An Introduction to Experimental Design ANOVA and ANCOVA

Andrew P Beckerman (with support from text and slides from Mark Rees and Gareth Phoenix)

2022-12-05

Contents

1 Introduction **5**
1.1 The Three Rs: The Foundation of Experimental Design. 6
1.2 The General Linear Model 6

2 Readings —- **9**
2.1 Install some extra packages —- 9

3 Introduction To Experimental Design **11**
3.1 Conepts associated with causation 12
3.2 Components of an Experiment 13
3.3 The holy grail of a control 13
3.4 So what does a good experimental design do? 13
3.5 How do we increase precision and reduce bias? 14
3.6 Experimental vs. Measurement units 15
3.7 Mini-Quiz 17
3.8 Replication: How many? 17

4 Examples and Challenge Questions **19**
4.1 Designing your first experiment 19

5 Design and Analysis of Experiments **23**
5.1 A CRD (Completely Randomised Design) Example 24
5.2 A priori vs. Post-Hoc Contrasts 29
5.3 THE RCBD - The Randomised Complete Block Design 33
5.4 An example of the RCBD 36
5.5 Analysing the CRBD 37

6 Designs for testing for interactions: the two-way ANOVA and factorial designs. **43**
6.1 Introducing Interactions 43
6.2 A Factorial Design and the Two-Way ANOVA 46

7 Interactions Part 2: Introducing the ANCOVA (analysis of covariance) **51**

7.1 What this chapter covers. 52
7.2 Setting up the various ideas. 52
7.3 Working through an ANCOVA example. 55
7.4 Building the model (and understanding it) 58
7.5 Some General Principles for ANOVA and ANCOVA modelling. . 69
7.6 A few parting tricks about specifying models in R 71

Chapter 1

Introduction

Welcome to An Introduction to Experimental Design ANOVA and ANCOVA

In this mini-module, you'll be learning about the principles of experimental design and analysis of a few classic designes, the 2x2 ANOVA and ANCOVA experiments.

This module is compulsory for all, because it forms the foundation for most of the more complex experiments you will do as a researcher. And it is the major step beyond the t-test, 1-way ANOVA, simple regression and chi-square contingency table analyses we've covered thus far.

The learning outcome for this mini-module are that you will understand the basic ideas about

- Replication, Randomisation and Reducing Noise
- Precision, Bias and Systematic Error
- The Completely Randomised Design
- The Randomised Block Design
- The 2-way ANOVA
- The ANCOVA Design

In order to be successful with this final section of the course, you need to feel comfortable with the 1-way ANOVA and the Regression model. Please review these concepts. You can also refer to Chapter 5 and 6 in Getting Started with R (available as an online Resource via STARPlus) which covers a great deal of the mechanics of using R to do these types of models. Finally, you will also need to feel comfortable with dplyr and ggplot - we'll be reinforcing the old stuff and introducing a few new tricks.

1.1 The Three Rs: The Foundation of Experimental Design.

Before we get started, it's vital that you understand that there are some very basic principles needed to ensure that your experiments can provide robust and reliable inference (answers to your questions). The "3 R's".

- **Randomisation**: the random allocation of treatments to the experimental units, to avoid confounding between treatment effects and other unknown effects.
- **Replication**: the repetition of a treatment within an experiment, to quantify the natural variation between experimental units and increase accuracy of estimated effects.
- **Reduce noise:** by controlling as much as possible the conditions in the experiment, e.g. by grouping of similar experimental units in blocks.

At this point, you may want to revisit, again, the following section of the APS 240 book on Randomisation

1.2 The General Linear Model

This section of the course is focused on a class of model called the General Linear Model. It is not a **GLM**. The **GLM** is a generalised linear model. I know, right?

The general linear model is, as we learned in the past few weeks, a model fit in R with the `lm()` function. It includes regression, ANOVA, ANCOVA and variations of these.

There are a few key characteristics to remember about these models. The general linear model has the following form:

$y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \epsilon$

Where the y is the response variable, the β's are estimated parameters, the X's are the predictor variables and the ϵ comes from a Gaussian distribution with zero mean and constant variance.

Let's decompose that a bit more.

There are two types of predictor variable:

Metric predictor variables are measurements of some quantity that may help to predict the value of the response. For example, if the response is the blood pressure of patients in a clinical trial, then age, fat mass and height are potential metric predictor variables. You may know these as **continuous explanatory (independent) variables**

Factor variables are labels that serve to categorize the response measurements into groups, which may have different expected values. Continuing the blood

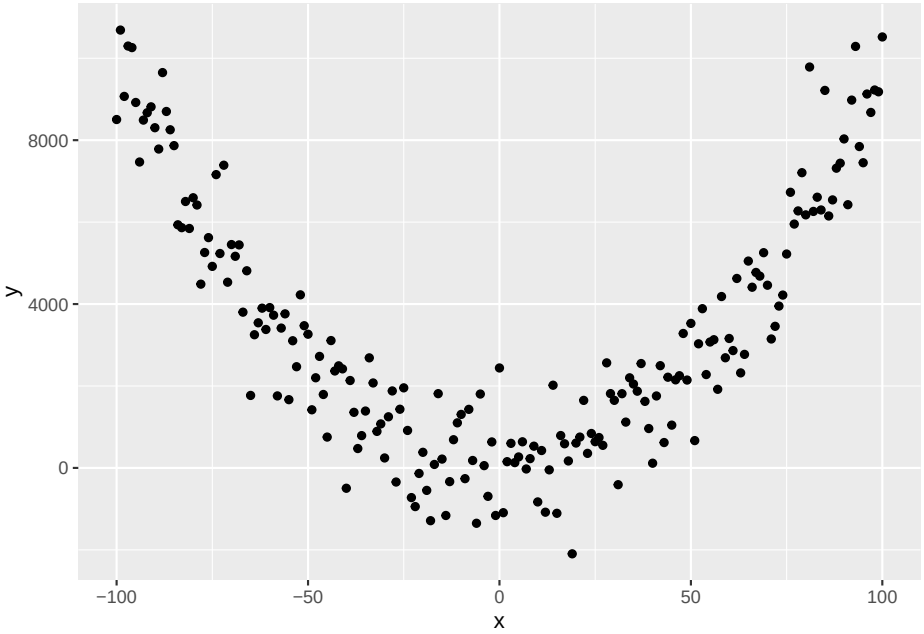
pressure example, factor variables might be sex and drug treatment received (drug A, drug B or placebo, for example). You may also know these as **categorical explanatory (independent) variables**.

So, you hopefully can see how this *general* linear model is capable of representing

1. *ANOVA* – Analysis of variance -> Predictors are factors.
2. *Regression* -> Predictor is a metric variable (continuous variable).

3. *Multiple regression* -> Predictors are metric variables (continuous variables).
4. *ANCOVA* - Analysis of co-variance -> Predictors are a mixture of metric variables (continuous variables) and factors.

Finally, it is important to understand that non-linear relationships such as these data below can be modelled with a linear model:



How, you ask!? Well…. consider this equation:

$y = 0.01 + x + x^2 + \epsilon$

Referring to our generic model structure above,

$y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \epsilon$

we hopefully can see that $\beta_0 = 0.01$, $\beta_1 = 0$ and $\beta_2 = 1$, where $X_2 = X^2$!

Linear models are perfectly capable of being used to estimate non-linear relationships!

Here is the code to make that figure.

```
# set x range
x <- -100:100
# define y without error
y_det <- 0.01+x^2
# add some random variation
y <- y_det+rnorm(length(x),0,1000)

# create dataframe and plot
df <- data.frame(x, y)
ggplot(df, aes(x = x, y = y))+
  geom_point()
```

Chapter 2

Readings —-

There are several resources that will help with this section of the stats course, and onwards

- Getting Started with R - An Introducton for Biologists, Second Edition (available as an electronic online resource via StarPlus). Specifically Chapter 5 and 6.
- Experimental Design for the Life Sciences - Nick Colegrave and Graham Ruxton (seen on eBay for £2.50!)
- Of course, the venerable coursebook for APS 240: https://dzchilds.github.io/stats-for-bio/index.html

2.1 Install some extra packages —-

In order to make this module more effective, we are going to use some additional resources from CRAN.

Please install these packages, if you have not already, using the install packages tab in RStudio:

- `tidyverse`
- `ggfortify`
- `agricolae`
- `car`
- `gmodels`
- `visreg`
- `patchwork`

Chapter 3

Introduction To Experimental Design

Experiments help us answer questions, but there are also non-experimental techniques. What is so special about experiments?

One of the central features of an experiment is the *treatment* - a manipulation of some variable of interest that should have an effect on the response variable we are investigating. Whether you are manipulating the levels of a hormone to explore it's impact on a cell/organ or embryo development, the concentration of a drug to explore it's efficancy in treating a disease or the levels of nitrogen in soil to explore the impacts on plant growth, a treatment is a deliberate manipulation.

It is also important to remember that there can be *natural* treatments - there may be natural variation among cells, organisms or gradients in the environment that you can use to represent treatments.

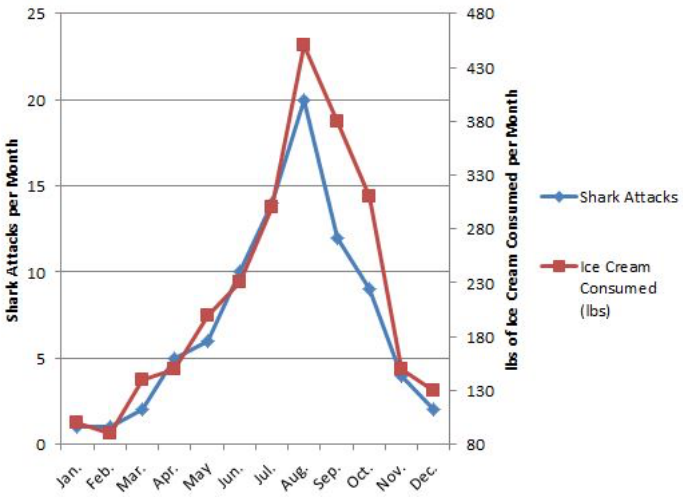
So, to be very clear:

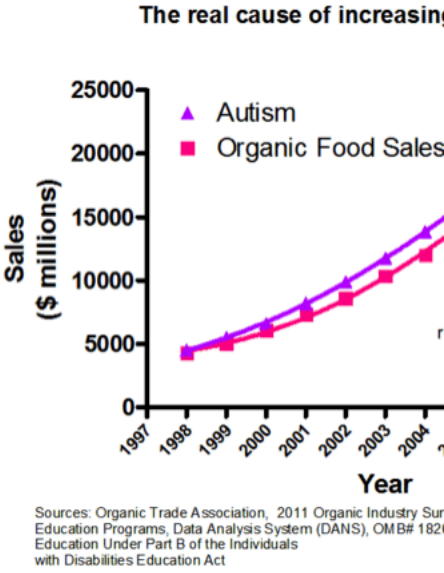
1. Experiments allow us to set up a direct comparison among the *levels* or *values* of *treatments* of interest.
2. We can design experiments to minimize any bias in the comparison.
3. We can design experiments so that the error in the comparison is small.
4. We design experiments to be in control, and having that control allows us to make stronger inferences about the nature of differences that we see in the response variable.

   Experiments allow us to move towards making inferences about causation.

This last point distinguishes an experiment from an observational study. In an observational study we merely observe which units are in which treatment

groups; we don't get to *control that assignment*. This underpins the classic issue with assigning *causation to correlation* - in the following two examples, there is a strong association between the variables, but there has been no control/manipulation.





3.1 Conepts associated with causation

Mosteller and Tukey (1977) list three concepts associated with causation and state that at least two (preferably all three) are needed to support a causal relationship:

- *Consistency* – make a change and the response is in the same direction or the amount of response is *consistent* across populations
- *Responsiveness* – make a change and the response changes according to theory
- *Mechanism* – make a change and we can monitor/identify a mechanism leading from cause to effect

Let's look at a classic example. Smoking and lung cancer – from 1922 to 1947 annual deaths for lung cancer went from 612 to 9287 (Observation). This was thought in the 1950s to be either an effect of smoking tobacco or atmospheric pollution (Hypothesis). Numerous studies showed that lung cancer was more prevalent in smokers (Observation: *consistency*). Chemical analyses of tobacco showed it contained carcinogens (Association: *mechanism*). Public health programs resulted in a reduction in smoking and lung cancer rates decreased (Intervention: *responsiveness*).

Note the initial study was an observational study and in this case it was not ethical to do the experiment per se!

3.2 Components of an Experiment

An experiment has *treatments, experimental units, responses, and a method to assign treatments to units*. These four things specify the experimental design.

Not all experimental designs are created equal. A good experimental design must adhere to the 3Rs. It should reveal consistency, responsiveness and mechanism. The way this happens is by avoiding systematic error in measuring things, and allow estimation of error in measurements with precision.

3.3 The holy grail of a control

At this point, it would be good to revisit the APS 240 sections on controls and procedural controls

3.4 So what does a good experimental design do?

In short, a good experimental design must:

- Avoid systematic error
- Allow estimation of error
- Be precise
- Have broad validity.

Lets walk through some definitions.

If our experiment has *systematic error*, then our comparisons will be biased, no matter how precise our measurements are or how many experimental units we use. **Randomisation** is our tool to combat *systematic error*.

Even without *systematic error*, there will be random error in the responses - this is what we call variation in what we are measuring or more formally variance. Such variation in responses invariably leads to random error in the treatment comparisons. When we compared two means in the t-test, we had to deal with the variation in both groups!

Experiments are precise when this random error in the treatment comparisons is small. Precision depends on the size of the random errors in the responses, the number of units used (**replication**), and the experimental design used.

Experiments must be designed so that we have an estimate of the size of random error. This permits statistical inference: for example, confidence intervals

(which arise from standard errors) or tests of significance based on t- or F-statistics.

We cannot do inference without an estimate of this variation. We would like our conclusions to be valid for a wide population, so we need to *randomise* our subjects or objects we are measuring - for example, we may need to be aware of both sexes and of young and old individuals. But there are always compromises - for example, broadening the scope of validity by using a variety of experimental units may decrease the precision of the responses.

3.5 How do we increase precision and reduce bias?

There are several key concepts

3.5.1 Blinding

Blinding occurs when the evaluators of a response do not know which treatment was given to which unit. Blinding helps prevent bias in the evaluation, even unconscious bias from well-intentioned evaluators. Double blinding occurs when both the evaluators of the response and the (human subject) experimental units do not know the assignment of treatments to units. Blinding the subjects can also prevent bias, because subject responses can change when subjects have expectations for certain treatments.

3.5.2 Placebos

Placebo is a null treatment that is used when the *act* of applying a treatment—any treatment — has an effect. Placebos are often used with human subjects, because people often respond to the process of receiving any treatment: for example, reduction in headache pain when given a sugar pill. Blinding is important when placebos are used with human subjects. Placebos are also useful for nonhuman subjects. The apparatus for spraying a field with a pesticide may compact the soil. Thus we drive the apparatus over the field, without actually spraying, as a placebo treatment.

3.5.3 Confounders

Confounding occurs when the effect of one factor or treatment cannot be distinguished from that of another factor or treatment. The two factors or treatments are said to be confounded. Except in very special circumstances, confounding should be avoided. Consider the following example. We plant corn variety A in Yorkshire and corn variety B in Lancashire. In this experiment, we cannot distinguish location effects (Yorkshire vs. Lancashire) from variety effects (cornA vs. cornB) — the variety factor and the location factor are confounded.

This is despite the fact that we know that Yorkshire will be better…. (that's a joke)

3.6 Experimental vs. Measurement units

A common source of difficulty in designing experiments is the distinction between experimental units and measurement units. We need to know the experimental units, as this is the key value used to generate our inference.

Now is a good time to re-look at the short section on Jargon Busting from the APS 240 book.

3.6.1 Experimental and measurement units: an example

Consider an educational study, with six classrooms of 25 pupils. Each classroom of students is then assigned, at random, to one of two different reading programmes.

At the end of a six-week term, all the students are evaluated via a common reading exam.

3.6.1.1 The challenge question

> Are there six experimental units (the classrooms) or 150 (25*6; the students)? We measured the reading ability of the students… but they were in classroom sets of 25….

3.6.2 Identifying the experimental unit - an example of pseudo-replication

To identify the experimental units the key question is: To which *thing* (students or classrooms) did we randomly allocate our treatments?

If we randomly allocated reading programmes to students, then students would be the experimental units. But we didn't, so the classroom is the experimental unit – it is the classroom to which we randomly allocated treatments.

The classroom is the experimental unit.

However, you are right - we don't *measure* how a classroom reads; we measure how students read. Thus *students are the measurement units* for this experiment.

3.6.3 Psudo-replication

Confusing these two things can lead to **pseudo-replication**. Treating measurement units as experimental usually leads to overoptimistic analysis — we will reject null hypotheses more often than we should, and our confidence intervals

will be too narrow. The usual way around this is to determine a single response for each experimental unit.

There is additional information on Independence and Pseudoreplication in the the APS 240 book.

3.6.3.1 Independence: and example

Consider an experiment with two growth chambers each containing 100 plants. One of the chambers received enhanced C02. One night after collecting data you leave the door open on the C02 chamber and the temperature drops and so the plants grow more slowly. When you come to analyze the data you get a highly significant effect of slow growth. However, that C02 results in reduced plant growth not what you expect (CO2 is good for photosynthesis…).

This is an entirely plausible outcome caused by misallocating plants as the experimental unit - it was really the CO2 chamber… to avoid such problems, one needs many chambers.

Consider a second experiment where you have 200 growth chambers and randomly allocate plants to each. If you *forget to close one door* it really has no effect as just one plant is affected. In fact, to get the same effect as in the first experiment you would have to accidentally leave the doors open on all 100 of the elevated C02 chambers. This is very unlikely indeed!!!

> There are 9 x 1058 ways selecting 100 chambers from 200 chambers
> so the chance of accidentally picking all the elevated C02 chambers
> is 1/ 9x1058 0 (stars in universe 7 x 1022).

Proper **randomization** and **replication** is very different from **pseudoreplication**.

3.6.4 Randomization with Replication protects against Confounding

An experiment is properly randomized if the method for assigning treatments to units involves a known, well-understood probabilistic scheme. The probabilistic scheme is called a randomization.

> In general, more experimental units with fewer measurement units
> per experimental unit works better.

No matter which features of the population of experimental units are associated with our response, our randomizations should put approximately *half the individuals with these features* into *each treatment group.*

Recall our example above of considering sex and age of subjects and imagine a treatment with two levels (hot and cold). Done well, proper randomisation will put approximately half the males, half the females, half the older, half the younger etc into each of the treatment levels.

The beauty of randomization is that it helps prevent *confounding*, even for factors that we do not know are important.

3.6.5 haphazard is NOT randomized - beware the non-randomized experiment –

A company is evaluating two different word processing packages for use by its clerical staff. Part of the evaluation is how quickly a test document can be entered correctly using the two programs. We have 20 test secretaries, and each secretary will enter the document twice, using each programme once.

Suppose that all secretaries did the evaluation in the order A first and B second. Does the second programme have an advantage because the secretary will be familiar with the document and thus enter it faster? Or maybe the second programme will be at a disadvantage because the secretary will be tired and thus slower?

Randomization generally costs little in time and trouble, but it can save us from `disaster`. The experiment above needs secretaries randomly assigned to A first -> B second and B first -> A second (50% in each!).

Anything that might affect your responses should be *randomized*! For example

- If the experimental units are not used simultaneously, you can (should) randomize the order in which they are used.
- If the experimental units are not used at the same location, you can (should) randomize the locations at which they are used.
- If you use more than one measuring instrument for determining response, you can (should) randomize which units are measured on which instruments.

3.7 Mini-Quiz

> A PhD student want to determine the effects of protein on beetle reproduction, so they design an experiment with a control and protein enhanced diet. To assign beetles to each of the treatments they pour a culture onto the table and catch the first 30 beetles that run to the edge of the table, these receive the protein enhanced diet. The next 30 beetles go in the control. **Is this randomized?**

(Hmm …. is there anything about the first 30 beetles that reach the edge of the table that could *bias* your inference?)

3.8 Replication: How many?

There is a really common question that people ask. How many replicates do I need? Unfortunatly, there are no simple rules… it depends… on….

- Resources available ($/£/€ and equipment and time)
- Variability of experimental units
- Treatment structure
- Size of effect (response)
- Relative importance of different comparisons

There is, however, a set of tools that can help with estimating sample sizes. It's called power analysis and requires that you have some a priori estimate of the expected variation in your response variable.

Chapter 4

Examples and Challenge Questions

In this section, we will review a field based experimental design example. There are challenge questions to answer.

We will also introduce tools to generate randomised experimental designs - this is a good trick to have up your sleeves!

4.1 Designing your first experiment

You are challenged to design an Arctic field manipulation experiment to evaluate UV-B radiation and increased CO2 impacts on plant growth.

Context: an arctic tundra study Increasing ultraviolet-B (+UV-B) radiation from ozone depletion (the arctic ozone hole) Increasing atmospheric CO2 (+CO2) from anthropogenic emissions For plants: +UV-B potentially harmful, +CO2 potentially beneficial

Hypotheses Elevated (+) UV-B radiation will reduce the growth of arctic plants Elevated (+) CO_2 will increase the growth of arctic plants

The resources available to you are constrained. The arctic research station has given permission for 16 plots (each 2m x 2m) in the natural vegetation nearby.

One +UV-B plot (2m x 2m) costs £4000 (this provides the UV-B lamps, frame, power and control system, wooden walkways around the plots) One $+CO_2$ plot (2m x 2m) costs £6000 (this provides a CO2 release system, CO2 control and covers CO2 purchase costs, wooden walkways around the plots) One control plot (2m x 2m) costs £200 (marking posts, wooden walkways around the plots)

You have a budget of £61,000

Design an experiment to test the hypotheses (i.e. the design of the plots and treatments including replication, not what measurements you will take - which will be plant growth rates....). Think hard about this. How many treatments do you have? How many plots/treatment would you like to allocate? Is this

possible? Will this be a balanced design, given your max budget? If it isn't, what rule can you use to allocate them replicates to treatments?

Write an answer down, before moving to the next section. We'll provide the answer separately!

Chapter 5

Design and Analysis of Experiments

In this section we are going to learn about how to implement two experimental designs:

- CRD: the completely randomised design
- CRBD: the completely randomised block design.

These two designs are valuable in dealing with two things that make it hard to make strong inference from experiments: *noise* and *counfounding effects*.

These unwated sources of variation comes in two forms (see APS 240 reading)

1. The first is *confounding variation*. This occurs when there are one or more other sources of variation that work in parallel to the factor we are investigating and make it hard, or impossible, to unambiguously attribute any effects we see to a single cause. Confounding variation is particularly problematic in observational studies because, by definition, we don't manipulate the factors we're interested in.

2. The second is *noise*. This describes variation that is unrelated to the factor we are investigating but adds variability to the results so that it is harder to see, and detect statistically, any effect of that factor. As noted above, much of experimental design is about improving our ability to account for noise in a statistical analysis.

We will consider these together, as some of the techniques for dealing with them are be applicable to both. The primary tools for dealing with them are

1. randomisation
2. blocking
3. appropriate controls

4. additional treatments.

In the following sections, we are going to focus on the first three. In doing so, we will also revisit how we make inference from 1-way ANOVA experiments and introduce a more generalised approach to making *contrasts* we want. If you recall from the 1-way ANOVA work you did in the previous module (Week 7), we learned about *treatment contrasts*, the default comparison of means to a reference level, and then the *Tukey test*, which makes all pairwise comparisons. Here will will find an intermediate zone.

5.1 A CRD (Completely Randomised Design) Example

The experiment is about plant crop biomass yield under several herbicide treatments - the herbicide targets weeds and not our target plant (e.g. Glyphosate): a control and two herbicides, and a third treatment that is a placebo - applied water but no herbicide.

The data for this example are called `plantYield.csv`.

For each treatment, we have n = 30 plants in separate pots in standarised conditions.

Can you explain why we are using the placebo? Do you know the measurement and experimental unit? As this is a 1-way ANOVA, what is the baseline hypothesis? Given that there is a control and two herbicides, are there alternative hypotheses you might test?

Let's look at the structure of the design. As hoped, we have 30 replicates of each treatment. The second view reveals that the replicates are allocated randomly among the replicate plants. There is no order to the values in the `treat` column.

```
## treat
##    Cont   Herb1   Herb2 Placebo
##      30      30      30      30
```

```
##    plots r   treat
## 1      1 1 Placebo
## 2      2 2 Placebo
## 3      3 1    Cont
## 4      4 3 Placebo
## 5      5 1   Herb2
## 6      6 2    Cont
## 7      7 1   Herb1
## 8      8 2   Herb2
## 9      9 3    Cont
## 10    10 4 Placebo
```

We have now added data to this design in order to start doing statistics. The *TRUTH* of these data are that on average, the controls have a yield of 20, Herbicide 1 increases yield by 5, Herbicide 2 by 6 and the placebo by 1 unit. These data are also quite variable. The standard deviation around the yields is large. We are going to analyse these data and

1. see if we can recover these estimates of 'known' yield.
2. test the null hypothesis.
3. test the hypothesis that herbicides, on average, increase yield.
4. test whether the two herbicides are different.
5. test whether the placebo is different from the control.

And there is the answer to one of the questions above!

```
# look at the design now.
head(plantYield)
```

```
##   plots r   treat      obs
## 1     1 1 Placebo 27.54449
## 2     2 2 Placebo 22.39018
## 3     3 1    Cont 18.23051
## 4     4 3 Placebo 19.79473
## 5     5 1   Herb2 23.46557
## 6     6 2    Cont 21.65736
```

5.1.1 The dplyr and ggplot pipeline for inference.

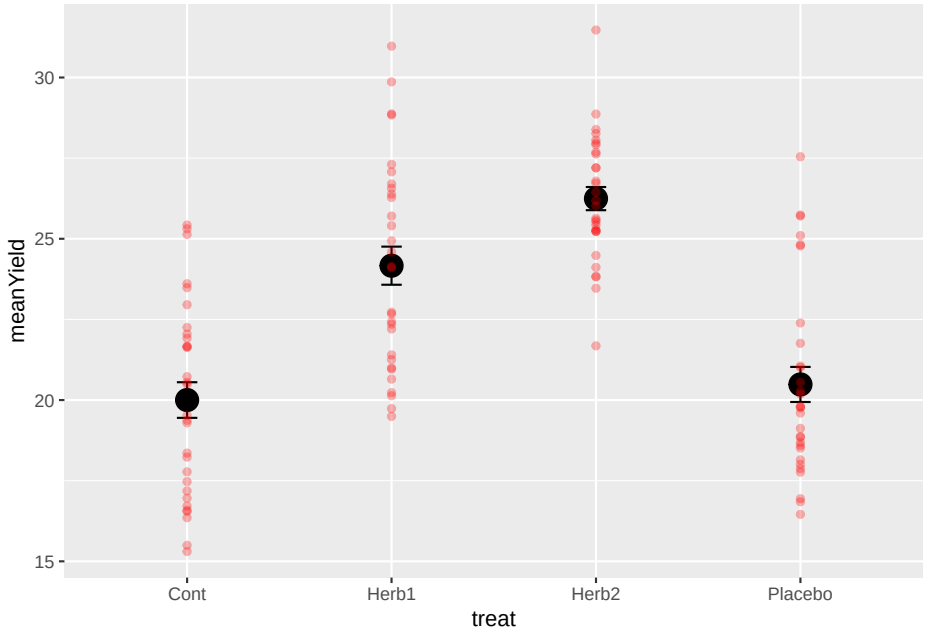
Now we can move to our standard data management and visualisation pipeline.

1. review the data (the `plantYield.csv` file contains the data)
2. sumamrise the data with dplyr - generate means and se's for the treatments
3. visualise with ggplot2

```
# check the data
# note
# obs == yield
# treat == treatment
# r = replicate (there are 30 of each treatment)

glimpse(plantYield)
```

```
## Rows: 120
## Columns: 4
## $ plots <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 1~
## $ r     <int> 1, 2, 1, 3, 1, 2, 1, 2, 3, 4, 5, 4, 6, 5, 3, 2, 6, 7, 8, 4, 5, 3~
## $ treat <chr> "Placebo", "Placebo", "Cont", "Placebo", "Herb2", "Cont", "Herb1~
## $ obs   <dbl> 27.54449, 22.39018, 18.23051, 19.79473, 23.46557, 21.65736, 26.5~
```

```
# let's force treat to be a factor.  This will make life easier later...
plantYield <- plantYield %>%
  mutate(treat = factor(treat))

# summarise to get means and ses
sumDat <- plantYield %>%
  group_by(treat) %>%
  summarise(
    # calculate the means
    meanYield = mean(obs),
    # calculate the se
    seYield = sd(obs)/sqrt(n())
  )

# plot the raw data and the mean±se
# start with the mean±se and then add the raw data
ggplot(sumDat, aes(x = treat, y = meanYield))+
  geom_point(size = 5)+
  geom_errorbar(data = sumDat, aes(ymin = meanYield - seYield, ymax = meanYield+seYield
                width = 0.1)+
  geom_point(data = design, aes(x = treat, y = obs), colour = 'red', alpha = 0.3)
```



A few things to notice.

1. The data are quite variable and the means of the herbicide treatments are roughly 5 and 6 units higher than the control. GOOD. This is as we

expected….

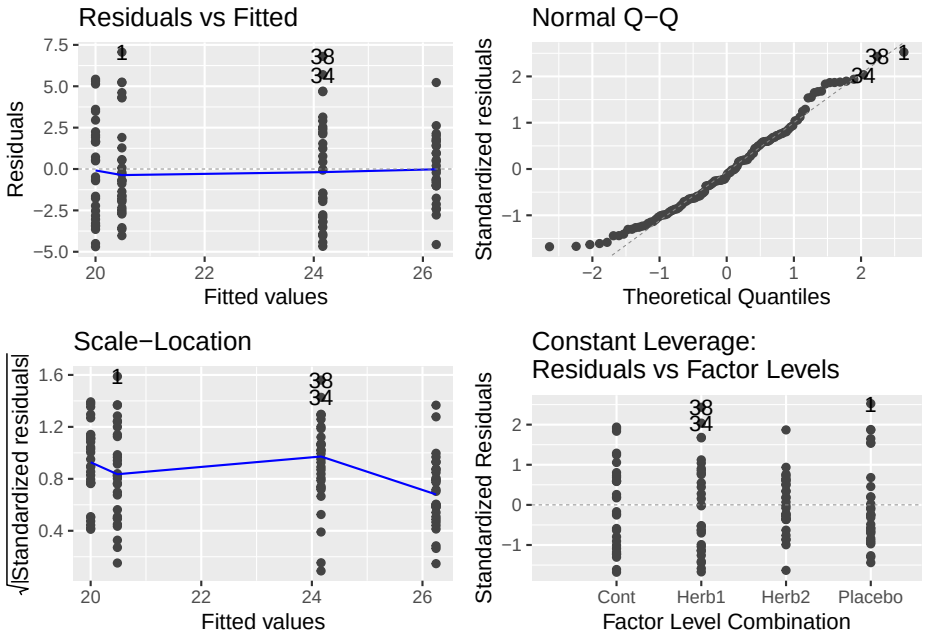
2. The standard errors are quite small, even though the variation is large! Why is that!?
3. The two herbicides don't look very different, especially given the variation around each treatment. Neither do the placebo and control. We need some stats.
4. For those of you interested in some extra reading and thinking, the 95% Confidence Interval around the means can be calculated using $1.96 * SE` ==` 1.96 * sd(obs)/sqrt(n())$. Go ahead and do that and look into that if you want…

5.1.2 The One-Way ANOVA.

If you've been paying attention, we've essentially designed and plotted the data for a 1-way ANOVA. These data are very similar to the daphnia parasite data we finished semester 1 with.

To analyse these data, we use the `lm()` function to build the model, check assumptions, and then make inference. Let's go.

```
# the model
modYield <- lm(obs ~ treat, data = plantYield)

# assumptions
autoplot(modYield)
```



```
# inference: anova
anova(modYield)
```

```
## Analysis of Variance Table
##
## Response: obs
##            Df Sum Sq Mean Sq F value    Pr(>F)
## treat       3 807.49 269.164  33.257 1.383e-15 ***
## Residuals 116 938.85   8.094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# contrasts
summary(modYield)
```

```
##
## Call:
## lm(formula = obs ~ treat, data = plantYield)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6952 -2.1542 -0.3872  1.8383  7.0609
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.0004     0.5194  38.506  < 2e-16 ***
## treatHerb1     4.1655     0.7346   5.671 1.06e-07 ***
## treatHerb2     6.2449     0.7346   8.502 7.43e-14 ***
## treatPlacebo   0.4832     0.7346   0.658    0.512
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.845 on 116 degrees of freedom
## Multiple R-squared:  0.4624, Adjusted R-squared:  0.4485
## F-statistic: 33.26 on 3 and 116 DF,  p-value: 1.383e-15
```

5.1.3 Making insight and inference

Lets walks through things very discretely.

1. Our graph suggests that herbicide treatments have an effect of increasing yield.
2. Our model is designed to test this hypothesis - are any of the differences among means non-zero?
3. Our hypothesis is probably really about whether the herbicide and placebos are different than the controls. All Hail the *treatment contrast*!
4. Our diagnostics are fantastic… the best you've ever seen.

5. The Anova Table confirms that there are differences - we can reject the null hypothesis
6. The summary table confirms that Herb1 and Herb2 are both larger than controls and the Placebo is not.

How do we interpret even more?

The estimate associated with Control is 20! Just where it should be.

The estimates associated with Herb1, Herb2 and Placebo are the differences between the mean of these treatments and the control (the reference level!). These differences are positive for Herb1 and Herb2, close to 5 and 6 respectively (as expected) and this positive difference is not 0 via the statistical test.

However, the difference for Placebo is close to 0 and therefore we can not reject the null hypothesis test that it differs from control. GENUIS!

5.2 A priori vs. Post-Hoc Contrasts

As we discussed above, there are likely several other questions we might have wanted to answer when designing this experiment. For example, are the two herbicides different in their effects?

5.2.1 Custom contrasts versus the Tukey Test

In the semester 1, we introduced how to do a Tukey Test. This is known as an *a posteriori* test – testing the significance of things suggested by the experiment, also known as data snooping or data dredging. These are multiple comparison methods (Bonferroni, Scheffe method, Tukey honest significant difference, Duncan's multiple range test) which try to control the chance of getting a significant result by chance.

To understand the risks of these, consider this experimental design. We have 7 treatments. With 7 treatments, there are 21 pairwise comparisons. With p-value threshold of 0.05 we expect 1/20 (5/100) tests to be significant. So with this 7 treatment and 21 comparison design, would you expect a signficant result by chance? You betyja.

This is why, unless a priori (in advance) you can justify ALL pairwise comparisons, a Tukey Test may not be appropriate.

Some statisticians really don't like them:

> "In my view multiple comparison methods have no place at all in the interpretation of data" -Nelder (a very very very well respected statistician).

5.2.1.1 The more appropriate approach - custom contrasts

The *more appropriate* approach is to specify *a priori* (before the experiment) a set of hypotheses you want to test, and then test them using contrasts.

For our experiment, as noted above, we were probably interested in what our treatment contrasts provided - tests of difference with the control. But we had a few others too.

Specifying specific contrasts is easy once you get your head around the 'structure' of the syntax.

Lets have a go with specifying a comparison JUST between Herbicide 1 and the control. Remember that your model is called `modYield` and your data is called `design`.

```
# check the levels and ORDERING of the treatments
# this function, levels(), tells you this
# note the ORDER: it is alphabetical, and control comes first
# the words fill in four slots c(X,X,X,X).
# we will use these slots....
levels(plantYield$treat)
```

```
## [1] "Cont"    "Herb1"   "Herb2"   "Placebo"
```

```
# define the contrast you want using -1, 1 and 0's
# this says compare control with herbicide 1.... and ignore the Herb2 and Placebo
# we give the reference -1 to the control slot
# and the reference 1 to the Herbicide 1 slot.
contrast <- c(-1,1,0,0)

# use the fit.contrast function from gmodels
fit.contrast(modYield, "treat", contrast)
```

```
##                      Estimate Std. Error t value     Pr(>|t|)
## treat c=( -1 1 0 0 ) 4.165507   0.734555 5.67079 1.058335e-07
## attr(,"class")
## [1] "fit_contrast"
```

So, this says that the difference between the control and Herbicide 1 is ~5 and that this is different from 0. Does that number `4.16` look familiar? It should. It is the same number from the `summary()` table of the full model. This is because we just specified one of the three treatment contrasts that `summary()` uses.

```
# remind ourselves of the contrast from the summary table
summary(modYield)
```

```
##
## Call:
## lm(formula = obs ~ treat, data = plantYield)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6952 -2.1542 -0.3872  1.8383  7.0609
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.0004     0.5194  38.506  < 2e-16 ***
## treatHerb1     4.1655     0.7346   5.671 1.06e-07 ***
## treatHerb2     6.2449     0.7346   8.502 7.43e-14 ***
## treatPlacebo   0.4832     0.7346   0.658    0.512
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.845 on 116 degrees of freedom
## Multiple R-squared:  0.4624, Adjusted R-squared:  0.4485
## F-statistic: 33.26 on 3 and 116 DF,  p-value: 1.383e-15
```

5.2.1.2 A different contrast - Herbicide 1 vs Herbicide 2.

If we want to compare the two herbicides we can use this approach. Note in advance that this contrast DOES NOT exist in the summary table!

```
# define the contrast you want using -1, 1 and 0's
# this says compare herb1 with herb2, ignoring the control and placebo.
# we give the slot for herbicide 1 a "-1" and the slot for herbicide 2 a "1".
contrast <- c(0,-1,1,0)

# use the fit.contrast function from gmodels
fit.contrast(modYield, "treat", contrast)
```

```
##                      Estimate Std. Error  t value    Pr(>|t|)
## treat c=( 0 -1 1 0 ) 2.079439   0.734555 2.830882 0.005473189
## attr(,"class")
## [1] "fit_contrast"
```

Isn't this cool? And quite surprising, right? We did not expect this. This says that despite the difference we created of ~1 unit of yield between Herb1 and Herb2, and even with the big variation, the statistics detect a significant difference.

Note that the difference reported is the difference between the two means that we calcuated from the sumDat calculation above!:

```
# check our summary data
sumDat
```

```
## # A tibble: 4 x 3
##   treat   meanYield seYield
```

```
##   <fct>       <dbl>   <dbl>
## 1 Cont         20.0   0.552
## 2 Herb1        24.2   0.591
## 3 Herb2        26.2   0.359
## 4 Placebo      20.5   0.544
```

Here it is: $26.2 - 24.2 = 2$

5.2.1.3 A more complex contrast: comparing the average of the herbicide effect with the control.

This might be a comparison you intended to make also… the average effect of herbicides in general. To do this, we expand the idea of -1,1 and 0's to include 1/2s (yes, 1/3's and more are possible):

```
# define the contrast you want using -1, 1 and 0's
# this says compare control with the average of herbicide 1 and 2, ignoring the placeb
# we give the control slot a -1 and the two herbicide slots a 1/2 each.
contrast <- c(-1, 1/2, 1/2, 0)

# use the fit.contrast function from gmodels
fit.contrast(modYield, "treat", contrast)
```

```
##                          Estimate Std. Error  t value     Pr(>|t|)
## treat c=( -1 0.5 0.5 0 ) 5.205226  0.6361433 8.182475 4.015904e-13
## attr(,"class")
## [1] "fit_contrast"
```

How very cool. This custom contrast delivers the inference that herbicides on average increase yield by five units.

Again, checking sumDat, we can see where this result comes from.

```
sumDat
```

```
## # A tibble: 4 x 3
##   treat   meanYield seYield
##   <fct>       <dbl>   <dbl>
## 1 Cont         20.0   0.552
## 2 Herb1        24.2   0.591
## 3 Herb2        26.2   0.359
## 4 Placebo      20.5   0.544
```

$(24.2 + 26.2)/2 = 25.2$ –> $25.2 - 20 = 5.2$

5.2.2 The Write Up using contrasts.

Fill in these blanks using the various contrasts you made above!

> We conclude that herbicides on average cause an _____ gram increase in yield (t = _____ , p = _____). We also note that there was a significant difference of _____ grams between the herbicides (t = _____ p = _____). The additional placebo treatment had no effect on yield (t = _____ p = _____).

5.2.3 Coming Back to Randomisation

We have worked here with a CRD where the measurement units are completely randomised to the experimental treatments. This simple effort is super valuable. As you've read.

> Randomisation guards against a variety of possible biases and confounding effects, including the inadvertent biases that might be introduced simply in the process of setting up an experiment…. Randomisation is a critical method for guarding against confounding effects. It is the best insurance we have against unwittingly getting some other factor working in parallel to a treatment.

But what if we know there is a gradient, or a feature of the environment or lab system that we KNOW could confound the design. Is there any way we can remove this known pattern? Yes…. of course there is.

5.3 THE RCBD - The Randomised Complete Block Design

Blocking allows us to reduce known experimental error.

A block is a group of experimental units that are homogeneous in some sense – in the same place, or measured at the same time, or by the same person. They may experience a similar temperature, or hormone concentration. They may simply be a position in the incubator where light varies from front to back.

So when constructing blocks we try and select experimental units that are homogeneous within blocks but where the blocks, and thus units within them, may be dissimilar.

Why block? When we use a completely randomised design, the location or timing of our treatment 'plots' (patches with different N or soil-moisture, incubators, locations in a 96 well plate) can generate *heterogeneity* in experimental error (variation).

This has consequences for our ability to detect effects. As the variance of the Experimental Error increases, confidence intervals get wider and the power of our analysis decreases - it's harder to detect effects of our treatments against the background noise. Ideally we would like to use experimental units that are homogeneous so the experimental error will be small. Blocking does this.

The simplest blocked design is the **Randomized Complete Block design (RCB)**

We have one complete set of treatments in each block. For the sake of example, lets imagine we identify three 'blocks' - soil moisture zones. In the design above, we would allocate 10/30 replicates of each treatment to each block.

In the first block, we randomly assign the 10 treatments to n locations in the block. We do an *independent randomization* in each block. This is the RCB design.

For example, consider the following matrix: the rows are the blocks, the letters the different treatments. In each block, each treatment is represented, but it is in a different location in the block (randomisation of the g treatments in the n units). The blocks are in a sequence - left to right - this could be different days, different locations or different positions on a hillside, for example representing an elevation or soil moisture gradient.

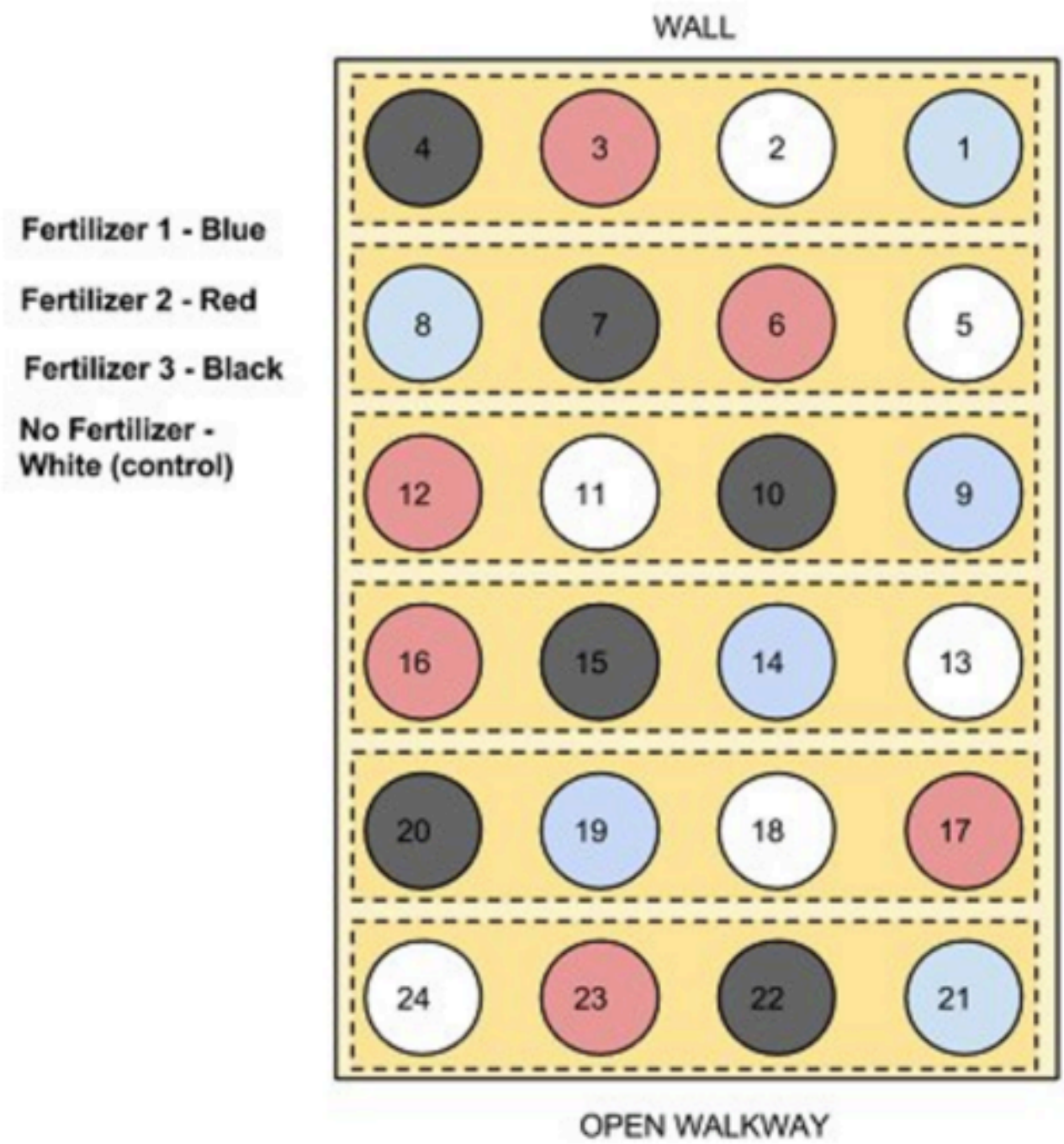
The Blocks are designed to 'capture' that underlying source of variability and allow us to detect among treatment differences more effectively.

For example, consider the following matrix: the rows are the blocks, the letters the different treatments. In each block, each treatment is represented, but it is in a different location in the block (randomisation of the g treatments in the n units). The blocks are in a sequence - left to right - this could be different days, different locations or different positions on a hillside, for example representing an elevation or soil moisture gradient.

The Blocks are designed to 'capture' that underlying source of variability and allow us to detect among treatment differences more effectively.

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,] "A"  "B"  "A"  "E"  "D"
## [2,] "C"  "A"  "D"  "C"  "A"
## [3,] "D"  "E"  "B"  "D"  "C"
## [4,] "E"  "C"  "E"  "B"  "E"
## [5,] "B"  "D"  "C"  "A"  "B"
```

Here is another picture of a block design that moves from just letters to something more literal.


WALL
4
3
2
1
8
7
6
5
12
11
10
9
16
15
14
13
20
19
18
17
24
23
22
21
OPEN WALKWAY
Fertilizer 1 - Blue
Fertilizer 2 - Red
Fertilizer 3 - Black
No Fertilizer -
White (control)

The blocks are arranged along a gradient, say along the side of a hill, so represent low and high elevation and associated soil moisture. The blocks capture this background variation. THEN, each treatment level (1-4) is allocated a random position in each block. In the end, each treatment level is replicated across blocks (n = 6!). From: https://www.researchgate.net/publication/322369242_Randomized_Block_Design_probiotic_example/figures?lo=1

It is important to note that blocks exist at the time of the randomization of treatments to units. We cannot impose blocking structure on a completely randomized design after the fact; either the randomization was blocked or it was not.

We use an RCB to increase the power and precision of an experiment by decreasing the error variance. This decrease in error variance is achieved by finding groups of units that are homogeneous (blocks) and, in effect, repeating the experiment independently in the different blocks. The RCB is an effective design when there is a single source of extraneous variation in the responses that we can identify ahead of time and use to partition the units into blocks.

In short ALWAYS block your experiment, if you can.

You can have spatial blocks, or temporal blocks where you repeat the experiment at different times, or block by batch.

In general, any source of variation that you think may influence the response and which can be identified prior to the experiment is a candidate for blocking.

5.4 An example of the RCBD

Lets modify our previous example to including blocking. If you wish to replicate the analysis, the data are `plantYield_Blocked.csv`. In these data, the means are similar to `plantYield` above, but Herbicide 1 is 10 units higher than the control and Herbicide 2 is 9 units higher. Furthermore, block 1 is supposed to be ~10 units higher than blocks 2,3,4 while block 5 is ~10 units lower.

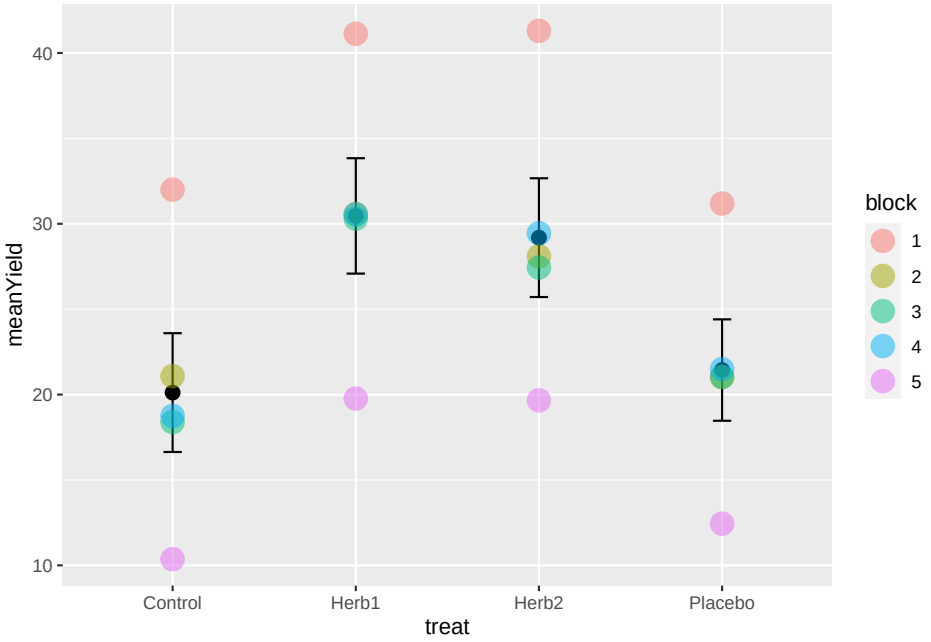
```
##    plots block   treat      obs
## 1     11     1 Placebo 31.18707
## 2     12     1 Control 31.99603
## 3     13     1   Herb2 41.29937
## 4     14     1   Herb1 41.12746
## 5     21     2 Control 21.08111
## 6     22     2   Herb1 30.56917
## 7     23     2   Herb2 28.09290
## 8     24     2 Placebo 21.02990
## 9     31     3   Herb2 27.44097
## 10    32     3   Herb1 30.29749
```

5.5 Analysing the CRBD

I'll leave it to you now to generate the following plot of the means ± standard errors from the `plantYield_Blocked.csv` file.

This requires thinking hard about the use of dplyr tools (`group_by()` and `summarise()`) and ggplot (adding more than one layer from two different sources of data - the summary data and the raw data). You need to make a *sumDat* object for the means and se's. Then you need to plot the raw data, and overlay the mean±se info from the sumDat.

Can you see the variation between block 1 and 5? Block 2-4 are all similar…. Block 1 is 10 units more, and Block 5 is 10 units less.



5.5.1 Building the model

In order to understand what's going on with blocking, and it's importance, lets build two models. This is a good trick and a good process to learn. The first model is a *naive* model that ignores block - treating this as a CRB. The second model is the *correct* model, letting block absorb the variation we can see among the blocks 1, 2-4 and 5.

```
# models
naive_model <- lm(obs ~ treat, plantYield_Block)
# note the order of these factors is important
# put block first.... so we can absorb this variation first
# the anova() table is a SEQUENTIAL table!
```

```
block_model <- lm(obs ~ block + treat, plantYield_Block)

# anova tables
anova(naive_model)
```

```
## Analysis of Variance Table
##
## Response: obs
##           Df Sum Sq Mean Sq F value  Pr(>F)
## treat      3 417.82 139.274  2.5085 0.09579 .
## Residuals 16 888.34  55.521
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(block_model)
```

```
## Analysis of Variance Table
##
## Response: obs
##           Df Sum Sq Mean Sq F value    Pr(>F)
## block      4 877.08 219.270  233.68 2.871e-11 ***
## treat      3 417.82 139.274  148.43 9.469e-10 ***
## Residuals 12  11.26   0.938
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first important thing to focus on here is the difference in the Mean Sq Residual Errors - in the `naive_model`, it is 55.52. In the `block_model`, it is 0.94. Wow…. a massive reduction in the residual error…. where has it gone?

The second important thing to notice is that having allocated variation to block in the `block_model`, and thus reducing the error variation, the *treatment* effect shifts from being insignificant to significant. At this point you should try and recall how F-tests are generated (what is the equation!) to really understand how blocking has made such a different.

5.5.2 Are the estimates of the parameters what we expect?

Lets check that the model is estimating differences as we might have expected. We can do this using the summary table.

Let's remember that, for example, the mean of Herb1 is expected to be 10 units higher than control with a yield of 20, and block 1 is supposed to be ~10 units higher than 2,3,4.

```
summary(block_model)
```

```
##
```

```
## Call:
## lm(formula = obs ~ block + treat, data = plantYield_Block)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3505 -0.7196  0.2147  0.6396  1.0719
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   31.2184     0.6126  50.957 2.14e-15 ***
## block2       -11.2092     0.6850 -16.365 1.43e-09 ***
## block3       -12.1132     0.6850 -17.685 5.84e-10 ***
## block4       -11.3415     0.6850 -16.558 1.25e-09 ***
## block5       -20.8449     0.6850 -30.433 9.94e-13 ***
## treatHerb1    10.3450     0.6126  16.886 9.96e-10 ***
## treatHerb2     9.0721     0.6126  14.808 4.50e-09 ***
## treatPlacebo   1.3192     0.6126   2.153   0.0523 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9687 on 12 degrees of freedom
## Multiple R-squared:  0.9914, Adjusted R-squared:  0.9864
## F-statistic: 197.1 on 7 and 12 DF,  p-value: 2.009e-11
```

In this table, the *INTERCEPT* is specifying the **FIRST BLOCK** and the **CONTROL TREATMENT LEVEL**. We know this because it's these words that are missing from the rest of the table, and they are each the first alphanumerially in the list of blocks and treatments. Make sure you understand this. It's tricky, but once you get it, it becomes obvious…. look for what is missing from the rest of the table!

- The value of the control, block 1 is approximately 30! Which is 20+10, which is what we expected.
- The value of Herb1 is ~10 units higher than this (remember, the value 9.84 is the DIFFERENCE between the control and treatment).
- And the value of block 5 is reported as 20 unites lower than block 1 control. This too is correct because, as above, block 1 control is 10 units higher than the control mean (20+10) and block 5 is 10 units lower….

Make sure you get this logic!

The take home message here is that these numbers from the model make complete sense with respect to the actual data. Furthermore, controlling for the among block variation *gave us more power to detect a treatment effect*, something we would have missed had we not estimated the block source of variation.

5.5.3 Correct Standard Errors for a Figure

When we made our initial plot above, we calculated the standard error based on all observations among blocks. However, the variation we really wish to represent is the variation after having controlled for the blocking effects. This means that the standard deviation we should probably use is of the error variance from the correct model: 0.94. Can you see where this comes from? The `Mean Sq` column and `Residuals` row from the `anova()` table.

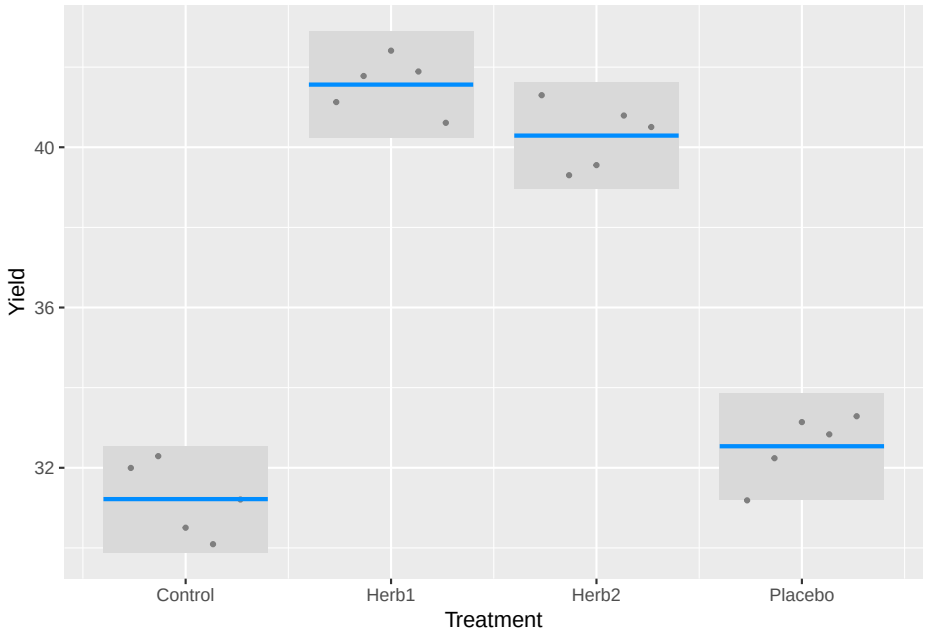
The standard deviation is the $\sqrt{Var}$ and thus, our correct standard errors from the model are $\sqrt{0.94}$

5.5.3.1 `visreg` - a helpful package for automating this.

There is a very nice plotting function in the package *visreg* that delivers these proper standard errors in a nice ggplot framework.

It presents points that are the partial residuals (deviation from the mean for each replicate), lines depicting the means, and shaded area as a 95% confidence interval, calculated as $1.96 * SE$, where *the SE is estimated from the model error variance* (just above). Compare this to your first graph.

```
visreg(block_model, "treat", gg=TRUE)+
  ylab("Yield") +
  xlab("Treatment")
```



5.5.4 Making inference in a blocked model: confidence intervals and contrasts

We are now in a very strong position to make inference.

Let's start with a rule of thumb linked to the 95% confidence interval (CI). If the CIs in the figure above don't overlap, they are different; if they do, they are not. This indicates that Cont and Placebo are not significantly different (95% confidence intervals overlap). Herb1 and Herb 2 are significantly different from these, but not each other.

This is OK. But it is not robust. Instead, let's revisit our *post-hoc* and *a priori* methods for evaluating differences among treatments. We can apply a tukey test and calculate all pairwise differences. This is not a good idea, but let's do it, using *agricolae* and the `HSD.test()` function. Living large!

```
# use agricolae HSD.test()

tukey_out <- HSD.test(block_model, "treat", group = TRUE)
tukey_out$groups
```

```
##              obs groups
## Herb1   30.46167      a
## Herb2   29.18874      a
## Placebo 21.43581      b
## Control 20.11665      b
```

This confirms our intuition and 95% Confidence Interval insights. But is it correct?

Let's make a formal test, using the `contrast()` and `fit.contrast()` functions for of one of the pairwise tests that looks obvious - between Herb1 and Herb2. Even with block in the model, the second argument for `fit.contrast()` is the treatment for which the contrast is made.

```
# fit.contrast from gmodels package
# see that even with the block in the model
contrast <- c(0,-1,1,0)
fit.contrast(block_model, "treat", contrast)
```

```
##                       Estimate Std. Error   t value   Pr(>|t|)
## treat c=( 0 -1 1 0 ) -1.272934  0.6126423 -2.077777 0.05985811
## attr(,"class")
## [1] "fit_contrast"
```

Amazing. The contrast defining a specific test provides a different answer than the post-hoc Tukey test. This is important… the Tukey Test makes lots of tests and they are penalised for so many tests. But, the contrast is the correct and most reliable result. While both *fit.contrast* and *HSD.test* both manage the model complexity and variance estimates properly, only the contrast reduces

the probability of finding a significant difference by chance or failing to find one.

Chapter 6

Designs for testing for interactions: the two-way ANOVA and factorial designs.

6.1 Introducing Interactions

In the previous sections, and the 1-way ANOVA module in week 7, we have focused on a single explanatory variable. In week 7, it was the `parasite` factor. In the previous chapters focusing on yield, it was the `herbicide` treatment. In these cases, the question we are asking is about a single, main effect. However, there are many cases where we design experiments that might have two variables. For example, we may be interesting in the effect of `parasite` on growth but also whether this effect varies by the amount of food available. Or we might be interested in the yeild of a crop as a function of the herbicide, but also of soil water content. Or we might be interested in the rate of cell division as a function of a growth hormon, but also of whether this varies by the presence of a calcium blocker. In all of these cases we are asking a question about whether the effect of one treatment varies by the level of another. Thus, our question here is about an *interaction*.

Before jumping into an example, let's introduce a simple statement that forms the core of both asking and interpreting *interactions*.

If there is an interaction between two explanatory variables, X and Z, on our response variable Y, then:

> The effect of X on Y varies by Z. **OR** The effect of X on Y depends

on Z.

This use of *varies by* **OR** *depends on* defines context dependency and that's what defines interactions.

6.1.1 An Example with CO2 and UV-B solar radiation impact on plant growth in the artic.

Context: an arctic tundra study Increasing ultraviolet-B (+UV-B) radiation from ozone depletion (the arctic ozone hole) Increasing atmospheric CO2 (+CO2) from anthropogenic emissions For plants: UV-B potentially harmful, +CO2 potentially beneficial Therefore +CO2 could alleviate UV-B damage impacts.

Hypothesis

The effect of UV-B radiation on growth will depend on levels of CO2

Predictions:

+UV-B radiation will reduce the growth of arctic plants +CO2 will increase the growth of arctic plants +UV-B radiation impacts will be less under +CO2

What is unique about this context, hypothesis, and predictions?

It's the presentation of CO2 and UVB in the same statement and the use of words like "the effect of X will alleviate the impacts of Y" and words like "the effects of X will be less under Y". These words and phrases reflect the *context dependency* of the effects of treatment levels.

Thus, to restate what we introduced above…. when we talk about *interactions*, we can rely on a very simple vocabulary that is independent of the actual treatment levels: we can always describe an interaction like this:

> The effect of X on Y depends on Z.

or

> The effect of X on Y varies by Z.

In this 'rubric', Y is the response variable, and X and Z are explanatory, independent variables. So, in our example above,

> the effect of CO2 on plant biomass yield depends on UV-B radiation levels.

OR

> the effect of CO2 on plant biomass yield varies with UV-B radiation levels.

This simple phrasing describes any interaction.

6.1.1.1 Numerical example for emphasis

Let's imagine the following situation. The folowing three numbers are what we call *Main Effects*:

Control = 20g Yield UV-B = 10g Yield C02 = 29g Yield

These numbers allow us to calculate the *Additive Effect*, where the effect of both CO2 and UV-B is simply estimated by adding the two independent effects together:

ADDITIVE RESULT: C02 + UV-B = 39g Yield

However, this additive outcome may not be what happens. Imagine if there were Synergy - the effects are more than the sum of the independent effects - or Antagonism - the effects are less than expected because one offsets the other.

SYNERGISTIC RESULT: C02 + UV-B = 60g Yield ANTAGONISTIC RESULT: C02 + UV-B = 19g Yield

6.1.1.2 The Factorial Design: Why Study Interactions?

The UV-B and C02 experiment could be thought of as two experiments – a Control vs UV-B and a Control vs C02 experiment. If we combine these we get a Factorial Experiment where we can actually estimate whether there is an interaction, and whether it is synergy or antagonism. In the factorial design for this example, there are four treatment levels and all combinations of all treatments.

- Control
- UV-B
- C02
- UV-B + C02

Some of you may be thinking that we could simply treat each of these treatment levels indepdently, as unique treatments levels - e.g. the one-way ANOVA. But we don't. We design and analyse the data we collect in a two-way analysis - a factorial design. Two-way ANOVA is the design

Factorial treatments have two main advantages over the 1-way approach.

When factors *DO interact* – so the effect of C02 depends on UV-B – then we can estimate the interaction - the dependency. *One-way designs* cannot do this, and can lead to serious misunderstandings (because we are assuming that the effect of one thing DOES NOT depend on the other).

Furthermore, when factors *DON'T interact*, *factorial designs* are more precise (smaller error variance) at estimating the main (non-interacting effects) than one-way designs experiments.

Hence ALWAYS use factorial designs when your experimental design contains the interaction (when you are asking the question that includes the word(s)

depends or *varies by*!

6.2 A Factorial Design and the Two-Way ANOVA

The following dataset `plantYield_factorial` contains two observation columns - `yield_ind` are data where there is NO INTERACTION. `yield_int` are data where there are an interaction. We will use both of these to showcase how to work the 2-way ANOVA and the analysis of a factorial design.

We note that these data have replication and randomisation. There are *four* replicate plants/plots allocated randomly to each of the four treatment combinations.

Here is the evidence of the factorial design

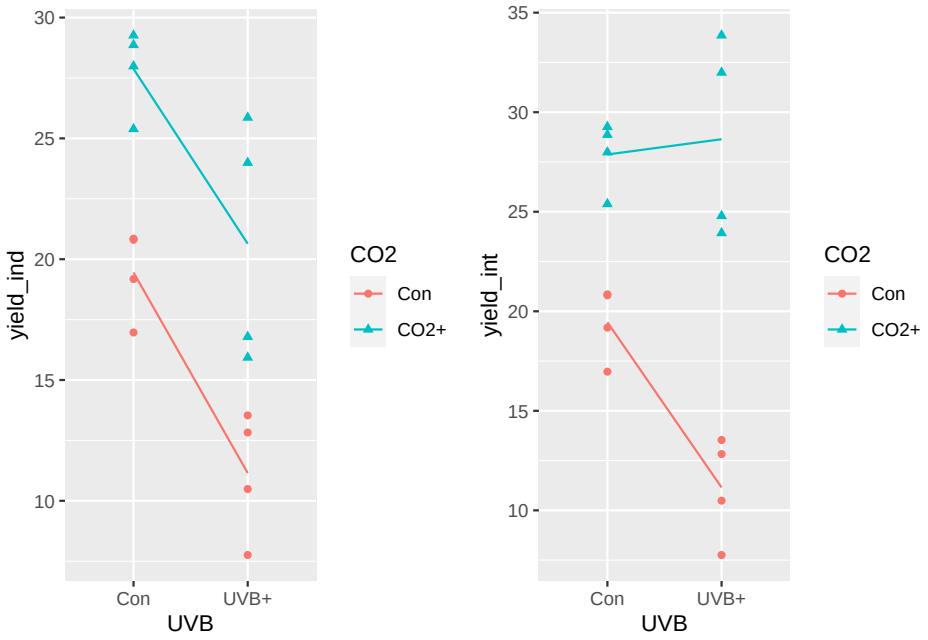
```
xtabs(~UVB+CO2, data = plantYield_factorial)
```

```
##       CO2
## UVB    Con CO2+
##   Con    4    4
##   UVB+   4    4
```

6.2.1 Plot the factorial design!

Here we combine some dplyr magic (calculating means in each group - there are two grouping variables!), some ggplot magic (adding the lines connecting the means on top of the raw data) and the beauty of patchwork, the package for plot layouts.

Our goal here is to plot the data that does not have the interaction next to a plot of the data that does have the interaction.



Given what you've read above, you should be able to fill in these blanks:

What you see above on the left, is a pattern that suggest that the effect of _____ on _____ does not vary by _____.

In contrast, on the right the pattern suggests that the effect of _____ on _____ _____ by _____:

6.2.2 How to analyse and interpret the factorial design

As with the plotting, we now analyse each data set.

We also make a mistake of analysing the data that has an interaction with a model that does not specify this interaction.

We are thus making three models. A model of the no interaction data without specifying and interaction, a model of the interaction data without specifying and interaction and a model of the interaction data with and interaction specified. This second model is incorrect, but fitting it helps illuminate why you should fit the interaction if your design contains the potential for an interaction (a factorial design).

> An Important Note on what + and * mean in models. In the below model specification, you will see CO2 + UVB and CO2 * UVB. Using the vocabulary from above, CO2 + UVB is specifying independent additive effects. In the ANOVA table produced by anova(), there will be three lines: one for CO2, one for UVB and one for residuals. In contrast, CO2 * UVB produces four lines of output in

the ANOVA table. We say that CO2 * UVB expands to include the main, independent effects of CO2 and UVB, but also the interaction between them. So CO2 * UVB == CO2 + UVB + CO2:UVB where the last term is the interaction. Thus, there are four lines reported: one for CO2, one for UVB, one for the interaction and one for residuals.

```
# A Correct model with no interaction on data with no interaction
int_mod_1 <- lm(yield_ind ~ CO2 + UVB, data = plantYield_factorial)

# A Correct model with interaction where the should be an interaction
int_mod_2 <- lm(yield_int ~ CO2 * UVB, data = plantYield_factorial)

# THE WRONG MODEL: model without interaction, on data that should use the interaction
int_mod_3 <- lm(yield_int ~ CO2 + UVB, data = plantYield_factorial)

anova(int_mod_1) # Good model, no interaction in the data and none in the model
```

```
## Analysis of Variance Table
##
## Response: yield_ind
##           Df Sum Sq Mean Sq F value    Pr(>F)
## CO2        1 321.00  321.00  35.905 4.504e-05 ***
## UVB        1 241.27  241.27  26.987 0.0001726 ***
## Residuals 13 116.22    8.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(int_mod_2) # Good model, interaction in the data and interaction in the model
```

```
## Analysis of Variance Table
##
## Response: yield_int
##           Df Sum Sq Mean Sq F value    Pr(>F)
## CO2        1 671.67  671.67 70.0311 2.364e-06 ***
## UVB        1  56.75   56.75  5.9165   0.03159 *
## CO2:UVB    1  82.15   82.15  8.5654   0.01268 *
## Residuals 12 115.09    9.59
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(int_mod_3) # Bad model, interaction in the data, but failure to specify in the m
```

```
## Analysis of Variance Table
##
## Response: yield_int
##           Df Sum Sq Mean Sq F value   Pr(>F)
## CO2        1 671.67  671.67  44.269 1.58e-05 ***
```

```
## UVB         1  56.75   56.75   3.740  0.07519 .
## Residuals 13 197.24   15.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's focus on model 2 and how we interpret this. The ANOVA table now has multiple rows. We saw this before with the block design. The important thing to note here is that the table is now read *sequentially*. We first note that CO2 explains 507.19 units of variation (Mean Sq) in plant Yield. Having captured this variation, we then note that UVB captures 216.29 additional units. Then, having capture the variation caused directly by CO2 and UVB, we now see that the interaction - asking whether the effect of CO2 varies by UVB - explains an additional 127.72 units. And… there are 11.11 units of unexplained variation.

Great. So, remember as well how we calculate the F-value. In the ANOVA (categorial variable) framework, the F-value is the ratio of variance explained by the factor relative to the residual variation. So that's where the F values come from…. And recall that BIG F-values indicate that more variation is allocated to the treatment levels, versus what's left over. The bigger the numerator value relative to the residual denominator, the more variation this term has explained.

In this experimental design, and any like it, one must remember that there is actually a single question, and it does not related to the independent main effects.It relates only to the interaction term: we designed this experiment to test the hypothesis that the effect of CO2 on yield varies by UVB. There is only a single choice of answer - yes or no. In this case, having captured variation with each independent effect, there is still a large amount of variation captured by 'allowing' the effects of each to vary by the other. Under a null hypothesis that these two variables to not interact (do not depend on each other), getting a Mean square estimate of 127.72 relative to the residual of 11.11 is very unlikely. So we reject the null!

Spend some time investigating what has happened with model 2 and model 3 and recall that ONLY ONE OF THEM IS CORRECT, regardless of whether the data look like the left or right panel above. Note the differences in the outputs. Note what we infer if we model the interaction data incorrectly.

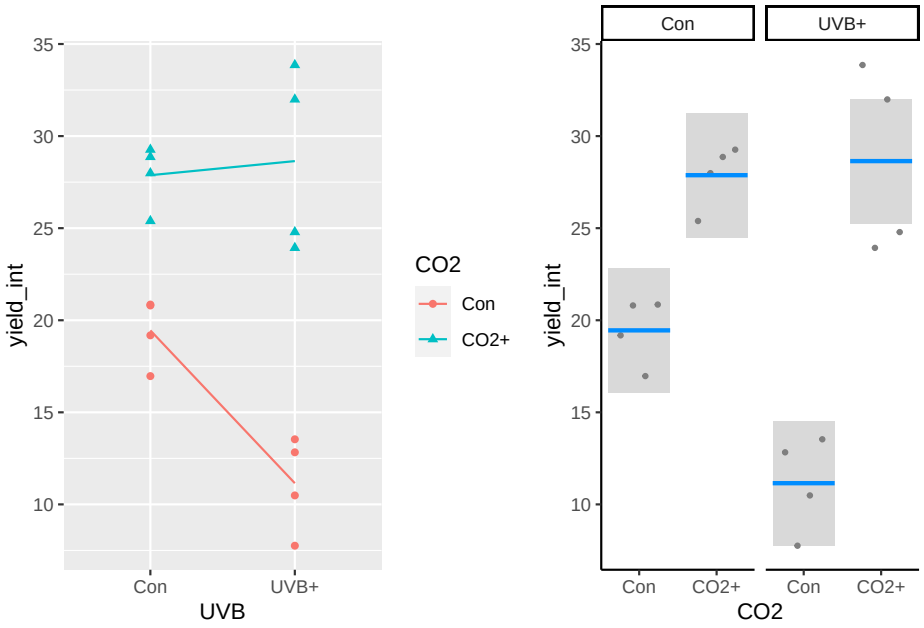
There is no free lunch.

You must understand your data and your question and you must fit the model that correctly matches your design! In this case, there really is only one model you should fit. It is model 2. EVEN if the data look like the left panel, with no interaction, this experiment was designed to test the hypothesis that the effect of CO2 varies by UVB. You can only accept or reject the null hypothesis of no interaction by fitting model 2. You can not guess the right model from the picture. The design dictates the model. But, you can guess the answer….

6.2.3 visreg and the 2-way ANOVA

Don't forget that the correct standard error for the 'result' is the residuals mean squared. You can use the dplyr + ggplot2 method, or visreg to estimate these. Here we use the `visreg` package and put the figure next to our original ggplot for the interaction data. Either of these would be OK for presentation, and most would opt for the left panel. Note the use of the the correct model for the design and question (model 2).

```
p3 <- visreg(int_mod_2,"CO2",by="UVB", gg=TRUE) +
  theme_classic()

p2+p3
```



Chapter 7

Interactions Part 2: Introducing the ANCOVA (analysis of covariance)

In the previous chapters, we introduced classic experimental design principles, and in doing so, we focused on revisiting the 1-way ANOVA and introduced the 2-way ANOVA. In these models, the explanatory variables are categorical (they are factors). This means explanatory variable has distinct, discrete categories or *levels*.

In many of our experiments, however, we might combine a categorical variable with a continuous variable. For example, we might estimate eggs produced as a function of body size (continuous) and season (categorical). Or we might estimate the effect of a mutagen (continuous) on tumour formation in wild type vs. genetically modified (categorical) fruit flies.

These simple examples describe an ANCOVA, where the one explanatory variable is continuous (e.g. body size or mutagen) and the other is categorical (e.g. season or fly-type). When this is the case, we are essentially combining the ANOVA with the Regression model! If we recall that regression is largely about estimating slopes and intercepts, we might think, hey, COOL, so in an ANCOVA, we can ask if the categorical variable alters the slope or not….. and that'd be correct.

What is very important to understand is that our core statement about interactions does not change.

> The effect of X on Y varies by Z, translates to a) The effect of body size on egg production varies by season; or b) the effect of mutagen on tumour formation varies by fly-type.

When written like this, the previous statements about slopes should be even more clear. The effect of body size (continuous) on egg production is a regression and we estimate a slope. The effect of the mutagen concentration on tumour formation is a regression and we estimate a slope. We then can ask whether this slope is different between seasons or fly-type.

7.1 What this chapter covers.

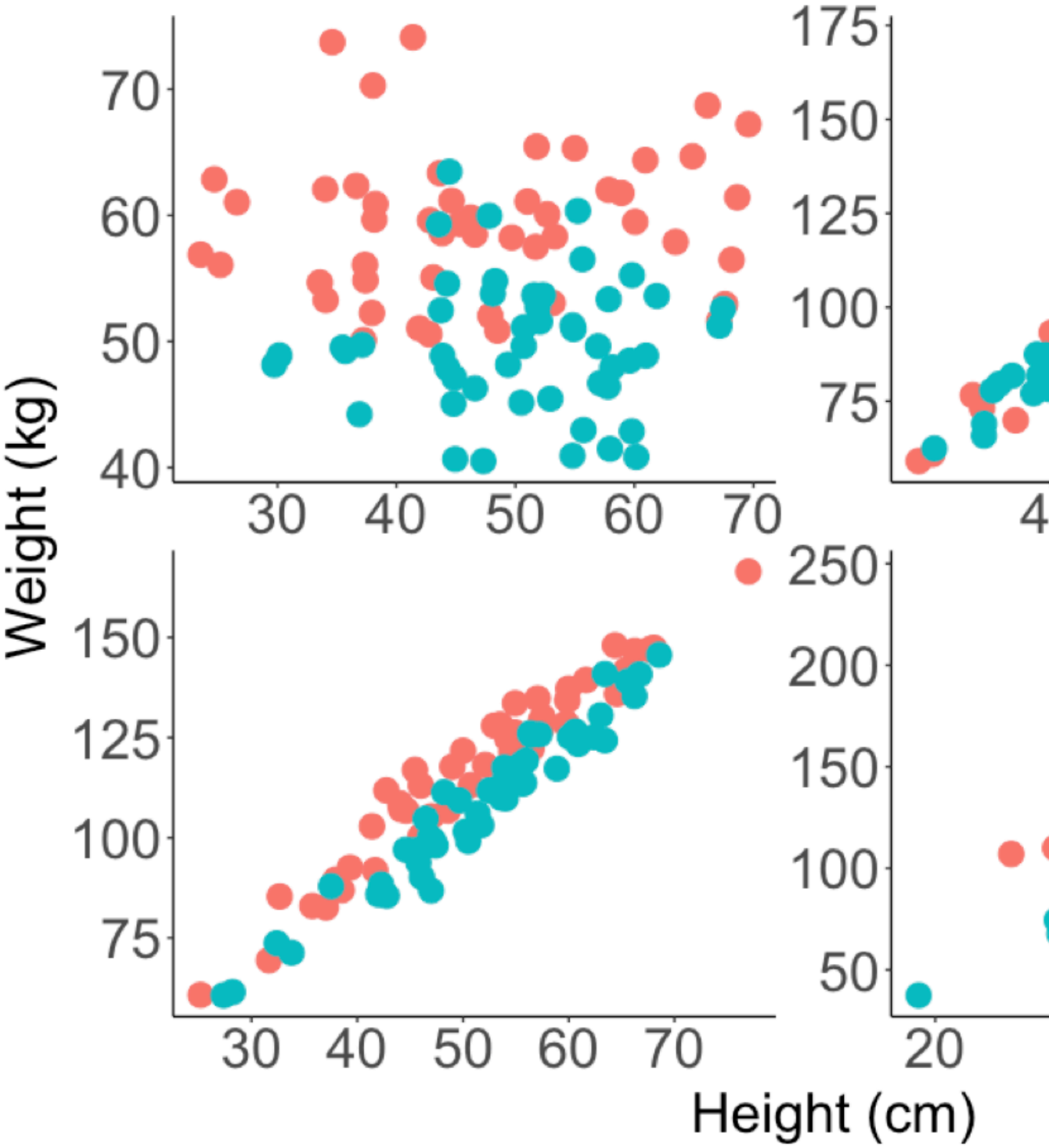
This chapter will focus on the ANCOVA, but it also touches on additional concepts. The primary foci are thus:

1. The ANCOVA model
2. transformations
3. plotting model results when you've made transformations.

7.2 Setting up the various ideas.

Let's start by looking at an example where the effect of Height on Weight varies by Sex. This is a classic set of data from numerous organisms…. it captures biologically the question about sexual dimorphism - does the relationship between Height and Weight (a regression with the continuous variable Weight) vary by Sex (the categorical variable [e.g. ANOVA], in this case M/F but it can be more depending on organism)?

This relationship can take on many patterns.


Weight (kg)
70
60
50
40
30
40
50
60
70
175
150
125
100
75
150
125
100
75
30
40
50
60
70
250
200
150
100
50
20
Height (cm)

- In the upper left we see a pattern where Males are heavier than Females, but there is no effect of Height on Weight.
- In the upper right, we see that there is a positive relationship between Weight and Height, but no differences between Males and Females.
- In the lower left, we might argue that there is a positive relationship between Weight and Height, that the relationship between Weight and Height does not vary (the slope is the same) and that Males are heavier (the red dots are mostly above the blue)
- In the lower right, we see evidence of an interaction - the effect of Height on Weight (the slopes), varies by Sex.

It's quite important to recognise that each of these patterns (predictions) is possible outcome to testing the SAME null hypothesis. The key thing to remember is that we have an *a priori* (in advance hypothesis) that there is no difference between the slopes for males and females. Regardless of the pattern, we should specify the appropriate model to test your the hypothesis and answer the question that motivated the design of the experiment and data collection.

Let's be super clear: in these data we start with the fundamental question that uses the same vocabulary that we used in the previous chapter - does the effect of Height on Weight vary by Sex. The data might look like any one of the above patterns. BUT, there is only one model syntax in the above figure that tests the null hypothesis and gives us a chance to evaluate the predictions: `lm(Weight ~ Height * Sex, data = data.frame.name)`.

> If you need a refresher on what the * symbol means, pop back to chapter 5 where we explain this when we fit the models; we also review this below.

7.2.1 Anticipating the ANOVA table for an ANCOVA model

Let us also recall the sequential nature of the ANOVA table output. This is about what we expect to see, having fit a model, which we will do below. In contrast to the ANOVA, with an ANCOVA, we will see first an estimate of variation explained by the continuous variable, then the categorical variable and finally, after seeing the report on both of those, an estimate of the variation attributable to the interaction. Another way to think of this is as follows (explore the figures above along with this):

What is the estimate of the intercept and slope, ignoring the different categories, and how much variation does this explain. What happens if we let there be two lines, one for each category, but with the same slope (e.g. different intercepts, but same slope). Does this explain more variation? Finally, what if we let there be two lines, and different slopes (the interaction exists) - does this explain additional variation?

If we get to the last scenario, and the answer is yes, after all the other options,

the *different slopes* does capture variation in our response, we have evidence of an interaction.

7.3 Working through an ANCOVA example.

Let's work with a built in dataset in R - the Davis Study, which is exactly these data. The associated assignment is another example.

The process of doing this will follow a tried and true approach to analysing data. You should have this workflow fully embedded in your head now:

1) PLOT the Data
2) Build the model to test the hypothesis
3) Check the Assumptions
4) Make Inference
5) Update the Figure for Publication

7.3.1 Organise the packages and get the data and make your picture

The data we are using is an example dataset built into R, but embedded in the `carData` package, which is installed with the `car` package - download and install this and they you can use the following code (don't forget the `tidyverse`, `agricolae`, `ggfortify`, `visreg` etc).

These data ARE NOT on the blackboard site.

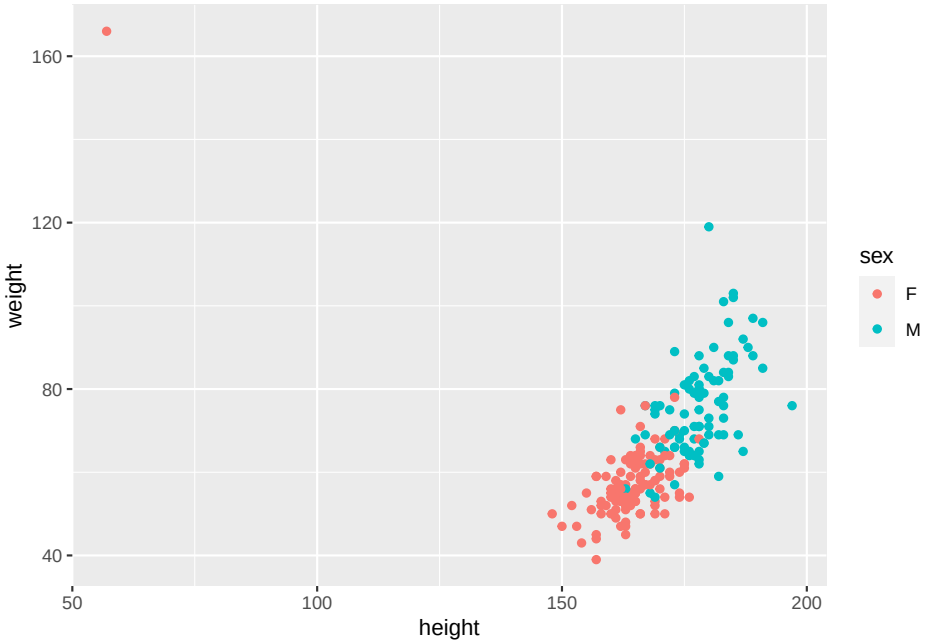
Note also that if you are building your own script, you need the following packages: `tidyverse`, `ggfortify`, `visreg`, `patchwork`

```
library(tidyverse)
library(ggfortify)
library(visreg)
library(patchwork)

# this creates a working version of the davis data for you.
# the carData:: syntax is a way to use this package without using
# the library() function.
Davis <- carData::Davis

# check it out
glimpse(Davis)
```

```
## Rows: 200
## Columns: 5
## $ sex    <fct> M, F, F, M, F, M, M, M, M, M, M, F, F, F, F, F, M, F, M, F, M, ~
## $ weight <int> 77, 58, 53, 68, 59, 76, 76, 69, 71, 65, 70, 166, 51, 64, 52, 65~
## $ height <int> 182, 161, 161, 177, 157, 170, 167, 186, 178, 171, 175, 57, 161,~
```

```
## $ repwt  <int> 77, 51, 54, 70, 59, 76, 77, 73, 71, 64, 75, 56, 52, 64, 57, 66,~
## $ repht  <int> 180, 159, 158, 175, 155, 165, 165, 180, 175, 170, 174, 163, 158~
```

```
# make the most basic of exploratory plots.
ggplot(Davis, aes(x = height, y = weight, col = sex))+
  geom_point()
```



Whoa!... what's going on here? Well,here is a massive justification of why we say *"always plot your data"* before you do anything. It looks like one of the data points has the height and weight data entered incorrectly, the wrong way around. This is a 'feature' of this dataset. Of course you could go back to your master spreadsheet and make the corretion, which is a good idea in real life, but let's see how to fix this using R. Let's fix that. Lets find the row with the mistake by looking for height values <100, and then make the switch.

```
#which row? 12
Davis %>% filter(height < 100)
```

```
##    sex weight height repwt repht
## 12   F    166     57    56   163
```

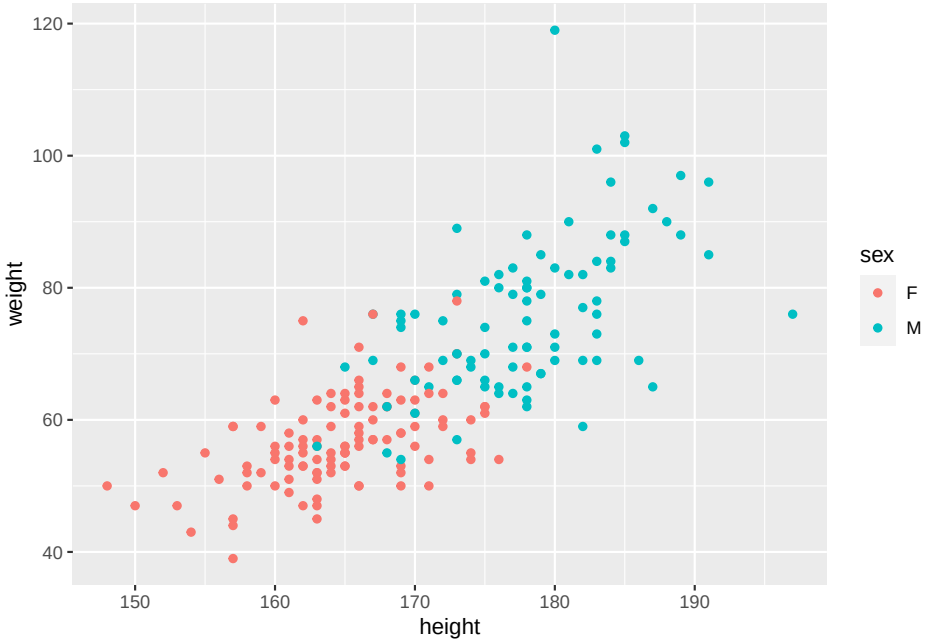
```
# BASE R syntax to see each observation
Davis$weight[12]
```

```
## [1] 166
```

```
Davis$height[12]
```

```
## [1] 57
```

```
# BASE R syntax to change the specific values in the specific row
Davis$weight[12] <- 57
Davis$height[12] <- 166

# replot
ggplot(Davis, aes(x = height, y = weight, col = sex))+
  geom_point()
```



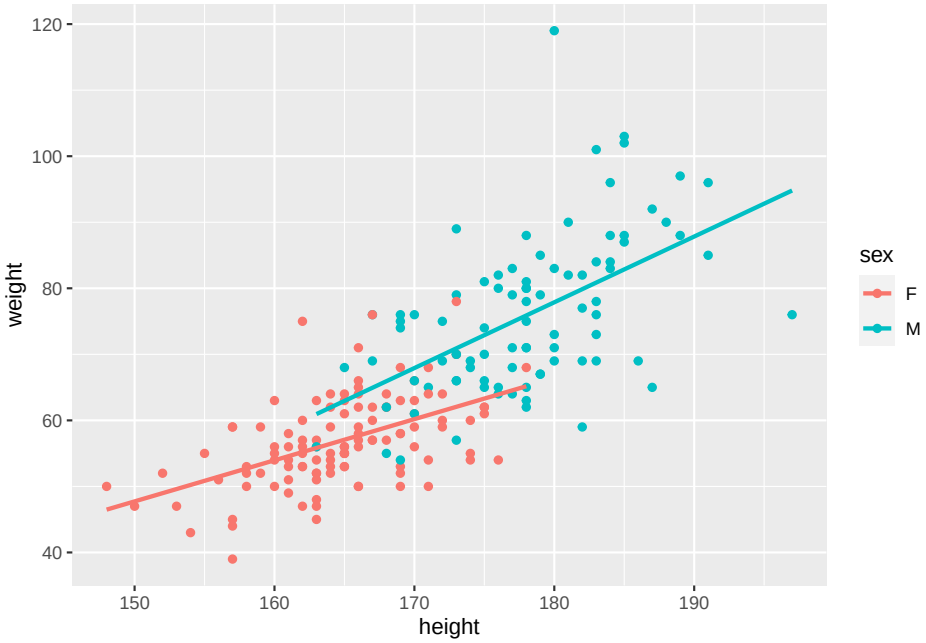
Excellent. Let's look at this picture and ask ourselves the null hypothesis question: does the effect of Height on Weight vary by Sex? What do you think?

We can actually get a bit of help here. We can use some ggplot magic to help us make this GUESS. Let's use `geom_smooth()` to help us guess the answer.

> NOTE: this is not doing statistics. This is using graphics to help guide insight and our expectation of the outcome before doing the statistical test of our hypothesis.

```
ggplot(Davis, aes(x = height, y = weight, col = sex))+
  geom_point()+
  # add a best fit line to each group (the sex category)
  geom_smooth(method = lm, se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Again, we have not 'proven' anything or tested our hypothesis. What we have is a good guess though. We can guess that

1. Males are heavier than females (what is the difference in the intercepts?)
2. The effect of height on weight is positive (the slope(s) are positive, not negative)
3. There *might* be a difference in slopes - the Male line looks steeper.

We might even go as far to estimate by eye the overall slope, assuming that there is no effect of Sex. Recalling that the slope is the rise over run or the change in y over the change in x, we can guess that the slope is $\sim (100-40)/(200-140) = (60/60) = 1$. Can you guess a slope for each Sex?

7.4 Building the model (and understanding it)

The next step is to build the model to test our hypothesis. As we declared above, the model to test the interaction is `lm(weight ~ height * sex, data = Davis)`. Let's discuss this a bit.

First, as described above, and in Chapter 5, this model with the `*` expands to the following full model of two main effects and an interaction:

```
lm(weight ~ height + sex + height:sex, data = Davis)
```

This reads as "weight is a function of the effect of height (slope), sex (intercept) and the interaction between height and sex (do the slopes vary?). The `height`

`* aex` syntax always expands to this *full model* - the model containing the main effects and the interaction.

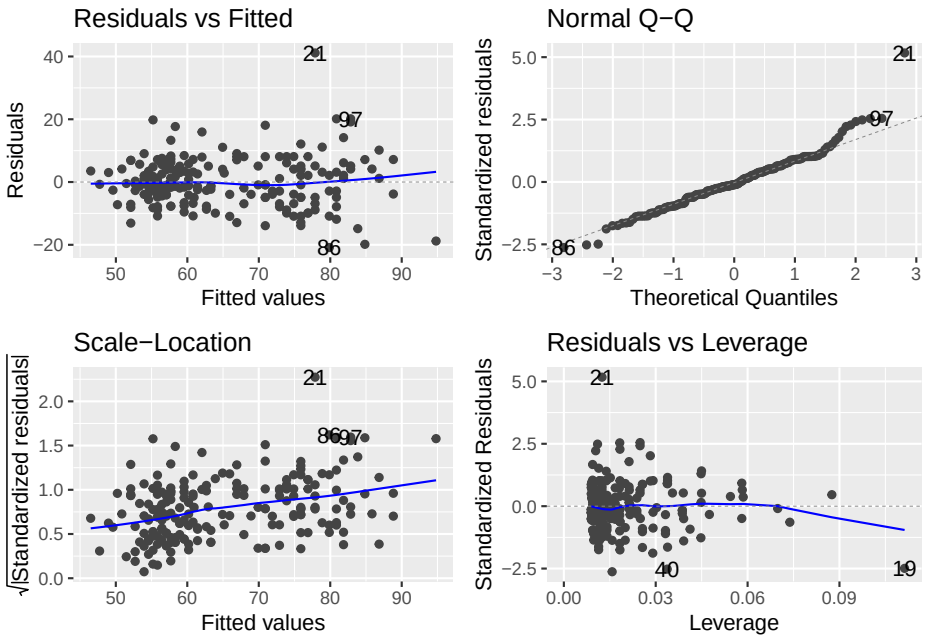
OK. Let's fit the model

```
# we call the model mod_Davis.
mod_Davis <- lm(weight ~ height*sex, data = Davis)
```

That's it. We've got the model. But before we make any attempts at actually evaluating our test of the hypothesis, we have to check the assumptions!

To do this, we use the `autoplot` function from the `ggfortify` package.

```
autoplot(mod_Davis)
```



OK. Let's walk through the three core diagnostics.

In the upper left, we are evaluating the systematic part of the model - are there any systematic departures from the linear model we've speficied? Are there interactions missing or specific non-linearities? Nope.

In the upper right, we are evaluating the normality of the residuals. These too look pretty good. If you notice, event though there are some deviations, but the pattern is that the deviations move away from the line but 'come back' at the very ends. This is typical of normally distributed residuals.

In the lower left, we are evaluating the assumption of a constant mean-variance relationship. Oops. Remember, we are expecting something that does not show a trend at all. Even without the panic lines, this looks like the variance (y-axis)

is increasing with the mean (x-axis).

7.4.1 Dealing with the mean-variance breakdown.

The above issue is one we can deal with via a *transformation*.

Let's start with the facts.... When the *variance increases with the mean*, a transformation that can work is the *logarithm of the continuous variables*.

There is a short section in the (APS 240 reading)[https://dzchilds.github.io/stats-for-bio/data-transformations.html#trans-types] to explore. Please do read this summary about transformations, and bookmark it for the future. It is very handy.

A question that routinely comes up is whether transformations are 'changing the data'. Rather than focusing on this, we offer the following interpretation. Linear models fit by `lm()` carry a set of assumptions (our focus on the three panels using autoplot). If these are not met, we can not trust our inference (F-testing).

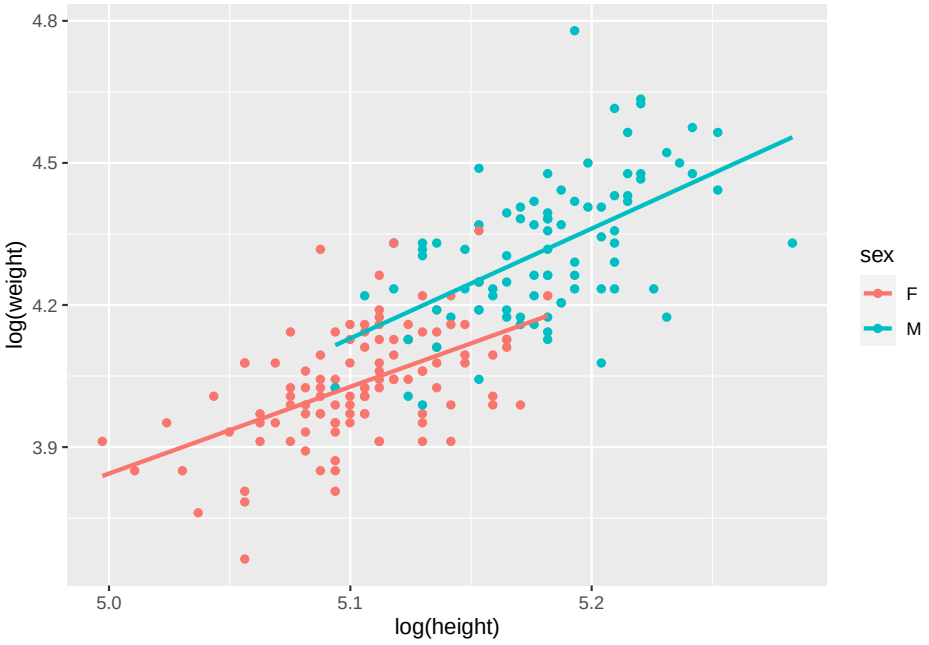
> Most biological data are unlikely to conform perfectly to all the assumptions, but experience has shown (fortunately) that t-tests, ANOVAs and regressions are generally quite robust—they perform reasonably well with data that deviate to some extent from the assumptions of the tests. However, in some cases residuals are clearly very far from normal, or variances change a lot across groups - from (APS 240)[https://dzchilds.github.io/stats-for-bio/data-transformations.html#transforms-introduction].

But if there are real problems, we can use transformations. They do not 'change the data' instead, they retain the relationships between the data points, but put everything on a difference scale so that we can use the power of the linear model to answer the question. The tricky thing is how do we report our insights to the reader back on the original 'response' scale. We'll get to that. But rest assured, transformations don't change the data.

So, here are the data where we transform both continuous variables - the continuous response variable and the continuous explanatory variable.

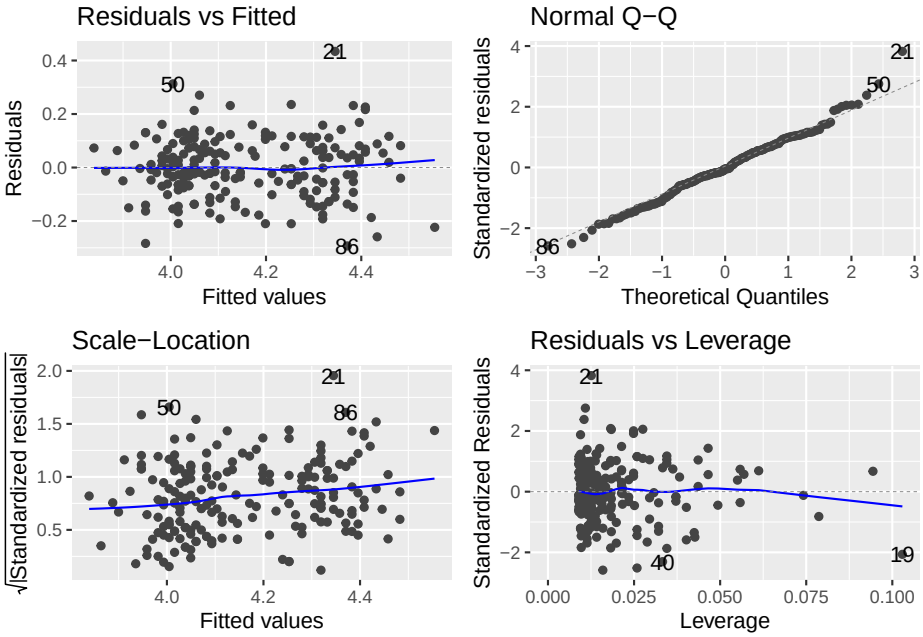
```
# re-plot with log-log
ggplot(data = Davis, aes(x = log(height), y = log(weight), col = sex))+
  geom_point()+
  geom_smooth(method = lm, se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



OK. The patterns are still there. And let's see what happens with the model and assumptions now!

```
# fit the model - note we can specify the transformation RIGHT IN THE MODEL
mod_Davis_log <- lm(log(weight) ~ log(height) * sex, data = Davis)

# check the new residuals
autoplot(mod_Davis_log)
```



The scale-location plot is definitely flatter - you can tell by the range on the y-axis of the scale-location plot. So, with this transformation, we can now move to evaluating the model. You are probably thinking about what it means to analyse the data on the log-log axis…. we will come to that shortly.

7.4.2 Making inference on the ANCOVA

The next step is to use the detail in the ANOVA table - yes, we use the same `anova()` function to build an anova table to explore the results of an ANCOVA.

Let's revisit our initial guesses and work through what the model is telling us:

1. Males are heavier than females (what is the difference in the intercepts?)
2. The effect of height on weight is positive (the slope(s) are positive, not negative)
3. There might be a difference in slopes - the Male line is steeper.

Together, the graph and these guesses lead to a feeling that *the effect of height on weight varies by sex*.

```
anova(mod_Davis_log)
```

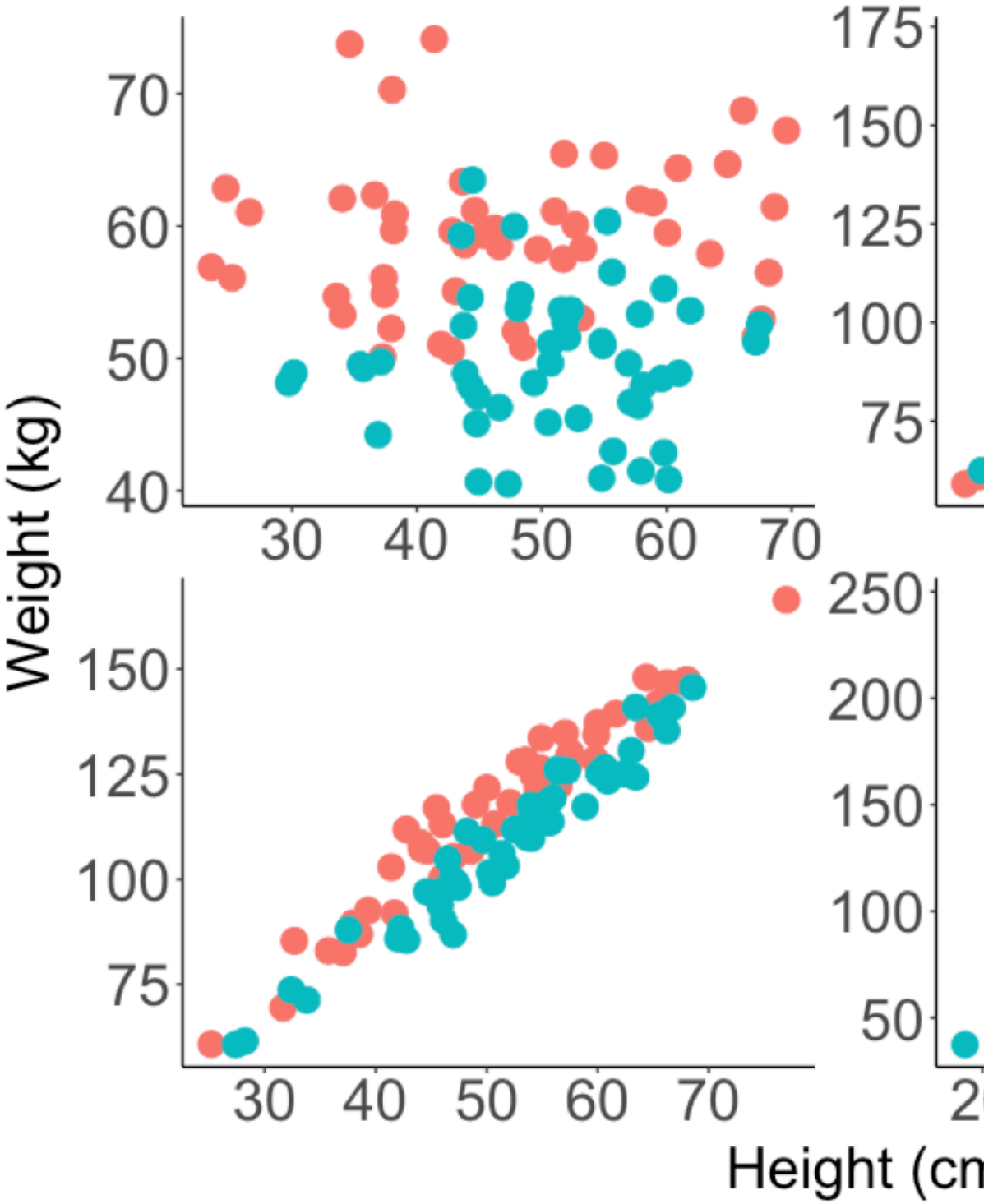
```
## Analysis of Variance Table
##
## Response: log(weight)
##                  Df Sum Sq Mean Sq  F value    Pr(>F)
## log(height)       1 4.6643  4.6643 357.4735 < 2.2e-16 ***
## sex               1 0.3446  0.3446  26.4115 6.647e-07 ***
```

```
## log(height):sex   1 0.0144  0.0144   1.1038    0.2947
## Residuals       196 2.5574  0.0130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Howwever…. The table tells us that there is *no evidence for an interaction*. The last line of this table contains the p-value for the interaction. It is testing the null hypothesis that the slopes are the same. We can not reject this null hypothesis. There is no evidence that allowing the slopes for Males and Females be different is supported. That's the answer to guess 3 above. It is also the answer to our question: does the effect of Height on Weight vary by Sex. *Nope.*

However, we do detect *main effects* of height and sex. This means that each of these DOES have an effect - but we say the effects are additive.

Great. What this means is that the data, despite the pattern our eye sees, can not be distinguished from the pattern in the picture in the lower left.


Weight (kg)
70
60
50
40
30
40
50
60
70
175
150
125
100
75
150
125
100
75
30
40
50
60
70
250
200
150
100
50
Height (cm

There is a slope associated with Height (we don't know whether it's positive or not, but we think it is), and there is an effect of sex (we don't know whether Males are heavier yet). What we can say, with confidence, is this.

> The effect of height on weight did not vary by sex (F = 1.01, df = 1,196, p = 0.29). The effect of sex (F, df, p) and height (F, df, p) are thus additive on Weight.

Let's work sequentially to read the table now:

When we allow for a common slope associated with `log(Height)`, we capture 4.66 units of variation in `log(weight)`, which when compared to the 0.013 units of variation left over (residual MSE), is a lot. The F-value is thus BIG and the p-value small. When we allow for an effect of Sex to define different average `log(weight)`, this too is important as it captures 0.344 units of variation, which compared to the 0.013 units in the residual, is also big (Big F, small P).

However, when we allow for the slopes caused by `log(height)` to vary with `sex`, this captures very little additional variation with respect to the residual variation (0.014 comprared to 0.013) hence a small F and a large P for this term.

7.4.3 The summary table, the actual slopes and making sense of the log stuff.

Let's start investigating the details via the `summary` table:

```
summary(mod_Davis_log)
```

```
##
## Call:
## lm(formula = log(weight) ~ log(height) * sex, data = Davis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29310 -0.06543 -0.00592  0.07420  0.43410
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5.3202     1.6016  -3.322  0.00107 **
## log(height)        1.8329     0.3138   5.841 2.13e-08 ***
## sexM              -2.3724     2.3765  -0.998  0.31936
## log(height):sexM   0.4852     0.4618   1.051  0.29473
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1142 on 196 degrees of freedom
## Multiple R-squared:  0.6626, Adjusted R-squared:  0.6575
## F-statistic: 128.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

This table reports on the *treatment contrasts* we introduced earlier this semester. Let's walk through the details.

The first two rows correspond to one of the sexes. Can you identify which one? The hint is looking at the other rows. They have an M after them. M comes after F (F is first in the alphabet). So the first two rows correspond to *the intercept and slope of the Females*.

You may be asking why the intercept is a negative number. A simple look back at the range of the x-axis should answer this… the y intercept occurs at x = 0, right?

You should understand, from the principles of an equation for a line, the following:

> the effect of height on weight for females is `log(weightF) = -5.32 + 1.82*log(heightF)`.

OK… let's interpret what the next two lines mean. The third line is labelled `SexM`. The fourth is labelled `logHeight:SexM`. Any guesses as to which is associated with the slope and which the intercept?

One trick, again, is to recognise the syntax -> the presence of the `:` is indicative of the interaction which is the thing that lets slopes vary. So, `SexM` is about intercepts and `logHeight:SexM` is about slopes. But what are they?

Recall from previous work that *treatment contrasts are about DIFFERENCES*. Thus, the `sexM` term is a difference between the female and male intercept! The `logHeight:SexM` is about the difference between the female and male slopes!

This allows the following maths to be worked out ….

```
log(weightF) = -5.32 + 1.82*log(heightF)
```

```
log(weightM) = (-5.32-2.37) + (1.82+0.48)*(log(heightF))
```

However, we know that the 0.48 increase in the slope is not signficant, and as a result, the -2.37 increase in the intercept moving from Female to Male is also not a significant change

7.4.4 Specifying the equations supported by the model.

In order to identify the equations supported by the model - e.g. the equations associated with the estimation of teo lines with different intercepts and the same slope, we can refit the model to reflect the ADDITIVE effects only, replacing the `*` with the `+`.

```
mod_Davis_log2 <- lm(log(weight) ~ log(height) + sex, data = Davis)
summary(mod_Davis_log2)
```

```
##
## Call:
```

```
## lm(formula = log(weight) ~ log(height) + sex, data = Davis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28714 -0.06906 -0.00978  0.07615  0.43717
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.46349    1.17541  -5.499 1.18e-07 ***
## log(height)  2.05688    0.23030   8.931 2.99e-16 ***
## sexM         0.12417    0.02417   5.138 6.66e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1143 on 197 degrees of freedom
## Multiple R-squared:  0.6607, Adjusted R-squared:  0.6573
## F-statistic: 191.8 on 2 and 197 DF,  p-value: < 2.2e-16
```

Right, now we get our expected pattern where the intercepts are different but the slope is the same:

```
log(weightF) = -6.46 + 2.05*log(heightF)
```

```
log(weightM) = (-6.46 + 0.12) + 2.05*log(heightF)
```

We can see the common slope and the slightly higher average weight (0.12 units!) of males.

7.4.5 Making even more inference and drawing the picture

OK, lets use a variation on our plotting tricks from the regression module (week 6) to get a nice picture.

```
# first create a data frame to use that combines the raw data
# with the predictions from the model.

# note that the model is the ADDITIVE model, for plotting
# step 1: make a copy of the raw data
# step 2: use mutate to add the fitted values (at each value of height (x-axis) where we have mea
plotThis <- Davis %>%
  # ... a column of numbers of predicted weight using the model
  mutate(predWeight = predict(mod_Davis_log2))

# check it out.
head(plotThis)
```

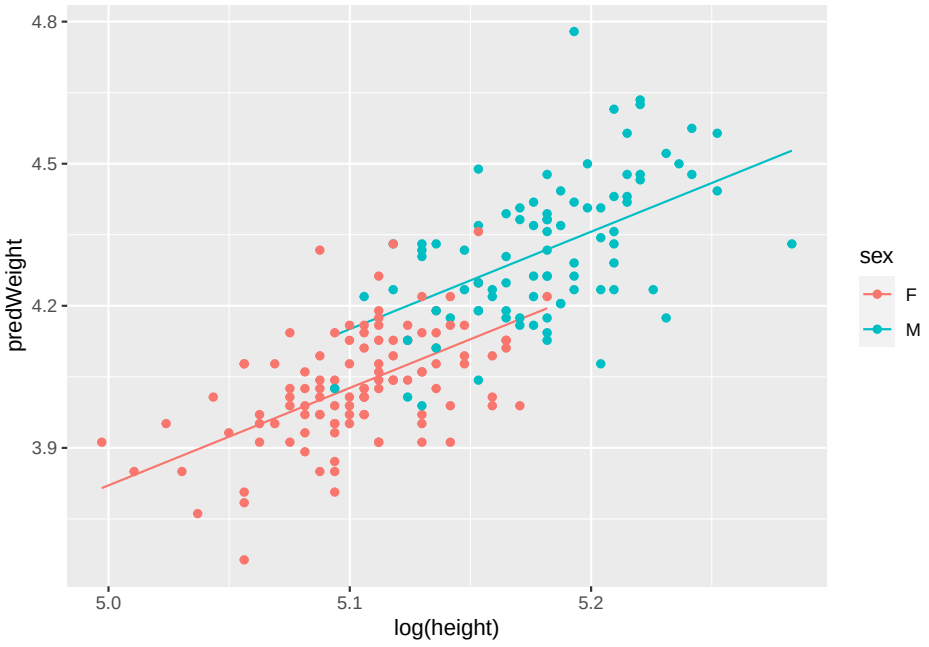
```
##   sex weight height repwt repht predWeight
## 1   M     77    182    77   180   4.364677
```

```
## 2   F     58    161    51    159    3.988326
## 3   F     53    161    54    158    3.988326
## 4   M     68    177    70    175    4.307379
## 5   F     59    157    59    155    3.936578
## 6   M     76    170    76    165    4.224381
```

You can see we have a fourth column now.... these are predictions (also known as fitted values) of `log(weight)` as a function of `log(height)` using the additive model, *at each of the originally measured height values in the data frame.*

We can actually use this data frame to generate the picture that represents the output of the statistical model.

```
# Step 3: make a figure with lines from the model predictions
# Step 4: add the raw data points from the original dataset
# NOTE! pay attention to where we are using log()!
ggplot(plotThis, aes(x = log(height), y = predWeight, col = sex))+
  geom_line()+
  # now add the raw data note
  geom_point(aes(x = log(height), y = log(weight)))
```



Great. Now we can see that the full model (mod_Davis_log) told us. The lines are parallel. There is no interaction. In fact, we can do a bit more here. Remember that the model says

For women $log(weight) = -6.46 + 2.05 * log(height)$ For men $log(weight) = (-6.46 + 0.12) + 2.05 * log(height)$

Because we are in log land, we can do some clever maths: We know that

$log(weightFemale) - log(weightMale) = -0.12$

(this is the treatment contrast for the intercept). We also know (please memorise) that the *difference between logs is equal to the log of the ratio*:

$log(weightF/weightM) = -0.12$

Finally, we can convert both sides to non-log by exponentiating:

$exp(log(weightF/weightM)) = exp(-0.12) = 0.88$

What does this mean? At any given height, female weight is 88% of a males (e.g. 12% lower). Cool! That's one of the benefits of log-log linear relationships. *The ratio of the categories in log-log land directly translates to a percent difference between the groups.* It's also worth noting that the height slope is ~2, which suggests… yes… because we are in log-land, that weight scales with $height^2$.

7.5 Some General Principles for ANOVA and ANCOVA modelling.

7.5.1 ANCOVA (and 2-way ANOVA) is always comparing the interaction model against the additive model!

This figure highlights the various `lm()` models and their interpretations/assumptions associated with the pieces of an ANCOVA dataset. This is a way to envision all of the potential linear relationships among the variables that might arise in an experiment with a continuous and categorical explanatory variable.

> Note that just because your data looks like a particular pattern doesn't mean you should use that model. Your statistical model should be driven by your question and original design.

For example, if you design a study with an ANCOVA question in mind, you are always starting your analysis with the question of whether the effect of X on Y varies by Z. Thus, ANCOVA always starts with comparing the interaction (model E) against the additive model (model D).

> Proof: If we compare the additive model and the model with the interaction, we are *specifically* testing for whether allowing the slopes to vary by sex *explains additional variation* above and beyond the main, additive effects of height and sex. To do this, we use the `anova()` function, but pass it two models instead of one! This does what is called a *Liklihood Ratio Test* comparing the two models. Remembering that the difference between the `*` model and the `+` model is just the presence/absence of the interaction term, this `anova()` comparison evaluates how much more variation is explained by the

	VERBAL HYPOTHESIS	INTERCEPT/ SLOPE	MODE
A	The number of eggs produced by limpets does not vary with density or season	Common intercept Zero slope(s)	lm(Eggs~1,
B	The number of eggs produced by limpets does not vary with density but is reduced in the summer season (dashed)	Different intercepts Zero slope(s) Parallel horizontal line	lm(Eggs~Seas
C	The number of eggs produced declines with density, but the maximum number of eggs (intercept) and the rate (slope) do not vary with season	Same intercept Same slope Same lines	lm(Eggs~Dens
D	The number of eggs produced declines with density and the number of eggs at density zero (intercept) differs between seasons but the rate (slope) does not vary with season	Different intercepts Same (negative) slope Parallel lines	lm(Eggs~Den data=
E	The number of eggs at density zero (intercept) and the rate (slope) vary with season	Different intercepts Different slopes 'Crossing' lines	lm(Eggs~Den data=

Figure 7.1: From Getting Started with R, 2nd Edition, OUP

interaction, after the additive effects have been modelled. In the follow code, compare the result of the first test to the *last line* of the second. This proves that when we fit model E, the test for the interaction is actually a comparison of model E with model D!

```
anova(mod_Davis_log, mod_Davis_log2)
```

```
## Analysis of Variance Table
##
## Model 1: log(weight) ~ log(height) * sex
## Model 2: log(weight) ~ log(height) + sex
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    196 2.5574
## 2    197 2.5718 -1 -0.014402 1.1038 0.2947
```

```
anova(mod_Davis_log)
```

```
## Analysis of Variance Table
##
## Response: log(weight)
##                  Df Sum Sq Mean Sq  F value    Pr(>F)
## log(height)       1 4.6643  4.6643 357.4735 < 2.2e-16 ***
## sex               1 0.3446  0.3446  26.4115 6.647e-07 ***
## log(height):sex   1 0.0144  0.0144   1.1038    0.2947
## Residuals       196 2.5574  0.0130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7.6 A few parting tricks about specifying models in R

We can add terms `weight ~ sex + height`

Have can crossed terms (interactions) `weight ~ sex * height` which is `weight ~ sex +  height + sex : height`

We can remove terms `weight ~ sex * height - sex : height` which is `weight~ sex +  height`

We can specify interations of order k among 2 or more terms: `weight~ (sex + height + age)^2` which expands to `sex + height + age + sex : height + sex : age + age : height` (e.g. all two way interactions among the three variables)