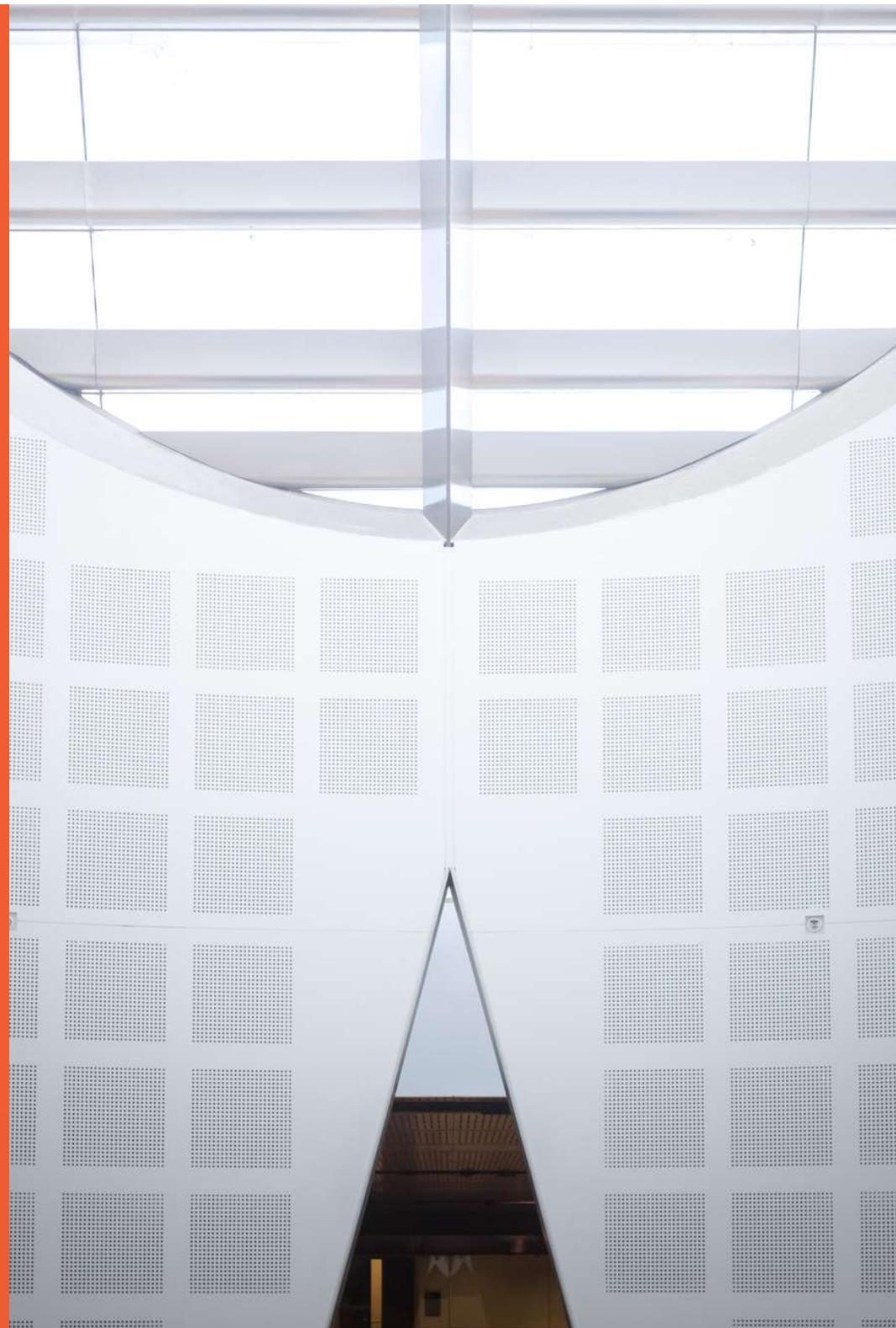


Linear Models 1: Linear regression, ANOVA, ANCOVA and repeated measures (a simple mixed model)

Presented by
Chris Howden
Sydney Informatics Hub
Core Research Facilities
The University of Sydney



THE UNIVERSITY OF
SYDNEY



Acknowledging SIH



All University of Sydney resources are available to Sydney researchers **free of charge**. The use of the SIH services including the Artemis HPC and associated support and training warrants acknowledgement in any publications, conference proceedings or posters describing work facilitated by these services.

The continued acknowledgment of the use of SIH facilities ensures the sustainability of our services.

Suggested wording:

General acknowledgement:

"The authors acknowledge the technical assistance provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

Acknowledging specific staff:

"The authors acknowledge the technical assistance of (name of staff) of the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

For further information about acknowledging the Sydney Informatics Hub, please contact us at sih.info@sydney.edu.au.

We value your feedback



- We aim to help HDR students and researchers in a wide range of fields across different faculties
- We want to hear about **you** and whether this workshop has helped you in your research.

- Later in this workshop there will be a link to a survey
- It only takes a few minutes to complete (*really!*)
- Completing this survey will help us create workshops that best meet the needs of researchers like you

During the workshop

- Ask short questions or clarifications during the workshop. There will be breaks during the workshop for longer questions.
-  – Slides with this blackboard icon are mainly for your reference, and the material will not be discussed during the workshop.

Challenge Question

- A wild boar is coming towards you at 200mph. Do you:
 - A. Ask it directions
 - B. Wave a red flag
 - C. Wave a white flag
 - D. Begin preparing a trap



After the workshop

These slides should be used after the workshop as **Workflows** and reference material.

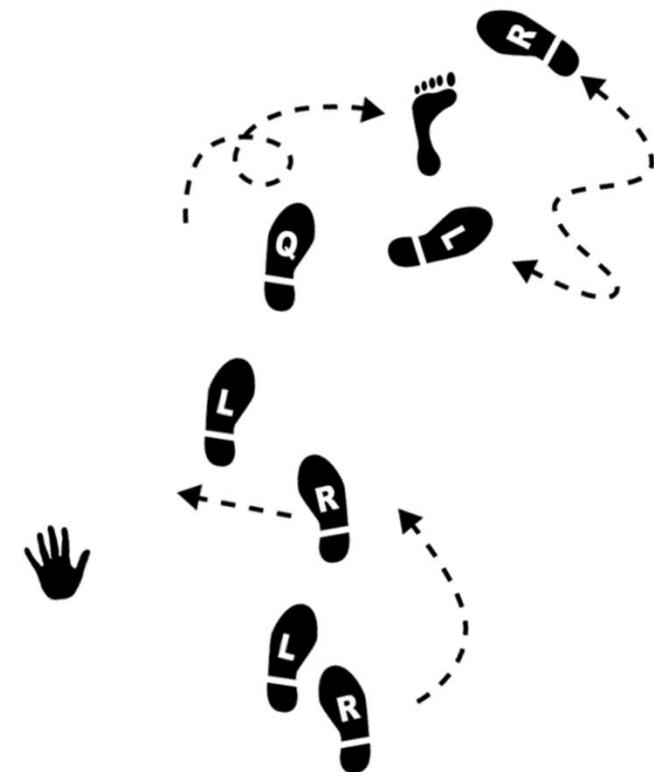
- Todays workshop gives you the **statistical workflow**, which is software agnostic in that they can be applied in any software.
- There are also accompanying **software workflows** that show you how to do it. We won't be going through these in detail. But if you have problems we have a monthly hacky hour where people can help you.

1 on 1 assistance

- You can email us about the material in these workshops at any time
- Or request a consultation for more in-depth discussion of the material as it relates to your specific project. Consults can be requested via our Webpage (link is at the end of this presentation)

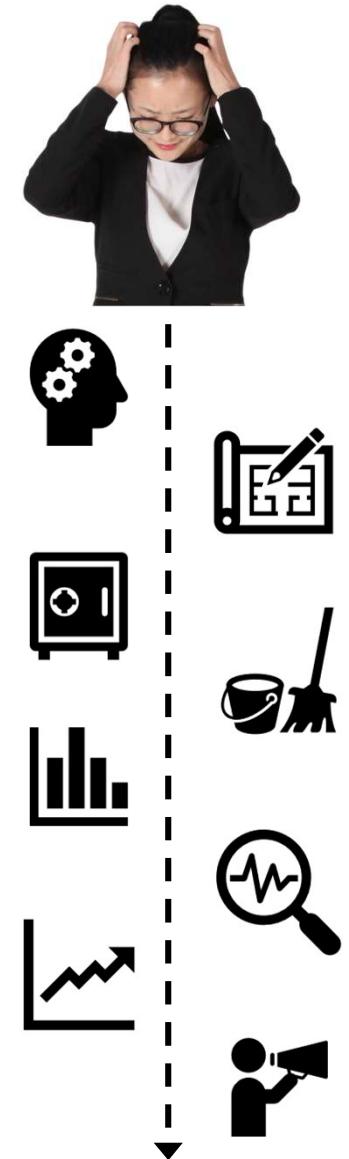
Research Workflow

- Why do we use a research workflow?
 - As researchers we are motivated to find answers quickly
 - This drive can cause problems if we don't think systematically
 - ... and we need to in order to:
 - Find the right method
 - Use it correctly
 - Interpret and report our results accurately
 - The payoff is huge, we can avoid mistakes that would affect the quality of our work and get to the answers sooner
- So... what is a workflow?
 - The process of doing a statistical analysis follows the same general “shape”.
 - We provide a general research workflow, and a specific workflow for each major step in your research
(currently experimental design, power calculation, analysis using linear models/survival/multivariate/survey methods)
 - You will need to tweak them to your needs



General Research Workflow

1. **Hypothesis Generation** (Research/Desktop Review)
2. **Experimental and Analytical Design** (sampling, power, ethics approval)
3. **Collect/Store Data**
4. **Data cleaning**
5. **Exploratory Data Analysis (EDA)**
6. **Data Analysis aka inferential analysis**
7. **Predictive modelling**
8. **Publication**



THE UNIVERSITY OF
SYDNEY

Sydney
Informatics Hub

CONTENTS: Linear Models I - An Introduction

A Statistical Workflow for most Linear Models, software agnostic

- Applicable in any software
- There is accompanying R code if you wish to do it in R. Plots are done using a combination of default plotting functions and ggplot functions. You will know the difference since ggplot functions start with `ggplot()`.

Applied workflows to 4 of the most common analyses:

- Simple Linear Regression
- ANOVA (Control vs Treatment)
- Repeated Measures
- ANCOVA

The first example introduces the basic concepts and workflow so we don't show you how to do it in R or SPSS. Subsequent examples will have R code.

What are Linear Models?

ANOVA

Linear Regression

ANCOVA

Logistic regression

Before After Control
Impact (BACI) Studies

Count regression

Repeated measures

Randomised Control
Trials (RCT's)

Plus Many More!!

A single unifying Theory

Regression and ANOVA are often taught as different things. Yet they aren't!

An easier way to understand them is with the single unifying Linear Models theory.

This allows us to apply them using the same workflow.

Model Fitting Workflow

Step 0) Clean and check data.

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA).

Step 2) Fit the Model

Step 3) Check Model Assumptions via Diagnostics: Residual Analysis

Step 4) Goodness of Fit: Plots and Statistics

Step 5) Interpret Model Parameters and reach a conclusion

Step 6) Reporting

Linear Models 3 and Model Building Workshops have more detail on many of these steps.

Step 0) Clean and check data

- Is covered in “Research Essentials”, not this workshop.
- Is very important, so ensure you do it!
- Get in the habit of checking the data every time you open it by looking at the **corners** i.e. start at the top left corner, then scroll to the far right corner, scroll down to the bottom right corner, scroll left to the bottom left corner, then finish by scrolling back up to the beginning top left corner.
 - Weird things can happen. New versions, a stray cosmic ray. I have literally opened data to find it corrupted, and then reopened it and it's fine. Similarly I have seen weird results only to rerun them to find them OK.

Simple Linear Model

Continuous response and predictor

Workflow Suitable for:

- Modelling continuous predictors (workflow shown is for 1 predictor, there are additional considerations when more than 1 e.g. multicollinearity, these are discussed in our Model Building workshop)
- Least Squares Regression
- Simple Linear Regression

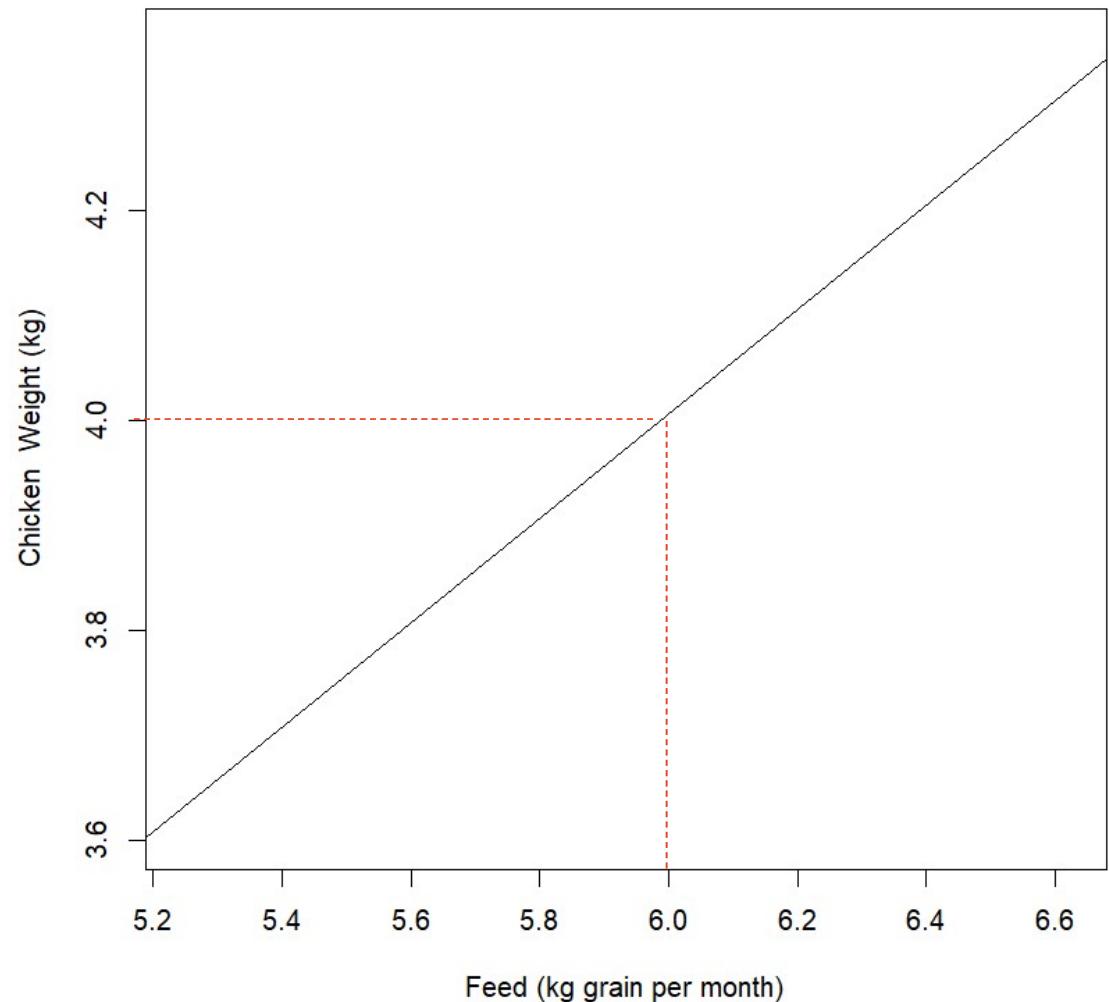
Simple Linear Model

Your Turn: Draw a linear model for the weight of chicken compared to the amount of feed it eats in its first month.

So in this example a chicken that eats 6 kg of Feed will weigh about 4kg



Linear Model aka Regression



So we know it's linear. Is that all we need to know?

NO! We want to know exactly how our Predictor (feed) affects our Response (weight).

And for that we need to fit an equation to the pictorial model you just drew so we can pull out the parameter that represents the Predictors affect on our Response.

High School Equation for a line

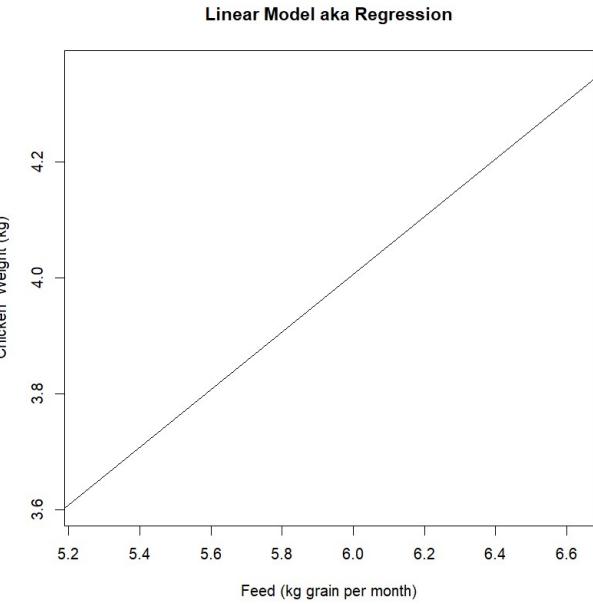
$Y = \text{slope (aka gradient)} * X + \text{Constant (aka Y intercept)}$

$$Y = mX + b$$

Statistical Equation for a line (puts the constant first)

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

So we want to find β_1 , which is the slope(gradient) of the line and represents the effect Feed has on Weight. (β_0 is the constant)



But we're still missing something?

THE DATA!!!!

Each datum has its **own natural variance** from the line since each chicken is a bit different!

Another name for the Natural Variance is the “Error” of the model. Which is why we usually represent it as an ε in the model.

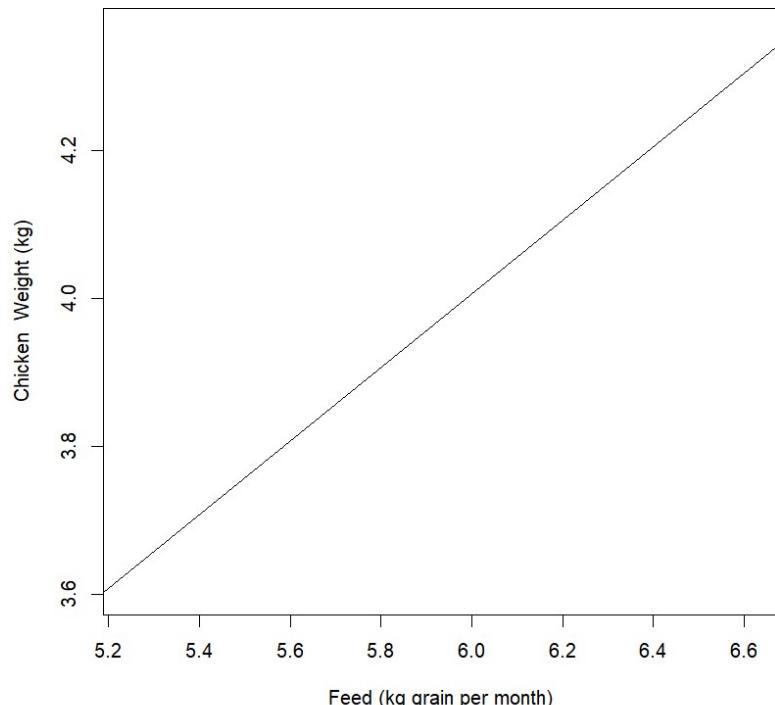
\hat{Y} ~ The “hat” over the \hat{Y} tells us that it's a **prediction** of Y for those specific predictor values for X .

Y ~ Is the **actual value** of Y , so it's the prediction + error.

MODEL FOR A LINE

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

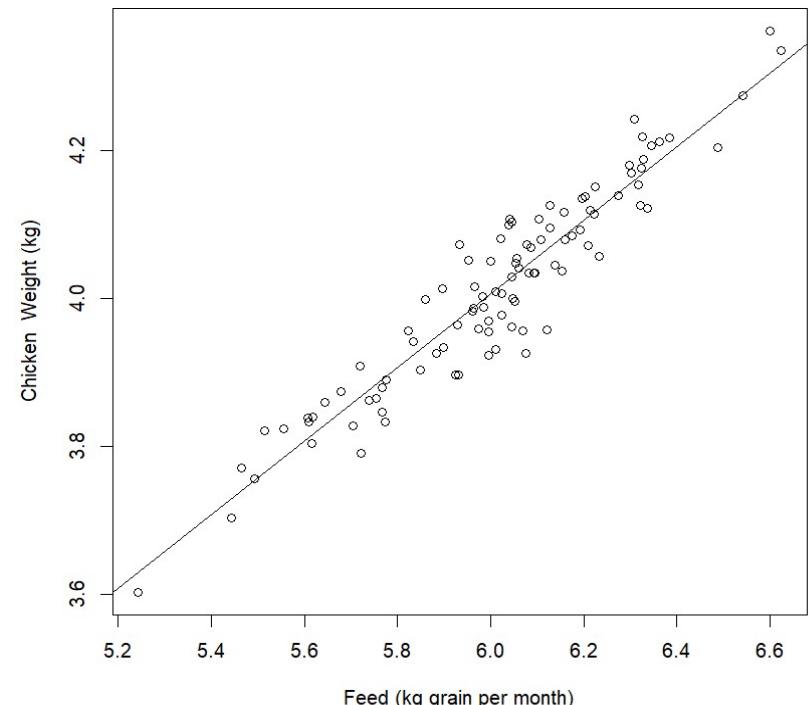
Linear Model aka Regression



MODEL FOR OUR DATA

$$Y_i = \hat{Y}_i + \varepsilon_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Linear Model aka Regression



THE UNIVERSITY OF
SYDNEY

Sydney
Informatics Hub

So how do we use this equation to understand the relationship between our predictor and response?

We look at the Parameter estimates of the model.

Parameter	Estimate	SE	T score	P value	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant / Intercept (β_0)	1.03	0.136	7.6	2.24e-11	0.8	1.3
Feed (β_1)	0.50	0.023	21.8	<2e-16	0.45	0.54
Model Fit is => $Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i \Rightarrow \text{Weight} = 1.03 + 0.50 * \text{Feed} + \varepsilon_i$						

Notation 2.24e-11 means move the decimal place to the left 11 places i.e. $2.24\text{e-}11 = 0.000000000224$. It is done so we can write small numbers concisely.

So how do we use this equation to understand the relationship between our predictor and response?

We look at the Parameter estimates of the model.

Parameter	Estimate	SE	T score	P value	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant / Intercept (β_0)	1.03	0.136	7.6	2.24e-11	0.8	1.3
Feed (β_1)	0.50	0.023	21.8	<2e-16	0.45	0.54
Model Fit is $\Rightarrow Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i \Rightarrow \text{Weight} = 1.03 + 0.50 * \text{Feed} + \varepsilon_i$						

First we look at the constant (β_0), to ensure it's needed and there is nothing weird going on. So we can say:

- It is likely different to 0 (since $p=2.24e-11$ which is very small so it is very unlikely we are making the wrong decision if we say this).
- It is likely somewhere between 0.8-1.3.

So how do we use this equation to understand the relationship between our predictor and response?

We look at the Parameter estimates of the model.

Parameter	Estimate	SE	T score	P value	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant / Intercept (β_0)	1.03	0.136	7.6	2.24e-11	0.8	1.3
Feed (β_1)	0.50	0.023	21.8	<2e-16	0.45	0.54
Model Fit is $\Rightarrow Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i \Rightarrow \text{Weight} = 1.03 + 0.50 * \text{Feed} + \varepsilon_i$						

Next, let's look at the likely association between Feed and Weight represented by β_1

- It is likely different to 0, (since $p < 2e-16$ which is very small so it is very unlikely we are making the wrong decision if we say this).
- The effect is likely somewhere between 0.45-0.54. Or in other words for each extra kg of Feed eaten we expect a chicken to weigh between 0.45-0.54 kg more.

**So, is that all we need to do? Is our Analysis finished,
can we now write up our conclusions?**

NO, because Computers are Stupid!!

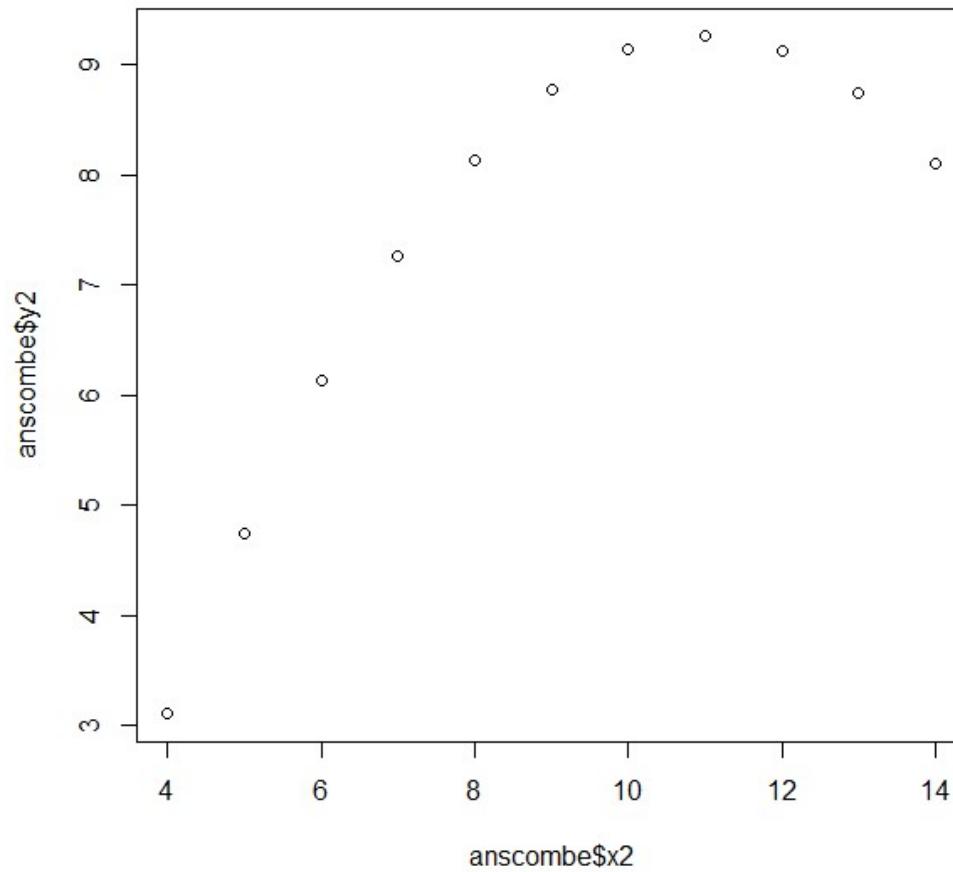
Because a computer will fit any model you tell them to even if:

- It's a bad fit to the data
- It's a stupid fit to the data

So it's up to YOU to decide if the model you are asking the computer to fit to your data is the right type and a good fit.

Because if it's a bad fit, then the parameters and conclusions we draw from them will be wrong. And there is little in the previous parameter table to warn you of this!!!! So we need to look at other things.

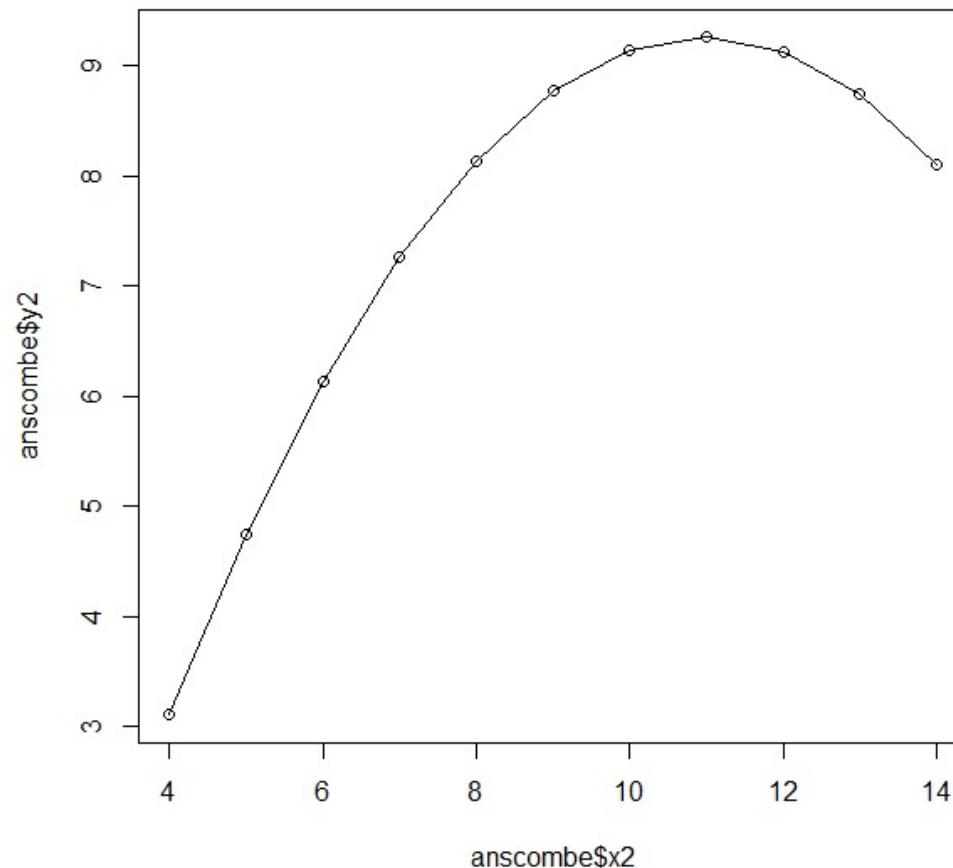
Your Turn: Draw the best fit to this data



Your Turn: Draw the best fit to this data

Hopefully You drew this.

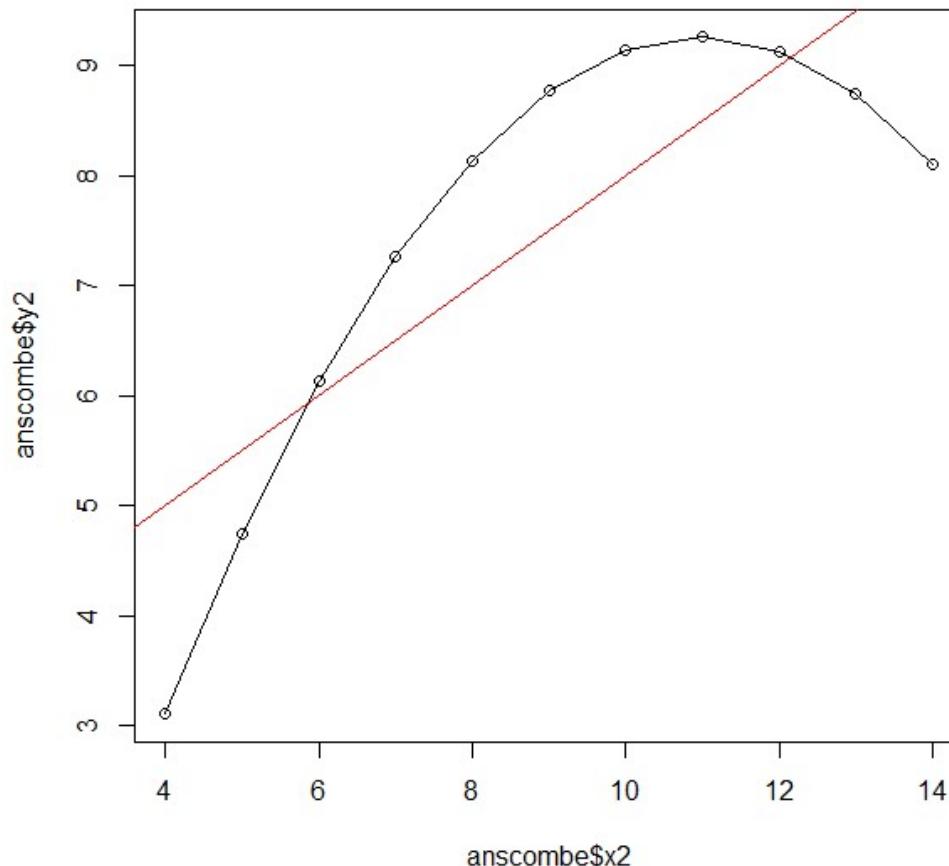
Now draw a linear fit.



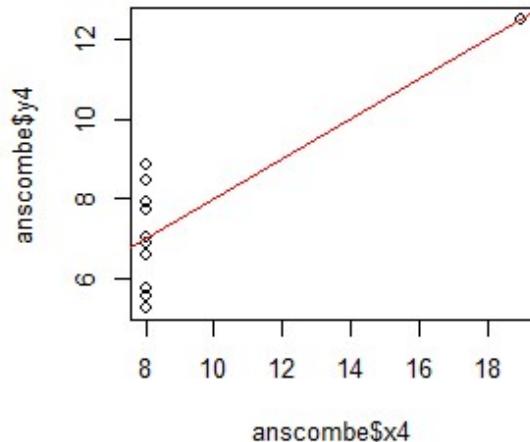
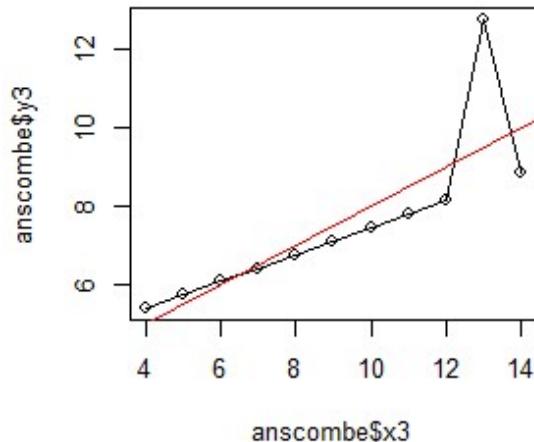
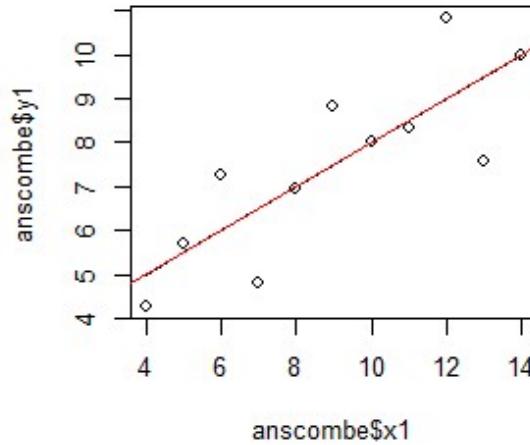
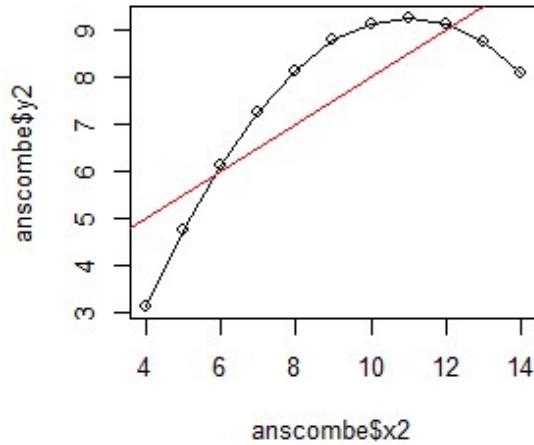
Your Turn: Draw the best fit to this data

Which model do you think a computer fits if you ask it to do a linear regression?

The wrong (linear) one! Because Computers are Stupid.



This is just one example from Anscombe's Quartet, 4 data sets all with the same linear fit. But only one is actually linear.



So how do we decide if the model we are asking the computer to use is a good enough fit to the data that the parameters, and the conclusions we make from them, make sense?

1) EDA

- 1) Plot the data to look for linearity (response vs predictor), correlation (serial plots), non-normality (histograms/kernel density plots), etc.

2) Check Model Assumptions via Diagnostics

- 1) Linearity
- 2) Normal Error
- 3) Independence

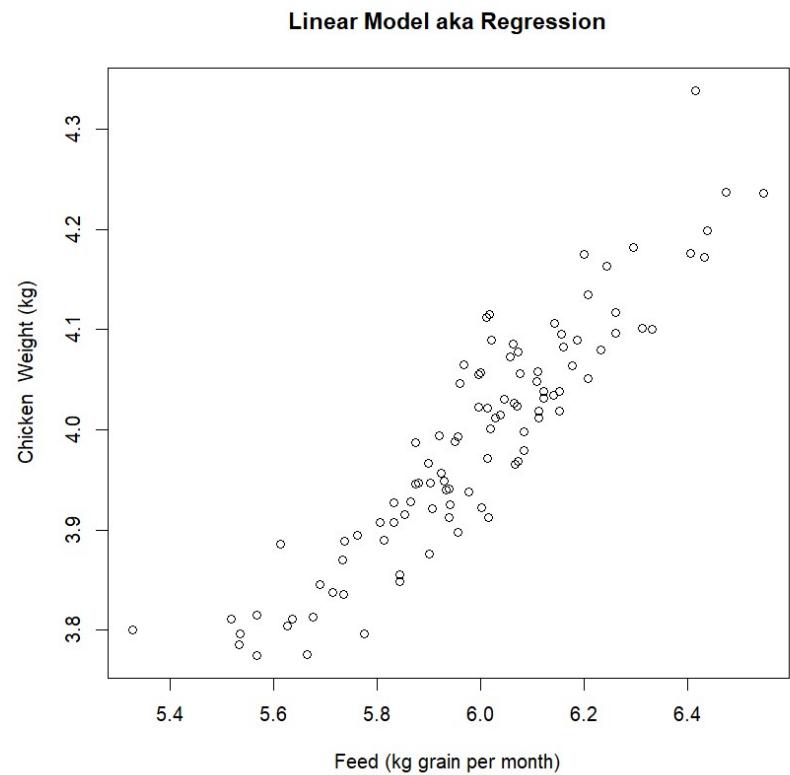
3) Check Model Goodness of Fit

- 1) How much of the response variance does the model explain?
- 2) Is the model a good fit of the data overall, or is it biased towards explaining just a couple datum?

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Linearity: Draw A Graphical model of the data

1. Simply plot the data and have a look. Is a linear model a good fit to the data?
2. Try to write down the model you want to fit as well. This will **help you interpret what the β Parameters mean**, particularly for complicated models.
 - $Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i$



Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Independence: Consider your experimental design

Is there anything about it that might lead to datum being correlated with each other. For example, if we had repeated measures on the same patient (chicken) then we would expect these to be correlated i.e. dependant on each other.

Modelling independence correctly is important for 3 main reasons:

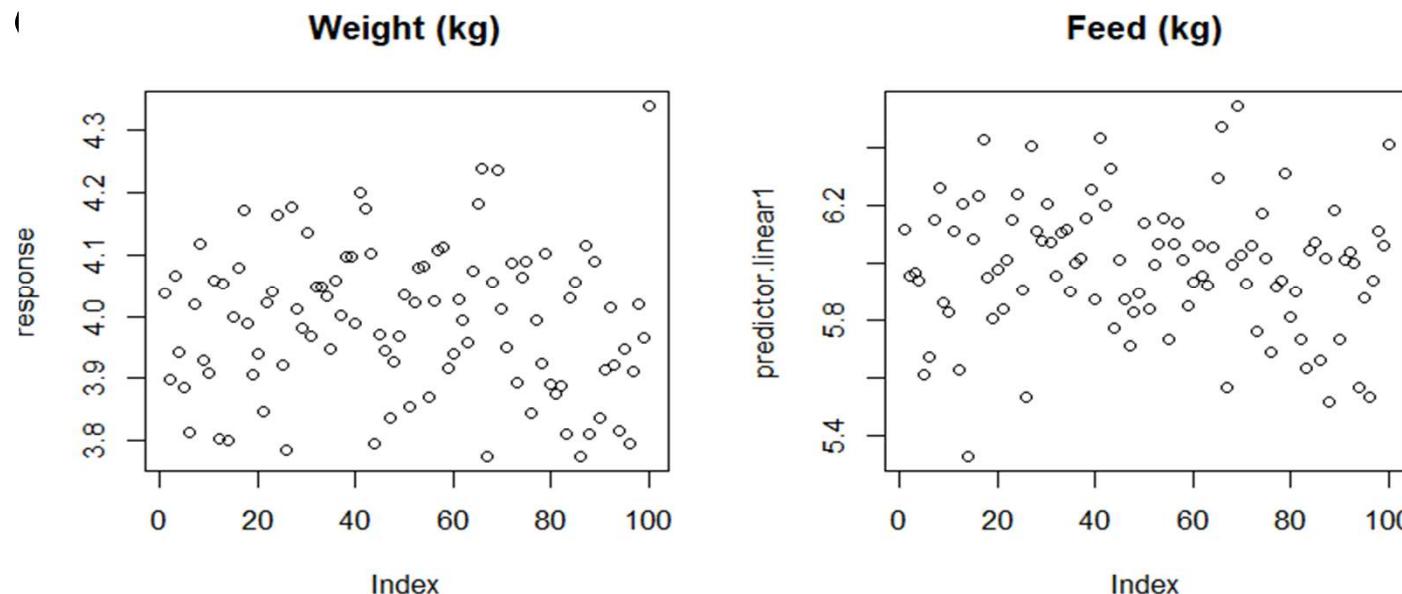
- 1) Ensures the correct sample size is used.** For example if I measured the chickens weight 100 times a second for 60 seconds do I really have 6000 samples per chicken? NO, of course not. Because the 6000 samples aren't independent. This is known as **Pseudo Replication** and inflates our sample size, lowering our SE's and making our p-values too low and confidence intervals too narrow.
 - 1) This is one reason for the replication crisis i.e. artificially low p-values.
- 2) Partitioning extra sources of error/noise which makes our analysis more accurate,** which is done using mixed models for designs such as split-plots, blocked, repeated measures.
- 3) Structural Correlation that should be added to the model** e.g. serial correlation such as Auto-regressive correlation.
 - 1) Stock prices are independent day to day (since something can happen to change their price) but are heavily dependant on the prior days price.

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Independence: Plot the data using a “Serial Plot” i.e. data plotted 1 after each other

This is simply a plot of the data, one after each other, as recorded in your data. You are looking for unexplained sequences of high or low values i.e. unexplained correlations.

- You can also organise your data into different structures to look for different types of Dependence e.g. if repeated measures then organise so each persons (chickens) data is sequential.
- You



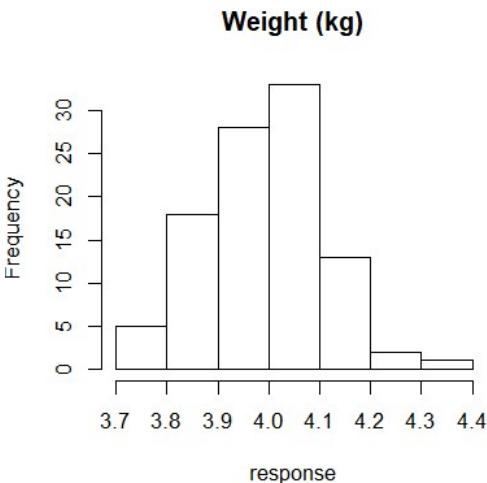
Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Normality

This is a very poorly understood assumption. The Assumption is that the **Error, not the Response** is normal. Meaning we can't test it until we fit a model. So don't make the mistake of thinking just because your data isn't normal this assumption has been violated.

What we can do is consider exactly what it is we are modelling and also look at the response using a histogram to see if a normal error might not fit. Obviously if the response looks normal there is a good chance the errors will be too. However a non-normal response can have a normal error (which I will show you when we look at ANOVA).

The main thing we are looking for here are things that usually prevent the error from being normal and are better fit using different models such as the response being non continuous e.g. binary, counts, extreme outliers, extreme skewness, truncation.

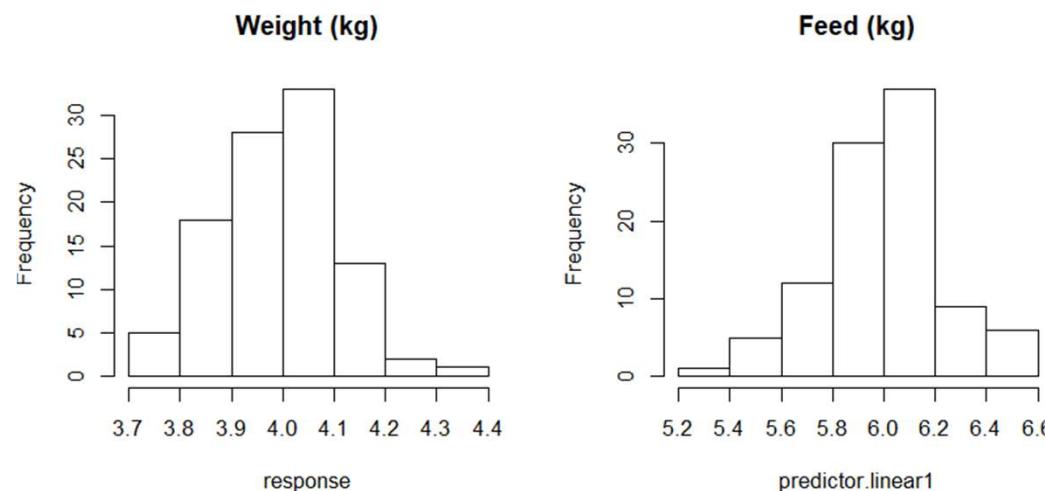


It's worth noting that discrete data can be modelled using a normal error under some circumstances e.g. weight rounded to the nearest gram is technically discrete, but can be fit using a normal error. Counts can also be fit using a normal error if large enough.

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Outliers

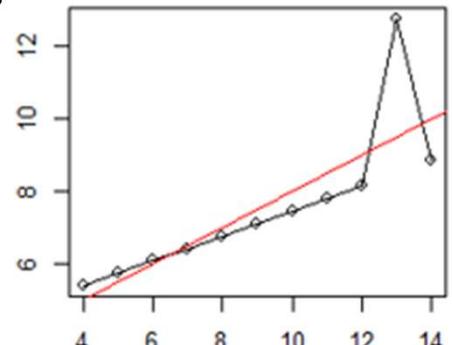
This is also a very poorly understood assumption. We want a model to represent the bulk of the data. We don't want it biased towards 1 or 2 outlying influential points. Just like checking the normality assumption we can only test this for sure once we have fit a model. However it is always worth looking at all our data to see if there are any outliers we might need to deal with. The best way to do this is via histograms.



Fixing Model Assumption Problems

Outliers

1. Check to see if they are a data entry or collection mistake and can be removed.
2. Consider transformations that reduce their influence e.g. log transforms will reduce the influence of large outliers.
3. Consider removing them to get a model that is a better fit to the majority of the data. If this is done one **must** say so in any reporting. For example: looking at the Anscombe example on the right. What is a better model. A line through the datum in a straight line, while saying there was a single large outlier. Or the red line shown?
4. Consider other models that can handle the outliers.



Fixing Other Model Assumption Problems

This is a complex business and is beyond the scope of this workshop. It is covered in more detail in other Linear Model courses we give. The quick answer is that you will usually need to use a different model. In brief:

Non linear fit

1. Add in quadratic and non linear terms for either the predictors or the response (GLM's can add such terms for the response via the link function as [Discussed in Linear Models II.](#)).
2. Use a non linear model such as a General Additive Model (GAM).

Normal error is inappropriate

1. Use a different type of linear model. A Generalised Linear Model (GLM) with a different Error distribution often works e.g. binomial for binary data (logistic regression), Poisson for count data. [Discussed in Linear Models II.](#)

Lack of Independence

1. Fit a mixed model that accounts for the correlation structure. [Discussed in Linear Models I and III.](#)
2. Remove datum until they are independent (also known as censoring).
3. Average the independent data e.g. average the 6000 chicken weights so we have a single score. Has the advantage of also usually making the data normally distributed, by invoking the Central Limit Theorem (CLT)

Step 2) Fit the Model

Use your software of preference to fit the model.

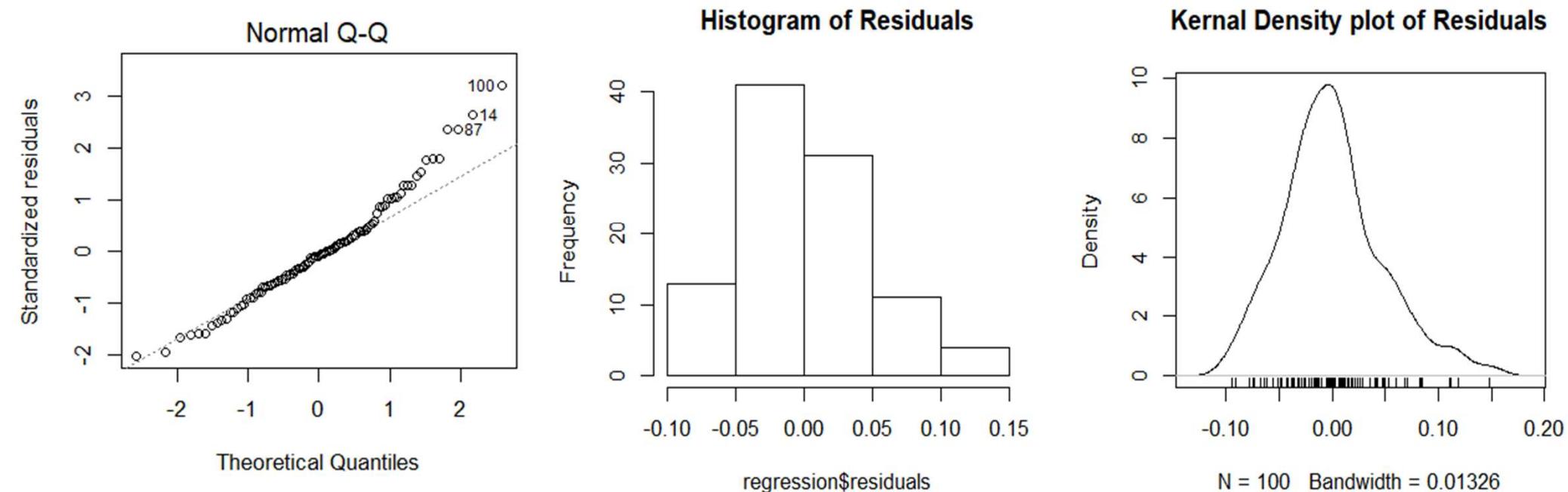
In R you'd use something like this:

```
> regression <- lm(response~predictor.linear1)
> regression <- lm(weight~feed, data=data)
```

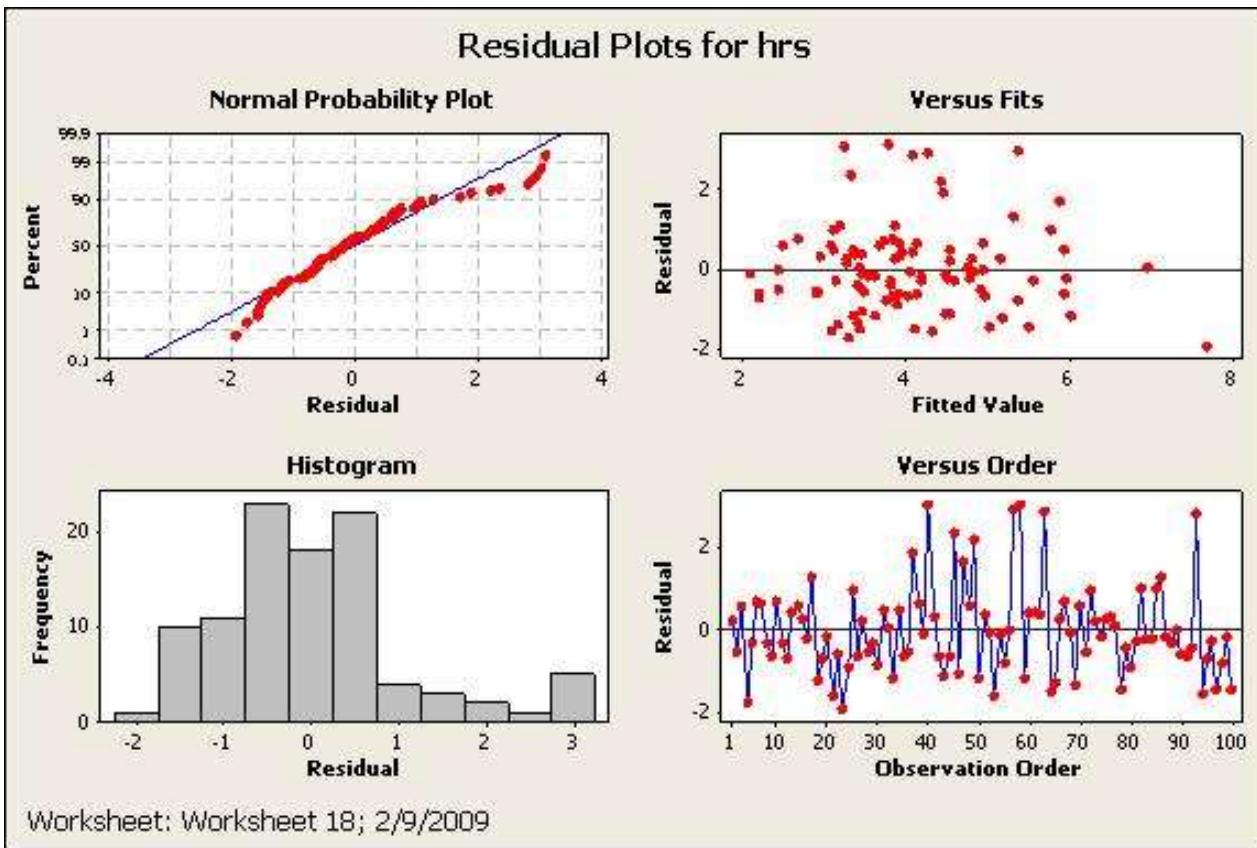
Step 3) Check Model Assumptions via Diagnostics: Residual Analysis

Normality

- The QQ plot is pretty standard, if normal data should be along the straight 1:1 line. I also like a histogram and density plot since these are easier to see the actual distribution and diagnose problems.
- Note that QQ plots are very sensitive. In this example we know the underlying error is normal (since we simulated it) yet one might not think that from the QQ plot.
- Linear models are very robust to the normality assumption.



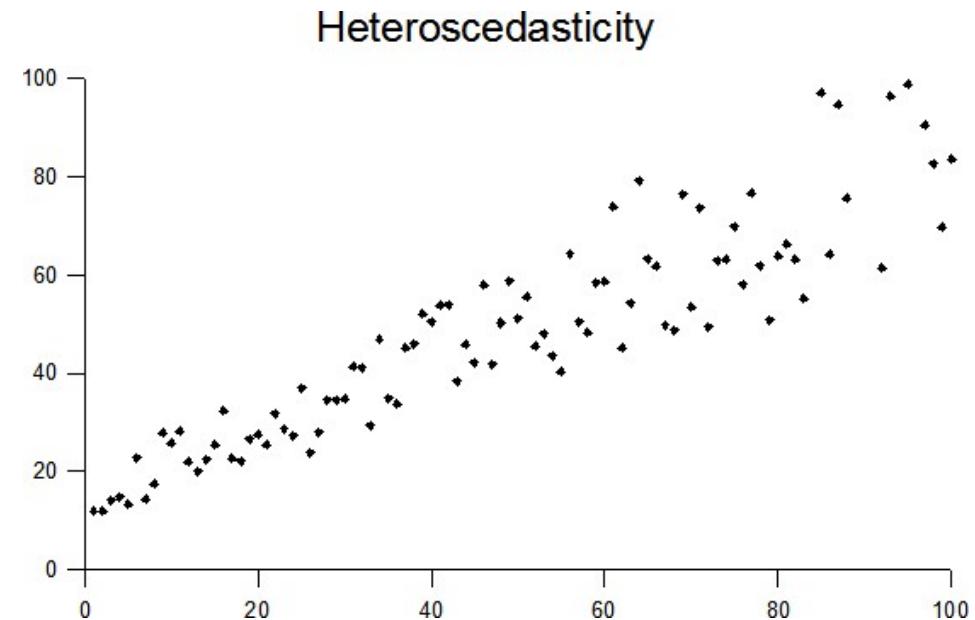
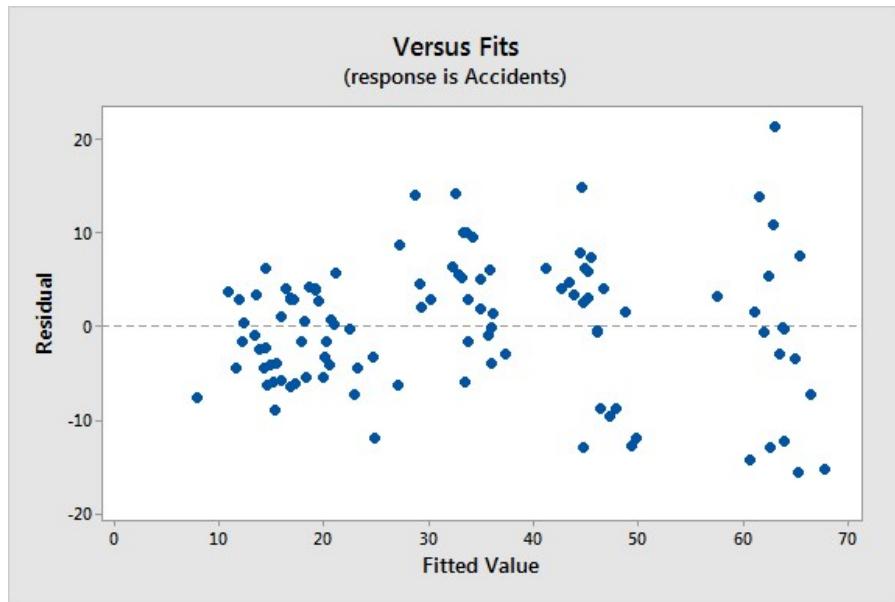
Non-normal Error



Often fixed by fitting a non normal error, transformations or adding new predictors that account for the non normality. In this example a natural log of the response fixed the problem.

<https://smartersolutions.com/multiple-regression-dealing-with-non-normal-residuals.html/>

Non normal error: Heteroscedasticity i.e. variance not constant



Left graph: residuals showing heteroscedasticity.

<https://statisticsbyjim.com/regression/heteroscedasticity-regression/>

Right graph: raw data with heteroscedasticity, and a linear model.

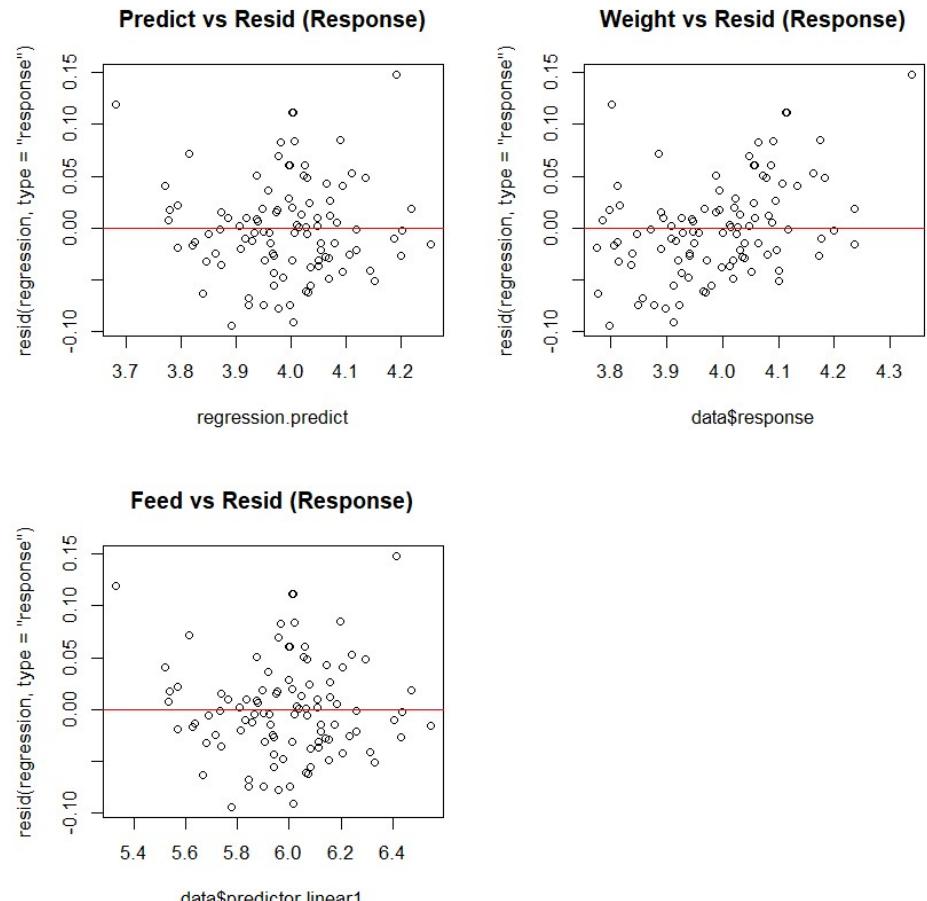
https://en.wikipedia.org/wiki/Homoscedasticity_and_heteroscedasticity

Common with count, rate and concentration data as we expect a count of 10000 to have higher variance than a count of 1. Often fit with a GLM and Poisson distribution as [discussed in Linear Models II](#). Can also be fixed by log transforming the response when fitting a normal error.

Step 4) Goodness of Fit: Residual Analysis

Is there any unexplained structure, non linearity or non constant variance?

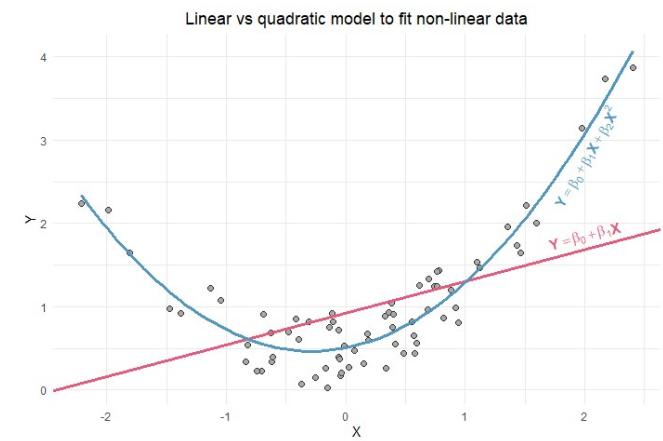
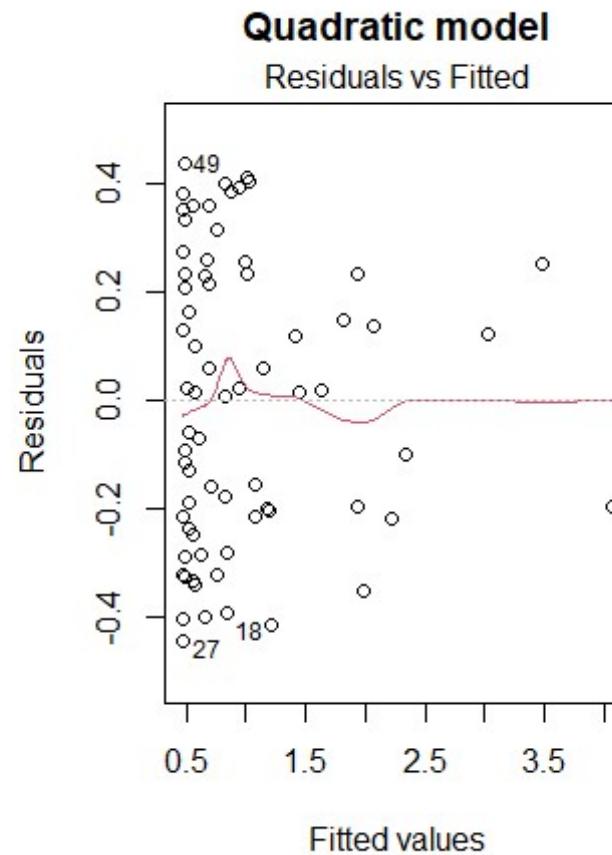
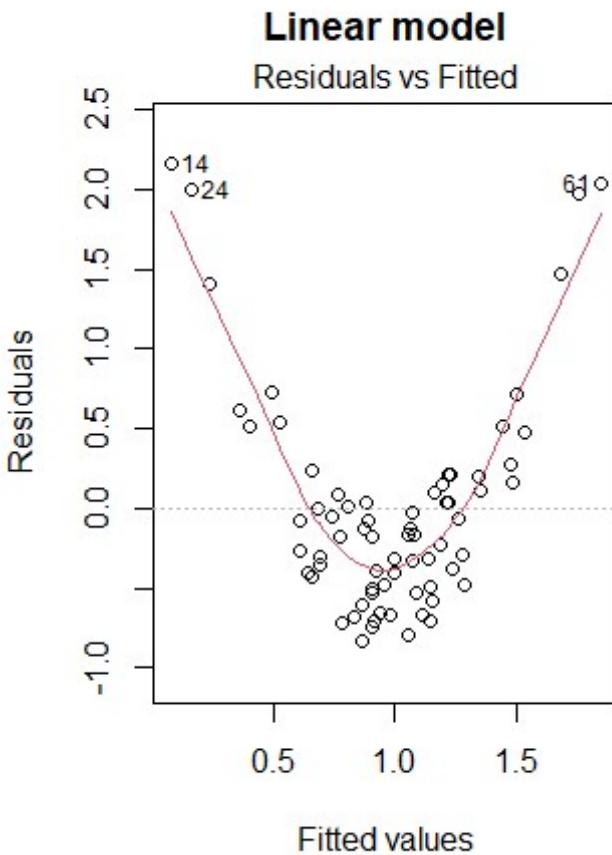
- We want to see our residuals randomly scattered about zero since this indicates a fit that is:
 - consistent across the different predicted, response and predictor values.
 - with no unexplained structure our model has missed.
- Patterns can indicate:
 - Missing predictors
 - Incorrect Error
 - Non linear fit e.g. quadratic
- No evidence of non constant Variance i.e. heteroscedasticity.



Any Outliers

No

Non Linear Fit: Residuals when a linear term is fit to a quadratic pattern



Left graph: residuals with a linear fit. Problematic as not random about zero.

Middle graph: residuals with quadratic fit. Good fit as random about zero.

Right graph: EDA on raw data, which shows the need for a quadratic fit.

Influential Outliers

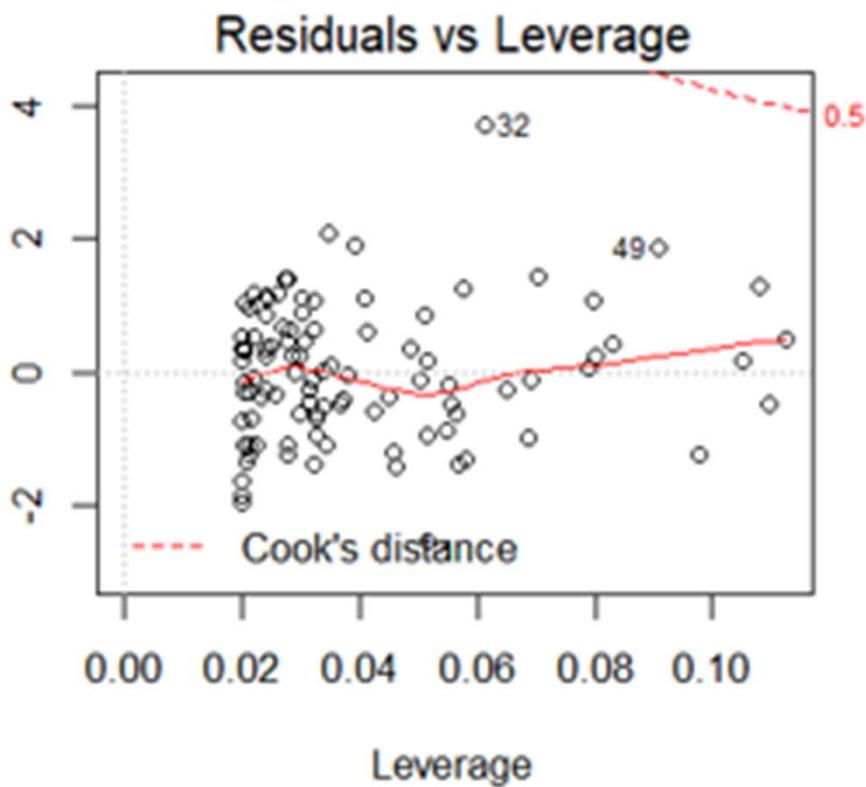
Some outliers have a greater **influence** on the model than others. These are known as **influential outliers**. They are outliers which have:

- High error i.e. when not used in the model their prediction is very different.
- High leverage i.e. they have a large impact on the model parameters.

Cooks Distance: a large cook's (d) indicates that the data point strongly influences the fitted values.

- To compute:
 1. Delete observations one at a time.
 2. Refit the regression model on remaining ($n-1$) observations
 3. Examine how much all of the fitted values change when the i^{th} observation is deleted.
- In terms of what values are high enough to warrant concern.
 - A general rule of thumb for ‘large n’ based on cooks distance following the F distribution is to keep an eye on values > 0.5 and view those > 1 with concern. However there are other thoughts on this.

Influential Outliers



All points within the Cooks distance red dotted lines - so no evidence of influential outliers.

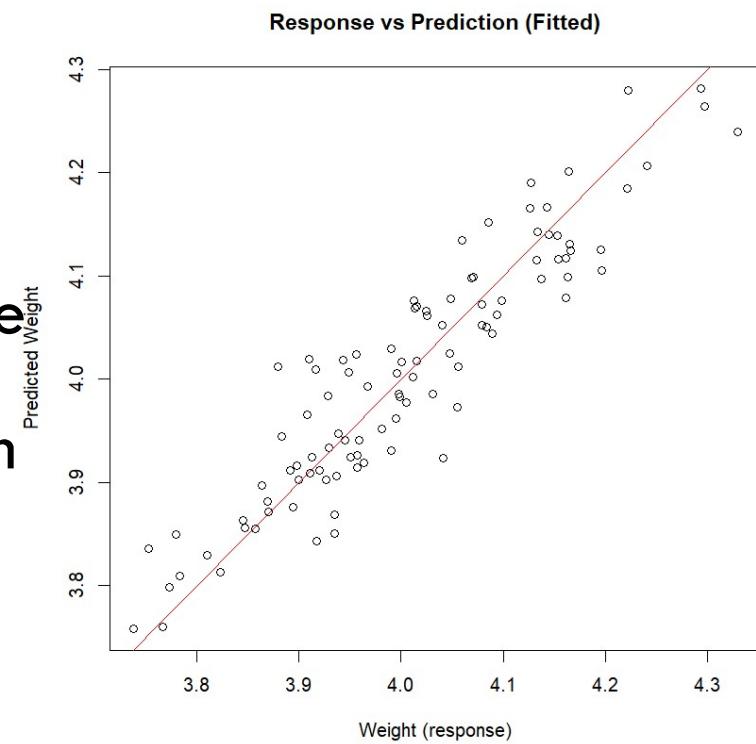


Step 4) Goodness of Fit: Plots and Statistics

So far we have established our model is a **good fit to our data** and there is **nothing obvious we have missed**. Next question. **How well does it predict i.e. fit, the data?**

This plot is a good visual representation of model fit. If the response is being exactly predicted than we expect it to fall along the 1:1 line.

The correlation along this line is the most commonly used Goodness of Fit Statistic: called R^2 . It is literally the correlation of the response and prediction squared. And represents the % of the responses variation the prediction i.e. model, explains. In this example it is 88%.



What is a 'good' R²?

It's totally domain specific, so take your benchmark from similar published work. It depends on how much natural variation we expect in the system. For example:

- Market Research Consumer Purchase Intent and Liking: 70-90%
- Ecological Communities: anything over 20% is fabulous!!

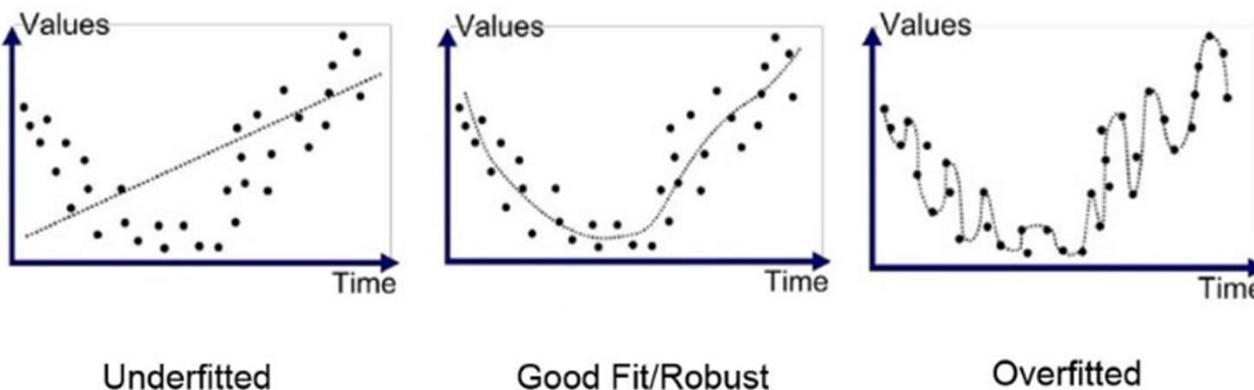
Why is a high R² bad: Overfitting leads to poor predictive power

We want our model to be a good representation of the underlying population so we can infer what is happening outside our sample. And when doing prediction for the predictions to be accurate.

When R² is too big it suggests we have fit some of this samples unique noise/error/variation along with the population signal. So although it is a good fit to this sample, it will be a poor fit to other samples **and the population**.

This is called **Overfitting**.

I question anything with an R² of greater than 90%. But again its domain specific, if I had ecological community data model with an R² of greater than 80% I'd be checking for overfitting.



Minimum Sample Size: 10 data points per parameter

A common cause of overfitting is having too many predictors compared to data points. This can also lead to unstable parameters with high SE.

A common rule of thumb to prevent this is to have at least 10 data points per parameter. Don't forget the intercept is a parameter too!

EG: A simple linear regression with 1 predictor has 2 parameters (constant plus the predictors slope parameter) so usually requires 20 observations.

Refer to our Power and Sample Size workshop for more information on sample size considerations

Step 5) Interpret Model Parameters and reach a conclusion

FINALLY!! We can actually have a look at our model and see what it is telling us.

Realistically most people, including me, often do the EDA plots, pick the model they think best suits the data, plot it and then look at this model summary first.

And then go back to do all the above due above model due diligence.

Which is understandable, but just make sure you do it!!

So how do we use this equation to understand the relationship between our predictor and response?

We look at the Parameter estimates of the model.

Parameter	Estimate	SE	T score	P value	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant / Intercept (β_0)	1.03	0.136	7.6	2.24e-11	0.8	1.3
Feed (β_1)	0.50	0.023	21.8	<2e-16	0.45	0.54
Model Fit is => $Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i \Rightarrow \text{Weight} = 1.03 + 0.50 * \text{Feed} + \varepsilon_i$						

Step 6) Reporting: Overall Conclusion suitable for publication

Always include these 2 things

1) Interpret the model and what it means for your research

There is strong evidence to show that feed influences weight ($p < 2e-16$), with each kg of feed adding between 0.45-0.54 kg of weight (95% CI). This effect on weight has been estimated very accurately [as 95% CI is quite narrow].

2) Show that the model is a good fit and assumptions have been tested and met

The model is a good fit to the data with an $R^2 = 88\%$. There were no outliers or unexplained structure. The error was normal.

And when giving a p-value always give an estimate of the effect size as well i.e. the 95% CI.

Step 6) Reporting: Overall Conclusion suitable for publication

So a suitable write up would be as follows

“There is strong evidence to show that feed influences weight ($p < 2e-16$), with each kg of feed adding between 0.45-0.54 kg of weight (95% CI). This effect on weight has been estimated very accurately [as 95% CI is quite narrow].

The model is a good fit to the data with an $R^2 = 88\%$. There were no outliers or unexplained structure. The error was normal”

A Conversation is
better than a
Presentation



So please speak up and ask questions!

People think differently.
So I may need to explain
things in 2 or 3 different
ways!



THE UNIVERSITY OF
SYDNEY

Sydney
Informatics Hub

ANOVA: ANalysis Of VAriance

Continuous response, categorical predictor

Workflow Suitable for:

- **Modelling discrete predictors** (workflow shown is for 1 predictor, there are additional considerations when more than 1 e.g. Confounding, , these are discussed in our Model Building workshop)
- **Control vs Treatment designs**
- **Randomised Control Trials (RCT)**



THE UNIVERSITY OF
SYDNEY

Model Fitting Workflow

Step 0) Clean and check data.

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA).

Step 2) Fit the Model

Step 3) Check Model Assumptions via Diagnostics: Residual Analysis

Step 4) Goodness of Fit: Plots and Statistics

Step 5) Interpret Model Parameters and reach a conclusion

Step 6) Reporting

Linear Models 3 and Model Building Workshops have more detail on many of these steps.

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Your Turn:

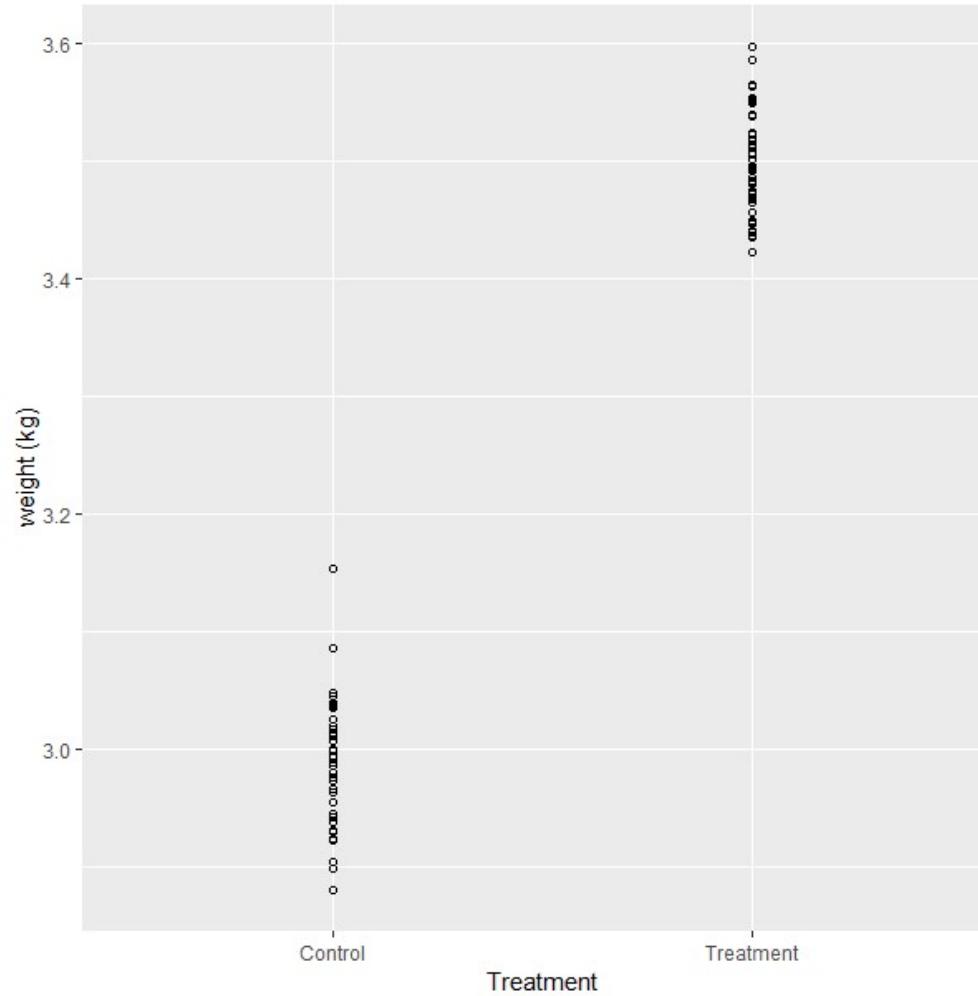
We have a chicken feed experiment where we added a protein supplement.

We expect the Control to have an average of 3kg, and the Treatment to add 0.5kg to weight. With a $SE(\text{mean}) = 0.05$ i.e. 95% of data is within 0.10 of the mean.

Plot the data!!



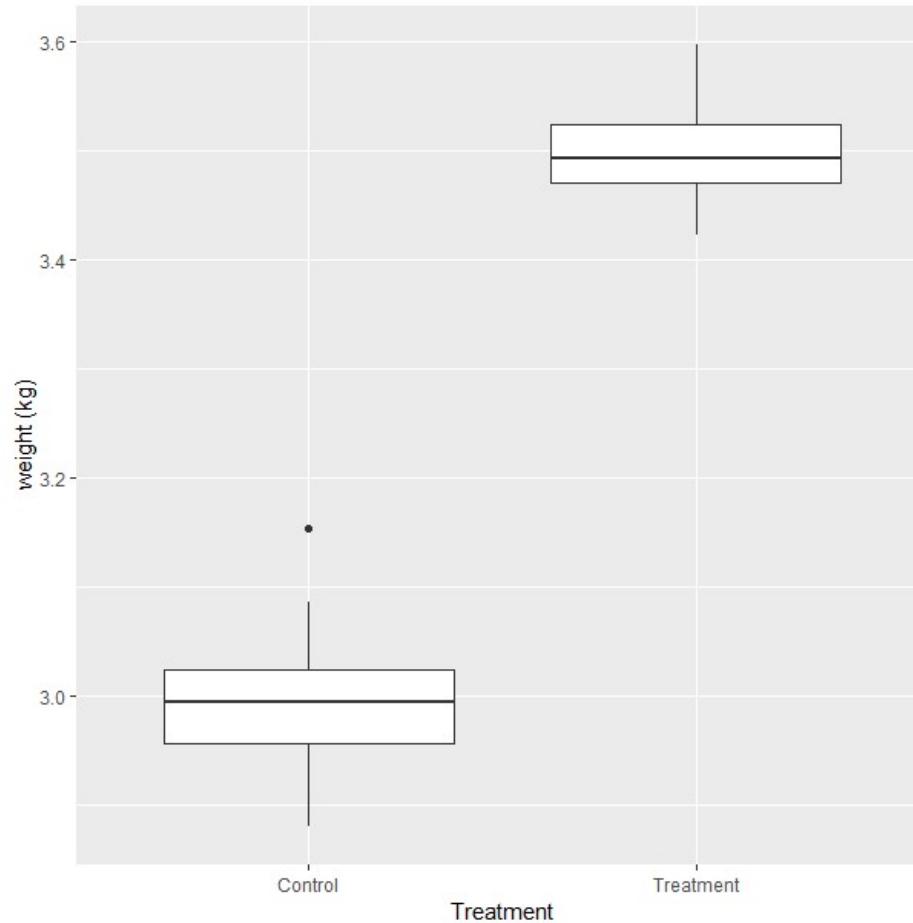
Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)



```
ggplot(data = data1, aes(x=treatment, y=response)) +  
  geom_point(pch=21) +  
  labs(x="Treatment", y="weight (kg)")
```

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Boxplots need at least 12 datum. If not enough use dot plots as per the previous slide.



This is an ANOVA: which is a type of Linear Model.

This may not look like a linear relationship, but it is a linear model. I will explain why later.

```
ggplot(data=data1, aes(x=treatment, y=response))+  
  geom_boxplot() +  
  labs(x="Treatment", y="weight (kg)")
```

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Independence: Consider your experimental design

Your Turn:

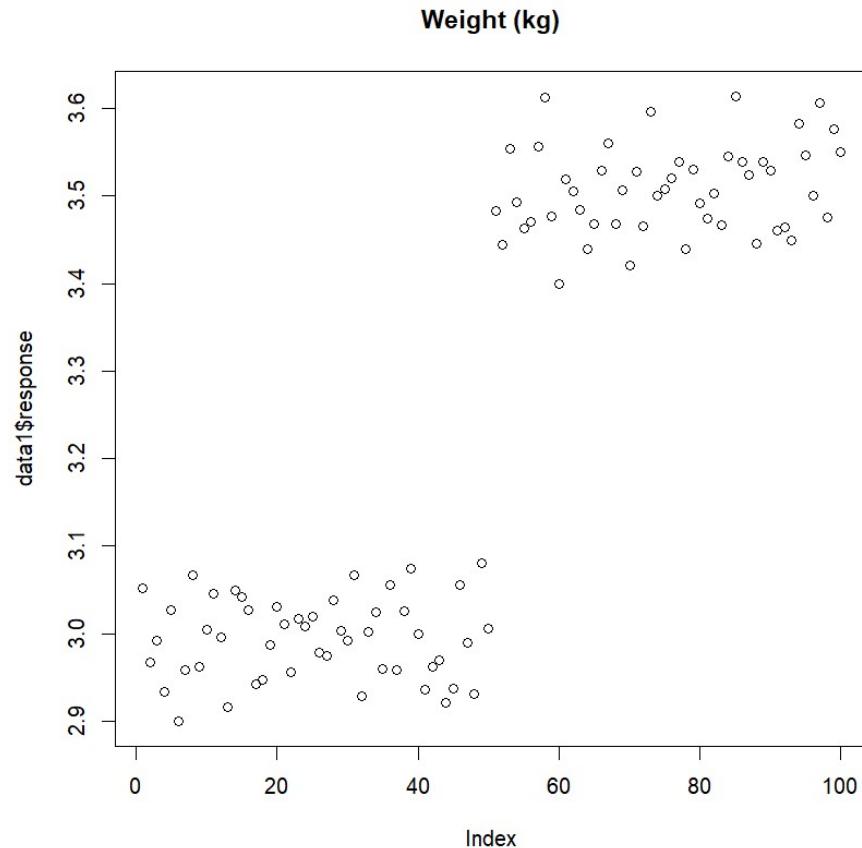
Is there anything about this design that might lead to datum being correlated with each other? For example, if we had repeated measures on the same patient (chicken) then we would expect these to be correlated i.e. dependant on each other.

YES! Chickens in the same treatment might be correlated, but our model will account for that since it's fitting a different mean to the control vs treatment.

Other reasons might be: Blocked design, Split-Plot design, etc.

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Independence and Outliers: Plot the data using a Serial Plot



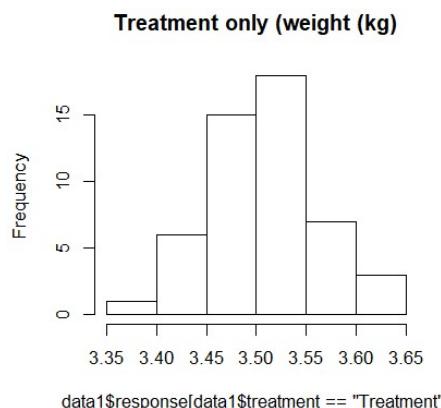
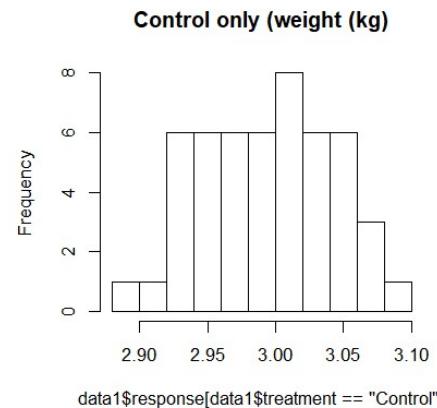
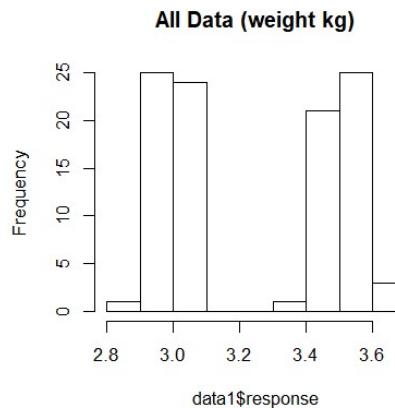
Notice the serial correlation i.e. data at the start are more similar to those at the end. As control data are at the beginning and Treatment at the end this is expected. And our model will account for that.

No Outliers

```
plot(data1$response, main="Weight (kg)")
```

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Normality and Outliers



```
par(mfrow=c(2,2))
```

```
hist(data1$response, main="All Data (weight kg)")
```

```
hist(data1$response[data1$treatment=="Control"], main="Control only (weight (kg))")
```

```
hist(data1$response[data1$treatment=="Treatment"], main="Treatment (weight (kg))")
```

The combined data is clearly bimodal and is certainly not normal!!!

YR Turn: So do we have a problem??

NO: The error needs to be normal, not the response. And as we can see here the error about the mean of each treatment is roughly normal.

(Even though the control might not look like it we know it is since its simulated data. A good example of just how non-normal something can look and we're still OK).

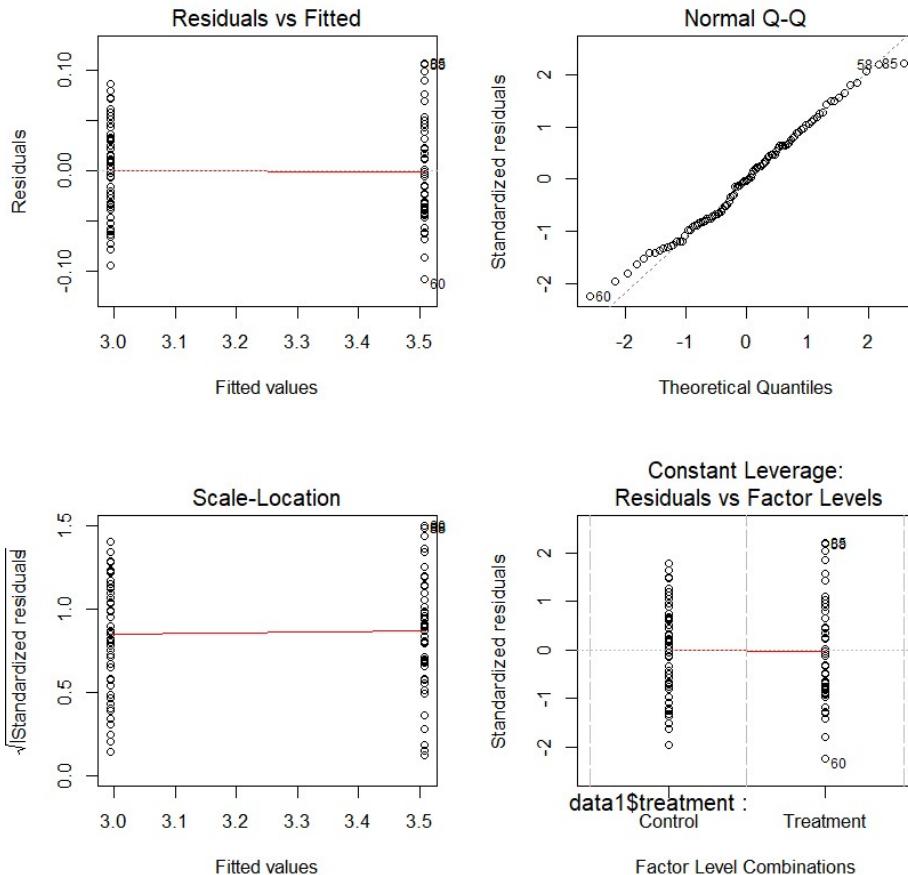
Step 2) Fit the Model

R Code:

```
lm.anova <- lm(data1$response~data1$treatment)
```

Step 3) Check Model Assumptions via Diagnostics: Residual Analysis

Normality



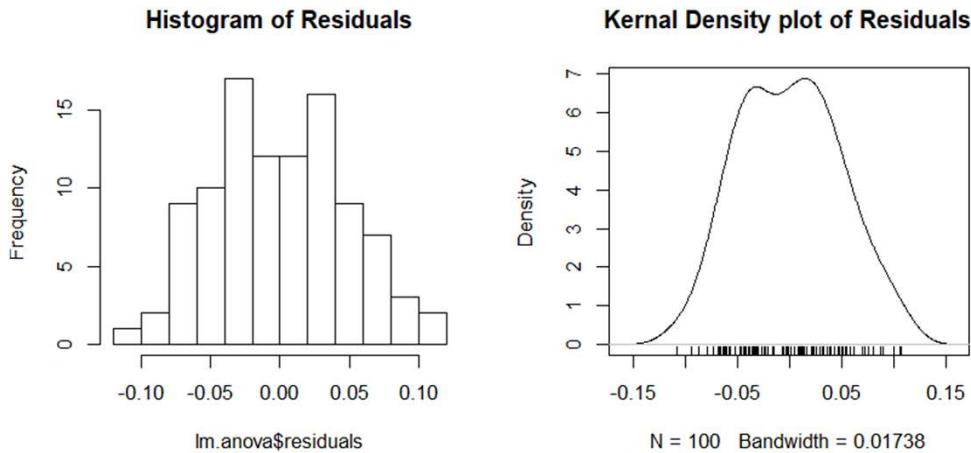
Residuals appear normal.

```
par(mfrow=c(2,2))  
plot(lm.anova)
```



Step 3) Check Model Assumptions via Diagnostics: Residual Analysis

Normality

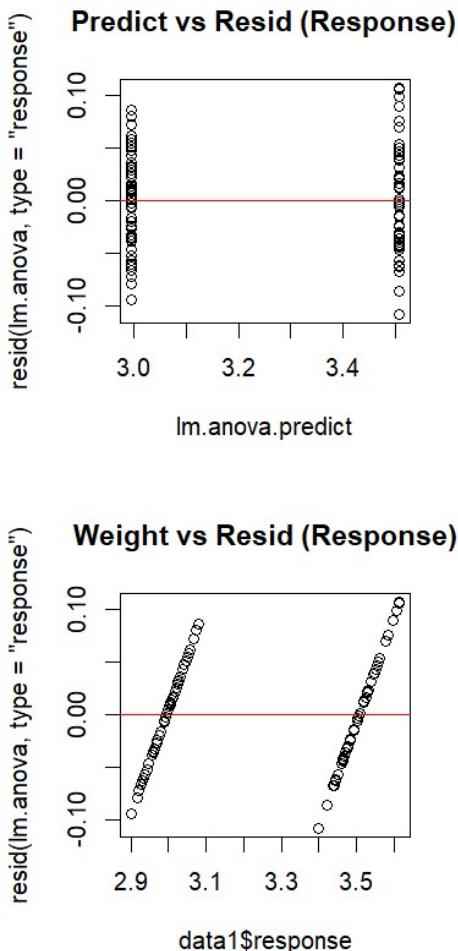


Residuals appear normal.

```
par(mfrow=c(2,2))
hist(lm.anova$residuals, main="Histogram of Residuals")
plot(density(lm.anova$residuals), main="Kernal Density plot of Residuals")
rug(lm.anova$residuals)
```

Step 4) Goodness of Fit: Residual Analysis

Outliers and unexplained structure or non linearity

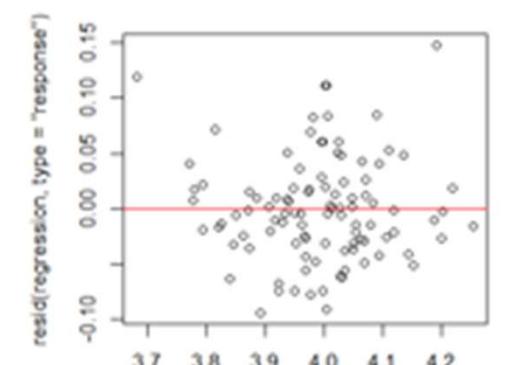


No evidence of outliers, or unexplained structure or non linearity.

We expect the ‘lines’ of data rather than a random ‘cloud’ of data which we saw in the regression (bottom right chart). This is because rather than a range of predictions for each different value of the predictor (feed) we only get 1 prediction for control and another for treatment, hence 2 vertical lines in the upper chart.

And 2 diagonal lines in the bottom chart when the x axis is the actual response since these are different.

The greater the difference between the groups the further these lines are apart.



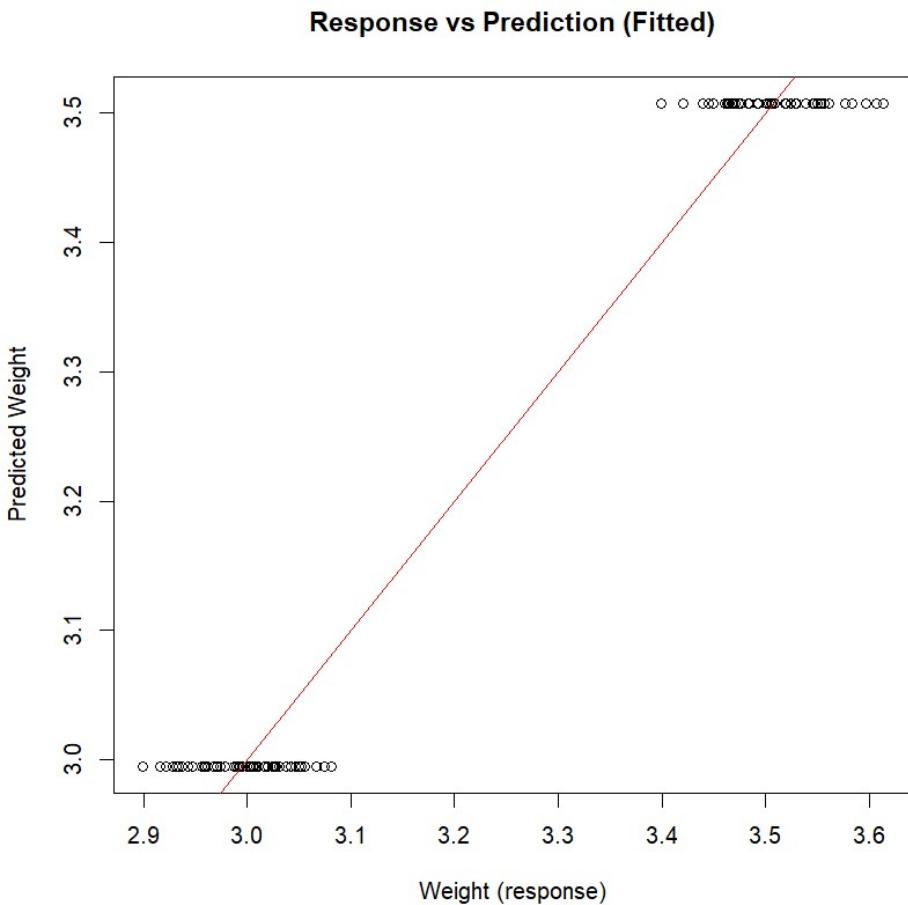
```
par(mfrow=c(2,2))
plot(lm.anova.predict, resid(lm.anova, type="response"), main="Predict vs Resid (Response)") # response residuals
abline(h=0, col="red")
plot(data1$response, resid(lm.anova, type="response"), main="Weight vs Resid (Response)") # response residuals
abline(h=0, col="red")
```

CQ: If I had 4 treatments, how many lines would I have?

- A. 2 lines
- B. 4 lines
- C. 8 lines
- D. 12 lines

Correct, 1 line for each treatment

Step 4) Goodness of Fit: Plots and Statistics

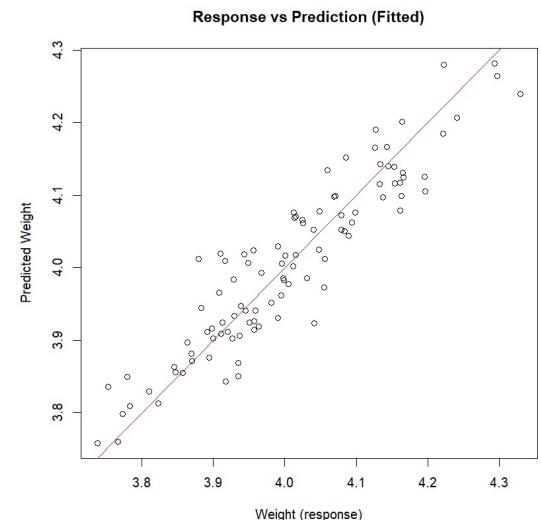


For the same reason used previously we expect 2 lines of data here, not a cloud of points i.e. we only have 2 prediction.

We expect the 2 lines of data to be centred on the red line.

If they aren't this suggests there is some bias to the fit worth investigating further.

For comparison Linear Regression looks like this



```
plot(data1$response, lm.anova.predict, main="Response vs Prediction  
(Fitted)", xlab="Weight (response)", ylab="Predicted Weight")  
abline(a=0, b=1, col="red")
```

Step 5) Interpret Model Parameters and reach a conclusion

R CODE and output used to create Tables

```
> summary(lm.anova)
```

Call:

```
lm(formula = data1$response ~ data1$treatment)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.108335	-0.036977	-0.001368	0.032338	0.106723

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.994670	0.006894	434.36	<2e-16 ***
data1\$treatmentTreatment	0.512748	0.009750	52.59	<2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04875 on 98 degrees of freedom

Multiple R-squared: 0.9658, Adjusted R-squared: 0.9654

F-statistic: 2766 on 1 and 98 DF, p-value: < 2.2e-16

```
> confint(lm.anova)
```

	2.5 %	97.5 %
(Intercept)	2.9809884	3.0083521
data1\$treatmentTreatment	0.4933985	0.5320966

Step 5) Interpret Model Parameters and reach a conclusion

Parameter	Estimate	SE	T score	P value	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant / Control (β_0)	3.00	0.0069	434	<2e-16	2.98	3.01
Treatment Effect (β_1)	0.51	0.0098	53	<2e-16	0.49	0.53

Model Fit is:

$$Y_i = \beta_0 + X_i \beta_1 + \varepsilon_i \text{ (same as the previous linear regression)}$$

$$\text{Weight} = 3.00 + 0.51(\text{if treatment}) + \varepsilon_i$$

Step 6) Reporting: Overall Conclusion suitable for publication

“There is strong evidence to show that the Treatment influences weight ($p < 2e-16$). It increases weight by between 0.49-0.53 kg (95% CI), from an average of approximately 3 (95% CI=2.98-3.01). This effect on weight has been estimated very accurately [as 95% CI is quite narrow].

The model is a good fit to the data with an $R^2 = 97\%$. There were no outliers or unexplained structure. The error was normal”

When giving a p-value always give an estimate of the effect size as well i.e. the 95% CI.

NB: In the real world since $R^2 = 97\%$ this is very likely a poor model due overfitting.



THE UNIVERSITY OF
SYDNEY

Sydney
Informatics Hub

Combination of ANOVA and Regression

Continuous response, categorical and continuous predictors

Workflow Suitable for:

- **Modelling a combination of discrete and continuous predictors** (workflow shown is for 1 of each type of predictor, there are additional considerations when more than 1 e.g. confounding and multicollinearity, , these are discussed in our Model Building workshop)
- Modelling more than 1 regression line
- To test if multiple regression lines are the same, or different.
- ANCOVA: ANalysis of COVAriance
- BACI (Before After Control Impact Designs)

Model Fitting Workflow

Step 0) Clean and check data.

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA).

Step 2) Fit the Model

Step 3) Check Model Assumptions via Diagnostics: Residual Analysis

Step 4) Goodness of Fit: Plots and Statistics

Step 5) Interpret Model Parameters and reach a conclusion

Step 6) Reporting

Linear Models 3 and Model Building Workshops have more detail on many of these steps.

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Your Turn:

Say we wanted to do the previous 2 experiments **at the same time**.

Plot the data!

Reminder:

Experiment 1

A linear model for the weight of chicken compared to the amount of feed it eats in its first month.

Experiment 2

We added a protein supplement. We expect the Control to have an average of 3kg, and the Treatment to add 0.5kg to weight. The SD = 0.05 i.e. 95% are within 0.10 of the mean.

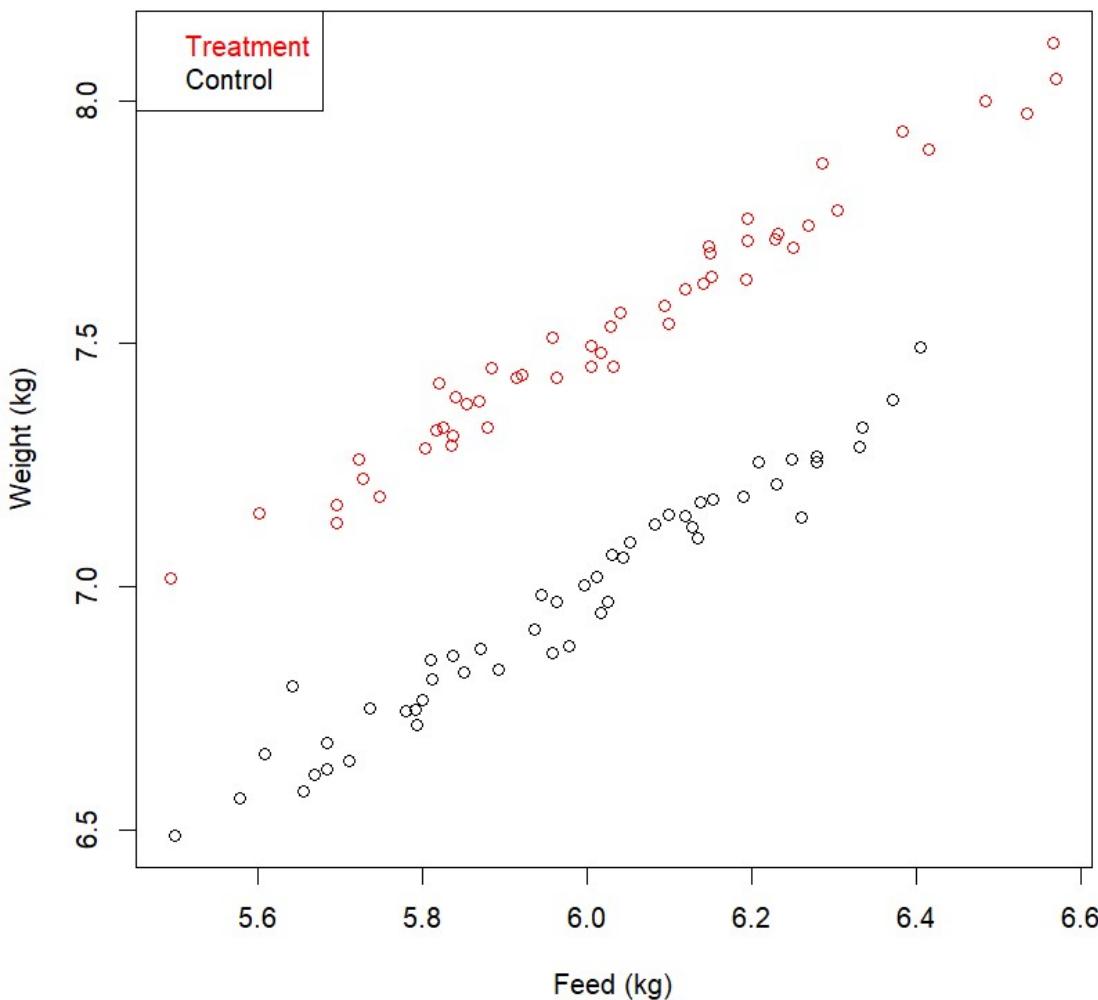


“Graphs allow us to view complex mathematical models fitted to data, and they allow us to assess the validity of such (statistical) models”

(Cleveland 1994, author of “*The elements of graphing data*” and “*Visualising data*”).

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Supplement has no impact on feed's relationship with weight



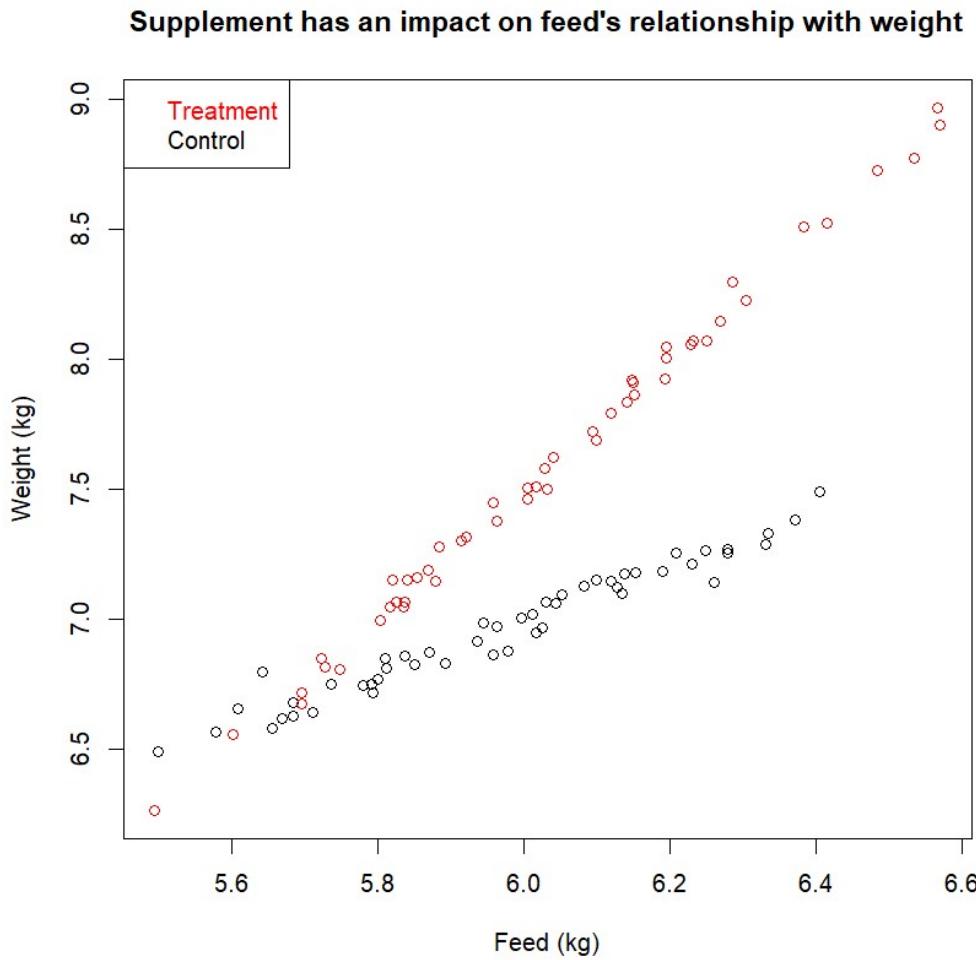
So effect of feed is the same across treatment and control.

But what if they “interact”?

How do we fit this?



Yr Turn. But what if the protein supplement boosted the impact of feed. What would we see then? Draw it.



Now we see the treatment has little impact at the lower end feeding.

But as the amount we feed them increases it starts to have an impact.

Maybe because at the lower end they are only getting enough for basic development and they need more feed to really grow.

```
plot(data3$predictor.linear1, data3$response, xlab="Feed (kg)", ylab="Weight (kg)", main = "Supplement has an impact on feed's relationship with weight")
points(data3$predictor.linear1[data3$treatment=="Treatment"], data3$response[data3$treatment=="Treatment"], col="red")
legend(x="topleft", legend=c("Treatment", "Control"), text.col=c("red", "black"))
```

Different Interpretation

Parallel lines (ANCOVA)

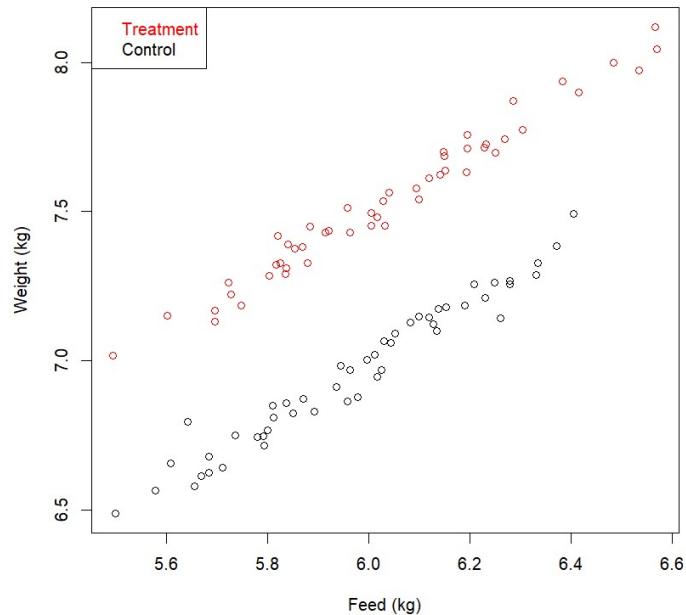
We can talk about the **consistent** impact of the:

- Protein treatment, in terms of the extra amount of weight it adds compared to the control
- Feed, in terms of the extra amount of weight it adds for each kg of feed.

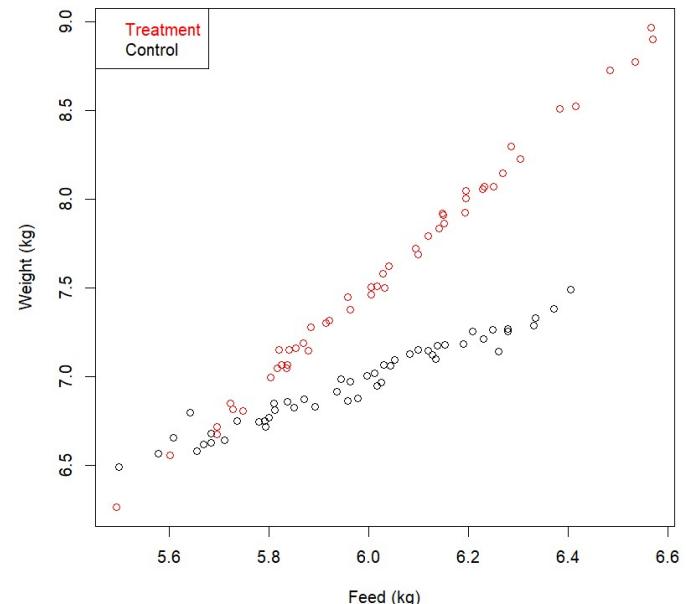
Non Parallel lines

As there is no consistent impact we need to talk about 2 different regression lines, each with a different impact of feed on weight.

Supplement has no impact on feed's relationship with weight



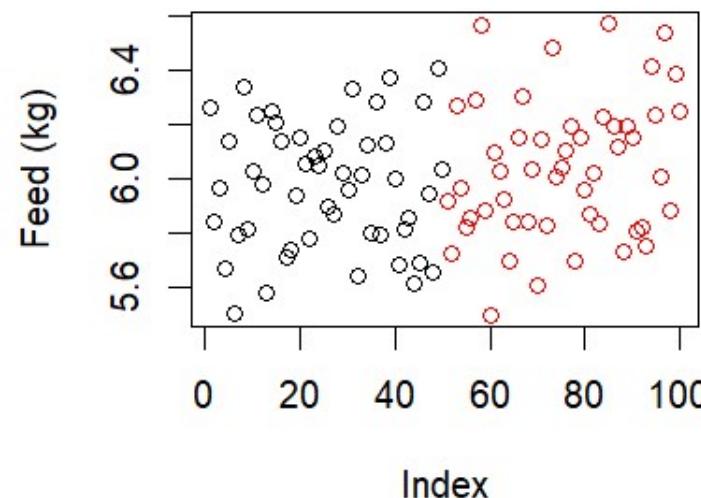
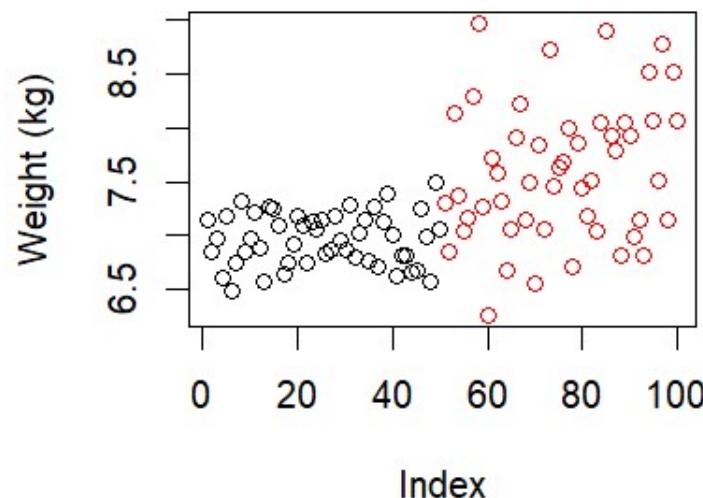
Supplement has an impact on feed's relationship with weight



Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Independence: Consider your experimental design and serial plot

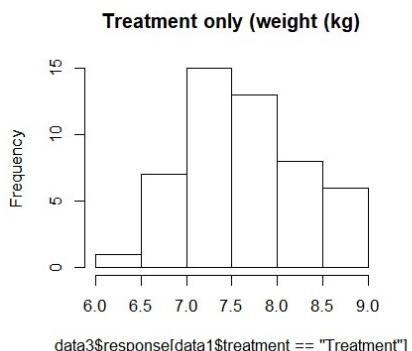
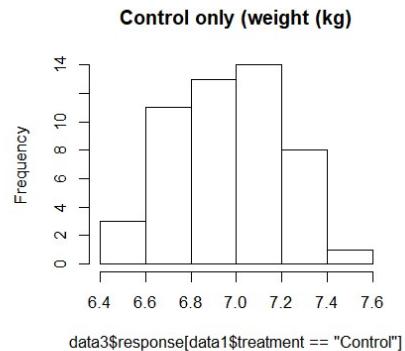
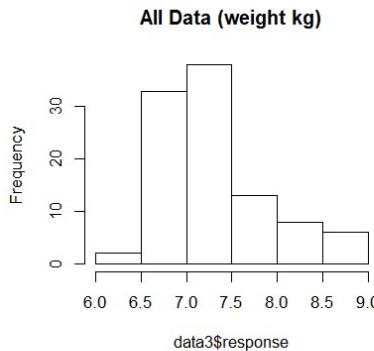
As with the ANOVA we expect there might be dependence within each treatment for the response. However the linear predictors (feed) should be independent, if they're not then we have a big problem!



```
par(mfrow=c(1,2))
plot(data3$response, col=ifelse(data3$treatment=="Treatment","red", "black"), ylab="Weight (kg)")
plot(data3$predictor.linear1, col=ifelse(data3$treatment=="Treatment","red", "black"), ylab="Feed (kg)")
```

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Normality and Outliers



The combined data is clearly skewed and is certainly not normal!!!

Which is what we would expect given that both treatments have the same response at low Feed, but one of them has higher weight at a higher Feed.

If we didn't include treatment this is an example of where our residuals might not be normal and it's **because of missing structure i.e. treatment**.

```
par(mfrow=c(2,2))
hist(data3$response, main="All Data (weight kg)")
hist(data3$response[data1$treatment=="Control"], main="Control only (weight (kg))")
hist(data3$response[data1$treatment=="Treatment"], main="Treatment (weight (kg))")
```

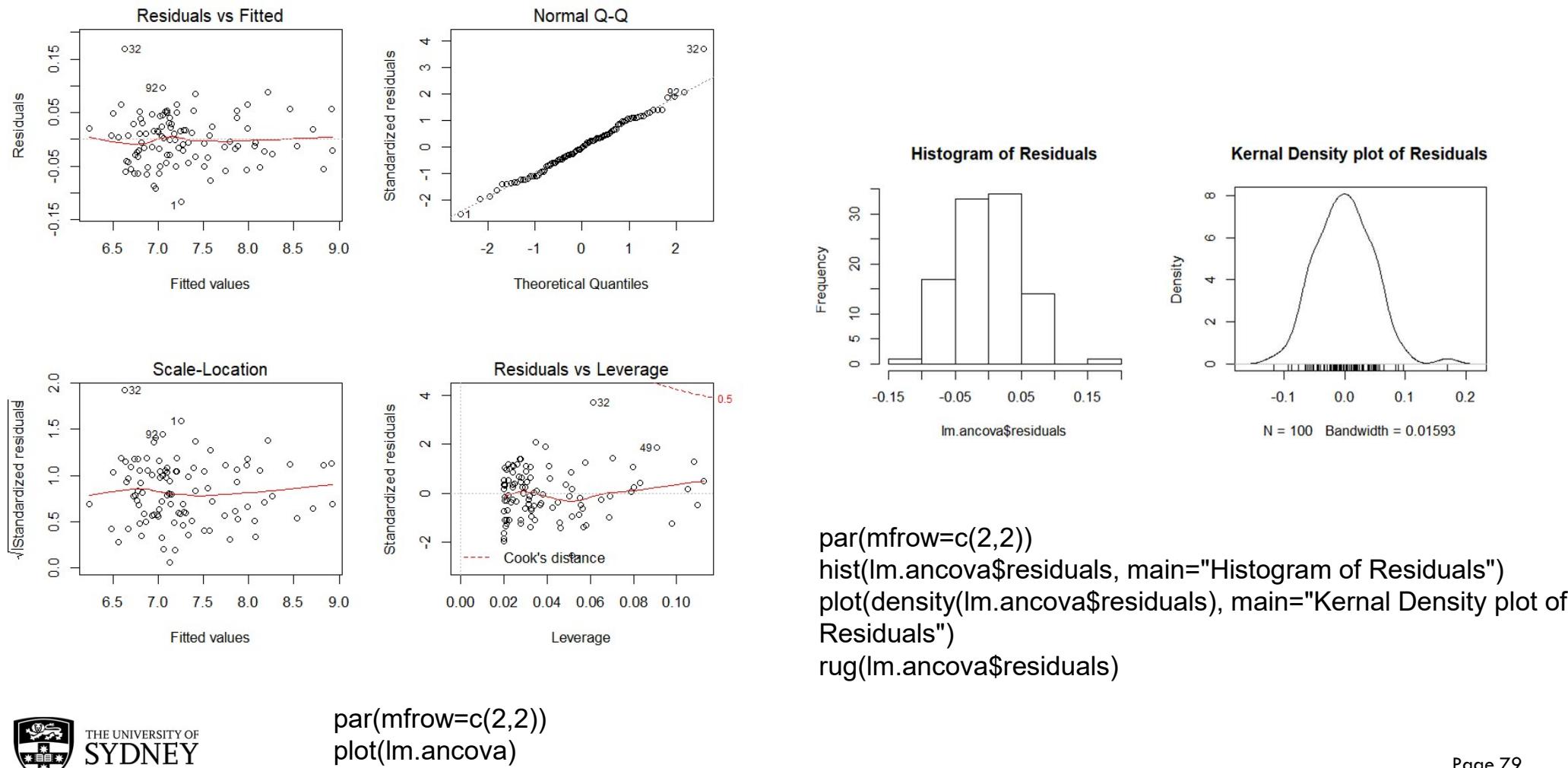
Step 2) Fit the Model

R Code:

```
lm.ancova <-
lm(data3$response~data3$treatment*data3$predictor.linear1)
```

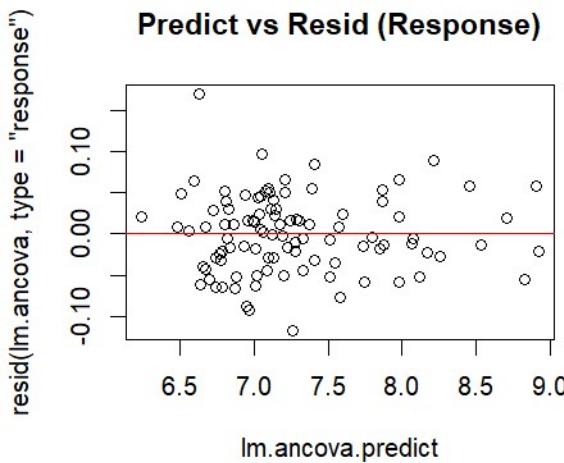
Step 3) Check Model Assumptions via Diagnostics: Residual Analysis

Normality Residuals appear normal.



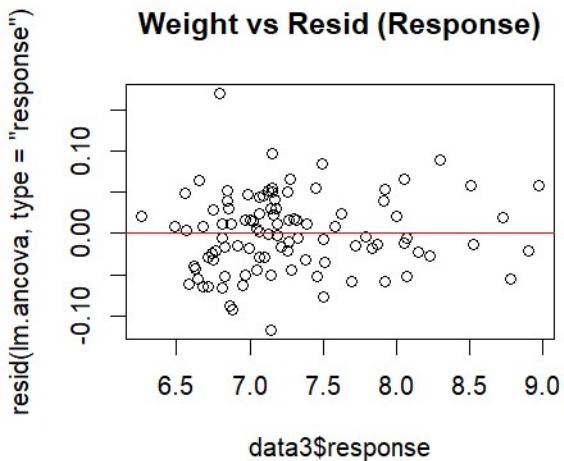
Step 4) Goodness of Fit: Residual Analysis

Outliers and unexplained structure or non linearity



No evidence of outliers, or unexplained structure or non linearity.

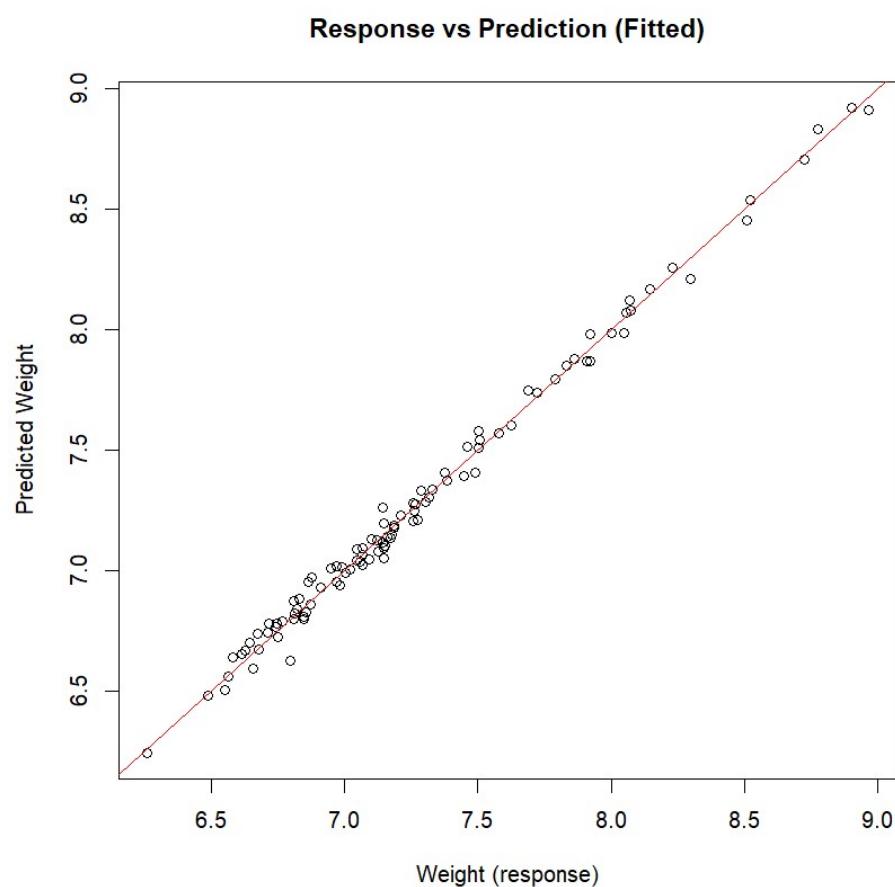
Although we don't have the diagonal lines we saw in ANOVA it is possible. It occurs when the treatment has a much bigger effect than the linear predictor.



And notice that the data gets a little sparse on the right, that's because only the treatment has these high predictions, while both of them have the low ones.

```
par(mfrow=c(2,1))
plot(lm.ancova.predict, resid(lm.ancova, type="response"), main="Predict vs Resid (Response)" # response residuals
abline(h=0, col="red")
plot(data3$response, resid(lm.ancova, type="response"), main="Weight vs Resid (Response)" # response residuals
```

Step 4) Goodness of Fit: Plots and Statistics



Looks like a good fit!

```
plot(data3$response, lm.ancova.predict, main="Response vs Prediction (Fitted)",  
xlab="Weight (response)", ylab="Predicted Weight")  
abline(a=0, b=1, col="red")
```

Step 5) Interpret Model Parameters and reach a conclusion

R CODE and output used to create Tables

```
> summary(lm.ancova)

call:
lm(formula = data3$response ~ data3$treatment * data3$predictor.linear1)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.11675	-0.02979	-0.00096	0.02979	0.16921

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.85896	0.17325	4.958	3.07e-06 ***
data3\$treatmentTreatment	-8.32034	0.23573	-35.296	< 2e-16 ***
data3\$predictor.linear1	1.02220	0.02898	35.269	< 2e-16 ***
data3\$treatmentTreatment:data3\$predictor.linear1	1.47117	0.03924	37.490	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04715 on 96 degrees of freedom

Multiple R-squared: 0.9934, Adjusted R-squared: 0.9932

F-statistic: 4846 on 3 and 96 DF, p-value: < 2.2e-16

```
> confint(lm.ancova)
```

	2.5 %	97.5 %
(Intercept)	0.5150596	1.202870
data3\$treatmentTreatment	-8.7882616	-7.852416
data3\$predictor.linear1	0.9646711	1.079732
data3\$treatmentTreatment:data3\$predictor.linear1	1.3932757	1.549063

Step 5) Interpret Model Parameters and reach a conclusion

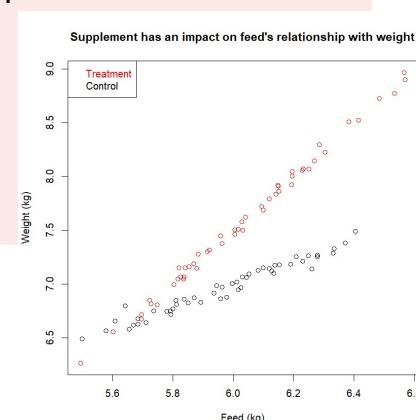
Parameter	Estimate	SE	T score	P value	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant Control (β_0)	0.86	0.17	5	<3e-6	0.51	1.2
Constant Adjustment Treatment (β_1)	-8.32	0.24	-35	<2e-16	-8.8	-7.9
Slope Control (β_3)	1.0	0.029	35	<2e-16	0.96	1.08
Slope Adjustment Treatment (β_4)	1.5	0.039	37	<2e-16	1.39	1.55

Model Fit is $\Rightarrow Y_i = \beta_0 + X_i\beta_1 + X_i\beta_3 + X_i\beta_4 + \varepsilon_i \Rightarrow$

Weight = $0.86 + 1.0 * \text{Feed} - 8.32(\text{if treatment}) + 1.5 * \text{Feed}(\text{if treatment}) + \varepsilon_i$

Weight of Control (**black** data in chart) = $0.86 + 1 * \text{Feed} + \varepsilon_i$

Weight of Treatment (**red** data in chart) = $-7.46 + 2.5 * \text{Feed} + \varepsilon_i$



Step 6) Reporting: Overall Conclusion suitable for publication

“There is strong evidence to show that feed impacts weight ($p < 2e-16$), with each kg of feed adding between 0.96-1.08 kg of weight (95% CI).

There is strong evidence that Protein supplements have a positive effect on the impact of Feed ($p < 2e-16$), increasing its effect by between 1.39-1.55 (95% CI), for a total average effect of 2.5kg weight increase for each kg of extra Feed.

This effect of feed on weight has been estimated very accurately [as 95% CI is quite narrow].

The model is a good fit to the data with an $R^2 = 99\%$. There were no outliers or unexplained structure. The error was normal”

When giving a p-value always give an estimate of the effect size as well i.e. the 95% CI.

NB: In the real world since $R^2 = 99\%$ this is almost certainly a poor model due overfitting, or some other problem.

ANCOVA: is a special case of this model

Adjusts for continuous covariates so we get a clean read on the discrete predictors impacts. Often used in observational studies to help remove the effect of covariates.

For example: To understand the effect of the protein supplement after accounting for the different amount of feed each chicken ate we can add feed is a covariate in an ANCOVA. This would account for the scenario where chickens that had the supplement happened to eat more food and as such weighed more for that reason, not due to the supplement.

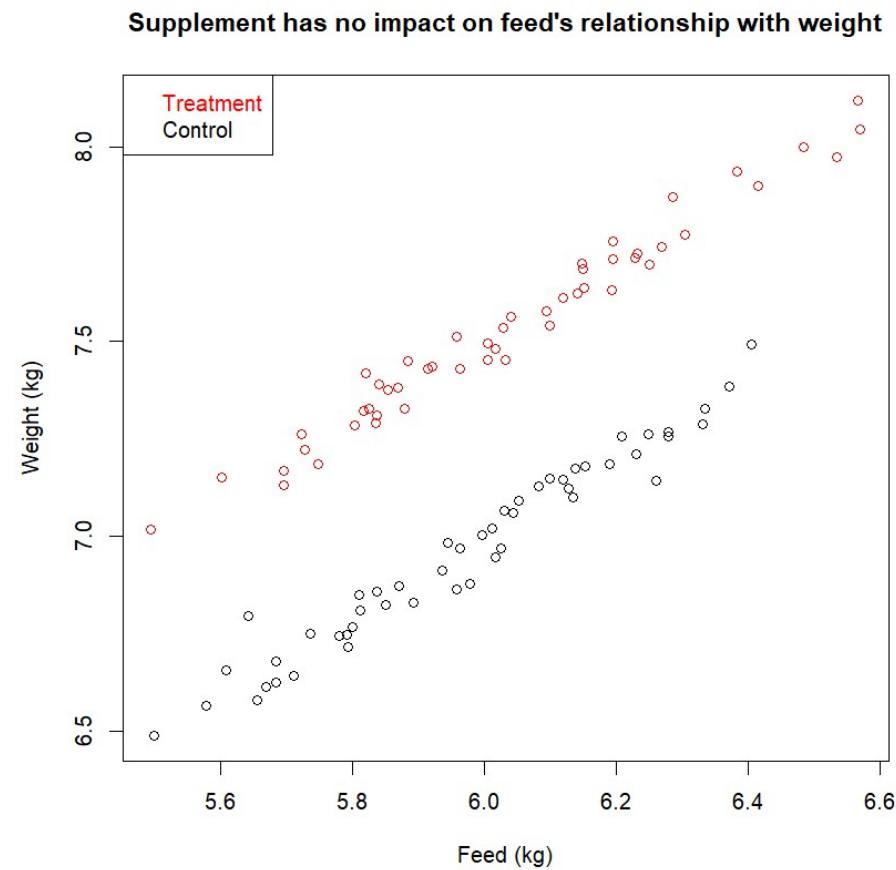
The key difference is that an ANCOVA makes an additional assumption called **Homogeneity of covariate regression coefficients; i.e. “parallel lines model”**. Which states that the regression lines must be parallel, i.e. the covariate has the same effect for each treatment.

ANCOVA: is a special case of this model

This allows us to measure the effect of each discrete parameter after accounting for the continuous covariate.

For example: The below model shows that the protein supplement increases the chickens weight by 0.5 kg, irrelevant to amount of feed it ate.

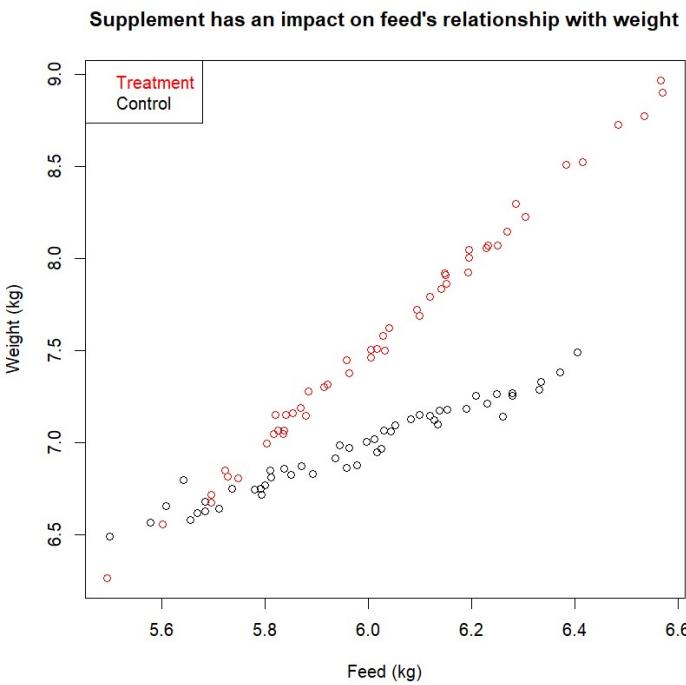
Statistically the Homogeneity of covariate regression coefficients; i.e. “parallel lines model” means the ***interaction is not required in the model.***



ANCOVA: what happens when the homogeneity of regression covariates is failed?

Don't worry! It's not a big deal. It just means that the covariate doesn't have a consistent effect overall treatments. Meaning we can't directly compare the treatments overall effects with each other and instead need to look at each treatments regression line.

Statistically it's the same model, but we also include an interaction.



So rather than the protein supplement consistently increasing weight by 0.5kg we see it has little impact at the lower end feeding.

But as the amount we feed them increases it starts to have an impact.



THE UNIVERSITY OF
SYDNEY

Sydney
Informatics Hub

Mixed Models: Random Intercept Model

Response is measured more than once on each respondent (sampling unit)

Workflow Suitable for:

- Modelling the variance associated with the respondents (sampling units). Usually gives a more accurate analysis by partitioning out the noise/variance associated with the respondents (sampling units).
- Repeated Measures
- Longitudinal Analysis
- More advanced workflows suitable for:
 - Cluster Designs
 - More complex designs with repeated measures on clusters of sampling units and experimental units
 - Variance Decomposition
 - Random Slopes



THE UNIVERSITY OF
SYDNEY

Model Fitting Workflow

Step 0) Clean and check data.

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA).

Step 2) Fit the Model

Step 3) Check Model Assumptions via Diagnostics: Residual Analysis

Step 4) Goodness of Fit: Plots and Statistics

Step 5) Interpret Model Parameters and reach a conclusion

Step 6) Reporting

Linear Models 3 and Model Building Workshops have more detail on many of these steps.

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

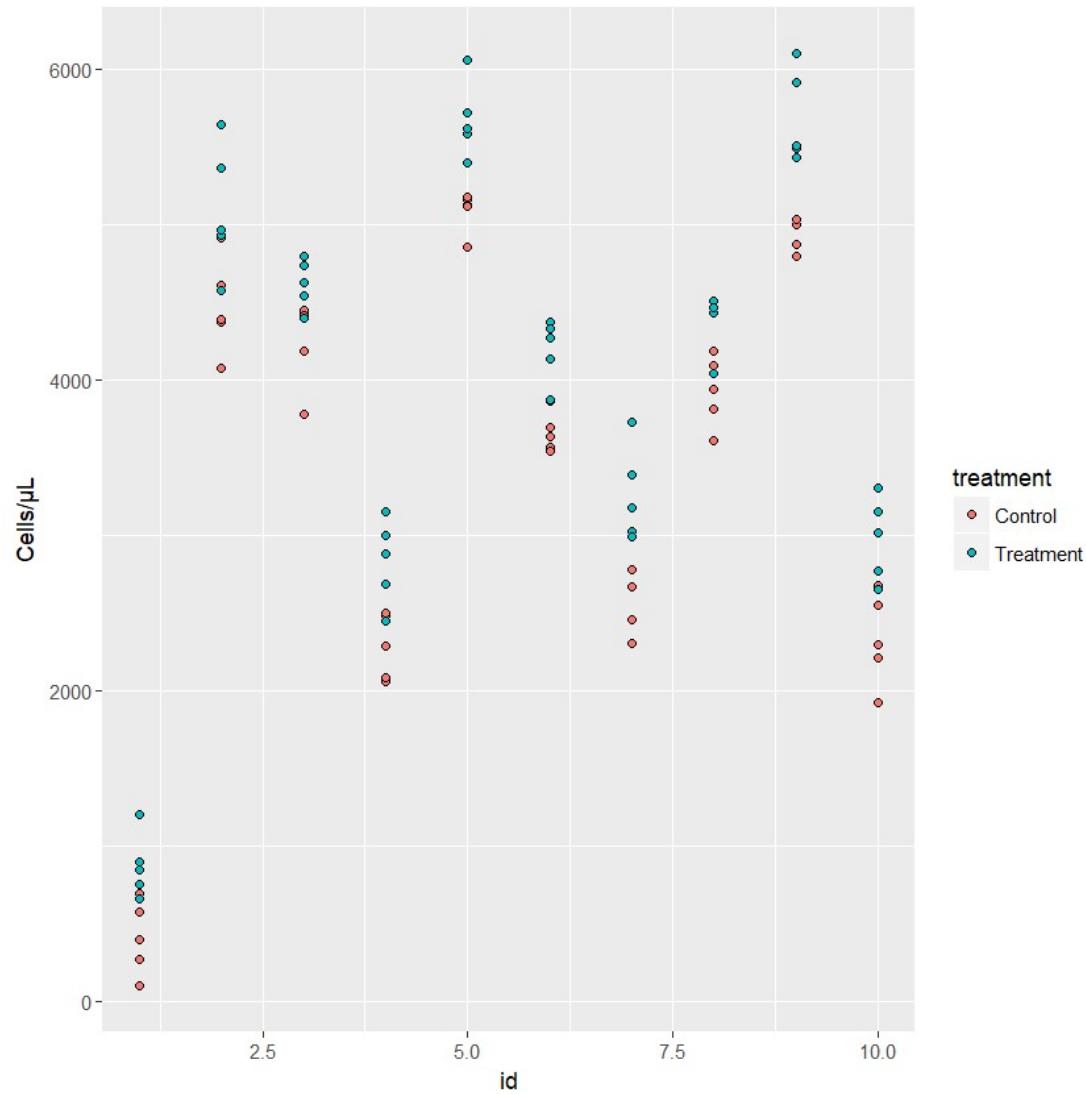
Your Turn:

Say we wanted to test the impact of a new drug on white blood cell counts in immune deficient people/dogs/Tasmanian tigers/chickens. We have 10 “people”, we take 5 measurements before the treatment and 5 after.

The white blood cell count is between 1000-7000 cells/micro litre (cells/ μ L). We expect the drug to increase white blood cell count by about 500 (cells/ μ L) to get it into the normal range. And within person variance is about 500 (cells/ μ L)

Plot the data!

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)



Notice how the difference between people is much bigger than the effect of the drug?

The models so far ignore this information.

A **Mixed Model** that includes person as a random effect accounts for this. Effectively removing this extra variance and making the model more accurate.

This is a classic example of where mixed models out perform those that ignore this extra info i.e. when the difference between sampling units is bigger than the effect we are looking for.

```
ggplot(data = data6, aes(x=id, y=response, fill=treatment)) +  
  geom_point(pch=21) +  
  labs(x="id", y="Cells ( $\mu$ L)")
```

Random Effects

Often reduce the noise by explicitly accounting for some of it.

- It's all about **noise:signal ratio**. If we can reduce the noise than we can detect a smaller signal. Giving a more accurate model when added to a fixed effects model meaning smaller effect sizes can be detected, smaller p-values for the same size effect and narrower Confidence Intervals.

Require a categorical variable (not a continuous one-if your model is failing to converge or looks wrong check that this variable has been defined as a categorical one, not doing so is a common mistake) with a unique level for each sampling/experimental unit e.g. a variable called ID where each respondent has its own code (usually numeric such as ID1, ID2, ID3, etc)

Require multiple sampling units and repeated measures for each. We generally need at least:

- 2 repeated measures within each sampling unit
- 5 sampling units (so we can estimate their variance)
- So a minimum of 10

Random Effects: A more efficient use of your Data

Say we want to understand the effects of new fishing nets on bycatch, but remove the effect of different boats due to various levels of experience, technology and size. We have randomly sampled 10 boats from the entire NSW East Coast Fleet. And we have measured the amount of bycatch from each for 2 treatments (existing and new nets).

We will always need at least 20 datum to estimate the effect of the 2 treatments (using the rule of thumb of 10 observations per fixed parameter).

The **question is how much sample do we need to remove the effect of the boats?**

- If we include each boat as a **fixed effect** i.e. old school Blocking, then we need approximately 10 additional parameters (1 for each boat).
 - For a total of **120 datum**. Using the rule of thumb of 10 datum per fixed parameter we would need 100 for the 10 boats and 20 for the 2 treatments. However, technically we only need 11 parameters and 110 datum since we need 1 for the constant/control for Boat A, 1 for the treatment, and 9 for the other boat interactions.
- But if we simply want to estimate the variance between the boats than all we need is 1 extra parameter, their variance as a **random effect**. We can do this by treating these boats as a random sample of all boats and since we don't really care what each got.
 - For a total of **20 datum** (using the rule of thumb for 10 datum per fixed parameter we need 20 for the 2 treatments, and using the rule of thumb of a minimum of 5 boats with 2 repeated measures to estimate the variance). Note that this is a simple non additive example, some might say we actually need **30 datum** as we are effectively estimating 3 parameters.

The point is that the Random effects method can be used on a **much smaller sample size** than the Blocking Method using Fixed Effects. (Irrelevant to exactly how we calculate the number of datum needed for each method)

Mixed models: Random & Fixed Effects

Fixed Effects

- Standard models you are used to.
- Measure a single **fixed effect** for **each** factor level i.e. 1 parameter for each factor level.
- So if we had 50 people and we want to understand the differences between them we need to estimate 50 fixed effects, 1 for each person. Which requires 50 parameters and using the rule of thumb of 10 observations per parameter means we need a sample size of 500.

Random Effects

- Measure the **randomness** of **all** factor levels e.g. if we had 50 people and we want to understand the amount of difference between all of them we could estimate the variance of their effects.
- 1 parameter for all people i.e. their variance. So 1 parameter in total.
 - Meaning we need a **much smaller sample size** (at least 10 as per previous slide)
- Usually added to a fixed effects model to make a mixed effects model. Or less often used by themselves to partition the Variance.



Challenge Question

- A Random effect is a Variance Estimate, and what do you need to estimate a Variance?
 - A) At least 1 data point

Wrong, since $n-1 = 0$ and we can't divide by 0. Also the variance is the average of the squared deviations from the average, and we can't have an average of 1!
 - B) At least 2 data points

Technically correct, BUT it won't be very stable or accurate. Trying to estimate random effects with only 2 datum per sampling unit will often fail to converge.
 - C) At least 5 data points

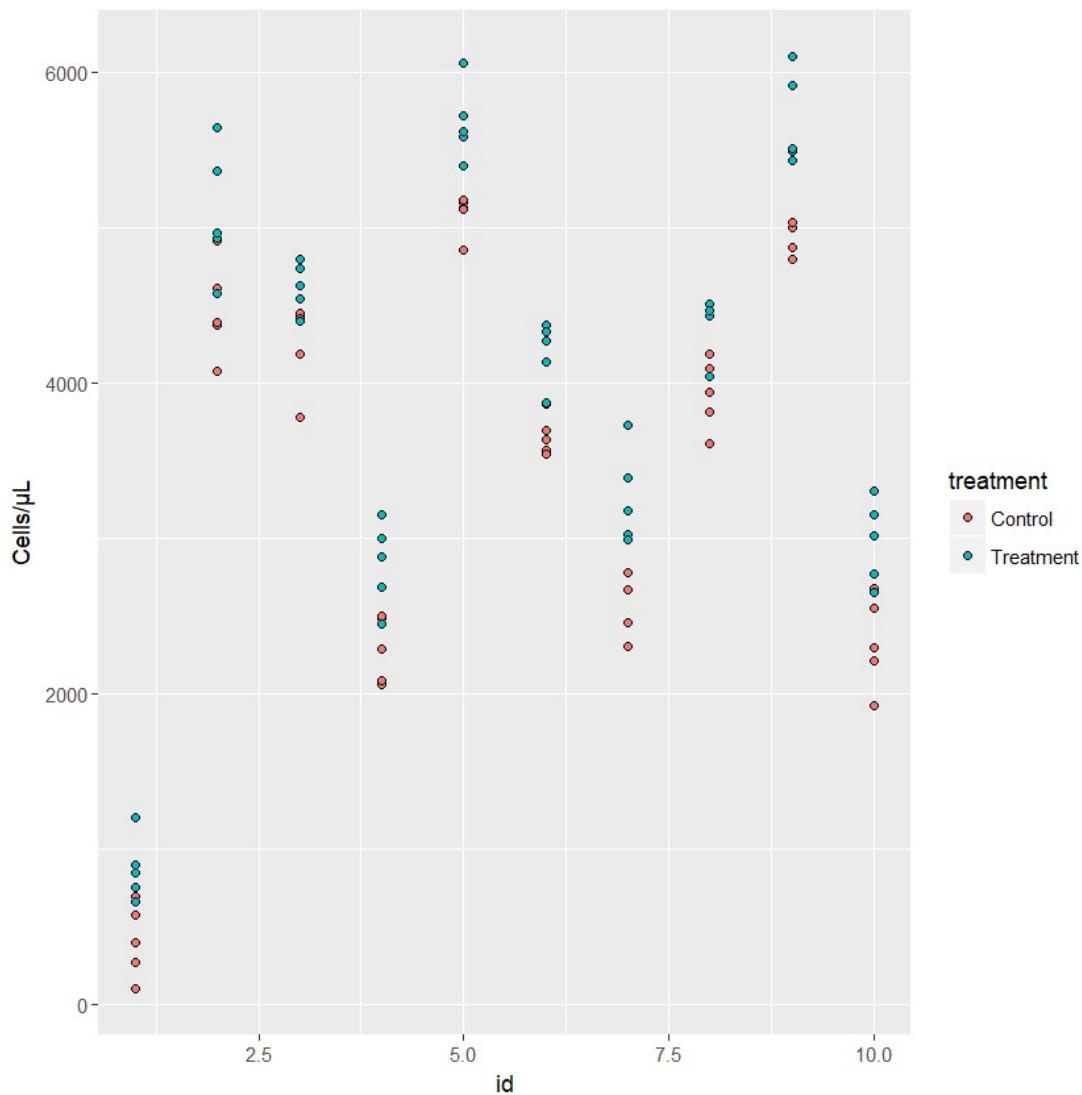
Often stated as the minimum # of sampling units for the model to converge to a stable result.
 - D) At least 30 data points

Often used as the minimum sample size required to invoke the Central Limit Theorem to assume averages are normal. However not needed for random effects.
 - E) At least 100 data points

Don't need this many

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA): Independence



If we started with a plot that factored in our design then we can clearly see the lack of independence between the patients and treatments.

Notice how the difference between people is much bigger than the effect of the drug?

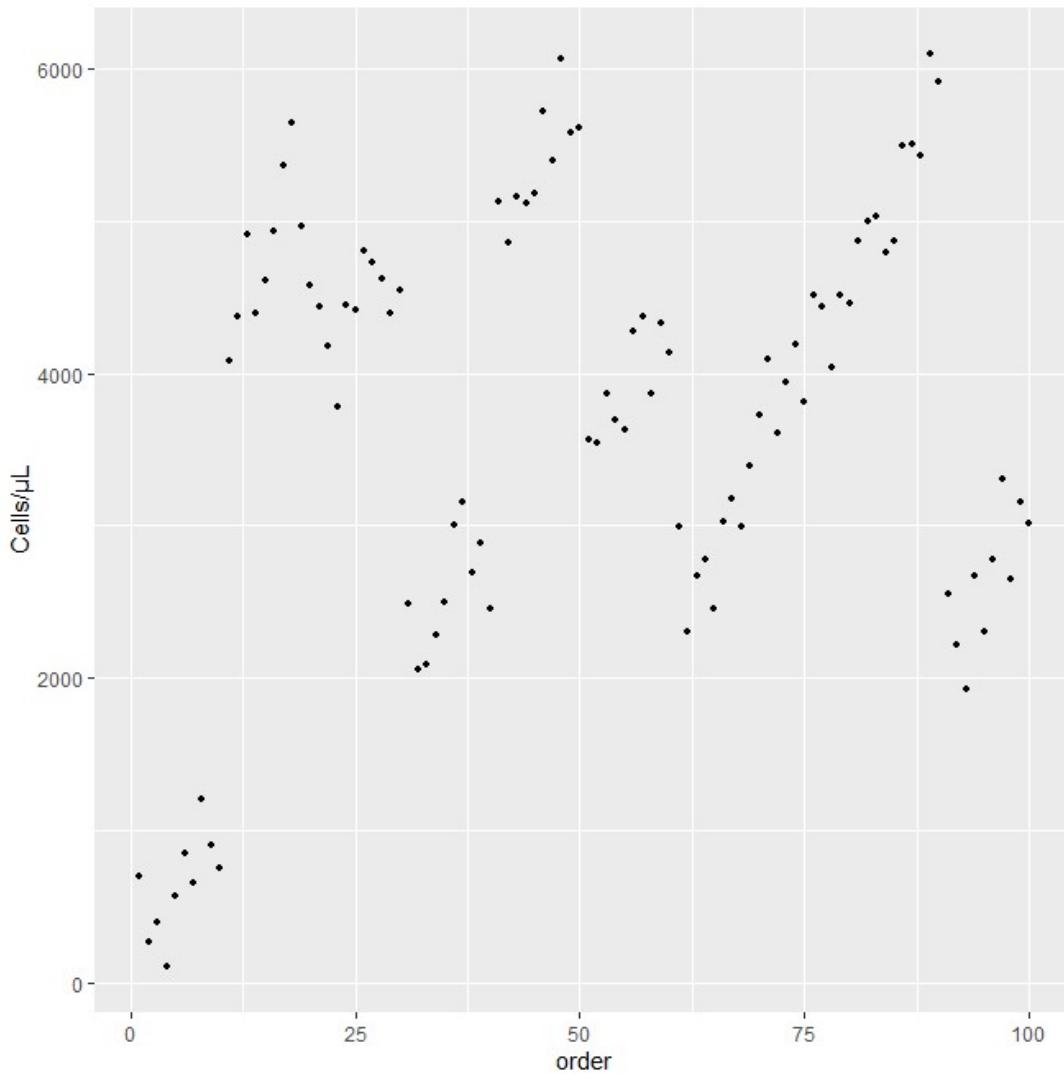
The models so far ignore this information.

A **Mixed Model** that includes person as a random effect accounts for this. Effectively removing this extra variance and making the model more accurate.

This is a classic example of where mixed models out perform those that ignore this extra info i.e. when the difference between sampling units is bigger than the effect we are looking for.

```
ggplot(data = data6, aes(x=id, y=response, fill=treatment)) +  
  geom_point(pch=21) +  
  labs(x="id", y="Cells (µL)")
```

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA): Independence



But say we hadn't factored in our experimental design like the preceding plot, maybe know one told us!?

If we followed this workflow we would have started with a **serial plot** to consider Independence.

And it would clearly show this clustering.

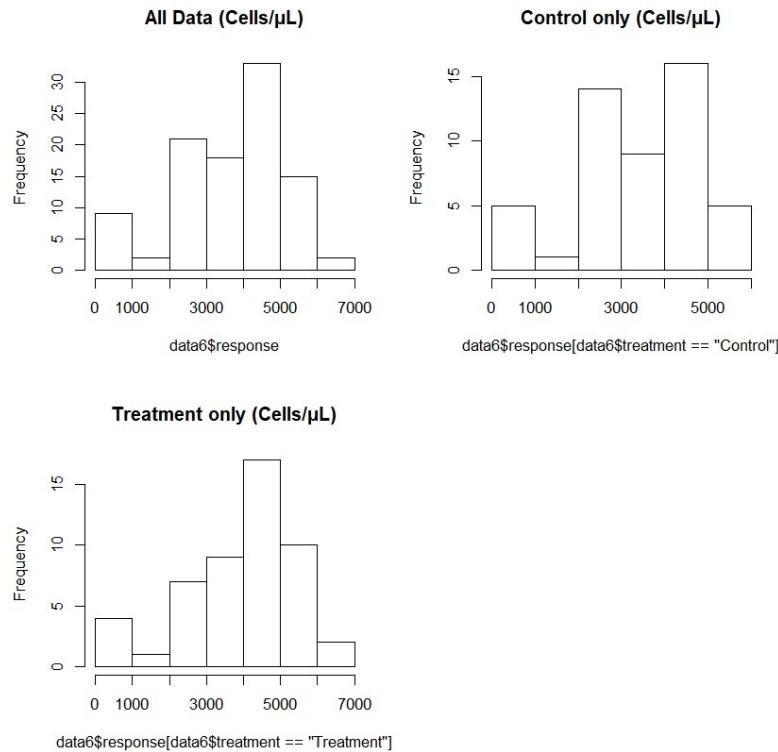
And if we saw something like this we would investigate and then realise it was due to the repeated measure and this is something we should include in our model.

Shows the importance of the serial plot and sticking to a workflow that starts with EDA

```
ggplot(data = data6, aes(x=c(1:nrow(data6)), y=response)) +  
  geom_point(pch=21, fill="black", size=1) +  
  labs(x="order", y="Cells/µL")
```

Step 1) Pick a suitable model to fit to the data via Exploratory Data Analysis (EDA)

Normality and Outliers



Could be normal, however there does look like there might be a bit of a negative skew.

But as the assumption is the model Errors are normal, not the response, we aren't too worried about this. But it's worth remembering and paying special attention to whether our model errors are normal.

```
windows()  
par(mfrow=c(2,2))  
hist(data6$response, main="All Data (Cells/μL)")  
hist(data6$response[data6$treatment=="Control"], main="Control only (Cells/μL)")  
hist(data6$response[data6$treatment=="Treatment"], main="Treatment only (Cells/μL)")
```

Sample Size calculations of repeated measures

- Even though we have n=100 this isn't really a lot of data.
- Keep in mind that we only have 10 subjects, and then 10 repeated measures on each.
- So using the n=10 per parameter rule even though we have n=100 this doesn't mean we can have 10 **subject level** parameters such as age. With only 10 subjects you'd be hard pressed to have even 2 groups with 5 in each.
- Repeated measures might allow us to evaluate more complex models with a few parameters for **longitudinal measures** as we have more points to model a line to e.g. white blood Cells/ μ L = virus load + cholesterol.

Step 2) Fit the Model

R Code:

```
lm.mm2 <- lmer(response~treatment + (1 | id), data=data6)
```

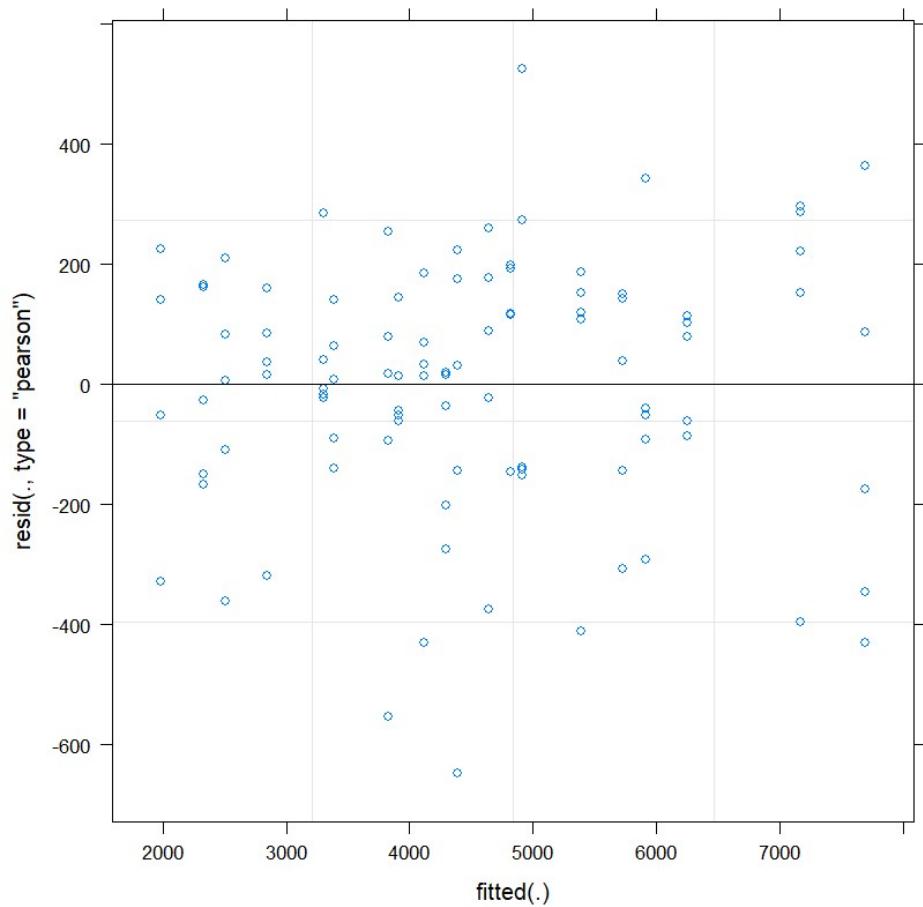
Where

treatment ~ is the fixed effect

(1 | id) ~ is the random effect representing each person. It is often
1,2,3,4,5 i.e. 1=1st person, 2=2nd person. BUT must be a categorical
variable (factor in R), **not continuous**.

Step 3) Check Model Assumptions via Diagnostics: Residual Analysis

Normality Residuals appear normal.

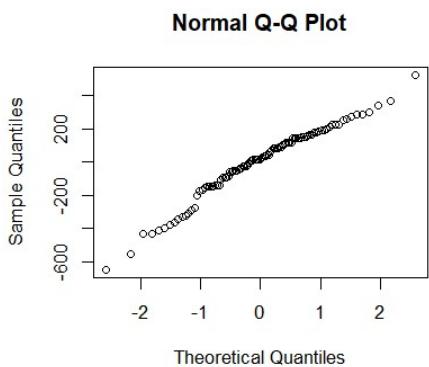
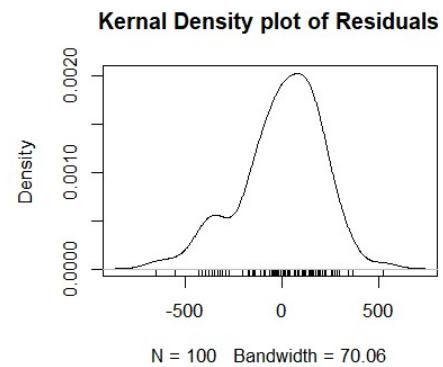
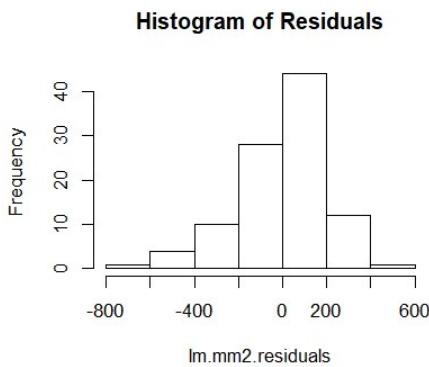


Standard plots we get from the R function `lmer()` to fit the model are different to what we get when we use `lm()`, which is what we have been using previously.

So we are missing the QQ plot, amongst others.

Step 3) Check Model Assumptions via Diagnostics: Residual Analysis

Normality Residuals appear normal.



Looks pretty good.

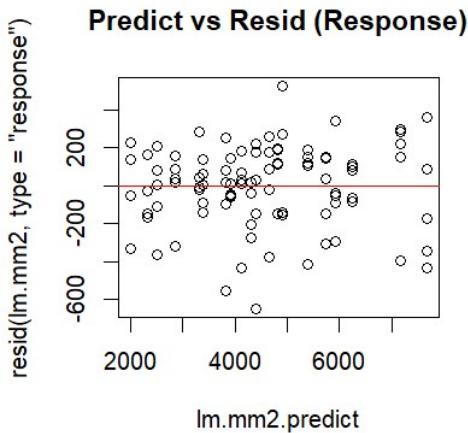
Might be a bit of a left skew
still, but likely not enough to
worry about.

```
lm.mm2.residuals <- resid(lm.mm2, type="response")
windows()
par(mfrow=c(2,2))
hist(lm.mm2.residuals, main="Histogram of Residuals")
plot(density(lm.mm2.residuals), main="Kernal Density plot of Residuals")
rug(lm.mm2.residuals)
qqnorm(lm.mm2.residuals)
```

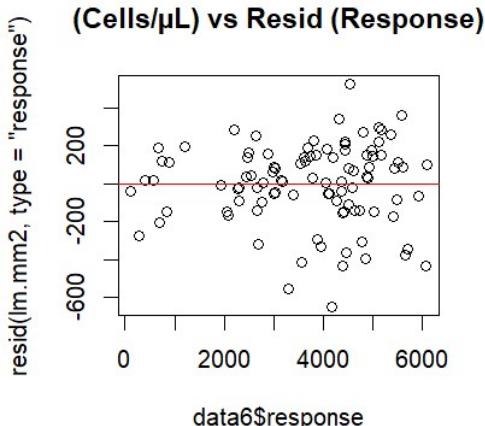


Step 4) Goodness of Fit: Residual Analysis

Outliers and unexplained structure or non linearity



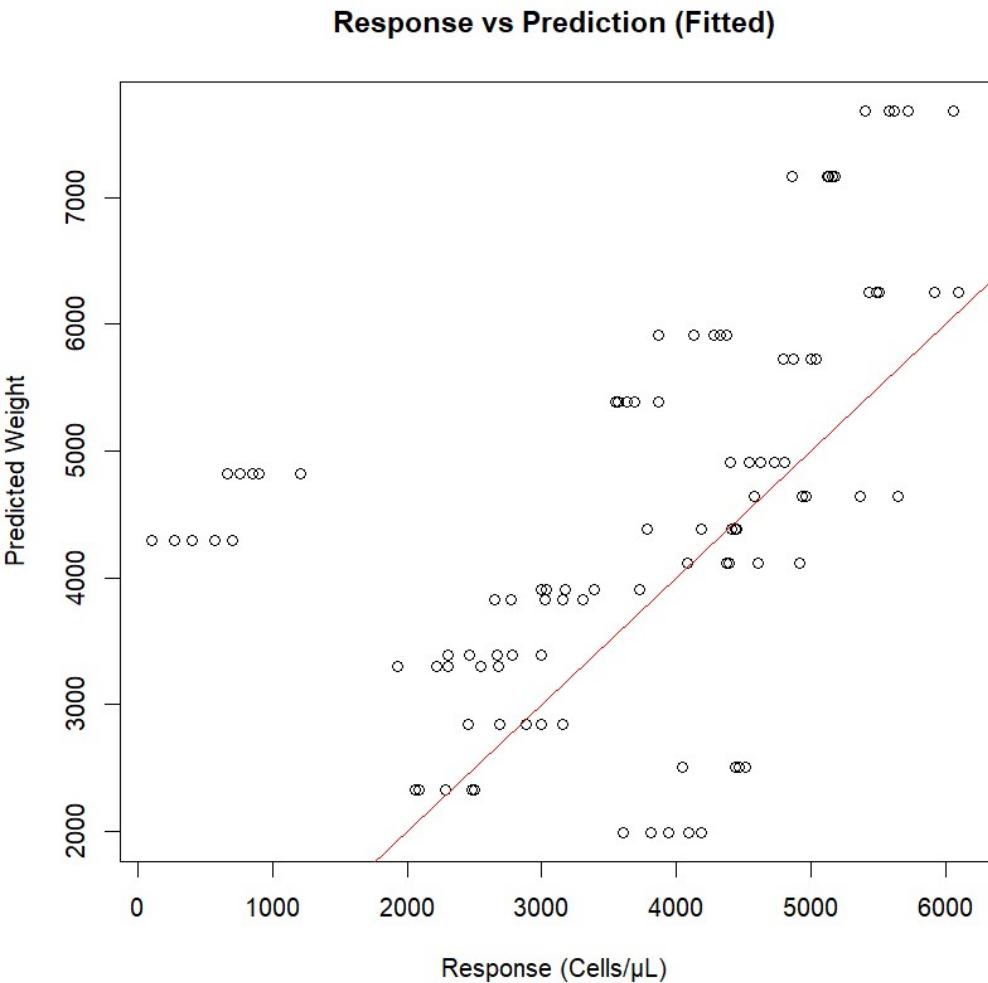
No evidence of outliers, or unexplained structure or non linearity.



Notice the predicted scores are falling out into 20 discrete vertical patterns of 5 points. This is expected since we had 5 repeated measures for 10 patients over 2 treatments.

```
par(mfrow=c(2,1))
plot(lm.mm2.predict, resid(lm.mm2, type="response"), main="Predict vs Resid (Response)" # response residuals
abline(h=0, col="red")
plot(data6$response, resid(lm.mm2, type="response"), main="Weight vs Resid (Response)" # response residuals
abline(h=0, col="red")
```

Step 4) Goodness of Fit: Plots and Statistics



Whoa... That doesn't look right!!

This plot **doesn't work for mixed models** since it ignores the random effect we added.

This random effect is a different baseline intercept for each patient. Hence why it's called the Random Intercept Model.

```
plot(data6$response, lm.mm2.predict, main="Response vs Prediction (Fitted)",  
xlab="Weight (response)", ylab="Predicted Weight")  
abline(a=0, b=1, col="red")
```

Step 5) Interpret Model Parameters and reach a conclusion

R CODE and output used to create Tables

```
> summary(lm.mm2)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: response ~ treatment + (1 | id)
Data: data6

REML criterion at convergence: 1400.8

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-2.91712 -0.52589  0.08103  0.65545  2.36502

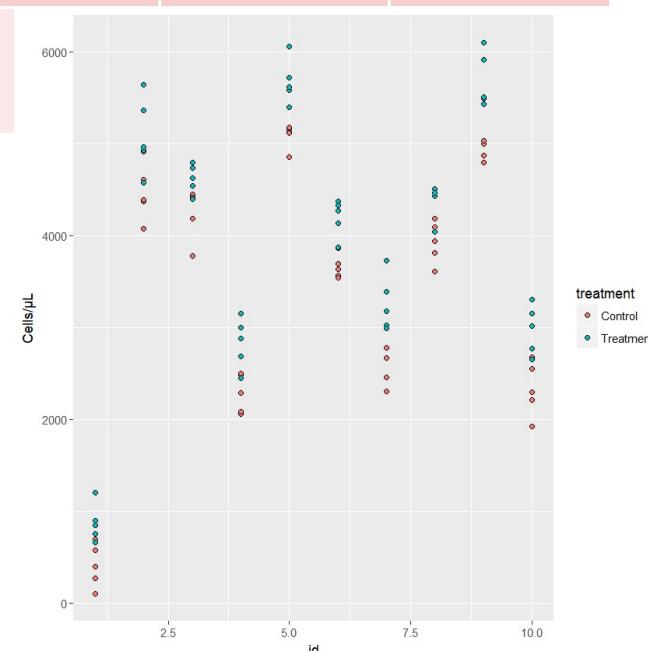
Random effects:
 Groups   Name        Variance Std.Dev. 
 id       (Intercept) 2494845  1579.5  
 Residual           49198    221.8  
Number of obs: 100, groups: id, 10

Fixed effects:
            Estimate Std. Error    df t value Pr(>|t|)    
(Intercept) 4209.40    500.47    9.04 8.411 1.44e-05 ***
treatmentTreatment 521.98     44.36   89.00 11.767 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
              (Intr)
trtmntTrtmn -0.044
> confint(lm.mm2)
Computing profile confidence intervals ...
          2.5 % 97.5 %
.sig01      1020.7097 2503.7810
.sigma      191.8871 257.1797
(Intercept) 3182.1446 5236.6612
treatmentTreatment 434.5836 609.3690
```

Step 5) Interpret Model Parameters and reach a conclusion

Parameter	Estimate	SE	T score	P value	95% Confidence Interval	
					Lower Bound	Upper Bound
Control (β_0)	4209	500	8	1.4e-5	3182	5237
Adjustment Treatment (β_1)	522	44	12	<2e-16	435	609
SD between patients	1580				1021	2504
SD within patients	222				192	257



Step 6) Reporting: Overall Conclusion suitable for publication

“There is strong evidence to show that the Treatment influences white blood cell count ($p < 2e-16$). It increases # of white blood cells by between 435-609 cells/ μL (95% CI), from an average of approximately 4209 cells/ μL (95% CI=3182-5237). This effect has been estimated fairly accurately [as 95% CI isn’t too wide].

There was much larger variation between patients ($sd = 1580$) than within ($sd = 222$), meaning it was worthwhile partitioning it out for a more accurate model.

There were no outliers or unexplained structure. The error was normal”

When giving a p-value always give an estimate of the effect size as well i.e. the 95% CI.

Was it worth fitting the more complex model?

If we fit a simple ANOVA model like we did previously it shows marginal support that the treatment has an impact (treatment $p=0.052$) while the random model has strong support ($p < 2e-16$). This is because the effect of treatment has been hidden by the noise in the data set (residual=1435), while the residual for the random model is much smaller (222) meaning it has more power. This is because the differences between subjects is included in the fixed effects residual, but is partitioned out in the random effects as the id-intercept SD (1580).

So fitting the more complex repeated measures model has shown us something the simpler ANOVA model cannot.

FIXED MODEL

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3398.2	203.0	16.743	<2e-16 ***
treatmentTreatment	563.7	287.0	1.964	0.0524 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1435 on 98 degrees of freedom

RANDOM MODEL

Random effects:

Groups	Name	Variance	Std.Dev.
id	(Intercept)	2494845	1579.5
Residual		49198	221.8

Number of obs: 100, groups: id, 10

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	4209.40	500.47	9.04	8.411	1.44e-05 ***
treatmentTreatment	521.98	44.36	89.00	11.767	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

TAKE HOME

A mixed model can be the difference between a p-value low enough to publish, and one so high publication is not possible (0.05 vs <2e-16).

It does this by reducing the noise (dropped from 1435 vs 222) and making the signal easier to detect (which is about the same 564 vs 522).



THE UNIVERSITY OF
SYDNEY

Sydney
Informatics Hub

Page 110

1:25

Other Resources



THE UNIVERSITY OF
SYDNEY



Further Assistance: Sydney University

SIH

- **1on1 Consults** can be requested on our website:
www.sydney.edu.au/research/facilities/sydney-informatics-hub.html OR Google “Sydney Informatics Hub” with the “I’m feeling lucky” button
- **Training** Sign up to our mailing list to be notified of upcoming training:
<https://signup.e2ma.net/signup/1945889/1928048/>
 - Research Essentials
 - Experimental Design
 - Power Analysis
- **Online library.** Useful links and the most recent version of all our workshops.
 - <https://sydney-informatics-hub.github.io/stats-resources/>
- **Hacky Hour**
www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training/hacky-hour.html OR Google “Sydney Hacky Hour”

OTHER

- **Open Learning Environment (OLE) courses**
 - **Science:** OLET5608 Linear Modelling: Exploratory data analysis, sampling, simple linear regression, t-tests and confidence intervals. Ability to perform data analytics with coding, basic linear algebra.
 - **Business:** BSTA5007 Linear Models
 - Many others, and constantly changing, so have a look at what is available by getting the list and searching for key words such as linear, regression, GLM, ANOVA, etc.
- **Linkedin Learning:** <https://linkedin.com/learning/>
 - **SPSS** <https://www.linkedin.com/learning/machine-learning-ai-foundations-linear-regression/welcome?u=2196204>

Other SIH workshops

Linear Models 1: Basic intro to *Linear models* with a normal (gaussian) error. Example workflows for Simple Linear Regression, ANOVA, ANCOVA, mixed models.

Linear Models 2: Extends the Linear Model framework introduced in LM1 to *Generalised Linear Models* which allow non normal errors and responses. Example workflows for Poisson (Count) and Logistic (Binary) regression.

Linear Models 3: *Tricks of the Trade* including Interpretation, Reporting and different ways to code categorical data (parametrising the data)

Model Building: LM workshops use simple 1 or 2 predictor examples. More than this requires additional Workflow steps and possibly different Methods to account for things like Multi-Collinearity. These additional topics are covered in this workshop.



Further Assistance

VIDEOS

- StatsQuest with Josh Starmer
 - Linear Models:
<https://www.youtube.com/playlist?list=PLbIh5JKOoLUIzaEkCLIUxQFjPllapw8nU>
 - What is a Statistical Model https://www.youtube.com/watch?v=yQhTtdq_y9M
- Zedstatistics, longer videos than StatsQuest. <https://www.youtube.com/c/zedstatistics>

WEBSITES

- R GLMM FAQ <https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html>

BOOKS AND PAPERS

- Faraway, Julian James. (2016) Extending the Linear Model with R : Generalized Linear Mixed Effects and Nonparametric Regression Models.
- Fox, John. (2016) Applied Regression Analysis and Generalized Linear Models.

Tricks to learning – R, linear models, SPSS, etc

- The trick is doing a little bit everyday and getting really good at it so by the time you get to actually needing R you are comfortable in it.
- When working an actual problem let yourself ‘process’ problems overnight. I’ve lost count of the time times I have battled for hours only to wake up the next day and nail it.
- As tempting as it is. Don’t just google stuff, if you get to know your books and references it will give you a broader understanding, which will help you in the long run.
- Create an R script with your ‘training code’. So as you read the book jump into R and try stuff out. Get used to creating sample data to test stuff out.
- And I’ll leave you with a paraphrased quote from one of the R guru’s Hadley Wickham “Frustration is good, it means you’re at the edges of your understanding and are learning!!”

R: Where to start

BOOKS

- Find an intro R book
 - Read it a little bit everyday, try and get a routine going such as a little at breakfast, before bed, whatever.
- I like this one for a good intro that includes a lot of statistical methods
 - **Kabacoff, Robert (2015) R in Action: Data Analysis and Graphics with R.** It also has a great web page resource which is a good first port of call too
 - <https://www.statmethods.net/>
 - Buy through Web site for a discount
- Only downside is that it doesn't use **Hadley Wickhams** packages, so I would also recommend one of his. In particular **R for Data Science** gives a great intro to data wrangling and visualisation using his packages. (Wickham, Hadley, and Garrett Grolemund (2017) R for Data Science Import, Tidy, Transform, Visualize, and Model Data)
- Finally I recommend **MASS (Modern Applied Statistics with S-Plus)** by Venables and Ripley. The 'Yellow Bible'. It has at least a little bit on pretty much any statistical method you can think of. I tend to start here to get an intro on what R can do and then research outwards. (Venables, W. N, and B. D Ripley (2013) Modern Applied Statistics with S-Plus)

ONLINE

- Lots of short (and long) YouTube courses
 - EXPLORE, find a style you like and watch a little each day if too long.



Acknowledging SIH



All University of Sydney resources are available to Sydney researchers **free of charge**. The use of the SIH services including the Artemis HPC and associated support and training warrants acknowledgement in any publications, conference proceedings or posters describing work facilitated by these services.

The continued acknowledgment of the use of SIH facilities ensures the sustainability of our services.

Suggested wording:

General acknowledgement:

"The authors acknowledge the technical assistance provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

Acknowledging specific staff:

"The authors acknowledge the technical assistance of (name of staff) of the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

For further information about acknowledging the Sydney Informatics Hub, please contact us at sih.info@sydney.edu.au.

We value your feedback



- We will email you a link to the survey shortly
- It only takes a few minutes to complete (*really!*)
- Completing this survey is another way to help us keep providing these workshop resources free of charge

