MATH3823 Generalized Linear Models

Robert G Aykroyd

2024-03-14

Table of contents

Weekly schedule

Items will be added here week-by-week and so keep checking when you need up-to-date information on what you should be doing.

Week 8 (18 - 22 March)

- **Before next Lecture:** Re-read Chapters 4 and 5 ready for practical.
- **Lecture on Tuesday:** Start *Chapter 6: Loglinear Modelling* with *Sections 6.1-6.3.*
- **Computer Practical on Tuesday:** Work on Assessed Practical.
- **Computer Practical on Wednesday:** Work on Assessed Practical.
- **Lecture on Thursday:** Complete *Chapter 6: Loglinear Modelling* with *Section 6.4-6.5.*
- **Weekly feedback:** Start Exercises for Chapter 6 and check answers with solutions.

Advanced notice

- **Module Assessment:** Set on 12 March with submission deadline 23 April (that is after the break). You will be expected to write a short report based on an RStudio practical.
- **Computer classes:** Supervised session on 19/20 March – check your timetable.
- **Generative AI usage within this module:** The assessments for this module fall in the red category for using Generative AI which means you must not use Generative AI tools. The purpose and format of the assessments makes it inappropriate or impractical for AI tools to be used.

Week 7 (11 - 15 March)

- **Before next Lecture:** Re-read *Chapter 5: Sections 5.1-5.3.*
- **Lecture on Tuesday:** Complete *Chapter 5:* by looking at *Section 5.5 - Application to dose-response experiments.*
- **Lecture on Thursday:** Consider selected Exercises from previous Chapters and discuss Assessed Practical. If time start *Chapter 6: Log-linear Modelling.*
- **Weekly feedback:** Complete Exercises for Chapter 5.

Week 6 (4 - 8 March)

- **Before next Lecture:** Re-read *Chapter 4: Sections 4.1-4.2.*
- **Lecture on Tuesday:** Complete *Chapter 4* with *Sections: 4.3 Model deviance, 4.4 Model Residuals & 4.5 Fitting GLMs in R*
- **Lecture on Thursday:** Start *Chapter 5: Modelling Proportions* with *Sections 5.1 & 5.2.*
- **Weekly feedback:** Complete Chapter 4 Exercises.

Week 5 (26 February - 1 March)

- **Before next Lecture:** Re-read all of *Chapter 3.*
- **Computer Practical:** Either on Tuesday or Wednesday, attend supervised computer class – see your timetable. For information see *Week 5 Computer Practical* folder on Minerva.
- **Lecture on Tuesday:** Start *Chapter 4* by covering *Section 4.1: The identically distributed case.*
- **Lecture on Thursday:** Continue *Chapter 4* with *Section 4.2: The general case.*
- **Weekly feedback:** Check your answers on Chapter 3 Exercises with online solutions. Start Chapter 4 Exercises.

Week 4 (19 - 23 February)

- **Before next Lecture:** Be confident with all material up to, and including, *Section 3.2: The GLM structure.*
- **Lecture on Tuesday:** We will cover *Section 3.3: The random part of a GLM*, *Section 3.4: Moments of exponential-family distributions* and, if time permits, *Sections 3.5: The systematic part of the model.*
- **Lecture on Thursday:** Complete Chapter 3 by covering *Section 3.6: The link function.* Then, we will consider selected Exercises from *Section 3.7.*
- **Weekly feedback:** Complete Exercises in *Chapter 3.*

Week 3 (12 - 16 February)

- **Before next Lecture:** Please re-read *Section 2.5: Model shorthand notation* and *Section 2.6 Fitting linear models in R*, and read *Section 2.7: Ethics in statistics and data science.*
- **Lecture on Tuesday:** We will start Chapter 3 with *Section 3.1: Motivating examples* and *Section 3.2: The GLM structure*
- **Before next Lecture:** Please re-read *Sections 3.1* and *3.2* carefully.
- **Lecture on Thursday:** CANCELLED due to illness.
- **Weekly feedback:** Complete the first two Chapter 3 Quizzes and start the Exercises in *Section 3.7.*

Week 2 (5 - 9 February)

- **Before next Lecture:** Please re-read *Section 2.1: Overview* and *Section 2.2: Linear models*, and read *Section 2.3: Types of normal linear model*.
- **Lecture on Tuesday:** We will cover *Section 2.4: Matrix representation of linear models* and briefly *Section 2.5: Model shorthand notation*.
- **Before next Lecture:** Please re-read *Sections 2.4* and *2.5* carefully.
- **Lecture on Thursday:** We will cover *Section 2.6: Fitting linear models in R* then discuss selected Exercises from Chapters 1 and 2.
- **Weekly feedback:** Complete the Chapter 2 Quizzes and complete the Exercises in *Section 2.8*.

Week 1 (29 January - 2 February)

- **Before first Lecture:** Please read the *Overview*.
- **Lecture on Tuesday:** We will briefly cover all material in *Chapter: Introduction*.
- **Before next Lecture:** Please re-read *Chapter 1* carefully, especially any sections not covered in Lectures.
- **Lecture on Thursday:** Start *Chapter 2: Essentials of Normal Linear Models* with *Section 2.1: Overview* & *Section 2.2: Linear models*.
- **Weekly feedback:** Complete the Chapter 1 Quizzes and self-study the Exercises in *Section 1.5*. If you have time, start Exercises in *Section 2.8*.

Provisional Weekly Lecture Schedule[a]

[a]Some sections will be left as *directed reading*, but please note that material in all sections is examinable.

Week 1	Chapter 1	All
	Chapter 2	Sections 2.1-2.2
Week 2		Sections 2.3-2.7
	Exercises	Chapter 1 & 2
Week 3	Chapter 3	Sections 3.1-3.4
Week 4		Sections 3.5-3.6
	Exercises	Chapter 3
Week 5	Chapter 4	Sections 4.1-4.2
Week 6		Sections 4.3-4.5
	Chapter 5	Sections 5.1-5.3
Week 7		Sections 5.4-5.5
	Exercises	Chapters 3 & 4
Week 8	Chapter 6	All
Easter		
Week 9	Chapter 7	Sections 7.1-7.2

Week 10		Sections 7.3-7.4
	Exercises	Chapter 6 & 7
Week 11	Revision	

Overview

Welcome!

Here is a short video [4 mins] to introduce the module.

Preface

These lecture notes are produced for the University of Leeds module *MATH3823 - Generalized Linear Models* for the academic year 2023-24. Please note that this material also forms part of the module *MATH5824 - Generalized Linear and Additive Models*. They are based on the lecture notes used previously for this module and I am grateful to previous module lecturers for their considerable effort: Lanpeng Ji, Amanda Minter, John Kent, Wally Gilks, and Stuart Barber. This year, again, I am using Quarto (a successor to RMarkdown) from RStudio to produce both the html and PDF, and then GitHub to create the website which can be accessed at rgaykroyd.github.io/MATH3823/. Please note that the PDF versions will only be made available on the University of Leeds Minerva system. Although I am a long-term user of RStudio, I am a novice at Quarto/RMarkdown and a complete beginner using Github and hence please be patient if there are hitches along the way.

RG Aykroyd, Leeds, January 3, 2024

Changes since last year

Feedback from the students last year was very positive, but there were consistent comments regarding two issues: (1) a shortage of practice exercises and the opportunity to discuss these in class, and (2) limited RStudio support in preparation for the assessment. For the first of these, additional exercises have been prepared and are included in the learning material. Also, I am trying some short quizzes so that you can check your basic knowledge. Further, I intend to set-aside some lecture time for us to discuss selected exercises. For the second, an additional computer session has been added, in Week 5 (26 February - 1 March), this is 3 weeks before the assessed practice in Week 8 (18 - 22 March). Further, a few new instructional videos will be available addressing some RStudio topics. Together,

these represents a considerable about of extra work for me, but I hope that they are helpful and so please give your feedback whenever there is an opportunity.

Generative AI usage within this module

The assessments for this module fall in the red category for using Generative AI which means you must not use Generative AI tools. The purpose and format of the assessments makes it inappropriate or impractical for AI tools to be used.

 Warning

Statistical ethics and sensitive data
Please note that from time to time we will be using data sets from situations which some might perceive as sensitive. All such data sets will, however, be derived from real-world studies which appear in textbooks or in scientific journals. The daily work of many statisticians involves applying their professional skills in a wide variety of situations and as such it is important to include a range of commonly encountered examples in this module. Whenever possible, sensitive topics will be signposted in advance. If you feel that any examples may be personally upsetting then, if possible, please contact the module lecturer in advance. If you are significantly effected by any of these situations, then you can seek support from the Student Counselling and Wellbeing service.

Official Module Description

Module summary

Linear regression is a tremendously useful statistical technique but is very limited. Generalised linear models extend linear regression in many ways - allowing us to analyse more complex data sets. In this module we will see how to combine continuous and categorical predictors, analyse binomial response data and model count data.

Objectives

On completion of this module, students should be able to:

a) carry out regression analysis with generalised linear models including the use of link functions;
b) understand the use of deviance in model selection;
c) appreciate the problems caused by overdispersion;
d) fit and interpret the special cases of log linear models and logistic regression;
e) use a statistical package with real data to fit these models to data and to write a report giving and interpreting the results.

Syllabus

Generalised linear model; probit model; logistic regression; log linear models.

University Module Catalogue

For any further details, please see MATH3823 Module Catalogue page

1 Introduction

Here is a short video [3 mins] to introduce the chapter

1.1 Overview

In previous modules you have studied linear models with a normally distributed error term, such as simple linear regression, multiple linear regression and ANOVA for normally distributed observations. In this module we will study **generalized** linear models.

Outline of the module:

1. Revision of linear models with normal errors.
2. Introduction to generalized linear models, GLMs.
3. Logistic regression models.
4. Loglinear models, including contingency tables.

> **!** Important
>
> This module will make extensive use of **R** and hence it is very important that you are comfortable with its use. If you need some revision, then material is available on Minerva under *RStudio Support*.

The purpose of a generalized linear model is to describe the dependence of a *response* variable y on a set of p *explanatory* variables $x = (x_1, x_2, \ldots, x_p)$ where, conditionally on x, observation y has a distribution which is **not necessarily** normal. Note that the normal distribution situation is a special case of the general framework and we will study that in the next Chapter.

Please be aware that in this learning material we may use lowercase letters, for example y or y_i, to denote both observed values or random variables, which is being considered should be clear from the context.

> **!** Important
>
> This module will make extensive use of many basic ideas from statistics. If you need some revision, then see *Appendix A: Basic material* on Minerva under *Basic Prerequisite Material*.

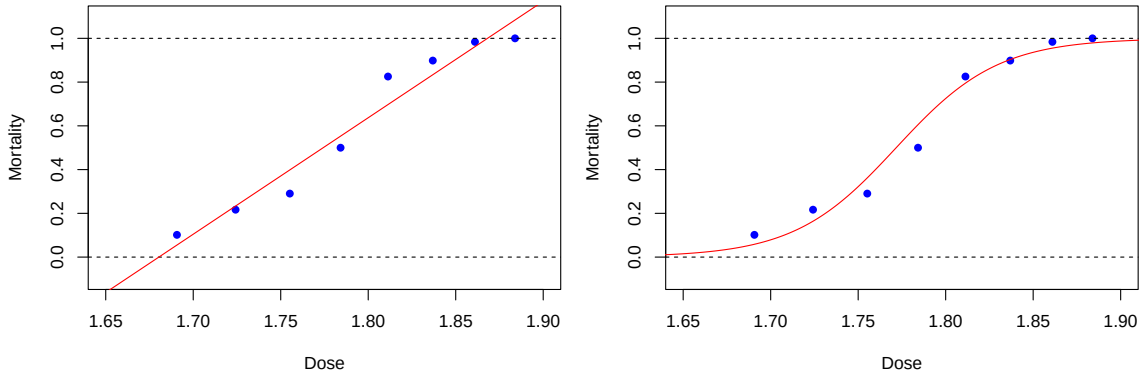
1.2 Motivating example

Table 1.1 shows data[1] on the number of beetles killed by five hours of exposure to 8 different concentrations of gaseous carbon disulphide.

Table 1.1: Numbers of beetles killed by five hours of exposure to 8 different concentrations of gaseous carbon disulphide

Dose x_i	No. of beetle m_i	No. killed y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Figure 1.1a shows the same data with a linear regression line superimposed. Although this line goes close to the plotted points, we can see some fluctuations around it. More seriously, this is a stupid model: it would predict a mortality rate of greater than 100% at a dose of 1.9 units, and a negative mortality rate at 1.65 units!



(a) Linear model

(b) Logistic model

Figure 1.1: Beetle mortality rates with fitted dose- response curves.

A more sensible dose–response relationship for the beetle mortality data might be based on the *logistic* function (to be defined later), as plotted in Figure 1.1b. The resulting curve is a closer, more-sensible, fit. Later in this module we will see how this curve was fitted using maximum likelihood estimation for an appropriate generalized linear model.

[1]Dobson and Barnett, 3rd edn, p.127

This is an example of a dose-response experiment which are widely used in medical and pharmaceutical situations.

> ⚠ Warning
>
> **Warning of potentially sensitive material.** For further information on dose-response experiments see, for example, www.britannica.com/science/dose-response-relationship.

Focus on correlation quiz

Test your knowledge recall and comprehension to reinforce idea of linear relationships and correlation.

For each situation, choose one of the following statements which you think is **most likely** to apply.

1. The diastolic blood pressure and the weight of patients attending a heart health clinic at the Leeds General Infirmary.

- (A) positive correlation
- (B) uncorrelated
- (C) negative correlation
- (D) other

2. The daily stock market closing prices of British Telecom and Virgin Media shares on the London Stock Exchange.

- (A) positive correlation
- (B) uncorrelated
- (C) negative correlation
- (D) other

3. The daily rainfall and hours of sunshine collected at a weather monitoring station in the Pennines of Yorkshire.

- (A) positive correlation
- (B) uncorrelated
- (C) negative correlation
- (D) other

4. The number of road accidents occurring at a busy roundabout and the UK Retail Prices Index.

- (A) positive correlation
- (B) uncorrelated
- (C) negative correlation
- (D) other

Click here to see explanations

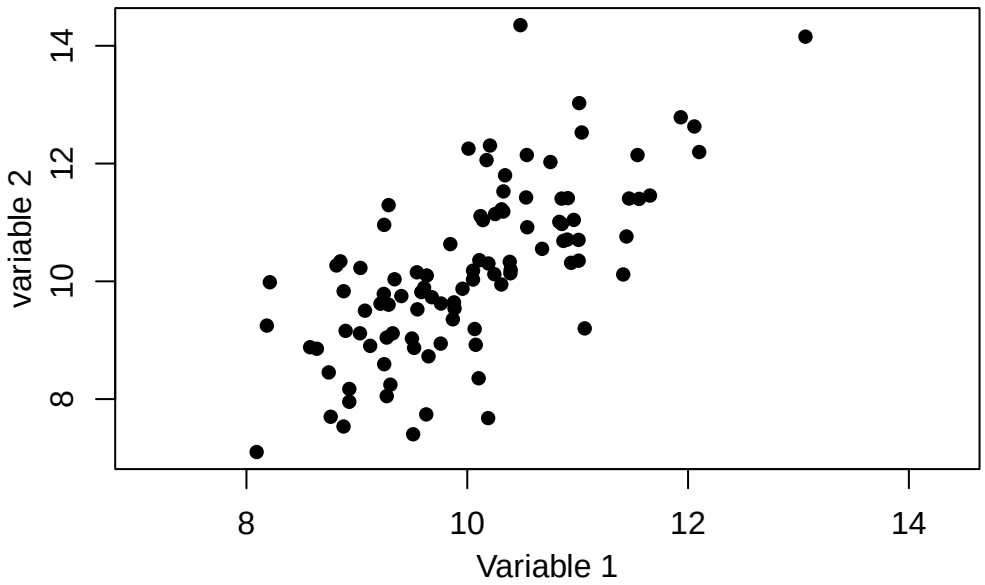
1. It is well documented that people who are over-weight (or at least high BMI) are more likely to have high blood pressure. This is true for systolic and diastolic blood pressure as well as for men and women. It is not certain if the relationship will be linear, but it will lead to a strong positive correlation value. This is likely to be a causal relationship, excess weight causes high blood pressure.
2. Although these two companies are in the same business sectors, technology and entertainment, it is unlikely that direct competition will be the main factor in the relative behaviour – if it were, then we might expect a negative correlation, that is one does well if the other does badly. Instead, they are both likely to be driven by the same economic and social trends. This is not a causal relationship but both are being driven by an (unseen) third variable. It does, however, lead to a correlation.
3. Although both weather features, rainfall and sun, can happen at the same time – perhaps leading to a rainbow – and there are very many cases when neither occurs – for example a cloudy but dry day – there is a general pattern that it will not rain when it is sunny and it will not be sunny when it rains. This leads to a negative correlation and a moderate value is likely even if the relationship is not linear. Again, this is not a causal relationship but is being driven, perhaps, by the presence of clouds.
4. It is hard to imagine that there will be a relationship between the number of road accidents and an inflation measure, hence a value of correlation close to zero – though do not expect to ever get a value of exactly zero. It is not completely, inconceivable that the number of road accidents will be higher when the economy is active, but this

is unlikely to lead to a substantial correlation value – if you know otherwise, let me know!

From each of the following scatter plots, choose from the drop down list which correlation value is **most likely**.

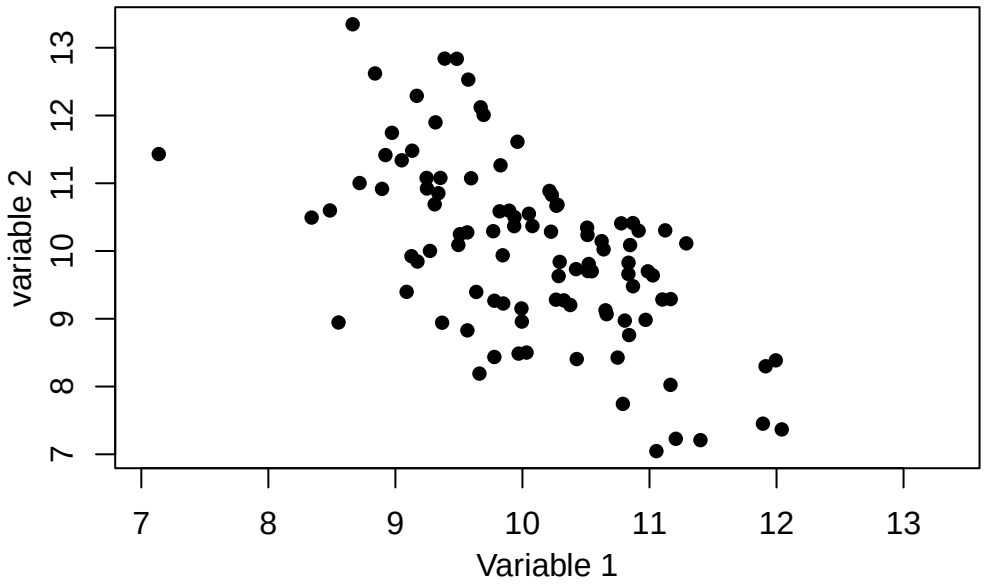
5. What is the most likely correlation for the data below?

- (A) -1.0
- (B) -0.7
- (C) -0.3
- (D) 0.0
- (E) +0.3
- (F) +0.7
- (G) +1.0



6. What is the most likely correlation for the data below?

- (A) -1.0
- (B) -0.7
- (C) -0.3
- (D) 0.0
- (E) +0.3
- (F) +0.7
- (G) +1.0

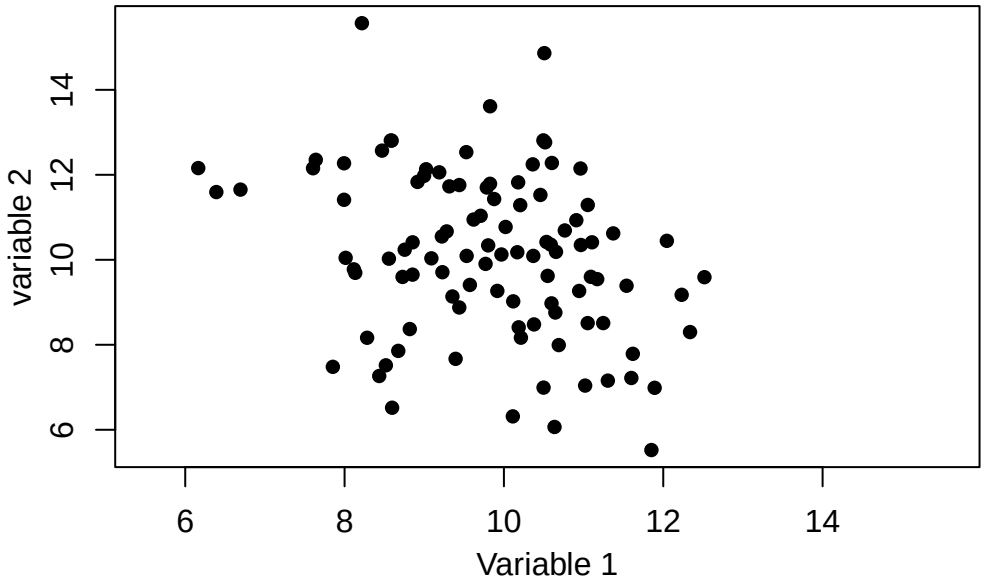


7.

What is the most likely correlation for the data below?

- (A) -1.0
- (B) -0.7
- (C) -0.3

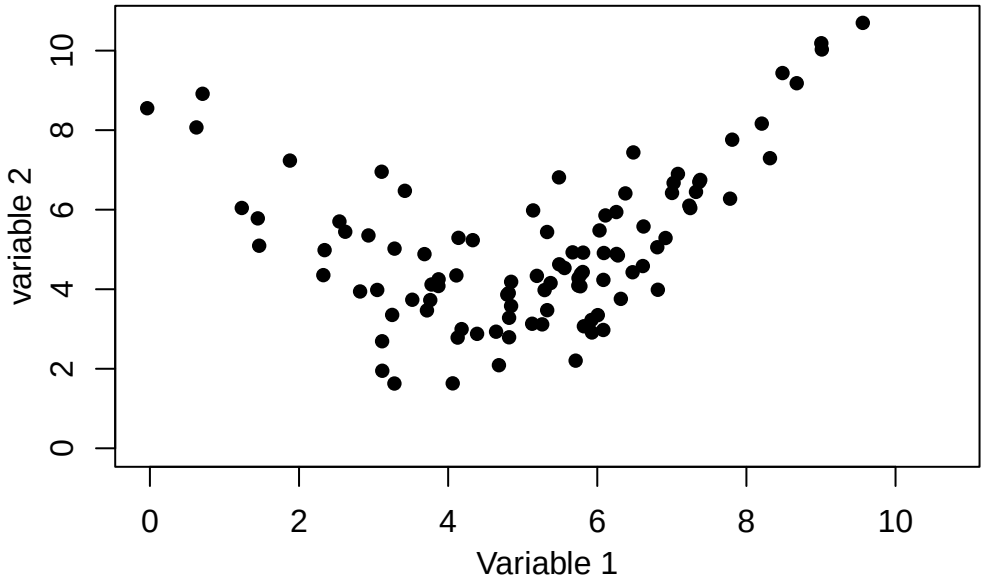
- (D) 0.0
- (E) +0.3
- (F) +0.7
- (G) +1.0



8.

What is the most likely correlation for the data below?

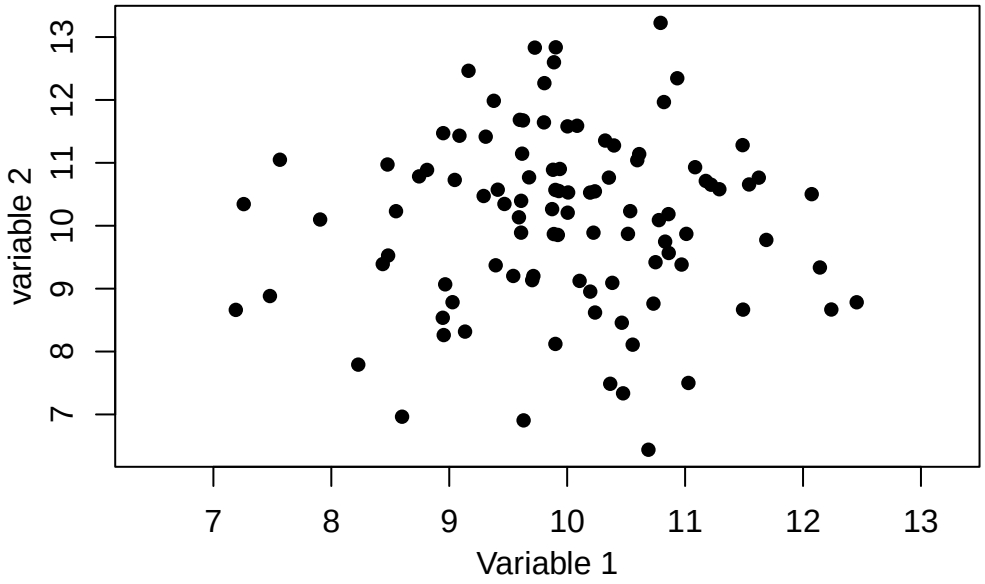
- (A) -1.0
- (B) -0.7
- (C) -0.3
- (D) 0.0
- (E) +0.3
- (F) +0.7
- (G) +1.0



9.

What is the most likely correlation for the data below?

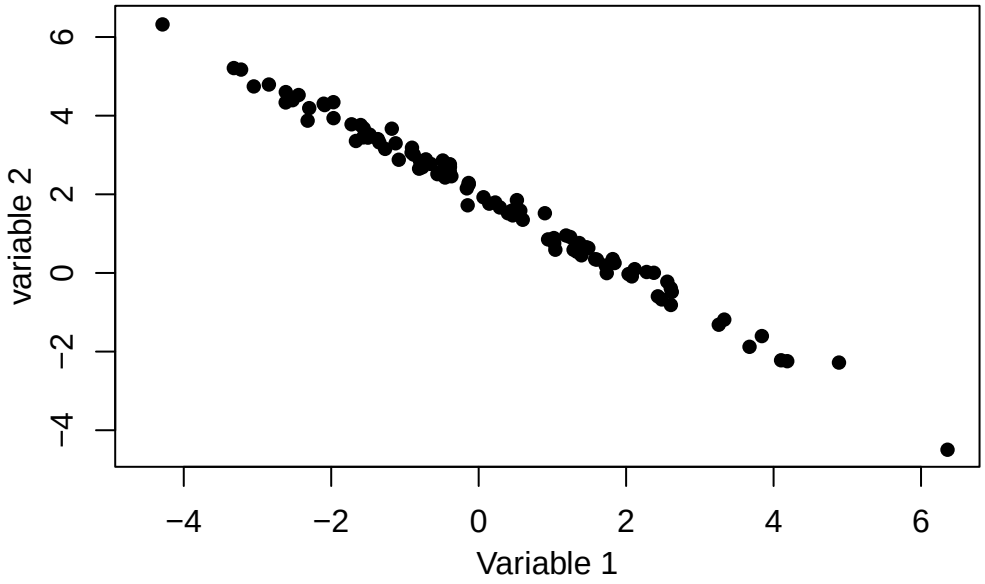
- (A) -1.0
- (B) -0.7
- (C) -0.3
- (D) 0.0
- (E) +0.3
- (F) +0.7
- (G) +1.0



10.

What is the most likely correlation for the data below?

- (A) -1.0
- (B) -0.7
- (C) -0.3
- (D) 0.0
- (E) +0.3
- (F) +0.7
- (G) +1.0



Click here to see explanations

5. Imagine dividing the plot by horizontal and vertical lines at the respective mean values. There would be a majority of points in the top-right and bottom-left indicating a positive correlation, but there are still some on the other quadrants. Hence, +1 is too high and 0.3 is too low, but would be suitable if the points were closer to a line or more dispersed, respectively. For information, the exact value is 0.70.

6. Again, imagine dividing the plot by horizontal and vertical lines at the respective mean values. This time the majority of points would be in the top-left and bottom-right quadrants, and there is moderate spread. A value -1 is too extreme and -0.3 is too close to zero, but would be suitable if the points were closer to a line or more dispersed, respectively. For information, the exact value is -0.60.

7. A similar situation to Question 6, but there is noticeably more spread. Although the majority of points are in the top-left and bottom-right quadrants there are substantial numbers in the other quadrants. It would be inaccurate to say that this shows uncorrelated variables as there is a definite negative slope to the pattern. For information, the exact value is -0.30.

8. This is a difficult one as there is a clear relationship, but it is quadratic rather than linear. For information, the exact value is 0.30. Perhaps without the accompanying graph, this correlation value would be misleading.

9. Dividing the plot by horizontal and vertical lines at the respective mean values leaves very similar numbers of points in al four quadrants. This indicates that the correlation will be close to zero – here is no relationship. For information, the exact value is 0.01.

10. Almost all of the points are in the top-left and bottom-right quadrants indicating a negative correlation. The points are very close to the linear and hence a value close to -1 is likely – such extreme cases are rare. For information, the exact value is -0.995.

1.3 Revision of least-squares estimation

Suppose that we have n paired data values $(x_1, y_1), \ldots, (x_n, y_n)$ and that we believe these are related by a linear model

$$y_i = \alpha + \beta x_i + \epsilon_i$$

for all $i \in \{1, 2, \ldots, n\}$, where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed (iid) with $\mathrm{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2$. The aim will be to find values of the model parameters, α, β and σ^2 using the data. Specifically, we will estimate α and β using the values which minimize the residual sum of squares (RSS)

$$RSS(\alpha, \beta) = \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2 . \tag{1.1}$$

This measures how close the data points are around the regression line and hence the resulting estimates, $\hat{\alpha}$ and $\hat{\beta}$, will give us a fitted regression line which is *closest* to the data.

Figure 1.2 illustrates this process. The data points are fixed but we are free to choose values for α and β, that is to move the line up and down and to rotate it as needed. In general, the data points, however, do not sit exactly on any line. For any values of α and β the value on the line $y = \alpha + \beta x$ can be calculated and the discrepancy, as measured in the vertical direction, is defined as the residual, $r_i = y_i - (\alpha + \beta x_i)$. Then, the residual sum of squares is formed as the sum of the squares of these individual residuals.

It can be shown that Equation 1.1 takes its minimum when the parameters are given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \text{and} \quad \hat{\beta} = \frac{s_{xy}}{s_x^2} \tag{1.2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means,

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

is the sample covariance and

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

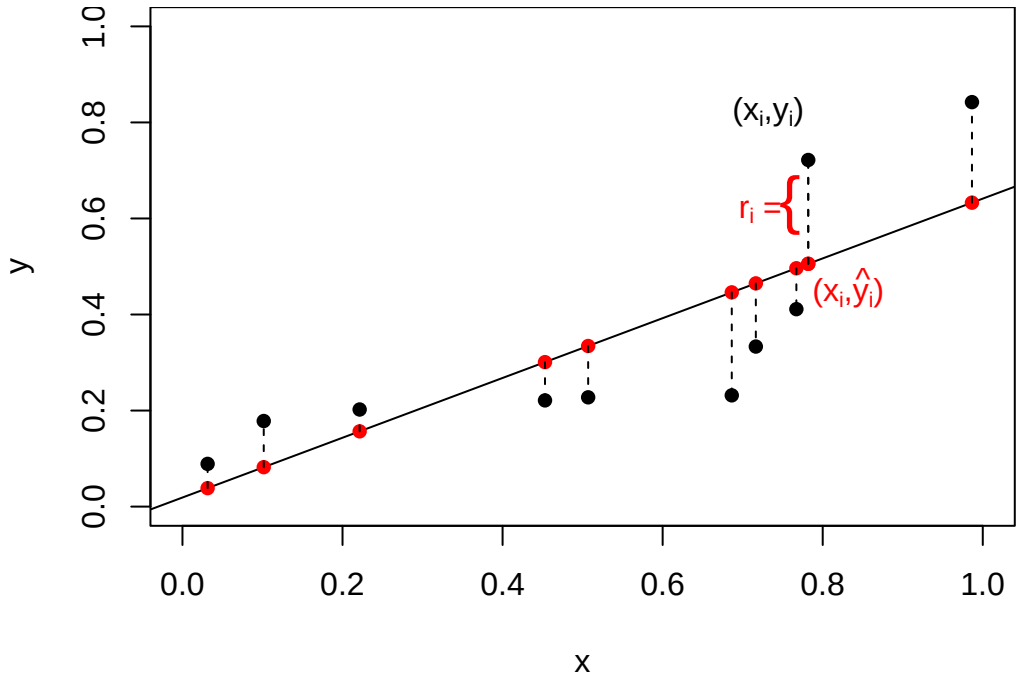


Figure 1.2: Diagram showing linear regression method

is the sample variance of the x values. It can be shown that these estimators are unbiased, that is $\mathrm{E}[\hat{\alpha}] = \alpha$ and $\mathrm{E}[\hat{\beta}] = \beta$ – see Section 1.5.

The fitted regression lines is then given by $\hat{y} = \hat{\alpha} + \hat{\beta}x$, the fitted values by $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$, and the model residuals by $r_i = \hat{\epsilon}_i = y_i - \hat{y}_i$ for all $i \in \{1, \ldots, n\}$.

To complete the model fitting, we also estimate the error variance, σ^2, using

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} r_i^2. \tag{1.3}$$

Note that, by construction, $\bar{r} = 0$ and, further, it can be shown that $\hat{\sigma}^2$ is an unbiased estimator of σ^2, that is $\mathrm{E}[\hat{\sigma}^2] = \sigma^2$.

Returning to the above beetle data example, we have $\hat{\alpha}$ =-8.947843, $\hat{\beta}$ = 5.324937, and $\hat{\sigma}^2 = 0.0075151$.

We will interpret the output later, but in *R*, the fitting can be done with a single command with corresponding fitting output from a second command:

```
Call:
lm(formula = mortality ~ dose)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-0.10816 -0.06063  0.00263  0.05119  0.12818

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.9478     0.8717  -10.27 4.99e-05 ***
dose          5.3249     0.4857   10.96 3.42e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08669 on 6 degrees of freedom
Multiple R-squared:  0.9524,    Adjusted R-squared:  0.9445
F-statistic: 120.2 on 1 and 6 DF,  p-value: 3.422e-05
```

> **!** Important
>
> You should have met *R* output like this in previous statistics modules, but if you need some revision then see *Appendix-C: Background to Analysis of Variance* on Minerva under *Basic Pre-requisite Material*.

1.4 Types of variables

The way a variable enters a model will depends on its type. The most common five types of variable are:

1. Quantitative
    a. Continuous: for example, height; weight; duration. Real valued. Note that although recorded data is rounded it is still usually best regarded as continuous.
    b. Count (discrete): for example, number of children in a family; accidents at a road junction; number of items sold. Non-negative and integer-valued.
2. Qualitative
    a. Ordered categorical (ordinal): for example, severity of illness (Mild/ Moderate/Severe); degree classification (first/ upper-second/ lower-second/ third).
    b. Unordered categorical (nominal):
        - Dichotomous (binary): two categories: for example sex (M/ F); agreement (Yes/ No); coin toss (Head/ Tail).
        - Polytomous (also known as polychotomous): more than two categories: for example blood group (A/ B/ O); eye colour (Brown/ Blue/ Green).

Note that although dichotomous is clearly a special case of polytomous, making the distinction is usually worthwhile as it often leads to a simplified modelling and testing approach.

Focus on data type quiz

Test your knowledge recall and comprehension to reinforce ideas ready for later in the module

For each of the following situations what is the **most appropriate** data type: nominal, ordinal, discrete, or continuous?

1. The eye colour of 100 patients visiting the Yorkshire Cancer Research Centre, for example, grey, green, brown, blue….

- (A) nominal
- (B) ordinal
- (C) discrete
- (D) continuous

2. The nationality of students at the University of Leeds, for example, British, Chinese, Greek, Indian….

- (A) nominal
- (B) ordinal
- (C) discrete
- (D) continuous

3. The five-star ratings submitted by 50 customers on TripAdvisor for the Leeds Queens Hotel, for example 1 star, 2 star,… 5 star.

- (A) nominal
- (B) ordinal
- (C) discrete
- (D) continuous

4. The diastolic blood pressure of 20 male and 20 female patients attending a heart health clinic at the Leeds General Infirmary in a study to investigate differences between men and women, for example, 80 mm Hg, 130 mm Hg,…

- (A) nominal
- (B) ordinal
- (C) discrete
- (D) continuous

5. The daily stock market closing price of British Telecom shares on the London Stock Exchange over a year to study the change over time, for example, 114.75p, 115.10p,…

- (A) nominal
- (B) ordinal
- (C) discrete
- (D) continuous

6. The January monthly rainfall collected since 1961 at a weather monitoring station in the Pennines of Yorkshire, for example, 8 mm , 12 mm,…

- (A) nominal
- (B) ordinal
- (C) discrete
- (D) continuous

7. The level of satisfaction of 100 randomly chosen voters with the policies of a political party, for example, agree, fully agree, neither agree nor disagree, disagree, and fully disagree.

- (A) nominal
- (B) ordinal

- (C) discrete
- (D) continuous

8. The number of new people following the *TheRoyalFamily* twitter page per day over a year, for example 459, 700,… to study the change due to a royal wedding.

- (A) nominal
- (B) ordinal
- (C) discrete
- (D) continuous

9. The number of road accidents occurring per month, at a busy roundabout over a 10-year period to study the change over time, for example, 0, 1, 2,…

- (A) nominal
- (B) ordinal
- (C) discrete
- (D) continuous

10. The number of Scottish strawberries in 50 randomly selected boxes bought from ASDA supermarket, for example, 50, 58, 68,…

- (A) nominal
- (B) ordinal
- (C) discrete
- (D) continuous

Click here to see explanations

1. Eye colour is qualitative and can take any one of an unordered set of categories. Although the eye colours are categories, there is no clear ordering to the colours.
2. Nationality is qualitative and can take any one of an unordered set of categories. Although the nationality are categories, there is no clear ordering to the countries.

3. The rating is qualitative and can take any one of set of categories but the categories are clearly ordered, 5 star is better than 4 start etc. Although the ratings are represented by integers, there is no reason why the difference between 1 and 2 stars has the same interpretation as between 4 and 5 stars and hence it cannot be discrete.
4. Although the recorded values might take only integer values, blood pressure is a measurement and could take be any real number.
5. Share price is a measurement and could be any real number, even though in practice it will be rounded.
6. Rainfall is a measurement and although the recorded values might take only integer values, rainfall could be any real number.
7. Satisfaction score is qualitative and can take any one of set of categories but the categories are clearly ordered, fully agree is better than agree etc. Although the scores could be represented by numerical values, e.g. 1,2,3,4,5, there is no reason why the difference between 1 and 2 has the same interpretation as between 4 and 5 and hence it is not discrete.
8. The number of people is a quantitative count which is limited to the non-negative integers. The variable is discrete.
9. The number of accidents is a quantitative count, being limited to the non-negative integers and hence is discrete.
10. The number of strawberries is a quantitative a count which is limited to the non-negative integers. The variable is discrete.

1.5 Exercises

> **!** Important
>
> Unless otherwise stated, data files will be available online at: rgaykroyd.github.io/MATH3823/Datasets/*filename.ext*, where *filename.ext* is the stated filename with extension. For example, for the first exercise:
> beetle = read.table("https://rgaykroyd.github.io/MATH3823/Datasets/beetle.txt", header=T)

1.1 Consider again the beetle data in Table 1.1. Perform the calculations by hand and then check the answers using R – a copy of the data is available in the file beetle.txt. Finally plot the fitted regression line on a scatter plot of the data.

Click here to see hints.

See the code chunk used to produce Figure 1.1.

1.2 Consider the following synthetic data:

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$
x_i	-1	0	1	2	2.5	3	4	6
y_i	-2.8	-1.1	7.2	8.0	8.9	9.2	14.8	24.7

Plot the data to check that a linear model is suitable and then fit a linear regression model. Do you think that the fitted model can be reliably used to predict the values of y when $x = 5$ and $x = 10$? Justify your answers.

Click here to see hints.

Which is more reliable prediction for a value within the range of the data (interpolation) or outside the range of the data (extrapolation)?

1.3 Starting from Equation 1.1, derive the estimation equations given in Equation 1.2. Show that $\widehat{\alpha}$ and $\widehat{\beta}$ are unbiased estimators of α and β. What can be said about $\widehat{\sigma}^2$ as an estimator of σ^2?

Click here to see hints.

For unbiasedness of intercept and slope check your MATH1712 lecture notes. For the latter, there is a careful theoretical proof about the variance parameter, but here only an intuitive explanation is expected.

1.4 The *Brownlee's Stack Loss Plant Data*[2] is already available in **R**, with background details on the help page, `?stackloss`.

After plotting all pairs of variables, which of `Air.Flow`, `Water.Temp` and `Acid.Conc` do you think could be used to model `stack.loss` using a linear regression? Justify your answer.

Perform a simple linear regression with using `stack.loss` as the response variable and your chosen variable as the explanatory variable. Add the fitted regression line to a scatter plot of the data and comment.

Click here to see hints.

You already met this example in MATH1712 – check your lecture note for guidance.

1.5 In an experiment conducted by de Silva et al. in 2020[3] data was obtained to investigate falling objects and gravity, as first consider by Galileo and Newton. A copy of the data is available in the file physics_from_data.csv.

Suppose that we wish to develop a method to predict the maximum Reynolds number from a single explanatory variable. Which of the variables do you think helps explain Reynolds number the best? Why do you think this?

Click here to see hints.

Read the data into R and perform a simple linear regression of the maximum Reynolds number as the response variable and, in turn, each of the other variables as the explanatory variable. Plot the data, with and add the corresponding fitted linear models.

[2] Brownlee, K. A. (1960, 2nd ed. 1965) *Statistical Theory and Methodology in Science and Engineering.* New York: Wiley. pp. 491–500.

[3] de Silva BM, Higdon DM, Brunton SL, Kutz JN. Discovery of Physics From Data: Universal Laws and Discrepancies. Front Artif Intell. 2020 Apr 28;3:25. doi: 10.3389/frai.2020.00025. PMID: 33733144; PMCID: PMC7861345.

> **Note**
>
> Exercise 1 Solutions can be found here.

2 Essentials of Normal Linear Models

Here is a short video [3 mins] to introduce the chapter

2.1 Overview

In many fields of application, we might assume the response variable is normally distributed. For example: heights, weights, log prices, etc.

The data[1] in Table 2.1 record the birth weights of 12 girls and 12 boys and their gestational ages (time from conception to birth).

A key question is, can we predict the birth weight of a baby born at a given gestational age using these data. For this we will need to make assumptions about the relationship between birth weight and gestational age, and any associated natural variation – that is we require a model.

First we should explore the data. Figure 2.1a shows a histogram of the birth weights indicating a spread around modal group 2800-3000 grams; Figure 2.1b indicates slightly higher birth weights for the boys than the girls; and Figure 2.1c shows an increasing relationship between weight and age. Together, these suggest that gestational age and sex are likely to be important for predicting weight.

Before considering possible models, Figure 2.2 again shows the relationship between weight and age but this time with the points coloured according to the baby's sex. This, perhaps, shows the boys to have generally higher weights across the age range than girls.

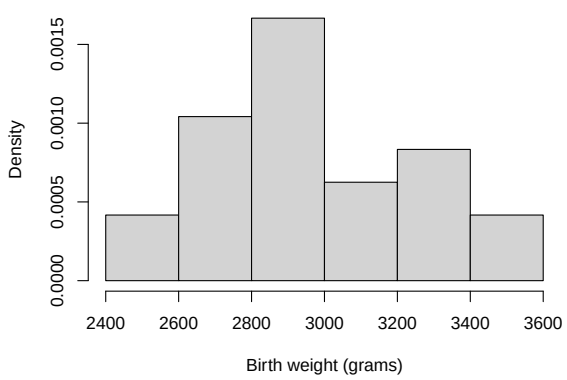
Focus on modelling quiz

Test your knowledge recall and application to reinforce basic ideas and prepare for similar concepts later in the module.

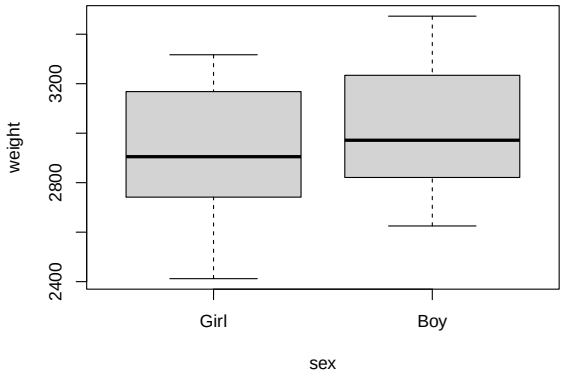
For each situation, choose one of the following statements which you think is **most likely** to apply.

1. What is the most useful graphical summary for identifying a potential relationship between two variables?

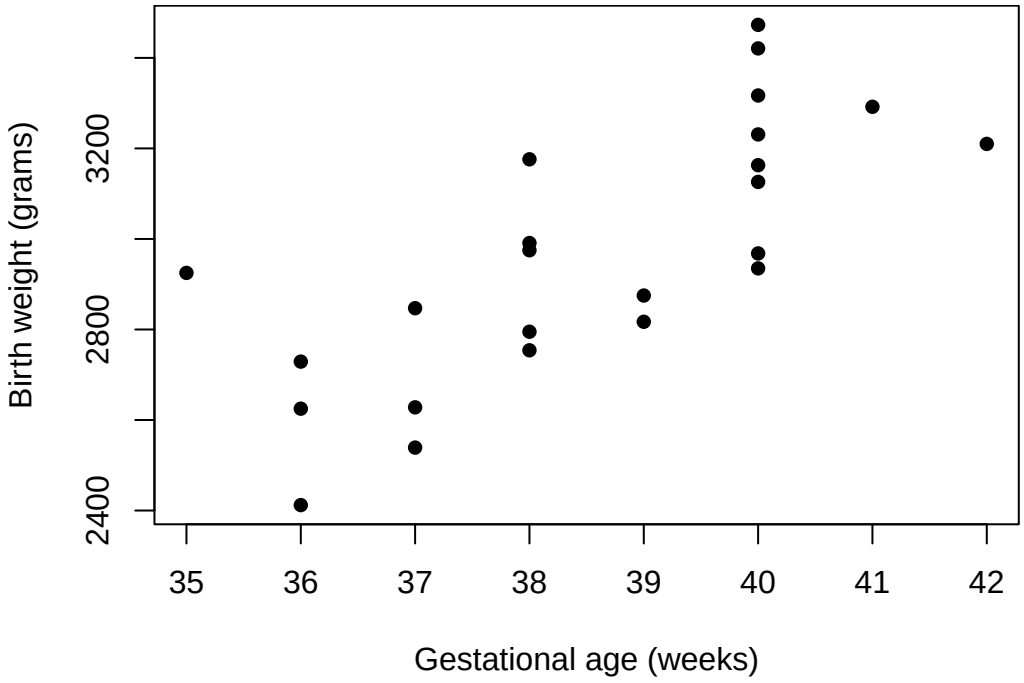
[1] Dobson and Barnett, 3rd edition, Table 2.3.



(a) Weight distribution



(b) Weight sub-divided by Sex



(c) Relationship beween variables

Figure 2.1: Birthweight and gestational age for 24 babies.

Table 2.1: Gestational ages (in weeks) and birth weights (in grams) for 24 babies (12 girls and 12 boys).

(a) Girls

Gestational Age	Birth weight
40	3317
36	2729
40	2935
37	2754
42	3210
39	2817
40	3126
37	2539
36	2412
38	2991
39	2875
40	3231

(b) Boys

Gestational Age	Birth weight
40	2968
38	2795
40	3163
35	2925
36	2625
37	2847
41	3292
40	3473
37	2628
38	3176
40	3421
38	3975

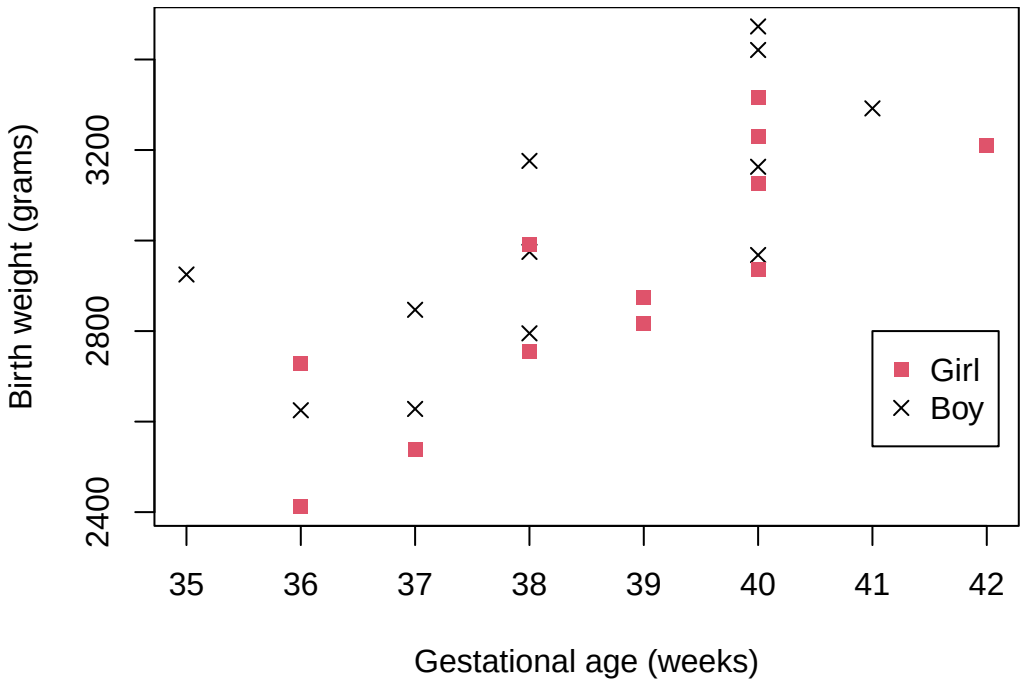


Figure 2.2: Birthweight and gestational age for 12 girls (red squares) and 12 boys (black crosses).

- (A) histogram
- (B) scatter plot
- (C) boxplot
- (D) kernel density plot
- (E) other

2. What is the most useful numerical summary for identifying a potential linear relationship between two variables?

- (A) sample means
- (B) skew
- (C) variances
- (D) correlation
- (E) other

3. Which of the following is a true statements about the correlation coefficient? (Choose any that apply.)

- (A) Can take any real number
- (B) Can be positive or negative
- (C) A value close to -1 means uncorrelated and close to +1 indicates a high correlation
- (D) Is always between -1 and +1
- (E) Is always a positive number

4. Which of the following is a true statements about regression? (Choose any that apply.)

- (A) All regression models describe linear relationships

- (B) A linear model can be fitted to any data set
- (C) After fitting a regression model we should always consider residuals
- (D) A linear regression can sometimes be used to approximate a non-linear relationship
- (E) A linear regression model will fit well if the correlation is close to zero

5. Which of the following is a true statements about statistical modelling? (Choose any that apply.)

- (A) Models only involve a statistical distribution
- (B) Only use a model that is 100% correct
- (C) Only use a linear model
- (D) All models are approximations
- (E) Models are part deterministic and part random

2.2 Linear models

Continuing the birth weight example. Of course, there are very many possible models, but here we will consider the following:

`Model 0` :	$\texttt{Weight} = \alpha$
`Model 1` :	$\texttt{Weight} = \alpha + \beta.\texttt{Age}$
`Model 2` :	$\texttt{Weight} = \alpha + \beta.\texttt{Age} + \gamma.\texttt{Sex}$
`Model 3` :	$\texttt{Weight} = \alpha + \beta.\texttt{Age} + \gamma.\texttt{Sex} + \delta.\texttt{Age.Sex}$

In these models, `Weight` is called the *response* variable (sometimes called the *dependent* variable) and `Age` and `Sex` are called the *covariates* or *explanatory* variables (sometimes called the *predictor* or *independent* variables). Here, `Age` is a continuous variable whereas `Sex` is coded as a *dummy* variable taking the value 0 for girls and 1 for boys; it is an example of a *factor*, in this case with just two *levels*: Girl and Boy.

Note that `Model 0` is a special case of `Model 1` (consider the situation when $\beta = 0$) and that `Model 1` is a special case of `Model 2` (consider the situation when $\gamma = 0$) and finally

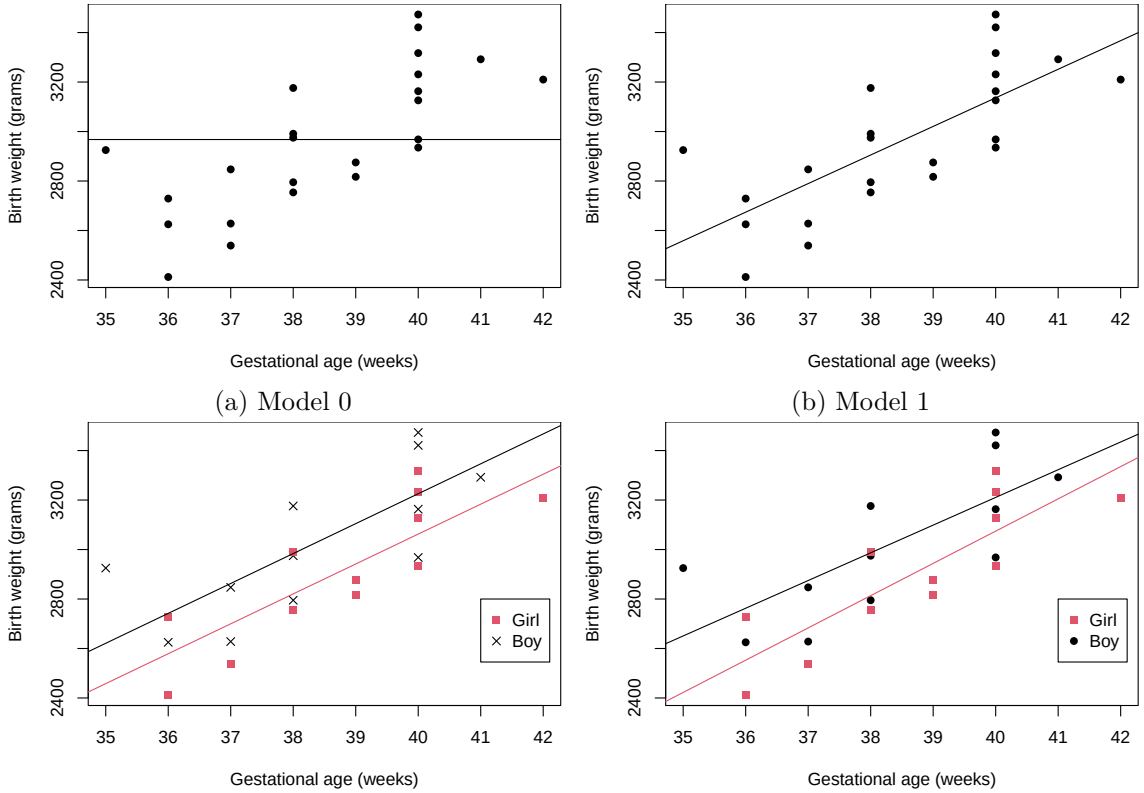
that `Model 2` is a special case of `Model 3` (consider the situation when $\delta = 0$) – such models are called *nested*.

In these models, α, β, γ and δ are *model parameters*. Parameter α is called the *intercept* term; β is called the *main effect* of `Age`; and is interpreted as the effect on birth weight *per week* of gestational age. Similarly, γ is the main effect of `Sex`, interpreted as the effect on birth weight of being a boy (because girl is the *baseline* category).

Parameter δ is called the *interaction effect* between `Age` and `Sex`. **Take care when interpreting an interaction effect.** Here, it does not mean that age somehow affects sex, or vice-versa. It means that the effect of gestational age on birth weight depends on whether the baby is a boy or a girl.

These models can be fitted to the data using (Ordinary) *Least Squares* to produce the results presented in Figure 2.3.

Which model should we use?



(a) Model 0

(b) Model 1

(c) Model 2

(d) Model 3

Figure 2.3: Birthweight and gestational age data with superimposed fitted regression lines from various competing models.

We know from previous modules that statistical tests can be used to check the importance of regression coefficients and model parameters, but it is also important to use the graphical results, as in Figure 2.3, to guide us.

`Model 0` says that there is no change in birth weight with gestational age which means that we would use the average birth weight as the prediction whatever the gestational age – this makes no sense. As we can easily see from the scatter plot of the data, the fitted line in this case is clearly inappropriate.

`Model 1` does not take into account whether the baby is a girl or a boy, but does model the relationship between birth weight and gestational age. This does seem to provide a good fit and might be adequate for many purposes. Recall from Figure 2.1b and Figure 2.2, however, that for a given gestational age the boys seem to have a higher birth weight than the girls.

`Model 2` does take the sex of the baby into account by allowing separate intercepts in the fitted lines – this means that the lines are parallel. By eye, there is a clear difference between these two lines but it might not be important.

`Model 3` allows for separate slopes as well as intercepts. There is a slight difference in the slopes, with the birth weight of the girls gradually catching-up as the gestational age increases. It is difficult to see, however, if this will be a general pattern or if it is only true for this data set – especially given the relatively small sample size.

Here, it is not clear by eye which of the fitted models will be the best and hence we should use a statistical test to help. In particular, we can choose between the models using F-tests.

Let y_i denote the value of the dependent variable `Weight` for individual $i = 1, \ldots, n$, and let the four models be indexed by $k = 0, 1, 2, 3$.

Let R_k denote the *residual sum of squares* (RSS) for Model k :

$$R_k = \sum_{i=1}^{n} (y_i - \hat{\mu}_{ki})^2, \tag{2.1}$$

where $\hat{\mu}_{ki}$ is the fitted value for individual i under `Model` k. Let r_k denote the corresponding *residual degrees of freedom* for `Model` k (the number of observations minus the number of model parameters).

Consider the following hypotheses:

$$H_0 : \texttt{Model } 0 \text{ is true}; \qquad H_1 : \texttt{Model } 1 \text{ is true}.$$

Under the null hypothesis H_0, the difference between R_0 and R_1 will be purely random, so the between-models mean-square $(R_0 - R_1)/(r_0 - r_1)$ should be comparable to the residual mean-square R_1/r_1. Thus our test statistic for comparing `Model` 1 to the simpler `Model` 0 is:

$$F_{01} = \frac{(R_0 - R_1)/(r_0 - r_1)}{R_1/r_1}. \tag{2.2}$$

It can be shown that, under the null hypothesis H_0, the statistic F_{01} will have an F-distribution on $r_0 - r_1$ and r_1 degrees of freedom, which we write: $F_{r_0 - r_1, r_1}$. Under the

alternative hypothesis H_1, the difference $R_0 - R_1$ will tend to be larger than expected under H_0, and so the observed value F_{01} will probably lie in the upper tail of the $F_{r_0-r_1,r_1}$ distribution.

Returning to the birth weight data, we obtain the following output from **R** when we fit `Model 1`:

```
(Intercept)         age
 -1484.9846    115.5283

Analysis of Variance Table

Response: weight
          Df  Sum Sq Mean Sq F value   Pr(>F)
age        1 1013799 1013799   27.33 3.04e-05 ***
Residuals 22  816074   37094
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus we have parameter estimates: $\hat{\alpha} = -1484.98$ and $\hat{\beta} = 115.5$. The Analysis of Variance (ANOVA) table gives: $R_0 - R_1 = 1013799$ with $r_0 - r_1 = 1$ and $R_1 = 816074$ with $r_1 = 22$.

If we wanted R_0 and r_0 then we can either fit `Model 0` or get them *by subtraction.*

The F_{01} statistic, Equation 2.2, is then

$$F_{01} = \frac{103799/1}{816074/22} = 27.33,$$

which can be read directly from the ANOVA table in the column headed 'F value'.

Is $F_{01} = 27.33$ in the upper tail of the $F_{1,22}$ distribution? (See Figure 2.4 and note that 27.33 is very far to the right.) The final column of the ANOVA table tells us that the probability of observing $F_{01} > 27.33$ is only 3.04×10^5 – this is called a p-value. The *** beside this p-value highlights that its value lies between 0 and 0.001. This indicates that we should reject H_0 in favour of H_1 – there is very strong evidence for the more complicated model. Thus we would conclude that the effect of gestational age is statistically significant in these data.

Next, consider the following hypotheses:

$$H_0 : \texttt{Model 1 is true}; \qquad H_1 : \texttt{Model 2 is true}.$$

Under the null hypothesis H_0, the difference between R_1 and R_2 will be purely random, so the between-models mean-square $(R_1 - R_2)/(r_1 - r_2)$ should be comparable to the residual mean-square R_2/r_2. Thus our test statistic for comparing `Model 2` to the simpler `Model 1` is:

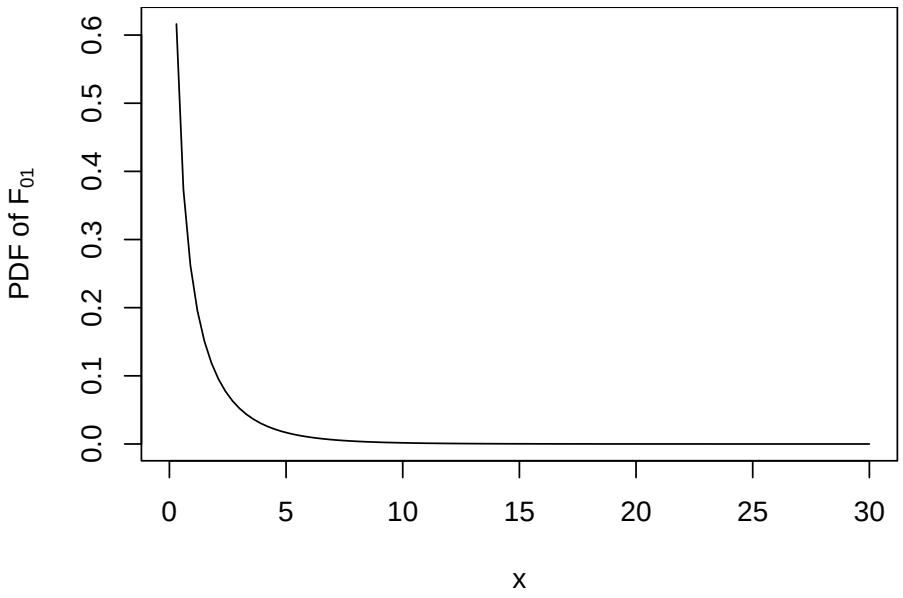


Figure 2.4: Probability density function of F_{01} distribution.

$$F_{12} = \frac{(R_1 - R_2)/(r_1 - r_2)}{R_2/r_2}. \tag{2.3}$$

Under the null hypothesis H_0, the statistic F_{12} will have an F-distribution on $r_1 - r_2$ and r_2 degrees of freedom, which we write: $F_{r_1-r_2,r_2}$. Under the alternative hypothesis H_1, the difference $R_1 - R_2$ will tend to be larger than expected under H_0, and so the observed value F_{12} will probably lie in the upper tail of the $F_{r_1-r_2,r_2}$ distribution.

Returning to the birth weight data, we obtain the following output from **R** (where `sexM` denotes Boy):

```
(Intercept)         age        sexM
 -1773.3218    120.8943    163.0393

Analysis of Variance Table

Response: weight
          Df  Sum Sq Mean Sq F value    Pr(>F)
age        1 1013799 1013799 32.3174 1.213e-05 ***
sex        1  157304  157304  5.0145   0.03609 *
Residuals 21  658771   31370
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus we have parameter estimates: $\widehat{\alpha} = -1773.3$, $\widehat{\beta} = 120.9$ and $\widehat{\gamma} = 163.0$, the latter being the effect of being a boy compared to the baseline category of being a girl.

The Analysis of Variance (ANOVA) table gives: $R_1 - R_2 = 157304$ with $r_1 - r_2 = 1$, and $R_2 = 658771$ with $r_2 = 21$. The F_{12} statistic, Equation 2.3, is then

$$F_{12} = \frac{157304/1}{658771/21} = 5.0145,$$

which can be read directly from the ANOVA table in the column headed 'F value'. Is $F_{12} = 5.01$ in the upper tail of the $F_{1,21}$ distribution?

The final column of the ANOVA table tells us that the probability of observing $F_{12} > 5.01$ is only 0.03609 – this is called a p-value. The * beside this p-value highlights that its value lies between 0.01 and 0.05. This indicates that we should reject H_0 in favour of H_1 – there is evidence for the more complicated model. Thus we would conclude that the effect of the sex of the baby, after controlling for gestational age, is statistically significant in these data.

To complete the analysis, we should now compare `Model 2` with `Model 3` – see Exercises.

Focus on regression quiz

Test your knowledge recall and application to reinforce basic ideas and prepare for similar concepts later in the module.

For each situation, choose one of the following statements which you think is **most likely** to apply.

1. Which of the following is a true statement about correlation and linear regression? (Choose any that apply.)

- (A) The correlation and the regression slope parameter will have the same sign
- (B) The correlation and the slope parameter will have the opposite sign
- (C) There is no relationship between correlation and linear regression
- (D) If the correlation is close to zero then a linear regression will not be useful
- (E) If the correlation is away from zero then a linear regression should always be used

2. Which of the following is NOT a true statement about model residuals? (Choose any that apply.)

- (A) Can help to identify unusual data points

- (B) Will always sum to zero
- (C) All should be zero if the model is a good fit
- (D) Can be used to judge how well the model fits the data
- (E) Should always be plotted as part of a data analysis

3. Which of the following is a true statement about prediction using linear regression? (Choose any that apply.)

- (A) We should be careful about extrapolating
- (B) Interpolation and extrapolation are both types of prediction
- (C) Interpolation is performed using the fitted model
- (D) Interpolation should only be used if we know the true relationship is linear
- (E) Prediction is always reliable

4. Which of the following is NOT an important part of regression model fitting? (Choose any that apply.)

- (A) The command **lm** to fit a linear model
- (B) A scatter plot of the residuals
- (C) The command **cor** to check for a relationship between variables
- (D) A histogram of the data
- (E) A scatter plot of the data

5. Which of the following is NOT important when performing data analysis? (Choose any that apply.)

- (A) The analysis should be done professionally. We must not involve our personal views to influence our analysis and conclusions

- (B) Do we need authorization from the person collecting the data before we can use it
- (C) Background information is not needed before analyzing a new data set – it's only the numbers that matter
- (D) Are the data representative of what we are aiming to investigate
- (E) Is the data reliable and have suitable data checks been performed

2.3 Types of normal linear model

Here we consider how normal linear models can be set up for different types of explanatory variable. The dependent variable y is modelled as a linear combination of p explanatory variables $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ plus a random error $\epsilon \sim N(0, \sigma^2)$, where '$\sim$' means 'is distributed as'. Several models are of this kind, depending on the number and type of explanatory variables. Table 2.3 lists some types of normal linear models with their explanatory variable types.

Table 2.3: Types of normal linear model and their explanatory variable types where indicator function $I(x = j) = 1$ if $x = j$ and 0 otherwise.

p	**Explanatory variables**	**Model**
1	Quantitative	Simple linear regression $y = \alpha + \beta x + \epsilon$
>1	Quantitative	Multiple linear regression$y = \alpha + \sum_{i=1}^{p} \beta_i x_i + \epsilon$
1	Dichotomous ($x = 1$ or 2)	Two-sample t-test $y = \alpha + \delta\ I(x = 2) + \epsilon$
1	Polytomous, k levels ($x = 1, \ldots, k$)	One-way ANOVA$y = \alpha + \sum_{j=1}^{k} \delta_j\ I(x = j) + \epsilon$
>1	Qualitative	p-way ANOVA

For the two-sample t-test model[2], observations in the two groups have means $\alpha + \beta_1$ and $\alpha + \beta_2$. Notice, however, that we have three parameters with only two group sample means and hence parameter estimation is not possible. To avoid this identification problem, we either impose a 'corner' constraint: $\beta_1 = 0$ and then β_2 represents the difference in the Group 2 mean relative to a baseline of Group 1. Alternatively, we may impose a 'sum-to-zero' constraint: $\beta_1 + \beta_2 = 0$, the values $\beta_1 = -\beta_2$ then give differences in the groups means relative to the overall mean. Table 2.4 shows the effect of the parameter constraint on the group means.

[2]Notice that this is a special case of the one-way ANOVA when there are only two-groups.

Table 2.4: Parameters in the two-sample t-test model after imposing parameter constraint to avoid the identification problem.

Constraint	Group 1 mean	Group 2 mean
$\beta_1 = 0$	α	$\alpha + \beta_2$
$\beta_1 + \beta_2 = 0$	$\alpha - \beta_2$	$\alpha + \beta_2$

For the general one-way ANOVA model with k groups, observations in Group j have mean $\alpha + \delta_j$, for $j = 1, \ldots, k$ – that leads to $k+1$ parameters describing k group means. Again we can impose the 'corner' constraint: $\delta_1 = 0$ and then δ_j represents the difference in means between Group j and the baseline Group 1. Alternatively, we may impose a 'sum-to-zero' constraint:$\sum_{j=1}^{k} \delta_j = 0$ and again $(\delta_1, \delta_2, \ldots, \delta_k)$ then represents an individual group effect relative to the overall data mean.

2.4 Matrix representation of linear models

All of the models in Table Table 2.3 can be fitted by least squares (OLS). To describe this, a matrix formulation will be most convenient:

$$\mathbf{Y} = X\beta + \epsilon \tag{2.4}$$

where

- $\mathbf{Y}$ is an $n \times 1$ vector of observed response values with n being the number of observations.
- X is an $n \times p$ *design* matrix, to be discussed below.
- β is a $p \times 1$ vector of parameters or coefficients to be estimated.
- ϵ is an $n \times 1$ vector of independent and identically distributed (IID) random variables, which here $\epsilon \sim N(0, \sigma^2)$ and is called the "error" term.

Creating the design matrix is a key part of the modelling as it describes the important structure of investigation or experiment. The design matrix can be constructed by the following process.

1. Begin with an X containing only one column: a vector of ones for the overall mean or intercept term (the α in Table 2.3).
2. For each explanatory variable x_j, do the following:
    a. If a variable x_j is quantitative, add a column to X containing the values of x_j.
    b. If x_j is qualitative with k levels, add k "dummy" columns to X, taking values 0 and 1, where a 1 in the ℓth dummy column identifies that the corresponding observation is at level ℓ of factor x_j . For example, suppose we have a factor $\mathbf{x}_j = (M, M, F, M, F)$ representing the sex of $n = 5$ individuals. This information can be coded into two dummy columns of X:

$$\begin{matrix} F & M \end{matrix}$$
$$\begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

3. When qualitative variables are present, X will be singular – that is, there will be linear dependencies between the columns of X. For example, the sum of the two columns above is a vector of ones, the same as the intercept column. We resolve this identification problem by deleting some columns of X. This is equivalent to applying the corner constraint $\delta_1 = 0$ in the one-way ANOVA.

   In the above example, after removing a column, we get:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

4. Each column of X represents either a quantitative variable, or a level of a qualitative variable. We will use $i = 1, \ldots, n$ to label the observations (rows of X) and $j = 1, \ldots, p$ to label the columns of X.

Example: Simple linear regression

Consider the simple linear regression model $y = \alpha + \beta x + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$. Given data on n pairs $(x_i, y_i), i = 1, \ldots, n$, we write this as

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{2.5}$$

where the ϵ_i are IID $N(0, \sigma^2)$. In matrix form, this becomes

$$\mathbf{Y} = X\beta + \epsilon \tag{2.6}$$

with

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

The ith row of Equation 2.6 has the same meaning as Equation 2.5:

$$y_i = 1 \times \beta_1 + x_i \times \beta_2 + \epsilon_i = \alpha + \beta x_i + \epsilon_i, \quad \text{for } i = 1, 2, \ldots, n.$$

Example: One-way ANOVA

For one-way ANOVA with k levels, the model is

$$y_i = \alpha + \sum_{j=1}^{k} \delta_j \, I(x_i = j) + \epsilon_i, \quad \text{for } i = 1, 2, \ldots, n,$$

where x_i denotes the group level of individual i. So if y_i is from the jth group then $y_i \sim N(\alpha + \delta_j, \sigma^2)$. Here α is the intercept and the $(\delta_1, \delta_2, \ldots, \delta_k)$ represent the "main effects".

We can store the information about the levels of g in a dummy matrix $X^* = (x^*_{ij})$ where

$$x^*_{ij} = \begin{cases} 1, & g_i = j, \\ 0, & \text{otherwise.} \end{cases}$$

Then set $X = [1, X^*]$, where 1 is an n-vector of 1's. For the male–female example at (1.12), we have $n = 5$ and a sex factor:

$$g = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 1 \\ 2 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \alpha \\ \delta_1 \\ \delta_2 \end{bmatrix}.$$

Then the ith row of X becomes $\beta_1 + \beta_2 = \alpha + \delta_1$ if $g_i = 1$ and $\beta_1 + \beta_3 = \alpha + \delta_2$ if $g_i = 2$. That is, the ith row of X is

$$\alpha + \sum_{j=1}^{2} \delta_j I(g_i = j)$$

so this model can be written $Y = X\beta + \epsilon$. Here, X is singular: its last two columns added together equal its first column. Statistically, the problem is that we are trying to estimate two means (the mean response for Boys and the mean response for Girls) with three parameters (α, δ_2 and δ_2).

In practice, we often resolve this aliasing or identification problem by setting one of the parameters to be zero, that is $\delta_1 = 0$, which corresponds to deleting the second column of X).

Focus on matrix representations quiz

For each situation, choose one of the following statements which you think is **most likely** to apply.

1. What is the dimension of the design matrix?

- (A) n by p
- (B) 2
- (C) X
- (D) n by 1
- (E) p by 1

2. What is the distribution of the error term?

- (A) random variable
- (B) normal
- (C) IDD
- (D) n by 1
- (E) mean zero

3. Which quantity represents the model parameters?

- (A) epsilon
- (B) beta
- (C) X
- (D) Y
- (E) n

4. Which two terms in the model have the same dimensions?

- (A) Y and X
- (B) X and beta
- (C) Y and epsilon
- (D) Y and beta
- (E) beta and epsilon

5. Which of the following is a potential problem when using qualitative variables?

- (A) X is not square
- (B) X is singular
- (C) X is full rank
- (D) X can be zero
- (E) X can be negative

2.5 Model shorthand notation

In R, a qualitative (categorical) variable is called a *factor*, and its categories are called *levels*. For example, variable `Sex` in the birth weight data (above) has levels coded "M" for 'Boy' and "F" for 'Girl'. It may not be obvious to **R** whether a variable is quantitative or qualitative. For example, a qualitative variable called `Grade` might have categories 1, 2 and 3. If `grade` was included in a model, R would treat it as quantitative unless we declare it to be a factor, which we can do with the command:

```
grade = as.factor(grade)
```

A convenient model-specification notation has been developed from which the design matrix X can be constructed. Below, $E, F, ...$ denote generic quantitative (continuous) or qualitative (categorical) variables. Terms in this notation may take the following forms:

a. 1 : a column of 1's to accommodate an intercept term (the α's of Table 2.3). This is included in the model by default.

b. E : variable E is included in the model. The design matrix includes k_E columns for E. If E is quantitative, $k_E = 1$. If E is qualitative, k_E is the number of levels of E minus 1.

c. $E + F$: both E and F are included the model. The design matrix includes $k_E + k_F$ columns accordingly.

d. $E : F$ (sometimes $E \cdot F$) : the model includes an interaction between E and F; each column that would be included for E is multiplied by each column for F in turn. The design matrix includes $k_E \times k_F$ columns accordingly.

e. $E * F$: shorthand for $1 + E + F + E : F$: useful for crossed models where E and F are different factors. For example, E labels age groups; F labels medical conditions.

f. E/F : shorthand for $1+E+E : F$: useful for nested models where F is a factor whose levels have meaning only within levels of factor E. For example, E labels different hospitals; F labels wards within hospitals.

g. poly$(E; \ell)$: shorthand for an orthogonal polynomial, wherein x contains a set of mutually orthogonal columns containing polynomials in E of increasing order, from order 1 through order ℓ.

h. $-E$: shorthand for removing a term from the model; for example $E * F - E$ is short for $1 + F + E : F$.

i. $I()$: shorthand for an arithmetical expression (not to be confused with the indicator function defined above). For example, $I(E + F)$ denotes a new quantitative variable constructed by adding together quantitative variables E and F. This would cause an error if either E or F has been declared as a factor. What would happen in this example if we omitted the $I(\cdot)$ notation?

The notation uses "~" as shorthand for "is modelled by" or "is regressed on". For example,

- Weight is regressed on age-group and sex with no interaction between them:

  $$\texttt{Weight} \sim \texttt{Age} + \texttt{Sex}$$

  as for the birthweight data in Figure 1.2c.

- Well being is regressed on age-group and income-group, where income is thought to affect wellbeing differentially by age:

  $$\texttt{Wellbeing} \sim \texttt{Age} * \texttt{Income}$$

- Class of degree is regressed on school of the university and on degree subject within the school:

  $$\texttt{DegreeClass} \sim \texttt{School/Subject}$$

- Yield of wheat is regressed on seed-variety and annual rainfall:

  $$\texttt{Yield} \sim \texttt{Variety} + \texttt{poly}(\texttt{Rainfall}, 2)$$

- Profit is regressed on amount invested:

$$\texttt{Profit} \sim \texttt{Investment} - 1$$

  (no intercept term, that is a regression through the origin).

Focus on model notation quiz

For each situation, choose one of the following statements which you think is **most likely** to apply.

1. What **R** command can be used to covert numerical values into a nominal variable?

- (A) as.factor
- (B) as.nominal
- (C) factorial
- (D) as.numerical
- (E) as.ordinal

2. Which of the following defines a model where variable Y is regressed on variables V1 and V2, but without an interaction?

- (A) Y ~ V1 + V2
- (B) Y ~ V1*V2
- (C) Y ~ V1:V2
- (D) Y ~ V1 + V2 + V1:V2
- (E) V1 + V2 ~ Y

3. Which of the following defines a model where variable Y is regressed on variables V1 and V2, including a constant and an interaction?

- (A) Y ~ V1 + V2 + V1:V2 -1

- (B) Y ~ 1+V1:V2
- (C) Y ~ V1*V2
- (D) Y ~ 1+V1/V2
- (E) Y ~ 1 + V1 + V2

4. Which of the following defines a model regression Y on the product of V1 and V2?

- (A) Y ~ V1.V2
- (B) Y ~ poly(V1,V2)
- (C) Y ~ I(V1*V2)
- (D) Y ~ I(V1 + V2)
- (E) Y ~ V1:V2

5. Which of the following defines a model where variable Y is regressed on a second-order polynomial in V1?

- (A) Y ~ I(V1)
- (B) Y ~ V1*V1
- (C) Y ~ 1+V1
- (D) Y ~ V1^2
- (E) Y ~ poly(V1,2)

2.6 Fitting linear models in R

A commonly used command for fitting a linear model in **R** is

`lm(formula)`.

Let `x,y,z,a,b`... represent **R** vectors, all of the same length n (perhaps read in from a data file using the `read.table` and `attach` commands).
If `a,b` are qualitative variables, then they first need to be declared as factors by `a = as.factor(a)`, etc.

The `formula` argument of `lm` specifies the required model in compact notation, e.g. `y` $\sim$ `x+z` or `y` $\sim$ `x + z*a` where $\sim$, `+`, `*` have the same meaning as in Section 2.5.

To extract information about a fitted linear model, it is best to store the result of `lm` as a variable and then to use the following functions:

- To fit a linear model and store the result in `my.lm` (for example):
  `my.lm = lm(y` $\sim$ `x + a*b)`
- To print various pieces of information including deviance residuals, parameter estimates and standard errors, deviances, and (if specified) correlations of parameter estimates:
  `summary(my.lm, correlation=T)`
- To print the anova table of the fitted model:
  `anova(my.lm)`
- To print the residual degrees of freedom of the fitted model:
  `df.residual(my.lm)`
- To print the vector of fitted values under the fitted model:
  `fitted.values(my.lm)`
- To print the residuals from the fitted model:
  `residuals(my.lm)`

- To print the parameter estimates from the fitted model:
  `coefficients(my.lm)`
- To print the design matrix for a specified model formula: `model.matrix(y` $\sim$ `a*b)`

The functions `summary`, `anova`, and possibly `model.matrix` are the most useful for printing out information about the fitted model. The results of the other functions can be saved as variables for further computation. In particular, you should look for an ansence of structure in a plot the residuals and you may need to make predictions of the response variable for new values of the explanatory variable.

Example of fitting a linear model in R

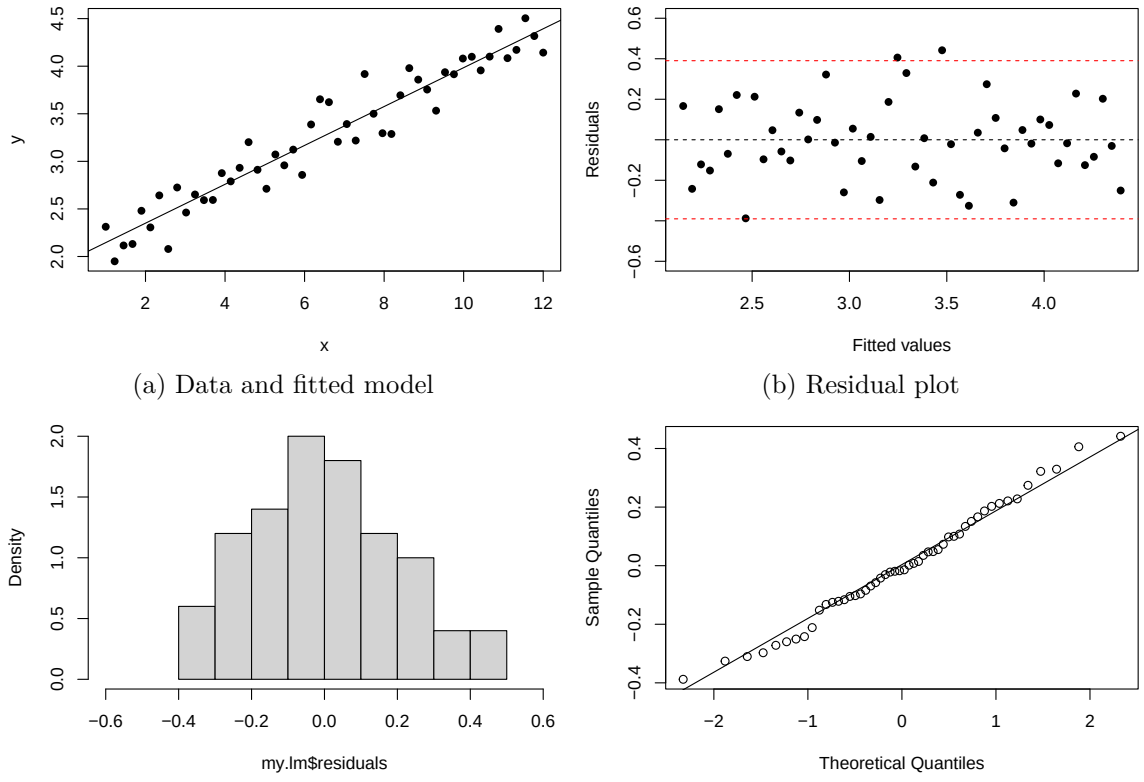
Here is a toy example of **R** commands for modelling a response in terms of one quantitative explanatory variable.

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38790 -0.12026 -0.01578  0.12751  0.44184

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.941552   0.062524   31.05   <2e-16 ***
x           0.204299   0.008609   23.73   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1972 on 48 degrees of freedom
Multiple R-squared:  0.9215,    Adjusted R-squared:  0.9198
F-statistic: 563.1 on 1 and 48 DF,  p-value: < 2.2e-16
```



(a) Data and fitted model

(b) Residual plot

(c) Histogram of residuals

(d) Normal distribution QQ plot

Figure 2.5: Illustration of model fitting on a toy example

A scatter plot of the data indicates that a linear model would be appropriate and the result-

ing fitted linear regression describes the data well. A residual plot shows that all residuals are within two standard deviations of zero, with the histogram showing a symmetric distribution and a QQ plot supporting normality of the residuals. All these diagnostic plots support that the model fits well.

Finally, we may wish to predict values of the response variable at new values of the response variable, for example at $x = 4$ and $x = 6$ – notice that the values of the explanatory variable must be converted into a **R** `data.frame`:

```
       1        2
2.758747 3.167345
```

2.7 Ethics in statistics and data science

A brief introduction to ethics and their relevance to statistics and data science.

In previous module you have already seen that professional and ethical consideration are important. Whether that be when we choose which graph to present or what action to take when data is missing or how to deal with suspected outliers. It is important for statisticians and data scientists to be aware of the ethical dimensions of their work and to be able to think these through.

Please note that the following was produce with the help of Dr Robbie Morgan from the Interdisciplinary Applied Ethics (IDEA) Centre at the University of Leeds.

Roughly speaking, ethics concerns what we ought to do (should do) and the kind of person that we ought to be. Of course, we're particularly interested in the kinds of ethical questions and challenges that you will encounter in your work as data scientists, such as:

- How should we collect data in a way that respects participants?
- Why is privacy important? How should data scientists protect privacy?
- How can algorithmic bias wrong members of the public? How should data scientists respond to this?
- To what extent are data scientists responsible for the impact of their work?
- When and how should data scientists challenge authority in the workplace?

The work that you do as data scientists can be hugely beneficial; it develops solutions for pressing real-world problems, enables organisations to carry out important functions more effectively and efficiently, and can improve the well being of the public by enacting innovation and technological advancement. At the same time, work in data science presents risk of significant harm and injustice. Widespread data collection threatens the privacy and safety of data subjects. Facial recognition and other surveillance technologies are the latest site of long-running debates over the values of privacy and freedom versus security. Biased algorithms can inflict injustices and exacerbate entrenched inequality. Automation can cause unemployment and instability, with unpredictable results as it affects ever more areas of our lives. So, it is very important to be clear about the obligations and responsibilities

that data scientists have to their employers, colleagues, and, most importantly, to society as a whole.

Please keep these issues in mind throughout any data analysis and modelling work, especially if you follow such a career path after graduation.

2.8 Exercises

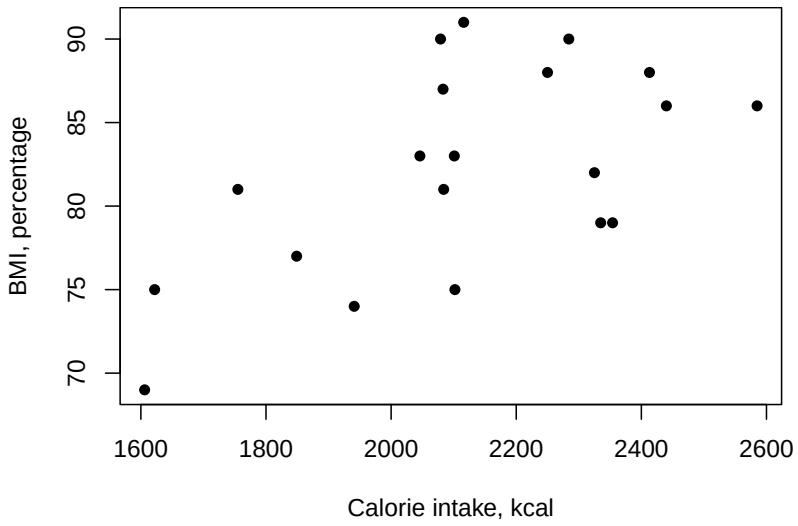
> **!** Important
>
> Unless otherwise stated, data files will be available online at: rgaykroyd.github.io/MATH3823/Datasets/*filename.ext*, where *filename.ext* is the stated filename with extension.

2.1. For each situation, consider the data description, correlation value, and data visualization. Then, identify whether the variables are related and if a linear model would be suitable.

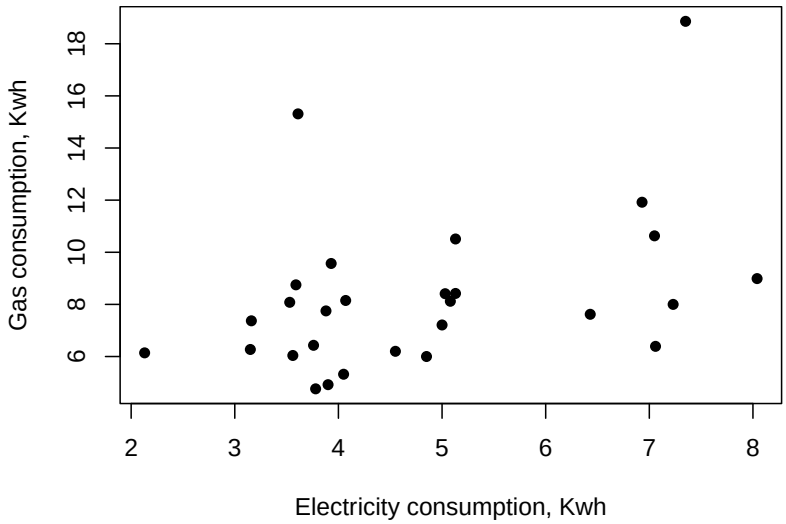
Click here to see hints.

For each description, think about what you expect to see and then confirm this, or otherwise, with the scatter plot. Does a linear relationship seem appropriate? Also, does variation in the scatter plot and correlation value suggest a strong relationship.

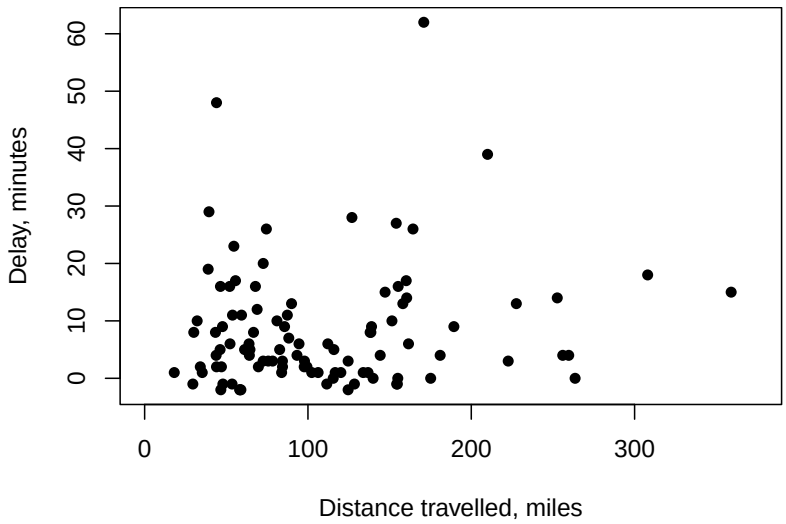
a. Childhood obesity is a serious medical condition which can lead to long-term health problems. A mixed-sex secondary school for ages 11-18 measures height and weight of all its new students (mostly aged 11 years old) on the first day of term and calculates their body mass index (BMI) and the average daily calorie intake was recorded. The correlation between calorie intake and BMI was 0.6 with the data shown in the scatter plot.



b. Domestic UK "smart meters" aim to record cumulative electricity and gas usage once per day. They use a wireless internet connection to transmit information to the energy providers, but they do not save previous values. The central system records readings as they are received and calculates the difference between successive readings. The scatter plot shows the daily electricity and gas consumption of a typical house collected using an automatic smart meter. The correlation between electricity and gas consumption is 0.4.

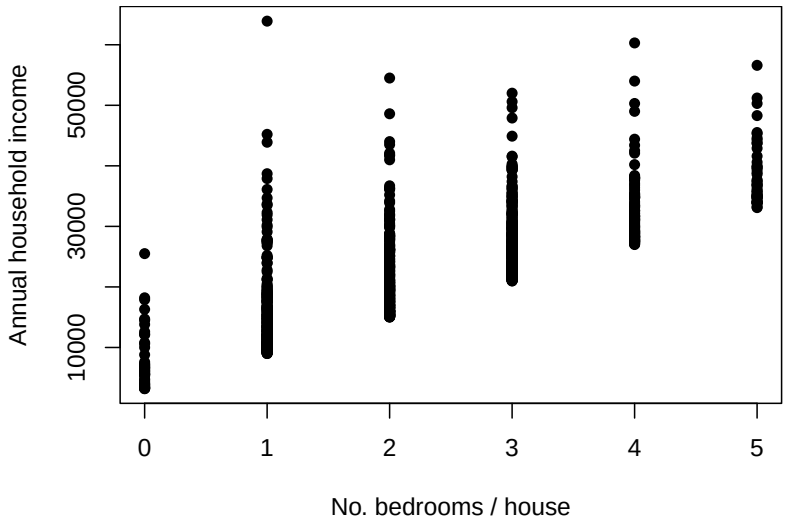


c. To investigate the punctuality of trains at Leeds City Station, a platform attendant records the number of minutes late, or early, that each train arrives at a particular platform relative to the timetable. The journey distance of each train arriving was later determined and recorded. The scatter plot shows the Distance Traveled, in miles, and the Delay, in minutes with a correlation of 0.1.



d. The 2022-23 UK Housing Survey, included questions regarding household income and the number of rooms. The data are shown in a scatter plot with a correlation of 0.7

between the variables.



2.2. An extra model which could have been considered for the Birth weight data example would be one that says that `Weight` is different for girls and boys, but does not depend on gestational age. Investigate this model.

Click here to see hints.

Write down the equation corresponding to this model. Then, load the birth weight data into RStudio and fit the model. How are the fitted model parameters related to the overall birth weight mean and the mean birth weights of the girls and boys? Is this a good fit to the data? Is Sex statistically significant?

2.3. For each given situation, consider the description and then investigate the suitability of a linear model.

Click here to see hints.

For each given data set, produce an appropriate graph within RStudio, fit a linear regression model and add the fitted model to the graph. Comment on the quality of fit.

a. Continuing the childhood obesity example, use the data in file *schoolstudy.csv* to model the relationship between BMI and calorie intake in 11-year old children.

b. To study the profitability of several iron ore (hematite) extraction quarries, small samples are taken from lorries arriving at an iron purification site. The lorries are open-topped with most travelling less than 20 Km but one quarry is more than 100 Km away. A chemical analysis provides a percentage of pure iron in the sample. Quarries with iron content less than 30% are not considered economically viable. The data file *iron.csv* contains measurements of percentage pure iron arriving at an iron purification site recorded over a 50 year period. Model the relationship between iron purity and time.

c. A study aims to investigate osteoporosis in women before and after menopause. The X-rays of a randomly selected sample of patients taking routine mammograms are analysed. The age of the patient and their menopause status are recorded, along with a measure of bone density calculated from the X-ray. Use the data in the file *bmd.csv* to model the relationship between age and Tscore, noting that a value of below -2.5 indicates osteoperosis, between -2.5 and -1.0 indicates osteopenia whereas above -1.0 is normal.

d. A primary school head teaching wishes to investigate the relationship between social skills of children and the ages of their brothers and sisters. The hypothesis is that those with older siblings will be better able to deal with social interaction. The file *siblings.csv* contains data on the age of the eldest sibling of a class of 6-year old school children along with a *social skills* score for each child assessed during the school lunch break. Model the relationship between sibling age and social skills.

2.4. In an experiment to investigate Ohm's Law, $V = IR$ where V is Voltage, I is current and R is resistance of the material, the following data[3] were recorded:

Table 2.5: Experimental verification of Ohm's Law

Voltage (Volts)	4	8	10	12	14	18	20	24
Current (mAmps)	11	24	30	36	40	53	58.5	70

Does this data support Ohm's Law? What is the resistance of the material used?

Click here to see hints.

There is no data file prepared for this and so create your own variables, then perform a linear regression. Comment on the quality of fit. Note that Ohm's Law is a linear function but without intercept and that the resistance is a constant multiplying the current.

2.5 In an investigation[4] into the effect of eating on pulse rate, 6 men and 6 women were tested before and after a meal, with the following results:

Table 2.6: At rest pulse rate before and after a meal for men and women

Men	before	105	79	79	103	87	97
	after	109	87	86	109	100	101
Women	before	74	73	82	78	86	77
	after	82	80	90	90	93	81

Suggest a suitable model for this situation and write down the corresponding design matrix. Calculate the parameter estimates using the matrix regression estimation equation.

[3] Aykroyd, P.J. (1956). Unpublished.
[4] Source unknown.

Perform an appropriate analysis in R to find out if there is evidence to suggest that the change in pulse rate due to a meal is the same for men and women.

Click here to see hints.

Beware! This time we have categorical variables: "before"/"after" and "Men/Women". To create the design matrix think of writing down the terms needed to produce each data value and don't forget to remove columns to make the solution identifiable. Also, don't do the matrix multiplication/inversion by hand but use R to solve the matrix calculation. If the change in pulse rate is different for men and women then the interaction should be significant.

2.6 A laboratory experiment[5] was performed into the effect of seasonal floods on the growth of barley seedlings in a incubator, as measured by their height in mm. Three types of barley seed (Goldmarker, Midas, Igri) were used with two watering condition (Normal and Waterlogged). Further, each combination was repeated four times on different shelves in the laboratory incubator (Top, Second, Third and Bottom shelf). The data are available in the file barley.csv

Suggest a suitable model for this situation. Identify the response and explanatory variables and list the levels for any qualitative variables. Write down the design matrix for each model you consider.

Perform appropriate analyses to test if each of the following are important: (a) watering condition, (b) type of barley seed, and (c) shelf position.

In the analysis, do not include any interactions involving shelf position. If you find a significant interaction between watering condition and type of barley seed, carefully interpret the parameter estimates.

Click here to see hints.

Beware! This example has categorical explanatory variables and so don't forget to use *as.factor*. After that you need to carefully form the model in the *lm* command and interpret the ANOVA table.

> **Note**
>
> Exercise 2 Solutions can be found here.

[5] Source unknown.

3 GLM Theory

Here is a short video [3 mins] to introduce the chapter

3.1 Motivating examples

We cannot always assume that the dependent variable Y is normally distributed. For example, for the beetle mortality data in Table 1.1, suppose each beetle subjected to a dose x_i has a probability p_i of being killed. Then, the number of beetles killed Y_i out of a total number m_i at dose-level x_i will have a $\text{Bin}(m_i, p_i)$ distribution with probability mass function

$$\Pr(y_i;\ p_i, m_i) = \left(\begin{array}{c} m_i \\ y_i \end{array} \right) p_i^{y_i} (1-p_i)^{m_i - y_i} \tag{3.1}$$

where y_i takes values in $\{0, 1, \ldots, m_i\}$.

Table 3.1 contains seasonal data on tropical cyclones for 13 seasons. Suppose that, within season i, there is a constant probability $\lambda_i dt$ of a cyclone occurring in any short time-interval dt. Then the total number of cyclones Y_i during season i will have a Poisson distribution with mean λ_i, that is $Y_i \sim \text{Po}(\lambda_i)$, with probability mass function

$$\Pr(y_i;\ \lambda_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \tag{3.2}$$

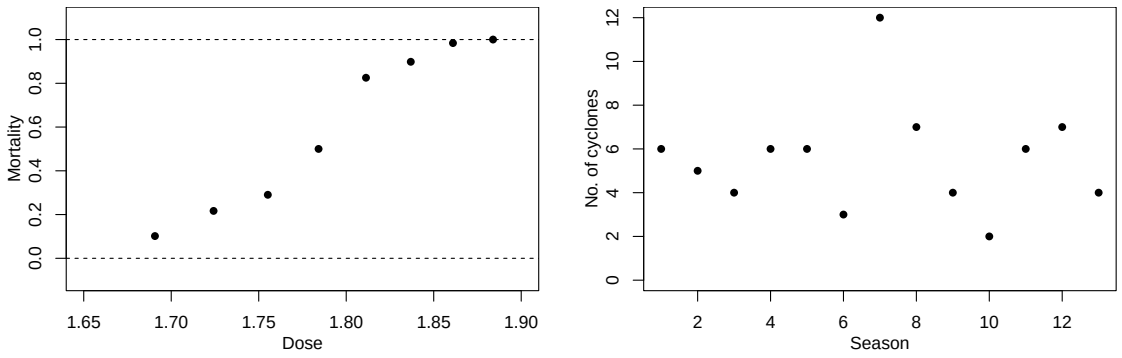
where y_i takes values in $\{0, 1, 2, \ldots\}$.

Table 3.1: Numbers of tropical cyclones in $n = 13$ successive seasons[1]

Season	1	2	3	4	5	6	7	8	9	10	11	12	13
No of cyclones	6	5	4	6	6	3	12	7	4	2	6	7	4

In these two examples, we have non-normal data and would like to know whether and how the dependent variable Y_i depends on the covariate dose or season.

[1] Dobson and Barnett, 3rd edn, Table 1.2



(a) Binomial model

(b) Poisson model

Figure 3.1: Examples of non-normally distributed data.

Generalized linear models provide a modelling framework for data analysis in the non-normal setting. We will revisit the beetle mortality and cyclone data sets after describing the structure of a generalized linear model.

Focus on distributions quiz

Test your knowledge recall and application to reinforce basic ideas and prepare for similar concepts later in the Chapter.

1. Which of the following best describes the sample space of a binomial random variable?

- (A) Positive real numbers
- (B) Non-negative integers
- (C) All real numbers
- (D) An integar values between 0 and m
- (E) Any real number between 0 and m
- (F) None of the above

2. Which of the following best describes the sample space of a Poisson random variable?

- (A) Non-negative integers

- (B) An integar values between 0 and m
- (C) Positive real numbers
- (D) All real numbers
- (E) Any real number between 0 and m
- (F) None of the above

2. Which of the following best describes the sample space of a normal random variable?

- (A) Any real number between 0 and m
- (B) All real numbers
- (C) An integar values between 0 and m
- (D) Non-negative integers
- (E) Positive real numbers
- (F) None of the above

2. For the binomial distribution, which of the following best describes the range of values possible for log(p/(1-p))?

- (A) All real numbers
- (B) Any real number between 0 and 1
- (C) Non-negative integers
- (D) An integar values between 0 and 1
- (E) Positive real numbers
- (F) None of the above

2. For the Poisson distribution, which of the following best describes the range of values possible for log( )?

- (A) All real numbers
- (B) Positive real numbers
- (C) An integar values between 0 and 1
- (D) Non-negative integers
- (E) Any real number between 0 and 1
- (F) None of the above

2. For the normal distribution, which of the following best describes the range of values possible for μ?

- (A) An integar values between 0 and 1
- (B) Non-negative integers
- (C) All real numbers
- (D) Any real number between 0 and 1
- (E) Positive real numbers
- (F) None of the above

3.2 The GLM structure

A *generalized linear model* relates a continuous or discrete response variable Y to a set of explanatory variables $\mathbf{x} = (x_1, \ldots, x_p)$. The model contains three parts:

Random part: The probability (mass or density) function of Y is assumed to belong to the *two-parameter exponential family* of distributions with parameters θ and ϕ :

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}, \tag{3.3}$$

where $\phi > 0$. Here, θ is called the *canonical* or *natural* parameter of the distribution and ϕ is called the *scale* parameter. We show below that the mean $\mathrm{E}[Y]$ depends only on θ, and $\mathrm{Var}[Y]$ depends on ϕ and possibly also θ. Various choices for functions $b(\cdot)$ and $c(\cdot)$ produce a wide variety of familiar distributions (see below). Sometimes we may set $\phi = 1$; then Equation 3.3 is called the *one-parameter exponential family*.

Most of the common statistical distributions are member of the regular exponential family, such as binomial, Poisson, geometric, gamma, normal, but not the Student's t and uniform with unknown bounds.

Further, note that in some references to generalized linear models (such as Dobson and Barnett, 3rd edn.), ϕ does not appear at all in the exponential family formula Equation 3.3, instead it is absorbed into θ and $b(\theta)$.

In this module, we will generally assume that each observation Y_i, $i = 1, \ldots, n$, is *independently* drawn from an exponential family where θ depends on the covariates. Thus we write

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}.$$

Note the subscripts on both y and θ, and hence the observations may not be identically distributed.

Systematic part: This is a *linear predictor*:

$$\eta = \sum_{j=1}^{p} \beta_j x_j. \tag{3.4}$$

Note that the symbol η is pronounced *eta*.

Link function: This is a one-to-one function providing the link between the linear predictor η and the mean $\mu = \mathrm{E}[Y]$:

$$\eta = g(\mu), \quad \text{and} \quad \mu = g^{-1}(\eta) = h(\eta). \tag{3.5}$$

Here, $g(\mu)$ is called the *link function*, and $h(\eta)$ is called the *inverse link function*.

We will now discuss each of these parts in more detail.

Focus on GLM structure quiz

Test your knowledge recall and application to reinforce basic ideas and prepare for similar concepts later in the Chapter

1. What symbols are usually used for the parameters in the definition of the two-parameter exponential family?

- (A) y, and

- (B) y and
- (C) and
- (D) b and c
- (E) y and x
- (F) None of the above

2. Which of the following is NOT a member of the two-parameter exponential family?

- (A) Student's t
- (B) Poisson
- (C) binomial
- (D) normal
- (E) Bernoulli
- (F) None of the above

3. Which of the following is NOT part of the usual definition of a generalised linear model?

- (A) Random part
- (B) Systematic part
- (C) Design matrix
- (D) Link function
- (E) None of the above

4. In the definition of a GLM, which of the following best describes an assumption about the distribution of the response variable?

- (A) Link function
- (B) Linear function of the explanatory variables
- (C) Member of the exponential family
- (D) Systematic
- (E) None of the above

5. Which part of the GLM definition involves the mean of the response variable and the explanatory variables?

- (A) Link function
- (B) Linear predictor
- (C) Regression parameters,
- (D) Expectation of Y
- (E) None of the above

3.3 The random part of a GLM

We begin with some examples of exponential family members.

Example: Poisson distribution

If Y has a Poisson distribution with parameter λ, that is $Y \sim \text{Po}(\lambda)$, then Y takes values in $\{0, 1, 2, ...\}$ and has probability mass function:

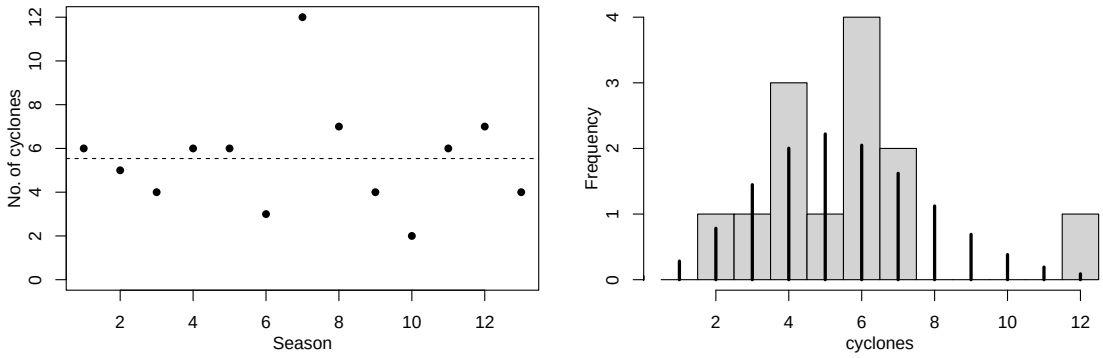
$$f(y) = \frac{e^{-\lambda}\lambda^y}{y!} = \exp\{y \log \lambda - \lambda - \log y!\}, \tag{3.6}$$

which has the form of Equation 3.3 with components as in Table 3.2.

Table 3.2: Exponential model components for the Poisson

θ	ϕ	$b(\theta)$	$c(y, \phi)$
$\log \lambda$	1	$\lambda = e^\theta$	$-\log y!$

For example, to model the cyclone data in Table 3.1, we might simply assume that the number of cyclones in each season has a Poisson distribution, assuming a constant rate λ across all seasons i. That is $Y_i \sim \text{Po}(\lambda)$. The parameter would be simply estimated by the sample mean, $\widehat{\lambda} = \bar{y} = 5.5384615$.



(a) No. of cyclones against season with mean (b) Fitted Poisson model assuming constant rate

Figure 3.2: Poisson model fitted to cyclone data.

Example: Binomial distribution

Let Y have a Binomial distribution, that is $Y \sim \text{Bin}(m, p)$, with m fixed. Then Y is discrete, taking values in $\{0, 1, \ldots, m\}$, and has probability mass function:

$$f(y) = \binom{m}{y} p^y (1-p)^{m-y} = \binom{m}{y} \left(\frac{p}{1-p}\right)^y (1-p)^m$$

which can be re-written as

$$f(y) = \exp\left\{ y \text{ logit } p + m \log(1-p) + \log \binom{m}{y} \right\}, \tag{3.7}$$

which has the form of Equation 3.3 with,

$$\theta = \text{logit } p = \log\left(\frac{p}{1-p}\right),$$

and with components as in Table 3.3.

Table 3.3: Exponential model components for the Binomial

θ	ϕ	$b(\theta)$	$c(y, \phi)$
logit p	1	$m \log(1 + e^\theta)$	$\log \binom{m}{y}$

Note that it can be shown that $-m \log(1-p) = m \log(1 + e^\theta)$ – see Exercises.

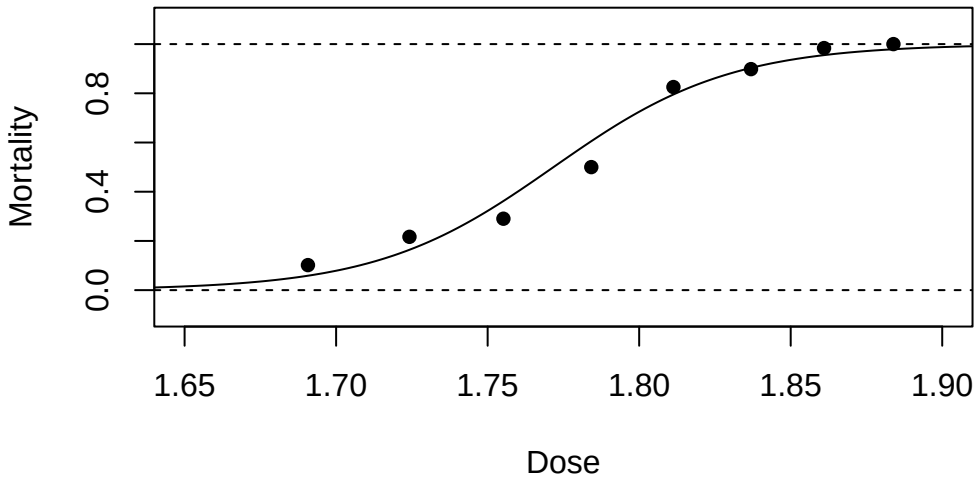


Figure 3.3: Binomial model fitted to beetle data.

Example: Normal distribution

Let Y have a Normal distribution with mean μ and variance σ^2, that is $Y \sim N(0, \sigma^2)$. Then Y takes values on the whole real line and has probability density function

$$\begin{aligned} f(y; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(y-\mu)^2\right\}, \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} + \left[\frac{-y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right]\right\}, \end{aligned}$$

which has the form of Equation 3.3 with components as in Table 3.4.

Table 3.4: Exponential model components for the Gaussian

θ	ϕ	$b(\theta)$	$c(y, \phi)$
μ	σ^2	$\theta^2/2$	$-\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)$

From the usual regression point of view, we write $y = \alpha + \beta x + \epsilon$, with $\epsilon \sim N(0, \sigma^2)$. From the point of view of a generalized linear model, we write $Y \sim N(\mu, \sigma^2)$ where $\mu(x) = \alpha + \beta x$.

3.4 Moments of exponential-family distributions

It is straightforward to find the mean and variance of Y in terms of $b(\theta)$ and ϕ. Since we want to explore the dependence of $\mathrm{E}[Y]$ on explanatory variables, this property makes the exponential family very convenient.

Proposition 3.1. *For random variables in the exponential family:*

$$E[Y] = b'(\theta), \quad \textit{and} \quad Var[Y] = b''(\theta)\phi. \tag{3.8}$$

Proof We give the proof for a continuous random variables. For the discrete case, replace all integrals by sums – see Exercises.

Starting with the simple property that all probability density functions integrate to 1, we have

$$1 = \int \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\} dy$$

and then differentiating both sides with respect to θ gives

$$0 = \int \left[\frac{y - b'(\theta)}{\phi}\right] \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}\ dy. \tag{3.9}$$

Next, using the definition of the exponential family to simplify the equation gives

$$0 = \int \left[\frac{y - b'(\theta)}{\phi}\right] f(y;\theta)\ dy$$

and expanding the brackets leads to

$$0 = \frac{1}{\phi}\left(\int y f(y;\theta) dy - b'(\theta) \int f(y;\theta)\ dy\right).$$

The first integral is simply the expectation of Y and the second is the integral of the probability density function of Y, and hence

$$0 = \frac{1}{\phi}\left(\mathrm{E}[Y] - b'(\theta)\right)$$

which implies that

$$\mathrm{E}[Y] = b'(\theta), \tag{3.10}$$

which proves the first part of the proposition.

Differentiating Equation 3.9 by parts and then using the definition of the exponential family to simplify again yields

$$0 = \int \left\{-\frac{b''(\theta)}{\phi} + \left[\frac{y - b'(\theta)}{\phi}\right]^2\right\} f(y;\theta)\ dy$$

and using Equation 3.10 gives,

$$0 = -\frac{b''(\theta)}{\phi} + \int \left[\frac{y - \mathrm{E}[Y]}{\phi}\right]^2 f(y;\theta)\, dy$$

$$0 = -\frac{b''(\theta)}{\phi} + \frac{\mathrm{Var}[Y]}{\phi^2}$$

which implies that

$$\mathrm{Var}[Y] = \phi\, b''(\theta).$$

which proves the second part of the proposition.

Together, these two results allow us to write down the expectation and variance for any random variable once we have shown that it is a member of the exponential family.

Example: Poisson and normal distribution moments

Table 3.5: Summary of moment calculations via exponential family properties

	θ	$b(\theta)$	ϕ	$\mathrm{E}[Y] = b'(\theta)$	$b''(\theta)$	$\mathrm{Var}[Y] = b''(\theta)\phi$
Poisson, $Po(\lambda)$	$\log\lambda$	e^θ	1	$e^\theta = \lambda$	e^θ	$e^\theta \times 1 = \lambda$
Normal, $N(\mu, \sigma^2)$	μ	$\theta^2/2$	σ^2	$\theta = \mu$	1	$1 \times \sigma^2 = \sigma^2$

3.5 The systematic part of the model

The second part of the generalized linear model, the linear predictor, is given in as $\eta = \sum_{j=1}^{p} \beta_j x_j$, where x_j is the jth explanatory variable (with $x_1 = 1$ for the intercept). Now, for each observation y_i, $i = 1, \ldots, n$, the explanatory variables may differ. To make explicit this dependence on i, we write:

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, \tag{3.11}$$

where x_{ij} is the value of the jth explanatory variable on individual i (with $x_{i1} = 1$). Rewriting this in matrix notation:

$$\eta = X\beta, \tag{3.12}$$

where now $\eta = (\eta_1, \ldots, \eta_n)$ is a vector of linear predictor variables, $\beta = (\beta_1, \ldots, \beta_p)$ is a vector of regression parameters, and X is the $n \times p$ design matrix.

Recall from Section 1.4 and Section 2.3 that we are concerned with two kinds of explanatory variable:

- Quantitative — for example, $x \in (-\infty, \infty)$ etc.
- Qualitative — for example, $x \in \{A, B, C\}$ etc.

As discussed in Section 2.4, each quantitative variable is represented in X by an $n \times 1$ column vector. Each qualitative variable, with $k+1$ levels, say, is represented by a dummy $n \times k$ matrix (one column, usually the first, being dropped to avoid identification problems) of 0's and 1's .

3.6 The link function

In Section 3.2 we saw that the random part of an observation, y, might be described by a member of the exponential family. We also saw, that the systematic part of y might be described using a linear predictor, η, of the explanatory variables. Further, we introduced the notion of a link function $\eta = g(\mu)$ to bring these two parts together, where μ is the mean of y.

Occasionally, the choice of link function $g(\mu)$ is motivated by theory underlying the data at hand. For example, in a dose–response setting, the appropriate model might be motivated by the solution to a set of partial differential equations describing the flow through the body of a dose of a drug.

When there is no compelling underlying theory, however, we typically choose a link function that will transform a restricted range of the dependent variable onto the whole real line. For example, when observations are typically positive, so we have $\mu > 0$, we might choose the logarithmic link:

$$g(\mu) = \log(\mu). \tag{3.13}$$

When observations are binomial counts from $B(m, p)$, $0 < p < 1$, with mean $\mu = mp$, we might choose the *logit* link from

$$\eta = g(\mu) = \text{logit}(\mu/m) = \text{logit}(p) = \log\{p/(1-p)\} \tag{3.14}$$

or the *probit* link which is the inverse of the cumulative distribution function of the $N(0, 1)$ distribution:

$$\eta = g(\mu) = \Phi^{-1}(\mu/m) = \Phi^{-1}(p), \tag{3.15}$$

or the *complementary log-log (cloglog)* link:

$$\eta = g(\mu) = \log(-\log(1-\mu/m)) = \log(-\log(1-p)), \tag{3.16}$$

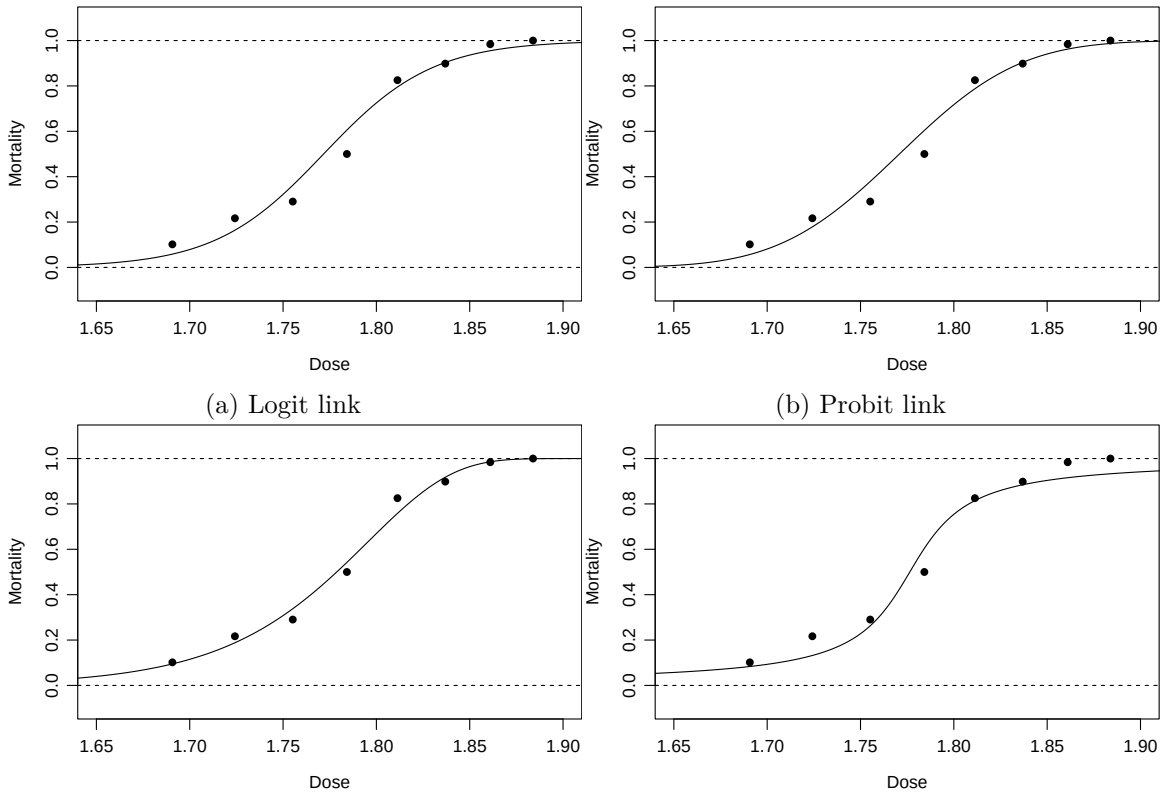
or the *cauchit* link which is the inverse of the cumulative distribution function of the Cauchy (t_1) distribution:

$$\eta = g(\mu) = \tan(\pi(\mu/m - \tfrac{1}{2})) = \tan(\pi(p - \tfrac{1}{2})). \tag{3.17}$$

Figure 3.4 shows these link functions for proportions fitted to the beetle mortality data. This demonstrates that the logit and probit links are very similar, that the complementary log-log link fits these data slightly better in the extremes, but that the cauchit link fits these data quite poorly in the extremes.

A mathematically and computationally convenient choice of link function $g(\mu)$ can be constructed by setting:

$$\theta = \eta, \tag{3.18}$$



(a) Logit link

(b) Probit link

(c) cloglog link

(d) Cauchit link

Figure 3.4: Dose–response curves fitted to the beetle mortality data from Table 1.1 with different choices of link function.

where θ is the canonical parameter of the exponential family as defined in Equation 3.3. Then, Equation 3.8 shows that the mean μ is a function of θ and therefore, Equation 3.18 indirectly provides a link between μ and η. That is, Equation 3.18 implicitly defines a link function $\eta = g(\mu)$. But what is the form of this $g(\cdot)$?

From Equation 3.8,

$$\mu = b'(\theta).$$

So, provided function $b'(\cdot)$ has an inverse $(b')^{-1}(\cdot)$, we may write

$$\theta = (b')^{-1}(\mu). \tag{3.19}$$

Now, from Equation 3.5, $g(\mu) = \eta$, so using Equation 3.18:

$$g(\mu) = \theta = (b')^{-1}(\mu), \tag{3.20}$$

from Equation 3.19. This makes explicit the $g(\mu)$ that is implicitly asserted by Equation 3.18. The function produced form Equation 3.20 is called the *canonical* link function.

Proposition 3.2. *For the canonical link function,*

$$g'(\mu) = 1/b''(\theta).$$

Proof: From Proposition 3.1, $\mu = E[Y] = b'(\theta)$, so

$$\frac{\mathrm{d}\mu}{\mathrm{d}\theta} = b''(\theta).$$

From Equation 3.20, for the canonical link function, we have $\theta = g(\mu)$, so

$$\frac{\mathrm{d}\theta}{\mathrm{d}\mu} = g'(\mu).$$

Now $\mathrm{d}\theta/\mathrm{d}\mu = (\mathrm{d}\mu/\mathrm{d}\theta)^{-1}$ and hence

$$g'(\mu) = 1/b''(\theta).$$

Which proves the proposition.

Example: Poisson canonical link function

For the Poisson distribution $\mathrm{Po}(\lambda)$, we have from Table 3.2 that $b(\theta) = e^{\theta}$. Therefore,

$$b'(\theta) = e^{\theta},$$

so the inverse of function $b'(\cdot)$ exists and is the inverse of the exponential function, which is the logarithmic function. Then, applying Equation 3.20

$$g(\mu) = \log(\mu)$$

Thus the canonical link for the Poisson distribution is log.

Example: Normal canonical link function

For the Normal distribution $N(\mu, \sigma^2)$, we have from Table 3.4 that $b(\theta) = \theta^2/2$. Therefore

$$b'(\theta) = \theta$$

so the inverse of function $b'(\cdot)$ exists and is the inverse of the identity function, which is the identity function. (The identity function is that which maps a value onto itself.) Then, applying Equation 3.20,

$$g(\mu) = \mu.$$

Thus the canonical link for the Normal distribution is the identity function.

For many models, μ has a restricted range, but we would like η to have unlimited range. It turns out, for several members of the exponential family, that the canonical link function provides η with unlimited range. However, Table 3.6 shows that this is not always so.

Table 3.6: Canonical link functions and their ranges (see McCullagh and Nelder, 2nd Edn., p291 with †binomial distribution with index m and mean μ and ‡gamma distribution with mean μ (see Exercises for details).

$f(y)$	Range of μ	$b(\theta)$	$\mu = b'(\theta)$	Canonical link, $g(\mu)$	Range of η
Normal	$(-\infty, \infty)$	$\frac{1}{2}\theta^2$	θ	μ	$(-\infty, \infty)$
Poisson	$(0, \infty)$	e^θ	e^θ	$\log \mu$	$(-\infty, \infty)$
Binomial†	$(0, m)$	$m \log(1 + e^\theta)$	$m/(1 + e^{-\theta})$	$\text{logit}(\mu/m)$	$(-\infty, \infty)$
Gamma‡	$(0, \infty)$	$-\log(-\theta)$	$-\theta^{-1}$	$-\mu^{-1}$	$(-\infty, 0)$

Why is the canonical link function Equation 3.20 convenient? The assertion Equation 3.18 means that, in the exponential-family formula Equation 3.3, we can simply substitute the linear predictor

$$\eta = \sum_j \beta_j x_j$$

from Equation 3.4 in place of θ, to give:

$$f(y; \mathbf{x}, \beta, \phi) = \exp\left\{ \frac{y\left[\sum_j \beta_j x_j\right] - b\left(\left[\sum_j \beta_j x_j\right]\right)}{\phi} + c(y, \phi) \right\}, \tag{3.21}$$

where $\mathbf{x} = \{x_j, j = 1, \ldots, p\}$ and $\beta = \{\beta_j, j = 1, \ldots, p\}$.

Further, suppose we have n independent observations, $\{y_i,\ i = 1, \ldots, n\}$. As discussed in Section 3.5, the explanatory variables $(x_1, \ldots, x_p)$ will depend on i, and so η will also depend on i. Therefore, we attach subscript i to y and to each x_j, giving:

$$f(y_i; \mathbf{x}_i, \beta, \phi) = \exp\left\{ \frac{y_i\left[\sum_j \beta_j x_{ij}\right] - b\left(\left[\sum_j \beta_j x_{ij}\right]\right)}{\phi} + c(y_i, \phi) \right\}. \tag{3.22}$$

where $\mathbf{x}_i = \{x_{ij}, j = 1, \ldots, p\}$ and $\beta = \{\beta_j, j = 1, \ldots, p\}$.

By independence, the joint distribution of all observations $\mathbf{y} = \{y_i,\ i = 1, \ldots, n\}$, with design matrix $X = \{x_{ij},\ \ i = 1, \ldots, n; j = 1, \ldots, p\}$, is:

$$f(\mathbf{y}; X, \beta, \phi) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi),$$

so

$$\log f(\mathbf{y}; X, \beta, \phi) = \sum_{i=1}^{n} \log f(y_i; \theta_i, \phi)$$

then substituting in using Equation 3.22 gives

$$\log f(\mathbf{y}; X, \beta, \phi) = \sum_{i=1}^{n} \left\{ \frac{y_i \left[\sum_j \beta_j x_{ij}\right] - b\left(\left[\sum_j \beta_j x_{ij}\right]\right)}{\phi} + c(y_i, \phi) \right\}$$

and finally simplifying to give

$$\log f(\mathbf{y}; X, \beta, \phi) = \frac{\sum_j \beta_j S_j - \sum_i b\left(\left[\sum_j \beta_j x_{ij}\right]\right)}{\phi} + \sum_i c(y_i, \phi) \tag{3.23}$$

where

$$S_j = \sum_{i=1}^{n} y_i x_{ij}.$$

Thus, in the log-likelihood Equation 3.23, it is only the first term that involves both the observations $\mathbf{y} = \{y_i,\ i = 1, \ldots, n\}$ and the parameters $\beta = \{\beta_j, j = 1, \ldots, p\}$, and this term depends on the observations only through the statistics $\mathbf{S} = \{S_j, j = 1, \ldots, p\}$ – these are called *sufficient statistics*, and their appearance in Equation 3.23 confers both theoretical and practical advantages.

Focus on link functions quiz

Test your knowledge recall and application to reinforce basic ideas and prepare for similar concepts later in the module.

1. If the reponse variable Y can take only positive values, then which of the following transformations would produce an unrestricted range?

- (A) 1/Y
- (B) logit(Y)

- (C) exp(Y)
- (D) log(Y)
- (E) None of the above

2. If the range of the reponse variable Y is restricted to the interval (0,1), then which of the following transformations would produce an unrestricted range?

- (A) logit(Y)
- (B) exp(Y)
- (C) 1/Y
- (D) log(Y)
- (E) None of the above

2. If response variable Y follows a normal distribution, then which of the following transformations would produce an unrestricted range?

- (A) log(Y)
- (B) exp(Y)
- (C) logit(Y)
- (D) $\sqrt{Y}$
- (E) None of the above

2. If the range of the reponse variable Y is a percentage, then which of the following transformations would produce an unrestricted range?

- (A) exp(Y)
- (B) log(Y)
- (C) logit(Y/100)

- (D) 1/Y

- (E) None of the above

2. If the reponse variable Y can take an real value, then which of the following transformations would produce an unrestricted range?

- (A) Log

- (B) Square-root

- (C) Logistic

- (D) Identity

- (E) None of the above

3.7 Exercises

3.1 Consider the beetle data again, see Table 1.1, but suppose that we had only been given the y values, that is the number killed, and misinformed that each came from a sample of size 62. Further, suppose that we did not know that different doses had been used. That is, we where given data: $\mathbf{y} = \{6, 13, 18, 28, 52, 53, 61, 60\}$ and led to believe that the model $Y \sim \text{Bin}(62, p)$ was appropriate. Use the given data to estimate p. Then, calculate the fitted probabilities and superimpose them on a histogram of the data.

Click here to see hints.

See the code chunk used to create Figure 3.2.

3.2 Verify that, in general, if $q = 1/(1 + e^{-x})$ then $x = \log(q/(1 - q))$ and then for the binomial distribution, $Y \sim \text{Bin}(m, p)$, show that

$$-m \log(1 - p) = m \log(1 + e^{\theta})$$

where $\theta = \text{logit } p = \log(p/(1 - p))$.

Click here to see hints.

Each deviation requires algebraic rearrangement and simplification only.

3.3 Suppose that Y has a gamma distribution with parameters α and β, that is $Y \sim \text{Gamma}(\alpha, \lambda)$, with probability density function

$$f(y; \alpha, \lambda) = \frac{\lambda^{\alpha} y^{\alpha-1} e^{-\lambda y}}{\Gamma(\alpha)} \qquad y > 0; \alpha, \lambda > 0.$$

Write this in the form of the exponential family and clearly identify θ, ϕ, $b(\theta)$ and $c(y, \phi)$ – as was done for the Poisson and binomial in Table 3.2 and Table 3.3.

Click here to see hints.

This is a difficult one and you might need to try a few rearrangements. Consider treating $1/\alpha$ as a scale parameter and let θ be a suitable function of λ and α.

3.4 Express the geometric distribution, $Y \sim \text{Geom}(p), 0 < p < 1$, with probability mass function

$$f(y) = (1-p)^{y-1}p; \qquad y = 1, 2, 3 \ldots$$

as an exponential family distribution.

Click here to see hints.

Follow the same process as usual and you might get an unexpected $c(y, \phi)$!

3.5 Prove Proposition 3.1 assuming that Y follows a discrete distribution. Use the proposition, starting from the given $b(\theta)$, to verify the results for expectation and variance of the Poisson in Table 3.5 and then derive similar results for the binomial, $Y \sim \text{Bin}(m, p)$.

Click here to see hints.

Replace the integration in the continuous case by summation for the discrete – we know that the sum of the probabilities for a discrete random variable equals 1. Remember, however, that θ is still a real number and so differentiating with respect to θ still makes sense. For the binomial, go through the steps of finding $b(\theta)$ and then it's first and second derivative.

3.6 Use the properties of exponential families to find the expectation and variance of each of the geometric and gamma distributions considered above.

Click here to see hints.

Use the results from question 3.3 and 3.4, and apply Proposition 3.1.

3.7 Using Equation 3.20, verify the canonical link function, $g(\mu)$, for the binomial and gamma distributions shown in Table 3.6.

Click here to see hints.

Use the appropriate $b'(\theta)$ and, using $\mu = b'(\theta)$ and Equation 3.20, rearrange to find θ and hence $g(\mu)$.

> **Note**
>
> Exercise 3 Solutions can be found here.

4 GLM Estimation

Here is a short video [2 mins] to introduce the chapter.

Throughout this module we use the principle of maximum likelihood estimation (MLE) to estimate model parameters and will consider two cases.

4.1 The identically distributed case

4.1.1 Maximum likelihood estimation

Suppose we have n independent and identically distributed (i.i.d.) observations $\mathbf{y} = \{y_i,\ i = 1, \ldots, n\}$, where each y_i is sampled from the same exponential family density

$$f(y_i; \theta, \phi) = \exp\left\{\frac{y_i\theta - b(\theta)}{\phi} + c(y_i, \phi)\right\}, \tag{4.1}$$

for $i = 1, \ldots, n$. In this case, the canonical parameter θ does not depend on i as the observations are identically distributed and hence must have the same parameter.

By independence, the joint distribution of all the observations $\mathbf{y}$ is:

$$f(\mathbf{y}; \theta, \phi) = \prod_{i=1}^{n} f(y_i; \theta, \phi).$$

So, taking logs and then substituting for the probability function using the exponential family form, Equation 3.3, gives

$$\log f(\mathbf{y}; \theta, \phi) = \sum_{i=1}^{n} \log f(y_i; \theta, \phi) = \sum_{i=1}^{n} \left[\frac{y_i\theta - b(\theta)}{\phi} + c(y_i, \phi)\right].$$

Regarding the observations $\mathbf{y}$ as constants (which they are, once we have data) and the scale parameter ϕ as a fixed *nuisance* parameter (whose value we may not know), the log-likelihood as a function of the parameter of interest θ is:

$$l(\theta; \mathbf{y}, \phi) = n\left(\frac{\bar{y}\,\theta - b(\theta)}{\phi}\right) + \text{constant}, \tag{4.2}$$

where $\bar{y} = \sum y_i/n$.

We estimate θ by maximizing the log likelihood – i.e. given the data $\mathbf{y}$, we estimate the value of θ to be that value for which the likelihood, and hence the log-likelihood, is greatest.

We maximize the log-likelihood by differentiating it and setting it to zero:

$$\frac{dl(\theta; \mathbf{y}, \phi)}{d\theta} = n\left(\frac{\bar{y} - b'(\theta)}{\phi}\right)$$

and hence the MLE for θ, which we denote $\hat{\theta}$, satisfies

$$b'(\hat{\theta}) = \bar{y} \tag{4.3}$$

and hence

$$\hat{\theta} = (b')^{-1}(\bar{y}). \tag{4.4}$$

Further, we showed in Proposition 3.1 that $\mathrm{E}[Y] = \mu = b'(\theta)$ and if we let $\hat{\mu}$ denote the MLE of μ, then $\hat{\mu} = b'(\hat{\theta})$[1], hence we have $\hat{\mu} = \bar{y}$. So we find that $\hat{\theta}$ is the value of θ for which the theoretical mean $\hat{\mu} = b'(\hat{\theta})$ matches the sample mean $\bar{y}$.

Example: MLE of the Poisson distribution

For the Poisson distribution, $\mathrm{Po}(\lambda)$, we have found that $b(\theta) = e^{\theta}$ and therefore $b'(\theta) = e^{\theta}$. Hence, the MLE of natural parameter θ is found as the solution of $b'(\hat{\theta}) = e^{\hat{\theta}} = \bar{y}$, that is $\hat{\theta} = \log(\bar{y})$.

4.1.2 Estimation accuracy

For our i.i.d. sample $\mathbf{y} = \{y_i,\ i = 1, \ldots, n\}$, we have $b'(\hat{\theta}) = \hat{\mu} = \bar{y}$. Let θ_0 be the true value of θ with corresponding mean μ_0, i.e.

$$b'(\theta_0) = \mu_0. \tag{4.5}$$

How accurate is $\hat{\theta}$? We know that

$$\mathrm{E}[\bar{Y}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[Y_i] = \mu_0 = b'(\theta_0), \tag{4.6}$$

using Equation 4.5. Also,

$$\mathrm{Var}[\bar{Y}] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[Y_i]$$

because the observations are independent. Then, using the result Equation 3.8 gives

$$\mathrm{Var}[\bar{Y}] = \frac{1}{n}\, b''(\theta_0)\phi. \tag{4.7}$$

We can use Taylor's theorem to expand $b'(\hat{\theta})$ about θ_0:

$$\bar{y} = b'(\hat{\theta}) \approx b'(\theta_0) + (\hat{\theta} - \theta_0)b''(\theta_0),$$

[1] Using the result that the MLE of any function of a parameter is given by the same function applied to the MLE of the parameter.

which implies that

$$(\hat{\theta} - \theta_0) \approx b''(\theta_0)^{-1}\{b'(\hat{\theta}) - b'(\theta_0)\} = b''(\theta_0)^{-1}(\bar{Y} - \mu_0), \tag{4.8}$$

using Equation 4.3 and Equation 4.5. We can use Equation 4.8 to get approximations to the mean and variance of $\hat{\theta}$:

$$\mathrm{E}[\hat{\theta} - \theta_0] \approx b''(\theta_0)^{-1}\mathrm{E}[\bar{Y} - \mu_0] = 0,$$

using Equation 4.6, so

$$\mathrm{E}[\hat{\theta}] \approx \theta_0, \tag{4.9}$$

and

$$\mathrm{Var}(\hat{\theta}) \approx \mathrm{E}\left[(\hat{\theta} - \theta_0)^2\right]$$

using Equation 4.9,

$$\mathrm{Var}(\hat{\theta}) \approx \mathrm{E}\left[\left(b''(\theta_0)^{-1}(\bar{Y} - \mu_0)\right)^2\right]$$

using Equation 4.8,

$$\mathrm{Var}(\hat{\theta}) \approx \left(b''(\theta_0)\right)^{-2}\mathrm{Var}[\bar{Y}]$$

using Equation 4.6,

$$\mathrm{Var}(\hat{\theta}) = \frac{\phi}{n\ b''(\theta_0)} \tag{4.10}$$

using Equation 4.7.

Thus we see that the first two derivatives of $b(\theta)$ play a key role in inference.

Example: Accuracy for the Poisson distribution

For the Poisson distribution, $\mathrm{Po}(\lambda)$, we have found that $\hat{\theta} = \log(\bar{Y})$. Using Equation 4.9 we know that $\hat{\theta}$ is, at least, approximately unbiased. Then, using that $\phi = 1$ (Table 3.2), $b'(\theta) = e^{\theta}$ (Table 3.6) hence $b''(\theta) = e^{\theta}$, and using Equation 4.10, leads to the result

$$\mathrm{Var}(\hat{\theta}) = \frac{\phi}{n\ b''(\theta_0)} = \frac{1}{ne^{\theta_0}}.$$

Focus on maximum likelihood quiz

Test your knowledge recall and application to reinforce basic ideas and prepare for similar concepts later in the Chapter

1. Which of the following statements best describes the purpose of maximum likelihood estimation?

- (A) To find good values of unknown model parameters

- (B) To calculate the mean of a distribution
- (C) To estimate probabilities of unknown events
- (D) To summarise a data set
- (E) To find the true values of model parameters
- (F) None of the above

2. Which of the following mathematical ideas is NOT used in the above maximum likelihood estimation?

- (A) Taking logs
- (B) By independence
- (C) The parameter depends on i
- (D) Substituting for the probability function
- (E) Identically distributed
- (F) None of the above

3. Which of the following terms is used to refer to the fixed parameter Φ?

- (A) Annoyance
- (B) Canonical
- (C) Normally distributed
- (D) Nuisance
- (E) Independent
- (F) None of the above

4. Which of the following is NOT a good property of a parameter estimator?

- (A) Has fixed variance for all sample sizes
- (B) Has expectation equal to the true value
- (C) Unbiased
- (D) Small variance
- (E) Variance decreases as sample size increases
- (F) None of the above

5. Which of the following does NOT influence the variance of the estimator of ?

- (A) Φ
- (B) c(y,Φ)
- (C) The true parameter value
- (D) The sample size
- (E) b( )
- (F) None of the above

4.2 The general case

4.2.1 MLE Estimation

Suppose that now the n independent observations $\mathbf{y} = \{y_i,\ i = 1, \ldots, n\}$ are not identically distributed. They are, however, sampled from the same exponential family density but with differing parameters, that is

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\},$$

for $i = 1, \ldots, n$. In this case, the canonical parameter does depend on i – but we assume that the scale parameter ϕ does not – and let $\theta = \{\theta_i,\ i = 1, \ldots, n\}$.

In most applications, we are not interested in estimation of θ but instead we are interested in the linear predictor parameters $\beta = \{\beta_1, \ldots, \beta_p\}$. Note, however, that each θ_i will depend on all $\beta_1, \ldots, \beta_p$. This is most obvious for the canonical parameter case where a convenient choice of link function is obtained using $\theta = \eta = X\beta$, hence $\theta_i = \mathbf{x}_i^T\beta = \sum_{j=1}^{p} x_{ij}\beta_j$ and each θ_i clearly depends on all $\beta_1, \ldots, \beta_p$.

The principle of maximum likelihood will be used to estimate the model parameters β. Using the independence of $\mathbf{y} = \{y_i,\ i = 1, \ldots, n\}$, given the parameters β, the likelihood function is:

$$L(\beta; \mathbf{y}, \phi) = f(\mathbf{y}; X, \beta, \phi) = \prod_{i=1}^{n} f(y_i; \mathbf{x}_i, \beta, \phi) \tag{4.11}$$

and the log-likelihood by

$$l(\beta; \mathbf{y}, \phi) = \log L(\beta; \mathbf{y}, \phi) = \sum_{i=1}^{n} \log f(y_i; \mathbf{x}_i, \beta, \phi). \tag{4.12}$$

Then we wish to find the value of β which maximizes the log-likelihood function,

$$\widehat{\beta} = \max_{\beta} l(\beta; \mathbf{y}, \phi).$$

For generalized linear models there is usually no closed-form expression for the MLE $\widehat{\beta}$. Instead, an iterative approach based on Newton's Method is often used.

For the normal linear regression model, however, that is where the exponential family is Gaussian and the link function $g(\mu)$ is the identity function, we have the familiar closed-form expression

$$\widehat{\beta} = \left(X^T X\right)^{-1} X^T \mathbf{y}. \tag{4.13}$$

4.2.2 The score function and Fisher information

We define the *score function*:

$$U(\beta) = \frac{\partial l(\beta; \mathbf{y}, \phi)}{\partial \beta}, \tag{4.14}$$

which is a $p \times 1$ vector. We define the *observed Fisher information*:

$$I(\beta) = -\frac{\partial U(\beta)}{\partial \beta^T} = -\frac{\partial^2 l(\beta; \mathbf{y}, \phi)}{\partial \beta\, \partial \beta^T}, \tag{4.15}$$

which is a $p \times p$ matrix whose (j,k)th element is: $-\frac{\partial^2 l(\beta; \mathbf{y}, \phi)}{\partial \beta_j \partial \beta_k}$. We also define the *expected Fisher information*:

$$J(\beta) = \mathrm{E}\left[-\frac{\partial^2 l(\beta; \mathbf{y}, \phi)}{\partial \beta\, \partial \beta^T}\right], \tag{4.16}$$

which is also a $p \times p$ matrix.

Newton's Method then involves choosing an initial value for the parameter vector, $\widehat{\beta}_0$ say, and then repeatedly updating, until convergence, the value using

$\hat{\beta}_t = \hat{\beta}_{t-1} - I^{-1}(\hat{\beta}_{t-1})\ U(\hat{\beta}_{t-1})$. A related method, called Fisher Scoring, replaces I in the iterative step by J – calculating the expected Fisher information is more work but does lead to better numerical properties.

Proposition 4.1. *With definitions Equation 4.14, Equation 4.15, Equation 4.16 above,*

- $E[U(\beta)] = 0$,
- $J(\beta) = E\left[U(\beta)U^T(\beta)\right]$.

Proof We give the proof for continuous random variables. For the discrete case, replace integration by sums – see Exercises.

To start the proof, notice that we can re-write the joint density of the data given the parameters as

$$f(\mathbf{y}; X, \beta, \phi) = L(\beta; \mathbf{y}, \phi) = \exp\left\{l(\beta; \mathbf{y}, \phi)\right\}$$

and then

$$1 = \int f(\mathbf{y}; X\beta, \phi)d\mathbf{y} = \int \exp\{l(\beta; \mathbf{y}, \phi)\}d\mathbf{y},$$

where $d\mathbf{y} = dy_1 \cdots dy_n$. Differentiating this with respect to $\beta = (\beta_1, \ldots, \beta_p)^T$ gives:

$$\begin{aligned} 0 &= \int \frac{\partial l(\beta; \mathbf{y}, \phi)}{\partial \beta} \exp\{l(\beta; \mathbf{y}, \phi)\}d\mathbf{y} \\ &= \int U(\beta) f(\mathbf{y}; X, \beta, \phi)d\mathbf{y} \qquad (\star) \\ &= \mathrm{E}\left[U(\beta)\right]. \end{aligned}$$

Proving the first part.

Next, differentiating $(\star)$ by parts with respect to β^T,

$$\begin{aligned} 0 &= \int \frac{\partial U(\beta)}{\partial \beta^T} f(\mathbf{y}; X, \beta, \phi) + \frac{\partial l(\beta; \mathbf{y}, \phi)}{\partial \beta} \frac{\partial l(\beta; \mathbf{y}, \phi)}{\partial \beta^T} f(\mathbf{y}; X, \beta, \phi)d\mathbf{y} \\ &= \int -I(\beta) f(\mathbf{y}; X, \beta, \phi)d\mathbf{y} + \int U(\beta)U^T(\beta) f(\mathbf{y}; X, \beta, \phi)d\mathbf{y} \\ &= -J(\beta) + \mathrm{E}\left[U(\beta)U^T(\beta)\right]. \end{aligned}$$

proving the second part.

Proposition 4.2. *Under some regularity conditions, the MLE $\hat{\beta}$ of β has the following asymptotic properties:*

- *$E(\hat{\beta}) = \beta$; i.e. $\hat{\beta}$ is unbiased for β.*
- *$Var(\hat{\beta}) = J^{-1}(\beta)$.*
- *$\hat{\beta}$ follows a p-dimensional Normal distribution.*

Combining these three statements we have that asymptotically

$$\hat{\beta} \sim N_p(\beta, J^{-1}(\beta)). \tag{4.17}$$

Proof: Omitted. Similar to the proof using Taylor's theorem in Section 4.1.

From Equation 4.17, variances $\text{Var}(\hat{\beta}_k)$, standard errors $\text{se}(\hat{\beta}_k)$ and correlations between parameter estimates $\text{Corr}(\hat{\beta}_k, \hat{\beta}_h)$ can be estimated.

4.2.3 The saturated case

Again we assume the observations $y_i,\ i = 1, \ldots, n$ are independent but now we assume that y_i is sampled from an exponential family probability function with canonical parameter θ_i,

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\},$$

for $i = 1, \ldots, n$. We can form the log-likelihood in the usual way to give

$$l(\theta; \mathbf{y}, \phi) = \sum_{i=1}^{n} \log f(y_i; \theta_i, \phi) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right]$$

and so we find the the MLE of θ using

$$\hat{\theta}_i = (b')^{-1}(y_i), \quad i = 1, \ldots, n.$$

Note that each parameter is only a function of the corresponding observation.

Further, from Proposition 3.1, $\text{E}[Y_i] = \mu_i = b'(\theta_i)$ and if we again let $\hat{\mu}_i$ denote the MLE of μ_i, then $\hat{\mu}_i = b'(\hat{\theta}_i)$, hence we have $\hat{\mu}_i = y_i$. Thus we see that, under the saturated model, the mean of the distribution of y_i is estimated to be equal to y_i itself. That is, the data are fitted exactly by the model. Of course this model is quite useless for explanation or prediction, since it misinterprets random variation as systematic variation. Nevertheless, the saturated model is useful as a benchmark for comparing models, as we will see later.

It is worth noting that the same situation can occur even when modelling in terms of the regression parameters $\beta_j,\ j = 1, \ldots, p$. When $p \geq n$, and if the covariates are linearly independent, each θ_i can take on any value independently of the others and so estimating the β_j's is equivalent to estimating the θ_i's. That is we are also considering the *saturated* or *full* model. This highlights the danger of putting too many covariates into the model. There is a big literature on how to deal with more parameters then data using techniques of *regularized regression*.

Focus on general maximum likelihood quiz

Test your knowledge recall and application to reinforce basic ideas and prepare for similar concepts later in the Chapter

1. Which of the following statements best describes the purpose of maximum likelihood estimation?

- (A) To summarise a data set
- (B) To find the true values of model parameters
- (C) To find good values of unknown model parameters
- (D) To calculate the mean of a distribution
- (E) To estimate probabilities of unknown events
- (F) None of the above

2. Which of the following mathematical ideas is NOT used in the general case of maximum likelihood estimation?

- (A) The log-likelihood
- (B) The parameter does not depends on i
- (C) Usually no closed-form expression
- (D) Using the independence
- (E) Not identically distributed
- (F) None of the above

3. Which of the following terms is NOT used in maximum likelihood theory?

- (A) The log-likelihood function
- (B) The moment generating function

- (C) Fisher information
- (D) Score function
- (E) The likelihood function
- (F) None of the above

4. Which of the following is a true statement about the saturated case?

- (A) The parameters and unrelated to the values of
- (B) The model fits the data exactly
- (C) The values of will be exactly the same as the values of
- (D) It should never be considered
- (E) Has fewer parameters than data observations
- (F) None of the above

5. Which of the following is NOT a true statement about the asymptotic properties of the maximum likelihood estimator of ?

- (A) Follows a normal distribution
- (B) Will always be equal to the true value.
- (C) The asymptotic properties are when n tend to infinity
- (D) The variance can be found form the expected Fisher information
- (E) Is unbiased
- (F) None of the above

4.3 Model deviance

The *deviance* is a quantity we use to assess the fit of a model to the data. Let M be a model of interest with fitted parameters $\hat{\theta}$ and corresponding fitted values $\hat{\mu}$. Also consider the saturated model with fitted parameters $\tilde{\theta}$ and fitted values $\tilde{\mu}$.

The *deviance* of model M is defined as twice the difference between the log-likelihood of the saturated model, $l(\tilde{\theta}; \mathbf{y}, \phi)$, and the log-likelihood of model M, $l(\hat{\theta}; \mathbf{y}, \phi)$, multiplied by ϕ,

$$
\begin{aligned}
D &= 2\phi \left\{ l(\tilde{\theta}; \mathbf{y}, \phi) - l(\hat{\theta}; \mathbf{y}, \phi) \right\} \qquad (4.18) \\
&= \begin{cases}
\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 = \text{Residual sum of squares} & \text{Normal} \\
2 \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i) \log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right) \right\} & \text{Binomial} \\
2 \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i + \hat{\mu}_i \right\} & \text{Poisson}
\end{cases}
\end{aligned}
$$

Note that Dobson, and others, call $D^* = D/\phi = 2\{l(\tilde{\theta}; y, \phi) - l(\hat{\theta}; y, \phi)\}$ the *scaled* deviance.

We now consider two situations:

Scale parameter ϕ known. For some data-types (e.g. Poisson, Binomial), we know $\phi = 1$. Consider two nested models M_1 and M_2 with r_1 and r_2 parameters respectively where the parameters in M_1 are a subset of those in M_2 and hence $r_1 < r_2$. Further, let D_1 and D_2 be the deviances of model M_1 and M_2 respectively.

Then, asymptotically,

- the log likelihood-ratio statistic $D_1 - D_2 \sim \chi^2_{r_2 - r_1}$ can be used to test the importance of the extra parameters in M_2 not included in M_1;
- a goodness-of-fit test for M_2 can be done based on $D_2 \sim \chi^2_{n - r_2}$.

The quality of the approximations involved depends on there being a large amount of *information*, for example, large counts for Binomial and Poisson data, or a large sample size for Normal data.

Scale parameter ϕ unknown. For some data-types (e.g. Normal, Gamma), ϕ is not known (typically $\phi = \sigma^2$). We must find a model M_3 *big* enough to be believed, then estimate ϕ by the residual mean square:

$$
\hat{\phi} = \frac{D_3}{n - r_3}. \qquad (4.19)
$$

Then test M_1 against M_2 using

$$
\mathrm{F} = \frac{(D_1 - D_2)/(r_2 - r_1)}{\hat{\phi}} = \frac{(D_1 - D_2)/(r_2 - r_1)}{D_3/(n - r_3)} \qquad (4.20)
$$

with

$$\mathrm{F} \sim F_{r_2-r_1, n-r_3}. \tag{4.21}$$

So if the observed value of the statistic Equation 4.20 was within the upper (say 5%) tail of the F-distribution Equation 4.21, we would infer that Model M_2 is better than Model M_1.

4.4 Model residuals

Consider a generalized linear model with observed values $y_i, i = 1, \ldots, n$ and fitted values $\hat{\mu}_i$. Then the *raw* or *response* residuals are defined by

$$e_i^{\text{raw}} = y_i - \hat{\mu}_i.$$

More useful are the *standardized* or *Pearson* residuals defined by

$$e_i^{\text{std}} = e_i^{\text{P}} = \frac{y_i - \hat{\mu}_i}{\sqrt{b''(\theta_i)}}.$$

Recall from Equation 3.8 that $\mathrm{Var}(Y_i) = \phi\, b''(\theta_i)$.

Deviance residuals are defined so that the sum of squared deviance residuals equals the total deviance. Thus we set

$$e_i^{\text{dev}} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i},$$

where d_i is the contribution of observation i to the deviance, D. For example, when y_i has a Poisson distribution with estimated mean $\hat{\mu}_i$, we have

$$e_i^{\text{dev}} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i + \hat{\mu}_i\right]}.$$

Residuals are useful for assessing the overall fit of a model to the data, and for identifying where the model might need to be improved.

4.5 Fitting generalized linear models in R

4.5.1 GLM-related R commands

The function used to fit a generalized linear model in **R** is

`glm(formula, family)`.

Let `x,y,z,a,b,c,` ... be a set of vectors all of the same length n (perhaps read in from a data file using the `read.table` and `attach` commands). If `a,b,c` are qualitative variables, then they first need to be declared as factors by `a = as.factor(a)`, etc.

The `formula` argument of `glm` specifies the required model in compact notation, e.g. `y` $\sim$ `x*a` or `y` $\sim$ `x + z*a` where $\sim$, `+`,`*` have the same meaning as in Section 2.5.

The `family` argument specifies which exponential family is to be used. We shall use `gaussian, poisson` and `binomial`; `gaussian` is the default. Other options are available; see `help(family)` for further information.

Along with the family, a link function can be specified. The possible choices are:

- `gaussian` – `"identity"` (default)
- `poisson` – `"log"` (default), `"sqrt"`, `"identity"`
- `binomial` – `"logit"` (default), `"probit"`, `"cloglog"`.

R assumes the default options unless we state otherwise. For example,

`glm(y` $\sim$ `a+b)` # Gaussian errors, identity link
`glm(y` $\sim$ $a + b$`, poisson)` # Poisson errors, log link
`glm(y` $\sim$ `a+b, poisson("sqrt"))` # Poisson errors, sqrt link

Note that for the binomial case, the response variable should be an $n \times 2$ matrix `ym`, say, not a vector, where the first column contains the numbers of successes and the second column the numbers of failures, for example:

`glm(ym` $\sim$ `a+b, binomial)` # binomial errors, logit link

To extract information about a fitted generalized linear model, it is best to store the result of `glm` as a variable and then to use the following functions:

- To fit a GLM and store the result in `y.glm` (for example):
  `y.glm = glm(y` $\sim$ `a*b, poisson("sqrt"))`
- To print various pieces of information including deviance residuals, parameter estimates and standard errors, deviances, and (if specified) correlations of parameter estimates:
  `summary(y.glm, correlation=T)`
- To print the anova table of the fitted model:
  `anova(y.glm)`
- To print the deviance of the fitted model:
  `deviance(y.glm)`
- To print the residual degrees of freedom of the fitted model:
  `df.residual(y.glm)`
- To print the vector of fitted values under the fitted model:
  `fitted.values(y.glm)`
- To print the residuals from the fitted model:
  `residuals(y.glm, type)`
  Note: `type` should be `"deviance"` (default), `"pearson"`, or `"response"`
- To print the parameter estimates from the fitted model:
  `coefficients(y.glm)`
- To print the design matrix for a specified model formula: `model.matrix(y` $\sim$ `a*b)`

The functions `summary` and `anova` are the most useful for printing out information about the fitted model. The results of the other functions can be saved as variables for further computation, if desired.

4.5.2 Example of fitting Poisson GLM in R

Here is a toy example of **R** commands for modelling a response in terms of two qualitative explanatory variables (that is *factors*). The model assumes the data are Poisson-distributed and uses the logarithmic link function.

```
Call:
glm(formula = y ~ a + b, family = poisson)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1408     0.3351   3.404 0.000663 ***
a2           -0.3054     0.3522  -0.867 0.385934
a3            0.1466     0.3132   0.468 0.639712
a4            0.4568     0.2932   1.558 0.119269
b2            0.9163     0.3162   2.898 0.003761 **
b3            0.9445     0.3150   2.999 0.002712 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 31.725  on 11  degrees of freedom
Residual deviance: 13.150  on  6  degrees of freedom
AIC: 68.227

Number of Fisher Scoring iterations: 5

Correlation of Coefficients:
   (Intercept) a2    a3    a4    b2
a2 -0.45
a3 -0.50        0.48
a4 -0.54        0.51  0.57
b2 -0.67        0.00  0.00  0.00
b3 -0.68        0.00  0.00  0.00  0.72


Analysis of Deviance Table

Model: poisson, link: log
```

```
Response: y

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev
NULL                    11     31.725
a     3   6.2793         8     25.445
b     2  12.2947         6     13.150


[1] 13.15047


[1] 6


       1         2         3         4         5         6         7         8
3.129412  2.305882  3.623529  4.941176  7.823529  5.764706  9.058824 12.352941
       9        10        11        12
8.047059  5.929412  9.317647 12.705882
```

Focus on GLM fitting in R quiz

Test your knowledge recall and application to reinforce basic ideas and prepare for similar concepts later in the Chapter

1. Which of the following is the main R command used to fit generalised linear models?

- (A) fitglm
- (B) lm
- (C) formula
- (D) family
- (E) glm
- (F) None of the above

2. Which of the following is NOT an option for the model family?

- (A) Gamma

- (B) binomial
- (C) poisson
- (D) logit
- (E) gaussian
- (F) None of the above

3. Which of the following is NOT a link option in a GLM?

- (A) probit
- (B) logit
- (C) constant
- (D) log
- (E) identity
- (F) None of the above

4. Which of the following is NOT a link option with the binomial family?

- (A) log
- (B) inverse
- (C) logit
- (D) cloglog
- (E) cauchit
- (F) None of the above

5. Which of the following is NOT a n R command used to check model fit?

- (A) residuals
- (B) deviance
- (C) summary
- (D) glm
- (E) anova
- (F) None of the above

4.6 Exercises

4.1 Use Equation 4.4 to obtain estimation equations for the natural parameter θ, based on a sample $\mathbf{y} = \{y_1, \ldots, y_n\}$, for each of the following situations:

(a) the binomial, $Y \sim \text{Bin}(m, p)$,
(b) the geometric, $Y \sim \text{Ge}(p)$,
(c) the exponential, $Y \sim \text{Exp}(\lambda)$.

Assuming that n is very large, use Equation 4.9 to comment on possible bias of the estimator for large samples. What is the corresponding variance of the estimator?

Click here to see hints.

For each situation equate the theoretical expectation, in terms of the derivative of b, to the sample mean – that is start from Equation 4.3 and solve. For bias and variance you should consider the asymptotic results.

4.2 For a sample of size n from the normal distribution, $Y \sim N(\mu, \sigma^2)$, how do the results produced using Equation 4.9 and Equation 4.10 compare with the familiar results $\hat{\mu} = \bar{Y}$, $\text{E}[\hat{\mu}] = \mu$, and $\text{Var}[\hat{\mu}] = \sigma^2/n$?

Click here to see hints.

Identify, θ, ϕ and $b''(\theta)$ then substitute into the equations.

4.3 For the normal linear regression model, $\mathbf{Y} = X\beta + \epsilon$ where $\epsilon \sim N_n(0, \sigma^2 I_n)$ use the principle of maximum likelihood to show that the MLE has the closed form given in Equation 4.13.

Click here to see hints.

Apply the rules for matrix differentiation of a quadratic form. If needed, refer to *Appendix C* in the *Basic Pre-requisite Material folder* in *Minerva*.

4.4 Check that you can derive the formulas given for deviance of normal, binomial, and Poisson models at the start of Section 4.3.

Click here to see hints.

Form the likelihoods of the saturated model, recalling that the fitted values are exactly equal to the observed y values, and substitute them into the general deviance equation. Then, simplify.

4.5 How do the deviance tests defined in Section 4.3 related to the tests used to find the best fit model for the birthweight example in Section 2.1?

Click here to see hints.

Start by comparing the notation. Then, replace the R and r in the F-statistic with D and adjust the notation for degrees of freedom. You should then get very expressions. The only difference being from which model σ^2 is estimated.

4.6 Consider a clinical trial into the effectiveness of Amoxicillin (a widely used antibiotic) against bacterial chest infections in children. A sample of 200 children with chest infections were randomly assigned to one of five groups with Amoxicillin dose levels of 20-100 mg/kg per day. Four weeks later it was recorded whether the child had required a re-treatment with a further dose of antibiotic, or not, with results summarized in Table 4.1

Table 4.1: Effectiveness of Amoxicillin against bacterial chest infections in children

	20mg	40mg	60mg	80mg	100mg
Group size	23	18	14	23	22
Re-treatment	20	14	5	6	3

First fit a normal linear model and then fit a generalized linear model assuming binomial errors and a logistic link. Superimpose both of fitted model onto a scatter plot of the data. Which do you think is the better model? Justify you answer.

Click here to see hints.

The linear model part would be straightforward, look at earlier examples if not. For the glm case, check the R code block for Figure 1.1b. Simply compare the fit by eye.

Note

Exercise 4 Solutions can be found here.

5 Modelling Proportions

Here is a video [8 mins] to introduce the chapter and including the demonstration of a short example in R.

5.1 Introduction

In this chapter we will focus on applications of generalized linear modelling where the response variable follows a binomial distribution. This can arise when the outcome is binary, that is it can take one of only two possible values, or is the sum of a set of such binary outcomes. These two outcomes record whether some event of interest has occurred or not. In the simplest case, the response variable, B say, is defined as

$$B = \begin{cases} 1 & \text{if the event has occurred} \\ 0 & \text{if the event has not occurred} \end{cases} \tag{5.1}$$

and we set $Pr(B = 1) = p$, and hence $Pr(B = 0) = 1 - p$. This is, of course, the definition of a Bernoulli trial leading to a Bernoulli random variables, $B \sim \text{Bernoulli}(p)$.

Suppose now that there are m similar binary outcomes, $B_i, i = 1, \ldots, m$ with $Pr(B_i = 1) = p$, and that the total number of times that the event occurred is recorded as the response Y. Assuming that m is known before the trials start, that p is fixed and that the individual Bernoulli trials are independent then Y follows a binomial distribution, $Y \sim \text{B}(m, p)$ with probability mass function

$$f(y) = \binom{m}{y} p^y (1-p)^{m-y}. \tag{5.2}$$

Note that the binomial random variable can be thought of as the sum of i.i.d Bernoulli random variables, $Y = B_1 + \cdots + B_m$, if that is helpful.

The Binomial distribution $\text{B}(m, p)$ is often described in terms of *success* and *failure* and a binomial distribution in terms of the number of successes in m independent trials, where p is the probability of success in each trial. The term *success* need not correspond to a favourable outcome; it is merely the language traditionally used in connection with this model. For example, success might correspond to death.

Of course, the special case with $m = 1$ reduces to the Bernoulli distribution. Further, if $Y_1 \sim \text{B}(m_1, p)$ and independently $Y_2 \sim \text{B}(m_2, p)$, then $Y_1 + Y_2$ also follows a binomial distribution, $\text{B}(m_1 + m_2, p)$ – note that is only valid when p is common.

Recall that the binomial can be re-written in the exponential family form of Equation 3.3 with

$$f(y) = \exp\left\{y \text{ logit } p + m\log(1-p) + \log\binom{m}{y}\right\},$$

with natural or canonical parameter $\theta = \text{logit } p = \log\left(p(1-p)\right)$, scale parameter $\phi = 1$, $b(\theta) = m\log(1+e^{\theta})$ and $c(y,\phi) = \log\binom{m}{y}$ as in Table 3.3.

5.2 The logistic model

Throughout this module we have assumed that a response variable, Y, depends on a set of explanatory variables, $\mathbf{x} = \{x_1, \ldots, x_p\}$. In particular, for binomial counts from $\text{B}(m,p)$, $0 < p < 1$ with mean $\mu = mp$ we set the link function, $g(\mu)$, equal to the linear predictor $\eta = \sum \beta_j x_j$ and hence have

$$g(\mu) = \sum_{j=1}^{p} \beta_j x_j = \mathbf{x}^T\beta$$

where $\beta = \{\beta_1, \ldots, \beta_p\}$ are the linear predictor parameters. Recall that the canonical logit link Equation 3.14 for the binomial leads to the systematic part of the model

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \mathbf{x}^T\beta \tag{5.3}$$

but that other links function are possible, such as the probit, Equation 3.15, the complementary log-log, Equation 3.16, and the cauchit Equation 3.17.

This model can alternatively be written

$$Y \sim \text{B}(m,p), \quad \text{where } p = \frac{\exp\{\mathbf{x}^T\beta\}}{1+\exp\{\mathbf{x}^T\beta\}} \tag{5.4}$$

which makes the dependence of Y on $\mathbf{x}$, and β, more explicit.

In general, we might want to consider n independent binomial random variables representing subgroups of the sample with $Y_1 \sim \text{B}(m_1, p_1), \ldots Y_n \sim \text{B}(m_n, p_n)$, see Table 5.1 and hence

$$\mathbf{Y} \sim \text{B}(\mathbf{m}, \mathbf{p}), \quad \text{and } \mathbf{p} = \frac{\exp\{X\beta\}}{1+\exp\{X\beta\}},$$

with response $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$, $\mathbf{m} = \{m_1, \ldots, m_n\}$, $\mathbf{p} = \{p_1, \ldots, p_n\}$, X being the $n \times p$ design matrix and $\beta = \{\beta_1, \ldots, \beta_p\}$ the linear predictor parameters.

Table 5.1: Linear logistic model

	Group 1	Group 2	...	Group n
Successes	Y_1	Y_2	...	Y_n
Failures	$m_1 - Y_1$	$m_2 - Y_2$	...	$m_n - Y_n$
Total	m_1	m_2	...	m_n

Group 1	Group 2	...	Group n

Then we can write down the log likelihood of β using Equation 4.12 and Equation 5.2 as

$$l(\beta; \mathbf{y}) = \sum_{i=1}^{n} \left\{ y_i \log(p_i) + (m_i - y_i) \log(1 - p_i) + \log \binom{m_i}{y_i} \right\} \tag{5.5}$$

where

$$p_i = \frac{\exp\{\mathbf{x}_i^T \beta\}}{1 + \exp\{\mathbf{x}_i^T \beta\}}.$$

We would then use the principle of maximum likelihood to estimate β

$$\hat{\beta} = \arg\max_{\beta} \, l(\beta; \mathbf{y}).$$

Then, the fitted probability are given by

$$\hat{p}_i = \frac{\exp\{\mathbf{x}_i^T \hat{\beta}\}}{1 + \exp\{\mathbf{x}_i^T \hat{\beta}\}}$$

and corresponding fitted values

$$\hat{y}_i = m_i \, \hat{p}_i.$$

The Pearson residuals for binomial data take the form

$$e_i^P = \frac{y_i - m_i \hat{p}_i}{\sqrt{m_i \hat{p}_i (1 - \hat{p}_i)}}.$$

From the general result $\text{Var}(Y_i) = \phi b''(\theta)$ in Proposition 3.1, it can be shown that $\text{Var}(Y_i) = m_i p_i (1 - p_i)$. Then, using the estimate of p_i leads to the denominator above.

For large m_i, the usual Normal approximation to the Binomial means that the Pearson residuals are approximately $N(0, 1)$ distributed.

It can be shown that the deviance for the fitted model is

$$D = 2 \sum_{i=1}^{n} \left\{ y_i \log \left(\frac{y_i}{m_i \hat{p}_i} \right) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i (1 - \hat{p}_i)} \right) \right\}, \tag{5.6}$$

which is approximately χ^2_{n-r} distributed if the model is correct, where r is the number of degrees of freedom in the model (that is the number of parameters or columns of the design matrix).

This formula can be shown to be equivalent to

$$D = 2 \sum_{j=1}^{2} \sum_{i=1}^{n} O_{ji} \log \frac{O_{ji}}{E_{ji}}$$

where O_{ji} denotes the observed value and E_{ji} denotes the expected value in cell (j, i) of the $2 \times n$ table of successes and failures:

Table 5.2: Observed and expected frequencies in the linear logistic model

	Group 1	Group 2	...	Group n
Successes, $j = 1$	O_{11}	O_{12}	...	O_{1n}
	E_{11}	E_{12}	...	E_{1n}
Failures, $j = 2$	O_{21}	O_{22}	...	O_{2n}
	E_{21}	E_{22}	...	E_{2n}

Another goodness-of-fit statistic is the Pearson chi-squared statistic:

$$X^2 = \sum_{j=1}^{2} \sum_{i=1}^{n} \frac{(O_{ji} - E_{ji})^2}{E_{ji}}.$$

This is asymptotically equivalent to the deviance Equation 5.6 (proof is by Taylor series expansion; omitted). Thus, asymptotically, X^2 is also approximately χ^2_{n-r} distributed. Both approximations can be poor if the expected frequencies are small, but X^2 copes slightly better with this problem. See Dobson, p.136 for more details.

Focus on model choice quiz

Test knowledge recall and comprehension to reinforce ideas of linear and non-linear models. More questions to be added later.

1. Which of the following is NOT a true statements about modelling?

- (A) The context of the problem and description of the data can be ignored – it's only the data that matter
- (B) Some situations might suggest more than one model to try.
- (C) Goodness of fit can be assessed using the model residuals.
- (D) It is important to plot the fitted model on the same graph as the data.
- (E) All model assumptions should be checked.

2. Which of the following is a true statements about linear regression?

- (A) It is always possible to use the fitted model for extrapolation.

- (B) It is always important to check model residuals.
- (C) A linear model will fit best when the correlation is close to zero.
- (D) When there is a moderate correlation then the relationship between variables is always linear.
- (E) A linear model should always be the first model considered.

3. Which of the following is a true statements about logistic regression?

- (A) The explanatory variables follow a binomial distribution
- (B) Should only be used if a linear model is not a good fit to the data.
- (C) A logisitc model should always be used if the correlation is close to zero.
- (D) It is always possible to use the fitted model for extrapolation.
- (E) The response variable follows a binomial distribution.

Click here to see explanations

More detail will be added later.

1. The starting point should always to learn about the background to a data set as this often suggests models to consider, appropriate distributions and any potential problems. Hence saying that it is only the data that maters is inappropriate.
2. Of course there are many important things to check, but looking at residuals is one of the most important.
3. Logistic regression assumes that the data follow a binomial distribution.

5.3 Overdispersion

Examination of residuals and deviances may indicate that a model is not an adequate fit to the data. One possible reason is *overdispersion*. In particular, it is advisable to compare the residual deviance of the proposed model with the corresponding degrees of freedom. If the ratio, deviance divided by degrees of freedom, is substantially greater than 1, then there is evidence for overdispersion – substantially less than 1 indicates under dispersion but that is very rare. Overdispersion can occur for any error distribution where the variance is

linked to the mean — e.g. Binomial, Poisson. In the binomial case, overdispersion is called *extra-Binomial* variation.

Recall that if $Y_i \sim \text{Bin}(m_i, p_i)$, $\text{Var}(Y_i) = m_i p_i(1-p_i)$. Overdispersion occurs if observations which have been modelled by a $\text{Bin}(m_i, \widehat{p}_i)$ distribution have substantially greater variation than $m_i\widehat{p}_i(1-\widehat{p}_i)$. This will lead to a value of D substantially greater than the expected value of $n-r$. This can occur if the model is missing appropriate explanatory variables or has the wrong link function, or if the Y_i are not independent.

One solution is to include an extra parameter τ in the model so that $\text{Var}(Y_i) = \tau \times m_i p_i(1-p_i)$. For more details, see Section 7.7 of Dobson or Chapter 6 of Collett (1991) *Modelling Binary data*, Chapman & Hall.

The `glm` function in **R** allows for *extra-Binomial* variation by setting `family=quasibinomial()` with the usual link functions available.

5.4 Odds ratios for 2x2 tables

In the special case of the data forming a 2×2 table, then a commonly calculated summary statistic is the *odds ratio*. This is a measure of association between the presence or absence of some property and an outcome. It represents the odds that an outcome will occur given presence, compared to the odds of the outcome occurring in the absence of that property.

The following 2×2 table summarizes the different possibilities.

	Success	Failure
Present	y_1	$m_1 - y_1$
Absent	y_2	$m_2 - y_2$

Similarly, by dividing by the corresponding row total, the probabilities are summaries as follows.

	Success	Failure
Present	π_1	$1 - \pi_1$
Absent	π_2	$1 - \pi_2$

There are conditional probabilities, for example Pr(Success|Present) etc., and form a *stochastic matrix* where sum of each row is now 1.

For each row, the *odds* of success is defined by

$$O_i = \pi_i/(1-\pi_i), \quad i = 1, 2,$$

and a comparison between these two probabilities is given by the *odds ratio*

$$\phi = \frac{O_1}{O_2} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}.$$

If $\phi = 1$ then the property does not affect the odds of success, $\phi > 1$ indicates that the property is associated with higher odds of success, and $\phi < 1$ indicates the property is associated with lower odds of success.

Note that the *log-odds*, $\log O_i = \log\left(\pi_i/(1-\pi_i)\right)$, which is the $\text{logit}(\pi_i)$, and hence is the natural parameter in the regular exponential family form of the binomial distribution.

Finally, an approximate 100(1-α)% confidence interval, $\left(\phi_{\alpha/2}, \phi_{1-\alpha/2}\right)$, for the odds ratio is calculated using the formulas:

$$\phi_{\alpha/2} = \phi \times \exp\left(-z_{1-\alpha/2}\sqrt{\frac{1}{y_1} + \frac{1}{(m_1-y_1)} + \frac{1}{y_2} + \frac{1}{(m_2-y_2)}}\right)$$

and

$$\phi_{1-\alpha/2} = \phi \times \exp\left(+z_{1-\alpha/2}\sqrt{\frac{1}{y_1} + \frac{1}{(m_1-y_1)} + \frac{1}{y_2} + \frac{1}{(m_2-y_2)}}\right)$$

where $z_{1-\alpha/2}$ is such that $\Pr(Z \leq z_{1-\alpha/2}) = 1-\alpha/2$. In particular, for a 95% confidence interval $z_{1-\alpha/2} = z_{0.975} = 1.96$.

If the confidence interval for the odds ratio includes the value 1 then the calculated odds ratio would not be considered statistically significant.

5.5 Application to dose–response experiments

A variable dose of some reagent is administered to each study subject, and the occurrence of a specific response is recorded. This is a *dose–response* experiment, one of the first uses of regression models for Bernoulli (or Binomial) responses.

For example, the Table 5.5 (including the information from Table 1.1) gives the number of beetles killed y_i and the number not killed $(m_i - y_i)$ out of a total number m_i that were exposed to a dose x_i of gaseous carbon disulphide, for $n = 8$ dose levels $i = 1, \ldots, 8$ (Dobson: pp.109 in 1st edn; pp.119 in 2nd edn; pp.127 in 3rd edn). The proportion killed $p_i = y_i/m_i$ at each dose level i is also shown in Table 5.5 and plotted in Figure 5.1.

Table 5.5: Numbers of beetles killed by five hours of exposure to 8 different concentrations of gaseous carbon disulphide

Dose x_i	No. of beetle m_i	No. killed y_i	No. not killed $m_i - y_i$	Proportion $p_i = y_i/m_i$
1.6907	59	6	53	0.10
1.7242	60	13	47	0.22

Dose x_i	No. of beetle m_i	No. killed y_i	No. not killed $m_i - y_i$	Proportion $p_i = y_i/m_i$
1.7552	62	18	44	0.29
1.7842	56	28	28	0.50
1.8113	63	52	11	0.83
1.8369	59	53	6	0.90
1.8610	62	61	1	0.98
1.8839	60	60	0	1.00

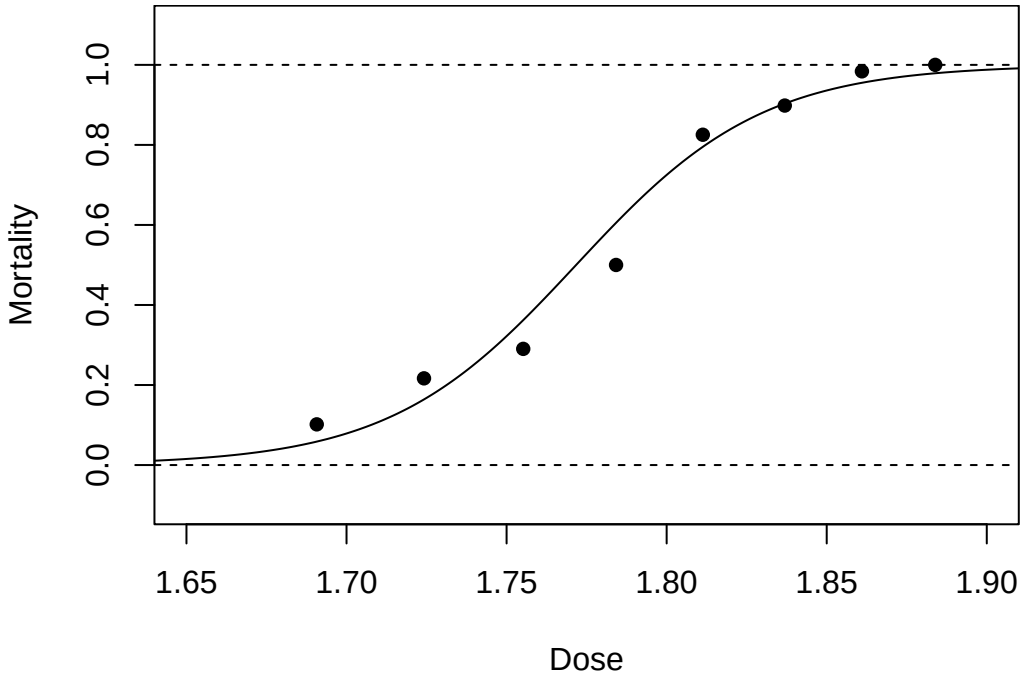


Figure 5.1: Beetle mortality rates with fitted dose-response curves.

Now Equation 5.4 motivates modelling the beetle data as

$$Y_i \sim \mathrm{B}(m_i, p_i), \quad \text{for } i = 1, \ldots, n = 8$$

where

$$\eta_i = \alpha + \beta x_i$$

and so

$$p_i = \frac{\exp\{\alpha + \beta x_i\}}{(1 + \exp\{\alpha + \beta x_i\})}.$$

To fit this model in **R**, a matrix with columns containing the numbers killed y_i and the numbers not killed $m_i - y_i$ is first calculated

```
     [,1] [,2]
```

```
[1,]    6   53
[2,]   13   47
[3,]   18   44
[4,]   28   28
[5,]   52   11
[6,]   53    6
```

which is then used in the `glm` command

The model parameter estimates are given by

```
(Intercept)        dose
  -60.71745    34.27033
```

and hence

$$\hat{p}_i = \frac{\exp\{-60.7 + 34.3x_i\}}{(1 + \exp\{-60.7 + 34.3x_i\})}. \tag{5.7}$$

Further, the deviance if given by

```
[1] 11.23223
```

or more helpfully a full summary, which contains parameter estimates and deviance values using

```
Call:
glm(formula = y ~ dose, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
dose          34.270      2.912   11.77   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4
```

or the anova table

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev
NULL                     7    284.202
dose  1   272.97         6     11.232
```

Clearly, many of the results are repeated in the various **R** output.

When considering testing model goodness of fit, it is important to distinguish between deviance values which compare a given model with the saturated model and deviance values which compare two competing models. For example, in the Analysis of Deviance Table, the right hand values compare the Null model with the saturated and the model including dose with the saturated. Whereas the left hand value is relevant to the comparison of Null model with the mode including dose.

Summary of results:
From the output, we can first test if the Null model, which contains only a constant term, is a good fit to the data using the hypotheses:

$$H_0 : \text{Null model is true against } H_1 : \text{Null model is false.}$$

From the output, note that the Null deviance is $D_0 = 284.202$ and that this model has $r_0 = 1$ parameters. Therefore, D_0 follows a χ^2 distribution with $n - r_0 = 8 - 1 = 7$ degrees of freedom and hence the p-value is:

```
[1] 1.425247e-57
```

This means that we reject H_0 as the observed value of D is in the upper tail of the χ^2 distribution.

If we had obtained a non-significant result, then we would stop the analysis and conclude that the Null model is an adequate fit.

Next, consider testing the model which includes the explanatory variable where the the hypotheses are

$$H_0 : \text{Null model is true against } H_1 : \text{Model with dose is true.}$$

From the output, the deviance of the proposed model is $D_1 = 11.232$ with $r_1 = 2$ parameters. The test statistics is then $D_0 - D_1 = 284.202 - 11.232 = 272.97$ which follows a χ^2 distribution with $r_1 - r_0 = 2 - 1 = 1$ degrees of freedom. The p-value is then

```
[1] 2.556369e-61
```

and we reject H_0 in favour of H_1 and conclude that dose is important.

We can finish the testing by checking if this model is a good fit to the data. That is,

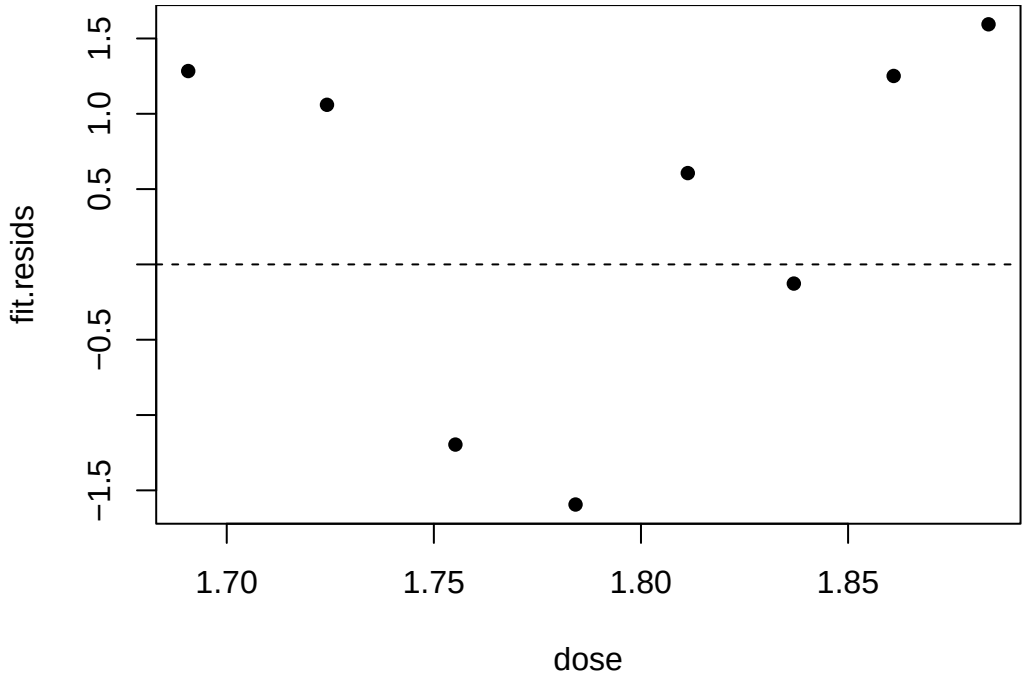
$$H_0 : \text{The model with dose is true against } H_1 : \text{The model with dose is false}$$

The deviance of this is $D_1 = 11.232$ with $r_1 = 2$ parameters which follows a χ^2 distribution with $n - r_1 = 8 - 2 = 6$ degrees of freedom and has p=value

```
[1] 0.08146544
```

which means that we accept H_0 at the 5% level and conclude that this model is an adequate fit to the data.

Finally, we can look at the residuals:



These show a very slight *u-shaped* pattern and hence it would be work exploring fitting using the other link functions. For this data set, the complementary log-log (cloglog) link has a slightly better fit, see Figure 3.4, but a slight pattern remains.

Before finishing, suppose that we wish to predict, by hand, the probability for dose level $x_0 = 1.85$, say, and the expected number of beetles killed when $m_0 = 60$ beetles, say, are exposed.

The probability is predicted as

$$\hat{p}_i = \frac{\exp\{-60.7 + 34.3x_0\}}{(1 + \exp\{-60.7 + 34.3x_0\})} = 0.936$$

and the expected number of beetles

$$m_0 \, \widehat{p}_i = 56.2$$

Finally, suppose that we wish to find the minimum dose which kills 95% of the beetles, that is $p_0 = 0.95$ which requires us to invert the logistic equation.

In general, suppose we wish to find x_p which corresponds to proportion p then

$$x_p = \frac{1}{\beta} \left\{ \log \left(\frac{p}{1-p} \right) - \alpha \right\}$$

and hence here $\widehat{x}_p = 1.856$.

5.6 Exercises

5.1. Which of the following descriptions suggest that a logistic regression should be considered. Justify your answers.

a. A pharmaceutical company perform clinical trials where volunteers are given varying doses of an experimental drug to study the chances of potential negative side effects. From historical data 4 previous volunteers out of a total of 120 have had negative side effects.

b. A credit card company monitors sales to identify fraudulent transactions. On the previous five days, 26, 20, 18, 19 and 21 regular transactions occurred before the first fraudulent transaction was identified.

c. A national internet service provider monitors the frequency of helpline phone calls regarding internet failures. From records, there are an average of 0.27 calls per day.

d. The University Student Education Services monitor the success of registered students every year. In particular, they record whether students pass their examinations, or not, and what *application score* they were given when they applied for a place at University.

Click here to see hints.

For each, consider what is response, is it binomial and what is the explanatory variable. In cases which are not binomial, suggest an alternative distribution.

5.2. For each of the situation produce appropriate graphs within RStudio. Describe the type of relationship, if any, shown in the plots. Is a linear, a logistic, or some other model appropriate? If linear or logistic, go ahead with model fitting and perform a residual analysis. Do you think that prediction from the model would be reliable? Justify each step of your analysis.

a. A UK loan company requires a detailed questionnaire to be completed for any loan application including such information as income, regular expenditure, previous loan details and other information from data analytics and consumer credit reporting companies. All the information is combined into a single *credit score* from 0 to 1000. A large sample of credit applicants were contacted by an independent market research company to discover the outcome of their application. The data file *loans.csv* gives the number of successful application, number declined and the credit scores.

b. A small estate agent company is interested in studying the changes in sales of houses due to targeted advertising campaigns and in particular to determine if additional spending is worthwhile. The company record the weekly sales before each advertising campaign and then again for one week a full month later, along with the cost of each campaign. The datafile *advertising.csv* contains the advertising budget values in pounds and corresponding changes in sale, also in pounds.

c. The yield of arable crops is sensitive to the nutrient levels in the soil. A key fertilizer used to boost production is the mineral phosphorus which is applied as phosphorus pentoxide, P_2O_5. At a wheat large farm, various amounts from 0 kg/ha to 80 kg/ha were applied to investigate the effect on wheat yield, in kg/ha. Use the data in file *wheat.csv* to investigate this situation.

d. Cardiac troponin (cTn) is a human protein which is commonly measured in the diagnosis of heart attacks. A small sample of blood is taken from patients suspected of having heart attacks and the level of cardiac troponin is recorded The cardiac troponin levels, in ng/mL, of 100 patients arriving at a hospital Accident & Emergency department with significant chest discomfort along with the later confirmed diagnosis (0=No and 1=Yes).

Click here to see hints.

For each, consider what is response, is it binomial and what is the explanatory variable. If the plot reveals a relationship which is either linear or logistic, then follow the appropriate steps previously discussed – check the code blocks in various examples. For the residual analysis, plot a graph of residuals against fitted values and check to a lack of pattern. For the linear cases, also consider the assumption of normality. For example using a histogram.

5.3 In the beetle example, use the fitted logistic regression model, Equation 5.7, to predict what dose of gaseous carbon disulphide would kill 90% of beetles. Then, fit alternative link functions to the data and plot the results. Do you think that the choice of link function makes a difference to your conclusions? Justify your answer.

Click here to see hints.

For the prediction, rearrange the usual expression for p in terms of x and the model parameters to give an expression for x. To fit alternative link functions, just change the argument in the *family* option.

5.4 (Based on Dobson & Barnett, pp 144–145). Suppose there are 2 groups of people: the first group is exposed to some pollutant and the second group is not. In a prospective study, each group is followed for several years and categorized according to the presence or absence of some disease. Let π_i denote the probability that a person in group i contracts the disease, $i = 1, 2$. The following 2×2 table summarizes the different possibilities.

	Diseased	Not diseased
Exposed	π_1	$1 - \pi_1$
Not exposed	π_2	$1 - \pi_2$

Note that the sum of each row is 1. For each $i = 1, 2$, the *odds* of contracting the disease is defined by

$$O_i = \pi_i/(1 - \pi_i),$$

and a comparison between these two probabilities is given by the *odds ratio*

$$\phi = \frac{O_1}{O_2} = \frac{\pi_1(1 - \pi_2)}{\pi_2(1 - \pi_1)}.$$

a. Show that $\phi = 1$ if and only if there is no difference between the control and exposed groups. What does it mean if $\phi > 1$?

b. Consider now m tables where each is of this form, with probabilities π_{ij} represented by a logistic model

$$\text{logit}(\pi_{ij}) = \alpha_i + \beta_i x_j, \ i = 1, 2, \ j = 1, \ldots, m,$$

where x_j is some specified quantitative explanatory variable. Interpret the parameters α_i and β_j, and give their effect on the log odds ratio $\log \phi_j$, say, for each table. Show that $\log \phi_j$ is constant across the m tables if $\beta_1 = \beta_2$.

c. Give a practical example where such a model might be appropriate.

d. How would you express this model in the R computer language?

Click here to see hints.

In (a), one way is easy. For the other way, rearrange the odds ration and consider the ratios of the event occurring and the ratio that the events not occurring. These are only equal if the probabilities are the same. In (b), note that the log odds ratio is exactly the logistic transform. Substitute in and simplify. An example should be easy, and check ealier for the R command.

5.5. Consider again the beetle example, but suppose that rather than the actual dose concentration had been measured, instead only whether a *Low dose* or *High dose* was used recorded. This would lead to the following table:

	Died	Not died	Total
Low dose	65	172	237
High dose	226	18	244

What are the respective probability of death values in the two dose groups? The *relative risk* is defined as the ratio of these two values. Calculate the risk of death due to High dose exposure relative to Low dose for this example. How could this value be used to measure the association between dose and death?

What are the respective odds of death in each of the two groups? Calculate the *odds ratio*, defined as the ratio of these two values. How could this value be used to interpret the data?

Click here to see hints.

If there is no difference then the relative risk should be 1 and a value substantially different indicates an association. The odds and the odds ratio have a similar interpretation. All measures are do the same thing, but each is standard in particular application areas.

> **Note**
>
> Exercise 5 Solutions can be found here.

6 Loglinear Models

Here is a video [4 mins] to introduce the chapter.

6.1 Overview

In this chapter, we deal with data sets in which the response variable y is a count and the explanatory variables are all factors – i.e. qualitative variables. Initially we assume y has a Poisson distribution.

See Chapter 9 of Dobson and Barnett (2008). See also Agresti (1996) *An introduction to categorical data analysis*.

We assume a generalized linear model with

- responses (counts) having independent Poisson distributions;
- a logarithmic link function (hence the name *log-linear model*).

Consider, for example, the two-way contingency table in Table 6.1}:

Table 6.1: A two-way contingency table with k_1 rows and k_2 columns, where each entry y_{ij} is a count.

	1	2	...	j	...	k_2	Total
1	y_{11}	y_{12}	...	y_{1j}	...	y_{1k_2}	y_{1+}
...	...	...	...	...	...	...	...
i	y_{i1}	y_{12}	...	y_{1j}	...	y_{ik_2}	y_{i+}
...	...	...	...	...	...	...	...
k_1	$y_{k_1 1}$	$y_{k_1 2}$	...	y_{1j}	...	$y_{k_2 k_2}$	y_{1k_1}
Total	y_{+1}	y_{+2}	...	y_{1j}	...	y_{+k_2}	y_{++}

In row i and column j of Table 6.1} we assume

$$Y_{ij} \sim \text{Po}(\lambda_{ij})$$

where

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}. \tag{6.1}$$

Here μ is called the *main effect*; α_i is a *row effect*; β_j is a *column effect*; and $(\alpha\beta)_{ij}$ denotes an *interaction effect* parameter. Some or all of these effects must be present in the model,

with constraints to ensure that the model is identifiable – this is sometimes refered ot as the *identifiability* or *aliasing problem*. Generally, **R** function `glm` automatically sets the first level of each effect to zero to achieve model identification. Thus, for the two-way model,

$$\alpha_1 = 0, \quad \beta_1 = 0, \quad (\alpha\beta)_{11} = (\alpha\beta)_{1j} = (\alpha\beta)_{k_1} = 0.$$

6.2 Motivating examples

6.2.1 Malignant melanoma

The data in Table 6.2 are from a study of 400 patients with malignant melanoma, a particular form of skin cancer (see Dobson and Barnett, p.172). For each tumour, its type and site were recorded. The data in Table 6.2 comprise the numbers of tumours y in each combination of site and tumour-type. This is a two-way *contingency table*. We want to know how melanoma frequency depends on site and type.

Table 6.2: Melanoma counts by type and site.

Type	Head	Trunk	Extremities	Total
Hutchinson's melanotic freckle	22	2	19	34
Superficial spreading melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

Table 6.3: Percentages across columns within rows for melanoma data.

Type	Head	Trunk	Extremities	Total
Hutchinson's melanotic freckle	64.7	5.9	29.4	100
Superficial spreading melanoma	8.6	29.2	62.2	100
Nodular	15.2	26.4	58.4	100
Indeterminate	19.6	30.4	50.0	100
Total	17.0	26.5	56.5	100

Table 6.4: Percentages across rows within columns for melanoma data.

Type	Head	Trunk	Extremities	Total
Hutchinson's melanotic freckle	32.4	1.9	4.4	8.5
Superficial spreading melanoma	23.5	50.9	50.9	46.3
Nodular	27.9	31.1	32.3	31.3
Indeterminate	16.2	16.0	12.4	14.0

Type	Head	Trunk	Extremities	Total
Total	100.0	99.9	100.0	100.0

Although we have two factors, `Type` and `Site`, that we may use as predictors, standard ANOVA regression methods are inappropriate here as the dependent variable is not continuous but is instead a count. We will use *log-linear regression*, a type of generalized linear model, to analyse these data.

Table 6.3 shows row and Table 6.4 column percentages for these data. For example, 15.2% of nodular melanomas occurred in the head and neck, 26.4% in the trunk, and 58.4% in the extremities. Compare this to the equivalent figures for Hutchinson's melanotic freckles: 64.7%, 5.9%, and 29.4% - strikingly different. So different types of melanomas are more likely to occur in different locations.

6.2.2 Flu vaccine

The data in Table 6.5 are from a randomized controlled trial in which 73 patients were randomized into two groups (see Dobson and Barnett, 2008, p.173). The treatment group was given a flu vaccine, while the control group was given a placebo. Levels of an antibody (HIA) were measured after 6 weeks and classified into three groups: Low, Moderate, and High.

Table 6.5: Antibody responses to flu vaccine from a randomized controlled trial.

Group	Low	Moderate	High
Placebo	25	8	5
Vaccine	6	18	11
Total	31	26	16

Is the pattern of response the same for each treatment group? The percentages in Table 6.6 suggest not - row percentages indicate lower responses in the placebo group.

Table 6.6: Percentages across rows within columns for antibody responses to flu vaccine.

Group	Low	Moderate	High	Total
Placebo	65.8	21.1	13.2	100
Vaccine	17.1	51.4	31.4	100
Total	42.4	35.6	22.0	100

Table 6.7: Percentages across rows within columns for antibody responses to flu vaccine.

Group	Low	Moderate	High
Placebo	80.6	30.8	31.2
Vaccine	19.4	69.2	68.8
Total	100	100	100

6.3 Maximum likelihood estimation

Recall that for each cell $Y_{ij} \sim \text{Po}(\lambda_{ij})$ so $\text{E}[Y_{ij}] = \lambda_{ij}$. However, we estimate $\hat{\lambda}_{ij} = y_{ij}$ only for the *saturated model* given by Equation 6.1. In general, for non-saturated models, the estimate $\hat{\lambda}_{ij} \neq y_{ij}$.

Consider the independence model for a 2-way table, that is where $(\alpha\beta)_{ij} = 0$ for all i, j. Here,

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j \tag{6.2}$$

and y_{ij} is the observed count for cell (i, j), so the likelihood is given by

$$L(\lambda; y) = \prod_{i,j} \frac{e^{-\lambda_{ij}} \lambda_{ij}^{y_{ij}}}{y_{ij}!}$$

and the log-likelihood is

$$l(\lambda; y) = \sum_{i,j} \left\{ y_{ij} \log \lambda_{ij} - \lambda_{ij} - \log y_{ij}! \right\}.$$

Next, using Equation 6.2, gives

$$l(\lambda; y) = \sum_{i,j} \left\{ y_{ij}(\mu + \alpha_i + \beta_j) - \exp(\mu + \alpha_i + \beta_j) - \log y_{ij}! \right\}$$

which can be re-written as

$$l(\lambda; y) = \mu y_{++} + \sum_i \alpha_i y_{i+} + \sum_j \beta_j y_{+j} - e^{\mu} \left(\sum_i e^{\alpha_i} \right) \left(\sum_j e^{\beta_j} \right) - \sum_{ij} \log y_{ij}!. \tag{6.3}$$

The maximum likelihood estimates of the model parameters are obtained in the usual way. Differentiating Equation 6.3 with respect to μ and setting the result to zero gives, at the MLE,

$$y_{++} = e^{\hat{\mu}} \left(\sum_i e^{\hat{\alpha}_i} \right) \left(\sum_j e^{\hat{\beta}_j} \right).$$

Differentiating Equation 6.3 with respect to α_i (where $i \neq 1$ because $\alpha_1 = 0$ to avoid an identifiability problem) and setting the result to zero gives, at the MLE,

$$y_{i+} = e^{\widehat{\mu}} e^{\widehat{\alpha}_i} \left(\sum_j e^{\widehat{\beta}_j} \right).$$

Differentiating Equation 6.3 with respect to β_j (where $j \neq 1$ because $\beta_1 = 0$) and setting the result to zero gives, at the MLE,

$$y_{+j} = e^{\widehat{\mu}} e^{\widehat{\beta}_j} \left(\sum_i e^{\widehat{\alpha}_i} \right).$$

Then

$$\frac{y_{i+}\ y_{+j}}{y_{++}} = e^{\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j} = \hat{\lambda}_{ij}. \tag{6.4}$$

It follows that

$$\hat{\lambda}_{i+} = y_{i+} \quad \hat{\lambda}_{+j} = y_{+j} \quad \hat{\lambda}_{++} = y_{++}.$$

Thus, the total fitted count in row i is identical to the total observed count in row i. Further, the total fitted count in column j is equal to the total observed count in column j and the total fitted count equals the total observed count.

6.4 Model fitting in R

Consider an analysis of Melanoma data introduced in Section 6.2.1.

To test the independence of `Type` and `Size`, we fit the model

$$\texttt{Count} \sim \texttt{Type} + \texttt{Site}$$

assuming Poisson counts and a logarithmic link function, using the commands:

```
Call:
glm(formula = mcount ~ type.F + site.F, family = "poisson")

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7544     0.2040   8.600  < 2e-16 ***
type.F2       1.6940     0.1866   9.079  < 2e-16 ***
type.F3       1.3020     0.1934   6.731 1.68e-11 ***
type.F4       0.4990     0.2174   2.295  0.02173 *
site.F2       0.4439     0.1554   2.857  0.00427 **
site.F3       1.2010     0.1383   8.683  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 295.203  on 11  degrees of freedom
Residual deviance:  51.795  on  6  degrees of freedom
```

```
AIC: 122.91

Number of Fisher Scoring iterations: 5
```

We see that the residual deviance for this model is 51.795 on 6 degrees of freedom. If the model is true, the residual deviance will have an approximate χ^2 distribution on 6 degrees of freedom. If the model is not true, the residual deviance will probably be too large to correspond to this distribution. Thus we calculate the p-value in the upper tail of the χ^2_6 distribution, which can be computed using the command:

```
[1] 2.050465e-09
```

This strongly indicates that the independence model is inadequate. Therefore, unless we can spot alternative simplifications, we will have to use the saturated model.

We can look at residuals from the model see where the departures from independence occur. The largest residual is for Hutchinson's freckle on the head and neck, which occurs more often than would be expected under independence.

For the saturated model, $\hat{\lambda}_{ij} = y_{ij}$. In this example, $\sum_{ij} \hat{\lambda}_{ij} = \sum_{ij} y_{ij} = 400$, so the probability of a tumour being in category (i, j) is $y_{ij}/400$ — just the observed proportions.

Note that in Table 6.2 we have a total of $y_{++} = 400$ observations. If the data were truly Poisson, this would be a suspiciously round number. In reality, this total was fixed by design, so we should take into account the fact that $y_{++} = 400$ and fit a more suitable model, such as the *multinomial* model which we will meet in the next chapter.

Overdispersion can occur for the Poisson model, just as in the binomial case – see Section 5.3. This is called *extra-Poisson* variation. The `glm` function in **R** can take this into account by including an extra parameter τ in the model using `family=quasipoisson()`.

6.5 Multi-way contingency tables

We can generalize the model notation introduced in Equation 6.1 to accommodate multi-way contingency tables; that is tables of counts indexed by multiple factors. For example, for a 3-way table of cell counts y_{ijk}, the saturated model would be written:

$$Y_{ijk} \sim \text{Po}(\lambda_{ijk})$$

where

$$\log \lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}.$$

Here, α_i, β_j and γ_j are *main effects*; $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, and $(\beta\gamma)_{jk}$ are two-way *interaction effects*; and $(\alpha\beta\gamma)_{ijk}$ is a *three-way interaction.*

This approach is easily extended to tables of any number of factors. Note, however, that not all terms need be present in a model, thought 3-way or higher-order interactions might sometimes be required.

A *hierarchical* model is one in which each term in the model is accompanied by all lower-order terms. For example, including the term $(\alpha\beta\gamma)_{ijk}$ would require inclusion of each of the terms $\alpha_i, \beta_j, \gamma_k, (\alpha\beta)_{ij}, (\alpha\gamma)_{ik}, (\beta\gamma)_{jk}$. We will be concerned only with *hierarchical* contingency table models. Non-hierarchical models are sometimes appropriate, depending on the nature of the factors involved, but hierarchical models are more generally applicable and easier to interpret.

6.6 Exercises

Hints to be added later.

6.1 Overdispersion is an occasional problem when fitting generalized linear models with a known scale parameter in situations where there is unexplained variation. In this exercise we illustrate the problem in Poisson regression. The starting point is the observation that if $Y \sim P(\lambda)$, then $\mathrm{E}[Y] = \lambda$ and $\mathrm{Var}[Y] = \lambda$.

a. Consider joint random variables (X, Y) where X takes two possible values with equal probabilities, $Pr(X = 1) = Pr(X = 2) = 1/2$.

   Suppose the conditional distribution of Y given X is Poisson,

   $$Y|X = 1 \sim P(\lambda_1), \quad Y|X = 2 \sim P(\lambda_2), \qquad (*)$$

   where $\lambda_1 < \lambda_2$. Let $\lambda = (\lambda_1 + \lambda_2)/2$ denote the average value. Thus the marginal distribution of Y is a mixture of two Poisson distributions. Show that

   $$\mathrm{E}[Y] = \lambda, \quad \mathrm{Var}[Y] = \lambda + (\lambda_1 - \lambda_2)^2/4,$$

   that is, although the mean of Y is the same under the mixture (or conditional Poisson) model as under the Poisson model, the variance is larger.

b. This phenomenon might be observed in data as follows. Let $n = 60$ and let an explanatory variable x_i take the value $x_i = 1$ for $i = 1, \ldots, 30$ and $x_i = 2$ for $i = 31, \ldots, 60$. Suppose that the observations $y_i|x_i$ come from the above conditional Poisson model $(*)$.

   Consider fitting the following two models in R with Poisson errors and a log link function:

   $$\text{(i)} \;\; y \sim 1, \qquad \text{(ii)} \;\; y \sim x.$$

   Since model (ii) is the correct model, it should yield a good fit to the data. But if the experimenter does not know about the variable x, it will only be feasible to fit model

(i). Let $\overline{Y}$ and S^2 denote the sample mean and variance of the $Y_i,\ i = 1, \ldots, 60$. Show that

$$\mathrm{E}[\overline{Y}] = \lambda, \quad \mathrm{E}[S^2] = \lambda + \frac{60}{59}(\lambda_1 - \lambda_2)^2/4.$$

Hence show that the Pearson χ^2 goodness of fit statistic for model (i) will indicate a poorly fitting model if λ_1 and λ_2 are far apart.

c. This example is very simple, but overdispersion can occur much more widely. Why is overdisperson not a problem for generalized linear models in which the response distribution includes a scale parameter?

6.2 (From Dobson & Barnett, p 163) This question should be done in a computer package such as R. You should think carefully about which variables, if any, to condition on in your analysis: HOME = 1,2,3, CONTACT = 1,2, or SATISFACTION = 1,2,3.

The data relate to an investigation into satisfaction with housing conditions in Copenhagen. Residents of selected areas living in rented houses built between 1960 and 1968 were questioned about their satisfaction and their degree of contact with other residents. The data were tabulated by type of housing. Investigate the associations between satisfaction, contact with other residents and type of housing.

Low Contact:

Satisfaction:	Low	Medium	High
Tower blocks	65	54	100
Apartments	130	76	111
Houses	67	48	62

High Contact:

Satisfaction:	Low	Medium	High
Tower blocks	34	47	100
Apartments	141	116	191
Houses	130	105	104

a. Produce appropriate tables of percentages to gain initial insights into the data; for example, percentages in each contact level by type of housing and level of satisfaction, or percentages in each level of satisfaction by contact and type of housing.

b. Using e.g. R, fit various log-linear models to investigate interactions between the variables.

c. For some model that fits (at least moderately) well, calculate the Pearson residuals and use them to find where the largest discrepancies are between the observed and expected values.

> **Note**
>
> Exercise 6 Solutions can be found here.