ANOVA & ANCOVA

Lecture 5

LSM3257

AY22/23; Sem 2 | Ian Z.W. Chan

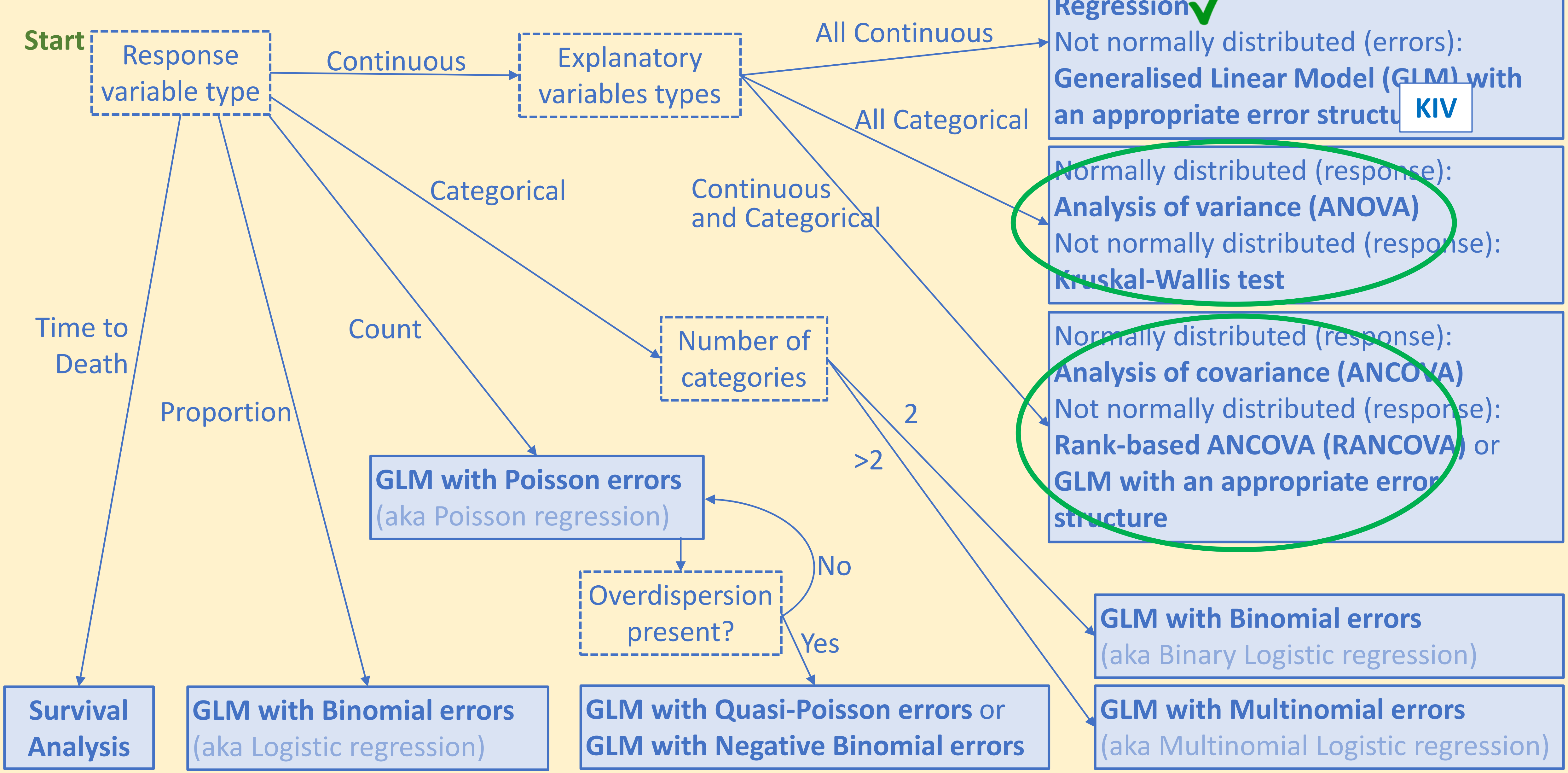
Last week…

Advanced analyses: when to use and Decision tree

Regression

- What is it?: Maximum Likelihood, slope (b), coefficient of determination (r^2)
- Linear Regression: Assumptions, Power analysis, Fit, Check, Predict
- Robust Regression
- Polynomial Regression
- Multiple Linear Regression: Model simplification, Model comparison, Multicollinearity

Advanced analyses – Analysis decision tree (will be modified further)



Start

Response variable type

Continuous

Explanatory variables types

All Continuous

Normally distributed (errors):
Regression ✓
Not normally distributed (errors):
Generalised Linear Model (GLM) with an appropriate error structu

KIV

All Categorical

Normally distributed (response):
Analysis of variance (ANOVA)
Not normally distributed (response):
Kruskal-Wallis test

Continuous and Categorical

Normally distributed (response):
Analysis of covariance (ANCOVA)
Not normally distributed (response):
Rank-based ANCOVA (RANCOVA) or **GLM with an appropriate error structure**

Categorical

Number of categories

2

GLM with Binomial errors
(aka Binary Logistic regression)

>2

GLM with Multinomial errors
(aka Multinomial Logistic regression)

Count

GLM with Poisson errors
(aka Poisson regression)

Overdispersion present?

No

Yes

GLM with Quasi-Poisson errors or **GLM with Negative Binomial errors**

Proportion

GLM with Binomial errors
(aka Logistic regression)

Time to Death

Survival Analysis

Summary (Learning Objectives)

Analysis of variance (ANOVA)

- Assumptions, fitting, checking and interpreting
- Alternatives: Welch's one-way ANOVA, Kruskal-Wallis test
- Repeated measures ANOVA (and Friedman test)
- Factorial vs. Split plot designs

  Factorial experiments: 2-way and 3-way ANOVA

  Nested ANOVA

Analysis of covariance (ANCOVA)

- Assumptions, fitting, checking and interpreting
- Alternative: Rank-based ANCOVA (RANCOVA)
- Factorial experiments: 2-way and 3-way ANCOVA

ANOVA

What is Analysis of Variance (ANOVA)?

Used when your one response variable is continuous and all your **explanatory variables are categorical**.

- T-tests can handle only 2 categories of 1 categorical explanatory variable → ANOVA can handle >2 categories and >1 variable.

Partitions the total variance in the dependent variable (SSY) into variance that can be explained by the different levels in explanatory variable A (SSA), explanatory variable B (SSB), etc., and finally the remaining unexplained variance (SSE):

$$SSY = SSA\ (+\ SSB + SSC \ldots) + SSE$$

Degrees of freedom = the number of datapoints (i.e. number of levels (k)) x replicates per level (n) – number of levels (k) (because we have to estimate the mean for each level). General formula:

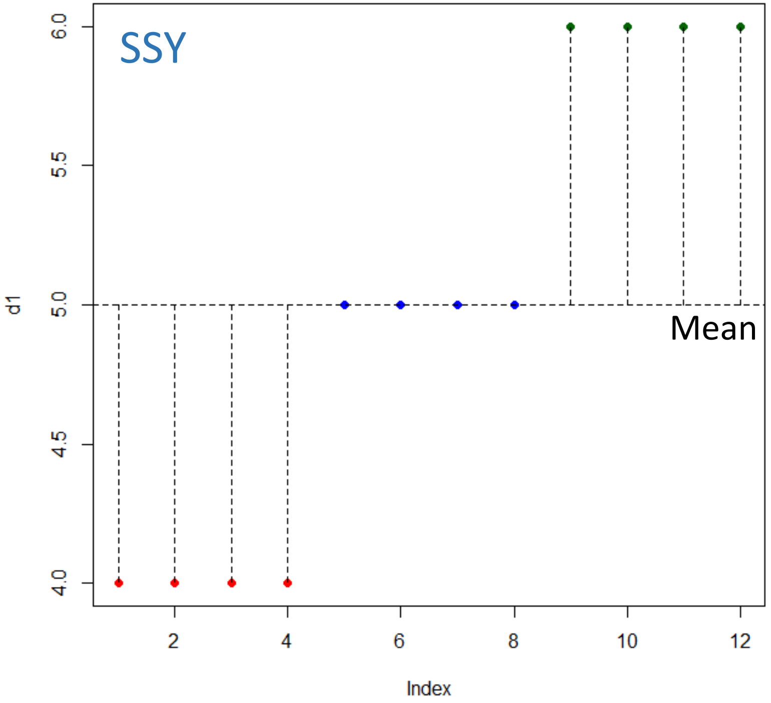
$$Df = kn - k = k(n - 1)$$

- Example: if you have 3 levels and 10 replicates per level: df = 3(10 – 1) = 27.

What is Analysis of Variance (ANOVA)?

Explained graphically…

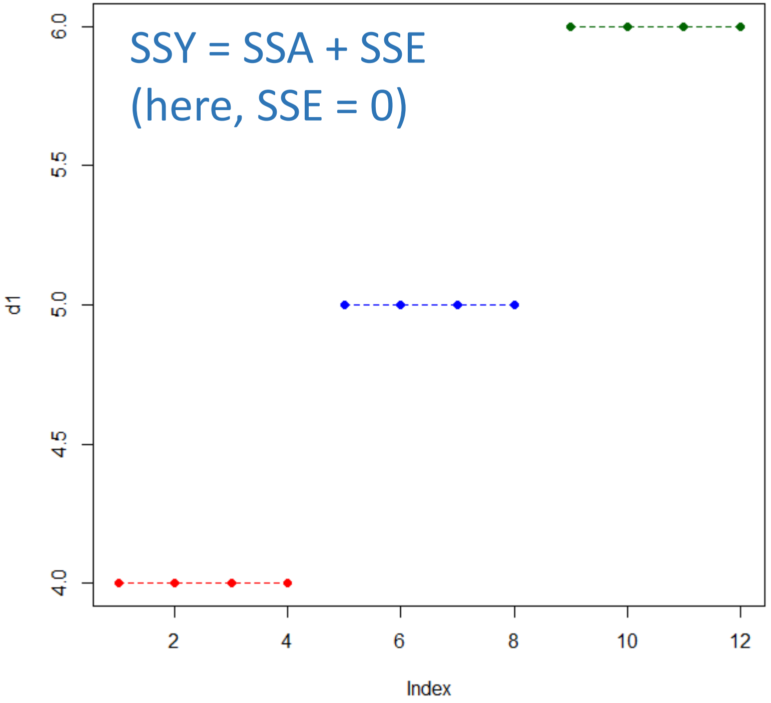
Null Model



SSY

ANOVA

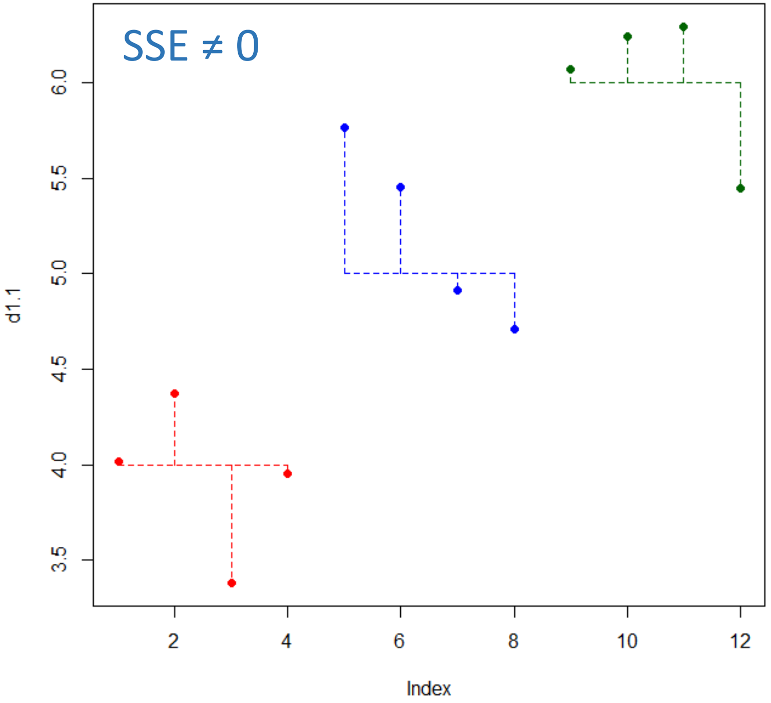
(1 explanatory variable with 3 levels)



SSY = SSA + SSE

(here, SSE = 0)

In reality



SSE ≠ 0

Examples

Continuous response variable	Categorical explanatory variable
Weight	Diet: Normal, Drought (2 levels)
Biodiversity	Pollution: Absent, Present (2 levels)
Carbon storage potential	Habitat: Forest, Grassland, Tundra (3 levels)
Good-looking Index (1 to 1000)	Faculty: Science, Business, Arts (3 levels)

Assumptions

The response variable is normally distributed **within each level**.

- Check by testing the normality of either (i) the datapoints in each level individually (easy to do if you have few groups; or (ii) the residuals of the model. Both test the same thing.

The variances within each level are equal (i.e. homogeneity or equality of variances).

Each datapoint is independent.

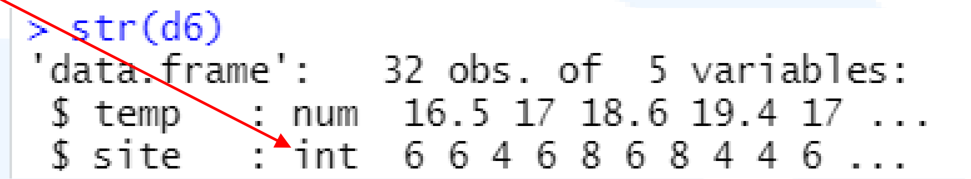
Absence of significant outliers.

Explore your data

#Read in and visualise data

```
d6=read.csv("temperatureData.csv")
str(d6)
```

If we run an lm(), R will do a Regression instead of an ANOVA



```
> str(d6)
'data.frame':   32 obs. of  5 variables:
 $ temp    : num  16.5 17 18.6 19.4 17 ...
 $ site    : int  6 6 4 6 8 6 8 4 4 6 ...
```

#Convert <site> into a factor (because it is a categorical variable)

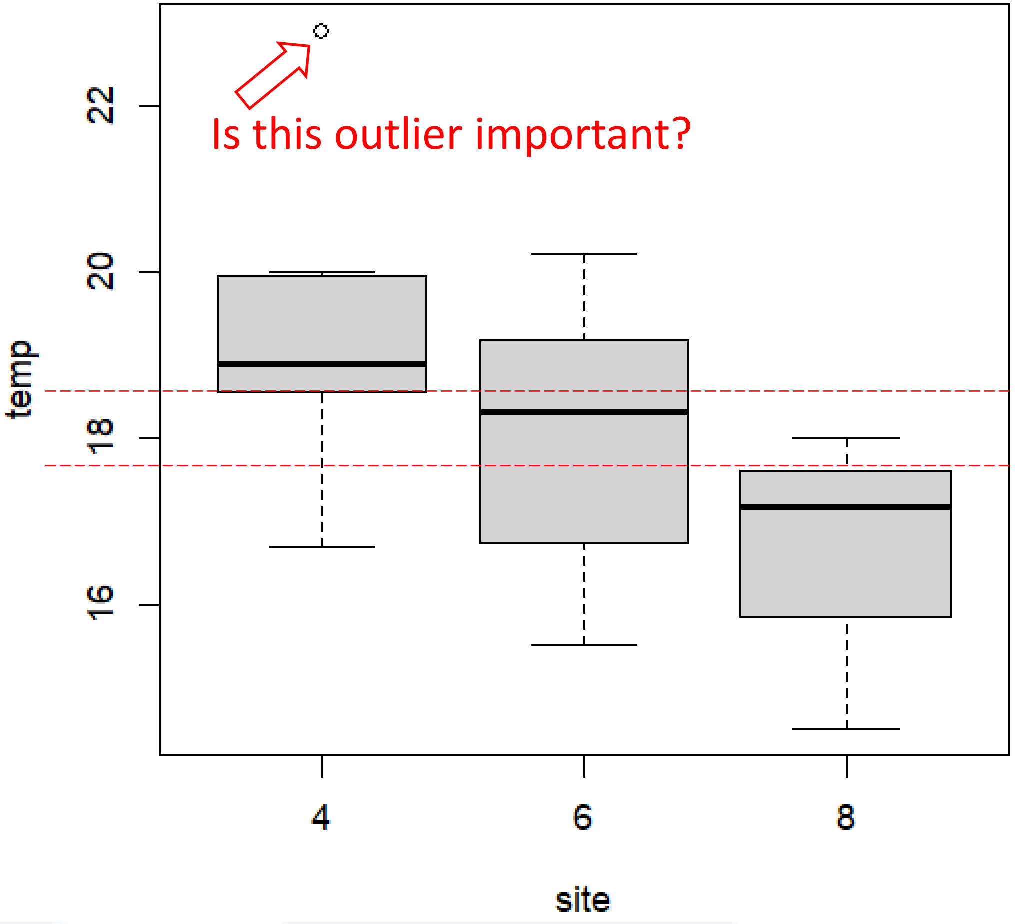
```
d6$site=as.factor(d6$site)
levels(d6$site)
```

```
> levels(d6$site)
[1] "4" "6" "8"
```

No overlap between the boxes of levels 4 and 8, so there is likely a significant difference. We need to do the ANOVA to confirm this.

#Visualise how <temp> varies with <site>

```
boxplot(temp~site,data=d6)
```



Is this outlier important?

Fitting the ANOVA

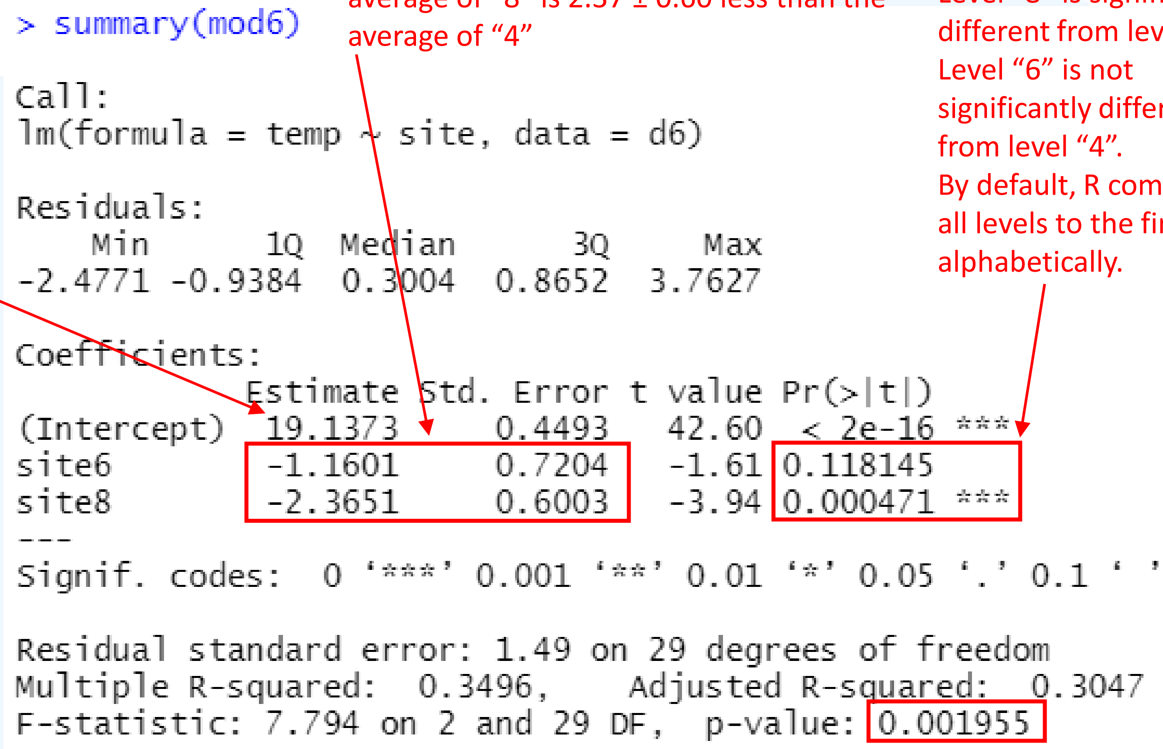
#Using either aov() or lm()

```
mod6=lm(temp~site,data=d6)
summary(mod6)
summary.aov(mod6)
```

Looks like there’s a significant effect!

Note 1: if you run aov(), then summary() will give you the summary.aov() output instead. To get the output you see above, you will need to call summary.lm().

Note 2: These two commands show you the same results, formatted to show different things. The “summary.lm” results are better for looking at effect sizes at different levels. The “summary.aov” results are better for looking at how the variance is partitioned.

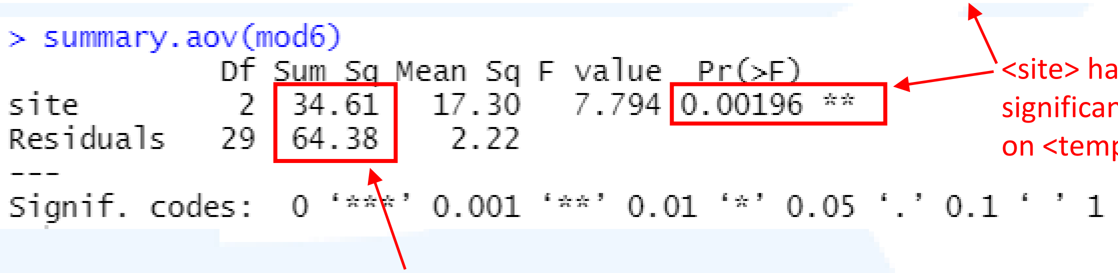


```
> summary(mod6)

Call:
lm(formula = temp ~ site, data = d6)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4771 -0.9384  0.3004  0.8652  3.7627 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  19.1373     0.4493   42.60  < 2e-16 ***
site6        -1.1601     0.7204   -1.61 0.118145    
site8        -2.3651     0.6003   -3.94 0.000471 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.49 on 29 degrees of freedom
Multiple R-squared:  0.3496,    Adjusted R-squared:  0.3047 
F-statistic: 7.794 on 2 and 29 DF,  p-value: 0.001955
```

Average <temp> of <site> level “4”

Effect sizes: the average of “6” is 1.16 ± 0.72 less than the average of “4”; the average of “8” is 2.37 ± 0.60 less than the average of “4”

Level “8” is significantly different from level “4”. Level “6” is not significantly different from level “4”. By default, R compares all levels to the first level alphabetically.



```
> summary.aov(mod6)
            Df Sum Sq Mean Sq F value  Pr(>F)   
site         2  34.61   17.30   7.794 0.00196 **
Residuals   29  64.38    2.22                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

<site> has a significant effect on <temp>

Variation explained by <site> and remaining unexplained. Note that Residuals is greater than <site> (that’s why $r^2 < 0.5$), but it’s OK! Our model still helped us to understand the data better.

Comparing between groups

Commonly confused for each other…

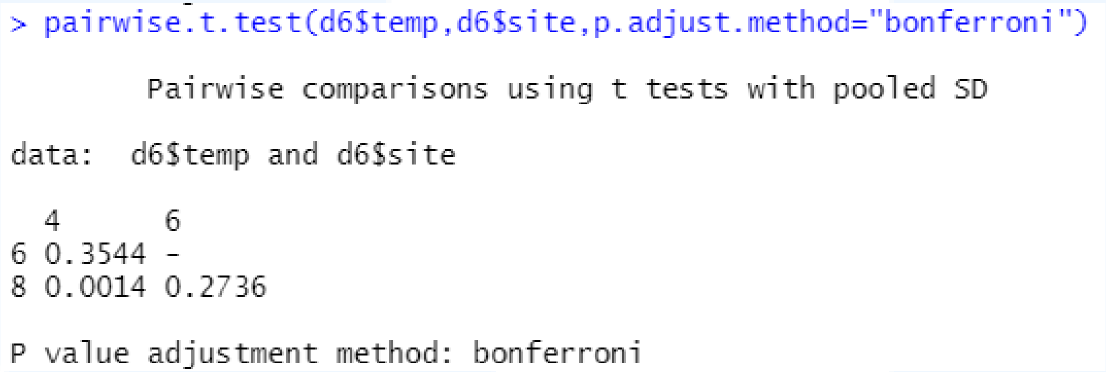
Pairwise t-tests ≠ Paired t-tests

- Pairwise t-tests compare all levels within a categorical variable.
- Paired t-tests compare two related groups/datasets.

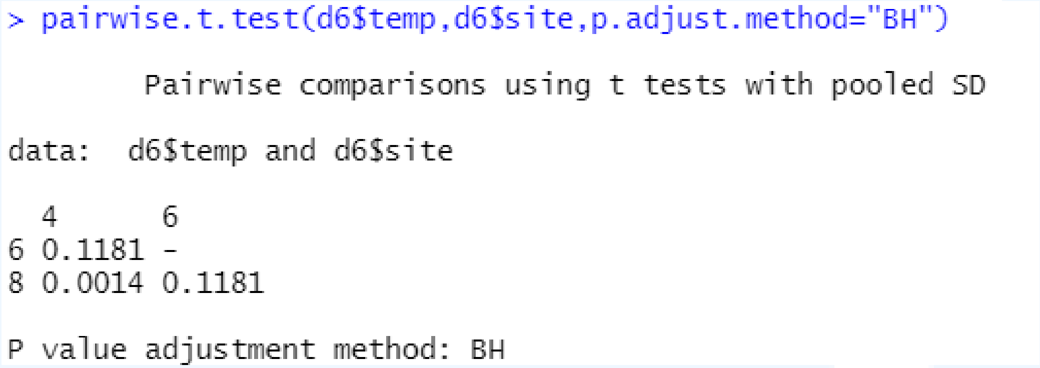
Do t-tests between all possible pairs of levels

- Remember if we are doing multiple comparisons, we need to correct our p-values. Which of these two corrections is more strict?

```
pairwise.t.test(d6$temp,d6$site,p.adjust.method="BH")
```



```
> pairwise.t.test(d6$temp,d6$site,p.adjust.method="bonferroni")

	Pairwise comparisons using t tests with pooled SD

data:  d6$temp and d6$site

  4      6
6 0.3544 -
8 0.0014 0.2736

P value adjustment method: bonferroni
```



```
> pairwise.t.test(d6$temp,d6$site,p.adjust.method="BH")

	Pairwise comparisons using t tests with pooled SD

data:  d6$temp and d6$site

  4      6
6 0.1181 -
8 0.0014 0.1181

P value adjustment method: BH
```

Alternative: we can use `relevel()` to change the reference level and re-run the ANOVA. (Very tedious if you have many levels.)

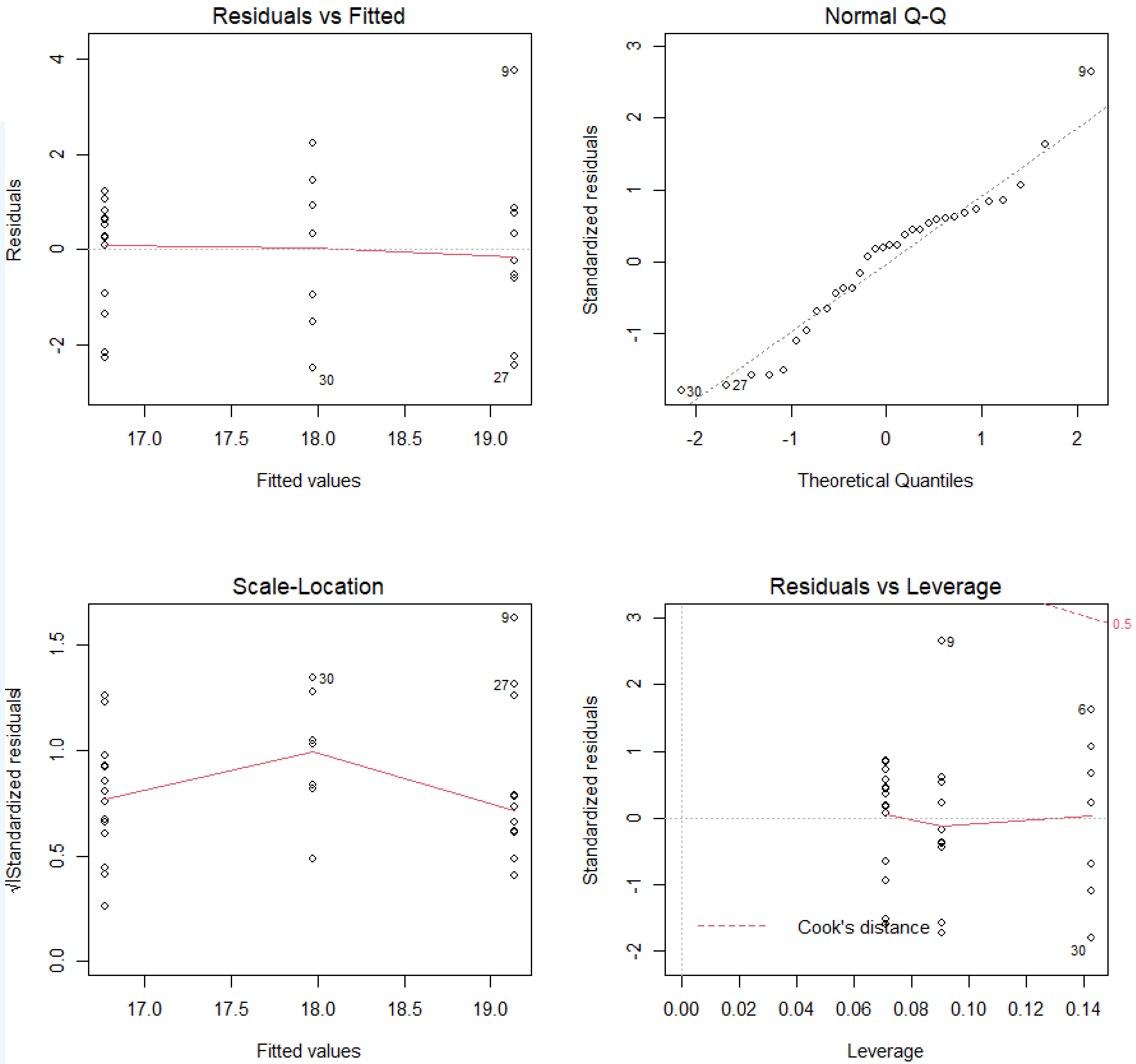
Interpretation: “Average temperatures at site 8 are significantly lower than at site 4 by 2.37 ± 0.6 degrees Celcius (mean ± SE; *P* < 0.001). There is no significant difference between site 4 and site 6, and between site 6 and site 8.”

Checking assumptions

```
#Plot diagnostic plots
par(mfrow=c(2,2))
plot(mod6)

#Test normality of residuals
shapiro.test(resid(mod6))

#Test equality of variance
install.packages("lawstat")
library(lawstat)
levene.test(d6$temp,d6$site)
#p-value > 0.05 is good
```

Response variable

Explanatory (grouping) variable



Everything looks OK-ish! Buy 4D today!!

If the normality assumption is violated…

Option 1: transform response variable (e.g. log()), then run the ANOVA again and hope for the best.

Option 2: Kruskal-Wallis test (can only handle 1 explanatory variable).

#Global test for all levels

```
mod2k=kruskal.test(temp~site,data=d6)
mod2k
```

For kruskal.test() objects, to see the results you have to call the object instead of using summary(). It's just the way the author wrote it. Alternatively, you can just don't save it to an object.

#Pairwise comparison between all pairs of levels

```
pairwise.wilcox.test(d6$temp,d6$site,p.adjust.method="BH")
```

Option 3: Use a GLM with a different error distribution (covered in later lectures).

If equality of variance is violated…

Option 1: Transform response variable.

Option 2: Welch’s one-way ANOVA.

#Global test for all levels

```
install.packages("rstatix")
library(rstatix)
welch_anova_test(temp~site,data=d6)
```

#Here I didn’t save it to an object so the results are displayed automatically.

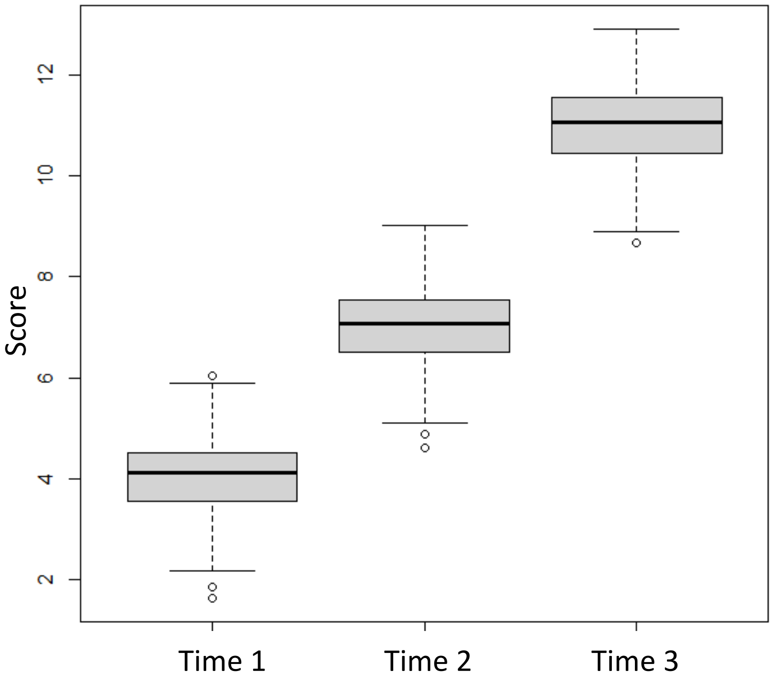
#Pairwise comparison between all pairs of levels

```
pairwise.wilcox.test(d6$temp,d6$site,p.adjust.method="BH")
```

Option 3: Use a GLS or GLM (covered in later lectures).

Repeated measures (i.e. “paired”) experiments: Repeated measures ANOVA

Example: you measure the same 25 participants at three time points. The datapoints are not independent, so we cannot do a normal ANOVA.



We need to do a Repeated Measures ANOVA.

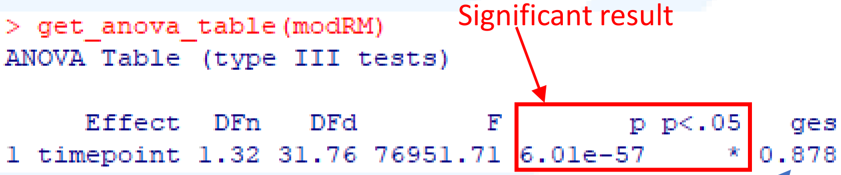
```
#Install the rstatix package
#Load the dataset
d3=read.table("scoreTimes.txt",header=T)
#Run the repeated measures ANOVA
modRM=anova_test(data=d3,dv=score,wid=subject,
within=timepoint)
get_anova_table(modRM)
```

Response variable

Dataset

Variable identifying the “paired repeats”

Explanatory variable, to run a 2-way (or more) repeated measures, input the explanatory variables like this: “within=c(variable1,variable2,variable3)”.



```
> get_anova_table(modRM)
ANOVA Table (type III tests)

     Effect  DFn   DFd        F        p p<.05   ges
1 timepoint 1.32 31.76 76951.71 6.01e-57     * 0.878
```

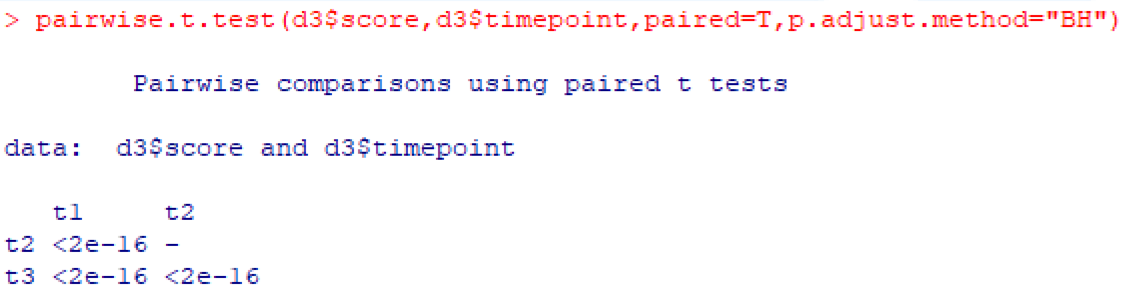
Significant result

This is the result of the Mauchly test of sphericity for the Sphericity assumption (i.e. the variances of the difference between groups should be equal). If they are equal, ges will be close to 1. If ges < 0.75, this test will automatically apply the Greenhouse-Geisser sphericity correction, so you don’t need to worry about it. Just report it.

Repeated measures (i.e. “paired”) experiments: Friedman test

#Do pairwise comparisons using paired t-tests

```
pairwise.t.test(d3$score,d3$timepoint,paired=T,p.adjust.method="BH")
```



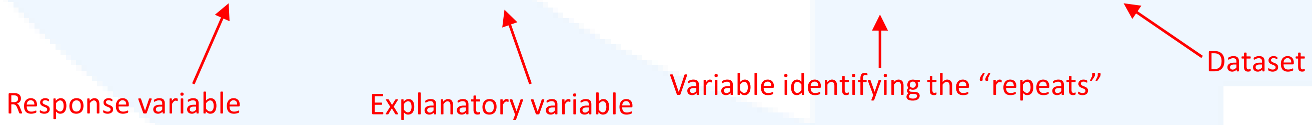
```
> pairwise.t.test(d3$score,d3$timepoint,paired=T,p.adjust.method="BH")

        Pairwise comparisons using paired t tests

data:  d3$score and d3$timepoint

   t1     t2
t2 <2e-16 -
t3 <2e-16 <2e-16
```

What if the normality assumption is violated?

Option 1: transform the y-variable.

Option 2: Use a Friedman test (only available for 1-way)

```
friedman.test(score~timepoint|subject,data=d3)
```

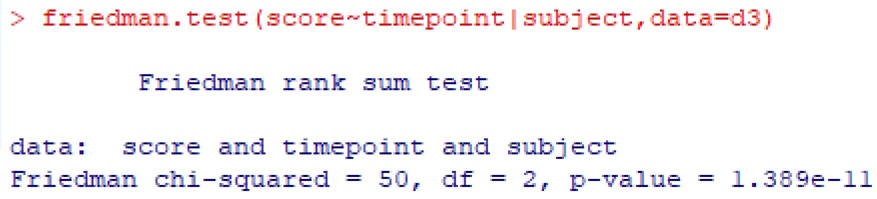


Response variable

Explanatory variable

Variable identifying the “repeats”

Dataset



```
> friedman.test(score~timepoint|subject,data=d3)

        Friedman rank sum test

data:  score and timepoint and subject
Friedman chi-squared = 50, df = 2, p-value = 1.389e-11
```

Option 3: Use a GLM.

Factorial experiments

2 or more categorical explanatory variables, each with 2 or more levels

Continuous response variable	Categorical explanatory variables
Weight	Diet: Normal, Drought (2 levels)
	Sex: Male, Female (2 levels)
	(4 unique level combinations)
Biodiversity	Country: Singapore, Malaysia, Indonesia (3 levels)
	Pollution: Low, Medium, High (3 levels)
	Treatment: A, B, C, D (4 levels)
	(36 unique level combinations)

Factorial experiments vs. Split Plot experiments

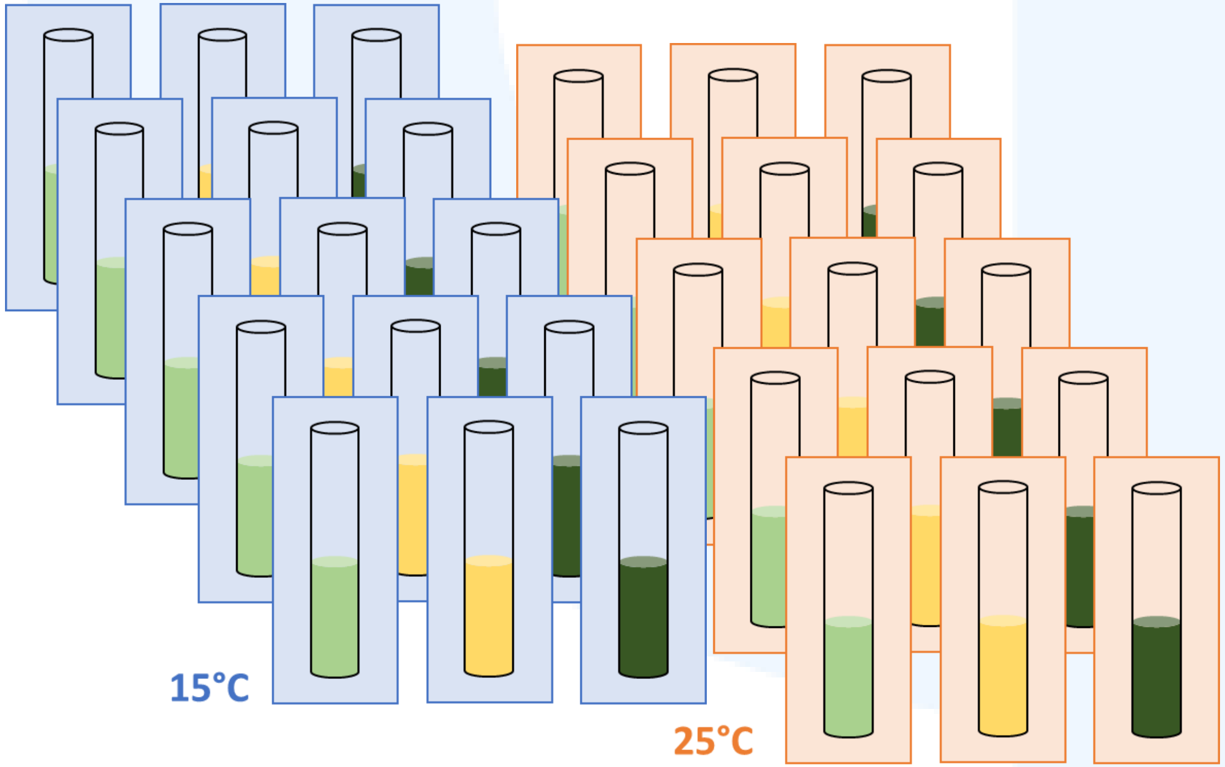
Imagine we have 2 categorical explanatory variables

temperature (xvar1): 2 levels (15°C, 25°C)

growthMedium (xvar2): 3 levels (dark green, yellow, light green)

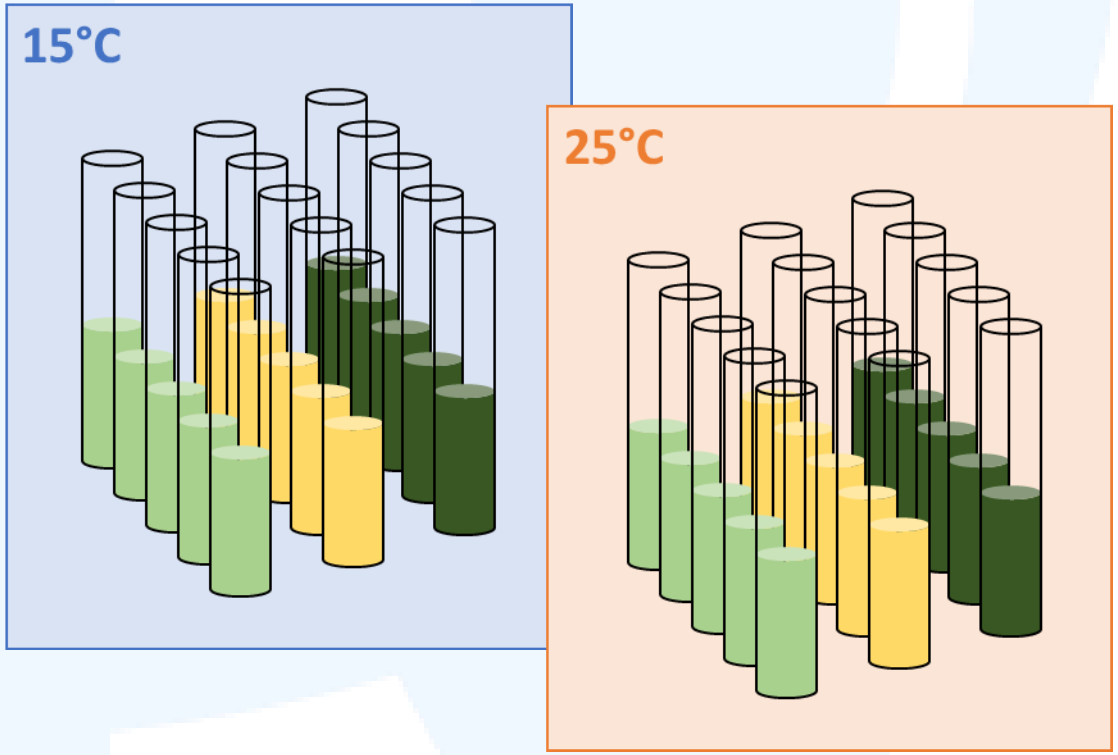
Factorial

15 diff rooms for each treatment,
the temperature level is independent



Split Plot (aka Nested)

1 big room, then temp
level is not independent
its nested now



Factorial experiments vs. Split Plot experiments

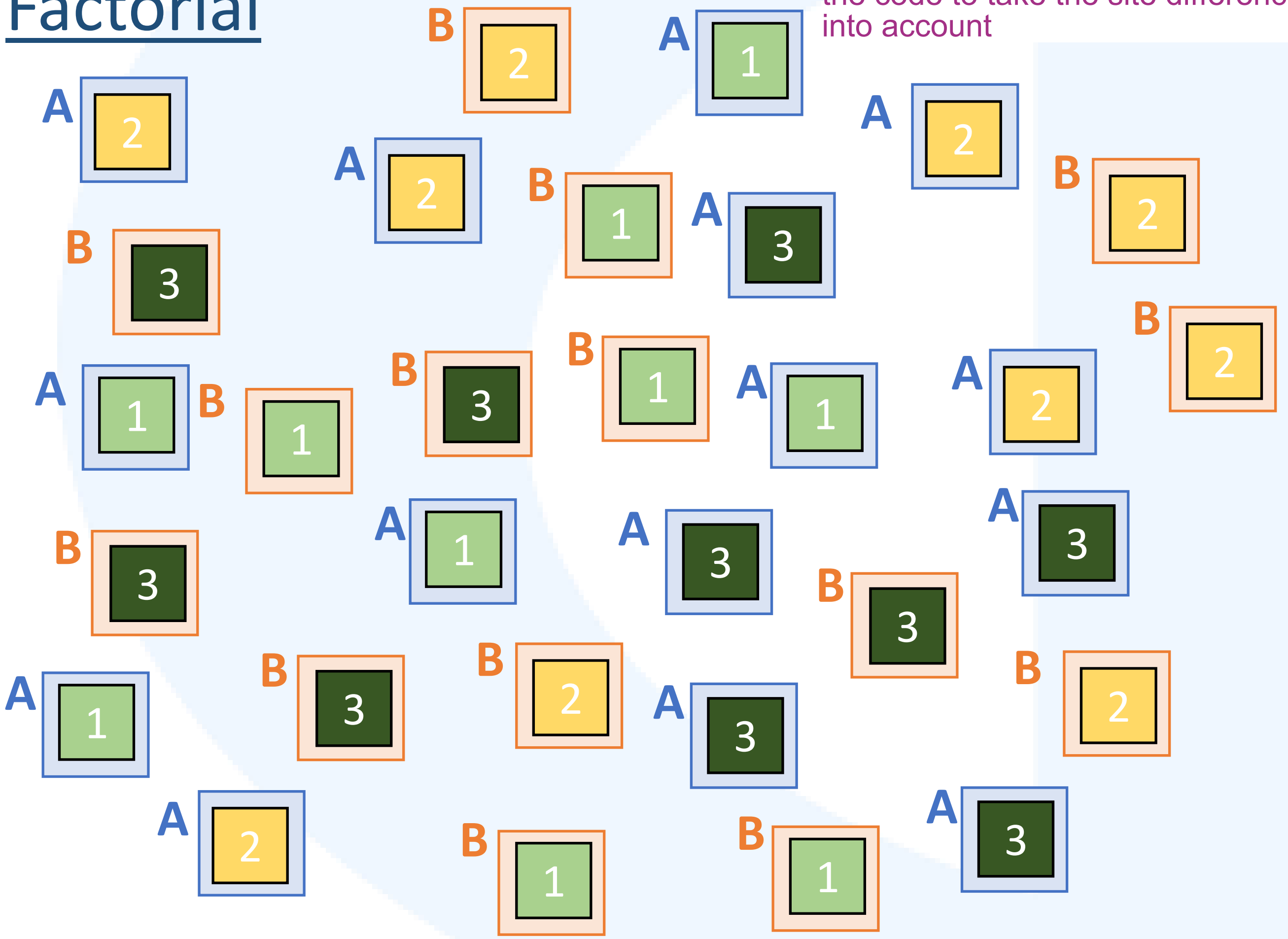
we may have multiple sites not because sites are important but because we want to reduce the bias. In this case, we may not want to code it as nested event

Nutrient level: 🟩1 🟨2 ■3

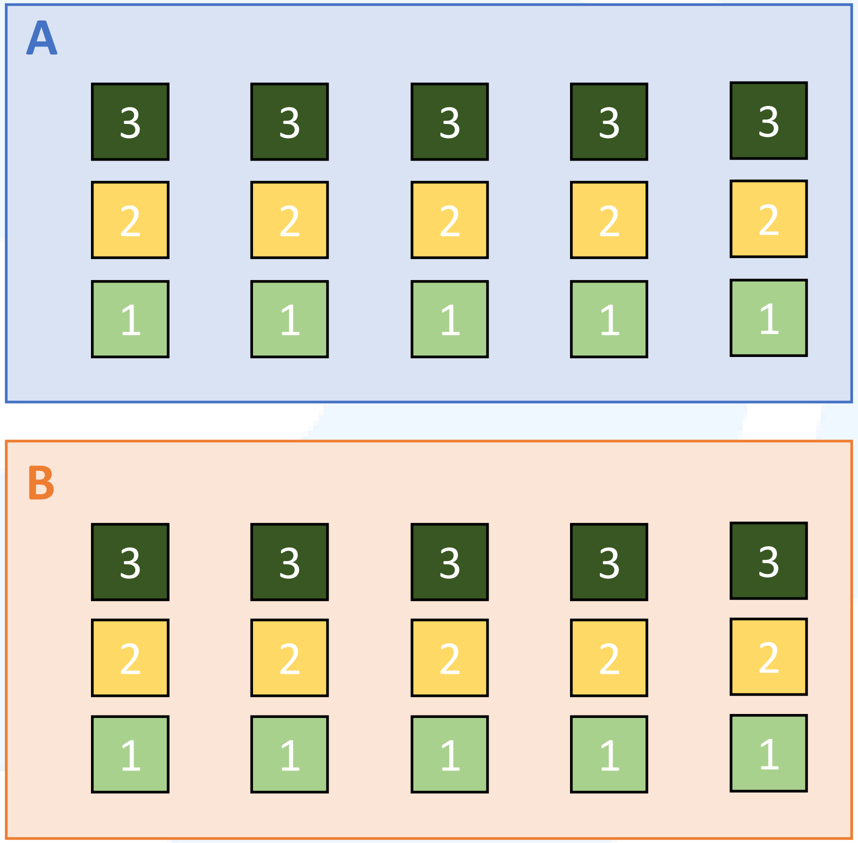
Temperature: A ▢ B ▢

Factorial

because we dont want the code to take the site difference into account



Split Plot (aka Nested)



Coded in R as:

y~x1+x2 (or x1*x2)

everything is randomized and independent

Coded in R as: x1 is going to be temperature level

y~x1/x2

x2 is the smaller variale, nested within in this case, nutrient level

Factorial experiments: 2-way ANOVA

We want to use <site> and <treated> to explain <temp>

- 2 categorical explanatory variables with a factorial design

- We use a 2-way ANOVA

#Visualise: quick (and ugly) plot

```
interaction.plot(d6$site,d6$treated,d6$temp)
```

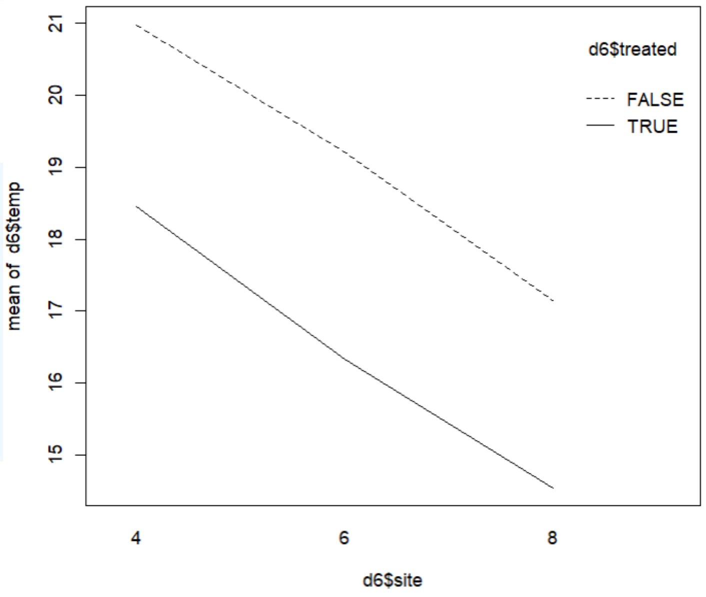
Main x-axis variable

Secondary explanatory variable

Y-axis variable

Doesn't look like there's an interaction between cyl and am. If there is an interaction, the two lines would cross (or at least not be parallel).

Interaction plot (Base R)



#Visualise: ggplot2

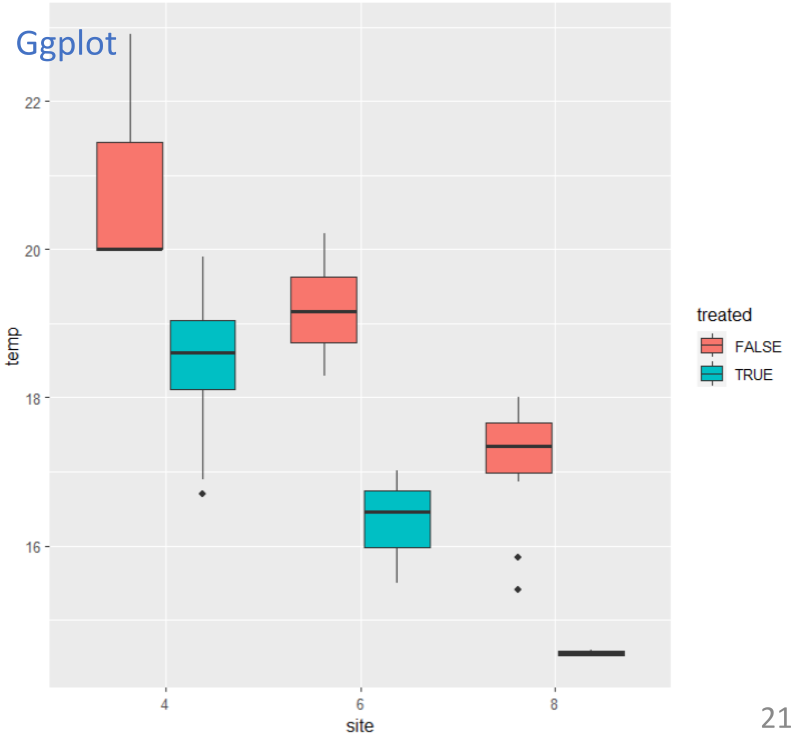
```
ggplot(d2,aes(x= Ans , y= Ans , fill= Ans ))+
geom_ Ans ()
```

Main x-axis variable

Y-axis variable

Secondary explanatory variable

Ggplot



Factorial experiments: 2-way ANOVA

#Run the 2-way ANOVA

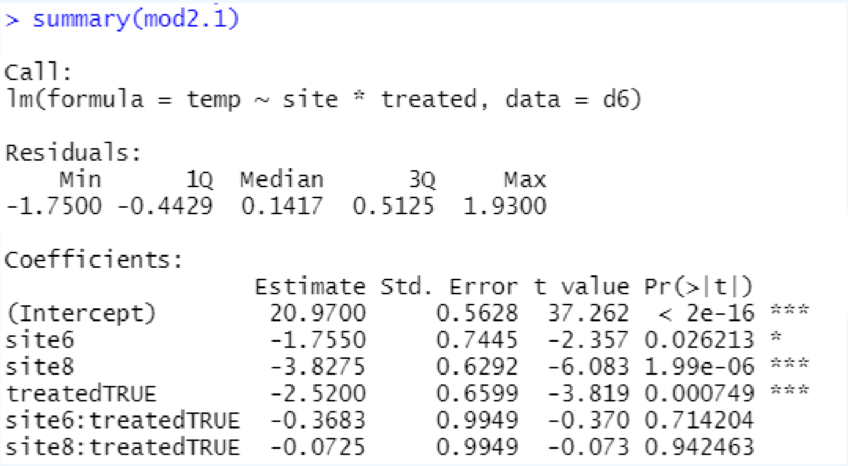
```
mod2.1=lm(temp~site*treated,data=d6)
summary(mod2.1)
```

#Start with the interaction: no interaction is significant so we can remove it

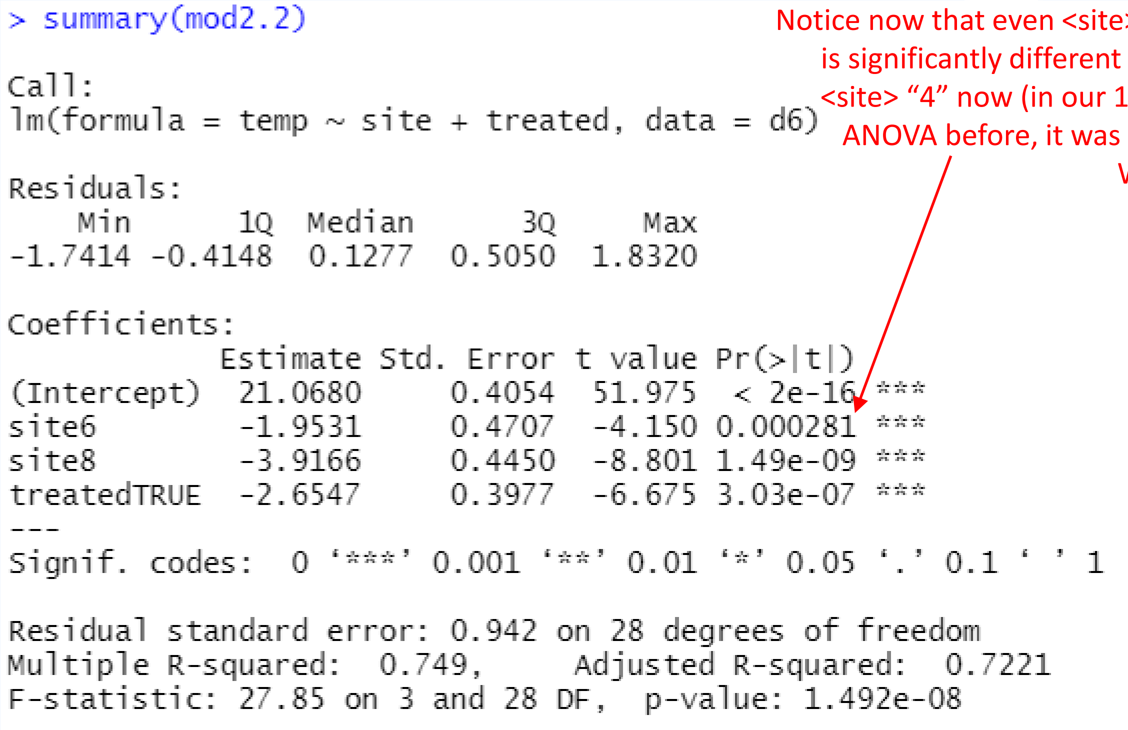
#Simplify the model

```
mod2.2=update(mod2.1,~.-site:treated)
summary(mod2.2)
```

#All significant: looks like we have our minimum adequate model

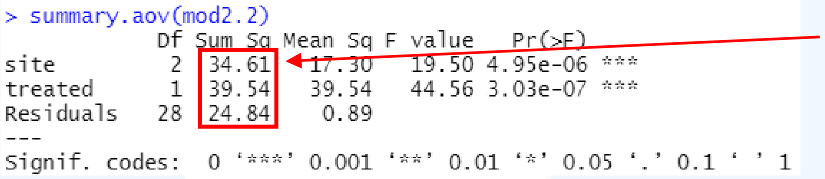


```
> summary(mod2.1)

Call:
lm(formula = temp ~ site * treated, data = d6)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7500 -0.4429  0.1417  0.5125  1.9300 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       20.9700     0.5628  37.262  < 2e-16 ***
site6             -1.7550     0.7445  -2.357 0.026213 *  
site8             -3.8275     0.6292  -6.083 1.99e-06 ***
treatedTRUE       -2.5200     0.6599  -3.819 0.000749 ***
site6:treatedTRUE -0.3683     0.9949  -0.370 0.714204    
site8:treatedTRUE -0.0725     0.9949  -0.073 0.942463    
```



```
> summary(mod2.2)

Call:
lm(formula = temp ~ site + treated, data = d6)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7414 -0.4148  0.1277  0.5050  1.8320 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  21.0680     0.4054  51.975  < 2e-16 ***
site6        -1.9531     0.4707  -4.150 0.000281 ***
site8        -3.9166     0.4450  -8.801 1.49e-09 ***
treatedTRUE  -2.6547     0.3977  -6.675 3.03e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.942 on 28 degrees of freedom
Multiple R-squared:  0.749,	Adjusted R-squared:  0.7221 
F-statistic: 27.85 on 3 and 28 DF,  p-value: 1.492e-08
```

Notice now that even <site> "6" is significantly different from <site> "4" now (in our 1-way ANOVA before, it was not). Why?



```
> summary.aov(mod2.2)
            Df Sum Sq Mean Sq F value   Pr(>F)    
site         2  34.61   17.30   19.50 4.95e-06 ***
treated      1  39.54   39.54   44.56 3.03e-07 ***
Residuals   28  24.84    0.89                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our model explains more variation than our residuals now. This is great!

Factorial experiments: 2-way ANOVA

#Always check assumptions!

```
par(mfrow=c(2,2))
plot(mod2.2)
```

#Test for normality

```
shapiro.test(resid(mod2.2)) #p=0.32
```

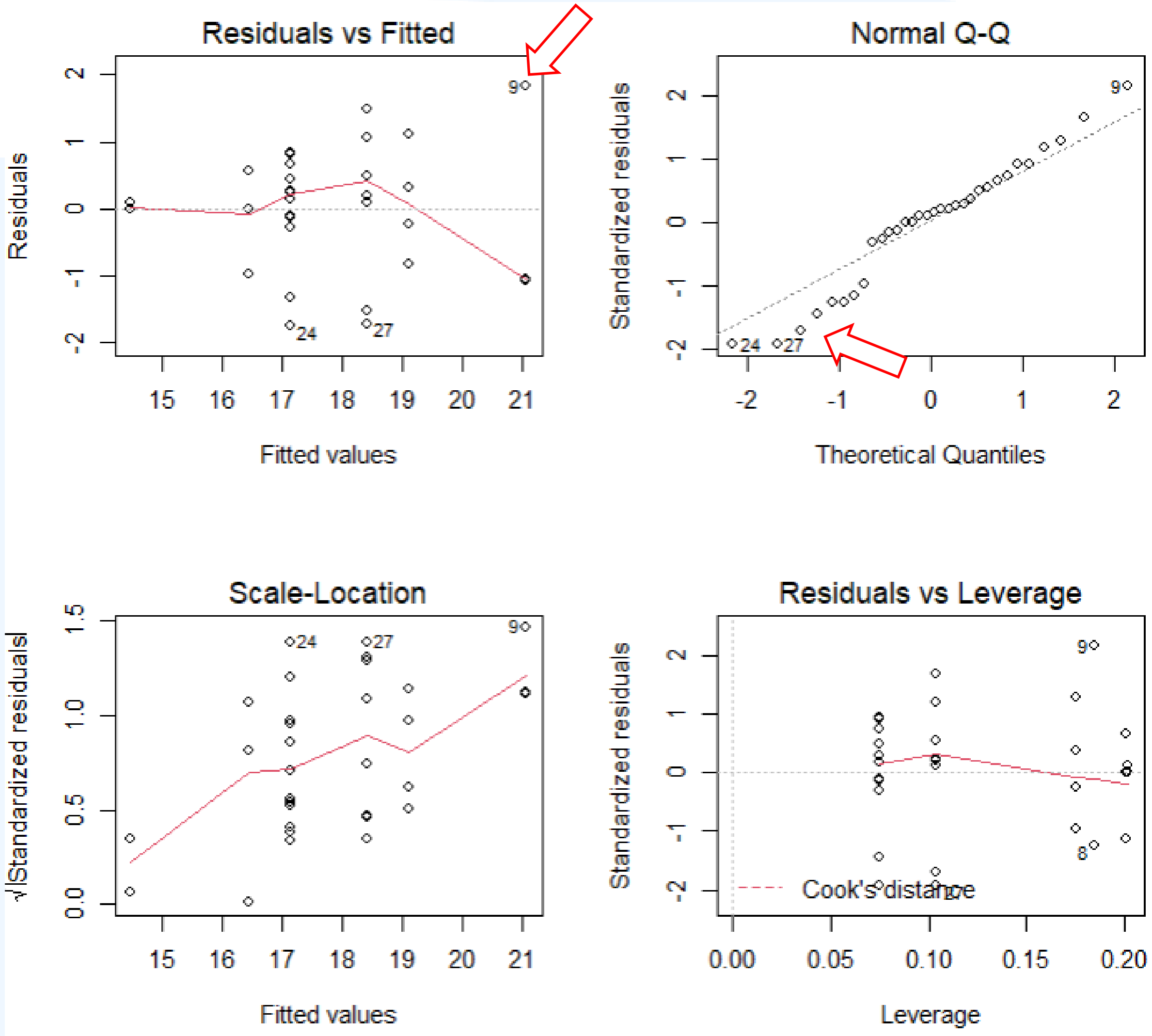
#Test for homogeneity of variance

```
levene.test(d2$temp,gp2.2) #p=0.81
```

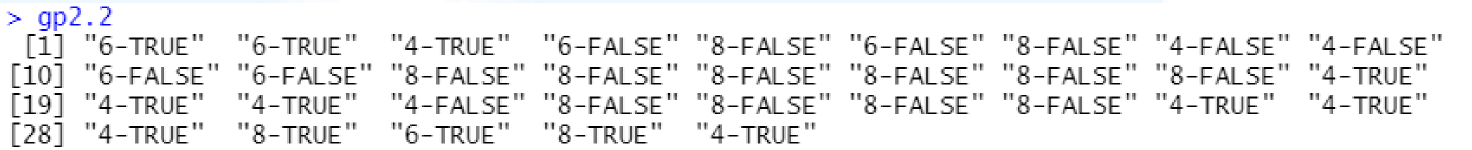
#Note: gp2.2 on next slide

#Everything seems OK

Note: for 2-way (and more) ANOVA, if any of the assumptions are violated, we cannot use Kruskal-Wallis and Welch's ANOVA. We will need to use a GLM
(later lectures).



Factorial experiments: 2-way ANOVA



```
> gp2.2
 [1] "6-TRUE"  "6-TRUE"  "4-TRUE"  "6-FALSE" "8-FALSE" "6-FALSE" "8-FALSE" "4-FALSE" "4-FALSE"
[10] "6-FALSE" "6-FALSE" "8-FALSE" "8-FALSE" "8-FALSE" "8-FALSE" "8-FALSE" "8-FALSE" "4-TRUE"
[19] "4-TRUE"  "4-TRUE"  "4-FALSE" "8-FALSE" "8-FALSE" "8-FALSE" "8-FALSE" "4-TRUE"  "4-TRUE"
[28] "4-TRUE"  "8-TRUE"  "6-TRUE"  "8-TRUE"  "4-TRUE"
```

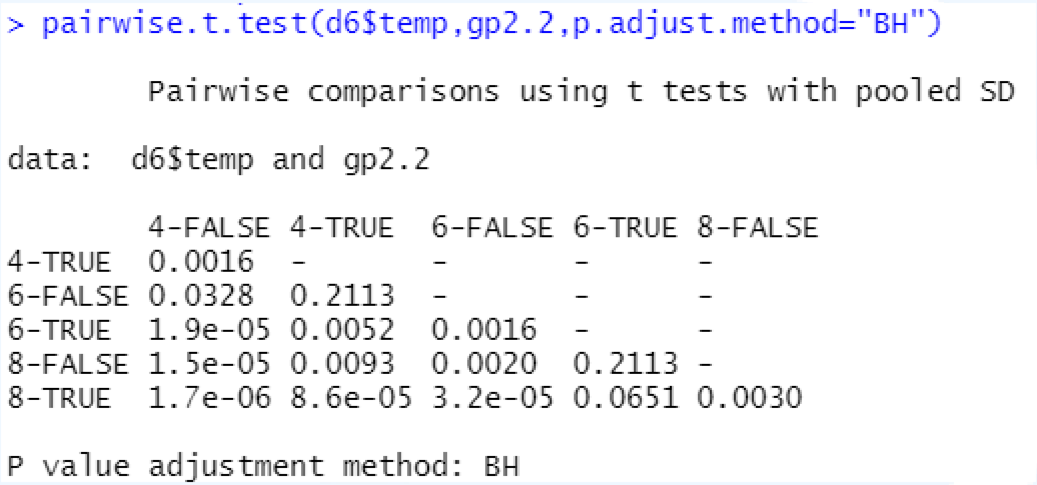
#Pairwise comparisons

```
gp2.2=paste(d6$site,d6$treated,sep="-")
```

This creates a new grouping variable and stores it in “gp2.2”. I am creating this so that I can use it to compare each unique level combination

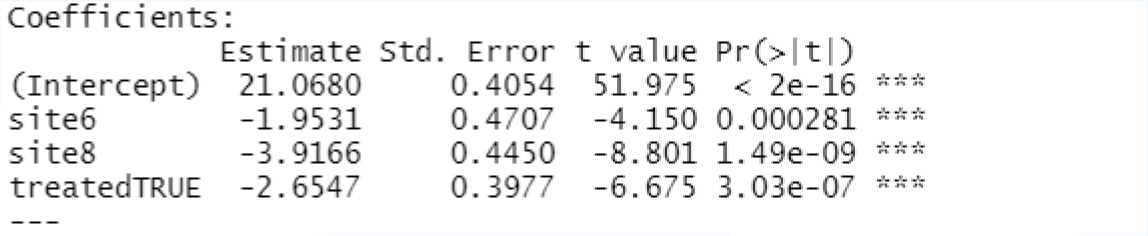
For each observation, I join the value of <site> (e.g. “6”) to the value of <treated> (e.g. “TRUE”), separated by a hyphen (becomes “6-TRUE”)

```
pairwise.t.test(d6$temp,gp2.2,
p.adjust.method="BH")
```



```
> pairwise.t.test(d6$temp,gp2.2,p.adjust.method="BH")

        Pairwise comparisons using t tests with pooled SD

data:  d6$temp and gp2.2

        4-FALSE 4-TRUE  6-FALSE 6-TRUE 8-FALSE
4-TRUE  0.0016  -       -       -      -
6-FALSE 0.0328  0.2113  -       -      -
6-TRUE  1.9e-05 0.0052  0.0016  -      -
8-FALSE 1.5e-05 0.0093  0.0020  0.2113 -
8-TRUE  1.7e-06 8.6e-05 3.2e-05 0.0651 0.0030

P value adjustment method: BH
```

Interpretation: “Site has a significant effect on temperature. Site 6 is 1.95 ± 0.47 (mean ± SE) degrees cooler than Site 4 (*P* < 0.001), and Site 8 is 3.92 ± 0.45 degrees cooler than Site 4 (*P* < 0.001). In addition, sites which are treated are 2.65 ± 0.40 degrees cooler than sites that are not treated (*P* < 0.001).”



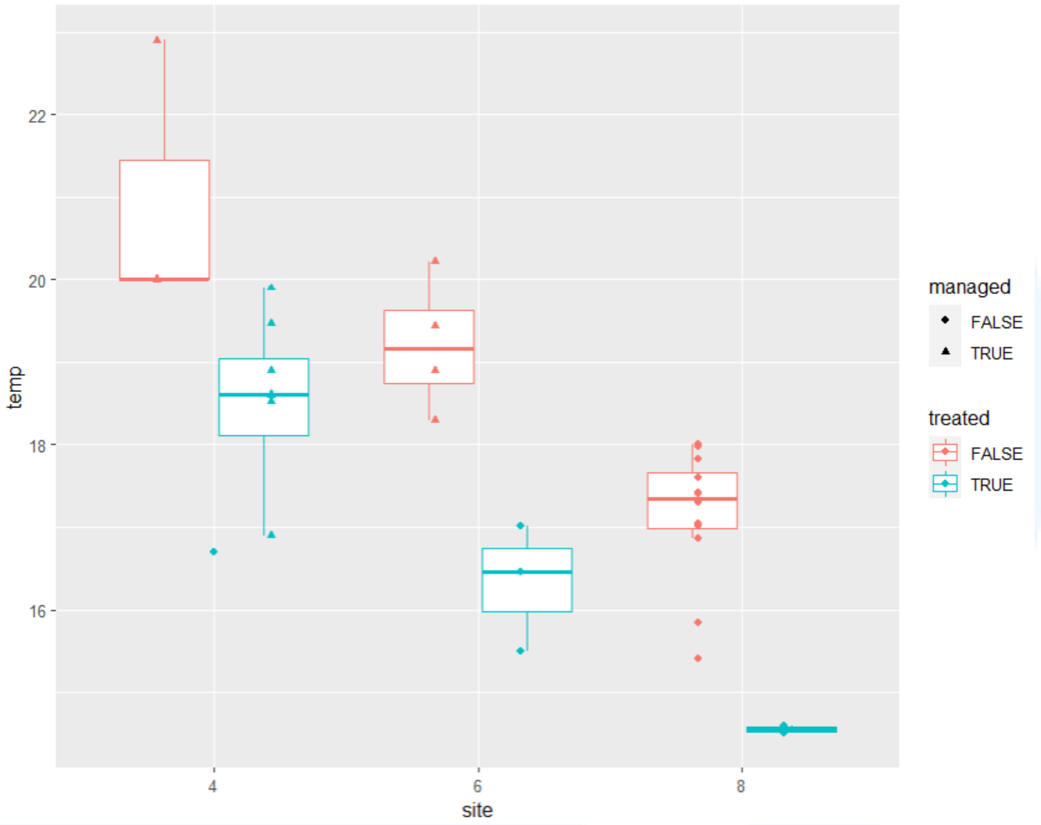
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  21.0680     0.4054  51.975  < 2e-16 ***
site6        -1.9531     0.4707  -4.150 0.000281 ***
site8        -3.9166     0.4450  -8.801 1.49e-09 ***
treatedTRUE  -2.6547     0.3977  -6.675 3.03e-07 ***
---
```

Factorial experiments: 3-way (and more-way) ANOVAs

#To do a 3-way (or more) ANOVA

#Plotting using ggplot

```
ggplot(d6,aes(x=site,y=temp,col=treated))
+geom_boxplot(outlier.shape=NA)+
geom_point(aes(col=treated,shape=managed),
position=position_dodge(width=0.65))
```



Just keep adding variables to the formula using * or +.
But try not to have more than 3 interacting variables, it gets too complicated to explain intuitively (even interpreting the graph, which is supposed to make things easier, becomes difficult!).

#Fitting the model

```
mod2.3=lm(temp~site*treated*managed,data=d6)
```

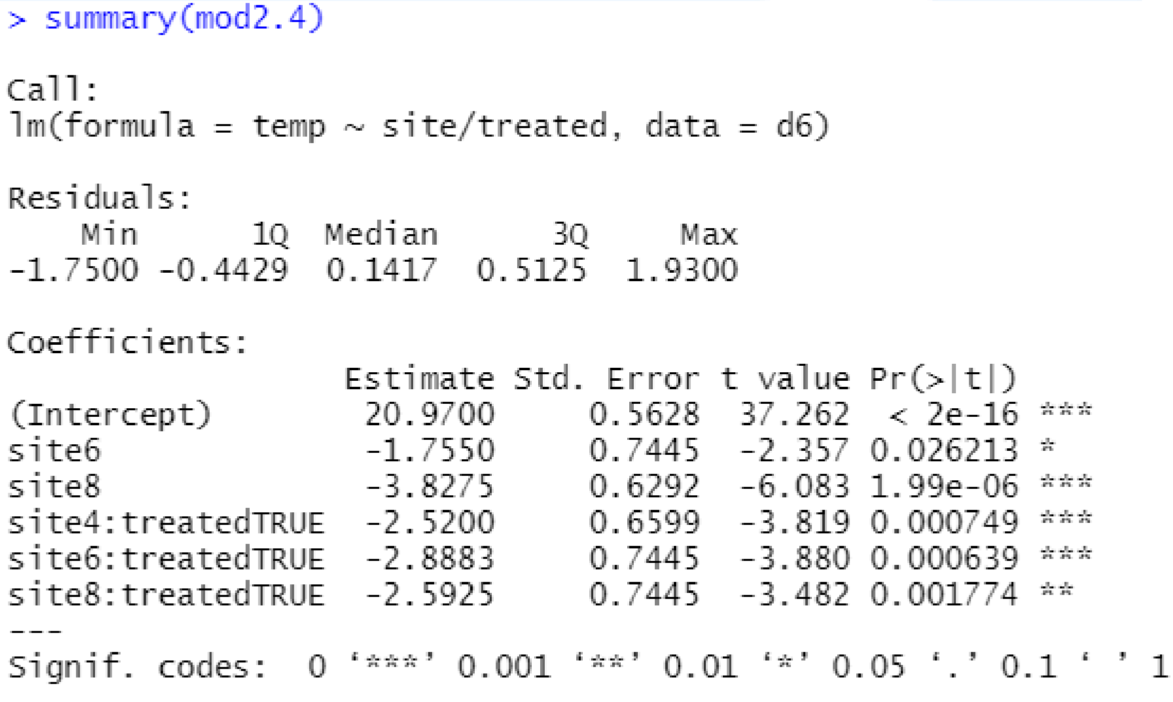
#Followed by simplification, checking and pairwise testing (if required)

Split Plot experiments: Nested ANOVA

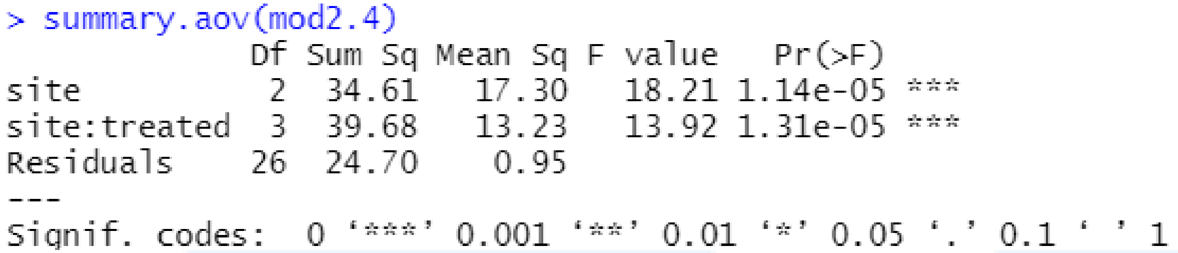
note that there is no result for treated alone. treated is only nested within the site, we can only analyze the treated within each site now, cannot combine and see.

#Assuming <treated> is nested inside <site>

```
mod2.4=lm(temp~site/treated,data=d6)
```



```
> summary(mod2.4)

Call:
lm(formula = temp ~ site/treated, data = d6)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7500 -0.4429  0.1417  0.5125  1.9300 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       20.9700     0.5628  37.262  < 2e-16 ***
site6             -1.7550     0.7445  -2.357 0.026213 *  
site8             -3.8275     0.6292  -6.083 1.99e-06 ***
site4:treatedTRUE -2.5200     0.6599  -3.819 0.000749 ***
site6:treatedTRUE -2.8883     0.7445  -3.880 0.000639 ***
site8:treatedTRUE -2.5925     0.7445  -3.482 0.001774 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

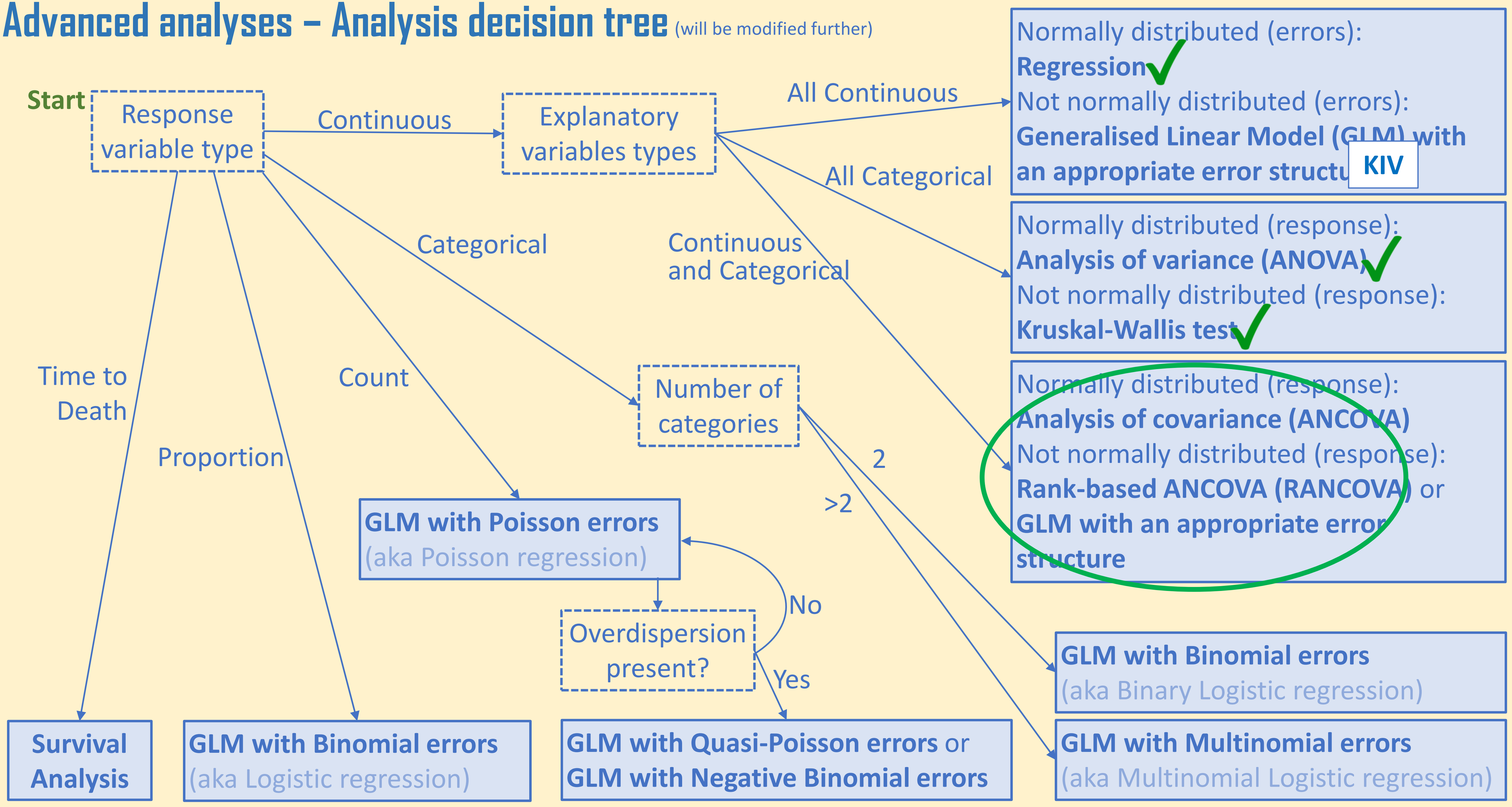


```
> summary.aov(mod2.4)
             Df Sum Sq Mean Sq F value   Pr(>F)    
site          2  34.61   17.30   18.21 1.14e-05 ***
site:treated  3  39.68   13.23   13.92 1.31e-05 ***
Residuals    26  24.70    0.95                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note: We are forcing the model to partition the variation using <site> FIRST, and THEN (within each level of <site>) further partition using <treated>. That’s why there’s no p-value for <treated> on its own. We know this is wrong—that’s why it’s important to know whether your experimental design is Factorial or Split Plot.

ANCOVA

Advanced analyses – Analysis decision tree (will be modified further)



Start

Response variable type

Continuous

Explanatory variables types

All Continuous

Normally distributed (errors):
Regression ✓
Not normally distributed (errors):
Generalised Linear Model (GLM) with an appropriate error structu

KIV

All Categorical

Normally distributed (response):
Analysis of variance (ANOVA) ✓
Not normally distributed (response):
Kruskal-Wallis test ✓

Continuous and Categorical

Normally distributed (response):
Analysis of covariance (ANCOVA)
Not normally distributed (response):
Rank-based ANCOVA (RANCOVA) or **GLM with an appropriate error structure**

Categorical

Number of categories

2

GLM with Binomial errors
(aka Binary Logistic regression)

>2

GLM with Multinomial errors
(aka Multinomial Logistic regression)

Count

GLM with Poisson errors
(aka Poisson regression)

Overdispersion present?

No

Yes

GLM with Quasi-Poisson errors or
GLM with Negative Binomial errors

Proportion

GLM with Binomial errors
(aka Logistic regression)

Time to Death

Survival Analysis

What is Analysis of Covariance (ANCOVA)?

Used when your **ONE response variable is continuous** and you have **both continuous and categorical explanatory variables**.

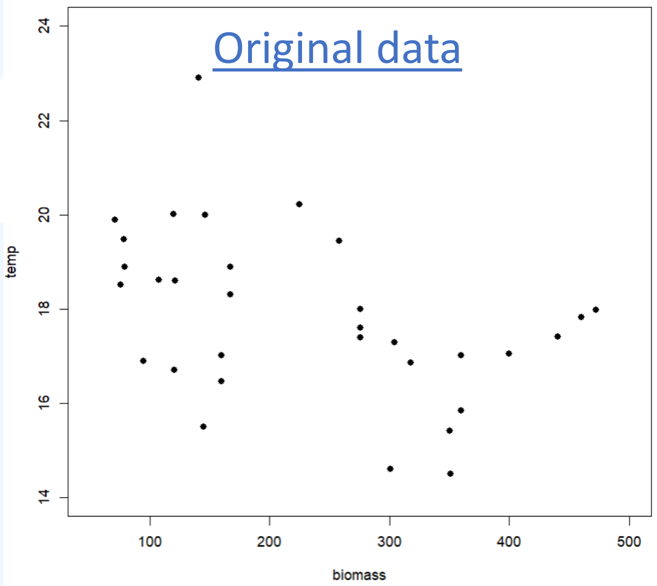
ANCOVA is a combination of Regression and ANOVA:

1) It will fit a model between the response variable and the continuous explanatory variable (aka the covariate) for each level of the categorical explanatory variable.

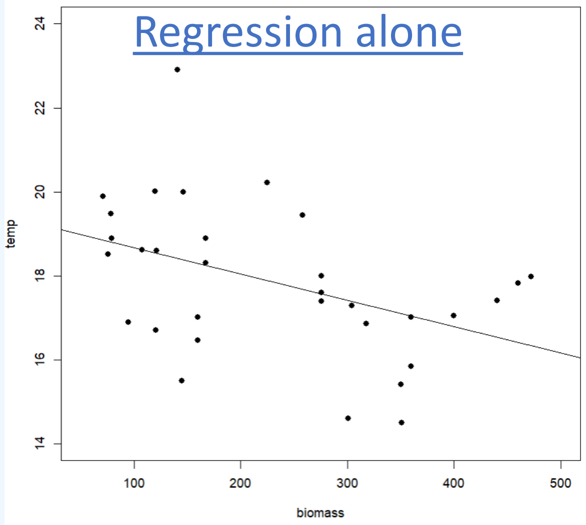
2) It will then give you the intercepts and slopes for each level.

What is Analysis of Covariance (ANCOVA)?

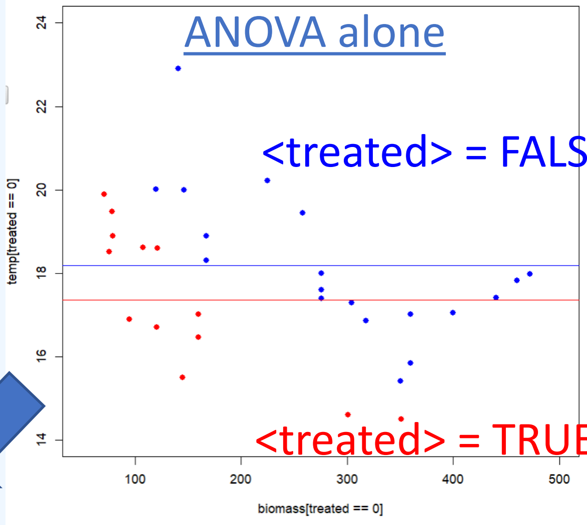
The Power of ANCOVA…



There is still a lot of overlap between the two levels

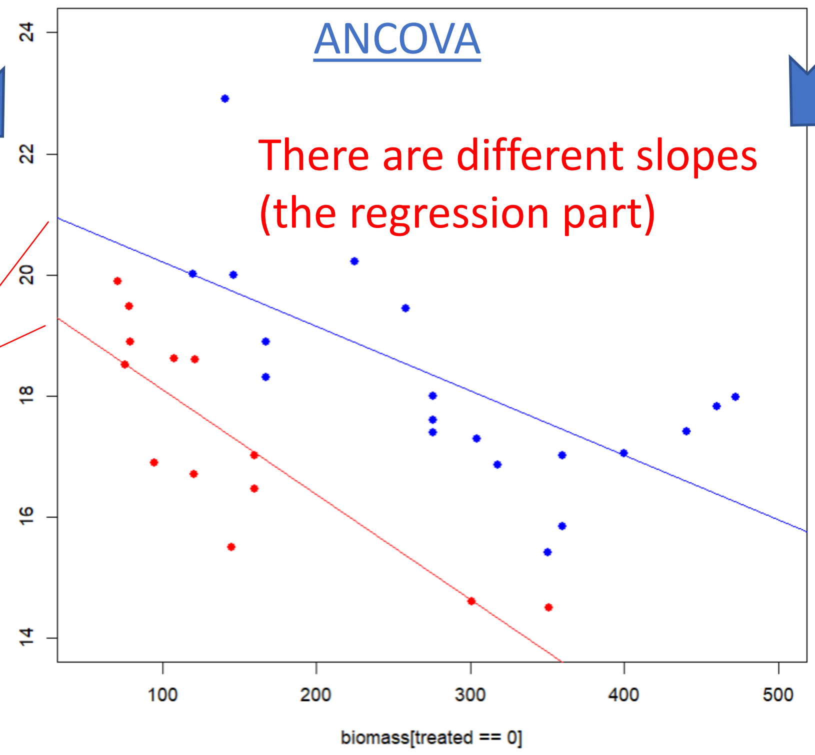


There is maybe only a very weak negative correlation.



<treated> = FALSE

<treated> = TRUE



There are different slopes (the regression part)

There are different intercepts (the ANOVA part)

If you do the regression for each level: now there is a very clear negative relationship between <temp> and <biomass>!

Examples

Continuous response variable	Explanatory variables
Biodiversity	Country: Singapore, Malaysia, Indonesia (categorical – 3 levels) Biomass: in kg (continuous) [Covariate]

1-way ANCOVA

Species population	Country: Singapore, Malaysia, Indonesia (categorical – 3 levels) Site: A, B, C (categorical – 3 levels) Biomass: in kg (continuous) [Covariate]

2-way ANCOVA

Assumptions (same as ANOVA)

The response variable is normally distributed within each level.

> Check by testing the normality of either (i) the datapoints in each level individually (easy to do if you have few groups; or (ii) the residuals of the model. Both test the same thing.

The variances within each level are equal (i.e. equality of variances).

Each datapoint is independent.

Absence of significant outliers.

Explore your data

We want to see whether <site> or <treated> can make sense of the relation between <temp> and <biomass>.

#Plot <temp> against <biomass>, by the levels of <site>

```
plot(temp[site==4]~biomass[site==4],data=d6,pch=16,
col="darkgreen",xlim=c(50,500),ylim=c(14,24))
```

```
points(temp[site==6]~biomass[site==6],data=d6,pch=16,
col="orange")
```

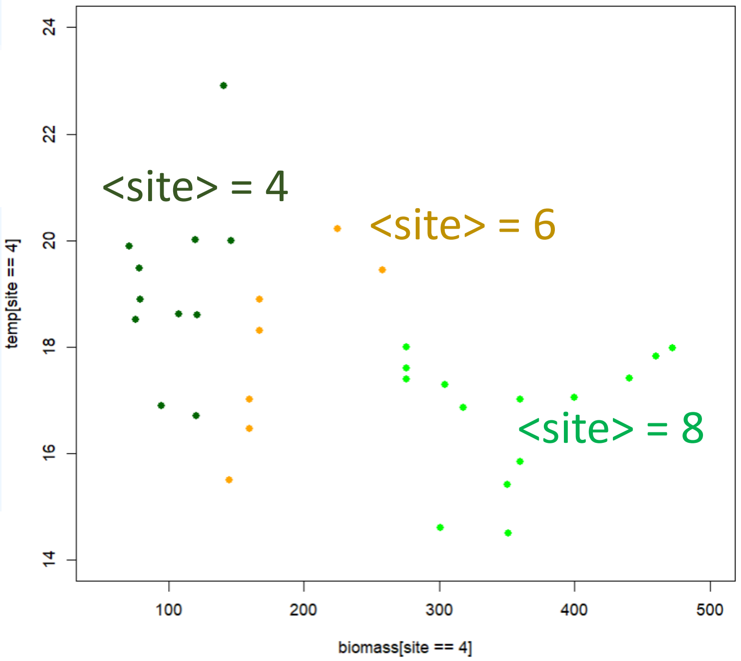
```
points(temp[site==8]~biomass[site==8],data=d6,pch=16,
col="green")
```

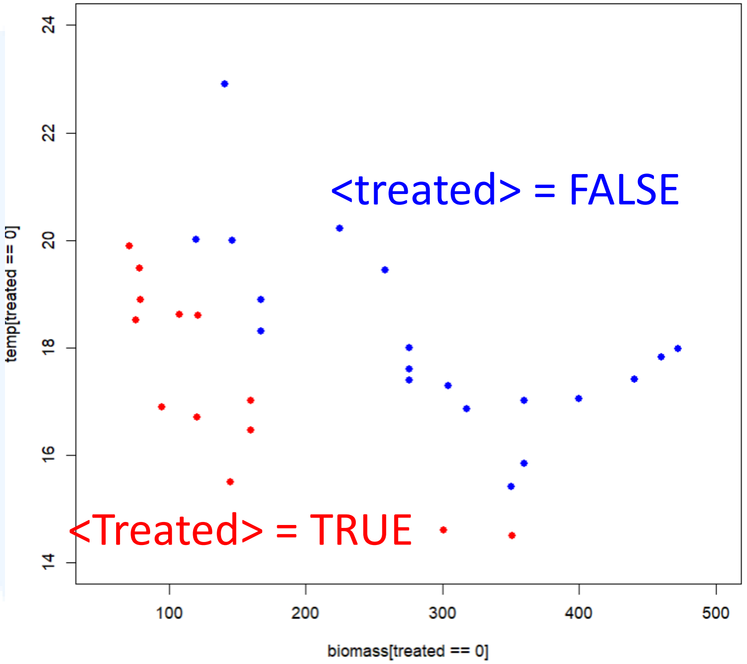
#Plot <temp> against <biomass>, by the levels of <treated>

```
plot(temp[treated==0]~biomass[treated==0],data=d6,pch
=16,col="blue",xlim=c(50,500),ylim=c(14,24))
```

```
points(temp[treated==1]~biomass[treated==1],data=d6,p
ch=16,col="red")
```

It looks like both reveal hidden structure in the dataset but <treated> is more straightforward to interpret.





Fit the ANCOVA

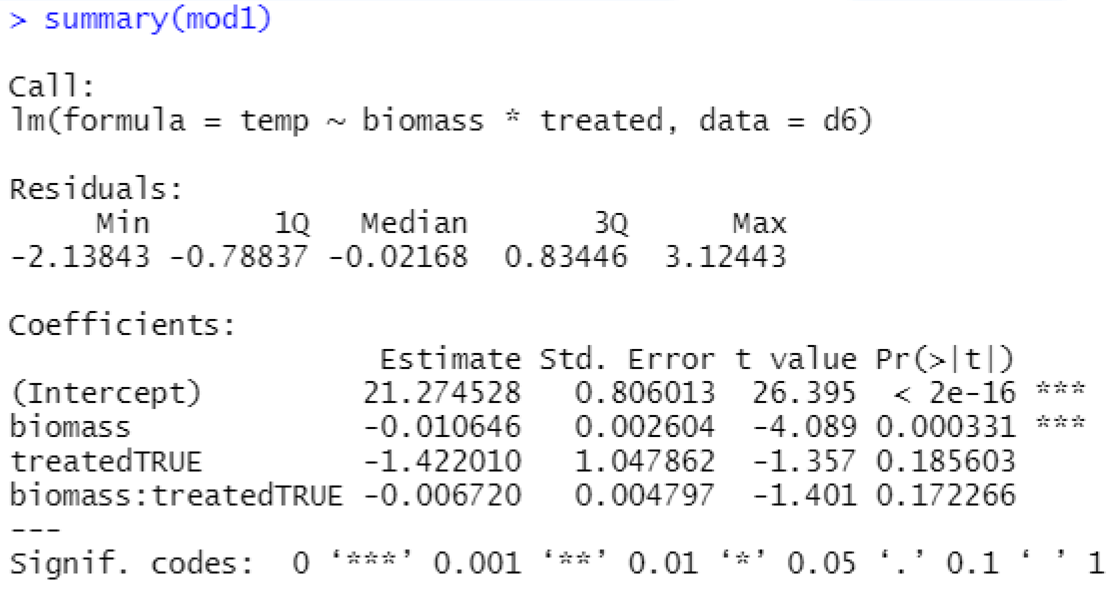
Debugging Time!

Can you spot the error in this slide?

#Run an ANCOVA with <biomass> and <treated>, allowing them to interact

```
mod1=lm(temp~biomass*treated,data=d6)
summary(mod1)
```

#interaction is non-significant



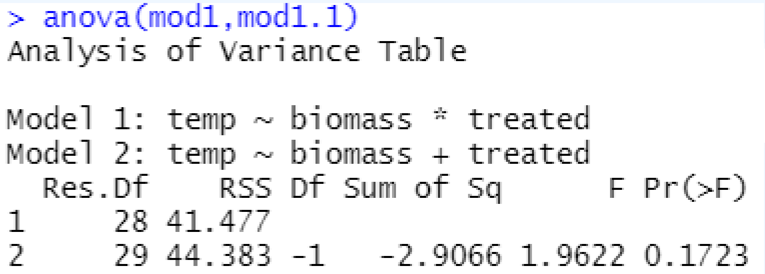
```
> summary(mod1)

Call:
lm(formula = temp ~ biomass * treated, data = d6)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.13843 -0.78837 -0.02168  0.83446  3.12443 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          21.274528   0.806013  26.395  < 2e-16 ***
biomass              -0.010646   0.002604  -4.089 0.000331 ***
treatedTRUE          -1.422010   1.047862  -1.357 0.185603    
biomass:treatedTRUE -0.006720   0.004797  -1.401 0.172266    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#Simplify by stepwise deletion

```
mod1.1=update(mod1,.-biomass:treated)
```

#Compare the 2 models using anova()

```
anova(mod1,mod1.1)
```



```
> anova(mod1,mod1.1)
Analysis of Variance Table

Model 1: temp ~ biomass * treated
Model 2: temp ~ biomass + treated
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     28 41.477                           
2     29 44.383 -1   -2.9066 1.9622 0.1723
```

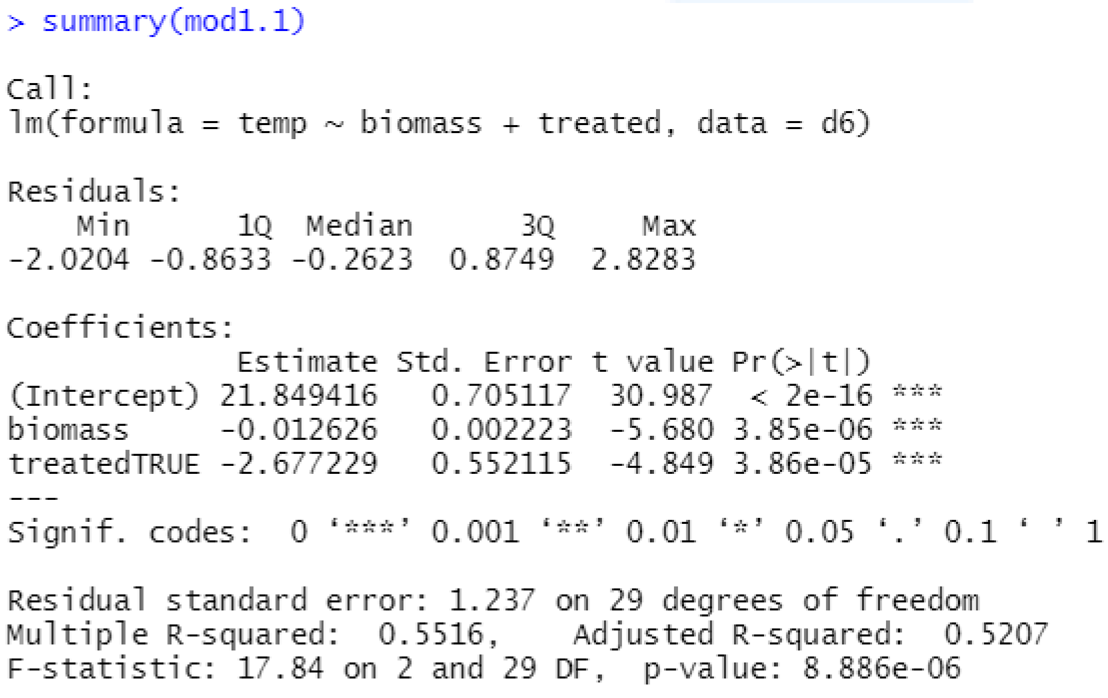
#There is no significant reduction in predictive power ($P > 0.05$), so we prefer the simpler model (i.e. mod1.1, without the interaction term).

View the results

#Summary

```
summary(mod1.1)
```

#Both significant: looks like our final model



```
> summary(mod1.1)

Call:
lm(formula = temp ~ biomass + treated, data = d6)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0204 -0.8633 -0.2623  0.8749  2.8283 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 21.849416   0.705117  30.987  < 2e-16 ***
biomass     -0.012626   0.002223  -5.680 3.85e-06 ***
treatedTRUE -2.677229   0.552115  -4.849 3.86e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.237 on 29 degrees of freedom
Multiple R-squared:  0.5516,	Adjusted R-squared:  0.5207 
F-statistic: 17.84 on 2 and 29 DF,  p-value: 8.886e-06
```

#Check assumptions

```
par(mfrow=c(2,2))
plot(mod1.1)
```

#All look good!

Interpretation: “Both treated (*P* < 0.001) and biomass (*P* < 0.001) have significant effects on temperature. Temperature decreases by 0.01 ± 0.002 degrees Celsius (mean ± SE) for every 1kg increase in biomass. Specimens that have been treated are on average 2.68 ± 0.55 degrees Celsius cooler than those that have not.”

What if residuals are not normal?

Option 1: Transform your variable(s).

Option 2: Use rANCOVA (by Thomas Forstner).

```
#Define the function and run the test
rancova=function(y,cov1,treatment){
      ry=rank(y)
      rcov1=rank(cov1)
      e=lm(ry~rcov1)$residuals
      m=aov(e~treatment)
      summary(m)
}
rancova(y=d6$temp,cov1=d6$biomass,treatment=d6$treated)
```

Continuous x-variable (covariate)

Categorical x-variable

Function if you want to test 2 covariates

```
rancova=function(y,cov1,cov2,treatment){
        ry=rank(y)
        rcov1=rank(cov1)
        rcov2=rank(cov2)
        e=lm(ry~rcov1+cov2)$residuals
        m=aov(e~treatment)
        summary(m)
}
rancova(y=d6$temp,cov1=d6$biomass,cov2=d6$numSen,treatment=d2$treated)
```

Note: at the moment, it’s not possible to test multiple categorical x-variables.

Option 3: use a GLM (later lectures).

What if equality of variance is violated?

Option 1: Transform response variable.

Option 2: Use GLS or a GLM (later lectures).

2-way ANCOVA with 1 covariate

#Explaining <temp> using (i) <biomass> (covariate) and (ii) <treated> interacting with <site>

```
mod4=lm(temp~biomass+treated*site,data=d6)
```

#If we think <biomass> may also interact with one of the other variables

```
mod4=lm(temp~biomass*treated*site,data=d6)
```

3-way ANCOVA with 2 covariates

#Explaining <temp> using <numSen> and <biomass> (covariates); <treated>, <site> and <managed>; and two 2-way interactions between <managed>:<treated> and <managed>:<site> only

```
mod5=lm(temp~numSen+biomass+treated+site+managed+managed:treated+managed:site,
data=d6)
```

Tis’ the time to…

Kahoot!

Summary (Learning Objectives)

Analysis of variance (ANOVA)

- Assumptions, fitting, checking and interpreting
- Alternatives: Welch’s one-way ANOVA, Kruskal-Wallis test
- Repeated measures ANOVA (and Friedman test)
- Factorial vs. Split plot designs

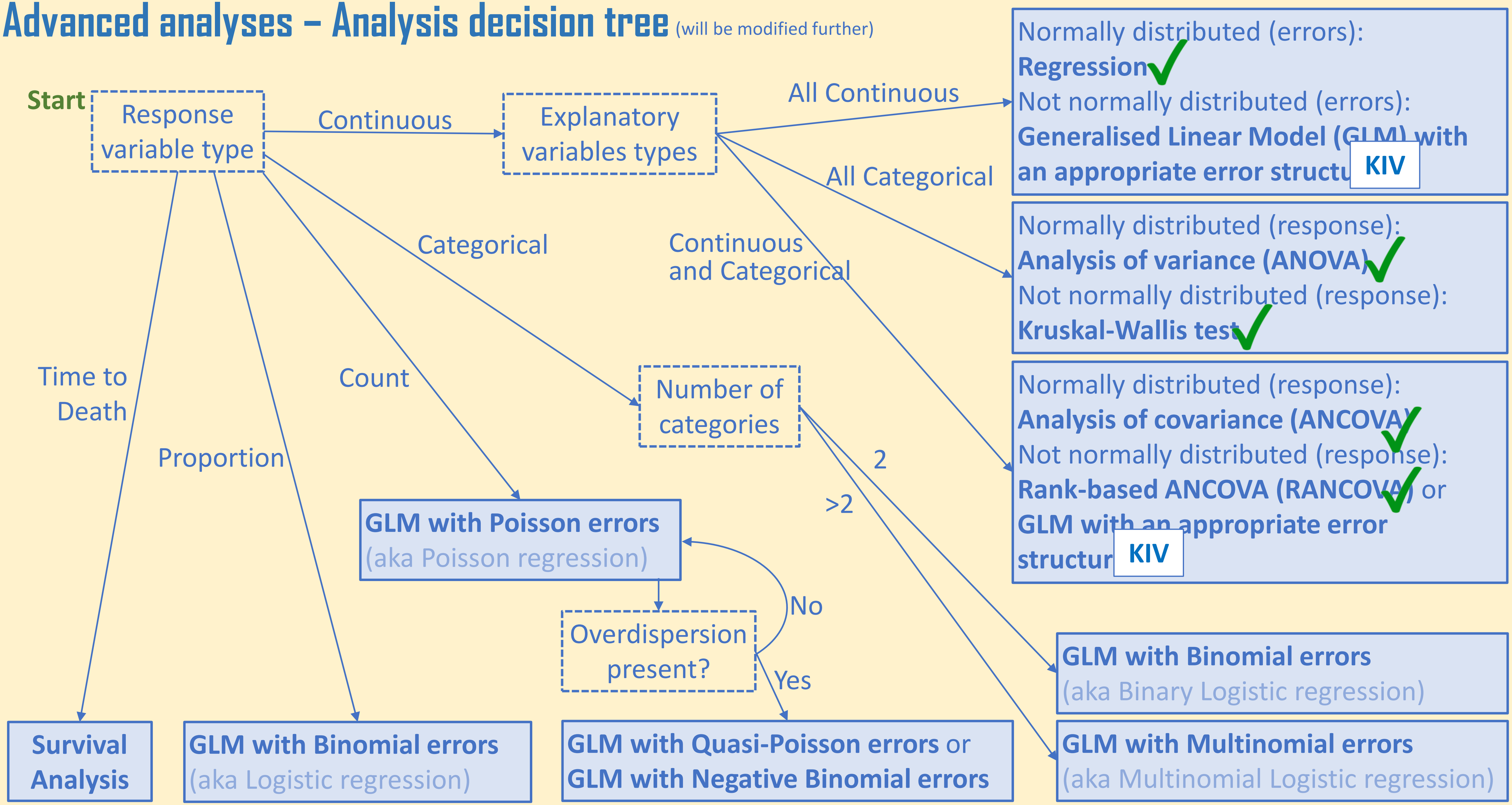
  Factorial experiments: 2-way and 3-way ANOVA

  Nested ANOVA

Analysis of covariance (ANCOVA)

- Assumptions, fitting, checking and interpreting
- Alternative: Rank-based ANCOVA (RANCOVA)
- Factorial experiments: 2-way and 3-way ANCOVA

Advanced analyses – Analysis decision tree (will be modified further)



Start

Response variable type

Continuous

Explanatory variables types

All Continuous

Normally distributed (errors):
Regression ✓
Not normally distributed (errors):
Generalised Linear Model (GLM) with an appropriate error structu KIV

All Categorical

Normally distributed (response):
Analysis of variance (ANOVA) ✓
Not normally distributed (response):
Kruskal-Wallis test ✓

Continuous and Categorical

Normally distributed (response):
Analysis of covariance (ANCOVA) ✓
Not normally distributed (response):
Rank-based ANCOVA (RANCOVA) ✓ or
GLM with an appropriate error structur KIV

Categorical

Number of categories

2

GLM with Binomial errors
(aka Binary Logistic regression)

>2

GLM with Multinomial errors
(aka Multinomial Logistic regression)

Count

GLM with Poisson errors
(aka Poisson regression)

Overdispersion present?

No

Yes

GLM with Quasi-Poisson errors or
GLM with Negative Binomial errors

Proportion

GLM with Binomial errors
(aka Logistic regression)

Time to Death

Survival Analysis