A pragmatist's guide to General and Generalised Linear (Mixed) Models and inference

Dan Haydon & Darren Shaw

Check for updated versions!
(Version 1.4.11)

About the authors

Dan Haydon is an ecologist and epidemiologist at the University of Glasgow
Darren Shaw is a Reader in Comparative Epidemiology at the University of Edinburgh

Acknowledgements

Many thanks to those who have proof-read and provided feedback on the text to date including Barbara Mable, Max Hadoke, and Mercedes Burgueno in particular, and the class of 2023-24 many of whom made so many helpful suggestions. We would be grateful for any comments or feedback on this text, major or minor, positive or negative.

Comments can be sent to Daniel.Haydon@glasgow.ac.uk

Cover photograph: First light on the Buchaille, Glen Coe, Scotland.

Contents

Appendices

Introduction

(back to Contents)

(Terms in bold are defined in the accompanying glossary)

For a substantial number of you, doing and understanding statistics are your worst nightmares. Your room 101. You may have no interest in studying statistics, and no aspiration to use statistics.

However, at some point, usually after you have undertaken some sort of research project and generated some of your own data, you are forced to confront 'statistics'. You seek guidance, begin to understand that perhaps some of the skills might be a bit useful, and if the guidance is provided in the right form, you will begin to understand how to apply them. You might even quietly concede that it wasn't quite as impenetrable or impossible as you first thought, and indeed, a little bit satisfying to finally have some agency and understanding over the analysis of your own data. There may even be a bit of a 'Eureka' moment when what was previously a whole bunch of piece-wise gibberish falls into place, and some of the smoke clears.

If any of this sounds familiar then this text is written for you. We find it a little sad that so many perfectly smart students have to go through this fear and loathing completely unnecessarily – when in fact you might even quite enjoy the subject.

In this text we are only going to talk about **general linear models** and **generalised linear models**. We're not going to make a distinction, we'll refer to them both as GLMs. Don't be freaked out by the name. A GLM is basically a sum. The maths is about as complicated as 12 = 4 + (3 x 2) -1 + 5 – 2. You've almost certainly encountered GLMs before. You are likely to have run into the words **T-test**, **anova** (**one-way** and/or **two way**), perhaps **regression**, and **multiple regression**. You may even have heard of **ancova**, **repeated measures** analysis, or **logistic regression**. *These are all types of GLMs.* And rather than use all this language and try to explain all these different things separately, we're going to describe a single unifying framework which covers all of these things and more. Without doubt – GLMs provide the most for the least when it comes to learning about using and interpreting statistics. They are certainly not the whole story – for example they don't include non-parametric statistics, but they are a hugely important chunk of modern statistical practice.

There are a lot of excellent statistics text books out there, so why are we producing this book? Because the feedback we've received over many years is that our chosen approach works for many students, and it works better than any other approach we know for students who don't think they want to know about the subject, and/or have forgotten so much that is usually required for a course on advanced statistics (for example, what a **logarithm** is, or whether 2.061154e-09 is a large or a small number - if you have, check out Appendix A and/or B for a reminder). We think the single framework approach is an unusually fast way to develop a relatively sophisticated understanding of modern statistics. This comes with some risk (some would say too fast, with insufficient development of the underlying theory), but we hope to get you to a place where deeper understanding can append easily on to the basic understanding of GLMs we aim to convey to you here.

GLMs are models. The clue is in the name. GLMs comprise **variables** (data) and **parameters** or **coefficients** (we use these two terms interchangeably – they are completely synonymous). Because GLMs contain parameters, they are a subset of **parametric statistics** – in contrast to a different suite of tests that don't use parameters in the same way and are called **non-parametric statistics**. We won't discuss non-parametric statistics at all. We have nothing against them, people use non-parametric statistics all the time, but our focus is on modelling data using the diverse varieties of distributions available to describe different data types, and the relatively more powerful inferences that can be made from such models, and so non-parametric statistics just aren't the focus of this text.

GLMs are actually relatively simple. As we shall see, GLMs are little more than arithmetic really. If you can 'get' the single framework that are GLMs, you'll be in a strong position to relatively easily develop a quite sophisticated understanding of how to analyse a wide range of statistical problems. One of the main advantages of using a GLM type approach is that it allows you to "build" statistical models to explain variation in the thing you are interested in by the things you may think are responsible for giving rise to the variation, which can be few or many. From this point you should be able to move relatively smoothly onwards to think about more advanced ideas.

However, the detail here is not the important thing. We really have three goals:

1) To persuade you that GLMs are useful

2) To persuade you that they are in fact straightforward and within your reach – that there is really nothing to fear

and

3) To convince you that when the time is right, and you need to use them – that you can confidently revisit your notes and get stuck in

We have chosen to use R (© 2023 The R Foundation for Statistical Computing) throughout this text. However, this is a text about GLMs and not about R. There is a diverse range of statistical software out there, and they all have their advantages and disadvantages. We have chosen to use R as: 1) it is open source and therefore does not incur any financial cost; 2) it is professionally authentic, and used by the majority of scientific researchers; and 3) consequently there is a very large active online community providing answers to queries (*e.g.* the website stackoverflow.com).

However, it is also true to say that there is a learning curve (and that includes us - we are still learning new things about R most days) and it is code not menu based. In addition, commands and packages do change. So whilst we are going to assume the reader knows what R is, our use of R will be relatively and deliberately simplistic. We do this for two reasons because we don't want R-hassles to distract from the statistical principles; and implementing environments and different R packages come and go, they certainly evolve.

None-the-less we will refer to more detailed uses of R in boxes where we think it's helpful. On the whole we've tried to work to two or three decimal places (if you are unclear how to round numbers check out Appendix C). We have called our data sets (known as dataframes in R) `my_data` if we need to refer to them, and some of the outputs will be edited down so you can focus on the most important parts we are trying to explain.

While the arithmetic of these GLMs is relatively simple, the interpretation can be subtle in places and there is a diversity of different philosophies out there as to what is 'best practice'. Our approach is quite pragmatic, but not everyone will agree with it. Once the basics have bedded-in you will begin to form your own judgements and opinions about what is best, and as the ideas become more routine, you'll have more bandwidth to think about more advanced and nuanced aspects of how to approach inference. You may come to disagree with us – if so – good. Our aim is to provide you with an entry-level understanding .. not an exclusively correct approach.

The text is in two parts. Part 1 focuses on building the GLMs. Part 2 addresses how to make statistical inferences from the GLMs. You may be frustrated that the word 'statistics' is hardly used until Chapter 17. This is entirely deliberate. We are convinced that if the process of *modelling* the data is well understood, learning what we can conclude from the model parameters - statistical *inference* - is made a good deal easier. While there is an intended logical flow that builds throughout the text, with some acquaintance with this material, any of the chapters should make reasonable sense if read individually.

Lastly, as instructors we find we need to lead you along a fine knife-edged ridge – on one side are cliffs that plunge down into the depths of seemingly abstract theory that will kill off any potential curiosity. On the other side is a perilous abyss in which students might acquire access to powerful methodologies without an appropriate understanding of how they work, or what they really tell you. And with the best will in the world, most of you won't find all of this *that* interesting! So, we will try to be as concise as possible (basic and more advanced digressions are covered in the extensive appendices). For more advanced students and colleagues that recognize when we are being overly simplistic - we ask for your forbearance.

It takes most people many years to develop a comprehensive knowledge of GLMs, so give yourself credit for what you do understand, and give yourself time to become more familiar with what you don't.

Part 1

Questions, data,
and
model construction

Chapter 1

Why are we doing this?

(back to Contents)

This chapter briefly outlines what GLMs can do for you, and provides a short breakdown of what we have to think about to understand and apply them.

When you think about it, most questions in science require us to study differences between things, or to put it another way: variation. Why is that population larger than this one? Why are there more species here than over there? Why do these cells migrate faster than those? Why is disease prevalence higher in this population than in that one? Why do these individuals live longer than those? Why are things different? After all, if things are all the same, then we have nothing much to understand.

There is always a small amount of random variation in anything, so we need methods to determine when things are sufficiently different that an explanation is required. And when we do see meaningful differences, we need methods that allow us to relate these differences to something else that might be related to or causing the variation. Simply speaking – this is what GLMs are good for.

We can hypothesize - that is - to make more-or-less informed guesses - about what might be responsible for the variation we observe. If you can phrase your research question to fit the following template:

> Can I explain variation in this thing – using variation in these things

Then GLMs might be what you need.

For example, you might ask:

> Can I explain variation in the density of predators in different areas using variation in their prey density?
>
> Can I explain variation in the number of earthworms in different soil types by variation in the soil pH?
>
> Can I explain variation in the incidence of disease among different people using variation in whether they were vaccinated or not?

These hypotheses might be helpful in explaining the variation, or they may not – but GLMs will enable us to formally evaluate these hypotheses. If they are not helpful – we may need to look for an alternative explanation. Constructing hypotheses early in your investigation is useful and important, as it allows us to know what data we will need to construct our GLMs.

Of course, the world is a complicated place and it might be that you seek to explain variation in something using variation in several other things. For example:

> Can I explain variation in whether a habitat patch is occupied or not using patch area, patch quality, and how close-by other patches are?
>
> Can I explain variation in the life-span of individuals using variation in how much alcohol they drink, how much tobacco they smoke, and the average lifetime of their grandparents?

Not all questions can be phrased like this – but most can, and with practice you'll quickly learn how to recognize how to do this. In these types of question, the thing in blue is termed the **response variable** (*i.e.* we are hypothesising that it 'responds' to variation in the things in red), while the things in red we term the **explanatory variables** (*i.e.* we hypothesize they explain variation in the response variable).

Note that in any particular GLM there is always *only one* response variable, but there may be *one or more* explanatory variables. It is essential that you are clear which variable is the response variable and which the explanatory variables. This critical information stems from the question you are asking – it isn't information in the data, or a feature of the data. If you don't know your research question, you cannot look at your data and determine which of the variables is the response variable. It depends on your motivation for collecting the data and is 'in *your* head'.

We answer questions that fit with this template by constructing a model of the form:

> Response variable - modeled by - some combination of explanatory variables

How does this work? You may have heard of the acronym **ANOVA**. It stands for ANalysis Of VAriance. It is a term that often is used to convey a variety of slightly different things. However, the general idea that we need to analyse variance makes a lot of sense because all of these questions require us to study variation. So, we will use GLMS to develop ways to model the observed variation in the response variable, using the variation in each of the explanatory variables.

Because GLMs have just one response variable, these models are termed **univariate**. There are methods that simultaneously analyse variation in more than one response variable. Such methods use **multivariate models** (for example, can I explain variation in *both* people's height and weight using variation in their age). However, we will not be discussing multivariate models here. If you *are* interested in the variation in more than one response variable a pragmatic way to proceed may be with a number of different (univariate) GLMs – one for each of the different response variables.

We can use GLMs to determine if there are likely to be relationships between each of the explanatory variables and the response variable. We can test the hypotheses we have constructed about the relationships between the response and explanatory variables, and assess the direction (does the response variable go up or down as the explanatory variable changes) and magnitude (by how much does the response variable go up or down as the explanatory variable changes) of the effects of each of the explanatory variables on the response variable.

This process is called **inference** and is extremely useful!

Essentially, GLMs split up the variation into *explained* variation and *unexplained* variation. The unexplained variation is often quite a high proportion of the overall variation in the response variable (after all the world is a complicated place, and we are unlikely to capture all of this complexity in such a simple model). Our GLMs needs to account for and quantify this unexplained variation (while of course recognizing that it cannot be explained!).

You may have come across a basic GLM before in the form of simple linear regression (a straight line fitted through a cloud of points with a slope and an intercept). You may remember an equation '$y = mx + c$'. Here y denotes the response variable, x the explanatory variable, and the parameters m and c denote the slope (how on average the values in y change with increasing units of x) and intercept (the average value of y when $x = 0$) of the straight line, respectively. As we will show in later chapters, we can extend this idea in a variety of ways to address more complicated situations.

As we shall also see, there are a number of stages to working with GLMs:

1. Formulate the research question according to the template described above and identify the key hypotheses
2. Acquire the data
3. Layout the data in ways we describe in section 2.1.
4. Formulate an appropriate GLM, comprising the requisite variables and parameters.
5. Fit this model, and check the data fit the model well enough.
6. Interpret the output of the model
7. Use the model to evaluate our various hypotheses (inference).

We may even use the model to predict values of the response variable under various combinations of the explanatory variables – that maybe we didn't even observe.

These are powerful and useful things to be able to do, and lie at the heart of the scientific method. And this is why we are studying them.

Important ideas to take-away

- The 'template' of the question that GLMs can be used to address: Can I explain variation in this thing – using variation in these things
- What we mean by response and explanatory variables
- What we mean by a univariate model
- Steps in working with GLMs

Chapter 2

Your data

(back to Contents)

If you can lay your data out correctly, understanding how to construct GLMs will become much simpler. Furthermore, the data will be laid out in a way that software packages can use to explore your data. In addition, you will quickly become familiar with slightly different forms that your response and explanatory variables may take. Here we briefly introduce you to data layout and different types of data.

You need to be absolutely crystal clear what your variables are, and which one you want to designate as the response variable, and which others (one or more) treated as explanatory variables. If you are not – you should return to Chapter 1!

2.1 Data layout

You should be able to count the total number of observations of your response variable. We will denote this number by *n*. For each observation of the response variable, there will be additional information that will constitute the explanatory variables – of which there may be one or more.

For example,

> You may have estimates of the density of predators in each area, and the corresponding prey density for each estimate.
>
> You may have counts of the number of earthworms in samples of different soil types and the corresponding soil pH for each count
>
> You may have records of how many people caught a certain infectious disease and information on whether each person was vaccinated against this disease or not.
>
> You may know whether a habitat patch is occupied or not and the area, quality, and a measure of the proximity of other patches of each patch.
>
> You may have the life-span of individuals and how much alcohol they drink, how much tobacco they smoke, and the average lifetime of their grandparents.

Each observation of the response variable, be it density of predators, number of earthworms, cases of disease, observation of a habitat patch, or life-span of an individual, forms a **record** (or row in a spreadsheet) comprising the observation of the response variable and all of its associated explanatory variables. If you have *n* observations of your response variable, you will have *n* such records.

For example, you might have performed an experiment with say 3 different treatments and a control group (so 4 groups in total). You might have 10 observations of your response variable from each of your 4 groups. In total, you will have 10 x 4 = 40 observations of your response variable, and therefore 40 records. Each record perhaps will comprise: an observation number; a sample number (say 1); a corresponding treatment, say ‘T1’; and an experimental observation, say 7.66. The data might look like Table 2.1:

Table 2.1. A ‘flat’ data layout. Different shading indicates the different treatments. The rows with dots indicate what we’ve missed out just so the table doesn’t take up too much space (i.e. observation numbers 3 through 9 exist … but we are not showing them here.

Obs no.	**Sample**	**Treatment**	**Response**
1	1	T1	7.66
2	2	T1	6.45
.	.	.	.
10	10	T1	9.45
11	1	T2	4.79
12	2	T2	3.78
.	.	.	
21	10	T2	5.79
22	1	T3	3.19
23	2	T3	6.12
.	.	.	
30	10	T3	4.87
31	1	Control	2.13
32	2	Control	3.01
	.	.	.
40	10	Control	1.98

So the 10th record is: 10 T1 9.45

The 31st record is: 1 Control 2.13

And so on.

Note there are 40 observations of the response variable, 40 records, and 40 rows of data. You might not even have different ‘treatment groups’ – perhaps you just have ‘groups’ that you wish to compare, this is fine so long as you include which group each observation of the response variable comes from in each record.

Or, you might have taken 5 vegetation samples from 5 sites in each of 3 meadows. You have an estimate of an isotope ratio from each sample. You will have 5 x 5 x 3 = 75 records. Each record will comprise an isotope ratio measurement, the sample number, the site and the meadow. The data might look like Table 2.2:

Table 2.2. A second example data set in ‘flat’ format. The different shadings represent different sites, and the different colours - different meadows.

Sample	Site	Meadow	Isotope ratio
1	S01	A	8.5
2	S01	A	7.9
.	.	.	.
5	S01	A	9.3
6	S02	A	5.6
7	S02	A	6.1
.	.	.	.
10	S02	A	6.3
.	.	.	.
.	.	.	.
21	S05	A	7.7
22	S05	A	6.9
.	.	.	.
25	S05	A	6.4
26	S06	B	4.1
27	S06	B	3.9
.	.	.	.
30	S06	B	4.2
31	S07	B	5.4
32	S07	B	5.1
.	.	.	.
35	S07	B	4.9
.	.	.	.
.	.	.	.
46	S10	B	2.9
47	S10	B	3.1
.	.	.	.
50	S10	B	3.3
51	S11	C	12.3
52	S11	C	12.1
.	.	.	.
55	S11	C	11.7
56	S12	C	11.9
57	S12	C	13.2
.	.	.	.
60	S12	C	9.8
.	.	.	.
.	.	.	.
71	S15	C	10.4
72	S15	C	10.1
.	.	.	.
75	S15	C	9.6

So the 1st record is: sample 1, site 1, meadow A, isotope ratio 8.5

The 72nd record is: sample 72, site 15, meadow C, isotope ratio 10.1.

Note: There are 75 observations of the response variable, 75 records, and 75 rows of data.

The golden rule is one record for each observation of your response variable, and one row in your data table for each record. If you get the data layout right, then it will be much more obvious to you how to explore your data and construct the GLMs which will apply to the data. So, this is really worth thinking about in advance. It is well worth simply 'making up' a small amount of data based on what parameters and variables you are interested in and typing it into (something like) Excel to confront the reality of what you are likely to end up with after you've collected the data for real.

We strongly recommend you *do not* 'pre-process' your data by say averaging sets of observations of your response variable. By doing so you are not only throwing away information from your study – removing some of the natural observed variation - you are going to be making your sample size smaller, which has an impact on inference. Each observation is hard-won, strive to preserve and analyze them in the same form as you observed them!

2.2 The Response variable

Your response variable will most likely be one of three possible 'types'.

Continuous: Observations of your response variable are real numbers with decimal places – or if they don't have decimal places - *they could have had decimal places*. For example, they may be concentrations, say 1.17 g/ml^3; or they might be say - height, for example 164 cm. This number doesn't have decimal places but it might have had. Afterall, height is fundamentally a continuous quantity. Furthermore, in general there are no obvious upper or lower limits to the observations, *or if there are, they very rarely arise close to these limits*. (There is a more detailed discussion of limited range continuous data in Appendix D).

Discrete: Observations of your response variable are non-negative integers: there is no sense in which they might have decimal places. For example, your data may be count data. You cannot see 3.5 wildebeest (the test of something truly discrete is if you can't make sense of anything 'in between'), you cannot see -3 wildebeest (although 0 wildebeest is entirely possible). And there is no defined upper limit on your counts.

Binary: Each observation of your response variable is of only two values or categories: 1 or 0, alive or dead, plus or minus, positive or negative, present or absent - nothing else is possible.

Other types of categorical data are possible. Your data may be 'trinary' – a bit like binary, but one of three possible values or categories (perhaps say positive, negative or neutral; or high, medium or low). Or, perhaps even 4 or 5 different values or categories.

Some data are 'circular' – like time, month or compass bearings (circular in the sense that 359 degs is very close to 1 deg.). We won't discuss these alternative response variable 'types' in the main text but they absolutely fall within the all-powerful GLM framework and we provide some guidance in Appendix E on how to think about these data types.

Alternatively, your data may be 'paired' (for example, measurements on the same subject before and after a treatment of some sort) and these are also amenable to analysis using the GLM framework as discussed in Appendix F.

It is critical to recognize the form of your response variable data (continuous, discrete, binary etc) as this will inform exactly how to construct and interpret your GLMs. Doing so should quickly become second nature to you.

2.3 Explanatory variables

Explanatory variables come in two superficially different forms.

We may regard them as numbers - numerical, by which we mean we could multiply them by something. For example 8.356, 1.2, or 12. These numbers might be heights, concentrations, temperatures, densities, pH's, volumes, lengths, areas ... and all sorts of other things. We'll call explanatory variables of this form **continuous** explanatory variables, or **covariates**. (And just in case you are wondering, we are not at all interested or concerned in the how these explanatory variables are distributed).

Alternatively, the explanatory variable really may be a label - like the treatment in an experiment (Table 2.1), or site number, or which meadow (Table 2.2). Where it's a word, or a number that we *interpret* as a label, we can't multiply it by anything. Meadow number 1 x 5 doesn't work. Nor does T2 x 2. The explanatory variable refers to one of a number of categories, or **levels**, and we define such explanatory variables to be **categorical**. We should be aware of how many levels such categorical explanatory variables have. Usually, these levels don't have any natural order to them (we can't easily place the meadows in any meaningful order); thus, we sometimes call them **nominal**. Sometimes the levels do have an order. For example, we might have defined temperature to be cold, medium and hot which obviously can be ordered, and we may then refer to them as **ordinal**. It is clear from this last example that there are some explanatory variables that may be treated as either categorical or continuous (if we'd remembered the thermometer we could have measured temperature numerically in degrees) and the choice of which can be up to you. We'll come back to this much later.

Important ideas to take-away

- A data record is a line in a data file that comprises an observation of the response variable and all of its associated explanatory variables
- How to lay out data one record per row
- How to recognize whether your response variable is continuous, discrete or binary
- How to recognize whether your explanatory variables are continuous or categorical
- If they are categorical, to be clear how many levels they have

Chapter 3

What are data - really, and how will we model them?

(back to Contents)

This chapter is a discussion about the connection between data and probability density functions. We won't talk much about specific probability density functions here (we do this more in Chapter 4 and 7), but we focus on the notion that we can think of data arising from them, and that probability density functions can therefore be useful in modelling variation we cannot account for with explanatory variables.

We have already recognized that the world is a complicated place, and our simple GLMs are likely only to be able to account for a fraction (often a small fraction) of the variation we observe in our response variable. After all, GLMs are only relatively simple models, and probably contain just a handful of explanatory variables. We know that in reality there are many more influences on the outcomes we observe in the world – we just didn't happen to, or were not able to measure them and record them as explanatory variables. So, we model this *unexplained* variation in our GLMs as random variation using **probability density functions (pdfs).**

Our goal in this chapter is to introduce the idea of how data *can be thought of* as arising from pdfs. To be clear – they don't! But perhaps we may not go far wrong by assuming that they could appear to do so! The most suitable pdf depends on our data, and we'll introduce the most common pdfs in more detail in Chapter 4, and how GLMs use pdfs in Chapter 7.

3.1 Probability density functions

The term probability density function or pdf may be new to you, but the idea is not. For example, when you wish to decide which side kicks off a sports match, you can toss a coin to decide. The outcome is assumed to be determined by 'chance', with each team having a 50% probability of winning the toss. The outcome of the coin toss is considered to be random. Of course, it's really not at all random. When you flip a coin, the result actually depends on the upward force in the 'toss', the torque you applied to the edge of the coin, perhaps the viscosity of the air, and whether you catch it, or let it fall to the ground. It's not chance at all … it all depends on physics. But we didn't collect any data on force, torque, or viscosity, so we call it chance.

In statistical language, we can say that the outcome of a coin toss is a **random variable** generated from a pdf called a **Bernoulli** distribution. Bernoulli distributions generate only two outcomes – in this case either 'heads' or 'tails', where the probability of either of the two outcomes is p (heads) and 1-p (not heads, *i.e.* tails) The sum of the two probabilities must equal one as no other result is possible. In this case, p might be 0.5, and so 1-p will also be 0.5. Of course, the coin might not be balanced ($p \neq 0.5$) and one outcome might be more probable than the other.

There’s another pdf you’ve likely come across. Many board games use the throw of a dice (let’s assume 6 sided, but sometimes more or less) to determine progress of a game. Which of the 6 numbers you roll are again thought to be random variables. And again, of course they aren’t really random. It all depends on how you throw the dice, but it’s (fortunately) extremely difficult to throw a dice so that you can produce any number you want on demand, so we are all content to regard it as chance. In statistical parlance we call this a **Uniform** distribution, defined from 1 to 6. There is an *equal* probability of 1/6 of generating each number between 1 and 6 (hence the name ‘uniform’). Of course, the dice may be a 30-sided rhombic triacontahedron, in which case it generates random variables from a uniform distribution defined between 1 and 30.

So, you have come across pdf’s before – both Bernoulli distributions and Uniform distributions are well known to you, even if these names were not. You’ve probably also heard of **Normal** distributions**,** which are also known as **Gaussian** distributions (we’ll use Normal, but the two terms are entirely synonymous). Mathematically, the Normal distribution is a bit more complicated to describe, but it’s the famous ‘bell-shaped curve’, with numbers towards the middle of the ‘bell’ being more likely than numbers from the far left or right ‘tails’ of the bell (see Chapter 4). There are in fact dozens of different pdfs: the **Poisson** distribution, the Bernoulli distribution, which is a special case of the related **Binomial** distribution, the **Negative Binomial** distribution (of which the Poisson distribution is a special case, see Chapter 10.1), the **Log-Normal** distribution, the **Gamma** distribution, the **Weilbull** distribution, the **Exponential** distribution … the list is pretty endless. They are all capable of generating ‘random numbers’ of various different types, and that have different distributions. When it comes to building your GLMs, you will have to choose which is most appropriate depending on your response variable. Here we are going to present the three most common pdfs used in the vast majority of GLMs – Normal, Bernoulli, and Poisson (which as we will see is a special case of a Negative Binomial).

3.2 Example: The Normal distribution

Suppose we are interested in variation in human height. Obvious ways of explaining this variation might be with the age and sex of the individuals. But even after including both these variables, we can probably account for perhaps only half of the variation in height. If we look at a sample of students who are all the same age (say 25) and sex (all male) we’ll still observe quite a bit of variation (Figure 3.1):

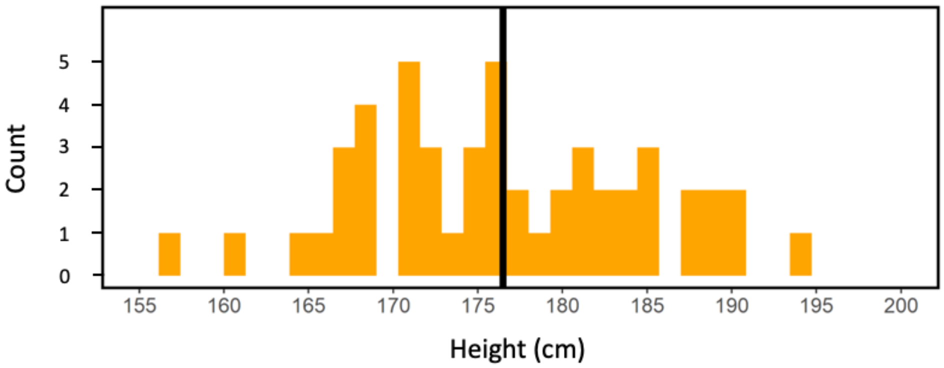


Figure 3.1. The frequency distribution of a sample of 50 Scottish male 25-year old students. Their heights ranged from 156 cm to 194, and the average is 176 cm.

Where does the remaining variation come from? Well, height is thought to be controlled by about 50 different genes in our genomes. Also, height may relate to early life nutrition, and perhaps some other things we haven't yet discovered. How can we deal with all this unexplained variation? We haven't collected the relevant data for these other sources of variation, so we can't *explain* it, but we'll *account* for it as coming from a pdf. This remaining variation we will regard as 'chance' or 'random', by which we mean that the heights of these 25-year old Scottish males might as well be regarded as random variables from a Normal distribution, with a shape and position that depends on the fact they are 25 year old Scottish males. We know the unexplained variation isn't really random, and that it depends on genes and nutrition and perhaps other things, but like the coins and the dice, we don't have the real explanations for the variation in the data so we treat it as a form of randomness.

Note that if they were 25-year old Scottish females, or 10 year old Scottish school girls, or 40 years old Maasai men we'd use a different shaped Normal distribution to model their heights (perhaps distributions with different averages, as females are generally a little less tall than males of the same age, school children tend to be smaller than adults, and Massai are famously tall!) although there would of course still be variation in each of these groups we couldn't account for.

3.3 Example: The Poisson distribution

Imagine you have done an experiment where you divided a class of 20 students into two groups, one group was asked to run up and down the stairs for 5 minutes, and the other to sit quietly at their desks. You then count the number of heart-beats in a minute for each student. Your data might look like this: 63 57 46 56 66 76 67 56 51 54 82 88 66 77 85 69 71 91 92 79. Quite a lot of variation! The first 10 are from the resting group, and the second 10 from the exercising group (Figure 3.2).

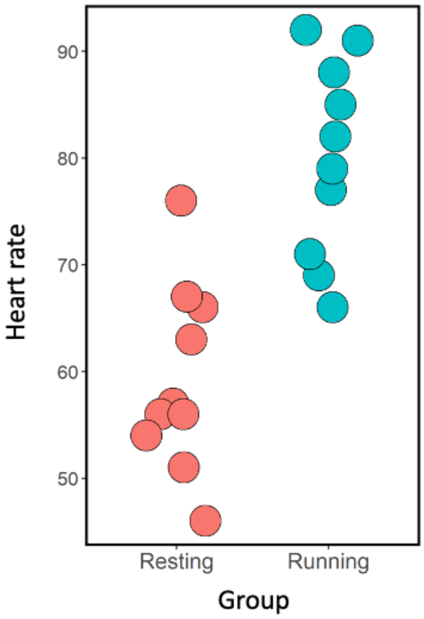


Figure 3.2. The heart rates (beats per minutes) of two groups of students, one following a resting period (red) and one following exercise (blue). The horizontal 'jitter' is just to avoid too much superimposition of the points.

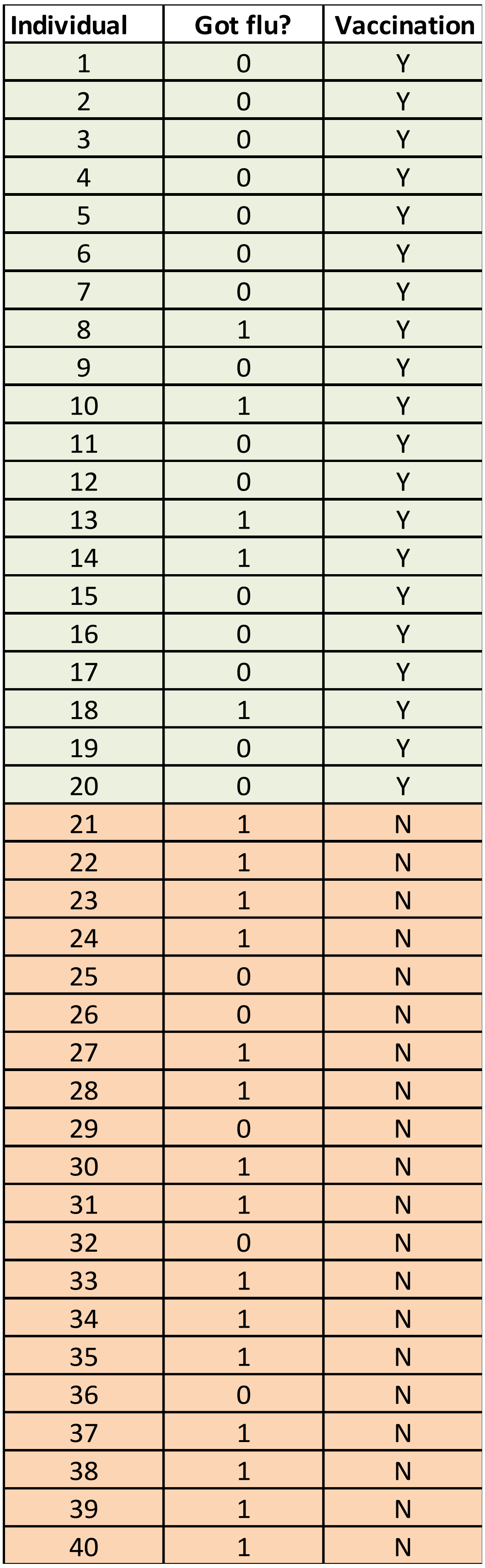
It makes sense that the resting group rates should (mostly) be lower than the exercising group, but why is there variation within each of the two groups? All kinds of reasons of course ... people are different (and of course there is the possibility you lost count of the pulses and some of this variation is 'researcher error'!). Had you done this experiment again, with different students, or even with the same students you might get a different answer (say 59 72 52 54 53 51 55 5967 66 81 85 87 91 93 90 81 67 100 74). We can understand the 'between group variation' because there is an obvious explanation for it (exercise!), and we can create a model (a GLM) that contains the explanatory variable 'have you just exercised?' to explain the difference between the two groups. Furthermore, we are not surprised by the 'within group' variation because we know people are different. But we can't really *explain* the within group variation unless perhaps we'd taken more details from each student (their height, weight, fitness, what they had for breakfast …). We didn't do that … although it really does have an explanation, we can't know what it is, so –again - we can only *account* for it, putting it down to 'chance' – 'random variation'.

What type of pdf should we use to *account* for the random variation in this example? Just as statisticians often think about binary outcomes coming from a Bernoulli distribution, and dice throws from a Uniform distribution, and people's heights as deriving from Normal distributions, they think of counts (as in heartbeats in a minute) as coming from something called a Poisson distribution. Poisson distributions generate non-negative integers (whole numbers without decimal points), so are ideal for many types of count data. In this example, the first 10 observations came from a Poisson distribution with an average of 60, and the second 10 from a Poisson distribution with an average of 80. Both Poisson distributions – but with different averages.

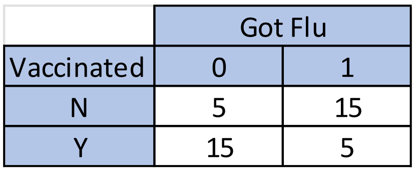
3.4 Example: The Bernoulli distribution

You have a sample of 40 people. Some of them (20) got flu over the winter and some of them didn't. There is variation in whether they got flu or not. However, 20 of them got the flu shot, and 20 of them didn't. Denoting individuals who contracted flu with a 1 and those that didn't with a 0, it might look like Table 3.1a. And we can summarise these data as in Table 3.1b

Table 3.1a. A binary data set in 'flat' format; and 3.1b a more concise table summarising the data.



Individual	Got flu?	Vaccination
1	0	Y
2	0	Y
3	0	Y
4	0	Y
5	0	Y
6	0	Y
7	0	Y
8	1	Y
9	0	Y
10	1	Y
11	0	Y
12	0	Y
13	1	Y
14	1	Y
15	0	Y
16	0	Y
17	0	Y
18	1	Y
19	0	Y
20	0	Y
21	1	N
22	1	N
23	1	N
24	1	N
25	0	N
26	0	N
27	1	N
28	1	N
29	0	N
30	1	N
31	1	N
32	0	N
33	1	N
34	1	N
35	1	N
36	0	N
37	1	N
38	1	N
39	1	N
40	1	N



	Got Flu	
Vaccinated	0	1
N	5	15
Y	15	5

We see that only 5/20 of the vaccinated individuals contracted flu, while 15/20 of the unvaccinated individuals contracted flu. This makes some sense: it appears that getting a flu shot reduces the probability you'll get flu from about 0.75 to 0.25. But how can we explain why 5 of the vaccinated individuals got flu? Again, we don't really know. Perhaps they were a bit run down, or worked in environments where they were exposed to a lot more transmission, or were slightly immunosuppressed for some reason. But – again – we didn't collect these data. There is an explanation, but we don't have the data to explain it. So ... we treat this variation as if the outcomes from the vaccinated group were generated from essentially a coin toss – heads you got flu, tails you didn't, but the probability of the coin toss generating a tail is 0.75. The data are *as if they came from a* **Bernoulli** distribution with p = 0.75. And likewise, as if the outcomes from the unvaccinated group were generated from a similar coin toss – heads you got flu, tails you didn't, but the probability of the coin generating a tail is this time 0.25. These data might be modelled as coming from a Bernoulli distribution with p = 0.25.

We've now run into three important distributions: the Normal distribution (good for continuous data), the Poisson distribution (good for count data), and the Bernoulli distribution (good for binary data). As noted before there are many other pdfs – all of which have their uses depending on your response data: For example, the Negative Binomial distribution; the Log-Normal distribution, the Gamma distribution, the Weilbull distribution, the Exponential distribution, but once you understand how to deploy these first three common distributions, how and when to use these others will be clearer.

Important ideas to take-away

- A lot of the variation that we observe in data that we can't explain with explanatory variables, *can be accounted for* quite tidily *as if* it was randomly generated
- This doesn't *explain* it, but it does allow us to *account* for it, i.e. to model it.
- Recognizing that our response variable may be continuous, or discrete (count) or binary
- Different types of response variable require us to consider different forms of 'randomness' that can arise from different probability density functions
- We are not concerned at all with the distribution of explanatory variables … the distribution of the response variable is what is important.

Chapter 4

A closer look at probability density functions

(back to Contents)

Here we discuss the specific details and properties of a range of commonly encountered probability density functions (pdfs). We discuss them in general terms, and not in specific relation to GLMs which is the subject of Chapter 7. More technical details are provided in Appendix G.

A GLM will assume that after all the information supplied by the explanatory variables has been used to model a response variable, there will remain unexplained or **residual variation** that can only be accounted for by appropriately fitted **probability density functions (pdfs)**. We will explain in more detail what we mean by this. But ... for current purposes, it is important to understand more about pdfs because every time we fit a GLM we have to decide which one of a number of different possible pdfs we're going to use to account for this unexplained or residual variation. Bear with us.

A pdf enables us to calculate how likely any particular number is to arise from it. A number generated from a pdf is called a **variate**. Probability density functions are slightly more subtle than we describe below, but this will do for the time-being.

Whatever pdf we are talking about we can say:

- They are defined by **arguments** (numbers) that completely specify the pdf
- They all have **means** (or **averages** – the two terms are synonymous) and variances (standard deviations are the square root of variances) but these are not necessarily the arguments – it depends on the distribution (mostly the mean is an argument). Regardless, the means and variances of distributions can be calculated from their arguments
- They have ranges over which the pdfs are defined – the minimum and maximum variates that can be generated (although these ranges may be infinite!)
- The area 'underneath the curve' of a pdf is always equal to one
- They can be used to calculate how likely (or in some cases probable) certain numbers are to come from them.

There are four pdfs that you are most likely to need but they all can be described in the same sort of ways. Let's start with the most common one.

4.1 The Normal distribution (or Gaussian distribution)

This is the well-known bell-shaped curve. It is defined entirely by just two arguments (or numbers): the mean of the distribution and its **variance** (or **standard deviation**). Regardless of the values of the mean and standard deviation, the Normal distribution always has a range that stretches from –infinity to +infinity, and because it describes **real numbers** (that can have decimal places) we refer to this as a **continuous distribution**. We often denote a normal distribution as: $N(\mu,\sigma)$, where N indicates we are talking about a Normal distribution, μ (pronounced 'mu') denotes the mean, and σ (pronounced 'sigma') denotes the standard deviation (e.g. Fig. 4.1).

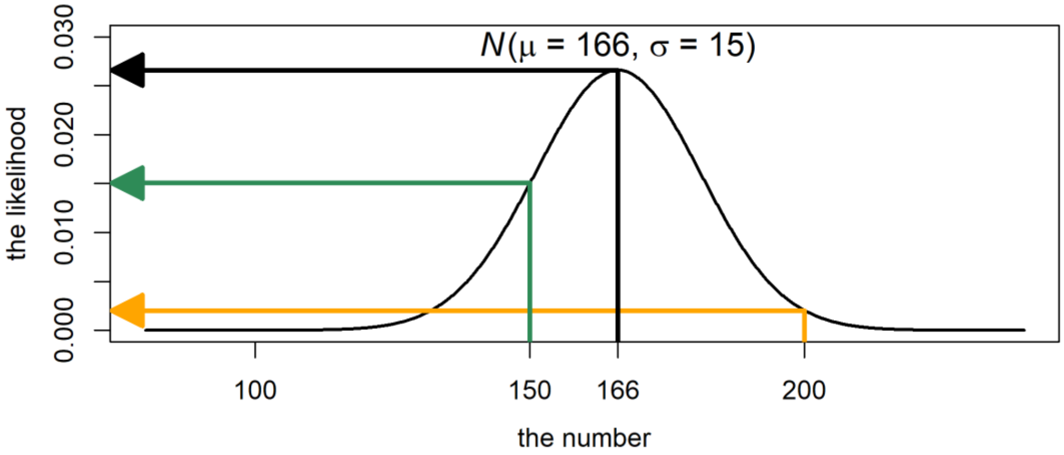


Figure 4.1. A Normal distribution with $\mu = 166$ and $\sigma = 15$, or $N(166,15)$. The y-axis indicates how likely we are to encounter a number on the x-axis. So, 100 or 250 are very unlikely (because they are in the tails of the distribution) but the likelihoods of say **200**, **150**, and **166** are increasingly larger.

The more likely values are closer to the mean. For example, the *likelihood* of the number 166 is just less that 0.03 (in fact its 0.026596), and the *likelihood* of 200 is 0.002038 (see Fig. 4.1). We'll be using this idea of *likelihood* a lot in Part 2 of this book.

It is straightforward to generate random variates from a pdf if the arguments are supplied (see Appendix G2). For example, here are 20 random variates generated (using the `rnorm()` command in R) from the distribution shown in Fig 4.1 ($\mu = 166$ and $\sigma = 15$):

```
>rnorm(n=20,mean=166,sd=15)
```

```
189.9792 174.9613 160.9644 146.1806 163.3794 156.3430 154.9649
169.8372 151.4414 166.7334 197.5020 178.7423 175.8785 170.9841
159.6283 141.2010 168.8701 137.4293 138.7290 147.6990
```

Figure 4.2 shows some examples of other Normal pdfs.

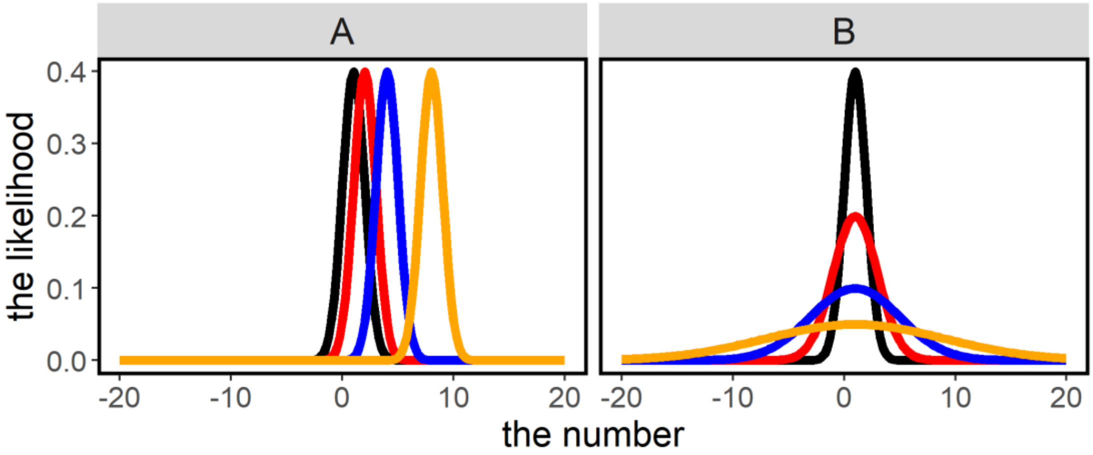


Fig. 4.2. A - Four different Normal distributions with increasing means from 1 to 8 but constant standard deviations (of 1): ***N*(μ=1, σ=1)**, ***N*(2,1)**, ***N*(4,1)**, and ***N*(8,1)**. B - Four different Normal distributions with increasing standard deviations from 1 to 8 but constant means (of 1),: ***N*(1,1)**, ***N*(1,2)**, ***N*(1,4)**, and ***N*(1,8)**.

A key point is that the mean and standard deviation can be controlled independently of each other. We can increase the mean and keep the standard deviation the same (Fig 4.2a), or we can increase the standard deviation and keep the mean the same (Fig 4.2b). Or we can do both. Why? *Because these are controlled by two separate arguments*.

Note that the area under each of these curves must always be one. So, the bigger the standard deviation (the more variation there is and the 'wider' the distribution), the flatter (or 'lower') the curve must go.

4.2 The Bernoulli distribution

As described in Chapter 3, this is a simple distribution that generates only two values – for simplicity we'll refer to them here as 0 and 1, but of course 0 and 1 can take on different meanings (negative, positive; fail, pass; no, yes; etc). The Bernoulli distribution has just one argument, denoted *p*, the probability of a '1'. The probability of a zero is thus 1-*p* (as the 'area under the pdf' must sum to 1). The mean of the distribution is *p*, and the variance (as it happens) is *p*(1-*p*). We might denote a Bernoulli distribution as Bern(*p*). The Bernoulli distribution is termed a discrete distribution because it can only generate discrete numbers – in this case, just 0 or 1. Examples are shown in Fig. 4.3.

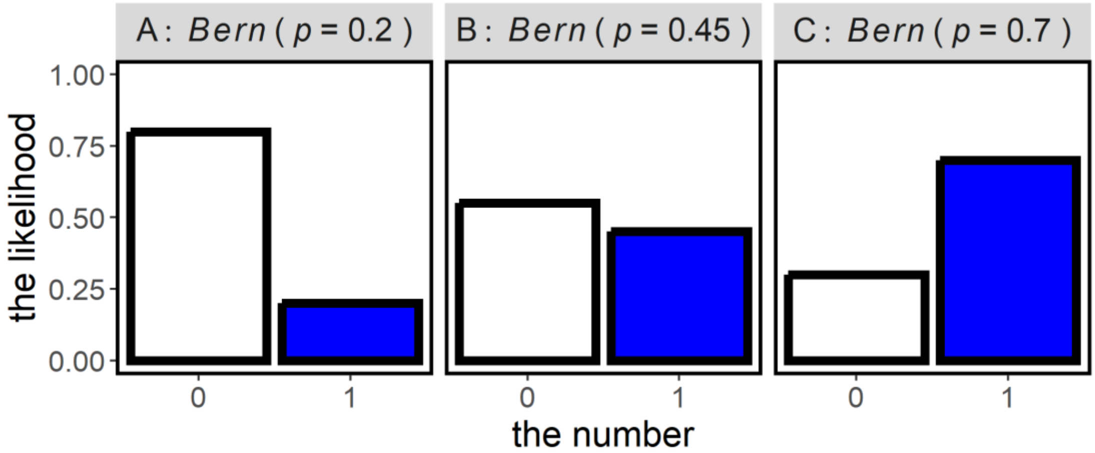


Figure 4.3. Three Bernoulli distributions with p = 0.2, 0.45, and 0.7, respectively. Because it's a discrete distribution the likelihood (or probability) changes in a step-like way.

The Bernoulli distribution is a special case of the Binomial distribution (see Appendix G.6) which is why it's often referred to as 'Binomial', but we shall persist with calling it Bernoulli despite how R references it.

Here are 20 random variates (generated from the `rbinom()` command in R using the pdf in Fig 4.3C (*p* = 0.7):

```
rbinom(n=20,prob=0.7)
```

```
1 1 0 0 1 1 0 1 1 0 1 0 0 1 1 0 0 0 1 1
```

The probability of generating a 1 from a from a Bern(0.7) distribution is … 0.7. The probability of generating a 0 from a Bern(0.7) distribution is … 0.3.

4.3 The Poisson distribution

The Poisson distribution generates non-negative integers between 0 and +infinity. It is therefore a discrete distribution. It is defined entirely by just one argument: the mean of the distribution, which happens also to be its variance. This mean (and variance) is conventionally termed λ (pronounced 'lam-dah'), and a Poisson distribution denoted Pois(λ). *Because both the mean and the variance are determined by a single number they cannot be varied separately*. If the mean is small, the variance must be small, and if the mean is large the variance must be equally large. Because Poisson variates look a lot like counts, we often start off with Poisson distributions to model count data.

Some examples are shown in Fig. 4.4.

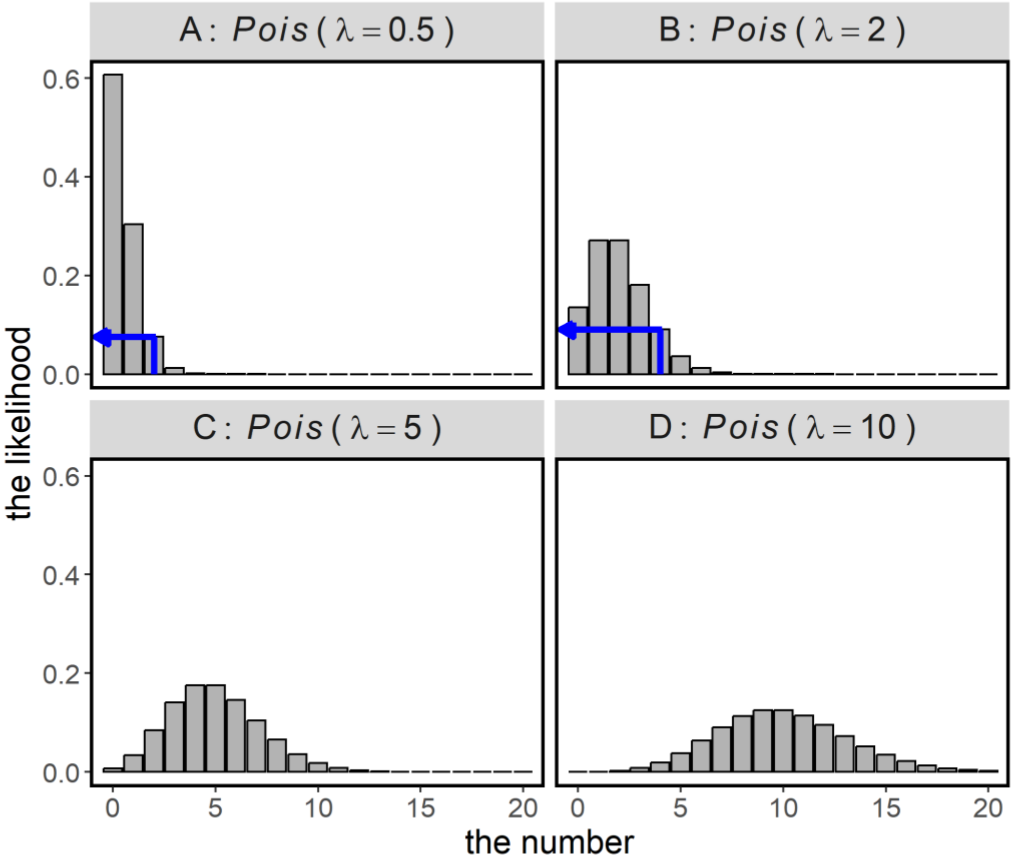


Figure 4.4. Four Poisson distributions with A) λ = 0.5, B) λ = 2, C) λ = 5, D) λ = 10 respectively. Because it is a discrete distribution the likelihood (or probability) changes in a step-like way.

Note how as the mean increases, so does the 'width' (the variance), and as the distributions get 'wider' they necessarily get 'flatter', as the area under the pdf must sum to one. Note also that when the mean is low the distribution cannot be symmetric (as negative numbers are not possible), but as the mean increases, the Poisson distribution becomes more symmetric and 'Normal' looking.

Here are 20 random variates generated (using the `rpois()` command in R) from the pdf in Fig 4.4B (λ = 2):

```
>rpois(n=20, lamnda=2)
2 2 1 2 1 2 0 3 4 2 3 3 2 1 3 2 3 1 5 2
```

The probability of generating a 2 from a Pois(0.5) distribution is 0.076 (as you can just about see from Fig. 4.4A), and the probability of generating a 4 from a Pois(2) distribution is a little higher: 0.090 (as you can just about see from Fig. 4.4B).

4.4 The Negative Binomial distribution

Like the Poisson distribution the Negative Binomial distribution generates non-negative integers between 0 and +infinity. It is therefore also a discrete distribution, and often used as an alternative to the Poisson distribution to model count distributions. However, it is defined entirely by two arguments, not one, and together these determine the mean and variance, *but because there are two arguments, like the Normal distribution these can be controlled independently of* *each other.* This will be enormously useful when we have count data that has 'more variance' than the mean (recall that for a Poisson distribution the mean and variance are locked to each other).

Negative Binomial distributions can be denoted: *NB*(μ, *k*). μ is the mean and *k* relates (inversely) to the variance (as shown in Appendix G.5), and is sometimes known as the 'size' parameter or 'theta' (in GLM outputs that implement this distribution). If *k* is very large the Negative Binomial converges on a Poisson distribution.

Figure 4.5 shows examples of the Negative Binomial distribution. They all have a mean of 5 but different variances:

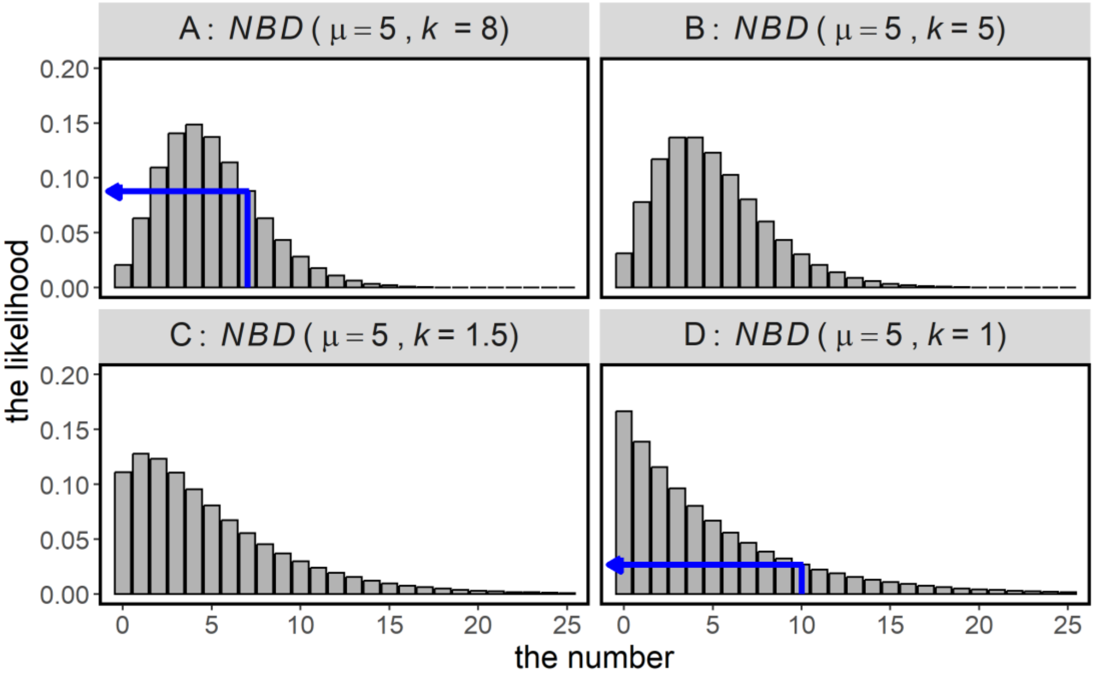


Figure 4.5. Four Negative Binomial distributions, all with a mean of 5, but decreasing *k* A) *NB*(5,80) , B) *NB*(5,5), C) *NB*(5,1.5), D) *NB*(5,1).

The probability of generating a 7 from an NB(5,80) is 0.103 (Fig. 4.5A), and the probability of generating a 10 from an NB(5,1) is 0.027 (Fig. 4.5D).

Here are 20 random variates (generated from the `rnbinom()` command in R) from the pdf shown in Fig 4.5A:

```
>rnbinom(n=20,mu=5,size=8)

9 5 4 5 6 4 5 2 9 4 2 4 1 4 3 4 3 5 7 3
```

and 20 random variates from the pdf shown in Fig 4.5D:

```
>rnbinom(n=20,mu=5,size=1)

7 16 1 2 1 4 3 3 4 4 9 2 8 6 5 0 2 2 8 17
```

Note the increased variance in the second set of random variates.

Important ideas to take-away

- Probability density functions have arguments that fully specify their position (where the mean is) and shape
- They can be used to generate random variates, and calculate how likely different numbers are to come from them
- There are many different sorts of pdf, some continuous and some discrete
- In principle there are many pdfs that could be used to model data, but we have described the four most commonly used ones
- Probability density functions can be used to represent a range of different data types, for example: Normal – for continuous data, Bernoulli – for binary data, and Poisson or Negative Binomial for count data
- There are good fundamental reasons why we would choose these distributions to model different types of response variable (discussed in greater depth in Appendix G)
- It is worth remembering the arguments that go with each of these four common distributions

Chapter 5

Our example data

(back to Contents)

This chapter introduces the data set we will use throughout the text.

Before proceeding further we'll introduce you to a data set that we will use to illustrate methodology throughout the rest of this text. Although the data set will be quite complex, we will use subsets of the data to make various points, and the use of one data set will mean you only need to recall one data context, and thereby minimize distractions unrelated to the central ideas.

The hypothetical data set we will call the *Four Rivers data set*, describes a large-scale survey of river water quality.

Water samples were taken from 12 different Sites along 4 different Rivers. Each sample is divided into 5 subsamples and each subsample is dispatched to one of 5 different laboratories (Labs). Each of the sites is designated as a river running through a Landscape that is either *Rural* or *Urban*, the Flow rate of the river at each site is recorded as either *High*, *Medium* or *Low*, and the Temperature of the water at the site is recorded. The water is subject to analyses for concentration of Phosphate and Nitrate at each of the 5 labs. These 8 variables are all potential explanatory variables.

We are interested in whether we can explain variation in 4 other variables (our response variables) based on different combinations of the 8 explanatory variables. There are 4 different Response variables that we may choose to examine (each one separately from the others). Each of these 4 variables was measured at each of the 5 different laboratories. The measures (Table 5.1) were: 1) the concentration of Chlorophyll in the samples; 2) counts of the number of zooplankton in the sample (ZooCount); 3) counts of the number of bacterial colonies in the sample (BacCount); and 4) a binary measurement of whether the zooplankton in the sample showed evidence of a fungal disease (Disease). Table 5.1 summarizes the variables in the data set.

Table 5.1. Classification of the variables in the Four Rivers data set.

	Type	Units
Response variable		
Chlorophyll	Continuous	μg/L
ZooCount	Discrete (count)	-
BacCount	Discrete (count)	
Disease	Binary	Presence (1) /absence (0)
Explanatory variables		
Site	Categorical (12 levels)	S01-S12
Lab	Categorical (5 levels)	L1-L5
Flow	Categorical (3 levels)	H, M, L
Landscape	Categorical (2 levels)	R(ural), U(rban)
River	Categorical (4 levels)	R1, R2, R3, R4
Temp	Continuous	°C
Phosphate	Continuous	μg/L
Nitrate	Continuous	mg/L

As you can see in Table 1, the response variables are of 3 different types – continuous, discrete and binary. Chlorophyll is a continuous variable (average = 68.3 μg/L – Fig. 5.1A). ZooCount and BacCount are both discrete counts, with averages of 4.9 and 82.7, respectively, (Fig. 5.1B,C). Finally, Disease presence is a binary variable (Figure 5.1D), with disease present in 60.4% of samples.

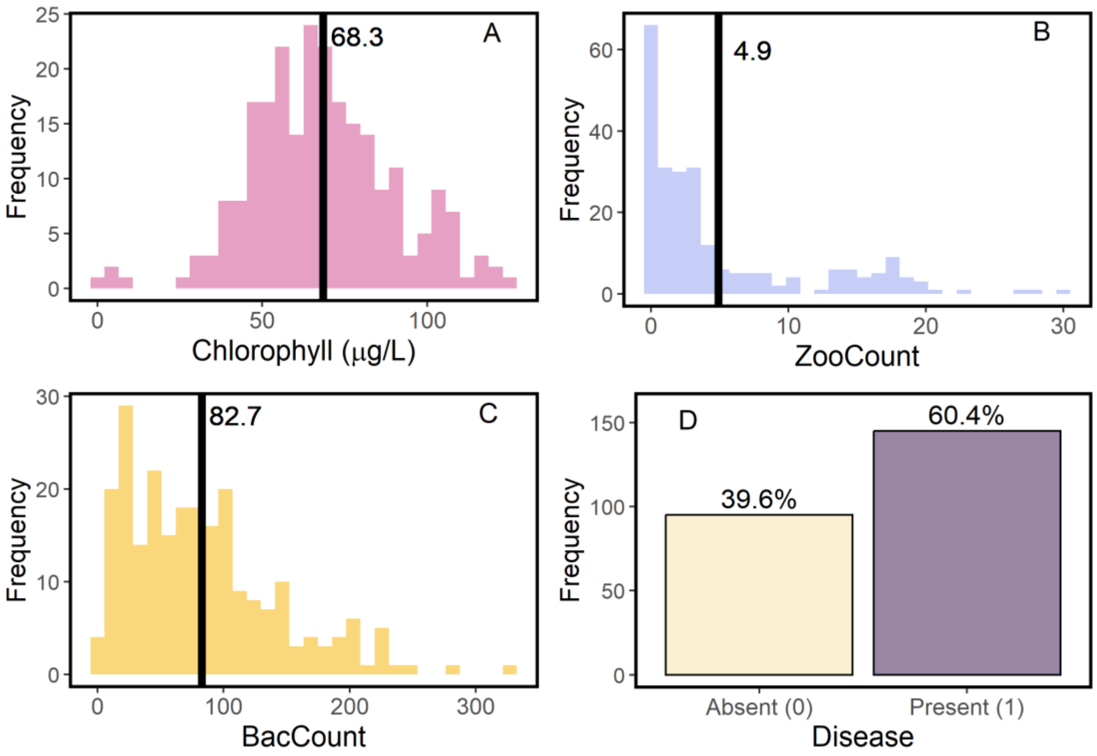


Figure 5.1. Summary of Four Rivers response variables: A – Chlorophyll, B – Zooplankton count, C– Bacterial count, D – Disease presence, with the mean values indicated for A – C and prevalence of disease presence for D.

The explanatory variables are of 2 different types – categorical and continuous. The 5 categorical explanatory variables can be summarised as:

```
 River     Site     Lab     Flow   Landscape
 R1:60    S01:20   L1:48    H:80   R:120
 R2:60    S02:20   L2:48    M:80   U:120
 R3:60    S03:20   L3:48    L:80
 R4:60    S04:20   L4:48
          S05:20   L5:48
          S06:20
          S07:20
          S08:20
          S09:20
          S10:20
          S11:20
          S12:20
```

In this dataset categorical explanatory variables are **balanced** – *i.e.* each level within a variable has the same number of observations. Balance is a desirable but not an essential property of a data set.

The data may look like Table 5.2.

Table 5.2. The structure of the Four Rivers data set, where the cells shaded olive (categorical) and green (continuous) are the explanatory variables and the pink, blue, orange and purple columns are the 4 possible response variables. Each row is a data record, and there are 4 rivers x 12 sites x 5 replicates sent to different labs = 240 records altogether.

Record #	River	Site	Lab	Flow	Landscape	Temp	Phosphate	Nitrate	Chlorophyll	ZooCount	BacCount	Disease
1	R1	S01	L1	L	Urban	11.49	38.91	13.1	56.30	3	13	0
2	R1	S01	L2	L	Urban	12.97	0.06	13.18	75.93	2	0	1
3	R1	S01	L3	L	Urban	12.03	38.71	14.37	68.75	5	8	1
4	R1	S01	L4	L	Urban	11.55	5.24	13.06	75.67	1	1	0
5	R1	S01	L5	L	Urban	11.62	68.37	10.38	57.82	3	25	0
6	R1	S02	L1	L	Urban	9.28	70.67	11.93	49.53	1	13	0
7	R1	S02	L2	L	Urban	10.62	85.94	11.84	67.07	3	14	0
8	R1	S02	L3	L	Urban	12.37	60.02	10.47	53.02	0	9	1
9	R1	S02	L4	L	Urban	11.03	99.92	12.39	70.59	1	35	1
10	R1	S02	L5	L	Urban	13.54	78.56	11.73	61.45	2	18	1
11	R1	S03	L1	M	Urban	11.62	100.68	11.31	54.33	3	43	0
12	R1	S03	L2	M	Urban	13.57	104.83	12	80.00	1	40	1
13	R1	S03	L3	M	Urban	12.65	122.11	13.62	76.88	0	55	0
14	R1	S03	L4	M	Urban	12.36	104.42	16.45	106.26	1	26	0
15	R1	S03	L5	M	Urban	10.81	105.07	13.57	83.78	4	23	0
16	R1	S04	L1	M	Urban	11.59	118	14.83	76.07	0	26	0
17	R1	S04	L2	M	Urban	11.74	170.42	13.97	87.65	1	64	0
18	R1	S04	L3	M	Urban	11.79	149.72	11.76	67.33	1	28	1
:	:	:	:	:	:	:	:	:	:	:	:	:
61	R2	S01	L1	L	Urban	11.93	139.94	9.6	35.86	1	53	0
62	R2	S01	L2	L	Urban	13.69	125.19	8.45	48.23	3	23	1
63	R2	S01	L3	L	Urban	13.32	154.51	8.99	40.19	3	47	1
64	R2	S01	L4	L	Urban	13.33	147.18	8.28	48.73	1	83	1
65	R2	S01	L5	L	Urban	12.6	166.25	15.25	70.14	2	19	1
66	R2	S02	L1	L	Urban	11.23	171.36	14.99	57.15	1	43	0
67	R2	S02	L2	L	Urban	13.18	182.93	15.39	72.64	4	43	1
68	R2	S02	L3	L	Urban	11.62	174.11	14.47	66.56	1	47	1
69	R2	S02	L4	L	Urban	12.59	132.77	12.91	68.53	1	49	1
70	R2	S02	L5	L	Urban	11.3	115.28	12.64	62.13	0	28	1
71	R2	S03	L1	M	Urban	11.4	130.39	14.42	66.38	2	28	0
72	R2	S03	L2	M	Urban	12.85	117.45	14.68	87.66	2	36	1
73	R2	S03	L3	M	Urban	12.19	102.21	10.84	55.98	3	14	0
74	R2	S03	L4	M	Urban	12.09	93.66	9.35	63.80	3	26	1
75	R2	S03	L5	M	Urban	13.19	109.48	8.53	48.89	3	22	0
:	:	:	:	:	:	:	:	:	:	:	:	:
228	R4	S10	L3	M	Rural	15.62	418.98	8.31	64.38	7	193	1
229	R4	S10	L4	M	Rural	14.24	365.57	8.16	77.03	6	151	1
230	R4	S10	L5	M	Rural	15.6	356.06	6.56	59.54	9	196	1
231	R4	S11	L1	H	Rural	14.59	368.83	7.45	59.17	1	198	1
232	R4	S11	L2	H	Rural	16.51	342.24	9.92	90.20	0	187	1
233	R4	S11	L3	H	Rural	14.63	398.97	6.65	57.90	0	120	1
234	R4	S11	L4	H	Rural	14.77	393.74	6.94	76.72	0	117	1
235	R4	S11	L5	H	Rural	14.32	391.69	5.26	53.88	1	89	1
236	R4	S12	L1	H	Rural	15.73	401.57	5.81	42.88	0	206	1
237	R4	S12	L2	H	Rural	14.03	389.38	10.64	100.21	0	277	1
238	R4	S12	L3	H	Rural	13.21	392.46	8.55	73.54	0	107	0
239	R4	S12	L4	H	Rural	14.19	376.48	12.91	114.90	0	93	0
240	R4	S12	L5	H	Rural	14.05	375.62	12.16	104.59	1	133	0

We would strongly recommend that you explore how both your response and explanatory variables range and co-vary with each other, prior to launching into any formal model building or inference. Doing so will help with your interpretation, both of the statistical models and the biological questions you are wanting to explore.

For the rest of this book we will make use of specified subsets of the Four Rivers data set to illustrate different types of possible analysis. We will be able to ask questions such as:

Can I explain variation in the Chlorophyll concentration in different samples using variation in their Phosphate concentration?

Can I explain variation in the Chlorophyll concentration in different samples using variation in their Nitrate concentration and Temperature?

Can I explain variation in the count of Zooplankton in different samples using variation in the Rivers from which they came?

Can I explain variation in the count of Zooplankton in different samples using variation in the Flow rate, Landscape and River from which they came?

Can I explain variation in the prevalence of diseased zooplankton in different samples using variation in the Temperature of water at these sites?

Can I explain variation in the prevalence of diseased zooplankton in different samples using variation in results of analyses from different Laboratories?

As we will see later, it will also be possible to ask more complex questions of a similar form.

Important ideas to take-away

- Recognize what type of response variable(s) you have in your data set – continuous, discrete, or binary
- Recognize what type of explanatory variable(s) you have in your data set - continuous, discrete, binary
- Construct your dataset to have one record per row

Chapter 6

GLM construction and fitted values

(back to Contents)

This chapter introduces how the models are constructed and used to generate fitted values (also known sometimes as predicted values) of the response variable. We introduce the basics of how to model the influences of one categorical or one continuous explanatory variable for simple models that assume individual observations of the response variable are Normally distributed. We also introduce GLMs that contain two explanatory variables. The modelling of scatter around these fitted values is considered in Chapter 7, models that contain more than two explanatory variables are considered in Chapter 8, and how these models are actually fitted to data is considered in Chapter 13.

GLMs take the following form:

the left-hand side ~ the right-hand side

The **left-hand side** is our best estimates of something closely related to observations of the response variable – these are called **fitted values**. The model will generate a fitted value for each observation of the response variable, thus one fitted value per record, and we will assume there are *n* such records.

Understanding the arithmetic structure of the models is fundamental, and key to full understanding of the output. To ensure this understanding requires some investment in algebraic notation, including the use of subscripts which we describe more fully in Appendix H. It can look a bit intimidating, but take the time to become comfortable with the basic principles because its fundamentally important.

Denote the i^{th} fitted value by f_i (note that *i* can therefore take on any integer value between 1 and *n*). We seek to develop estimates of the fitted values in terms of the explanatory variables we have chosen to include in the model. Remember, there is only one response variable in a univariate model but there may be several explanatory variables. The **right-hand side** (also known as the **linear predictor**) of the model will take the form of a somewhat arbitrary value of the response variable (which we will call a reference value) with a series of *adjustments* made that are dependent on the explanatory variables included in the model. *If we are confident that an adjustment for any given explanatory variable is different from zero, then we can infer that the adjustment is useful in determining the fitted value, and, in turn, the explanatory variable is useful in explaining variation in the response variable.*

For example, we may wish to ask whether we can explain variation (that is, identify differences) in the concentration of Chlorophyll in water samples depending on whether they were taken from an Urban or Rural Landscape:

$$\text{Chlorophyll}_i = \text{reference value} + \text{adjustment for Landscape}_i$$

Or we could ask whether the concentration of Chlorophyll in water samples varies depending on the amount of Nitrate in the same samples

$$\text{Chlorophyll}_i = \text{reference value} + \text{adjustment for Nitrate}_i$$

or perhaps combining both potential influences together:

$$\text{Chlorophyll}_i = \text{reference value} + \text{adjustment for Landscape}_i$$
$$+ \text{adjustment for Nitrate}_i$$

The explanatory variables – here Landscape and Nitrate are sometimes called **main effects**.

So now we need to do a number of things:

1) Build the model - deciding which explanatory variables to include in the model (which adjustments we want to include in the model)

2) Fit the model to the data (estimate what the adjustments are)

3) Determine how confident we are that the adjustments are different from zero (conduct inference)

6.1 Adjustments

We make an adjustment for each explanatory variable (each of these adjustments is sometimes referred to as a term). What form do these adjustments take? It depends on whether the explanatory variable is categorical or continuous.

6.2 Categorical adjustments

Landscape is a categorical variable referring to whether the river is running through an urban or rural environment. It has two levels: Rural and Urban. There are no numbers that can be used in place of the labels 'Urban' or 'Rural' .. they are the *names* of categories and obviously different to – say – Nitrate, which can be measured *numerically* as a concentration.

If we want to include an *adjustment* for whether the sample was taken from a rural or urban environment we will need to estimate the magnitude of these adjustments, and we'll make the same adjustment for all samples from Rural environments, and another different adjustment for all samples from Urban environments. We could choose a variety of ways of denoting these adjustments, but here we will use lower case Greek letters used in alphabetical order (α, β, γ, δ, ε, etc). So, the adjustment for Landscape might be denoted α_j, where j represents R for Rural (α_R), or U for Urban (α_U). The **algebraic structure** of the model would be:

$$f_i = c + \alpha_j \qquad \text{(model 6.1)}$$

In this case for every fitted value, α_j will take one of two values. One of these levels will always have an adjustment of zero, as it will be the reference level. Here, the Rural level is taken as the reference level, and all adjustments for other levels *are relative to this reference*. Hence, for Rural landscapes $\alpha_R = 0$, and there will be no adjustment because Chlorophyll concentrations in Rural landscapes can be represented by c. In Urban landscapes, α_u would be estimated when we fit the GLM to our data. Suppose we arrive at an estimate of α_u = -9.711, and suppose *c* is estimated to be 61.515. Thus, the model generates just two different fitted values for Chlorophyll concentration, one for all the samples from Rural landscapes and one for all the samples from Urban landscapes:

$$f_i = c + \alpha_R \quad = \quad 61.515 + 0 = 61.515 \ \mu\text{g/L}$$

$$f_i = c + \alpha_U \quad = \quad 61.515 + (-9.711) = 51.804 \ \mu\text{g/L}$$

The fitted values reflect the fact that Chlorophyll concentrations tend to be less in Urban landscapes by about one sixth of those from Rural landscapes.

A key point here is that there is no 'per unit adjustment' for categorical variables as the different levels don't by their nature have units. A second important point is that one of these levels will always have an adjustment of zero, as it is represented by the reference level. Here, the Rural level is taken as the reference level, and all adjustments for other levels *are relative to this reference*. This makes some sense - if we need to fit just two different numerical values (one for Rural landscapes and one for Urban landscapes) we don't need 3 different coefficients (*c*, $\alpha_{R,}$ and α_U) to do so ... two will be enough. If we are confident that α_u is not zero, then we might conclude that Chlorophyll concentrations are partly explained by the type of Landscape a river runs through.

How is the reference level chosen? It doesn't matter a whole lot, and it will depend on the software you are using. In R, it is whichever level-label starts with the letter earliest in the alphabet (here *R* comes before *U* so *R* is automatically selected to be the reference) although you can change this if you want to using the `relevel()` command.

6.3 Continuous adjustments

Nitrate is continuous, so we measure the adjustment *per unit* concentration of Nitrate (a unit will depend on the explanatory variable, a unit of Nitrate here is 1 mg/L, for Temp it is 1°C and for Phosphate it is 1 μg/L). We denote this unit adjustment usually by the letter *m*, in this case subscripted by *N* for Nitrate: m_N. And because it's per unit of Nitrate, we need to multiply it by the number of units of Nitrate in the i^{th} sample – which we might generically denote *x,* subscripted by *N* for Nitrate and *i* for the i^{th} sample: $x_{N,i}$ (check out Appendix H if you find the use of subscripts confusing). Our reference point will be denoted by *c*. Hence, we have the following algebraic structure for the model:

$$f_i = c + m_N \, x_{N,i} \qquad \text{(model 6.2)}$$

f_i denotes our fitted values of the Chlorophyll in each sample conditional on its Nitrate concentration. (The absence of any other sign between the m_N and $x_{N,i}$ denotes they should be multiplied together).

(This should look familiar to you—the simple regression we discussed in Chapter 2 is exactly this model, with *m* denoting the slope and *c* the intercept, which is our arbitrary reference value)

Because Nitrate is continuous, each of the $x_{N,i}$'s are potentially unique, and we will likely have a slightly different adjustment for each sample. We can see from examining the accompanying Four Rivers data (Table 5.2) that, for example, the explanatory variable for Nitrate concentration from record 15 ($x_{N,15}$) was 13.57 mg/L, and from record 64, $x_{N,64}$ = 8.28 mg/L, and from record 236, $x_{N,236}$ = 5.81 mg/L. Suppose that we had estimated the coefficient m_N to be say 4.759, and the coefficient for the reference point, *c*, to be 11.195 (and these values would be generated by fitting our model to data as we shall see later) we'd be predicting values of Chlorophyll to be as follows:

record 15: 11.195 + 4.759 x *13.57* = 75.774 μg/L

record 64: 11.195 + 4.759 x *8.28* = 50.600 μg/L

record 236: 11.195 + 4.759 x *5.81* = 38.845 μg/L

(here the explanatory data are in *red italics*, and coefficients are in black).

We actually observed Chlorophyll at these sites to be: 83.78, 48.73, and 42.88 μg/L (so perhaps the model is not doing too badly!)

We typically call *m* a slope, and adjustments for continuous explanatory variables are invariably *per unit* estimates of the effect of the explanatory variable on the response variable. Note that in this case our arbitrary reference value is the value predicted (or fitted) by the model in the event of zero nitrate (11.195 + 4.759 x *0* = 11.195 μg/L) – commonly known as the (y-axis) 'intercept'. It may be that there are no samples with zero Nitrate, but the adjustments need to be made relative to some Nitrate concentration and zero is the simplest choice.

6.4 The structure of a GLM

So – the influence of categorical explanatory variables on response variables is captured by adjustments for the different levels (not counting the one level that is assigned to be the reference) and the influence of continuous explanatory variables on response variables is captured by slopes. And we can include as many adjustments in the model as we want, depending on how many explanatory variables we choose to include in the model. Thus, GLMs are comprised of these two slightly different components: adjustments for continuous explanatory variables and adjustments for categorical explanatory variables. The numerical values of these adjustments are included in the output from fitting the model to the data. We are specifically interested in our confidence that these adjustments are different to zero, and from this we can determine whether the explanatory variables help to explain variation in the response variable.

The right-hand sides of GLMs are always constructed in the same way (*c + adjustments*), regardless of the distribution you choose to model the response variable with (i.e. Normal, Poisson, Negative Binomial, Bernoulli, etc). However, *exactly* how we interpret these 'right hand sides' *does* depend a bit on this choice of distribution. In the following sections we'll provide you with examples of different models, fitted to different subsets of the Four Rivers data set.

6.5 Example

So, how do you actually fit these models in R? It is very simple indeed. Let's consider the subset of data from Lab 1, and fit the 3 models we have discussed so far.

Model 6.1: Chlorophyll *(depends on)* Landscape

Model 6.2: Chlorophyll *(depends on)* Nitrate

And

Model 6.3: Chlorophyll *(depends on)* Nitrate *(and)* Landscape

There are 4 rivers, and 12 sites per river so 48 different samples that were sent to Lab 1. We could make two plots (Fig 6.1), with the different explanatory variables on the x-axis:

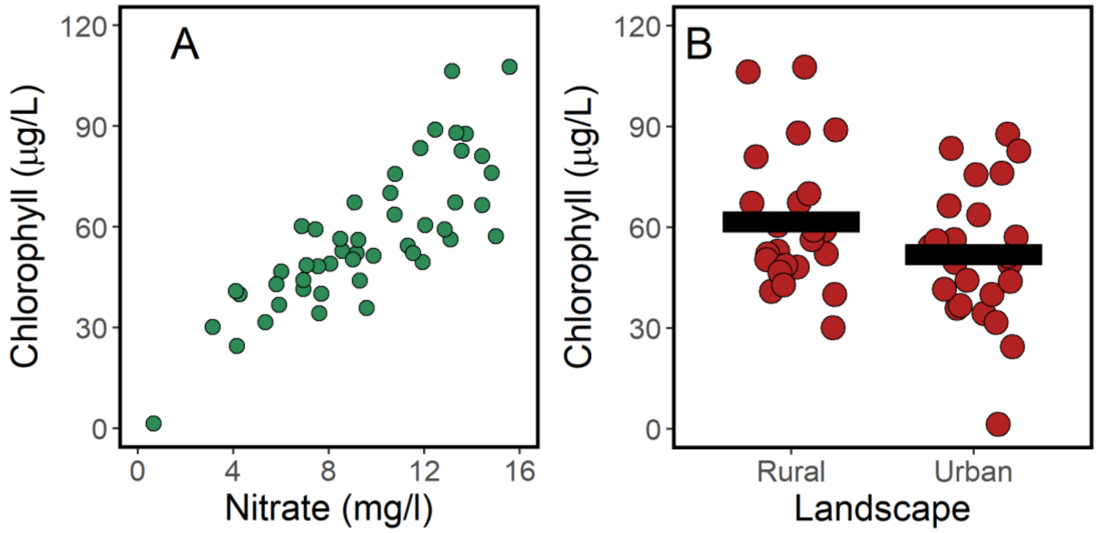


Fig. 6.1. Plots of Chlorophyll against the explanatory variables Nitrate and Landscape. A) The scatter plot shows the distribution of Chlorophyll values (y axis) in relation to Nitrate (x axis). B) The dot plot on the right shows the mean (bar), and individual Chlorophyll values (y axis) in relation to Rural or Urban landscapes (x axis).

6.6 Including a categorical explanatory variable

We'll read the data into R using `read.csv()` and subset the data to focus on samples sent to laboratory 1:

```
> Four_Rivers_data <- read.csv('Four_River_data.csv')
> my_data <- subset(Four_Rivers_data, Lab == 'L1')
```

The R-command instructing R to fit model 6.1 would be:

```
> model_6.1 <- glm(Chlorophyll ~ Landscape,
           data = my_data, family = gaussian)
```

The command tells R to use a single categorical explanatory variable, Landscape, to model the response variable, Chlorophyll, as a Normally distributed observation (denoted by `family = gaussian`). We do not actually need to put the family here as it is assumed to be gaussian by default but we include it here for completeness.

As discussed in section 6.2, the algebraic structure of the model will take the form:

$$f_i = c + \alpha_j \qquad (i = 1 \,..\, 48,\ j = R \text{ or } U)$$

where fitted Chlorophyll values depend on a reference level (which happens to be the average Chlorophyll concentration in Rural Landscapes) + an adjustment for when the Landscape is Urban.

We could examine the output using the `summary()` command

```
> summary(model_6.1)
```

which yields:

```
Call: glm(formula = Chlorophyll ~ Landscape,
          data = my_data, family = gaussian)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      61.515      4.152  14.817   <2e-16 ***
LandscapeU       -9.711      5.871  -1.654    0.105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 413.6859)

    Null deviance: 20161  on 47  degrees of freedom
Residual deviance: 19030  on 46  degrees of freedom
AIC: 429.38

Number of Fisher Scoring iterations: 2
```

Focus on the bold – we are given "(Intercept)" – this is *c*. The estimate indicates that the value of *c* is 61.515, which represents the average Chlorophyll concentration in samples from Rural landscapes. We are also given a **`LandscapeU`** estimate - we make an adjustment of -9.711 to obtain the average Chlorophyll concentration in samples from Urban Landscapes (note the U after Landscape – **`LandscapeU`** - indicating that this is the coefficient for Urban). Where is the adjustment for Rural? Because the Rural level is the reference level (recall that *R* precedes *U* in the alphabet!), it is by definition zero, and so it is not reported. (We also are provided with other output (coloured grey) which we are going to ignore for the rest of Part 1 of this book and will not even discuss until Part 2)

It is important to recognize that the model only generates two fitted values for Chlorophyll concentration:

$$f_i = 61.515 + 0 \;=\; 61.515\ \mu\text{g/L (all Rural landscape samples)}$$

and

$$f_i = 61.515 - 9.711 \;=\; 51.804\ \mu\text{g/L (all Urban landscape samples)}$$

This difference can be observed in Fig. 6.2. The fitted values capture the average for each level of landscape, but of course there is 'scatter' not explained by landscape.

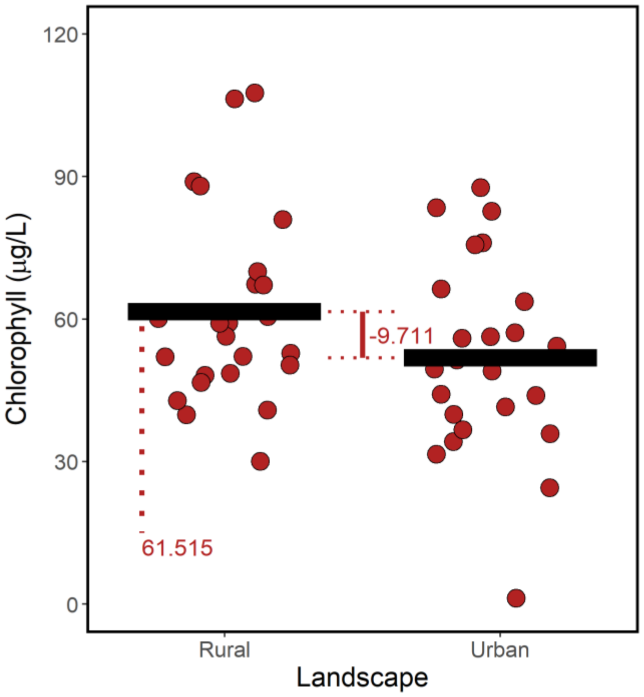


Figure 6.2. Plot of Chlorophyll vs the explanatory variable Landscape, with the fitted values indicated by horizontal bars.

The reason we are fitting the GLM is so we can determine the effect of the different levels of landscape (which we have established to differ by -9.711 ug/L). In Part 2 of the text we will establish how confident we are that the effect of the landscapes really are different, and this is achieved by determining whether -9.711 is '*significantly'* different from zero.

6.7 Including a continuous explanatory variable

The R-command instructing R to fit model 6.2 would be:

```
> model_6.2<-glm(Chlorophyll~Nitrate,data=my_data,family=gaussian)
```

The command tells R to use a single continuous explanatory variable, Nitrate, to model the response variable, Chlorophyll, as a Normally distributed observation.

As discussed in section 6.3, the algebraic structure of the model will take the form:

$$f_i = c + m_N \, x_{N,i}$$

where the fitted value for every observation of Chlorophyll concentration is the reference point '*c*' adjusted by the product $m_N \, x_{N,i}$, which of course depends on the observed Nitrate value $x_{N,I}$ in each sample.

We examine the output using the `summary()` command

```
> summary(model_6.2)
```

It generates quite a lot of unnecessary information for our purposes, so we reproduce only the bits we're currently focusing on:

```
Call: glm(Chlorophyll ~ Nitrate, data = my_data,
            family = gaussian)
  Coefficients:
                   Estimate Std. Error
  (Intercept)        11.195      5.119
  Nitrate             4.759      0.503
```

The model output shows that the per unit effect of Nitrate on Chlorophyll (the slope m_N) is estimated to be 4.759 and the value of *c* (the y-axis intercept) to be 11.195. We can plot this line on top of the data, as in Fig. 6.3:

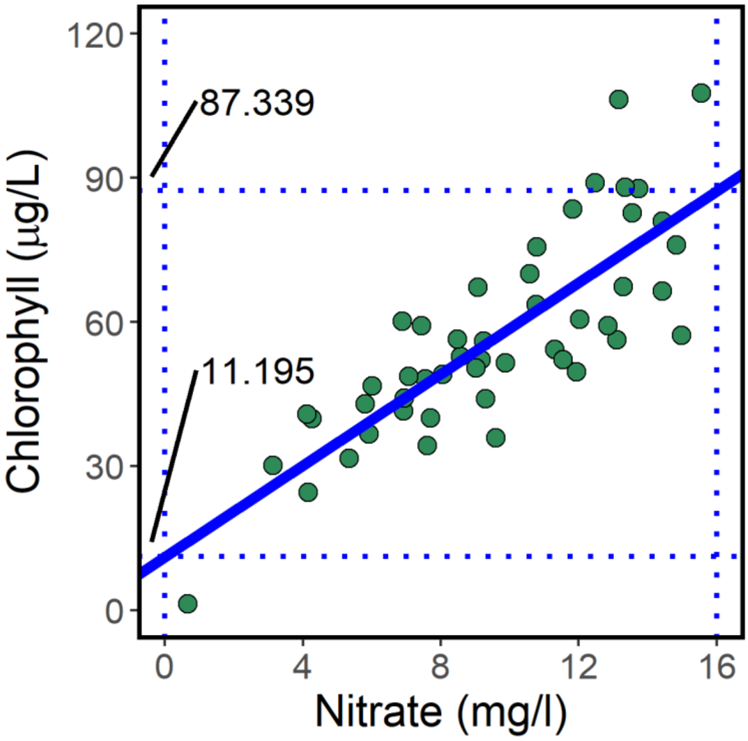


Figure 6.3. Plot of Chlorophyll vs the explanatory variable Nitrate, with the line of best fit (solid blue), and dotted lines to guide the estimation of the slope.

Note how the slope (solid blue line) is positive - it rises from just over 11 at 0 Nitrate to just over 87 on 16 units of nitrate. The slope is the change in y divided by the change in x: about 76/16, which is about 4.75. The fitted values capture the effect of nitrate on chlorophyll (the line) but of course there is 'scatter' not explained by nitrate.

The actual GLM is:

$$f_i = 11.195 + 4.759\, x_{N,i} \qquad (i = 1 .. 48)$$

and it will generate 48 different fitted values for each different observed value of $x_{N,i}$. Looking at Fig. 6.3 we can see the line appears to pass through the middle of the points, reflecting what looks like a reasonably consistent positive relationship between the concentration of nitrate and chlorophyll.

The reason we are fitting the GLM is so we can determine the significance of the unit adjustment for nitrate (estimated to be 4.759 ug/L of chlorophyll per mg/L of nitrate). We are specifically interested in how confident we can be that 4.759 is different from zero (a slope that would reflect a horizontal line on Fig. 6.3), but we'll come to that later.

6.8 Including more than one explanatory variable

We can combine models 6.1 (single categorial explanatory variable) and 6.2 (single continuous explanatory variable), and there are very good reasons for doing so (which we'll discuss later). Fortunately, being a GLM we can put multiple explanatory variables into the same model – for example, both Nitrate and Landscape:

The R-command instructing R to fit the combined model would be:

```
> model_6.3<-glm(Chlorophyll~ Nitrate + Landscape,data=my_data,
family=gaussian)
```

The command tells R to use a both a categorical (Landscape) and continuous (Nitrate) explanatory variable to model the response variable, Chlorophyll, as a Normally (Gaussian) distributed observation.

The algebraic structure of the model will take the form:

$$f_i = c + \alpha_j + m_N\, x_{N,i} \qquad (i = 1 .. 48, j = R \text{ or } U)$$

(model 6.3)

Where every fitted value for Chlorophyll depends on some reference value, *c*, an *adjustment* if the landscape is Urban and a per unit of Nitrate *adjustment*. We can think of this as the model actually containing two intercepts, $c + \alpha_R$ and $c + \alpha_U$, and one slope m_N.

We examine the output using the `summary()` command

```
> summary(model_6.3)
```

which would generate this output:

```
Call:
glm(formula = Chlorophyll ~ Landscape + Nitrate, data = my_data,
family=gaussian)

Coefficients:
            Estimate Std. Error
(Intercept)   16.122      4.800
LandscapeU   -11.053      3.160
Nitrate        4.822      0.451
```

The estimate indicates the reference value (*c*) of 16.122 μg/L - this is the mean Chlorophyll value for Rural samples when the Nitrate concentration is zero. For Urban Landscape samples, when Nitrate concentrations are zero, the adjustment to Chlorophyll is -11.053 μg/L. But we make the *same* per unit Nitrate adjustment of

4.822 µg/L of Chlorophyll per mg/L of Nitrate for both Rural and Urban landscapes. Thus, the fitted value for a sample that contains 9 mg/L of Nitrate from an Urban landscape would be:

$$f_i = 16.122 + (-11.053) + \mathit{4.828} \times \mathit{9} = 48.521\ \mu g/L$$

(continuous explanatory data are indicated here in *green italics*)

We can graph this model as in Fig. 6.4; here we plot separate lines for the two types of landscapes - the slopes are parallel (there is only one slope) but the observations of the response variable are generally lower in Urban Landscapes compared to Rural (reflected by the negative adjustment to the intercept):

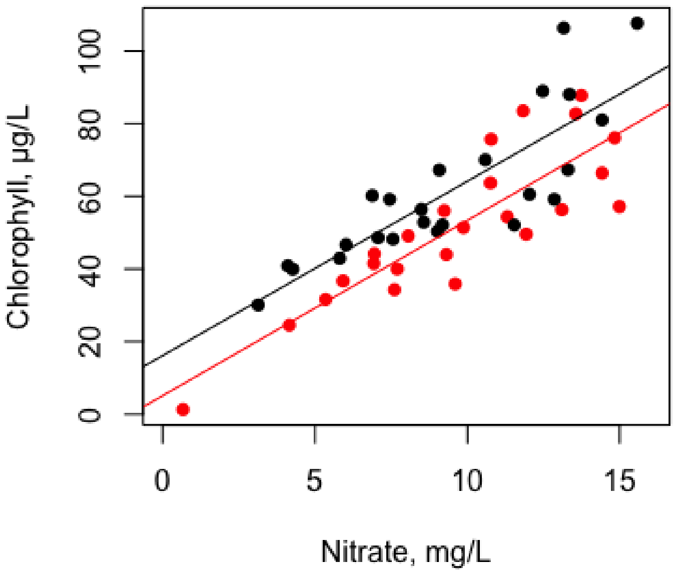


Figure 6.4. Plot of Chlorophyll concentration with Nitrate. The red points and line correspond to samples taken from rivers flowing through Urban landscapes whereas the black points and line represent rivers flowing through Rural landscapes.

The model corresponds to a well-known situation in which the different levels of the categorical explanatory variable cause the lines to intercept at different places on the y-axis, but the *effect* of a unit of the continuous explanatory variable has the *same* effect on the response variable, *regardless* of the level of the categorical explanatory variable, i.e. the lines are parallel. That is, the influence of the continuous explanatory variable does not depend on the categorical explanatory variable.

6.9 Continuous or categorical?

It is possible your explanatory variable could be treated as either continuous or categorical. For example, Flow rate could be regarded as continuous (it is after all **ordinal**), or categorical. The benefit of modelling explanatory variables as (simply) continuous is that they require only one coefficient, the slope, but we assume that the response variable changes linearly with changes in the explanatory variable. The benefit of modelling explanatory variables as categorical is that no linear relationship is assumed between the different (**nominal**) levels (for example, the response variable could be low for Low Flows, high for Medium Flows, and low again for High Flows); the downside is that this flexibility comes at a price: a coefficient is required for every level except the reference and this consumes degrees of freedom (as we'll see later) which is generally undesirable. If eye-balling the relationship suggests its linear, probably best to start by treating it as continuous.

6.10 Effect sizes

An **effect size** is how much the response variable changes in 'response' to changes in the explanatory variables. Usually we report effect sizes for continuous explanatory variables by how much the response variable changes in response to a unit change in the explanatory variable (i.e. from say a value of x_i to x_i+1). This of course is simply the value of the slope *m*, but there is no reason why effect sizes cannot be reported in response to multiple unit changes in the explanatory variable so long as this is fully explained. For categorical explanatory variables, the effect sizes are the changes in the response variable that result from a change from one level to another. If the change is from or to the reference level then these are simply the adjustments reported in the model summary. If the required effect size is the result of changing from one non-reference level to another non-reference level, then this could be calculated manually by comparison of the fitted values for the levels that are to be compared.

The size of the effect on its own is no guide to whether we can be confident its different to zero or not. You may encounter large effects sizes that might be indistinguishable from zero, or very small (and possibly biologically meaningless effect sizes) that we can be very confident are different to zero. In general it is a good idea to report an effect size with an estimate of its **confidence intervals**.

> There are various R commands that are useful for calculating summaries of fitted values like `ggeffects(model, 'term')` in the package `ggpredict`, or `emmeans(model ~ term)` in the package `emmeans`).

While it is usual to consider effect sizes from changing one explanatory variable at a time, the effect of any change of interest to the 'right hand side' of the GLM on the left hand side of the GLM is a legitimate 'effect size' so long as it is made clear exactly what the change (or combinations of changes) are being made. Effect sizes get a bit more complicated in the presence of interactions (see 11.5).

6.11 Further elaborations

If you've understood up to this point, you're mostly there! There are 4 important areas of further development when it comes to model *construction* (as opposed to understanding and interpreting all the output). These are:

1) What if you needed to fit lines that were not parallel? That is to say … what if the per unit effect of Nitrate on Chlorophyll (*i.e.* the slopes) were *different* depending on whether the Landscape was Rural or Urban? More generally, this is quite a common situation: the effect of one explanatory variable on

the response variable is not simply additive as in model 6.3, but *depends* on another explanatory variable. This is called an **interaction** and we address this situation in Chapter 11.

2) What if we wanted to add more explanatory variables (more main effects - perhaps Phosphate and Flow and Laboratory in addition to Nitrate and Landscape?). This is straightforward and we'll look at more complex models of this type in Chapter 8.

3) What if the response variable was not continuous, as Chlorophyll is, but perhaps count data, or binary data. In other words, how do we model data that might be distributed in ways other than Normal (or Gaussian). We'll come to this in chapters 9 and 10.

4) There may be occasions to include a (usually) continuous explanatory variable but force the slope to be one. Suppose for example that the volume of river water collected in each sample was not exactly the same, but varied between say 3 and 7 millilitres. We'd expect the counts of zooplankton in these samples to be directly proportional to the volume of water collected. Rather than dividing the zooplankton count by the volume and using this standardized 'density per unit volume' as our response variable, we could simply include the volume as an **offset** and model the counts – a preferable strategy as we remain closer to the raw data we actually collected. Offsetting is a useful trick, and described in more detail in Appendix J.

But having discussed the fitted values a bit, we should now think about the scatter around these fitted values (Chapter 7) before addressing these other issues.

Important ideas to take-away

- The right hand side containing our explanatory variables (also known as the linear predictor) can be expressed formally by what we will call the algebraic structure of the model

- Fitted values of our response variable are calculated from the right-hand side of the GLM, by making a series of simple adjustments to a reference value.

- Adjustments to the fitted values are made for each explanatory variable. These adjustments are products of the slope and explanatory variable when the explanatory variable is continuous, and simple additions/subtractions for different levels of a categorical explanatory variable

- We can include as many explanatory variables as we want and the adjustments for each are simply summed together

Chapter 7

What about the scatter?

(back to Contents)

Having considered how to model explainable variation (using the explanatory variables) this chapter focuses on the unexplained variation – the scatter around the fitted values, and how we can use probability density functions to account for this unexplained variation. Unexplained variation arising when modelling count and binary data is considered in Chapter 9 and 10 but the basic ideas are exactly the same.

7.1 Residuals

In Chapter 6 we described some simple General Linear Models. We want to emphasise that they are models. Simplifications of reality. Caricatures if you will, that capture some primary features of the data, but quite possibly failing to explain fine detail. For example, when we try to describe variation in Chlorophyll with Nitrate (Chapter 6, model 6.2), it's clear there is obvious 'vertical scatter' around the line, which is to say that the observed values of Chlorophyll concentration (the black filled circles) deviate from the fitted values – as represented by the diagonal line (Fig. 7.1).

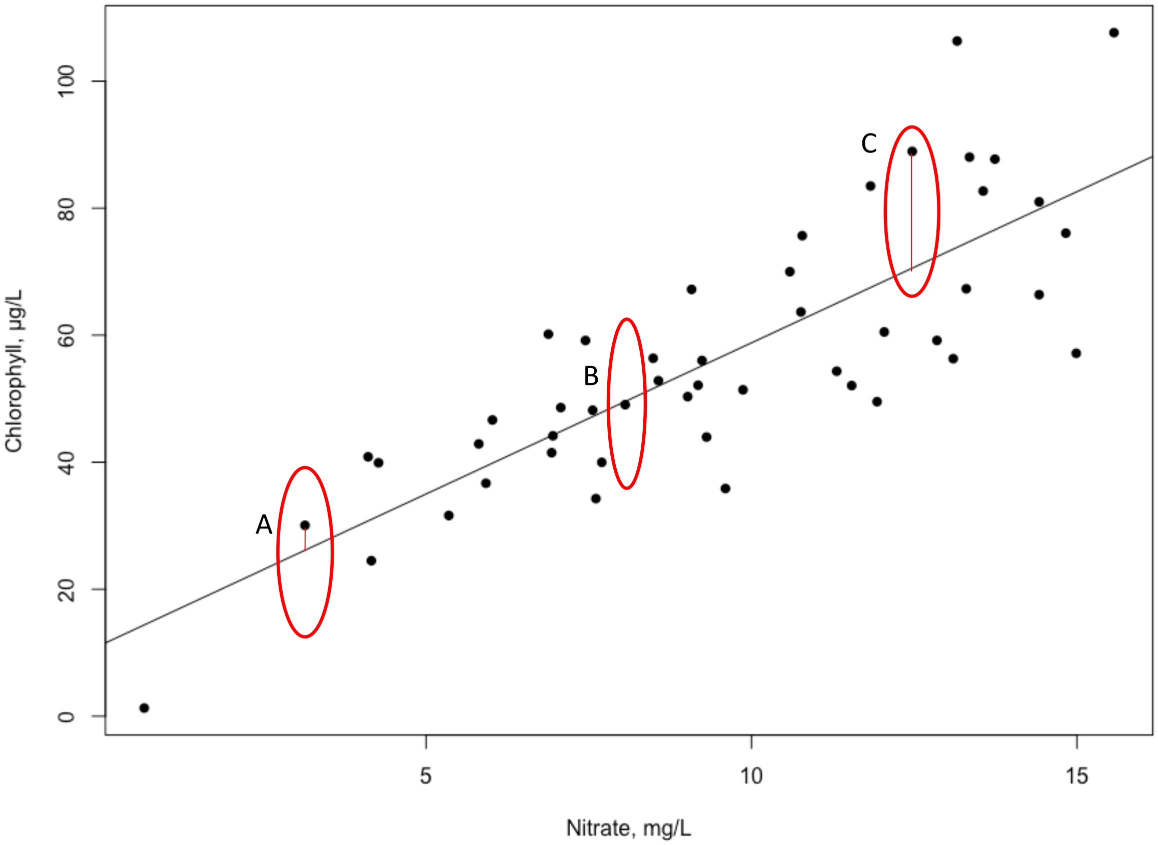


Figure 7.1. Scatterplot of Chlorophyl concentration (y axis) against Nitrate concentration (x axis). Each data point is associated with a residual that measures its deviation from the value predicted by the model (i.e. the fitted line). A: a fairly positive residual; B: a very slightly negative residual; C: a larger positive residual.

We can see from Fig 7.1 that the data point labelled A falls above the best-fitting line, and the fitted value (predicted by the line for that value of Nitrate) underestimates the data point. The point labelled B is fitted almost exactly right, and the point labelled C is underestimated by the model by a fair margin. The difference (indicated by the thin red lines in Fig. 7.1) is referred to as the residual:

the residual = the observed value – the fitted value

We usually refer to the i^{th} *observation* (as opposed to fitted value) of the response variable by y_i (because we usually plot the response variable on the *y* axis), and the fitted value by f_i (as we saw in Chapter 6), and the residual by ε_i, hence:

$$\varepsilon_i = y_i - f_i$$

So, note that the residuals associated with points A and C are positive, and that with point B just a tad negative. Note also that the line fits the data pretty well; most of the points fall close to the line and are associated with quite small either positive or negative residuals. But there are a few cases where the model is out by more ... (point C being an example). The fitted model (the slope and intercept of the line) is selected to minimize the collective magnitude of the residuals across the data set (as discussed in Chapter 13).

Here are all 48 residuals:

```
> model_6.2$residuals
[1]  -17.237 -18.439 -10.689  -5.701  -0.492  11.137  -7.179  -7.983
[9]    0.861   1.181  22.348  18.401 -21.021 -25.382 -13.439 -13.141
[17]  -7.859   6.973 -14.043 -13.168  -2.762  -3.791   8.408  13.313
[25]  -6.766  -5.045   1.259   0.842  13.173  15.996   4.771   8.394
[33]  12.804  10.096  16.223  32.497  -6.502 -11.541  -2.678  -2.675
[41] -13.093  -0.110   1.017   3.922   3.759   6.806  12.521   4.035
```

And here is a frequency histogram of these 48 values (Fig 7.2):

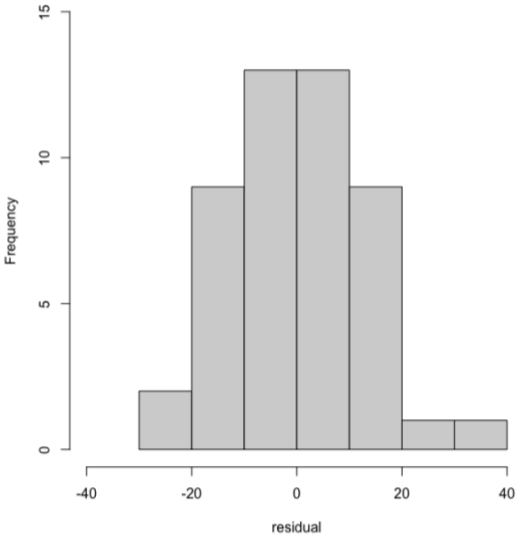


Figure 7.2. The frequency histogram of residuals from the data and model shown in Fig. 7.1.

The residuals collectively appear to conform to something very close to a Normal distribution. This is just as well because this is an assumption of the GLM that we fitted wherein we assumed each observation of the response variable is Normally distributed. Recall the command for model 6.2 from Chapter 6:

```
  glm(formula = Chlorophyll ~ Nitrate, data = my_data, family
  = gaussian)
```

The 48 fitted values are assumed to be the means of Normal (Gaussian) distributions that describe the possible distribution of observations of the response variable around *a particular* mean *tailored* for a specific value of Nitrate (this is what 'family = gaussian' does). Or in other words, each of the *i* (*i* = 1 .. 48) observations of the response variable, y_i, (in this example Chlorophyll) is modelled as coming from a Normal distribution with its own mean f_i (conditioned on Nitrate) and a particular standard deviation (which is the same for every data point) (see Fig. 7.3). Indeed, we could say: $y_i \sim N(f_i, \sigma)$, where *N* denotes a Normal distribution with arguments (mean and standard deviation) f_i and σ, respectively.

For example, point A in Fig. 7.3 (record number 216 in the Four Rivers data set) has a Nitrate concentration of 3.14 mg/L. The slope and intercept (from model 6.2) are 4.76 and 11.19, respectively, so the fitted value is:

$$f_{216} = 11.195 + 4.759 \times 3.14 = 26.138$$

which is a little lower than the observed value for Chlorophyll for record 216, which is 30.16 mg/L. But 26.138 should be as interpreted as the mean of a Normal distribution from which 30.16 'could have come'.

Likewise, point B (record number 21 in the Four Rivers data set) has a Nitrate concentration of 8.06 mg/L. The slope and intercept are 4.759 and 11.195, respectively, so the fitted value is:

$$f_{21} = 11.195 + 4.759 \text{ x } 8.04 = 49.457$$

which is very close to the observed value for Chlorophyll for record 21, which is 49.06 mg/L. But 49.457 should be as interpreted as the mean of a Normal distribution from which 49.06 'could have come'.

Lastly, point C (record number 56 in the Four Rivers data set) has a Nitrate concentration of 12.47 mg/L. The slope and intercept are 4.759 and 11.195, respectively, so the fitted value is:

$$f_{56} = 11.195 + 4.759 \text{ x } \mathit{12.47} = 70.540$$

which is quite a bit less than the observed value for Chlorophyll for record 56, which is 88.94 mg/L. But 70.540 should be as interpreted as the mean of a Normal distribution from which 88.94 'could have come'.

Of course, just how *likely* these observations are to have come from these Normal distributions depends on how close they are to the mean, and what the standard deviation is. In Fig 7.3, we can see that point A lies just above the mean of the distribution that is being used to model it, but is still 'quite likely', whereas point B is right under the mean, and close to being the most likely value given this distribution. Point C is way out in the upper tail, and really quite unlikely given this distribution. However, remember the position of the line – defined by just the intercept and slope is a sort of compromise ... trying to find fitted values that in some sense make all 48 observations of the response variable 'as collectively likely as possible' (or 'maximally *likely'*) given the model we are fitting. If there are many such very poorly fitting points we might become concerned that although the model is the best fitting, it just doesn't fit very well – we'll return to this concern in Chapter 16.

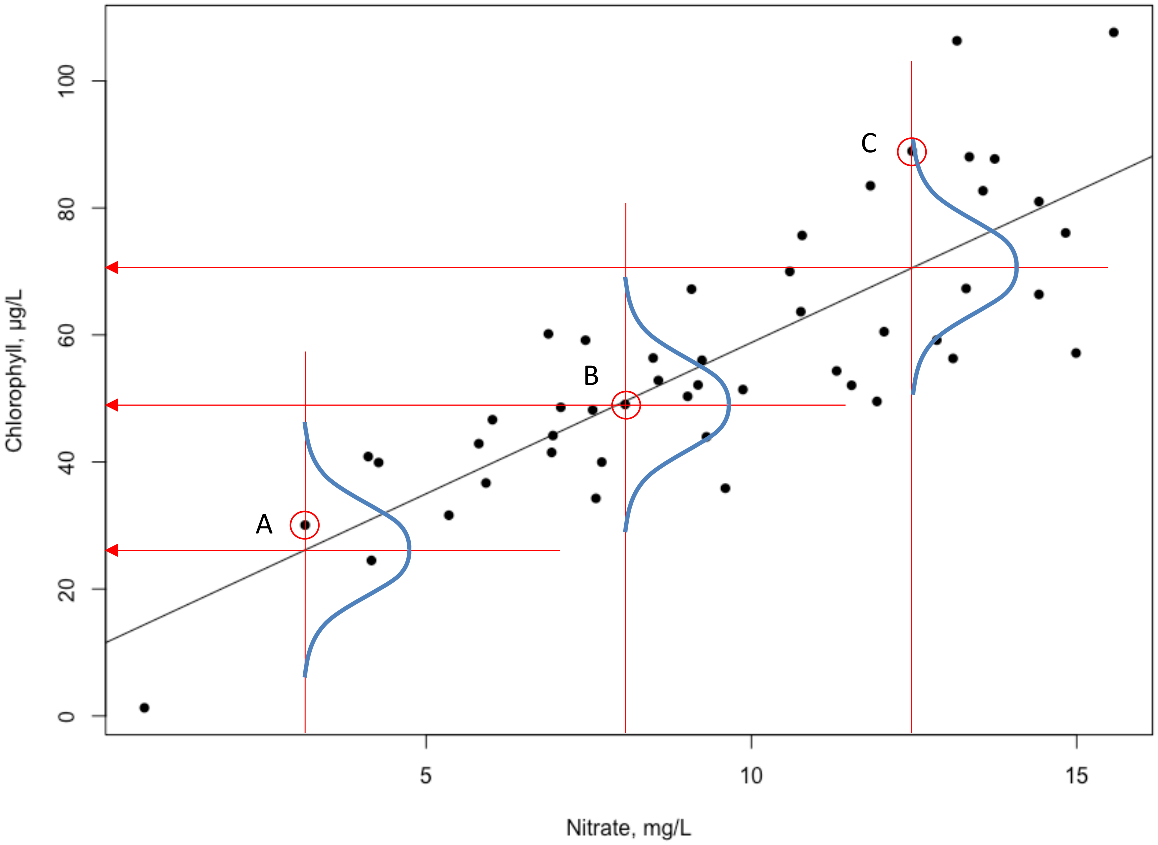


Figure 7.3. The relationship between Chlorophyll and Nitrate. Each data point is being modelled using a Normal distribution centred on the value predicted by the model. Three distributions are shown for the data points circled in red.

The key point – widely misunderstood by so many people, is that it is *not* that all the observations of a response variable when plotted as a frequency histogram look like one Normal distribution, but that *after taking into account all of our explanatory variables (i.e.* having made all the 'adjustments' to whatever reference value we have adopted), the *remaining* variation in the response variable is approximately Normally distributed.

If we 'collapse' the 48 normal distributions we have used to model the 48 data points over to the *y*-axis, we see that *taken overall* the observations of the response variable are *not* normally distributed (see the red line in Fig 7.4). They are distributed according to some superimposition of 48 Normal distributions, each with its own mean determined by an observation specific Nitrate concentration.

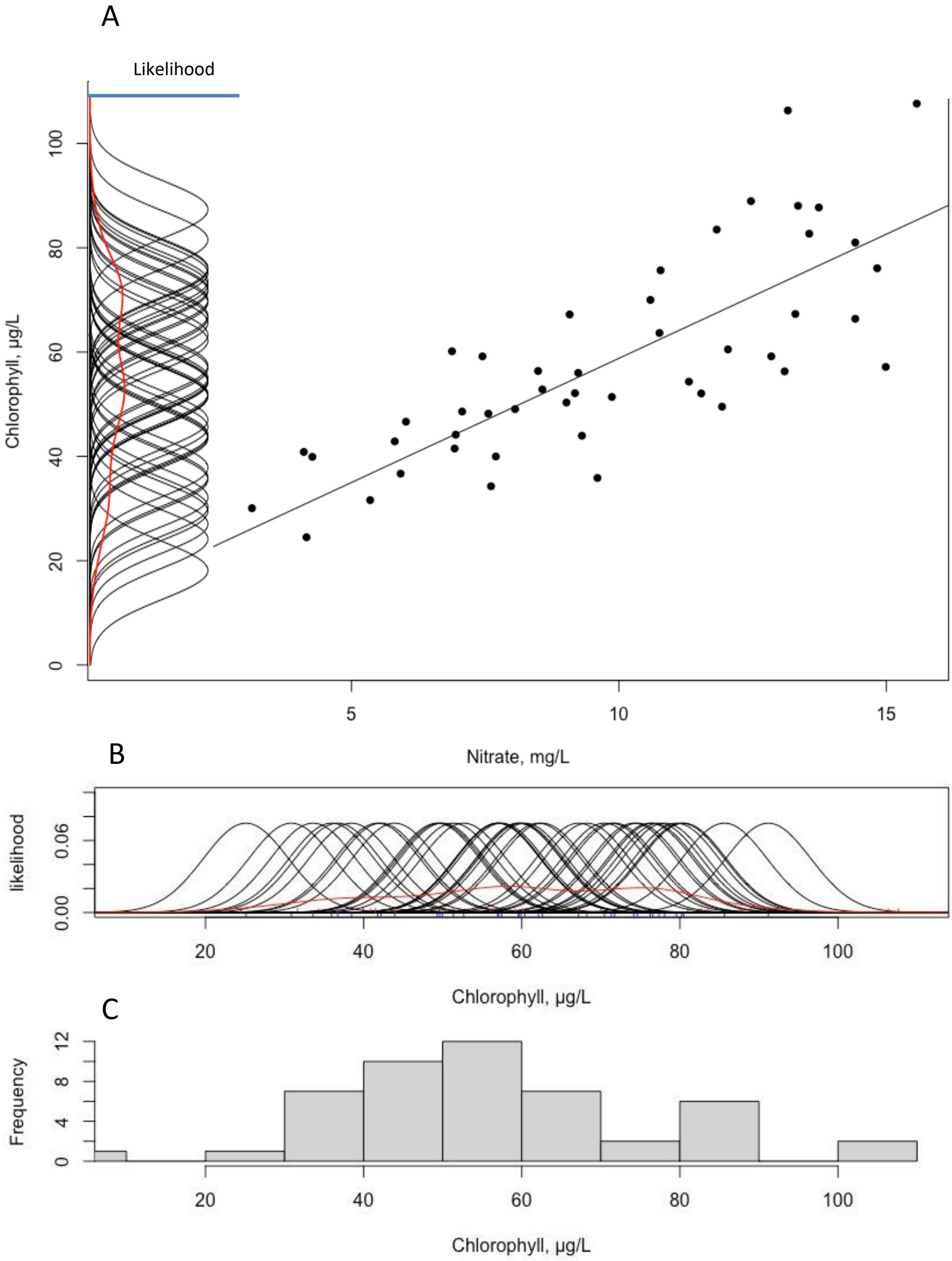


Figure 7.4. A) Collapsing all these Normal distributions onto the y-axis show us the actual distribution of collective observations of the response variable, which is a complex superimposition of Normal distributions (red line) indicating the likelihoods (blue axis); B) the same distribution as on the left-hand side of A, but rotated 90 degrees; C) the actual observed collective frequency distribution of observations of the response variable. *The important point is that the distribution in C is not informative of how we choose the distribution to model variation around each point, it could look like almost anything and we might still choose to model variation around each point as Normally distributed*.

You might ask .. how do we know what the variance of the Normal distribution in these plots is? It is estimated at the same time as the intercept and slope and reported in the output in the summary command (here in green):

```
> summary(model1)

Call:

glm(formula = Chlorophyll ~ Nitrate, family = gaussian, data =
my_data)


Coefficients:

             Estimate Std. Error

(Intercept)   11.1948     5.1191

Nitrate        4.7590     0.5031

---

(Dispersion parameter for gaussian family taken to be 148.8343)
```

The variance is 148.834 (in green), and so the standard deviation is the square root of this number, which is 12.200. It determines the 'width' of the Normal distribution that accounts for the residual variation in these Chlorophyll measurements – it cannot be too 'tight', or too 'loose' ... just right to provide a snug fit to the residual variation. This standard deviation is the σ we have used to define our model: $y_i \sim N(f_i, \sigma)$ and is another coefficient estimated from the data.

So – you might be able to see that once we define a *type* of distribution (in this case a Normal distribution) and we define a model that generates a potentially different distribution (of the same type but with a different mean say) from which every observation 'could have come' – we might say

$y_i \sim$ *some distribution(arguments depending on the explanatory variables)*

we can use these distributions to generate the likelihood of every observation of the response variable, and in principle, the likelihood of *all* the observations of the response variable (see Chapter 4). We 'choose' the coefficients of the model, in this case the intercept, slope and variance (or standard deviation), to maximize the likelihood of all the observations of the response variable (R will do this for us). Indeed ... this is how we arrived at the particular values of the intercept, slope and variance (11.195, 4.759, and 148.834, respectively).

And if we can apply this approach with a simple model with just an intercept, slope and variance, using a Normal distribution, we could use it also for more complex models with more adjustments, and we could choose different distributions (say Poisson or Bernoulli or Negative Binomial) and pretty much fit any data using any distribution. This is exceedingly useful!

Important ideas to take-away

- GLMs model the response variable as variates from probability density functions (pdfs)

- Each observation of the response variable is modelled with a pdf conditioned on the explanatory variables
- This enables the likelihood of each observation of the response variable to be calculated, and the coefficients of the model to be chosen so that the model ensures the data are (collectively) maximally likely
- It is critical to appreciate that the frequency histogram of the collective observations of the response variable will not look like the distribution you have assumed *each single* observation will come from. But will look like a complex superimposition of such distributions

Chapter 8

Constructing models with more explanatory variables

(back to Contents)

This chapter describes how to extend a GLM by adding more variables. Here we focus only on additional main effects. Interactions (including quadratics) are considered in Chapter 11.

In Chapter 6 we constructed a model with two main effects (model 6.3):

Chlorophyll$_i$ = reference value + adjustment for Nitrate$_i$ + adjustment for Landscape$_i$

which asks if we can explain variation in the Chlorophyll in our samples using variation in Nitrate in the samples, and the Landscape that a river flows through.

The R-command instructing R to fit model 6.3 was:

```
> model_6.3<-glm(Chlorophyll~ Landscape + Nitrate,
data=my_data,family=gaussian)
```

The command tells R to use a both a categorical (Landscape), and continuous, (Nitrate) explanatory variable to model the response variable, Chlorophyll, as a Normally distributed observation.

The algebraic structure of the model took the form:

$$f_i = c + \alpha_j + m_N\, x_{N,i} \qquad (i = 1 \,..\, 48, j = R \text{ or } U)$$

The model contained two intercepts, $c + a_R$ and $c + a_U$, and one slope m_N.

8.1 Adding categorical explanatory variables

We can explore whether additional explanatory variables help to explain variation in our response variable simply by adding them in. Suppose we wanted to add the categorical variable Flow. Flow is categorical with 3 levels – L(ow), M(edium), and H(igh).

So we'd have

Chlorophyll$_i$ = reference value + adjustment for Landscape$_i$

+ adjustment for Flow$_k$

+ adjustment for Nitrate$_i$

We note that k might be L, M or H.

The R-command instructing R to fit the model would be:

```
> model_8.1<-glm(Chlorophyll~ Landscape
                                  + Flow
```

```
                            + Nitrate, data=my_data, 
family=gaussian)
```

And the algebraic structure would be:

$$f_i = c + \alpha_j + \beta_k + m_N\, x_{N,i}$$

(i = 1 .. 48, j = R or U, k = L, M, H)

(model 8.1)

α_j generates adjustments for landscape (in this case Urban since the reference level is Rural and its adjustment is zero), β_k generates adjustments for Flow (in this case Low and Medium since the reference level is High and its adjustment is zero), and m_N is the adjustment per unit of Nitrate.

The relevant part of the output would like this:

```
> summary(model_8.1)

Call:

glm(formula = Chlorophyll ~ Landscape + Flow + Nitrate, family = 
gaussian, 

    data = my_data)

Coefficients:

            Estimate Std. Error

(Intercept)   26.877      3.961

LandscapeU   -11.020      2.303

FlowL        -18.131      2.822

FlowM        -10.708      2.826

Nitrate        4.700      0.330

---

(Dispersion parameter for gaussian family taken to be 63.5517)
```

The model actually generates 2 (Landscapes) x 3 (Flows) = 6 different intercepts: $c + \alpha_R + \beta_L$, $c + \alpha_R + \beta_M$, $c + \alpha_R + \beta_H$, $c + \alpha_U + \beta_L$, $c + \alpha_U + \beta_M$, and $c + \alpha_U + \beta_H$, and a single slope m_N (so 6 parallel lines). And Rural and High are the reference levels, so remember that α_R and $\beta_H = 0$, and are not reported in the output above.

So the graph would look like Fig. 8.1:

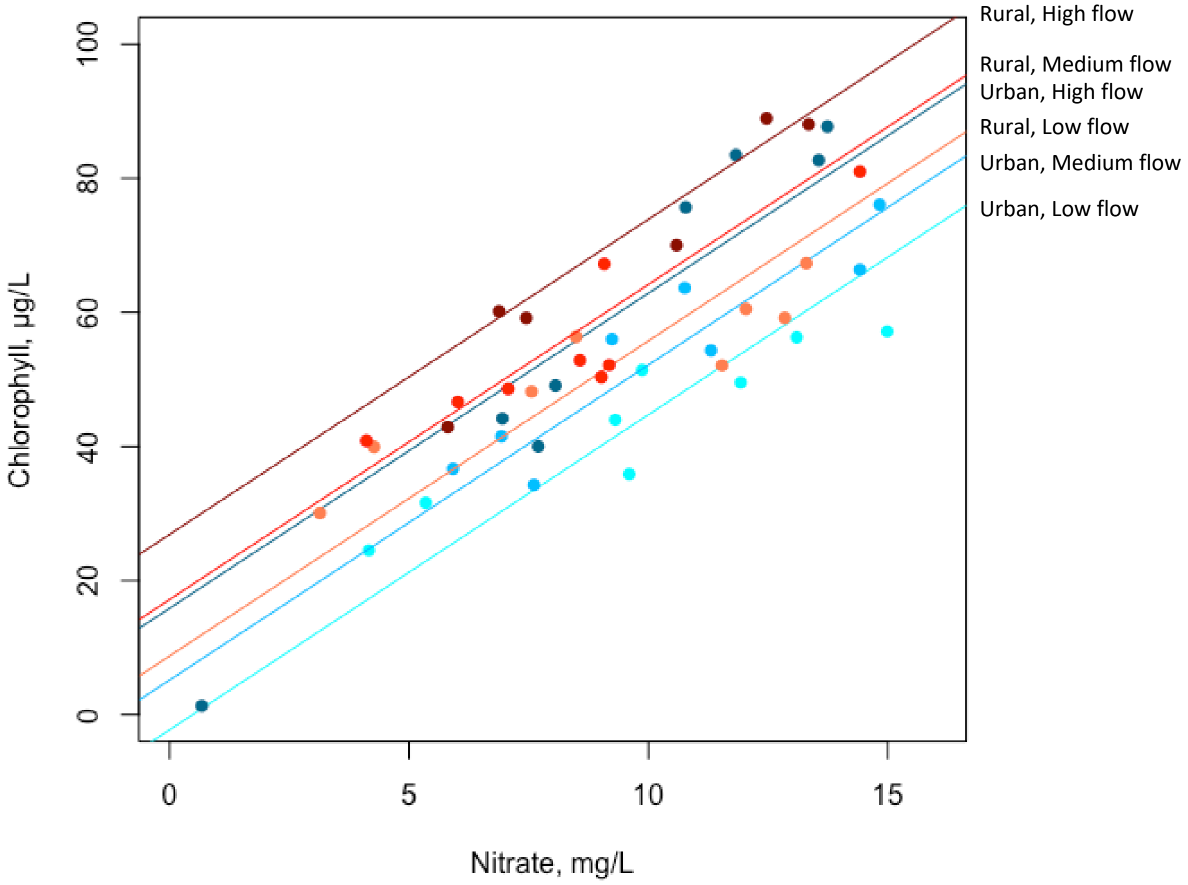


Figure 8.1. Plot of Chlorophyll with Nitrate, with 6 different intercepts for each 2 x 3 combination of levels of Landscape and Flow.

The fitted value for a sample from a Rural landscape with Low Flow and a Nitrate concentration of say, 8.57, would be:

$$f_i = 26.877 + 0 + (-\ 18.141) + 4.700 \text{ x } 8.57 = 49.015$$

and a sample from an Urban landscape with High Flow and a Nitrate concentration of say 13.16 would be:

$$f_i = 26.88 + (-11.02) + 0 + 4.700 \text{ x } 13.16 = 77.712$$

(we've included the zero's in here to indicate the absence of an adjustment for the reference levels).

8.2 Adding continuous explanatory variables

Additional continuous explanatory variables can be added in exactly the same way.

Suppose we wanted to add the continuous variable Phosphate.

We'd have:

Chlorophyll$_i$ = reference value + adjustment for Nitrate$_i$

+ adjustment for Phosphate$_i$

The R-command instructing R to fit model 4 would be:

```
> model_8.2a<-glm(Chlorophyll~ Nitrate
                           + Phosphate, data=my_data,
                           family=gaussian)
```

And the algebraic structure would be:

$$f_i = c + m_N x_{N,i} + m_P x_{P,i}$$

(model 8.2a)

We now have two slopes, m_N capturing the per unit effect of Nitrate on Chlorophyll and m_P capturing the per unit effect of Phosphate on Chlorophyll. The relevant part of the output would like this:

```
Call:
glm(formula = Chlorophyll ~ Nitrate + Phosphate, data = my_data)
Coefficients:
             Estimate Std. Error
(Intercept) -4.98017   11.06891
Nitrate      5.89696    0.77338
Phosphate    0.04845    0.02110
---
(Dispersion parameter for gaussian family taken to be 106.2823)
```

The fitted value for – say - a sample containing 13.30 mg/L of Nitrate and 24.25 mg/L of Phosphate would be:

f_i = -4.980 + 5.897 x *13.30* + 0.048 x *24.25* = 74.614

(continuous explanatory data are indicated in italics)

We can graph this but only using a 3-dimensional plot (Fig. 8.2). We don't advocate the use of 3-dimensional graphs for the formal presentation of data but it may help you to better understand how the data are being modelled. The plane is defined by one point where it intercepts with the y-axis, and two slopes – both of which are positive in this case. The intercept is the expected concentration of Chlorophyll in samples with zero Nitrate and zero Phosphate.

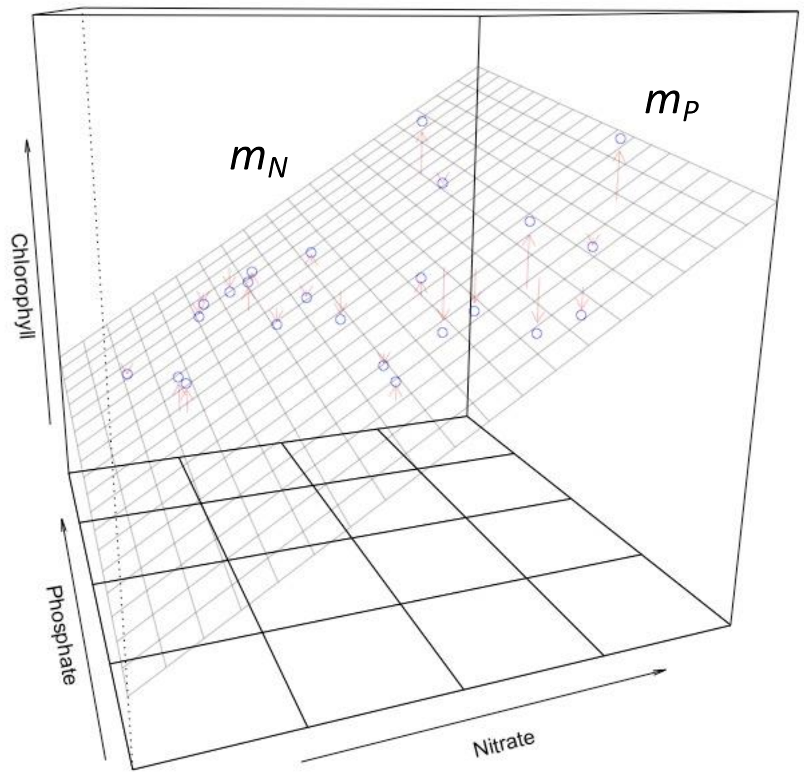


Figure 8.2. 3-dimensional plot of Chlorophyll in relation to variation in Nitrate and Phosphate concentrations. Residuals are indicated by arrows.

3-D plots like Fig. 8.2 are very easy to do using the package `rockchalk`,

```
plotPlane(model, plotx1 = "Nitrate", plotx2 = "Phosphate", drawArrows
= TRUE)
```

We could add temperature as well: We'd have

Chlorophyll$_i$ = reference value + adjustment for Nitrate$_i$

+ adjustment for Phosphate$_i$

+ adjustment for Temperature$_i$

The R-command instructing R to fit the model would be:

```
 > model_8.2b<-glm(Chlorophyll~ Nitrate
                              + Phosphate
                              + Temp, data=my_data, family=gaussian)
```

And the algebraic structure would be:

$$f_i = c + m_N x_{N,i} + m_P x_{P,I} + m_T x_{T,i} \qquad \text{(model 8.2b)}$$

We'd now have three slopes, m_N capturing the per unit effect of Nitrate on Chlorophyll, m_P capturing the per unit effect of Phosphate on Chlorophyll, and m_T capturing the per unit effect of Temperature on Chlorophyll. The relevant part of the output would like this:

```
Call:
glm(formula = Chlorophyll ~ Nitrate + Phosphate + Temp, data =
my_data)

Coefficients:
                Estimate Std. Error
(Intercept)    -44.26449   18.65404
Nitrate          5.93849    0.55124
Phosphate        0.03714    0.02001
Temp             2.80339    1.41257
---
(Dispersion parameter for gaussian family taken to be 115.9424)
```

The fitted value for – say - a sample containing 10.59 mg/L of Nitrate, 150.59 mg/L of Phosphate and a Temperature of 14.08˚C would be:

f_i = -44.264 + 5.938 x *10.59* + 0.037 x *150.59* + 2.803 x *14.08* = 63.657

(continuous explanatory data are indicated in italics)

Note how the coefficients have changed a bit compared to the simpler model with just two covariates. They are still indicating similar positive relationships but they've changed now that the model also takes temperature into account. This is not unexpected. The intercept has changed quite a bit as well – it's what the model predicts the concentration of Chlorophyll to be in samples with zero Nitrate, zero Phosphate, and at zero ˚C. We can't plot this relationship as it would require a 4-dimensional image (we discuss how to present such analyses in Chapter 22).

8.3 Adding both categorical and continuous explanatory variables

We could combine all these categorical and continuous variables into one super complicated model:

Chlorophyll$_i$ = reference value + adjustment for Landscape$_i$
+ adjustment for Flow$_k$
+ adjustment for Lab$_l$
+ adjustment for Nitrate$_i$
+ adjustment for Phosphate$_i$
+ adjustment for Temperature$_i$

The R-command instructing R to fit the model would be:

```
> model_8.3<-glm(Chlorophyll~ Landscape
                            + Flow
                            + Lab
                            + Nitrate
                            + Phosphate
                            + Temp, data=my_data, family=gaussian)
```

And the algebraic structure would be:

$$f_i = c + \alpha_j + \beta_k + \gamma_l + m_N\, x_{N,i} + m_P\, x_{P,i} + m_T\, x_{T,i}$$

(model 8.3)

The relevant part of the output would like this this:

```
Call:
glm(formula = Chlorophyll ~ Landscape + Flow + Lab + Nitrate +
    Phosphate + Temp, data = my_data)

Coefficients:
               Estimate Std. Error
(Intercept)   13.252838   6.471986
LandscapeU   -10.256562   1.003176
FlowL        -16.298490   1.253843
FlowM         -9.104391   1.229415
LabL2         16.438448   1.542810
LabL3          8.011488   1.544282
LabL4         20.248865   1.544390
LabL5         13.736490   1.542318
Nitrate        5.102351   0.180201
Phosphate      0.022003   0.006484
Temp           0.287365   0.460357
```

The fitted value for – say - a sample containing 12.85 mg/L of Nitrate, 68.24 mg/L of Phosphate, a Temperature of 14.00˚C, from an Urban Landscape with Medium Flow analysed in Lab 2 would be:

f_i = 13.253 + (-10.257) + (-9.104) + 16.438 + 5.102 x *12.85* + 0.022 x *68.24* + 0.287 x *14.00* = 81.410

(continuous explanatory data are indicated in italics)

<u>Important ideas to take-away</u>

- It is quite straightforward to add additional categorical and continuous explanatory variables to a GLM
- The various adjustments for each variable are summed to generate the fitted values

Chapter 9

Modelling count data

(back to Contents)

This chapter describes how to analyze count data. Binary data are considered in Chapter 10.

So far, we have assumed that each observation of your response variable can be modelled assuming it derives from a Normal distribution. But many times – this is obviously not an appropriate assumption. Count data can't have decimal places (you can't count 1.6 wildebeest), and they can't be negative (you can't observe -4 wildebeest). As we've discussed in Chapter 4, count data can be modelled using discrete distributions comprising non-negative integers - such as the Poisson or Negative Binomial distributions.

Fortunately, count data are straightforward to model – both in theory and practice! We have previously seen how the fitted values derived from the right-hand-side of a GLM represent the mean of a Normal distribution. We are going to construct the model in *exactly the same way as we've learned so far*, but when we model count data, we want the right-hand-side of the GLM to tell us something about the mean of a Poisson or Negative Binomial distribution, not a Normal distribution.

Unlike when we model data using Normal distributions where the right-hand-side of the GLM *is* the mean of a Normal distribution, when modelling data using Poisson or Negative Binomial distributions the right-hand-side of the GLM is *the natural logarithm of* the mean of a Poisson or Negative Binomial distribution (logarithms are discussed in Appendix A).

Everything else will remain the same.

This is part of an important 'liberation process'. If we can use the right-hand-side of a GLM to model the mean of a Normal distribution, we can use it to model the mean of any distribution, and once we can fit a wider variety of distributions, we can model a wider variety of types of data: count, binary, and other rarer sorts of data. That is - we move from what are called 'General Linear Models' (for Normally distributed data) to 'Generalised Linear Models' (that are modelled using other distributions). The beauty of it all is … we don't have to change the right-hand-side at all. We build these in exactly the same way, regardless of the distribution we choose to use.

9.1 Including a categorical explanatory variable

Consider the response variable ZooCount – a count of zooplankton per sample, and consider (as we did in section 6.6) the subset of the data that were sent to Lab 1. We might ask whether we could explain variation in the zooplankton count with the categorical explanatory variable Flow.

$$\log(\text{ZooCount}_i) = \text{reference value} + \text{adjustment for Flow}_j$$

j would be *L*, *M* or *H*.

The R-command instructing R to fit this Poisson model would be:

```
> model_9.1<-glm(ZooCount ~ Flow, data=my_data, family=poisson)
```

The only difference compared to when we were modelling Chlorophyll using the Gaussian family is that we are now adopting the Poisson family because of the discrete nature of count data – note the `family=poisson` in the above command.

And the algebraic structure would be:

$$\log(f_i) = c + \alpha_j \qquad (i = 1 .. 48, j = L, M, H) \qquad \text{(model 9.1)}$$

and α_j generates adjustments for Flow. Note that the right-hand side is constructed exactly as before, but the *left-hand side* is now different: we model the natural log of the means of Poisson distributions.

The relevant part of the output would be like this:

```
Call:
glm(formula = ZooCount ~ Flow, data = my_data, family = poisson)
Coefficients:
             Estimate Std. Error
(Intercept)   -0.6931     0.5000
FlowL          3.5904     0.5068
FlowM          2.8034     0.5149
```

The (natural) log of the fitted value for a sample from low flow would be -0.69 + 3.60 = 2.91.

The right-hand-side is also known as the **linear predictor**. In Gaussian models linear predictors and fitted values are the same thing, however when modelling data with different distributions this won't be the case. In R we can inspect these linear predictors with the `model$linear.predictors` command:

```
> model_9.1$linear.predictors
 [1]  2.90  2.90  2.11  2.11 -0.69 -0.69  2.90  2.90  2.11  2.11 -0.69
[12] -0.69  2.90  2.90  2.11  2.11 -0.69 -0.69  2.90  2.90  2.11  2.11
[23] -0.69 -0.69
```

From Figure 9.1 we can see that the observed values for Low Flow often seem to be between 15 and 20 so the 2.91 at first looks confusing – but of course it's the *natural log of the mean* of observed count. We need to exponentiate 2.91 to arrive at the actual mean count: exp(2.91) = 18.36.

The values fitted to the data (i.e. the exponentiated linear predictors) are referred to by R as the fitted values and inspected with the `model$fitted.values` command:

```
> model_9.1$fitted.values
 [1] 18.12 18.12  8.25  8.25  0.50  0.50 18.12 18.12  8.25  8.25  0.50
[12]  0.50 18.12 18.12  8.25  8.25  0.50  0.50 18.12 18.12  8.25  8.25
[23]  0.50  0.50
```

At Medium Flows the model predicts exp(-0.693 + 2.803) = 8.248, and at High Flows the model predicts: exp(-0.693) = 0.500. In Fig 9.1 we lay these Poisson distributions with 3 different means alongside the data (the distributions are step like because a Poisson distribution is discrete – and only defined for integer values, but of course *the mean* of a Poisson distribution does not need to be an integer).

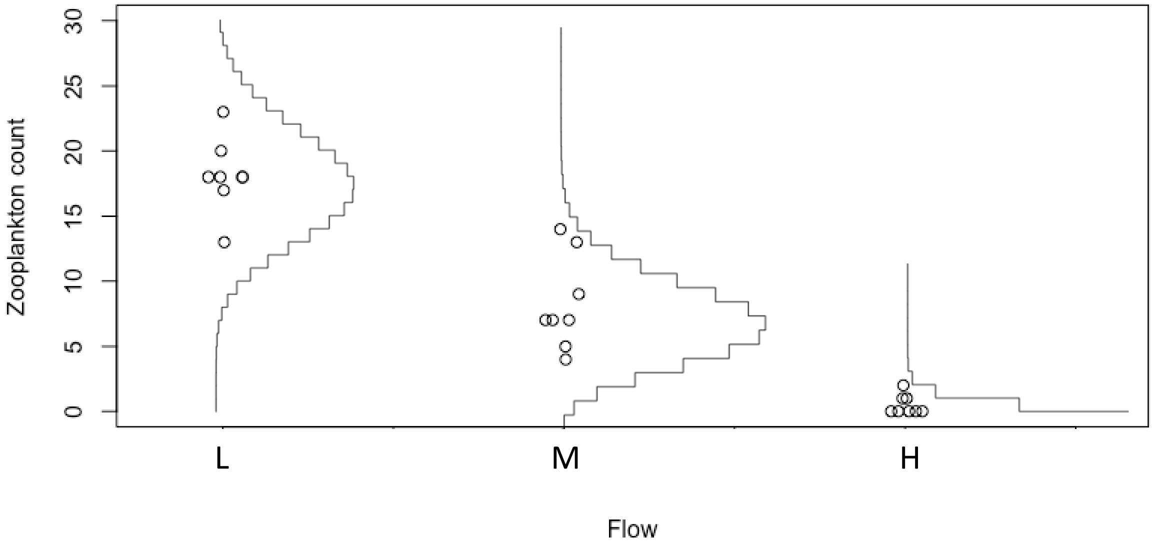


Figure 9.1. Plot of Zooplankton count against Flow rate. Data points are modelled assuming a Poisson distribution whose mean (and variance) are conditioned on the explanatory variables. In this case the explanatory variable is categorical with 3 levels.

In Fig. 9.1 we can see that the data points always fall on a horizontal line indicating the positions of integers on the y-axis. The data and the distributions show that as the means increase, so does the variance (recall that for Poisson distributions the variance is equal to the mean, see section 4.3), and because a Poisson distribution is confined to non-negative integers the distributions can become more symmetric and more ‘normal-looking’ as the mean is increased.

9.2 Including a continuous explanatory variable

BacCount represents the number of bacterial colony forming units per ml of sample, as a measure of disease load. It is also count data. We could, for example, ask if we can explain variation in BacCount with say Phosphate.

$$\log(\text{BacCount}_i) = \text{reference value} + \text{adjustment for Phosphate}_i$$

The R-command instructing R to fit this model would be:

```
> model_9.2<-glm(BacCount ~ Phosphate, data=my_data,
family=poisson)
```

And the algebraic structure would be:

$$\log(f_i) = c + m_P\, x_{P,i} \qquad \text{(model 9.2)}$$

As before, the slope m_p generates per unit adjustments for Phosphate. The relevant part of the output would be like this:

```
Call:
```

```
glm(formula = BacCount ~ Phosphate, family = poisson, data = dx)

Coefficients:
             Estimate Std. Error
(Intercept) 2.8968919  0.0452469
Phosphate   0.0061390  0.0001505
```

On Fig. 9.2 we have plotted Poisson distributions for colony forming unit (CFU) counts for phosphate values of 100, 200 and 300 ug/L given by the following equations:

$$\log(f_{100}) = 2.897 + 0.006 \times \mathit{100} = 3.497$$

$$f_{100} = \exp(3.497) = 33.016 \text{ CFU/ml}$$

$$\log(f_{200}) = 2.897 + 0.006 \times \mathit{200} = 4.097$$

$$f_{200} = \exp(4.097) = 60.159 \text{ CFU/ml}$$

$$\log(f_{300}) = 2.897 + 0.006 \times \mathit{300} = 4.697$$

$$f_{300} = \exp(4.697) = 109.618 \text{ CFU/ml}$$

The Poisson distributions look quite like Normal distributions because the means are quite large, however, note again how the variance of these distributions increases with the mean (which is just as well as there is a lot more vertical scatter on the right side of the plot than the left), and that the relationship between CFUs and Phosphate isn't actually linear. This is because while the natural log of CFU *is* a linear function of phosphate, when we exponentiate to 'unlog' the linear predictor, it takes on a curvilinear form.

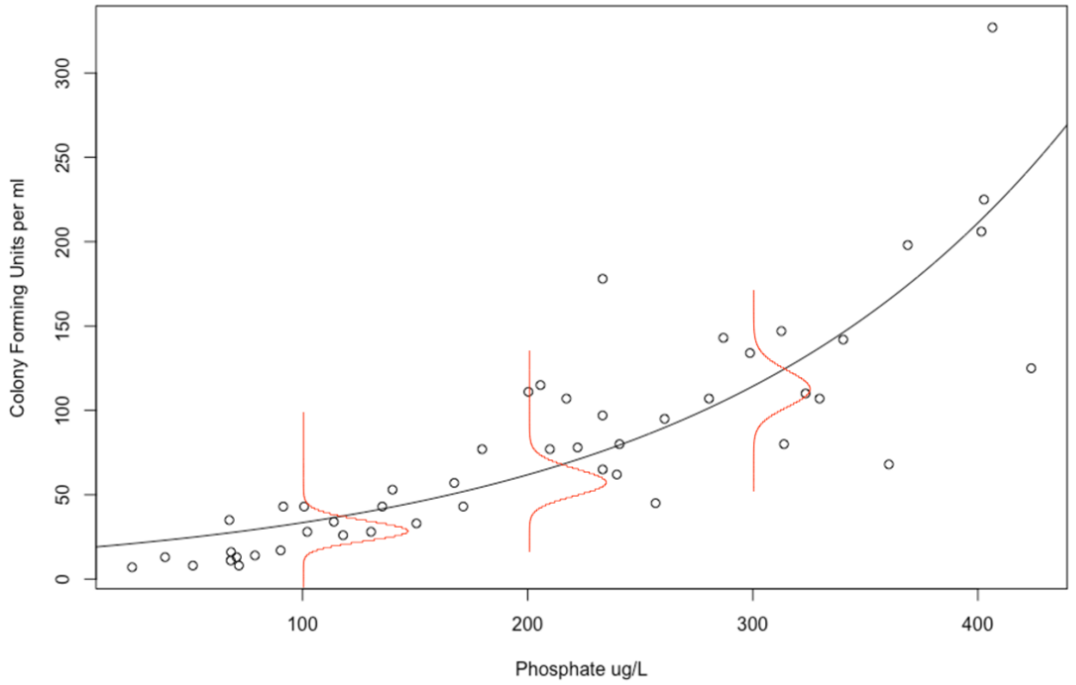


Figure 9.2. Plot of BacCount (colony forming units) against Phosphate concentration, with the exponentiated fitted line indicating the model fit. Data points are modelled assuming a Poisson distribution whose mean (and variance) are conditioned on the explanatory variables (the line). In this case the explanatory variable is continuous. The line is curved because the model is linear for log(BacCount) but not linear after it is back-transformed to BacCount. Note how the variance of the distributions increases with the mean from left to right.

9.3 Including more than one explanatory variable

We can build more complex models just as we describe in Chapter 8, remembering always that we are modelling the natural logarithm of the mean of a Poisson distribution when we do so.

If we model variation in BacCount using both Phosphate and Temperature we'd have

$$\log(\text{BacCount}_i) = \text{reference value} + \text{adjustment for Phosphate}_i$$
$$+ \text{adjustment for Temperature}_i$$

The R-command instructing R to fit this model would be:

```
> model_9.3<-glm(BacCount ~ Phosphate + Temperature, data=my_data,
family=poisson)
```

And the algebraic structure would be:

$$\log(f_i) = c + m_P\, x_{P,i} + m_T\, x_{T,i}$$

(model 9.3)

As before, the slope m_p generates per unit adjustments for Phosphate, the slope m_T generates per unit adjustments for temperature. The relevant part of the output would be like this:

```
glm(formula = BacCount ~ Temp + Phosphate, family = poisson,
     data = my_data)
Coefficients:
                Estimate Std. Error
(Intercept) 2.1077872   0.1829044
Temp        0.0706464   0.0157804
Phosphate   0.0054935   0.0002064
```

We can visualize these relationships in 3D; note how the exponentiation results in a 'warping' of the plane (Fig. 9.3). It is important to emphasize that the relationships between the response and continuous explanatory variables are linear with respect to the natural log of the average of the response variable, but this won't look linear once the log is removed through exponentiation.

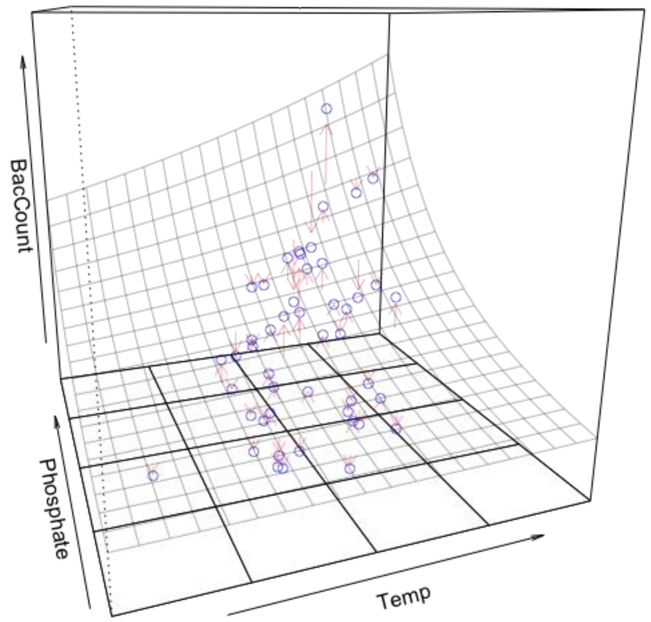


Figure 9.3. 3-D plot of (back-transformed) BacCount in relation to Phosphate and Temperature. Data points are modelled assuming a Poisson distribution whose mean (and variance) are conditioned on two explanatory variables. In this case both explanatory variables are continuous. The plane looks warped (it is!) not because there is an interaction but because the model in linear for log(BacCount) but not linear after its back-transformed. Residuals are indicated by red arrows.

For the bacterial count at 14˚C and 45 ug/L of phosphate the linear predictor would generate:

log(Bacterial Count) = 2.110 + 0.071 x *14* + 0.005 x *45* = 3.329

(continuous explanatory data are indicated in italics)

And so the fitted value would be:

Bacterial Count = exp(3.329) = 27.910 CFUs

You might be wondering *why* we model the log of the mean of the response variable when analyzing count data, when we often don't do so when modelling continuous data. The real reasons are beyond the scope of this text, but it helps to realize that the right-hand-sides of our GLMs can generate a wide range of values, including negative values. Counts cannot be negative – but the logarithm of an average count < 1 can be negative (for example, the natural log of 0.75 is -0.29; and the natural log of 0.025 is -3.69), so by modelling the log of the mean, both the left-hand-side and the right-hand-side of the GLM can in principle range from – infinity to + infinity.

Used in this way the log transformation is called a **link function** (i.e. the function that 'links' the response variable to our explanatory variables). For a Normal distribution the link function is the **identity function** (i.e. we don't need a function at all). The **log-link function** for the Poisson distribution is referred to as the 'canonical' function – we don't *have* to use it, there are other choices, but this is the most natural and common one.

9.4 Deviations from a Poisson distribution

In Chapter 16 we will discuss how to determine if a Poisson-based model is an acceptable fit to the data. The two most common problems encountered in analyzing count data are too much variation for a Poisson distribution – so-called **overdispersion** (Fig. 9.10B), or too many zero observations of the response variable – so-called **zero inflation** (Fig. 9.10C).

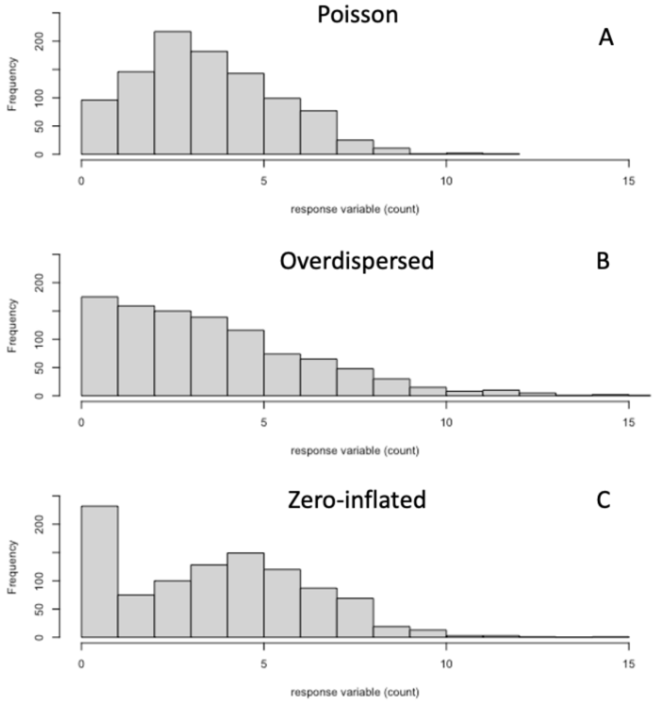


Figure 9.10. Three different forms of count distribution, all with a mean of 4. A) A regular Poisson distribution; B) Overdispersion - note the right hand tail is extended relative to (A); C) Zero-inflation – note the spike at zero, and the slight shift to the right to maintain the mean at 4.

Overdispersion is usually addressed by adopting a Negative Binomial distribution in place of the Poisson distribution (for example, by using the command `glm.nb()` or `glmer.nb` in the `MASS` package). As with the Poisson, you will be modelling the logarithm of the mean of the (Negative Binomial) distribution, and the output will look almost identical to a regular `family = poisson` model, but you will see an additional dispersion parameter estimated to capture the added variance of the Negative Binomial distribution. Zero-inflation is a bit more complicated but can be addressed using a hurdle model (for example the `Zeroinf()` or `Hurdle()` command in the package `pscl` package) as discussed in Appendix K.

Important ideas to take-away

- The principles for modelling count data are identical to modelling data assumed to be Normally distributed, the only difference is that instead of modelling means of Normal distributions we model the natural logarithm of a Poisson (or Negative Binomial) distribution
- Thus, output from the right-hand side of the model (the linear predictor) must be exponentiated in order to be quantitatively 'recognizable' and compared to your observed count data
- The use of the logarithm is an example of a link function that ensures the right-hand side of the model and left-hand side of the model behave in the same way (both can vary in principle at least from -infinity to +infinity)

Chapter 10

Modelling binary data

(back to Contents)

This chapter describes how to analyze binary data. Count data are considered in Chapter 9.

What if observations of your response variable are binary? Yes or no. Positive or negative. One or zero. Pass or fail. As we've discussed in Chapter 4, such binary data can be modelled using a distribution comprising just ones or zeros.

This is also a straightforward thing to do. We have previously seen how the fitted values derived from the right-hand-side of a GLM represent the mean of a Normal distribution, or the log of the mean of a Poisson distribution. We are going to construct the model in *exactly the same way as we've learned so far*, but when we model binary data, we want the right-hand-side of the GLM to tell us something about the mean of a Bernoulli distribution.

The only thing we need to bear in mind is that now when we model data using Bernoulli distributions the right-hand-side of the GLM *is* the logit transformed probability of observing a 'one' (as opposed to a zero). What's a logit transform? Just the probability divided by one minus the probability – logged. So, the GLM takes the form:

$$\log\left(\frac{p_i}{1-p_i}\right) = c + \text{the adjustments for explanatory variables}$$

Where $\log\left(\frac{p_i}{1-p_i}\right)$ is the logit transform of the response variable.

That's all.

This is another example of the 'liberation process'. If we can use the right-hand-side of a GLM to model the mean of a Normal distribution, or the log of the mean of a Poisson distribution, we can use it to model the mean of any other distribution – including a Bernoulli distribution. And as in all our previous examples the beauty of it is ... we don't have to change the right-hand-side at all. We build the right-hand-side in exactly the same way regardless of the distribution we choose to use to represent observations of the response variable.

10.1 Bernoulli or Binomial?

It is often said that binary data are modelled with a Binomial distribution. Why do *we* say Bernoulli? A Binomial variate is (say) the number of heads one would get from tossing a coin *N* times, if the probability of getting a head each time was *p*. *The same p*. We might write this as Binomial(*N*, *p*). It isn't unheard of that data arise as a result of a process like this ... but it is not common.

More often, each observation has a *different* probability of being one of the two possible things (say a one or a zero) – that is – each observation of the response variable has a probability p_i of being a one or a zero depending on its associated explanatory variables. This can be written as Bern(p_i). A Bernoulli distribution is just a special case of a Binomial distribution when the coin is tossed just *once*. The difference is not a big deal (it is just semantics), but we think it's simpler to think of binary data arising from a Bernoulli-like process, rather than a Binomial-like process, because most times each observation of the response variable will have its own probability of being one or the other of the two possible binary outcomes.

Model construction is just the same as it always has been ... except that the usual right-hand-side of the model represents the logit transform of the probability of being one or the other of the two possible binary outcomes.

10.2 The logit link function

Why this logit transform? You can reflect on the fact that while a probability *p* is bounded between 0 and 1, *p*/(1-*p*) can vary from zero to infinity, and remembering that the log of a number that is less than one is negative, log(*p*/(1-*p*)) can vary from minus infinity to plus infinity. So – by logit transforming the left-hand-side of the GLM, it ranges in principle in the same way as the right-hand-side (just as the log of the mean of a Poisson distribution did in the previous chapter).

Of course, we can recover the more recognizable probability from the logit transform with some basic algebra.

If

$$\log\left(\frac{p_i}{1-p_i}\right) = something$$

Then we can exponentiate both sides to remove the log:

$$\frac{p_i}{1-p_i} = \exp\,(something)$$

And then multiplying both sides by $1 - p_i$ and simplifying arrive at:

$$p_i = \frac{\exp\,(something)}{1 + \exp\,(something)}$$

And of course the '*something'* here is just whatever the right-hand-side of the GLM was ... the linear predictor.

10.3 Including a categorical explanatory variable

Consider the response variable Disease in the Four Rivers data set – it's coded as a 1 if any of the zooplankton in the sample are infected with a fungus, and a zero if not. As we did in section 6.6, we'll consider just those samples sent to one particular lab, in this case Lab 2. We can think of fungal infections being present in the i^{th} sample with probability p_i. And we can ask ... whether variation in this probability can be explained by our explanatory variables. For example – does the prevalence of fungal infections depend on, say, Flow? That is:

$\log\left(\frac{p_i}{1-p_i}\right)$ = reference value + adjustment for Flow$_j$

where *j* would be *L*, *M* or *H*.

The R-command instructing R to fit this model would be:

```
> model_10.1<-glm(Disease ~ Flow, data=my_data, family=binomial)
```

(Note the `family=binomial` in the above command).

The algebraic structure would be:

$$\log\left(\frac{p_i}{1-p_i}\right) = c + \alpha_j \qquad (i = 1 \,..\, 48,\ j = L, M, H)$$

(model 10.1)

As before, α_j generates adjustments for Flow.

The relevant part of the output would be like this:

```
Call:
glm(formula = Disease ~ Flow, family = binomial, data = dx)

Coefficients:
            Estimate Std. Error
(Intercept)  -0.2513     0.5040
FlowL         2.1972     0.9085
FlowM         1.3499     0.7664
```

So – we are modelling the *logit transformed* probability of disease presence as potentially depending on Flow. The linear predictor (the right-hand-side) will be: -0.25+0, -0.25+1.35, -0.25+2.20, depending on whether the Flow is High, or Medium or Low. They don't look like probabilities .. because they are logit transformed. To back transform them to probabilities we need to exponentiate and divide by 1 + the exponentiate. For High Flow (the reference) it would be:

$$\frac{\exp(-0.25 + \ 0)}{1 + \exp(-0.25 + \ 0)} = 0.44$$

For Medium flow it would be:

$$\frac{\exp(-0.25 + \ 1.35)}{1 + \exp(-0.25 + \ 1.35)} = 0.75$$

And for Low flow it would be:

$$\frac{\exp(-0.25 + \ 2.2)}{1 + \exp(-0.25 + \ 2.2)} = 0.87$$

(see Fig 10.1.)

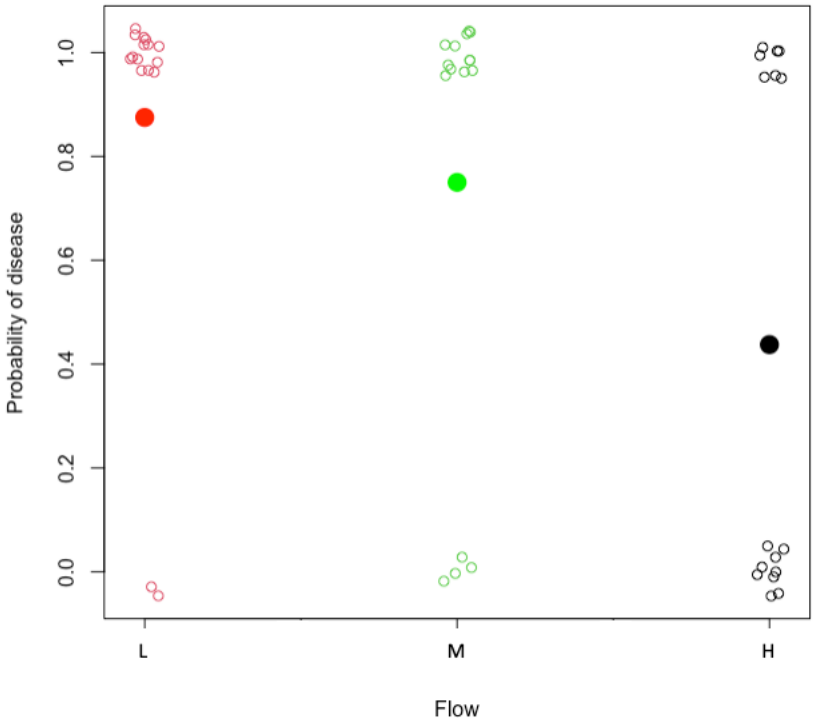


Figure 10.1. The probability of diseased zooplankton in samples from rivers with different flow rates. Some jitter has been added to the x-axis to avoid superimposing data points (unfilled circles). The probability of disease predicted by the model is indicated by filled circles.

10.4 Including a continuous explanatory variable

Alternatively, we can ask … whether variation in the probability of infection can be explained by Temperature. That is:

$$\log\left(\frac{p_i}{1-p_i}\right) = \text{reference value + adjustment for Temperature}_i$$

The R-command instructing R to fit this model would be:

```
  > model_10.2<-glm(Disease ~ Temp, data=my_data, family=binomial)
```

And the algebraic structure would be:

$$\log\left(\frac{p_i}{1-p_i}\right) = c + m_T \, x_{T,i}$$

(model 10.2)

Where m_T is the slope that represents the per unit effect of temperature, $x_{T,i}$.

The relevant part of the output would be like this:

```
Call:
glm(formula = Disease ~ Temp, family = binomial, data = my_data)
Coefficients:
            Estimate Std. Error
(Intercept) -12.1415     3.9110
Temp          0.9138     0.2967
```

We are modelling the logit of the probability of infection as a linear function of temperature (Fig 10.2):

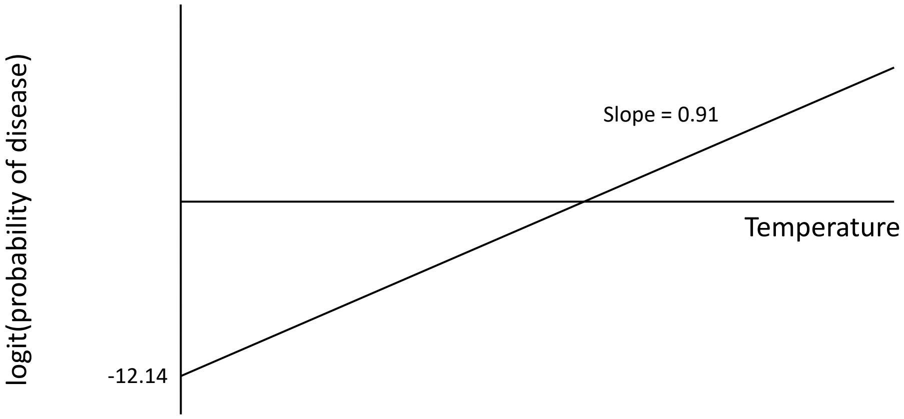


Figure 10.2. Logit transformed probability of disease in relation to temperature. In their simplest forms, GLMs model the logit(probability) as a linear function of a continuous explanatory variable.

Knowing that the actual probabilities are given by the back-transformed linear predictor:

$$p_i = \frac{\exp\,(c + m_T\ x_{T,i})}{1 + \exp\,(c + m_T\ x_{T,i})}$$

We can sketch out the curves that illustrate how the actual probabilities change with temperature and flow (Fig. 10.3).

$$p_i = \frac{\exp\,(-12.141 + \ 0.914\, x_{T,i})}{1 + \exp\,(-12.141 + \ 0.914\, x_{T,i})}$$

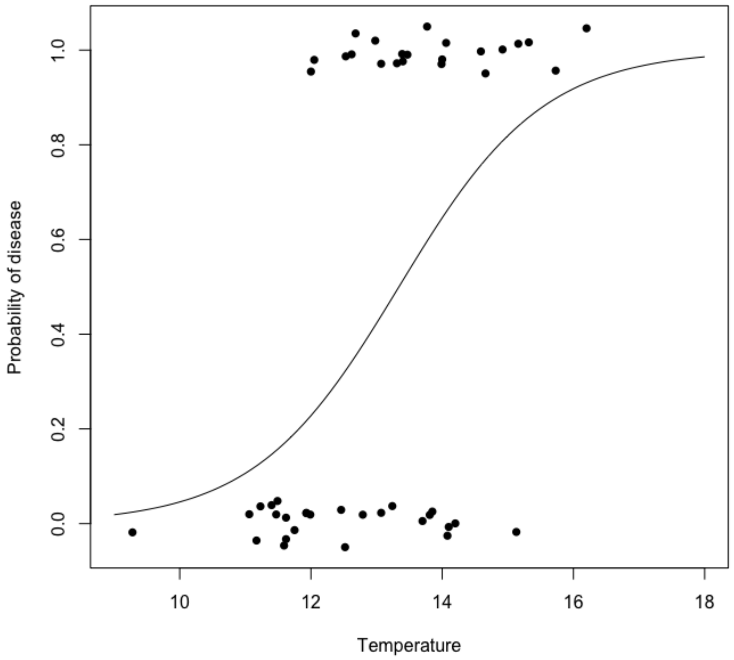


Figure 10.3. When the logit(probability) is back-transformed to a probability we see relationships that are asymptotically bounded between 0 and 1.

You can think of the slope in the usual way as indicative of slope of the 'middle-part' of the curve, and the intercept as where a line extrapolated down would intercept with the y-axis.

10.5 Including more than one explanatory variable

We can ask .. whether variation in the probability of infection can be explained by Temperature and Flow. That is:

$\log\left(\frac{p_i}{1-p_i}\right)$ = reference value + adjustment for Flow$_j$ + adjustment for Temperature$_i$

The R-command instructing R to fit this model would be:

```
> model_10.3<-glm(Disease ~ Flow + Temp, data=my_data,
family=binomial)
```

And the algebraic structure would be:

$$\log\left(\frac{p_i}{1-p_i}\right) = c + a_j + m_T\, x_{T,i} \qquad (i = 1 .. 48, j = L, M, H)$$

(model 10.3)

Where α_j is the adjustment for Flow and m_T is the slope that represents the per unit effect of Temperature, $x_{T,i}$.

The relevant part of the output would be like this:

```
Call:
```

```
glm(formula = Disease ~ Flow + Temp, family = binomial, data =
my_data)
Coefficients:
             Estimate Std. Error
(Intercept) -24.0338     6.9975
FlowL         3.9609     1.3803
FlowM         3.2137     1.2602
Temp          1.6276     0.4846
```

So – we are modelling the *logit transformed* probability of disease presence as a *linear* function of temperature (the logit probability of disease presence increases at 1.63 per degree – whatever that means!), and there are different intercepts to this linear relationship depending on flow, with the intercepts for low and medium flow being more positive than for high flow (Fig. 10.4).

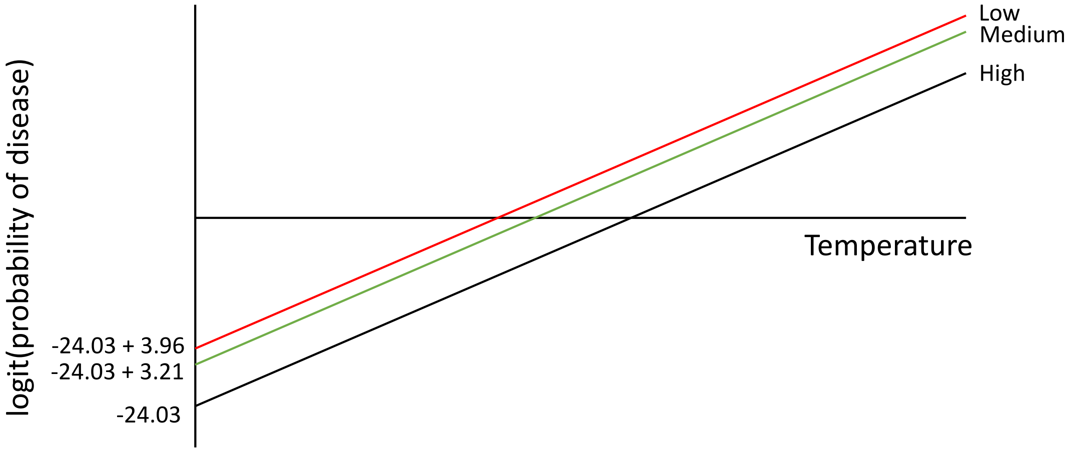


Figure 10.4. The logit model with parallel slopes but multiple intercepts.

We can see the logit values by requesting the linear predictor:

```
> model_10.3$linear.predictors
 [1] -1.37 -4.97 -1.91 -1.96 -5.12 -5.85  0.57  0.47 -2.15 -1.29 -1.49
[12] -2.76 -0.66 -1.79 -2.26  0.97 -6.03 -2.48 -0.56  2.71 -0.54  0.99
[23] -1.12 -1.74  1.85  0.31  0.84  3.85 -1.08 -4.91  4.86 -0.66  0.00
[34]  2.29  0.59 -1.56 -0.46  1.05  2.06  0.45  2.33 -3.64  2.70  2.34
[45]  3.04  3.46 -0.29  1.57
```

These don’t look like probabilities at all ... because they are logit transformed.

We can inspect the ‘back transformed’ probabilities by requesting the fitted values:

```
> model_10.3$fitted.values
[1] 0.20 0.01 0.13 0.12 0.01 0.00 0.64 0.61 0.10 0.22 0.18 0.06 0.34
[14] 0.14 0.09 0.73 0.00 0.08 0.36 0.94 0.37 0.73 0.25 0.15 0.86 0.58
[27] 0.70 0.98 0.25 0.01 0.99 0.34 0.50 0.91 0.64 0.17 0.39 0.74 0.89
[40] 0.61 0.91 0.03 0.94 0.91 0.95 0.97 0.43 0.83
```

These are probabilities that will all fall within the range 0-1.

Knowing that the actual probabilities are given by the back transformed linear predictor:

$$p_i = \frac{\exp\left(c + \ \alpha_j + m_T\ x_{T,i}\right)}{1 + \exp\left(c + \ \alpha_j + m_T\ x_{T,i}\right)}$$

We can sketch out the curves that illustrate how the actual probabilities change with temperature and flow (Fig 10.5).

$$p_{i,H} = \frac{\exp\left(-24.034 + \ 0 + 1.628 \ \ x_{T,i}\right)}{1 + \exp\left(-24.034 + \ 0 + 1.628 \ \ x_{T,i}\right)}$$

$$p_{i,M} = \frac{\exp\left(-24.034 + \ 3.214 + 1.628 \ \ x_{T,i}\right.}{1 + \exp\left(-24.034 + \ 3.214 + 1.628 \ \ x_{T,i}\right)}$$

$$p_{i,L} = \frac{\exp\left(-24.034 + \ 3.961 + 1.628 \ \ x_{T,i}\right.}{1 + \exp\left(-24.034 + \ 3.961 + 1.628 \ \ x_{T,i}\right)}$$

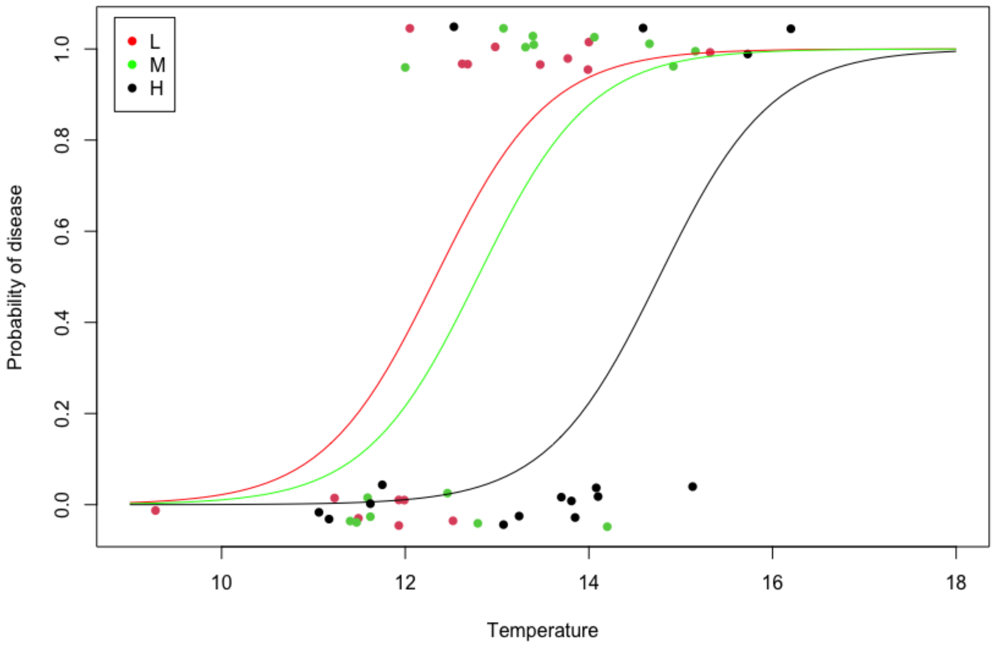


Figure 10.5. The back-transformed logit model with multiple intercepts.

These logit curves are a little hard to relate to the output from the GLM. You can think of the slope as relating to the gradient of the ‘middle-bit’ of the curve, and it is steep or shallow or positive or negative just like a regular slope. And likewise, the intercept as the point where an extrapolated line at that gradient would intercept with the y-axis (see Fig 10.6).

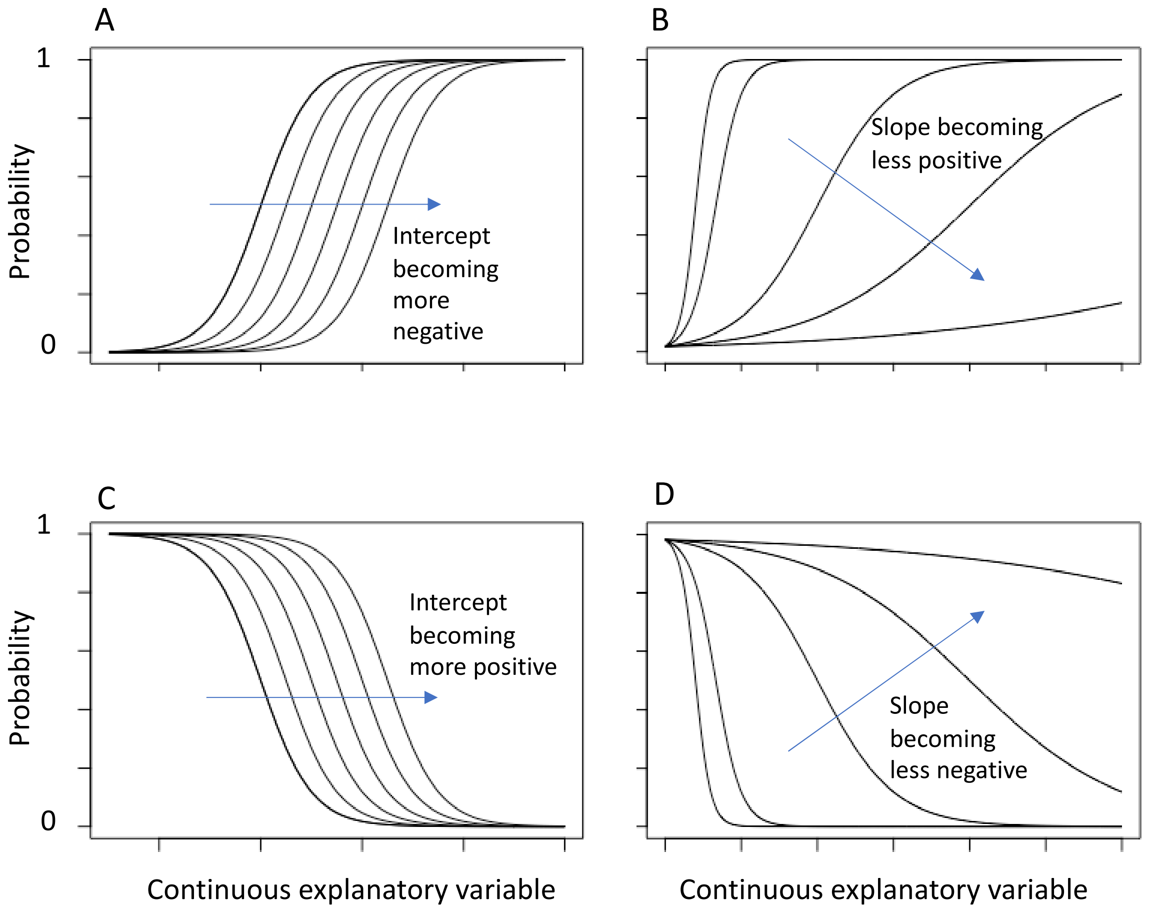


Figure 10.6. The effect of changing the intercept (A, C) and slope (B, D) of a logit function on the probability of a binary outcome.

10.6 Odds ratios

The results of GLMs fitted to binary data are often described in terms of odds ratios. If the probability of a ‘1’ is 0.75 and the probability of a ‘0’ is 0.25, we can say the odds of a ‘1’ is 0.75/0.25 = 3. That is to say, a ‘1’ is 3 times as likely as a ‘0’. We can use the fitted values from GLMs fitted to binary data to estimate the odds of a ‘1’ for any combination of explanatory variables in the model. We can compare two such odds through an odds ratio. The odds ratio indicates by what factor the odds change as the situation moves from that of the denominator to that of the numerator. We could for example, capture the way the odds of zooplankton disease change moving from High (denominator) to Low (numerator) Flows (or any pair of levels of Flow). We could also describe the way the odds change as Temperature is increased from any one value (denominator) to one unit (in this case deg C) higher (numerator).

This a good deal easier to do than it sounds. For reasons explained in Appendix L all it often requires is exponentiating the coefficient governing the effect change you are interested in.

For example, from the output from the model with Temperature and Flow, we can calculate the odds ratio of moving from High Flow (reference) to Medium Flow from exp(1.350) = 3.857 (see the R output in section 10.3 if you don’t remember where this number came from). That is, the *odds* of diseased zooplankton increases almost 4 fold moving from High to Medium Flow. Likewise, the *odds* of diseased zooplankton increases exp(2.197) = 8.900 fold moving from High (reference) to Low flow. (These trends are apparent from Fig. 10.1, but note that *odds are not probabilities* .. (you can see the probabilities don't change 4 or 9 fold going from High to Medium, or High to Low Flows). Odds are ratios of probabilities, and so very confusingly, odds ratios are ratios of ratios).

Likewise, the change in *odds* resulting from a unit increase in Temperature (using the model in section 10.4) would be exp(0.9138) = 2.494. The *odds* of infection more than double for every degree increase in temperature.

Exponentiation of coefficients works (in the absence of interactions) because of the way exponentiates simplify (as described in Appendix L). But the GLM can be used to calculate any two sets of odds which can then be used to calculate an odds ratio. For example, while there is no coefficient that compares Low and Medium Flows (neither are reference), we can calculate odds for each (at say 10˚C).

$$p_{i,M} = \frac{\exp\,(-24.034 + \ 3.214 + 1.628 \times 10)}{1 + \exp\,(-24.034 + \ 3.214 + 1.628 \times 10)}$$

= 0.011

So the odds of disease at Medium Flow at 10˚C = 0.011/0.989 = 0.011 (note the denominator here is 1-0.011), and

$$p_{i,L} = \frac{\exp\,(-24.034 + \ 3.961 + 1.628 \times 10)}{1 + \exp\,(-24.034 + \ 3.961 + 1.628 \times 10)}$$

= 0.022

So the odds of disease at Low Flow at 10˚C = 0.022/0.978 = 0.022.

Thus, the odds change by a factor of two moving from Medium (denominator) to Low (numerator) Flows at 10 deg C (or indeed at any fixed Temperature): 0.022/0.011 = 2.00. More details are supplied in Appendix L.

The are many packages that will calculate odd ratios for you, but it is good to understand how to do it 'manually' before using and interpreting the output from these packages.

10.7 The `cbind` trick

We have deliberately introduced the modelling of binary data in a way that anticipates every observation of the response variable might have a have different probability of being a '1', and the data would be formatted in the usual 'flat' way – one observation of the response variable and its accompanying explanatory variables (i.e. one record) per row of the data table. However, if all the explanatory variables are categorical there is an alternative format that is more concise, wherein we use the numbers of 1's and 0's for each combination of levels – and we describe this in more detail in Appendix M.

Important ideas to take-away

- The principles for modelling binary data are identical to modelling data assumed to be Normally distributed, the only difference is that instead of modelling means of Normal distributions we model the logit transformed probability

- Thus, output from the right-hand side of the model (the linear predictor) must be back-transformed in order to be quantitatively ‘recognizable’ and compared to your original binary data
- The use of the logit function is an example of a link function that ensures the right-hand side of the model and left-hand side of the model behave the same way (both can vary in principle at least from -infinity to +infinity)
- Results from analyses of binary data are often described using odds ratios

Chapter 11

Interactions

(back to Contents)

This chapter introduces interactions. An interaction between two explanatory variables exists when the effect of one of the explanatory variables on the response variable depends on the other. Interactions can arise between two continuous explanatory variables, between two categorical explanatory variables, or a categorical and a continuous explanatory variable. Although the fundamental interpretation of an interaction does not change, these three combinations necessarily look a bit different.

So far, we have built models that can represent any number of explanatory variables, also known as **main effects**, be they continuous or categorical. Each variable generates some sort of adjustment to a reference value. However, the adjustments for each explanatory variable are added entirely separately from each other. The adjustment – say – for Nitrate is completely unrelated to the adjustment for Phosphate; or the adjustment for Flow doesn't depend in any way on the adjustment for Temperature, or Landscape.

What if it was more complicated? We suspect that Nitrate has a positive influence on Chlorophyll concentration – as measured by the slope of a graph with Nitrate on the x-axis (it's the explanatory variable) and Chlorophyll on the y-axis (Chlorophyll being the response variable) (Fig. 11.1A).

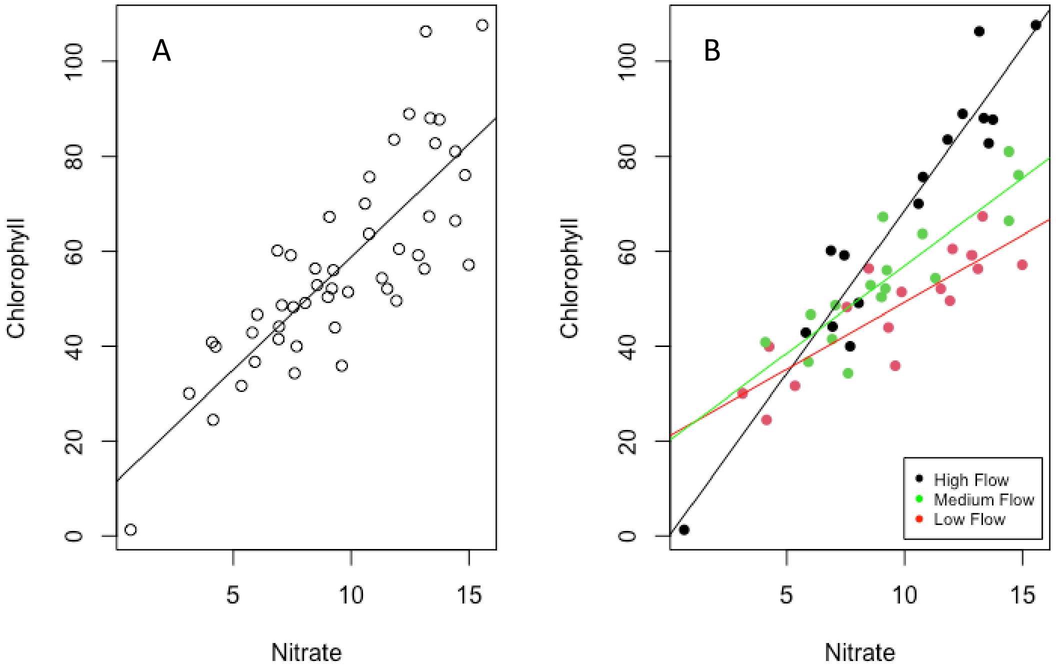


Figure 11.1. A) The simple effect of Nitrate on Chlorophyll; compared to B) when the effect of nitrate on chlorophyll *depends* on Flow (note how the slopes are different for the different levels of Flow). Here we are still only looking at the data from Lab 1.

But what if the effect of Nitrate was different at different levels of Flow? Perhaps the effect of Nitrate is greater at High Flow than at lower Flows? That is to say – what if the effect of Nitrate *depended* on the Flow? What if the slope reflecting the effect of Nitrate *depended* on the Flow level? (Fig. 11.1B).

This is an example of when *the effect of one explanatory variable on the response variable depends on another explanatory variable*. And this dependency is called an **interaction**. In principle, interactions can occur between any number of explanatory variables but we will limit our attention to interactions between just pairs of explanatory variables – so called 'two-way interactions'. Three-way interactions are complicated to interpret and best avoided unless completely necessary!

Interactions may exist between a continuous and categorical explanatory variable, between two continuous explanatory variables, and between two categorical variables. Regardless – they always represent a situation in which the effect of one explanatory variable on the response variable depends on another explanatory variable. But they look a bit different in terms of the algebraic structure of the model – as we will see.

However, it doesn't make any difference what distribution you assume observations of your response variable may come from (Normal, Poisson, Bernoulli, Negative Binomial) – interactions are always modelled the same way. As ever, the construction of the 'right-hand-side' of a GLM always follows the same principles.

11.1 Interactions between continuous and categorical explanatory variables

We've seen in Chapter 8 how we can include a continuous and categorical explanatory variable in the same model. We might have

$$f_i = c + \alpha_j + m_N x_{N,i} \qquad j = L, M, H;\ i = 1 \,..\, 48$$

So, now we just want to make an adjustment to the effect of Nitrate depending on the level of Flow – that is, an adjustment (β_j) to the *slope* m_N, so it is different for different levels of the variable Flow:

$$f_i = c + \alpha_j + (m_N + \beta_j)x_{N,i}$$

(model 11.1)

This model can generate 3 different intercepts: $c + 0$, $c + \alpha_L$, $c + \alpha_M$; and 3 different slopes: $m_N + 0$, $m_N + \beta_L$, $m_N + \beta_M$. A key point to remember is that the subscript on the adjustment to the intercept and the slope are paired. We don't want to combine the intercept for one level with the slope for another – so if *j* is say *L* for the intercept, it should be *L* for the slope also. This is why both α and β are subscripted by *j*.

Interactions are straightforward to code in R. We'd write:

```
> model_11.1<-glm(Chlorophyll ~ Flow + Nitrate + Flow:Nitrate,
data = my_data)
```

Where `Flow:Nitrate` represents the interaction between Flow and Nitrate.

And again, subsetting on samples we'd sent to Lab 1 (as we did in section 6.6), we'd get output that looked like this:

```
Call:
glm(formula = Chlorophyll ~ Flow + Nitrate + Flow:Nitrate,
    data = my_data)
Coefficients:
              Estimate Std. Error
(Intercept)    -0.2170     5.1255
FlowL          21.1690     7.2848
FlowM          20.1892     7.7292
Nitrate         6.8783     0.4832
FlowL:Nitrate  -4.0488     0.7037
FlowM:Nitrate  -3.1814     0.7638
(Dispersion parameter for gaussian family taken to be 53.33138)
```

Which would map on to Fig. 11.1B as shown in Fig. 11.2:

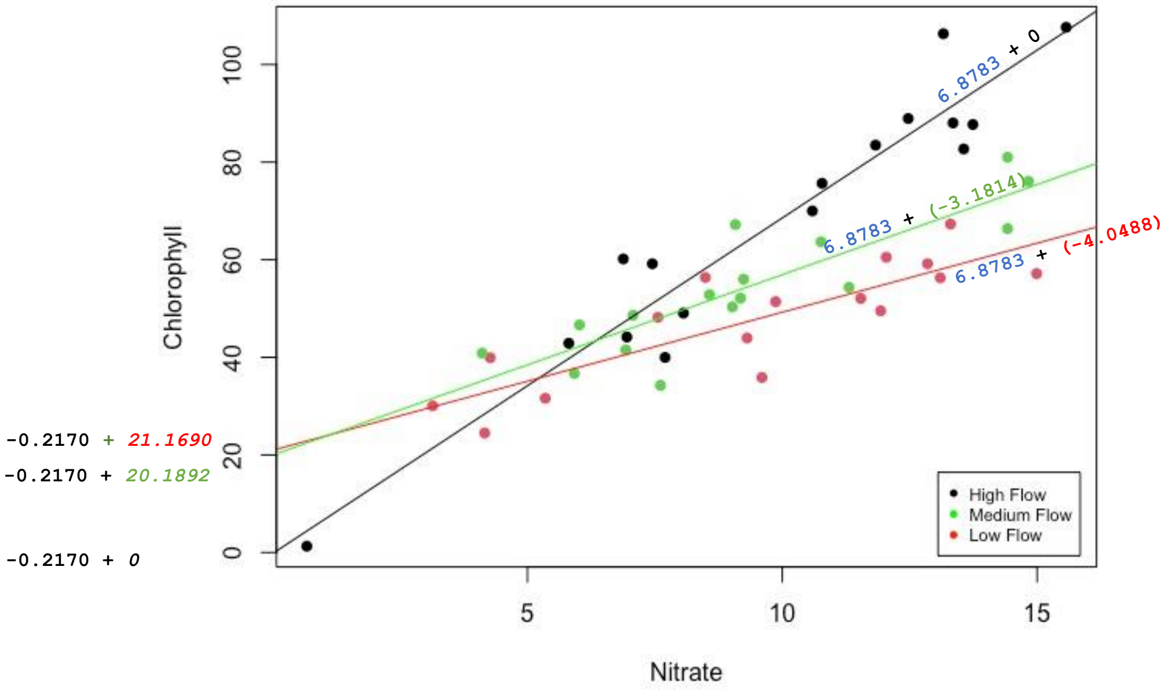


Figure 11.2. Annotated graph showing adjusted slopes and intercepts for the interaction of the continuous variable Nitrate and the categorical variable Flow.

We cannot state exactly what the effect of Nitrate is on Chlorophyll – it *depends* on the Flow level. When you have to answer questions like ‘what is the effect of Nitrate on Chlorophyll?’ by saying ... ‘well it depends on the Flow’ … you know you have an interaction on your hands.

Note how we don’t need additional adjustments for the reference level for either the intercept or slope as these can be represented by c and m_N. To get 3 different intercepts and 3 different slopes we only need 6 coefficients: c, m_N, α_L, α_M, β_L, and β_M. Any more would be redundant.

The number of additional coefficients required to model an interaction between a continuous and categorical explanatory variable with q levels will be (q-1). In this example, q = 3 and so 3-1 = 2 (β_L and β_M).

If we want to know whether the effect of Nitrate on Chlorophyll depends on Flow, we’ll be interested in how confident we are that the adjustments to the slopes are different to zero.

And, by-the-way, it doesn’t make any difference whether we talk about the effect of Nitrate on Chlorophyll depending on Flow; or the effect of Flow on Chlorophyll depending on Nitrate. Same thing.

We could (very often) have more than one interaction in the model, for example,

```
> model_11.1.1<-glm(Chlorophyll ~ Flow + Landscape + Nitrate + Phosphate
            + Flow:Nitrate + Landscape:Phosphate, data = my_data)
```

would have the algebraic structure

$$f_i = c + \alpha_j + \beta_k + (m_N + \gamma_j)x_{N,i} + (m_P + \delta_k)x_{N,i}$$

(model 11.1.1)

And of course one might have two interactions with the same continuous explanatory variable, for example:

```
> model_11.1.2<-glm(Chlorophyll ~ Flow + Nitrate
            + Flow:Nitrate + Landscape:Nitrate, data = my_data)
```

Would have the algebraic structure

$$f_i = c + \alpha_j + \beta_k + (m_N + \gamma_j + \delta_k)x_{N,i}$$

(model 11.1.2)

Note that this has two adjustments to the *same slope*, m_N.

11.2 Interactions between two continuous explanatory variables

The concept is exactly the same: the effect of one explanatory variable on the response variable depends on another explanatory variable – the slope governing how one continuous explanatory variable affects the response variable depends on the value of another continuous explanatory variable.

We've seen in Chapter 8 how we can include two continuous explanatory variables in the same model. We might be interested in the continuous explanatory variables Temperature and Nitrate:

$$f_i = c + m_T x_{T,i} + m_N x_{N,i} \qquad i = 1 .. 48$$

If we wanted to ask whether the effect of Nitrate on Chlorophyll depended on Temperature (or conversely and synonymously, effect of Temperature on Chlorophyll depended on Nitrate), we'd just include one additional term comprised of an additional parameter and the product of the values of the respective continuous variables (in this case Temperature and Nitrate): $m_{T:N}x_{T,i}\,x_{N,i}$, thus:

$$f_i = c + m_T x_{T,i} + m_N x_{N,i} + m_{T:N}x_{T,i}\,x_{N,i}$$

The reason this only requires one parameter is explained in Appendix N.

Again, this is easy to implement in R.

```
model_11.2<-glm(Chlorophyll ~ Temp + Nitrate + Temp:Nitrate, data
= my_data)
```

And the output would look like:

```
Call:

glm(formula = Chlorophyll ~ Temp + Nitrate + Temp:Nitrate, data =
my_data)

Coefficients:

                Estimate Std. Error
```

```
 (Intercept)    41.8308     38.0731
Temp           -2.6616      2.7643
Nitrate        -4.5339      3.7169
Temp:Nitrate    0.7578      0.2792
 (Dispersion parameter for gaussian family taken to be 107.081)
```

In the absence of Nitrate, Chlorophyll concentration decreases by m_T = -2.662 mg/L per degree C. But at high Nitrate concentrations (say $x_{N,i}$ = 15 mg/L) the effect is $m_T + m_{T:N}x_{N,i}$ = -2.662 + 0.758 x 15 = 8.708 mg/L per degree C. In fact – at higher Nitrate concentrations the effect of Temperature is completely reversed. We cannot simply state whether Temperature has a positive or negative effect on Chlorophyll – it *depends* on the Nitrate concentration. There is an interaction.

We can see this in the 'plane plot'. Fig 11.3A is the model without the interaction, and the plane is entirely flat, the slope (for Nitrate) is the same whether one looks at the right or the left, and the slope for Temperature the same whether one looks at the top or the bottom. However, in the presence of the interaction (Fig 11.3B), we can see that the slope for Temperature (at low Nitrate) is negative, and at high Nitrate it is positive. The plane is 'warped'. The effect of Temperature on Chlorophyll depends on Nitrate, and conversely, the effect of Nitrate on Chlorophyll depends on Temperature.

The number of additional coefficients required to model a continuous-continuous interaction is always one.

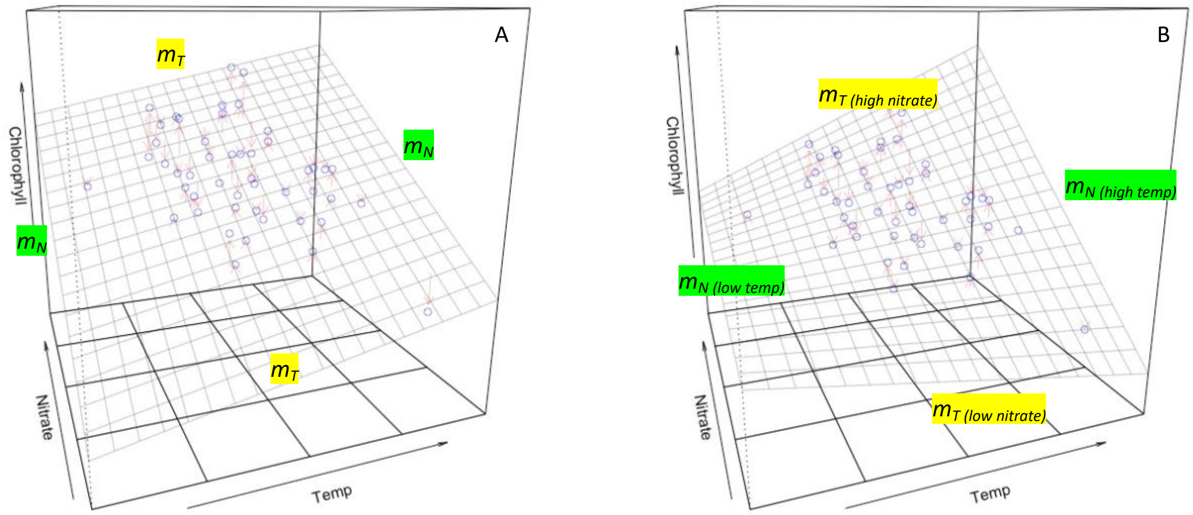


Figure 11.3. A) The influence of the continuous variables Nitrate and Temperature on Chlorophyll in the absence of an interaction; and B) with an interaction.

If we want to know whether the effect of Nitrate on Chlorophyll depends on temperature, we'll be interested in how confident we are that the adjustments to the slopes are different to zero.

11.3 Quadratic terms

A continuous explanatory variable can interact with itself! What can this mean? Noting that the overall adjustment to be made by say Nitrate, would usually be denoted by the product $m_N x_{N,i}$, and the per unit adjustment for Nitrate is the slope m_N, we could make the *per unit* adjustment for Nitrate *depend on* .. Nitrate. The *slope* would become $(m_N + m_{N2} x_{N,i})$ where the additional term m_{N2} determines the influence of Nitrate on the effect of Nitrate. The overall adjustment for Nitrate becomes: $(m_N + m_{N2} x_{N,i}) x_{N,i}$ or $m_N x_{N,i} + m_{N2} x_{N,i}{}^2$. Evidently a quadratic term will appear on the right-hand side of the GLM. If the coefficient m_{N2} was positive, this would mean that the effect of a unit change in Nitrate on (say Chlorophyll) would depend on Nitrate, really just like any other interaction (the greater Nitrate concentration, the greater the influence of a further unit increase in Nitrate). Quadratic terms work just like any other interaction, the main effect should also be included in the model. The algebraic structure would look like:

$$f_i = c + \alpha_j + m_N x_{N,i} + m_{N2} x_{N,i}^2$$

(Here, we've include a categorical explanatory variable – say Flow, represented by α_j. And again, this is easy to implement in R:

```
model_11.q<-glm(Chlorophyll ~ Flow + Nitrate + Nitrate:Nitrate,
data = my_data)
```

And the quadratic parameter will appear listed with the other coefficients in the output. If $m_N > 0$ and $m_{N2} < 0$ (or $m_N < 0$ and $m_{N2} > 0$) then the result can be non-monotonic (humped or U shaped) relationships between the response and continuous explanatory variable. The model generates distinctly non-linear looking relationships, but is still considered 'linear' in the sense that the quadratic adjustment is still part of a (linear) sum of adjustments.

11.4 Interactions between two categorical explanatory variables

The concept is exactly the same: the effect of one explanatory variable on the response variable depends on another explanatory variable – the adjustment we want to make that governs how one level of an explanatory variable affects the response variable – depends on the level of another categorical explanatory variable.

Remembering how we model two categorical variables (say, Flow with 3 levels and Landscape with 2 levels):

$$f_i = c + \alpha_j + \beta_k \qquad j = L, M, H;\ k = R, U;\ i = 1 .. 48$$

And that Landscape(Rural) and Flow(High) would be the reference levels (coming earlier in the alphabet than the other levels for these two categorical explanatory variables). We'd have adjustments for Low, Medium, and High Flows, and then an entirely separate adjustment for Rural or Urban Landscapes.

$$f_{HR} = c + 0 + 0$$

$$f_{MR} = c + \alpha_M + 0$$

$$f_{LR} = c + \alpha_L + 0$$

$$f_{HU} = c + 0 + \beta_U$$

$$f_{MU} = c + \alpha_M + \beta_U$$
$$f_{LU} = c + \alpha_L + \beta_U$$

Note how we can generate 6 different fitted values – one for each combination of levels, but we only use 4 coefficients (c, α_L, α_M, β_U). This is efficient, but it comes with a constraint: we have to make the *same* adjustment at say, Medium Flow, in both Rural and Urban Landscapes, and likewise, the same adjustment for Urban Landscapes regardless of the Flow. But what if the adjustment we wanted to make for Flow depended on whether it was a Rural or Urban Landscape? Or the adjustment we wanted to make for Urban Landscapes depended on Flow? We'd need to be able to generate 6 *different* adjustments for each 3 x 2 combinations of Flow and Landscape, and we'd need 6 *different* coefficients, not 4.

There are various ways of thinking about this but they all amount to the same thing. We could have something like this:

$$f_i = c + \gamma_{jk} \qquad j = L, M, H;\ k = R, U;\ i = 1 \,..\, 48$$

(model 11.3a)

Here we would have a reference combination and 5 non-zero adjustments captured by 5 different estimates of γ (say $\gamma_{LU}, \gamma_{MU}, \gamma_{HU}, \gamma_{LR}$ and γ_{MR}) and c could deal with the High Flow-Rural Landscape combination. The important point is we need 6 coefficients for 6 different fitted values.

Alternatively, we could just add two more coefficients – say γ_{MU} and γ_{LU} to what we had above and write:

$$f_i = c + \alpha_j + \beta_k + \gamma_{jk} \qquad j = L, M, H;\ k = R, U;\ i = 1 \,..\, 48$$

(model 11.3b)

$$f_{HR} = c + 0 + 0$$
$$f_{MR} = c + \alpha_M + 0$$
$$f_{LR} = c + \alpha_L$$
$$f_{HU} = c + 0 + \beta_U$$
$$f_{MU} = c + \alpha_M + \beta_U + \gamma_{MU}$$
$$f_{LU} = c + \alpha_L + \beta_U + \gamma_{LU}$$

And we have 4+2=6 coefficients to generate the 6 different fitted values. Obviously, we don't need 6 different γ's and 2 α's and a β and a c (10 coefficients) to generate just 6 different fitted values, so R just fits additional coefficients for each combination of levels that doesn't contain a reference level (in this case: Low Flow and Urban Landscape, and Medium Flow and Urban Landscape) (Fig 11.4).

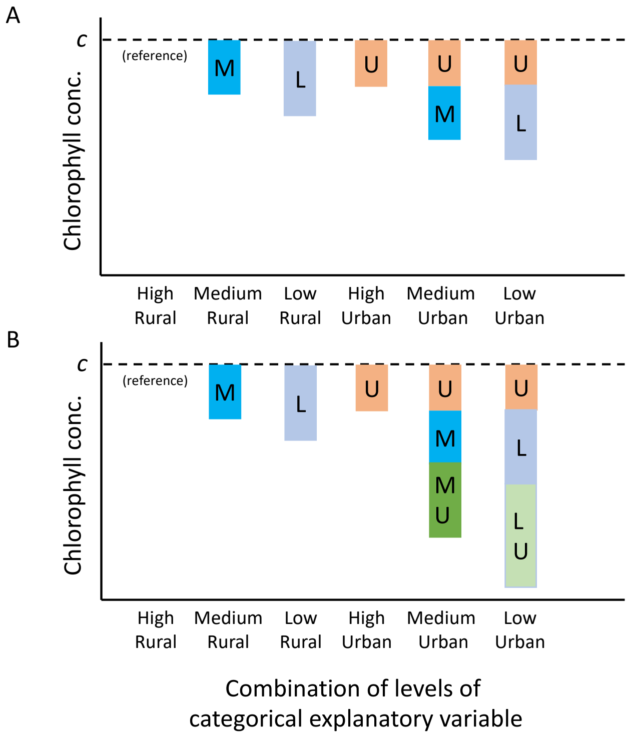


Figure 11.4. Schematic (not drawn to actual scale) conceptualizing categorical/categorical interactions. A) No interaction. Note how there are only 4 quantities to derive 6 estimates for the 6 combinations of the levels of Landscape and Flow. If any of one of these 4 quantities is changed, multiple estimates will be affected. B) Medium-Urban and Low-Urban are no longer constrained to be simple sums of the estimates of Medium and Urban effects; or Low and Urban effects, but have their own unique combinatorial adjustment (MU and LU respectively). With 6 separate quantities, the Chlorophyll concentrations can be estimated independently for each of the 6 combinations of levels.

It is a bit confusing but model 11.3a and b are the same thing ... R will figure it out, and label the adjustments so you know where to look.

We write:

```
model_3<-glm(Chlorophyll ~ Flow + Landscape + Flow:Landscape, data
= my_data)
```

And the output would look like:

```
Call:
glm(formula = Chlorophyll ~ Flow + Landscape + Flow:Landscape,
    data = my_data)
Coefficients:
                    Estimate Std. Error
(Intercept)           77.894      6.664
FlowL                -26.194      9.424
FlowM                -22.944      9.424
LandscapeU           -19.885      9.424
FlowL:LandscapeU      11.973     13.327
FlowM:LandscapeU      18.549     13.327
```

```
(Dispersion parameter for gaussian family taken to be 355.2302)
```

And the fitted values would be:

$$
\begin{aligned}
f_{HR} &= 77.894\ +\ 0 +\ 0\\
f_{MR} &= 77.894 + (-22.944) +\ 0\\
f_{LR} &= 77.894 + (-26.194)\\
f_{HU} &= 77.894 +\ 0 + (-19.885)\\
f_{MU} &= 77.894 + (-22.944) + (-19.885) + 18.549\\
f_{LU} &= 77.894 + (-26.194) + (-19.885) +\ 11.973
\end{aligned}
$$

Fitted this way, the number of additional coefficients required to model an interaction between a categorical explanatory variable with *q* levels and another with *r* levels will be (*q*-1) x (*r*-1). In this example, *q* = 3 and *r* = 2 and (3-1) x (2-1) = 2.

If we want to know whether the effect of Flow on Chlorophyll depends on Landscape, we'll be interested in how confident we are that these two additional adjustments are different to zero.

There are various different ways of instructing R to fit interactions. For simplicity, we suggest the format here:

```
main_effect_1 + main_effect_2 + main_effect_1:main_effect_2
```

But you can achieve essentially the same thing with

```
main_effect_1:main_effect_2
```

or

```
main_effect_1*main_effect_2.
```

11.5 Effect sizes in the presence of interactions

In the presence of an interaction, the effect size (6.10) associated with one explanatory variable will depend on another explanatory variable. There isn't anything you can do about this, the situation is just a bit more complicated. Calculate the fitted values for the combination of explanatory variables you want to compare, and then the differences between them are the effect sizes. It is probably best to change just one explanatory variable at a time, but of course by the definition of an interaction, you will need to recognize that whatever effect size you identify is conditional on what is assumed about the other explanatory variable in the interaction.

There are various R commands that are useful for calculating summaries of fitted values that account for interactions. For example `emmeans(model ~ term1|term2)` in the package `emmeans`).

Important ideas to take-away

- An interaction between two explanatory variables is always interpreted the same way: the effect of one explanatory variable on the response variable depends on another explanatory variable
- However, because explanatory variables may be continuous or categorical, and the adjustments take on slightly different algebraic forms, there are actually 3 different forms of algebraic adjustment, depending on whether the interaction is continuous/continuous, continuous/categorical, or categorical/categorical

Chapter 12

Random effects

(back to Contents)

This chapter is a very brief introduction to what random effects are, and how they are represented in GLMs. The statistical application and interpretation of random effects is left to Part 2.

Categorical explanatory variables may be of two types. These are called **fixed** and **random**. So far, we have only talked about fixed effects: all our categorical explanatory variables have been 'fixed'.

We're getting a bit ahead of ourselves with random effects but we think it's best to introduce them here as they are an important part of model construction. A model that contains both fixed and random effects is called a **mixed model**.

Recall that in the Four Rivers data set there were … 4 rivers. Each river was sampled at 12 different points. It might be that perhaps one of these rivers runs over chalk, another over clay, a third might be more acidic. For any number of reasons the rivers may be slightly different to each other – have their own idiosyncratic 'river-characteristics', and of course all 12 samples from each river will share this same 'river-characteristic'. That is to say – we have *repeatedly sampled* from the *same* river. We have **repeated measures** from the same river. This introduces an element of relatedness between each of the 12 samples taken from the same river, and left unaccounted for this is a problem as it breaks the assumption that the residual variation associated with each of the data points is independent of each other. So, we need to deal with it.

As it happens, we're not interested in the river the samples came from. We *were* interested in the relationships between the physical-chemistry of the water (Temperature, Phosphate, Nitrate, Flow rate etc) and some of its biological properties (Chlorophyll concentration, Bacterial count, Zooplankton abundance and Disease prevalence), but the name of the river didn't matter to us. However, because there is a possibility that 'river-ness' might have a consistent effect on our samples, we would include 'River' in our model.

There are 4 rivers, so it could be included as a categorical explanatory variable in the usual way.

For example:

```
> model_12.1<-glm(Chlorophyll~Nitrate+Flow+River,data=my_data)

> summary(model_12.1)

Call:
```

```
glm(formula = Chlorophyll ~ Nitrate + Flow + River, data =
my_data)
Coefficients:
            Estimate Std. Error
(Intercept)  11.1428     6.5950
Nitrate       5.4225     0.4672
FlowL       -17.8114     2.7868
FlowM       -10.2532     2.7947
RiverR2      -5.1843     3.2613
RiverR3      12.3535     3.6659
RiverR4       5.1035     4.4274
(Dispersion parameter for gaussian family taken to be 61.78849)
```

However, including river has resulted in the need to estimate 3 more coefficients, enabling an adjustment to be made for each river to account for the particular 'river-ness characteristic' that might lead to unwanted relationships between the samples (note that river 1 is the reference river). If we'd sampled from 100 rivers instead of 4, we'd need to estimate 99 more coefficients! Which is a very complicated model to account for something we're not very interested in.

Under these circumstances we might choose to call River a random effect. The model will be fitted with adjustments for each river, but the adjustments will all come from a single particular distribution (we will assume a Normal distribution but we could specify a different one). This Normal distribution will have a mean of zero, and a variance. The variance will be sufficient that the various adjustments needed for each river could have derived from this distribution. The more different the rivers are to each other, the greater the variance will need to be. Because we are not actually interested in the differences between the rivers (this was not part of our research question) we don't get to see what the individual adjustments actually are, but we do get to see the variance of the distribution from which they came.

To fit random effects we need a different R package – we'll use a package called `lme4` (but there are several to choose from) and a command called `lmer` which fits mixed models assuming observations of the response variable are modelled as coming from a Normal distribution.

The R command would be:

```
> model_mixed_12.2<-lmer(Chlorophyll ~ Nitrate
                                  + Flow
                                  + (1|River),data=my_data)
```

Note the bit in red bold – this instructs the model to include river as a random effect – an adjustment to the intercept for each river.

The algebraic structure might be written:

$$f_i = c + m_N x_{n,i} + \alpha_j + R_k$$

(i = 1.. 48, j = L, M, H, and k = R_1, R_2, R_3, R_4)

(model 12.2)

where the random effect for river is indicated here by an upper case letter.

The output might look like this:

```
Linear mixed model fit by REML.
Formula: Chlorophyll ~ Nitrate + Flow + (1 | River)
   Data: my_data
REML criterion at convergence: 327.4
Random effects:
 Groups    Name        Variance Std.Dev.
 River     (Intercept) 49.07    7.005
 Residual              61.79    7.860
Number of obs: 48, groups:  River, 4
Fixed effects:
            Estimate Std. Error
(Intercept)  15.5964     5.9662
Nitrate       5.2827     0.4452
FlowL       -17.8732     2.7860
FlowM       -10.3413     2.7932
```

The output looks familiar ... an intercept, a slope for Nitrate, two adjustments for Flow, but we now have a section for the random effect (in bold) and specifically a variance of 49.07 which is the variance of a Normal distribution from which the 4 adjustments for each river derive. Had there been 100 rivers, the output would look exactly the same (except perhaps the variance of the river effect would be larger!). In a formal sense, although we are accounting for 4 different rivers, we are actually only really estimating the one variance term.

You might wonder why the mean of this Normal distribution with variance of 49.07 is zero. Any non-zero tendency would be common to all rivers so the mean of the Normal distribution modelling 'effect of river' can be incorporated into the intercept (here 15.596).

We can include random effects in models in which observations of the response variable are not Normally distributed (perhaps because the response variable is count or binary data) using the `glmer` command (i.e. generalised linear mixed effect model) in the same package, `lme4`; we just need to say – as usual – which distribution we want to use (say Poisson or Binomial).

Random effects can be described using **Intra-class correlation coefficients (ICCs)**. The ICC captures the relative within group correlation of the random effect, or the 'repeatability' of observations made on the same level. The ICC for River in the analysis above would be given by the variance for the random effect of River, divided by all the sources of variation, so:

$$ICC_{River} = \frac{49.07}{49.07 + 61.79} = 0.44$$

We might want to include the interaction between River (a random effect) and Nitrate (a fixed effect or indeed any other fixed explanatory variable). In this case, the interaction would be fitting different slopes for the effect of Nitrate on Chlorophyll ... to each River – so we have 4 different slopes. These four 'adjustments' to the slope would require 3 additional coefficients, but since river is a random effect we can apply the same trick as for the random intercepts, and model the random slopes as coming from a Normal distribution with zero mean and a certain variance depending on how big the interaction effect is.

```
> model_mixed<-lmer(Chlorophyll~Nitrate+Flow+(1+Nitrate|River),
data=River_data)
> summary(model_mixed)
Linear mixed model fit by REML.
Formula: Chlorophyll ~ Nitrate + Flow + (1 + Nitrate | River)
   Data: River_data
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 River    (Intercept) 153.6376 12.3951
          Nitrate       0.8573  0.9259  -0.81
 Residual             109.6505 10.4714
Number of obs: 240, groups:  River, 4
Fixed effects:
            Estimate Std. Error
(Intercept)  25.9114     6.9140
Nitrate       5.4833     0.5359
FlowL       -17.7643     1.6993
FlowM        -9.9701     1.7189
```

The random slopes modelling the influence of Nitrate on Chlorophyll have a mean of 5.483 and a variance of 0.857.

In fact, with just 4 Rivers we only save estimating 2 coefficients with the mixed model compared to the 3 adjustments required for river in a purely fixed-effect model. Indeed, it is commonly suggested that random effects should really have at least 5 different levels in order to be robustly estimated.

Random effects are very useful because so long as the adjustments for each of the levels can be modelled as coming from a single distribution it doesn't matter how many levels there are – there could be hundreds ... or thousands of different levels, we'd still only estimate one variance. The price of this simplification is we don't get to see what the individual adjustments actually are, although there are situations in which estimating the variance of the random effects is of direct biological interest (often, for example in studies of genetic variation).

It is important to recognize that whether a categorical explanatory variable can be treated as a random effect or not depends on what we want to know from our data. It is not an inherent feature of a variable – it depends on your motive. It might very well have been that we *did* want to know which river was being adjusted in what way – but if we really want to study the differences between the levels of our random effect – it probably isn't a random effect, and we'd best call it a fixed effect.

Important ideas to take-away

- Random effects are a certain type of categorical explanatory variable
- A categorical explanatory variable may be treated as a random effect if we are not interested in examining the potentially different effects of the different levels
- Thus, whether an explanatory variable is a fixed or random effect doesn't depend on the data themselves, but really the investigator's motive for including the explanatory variable in the model
- Random effects are very useful for modelling data which can be viewed as repeated observations of some subject or object, that themselves are not of particular interest, but plausibly introduce dependencies between observations which we need to account for
- Random effects require only one coefficient to represent them – the variance of a distribution from which adjustments from each level are made
- An explanatory variable should have at least 4 and ideally more levels to treat it as a random effect

Chapter 13

Fitting the models

(back to Contents)

This chapter is a brief introduction into how models are fitted using maximum likelihood, and some of the reasons this can go wrong. A longer discussion of how to work with the log-likelihoods of models can be found in Part 2.

We're not going to talk in much detail about how the models are fitted. Once you have instructed R which model to fit, the fitting is done quicky and efficiently by R for you. However, it is useful to know the basic principles.

13.1 Least squares

You may recall hearing in some previous statistics course something called the method of **least squares**. Take a look at Fig 13.1. In a simple example like this one we'd be looking to fit a line that minimizes the sum of the squared length of the red lines – that is – the squared residuals. Square the residuals so they all become positive, and just add them all up ... and find the equation for the line (in this case the intercept and slope) that minimizes this sum of squared deviations. Clearly, if the line is chosen to minimize the lengths of the red lines squared, then the line will pass as closely it as can through them. For example, the line in Fig. 13.1A clearly performs better by this metric than the line in Fig. 13.1B.

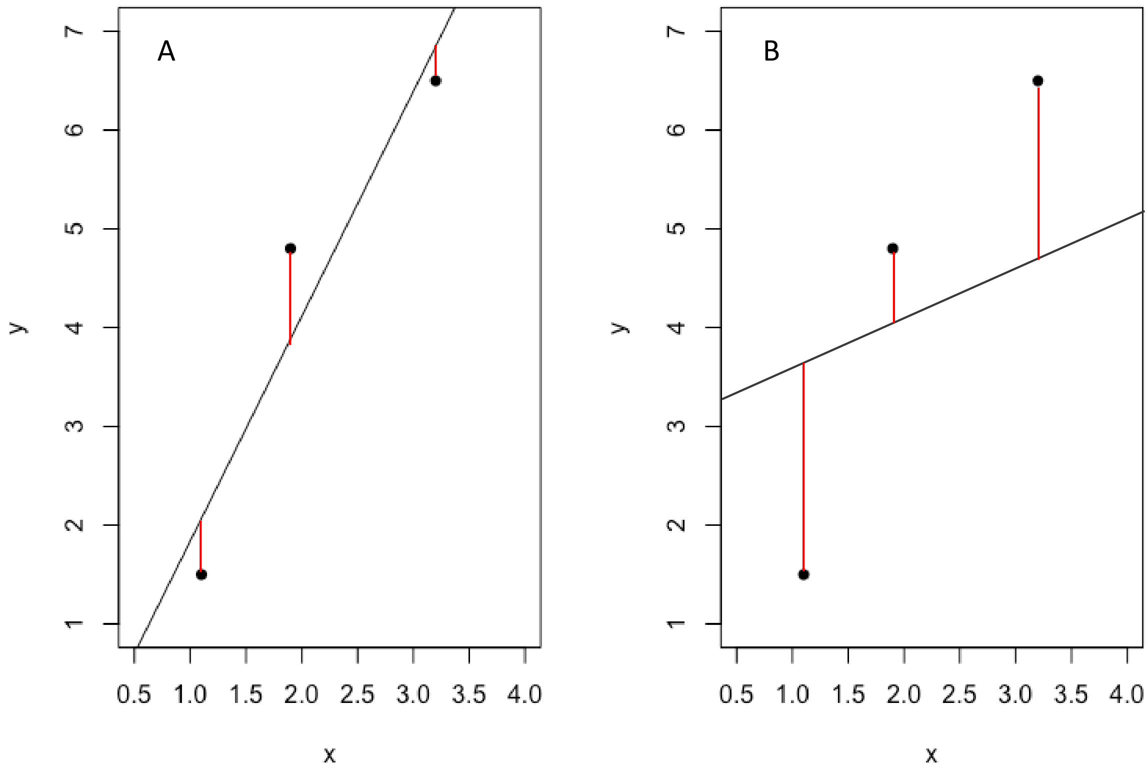


Figure 13.1. A) The best fitting line to 3 data points; compared to B) a poorly fitting line to 3 data points.

This 'least squares' approach works fine if we are assuming observations of the response variable are Normally distributed, but it doesn't work for other common and very useful distributions (such as Poisson or Bernoulli).

13.2 Maximum likelihood

A more general method uses **maximum likelihood**. Recall that the line models the means of the distributions that are assumed to model observations of the response variable.

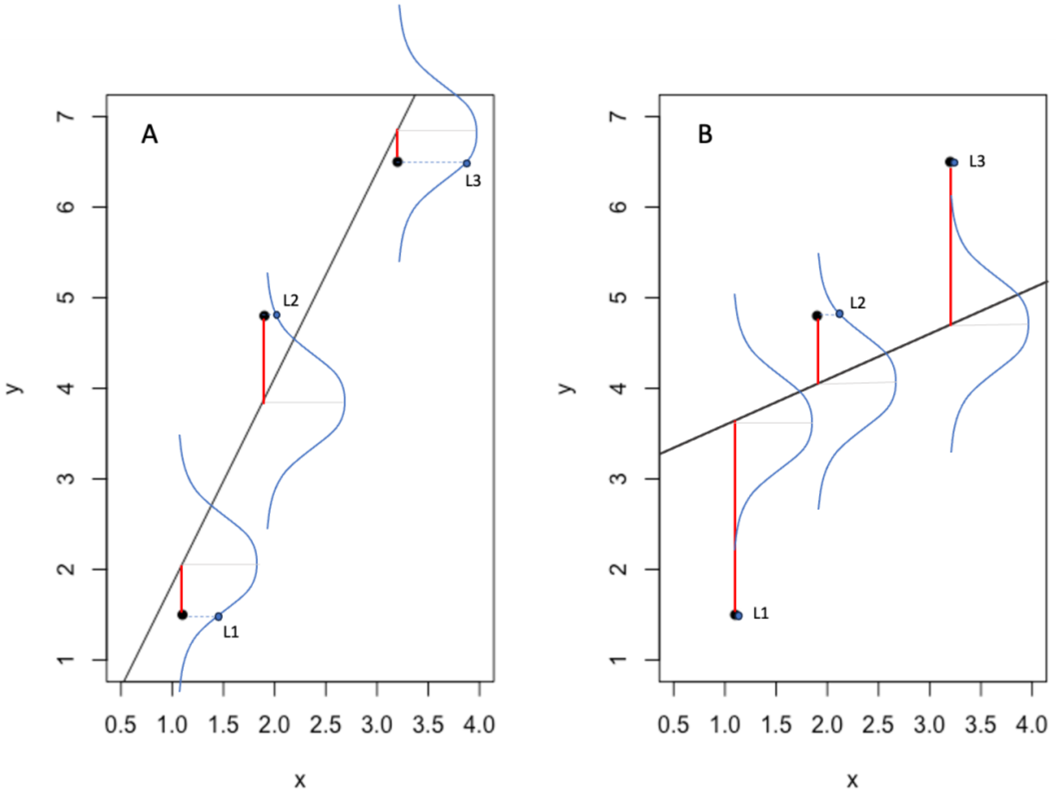


Figure 13.2. Visualizing the likelihood of each data point given a model for a best-fitting line (A), and a poorly-fitting line (B).

So … depending on the relative position of the data within each of these distributions that is generated to model it – there is a certain likelihood of the data point, given the mean generated by the model (Fig. 13.2). Going from left to right on Fig. 13.2A, the first point is only a little less than the mean fitted by the model, and the likelihood (L1) is reasonably high. The second point is quite a bit above the fitted mean and the likelihood is smaller (L2). And the third data point is again less than the mean, but closer to the mean than the first one, so has the highest likelihood (L3). The residuals from the first and third points in Fig. 13.2B are much further out in the tails or their distributions and are consequently much less likely than their counterparts in Fig. 13.2A. If we assume that the variation modelled by these three different distributions is independent for each observation, then just as the probability of 3 heads in a row in a coin tossing experiment is ½ x ½ x ½, the likelihood of all 3 data points is L1 x L2 x L3. All we need R to do is find the intercept, slope of the line, and the variance of the Normal distribution that maximizes this product. There are a range of clever ways of doing this that don't need to concern us just now.

However, there are some points worth emphasising. First, if we can do this for 3 points we can do it for any number of points – data sets both large and small. Second, if we can calculate the likelihood of data points using Normal distributions like this, we can also calculate the likelihood of data points using any distribution – for example, Poisson, Bernoulli, or Negative Binomial -–using exactly the same principles.

You might be wondering why not just fit distributions with very large variances so that even points quite distant from the fitted values are still relatively close to the middle of the distribution? This won't help – and in fact could easily make the data less likely - because recall that the 'area under the curve' must equal one. As we make a distribution 'broader' (for example, the blue distribution in Fig 13.3) we must lower the 'height' of the middle part to conserve the area under the curve, and so the likelihood of the most likely numbers becomes less, even if the likelihood of the less likely numbers becomes more, relative – say – to a distribution with a smaller variance (the red distribution in Fig 13.3).

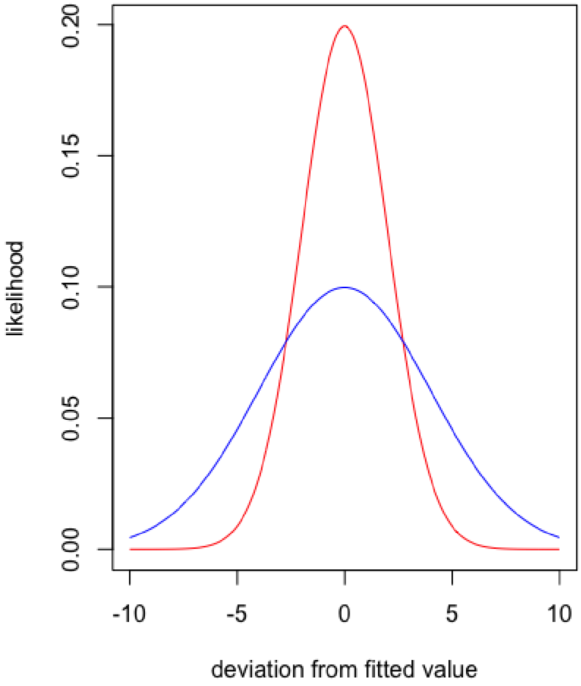


Figure 13.3. Two Normal distributions with the same mean and different variances. As the variance of a distribution is increased the likelihood of the most often encountered values will decrease, as the area under the curve must be conserved.

Therefore, the most likely variance of these distributions will be not too small, and not too large, but characterize distributions that fit the data 'snugly'.

If we multiply a lot of likelihoods that are less than one together – we get a very small number indeed. For example, in calculating the maximum likelihood fit of this model (Fig. 13.4):

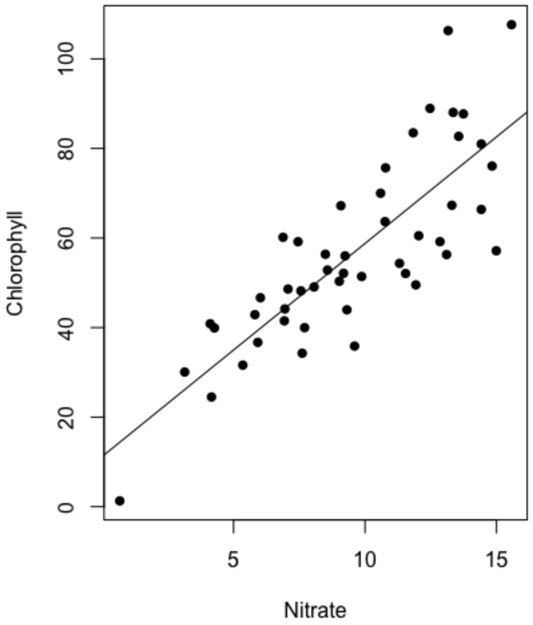


Figure 13.4. The relationship between Chlorophyll and Nitrate for the 48 samples sent to Lab 1.

We'd need to take the product of 48 different likelihoods. This turns out to be rather small:

0.000000000000000000000000000000000000000000000000000000000000000000
0000000000000005240294

Or 5.2240294e-82, which is a bit awkward for both us and the computer. Hence, we tend to work with the natural logarithm of this number, which is -187.156, and refer to this as the **log-likelihood**, remembering that the less negative (or more positive) a log-likelihood is, the larger the likelihood.

We need to remain very aware that by calculating the likelihood of the data set as the product of the likelihoods of each data point, we are assuming that the variation remaining in the response variable *after we have accounted for all the explanatory variables that may induce relationships between them*, is *independent* for each data point. This is a major assumption, and if we find any evidence that there is a non-random pattern in this remaining **residual variation**, we will know that the assumption has been violated and the model fit will be questionable. If we have any reason to believe data points may be related by something that isn't adequately represented in the model ... we will have a problem. We will return to this point in Chapter 16.

The likelihood (or the log-likelihood) of the all the data given the model turns out to be a very useful way of comparing different models fitted to the same response variable. Intuitively – other things being equal – we will favour models that make our observed data more likely, and tend to disfavour models which make our observed data less likely. We will introduce formal ways of comparing likelihoods in Chapters 19-21.

Finally, sometimes problems are encountered when fitting the more complicated GLMs in R and mysterious error or warning messages are triggered - and they can be very mysterious! Typically the explanation is one of the following:

1) The model is overly complex for the data you have, and try simplifying it by removing a term that isn't so central to your study
2) There is inadequate replication for some levels of a categorical variable – a more fundamental problem unless you can generate more data
3) Two different variables are really almost exactly the same thing, i.e. they are very highly correlated, in which case try identifying which pair this might be and removing one of them
4) The explanatory variables might have been measured on numerically very different scales (so the explanatory variables comprise both very large numbers and very small numbers) and you should consider centering them (subtracting the mean of the *explanatory* variable from each observation of *that* explanatory variable), and standardizing its standard deviation (by dividing each observation of the explanatory variable by its respective standard deviation). [There is a reasonable case to be made for always centering and standardizing your explanatory variables, but it does slightly complicate the interpretation of the coefficients which will then assume units of standard deviations of the explanatory variable].

In any case, you will likely need to systematically deconstruct (or construct) your model – in order to identify which terms are causing the problem. It is also possible to try fitting the model using different optimizers (check out the `control` argument in lmer and glmer). Or a different R package (e.g. `glmmTMB`) and the problem may appear to go away, but it might just be that different packages have different warning triggers (or none at all!) so it is best to understand what is really going on.

Important ideas to take-away

- Fitting models by choosing values for coefficients that maximize the likelihood of the data is a robust and general way of fitting models to data
- Fitting models by maximum likelihood and least squares generate exactly the same values for the coefficients of a model of response variables when observations are assumed to be Normally distributed. But the method of least squares cannot be applied to distributions other than the Normal distribution
- Simple maximum likelihood assumes that *residual* variation is independently distributed for each data point

Chapter 14

Degrees of freedom

(back to Contents)

This chapter is a short introduction to the concept of degrees of freedom, how to work out how many degrees of freedom a model requires, and how they can be interpreted as a measure of model complexity.

We've put this discussion off for as long as possible, but we are about to run into this concept and so what follows is designed to provide you with an informal sense of what degrees of freedom are.

14.1 What are degrees of freedom?

You can imagine that you are awarded a degree of freedom for every observation of your response variable that you collect. In essence, a degree of freedom for every data record you have (recall a data record is a 'row' of data assuming you've laid your data out as we recommend).

Each of (say) *n* observations of the response variable is a piece of information unrelated to other observations of the response variable, and each can in principle change without influencing any of the others, so that we can say each data point is *free*, and the data set as a whole has *n degrees of freedom*.

Degrees of freedom can be thought of as enumerating separate *pieces of information* that fully describe or define something. The thing might be our response variable – defined by *n* data points, or it might be something else ... say a model.

In order to fully define a single Normal distribution we need 2 pieces of information: the mean and the variance. We might say a Normal distribution is a model that requires 2 degrees of freedom. A simple GLM of the form say $f_i = c + mx$ is defined by a minimum of 2 pieces of information, the intercept (c), the slope (m) that together define the line, and thus the mean of the distribution we are using to represent the data. In fact, the model may require another coefficient to represent the variance (if the model is – say – a Gaussian (Normal) model) bringing the total to 3, but not if the model doesn't need a separate variance (if the model is – say - a Poisson model).

14.2 How many degrees of freedom do more complex models require?

Once we can write-down the algebraic structure of the model, we can easily count up how many coefficients it contains – each being a piece of information – and each accounting for one degree of freedom.

A degree of freedom is required for the baseline (the intercept), each adjustment for a level of categorical explanatory variable, each slope for continuous explanatory variables, any interaction terms, and the variances for any random effects, and dispersion terms if required by the distribution adopted to model residual variation. Table 14.1 contains some examples, and Appendix Z goes through more examples with more explanation.

Table 14.1. Models and their accompanying degrees of freedom. The +1 in brackets refers to the additional degree of freedom required for estimation of the dispersion term, and remember that a categorical explanatory variable with *q* levels only requires *q*-1 coefficients because one will be the reference level.

Algebraic structure of the model	Distribution of observations of the response variable	Number of levels	Df
$f_i = c$	Normal		1 (+1)
$f_i = c + \alpha_j$	Normal	*j* = 1..4;	4 (+1)
$f_i = c + \alpha_j + \beta_k$	Normal	*j* = 1,2; *k* = 1..3	4 (+1)
$\log(f_i) = c + \alpha_j + \beta_k + \gamma_l$	Poisson	*j* = 1,2; *k* = 1..3; *l* = 1..7	10
$f_i = c + m x_i$	Normal		2 (+1)
$f_i = c + \alpha_j + \beta_k + \gamma_l + m_1 x_{1,i} + m_2 x_{2,i} + m_3 x_{3,i}$	Normal	*j* = 1,2; *k* = 1..3; *l* = 1..7	13 (+1)
$\log(f_i) = c + \alpha_j + \beta_k + \gamma_{jk}$	Poisson	*j* = 1,2; *k* = 1..3	6
$\log(f_i) = c + m_1 x_{1,i} + m_2 x_{2,i} + m_3 x_{3,i}$	Negative Binomial		4 (+1)
$\log(p_i/(1-p_i)) = c + m_1 x_{1,i} + m_2 x_{2,i} + m_3 x_{3,i} + m_4 x_{5,i} + m_5 x_{5,i}$	Bernoulli		6
$f_i = c + \alpha_j + (m + \gamma_j)\, x_i$	Bernoulli	*j* = 1..4	8

Degrees of freedom are a good measure of how complicated something is. That is the number of separate pieces of information required to define it fully. A model with more coefficients is more complex than a model with less. A more complex model is a more complex explanation for the variation we encounter in our response variable and which we wish to understand. As scientists we seek the simplest explanation for variation, so the complexity of our model is going to matter.

However, there is a further important consideration that is the number of degrees of freedom in the data set that are *not* 'consumed' by the model. Suppose we have a complex data set (say 100 observations of a response variable) that therefore possesses 100 degrees of freedom and that contains patterns of variation that we seek to understand. We seek to explain our data with an explanation that is simpler ... that requires less information than the full data set … in fact ... we seek a good model. The model may be relatively complicated, perhaps containing – say – 6 coefficients, but if we can capture the key features of something that is really 100 pieces of information with an explanation that requires just 6 pieces of information we have done pretty well. And in some sense we have 100 – 6 = 94 'excuses' for the variation we have not explained. This partitioning of the available degrees of freedom into those *used* by the model to explain the variation, and those *remaining* to account for the unexplained variation is important. These remaining degrees of freedom are termed **residual degrees of freedom**, and you will see them reported in the output. They are simply the number of observations of the response variable less the number of coefficients required by the model.

Of course, if we constructed a model with the same number of coefficients as observations of our response variable (say 100), each coefficient could represent just one of our observations, and we'd have a model that could account for *all* the observed variation (such a model is known as a **saturated model**). We'd have a model that required 100 degrees of freedom, and we'd be left with zero residual degrees of freedom (with – as we will see later – catastrophic consequences for the estimates of our standard errors). But this model hasn't really explained anything – we wanted to explain something that comprised 100 pieces of information, and we required 100 pieces of information to understand it. The more coefficients we include in our models, the more complicated we make them, the more variation we will explain (for sure) but in fact what we are really trying to do is find the sweet-spot – that is to construct models that explain as much as possible, as simply as possible.

Intuitively, we can see that if we want the most parsimonious explanations for our data, we want to keep the residual degrees of freedom as high as possible, and choose models that are *as complicated as necessary but as simple possible.* We will see in section 22.1 that the reward for preserving as many residual degrees of freedom as we can is that our models will be more powerful, and able to identify more subtle influences of our explanatory variables.

Important ideas to take-away

- The total degrees of freedom you have to work with will equal the number of observations of our response variable
- The number of degrees of freedom our model will require is equal to the number of coefficients you estimate from the data. This may include an

estimate of the variance or dispersion for models where this is estimated separately from the mean

- The number of degrees of freedom your model requires is a measure of the model's complexity
- More complex models will always account for more of the variation in our response variable. But we need to balance accounting for variation with the need to keep the model as simple as possible
- Our choice of model is made in an attempt to explain the most variation with the fewest coefficients
- The remaining degrees of the freedom are considered residual. Maintaining the residual degrees of freedom as high as possible is important, as it will increase the power of our model to detect smaller effects

Chapter 15

Choosing the model

(back to Contents)

This chapter is a brief introduction to the principles of how to choose which model to fit to your data. More formal consideration of the process of model selection can be found in Part 2.

You now know how to add in continuous and categorical **fixed effects**, and **random effects** to the right hand sides of the model (the **linear predictor**); you know how to model interactions between pairs of **explanatory variables**; you know how to choose a distribution to model the unexplained variation in your **response variable**, and the **link functions** (**identity**, **log**, or **logit**) that you would use for these different distributions, and you know how to fit the models. But what are the guiding principles that underpin selecting which model to fit?

This is a complicated and nuanced issue. Obviously, the model needs to address the biological questions that motivated our study. And while we naturally wish to keep things as simple as we reasonably can, there are various good reasons why models might get complicated.

15.1 Single more complex models superior to multiple simple models

It is good and more effective practice to include all your research questions in one more complicated model if you can, as opposed to several models with fewer explanatory variables in each, for at least 4 reasons.

1) Each time you fit a different model to the same response variable you are repeatedly estimating coefficients such as the intercept and (depending on the distribution used) measures of dispersion. To repeatedly re-estimate these coefficients in different contexts from the same data is in some sense wasteful of degrees of freedom (Chapter 14), and diminishes your statistical power.
2) More complicated models that leave less variation unexplained will quite possibly be capable of establishing the significance of explanatory variables that have smaller effects on your response variable than simpler models. Simplistically speaking, unexplained variation acts as a sort of 'statistical fog' that obscures your ability to detect potentially explainable variation.
3) Simpler models that omit influential explanatory variables or interactions that exist between them are more likely to leave residuals that are not truly

independent of each other, in violation of what is assumed when fitting the model.

4) If substantial amounts of variation are left unexplained you are more likely to encounter **heteroscedasticity** (in Gaussian models) or **overdispersion** (in models where this can be a problem - for example, models that assume Poisson distributions where there is a fixed relationship between the mean and the variance).

So, in general it is sensible to include in your model all the variables *and potential interactions* that you think are likely to have a substantial influence on your response variable. You may have variables that you don't think are likely to have a substantial influence on your response variable, but you nonetheless wish to formally establish that this is indeed the case. And/or you may have variables that you are not the least bit interested in – so-called **nuisance variables** – but you none-the-less think may have an influence on the response variable, and perhaps you may be able to treat as random effects. However, this advice notwithstanding, *do try to keep things as simple as you reasonably can*, it will save on degrees of freedom, and make interpretation and description of your findings easier and simpler.

15.2 The 'most plausible complex model'

What we mean by this is a model that contains all the main effects and interactions that might (based on previous knowledge or expert opinion) plausibly contribute to explaining the variation in the response variable. These are likely to include terms (including two-way interaction terms) that are required in order to address your primary research question(s) (go back and recall what you wanted to find out when you went to all the trouble of collecting these data in the first place!), but they may also include additional terms (possibly including other two-way interactions) that are not of direct interest to you, but none-the-less you suspect should be accounted for if the model is to be optimized. The formulation of the **most complex plausible model** is inevitably motivated by a subjective combination of what you are really interested in finding out, and variables (and potentially their interactions) that you feel you can't afford to leave out for some reason. The most complex plausible model doesn't have to be complicated at all .. and certainly it shouldn't be any more complicated that you feel it needs to be.

15.3 Model selection

We advocate giving careful thought to the formulation of the most complex plausible model – and then fitting it. However, it may well turn out that some of the interactions or main effects in the most complex plausible model could be removed without significantly reducing the likelihood of the (response variable) data given the model – which is something you may or may not choose to do. If you do decide you wish to simplify the model you will be embarking on a process called **model selection**. There is something to be said for simplifying a model in this way. The existence of coefficients that are not modelling variation driven by influential explanatory variables with a real effect can lead to a problem called **over-fitting**. If the model is quite a complicated one with several terms and interactions, it is possible that the signal of the explanatory variables and interactions that *are* important is statistically obscured by the presence of a lot of terms that *are not* important (section 24.8 is an example of this). It is also wasteful of degrees of freedom and generates unnecessary complexity that will hinder a simple and easily understood description of your findings. However, it is a form of statistical **fishing**. If you fit many models to your data you are more likely to find one that fits the data well, simply by chance.

(We are not going to go into it here, but if the model you are building seeks to establish causality between your response and explanatory variables (and while this is a common motivation it is not the only one), you may wish to invest some time in understanding something about **causal inference**, where some useful principles have been established about what may or may not be helpful in including in your model, see for example Laubach et al. 2021 A biologist's guide to model selection and causal inference).

Simply speaking, once you have fitted your most complex plausible model there are two positions you could take:

1) Adopt a 'first-and-final' modelling approach: you will fit the initial model – the most complex plausible model, and simply work with whatever output you get as best you can (Chapter 21)**;**

or,

2) Undertake a process of model selection: you will fit the initial model, inspect the output and then proceed with a step-wise process of sequentially removing terms that are deemed to be unimportant, until only important terms remain (Chapter 20)**.**

Either is common and accepted practices, albeit much debated. Option 1 is simpler for sure. But if the model contains interactions that are not important, their presence will obstruct evaluation of the constitutive main effects, so may need to be removed. Model selection can lead to biases in the estimates of coefficients, and the order in which terms are removed and the model is simplified can influence the final model. But a complex model containing many unimportant terms and interactions may obscure the importance of more relevant terms. There is no 'one size fits all' answer, and a confusing range of methods to choose between. It is ultimately a judgement call (as the old aphorism goes 'Good judgment comes from experience; experience comes from bad judgment'). Understanding will come with practice! *The one thing you ought to decide in advance of your analysis is which option you will choose!*

And there are two things we definitely advocate against: 1) basing your conclusions on multiple models, each with a single explanatory variable (for the reasons outlined above); and 2) starting with a very simple model, and adding terms in. The reason for objecting to this latter practice (so-called **forward selection**) is that you will start with a very badly fitting model, and it may be *so* bad that the large amount of unexplained variation will actually obscure your ability to accurately determine whether some explanatory variables should be added to your model.

If you have a large number of explanatory variables (for example perhaps more that 8) you consider examining whether the 'dimensionality' of your explanatory data set might be reduced using **principal components analysis (PCA)** described in Appendix P. Alternatively, you might also consider more advanced forms of model selection known as parameter shrinking methods such as **lasso regression**.

Important ideas to take-away

- There is no right or wrong way to choose a model, but there are important principles that should underlie your approach to model building, and which you should understand

Chapter 16

Checking your model – diagnostic checks

(back to Contents)

This chapter introduces why and how to start thinking about how well your model fits your data, and checks to ensure important model assumptions have not been violated, a vital step between model construction and inference.

Between fitting our model and inferring anything from it, there is a vital further step. We need to make sure the model is a reasonable fit to the data, and that there is no compelling evidence that we have violated the assumptions the modelling process has made. If the model is a very poor fit to the data, and/or we've violated some assumptions of the modelling process, then any inference we make from the model could be seriously flawed. This process is often called conducting model **diagnostic analysis** or checks.

Common problems include:

I. Individual data points are not distributed in the way the model has assumed. Although we may have assumed each observation of the response variable came from, say, a Normal distribution, or a Poisson distribution, in fact ... it looks like they didn't.

II. A commonly related issue is that maybe the model has captured variation in the mean of the response variable, but perhaps there is also variation in the variance of the data that the model has failed to capture. This is known as **heteroscedasticity**.

III. There is a pattern to the residuals that indicates they cannot be regarded as independent of each other.

IV. Recognizing any correlation between the explanatory variables and factoring this into the inference. This is known as **collinearity**.

16.1 Residual analysis

Most of these problems are identified through **residual analysis**. Unfortunately, while residual analysis is relatively straightforward when observations of the response variable are assumed to come from Normal distributions, it is less straightforward when other distributions are assumed. It is worth briefly reflecting on why this is. First, recall that the residuals are the differences between the fitted values from the model and the data points. If the observations really do come from Normal distributions, then we would expect the unexplained variation to be Normally distributed also. This is because if we have a set of variates generated from Normal distributions with different means (essentially our response variable) and we subtract the mean of each of these Normal distributions from each of these variates, the resulting variates are still Normally distributed, but now the mean of these new distribution will be zero.

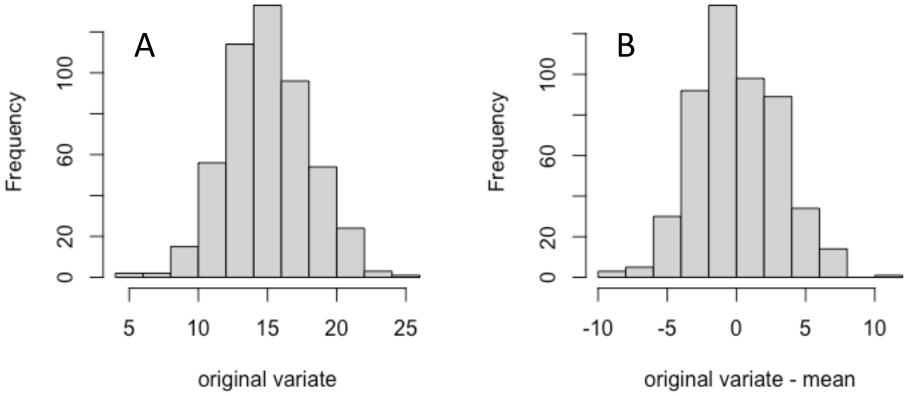


Fig. 16.1. A) 500 Normally distributed variates with mean = 15 and sd = 3. B) The same variates less the value of the mean.

This is most easily understood if the means are all the same (Fig 16.1). We have effectively just moved the distribution to the left, by the value of the mean. However, it applies just as well if the means are different (Fig. 16.2).

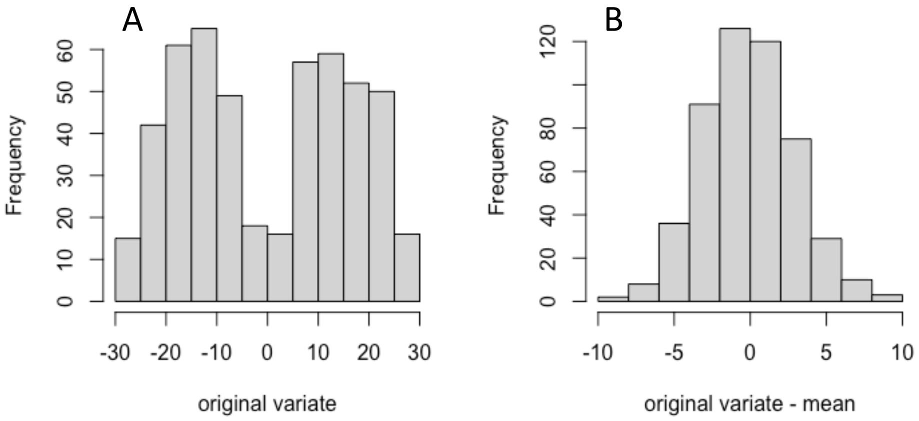


Fig. 16.2. A) 500 Normally distributed variates from 500 different Normal distributions each with an individual mean somewhere between -25 and +25, and sd = 3. B) The same variates less the value of their respective means.

[*This should remind you of an important point made earlier ... we don't expect the frequency histogram of a response variable to look Normally distributed even if each data point itself does come from a Normal distribution. So ... it's not a useful graph to look at in deciding what distribution to choose to model your response variable*.]

But, this is a rather unusual feature of a Normal distribution and it doesn't apply to all other distributions. We can appreciate this by inspecting Fig 16.3.

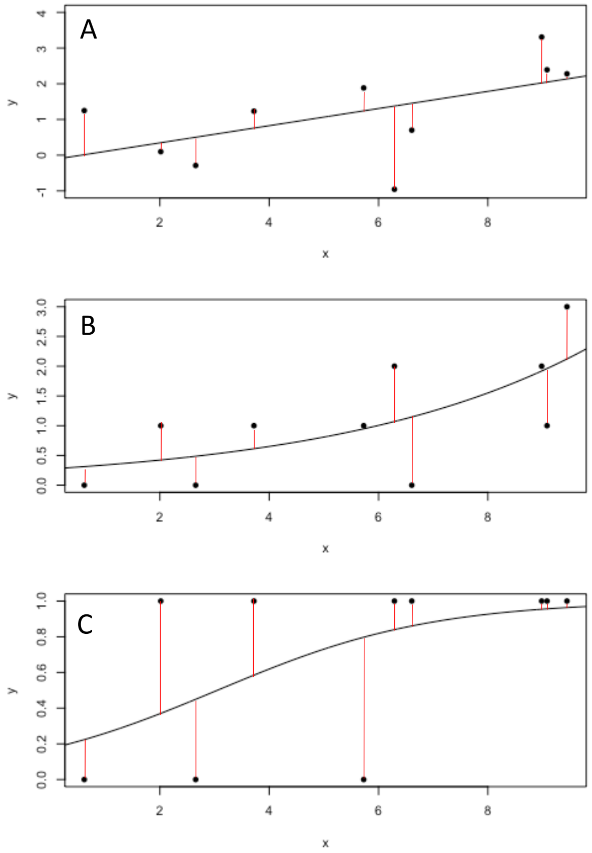


Figure 16.3. Raw residuals from different distributions used for observations of the response variable. The data in each panel are generated according to exactly the same model: *y* = 0.2*x*, but different distributions: A) Normal; B) Poisson, C) Bernoulli. The simple residuals are indicated by the red lines connecting the fitted values to the data points.

In Fig. 16.3A the data are Normally distributed and the (length of the) red lines (the residuals) are Normally distributed. So, we can simply plot them out, and inspect them to see if they look Normally distributed. In Fig. 16.3B the data are Poisson distributed but the residuals are clearly not integers so they can't be Poisson distributed. And in Fig. 16.3C the data are Bernoulli distributed but the residuals are clearly not 0's or 1's so they can't be Bernoulli distributed. For these non-Normal distributions, these simple residual values don't have a formally recognized distribution to compare them with, so we can't perform residual checking in the same way. This is why diagnostic checks for GLMs that don't assume Normal distributions (**Generalised Linear Models**) are tricky, and we are not going to go into residual analysis in great detail.

Being able to model variation around the fitted values with other continuous distributions (for example the Gamma, Lognormal, Weibull or Beta distributions) somewhat overcomes any need for **transforming** the response variable in the event your residuals are evidently not Normally distributed. However, such transformations (for example square-root, log, inverse or **Box Cox**) remain a perfectly acceptable solution to inappropriately distributed residuals if you want to keep things simple. Inference will remain entirely valid, so long as it is made clear the relationships identified exist between the *transformed* response variable and the explanatory variables.

We'll start with checks for when the assumed distribution is Normal.

16.2 Diagnostic checks for data assumed to be Normally distributed

We are interested in three types of plots. The frequency distribution of the residuals, the residuals plotted against the fitted values, and the residuals plotted against the explanatory variables.

As an example, consider the model from Chapter 6, which examined whether the Chlorophyll concentrations in samples sent to Lab 1 could be related to variation in Nitrate. The algebraic structure of the model took the form:

$$f_i = c + m_N\, x_{N,i} \qquad \text{(model 16.1)}$$

And we can plot the data and the line of best fit from the model.

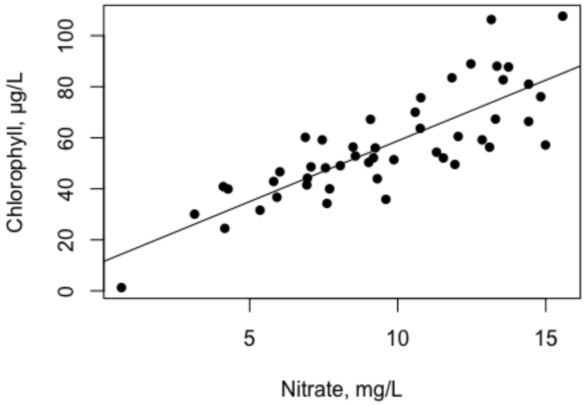


Figure 16.4. The relationship between Chlorophyll concentration and Nitrate from Lab 1.

We can see a slight 'flaring' of the variation towards the right-hand side of the plot. There is more variation in Chlorophyll concentration around the right-hand end of the line than the left. This suggests that the variance of the Normal distributions required to model the data on the right may be more than on the left. But the model assumed that although the mean Chlorophyll concentration increased with Nitrate, the variance stayed the same (this is the assumption of **homoscedasticity**). Recall that in this example there was only one dispersion parameter - estimated to be 148.83, regardless of Nitrate concentration. However, here the data do look somewhat **heteroscedastic**. This is mild heteroscedasticity and its likely not a serious problem. It could be because this is how it is, or it might be that we can account for this by including additional terms (as we did when we included an interaction between Flow and Nitrate – see Figure 11.1B that accounts for this flaring).

We can plot the frequency histogram of the residuals from this model and it all looks satisfactorily Normal, as required.

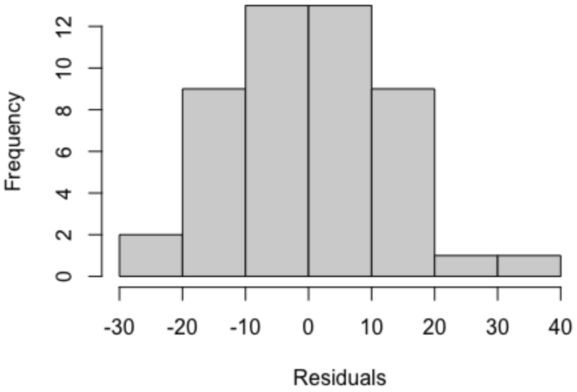


Figure 16.5. Frequency histogram of the residuals from model 16.1.

We recommend that you simply 'eye-ball' this plot, and check it is roughly symmetric, with sloping flanks. There are formal statistical methods for testing whether a distribution deviates *significantly* from Normality. *But, this is not what we want to know*. We don't care if there are *significant* but *very small* deviations from normality. GLMs are quite robust to small deviations from normality. And in any case, the more data you have, the more likely small deviations from normality will be shown to be significant. If you don't have much data, then large deviations from normality will not flag up as significant. Just because we don't have much data, doesn't mean it should be easier to conclude the assumptions of the model have been met. We are interested in *gross deviations* from non-normality – whether they are significant or not. So – we suggest you just take a look at the plot and visually examine it for major indications of non-normality.

We also advocate plotting residuals against fitted values and explanatory variables (in Fig. 16.5 the two plots are essentially identical because there is only one explanatory variable – but this will not be the case when there are multiple explanatory variables). The heteroscedasticity is evident scanning from left to right.

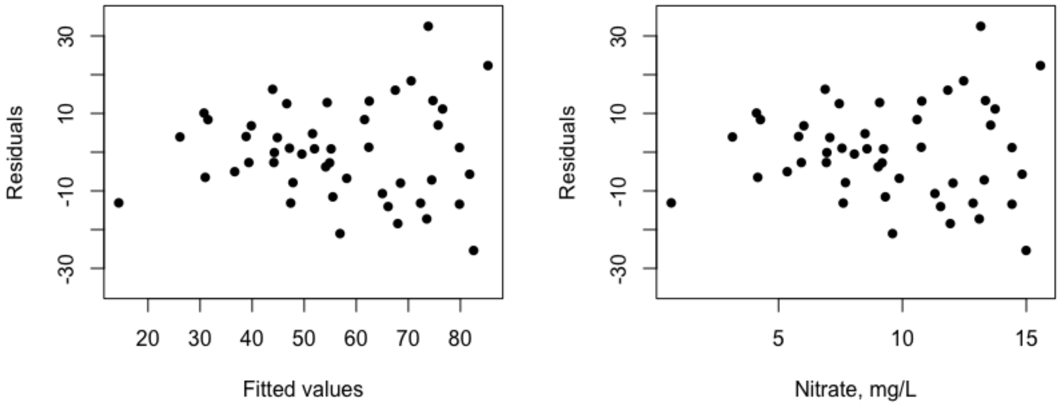


Figure 16.5. Scatter plots showing the relationship between the residuals from model 16.1 and the fitted values and explanatory variable (in this simplest of models these plots are equivalent to each other.

We can also examine something called a **quantile** plot (usually called a **qqplot** in R). A quantile refers to a position in a ranked set of numbers. If you had 100 numbers ranked from smallest to highest, the 1st number would correspond to the 1% quantile, the 5th number to the 5% quantile, and the 95th to the 95th quantile, and so on. By plotting the quantiles from the ranked residuals against the quantiles of a Normal distribution with mean zero and the same standard deviation – we'd expect to get a straight line – if the residuals were from this Normal distribution. qqplots are a useful general tool for indicating whether two sets of numbers come from the same distribution, and they can be constructed for residuals from any model (it doesn't matter what combination of fixed/random/categorical or continuous explanatory variables the model contains). The exact construction of a qqplot is described in Appendix Q.

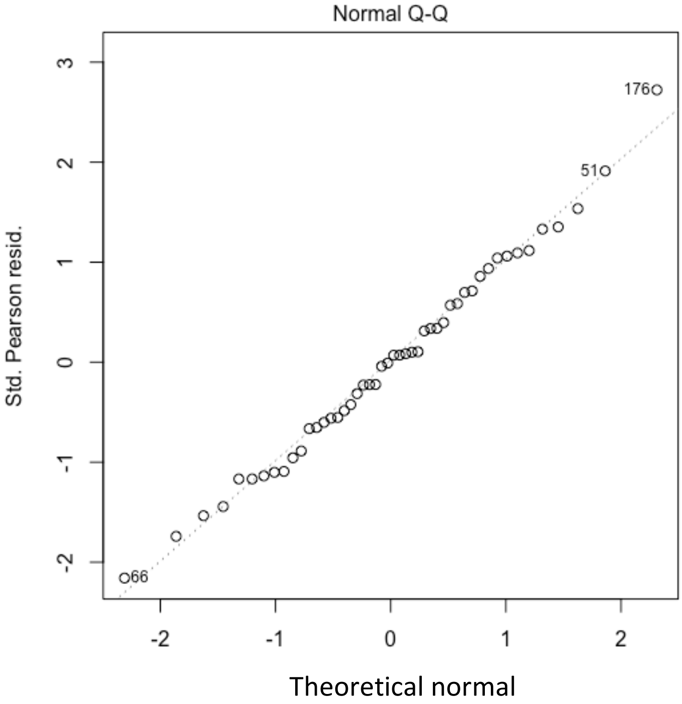


Figure 16.6. The qqplot for the residuals from model 16.1. That these points fall so close to the line x=y suggests the residuals can be regarded as being Normally distributed.

`plot(model)` generates a number of plots of residuals, one of which is the qqplot.

If we add in some categorical explanatory variables (Landscape and Flow) as we did in chapter 8), the algebraic structure would be:

$$f_i = c + \alpha_j + \beta_k + m_N \, x_{N,i}$$

(i = 1 .. 48, j = R or U, k = L, M, H)

(model 16.2)

α_j generates adjustments to the intercept for Landscape, and β_k generates adjustments to the intercept for Flow. The model and data look like this:

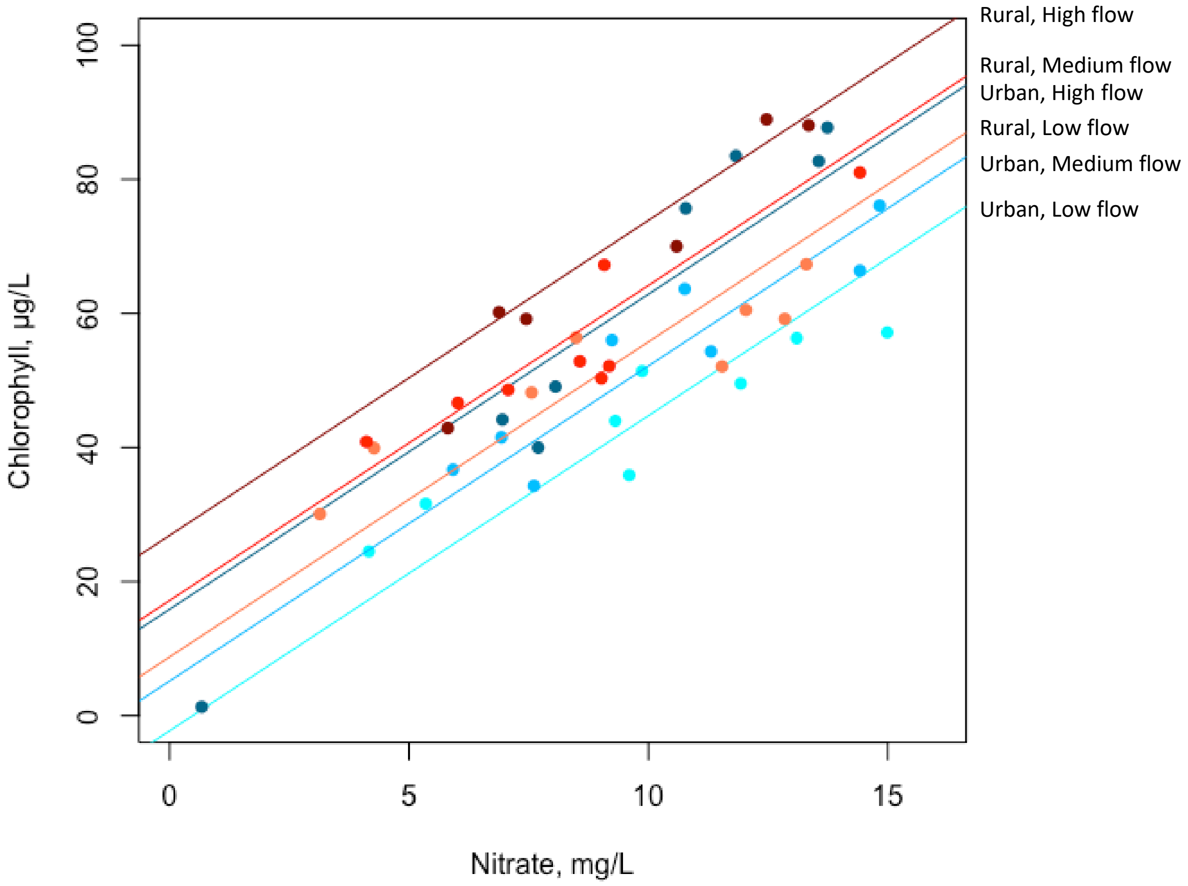


Figure 16.7. The plot for model 16.2 with 6 different intercepts for the 2 x 3 combinations of Landscape and Flow.

The model generates this distribution of residuals:

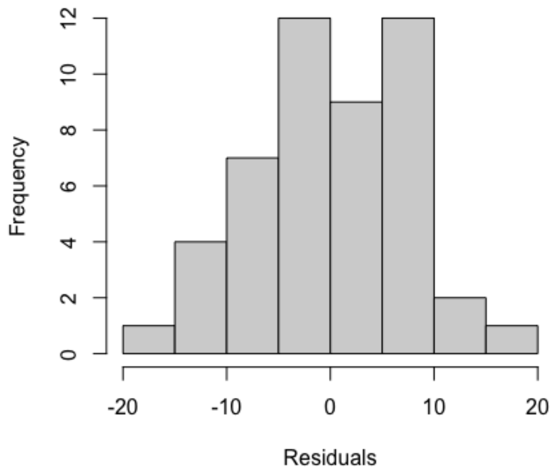


Figure 16.8. The frequency histogram of residuals from model 16.2.

There is a slight hint of bimodality (two peaks), but it might depend on how the histogram was constructed (for example the ‘bin’ boundaries defining each bar interval). We can similarly look at the plot of residuals against fitted values and explanatory variables. The behaviour seems generally good, although there is a slight uptick of residuals in Fig. 16.9A. Fig. 16.9C and D indicate similar variances in the response variable for the different levels (important for complying with the assumption of homoscedasticity).

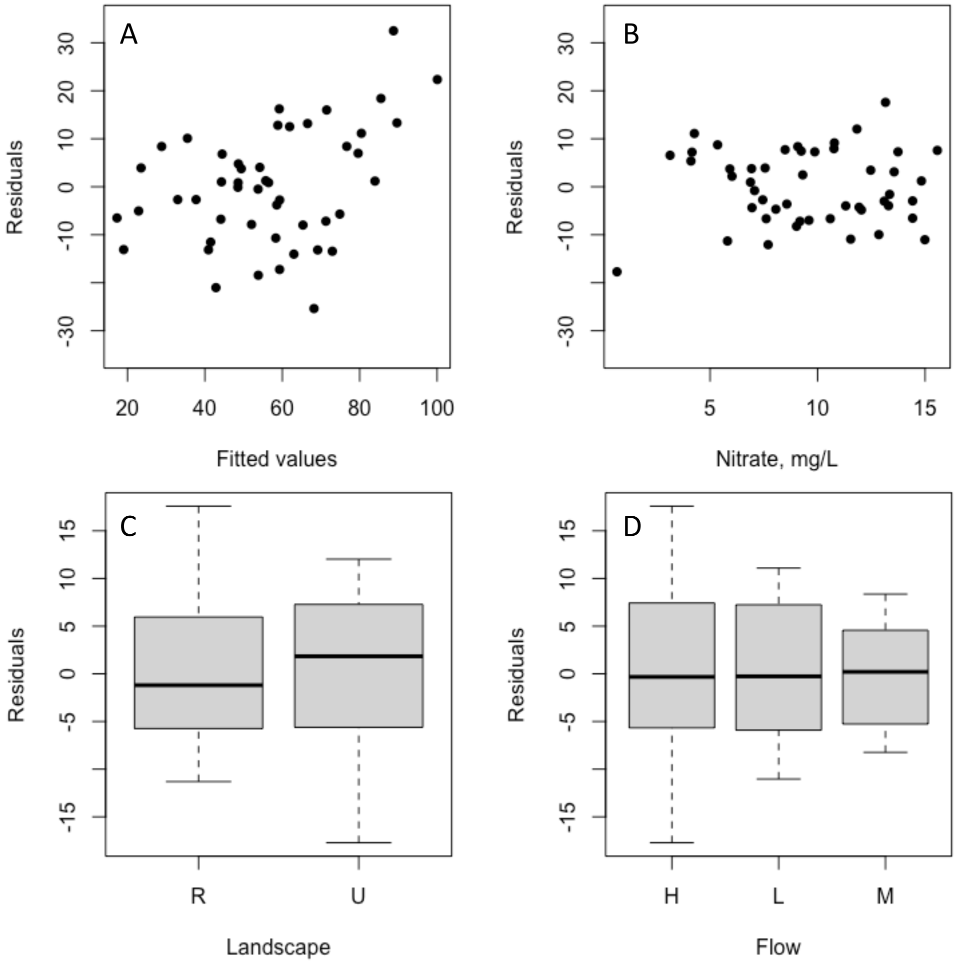


Figure 16.9. Residuals from model 16.2 plotted against fitted values (A) and the three explanatory variables (B,C,D).

The qqplot in Fig 16.10 appears to show some patterning, with sequentially ranked residuals positioned similarly, potentially indicative of the influence of a variable not included in the model.

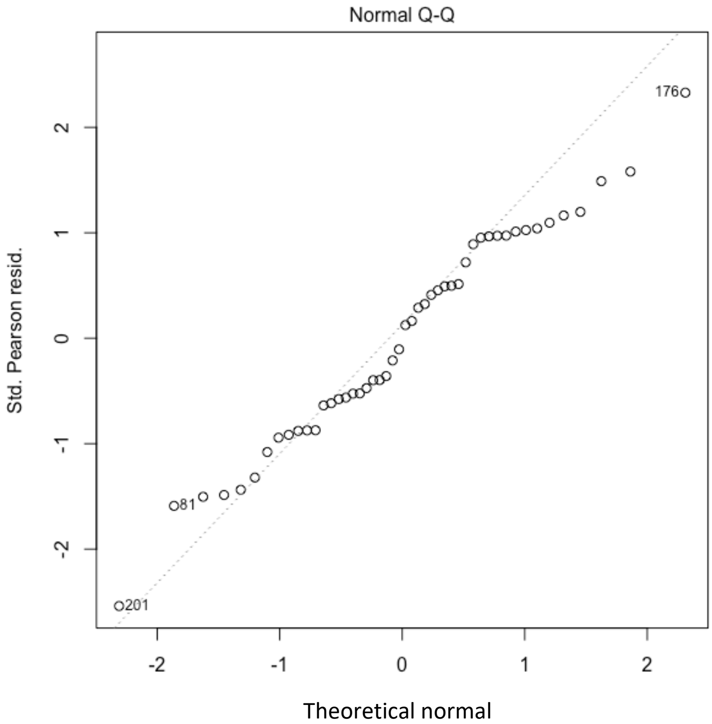


Figure 16.10. qqplot for residuals from model 16.2.

16.3 Diagnostic checks for data assumed to be other than Normally distributed

The residuals we used to check the fit of a model that assumes Normally distributed data are simply the differences between the data and fitted values. These are sometimes called **raw residuals**. As we say in Fig. 16.3, when we adopt other distributions (for example Poisson or Bernoulli) raw residuals don't have a formally defined distribution. So, we have a problem. What are these distributions supposed to look like? There are different ways of transforming the residuals (so called **standardized residuals** or **Pearson standard residuals**) and conducting the same visual checks on these, but various assumptions are being made, and the interpretation is notoriously difficult.

Perhaps a better and more general way of thinking about model fit is to ask if your model can generate data that looks like the real data. If your data are 'likely' given your model, then it makes sense that your model could generate simulated data like your real data. How can a model generate data? Easily. And understanding how, will probably help you understand what these models are really doing.

We should be familiar with the idea that we can generate random numbers according to any distribution in R (if not go back to Chapter 4). Briefly, commands like `rnorm()`, `rpois()`, and `rbinom()` generate Normal, Poisson and Binomial random numbers with means and variances dictated by arguments supplied to these commands. Our GLMs are estimating the means and variances for each data point from our dataset as a whole. So, just as we might recognize a random number *r* to come from a Normal distribution with mean (μ) and variance (σ^2): $r \sim N(\mu, \sigma^2)$, so we can generate 'pseudo data' as $y_i' \sim N(f_i, \sigma^2)$, where f_i is the fitted value for the i^{th} observation of the response variable estimated by our model, and σ^2 estimated similarly from the data (recall we usually refer to the observed values of our response variable as *y_i*,;here the superscripted prime (the apostrophe) indicates that the response variable y_i, is simulated and not observed). It doesn't matter what distribution our model assumes – the principle is the same: we have estimated everything we need to simulate data sets from the model. Were we to have fitted our model assuming the response variable was Bernoulli distributed we would simulate data using: $y_i' \sim Bern(p_i)$, and if Poisson using: $y_i' \sim Pois(f_i)$.

R makes data simulation very easy. `simulate(my_model,10)` will generate 10 replicate data sets from your model. We could generate say 1000 such replicate data sets, and use it to generate a distribution of what our model predicts for each single observation of our response variable and note the position of our response variables in these distributions. If the positions of the observations of the response variable fit into these simulated distributions in a way we would expect (and this can be done in clever ways) the model could be said to be a reasonable fit. The R package DHARMa makes this whole process very easy.

```
simulationOutput <- simulateResiduals(fittedModel =
my_model)

plotQQunif(simulationOutput)
```

[don't be put off by the reference to unif in the command – these QQ plots can be constructed for any distribution with this command]

Figure 16.11 shows these qqplots for these 3 models:

```
mA<-glm(Chlorophyll~Nitrate+Landscape+Flow,data=my_data)

mB<-glm(ZooCount~Flow,family=poisson,data=my_data)

mC<-glm(Disease~Flow,family=binomial,data=my_data)
```

The distributions indicated in A and C seem quite good, but B shows some departure that we should probably investigate further.

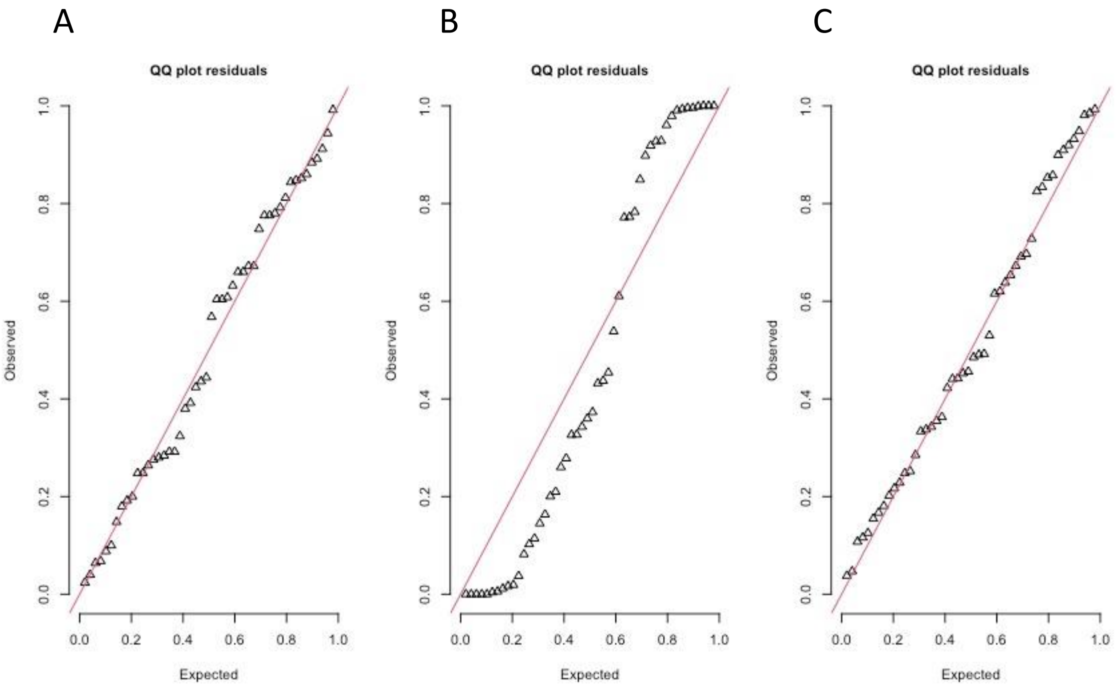


Figure 16.11. qqplots for residuals defined and simulated in the package DHARMa for: A) a Gaussian model; B) a Poisson model; and C) a Binomial model. The red line indicates a one-to-one correspondence between observed values and those expected based on the simulated data. The plotQQunif command will by default conduct formal tests for departure from the expected distribution (see ‘Plotting the scaled residuals’ in the DHARMa help section for more details of exactly how these plots are constructed).

The simulation approach to model fit is powerful and relatively straightforward. Further discussion is beyond the scope of this text, but it is well worth investigating the DHARMa package in greater depth:

(https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html).

16.4 When does a lot of unexplained variation matter?

We know these models essentially partition the variation in observations of the response variable into explained and unexplained variation. So, it’s a relatively straightforward matter to calculate the percentage of variation that is explained by the model. This is most straightforward when fitting a model using the method of **least squares**, and we can examine the % **sums of squares** that are explained, a metric known as **R-squared**. When we fit models using maximum likelihood we can’t calculate this quantity – as it depends on a fitting process only applicable when the model assumes a Normal distribution. But we can calculate something called **pseudo-R-Squared** using the **null deviance** and **residual deviance**.

The **null deviance** is twice the difference of the log-likelihood between a **saturated model** and an **intercept only model** (or **null model**) fitted to the data. An intercept-only model would be say: $f_i = c$, or $\log(f_i) = c$ (for count data), or $\text{logit}(p_i) = c$ (for binary data). A saturated model would be one in which there is a coefficient for each data point (basically a model in which the data are as likely as possible). The null deviance can be thought of as the amount of variation available to explain.

The residual deviance is twice the difference between the log-likelihood of a saturated model, and the model we are fitting to the data. The smaller this difference is, the better our model has performed, but it can't do any better than the saturated model. Thus:

$$1 - \frac{\text{residual deviance}}{\text{null deviance}}$$

behaves very like R-squared, and forms the basis for what we call pseudo-R-squared. Psueudo-R-squared is useful, as it can be calculated for any model, regardless of the distribution used to account for unexplained variation in the response variable.

An R package for calculation of Pseudo R-squared for fixed effects models would be `PseudoR2` in the `DescTools` package, for example:

```
>PseudoR2(my_fixed_effects_model)
```

and the `r.squaredGLMM` command in the `MuMIn` package for mixed effects models. For example:

```
> r.squaredGLMM(my_mixed_model)
```

This command returns marginal and conditional estimates depending on whether the random effects are included or not.

It is not unusual to find that our models account for rather little of the total variation. The world is a variable place, and we usually can't account for most of it! However, this does not mean that the explanatory variables we include in our models are necessarily unimportant. *We need to make a distinction between how confident we are that an explanatory variable is influencing the response variable, and whether this is 'biologically meaningful' or not*. Bacon sandwiches are known carcinogens, but the effect is very small indeed, and this is not a major factor in whether we decide to eat them or not. *Not everything that is significant will be important to us, and not everything that is important to us will be significant*. They are different things.

The bottom line is that models that explain a low percentage of the variation in the response variable are not necessarily problematic, so long as we carry out our inference correctly, and the biological magnitude of the effects of the explanatory variable are deemed interesting and/or useful. However, we think it's useful to be aware of the percentage of variation our models account for when interpreting and reporting findings.

16.5 Overdispersion

There is one type of model for which a lot of unexplained variation can be a problem, and that is when the model assumes a Poisson distribution. Recall that a Poisson distribution only has one argument that equals both the mean and the variance of the Poisson distribution. If these are not approximately the same– and there is no particular reason why they should be, we'll have a problem. It is entirely possible – indeed very common – that the unexplained variation in observations of our response variable exceeds that which can be plausibly accounted for with a Poisson distribution – which is to say … a Poisson model may not fit the data. *This situation of excess variation relative to the mean cannot be identified prior to fitting a model* – who knows, perhaps all the variation will be explained by the explanatory variables. But it can be identified retrospectively – in qqplots, or more simply, by examining the relative size of the residual deviance to the residual degrees of freedom. Ideally, we'd like the residual deviance and residual degrees of freedom to be about the same. Once the residual deviance becomes more than 20-30% larger than the **residual degrees of freedom** we should begin to worry about **overdispersion**.

Overdispersion is evident in the `ZooCount ~ Flow` model below, where 184.09 is four times the residual degrees of freedom, and it is this overdispersion that is evident in the qq plot in Fig. 16.11B.

```
glm(formula = ZooCount ~ Flow, family = poisson, data = d1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1632     0.4471  -2.601  0.00929 **
FlowL         3.4782     0.4540   7.661 1.84e-14 ***
FlowM         2.8094     0.4604   6.102 1.05e-09 ***
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 370.67  on 47  degrees of freedom
Residual deviance: 184.09  on 45  degrees of freedom
```

It is important to stress that while a high ratio of residual deviance to residual degrees of freedom indicates a lot of unexplained variation, this is only a *problem* for distributions in which there is a fixed relationship between a distribution mean and its variance. For distributions like a Normal distribution, there are separate arguments for the mean and variance, and they can be fitted independently of each other: each can be whatever the data require. But the Poisson distribution only has one argument that must serve as both the mean and the variance, and so for a Poisson distribution overdispersion can be a serious problem. If you encounter it, usual practice is to model the data with a Negative Binomial distribution instead, which, like a Normal distribution, has separate arguments for the mean and variance so overdispersion is not a technical problem. The `MASS` package has a commands `glm.nb` (or `glmer.nb` if you also have random effects) that will fit Negative Binomial models.

If you have a mixed model and using `lmer` or `glmer` in the `lme4` package the residual deviance will not be included in the output, but you can generate it using the `over.disp` command in the package `RVAideMemoire.`

16.6 Collinearity or non-orthogonality

Collinearity (also known as **non-orthogonality**) arises when two explanatory variables are correlated with each other. This means that the information they are providing about variation in the response variable is common to each. A classic example might be if we attempted to model variation in human height using both the length of an individual's left leg, and the length of their right leg. Both are useful for this purpose, but if you have one, there is no need for the other.

Collinearity creates a *problem* because if two explanatory variables – say – A and B – are collinear to each other, the presence of both in the model may be unnecessary. We can end up in a situation where there is no gain to adding B to the model if A is already in it (suggesting that B is not important) but also no point in adding A to the model if B is already in it (paradoxically suggesting that A is not important). This situation would be revealed by applying the relevant likelihood ratio tests (Chapters 19-21). Collinearity is common in biological data – particularly in observational data or natural experiments (less so in controlled experimental data), and while it doesn't violate any assumptions of the model, and it doesn't reduce the explanatory capability of the model, it does complicate interpretation, often widening the standard errors on the model coefficients, and making explanatory variables appear to be less significant than perhaps they actually are. It is not something we can do anything about, but we can at least understand it. Thus, we should be aware when it is present, and adapt our interpretations accordingly. *Collinearity between two explanatory variables and the existence of an interaction between two explanatory variables are two completely distinct and different phenomena – the presence of one neither precludes or makes the other more likely*.

Collinearity may be almost complete, or partial, and it may be asymmetric. That is to say, the information provided by – say – explanatory variable A, about the response variable may overlap 100% with that of – say – B. Or less than 100%. And just because A shares information with B, doesn't mean B shares information with A. It may still be worth having both in a model, but don't expect their joint presence to equal their contributions when present on their own (this and other examples are discussed in Fig 16.12).

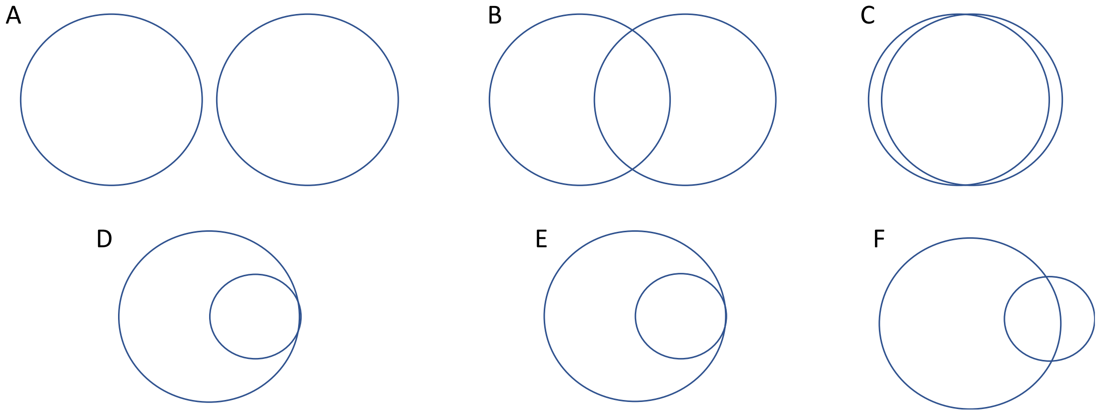


Figure 16.12. Different patterns of collinearity. Think if each circle as the set of information that the red or blue variable contains that is informative of the response variable. A) no collinearity, variable 'red' and 'blue' are completely uncorrelated with each other (orthogonal), and convey entirely separate information about the variation in the response variable, and both may be useful in the model; B) Some collinearity, i.e. some modest overlap in the information conveyed by red and blue in the variation in the response variable, and both may still be useful to retain in the model; C) Almost complete collinearity, i.e. almost complete overlap in the information conveyed by red and blue in the variation in the response variable, and while either red or blue might be useful in the model, no point in retaining both; D) Both blue and red explain variation in the response variable, but all the information conveyed by red is conveyed by blue, but the converse is not true, so retain blue in the model, and don't bother with red, unless blue is not available in which case red may still be useful; E) Both blue and red explain variation in the response variable, but all the information conveyed by blue is conveyed by red, but the converse is not true, so retain red in the model, and don't bother with blue, unless red is not available in which case blue may still be useful; F) Both blue and red explain variation in the response variable, but while red explains more variation in the response variable than blue, the majority, but not all of the

information conveyed by blue is conveyed by red. Both may be useful to retain in the model.

Some people use **Variance Inflation Factors (VIFs)** to help understand where collinearity might be a problem (VIFs are explained more in Appendix R).

It is important to stress that collinearity is a completely separate phenomenon to an interaction. An interaction is when the effect of one explanatory variable on the response variable depends on another explanatory variable. Collinearity is when the information provided by one explanatory variable is *also* provided by another.

Important ideas to take-away

- Model checking is not an optional extra, it's an imperative. Your inference will not be reliable if the model is in serious breach of its underlying assumptions
- Models assuming that observations of the response variable are normally distributed should have normally distributed (raw) residuals
- The raw residuals from models that assume other types of distribution have no formal definition so residual analysis is more complicated but remains important and necessary. Options are to examine some form of transformed residual, some form of simulated residual, or (less rigorously) to inspect the residual deviance
- GLMs are tolerant of minor violations of model assumptions. No model is perfect, indeed all models are wrong, but you need to make sure your model is not *importantly wrong*

Postscript (part 1)

(back to Contents)

Before getting stuck into inference, we'd like to emphasise two important principles of data modelling. The first (again), is that you should answer as many of your questions as possible in a single model. There is rarely justification for basing your inference relating to the same response variable on more than one GLM. In other words, fit one more complicated model, rather than multiple simpler models. This is more powerful, efficient, and concise. The second, is whenever possible try to model the original data. In general, try to avoid averaging them, smoothing them, summing them, adding one to them, or otherwise transforming them. The ability to model data with a diverse range of distributions removes the motive to 'normalize' the data. The use of random effects should enable the avoidance of issues relating to 'repeated measures' or **pseudo-replication**, and the appropriate accounting for of **nuisance variables**. It is possible to 'control' for the effects of an explanatory variable by simply including the explanatory variable in the model. Sometimes – especially in the older literature, you'll see the response variable represented as residuals from a previously fitted model, as analysts attempt to 'pre-control' for effects of this or that. There is no obvious need to do this when it can be accomplished in a single model.

While we have covered most of the common types of data you are likely to model, you will occasionally encounter other types. Instead of a binary response variable you may have unordered trinary data (for example 'agree', 'disagree', or 'don't know') which can be modelled using a multinomial as opposed to a binomial (or Bernoulli) distribution (Appendix S). You may have a response variable that is an ordered category (perhaps a **Likert scale**) in which case you may want to use an **ordinal GLM** (Appendix T). Or, you may have circular data (for example turning angles, compass bearings, calendar months, time of day) in which case you may want to explore wrapped or circular distributions (briefly described in Appendix E). If you have a single response variable, there will almost certainly be a suitable distribution out there (and an R package) that will do what you need.

Lastly, lets circle back and recognize some of the more traditional statistical models that are encompassed by the idea of the GLM family.

Old-speak	GLM-speak	Algebraic structure
T-Test	One categorical explanatory variable, 2 levels, `family = Gaussian`	$f_i = c + \alpha$
One-way Anova	One categorical explanatory variable, > 2 levels, `family = Gaussian`	$f_i = c + \alpha_j$
Two-way Anova	Two categorical explanatory variables, `family = Gaussian`	$f_i = c + \alpha_j + \beta_k$

Linear regression	One continuous explanatory variable, `family = Gaussian`	$f_i = c + mx_i$
Multiple regression	Two (or more) continuous explanatory variables, `family = Gaussian`	$f_i = c + m_1 x_{1,i} + m_2 x_{2,i}$
Analysis of covariance	One continuous explanatory, one categorical explanatory variable, and their interaction, `family = Gaussian`	$f_i = c + \alpha_j + (m + \gamma_j)x_i$
Logistic regression	One continuous explanatory variable `family = binomial`	$\log\left(\frac{p_i}{1-p_i}\right) = c + mx_i$
Poisson or count regression	One continuous explanatory variable `family = Poisson`	$\log(f_i) = c + mx_i$
χ^2 contingency test or log-linear analysis	Two (or more) categorical explanatory variables, `family = Poisson`	$\log(f_{jk}) = c + \alpha_j + \beta_k$
Repeated measures ANOVA	One (or more) explanatory variables and a random effect	(say) $f_i = c + \alpha_j + R_k$

If you have got to here – you have come a very long way.

Part 2

Inference

Chapter 17

Test statistics

(back to Contents)

A broad swath of statistics uses quantities called test statistics that have a simple and well understood general behaviour enabling a wide range of the most common questions to be answered using a wide range of different sorts of data. Here we introduce two test statistics we'll be using a lot. This material is an important pre-requisite for the rest of the chapters in Part 2. You'd best read this chapter.

Before we embark on a lengthier discussion of inference, it is worth describing the basic mechanics that applies to most common statistical tests.

Most of statistics is about studying data for some sort of pattern or relationship, and then trying to establish how likely it is that the pattern or relationship could have arisen by chance, and if they are sufficiently unlikely to have arisen by chance then we accept an alternative explanation. Of course, data sets are infinitely diverse in their natures, and we may be curious about any number of different patterns or relationships, so how can just a few statistical tests provide the analyses we want given this diversity of data and questions? The answer is that we can subject data to all kinds of seemingly bizarre manipulations (summing squares, looking at ratios of sums of squares, calculating the differences in log-likelihoods ... all kinds of weird stuff) ... in order to arrive at a metric (test statistic) that has some standard properties. So, it doesn't matter what the original nature of the data were, or even what the exact question was we wanted to address. We can just focus on a single metric, and determine from this if the pattern or relationship is consistent with having arisen by chance, or requires a more substantive explanation. These metrics are called **test statistics**, and they are often ingeniously developed to have very predictable qualities, that is – to be distributed in known ways. The names of these test statistics often indicate the distributions they are expected to conform to. For example, the ***z* statistic**, the **T statistic**, the **F statistic**, the **χ^2 statistic**, and so on.

For example, we might generate a test statistic that tests whether a certain pattern is present in our data. We can hypothesize two scenarios: 1) a **null hypothesis** that there is no basis for the pattern and any observed pattern has arisen purely by chance; and 2) the converse of the null hypothesis – the **alternate hypothesis** that the pattern does have a basis and hasn't arisen purely by chance. We'll know the distribution of the test statistic *on the assumption the null hypothesis applies*. The larger the test statistic is – and the more deviant it is from the expected distribution – the less likely it is that the null hypothesis applies, and the less likely it is that the observed pattern arose purely by chance. But how large *is large enough* that we conclude the pattern is *real* and hasn't arisen simply by chance? We need to know how probable it is that a test statistic this large (or more) could have arisen by chance – this is the famous **p-value**. If the test statistic is *unlikely* to be as large as we observe it to be by chance, we can reasonably infer that pattern didn't arise by chance, and most likely *does* have a substantive basis. What do we mean by unlikely? By convention we choose a threshold of less than 0.05 (i.e. a 1 in 20 chance of such an outcome arising by chance, although we are free to adopt a more or less stringent threshold depending on the consequences of rejecting the wrong hypothesis). We can term test statistics less likely than such a threshold to be **statistically significant**.

While there are many different test statistics, we will routinely use the two described below.

17.1 χ^2 statistics

You may have come across χ^2 statistics before in relation to the chi-squared contingency or goodness-of-fit test (a separate use of this test statistic that we are not going to discuss here). In fact, χ^2 statistics are very general, and widely used to test for a variety of different things, and it's useful to make a distinction between the different things we might test for, and the test statistics we might use in those tests.

For example, in section 16.5 we introduced the notion of overdispersion in Poisson-based models. Overdispersion was assessed by inspecting how large the residual deviance was, relative to the residual degrees of freedom. In 16.5 the model we fitted had a residual deviance of 184.09 with 45 residual degrees of freedom. So ... *is 184.09 too large?* Some clever mathematics can be used to show that the residual deviance in the absence of any non-random overdispersion is in fact χ^2 distributed, with (in this example) 45 degrees of freedom (the degrees of freedom is an argument of the χ^2 distribution, in the same way that the mean and standard deviation are arguments of a Normal distribution) and so the residual deviance itself can be used as a test statistic. According to this expected distribution anything more extreme than 61.66 is less probable than 0.05 (Fig. 17.1). So, an observed residual deviance of 185.09 is really very unlikely indeed under the null hypothesis, we therefore reject the null hypothesis that the overdispersion has been observed by chance alone, and conclude the overdispersion is almost certainly real.

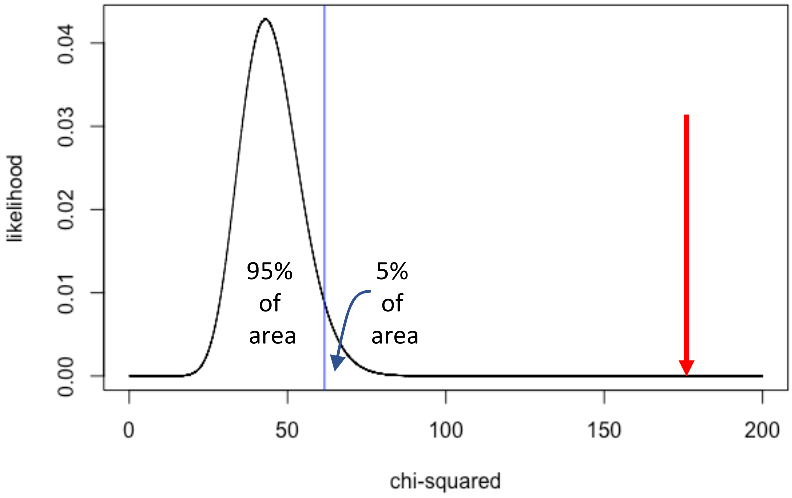


Figure 17.1. A χ^2 distribution with 45 degrees of freedom and the position of the 95th quantile shown in blue. The area under the curve to the left of the blue line accounts for 95% of the area under the curve. Observed values of the test statistic between 0 and 61.66 would not be regarded as improbable. Observed values to the right of the blue line would suggest that the residual deviance is not consistent with the null hypothesis that the overdispersion has arisen by chance alone. The p-value that would be routinely cited from this test would be the area under the curve to the right of the red arrow (which is very very small indeed, so $p < 0.0001$).

The details of exactly how and why this example works are not important just now. Our general point is that we can calculate test statistics from data (in this case the residual deviance) and knowing how this should be distributed under the assumption of our null hypothesis (no real overdispersion), use the position of the observed test statistic in the expected distribution of the test statistic – to reject (or fail to reject) our null hypothesis.

We can plot different χ^2 distributions with different degrees of freedom to determine how the 95th percentile changes (Fig. 17.2). The appropriate number of degrees of freedom depends on the details of the test being applied, but often is just a few – we'll discuss more later.

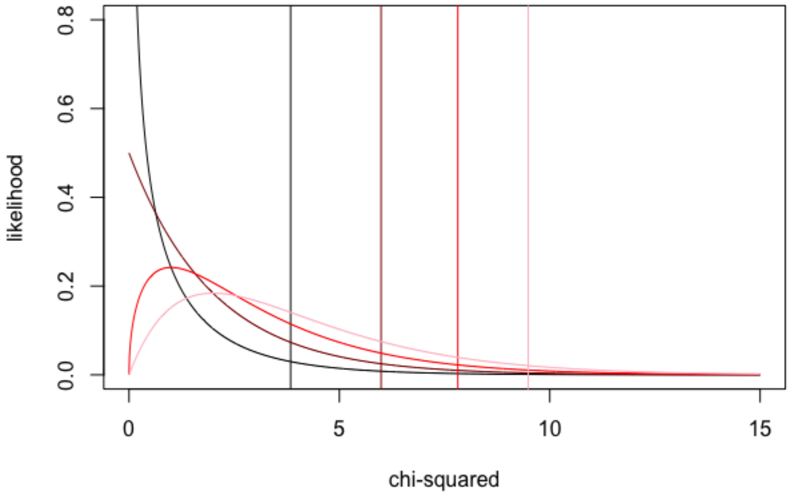


Figure 17.2. Four χ^2 distributions with different degrees of freedom: 1 (black); 2 (dark red); 3 (red), and 4 (pink). The 95th quantiles fall at: 3.84, 5.99, 7.81, and 9.48, respectively, as indicated by the vertical lines.

17.2 T-statistics

Here are 10 numbers:

6.08 2.29 4.15 5.99 2.62 3.74 3.14 -0.12 0.24 3.42

The observed mean is 3.15 and the standard deviation 2.06. How sure can we be that these numbers are not variates from a distribution with a mean of zero? The observed mean is greater than zero, and there is a pattern in the data ... (e.g. 9/10 of these numbers are greater than zero), but is this pattern sufficiently evident that I could reject the null hypothesis that 3.15 is meaningfully different to zero? On the face of it ... this feels like a rather complicated question. Fortunately, if we assume the numbers come from a Normal distribution, then there is a test statistic called a T-statistic which can be calculated by dividing – the difference between the mean of these 10 numbers and the value assumed under the null hypothesis, by their standard deviation. The observed mean is 3.15, the standard deviation is 2.06, so the relevant T statistic is (3.15 - 0)/2.06 = 1.53. Is this test statistic *unusually large*? According to the expected distribution (assuming the mean of the 10 numbers really was zero) there is a 95% chance the test statistic should be between -2.26 and +2.26. Since 1.53 falls in this interval ... we can conclude that – no, 1.53 is not unusually large, and therefore we cannot reject the hypothesis these 10 numbers have a mean that is different to zero (the p-value is 0.16) (Fig. 17.3).

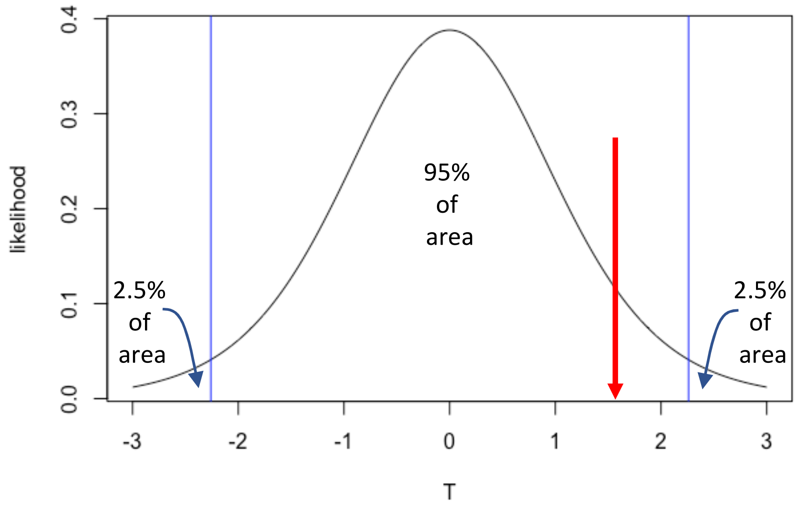


Figure 17.3. A T distribution with 9 degrees of freedom and the position of the 2.5th and 97.5th quantile shown in blue. The area under the curve between the blue lines accounts for 95% of the area under the curve. Observed values of the test statistic falling within this interval (-2.26 to +2.26) would not be regarded as improbable. Observed values of the test statistic in the left-hand or right-hand tail would suggest that the value of the mean hypothesized under the null hypothesis is inconsistent with the observed data. The p-value that would be routinely cited from this test would be twice the area under the curve to the right of the red arrow – there by accounting for both *the greater than* and *the less than* possibilities.

Our specific point is that if we know the mean and standard deviation of a distribution we can use a T statistic to determine if this mean differs from a particular value (zero, or indeed some other value). The T statistic reflects the number of standard deviations the observed mean is from the value proposed under the null hypothesis (by default most often assumed to be zero, but it could be any value of interest, see for example Fig. 17.4). Of course, the bigger the difference and the smaller the standard deviation the larger the T statistic will be. Usually, a T statistic in excess of 2 will indicate a significant difference from the mean proposed under the null hypothesis, but it depends on the number of degrees of freedom available (Fig. 17.5).

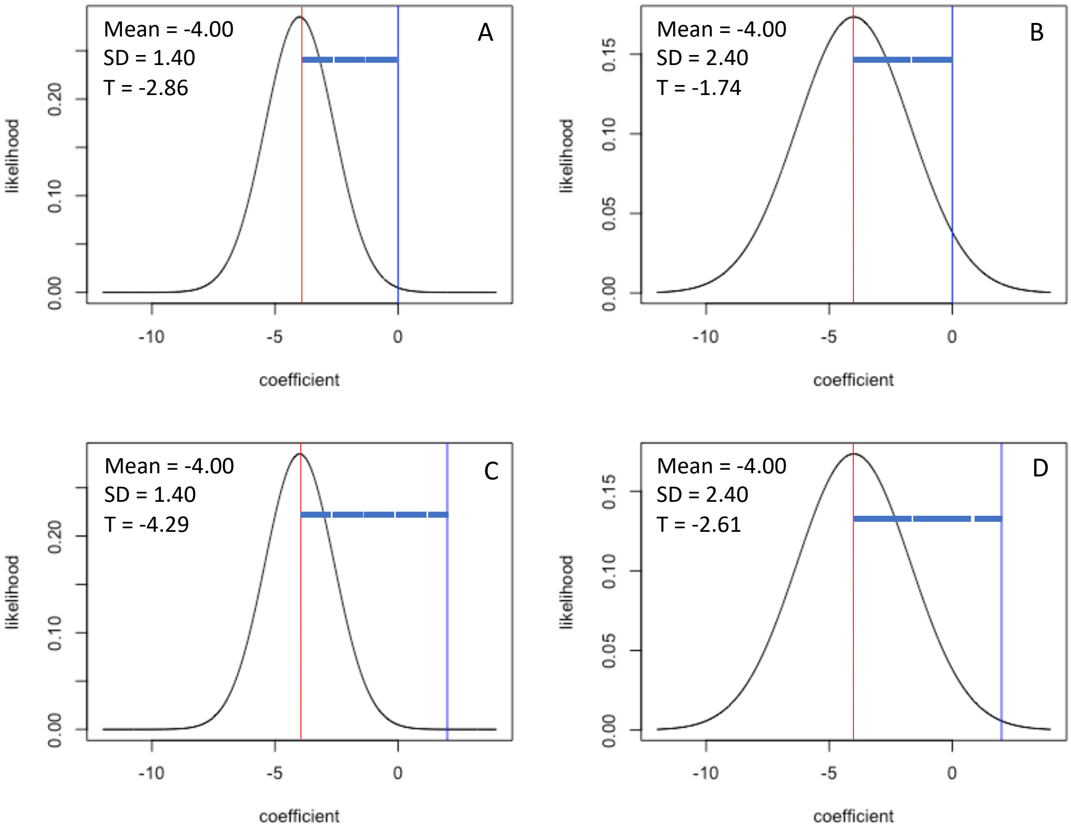


Figure 17.4. A coefficient with a mean = -4.0 and: A) SD = 1.4, and so (-4 - 0)/1.4 = -2.86 SDs from zero and significantly different from the null hypothesized value of 0 (p = 0.004). The dashed thick horizontal blue line indicates how many SDs there are between the mean of the distribution of the coefficient and the value assumed by the null hypothesis; B) SD = 2.4, and so (-4 - 0)/2.4 = -1.74 SDs from zero and not significantly different from the null hypothesized value of 0 (p = 0.082); C) SD = 1.4 and testing the null hypothesis that the coefficient is different to 2, and so (-4 - 2)/1.4 = -4.29 SDs from 2 and therefore significantly different from the null hypothesized value of two (p < 0.001); D) SD = 1.4, and testing the null hypothesis that the coefficient is different to 2, and so (-4 - 2)/2.4 = -2.61 SDs from two and therefore significantly different from the null hypothesized value of two (p = 0.009). Here we have assumed the degrees of freedom associated with the T test to be very large. If the degrees of freedom are modest, then the coefficients will have to be further (more SDs away) from the null value to achieve the same level of significance.

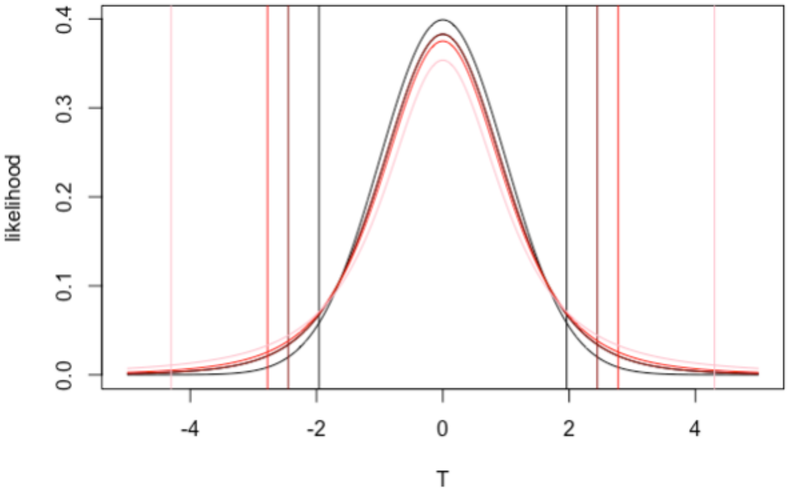


Figure 17.5. Four T distributions with 10,000 (black), 6 (dark red), 4 (red), and 2 (pink) degrees of freedom. With a large number of degrees of freedom the T distribution converges on a Standard Normal distribution (mean = 0, standard deviation =1) where the 2.5th and 97.5th quantiles are ±1.96, but as the degrees of freedom reduces to (say) 6, 4 and 2, the tails thicken and this quantile range expands ±2.45, ±2.78 and ±4.30, as indicated by the vertical lines. This makes some sense – the fewer data you have, the bigger the differences need to be to be confident they are real.

Our more general point is that we can calculate test statistics from data, and knowing how they should be distributed under the assumption of our null hypothesis, use the position of the observed test statistic in the expected distribution of the test statistic – to reject (or fail to reject) our null hypothesis.

There are lots of different test statistics that derive from lots of different possible distributions, but the principles governing their use is much like these two examples.

17.3 One or two tails?

You may have clocked that in Fig. 17.1 (χ^2 statistic) we drew one blue line (at the 95th quantile), and in Fig 17.3 (T statistic) we drew two, at the (at the 2.5th and 97.5th quantiles). Why two lines? Because the T statistic could have been positive or negative (in principle the mean from which our 10 numbers came from in the second example could have been less than zero or greater than zero), and thus could have fallen in the left-hand or the right-hand tail. Thus, the test is two-tailed, and we are alert for deviations to the left or the right, and leave 95% in the middle. Had our question been: is the observed mean less than zero (of greater than zero) we'd have drawn just one blue line at the 5th quantile and asked if the observed test statistic was to the left of the line (or the 95th quantile and asked if the observed test statistic was to the right of the line). Most T tests are by default two tail. But it's something to watch out for. Why one line in Fig 17.1? Because we are interested in whether our test statistic (the residual deviance) is too large, and not 'too large *or* too small'.

In Appendix G.2 we discuss how to work in R with distributions in the way discussed in examples 1 and 2, although most packages will generate p-values for you.

Important ideas to take-away

- Most statistical tests work by manipulating data in some way to generate test statistics that under the assumption the null hypothesis applies, have well understood distributions. This makes it possible to determine if the test statistic is unusually large or small. If it is – we can conclude the data we have are not consistent with the null hypothesis, and we for the time being at least – adopt the alternate hypothesis.
- We use p-values to make this call. The p-value is the probability of finding the observed, or more extreme value of the test statistic, on the assumption the null hypothesis applies.
- If the p-value is less than 0.05 we generally consider this grounds for rejecting the null hypothesis.
- The two most common statistics we will encounter are the χ^2 and T statistics.

Chapter 18

Frequentist or Bayesian statistics?

(back to Contents)

It is useful at this stage to briefly recognize we are only describing one statistical perspective in this text – the frequentist approach. It is a very common one indeed, widely applicable and effective, and the one most people start with, but there is another 'school' – the Bayesian approach, that is arguably superior in some respects. Here we just briefly compare and contrast.

There are two different approaches to inference, known as Bayesian and Frequentist. Part 2 of this text is only going to introduce and discuss the frequentist approach, but it is worth being aware of the difference.

We have introduced the idea of fitting a model that maximizes the likelihood of the *data given the model*. One way of viewing the coefficients of a model is that they embody hypotheses. For example, a slope measures the relationship between the response and explanatory variable. When we are sufficiently confident the magnitude of a slope is really different to zero, we may conclude that our data support a relationship between the response and explanatory variable. In fact, the existence of each coefficient in a model can be viewed as setting up two hypotheses: a null hypothesis that the coefficient is not different to zero, and an alternative hypothesis that it does differ from zero (together with whatever interpretation we place on that outcome). The null hypothesis is assessed using a test statistic, and an associated p-value that reflects the probability of there being a relationship at least as strong or stronger than that observed in the data, assuming the null hypothesis to apply. Thus, an inference is made based on the likelihood of the data given the model. Or put another way, the likelihood of our data given our null hypothesis. The p-value basically says … if we were to generate thousands of independent versions of our data, a certain proportion (or *frequency*) of the time the data would be consistent with our null hypothesis. If that frequency is very low, we reject the null hypothesis. Hence the name: **frequentist statistics**.

The idea that we would assess the support for a general hypothesis based on the likelihood of our particular dataset is a bit peculiar. We don't really want to know how likely our data are, we want to know how likely our hypothesis is. And this is not what the frequentists p-value tells you. It isn't completely bonkers, but it is a bit backwards and indirect. A similar thought occurred to the clergyman Thomas Bayes around 1750. Bayes Theorem showed how it was possible to turn this inference around and calculate the likelihood of a hypothesis given data. This seems a lot more generally useful than the likelihood of a data set given a hypothesis, and proponents of so-called **Bayesian statistics** approach data analysis from this different perspective.

It's important to point out that Bayesian approaches can be applied to the same types of models as those we'd work with using a frequentist approach, but they are fitted and assessed in slightly different ways (the R package `MCMCglmm` fits GLMs using a Bayesian approach).

Aside from perhaps a more logical approach, Bayesians takes a more nuanced approach to what the models really tell us. Bayesians don't test hypotheses, and they don't work with p-values. Instead, they focus on the estimates of particular parameters of interest, and how credible these estimates are. Methodologically the Bayesian approach has various advantages: it introduces a handy way of introducing pre-existing knowledge around the parameters through the use of prior distributions that can be more or less informative, depending on the quality of prior knowledge about a parameter. Bayesian analysis makes fewer assumptions about how a fitted model parameter may be distributed, and there is an extensive range of clever machinery (**Monte Carlo Markov Chains** or **MCMC**) available for fitting more complicated models (**hierarchical models**) where more or less data may be missing. However, they are more complicated to work with while being ultimately a more powerful approach.

There is a great deal more that could be said about these two schools of statistics (there is a good lecture on this subject here). You could go through your whole life as a researcher and use only one or the other of these approaches ... it would be perfectly fine and indeed quite common. However, it is usual to develop familiarity with the basic ideas in the frequentist context, and then decide for yourself which approach is best suited to your needs. So we'll continue with our exposition of the frequentist approach

Important ideas to take-away

- There are two importantly different schools of statistics – frequentist and Bayesian
- They have different philosophical underpinnings, and while both approaches can be used to fit simple models to data, the fitting process is different, and interpretation of results different also
- Both schools are perfectly legitimate, and commonly used
- Students mostly start with a Frequentist approach and may or may not decide to explore the Bayesian approach if they develop serious interests in data analysis.

Chapter 19

Likelihood Ratio Tests

(back to Contents)

Likelihood ratio tests or LRTs can be used in a couple of different ways in our account of GLMs, but here we just introduce the concept of the test, and provide more examples and context in Chapters 20 and 21. You'd be best reading Chapter 17 if you haven't already.

As frequentists, there are two ways we make inferences from our models.

One is to compare two models that are identical except that one includes additional explanatory variables or interactions of interest, and the other which does not. The *relative* performance (as assessed by their respective *likelihoods*) of these models in explaining the variation in our response variable tells us something about the importance of the variable(s) of interest that is (are) absent in the simpler model. For example, if a model with just Landscape in it performs much worse than a model with both Landscape and Nitrate in it, we can infer Nitrate is going to be an important part of our explanation of variation in Chlorophyll concentration. Conversely, if a model comprising – say – both our categorical explanatory variables Flow and Landscape doesn't explain the variation in Chlorophyll any better than a model with just Landscape in it ... obviously Flow isn't contributing much to the explanatory power of the model.

The other is to examine the *coefficients* associated with individual terms in *one* particular model. For example, if we are confident that the slope associated with Nitrate is very unlikely to be zero, then every unit increase in Nitrate will almost certainly be influencing chlorophyll concentration. Conversely, if the adjustments for levels of Flow are not distinguishable from zero then it doesn't matter what the flow is, there won't be a discernible influence on Chlorophyll concentration.

These two approaches usually give comparable answers, but not always. And where they differ will mostly be when the study wasn't designed to be quite sufficiently powerful enough to detect the influences of the explanatory variables it was designed to investigate. The first relative approach we'll call **model comparison**, and is more robust to collinearity, and the choice of reference level when modelling the influence of categorical explanatory variables with more than two levels. So we start with this. But in Chapter 22, we'll look at **coefficient analysis**. In practice, we will in any case need to think about both the significance of the terms and the direction and magnitude of the effects as modelled by the coefficients.

19.1 Likelihood ratio tests (LRTs)

Likelihood ratio tests are an essential part of the inference tool kit. We use LRTs to compare two *models fitted to the same observations of the response variable*, one that includes a particular term (i.e. a main effect or an interaction) on the right-hand side, and one that doesn't. Thus, we have what we call a (more) complex model (M_c) that includes one or more particular terms, and a simpler model (M_s) that does not. If the particular term(s) is (are) not useful in explaining variation in the response variable, then the simpler model will fit the data essentially as well as the more complex model, and we should select the simpler model as it is a simpler and equally good explanation of the data as the more complex model. Thus, we have a null hypothesis that the two models are essentially equally effective at explaining the variation in the response variable, and an alternative hypothesis that the more complex model does a better job.

We can count-up the degrees of freedom required by each of these models (df_c and df_s respectively, see Chapter 14 if you are not familiar with how to do this). We can fit both models to the data, and compute the log-likelihood of the data given each of the models (LL_c and LL_s respectively). We can then compute twice the difference between these two log likelihoods ($2\Delta LL$), which is our likelihood ratio test statistic, which happens to be approximately χ^2 distributed with the number of degrees of freedom by which the two models differ in the degrees of freedom each requires (df_c - df_s). Thus:

$$2\Delta LL = 2\ (LL_c - LL_s)$$

may be assumed to be χ^2 distributed with (df_c - df_s) degrees of freedom.

If $2\Delta LL$ is sufficiently large, i.e. there is a big difference between the likelihood of the data given the model with and without the particular term(s), we'll reject the null hypothesis that the two models are essentially as good as each other, and accept the alternate hypothesis that the complex model is better – strongly suggesting the particular term(s) must matter and we'd want to retain it (them).

We can add a little narrative to this technical description. A model with more coefficients will *always* lead to the data being more likely than a model with less coefficients. So we *know* that LL_c will always be more positive (or 'less negative') than LL_s (remember that smaller negative ('more positive') log likelihoods correspond to higher likelihoods). This is because even if there isn't a real effect for our particular term to capture, it will be used to account for some of the otherwise unexplained (residual) variation. The real question we are asking is: is the additional complexity, that is – the additional coefficients in our more complex model, increasing the likelihood of the data *enough* to warrant their retention in the model?

19.2 Nested models – a requirement of LRTs

To compare two models using an LRT they must be nested. A model is nested within another if the simpler model is a subset of the more complex one, i.e. we need to be able to construct the simpler model merely by zeroing out terms (main effects or interactions) in the more complicated model.

These pairs of models (more complex first, simpler second - after removal of the crossed-out term) are nested:

i. Chlorophyll~Landscape+Flow+Nitrate+Phosphate+Temp
Chlorophyll~Landscape+Flow+Nitrate+Phosphate~~+Temp~~

ii. Chlorophyll~Landscape+Flow+Nitrate+Phosphate
Chlorophyll~Landscape+Flow~~+Nitrate~~+Phosphate

iii. Chlorophyll~Landscape+Flow+Nitrate+Phosphate+Temp
Chlorophyll~Landscape+Flow+Nitrate~~+Phosphate~~+Temp

iv. Chlorophyll~Nitrate+Phosphate+Temp
Chlorophyll~~~Landscape+Flow+Nitrate+Phosphate+~~Temp

v. Chlorophyll~Landscape+Flow
Chlorophyll~Landscape~~+Flow~~

vi. Chlorophyll~Landscape+Flow
Chlorophyll~~~Landscape~~+Flow

vii. Chlorophyll~Landscape+Flow+Landscape:Flow
Chlorophyll~Landscape+Flow~~+Landscape:Flow~~

viii. Chlorophyll~Nitrate+Phosphate+Temp+Phosphate:Temp+Nitrate:Temp
Chlorophyll~Nitrate+Phosphate+Temp~~+Phosphate:Temp~~+Nitrate:Temp

The following are not nested (note the terms in red in the second model that are not present in the first model):

ix. Chlorophyll~Flow+Nitrate+Phosphate
Chlorophyll~Landscape+Nitrate+Phosphate

x. Chlorophyll~Landscape+Flow
Chlorophyll~Nitrate+Phosphate

xi. Chlorophyll~Flow+Nitrate+Phosphate+Temp
Chlorophyll~Landscape+Temp

xii. Chlorophyll~Nitrate+Phosphate+Temp+Phosphate:Temp
Chlorophyll~Nitrate+Phosphate+Temp+Nitrate:Temp

(This concept of 'nestedness' is distinct from the use of nested when talking about certain types of mixed model that are sometimes called hierarchical models.)

19.3 LRTs and interactions

Interactions can be treated just like any other term. We can compare (nested) models with and without interactions. The only complication is that in order to compare a model with and without an interaction, both models *must* contain the main effects that comprise the interaction (as in examples *vii* and *viii* above). In fact, we suggest you don't even construct models that include interactions without having the terms of the interaction represented as main effects. This will have some consequences for how we go about applying LRTs in Chapters 20 and 21.

19.4 LRTs are very general

LRTs can be applied to any pair of nested models (GLMs or indeed any model fitted by maximum likelihood) that fit the same distribution to the same response variable. *We strongly suggest that you only compare models that differ by one term.* Otherwise, it becomes difficult to determine which terms are causing the difference.

19.5 Limitations of LRTs

Strictly speaking, the distribution of the Likelihood Ratio Test statistic is only χ^2 distributed when you have a very large number of observations of the response variable. For smaller data sets, and particularly in mixed models, the distribution of the test statistic applied to fixed effects is only approximately χ^2 distributed, and this can lead to p-values from your LRTs being smaller than they 'should be'. Thus, inference is 'safe' if the LRT fails to reject the null hypothesis, but if it narrowly rejects the null hypothesis you'd want to proceed cautiously. There isn't all that much you can do about this. There are more complex approaches to inference in these situations but they are beyond the scope of this text (these issue are reviewed in Bolker et al 2009).

19.6 Example of an LRT

Supposing we had decided that our most complex plausible model was:

```
> M_mcpm <-glm(Chlorophyll~Landscape+Flow, data= my_data)
```

The algebraic structure would be:

$$f_i = c + \alpha_j + \beta_k \qquad \text{(model 19.1)}$$

$$i = 1..48; j = R, U; k = L, M, H$$

In R we can access the log-likelihood of the data given this model with the command:

```
> logLik(M_mcpm)
'log Lik.' -206.9627 (df=5)
```

The model requires 5 degrees of freedom (1 for *c*, 1 for Landscape, 2 for Flow, and one for the variance of the Normal distribution used to model the residual variation) and the log-likelihood of the data given this model is -206.96 (thus the likelihood is 1.159082e-92, a very small number indeed!).

Suppose we wanted to test whether the categorical explanatory variable Flow was contributing significantly to the likelihood of the observations of Chlorophyll concentration. We could construct a simpler model:

```
> M_s<-glm(Chlorophyll~Landscape, data= my_data)
```

The algebraic structure would be:

$$f_i = c + \alpha_j \qquad \text{(model 19.2)}$$

$$i = 1..48; j = R, U$$

The log-likelihood of the data given this model is:

```
> logLik(M_s)
'log Lik.' -211.6902 (df=3)
```

Note that the simpler model only requires 3 df, because we've discarded two coefficients representing the 3 different levels of Flow. We have:

$$2\Delta LL = 2\ (LL_c - LL_s)$$

$$2\Delta LL = 2\ (-206.9627 - -211.6902) = 9.455$$

is χ^2 distributed with (df_c - df_s) = (5 – 3) = 2 degrees of freedom. Fig. 19.1 shows how this observed value of the test statistic is greater than the 95th quantile of a χ^2 distribution with 2 dfs (which is 5.99) and that therefore we reject the null hypothesis that the more complex and simpler model is equally good at explaining the data, and adopt the more complex one.

The procedure generates a p-value that we can cite in support of the significance of Flow – the only term by which the simpler and more complex model differ. This can be calculated from R using a member of the p_ functions (`pchisq`) (see Appendix G.2, Fig. G.2).

```
> 1-pchisq(9.455,2)
[1] 0.008848565
```

It is the area to the right of the red arrow.

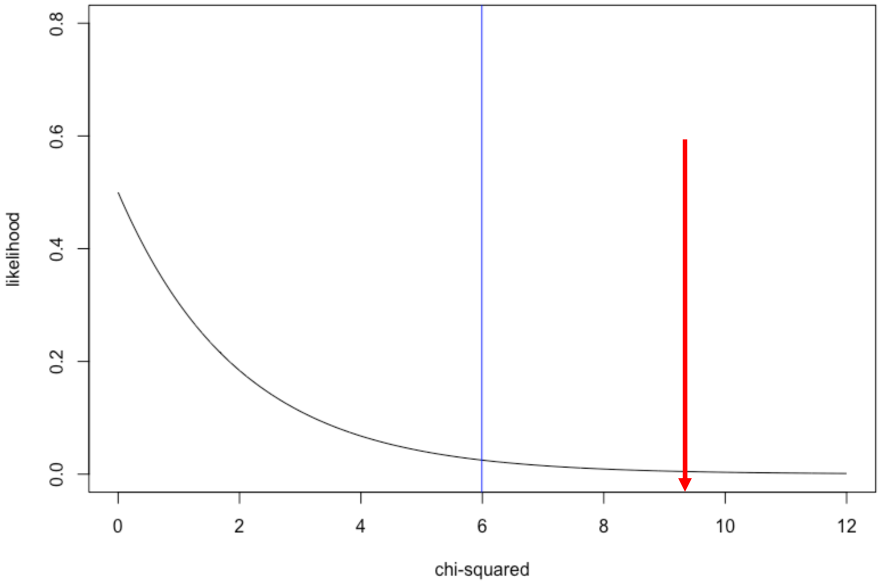


Figure 19.1. A χ^2 distribution with 2 dfs, showing the position of the 95th quantile (5.99, blue line), and the observed value of the test statistic (9.455, red arrow). The p-value for this test is 0.0088 and equates to the area to the right of the red arrow.

These are the bare bones of LRTs. We'll encounter a lot more of them in the next two chapters.

There are many different ways of performing LRTs in R. One useful package is the `lrtest` command in the `lmtest` package:

```
> lrtest(M_mcpm,M_s)
Likelihood ratio test

Model 1: Chlorophyll ~ Landscape + Flow
Model 2: Chlorophyll ~ Landscape
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   5 -206.96
2   3 -211.69 -2 9.4551   0.008848 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Important ideas to take-away

- A powerful way to learn what terms in our model are important and which are not is to compare two models, identical except that one includes an explanatory variable or interaction of interest, and the other which does not. The *relative* performance of these models in explaining the (same) variation in our response variable tells us something about the importance of the variable of interest that is absent in the simpler model.
- The analysis of relative performance involves comparing the likelihoods of the data given the two models using a likelihood ratio test (LRT), which generates a test statistic that has a χ^2 distribution. The test statistic is calculated as twice the absolute difference in the log-likelihoods of the data given the two models.
- LRTs can be used to compare any two models so long as they are fitted to the same response variable using the same distribution for the residual variation, and the simpler model is nested within the more complex one.
- An LRT tests the null hypothesis that the simpler and more complex model are indistinguishable in accounting for the observed variation in the response variable. If we can’t reject the null hypothesis we adopt the simpler model (because its more parsimonious). If we reject the null hypothesis, we conclude the complex model is a better explanation of the data, and that the term by which the two models differ must be significant.
- In practice, we’ll need to think about both the significance of the terms and the direction and magnitude of the effects as modelled by the coefficients.

Chapter 20

Using Likelihood Ratio Tests to make inferences from a final model

(back to Contents)

If you have fitted your most complex plausible model you can use LRTs to determine the importance of each of the terms in it.

If you have fitted your **most complex plausible model** you can use LRTs to assess the importance of each term is our recommended way to make inferences from it.

The logic is exactly as described in Chapter 19: if we compare two models and they perform indistinguishably well in account for the likelihood of the data (the null hypothesis) then we prefer the simpler one through the usual reasoning of parsimony. If the models are distinguishable then it will be the more complex model that performs better, and the additional term in the complex model can be judged to be making a statistically significant contribution to explaining the likelihood of the observations of the response variable. This is pre-requisite to determining whether we need to inspect the direction and magnitude of the effect.

20.1 An example of applying LRTs to a final model

Supposing that as before we had decided that our most complex plausible model was:

```
> M_mcpm<-glm(Chlorophyll~Landscape
                        +Flow
                        +Phosphate
                        +Nitrate
                        +Temp
                        +Phosphate:Landscape
                        +Nitrate:Flow,data=my_data)
```

The algebraic structure would be:

$$f_i = c + \alpha_j + \beta_k + (m_P + \gamma_j)\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 20.1)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

Here α_j represents landscape, β_k represents Flow, and γ_j and δ_k represent the adjustments to the relationship between Phosphate ($x_{P,i}$) and Nitrate ($x_{N,i}$) arising from their respective interactions with Landscape and Flow. We can access the log-likelihood of the data given this model with the command:

```
> logLik(M_mcpm)
'log Lik.' -140.0671 (df=11)
```

The model requires 11 degrees of freedom (1 for *c*, 1 for Landscape, 2 for Flow, one each of the slopes for Nitrate, Phosphate and Temperature, 1 for the adjustment to the slope governing the influence of Phosphate for different levels of Landscape, 2 for the adjustment to the slope governing the influence of Nitrate for 3 different levels of Flow, and one for the variance of the Normal distribution used to model the residual variation) and the log-likelihood of the data given this model is -140.07.

We start by testing the interactions. Is the interaction of Landscape and Phosphate making a significant contribution to explaining the variation in Chlorophyll? We create a simpler (nested) model by removing the interaction:

```
> M_s<-glm(Chlorophyll~Landscape
                        +Flow
                        +Phosphate
                        +Nitrate
                        +Temp
                        +Nitrate:Flow,data=my_data)
```

The algebraic structure would be:

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 20.2)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

We answer this question by conducting an LRT. The log-likelihood of the simpler model is:

```
> logLik(Ms)
'log Lik.' -140.5277 (df=10)
```

Note that the simpler model only requires 10 df, because we've removed the adjustment to the slope representing Phosphate dependent on Landscape. We have:

$$2\Delta LL = 2\ (LL_c - LL_s)$$

$$2\Delta LL = 2\ (-140.07- -140.53) = 0.92$$

is χ^2 distributed with (df_c - df_s) = (11 – 10) = 1 degree of freedom.

As before, the test statistic doesn't look very big … and indeed it is well inside the 95th quantile of a χ^2 distribution with 1 df (Fig. 19.1)

We observe that our test statistic of 0.92 is not larger than we would expect given the null hypothesis that the simple and complex models are equally effective at explaining the variation in our response variable, and so we fail to reject our null hypothesis, and adopt for the time being at least the simpler model as the more parsimonious explanation of the data. Or – to put it another way – the interaction between Phosphate and Landscape isn't helpful in explaining variation in our response variable.

We can generate a specific p-value for the null hypothesis the two models are equivalent:

```
> 1-pchisq(0.92,1)
[1] 0.337475
```

Confirming that we don't reject the null hypothesis. (Check Appendix G.2 (Fig. G.2) to see how the `p_` command works.)

We will discard the interaction between Phosphate and Landscape so that we can test the main effects. So, we now have as our new most complex plausible model:

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 20.2)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

We try dropping the other interaction, Nitrate with Flow:

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + m_N\, x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 20.3)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

```
> M_mcpm<-glm(Chlorophyll~Landscape+Flow+Nitrate+Phosphate+Temp
                        +Nitrate:Flow,data=my_data)
> logLik(M_mcpm)
'log Lik.' -140.5277 (df=10)
> M_s<-glm(Chlorophyll~Landscape+Flow+Nitrate+Phosphate
                        +Temp,data=my_data)
> logLik(M_s)
'log Lik.' -163.3734 (df=8)
```

and

$$2\Delta LL = 2\ (LL_c - LL_s)$$

$$2\Delta LL = 2\ (-140.53 - -163.37) = 45.68$$

is χ^2 distributed with $(df_c - df_s) = (10 - 8) = 2$ degrees of freedom.

This test statistic is large, and considerably to the right of the 95th quantile of 5.99 for the χ^2 distribution with 2 df (Fig. 18.2). There are 2 degrees of freedom required for this test because by removing the Nitrate:Flow interaction from the model we removed 2 coefficients (Flow having 3 levels). We thus reject the null hypothesis and conclude the interaction between Nitrate and Flow is highly significant.

As before, we can generate a specific p-value for the null hypothesis these two models are equivalent:

```
> 1-pchisq(45.68,2)
[1] 1.204242e-10
```

Thus, we still have as our most complex plausible model:

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 20.2)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

and we can move on to the main effects that are not represented in the retained interactions. We try dropping Temperature first:

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} \qquad \text{(model 20.3)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

```
> M_mcpm<-glm(Chlorophyll~Landscape+Flow+Nitrate+Phosphate+Temp
                        +Nitrate:Flow,data=my_data)
> logLik(M_mpcm)
'log Lik.' -140.5277 (df=10)
> M_s<-glm(Chlorophyll~Landscape+Flow+Nitrate+Phosphate
                        +Nitrate:Flow,data=my_data)
> logLik(M_s)
'log Lik.' -141.1152 (df=9)
```

and

$$2\Delta LL = 2\ (LL_c - LL_s)$$

$$2\Delta LL = 2\ (-140.53 - -141.12) = 1.18$$

is χ^2 distributed with $(df_c - df_s)$ = (10 – 9) = 1 degree of freedom.

This value of the test statistic is small – well to the left of the 95th quantile for a χ^2 distribution with 1 df, and so we fail to reject the null hypothesis the two models are equally good, and note that Temperature does not make a significant contribution to explaining variation in Chlorophyll concentration.

And the p-value would be:

```
> 1-pchisq(1.18,1)
[1] 0.277356
```

We'll keep Temperature in the model because we adopted the most complex plausible model as our final model (barring removal of non-significant interactions):

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 20.2)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

We will not test or consider dropping Flow or Nitrate because they appear in a retained interaction, and so this leaves only Phosphate or Landscape to drop. We drop Phosphate:

$$f_i = c + \alpha_j + \beta_k + (m_N + \delta_k)\, x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 20.4)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

```
> M_mcpm<-glm(Chlorophyll~Landscape+Flow+Nitrate+Phosphate+Temp
                        +Nitrate:Flow,data=my_data)
> logLik(M_mcpm)
'log Lik.' -140.5277 (df=10)
> Ms<-glm(Chlorophyll~Landscape+Flow+Nitrate+Temp
                        +Nitrate:Flow,data=my_data)
> logLik(Ms)
'log Lik.' -143.5805 (df=9)
```

and

$$2\Delta LL = 2\ (LL_c - LL_s)$$

$$2\Delta LL = 2\ (-141.12 - -146.04) = 6.11$$

is χ^2 distributed with (df_c - df_s) = (10 – 9) = 1 degree of freedom.

This value of the test statistic is large – well to the right of the 95th quantile for a χ^2 distribution with 1 df (check back to Fig. 19.1). We can confidently reject the null hypothesis the two models are equally good, and conclude Phosphate contributes significantly to explaining the variation in Chlorophyll concentration.

The p-value would be:

```
> 1-pchisq(6.1056,1)
[1] 0.01347542
```

Indicating the rejection of the null hypothesis.

Our most plausibly complex model remains the same (we are not dropping terms unless they are non-significant interaction terms):

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 20.2)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

and we have only Landscape to drop:

$$f_i = c + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 20.5)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

```
> M_mcpm<-glm(Chlorophyll~Landscape+Flow+Nitrate+Phosphate+Temp
                        +Nitrate:Flow,data=my_data)
> logLik(M_mcpm)
'log Lik.' -140.5277 (df=10)
> Ms<-glm(Chlorophyll~Flow+Nitrate+Phosphate+Temp
                        +Nitrate:Flow,data=my_data)
> logLik(Ms)
'log Lik.' -151.3979 (df=9)
```

and

$$2\Delta LL = 2\ (LL_c - LL_s)$$

$$2\Delta LL = 2\ (-140.5277 --151.3979) = 26.79$$

is χ^2 distributed with (df_c - df_s) = (10 – 9) = 1 degree of freedom.

This value of the test statistic is large – well to the right of the 95th quantile for a χ^2 distribution with 1 df (check back to Fig. 20.1) and so we can also confidently reject the null hypothesis these two models are equally good, and conclude Landscape contributes significantly to explaining the variation in Chlorophyll concentration.

The p-value testing the null hypothesis would be:

```
> 1-pchisq(21.7404,1)
[1] 3.121479e-06
```

Indicating that once again we emphatically reject it.

We could summarize these findings in a table:

Table 21.1. Summarizing the terms tested using LRT in the most complex plausible model. The degrees of freedom indicate the number of coefficients required to represent each of the terms.

Term	$2\Delta LL$	df	p
Phosphate:Landscape	0.92	1	0.337
Nitrate:Flow	45.68	2	< 0.001
Temperature	1.18	1	0.277
Phosphate	6.11	1	0.013
Landscape	26.79	1	< 0.001
Flow	Term not tested		
Nitrate	Term not tested		

This is a key point often not understood:

We don't test the significance of Flow and Nitrate as main effects because they are present in a significant interaction. *Main effects that are present in significant interactions must be significant.* If the effect of (say) Nitrate on Chlorophyll concentration depends significantly on Flow, then how can Flow not be significant?

When reporting the results of an LRT you should cite the test statistic, the degrees of freedom and the p-value. All three are important to report (see Chapter 23).

We can summarize the steps described here in the following work flow:

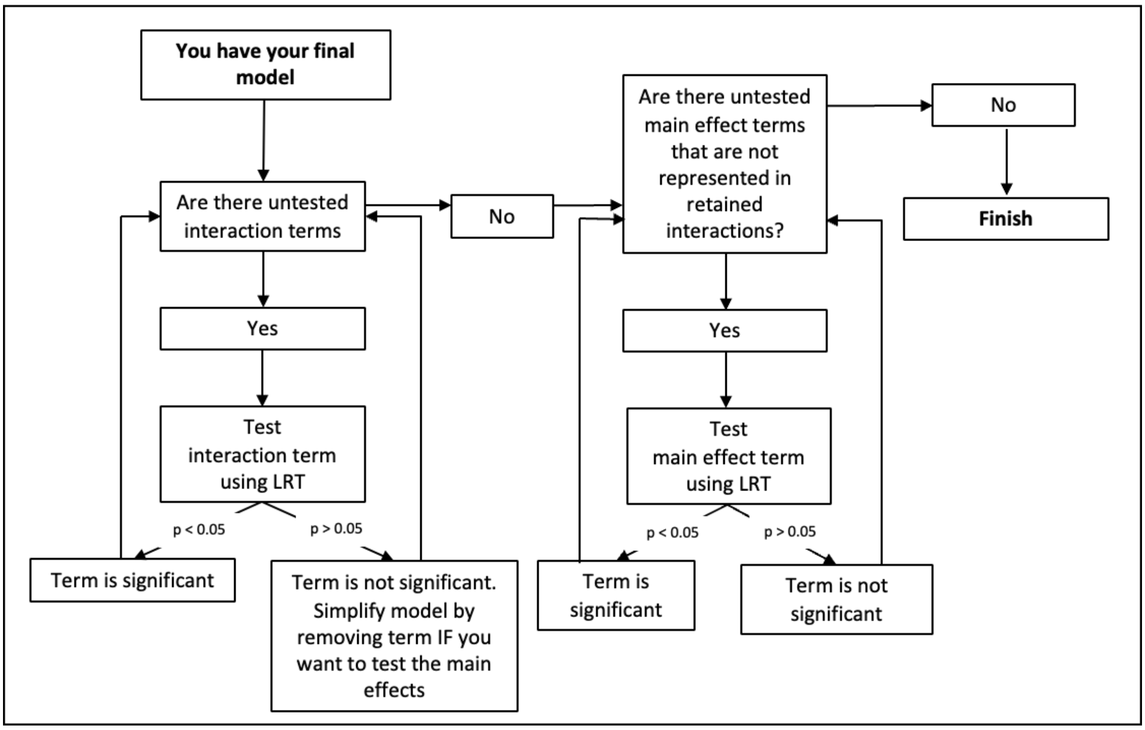

You have your final model
Are there untested interaction terms
No
Yes
Test interaction term using LRT
p < 0.05
p > 0.05
Term is significant
Term is not significant. Simplify model by removing term IF you want to test the main effects
Are there untested main effect terms that are not represented in retained interactions?
No
Finish
Yes
Test main effect term using LRT
p < 0.05
p > 0.05
Term is significant
Term is not significant

Figure 20.1. The process for using LRTs to make inferences on a final model. *Note that all terms are retained unless they are non-significant interactions.*

20.2 The ordering of the testing sequence

Aside from testing interactions first, the ordering of testing the terms in the final model generally doesn't matter as we don't remove the terms that are revealed not to be significant (unless they are interactions).

20.3 Random effects

In general, we suggest simply leaving random effects in models, and not testing them for significance. Random effects would be included in the first place because of concerns that there may be otherwise unrecognized dependencies between observations of the response variable, and whether they are significant or not, it is still a good idea to account for such dependencies. A random effect in any case only requires only 1 degree of freedom, so the cost in terms of degrees of freedom is modest.

In the special case that you are fitting mixed models assuming normally distributed response variables you may have the option of fitting your model using 'full' maximum likelihood (ML), or 'restricted' maximum likelihood (REML). In the `lmer` command `REML` is set to `TRUE` by default. `REML = TRUE` generates more accurate estimates of the random effect variances, but if you want to compare two models using an LRT you must set `REML = FALSE`.

Instead of using LRTs you may choose to use AIC to compare different models. The basic ingredients of LRTs and AIC are essentially identical, but in our view LRTs are slightly more exact and better justified. However, AIC is very commonly seen in the literature and we describe it briefly in Appendix U. It has particular value if for some reason you need to compare models that are not nested.

Important ideas to take-away

- However you have arrived at your final model (whether through ‘your first model is your final model’, or model selection – see the following chapter), we recommend testing the significance of each of the terms using likelihood ratio tests
- It doesn’t matter what sort of GLM you have fitted you can always use LRTs to compare any two models so long as they are fitted to the same response variable using the same distribution for the residual variation, and the simpler model is nested within the more complex one
- Testing starts with the interactions, and if the interaction is shown not to be significant and you wish to test the significance of the main effects it will be necessary to remove the non-significant interaction from the model.
- You can’t test the main effects that also occur in significant interactions. Main effects that occur in significant interactions are *de-facto* regarded as significant. No testing is required or is even appropriate
- We recommend not testing the significance of random effects unless you have a very specific reason for doing so

Chapter 21

Model selection

(back to Contents)

Model selection could be used to reduce a most complex plausible model to a most complex minimal model through a sequential series of likelihood ratio tests that retain influential explanatory variables and remove those that don't explain significant variation. You may very well choose not to conduct model selection and base your inference on the most complex plausible model – in which case you can skip to Chapter 22.

In Chapter 15 we discussed how we might choose the model we fit to our response variable. Having fit what we call the **most complex plausible model** and examined in more detail what matters and what perhaps doesn't, might you wish to simplify it? If there turns out to be explanatory variables in your model that are not important you could consider removing them. This is a process called **model selection**.

You may or may not choose to undertake model selection – the pros and cons are addressed in Chapter 15, but if you did, and you chose to use LRTs to do so, this is how you might do it. If you are not interested in model selection you could move straight to Chapter 22.

Model selection is a process that starts with the most complex plausible model that includes all the terms (main effects and interactions) you think ought to be in the model and applies a sequence of checks to determine whether each term can be judged significant based on an LRT. Following each test, the term is either retained or removed. The process is similar to that described in the previous chapter but this time *all* terms that are not significant are removed. After the process of model selection is completed *you then have to test all the retained terms a final time* (go through the Chapter 20 protocol). The process is summarized in Figure 21.1, and illustrated by the example in the following section.

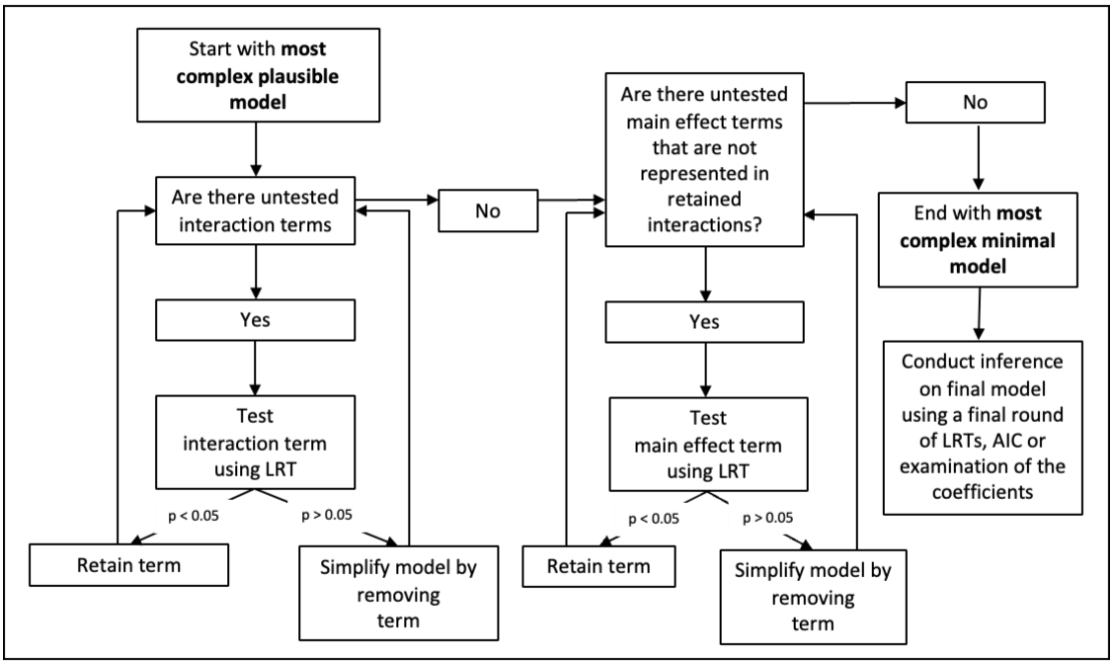

Start with most complex plausible model
Are there untested interaction terms
No
Yes
Test interaction term using LRT
p < 0.05
p > 0.05
Retain term
Simplify model by removing term
Are there untested main effect terms that are not represented in retained interactions?
No
Yes
Test main effect term using LRT
p < 0.05
p > 0.05
Retain term
Simplify model by removing term
End with most complex minimal model
Conduct inference on final model using a final round of LRTs, AIC or examination of the coefficients

Fig 21.1. The **stepwise-backward** process of model selection using LRTs.

21.1 An example of model selection using LRTs

Supposing that based on our research questions and understanding of the biology we had decided that our most complex plausible model was:

```
> M_mcpm<-glm(Chlorophyll~Landscape
                        +Flow
                        +Phosphate
                        +Nitrate
                        +Temp
                        +Phosphate:Landscape
                        +Nitrate:Flow,data=my_data)
```

The algebraic structure would be:

$$f_i = c + \alpha_j + \beta_k + (m_P + \gamma_j)\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 21.1)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

Here α_j represents landscape, β_k represents Flow, and γ_j and δ_k represent the adjustments to the relationship between Phosphate ($x_{P,i}$) and Nitrate ($x_{N,i}$), arising from their respective interactions with Landscape and Flow. We can access the log-likelihood of the data given this model with the command:

```
> logLik(M_mcpm)
'log Lik.' -140.0671 (df=11)
```

The model requires 11 degrees of freedom (1 for *c*, 1 for Landscape, 2 for Flow, one each of the slopes for Nitrate, Phosphate and Temperature, 1 for the adjustment to the slope governing the influence of Phosphate for different levels of Landscape, 2 for the adjustment to the slope governing the influence of Nitrate for 3 different levels of Flow, and one for the variance of the Normal distribution used to model the residual variation) and the log-likelihood of the data given this model is -140.07.

We start by testing the interactions. Is this model better without the interaction of Landscape and Phosphate?

```
> M_s<-glm(Chlorophyll~Landscape
                       +Flow
                       +Phosphate
                       +Nitrate
                       +Temp
                       +Nitrate:Flow,data=my_data)
```

The algebraic structure would be:

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 21.2)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

We answer this question by conducting an LRT. The log-likelihood of the simpler model is:

```
> logLik(Ms)
'log Lik.' -140.5277 (df=10)
```

Note that the simpler model only requires 10 df, because we've discarded the adjustment to the slope representing Phosphate dependent on Landscape. We have:

$$2\Delta LL = 2\ (LL_c - LL_s)$$

$$2\Delta LL = 2\ (-140.07- (-140.53)) = 0.92$$

is χ^2 distributed with $(df_c - df_s)$ = (11 – 10) = 1 degree of freedom.

The test statistic doesn't look very big … and indeed it is well (well) inside the 95th quantile of a χ^2 distribution with 1 df (Fig. 21.1).

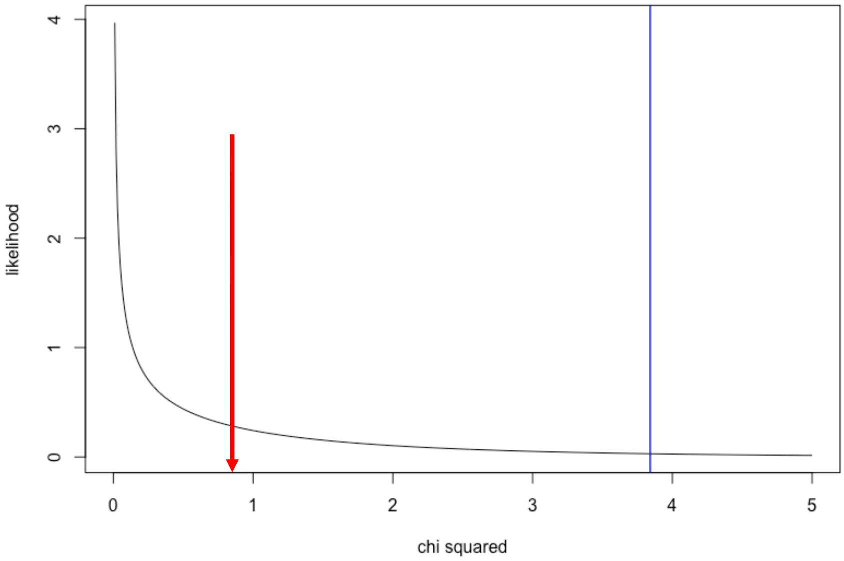


Figure 21.1. A χ^2 distribution with 1 df, with the 95th quantile indicated in blue (at 3.84), and the position of our test statistic in this example indicated by the red arrow, indicating that our test statistic is no different to what would be expected under the null hypothesis that the simpler and more complex models are as good as each other. The test therefore leads us to adopt the simpler model as the more parsimonious explanation of the data.

We observe that our test statistic of 0.92 is not larger than we would expect given the null hypothesis that the simple and complex models are equally effective at explaining the variation in our response variable, and so we fail to reject our null hypothesis, and adopt for the time being at least the simpler model as the more parsimonious explanation of the data. Or – to put it another way – the interaction between Phosphate and Landscape isn't helpful in explaining variation in our response variable.

We can generate a specific p-value for the null hypothesis the two models are equivalent:

```
> 1-pchisq(0.92,1)
[1] 0.337475
```

Confirming that we don't reject the null hypothesis. (Check Appendix G.2 (Fig. G.2) to see how the `p_` function works.)_

Discarding the interaction between Phosphate and Landscape, we now have as our new most plausibly complex model:

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 21.2)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

We try dropping the other interaction, Nitrate with Flow:

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + m_N\ x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 21.3)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

```
> M_mcpm<-glm(Chlorophyll~Landscape+Flow+Nitrate+Phosphate+Temp
                      +Nitrate:Flow,data=my_data)
> logLik(M_mcpm)
'log Lik.' -140.5277 (df=10)
> M_s<-glm(Chlorophyll~Landscape+Flow+Nitrate+Phosphate
                      +Temp,data=my_data)
> logLik(M_s)
'log Lik.' -163.3734 (df=8)
```

and

$$2\Delta LL = 2\ (LL_c - LL_s)$$

$$2\Delta LL = 2\ (-140.53 - (-163.37)) = 45.68$$

is χ^2 distributed with (df_c - df_s) = (10 – 8) = 2 degrees of freedom.

This test statistic is large, and considerably to the right of the 95th quantile of 5.99 for the χ^2 distribution with 2 df (Fig. 18.2). There are 2 degrees of freedom required for this test because by removing the Flow:Nitrate interaction from the model we removed 2 coefficients (Flow having 3 levels). We thus reject the null hypothesis and conclude we should retain the Nitrate:Flow interaction term in the more complex model to explain the data.

As before, we can generate a specific p-value for the null hypothesis these two models are equivalent:

```
> 1-pchisq(45.68,2)
[1] 1.204242e-10
```

Confirmation that we reject the null hypothesis.

Thus, we still have as our most plausibly complex model:

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} + m_T\, x_{T,i} \qquad \text{(model 21.2)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

and we can move on to the main effects that are not represented in the retained interactions. We try dropping Temperature:

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} \qquad \text{(model 21.3)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

```
> logLik(M_mcpm)
'log Lik.' -140.5277 (df=10)
> M_s<-glm(Chlorophyll~Landscape+Flow+Nitrate+Phosphate
                +Nitrate:Flow,data=my_data)
> logLik(M_s)
'log Lik.' -141.1152 (df=9)
```

and

$$2\Delta LL = 2\ (LL_c - LL_s)$$

$$2\Delta LL = 2\ (-140.53 - (-141.12)) = 1.18$$

is χ^2 distributed with (df_c - df_s) = (10 – 9) = 1 degree of freedom.

This value of the test statistic is small – well to the left of the 95th quantile for a χ^2 distribution with 1 df, and so we fail to reject the null hypothesis the two models are equally good, keep the simpler one, and discard Temperature.

And the p-value would be:

```
> 1-pchisq(1.18,1)
[1] 0.277356
```

So, our most plausibly complex model becomes:

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} \qquad \text{(model 21.3)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

We will not test or consider dropping Flow or Nitrate because they appear in a retained interaction, and so this leaves only Phosphate or Landscape to drop. We drop Phosphate:

$$f_i = c + \alpha_j + \beta_k + (m_N + \delta_k)\, x_{N,i} \qquad \text{(model 21.4)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

```
> M_mcpm<-glm(Chlorophyll~Landscape+Flow+Nitrate+Phosphate
                         +Nitrate:Flow,data=my_data)
> logLik(M_mcpm)
'log Lik.' -141.1152 (df=9)
> M_s<-glm(Chlorophyll~Landscape+Flow+Nitrate
                         +Nitrate:Flow,data=my_data)
> logLik(M_s)
'log Lik.' -146.0358 (df=8)
```

and

$$2\Delta LL = 2\ (LL_c - LL_s)$$

$$2\Delta LL = 2\ (-141.12 - -146.04) = 9.84$$

is χ^2 distributed with (df_c - df_s) = (9 – 8) = 1 degree of freedom.

This value of the test statistic is large – well to the right of the 95th quantile for a χ^2 distribution with 1 df (check back to Fig. 20.1). We can confidently reject the null hypothesis that these two models are equally good, and retain the more complex one with Phosphate present.

The p-value would be:

```
> 1-pchisq(9.84,1)
[1] 0.001707575
```

Indicating the rejection of the null hypothesis.

So again, our most plausibly complex model remains the same:

$$f_i = c + \alpha_j + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} \qquad \text{(model 21.3)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

and we have only Landscape to drop:

$$f_i = c + \beta_k + m_P\, x_{P,i} + (m_N + \delta_k)\, x_{N,i} \qquad \text{(model 21.4)}$$

$$i = 1..48;\ j = R, U;\ k = L, M, H$$

```
> M_mcpm<-glm(Chlorophyll~Landscape+Flow+Nitrate+Phosphate
                        +Nitrate:Flow,data=my_data)
> logLik(M_mcpm)
'log Lik.' -141.1152 (df=9)
> M_s<-glm(Chlorophyll~Flow+Nitrate+Phosphate
                        +Nitrate:Flow,data=my_data)
> logLik(M_s)
'log Lik.' -154.5109 (df=8)
```

and

$$2\Delta LL = 2\ (LL_c - LL_s)$$

$$2\Delta LL = 2\ (-141.1152\text{--}154.5109) = 26.79$$

is χ^2 distributed with $(df_c - df_s) = (9 – 8) = 1$ degree of freedom.

This value of the test statistic is large – well to the right of the 95th quantile for a χ^2 distribution with 1 df (check back to Fig. 20.1) and so we can also confidently reject the null hypothesis these two models are equally good, and retain the more complex one.

The p-value testing the null hypothesis would be:

```
> 1-pchisq(26.79,1)
[1] 2.26808e-07
```

Indicating that once again we emphatically reject it.

We are left with a model that includes Landscape, Flow, Phosphate, Nitrate and the interaction of Nitrate and Flow. We can't make the model any simpler without significantly reducing our ability to explain variation in our response variable, so this is our **most complex minimal model**, in which we know that all the terms are required and will be significant to at least the 0.05 level. However, you will now need to test all the terms one final time as described in Chapter 20 (since the background within which the terms were tested most probably has changed) and examines the coefficients for effect size and direction as described in Chapter 22.

21.2 The ordering of the testing sequence

It is sensible, and in some cases, necessary to test interactions first, and then move on to those main effects not represented in retained interactions. We cannot use LRTs to test the main effects that are also represented in retained interactions because we can't retain an interaction and at the same time remove either of the main effects within it.

The ordering becomes complicated because the final choice of most complex minimal model can depend on it, most often when some explanatory variables are collinear with each other. The most objective ordering is to remove the terms in order of their increasing impact on the residual deviance. This can be established using the `drop1` command available in base R, and lists the effects on the residual deviance of dropping each term from the most complex plausible model.

```
> drop1(Mc)
Single term deletions
Model:
Chlorophyll ~ Landscape + Flow + Nitrate + Phosphate + Temp +
    Phosphate:Landscape + Nitrate:Flow
                    Df Deviance
<none>                   962.42
Temp                 1   978.08
Landscape:Phosphate  1   981.07
Flow:Nitrate         2  2537.15
```

Assuming we test interactions first, this ordering would suggest we LRT `Landscape:Phosphate` first, then `Flow:Nitrate` (and on dropping `Landscape:Phosphate`):

```
> drop1(M_mpcm)
Single term deletions
Model:
Chlorophyll ~ Landscape + Flow + Nitrate + Phosphate + Temp +
    Nitrate:Flow
             Df Deviance
<none>            981.07
Landscape     1  1543.13
Phosphate     1  1114.15
Temp          1  1005.38
Flow:Nitrate  2  2541.60
```

LRT first `Temperature`, `Phosphate`, and then `Landscape`.

21.3 Random effects and model selection

In principle, it's possible to include random effects in a model selection process. There is however a constraint: if we have just one random effect in the model, its removal changes the model from a mixed model to a fixed effects only model, and most packages that fit random effects require at least one random effect to be present. This would require us to compare a more complex model fitted in one package (say `lme4`) with a simpler model fitted in another (say `glm` in base R). This requires caution because different packages may calculate log-likelihoods in different ways that make then meaningless to compare across packages. There is an additional complication that Likelihood Ratio Test statistics applied to random effect terms in mixed models (variance terms) are poorly approximated by a χ^2 distribution.

In general, we suggest simply leaving random effects in models, and not including them in any model selection process. Random effects would be included in the first place because of concerns that there may be otherwise unrecognized dependencies between observations of the response variable, and whether they are significant or not, it is still a good idea to account for such dependencies. A random effect requires only 1 degree of freedom, so there is little to be gained from removing them in any case.

In the special case that you are fitting mixed models assuming normally distributed response variables you may have the option of fitting your model using 'full' maximum likelihood (ML), or 'restricted' maximum likelihood (REML). In the `lmer` command `REML` is set to `TRUE` by default. `REML = TRUE` generates more accurate estimates of the random effect variances, but if you want to compare two models using an LRT you must set `REML = FALSE`. Thus, if you conduct model selection then fit the models with `REML = FALSE`, but once you have determined your final model refit the model with `REML = TRUE`.

21.4 AIC

Instead of using LRTs you may choose to use AIC to compare different models. The basic ingredients of LRTs and AIC are essentially identical, but in our view LRTs are slightly more exact and more clearly founded. However, AIC is very commonly seen in the literature and we describe it briefly in Appendix U. It has particular value if for some reason you need to compare models that are not nested.

Important ideas to take-away

- You may choose to apply model selection to explore how to simplify your most complex plausible model, but there is no requirement to do so, and it's often simpler not to
- Model selection would start by testing the interactions
- You can't test the main effects that also occur in retained interactions. Main effects that occur in significant interactions are de-facto regarded as significant. No testing is required or is even appropriate
- The order in which you test terms can influence the most complex minimum model
- We recommend not including random effects in model selection unless you have a very specific reason for doing so
- Model selection is only one way of arriving at your final model. Regardless, you will need to conduct inference on your final model as described in Chapters 20 and 22

Chapter 22

Coefficient analysis

(back to Contents)

LRTs are the best way of determining whether a model term (a main effect or interaction) is making a significant contribution to explaining the variation in your response variable, but this doesn't tell you what the magnitude or the direction of the effect of the term on the response variable is. To examine this, we need to look at the coefficients in the model.

It is important to distinguish between whether an effect of an explanatory variable (or an interaction in which it is involved) is significant or not in explaining variation in the response variable, and what the size and direction of its effect is if it is statistically significant. An effect might be highly statistically significant, but so small that it is biologically uninteresting. This is only likely to be true when you have a lot of data, which enables us to identify small effects that are statistically significant. Of course – what is interesting is in the eye of the beholder, but we still need to have an idea of effect size and direction.

A good understanding of your model and its algebraic structure makes coefficient analysis straightforward. If you understand how terms represent adjustments then you really already know how to interpret them. We'll make an adjustment for every unit change of a continuous explanatory variable, and an adjustment for every level of a categorical explanatory variable (except the reference level). Interactions will result in further adjustments based on the combination of two explanatory variables, as described in Chapter 11.

While it is useful to be aware of the magnitude of all the coefficients in your model, it is only those coefficients that are large relative to their standard errors that might be important.

22.1 Model coefficients are estimated with uncertainty

One might imagine collecting some data, fitting a GLM, and estimating a slope. Perhaps it would be +4.759 mg Nitrate/L (as in model 6.1), suggesting a strong positive relationship between concentration of Nitrate and Chlorophyll (Fig. 22.1A). You might imagine going back the following week and collecting all these samples all over again, and re-doing the analysis. You'd hope to get something similar, but you wouldn't expect the slope to be *exactly* 4.759. It might be 4.481 or 4.935, may be 4.667 or 4.828 but not *exactly* 4.759! We of course don't know the 'true' actual relationship between Nitrate and Chlorophyll, we can only estimate it based on a sample. The larger the sample the more precisely we can estimate it, but it's still only an estimate, and is associated with some uncertainty – evident from the fact that we don't anticipate being able to repeat *exactly* previous estimates. So, how much uncertainty is there?

Ironically, it is possible to calculate the uncertainty in our parameter estimates quite precisely. As you might expect intuitively, uncertainty in our parameter estimates goes up as our unexplained variation increases, and it goes down as the residual degrees of freedom increases. This is primarily why we prefer parsimonious models that estimate as few unnecessary parameters as possible (therefore leaving as many degrees of freedom as possible as **residual degrees of freedom**), while explaining as much variation as possible. This way we can estimate the coefficients in our model and the associated uncertainty in these estimates.

Regardless of what distribution we use to model the unexplained or residual variation (Normal, Poisson, Bernoulli etc), the distributions of the coefficients that we estimate are assumed to be Normal. This is not as strange as you might think. There is no reason why the distribution used to model observations of the response variable should be the same as the distributions assumed to represent the coefficients in our model. Coefficients and data are quite different things. These Normal distributions – like any Normal distribution – are defined by a mean and a standard deviation, and these correspond to the estimated value of the coefficient, and the accompanying standard error.

22.2 Inference from coefficients relating to continuous explanatory variables

For model 6.1 (the output of which is replicated below), the algebraic structure of the model is:

$$f_i = c + m_N x_{N},$$

and the full unexpurgated output like this (a subset of which you'll have already seen from Chapter 6):

```
> m1<-glm(Chlorophyll~Nitrate,data=my_data)
> summary(m1)
Call:
glm(formula = Chlorophyll ~ Nitrate, data = d1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.1948     5.1191   2.187   0.0339 *
Nitrate       4.7590     0.5031   9.458 2.32e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for gaussian family taken to be 148.8343)
    Null deviance: 20161.3  on 47  degrees of freedom
Residual deviance:  6846.4  on 46  degrees of freedom
```

```
AIC: 380.31
```

From this we see that the mean estimate of the slope, m_N, is 4.759 μg/L Chorophyll per mg/L of Nitrate and the standard error is 0.503 (in bold). A standard error is the standard deviation of a mean. Don't let this distract you! (more explanation in Appendix V).

We know quite a bit about Normal distributions by now ... we know that our slope is distributed as shown in Fig. 22.1B.

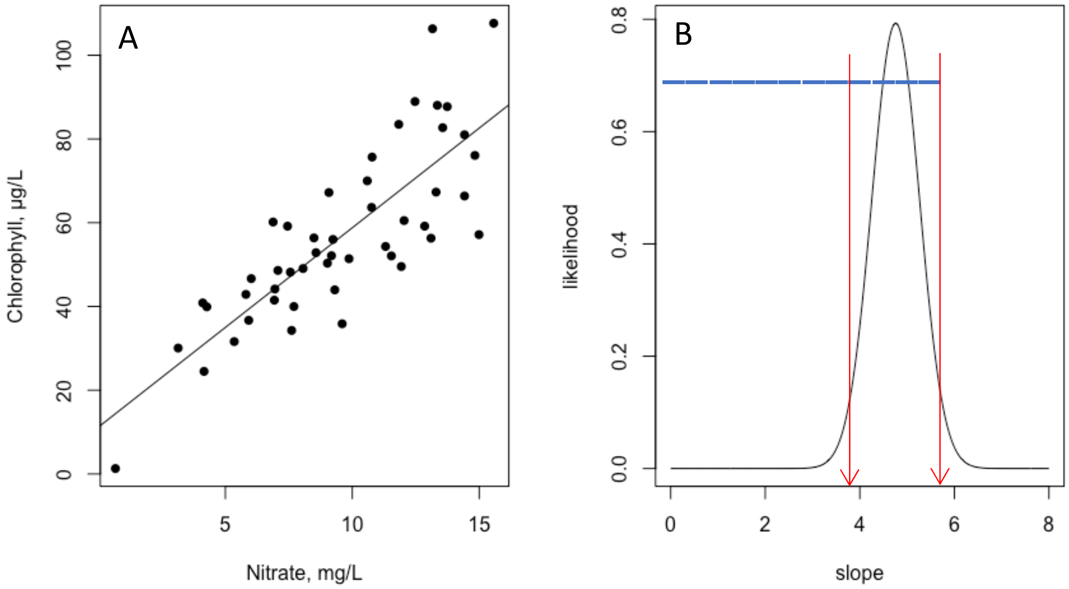


Figure 22.1. A) The relationship between Chlorophyll and Nitrate, and B) the distribution of the estimated slope. The blue lines indicate the width of the standard deviation (0.503) of the estimated mean (4.759), showing that there are more than 9 standard deviations 'between' the mean and zero; and the red lines indicate the position of the 95% confidence interval: 3.773 – 5.745, (see section 22.3).

What is clear from Fig. 22.1B is that the slope is really very unlikely to be zero because almost all the area under the curve is well to the right of zero. We can test the null hypothesis that we would find a relationship in these data like this (or more extreme) were there actually no relationship between Nitrate and Chlorophyll using a T test. 4.759 is (4.759-0)/0.503 = 9.458 standard deviations from zero (any more than 2 standard deviations from zero would likely signify a significant difference) so we can reject the null hypothesis emphatically (check back to Chapter 17 if you're not remembering how to interpret T statistics):

```
> 2*(1-pt(9.458,46))
[1] 2.322587e-12
```

(why 46 degrees of freedom? Because that is the number of residual degrees of freedom - emboldened in red in the output above).

Twice the area to the right of 9.458 is `2.322587e-12` which is tiny, so we'd say $p < 0.001$ (R does all this for us as you can see in the output, the T statistic is the number of standard deviations the parameter estimate is from zero, and R generates a p-value from this test statistic emboldened in blue). Thus, we are confident the effect of Nitrogen on Chlorophyll is positive and highly significant, with Chlorophyll concentration expected to increase by about 4.7 µg with every mg of Nitrate.

22.3 Calculating the confidence interval for a coefficient

We've calculated the p-value that is a test of our null hypothesis (that the slope is zero) and we've emphatically rejected it. But we can also place an interval on our confidence in our estimate of the slope. The idea is clear from Fig. 22.1B. Formally, a 95% **confidence interval** is the interval in which we expect the true value of a parameter to fall with probability 0.95. The percentage may be chosen to be anything you want, but the standard is a 95% confidence interval (CI). For a normal distribution 95% confidence intervals are generated (approximately) by adding (and subtracting) two standard errors to (and from) the mean. In fact, the exact multiplier is that corresponding to the 95% CIs on a T distribution with df equal to the residual degrees of freedom in the model, i.e.

```
> qt(0.025,46)
[1] -2.012896
> qt(0.975,46)
[1] 2.012896
```

(where `qt` is the command for a specified quantile of a T distribution)

So, the 95 CIs on the slope for Nitrate (m_N) would be:

4.759 + 2.013 x 0.503 = 3.746

4.759 - 2.013 x 0.5031 = 5.772

Thus, the true value of the slope is likely to lie within the interval 3.746 to 5.772, with probability 0.95. If we reject the null hypothesis with a p-value less than 0.05 we expect the 95% CI for the parameter to exclude zero. Put another way, if the 95% CI doesn't include zero, then we regard the coefficient to be significantly different from zero, with at least 95% confidence.

Confidence intervals for coefficients in a model can be obtained using the `confint` command in base R:

```
> confint(my_model)
                2.5 %    97.5 %
(Intercept) 1.161508 21.228167
Nitrate     3.772825  5.745134
```

We have proposed that inference is conducted using LRTs, but that assessment of the magnitude and effect size are based on analysis of the coefficients. However, the results of each will generally be consistent with each other. An LRT applied to model 6.1 would compare the complex model with Nitrate and the simpler model without Nitrate, i.e. with just the intercept *c*. There is nothing special about the intercept only model (or **null model** as it is sometimes known), it is just the model (in this case) where all the observations of the response variable are assumed to come from one single Normal distribution.

```
> m_c<-glm(Chlorophyll~Nitrate,data=d1)
> logLik(m_c)
'log Lik.' -187.1556 (df=3)
> m_s<-glm(Chlorophyll~1,data=d1)
> logLik(m_s)
'log Lik.' -213.0767 (df=2)
```

and

$$2\Delta LL = 2\ (LL_c - LL_s)$$

$$2\Delta LL = 2\ (-187.1556 - -213.0767) = 51.842$$

is χ^2 distributed with (df_c - df_s) = (3 – 2) = 1 degree of freedom (p < 0.001).

The LRT leads us to reject the null hypothesis that the null and more complex model are as good as each other, and adopt the more complex model with Nitrate represented. Thus, we conclude that Nitrate is explaining a significant amount of the variation in Chlorophyll. It is therefore to be expected that the slope governing the relationship between Nitrate and Chlorophyll is significantly different from zero. Indeed, we could regard the LRT as a test of the null hypothesis that the slope m_N = 0, so we are doing almost exactly the same thing with the LRT and the T test on the coefficient, but of course the coefficient conveys information about the effect size and direction also.

22.4 Inference from coefficients relating to categorical explanatory variables

Suppose we had added the categorical explanatory variable Flow, 3 levels: H(igh), L(ow), M(edium) into model 6.1. The algebraic structure of the model is:

$$f_i = c + \alpha_j + m_N\, x_{N,i} \qquad \text{(model 22.2)}$$

$$j = H, M, L$$

We'd see this output:

```
> m2<-glm(Chlorophyll~Nitrate+Flow,data=my_data)
> summary(m2)
Call:
glm(formula = Chlorophyll ~ Nitrate + Flow, data = d1)
Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.9916     4.6840   4.695 2.62e-05 ***
Nitrate       4.6374     0.4035  11.493 7.71e-15 ***
FlowL       -18.1583     3.4537  -5.258 4.10e-06 ***
FlowM       -10.7478     3.4585  -3.108   0.0033 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for gaussian family taken to be 95.1719)
    Null deviance: 20161.3  on 47  degrees of freedom
Residual deviance:  4187.6  on 44  degrees of freedom
AIC: 360.71
```

We see 2 adjustments for Flow (H being the reference) and both coefficients (-18.158 for Low Flow and -10.748 for High Flow) are significantly different to zero, being, respectively, 5.258 and 3.108 standard deviations to the left of zero. We'd get the following 95% CIs:

```
> confint(m2)
                 2.5 %     97.5 %
(Intercept)  12.811212  31.171981
Nitrate       3.846574   5.428250
FlowL       -24.927537 -11.389150
FlowM       -17.526308  -3.969303
```

All 4 confidence intervals exclude zero, corresponding with the results of our LRTs.

22.5 Relationship of the T-tests to the LRT

The LRT comparing the model with and without Flow would give:

```
> m_c<-glm(Chlorophyll~Nitrate+Flow,data=my_data)
> logLik(m_c)
'log Lik.' -175.3572 (df=5)
> m_s<-glm(Chlorophyll~Nitrate,data=my_data)
> logLik(m_s)
'log Lik.' -187.1556 (df=3)
```

and

$$2\Delta LL = 2\ (LL_c - LL_s)$$

$$2\Delta LL = 2\ (-175.3572 - -187.1556) = 23.597$$

is χ^2 distributed with (df_c - df_s) = (5 – 3) = 2 degrees of freedom ($p < 0.001$). The LRT leads us to reject the null hypothesis that the simpler and more complex model are as good as each other, and adopt the more complex model with Flow also represented. Thus, we conclude Flow is explaining a significant amount of the variation in Chlorophyll. It is therefore to be expected that at least one of the adjustments governing the relationship between Flow and Chlorophyll are significantly different to zero. Indeed, we could regard the LRT as a test of the null hypothesis that all the adjustments are the same and equal to zero, that is: $\alpha_L = \alpha_M = 0$, a hypothesis we emphatically reject in this example, and instead note that at Low and Medium flows Chlorophyll concentrations are significantly less than at High flows.

The important point here is that *it only takes one* of the adjustments to be significantly different to zero for the whole main effect to be regarded as significant. If even a single level of the categorical variable requires a significant adjustment then it has to be that the explanatory variable as a whole has a significant effect.

22.6 Wald statistics

Wald statistics are a bit like T-statistics but instead of reflecting how many standard errors distant a coefficient is away from the value under the null hypothesis (often chosen to be zero), the distance is measured in units of variance, and instead of following a T-distribution under the null hypothesis, the distribution of Wald statistics approximates that of a χ^2 distribution with 1 df. There is no pressing reason to use them at this stage!

22.7 Interaction terms

Interaction terms can be subjected to exactly the same examination – we can use T statistics to determine if they differ significantly from zero, and we can compute 95% CIs.

22.8 Missing p-values

Some packages don't provide p-values on estimated coefficients because the authors of the packages for various valid reasons believe inference is best done in other ways (this harks back to points made in Chapter 19).

> `lmer` and `glmer` don't display p-values for the coefficients. This is because the author of the package – a bit like us – believes there are better ways to conduct inference (e.g. inference based on 'whole model comparison' such as LRTs or related methods). But if you load the package `lmerTest` p-values can be generated.

22.9 Do we need to do both T-tests and LRTs?

No. But you do need to study your coefficients. We advocate establishing and reporting the significance of model terms using LRTs, but using the coefficients to describe the direction, magnitude, and confidence in the effect size. For example, one might say: The effect of Nitrate on Chlorophyll was highly significant ($2\Delta LL$ = 51.842, df = 2, p < 0.001), with average Chlorophyll concentration expected to increase by 4.7 μg (95% CIs: 3.746 - 5.772) with every mg of Nitrate.

While the inferences made from LRTs and coefficient analyses are often essentially the same, they may generate inconsistent results particularly when significance is borderline (and particularly for smaller data sets – see discussion in Chapter 19.5). Coefficient analysis doesn't generate a single p-value for the effect of a categorical explanatory variable with more than 2 levels, only the effects of each of the levels *relative to the reference level*. Furthermore, because each level is being compared to the reference level (assumed to have a zero adjustment) *the p-values for each level depend on the choice of the reference level*. Given this is determined by how the levels were labelled this is quite unsatisfactory. Indeed, it is entirely possible that for some choice of reference level it will appear that none of the adjustments are significantly different to zero, while for other choices of reference level they will appear significant. We explain how this can be in Appendix W. This can be very misleading, and is why we advocate establishing and reporting the significance of model terms using LRTs, and using the coefficients to describe the direction, magnitude, and confidence in the effect size.

22.10 Post-hoc tests

For categorical explanatory variables with more than 2 levels, you may be content to know simply that the categorical explanatory variable does or does not significantly influence the response variable. This, combined with the magnitude and direction of effects corresponding to the different levels conveyed by the corresponding adjustment parameters is often enough. However, the adjustment coefficients do only enable formal statistical testing of the effects of the different levels *relative to the reference level*. If you want to compare between the different (non-reference) levels you'll need to conduct post-hoc tests (discussed in Appendix X).

Important ideas to take-away

- While LRTs are excellent tools for determining whether a main effect or interaction is contributing significantly to explaining variation in our response variable, we need to look at the coefficients in our models to determine the direction, magnitude, and our confidence in the accuracy of the size of effects
- Just because a term is statistically significant doesn't mean it's biologically important or interesting
- Coefficients in GLMs are assumed to be Normally distributed with means equal to the estimated value of the coefficient, and standard deviations equating to the standard errors provided in the summary output
- Knowing the mean and standard deviation of a coefficient allows us to test if it statistically significantly different to any value (most often zero)
- We can also compute 95% confidence intervals (CIs) by adding and subtracting (approximately) 2 standard deviations to and from the mean.
- If 95% CIs do not include zero we can say we are at least 95% sure the estimated mean is different to zero
- The results of LRTs and coefficient analysis are usually consistent with each other, but p-values for coefficients relating to the adjustments for different levels of a categorical explanatory variable are sensitive the choice of

reference level, making inference based on these p-values potentially misleading.

Chapter 23

Statistical power

(back to Contents)

A basic understanding of statistical power is critically important for efficient experimental design, and ensuring your resources are not squandered through too little or too much sampling. Consideration of statistical power is also very important when interpreting 'negative results' - that is ... if you don't find an effect. Was it that there really isn't an effect? Or your study wasn't sufficiently powerful to find one?

23.1 Type 1 and Type 2 errors

Of course we'd hope inference will go right ... we'll detect effects if they are present, and fail to do so when they are not. However, there are two important ways inference can go wrong. The first is when you think you've detected an effect that doesn't actually exist. Or more formally, you reject a null hypothesis when in fact it's actually true. This is called (somewhat forgettably) a Type 1 error. The second is when you fail to detect an effect that actually does exist, or more formally, you fail to reject a hypothesis that is actually false. This is (also forgettably) termed a Type 2 error (it might help to note that in explaining a Type 2 error to someone, you'd need to use 2 f-words – fail and false).

Statistical power is the probability of avoiding a Type 2 error. Or put another way, the probability of detecting an effect (of at least a certain magnitude) if it is present. This makes sense ... a lot of statistical power means you are likely to spot an effect if it is present, even if it is quite small. Most studies settle for a probability of detecting an effect (sometimes known by the letter β) of around 0.8.

Of course, we have to also define what we mean by *detecting an effect*. With what confidence? As previously discussed, 95% confidence is usually regarded as acceptable. If we reject the null hypothesis with 95% confidence, there is a 5% chance its true (meaning the acceptable Type 1 error probability is set at 0.05). The acceptable Type 1 error probability is sometimes denoted by the letter α (defined in this context α and β have nothing to do with the notation we use for adjustments on the right-hand side of a GLM). The power parameters α and β are summarized in Table 23.1.

Table 23.1. Types of inferential error

Error type	Inferential mistake	Commonly accepted values
Type 1	rejecting a null hypothesis when it's actually true	α is the probability of making this mistake and is usually ≤ 0.05
Type 2	failing to reject a hypothesis that is actually false	β is the probability of avoiding this mistake and usually should be $\geq\ 0.8$

Some understanding of power is very important. First, if you don't have enough of it, even if there are real effects, you'll have little chance of identifying them with any confidence and your study will have been a waste of time, money and effort. If you have far more than you need, then again, time, money and effort will have been wasted on collecting data, all of which was not necessary. Second, without sufficient power, you'll be unlikely to reject your null hypotheses, and *you'll be unsure whether there actually are effects but you didn't have sufficient power to detect them; or there aren't effects* and it wouldn't have mattered how much power you have! In short, without an understanding of power, it's very hard to know what to make of 'a negative result', and negative results must be interpreted carefully.

So, power is worth studying in a bit more detail.

The parameter α is very much within our control. We can decide how large our test statistics should be before we use them to justify rejecting a null hypothesis. While 0.05 is a conventional standard for Type 1 error probability there are situations where you may wish to change it. If the consequences of believing you've detected something when it doesn't exist are very high (for example, invasive surgery for a relatively mild condition a patient doesn't actually have), it might be wise to demand higher confidence (lower α). Conversely, if the consequences of failing to detect an effect that actually does exist are very serious (say, diagnosing a potentially treatable cancer), then a result that would usually be considered only suggestive, might be taken more seriously (and perhaps α should be higher). Certainly, why the scientific community has adopted α = 0.05 is a good question that doesn't have a very good answer, and a reminder that there is nothing particularly objective about this level of significance.

But β is quite a bit trickier. β is determined by at least three different things:

1. The strength of effect we wish to detect: We do have *some* say in this. Obviously if we insist on being able to identify a very small difference between samples we are going to require a lot of data (residual degrees of freedom) to do so. So when designing a study, we should always be asking …

what is the minimum effect size we want this study to be able to detect? Make this size too large, and it might exceed the actual size of any effect that may exist, and of course we'd not detect anything smaller. Make it too small, and the sample sizes will become so large as to make the study too time consuming and expensive to gather the required data.

2. By the amount of unexplained variation (or noise) in the data (which of course depends on the model): the more noise there is, the bigger the sample size that will be required to identify a given effect size. What we call 'noise' might be partly potentially explainable variation (were we to have collected the most appropriate and useful explanatory variables), and partly truly unexplainable variation we'll just have to live with. So, again we have *some* control over this … but not a great deal. Noise can be reduced through careful consideration of what the most relevant explanatory variables might be; and model formulation to ensure they are most effectively represented (for example, including appropriate interactions). However, being able to quantify this unexplained variation in advance is often the hardest part of trying to power a study.
3. The size of the data set (more precisely, the number of observations of the response variable): Given an effect size, and some estimate of the variation we are likely to encounter, we can – in principle – calculate a sample size that would give us the required power – the probability of detecting such an effect were it to be present. While resources (time, field assistants, equipment, reagents etc) are generally limiting we can in principle have a lot more control over the sample size, even if in practice samples sizes are often determined by what is affordable or practically feasible.

Any thoughts about the power of a study prior to conducting data collection and analysis is a good thing (shockingly, it is often given almost no consideration even by experienced researchers) and it is a requirement in some fields to make the work publishable or to obtain grant funding (e.g. experiments involving human or animal subjects). It is possible to conduct formal quantitative power analyses for the simplest experimental designs, and there are many calculators on-line to help us do so; but simulation will likely be required for more complex designs.

For simple analyses there are power calculators available on-line, for example:

https://en.wikipedia.org/wiki/G*Power
https://clusterrcts.shinyapps.io/rshinyapp/
https://clincalc.com/stats/samplesize.aspx
https://www.gigacalculator.com/calculators/power-sample-size-calculator.php

In the simplest scenario, we might have two groups (say a control-C and a treatment-T), and a response variable modelled with (say) a Normal distribution, i.e.

$$f_i = c + \alpha_j \qquad i = 1 .. n, j = C, T$$

The effect size (the effect of the treatment) would be given by α_j. The T-statistic associated with the estimate of the coefficient α_j would be given by α_j / (std. err. α_j) and the absolute value of this will need to be (more or less) larger than 2 to be at least 95% sure that $\alpha_j \neq 0$. So, we can see that larger T statistics (more power) increases with effect size, and as the standard error of the coefficient is reduced (Fig. 23.1).

There is an important distinction to be made between standard deviations and standard errors. Standard deviations are intrinsic properties of distributions that are not in any way dependent on how we sample from them. The higher the standard deviation the more variability there will be in the variates sampled from the distribution. A standard error reflects the uncertainty in our estimate of the mean of a distribution. How sure are we that our estimate of the mean (which is based on a sample) ... is the mean? The more samples, the more certain we will be in our estimate of the mean. Estimated means of distributions (that have standard deviations), do themselves have distributions (that have standard errors). A standard error is a standard deviation of a mean but based on a particular sample. If we take *n* samples from a distribution with a standard deviation of σ, and calculate the mean of these samples, then the standard error associated with our estimate of this mean is $\sigma/\sqrt{n}$. As *n* increases, so the standard error decreases. We discuss this in more detail in Appendix V.

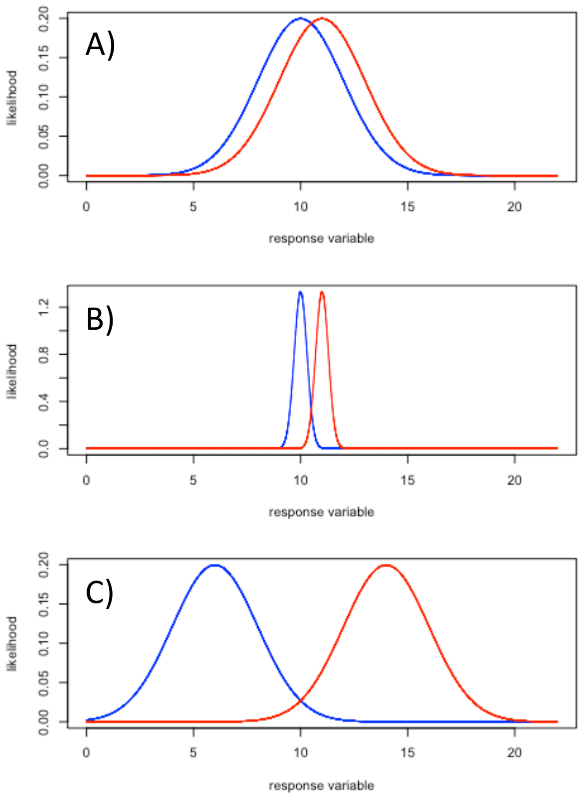


Figure 23.1. Effects of variance and effect size on the ability to detect differences between distributions. A) Small effect size (the difference between the averages of the red and blue distributions) and larger standard deviation means a lot of data would be required to distinguish between the blue and red distributions. B) The averages are the same as A, but a smaller standard deviation would make the two distributions easier to distinguish from each other. C) Increasing the effect size means that even with large standard deviations, differences could be distinguished, even without enormous sample sizes.

So standard errors are related to the sample size and the underlying variability in the data in a mathematically simple way. Consequently, given a certain sample size, we can calculate statistical power. Alternatively, given a certain required statistical power, we can calculate the sample size necessary to generate that power. So ‘power calculators’ and ‘sample size calculators’ are in a sense two sides of the same coin. They allow calculation of *one* of: the sample size, the effect size relative to underlying variation, the significance level, or the power – to be calculated if any three of these four quantities can be estimated. The underlying variation is often not straightforward to estimate but might be estimated from previous similar studies, or perhaps examined over a plausible range of values.

However, the simple formulas generally don’t apply to more complex situations with more explanatory variables, and the key to power analysis is really data simulation. Simple programmable loops can be written in R to simulate data, fit GLMs, examine p-values and explore the power of arbitrarily complex GLMs. This is not impossibly complex, but it’s beyond the scope of this text. If power analysis looks too complicated – you should ask for some advice, and/or closely examine the design of similar studies in the literature.

23.2 Power calculators in R

The `pwr.t.test` command in the package `pwr` can be used to estimate one of: the sample size, the effect size relative to variation, the significance level, or the power to be calculated if three of these four quantities are provided. Here, *d* is the effect size divided by the standard deviation, and `n=NULL` indicates it is the sample size to be estimated.

```
> pwr.t.test(n=NULL, d=0.25, sig.level = 0.05, power=0.8)

     Two-sample t test power calculation

              n = 252.1275
              d = 0.25
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

can be used to generate sample size estimates for the simple one explanatory variable – 2 level GLM (otherwise known as a two sample T test) .

Below we've set `n=252`, and `power=NULL`, and we'll estimate the power of a study with a sample size of 252 in each group:

```
> pwr.t.test(n=252, d=0.25, sig.level = 0.05, power=NULL)

     Two-sample t test power calculation

              n = 252
              d = 0.25
      sig.level = 0.05
          power = 0.7998008
    alternative = two.sided

NOTE: n is number in *each* group
```

Chapter 24

How to write up your analysis – methods and results

(back to Contents)

This chapter is not about how to write a scientific paper. But there are some basic guidelines about the organization of a paper that are important to adhere to.

The Introduction should conclude with a statement of the goals of the study. These might be posed as questions *or* hypotheses *or* objectives. It is probably redundant to describe the goals of your study using more than one of these frameworks.

However, the goals can be clearest if explicitly framed around the statistical analyses to be conducted. The introduction should also clearly motivate each of the variables considered in the study (e.g. what is and isn't known that led you to include them).

Some pointers then on constructing the methods, results, and discussion sections:

24. 1 Methods

- Should have a section on how the data were collected or acquired
- Should explicitly describe (and justify) sample sizes, experimental design and/or sampling strategy
- Should have a section on how the data were analysed
- Should map directly and transparently on to the questions/hypotheses/ objectives of the study as laid out at the end of Introduction. Maintaining this sequencing throughout the manuscript is important
- Should identify which variables were modelled as continuous and categorical, fixed and random, which distributions were used to model your response variable, and identify any link functions
- Should say which variables were included in which model, and what interactions were tested – and why
- Pay close attention to the units of your variables
- Should describe in general terms how the final model was identified (whether 'first and final' or some form of model selection adopted)
- Should develop some notation for your coefficients and data and define it and use it consistently
- Should provide the algebraic structure of the most important models
- Should define ranges on any subscripts you use.
- Should *not* use verbatim R commands to describe what you did,
- Should *not* say you used RStudio (it's only an editing environment)

- Should say which version of R you used, which R commands, and which version of any packages you used
- Should say how inference was conducted (AIC, LRT, T-tests on coefficients etc)
- Should indicate that appropriate diagnostic checks on the model were performed
- Reference power analyses if undertaken
- Should be written as concisely as possible

24.2 Results

- Should provide a brief and concise summary of the data
- Should map directly and transparently on to the questions/ hypotheses/objectives of the study as laid out at the end of Introduction and Methods *in the same order*
- Are unlikely to require more than 4-5 figures (if more needed consider panel figures)
- May contain tables but don't duplicate information with figures
- All tables and figures must be referenced from the text
- Should *not* contain verbatim R output
- Should *not* contain a blow-by-blow account of model selection
- Should *not* usually contain detailed output from diagnostic checks (unless these are for some reason critical) but should confirm that all assumptions made by the analysis were met
- Should convey the significance (or insignificance) of all major findings, supported by appropriate statistics (test statistic, degrees of freedom, p-value), or coefficients and their p-value or confidence intervals
- Should pay close attention to the units of your variables
- Should convey the effect size and direction where appropriate
- Reference power if appropriate

24.3 Discussion

- Should always start with a simple concise summary of the most important findings of your study, related in the same order as the questions /hypotheses/objectives of the study were laid out at the end of Introduction. By reading the last paragraph of the introduction and the first paragraph of the discussion a reader should come away with the fundamental narrative of your paper.
- Should not introduce any new results
- Need not refer back to previous figures and tables

- Should contextualize your findings in respect of previous studies
- Should consider caveats, qualifications and limitations of your own study (in the most positive way you reasonably can
- Should consider possible future work that might be conducted in light of your study.

24.4 The Figures

- Avoid representations of 3D graphs
- Keep response variables on the y-axis, and explanatory variables on the x-axis
- Ensure that fitted relationships shown on graphs are those you report in your results, and not some ggplot curve of a different origin
- Try to present model fits superimposed on relevant data whenever you can
- Be consistent in your use of axis scales (usually start the y-axes at zero if it makes sense to do so)

Examples of write up

In what follows we provide brief outlines of how you might convey an analysis of each of our 4 response variables. To keep things simple, we take them one at a time, and present 4 different write-ups, describing analyses with and without model selection.

24.5 Continuous response variable: Chlorophyll

Introduction [Chlorophyll]

Should conclude with a statement like:

In this study our overarching goal is to understand the physical and chemical determinants of a key biological indicator of water quality. Specifically, the influence of phosphate, nitrate, temperature, speed flow and landscape on chlorophyll concentration. [Each of these variables should have been motivated by critical analysis of the literature, earlier in the introduction.]

Methods [Chlorophyll]

Data collection *{this is just bare bones ... you would more complete details of rivers, sites, and how different flows and landscapes were defined}*

Data were acquired from 12 sites (S01-S12) distributed along 4 different rivers (R1-R4). Sites were stratified opportunistically between flow rates that were low, medium or high, and between stretches of river that flowed through urban and rural landscapes (you would need more detail on the sampling strategy). Temperature of the river water at a depth of 10 cm was recorded in situ, and five samples of river water (between 5 and 10 ml) were taken from each site and stored at 4˚C prior to dispatching one of each of the five samples to each of five laboratories for analysis. Each laboratory provided analysis of the concentration of chlorophyll (μg/L), nitrate (mg/L), and phosphate (μg/L) in each sample.

Statistical analysis {without model selection}

The data were analyzed using a General Linear Mixed Model where chlorophyll concentration (assumed to be Gaussian distributed) was the response variable, and temperature, nitrate and phosphate were treated as continuous explanatory variables, landscape (2 levels: urban and rural) and flow (3 levels: low, medium and high) as categorical, and lab (L1-L5) and river (A-D) as random effects. The models included all the fixed and random effects, and three interactions hypothesized to be important: flow x nitrate, flow x landscape, and temperature x nitrate. The fitted model was:

$$f_i = c + m_T x_{T,i} + (m_N + \gamma_j) x_{N,i} + m_P x_{P,i} + \alpha_j + \beta_k + m_{T:N} x_{N,i} x_{T,i} + \delta_{jk} + R_l + L_m$$

Where f_i (i = 1..240) is the fitted value for chlorophyll concentration, c the intercept, m_T, m_N and m_P the slopes relating temperature, nitrate and phosphate to chlorophyll, γ_j represents the continuous-categorical interaction between nitrate and flow (j = low, medium and high), α_j the adjustment to the intercept for flow, β_k the adjustment for landscape (k = rural, urban), $m_{T:N}$ the continuous interaction between nitrate and temperature, δ_{jk} the categorical interaction between landscape and flow, and R_l (l = *R1, R2, R3, R4*) and L_n (n = 1..5) the random effects of river and lab.

Models were fitted in R (v4.3.0) and lme4 (v1.1-33). Inference was conducted using Likelihood Ratio Tests (LRTs). Main effects were tested after removing non-significant interactions, and post-hoc tests conducted in emmeans (v1.8.5). Pseudo R^2 values were estimated using the r.squaredGLMM command in the MuMIn package. Standard tests to check the assumptions of the models were met were conducted using the DHARMa package (v0.4.6).

Results [Chlorophyll]

Residual analysis indicated no diagnostic anomalies with the model fit. Average chlorophyll concentration was 68.34 μg/L (range 1.29-126.35 μg/L). The fixed effects accounted for 81.4% of the variation. The Intraclass Correlation coefficients for the random effects river and lab were 0.401 and 0.552 respectively.

LRTs revealed that the interaction between nitrate and flow was significant (χ^2 = 405.1, df = 2, p < 0.0001). While mean levels of chlorophyll were highest at high flows (76.3 μg/L) and lowest at low flows (59.6 μg/L), the effect of nitrate on chlorophyll concentration was always positive but least at low flows (where chlorophyll increased at 3.87 μg/L per mg nitrate and most at high flows (where chlorophyll increased at 7.11 μg/L per mg nitrate). Landscape had a significant effect on chlorophyll concentration (χ^2 = 387.1, df = 1, p < 0.0001), with samples from rural landscapes containing on average 10.05 μg/L more (95% Confidence Intervals (CIs) 9.01 – 11.09 μg/L) than samples from urban landscapes. These findings are illustrated in Fig 24.1. LRTs showed that all other terms in the model were not significant (p>0.05).

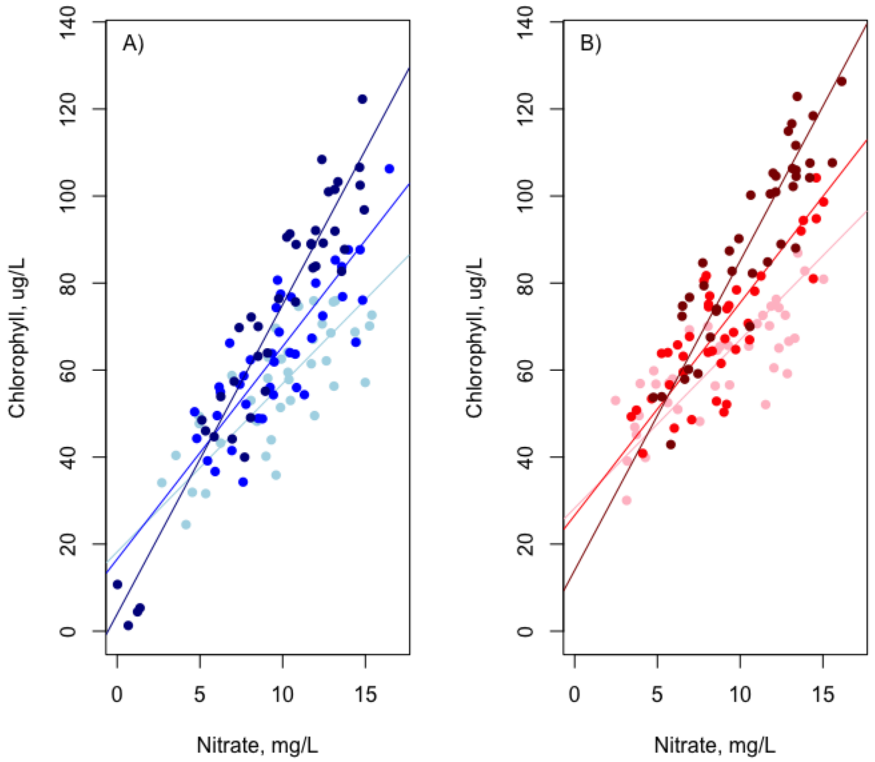


Figure 24.1. A) relationship between nitrate and chlorophyll concentrations in urban landscapes for low flows (light blue), medium (blue) and high flows (dark blue); and B) in rural landscapes for low flows (red blue), medium (red) and high flows (dark red).

Commentary: the figure describes the significant results from the analysis, with the data and the model fits appearing together. There will be all sorts of ways of creating of plots like these, but here we `subset()` the data into their different flows and landscape levels, use the plot command to establish the first `plot()`, and the `points()` command to add the different point layers with different colours. We then manually added the lines using the `abline()` command. Effect sizes are relatively easy to describe because there is no link function required.

24.6 Count data: Bacterial counts

Introduction [bacterial counts]

Should conclude with a statement like:

In this study our overarching goal is to understand the physical and chemical determinants of key biological indicators of water quality. Specifically, the influence of phosphate, nitrate, temperature, speed flow and landscape on bacterial counts. [Each of these variables should have been motivated by critical analysis of the literature, earlier in the introduction.]

Methods [bacterial counts]

Data collection *{this is just bare bones … you would more complete details of rivers, sites, and how different flows and landscapes were defined}*

Data were acquired from 12 sites (S01-S12) distributed along 4 different rivers (R1-R4). Sites were stratified opportunistically between flow rates that were low, medium or high, and between stretches of river that flowed through urban and rural landscapes (you would need more detail on the sampling strategy). Temperature of the river water at a depth of 10 cm was recorded in situ, and five samples of river water (between 5 and 10 ml) were taken from each site and stored at 4˚C prior to dispatching one of each of the five samples to each of five laboratories for analysis. Each laboratory provided analysis of the concentration of nitrate (mg/L), and phosphate (µg/L), and bacterial counts per ml in each sample.

Statistical analysis {with model selection}

The data were analyzed using a Generalised Linear Mixed Model where bacterial count (assumed initially to be Poisson distributed) was the response variable, and temperature, nitrate and phosphate were treated as continuous explanatory variables, landscape (2 levels: urban and rural) and flow (3 levels: low, medium and high) as categorical, and lab (L1-L5) and river (A-D) as random effects. The model was formulated by starting with a most plausibly complex model including all the fixed and random effects, and three interactions hypothesized to be important: flow x nitrate, flow x landscape, and temperature x nitrate:

$$\ln(f_i) = c + m_T x_{T,i} + (m_N + \gamma_j)x_{N,i} + m_P x_{P,i} + \alpha_j + \beta_k + m_{T:N} x_{N,i} x_{T,i} + \delta_{jk} + R_l + L_m$$

Where f_i (i = 1..240) is the fitted value for bacterial count, c the intercept, m_T, m_N and m_P the slopes relating temperature, nitrate and phosphate to chlorophyll, γ_j represents the continuous-categorical interaction between nitrate and flow (j = low, medium and high), α_j the adjustment to the intercept for flow, β_k the adjustment for landscape (k = rural, urban), $m_{T:N}$ the continuous interaction between nitrate and temperature, δ_{jk} the categorical interaction between landscape and flow, and R_l (l = *A, B, C, D*) and L_n (n = 1..5) the random effects of river and lab.

This initial model was subject to model selection using likelihood ratio tests (LRTs) with interactions tested first, and then main effects not represented in retained interactions in the order that they increasingly impacted on the deviance (using the drop1 command). Models were fitted in R (v4.3.0) and lme4 (v1.1-33). Inference was conducted on the final model using LRTs, and where appropriate post-hoc tests conducted in emmeans (v1.8.5). Standard tests to check the assumptions of the models were met were conducted using the DHARMa package (v0.4.6).

Results [bacterial counts]

Residual analysis indicated overdispersion relative to the Poisson distribution, so the models were refit assuming bacterial counts were distributed according to Negative Binomial distribution. Following the removal of non-significant terms (the landscape x flow and nitrate x flow interactions and the main effects flow and landscape) the final model was:

$$\ln(f_i) = c + m_T x_{T,i} + m_N x_{N,i} + m_P x_{P,i} + m_{T:N} x_{N,i} x_{T,i} + R_l + L_m$$

There were no further diagnostic anomalies with the model fit following the adoption of the Negative Binomial distribution (the dispersion parameter was estimated to 7.08).

Average bacterial count was 82.68 mL^{-1} (range 0-327 mL^{-1}). The fixed effects accounted for 79.3% of the variation in bacterial counts. The main driver of variation in bacterial count is phosphate concentration (χ^2 = 141.8, df = 1, p < 0.0001). On average, bacterial counts increase from 25 to 157 as Phosphate concentration increases from 50 to 350 μg/L. Bacterial counts are also influenced by the interaction between nitrate and temperature (χ^2 = 132.3, df = 1, p < 0.0001). Individually, increasing nitrate and temperature have a negative effect on bacterial count, but the interaction between the two is positive, so highest bacterial counts arise at high temperatures and nitrate concentrations. However, the combined effects of nitrate and temperature are modest compared to the dominant effect of phosphate (Fig. 24.2).

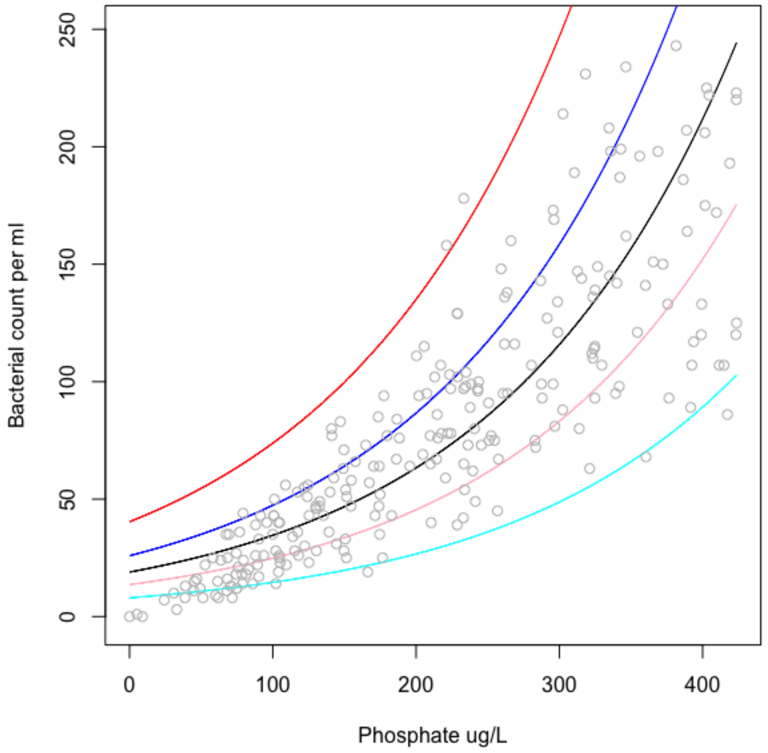


Figure 24.2. The relationship between phosphate and bacterial count for five different combinations of temperature and nitrate concentration. At mean temperature (13.16 degrees C) and nitrate concentration (9.52 mg/L): black line; at the highest temperature (16.51 degrees C) and highest nitrate concentration (16.45 mg/L): red line; at lowest temperature (9.03 degrees C) and lowest nitrate concentration (0.02 mg/L): blue line; at the highest temperature (16.51 degrees C) and lowest nitrate concentration (0.02 mg/L): pink line; at lowest temperature (9.03 degrees C) and highest nitrate concentration (16.45 mg/L): cyan line.

Commentary: This is a harder analysis to describe, first because the log-link function means the influences of the explanatory variables are not linear, so they are not well described by slopes on the un-logged scale, and second, because there are two continuous interacting explanatory variables. So, the figure is constructed to capture the range of relationships between the 3 continuous explanatory variables and the response variable. This can be achieved by constructing a hypothetical sequence of Phosphate values going from the smallest to the largest observed phosphate value: `> P_seq<-seq(0.06,423.7,0.1)`, calculating the bacterial counts corresponding to different combinations of Temperature and Nitrate (as shown in the legend to Fig 24.2) according to (say):

```
> Bac_seq_MM<-exp(4.0376160+(-0.2383358*mean_N) +
(0.0184482*mean_N*mean_T)+0.0060379*P_seq+(-0.0865801*mean_T))
```

And plotting the curves, using first the `plot()` command, and then the `lines()` command. The data can then be added using the `points()` command.

24.7 Count data: Invertebrate counts

Introduction [invertebrate counts]

Should conclude with a statement like:

In this study our overarching goal is to understand the physical and chemical determinants of key biological indicators of water quality. Specifically, the influence of phosphate, nitrate, temperature, speed flow and landscape on invertebrate counts. [Each of these variables should have been motivated by critical analysis of the literature, earlier in the introduction.]

Methods [invertebrate counts]

Data collection *{this is just bare bones ... you would more complete details of rivers, sites, and how different flows and landscapes were defined}*

Data were acquired from 12 sites (S01-S12) distributed along 4 different rivers (R1-R4). Sites were stratified opportunistically between flow rates that were low, medium or high, and between stretches of river that flowed through urban and rural landscapes (you would need more detail on the sampling strategy). Temperature of the river water at a depth of 10 cm was recorded in situ, and five samples of river water (between 5 and 10 ml) were taken from each site and stored at 4˚C prior to dispatching one of each of the five samples to each of five laboratories for analysis. Each laboratory provided analysis of the concentration of nitrate (mg/L), and phosphate (μg/L), and the number of zooplankton present in each sample.

Statistical analysis

The data were analyzed using a Generalised Linear Mixed Model where invertebrate count (assumed to be Poisson distributed) was the response variable, and Temperature, Nitrate and Phosphate were treated as continuous explanatory variables, Landscape (2 levels: Urban and Rural) and Flow (3 levels: Low, Medium and High) as categorical, and Lab (1-5) and River (1-4) as random effects.

The models included all the fixed and random effects, and three interactions hypothesized to be important: Flow x Nitrate, Flow x Urban, and Temperature x Nitrate.

The fitted model was:

$$\ln(f_i) = c + m_T x_{T,i} + (m_N + \gamma_j)x_{N,i} + m_P x_{P,i} + \alpha_j + \beta_k + m_{T:N} x_{N,i} x_{T,i} + \delta_{jk} + R_l + L_m$$

Where f_i (i = 1..240) is the fitted value for invertebrate count, c the intercept, m_T, m_N and m_P the slopes relating Temperature, Nitrate and Phosphate to Chlorophyll, γ_j represents the continuous-categorical interaction between Nitrate and Flow (j = Low, Medium and High), α_j the adjustment to the intercept for Flow, β_k the adjustment for Landscape (k = Rural, Urban), $m_{T:N}$ the continuous interaction between Nitrate and Temperature, δ_{jk} the categorical interaction between Landscape and Flow, and R_l (l = *R1, R2, R3, R4*) and L_n (n = 1..5) the random effects of River and Lab.

{with model selection}

This initial model was subject to model selection using likelihood ratio tests (LRTs) with interactions tested first, and then main effects not represented in retained interactions in the order that they increasingly impacted on the likelihood (using the drop1 command). Models were fitted in R (v4.3.0) and lme4 (v1.1-33). Inference was conducted on the final model using LRTs, and where appropriate post-hoc tests conducted in emmeans (v1.8.5). Standard tests to check the assumptions of the models were met were conducted using the DHARMa package (v0.4.6).

Results

The mean number of invertebrates in each sample was 4.91 (range 0-30). Following the removal of non-significant terms (nitrate, temperature and phosphate and the interactions they were included in) the final model was:

$$\ln(f_i) = c + \alpha_j + \beta_k + \delta_{jk} + R_l + L_m$$

There were no diagnostic anomalies with the model fit. The interaction of flow and landscape was significant (χ^2 = 24.38, df = 2, p < 0.0001), with fewer invertebrates in the samples obtained from sites with higher flow rates, however this reduction was much less marked in urban sites relative to rural sites (Fig. 24.3).

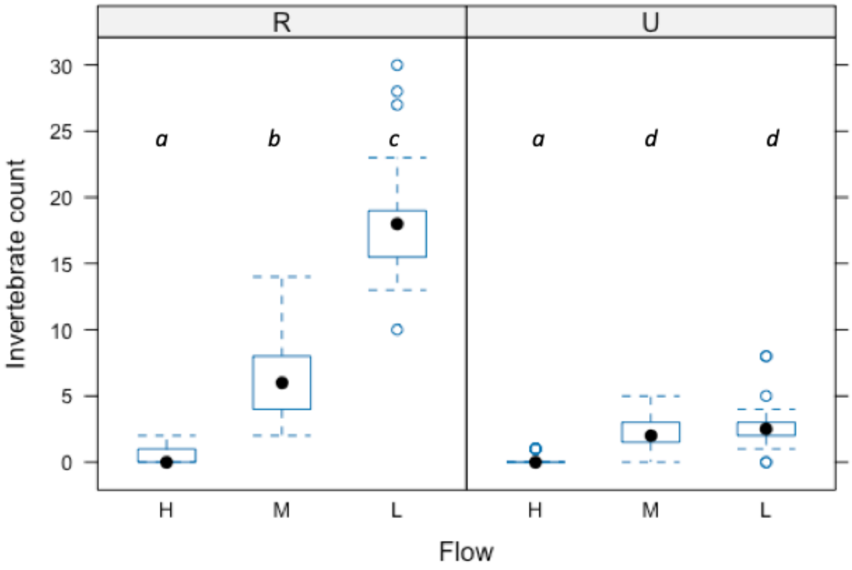


Figure 24.3. The relationship between invertebrate count for flows (High, Medium and Low) in Rural (R) and Urban (U) landscapes. The different letters (*a*-*d*) indicate which counts are different to which other counts (i.e. they are all significantly different to each other except the High flows, and the Medium and Low flows in Urban landscapes.

Post-hoc tests revealed that invertebrate counts from all combinations of flow and landscape differed significantly from each other except high flows, and medium and low flows in urban landscapes (Fig. 24.4).

Commentary: The figure was prepared using the package `lattice` and the `bwplot()` command. Model fitting worked better using the `glmmTMB` command in the `glmmTMB` package. Overdispersion can be checked using the `overdisp.glmer()` in the package `RVAidememoire`.

24.8 Binary data: Invertebrate health

Introduction [invertebrate health]

Should conclude with a statement like:

In this study our overarching goal is to understand the physical and chemical determinants of key biological indicators of water quality. Specifically, the influence of phosphate, nitrate, temperature, speed flow and landscape on invertebrate fungal infection, and whether river had an influence on infection prevalence. [Each of these variables should have been motivated by critical analysis of the literature, earlier in the introduction.]

Methods [invertebrate disease prevalence]

Data collection *{this is just bare bones ... you would more complete details of rivers, sites, and how different flows and landscapes were defined}*

Data were acquired from 12 sites (S01-S12) distributed along 4 different rivers (R1-R4). Sites were stratified opportunistically between flow rates that were low, medium or high, and between stretches of river that flowed through urban and rural landscapes (you would need more detail on the sampling strategy). Temperature of the river water at a depth of 10 cm was recorded in situ, and five samples of river water (between 5 and 10 ml) were taken from each site and stored at 4˚C prior to dispatching one of each of the five samples to each of five laboratories for analysis. Each laboratory provided analysis of the concentration of nitrate (mg/L), and phosphate (μg/L), and the number of zooplankton present in each sample. The zooplankton sampled were examined for evidence of fungal infection and recorded as either infected or uninfected.

Statistical analysis {with model selection}

The data were analyzed using a Generalised Linear Mixed Model where zooplankton infection status was treated as a binary variable (the sample contained infected zooplankton or it did not) and the response variable, and Temperature, Nitrate and Phosphate were treated as continuous explanatory variables, Landscape (2 levels: Urban and Rural) and Flow (3 levels: Low, Medium and High) as categorical, and Lab (1-5) and River (1-4) as random effects.

The model was formulated by starting with a most plausibly complex model including all the fixed and random effects, and three interactions hypothesized to be important: Flow x Nitrate, Flow x Urban, and Temperature x Nitrate.

The fitted model was:

$$\ln\left(\frac{p_i}{1-p_i}\right) = c + m_T x_{T,i} + (m_N + \gamma_j) x_{N,i} + m_P x_{P,i} + \alpha_j + \beta_k + m_{T:N} x_{N,i} x_{T,i} + \delta_{jk} + R_l + L_m$$

Where p_i (i = 1..240) is the fitted value for the probability governing the binary outcome describing invertebrate health (diseased or not diseased), c the intercept, m_T, m_N and m_P the slopes relating Temperature, Nitrate and Phosphate to Chlorophyll, γ_j represents the continuous-categorical interaction between Nitrate and Flow (j = Low, Medium and High), α_j the adjustment to the intercept for Flow, β_k the adjustment for Landscape (k = Rural, Urban), $m_{T:N}$ the continuous interaction between Nitrate and Temperature, δ_{jk} the categorical interaction between Landscape and Flow, and R_l (l = *R1, R2, R3, R4*) and L_n (n = 1..5) the random effects of River and Lab.

This initial model was subject to model selection using likelihood ratio tests (LRTs) with interactions tested first, and then main effects not represented in retained interactions in the order that they increasingly impacted on the deviance. Models were fitted in R (v4.3.0) and lme4 (v1.1-33). Inference was conducted on the final model using LRTs, and where appropriate post-hoc tests conducted in emmeans (v1.8.5). Standard tests to check the assumptions of the models were met were conducted using the DHARMa package (v0.4.6).

Results

The mean infection prevalence was 60.4%. The final model included only the main effects of temperature and flow and was:

$$\ln\left(\frac{p_i}{1-p_i}\right) = c + m_T x_{T,i} + \alpha_j + R_l + L_m$$

There were no diagnostic anomalies with the model fit.

Temperate has a highly significantly positive effect on infection prevalence (χ^2 = 48.52, df = 1, p < 0.0001) with an odds ratio of 4.06 (95% confidence intervals (CIs) 2.62-6.69). Flow rates also had a highly significant influence on prevalence (χ^2 = 87.03, df = 2, p < 0.0001), with the lowest fitted prevalence occurring at high flows (16.2% at average temperature), and medium flows, followed by medium flow (51%) and the highest prevalences at low flows (86.4%) as indicated in Fig. 24.4. The odds ratios for medium and low flow relative to high flow were 15.04 (95% CIs: 5.52 – 47.26) and 115.86 (95% CIs: 32.83-518.21) respectively.

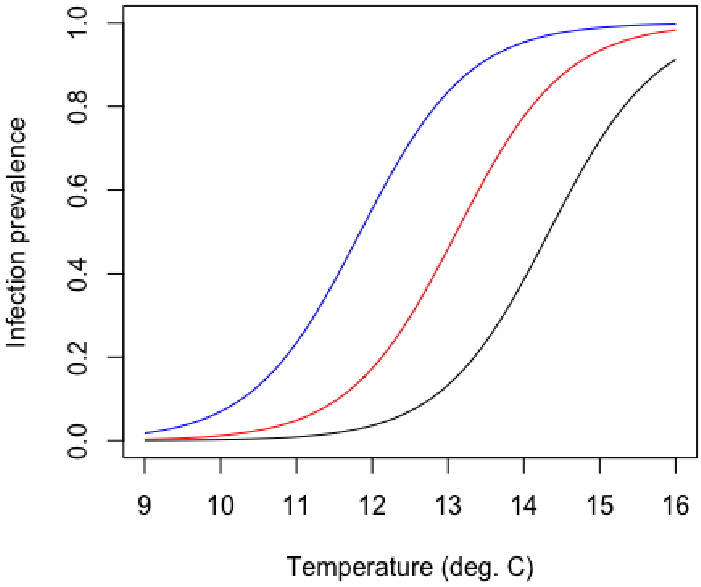


Figure 24.4. The relationship between Infection prevalence and temperature at different flow rates (High flow – black line; Medium flow – red line, Low flow – blue line).

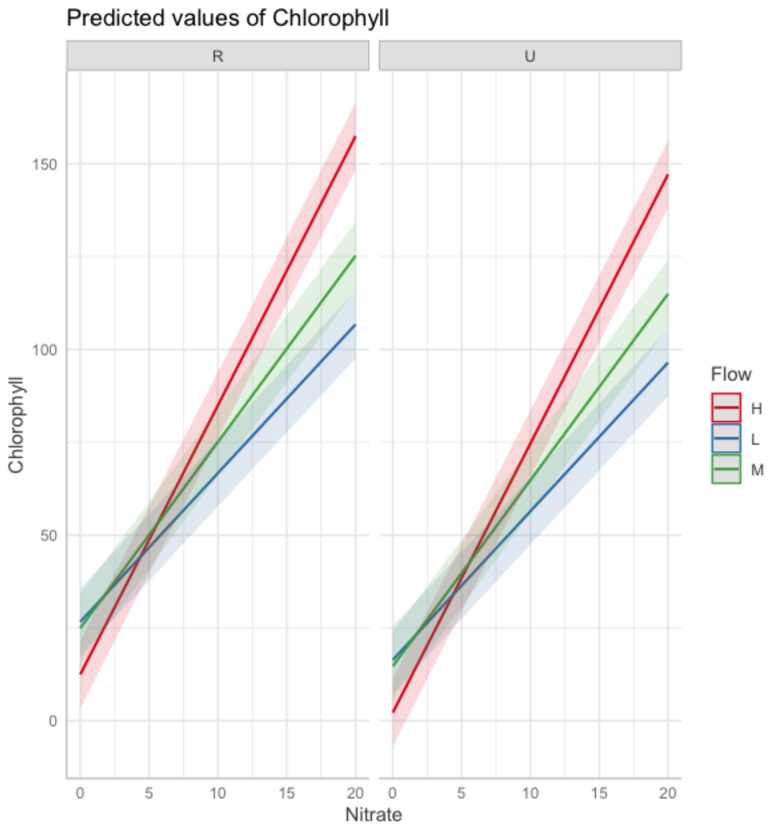
The random effect of river was not significant according to the LRT suggesting infection prevalence was not influenced by river (χ^2 = 1.23, df = 1, p < 0.268).

Commentary: Binary data suggests the results could be described using odds ratios. It is perhaps less important to show the raw data when they are binary, but they could be added as a `points()` layer with vertical jitter to avoid superimposition. This last analysis is an interesting example of the first most complex plausible model indicating none of the terms are significant, and the importance of flow and temperature only becomes apparent in the simpler models. The significance of the random effects can be explored in the usual way by dropping the term and comparing model likelihoods with an LRT. The figure is again generated by constructing a hypothetical sequence of temperatures from the minimum to the maximum observed (using the `seq()` command), and calculating the fitted values from the coefficients for each of the flow rates.

There will be a range of packages that facilitate the construction of figures from models, but care should be taken to ensure you know exactly what fit is being plotted. It is useful to be able to generate figures from a fundamental understanding of the algebraic structure of the model, even if the coding is a bit clunky.

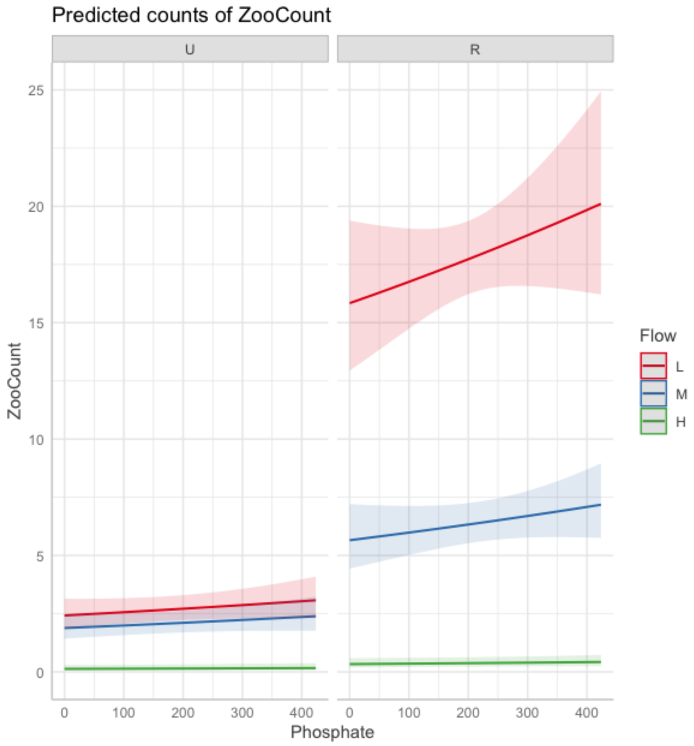
Plotting fitted values for complex models can be challenging. A very useful tool for making this easier is `ggpredict` in the package `ggeffects`.

```
m1<-lmer(Chlorophyll~Temp+Nitrate+Phosphate+Flow+Landscape
         +Flow:Nitrate+(1|River)+(1|Lab),
         REML=FALSE,data=my_data)
gp1 <- ggpredict(m1, terms = c("Nitrate","Flow","Landscape"))
plot(gp1)
```

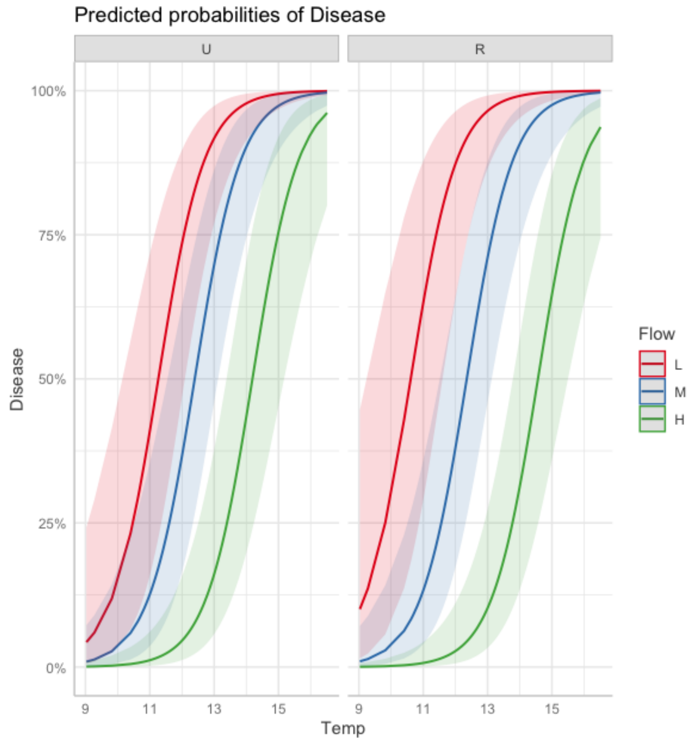


It may also be a good idea to plot the data on top of these model fits, this can be achieved with geom_points(), but this is not a text on R graphics, you can explore the functionality of `gg` packages elsewhere.

```
m2<-glmmTMB(ZooCount~Temp+Nitrate+Phosphate+Flow+Landscape
      +Flow:Nitrate+Flow:Landscape+(1|River)+(1|Lab),
      data=my_data,family=poisson)
gp2 <- ggpredict(m2, terms = c("Phosphate[all]","Flow","Landscape"))
plot(gp2)
```



```
m3<-glmmTMB(Disease~Temp+Nitrate+Phosphate+Flow+Landscape
      +Flow:Nitrate+Flow:Landscape+(1|River)+(1|Lab),
      family=binomial,data=my_data)
gp3 <- ggpredict(m3, terms = c("Temp[all]","Flow","Landscape"))
plot(gp3)
```



Postscript Part 2

There are a range of frequentist inferential frameworks you might use (for example, *z*- or *T*- tests on coefficients, LRTs to compare models, confidence intervals on coefficients, AIC), and correctly applied and interpreted they are all legitimate approaches to learning from and interpreting data. *On the whole, your conclusions will not depend much on what approaches you choose. In the event that it does make a difference, you should be very cautious indeed*, as this would suggest there is something borderline about the inference. You would not want to have the veracity of a scientific claim resting on something as marginal as the supposed superiority of one of these methods of analysis over another. Do not search too hard for 'statistical significance'. Even if your p-value is just below 0.05, there is still a surprisingly high chance of a 'false positive risk' (for a sobering analysis of the false positive risk check out Colquhoun, D. (2017) *The reproducibility of research and the misinterpretation of p-values.* R. Soc. Open Sci. https://doi.org/10.1098/rsos.171085). Likewise, if your p-value is just above 0.05, be mindful of the power of your study, before concluding the absence of any effect. Do not be concerned about so-called 'negative results', or concluding that more data or more powerful studies are required.

Remember, your study will not *prove* anything, and it never could. But it will add to the balance of evidence in support of a hypothesis, be it a null or an alternative one. This in itself is important .. it is how science works. No one can expect more of you!

Appendix A

A reminder about logarithms

We can express any number as a base raised to an exponent: $y = x^{z}$.

x is the base, and z the logarithm of y to base x.

So, using a base of 10, we'd have $\log_{10}$ of

$0.01 = 10^{-2}$ so $\log_{10}(0.01) = -2$

$0.1 = 10^{-1}$ so $\log_{10}(0.1) = -1$

$1 = 10^{0}$ so $\log_{10}(1) = 0$

$10 = 10^{1}$ so $\log_{10}(10) = 1$

$100 = 10^{2}$ so $\log_{10}(100) = 2$

$1{,}000 = 10^{3}$ so $\log_{10}(1{,}000) = 3$

$10{,}000 = 10^{4}$ so $\log_{10}(10{,}000) = 4$

$100{,}000= 10^{5}$ so $\log_{10}(100{,}000) = 5$

$1{,}000{,}000 = 10^{6}$ so $\log_{10}(1{,}000{,}000) = 6$

$16.773 = 10^{1.2246}$ so $\log_{10}(16.773) = 1.2246$

We can reverse this process by so-called exponentiation:

$\mathrm{Exp10}(-2) = 10^{-2} = 0.01$

$\mathrm{Exp10}(-1) = 10^{-1} = 0.1$

$\mathrm{Exp10}(0) = 10^{0} = 1$

$\mathrm{Exp10}(1) = 10^{1} = 10$

$\mathrm{Exp10}(2) = 10^{2} = 100$

$\mathrm{Exp10}(3) = 10^{3} = 1{,}000$

$\mathrm{Exp10}(4) = 10^{4} = 10{,}000$

$\mathrm{Exp10}(5) = 10^{5} = 100{,}000$

$\mathrm{Exp10}(6) = 10^{6} = 1000{,}000$

$\mathrm{Exp10}(1.2246) = 10^{1.2246} = 16.773$

For mathematical reasons we don't need to worry about just now, we usually use natural logs that use a base of 2.718282, a number which is referred to as '*e*'. So we'd have $\log_{e}$, or as is often written ln (which stands for natural logarithm).

$0.01 = e^{-4.60517}$ so $\log_{e}(0.01) = -4.60517$

$0.1 = e^{-2.302585}$ so $\log_{e}(0.1) = -2.302585$

$1 = e^{0}$ so $\log_{e}(1) = 0$

10 = $e^{2.302585}$ so $\log_e(10)$ = 2.302585

100 = $e^{4.60517}$ so $\log_e(100)$ = 4.60517

1,000 = $e^{6.907755}$ so $\log_e(1{,}000\ )$= 6.907755

10,000 = $e^{9.21034}$ so $\log_e(10{,}000)$ = 9.21034

100,000= $e^{11.51293}$ so $\log_e(100{,}000)$ = 11.51293

1,000,000 = $e^{13.81551}$ so $\log_e(1{,}000{,}000)$ = 13.81551

16.773 = $e^{2.81977}$so $\log_e(16.773)$ = 2.81977

And as before, we can reverse this process by so-called exponentiation:

$e^{-4.60517}$ = 0.01

$e^{-2.302585}$ = 0.1

e^{0} = 0

$e^{2.302585}$ = 10

$e^{4.60517}$ = 100

$e^{6.907755}$ = 1,000

$e^{9.21034}$ = 10,000

$e^{11.51293}$ = 100,000

$e^{13.81551}$ = 1,000,000

$e^{2.81977}$ = 16.773

We can write $e^{2.81977}$ or exp(2.81977) – they mean the same thing.

Note that in R, log() returns the natural log, and log10() returns the log to base 10.

Note also how when we take the log of a number less than one for any base ... the log will be negative. The more negative the logarithm of a number, the smaller the number will be. So more negative log likelihoods reflect smaller likelihoods.

Note that logs of numbers that are not positive don’t exist. Log(0) or log(-1) will generate an error regardless of the base used.

Recall also that log(a x b x c x d) = log(a) + log(b) + log(c) + log(d)

And that

exp(a + b + c + d) = exp(a) x exp(b) x exp(c) x exp(d)

(back to Contents)

Appendix B

A word about scientific notation

You will often see numbers expressed like this: 3.24e-08.

We can convert this representation to something that may look more familiar to you by moving the decimal point 8 places to the left: 0.0000000324.

If the number were 3.24e+08, we'd move the decimal place to the right: 324000000.0

(back to Contents)

Appendix C

Rounding numbers to a specified number of decimal places

We wouldn't usually want to write out a number in a scientific report to too many decimal places (3 or 4 usually), so we need to round them off. The rule being if the last digit to be reported is more than half-way to the next highest, we'd round it up to this higher number.

So if we want to round 0.035727 to 3 decimal places we note that the 5 in the 3rd decimal is followed by a 7, so the 5 is nearer to a 6 than a 5, so we round it to 0.036. The Table shows more examples.

Table C.1. Examples of numbers rounded to different numbers of decimal places.

Original number	**To 4 decimal places**	**To 3 decimal places**	**To 2 decimal places**
0.057592945	0.0576	0.058	0.06
0.015729148	0.0157	0.016	0.02
0.185229682	0.1852	0.185	0.19
6.372890528	6.3729	6.373	6.37
-4.999967835	-5.0000	-5.000	-5.00
0.000004725	< 0.0001	< 0.001	< 0.01
0.003639727	0.0036	0.004	< 0.01

When writing a report, determine how many decimal places to report and remain consistent throughout. A journal will have guidance or at least obvious practice about how they want numbers presented. Three is generally reasonable.

(back to Contents)

Appendix D

Limited range continuous data

It is perfectly obvious that a great variety of response variable data we call continuous and might choose to model with a Normal distribution is not. For example, it may be that negative values are not possible (note that if we choose to re-scale data by subtracting the mean – so the mean becomes zero, then about half the data may become negative). Or it may be that some upper limit exists; a common example might be percentage data which may be bounded between 0 and 100. None of this really matters so long as there are not many observations *right on* the boundaries. *The critical requirement is that having fitted the model, the residuals look approximately normally distributed*. So long as this is the case, you don't need to worry about limits to the range of the response variable. Although be very careful when making predictions that extrapolate beyond the range of the observed explanatory variables.

If the response variable data are continuous, and of limited range, and the data falls over the entire range, then one option is to model them using a beta distribution. Beta distributions are defined from 0-1, but there is no reason you couldn't rescale the data to be between 0-1.

Beta distributed data can be modelled using a variety of different R packages, for example `betareg`.

(back to Contents)

Appendix E

Wrapped (or circular) distributions

A surprising number of data types turn out to be circular in the sense that 11pm is close to 1am, December is close to January, a compass bearing of 355 degrees is close to one of 5 degrees (see for example the turning angles shown in Fig. E1). And just as these quantities are in some sense wrapped – with the end joining back up with the beginning, so we can wrap the left-hand of a distribution round to the right-hand end. The most common example would be a wrapped Normal distribution, known as a Von Mises distribution. Like a Normal distribution, Von Mises distributions are continuous distributions and have the same two arguments as a Normal distribution: a mean and a variance, and these arguments can be made to depend on various explanatory variables just as in a regular GLM. Other distributions are often used to model circular data – for example the wrapped Cauchy distribution (also continuous). They perform similarly.

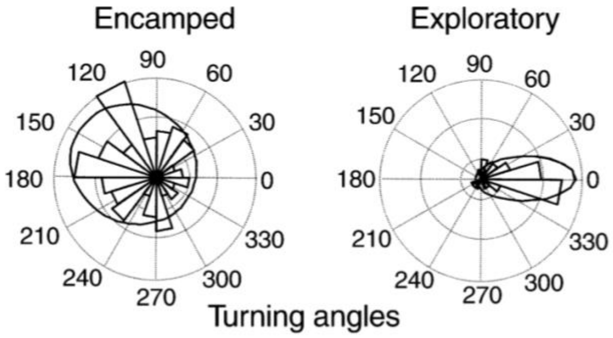


Figure E.1. Examples of circular data. The turning angles between daily steps taken by GPS collared elk (bars) and a circular distribution with different means depending on whether the elk is in an ‘encamped’ phase when the animal turns back on itself a lot, or an ‘exploratory’ phase when the turning angles are closer to zero, and the animal travels in straighter lines.

There are various R packages that model circular data, but the implementation isn’t quite as straightforward as a regular GLM command. Check out `lm.circular` in the `circular` package, or the `CircNNTSR` package.

(back to Contents)

Appendix F

Paired data

(what follows will make more sense once you have completed chapters 1-16)

It is not uncommon to have observations of the response variable that are somehow paired. For example, a measurement of an individual before and after some treatment (perhaps a diet, or course of medication). These sorts of experiments are powerful because by recognizing the paired nature of the data, the variability in the individuals that is *not* due to whatever the treatment is can be factored out. The slight complication is that this is a form of repeated measures – multiple observations from the same individual, and this introduces a potential correlation between observations of the response variable that must be accounted for.

For example, suppose there are 50 individuals observed before and after a course of treatment of some sort, so 100 observations altogether. In our notation, we'd have the response variable y_i (i = 1 .. 100), an explanatory categorical variable (say) ID with 50 levels – one for each individual, and an explanatory categorical variable BEF_AFT with 2 levels – before or after.

There are two ways this situation may be approached. We could model y_i directly and include BEF_AFT as a fixed effect with 2 levels (represented here by α_j , j = before or after) and ID as a random effect with 50 levels (represented here by I_k , k = 1 ..50):

$$y_i = c + \alpha_j + I_k$$

Alternatively, we could calculate the difference $y_{diff,j}$ between the measurements of the response variable for each individual before and after the treatment, our model might be:

$$y_{diff,j} = c \qquad\qquad j = 1 .. 50$$

There is now one observation of the response variable per individual, so no repeated measures, and the variation between individuals that is not related to the treatment is factored out by focusing only on the *difference* between before and after the treatment per individual. If the treatment has no effect we expect c = 0. If we are confident c is different to zero, the treatment is having some sort of effect.

(back to Contents)

Appendix G

Probability density functions – a bit more technical

Here are some further notes and observations on probability density functions in general, and more specifically.

G.1 Some generalities

Strictly speaking, only continuous variables are described by probability density functions, and discrete variables (like integers) are formally described by **probability *mass* functions (pmfs)**. Being discrete, pmfs are not characterized by curves in the same way as a normal distribution might be (Fig G.1A), but by discrete steps (Fig G.1B), such that each integer has a fixed 'likelihood'. The area 'under the curve' is in both cases conserved to be equal to 1. However, there is a difference.

Because non-negative integers are discrete variates the 'likelihood' of each is in fact a probability (for example, the probability of generating a '2' from a Poisson distribution with mean of 3 is 0.224 – as is evident from Fig G.1B), and the 'area under the curve' is the sum of the probabilities of all the possible variates added together – which of course is equal to 1. This is true of all discrete distributions, thus we can refer not to the likelihood of each discrete element but the probability.

Real numbers are continuous, and cannot by definition be 'discretized'. It isn't possible to read-off the probability of say 0.001682759230523 from Fig. G.1A, or compare the *probability* of generating 0.001682759230523 with the probability of generating 0.001682759230524. However, it is possible to read-off the likelihood of generating 0.001682759230523 from a Nomal distribution with (say) mean 0 and variance 0.0001 (its 39.33 - as is evident from Fig G.1A). Note how in Fig G.1A the likelihoods exceed 1; this is not common but it's perfectly possible if the variance is small enough (note that the narrower the distribution is … the taller it needs to be given the area is conserved to equal 1). Likelihoods are not probabilities and while they must be positive, they are not bounded between 0 and 1. However, we can generate probabilities from pdfs by *integrating* between two values: for example, the *probability* of generating a number between -infinity and 0 from a Normal distribution with mean of zero is ½. Of course, we can integrate discrete distributions as well (for example, the probability of generating a random number greater than 5 from a Poisson distribution with mean of 3 is 0.084).

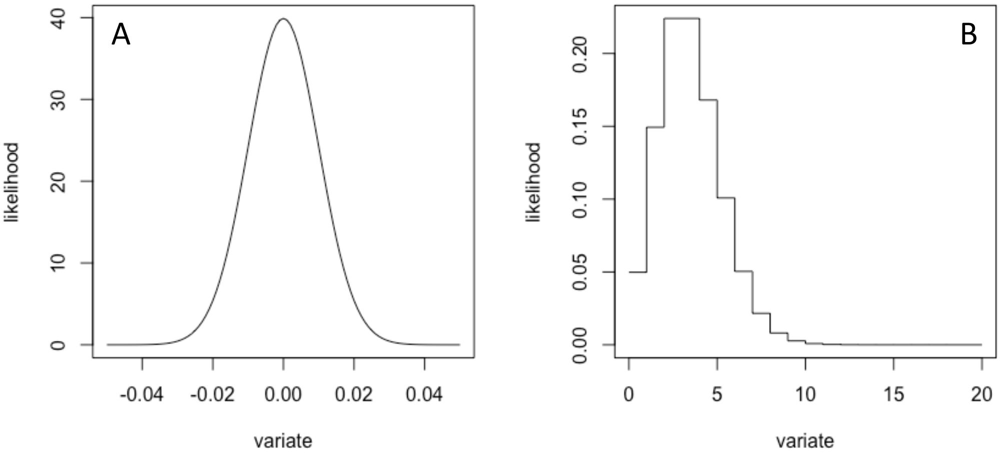


Figure G.1. A Normal distribution with mean = 0 and variance =0.0001. A Poisson distribution with mean and variance = 3.

G.2 Working with pdfs and pmfs in R

R has some useful functions for studying probability density and mass functions. They are organized into families prefixed with `r`, `d`, `q`, and `p`. `r_` generates random numbers, `d_` returns the likelihoods (or probabilities) of a specified value, `q_` returns the value corresponding to a specified area between the far left and the specified value, and `p_` returns the area to the left of a specified value. So, for a Normal distribution `rnorm(),dnorm(),qnorm() and pnorm()`. For a Poisson distribution `rpois(),dpois(),qpois() and ppois()`. For a Uniform distribution `runif(),dunir(),qunif() and punif()`, and so on. Fig. G.2 summarizes what these different functions do, and you can look up the arguments they require by prefixing these commands with a question mark (e.g. `> ?dnorm)` .

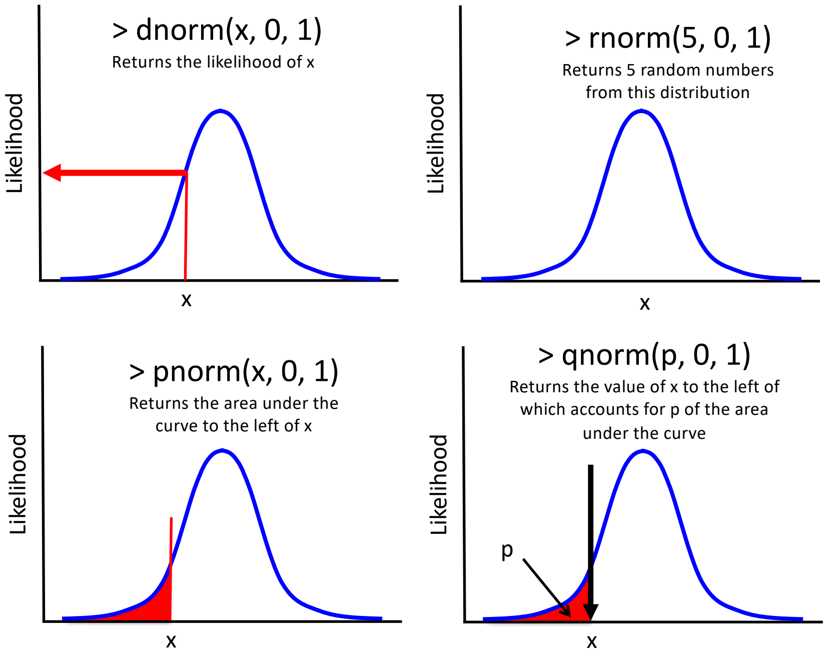

> dnorm(x, 0, 1)
Returns the likelihood of x
Likelihood
x
> rnorm(5, 0, 1)
Returns 5 random numbers from this distribution
Likelihood
x
> pnorm(x, 0, 1)
Returns the area under the curve to the left of x
Likelihood
x
> qnorm(p, 0, 1)
Returns the value of x to the left of which accounts for p of the area under the curve
Likelihood
p
x

Figure G.2. A schematic illustrating the different applications of four R commands to pdfs and pmfs.

R-code like this can be used to plot and explore different pdfs and pmfs

```
x=seq(-0.05,0.05, 0.001)
plot(x, y=dnorm(x,0,0.01), type='l', ylab='likelihood', xlab='variate')

x=seq(0,20, 1)
plot(x,y=dpois(x,3),type='s',ylab='likelihood',xlab='variate')
```

G.3 Why is the Normal distribution so common?

The Normal (or Gaussian) distribution is encountered so often in statistics because so many biological variables do seem to be Normally distributed. There is a reason for this, and it's called the **Central Limit Theorem** (CLT). The CLT shows how the sum of a number of variables ... regardless of how each of these variables is distributed, will be Normally distributed. For example, the sum of 8 variables, each from a different Uniform distribution, will be Normally distributed. The sum of 15 variables, 8 from different Gamma distributions and 7 from different Poisson distributions, will be Normally distributed. Which is a bit like saying that if we have a variable that itself depends on a lot of other things ... its likely to be Normally distributed too. Human height depends on about 50 genes, most likely in an additive way, and so unsurprisingly human height is well described by a Normal distribution.

There are lots of good applets on-line that demonstrate this nicely. For example: http://195.134.76.37/applets/AppletCentralLimit/Appl_CentralLimit2.html

G.4 Why is the Poisson distribution so common?

If events happen at a constant rate, then the number of events observed over a fixed period of time would be Poisson distributed. Famously, this applies to the number of alpha particles emitted from a radioactive source in a fixed time interval. But biologically, this result is important to us also. If wildebeest walk past a lookout post at a fixed rate, the number of wildebeest recorded should be a Poisson variate. Poisson events may be observations of something through time, or over space. If animals are distributed randomly (and this does not mean precisely uniformly) in a landscape, and you wander randomly around this landscape, or along a transect, the number of animals encountered should be Poisson distributed. The number of animals counted in fixed quadrats (of any scale) should be Poisson distributed if their density is constant per unit area.

G.5 Why is the Negative Binomial distribution more common?

If the rate that events happen is in fact not constant, but itself varies (say according to a Gamma distribution), then the number of events observed is Negatively Binomially distributed. That is to say, the argument of the conventional Poisson distribution is itself a Gamma variate (indeed, another name for the Negative Binomial distribution is the Gamma-Poisson distribution). The Poisson distribution is in fact a special case of the Negative Binomial distribution, but since in ecology and epidemiology rates are bound to vary a bit, often quite a lot (the term ‘aggregation’ is quite often used to describe this form of heterogeneity), the use of the Negative Binomial distribution is common (for example in describing parasite counts within hosts where some hosts contain *a lot* more parasites than others).

There are many different ways of expressing a Negative Binomial distribution, and it is a famously confusing issue. In Chapter 4 we define the arguments of a Negative Binomial distribution to be the mean (μ), and a parameter we call *k*, and which glm.nb (or glmer.nb) calls theta, and is also sometimes called the size parameter. The variance of the Negative Binomial distribution expressed this way is given by:

$$var\ NB(\mu,k) = \frac{\mu^2}{k} + \mu$$

As *k* becomes smaller the Negative Binomial distribution can account for higher levels of heterogeneity (more ‘clumpiness’ or aggregation of the counts). As *k* becomes very large the variance reduces to the mean, and the Negative Binomial distribution converges on a Poisson distribution.

G.6 What is the relationship between a Bernoulli and a Binomial distribution?

A binomial variate is generated when you toss a coin *N* times, each with an independent probability *p* of generating say a head. The number of heads will be Binomially distributed, and obviously is bounded between 0 and *N*. A Bernoulli variate is generated when you toss the coin just once (i.e. *N* = 1). Obviously the outcome must be either a head or a tail, so the number of heads is either 0 or 1. So a Bernoulli distribution is a special case of a Binomial distribution when *N* =1.

A simple Binomial distribution applies if the probability *p* of (say) a head is the same for all *N* tosses of the coin. When observing binary data, it is quite possible, likely indeed, that the probability of a yes, a pass, a plus, a success ... whatever, might be different for each observation of the response variable. Therefore, we suggest it makes more sense to think of observations of a binary response variable coming from multiple (by definition, single) Bernoulli trials with different probabilities of success, than a Binomial distribution with *N* trials.

G.7. Other useful pdfs

While not commonly encountered it is useful to know about:

Beta distributions. Should you be modelling probability (proportions or percentages) directly, Beta distributions are continuous distributions with two arguments that influence the mean and variance and are bounded on 0-1.

Gamma distributions. A distribution comprising non-negative continuous, defined by two arguments, pleasingly flexible, and often used for waiting times.

Log-normal distributions. What if instead of a quantity (say *x*) that was normally distributed, the logarithm of *x* was normally distributed? This would be a Log-normal distribution. While *x* must be positive (since we are considering the distribution of log(x) and we can't take a log of a negative number), log(*x*) is defined between minus infinity and plus infinity. Just as normal distributions arise when a variable is related to the sum of many random quantities, a log-normal distribution arises when a variable is related to the product of many random quantities (remember that the log of the product of *a***b***c***d***e* … = log(*a*) + log(*b*) + log(*c*) + log(*d*) + log(*e*) ... and the central limit theorem will apply to this sum - albeit a sum of logged quantities!).

(back to Contents)

Appendix H

Use of subscripts

It is critical that you understand how subscripts work.

We use subscripts when we have a variable or parameter that is required to be indexed to something. The subscripts occupy a fixed *place-holder position* (i.e. 'first subscript' or 'second subscript', more rarely 'third subscript').

For example, we often use *f* to indicate the fitted value on the left-hand side of a general linear model, but each use of *f* refers to a fitted value for a specific observation of the response variable (i.e. the 3rd one, or the 11th one …). Thus we introduce a single place-holder position, and it may be occupied by any integer value from one to the number of observations of the response variable in the data set (usually denoted '*n*'). So, for example, if we had 10 observations we'd need an *f* for f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, and f_{10}. This gets a bit tedious … so we write instead:

$$f_i \qquad i = 1 .. 10.$$

Which means the subscript *i* may take values from 1 to 10. ' *i* ' is a 'generic' *place-holder filler.* The important point is there is nothing special about the use of '*i*', We could have used *k* and it would have meant exactly the same thing. The letter you use is a matter of personal choice. It doesn't even have to be the same letter, the key point is a given place-holder position always refers to the same thing (in this case linking *f* to a particular observation of a response variable). So … denoting *f* this way changes nothing.

$$f_k \qquad k = 1 .. 10.$$

If I have a model with an explanatory variable with four levels, I'd need four adjustments in my model (one for each level but of course for the reference level the adjustment will be zero), which I might denote α_j:

$$f_i = c + \alpha_j \qquad \text{(Eq H.1)}$$

And now *j* could take values denoting the four different levels (they could be called say {red, white, blue, green}), *j* could take on any of the values in the set defined within { .. }. So Eq. A.1 can take on four different values to model 10 different observations of the response variable.

If we now have a second explanatory variable with say two levels, we're going to need an additional two adjustments – one for each of these levels. We're going to need a different letter (say β) to denote the different variable, and we would choose a different placeholder letter for the subscript so we can denote that it has only two possible values and not four (say *k*, but it could be anything so long as you haven't used it to refer to something else).

$$f_i = c + \alpha_j + \beta_k \qquad \text{(Eq. H.2)}$$

Where *k* takes two different values {male, female}. So Eq. A.2 can take on eight different values to model 10 different observations of the response variable.

Now imagine we also wanted to fit an interaction, i.e. a specific adjustment for every combination of each of the explanatory variables. We'd need eight different adjustments (4 x 2), a new letter to denote the interaction term (say γ) and two subscript place holders to denote which level of the first explanatory variable it refers to and which level of the second explanatory variable it refers to. But we have these place-holder fillers already defined (they are *j* and *k*), so as long as we define which explanatory variable the first place holder refers to (it doesn't matter which so long as we're consistent) and which the second refers to, all will be clear.

$$f_i = c + \alpha_j + \beta_k + \gamma_{j,k} \qquad \text{(Eq. H.3)}$$

It is good practice to define the range of each subscript, thus to indicate:

i = 1 .. *n*; *j* = {red, purple, blue, green}; *k* = {male, female}

And thus, we might say the 7th response variable (*i* = 7), associated with the first explanatory variable being 'green' (*j* = green) and the second 'male' (*k* = male) would be denoted:

$$f_7 = c + \alpha_{green} + \beta_{male} + \gamma_{green,male} \qquad \text{(Eq. H.4)}$$

We would see subscripts used in relation to continuous explanatory variables also. In the simplest case each fitted value of the response variable (f_i), is associated with a continuous explanatory variable x_i (and now it is notable that we use *i* for both, because the *ith* fitted value links directly to the i^{th} explanatory variable). For both the fitted and explanatory variable *i* must run from 1 to *n*. We might have:

$$f_i = c + m\,x_i \qquad \text{(Eq. H.5)}$$

(note: $m\,x_i$ indicates multiplication of *m* (the slope) and x_i the numerical value of the explanatory variable)

Things would get a little more complicated if we had two continuous explanatory variables. It is convention to use *m* for slopes, and *x* for continuous explanatory variables, but now that we have two of them, we have to distinguish between them with subscripts:

$$f_i = c + m_1\,x_{1,i} + m_2\,x_{2,I} \qquad \text{(Eq. H.6)}$$

So now the first place-holder for *x* refers to which explanatory variable it is (the first one or the second one), and the second place-holder refers to which observation of the response variable it is indexed to.

If we had a lot of continuous explanatory variables (say p of them) we could condense this notation by writing:

$$f_i = \mathrm{c} + \sum_{j=1}^{n} m_j x_{j,i} \qquad \text{(Eq. H.7)}$$

Where the capital sigma indicates 'sum' of the products to the right.

(back to Contents)

Appendix J

Off-setting

Off-setting is most often used to control for something that has a one-to-one influence on the response variable. Suppose we collected a slightly variable amount of water in each of our river samples, and the count of zooplankton would likely depend in a straightforward way on this volume and we'd obviously need to account for that. One option would be to divide the count by the volume and analyse the concentration as the response variable. But a better solution is to offset.

If all the samples were of the same volume we could just analyze count in the usual way, say (with explanatory variables Flow and Nitrate), something like:

$$\log{(f_i)} = c + \alpha_j + m_N x_{N,i}$$

But if the volume of the samples are not the same, we'd offset, which would require a model such as:

$$\log{(f_i)} = c + \alpha_j + m_N x_{N,i} + m_V \log{(x_{V,i})}$$

Where $x_{V,i}$ is the volume of the i^{th} sample, and m_V is required to be fixed at a value of 1. Note that we can rewrite Eq. J.2 as

$$\log{(f_i)} - \log{(x_{V,i})} = c + \alpha_j + m_N x_{N,i}$$

(because $m_V = 1$). And note that $\log{(f_i)} - \log\left(x_{V,i}\right) = \log{(f_i / x_{V,i})}$ which is the log of the count per unit volume.

To offset in R we'd simply write

```
> model = glm(Zoocount ~ offset(Volume) + Flow + Nitrate,
data=my_data, family = Poisson)
```

And we are modelling the data as it was collected, which is in general preferable.

(back to Contents)

Appendix K

Zero-inflated and hurdle models

Poisson and Negative Binomial models will generate zero's but once the arguments are specified (albeit conditioned on explanatory variables), the probability of generating zero's is determined. For example, if the argument (the mean and variance) of a Poisson distribution is 3.5, then the probability of a zero will be 0.03 as say obtained from `dpois(0,3.5)`. If the response variable contains more zero's than predicted from the best fitting distribution, it would be called zero-inflated. Extreme zero-inflation can be spotted by visual inspection of the data, but it more subtle cases should be identified through careful diagnostic checking and residual analysis. While Negative Binomial distributions can be effective at accommodating overdispersion (more variation than we'd expect from say a Poisson distribution), they are not in general suitable for dealing with zero-inflation.

Zero-inflation can be modelled with a mixture of two distribution: a Bernoulli distribution used to model excess zero's or 'something else', and a second discrete distribution (some sort of Poisson or Negative Binomial) to model the 'something else'. The arguments of both distributions can be conditioned on the explanatory variables. If the second distribution can also generate zero's (as regular Poisson and Negative Binomial distributions can) then the models are called **zero-inflated models**, but if they cannot (because they might be **zero truncated Poisson** or **zero truncated Negative Binomial** distributions) then they are termed **hurdle models**.

The `zeroinfl` command in the `pscl` package makes this easy to implement. For example, if we wanted to model possible zero inflation in ZooCount in this way, where the probability of generating a zero or a Poisson variate might be dependent on Temperature, and the mean of the Poisson distribution dependent on Landscape.

```
>m1<-zeroinfl(ZooCount ~ Landscape | Temp, dist = 'poisson', data = 
my_data)

>summary(m1)

Call:
zeroinfl(formula = ZooCount ~ Landscape | Temp, data = my_data, dist 
= "poisson")

Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)   2.4262     0.0682  35.573   <2e-16 ***
LandscapeU   -1.9086     0.2008  -9.503   <2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) -14.2995     7.7321  -1.849   0.0644 .
Temp          0.9361     0.5385   1.739   0.0821 .
---
Log-likelihood: -130.5 on 4 Df
```

Remember that the usual link functions apply: the logit link function for probability of generating a zero or a Poisson variate, and the natural log for the mean of the Poisson distribution.

(back to Contents)

Appendix L

Odds Ratios

L.1 Odds ratio's in the absence of any interactions

Odd ratios are only used with models fit to binary data, but are a good way of describing effect sizes. For a particular combination of explanatory variables (say situation 'A') the model will generate a probability p_A of a '1' (however a '1' was defined, it might have been a 'yes', 'positive', 'pass', 'survived', or whatever), and 1- p_A, the probability of the alternative outcome (the '0', or 'no', 'negative', 'died' etc). The odds of a '1' is defined as p_A /(1- p_A). If p_A = 0.9 then the odds of a '1' under situation 'A' would be 0.9/0.1 = 9. A '1' is 9 times more likely than a '0' under situation 'A'.

The odds ratio under situation 'B' would be given by p_B /(1- p_B). We can summarize our results by describing the ratio of the odds as we move from say situation 'B' to situation 'A'. That is, the odds ratio is given by:

$$\frac{\frac{p_A}{1-p_A}}{\frac{p_B}{1-\ p_B}}$$

Recall that the glm with logit link function will take the form:

$$\log\left(\frac{p}{1-p}\right) = c + adjustments$$

then

$$\frac{p}{1-p} = \exp(c + adjustments)$$

Which of course are the odds. If situation 'A' and 'B' correspond to different levels of a categorical explanatory variable (say situation B corresponds to the reference level), then we'd have:

$$\text{odds ratio} = \frac{\exp(c + \alpha_A + other\ adjustments)}{\exp(c + 0 + other\ adjustments)}$$

which equals

$$\frac{\exp(c)\ \exp(\alpha_A)\ \exp(other\ adjustments)}{\exp(c)\ \exp(0)\ \exp(other\ adjustments)}$$

which equals $\exp(\alpha_A)$.

We can perform a similar calculation for a continuous explanatory variable. For example, comparing the odds at a Temperature of say $x_{T,i}$ degrees and the odds at Temperature $x_{T,i}$+1 degrees:

$$\log\left(\frac{p}{1-p}\right) = c + m_T x_{T,i} + other\ adjustments$$

so

$$\frac{p}{1-p} = \exp(c + m_T x_{T,i} + other\ adjustments)$$

And comparing the odds of two temperatures that differ by 1 degree:

$$\text{odds ratio} = \frac{\exp\ (c + m_T(x_{T,i} + 1) + other\ adjustments)}{\exp\ (c + m_T x_{T,i} + other\ adjustments)}$$

which equals

$$\frac{\exp(c)\ \exp\left(m_T x_{T,i}\right)\ \exp(m_T)\ \exp(other\ adjustments)}{\exp(c)\ \exp\left(m_T x_{T,i}\right)\ \exp(other\ adjustments)}$$

which equals $\exp(m_T)$. Note how the odds ratio remains unaffected by what the increase in 1 degree is *from*, that is to say .. whether we are comparing 6 degrees to 5, or 21 degrees to 20.

It doesn't matter 'which way round' the odds ratio is calculated (what situation appears in the numerator and which in the denominator) so long as the interpretation is correct – are the odds increasing or decreasing with the change in 'situation'. The odds ratio corresponds to the change *from* the denominator 'situation' *to* the numerator 'situation'.

Inference can be performed on odds ratios by constructing and inspecting their confidence intervals. Various packages will do this, but of course the CIs of the odds ratio can be obtained by exponentiating the CIs of the relevant coefficients.

In R the command `or_glm` in the package `oddsratio` will do these calculations for you. For example were the fitted model to be:

```
model_3<-glm(Disease ~ Flow + Temp, family = binomial, data = my_data)

> or_glm(my_data, model_3, incr = list(Temp = 1), ci = 0.95)
  predictor oddsratio ci_low (2.5) ci_high (97.5)          increment
1     FlowL    52.504        4.894       1245.115 Indicator variable
2     FlowM    24.871        2.767        434.630 Indicator variable
3      Temp     5.092        2.276         15.896                  1
```

The odds of zooplankton being diseased at Low flow increase 52.5 fold relative to reference (High flow), and by 24.8 fold at Medium flows relative to reference (High flow); and 5.1 fold per 1 degree increase in Temperature.

(note that or_glm won't work for mixed models unless generated in `glmmPQL` (in the `MASS` package).

L.2 Odds ratios when there are interactions

Odds ratios are a concise way of capturing the effect sizes when modelling binary data, but things get a good deal messier in the presence of interactions. You will often see R packages reporting exponentiated interaction coefficients, but unlike exponentiated main effect coefficients these have no simple interpretation. Best to just think of an odds ratio as describing the change in odds relating to two different situations. For example, suppose we'd fitted the model:

```
model_4<-glm(Disease ~ Flow + Temp + Flow:Temp,
                   family = binomial, data = my_data)
```

We could calculate the odds ratio for any two situations:

Say when Flow was Low and Temperature was 8 degrees; and Flow was High and Temperature was 8 degrees.

Or when (say) Flow was Low and Temperature was 12 degrees; and Flow was High and Temperature was 12 degrees.

(Note these two odds ratios will not be the same because the effect of changes in Flow now *depends* on the temperature – *because* of the existence of the interaction)

Or a situation when (say) Flow was Low and Temperature was 8 degrees; and Flow was High and Temperature was 12 degrees.

We can compare any two situations using an odds ratio .. interaction or not. However, in the presence of an interaction it's probably easiest to just calculate the odds (the exponentiated linear predictor) for each of the respective situations and inspect the ratio directly rather than use a package.

(back to Contents)

Appendix M

The cbind trick

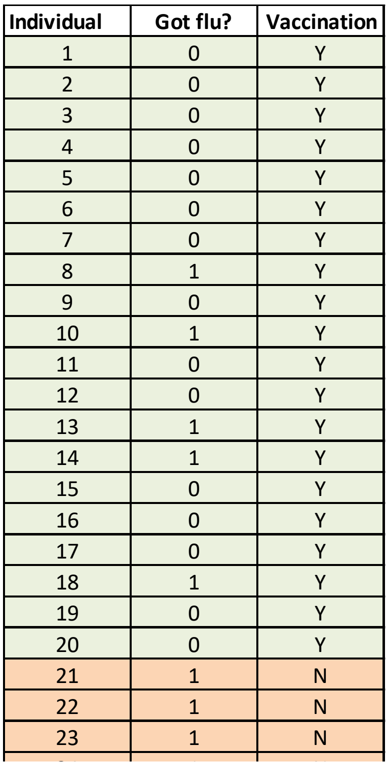
Suppose the response variable was binary and the explanatory variables all categorical. An example of such data would be those shown in Table 3.1a, where data for 40 individuals are shown, each row of the data table is a record indicating whether the person had contracted 'flu, and whether or not they have received a 'flu vaccination (categorical two levels). Rather than tediously entering 40 rows of data, we could summarize the data as:

Vaccinated	Infected	Not_infected
Y	5	15
N	15	5

These data could then be analyzed with the command:

```
m_short<-glm(cbind(Infected,Not_infected)~Vaccinated,
      family=binomial,data=my_data)
```

and the results would be exactly the same as if the data had been entered in long format. The cbind trick automatically weights the data appropriately according to the sample size.



Individual	Got flu?	Vaccination
1	0	Y
2	0	Y
3	0	Y
4	0	Y
5	0	Y
6	0	Y
7	0	Y
8	1	Y
9	0	Y
10	1	Y
11	0	Y
12	0	Y
13	1	Y
14	1	Y
15	0	Y
16	0	Y
17	0	Y
18	1	Y
19	0	Y
20	0	Y
21	1	N
22	1	N
23	1	N

etc …

(back to Contents)

Appendix N

The derivation of the interaction term for two continuous explanatory variables.

We saw in Chapter 8 how we can include two continuous explanatory variables in the same model. For example, we might be interested in the continuous explanatory variables Temperature and Nitrate:

$$f_i = c + m_T x_{T,i} + m_N x_{N,i} \qquad i = 1 .. 48$$

If we wanted to ask whether the effect of Nitrate on Chlorophyll depended on Temperature (or conversely and synonymously, effect of Temperature on Chlorophyll depended on Nitrate), we'd just want to make an adjustment to the effect of Nitrate depending on the Temperature – that is an adjustment to the slope *m_N*, so it is different for different values of Temperature, and reciprocally an adjustment to the effect of Temperature depending on the Nitrate concentration – that is an adjustment to the slope *m_T*, so it is different for different values of Nitrate

$$f_i = c + (m_T + (m_{N\_on\_T} x_{N,i}))\, x_{T,i} + (m_N + (m_{T\_on\_N} x_{T,i}))\, x_{N,i} \qquad i = 1 .. 48$$

We can multiply out the brackets:

$$= c + m_T x_{T,i} + m_{N\_on\_T} x_{N,i}\, x_{T,i} + m_N x_{N,i} + m_{T\_on\_N} x_{T,i}\, x_{N,i}$$

And then rewrite with $m_{T:N} = m_{T\_on\_N} + m_{N\_on\_T}$

$$= c + m_T x_{T,i} + m_N x_{N,i} + m_{T:N} x_{T,i}\, x_{N,i}$$

So, the interaction is represented by just one additional coefficient *m_{T-N}* which we multiply by the product $x_{T,i}\, x_{N,i}$.

(back to Contents)

Appendix P

What is Principal Components Analysis (PCA)?

PCA has a number of applications in quantitative analysis. Here we shall focus on its role in helping to simplify a situation where one has a large number of potentially correlated explanatory variables.

Suppose that in addition to *n* observations of the response variable (y_i*, i* = 1 .. *n*) one also has a series of *p* explanatory variables (x_{ij}, *j* = 1 ..*p*). These explanatory variables might be continuous or categorical, it doesn't matter, but easiest to imagine them as continuous. Suppose for the time being that *p* is just 2: x_1 and x_2 (note the response variable doesn't feature in this or what follows). Suppose these together look as in Fig Z1.A. Note x_1 and x_2 are correlated (positively in this case). We will 'centre' and standardize both these variables by subtracting off the respective means and dividing by the respective standard deviations to arrive at Fig. Z1.B. We can see that most of the variation in the explanatory data is explained by a combination of x_1 and x_2 that runs from 'south-west' to 'north-east' across the graph.

If we were Martians, we might have 'seen' this combination ourselves (with our rather different vision systems and perception of the world) as a single combined variable, but we are dumb humans and we see it as two. But even dumb humans can recognize that the 'real' things that matters here is the combination – shown by the red axis – y_1 – in Fig. Z1.C. In fact .. y_1 is a linear combination of x_1 and x_2 which we can write:

$$y_1 = m_{11}x_1 + m_{12}x_2 \qquad \text{Eq. 1}$$

Or for a specific value of y_1:

$$y_{1,i} = m_{11}x_{1,i} + m_{12}x_{2,i} \qquad \text{Eq. 2}$$

And we can choose a second axis, y_2, that must be 'at right angles' to y_1 (in this case being only a 2-dimensional data set, there is only once choice for y_2).

$$y_2 = m_{21}x_1 + m_{22}x_2 \qquad \text{Eq. 3}$$

Or for a specific value of y_1:

$$y_{2,i} = m_{21}x_{1,i} + m_{22}x_{2,i} \qquad \text{Eq. 4}$$

In other words, we can redefine the axes of our original graph (Fig. ZB) changing the reference from x_1 and x_2 to y_1 and y_2 (Fig. Z1.D). And the coordinates of all our data from x_{ij} to y_{ij} using Eqs 2 and 4. You can see that this amounts to a (linear) rotation of the data. We haven't changed the relative positioning of the data points .. simply their orientation with respect to a different references system. The new variables y_1 and y_2 are known as the first and second **Principal Components** or PC1 and PC2, and the *m*-coefficents are known as **loadings**. The y_{ij} values are known as scores. Because this is only a 2-dimensional data set there are only two Principal Components (PCs). If we had *p* variables, we'd have a *p*-dimensional data set, *p* PCs.

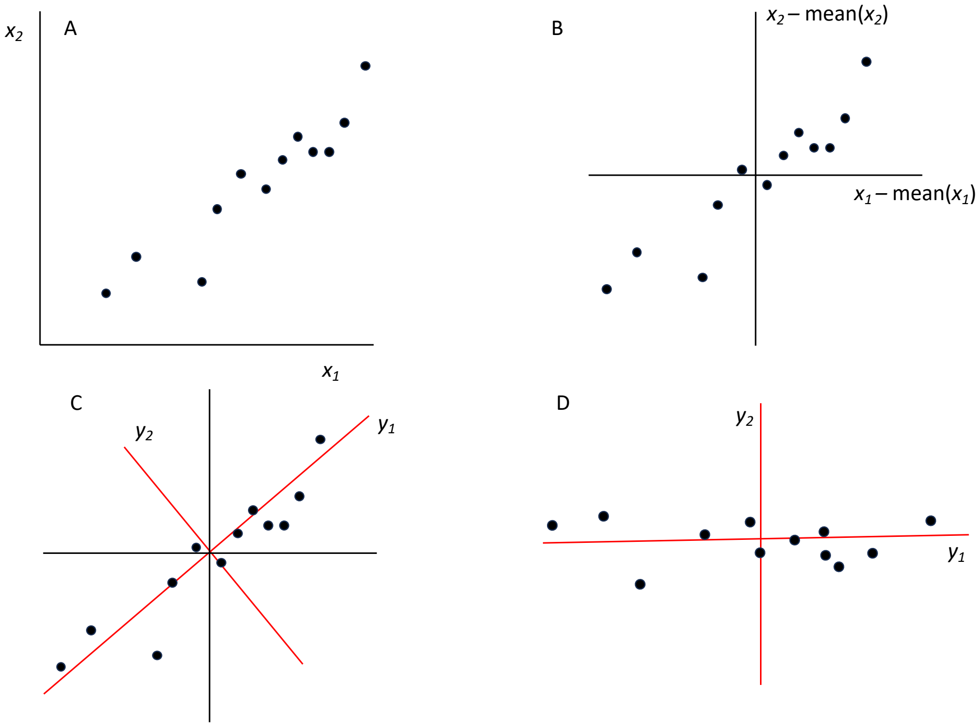


Figure Z1. A: The original two explanatory variables x_1 and x_2; B: after centering; C: identification of the Principal Components y_1 and y_2; D: the rotated data in their new reference frame.

A PCA analysis will deliver three important things:

First, the percentage of the variation explained by each PC. PC1 will by definition explain the most, then PC2, then PC3 .. and last PC*p*. If the explanatory variables are correlated a good deal of the total variation might be explained by just a few (usually between one and three) of the PCs. Then, rather than having to worry about a *p*-dimensional analysis (having to enter *p* explanatory variables into your GLM), you could just enter one, two, or three PCs instead, knowing that these capture the bulk of the variation in the explanatory variables.

Second, the downside of PCA Is that while it may mean we can worry about fewer dimensions, its less clear what these dimensions really mean, or how they relate to our original dimensions (the *x*'s). The loadings help us out here. The coefficients m_{11} and m_{12} can be plotted as a point on Fig. Z1.D and a line drawn from the origin through this point. This line will show how the original variable x_1, relates to our new variables y_1 and y_2. Likewise, the coefficients m_{21} and m_{22} can be plotted as a point on Fig. Z1.D and a line drawn from the origin through this point, and this line will show how the original variable x_2, relates to our new variables y_1 and y_2. These 'vector plots' are known as bi-plots. When the different vectors run very close to each other, it reflects a high positive correlation between the corresponding explanatory variables. When they run at right angles to each other, they indicate no correlation between the associated explanatory variables. When they head off in entirely opposite directions it reflects a high negative correlation between the associated explanatory variables. The length of these vectors are often scaled to reflect their importance in explaining variation.

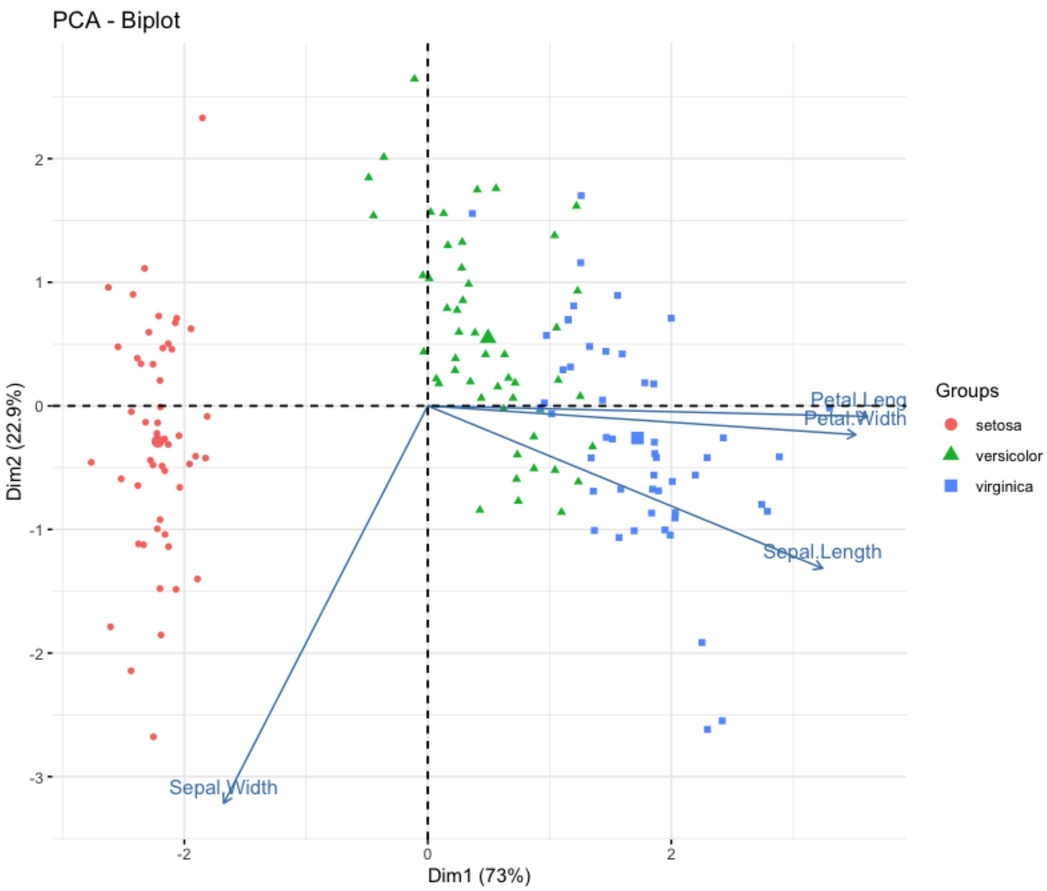
Third, we have the scores. It is hard to visualize the structure of the variation captured by your explanatory variables – we can't plot the data in any more than 2 or may be 3 dimensions .. and *p* may be a lot greater than 2 (or 3). But if most of the variation is captured by PC1 and PC2, then this is much more easily plotted, and you may find interesting groupings and relationships between your *n* observations of the sets (**records**) of explanatory variables.

To summarize:

- The original data will be a matrix, with *p*-columns (one for each variable) and *n*-rows (one for each observation of the response variable), plus column and row labels of course.
- For conventional PCA (princomp in R) $n > p$. If not, you can continue using prcomp but you'll only get the first *n* PCs.
- Output will include a breakdown of the individual and cumulative variation explained by the (up to *p*) PCs.
- A *p* by *p* matrix of loadings
- A *n*-rows and *p*-column matrix of scores.
- You may be offered a choice of whether you wish to use a correlation or covariance matrix for your PCA. If you centre and standardize the data, it won't make any difference. If you don't centre and standardize, then use of a correlation matrix will standardize for you, the choice of covariance matrix won't.

Confusingly, in prcomp the loadings are called **rotations**, and the score '**x**'.

Here is a somewhat typical output of PCA:



The data set comprised p = 4 continuous measurements of iris flower dimensions (petal and sepal length and width) from 3 different species of iris (50 observation of each species of iris so n = 150 in all). PC1 accounts for 73% of the variation in the explanatory variables, and PC2 22.9%. So PC1 and PC2 account for 95.9% of the variation in total. The figure shows the 'rotated' scores for the 150 iris flower observations plotted in PC1-PC2 space. As you can see the 3 different species fallout quite nicely into 3 groups (red, blue and green). The arrows indicate that PC1 (on the x-axis) is very closely associated with Petal length and width, which themselves are closely positively correlated (as you might expect). Sepal width is a more even mixture of both PCs, but 'at right angles' to Petal width and length indicating it is largely uncorrelated with these two measurements. Sepal length is largely independent of sepal width, but more closely correlated with petal dimensions.

(you can look at this example yourself using this website: https://statisticsglobe.com/biplot-pca-r)

(back to Contents)

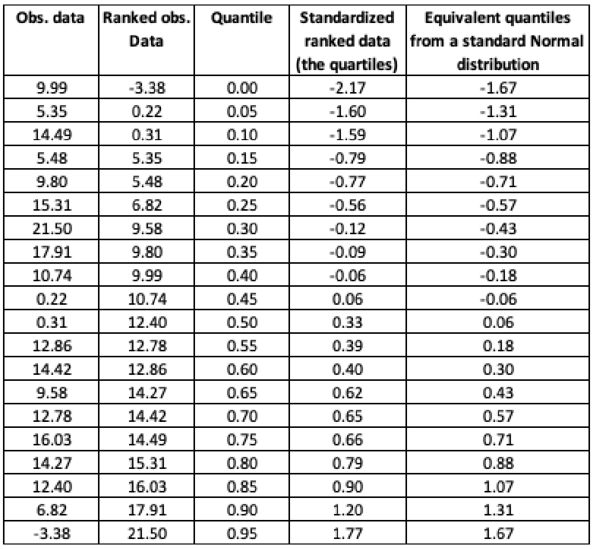
Appendix Q

qqplots – a closer look

A quantile-quantile plot compares the quantiles of one distribution with the quantiles of another. If the two distributions are the same, then the points should fall on the line x = y. Often one of these two distributions is chosen to be a purely statistical one (very often the Normal distribution for example), and the other may be your data, or more likely – your residuals (we will just call it data for current purposes).

Suppose you have *n* data points (frequency histogram shown in Fig P.2A). If the data are standardized by subtracting off the mean from each value, and dividing by the standard deviation, your standardized data would have a mean of zero and a standard deviation of one. If the standardized data (y_i) are now ranked from smallest to largest, their values become the $(i\text{-}1)/n^{\text{th}}$ quantiles (i = 1 .. n), and will be plotted on the y-axis of the qqplot. These manipulations are shown in the first 4 columns for Table P.1 where *n* = 20.

Table P.1. Col. 1: 20 data points; Col. 2: the same data ranked; Col. 3. The 20 quartiles when there is one quartile for every data point; Col. 4. The ranked data after standardization. Col. 5: the equivalent 20 quartiles for a Standard Normal distribution.



Obs. data	Ranked obs. Data	Quantile	Standardized ranked data (the quartiles)	Equivalent quantiles from a standard Normal distribution
9.99	-3.38	0.00	-2.17	-1.67
5.35	0.22	0.05	-1.60	-1.31
14.49	0.31	0.10	-1.59	-1.07
5.48	5.35	0.15	-0.79	-0.88
9.80	5.48	0.20	-0.77	-0.71
15.31	6.82	0.25	-0.56	-0.57
21.50	9.58	0.30	-0.12	-0.43
17.91	9.80	0.35	-0.09	-0.30
10.74	9.99	0.40	-0.06	-0.18
0.22	10.74	0.45	0.06	-0.06
0.31	12.40	0.50	0.33	0.06
12.86	12.78	0.55	0.39	0.18
14.42	12.86	0.60	0.40	0.30
9.58	14.27	0.65	0.62	0.43
12.78	14.42	0.70	0.65	0.57
16.03	14.49	0.75	0.66	0.71
14.27	15.31	0.80	0.79	0.88
12.40	16.03	0.85	0.90	1.07
6.82	17.91	0.90	1.20	1.31
-3.38	21.50	0.95	1.77	1.67

If we wanted to compare these quantiles to say those from a standard Normal distribution we would choose *n* quantiles that divided the Normal distribution into *n*+1 equally sized areas. If *n* was – say – 20 we'd have Fig. P.1, and where the red lines intercept with the x-axis would be the theoretical quantiles we'd compare with those of your observed data by plotting them on the x-axis of the qqplot (col 5 of Table P.1).

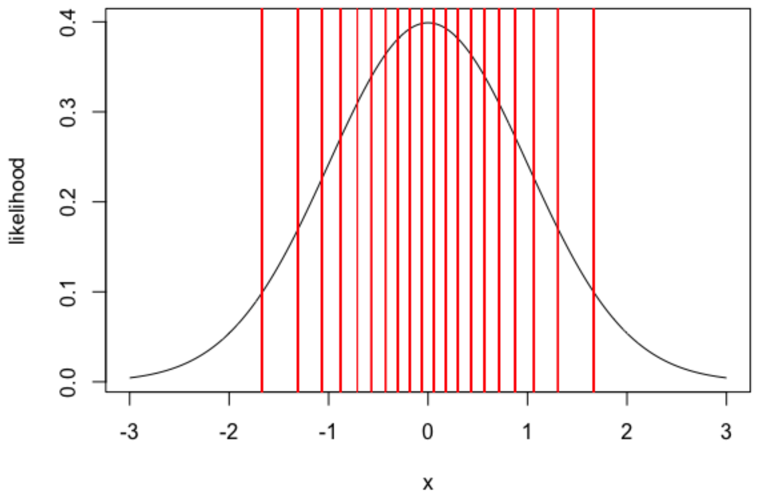


Figure P.1. 20 quantiles of the Normal distribution. The value of each quantile is where the red line intercepts with the x-axis.

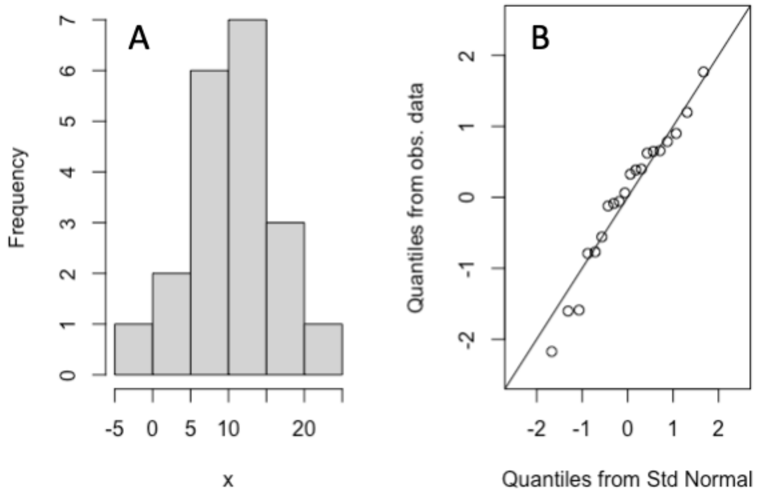


Figure P.2. A) The original observed data. B) qqplot after standardizing the observed data. The points mostly fall close to the x=y line because the data is quite Normally distributed.

Were the original data to be distributed in a way that was less Normal, the qqplot would reveal departure from the x=y relationship. For example Fig. P.3.

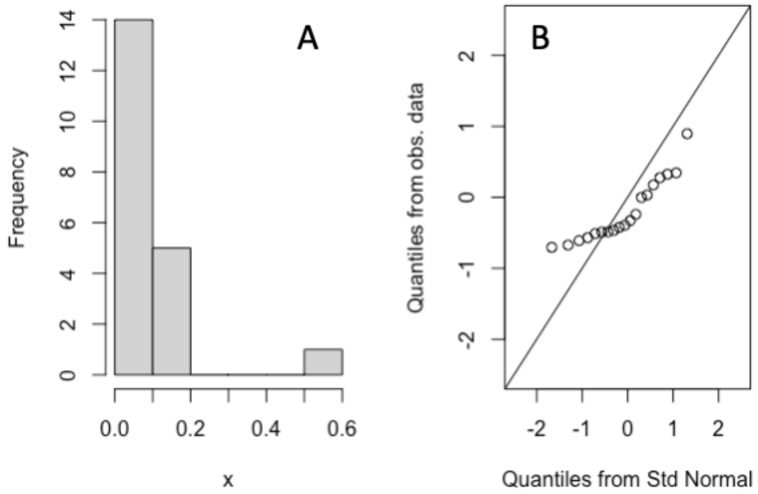


Figure P.3. A) The original observed data. B) qqplot after standardizing the observed data. The points depart substantially from the x=y line because the data is not well described by a Normal distribution.

DHARMa defines residuals in a cunning way so that something akin to qqplots compare the distribution of the response variable with the distribution of pseudo data simulated by the model. It is well described in the DHARMa package notes. If the model is sound, it should be capable for simulating data that looks quite like the data it purports to model! (see Chapter 16 for more details)

(back to Contents)

Appendix R

Variance Inflation Factors (VIFs)

Variance Inflation Factors (VIFs) are a way of measuring the intensity of **collinearity**. Collinearity (or **non-orthogonality**) arises when correlations exist between a models explanatory variables. One way to measure this is to examine the unadjusted R^2 value obtained by constructing a regression model where the i^{th} explanatory variable becomes the response variable and the other explanatory variables are used as explanatory variables (denoted $R^2{}_i$). The VIF_i is estimated as $1 / (1 - R^2{}_i)$. So, if the other explanatory variables are not in anyway correlated with the i^{th} explanatory variable VIF_i is 1, and as the correlation strengthens VIF_i will increase. Collinearity doesn't influence the explanatory power of the model, but it does cause the standard errors of the parameters to increase (hence the term 'variance inflation'), and cause a consequent reduction in associated T statistics, and broaden the confidence intervals, thereby reducing the power of our models to detect significant effects.

There are various 'rules of thumb' that can be applied. VIF_i's that are less than 3 are regarded as unproblematic. But – really VIF_i's are just one way of thinking about collinearity and they do nothing to 'solve' the problem – they just alert you to the fact collinearity is present. Various forms of model comparison (LRTs or AIC), or simply eyeballing the data should provide similar information.

(back to Contents)

Appendix S

The Multinomial distribution

Just as we can have a Binomial distribution in which we perform an ‘experiment’ *N* times, each time with a probability *p* of one outcome and probability 1-*p* of another (which we call a Bernoulli distribution if *N* = 1), so we can conceive of situations in which there are more than two outcomes per experiment. We would then transition from a Binomial distribution to a Multinomial distribution.

Just as when there are two outcomes, we need to define a single probability (and generating the two probabilities *p* and 1-*p* for the two outcomes), so in a multinomial distribution with *M* outcomes, we need *M*-1 probabilities. So we might have four possible outcomes per experiment, with outcome 1 occurring with probability p_1, outcome 2 with probability p_2, outcome 3 with probability p_3, and outcome 4 with probability $1-p_1-p_2-p_3$.

Multinomial GLMs will model the logit probability of the different possible outcomes of the response variable relative to some defined baseline category of the response variable.

$$\log\left(\frac{p_2}{p_1}\right) = c_2 + adjustments_2$$

$$\log\left(\frac{p_3}{p_1}\right) = c_3 + adjustments_3$$

$$\log\left(\frac{p_4}{p_1}\right) = c_4 + adjustments_4$$

The output will contain different adjustments for each of these different equations (i.e. the coefficients of the model are not fixed for the different equations).

Multinomial GLMs can be constructed using the `multinom` command in the package `nnet` using pretty much the usually formatted command.

(back to Contents)

Appendix T

Ordinal GLMs

Instead of a response variable that is binary, it might perhaps have several categories that could in some way be consider ordered. For example, the Likert scales often use 5 or 7 categories expressing some ordered response (from strongly disagree through to strongly agree). Analysis of ordinal data is a little trickier, and serious thought should be given to whether you can 'binarize' your data, but if you can't, you could consider the use of ordinal GLMs.

Suppose the i^{th} observation of your response variable is one of 5 ordered responses: $j = 1..5$, where $1 < 2 < 3 < 4 < 5$. We can model the probability that $y_i \leq j$ (as opposed to $y_i > j$) :

$$\log\left(\frac{p_{y_i \leq j}}{1 - p_{y_i > j}}\right) = c + adjustments$$

And we can proceed with the usual machinery of GLMs.

The interpretation requires care, but the output will enable construction of equations yielding the logit(probability) that $y_i \leq 1$, $y_i \leq 2$, $y_i \leq 3$, $y_i \leq 4$, and $y_i \leq 5$, in terms of your chosen explanatory variables. Inference can be conducted in the usual way.

Ordinal GLMs can be constructed using the typical format with the `polr` command in the `MASS` package.

(back to Contents)

Appendix U

Akaike's information criterion (AIC)

There is much that other people would say about the use of AIC. Here we just provide a few basic remarks. One thing for sure, don't combine LRTs and AIC in the analysis of the same data set. Choose one or the other. Our preference is for inference by LRT, but there isn't in our view, a great deal by which they differ. The one important exception being that unlike LRTs, AIC can be applied to models that are not nested.

AIC is a single number that can be calculated for any model fitted using likelihood, and can be used as a method to select or compare a model.

AIC is defined as -2*LL* + 2*k* where *LL* is the log-likelihood of the model, and *k* is the number of parameters required by the model. Because *LL* are generally negative, -2LL is a positive quantity, and the smaller it is, the better the model fits. Furthermore, other things being equal we prefer simpler models to more complex models, so low *k* is preferred over high *k*. Thus – models with low AIC are preferred to models with higher AIC. Of course, one way to generate models with higher log-likelihoods (lower -*LL*'s) is to increase their complexity (higher *k*'s). AIC balances model fit and model complexity, allegedly identifying models that are optimally complex.

As a rule of thumb, models within 2 AIC units of each other are not regarded as distinguishable, but a model with an AIC more than 2 less than another would be regarded as better supported.

Mathematically, AIC is essentially equivalent to LRTs

Consider two models: M_1, with p_1 parameters, and M_2 with p_2 parameters. M_2 is nested within M_1, and has (say – and without loss of generality) one less parameter.

Likelihood ratio test

The log-likelihood of the data given M_1 is LL_1, and given M_2, LL_2. A likelihood ratio test would use the test statistic:

$2 \times (LL_1 - LL_2)$

and test its significance using a chi-squared distribution with 1 df (for the one parameter difference between the models), and we'd reject the null hypothesis if the test statistic exceeded 3.84.

AIC

Recalling AIC = $-2LL + 2k$

We have $\mathrm{AIC}_{M1} = -2LL_1 + 2p_1$, and $\mathrm{AIC}_{M2} = -2LL_2 + 2p_2$

Compare the two AIC values:

$(-2LL_2 + 2p_2) - (-2LL_1 + 2p_1)$

$= -2LL_2 + 2p_2 + 2LL_1 - 2p_1$

$= 2\,(LL_1 - LL_2) - 2(p_1 - p_2)$

$= 2\,(\mathrm{LL}_1 - \mathrm{LL}_2) - 2 \qquad (\text{if } p_2 = p_1 - 1)$

And we consider M_1 to be better supported than M_2 if AIC_{M1} is smaller than AIC_{M2} by ~2 or more.

Note the similarity between the comparison using LRT and AIC. Using LRTs we'd be asking if:

2 x $(LL_1\text{-}LL_2)$ is more or less than 3.84

and using AIC we'd be asking if:

2 x $(LL_1\text{-}LL_2) - 2$, is more or less than 2?

Or put another way if:

2 x $(LL_1\text{-}LL_2)$, is more or less than 4?

They are essentially identical comparisons.

It is quite common to present a table of AIC values for different models, with the difference between the minimum AIC and the alternatives. For example, the models compared using LRTs in Chapter 20 might be tabled like Table S.1.

Table S.1. An example of a comparison of models undertaken using AIC.

Model	**AIC**	**ΔAIC**
Chlorophyll~Landscape+Flow+Phosphate+Nitrate+ Nitrate:Flow	300.230	-
Chlorophyll~Landscape+Flow+Phosphate+Nitrate+ Temp+ Nitrate:Flow	301.055	0.825
Chlorophyll~Landscape+Flow+Phosphate+Nitrate+ Temp+Phosphate:Landscape+Nitrate:Flow	302.134	1.904
Chlorophyll~Landscape+Flow+Nitrate+ Nitrate:Flow	308.072	7.842
Chlorophyll~Flow+Phosphate+Nitrate+ Nitrate:Flow	325.022	24.792
Chlorophyll~Landscape+Flow+Phosphate+Nitrate+ Temp+Phosphate:Landscape	344.663	44.433

Note the inference is identical to that made in Chapter 20 using LRTs, with the preferred model identified as including Landscape, Flow, Phosphate, Nitrate and the interaction of Nitrate and Flow, regardless of whether LRTs or AIC is used.

(back to Contents)

Appendix V

What is the difference between a standard deviation and a standard error?

All distributions have means and standard deviations. These are fundamental properties *of distributions* that do not depend on sampling. The distribution may be say a Normal probability density function, or a Poisson probability density function (more correctly termed a probability mass function) with defined means and standard deviations, or a set of *n* numbers, y_i, from which we could calculate the mean ($\bar{x}$) and standard deviation (*s*) of these *n* numbers using the following formulae:

$$\bar{x} = \sum_{i=1}^{n} x_i / n \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

The standard deviation reflects the 'width' of the distribution; the larger it is, the more variability the distribution embodies, the more variable the variates (or random numbers) that we could generate from the distribution. The formula for *s* pretty much calculates the square root of the average value of the squared deviation of each value from the overall mean. (Why is it *n*-1 in the denominator? Because the formula assumes the mean has already been estimated, and so there are – so to speak – only *n*-1 remaining degrees of freedom.)

When we estimate a parameter (or coefficient) from data (any parameter ... it could be a simple mean, or a parameter representing an adjustment in a GLM), the estimate is made with uncertainty. The estimated parameter is itself assumed to the mean of a distribution from which the parameter might have come. Most often these distributions are assumed (with good reason) to be Normal distributions, and we can estimate the standard deviations of these distributions *of parameters* using clever mathematics *but we call them standard errors* because the parameters are estimated from a sample.

Standard errors *are* standard deviations, but relate to distributions *of parameters*, not *data*. It would not be incorrect to refer to the standard deviation of a parameter, but standard error might avoid confusion.

Here are 5 random numbers from a normal distribution with mean = 100 and SD = 10.

```
92.565  90.089  73.703 102.910 110.128
```

The mean of this sample of 10 numbers is 93.879 and the standard deviation is 13.863. We'd expect 93.879 to be close to 100, and 13.86 to be close to 10. The estimated standard error of our estimate of the mean is given by $13.86 / \sqrt{5}$ = 6.200. Which is to say that our estimate of the mean of these 5 numbers is normally distributed with a mean of 93.879 and standard error of 6.200. We could construct

95% confidence intervals around this mean by adding and subtracting (approximately) two of these standard errors and we'd see the true mean fall inside of this interval.

Here is a larger sample of 100 random numbers from the same *N*(100,10) distribution

```
95.317 108.514 113.957 104.082 106.894  94.286  91.551  95.639
92.636 106.804 108.897  95.788 119.187 109.998 101.291 109.457
114.806 103.165 122.668  82.800  93.892 113.480  94.403 114.693
112.189 107.957 107.391 104.392  96.267 107.173  95.332  95.540
89.532  96.051  95.658  91.081  86.383  88.288  96.690 112.836
86.052 109.994  98.595 101.843 113.410 110.433  97.957 102.842
100.133  85.465 100.274 107.341  93.043  93.671 104.852  84.275
108.155  72.825 108.893 107.721  91.344  90.234  94.875  88.552
96.588 108.433 107.697 101.220  89.144 105.698  95.884 112.098
105.686 109.251  99.673  88.819  89.593 106.211 101.802 107.493
96.996  97.020  95.471  90.945  93.186 102.573 108.112  98.117
98.803 123.087  82.287 101.399  95.727  91.644 105.874  99.048
101.210  91.207  99.028  80.385
```

The mean of these 100 numbers is 100.071 and the standard deviation is 9.501. Based on this larger sample, we'd expect 100.071 to be (a lot) closer to 100 than the means of samples of just 5 observations, and 9.501 to be closer also to 10. The standard error of our estimate of the mean is given by $9.501/\sqrt{100}$ = 0.950. Which is to say that our estimate of the mean of these 100 numbers is normally distributed with a mean 100.071 and standard error of 0.950.

The standard error reduces as the square-root of the sample size (Fig. S.1), while standard deviations are fundamental inalterable properties not dependent on sample size.

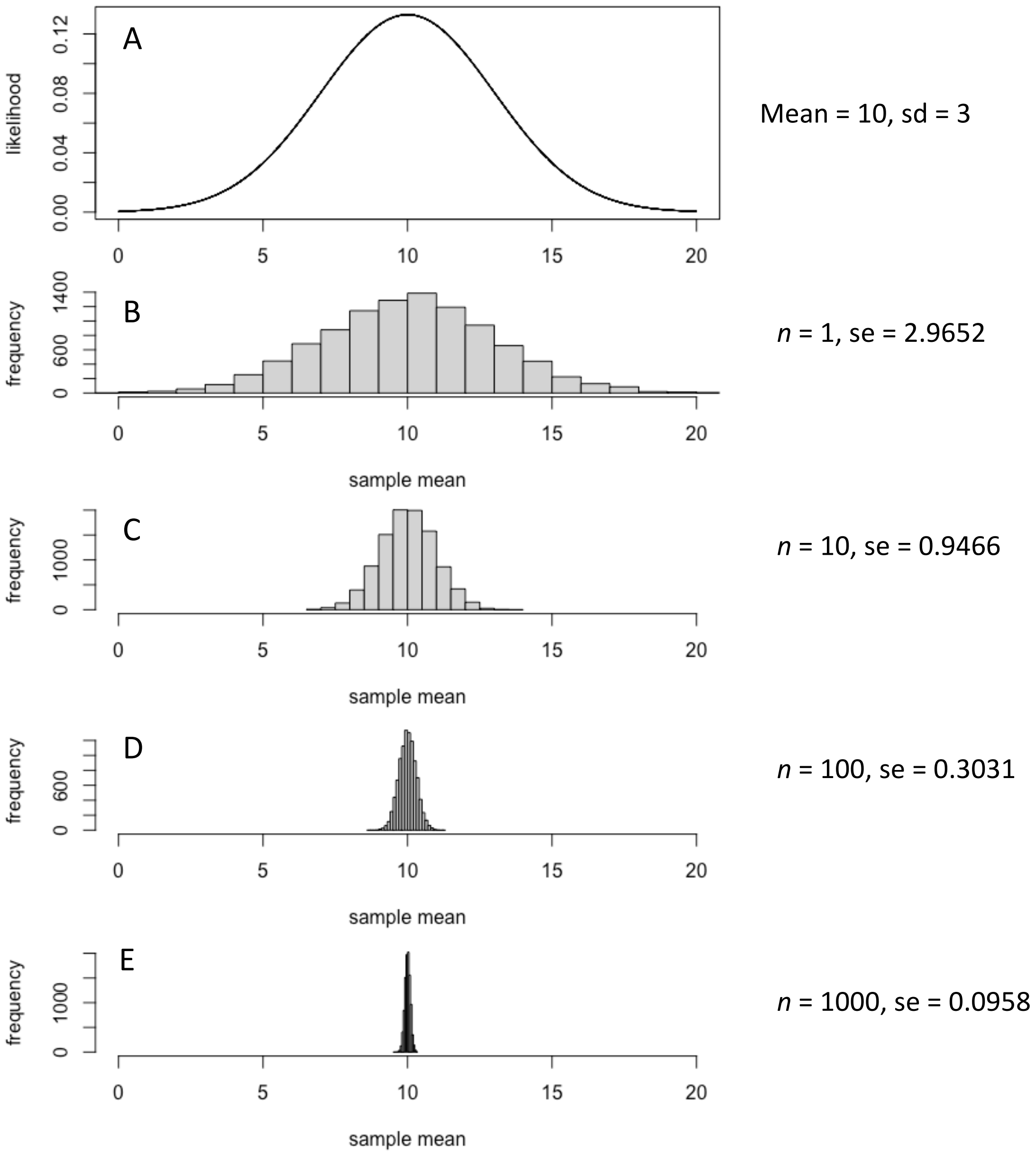


Figure S.1. Demonstration of a shrinking standard error. We take (10,000) samples of *n* variates from the parent distribution (A – Normal distribution, mean = 10, standard deviation = 3). Figures B-E show the distribution of the means of these samples of different sizes. B: the mean calculated from just a sample of 1. Unsurprisingly, this is just a manifestation of the parent distribution. C: means computed from a sample size of 10. D: means computed from a sample size of 100. E: means computed from a sample size of 1000. As *n* increases, the means of each sample become increasingly stable and closer to the mean of the parent distribution, resulting in reduction of the standard error (or standard deviation of the mean). The standard error of distributions in B – E is $s/\sqrt{n}$.

(back to Contents)

Appendix W

Why are LRTs preferred for testing significance of categorical explanatory variables?

Consider a model such as

$$f_i = c + \alpha_j \qquad (j = A, B, C).$$

Suppose the 3 fitted values for were $f_A = 4$, $f_B = 8$ and $f_C = 12$ and the reference level was A. We'd have

$f_A = 4 + 0$, $f_B = 4 + 4$, and $f_C = 4 + 8$.

So $\alpha_B = 4$ and $\alpha_C = 8$. Suppose the standard error on these estimates was 3. This would result in 95% CIs for these coefficients of approximately 4 ± 6 = -2 to 10 and 8 ± 6 = 2 to 14 for α_B and α_C respectively. We would conclude α_C was significantly different to zero, and the effect of our explanatory variable was significant.

However, suppose B had been chosen as the reference level. We'd have

$f_A = 8 - 4$, $f_B = 8 + 0$, and $f_C = 8 + 4$.

So $\alpha_A = -4$ and $\alpha_C = 4$. The standard error on these estimates would still be 3. This would result in 95% CIs of approximately -4 ± 6 = -10 to 2 and 4 ± 6 = -2 to 10 for α_A and α_C respectively. We would conclude that neither coefficient was significantly different to zero, and the effect of our explanatory variable was not significant.

Our inference depends on how we labelled our levels … how crazy is that? The LRT is completely unaffected by the choice of reference level (it depends only on the likelihoods – and note the fitted values aren't changing – so nor would the likelihood), and so provides a more robust approach to inference.

Problematic though this is, it is a problem that could only arise for categorical explanatory variables with more than 2 levels. If your model contains only continuous explanatory variables, or categorical explanatory variables with 2 levels, coefficient analysis and LRTs will generate consistent inference.

(back to Contents)

Appendix X

Post-hoc tests

The p-values that R generates for the coefficients in a GLM are testing the null hypothesis that the coefficients do not differ significantly from zero. For coefficients corresponding to adjustments for different levels of an explanatory variable this amounts to testing the difference between each level and the reference level (where the adjustment is assumed to be zero). But what if we are interested in comparing two different levels, neither of which are the reference? This is entirely possible – each coefficient is estimated with an associated standard error so its relatively trivial to determine if they differ significantly from each other. However, some serious thought should be given as to whether it is really necessary or useful to test all of these different hypotheses. Often just knowing that the main effect is an important driver of variation might be enough.

Post-hoc tests require the use of additional packages. Post-hoc tests can be carried out using many different packages, for example `emmeans` or `multcomp`.

Suppose the model was

```
> m1<-glm(Chlorophyll~Flow+Landscape,data=my_data)
> summary(m1)

Call:
glm(formula = Chlorophyll ~ Flow + Landscape, data = my_data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   72.807      5.440  13.383  < 2e-16 ***
FlowL        -20.207      6.663  -3.033  0.00405 ** 
FlowM        -13.669      6.663  -2.052  0.04620 *  
LandscapeU    -9.711      5.440  -1.785  0.08115 .  
```

So, the summary output compares L(ow) to the reference H(igh), and M(edium) to the reference H(igh), and U(rban) landscapes to R(ural). But not L(ow) to M(edium). The command

```
> emmeans(m1, pairwise ~ Flow)
```

Generates the contrast L(ow) – M(edium) which should be zero if the adjustment for levels L and M do not differ

```
$contrasts
 contrast estimate   SE df t.ratio p.value
 H - L       20.21 6.66 44   3.033  0.0111
 H - M       13.67 6.66 44   2.052  0.1120
 L - M       -6.54 6.66 44  -0.981  0.5924

Results are averaged over the levels of: Landscape 
P value adjustment: tukey method for comparing a family of 3 
estimates
```

And the post-hoc tests indicates that the only significant difference between Chlorophyll levels are between the High and Low Flows. The Tukey method adjusts the p-values for the multiple comparisons that are being made but note that it uses the pooled standard error from the original analysis (i.e. the standard error is the same for all of the comparisons).

```
> emmeans(m1, pairwise ~ Landscape)
```

Would contrast Rural and Urban landscape but there isn't any point in doing this as Rural is the reference level, and so the comparison is already being made in the output of the summary.

Where post-hoc tests are arguably the most useful is to determine the effect sizes and direction of pairwise comparisons when there are significant interactions retained in the model. For example, were the model to be:

```
m1<-glm(Chlorophyll~Flow+Landscape+Flow:Landscape,data=my_data)
```

the interaction coefficients could be compared using

```
emmeans(m1, pairwise ~ Flow:Landscape)

$emmeans
 Flow Landscape emmean   SE df lower.CL upper.CL
 H    R           77.9 6.66 42     64.4     91.3
 L    R           51.7 6.66 42     38.3     65.1
 M    R           55.0 6.66 42     41.5     68.4
 H    U           58.0 6.66 42     44.6     71.5
 L    U           43.8 6.66 42     30.3     57.2
 M    U           53.6 6.66 42     40.2     67.1

Confidence level used: 0.95

$contrasts
 contrast  estimate   SE df t.ratio p.value
 H R - L R    26.19 9.42 42   2.780  0.0809
 H R - M R    22.94 9.42 42   2.435  0.1678
 H R - H U    19.89 9.42 42   2.110  0.3020
 H R - L U    34.11 9.42 42   3.619  0.0096
 H R - M U    24.28 9.42 42   2.576  0.1258
 L R - M R    -3.25 9.42 42  -0.345  0.9993
 L R - H U    -6.31 9.42 42  -0.669  0.9844
 L R - L U     7.91 9.42 42   0.840  0.9582
 L R - M U    -1.91 9.42 42  -0.203  0.9999
 M R - H U    -3.06 9.42 42  -0.325  0.9995
 M R - L U    11.16 9.42 42   1.185  0.8417
 M R - M U     1.34 9.42 42   0.142  1.0000
 H U - L U    14.22 9.42 42   1.509  0.6605
 H U - M U     4.39 9.42 42   0.466  0.9971
 L U - M U    -9.83 9.42 42  -1.043  0.9006

P value adjustment: tukey method for comparing a family of 6
estimates
```

The contrasts compare the adjustments for all possible combinations of levels to each other, and tests whether the difference between these adjustments is significantly different to zero.

(back to Contents)

Appendix Y

What is a design matrix?

Consider a GLM with say two continuous explanatory variables:

$$f_i = c + m_1 x_{1,i} + m_{2,i} x_{2,i}$$

Suppose the data say was very small .. perhaps just 6 records. We could write an equation for each of the 6 fitted values:

$$f_1 = c + m_1 x_{1,1} + m_2 x_{2,1}$$

$$f_2 = c + m_1 x_{1,2} + m_2 x_{2,2}$$

$$f_3 = c + m_1 x_{1,3} + m_2 x_{2,3}$$

$$f_4 = c + m_1 x_{1,4} + m_2 x_{2,4}$$

$$f_5 = c + m_1 x_{1,5} + m_2 x_{2,5}$$

$$f_6 = c + m_1 x_{1,6} + m_2 x_{2,6}$$

We can write this in matrix vector form as:

$$\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ 1 & x_{1,3} & x_{2,3} \\ 1 & x_{1,4} & x_{2,4} \\ 1 & x_{1,5} & x_{2,5} \\ 1 & x_{1,6} & x_{2,6} \end{bmatrix} \begin{bmatrix} c \\ m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} c + m_1 x_{1,1} + m_2 x_{2,1} \\ c + m_1 x_{1,2} + m_2 x_{2,2} \\ c + m_1 x_{1,3} + m_2 x_{2,3} \\ c + m_1 x_{1,4} + m_2 x_{2,4} \\ c + m_1 x_{1,5} + m_2 x_{2,5} \\ c + m_1 x_{1,6} + m_2 x_{2,6} \end{bmatrix}$$

(remembering you multiply the rows of the matrix by the column vector to recover the full-form equations). The design matrix in this example is:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ 1 & x_{1,3} & x_{2,3} \\ 1 & x_{1,4} & x_{2,4} \\ 1 & x_{1,5} & x_{2,5} \\ 1 & x_{1,6} & x_{2,6} \end{bmatrix}$$

The parameter vector we might call $\boldsymbol{\theta} = \begin{bmatrix} c \\ m_1 \\ m_2 \end{bmatrix}$

We can then write the model as:

$$\mathbf{f} = \mathbf{X}\,\boldsymbol{\theta}$$

or if we want to model the data,

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0,\sigma)$

Design matrices can be created for any GLM. For example if the model were

$$f_i = c + \alpha_j + \beta_k + m_1 x_{1,i} + m_{2,i} x_{2,i}$$

j = 1 .. 2, *k* = 1 .. 3

the matrix vector equation would like this:

$$\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \end{bmatrix} = \begin{bmatrix} 1 & \delta_{\alpha_1,1} & \delta_{\alpha_2,1} & \delta_{\beta_1,1} & \delta_{\beta_2,1} & \delta_{\beta_3,1} & x_{1,1} & x_{2,1} \\ 1 & \delta_{\alpha_1,2} & \delta_{\alpha_2,2} & \delta_{\beta_1,2} & \delta_{\beta_2,2} & \delta_{\beta_3,2} & x_{1,2} & x_{2,2} \\ 1 & \delta_{\alpha_1,3} & \delta_{\alpha_2,3} & \delta_{\beta_1,3} & \delta_{\beta_2,3} & \delta_{\beta_3,3} & x_{1,3} & x_{2,3} \\ 1 & \delta_{\alpha_1,4} & \delta_{\alpha_2,4} & \delta_{\beta_1,4} & \delta_{\beta_2,4} & \delta_{\beta_3,4} & x_{1,4} & x_{2,4} \\ 1 & \delta_{\alpha_1,6} & \delta_{\alpha_2,6} & \delta_{\beta_1,5} & \delta_{\beta_2,5} & \delta_{\beta_3,5} & x_{1,5} & x_{2,5} \\ 1 & \delta_{\alpha_1,6} & \delta_{\alpha_2,6} & \delta_{\beta_1,6} & \delta_{\beta_2,6} & \delta_{\beta_4,6} & x_{1,6} & x_{2,6} \end{bmatrix} \begin{bmatrix} c \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ m_1 \\ m_2 \end{bmatrix}$$

Where the δ's are 1's or 0's, indicating which level of each of the two categorical variables each observation of the response variable is associated with.

This is useful since it provides very efficient ways to fit the model

(back to Contents)

Appendix Z

Counting degrees of freedom in a model

Here are some examples of how to write down the algebraic structure of different models. In all examples f_i is the fitted value predicted by the model for the *i*th observation of the response variable.

	The algebraic structure of the model	
Ex 0	$f_i = c$	The intercept model
Ex 1	$f_i = c + \alpha_j$	One cat. expl. variable
Ex 2	$f_i = c + \alpha_j + \beta_k$	Two cat. expl. variables
Ex 3	$f_i = c + \alpha_j + \beta_k + \gamma_l$	Three cat. expl. variables
Ex 4	$f_i = c + m \cdot x_i$	One cont. expl. variable
Ex 5	$f_i = c + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i}$	Two cont. expl. variables
Ex 6	$f_i = c + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i} + m_3 \cdot x_{3,i}$	Three cont. expl. variables
Ex 7	$f_i = c + \alpha_j + \beta_k + \gamma_l + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i} + m_3 \cdot x_{3,i}$	Three cat. expl. variables and three cont. expl. variable
Ex 8	$f_i = c + \alpha_j + \beta_k + \gamma_{jk}$	Two cat. expl. variables with interaction
Ex 9	$f_i = c + \alpha_j + \beta_k + \gamma_l + \delta_{jk} + \zeta_{kl}$	Three cat. expl. variables with two interactions
Ex 10	$f_i = c + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i} + m_3 \cdot x_{1,i} \cdot x_{2,i}$	Two cont. expl. variables with interaction
Ex 11	$f_i = c + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i} + m_3 \cdot x_{3,i} + m_4 \cdot x_{1,i} \cdot x_{2,i} + m_5 \cdot x_{2,i} \cdot x_{3,i}$	Three cont. expl. variables with two interactions
Ex 12	$f_i = c + \alpha_j + (m + \gamma_j) \cdot x_i$	One cat. and one cont expl. variable with a cat-cont interaction
Ex 13	$f_i = c + \alpha_j + \beta_k + \gamma_l + (m_1 + \delta_j) \cdot x_{1,i} + (m_2 + \zeta_l) \cdot x_{2,i} + m_3 \cdot x_{3,i}$	Three cat. and three cont expl. variables with two cat-cont interactions
Ex 14	$f_i = c + \alpha_j + \beta_k + \gamma_l + (m_1 + \delta_j + \zeta_l) \cdot x_{1,i} + m_3 \cdot x_{3,i}$	Three cat. and three cont expl. variables with two cat-cont Interactions but with the *same* continuous variable.

<u>Example 0.</u> The intercept (or 'null') model with no explanatory variables.

$$f_i = c$$

f_i indicates the fitted values. The subscript *i* runs from 1 .. *n* (the number of observations, i.e. rows of your data), each observation is assumed to come from *exactly* the same distribution parameterized by the intercept *c*. For a general linear model this would be one single normal distribution with mean *c*.

We are modeling the data as coming from a single distribution but we are not explaining variation (because we don't have any explanatory variables!).

The degrees of freedom required by the right-hand-side (or linear predictor) would be 1 (for *c*).

Example 1. One categorical explanatory variable with *p* levels.

$$f_i = c + \alpha_j$$

i = 1 .. *n*
j = 1 .. *p*

f_i indicates the fitted values. The subscript *i* runs from 1 .. *n* (the number of observations, i.e. rows of your data). So here we have a baseline (*c*) and an adjustment for each of the *p* levels of whatever α represents.

The degrees of freedom required by the right-hand-side (or linear predictor) would be 1 (for *c*) + (*p*-1) (for the categorical explanatory variable).

Example 2. Two categorical explanatory variables, one with *p* levels and one with *q* levels.

$$f_i = c + \alpha_j + \beta_k$$

i = 1 .. *n*
j = 1 .. *p*
k = 1 .. *q*

f_i indicates the fitted values. The subscript *i* runs from 1 .. *n* (the number of observations, i.e. rows of your data). So here we have a baseline (*c*), an adjustment for each of the *p* levels of whatever α represents, and an adjustment for each of the *q* levels of whatever β represents.

The degrees of freedom required by the right-hand-side (or linear predictor) would be 1 (for *c*) + (*p*-1) (for the first categorical explanatory variable) + (*q*-1) (for the second categorical explanatory variable).

Example 3. Three categorical explanatory variables, one with *p* levels, one with *q* levels, and one with *r* levels

$$f_i = c + \alpha_j + \beta_k + \gamma_l$$

i = 1 .. *n*
j = 1 .. *p*
k = 1 .. *q*
l = 1 .. *r*

f_i indicates the fitted values. The subscript *i* runs from 1 .. *n* (the number of observations, i.e. rows of your data). So here we have a baseline (*c*), an adjustment for each of the *p* levels of whatever α represents, an adjustment for each of the *q* levels of whatever β represents, and an adjustment for each of the *r* levels of whatever γ represents.

The degrees of freedom required by the right-hand-side (or linear predictor) would be 1 (for *c*) + (*p*-1) (for the first categorical explanatory variable) + (*q*-1) (for the second categorical explanatory variable) + (*r*-1) (for the third categorical explanatory variable).

Example 4. One continuous explanatory variable.

$$f_i = c + m \cdot x_i$$

i = 1 .. *n*

f_i indicates the fitted values. The subscript *i* runs from 1 .. *n* (the number of observations, i.e. rows of your data), *m* – is the slope that we multiply a numerical value (x_i) of a continuous variable by. So here we have a baseline (c) and an adjustment provided by the product of *m* and x_i.

The degrees of freedom required by the right-hand-side (or linear predictor)
would be 1 for (for *c*) + 1 (for *m*).

Example 5. Two continuous variables.

$$f_i = c + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i}$$

i = 1 .. *n*

f_i indicates the fitted values. The subscript *i* runs from 1 .. *n* (the number of observations, i.e. rows of your data), m_1 – is the slope that we multiply a numerical value ($x_{1,i}$) of a continuous variable by, and m_2 – is the slope that we multiply a numerical value ($x_{2,i}$) of a continuous variable by. So here we have a baseline (c), an adjustment provided by the product of m_1 and $x_{1,i}$, and an adjustment provided by the product of m_2 and $x_{2,i}$.

The degrees of freedom required by the model would be 1 (for *c*) + 1 (for m_1) + 1 (for m_2).

Example 6. Three continuous explanatory variables.

$$f_i = c + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i} + m_3 \cdot x_{3,i}$$

i = 1 .. *n*

f_i indicates the fitted values. The subscript *i* runs from 1 .. *n* (the number of observations, i.e. rows of your data), m_1 – is the slope that we multiply a numerical value ($x_{1,i}$) of a continuous variable by, m_2 – is the slope that we multiply a numerical value ($x_{2,i}$) of a continuous variable by, and m_3 – is the slope that we multiply a numerical value ($x_{3,i}$) of a continuous variable by. So here we have a baseline (c), an adjustment provided by the product of m_1 and $x_{1,i}$, an adjustment provided by the product of m_2 and $x_{2,I}$, and an adjustment provided by the product of m_3 and $x_{3,i}$.

The degrees of freedom required by the right-hand-side (or linear predictor) would be 1 (for *c*) + 1 (for m_1) + 1 (for m_2) + 1 (for m_3).

Example 7. Three categorical explanatory variables, one with *p* levels, one with *q* levels, and one with *r* levels, and three continuous explanatory variables.

$$f_i = c + \alpha_j + \beta_k + \gamma_l + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i} + m_3 \cdot x_{3,i}$$

i = 1 .. *n*
j = 1 .. *p*
k = 1 .. *q*
l = 1 .. *r*

f_i indicates the fitted values. The subscript *i* runs from 1 .. *n* (the number of observations, i.e. rows of your data). So here we have a baseline (*c*), an adjustment for each of the *p* levels of whatever α represents, an adjustment for each of the *q* levels of whatever β represents, an adjustment for each of the *r* levels of whatever γ represents*,* m_1 – is the slope that we multiply a numerical value ($x_{1,i}$) of a continuous variable by, m_2 – is the slope that we multiply a numerical value ($x_{2,i}$) of a continuous variable by, and m_3 – is the slope that we multiply a numerical value ($x_{3,i}$) of a continuous variable by.

The degrees of freedom required by the right-hand-side (or linear predictor) would be 1 (for *c*) + (*p*-1) (for the first categorical explanatory variable) + (*q*-1) (for the second categorical explanatory variable) + (*r*-1) (for the third categorical explanatory variable) + 1 (for m_1) + 1 (for m_2) + 1 (for m_3).

Interactions

Categorical with categorical

Example 8. Two categorical explanatory variables, one with *p* levels and one with *q* levels, and the interaction between the two.

$$f_i = c + \alpha_j + \beta_k + \gamma_{jk}$$

i = 1 .. *n*
j = 1 .. *p*
k = 1 .. *q*

f_i indicates the fitted values. The subscript *i* runs from 1 .. *n* (the number of observations, i.e. rows of your data). So here we have a baseline (*c*), an adjustment for each of the *p* levels of whatever α represents, an adjustment for each of the *q* levels of whatever β represents, and *p* x *q* interaction terms (γ_{jk}) of which (*p*-1) x (*q*-1) will be non-zero.

The degrees of freedom required by the right-hand-side (or linear predictor) would be 1 (for *c*) + (*p*-1) (for the first categorical explanatory variable) + (*q*-1) (for the second categorical explanatory variable) + (*p*-1) x (*q*-1) for the non-zero interaction terms.

Example 9. Three categorical explanatory variables, one with *p* levels, one with *q* levels, and one with *r* levels, and two interactions between the first (α) and second (β), and second (β) and third (γ) categorical explanatory variables

$$f_i = c + \alpha_j + \beta_k + \gamma_l + \delta_{jk} + \zeta_{kl}$$

i = 1 .. *n*
j = 1 .. *p*
k = 1 .. *q*
l = 1 .. *r*

f_i indicates the fitted values. The subscript *i* runs from 1 .. *n* (the number of observations, i.e. rows of your data). So here we have a baseline (*c*), an adjustment for each of the *p* levels of whatever α represents, an adjustment for each of the *q* levels of whatever β represents, an adjustment for each of the *r* levels of whatever γ represents, *p* x *q* interaction terms (γ_{jk}) of which (p-1) x (q-1) will be non-zero, and q x r interaction terms (ζ_{kl}) of which (q-1) x (r-1) will be non-zero.

The degrees of freedom required by the right-hand-side (or linear predictor) would be 1 (for c) + (p-1) (for the first categorical explanatory variable) + (q-1) (for the second categorical explanatory variable) + (r-1) (for the third categorical explanatory variable) + (p-1) x (q-1) for the non-zero interaction terms between whatever α and β represent + (q-1) x (r-1) for the non-zero interaction terms between whatever β and γ represent.

Continuous with continuous

Example 10. Two continuous variables and their interaction.

$$f_i = c + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i} + m_3 \cdot x_{1,i} \cdot x_{2,i}$$

i = 1 .. n

f_i indicates the fitted values. The subscript i runs from 1 .. n (the number of observations, i.e. rows of your data), m_1 – is the slope that we multiply a numerical value ($x_{1,i}$) of a continuous variable by, m_2 – is the slope that we multiply a numerical value ($x_{2,i}$) of a continuous variable by, and m_3 – is a coefficient that we multiply by the product of $x_{1,i}$ and $x_{2,I}$ representing the interaction of the two continuous explanatory variables. So here we have a baseline (c), an adjustment provided by the product of m_1 and $x_{1,i}$, an adjustment provided by the product of m_2 and $x_{2,i}$, and an adjustment provided by the product of m_3 , $x_{1,i}$, and $x_{2,i}$.

The degrees of freedom required by the right-hand-side (or linear predictor) would be 1 (for c) + 1 (for m_1) + 1 (for m_2) + 1 (for m_3).

Example 11. Three continuous explanatory variables with an interaction between the first (x_1) and second (x_2), and second (x_2) and third (x_3) variables.

$$f_i = c + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i} + m_3 \cdot x_{3,i} + m_4 \cdot x_{1,i} \cdot x_{2,i} + m_5 \cdot x_{2,i} \cdot x_{3,i}$$

i = 1 .. n

f_i indicates the fitted values. The subscript i runs from 1 .. n (the number of observations, i.e. rows of your data). So here we have a baseline (c), m_1 – is the slope that we multiply a numerical value ($x_{1,i}$) of a continuous variable by, m_2 – is the slope that we multiply a numerical value ($x_{2,i}$) of a continuous variable by, m_3 – is the slope that we multiply a numerical value ($x_{3,i}$) of a continuous variable by, m_4 – is a coefficient that we multiply by the product of $x_{1,i}$ and $x_{2,I}$ representing the interaction of continuous explanatory variables one and two, and m_5 – is a coefficient that we multiply by the product of $x_{2,i}$ and $x_{3,I}$ representing the interaction of continuous explanatory variables two and three.

The degrees of freedom required by the right-hand-side (or linear predictor) would be 1 (for c) + 1 (for m_1) + 1 (for m_2) + 1 (for m_3) + 1 (for m_4) + 1 (for m_5).

Continuous with categorical

Example 12. One categorical explanatory variable with p levels, one continuous explanatory variable, and their interaction.

$$f_i = c + \alpha_j + \left(m + \gamma_j\right) \cdot x_i$$

i = 1 .. *n*
j = 1 .. *p*

f_i indicates the fitted values. The subscript *i* runs from 1 .. *n* (the number of observations, i.e. rows of your data). We have a baseline (*c*), an adjustment for each of the *p* levels of whatever α represents, and *m* - a slope that *is itself adjusted* by γ_j depending on the level of the explanatory variable that applies. So here we have a baseline (*c*) and an adjustment for each of the *p* levels of whatever α represents, an adjustment provided by the product of the value of $(m + \gamma_j)$ and x_i.

The degrees of freedom required by the right-hand-side (or linear predictor) would be 1 (for *c*) + (*p*-1) (for the categorical explanatory variable) + 1 (for *m*) + (*p*-1) (for the interaction terms).

Example 13. Three categorical explanatory variables, one with *p* levels, one with *q* levels, and one with *r* levels, three continuous explanatory variables, and interactions between the first continuous (x_1) and first categorical variable (represented by α), and the second (x_2) continuous and third (represented by γ) categorical variable:

$$f_i = c + \alpha_j + \beta_k + \gamma_l + \left(m_1 + \delta_j\right) \cdot x_{1,i} + (m_2 + \zeta_l) \cdot x_{2,i} + m_3 \cdot x_{3,i}$$

i = 1 .. *n*
j = 1 .. *p*
k = 1 .. *q*
l = 1 .. *r*

f_i indicates the fitted values. The subscript *i* runs from 1 .. *n* (the number of observations, i.e. rows of your data). So here we have a baseline (*c*), an adjustment for each of the *p* levels of whatever α represents, an adjustment for each of the *q* levels of whatever β represents, an adjustment for each of the *r* levels of whatever γ represents, m_1 - a slope that is itself adjusted by δ_j depending on the level of the explanatory variable (represented by α) that applies, m_2 - a slope that is adjusted by ζ_l depending on the level of the other explanatory variable (represented by γ) that applies, and then a final adjustment provided by the product of m_2 and $x_{2,i}$.

The degrees of freedom required by the right-hand-side (or linear predictor) would be 1 (for *c*) + (*p*-1) (for the first categorical explanatory variable) + (*q*-1) (for the second categorical explanatory variable) + (*r*-1) (for the third categorical explanatory variable) + 1 (for m_1) + 1 (for m_2) + 1 (for m_3) + (*p*-1) (for the interaction with α) + (*r*-1) (for the interaction with γ),

Example 14. Three categorical explanatory variables, one with *p* levels, one with *q* levels, and one with *r* levels, two continuous explanatory variables, and two interactions – both with the first continuous variable (x_1) involving two categorical variables (α and γ):

$$f_i = c + \alpha_j + \beta_k + \gamma_l + \left(m_1 + \delta_j + \zeta_l\right) \cdot x_{1,i} + m_2 \cdot x_{2,i}$$

i = 1 .. *n*

$j = 1 \,..\, p$
$k = 1 \,..\, q$
$l = 1 \,..\, r$

f_i indicates the fitted values. The subscript *i* runs from 1 .. *n* (the number of observations, i.e. rows of your data). So here we have a baseline (*c*), an adjustment for each of the *p* levels of whatever α represents, an adjustment for each of the *q* levels of whatever β represents, an adjustment for each of the *r* levels of whatever γ represents, m_1 - a slope that is itself adjusted both by δ_j depending on the level of the explanatory variable (represented by α) that applies, *and* ζ_l depending on the level of the other explanatory variable (represented by γ) that applies, and then a final adjustment provided by the product of m_2 and $x_{2,i}$

The degrees of freedom required by the right-hand-side (or linear predictor) would be 1 (for *c*) + (*p*-1) (for the first categorical explanatory variable) + (*q*-1) (for the second categorical explanatory variable) + (*r*-1) (for the third categorical explanatory variable) + 1 (for m_1) + (*p*-1) (for the interaction with the explanatory variable represented by α) + (*r*-1) (for the interaction with the explanatory variable represented by γ) + 1 (for m_2)

Here are some examples of df counting:

Important Note: If we are using a Normal or Negative Binomial distribution to model the data we'd need one further degree of freedom for the model to represent the second argument of the distribution (the standard deviation in the case of the Normal distribution, and the dispersion parameter in the case of the Negative Binomial distribution).

	The algebraic structure of the model	**Range levels**	**Degrees of freedom (ignoring any dispersion terms e.g. the sd)**	**Model description**
Ex 0	$f_i = c$		1	The intercept model
Ex 1	$f_i = c + \alpha_j$	*j*=1..3	1+(3-1) = 3	One cat. expl. variable
Ex 2	$f_i = c + \alpha_j + \beta_k$	*j*=1..3, *k*=1..2	1+(3-1)+(2-1) = 4	Two cat. expl. variables
Ex 3	$f_i = c + \alpha_j + \beta_k + \gamma_l$	*j*=1..3, *k*=1..2, *l*=1..5	1+(3-1)+(2-1)+(5-1) = 8	Three cat. expl. variables
Ex 4	$f_i = c + m \cdot x_i$		1+1 = 2	One cont. expl. variable
Ex 5	$f_i = c + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i}$		1+1+1 = 3	Two cont. expl. variables
Ex 6	$f_i = c + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i} + m_3 \cdot x_{3,i}$		1+1+1+1 = 4	Three cont. expl. variables
Ex 7	$f_i = c + \alpha_j + \beta_k + \gamma_l + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i} + m_3 \cdot x_{3,i}$	*j*=1..3, *k*=1..2, *l*=1..5	1+(3-1)+(2-1)+(5-1)+1+1+1 = 11	Three cat. expl. variables and three cont. expl. variable
Ex 8	$f_i = c + \alpha_j + \beta_k + \gamma_{jk}$	*j*=1..3, *k*=1..2	1+(3-1)+(2-1)+ (3-1)x(2-1) = 6	Two cat. expl. variables with interaction
Ex 9	$f_i = c + \alpha_j + \beta_k + \gamma_l + \delta_{jk} + \zeta_{kl}$	*j*=1..3, *k*=1..2, *l*=1..5	1+(3-1)+(2-1)+(5-1) (3-1)x(2-1)+ (2-1)x(5-1) = 14	Three cat. expl. variables with two interactions
Ex 10	$f_i = c + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i} + m_3 \cdot x_{2,i} \cdot x_{3,i}$		1+1+1+1 = 4	Two cont. expl. variables with interaction
Ex 11	$f_i = c + m_1 \cdot x_{1,i} + m_2 \cdot x_{2,i} + m_3 \cdot x_{3,i} + m_4 \cdot x_{1,i} \cdot x_{2,i} + m_5 \cdot x_{2,i} \cdot x_{3,i}$		1+1+1+1+1+1 = 6	Three cont. expl. variables with two interactions
Ex 12	$f_i = c + \alpha_j + (m + \gamma_j) \cdot x_i$	*j*=1..3	1+(3-1)+1+(3-1)=6	One cat. and one cont expl. variable with a cat-cont interaction
Ex 13	$f_i = c + \alpha_j + \beta_k + \gamma_l + (m_1 + \delta_j) \cdot x_{1,i} + (m_2 + \zeta_l) \cdot x_{2,i} + m_3 \cdot x_{3,i}$	*j*=1..3, *k*=1..2, *l*=1..5	1+(3-1)+(2-1)+(5-1)+1+(3-1)+1+(5-1)+1 = 17	Three cat. and three cont expl. variables with two cat-cont interaction
Ex 14	$f_i = c + \alpha_j + \beta_k + \gamma_l + \left(m_1 + \delta_j + \zeta_l\right) \cdot x_{1,i} + m_2 \cdot x_{2,i}$	*j*=1..3, *k*=1..2, *l*=1..5	1+(3-1)+(2-1)+(5-1)+1+(3-1)+1+(5-1) = 16	Three cat. and three cont expl. variables with two cat-cont interaction but with the same continuous explanatory variable

The maximum range levels are intended to convey the range of the subscript, indicating the number of levels of each of the explanatory variables. In Ex. 1, because *j* runs from 1 to 3, the explanatory variable represented by α has 3 levels.

Glossary

AIC	AIC is one way of assessing the level of support for a model. AIC = -LL + 2p where LL is the log-likelihood of the model, and p is the number of parameters estimated by the model. Models with low AICs are favoured over models with higher AIC, but in general models that differ by less than 2 AIC units are considered equivalently well supported.
Algebraic structure of a model	A mathematical description of a model written in general algebraic terms. In its simplest form perhaps F(i) = c + m x(i) + c, or F(i) = c + m x(i) + a(j). It is important to always be aware of the algebraic structure of the model you have chosen to fit to data. [Here i refers to the ith observation of the response and explanatory variables, and j refers to the different levels of a categorical explanatory variable]
Alternative hypothesis	The hypothesis adopted if the null hypothesis is rejected. The null hypothesis usually hypothesizes that there is no effect of an explanatory variable on a response variable. The alternate hypothesis may be that there is an effect, or an effect in a specified direction.
ancova	Stands for analysis of covariance. Traditional term used to describe analyses of categorical explanatory variables while accounting for the effects of continuous explanatory variables. Or put more simply .. a GLM with both continuous and categorical explanatory variables including possibly interactions. https://keydifferences.com/difference-between-anova-and-ancova.html
anova (analysis of variance)	Stands for analysis of variance. Traditional term used to describe analyses of categorical explanatory variables. Or put more simply .. a GLM with only categorical explanatory variables. ANOVA is sometimes used to reference a very particular test (say "One way ANOVA" or "Two way ANOVA), but also to refer to a method of analysis (an ANOVA table) conducted with any sort of general linear model with any combination of both catagorical and continuous explanatory variables. Usually fitted using Least Squares. https://www.qualtrics.com/uk/experience-management/research/anova/
Approximate Bayesian Computation (ABC)	A numerically intensive approach to fitting a model to data, often used when it isn't possible to write down a likelihood function for the summary statistic that the model is trying to fit. https://towardsdatascience.com/the-abcs-of-approximate-bayesian-computation-bfe11b8ca341
Arguments	Usually this refers to information supplied to a function. For example, a Normal distribution is defined by two arguments, the mean and standard deviation, usually denoted N(mean, sd). Probability density functions may have one or more arguments. For example, a Bernoulli distribution has just one argument (p), the probability of a '1'.
Average	The sum of a set of numbers divided by the number of values in the sum. The mean of {2, 4, 12} = 6. Synonymous with mean.
Balance	A data set is balanced when there are (more or less) the same number of observations of the response variable for each level, or combination of levels of the categorical explanatory variables. Lack of balance can lead to difficulties fitting a model, and problems associated with non-orthogonality, although many model designs are quite robust to lack of balance.
Bayesian	An approach to statistical analysis based on Bayes theorem, in which the outcome is the support for a hypothesis, conditional on the data and the priors. An alternative philosophical approach to the often encountered 'frequentist' approach. https://www.britannica.com/science/Bayesian-analysis
Bernoulli distribution	A probability density function that generates just two discrete outcomes: '1' with probability p, and '0' with probability 1-p. A Bernoulli variate is a special case of a Binomial variate when only 1 trial is conducted. A Bernoulli distribution is defined by just one argument - the probability of a '1'. https://en.wikipedia.org/wiki/Bernoulli_distribution
Beta distribution	A probability density function that generates real variates bounded by 0 and 1. Ideal for modelling directly observed proportions or probabilities. https://en.wikipedia.org/wiki/Beta_distribution

Binary	A variable is binary if it can take on only two values (for example the outcome of a single coin toss)
Binomial distribution	A probability density function that generates discrete variates bounded by 0 and N, where N is the number of trials, each of which generates just one of two outcomes with probability p and 1-p respectively. So a binomial variate would be the number of heads from say 20 coin tosses. A binomial distribution is defined by two arguments, the probability of a '1', and the number of trials conducted to generate each variate. https://en.wikipedia.org/wiki/Binomial_distribution
Bonferroni correction	A method to counteract the risks of multiple hypothesis testing. If it was a requirement that a p-value < 0.05 for one hypothesis test, when conducting 10 tests we might require p < 0.05/10. https://en.wikipedia.org/wiki/Bonferroni_correction
Bootstrapping	Generating new data sets by sampling observations of the response variable and their associated explanatory variables (i.e. 'rows of data') with replacement from an existing original data set. The resampled data sets the same size as the original data set, and are typically subjected to some analysis that yields some estimate of interest. Distributions of these estimates can be generated by bootstrapping a data set say 1000 times. https://towardsdatascience.com/bootstrapping-statistics-what-it-is-and-why-its-used-e2fa29577307
Bounded	A bounded variate may only take on values in a certain range (0-1, or 0-N).
Box Cox (transform)	A Box Cox transform is a traditional transform that attempts to normalize a set of numbers (usually your response variable). Logging, or square-rooting your response variable are really only two points on a continuum of different ways that you may choose to transform your data. Such transforms are often not necessary with effective use of generalised linear models. https://towardsdatascience.com/box-cox-transformation-explained-51d745e34203
Categorical explanatory variable	An explanatory (or independent variable) that is categorical and has a certain number of levels. Sometimes a categorical explanatory variable is referred to as a 'factor'. Sex would be a categorical variable that usually has two levels: male and female.
Causal inference	Causal inference is an (increasingly formalized) approach that can be used to help determine the effect of a particular phenomenon (say a variable) that is a component of a larger system on another component of the same system. The approach introduces terms such as confounders, colliders, and modifiers that classify the way a third variable might interfere with inferring causal relationships between other variables. A good review is here.
Central Limit Theorem	The central limit theorem states that the mean of sample of random variates will tend to a normal distribution, regardless of which distributions the variates are sampled from.
Chi-squared distribution	A common distribution used to test a wide variety of different test statistics (for example chi-squared contingency tests, goodness of fit tests, likelihood ratio tests, Wald tests, over dispersion tests etc). Chi-squared distributions have one argument - the number of degrees of freedom. A Chi-squared distribution is in fact the distribution of k squared standard normal variates, where k is the 'degrees of freedom' argument.
Chi-squared statistic	A chi squared distribution with n degrees of freedom is the distribution of the sum of n squared standard normal variates. A chi-squared statistic is assumed to be distributed according to a chi squared distribution under the null hypothesis.
Coefficient	A coefficient is synonymous with a parameter. Usually refers to perhaps a slope, or a 'correction' for a certain level of a categorical explanatory variable, but may also be an estimated variance or dispersion coefficient. Generally speaking, anything that is estimated from your data.
Coefficient analysis	Analysis - usually inference - performed on the individual coefficients (or parameters) of a single model. This contrasts with analyses that are based on a comparison of the relative likelihoods of two closely related models.

Collinearity or non-orthogonality	Two variables (usually two different explanatory variables) are said to be collinear (or synonymously, non-orthogonal) if they are positively or negatively correlated with each other. Correlations can exist between variables that are continuous or categorical. Collinearity can lead to difficulties interpreting your output. Collinearity can be assessed using Variance Inflation Factors (VIFs).
Confidence interval	The confidence level is the percentage of times you expect to reproduce an estimate between the upper and lower bounds of a confidence interval. The percentage may be chosen to be anything you want, but the standard is a 95% confidence interval. For a Normal distribution 95% confidence intervals are generated (approximately) by adding (and subtracting) two standard errors to (and from) the mean. https://www.scribbr.com/statistics/confidence-interval/
Contingency test	Tests on contingency tables are used to evaluate the association and the independence between the rows and the columns of a contingency table as well as to calculate various association measures. In a 2 x 2 contingency table, we might ask 'are the distribution of the observations in rows independent of the distribution over the columns'. Such questions might be addressed with a Chi-squared (contingency) test, or a Fisher's exact test.
Continuous	A real number that can take on an infinite number of values, but may be bounded (for example a proportion or percentage).
Continuous distribution	One that generates continuous outcomes (or variates), for example, Gaussian (or Normal), Gamma, Log-normal, or Beta distributions.
Continuous explanatory variables	Explanatory variables that are continuous. Often synonymous with covariate. For example, temperature measured to a couple of decimal places could be an example of a continuous explanatory variable. Important to note that there is no requirement that continuous explanatory variables be distributed in any particular way.
Correlation	A simple measure of association between two variables, bounded between -1 and +1. The statistical significance of correlation coefficients can be assessed. Correlation is distinct from simple linear regression as no line is fitted to the data, and one variable is not assumed to 'depend' on the other. See Pearson and Spearman correlation.
Covariate	Covariates usually refer to Continuous explanatory variables.
Credible interval	Simply speaking, a credible interval is the Bayesian equivalent of a confidence interval. There are subtle differences that you should be aware of when you use them. https://en.wikipedia.org/wiki/Credible_interval
Cross Validation	A method to examine predictive ability in which a proportion of the data are excluded from the model fitting process, and the resulting model can then be used to predict the excluded data. In k-fold cross validation the data is divided randomly into k equal portions. The model is fitted to all but one of these portions, and used to predict the response variables in the excluded portion.
Data dredging	Fitting models until you get an answer you like. Often used in the same sense as fishing.
Degrees of Freedom	Degrees of freedom refers to the maximum number of logically independent values, which are values that have the freedom to vary in the data sample. This will often be - and cannot exceed - the number of observations of the response variable. Each parameter (or coefficient) estimated from these data 'uses up' a degree of freedom, so the error (or residual) degrees of freedom is usually the total degrees of freedom less the number of coefficients (or parameters) estimated from the data. Maintaining a large number of error (or residual) degrees of freedom is important because it reduces the standard errors of the models estimated parameters (or coefficients) and therefore increases their significance.
Density function	Another way of referring to a probability distribution or a probability density function

Design matrix	A design matrix is an advanced way of describing a GLM using matrix vector notation. You don't need to understand it just yet. https://en.wikipedia.org/wiki/Design_matrix
Deviance	Deviance can be viewed as a generalization of residual sums of squares in linear models. You can think of the deviance of a model as proportional to the negative log likelihood (technically it is twice the negative log likelihood of the data for a specified model plus a constant). Or, more conceptually, as proportional to the variation in the data that has not been explained by the model. Residual deviance and null deviance are special cases of deviance. https://statisticaloddsandends.wordpress.com/2019/03/27/what-is-deviance/
Diagnostic analysis	A general term for the tests it is necessary to conduct after fitting a model, to ensure the major assumptions made by the modelling process have not been importantly violated. Often called residual analysis. Usually involves looking for trends in the patterns of residuals when plotted against fitted values, and checking on the distribution of residuals (more straightforward for general than generalised linear models). The DHARMa package is very good for this.
Discrete	A number is discrete if it can only take on particular values (usually meaning integer, or binary values). Discrete numbers may be bounded (e.g. a Binomial variate).
Discrete distribution	One that generates discrete outcomes (or variates), for example a Bernoulli, Binomial, Poisson or negative binomial distribution
Dispersion coefficient	This is cited in the summary output of a glm. For a model that uses a Normal distribution it is the residual variance, or the variance of the residuals
Effect size	This is the effect on the response variable of particular (defined) changes to one or more explanatory variables. Common examples of effect sizes would be the slope that indicates the effect of a unit change in a continuous explanatory variable on the response variable; or the effect of changing from one level of an explanatory variable to another on the response variable. If the variable associated with the effect size is involved in a significant interaction, then the effect size of the variables will depend on each other.
Explanatory variable	These are variables that are used to explain the variation in your response (or dependent) variable. They appear on the right-hand side of your model. They are subject to no distributional assumptions.
Exponential distribution	A continuous monotonic probability density function with mode 0, comprising non-negative real numbers, with one argument. https://en.wikipedia.org/wiki/Exponential_distribution
Exponential (or exponentiating)	This is the opposite of taking the logarithm of a number. It reverses the log function. Exp(x) is the same as e^x. Exp(ln(x)) = x and ln(exp(x)) = x. Try it.
Extrapolation	In the context of a statistical model, extrapolation usually refers to making a prediction based on the model that assumes values of explanatory variables outside the range of those to which the model was fitted.
F statistic	This is a key statistic output from analysis conducted using models fitted using Least Squares. It is the ratio of the 'explained mean sums of squares'/'unexplained mean sums of squares'. It would only apply to response variables assumed to be Normally distributed. This is not a statistic that exists when models are fitted using likelihood, but important to understand what this commonly encountered statistic means. (Note: sometimes F is also used in our courses to denote Fitted values generated by a model so this is potentially confusing but the use should be clear from the context).
Factor	An explanatory variable (or independent variable) that is categorical and has a certain number of levels. Synonymous with and more often referred to as a 'categorical explanatory variable'. Sex would be a 'factor' that usually has two levels: male and female.
Fishers exact test	An alternative to a Chi-squared test used to analyse contingency tables ('do the rows depend on the columns'?). Better for small samples. https://en.wikipedia.org/wiki/Fisher%27s_exact_test

Fishing	A term used to describe a process whereby an excessive number of different models, or different explanatory variables are examined in search of any relationship with a response variable without appropriate prior justification for doing so. A desperate search for a positive result!
Fitted values	These are the values of the response variable that the model predicts for the various observed combinations of the explanatory variables. They will differ from the observed values of the response variable by the residual.
Fitting a model	Once a model has been specified by defining its algebraic structure, it will be 'fitted to data', in the sense that the coefficients (or parameters) that link the response variable to the various explanatory variables are chosen so that the right hand side of the model provides the closest possible fit - given this model structure - to the pattern of variation in the response variable represented on the left hand side. For example, we may specify that y = mx + c, but then we must fit this relationship to the data (y and x) in order to choose the best fitting values of the coefficients m and c. There are different methods of fitting models to data that include least squares, maximum likelihood, MCMC, ABC and many others.
Fixed effect	Fixed effects are explanatory variables that are constant. These variables, like age, sex, or ethnicity, don't change or change at a constant rate over time. They have fixed effects; in other words, any change they cause is always assumed to be the same. For example, any effects from being a woman, or a 17-year-old will not change over time. Generally speaking, all explanatory variables are fixed unless they are defined as random effects. Fixed effects may be continuous or categorical.
Forward selection	A process whereby explanatory variables are sequentially added to a simple model, and retained if they increase the explanatory power or likelihood of the model by a sufficient amount.
Frequentist	Generally speaking, if you are not using a Bayesian approach you are using a frequentist approach. The term frequentist is used because p-values and confidence intervals used in frequentist analysis provide an indication of how frequently your ouput (and output more divergent from the null hypothesis) would be observed were the null hypothesis to be 'true' if your data generation process was repeated a large number of times. It is hard to simply describe this paradigm, but frequentist approaches generally consider how likely a data set is, given a hypothesis (a model), where as a Bayesian approach considers how likely a hypothesis (a model) is, given data.
Friedman Test	A non-parametric test used for one-way repeated measures anova
Gamma distribution	A continuous probability density function that generates non-negative real numbers. Often used to model 'time to' or 'time since' an event. https://en.wikipedia.org/wiki/Gamma_distribution
Gaussian distribution	Synonymous with Normal. The same as a Normal distribution
General Additive Model (GAM)	A GAM is a generalised linear model in which the response variable depends on sums of smooth functions of some predictor variables, and interest focuses on inference about these smooth functions. While the distinction between a GLM and a GAM is technically a bit blurry, a GAM differs from a GLM in the nature of the 'smooth functions' used, and this endows them with great flexibility. They are useful for modelling response variables that are particularly 'wiggly' with respect to the explanatory variables. https://multithreaded.stitchfix.com/blog/2015/07/30/gam/
General Linear Model (GLM)	A model in which a Normally distributed response variable is modelled as a linear sum of the effects of explanatory variables and their products (interactions). Two key points: 1) the response variable is assumed to be Normal, and 2) 'terms' on the right hand side are assumed to be additive to each other.

Generalised Linear Model	A model in which a response variable is modelled as a linear sum of the effects of explanatory variables and their products (interactions). Two key points: 1) the response variable is assumed to be distributed according to some distribution *other than Normal* (common examples would be Bernoulli, Beta, Poisson, Negative Binomial), and 'terms' on the right hand side are assumed to be additive to each other.
Goodness-of-fit	Goodness of fit (GoF) metrics assess how much of the variation in the response variable has been captured by the model. Can be quantified by R-squared, or Pseudo R squared. Models with more parameters (or coefficients) will always have higher GoF, and higher likelihoods (less negative log-likelihoods). In general, we do not seek models with the highest GoF, but the most parsimonious.
Heteroscadisticity	Heteroscedasticity arises when the standard deviation(s) of a response variable, monitored over different values of associated explanatory variables are not constant. For example, we might observe increasingly wide scatter in the observed value of a response variable as a continuous explanatory variable increases in value (I've referred to this as a 'trumpet' in class), or very uneven variation for different levels of a categorical explanatory variable.
Hierarchical model	A hierarchical model is a model in which lower levels are sorted under a hierarchy of successively higher-level units. Data is grouped into clusters at one or more levels, and the influence of the clusters on the data points contained in them is taken account of in any statistical analysis. This text has not considered them.
Homoscadisticity	Homoscadisticity is when the standard deviations of a response variable, monitored over different values of associated explanatory variables are constant. The absence of heteroscedasticity
Hurdle model	A hurdle model is a two-part model that specifies one process for zero counts and another process for positive counts (contrast carefully with zero-inflated models where the second process is for non-negative counts)
Identifiability	Identifiability issues arise when the value of one parameter (or coefficient) essentially trades-off with the values of other parameters, so that the data doesn't permit the clear 'identification' of all the parameters in the model. More technically a model is identifiable if it is theoretically possible to learn the true values of this model's underlying parameters after obtaining an infinite number of observations from it.
Identity function	A link function in which f = linear predictor
iid	Stands for independently and identically distributed. Random numbers are iid if they are generated independently from an identical probability density function (with exactly the same arguments). `rnorm(20, 0, 1)` would generate 20 iid normal variates in R.
Independent variable	This is another term for an explanatory variable. (the use of independent and dependent variables is quite confusing and we greatly prefer the explanatory and response variable terminology)
inference/Inferential statistics	Use of data to infer something about a population by using a sample to generalise to a population. We make a distinction between modelling data (estimating the parameters of a model) and making a judgement as to how important the parameters are in influencing the response variable (inference).
Interaction	The inclusion of an [+ explanatory_variable1 x explanatory_variable2] term in a model that enables the effect of one explanatory variable on the response variable to depend on another. Interactions are represented a little differently in the algebraic structure of the model depending on whether they are between two continuous explanatory variables, two categorical explanatory variables, or one of each. It is possible to test for interactions between more than two explanatory variables (i.e. a 3-way interaction) but interpretation becomes challenging.

Intercept	A baseline parameter appearing on the right-hand side of a GLM to which 'adjustments' are made conditional on the explanatory variables. For example, in f = c + [adjustments]. f denotes the fitted values and c is the intercept. In the simplest glm: f = c + mX, then c is the value of f when X = 0, and literally the y-axis intercept on a graph of f on the y-axis and X on the x-axis.
Intercept only model	A model that doesn't have any explanatory variables, only the reference point or intercept. For example f = c. In R this would fitted with the command glm(y ~ 1)
Interpolation	In the context of a statistical model, interpolation usually refers to making a prediction based on the model that assumes values of explanatory variables within the range of those to which the model was fitted. For example, you may have measured the effects of behaviour at 5, 10 and 15 degrees C, but wish to predict it at 12 degrees C. This would be interpolation. If you chose to predict at 20 deg C, this would be extrapolation.
Intra-class coefficient (ICC)	A measure that captures the consistency of observations made at different levels of a random effect. It is estimated as the variance associated with a random effect divided by the sum of all other variances (variance of random effects and residual variance). So if a random effect accounts for the majority of unaccounted variance in the observations of the response variable the ICC will be high.
Kruskal Wallis tests	A non-parametric test equivalent to a one-way anova
Kurtosis	Kurtosis is a measure of the fatness of the tails of a distribution. Fatter tails correspond to a greater frequency of outliers. Kurtosis is assessed relative to a Normal distribution.
Lasso regression	LASSO (Least Absolute Shrinkage and Selection Operator) models enable the selection and shrinkage of parameters. This approach is very useful when analyzing data sets with a large number of explanatory variables (there is no definition of 'large' but think 10 or more). A good R package for Lasso regression is glmmLasso.
Least Squares	Fitting a model by minimizing the squared deviation of each data point from the fitted values generated by the model. Applicable only to response variables that derive from Normal (or Gaussian) distributions, and should generate best fitting parameters (or coefficients) that are exactly identical to those obtained by fitting the model using maximum likelihood.
Left hand side	Reference to the left hand side of an expression. For example, in the model F = c + mx , F is the left hand side, and c + mx the right hand side. The '=' sign divides the left hand side from the right hand side.
Levels	The number of different possible 'states' of a categorical explanatory variable. Sex would be a categorical variable that usually has two levels: male and female.
Likelihood	A probability density function will return the likelihood of a particular value of a variate. Likelihoods are by definition non-negative real numbers that are not bounded to fall between 0-1 (as probabilities are required to do). In R, dnorm(0.8, 0, 1) returns the likelihood of obtaining the value 0.8 from a Normal distribution with mean = 0, and sd = 1. The likelihood of a set of independent variates is the product of their individual likelihoods. Likelihoods are often very very small numbers and so we often work with the natural log of the likelihood (which will almost always be negative and computationally easier to handle).
Likelihood function	A function that returns a likelihood, for example, a probability density function is a simple likelihood function. When you define your model in R, and the 'family' that you are using to model the response variable, the R function you are using (for example glm or glm.nb or glmer) will construct a likelihood function. It will then identify the values of the models coefficients (or parameters) that generate the maximum likelihood of your observed response variable given the algebraic structure of the model.

Likelihood Ratio Test	A test that compares two nested models, a simpler one, and a more complex one. The test statistic is twice the difference in the log-likelihood of the two models, and this is assumed to be Chi-squared distributed with degrees of freedom determined by the difference in the number of parameters (or coefficients) estimated by each of the models.
Likert scale	A categorical ordinal scale often comprising 5 - 8 items, often used to specify a level of agreement or disagreement on a symmetric agree-disagree scale for a series of statements (for example, Strongly disagree, disagree, neither agree or disagree, agree, or strongly agree).
Linear predictor	The 'right hand side' of a glm. A linear sum of 'adjustments' determined from continuous and categorical explanatory variables, and the intercept.
Link function	A function that is used to transform a response variable so that it can be modelled by a linear predictor. Log-link functions are used for Poisson, Negative Binomial, and Gamma distributions, Logit-link functions are used for Binomial (or Bernoulli distributions).
Log link function	A link function in which f = log(linear predictor)
Log (logarithm)	A mathematical transformation of a number that makes it a lot smaller! Log(x) = base^x. Base is often chosen to be 10, but the natural log, usually denoted ln(x), is when the base is chosen to be 2.718 (for various mathematical reasons we won't go into). Logging a number is reversed by exponentiating it. And exponentiation is reversed by logging.
Log-likelihood	Likelihoods are often very very small numbers, and it is easier in many ways to work with the log of the likelihood (denoted LL) - which is often quite a large negative number. Helpful to remember that the more positive a LL is (the closer it is to zero), the higher the likelihood of the data given the model. The logs used to compute log-likelihoods are natural logs
Logistic regression	An alternative name for a glm that models binary data using a logit link function.
Logit link function or transform	logit transforms are applied to probabilities. Logit(p) = ln(p/(1-p)). The transformed probability can range from -infinity to +infinity. Used as the link function when modelling binary data.
Log Normal distribution	A lognormal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then Y = ln(X) has a normal distribution. It has two arguments - the mean and standard deviation of the logged variate. The variates are non negative.
Mann Whitney test	The non-parametric equivalent of a two-sample T test. Observations of each of the two samples are not required to come from particular distributions, but it is assumed the samples share the same shaped distribution (although the median values may well be different).
Main effect	A name of a categorical or continuous explanatory variable acting alone (i.e. not in an interaction).
MANOVA	This stands for multivariate analysis for variance and is used for examining variation between groups characterized by multiple response variables.
Maximum likelihood	When the coefficients (or parameters) of a model have been selected to maximize the value of the likelihood function - thereby maximizing the likelihood of the observed values of the response variable given the algebraic structure of the model. Coefficients fitted this way are said to be maximum likelihood estimates.

MCMC (Monte Carlo Markov Chain)	Markov chain Monte Carlo (MCMC) methods comprise a class of algorithms used to fit Bayesian models to data. Common algorithms include Metropolis-Hastings, and Gibbs Sampling. Simplistically speaking, they generate proposed parameter combinations that are used to evalulate the likelihood of a model, and these are accepted (or rejected) in a clever way that enables posterior distributions of the parameters in the model to be estimated from the accepted sample.
Mean	The sum of a set of numbers divided by the number of values in the sum. The mean of {2, 4, 12} = 6. Synonymous with average.
Median	The 'middle' value of a set of numbers, such that the same number of values are smaller and greater than this number. If x = {1,2,3,4,5,6,7,8,9} the median is 5. If x = {-1,2,3,4,5,6,7,8,900} the median is still 5.
Mixed model	A model (for example a general linear model, or generalised linear model) that contains both fixed and random effects.
Mode	The most likely value generated by a probability density function, i.e. the value corresponding to the 'highest' part of the pdf. This may or may not be the same as the mean or median depending on whether the pdf is symmetrical or not. A Normal (or Gaussian) distribution is always symmetric so the mean, median and mode are all identical. However, many other distributions are not symmetric in this sense.
Model Comparison	Any approach to inference based on the relative performance of two (often quite closely) related models that potentially might explain the same thing. Model comparison may be conducted in many different ways but comparisons of likelihoods or different information criteria are common.
Model Selection	A process whereby a number of different models are fitted to the same data in order to find the 'best' one. If models are nested then they may be compared using Likelihood Ratio Tests (LRTs), and if not nested using AIC. Views on the appropriateness of model selection vary, but care should be taken not to data dredge.
Model structure	A mathematical description of a model written in general algebraic terms. In its simplest form perhaps F(i) = c + m x(i) + c, or F(i) = c + m x(i) + a(j). It is important to always be aware of the algebraic structure of the model you have chosen to fit to data. Synonymous with algebraic structure of the model.
Moments	Moments are properties of probability density functions that include means, variances, skewness and kurtosis.
Monotonic	A monotonic function is a function which is either entirely nonincreasing or nondecreasing (i.e. it isn't at all wiggly!)
Most complex minimal model	A term we have coined to describe a model from which no explanatory variables or interactions could be removed without significantly reducing the likelihood of (response variable) data given the model.
Most complex plausible model	A term we have coined to describe a model that contains all the main effects and interactions that might (based on previous knowledge or expert opinion) plausibly contribute significantly to explaining the variation in the response variable. It may well turn out that some of these interactions or main effects could be removed without significantly reducing the likelihood of the (response variable) data given the model.
Multiple regression	Usually a name for a glm that assumes normally distributed observations of the response variable and multiple continuous explanatory variables.
Multivariate statistics	This usually refers to a familiy of statistics that considers variation in multiple different response variables simultaneously; glm's are univariate because there is only one response variable that is modelled on the left-hand side, even though of course there may be multiple explanatory variables on the right-hand side.
Natural log (logarithm)	A log that uses 2.718282 as its base. Natural logs are reversed by exponentiating using the exp(x) function.

Negative Binomial distribution	A discrete probability density function defining non-negative integers using two arguments (these arguments are not usually the mean and the standard deviation, but the mean and standard deviation can be deduced from these arguments). Poisson distributions are a special case of a Negative Binomial distribution. Negative Binomial distributions always have a larger standard deviation than a Poisson distribution and so are often used for count data that are 'overdispersed' relative to a Poisson distribution. https://en.wikipedia.org/wiki/Negative_binomial_distribution
Nested models	A simple model is nested within a more complex model if the simple model can be obtained by deleting terms from the complex model. Only nested models can be compared with Likelihood Ratio Tests. Nested is sometimes used in an entirely different sense to make reference to a hierarchical model when one (often random) effect may be nested within another.
Nominal	A nominal variable is a categorical variable the possible observations of which have no natural order.
Non-orthogonal	Two variables (usually two different explanatory variables) are said to be non-orthogonal (or synonymously, collinear) if they are positively or negatively correlated with each other. Correlations can exist between variables that are continuous or categorical. Non-orthogonality can lead to difficulties interpreting your output. Non-orthogonality can be assessed using Variance Inflation Factors.
Non-parametric test/statistics	A family of statistical tests that don't require parameters to be estimated, and are sometimes called 'model-free' approaches. They don't require an assumption that the data you are working with derive from a particular distribution, so generally make far fewer assumptions than so called parametric tests (such as glm's), but generally provide slightly reduced statistical power compared to equivalent parametric tests. Examples of non-parametric tests include Mann-Whitney, Wilcoxon Rank, Kruskall-Wallis and Friedman tests, and Spearman rank correlation.
Normal distribution	A Normal distribution defines the classical symmetric bell-shaped curve. It is defined by two arguments, the mean and the standard deviation (or variance = the square of the standard deviation). Normal distributions are continuous distributions that describe real numbers defined on the interval -infinity to +infinity.
Nuisance variables	Explanatory variables (including random effects) that are not of particular interest to the investigator other than that they may need to be accounted for. For example, if an experiment must be performed in a number of blocks, but the blocks are of no particular scientific interest, block will likely be included in the model and possibly referred to as a nuisance variable.
Null deviance	The deviance associated with the null model f = c. Technically, twice the negative log likelihood of the data for the intercept only model plus a constant. https://www.statology.org/null-residual-deviance/
Null hypothesis	A null hypothesis is a type of statistical hypothesis that proposes that no statistical significance exists in a set of given observations. Hypothesis testing is used to assess the credibility of a hypothesis by using sample data. In the simple model Y = c + mX, a null hypothesis might be that that m = 0 (and thus there is no relationship between Y and X). The null hypothesis can be tested by estimating m and determining whether the estimate is judged significantly different to zero. It is possible that every parameter estimated in a model is associated with a null hypothesis, so be careful not to test too many hypotheses (or consider undertaking Bonferroni corrections if you do).
Null model	An intercept only model, in which no explanatory variables are included at all. In R this is written as: `Null_model <- glm(y ~ -1)`
Offset	A continuous explanatory variable that is represented in a model by a slope that is fixed to be one. Often used when the response variable needs to be standardized by the continuous explanatory variable. In R this could be performed with: `glm(y~x+offset(w))`
One sample T-test	A test that uses a T-statistic to examine whether the mean of a set of observations differs significant from a particular fixed value. Assumes the observations are Normally distributed.

One way ANOVA	ANOVA in this context usually refers to a model (for example a glm) with just one categorical explanatory variable (and any number of levels).
Ordinal	An ordinal variable is a variable for which the possible observations of which have a natural order.
Ordinal GLM	A GLM constructed to test an ordinal categorical response variable (for example if the data were say: strongly dislike, dislike, neutral, like, strongly like).
Over-fitting	An overfitted model is one in which an excessive number of parameters (or coefficients) fit the noise in the data rather than the signal. The result is more variation explained, but low predictive ability, and less error (or residual) degrees of freedom, and consequently higher standard errors for the estimated parameters (or coefficients)
Over parameterized	A model with too many parameters, potentially leading to difficulties of model fitting (parameter estimation), interpretation (identifiability) or low predictive ability. This arises from an insufficient amount of information in the data to fit such a complex model.
Overdispersion	When there is more variation in your response variable than the distribution you are using to model it usually generates. Overdispersion can only be described in reference to a particular distribution, and only for those distributions where the variance is constrained (i.e. Poisson and Bernoulli). Distributions with more than one argument can usually model whatever variance you require, hence overdispersion isn't a relevant concept when using Normal or Negative Binomial distributions.
P-value	The p-value tells you how often you would expect to see a test statistic as extreme or more extreme than the one calculated by your statistical test if the null hypothesis of that test was true. The p-value gets smaller as the test statistic calculated from your data gets further away from the range of test statistics predicted by the null hypothesis. The p-value is a proportion: if your p-value is 0.05, that means that 5% of the time you would see a test statistic at least as extreme as the one you found if the null hypothesis was true. It is a common convention to use 0.05 as the threshold for statistical significance, but particularly when evaluating multiple p-values it should probably be less (see Bonferroni correction).
Parameter	A parameter usually refers to a coefficient in a model - perhaps a slope, or an 'adjustment' for a certain level of a categorical explanatory variable. Generally speaking, anything that is estimated from your data.
Parametric test/statistics	A family of statistical tests that require parameters to be estimated associated with a model with a defined algebraic structure. Such tests do require an assumption that the response variable you are working with derives from a particular distribution, and they generally provide greater statistical power compared to equivalent non-parametric tests.
Parsimony	The parsimony principle for a statistical model states that a simpler model with fewer parameters is favored over more complex models with more parameters, provided the models fit the data similarly well.
Pearson correlation	A statistic used to assess correlation between two Normally distributed variables
Pearson residuals	The raw residual divided by the standard deviation of the response variable. Pearson residuals can be used to identify outliers that are unusually large in an objective way.
Percentile	A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value (or score) below which 20% of the observations may be found.

Poisson distribution	A discrete probability density function defining non-negative integers defined by one argument which represents both the mean and the variance of the distribution. Poisson distributions are a special case of a Negative Binomial distribution. Often used for count data, but often lacking the required variance, in which case the Negative Binomial distribution is likely to be the next best option. https://en.wikipedia.org/wiki/Poisson_distribution
Polymomial	The sum of several terms that contain different powers of the same variable(s). For example: f = c + m1*X + m2*X^2 is a polynomial (containing a single quadratic term)
Posterior distribution	Exclusive to Bayesian statistics, a posterior distribution for each parameter (or coefficient) in the model is obtained that is conditional on the data and the priors for the parameters. Posteriors are often constructed from accepted proposals of parameter values generated through MCMC.
Post-hoc test	A post hoc test is usually used only after we find a statistically significant categorical explanatory variable and need to determine which levels differ from which other levels - as opposed to which differ from the reference level. https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/post-hoc/
Power	The statistical power of a hypothesis test is the probability of detecting an effect, if there is a true effect present to detect. It is generally accepted that the design of data collection and analysis should give you an 80% or greater chance of finding a statistically significant difference when there is one. https://www.scribbr.com/statistics/statistical-power/
Predictive ability	The ability of a model to predict observations of the response variable not included in the sample of observations used to fit the model. So the ability to predict 'new data'. One way to assess this is through cross-validation. Models that have been underfitted will perform poorly as they will not have captured all of the real signal in the available data, while models that have been overfitted will also perform poorly since the model will have fitted noise that will contaminate accurate prediction.
Principal Components Analysis (PCA)	A method by which a large number of correlated explanatory variables might be reduced to fewer while preserving the majority of the variation contained within them. Essentially a process of redefining the primary axes of variation in terms of linear combinations of the original variables, and rotating the data so that they align to these new axes. Useful for dimensionality reduction, and identifying groupings in a response variable that might otherwise be hard to recognize (see Appendix P).
Prior distribution	Exclusive to Bayesian statistics, any parameter (or coefficient) in a model will require a 'prior' describing the distribution of its possible values prior to consideration of the current data to which the model is being fitted. These priors may be informed by previous knowledge of the parameter (an 'informative prior') or not (less informative).
Probability Density Function (pdf)	A formal way of referring to a particular distribution of a random number. The function generates a likelihood corresponding to any variate in the range over which the pdf applies. Strictly speaking pdfs refer to continuous random variates (for example: Gaussian or Normal, Log-normal, Gama, Beta, etc)
Probability Distribution (pd)	A formal way of referring to a particular distribution of a random number (for example: Gaussian or normal, Binomial, Bernoulli, Beta, Poisson, etc). The function generates a likelihood (if the variate is continuous) or probability (if the variate is discrete) corresponding to any variate in the range over which the pd applies.
Probability Mass Function (pmf)	A form of Probability Distribution applying to a discrete distribution that gives the probability (as opposed to a likelihood) that a discrete random variable is exactly equal to some value.
Pseudo-replication	Pseudo-replication arises when the number of observations of the response variable is inflated by potentially correlated observations (often taken from repeated observations from the same subjects), and the correlation is not accounted for in the model.

Pseudo R squared	A measure that behaves like R-squared for general linear models. Pseudo R squared may be computed in a variety of different ways, but values are simplistically = (Null deviance - Residual deviance)/(Null deviance) x 100. It is a measure of how a particular model reduces the deviance relative to the null model. Pseudo R Squared value may be computed inclusive ('conditional') or exclusive ('marginal') of random effects.
qqplot	The Q-Q plot, or quantile-quantile plot, is a graphical tool to help assess if a set of data plausibly came from some distribution such as a Normal distribution. https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0
Quadratic terms	It might be that variation in the response variable is best captured by raising a continuous explanatory variable to a certain power - usually squaring it. The squared terms are referred to as quadratic terms. Useful for modellng relationships between the response and explanatory variables that are not monotonic.
Quantile	Each of any set of ranked values of a variate which divide a distribution into equally sized groups, each containing the same fraction of the total population. A median is a quantile (usually referred to as a 50% quantile) as the median divides the data into two equally sized groups. If each of n ranked observations, x_i (i = 1 .. n) is given its own quantile, then the i/n^{th} x 100 quantile (or percentile) = x_i.
Quartile	Quartiles are a special case of a quantiles, that divide a date set into 4 equally sized groups. For example, the lower quartile, or first quartile (Q1), is the value under which 25% of data points are found when they are arranged in increasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order.
R-squared	A concept that only applies when fitting models with least squares. R^2 is defined as the proportion or percentage of total variation in the response variable that has been explained by the model. The total variation is assumed to be the 'total sums of squares'. See also pseudo R squared.
Random effect	Explanatory variables may be fixed or random. Random effects are used for categorical explanatory variables, usually with at least 5 levels, when the investigator is not directly motivated to understand the differences between the different levels. Often used to handle repeated measures of a response variable on something (an individual, a location, etc). Requires only one degree of freedom (for the variance of the (usually) Normal distribution from which the random 'adjustments' for different levels of the random effect are assumed to derive), regardless of the number of levels.
Random number	Synonymous with random variate. Random numbers can be generated from probability density (or mass) functions.
Random variate/variable	A random variate is a random number - but important to be clear from which distribution a variate is assumed to come from - for example a random variate may be come from a Normal distribution, or a Bernoulli distribution .. or many many others!
Raw residual	The simple difference (observed - fitted) between an observed and fitted value.
Record	A term we use to refer to an individual observation of the response variable and all of its associated explanatory variables. A single row from a flat data sheet.
Real number	A number is real if it has decimal places (technically .. a very large number of decimal places). Real numbers are continuous.
Regression	This is a term that is often used to mean 'a model in which the explanatory variables are continuous', but in fact it's a great deal more general, and really refers to the general process of fitting one response variable to one or more explanatory variables (continuous or categorical) through a wide range of possible different functional forms (linear regression being the simplest and most often encountered). It does not assume a particular method of fitting the model.

REML	Stands for 'Restricted Maximum Likelihood'. Only encountered when working with Gaussian mixed models. It is an alternative to 'Maximum Likelihood' when fitting a model to data. Your final models should be fitted with REML=TRUE as this gives unbiased estimates of your variance components (random effects). However, when comparing models that differ only in their fixed effects (for example when using LRTs for model selection, or to assess the significance of fixed effect terms in your final model), the log-likelihoods of these models should be computed using full maximum likelihood (ML, REML=FALSE).
Repeated measures	Repeated measures arise when sequential observations of the response variable are made from (or by) the same subject. The subject might be an individual, a location, a time, an observer - many possible forms. The result is that these commonalities might induce a correlation between observations that would need to be accounted for by the model - often by the use of a random effect (the random effect might be individual id, location, time, observer etc).
Residual	The difference between an observation of the response variable and the fitted value for that observation. In Generalised Linear Models residuals may be transformed in various ways (standard, Pearson, Deviance etc). https://www.datascienceblog.net/post/machine-learning/interpreting_generalized_linear_models/
Residual analysis	A general term for the tests it is necessary to conduct after fitting a model, to ensure the major assumptions of the fitting process have not been importantly violated. Synonymous with diagnostic analysis. Usually involves looking for trends in the patterns of residuals when plotted against fitted values, and checking on the distribution of residuals (more straightforward for general than generalised linear models). The DHARMa package is very good for this.
Residual degrees of freedom	Usually, the number of observations of the response variable less the number of coefficients estimated by the model. Can only be established once a model is specified.
Residual deviance	A specific way of referring to the deviance of a particular model. Technically, twice the negative log likelihood of the data for a particular model plus a constant. https://www.statology.org/null-residual-deviance/
Residual variation	A general term to refer to variation that remains unexplained by the model
Response (or dependent) variable	The variable that we seek to account for variation in. Also referred to as the 'dependent variable' or sometimes the 'y-variable'. GLMs assume that observations of the response variable derive from particular distributions (Normal or Gaussian, Bernoulli, Poisson etc)
Right hand side	Reference to the right-hand side of an expression. For example, in the model f = c + mx , f is the left hand side, and c + mx the right hand side.
Saturated model	A model that has a parameter for every observation of the response variable. The degrees of freedom required by the model equals the number of observations of the response variable.
Skewness	A distribution is skewed if one of the tails is extended (stretched out) relative to the other. A distribution might be left-skewed, or right-skewed.
Spearman rank correlation	A statistic used to assess correlation between two variables, the variables are not assumed to come from any particular distribution.
Spline	A spline is a highly flexible function devised to put a 'wiggly line' through a set of points. More technically a spline is a type of piecewise polynomial function. Splines can be devised to track the wiggly-ness of data more-or-less closely depending on the number of nodes permitted and the complexity of the polynomial functions adopted.
Standard deviation	A quantity expressing by how much the members of a group differ from the mean value for the group. A measure of the 'spread' of a probability density function. More specifically, the square root of the average value of squared differences from the mean. The standard deviation is the square root of the variance.

Standard error	A standard error can be regarded as a special case of a standard deviation. We might talk about the standard error of a coefficient (or parameter) - but we could equally easily (and correctly) talk of the standard deviation of the same coefficient (or parameter) and we'd be referring to the same thing. Usually standard error refers to the standard deviation of a coefficient (or parameter). When using either term be clear exactly what the subject of the standard error (or deviation) is.
Standard Normal Distribution	A special case of a Normal or Gaussian distribution which has mean = 0 and standard deviation = 1.
Standardized residuals	The raw residuals divided by the true standard deviation of the residuals. As the true standard deviation is rarely known, a standardized residual is rarely used. A Studentized residual is the raw residual divided by the estimated standard deviation of the residuals.
Statistically significant	A result that is sufficiently unlikely to be explained by chance alone. A pattern in the data at least as clear as that observed (often assessed by a test statistic of some sort) that has less than a pre-specified probability (often 0.05) of arising by chance.
Stepwise backward	The process of model selection whereby a complicated model is reduced to a simpler one through the removal of less significant explanatory variables
Stepwise forward	The process of model selection whereby a simple model is made more complicated by the addition of terms that are retained if they are judged to be significant.
Sum of Squares	This is a rather imprecise term used when fitting models using Least Squares. The term is usually used in reference to the total sums of squares (a measure of the total variation in the response variable we wish to attempt to account for), the explained sums of squares that accounted for by the explanatory variables), and the unexplained or error sum of squares (that which the explanatory variables cannot account for). You will also see the mean explained sums of squares (the explained sums of squares divided by the number of coefficients in the model), and the mean error sum of squares (the error sum of squares divided by the error (or residual) degrees of freedom). These mean sums of squares are used to compute F statistics.
T distribution	A T distribution (or 'Students T distribution') is a bit like a Normal distribution but with fatter tails, the tails becoming thinner the larger the (residual) degrees of freedom - eventually converging on a normal distribution with about 30+ dfs. https://en.wikipedia.org/wiki/Student%27s_t-distribution
T statistic	A T statistic tells you how many standard deviations a parameter (or coefficient) is from a chosen value. The chosen value is often zero (corresponding perhaps to the null hypothesis for a coefficient (or parameter). T statistics are assumed to be distributed according to a T distribution. Although Z and T statistics are calculated the same way, Z statistics are assessed using a Z-distribution, and T statistics using a T-distribution. If the error (or residual) degrees of freedom is large, they will generate almost identical outputs (Z and T statistics with an absolute magnitude of 2 or more tend to be significant), but the T statistic and distribution is more accurate when the error (or residual dfs) are modest (~ < 30).
T-test	Technically a T-test is any test the involves the use of a T statistic (and used to test the significance of coefficients (or parameters) generated by general and generalised linear models), but often a T-test is used to refer to a simpler family of tests including the one and two-sample T-test, and the paired T-test.
Term	An adjustment in the right-hand side of a GLM. A term may be the intercept, an adjustment for a fixed categorical or continuous explanatory variable, an interaction, or a random effect.
Test statistic	A general term for a 'number' derived from a data set .. in such as a way as to have a known distribution. Common distributions that describe well known test statistics are the T distribution, the Z distribution (Standard Normal), the chi-squared distribution, the F distribution, and so on.

Two sample T-test	A test that uses a T-statistic to examine whether the mean of two sets of observations differs significantly from each other. Assumes both sets of observations are Normally distributed
Two sample (paired) T-test	A test that uses a T-statistic to examine whether the mean of two sets of paired observations differs significantly from each other. Assumes both sets of observations are Normally distributed. For example, the paired (repeated) observations might be of an animals weight before and after a period of dieting. Other things being equal paired designs are more powerful than unpaired designs as the natural variation in the subjects (in this example the original size of the animals) is factored out. Such designs may also be analysed through models constructed to deal with repeated measures.
Transform	Traditionally, it was not uncommon to transform a response variable in an attempt to make it 'more Normal'. Common transforms in increasing order of normalizing strength would be: square root, log, and inverse. Box Cox transformations enable you to select the 'optimum strength'. The inference is robust to these transformations, but remember that analysis applies to the transformed variable .. not the original untransformed variable. A good understanding of generalised linear models substantially lessens the need to utilize such transforms.
Two-way ANOVA	ANOVA in this context usually refers to a model (for example a glm) with just two categorical explanatory variables.
Underpowered	An experimental or data collection process combined with analysis that is unlikely to detect an effect of a size deemed to be of interest. Underpowered is often assumed to mean a power < 80%.
Uniform distribution	A bounded distribution in which all variates between a minimum and maximum value are equally likely.
Univariate model	A model that has only one response variable (as opposed to a multivariate approach which would consider multiple response variables - such as MANOVA.
Variance	A quantity expressing by how much the members of a group differ from the mean value for the group. A(nother) measure of the 'spread' of a probability density function. More specifically, the average value of squared differences from the mean. The variance is the square of the standard deviation
Variance Inflations Factors	A variance inflation factor (VIF) detects collinearity in regression analysis. Collinearity is when there's correlation between explanatory variables in a model; it's presence can adversely affect your results. The VIF estimates how much the variance (or standard deviation or standard error) of a models coefficients (or parameters) are inflated due to collinearity in the model. Some people regard VIFs > 5 as a serious issue that should be addressed. https://www.statisticshowto.com/variance-inflation-factor/
Variable	Could refer to an explanatory or a response variable. Models are comprised of just two things: variables (data), and parameters (or coefficients) that are estimated from data through the process of fitting the model to data
Variate	A term used to describe a 'draw' of a random number from a particular distribution. We might talk of a 'Normal variate' .. A random number chosen from a Normal distribution.
Wald statistic	A little like a Z or T statistic, a Wald statistic indicates how many variances (as opposed to standard deviations) a parameter (or coefficient) is from a chosen value. The chosen value is often zero (corresponding perhaps to the null hypothesis for a coefficient (for parameter). A Wald statistic has a Chi-squared distribution with 1 degree of freedom.
Weibull distribution	A continuous probability density function that generates non-negative real numbers. https://en.wikipedia.org/wiki/Weibull_distribution
Wiggly	A technical term (-:) for describing a relationship that goes up and down a lot!
Wilcoxon Rank test	The non-parametric equivalent of a 2-sample T test or a Mann-Whitney Test. Observations of the samples are not required to come from particular distributions.

Z distribution	A Z distribution is the so-called 'Standard Normal' distribution with mean zero and standard deviation 1, denoted N(0,1).
Z statistic	A Z statistic tells you how many standard deviations a parameter (or coefficient) is from a chosen value. The chosen value is often zero (corresponding perhaps to the null hypothesis for a coefficient (for parameter). Although Z and T statistics are calculated the same way, Z statistics are assessed using a Z-distribution, and T statistics using a T-distribution. If the error (or residual) degrees of freedom is large, they will generate almost identical outputs (Z and T statistics with an absolute magnitude of 2 or more tend to be significant), but the T statistic and distribution is more accurate when the error (or residual dfs) are modest (~ < 30).
Zero-inflation	When there are more zero's in your set of observations of the response variable than a model can account for. If a data set is zero-inflated in this way, you may choose to address the issue with a zero-inflated model or hurdle model.
Zero-inflated model	Zero-inflated models are two-part models that specify one process for zero counts and another process for non-negative counts (that may include zero's). They differ from hurdle models in the sense that both parts of the model are able to generate zero's.
Zero-truncated distribution	A discrete distribution that has the zero's removed and the probability masses re-normalized to sum to one.