)
medium       2 10.495  5.2473  6.1129 0.00646 **
Residuals   27 23.177  0.8584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Difference between intercept and third level

Using lm()

```
modl <- lm(diameter ~ medium, data = culture)
summary(modl)
lm(formula = diameter ~ medium, data = culture)

Residuals:
   Min     1Q Median     3Q    Max
-1.541 -0.700 -0.080  0.424  1.949

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     10.0700     0.2930  34.370  < 2e-16 ***
mediumwith sugar                 0.1700     0.4143   0.410  0.68483
mediumwith sugar + amino acids   1.3310     0.4143   3.212  0.00339 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9265 on 27 degrees of freedom
Multiple R-squared:  0.3117,	Adjusted R-squared:  0.2607
F-statistic: 6.113 on 2 and 27 DF,  p-value: 0.00646
```

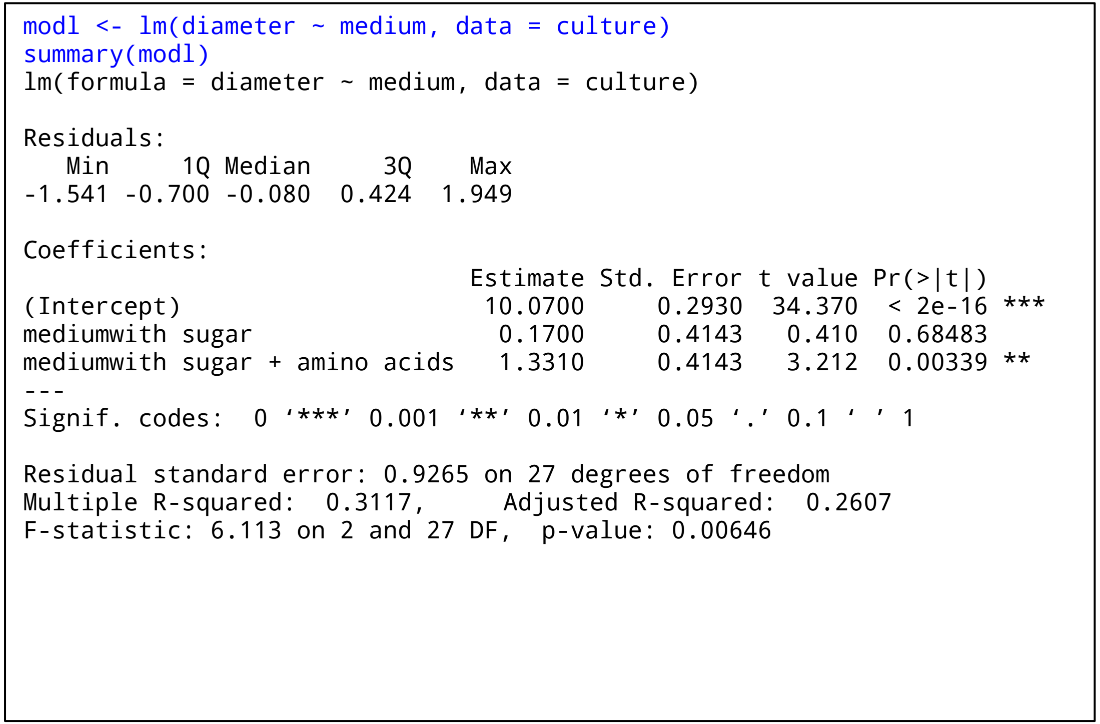
Usual steps in applying lm()

lm()

summary(mod1) - ‘estimates’ and direction of effects

+’ve bigger than intercept

-’ve smaller than intercept



```
modl <- lm(diameter ~ medium, data = culture)
summary(modl)
lm(formula = diameter ~ medium, data = culture)

Residuals:
   Min      1Q Median     3Q    Max
-1.541 -0.700 -0.080  0.424  1.949

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    10.0700     0.2930  34.370  < 2e-16 ***
mediumwith sugar                0.1700     0.4143   0.410  0.68483
mediumwith sugar + amino acids  1.3310     0.4143   3.212  0.00339 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9265 on 27 degrees of freedom
Multiple R-squared:  0.3117,     Adjusted R-squared:  0.2607
F-statistic: 6.113 on 2 and 27 DF,  p-value: 0.00646
```

Usual steps in applying lm()

anova(mod1)

Test of the ‘explanatory power’ of the model

For reference: it’s also how to compare models

```
anova(mod1)

Analysis of Variance Table

Response: diameter
          Df Sum Sq Mean Sq F value  Pr(>F)
medium     2 10.495  5.2473  6.1129 0.00646 **
Residuals 27 23.177  0.8584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Usual steps in applying lm()

Post hoc - which means differ

Use glht() from package multcomp

```
library(multcomp)
post <- glht(mod1, linfct = mcp(medium = "Tukey"))
summary(post)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = diameter ~ medium, data = culture)

Linear Hypotheses:
                                           Estimate Std. Error t value Pr(>|t|)
with sugar - control == 0                    0.1700     0.4143   0.410  0.91168
with sugar + amino acids - control == 0      1.3310     0.4143   3.212  0.00912 **
with sugar + amino acids - with sugar == 0   1.1610     0.4143   2.802  0.02442 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

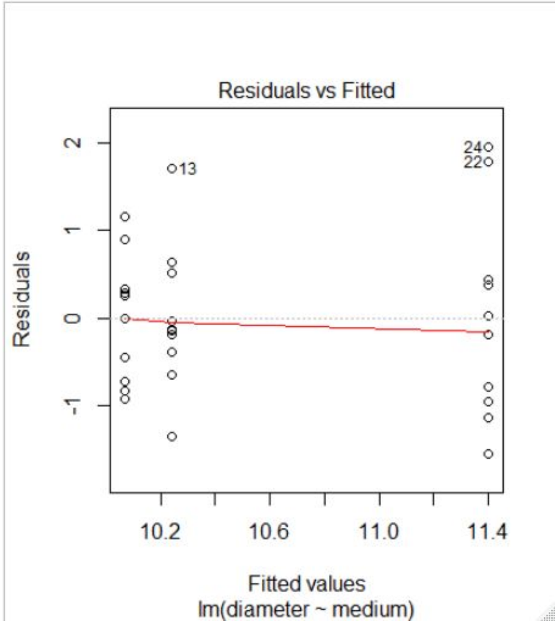
Torsten Hothorn, Frank Bretz and Peter Westfall (2008), Simultaneous Inference in General Parametric Models. *Biometrical Journal*, **50**(3), 346--363

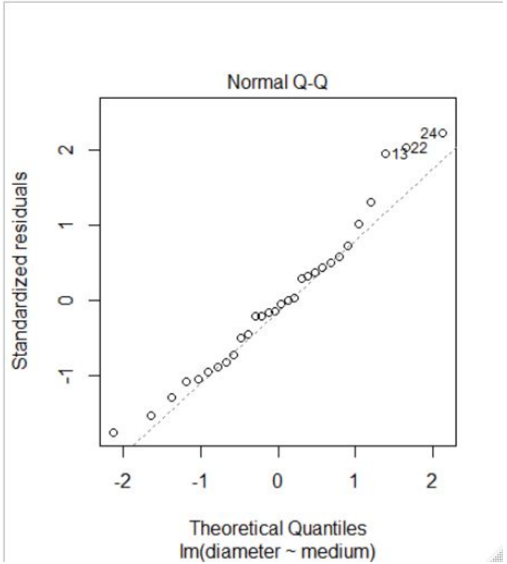
Assumptions - exactly as stage 1

```
shapiro.test(mod1$residuals)

	Shapiro-Wilk normality test

data:  mod1$residuals
W = 0.96423, p-value = 0.3953
```

```
plot(mod1)
```

These look fine





Key points

T-tests, ANOVA and regression are fundamentally the same, collectively called ‘general linear models’. They can be carried out in R with lm()

The concept can be extended to ‘generalised linear models’ for different types of response. Generalised linear models are carried out in R with glm()

The output of lm() looks more complex, at first, than the outputs of t.test() and aov()

The output of glm() is like that for lm(). So we will revisit regression, t-tests and ANOVA using lm() to help you understand the output