MATH3823 Generalized Linear Models

Robert G Aykroyd

2024-02-04

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 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*.

Advanced notice

- **Module Assessment:** Set on 14 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:** 27/28 February for Practice and 19/20 March for Assessment – 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.

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
	Exercises	Chapters 3 & 4
Week 7	Chapter 5	All
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*.

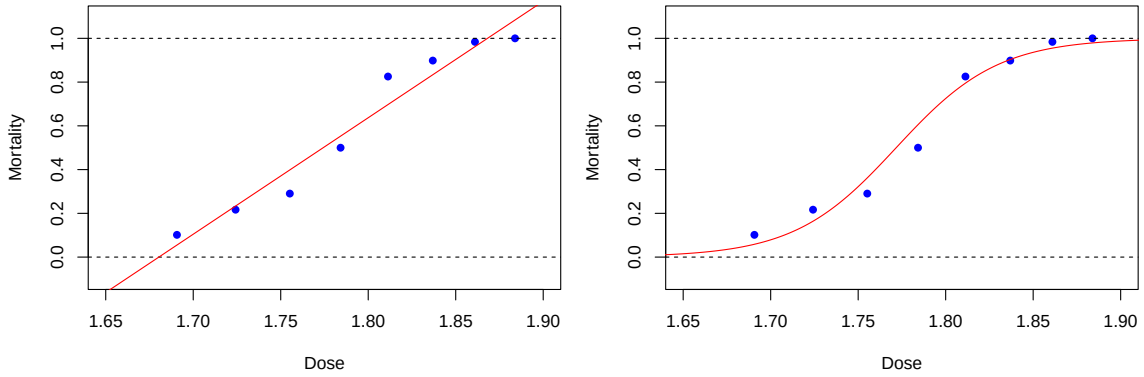
1.2 Motivating example

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

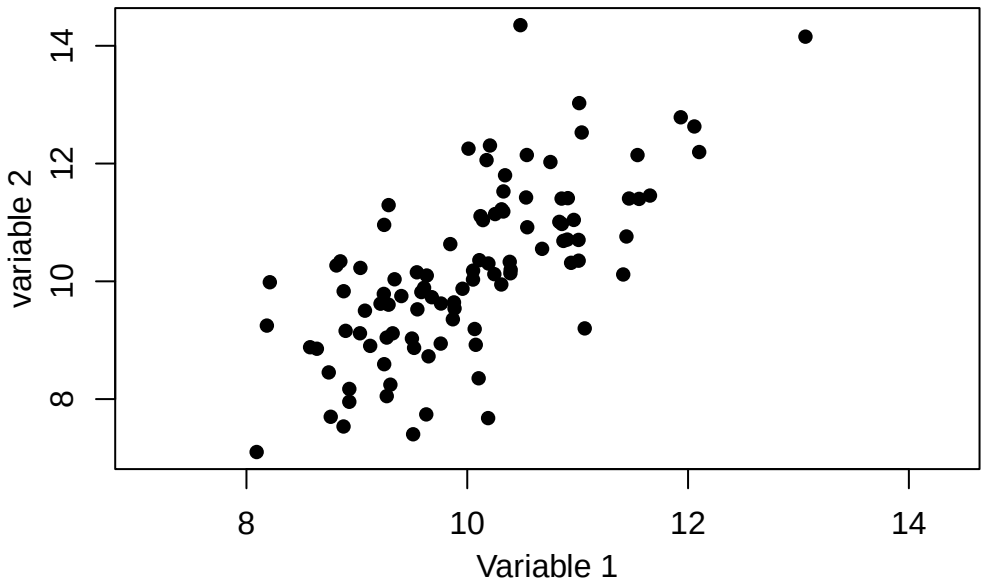
1. It is well documented that people who are over-weight (or at least high BMI) are more likely to have high blood pressure. This is true for systolic and diastolic blood pressure as well as for men and women. It is not certain if the relationship will be linear, but it will lead to a strong positive 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**.

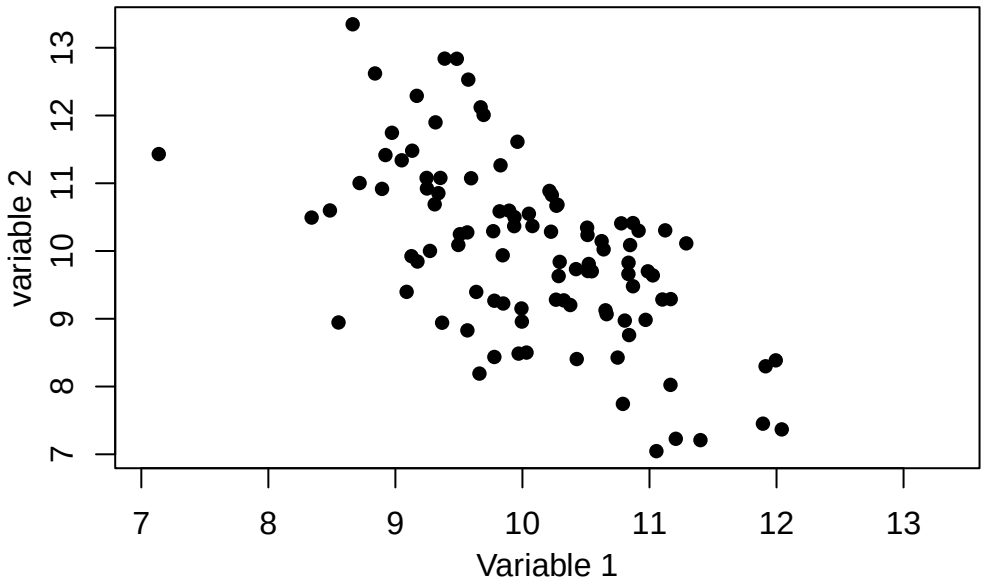
5. What is the most likely correlation for the data below?

- (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?

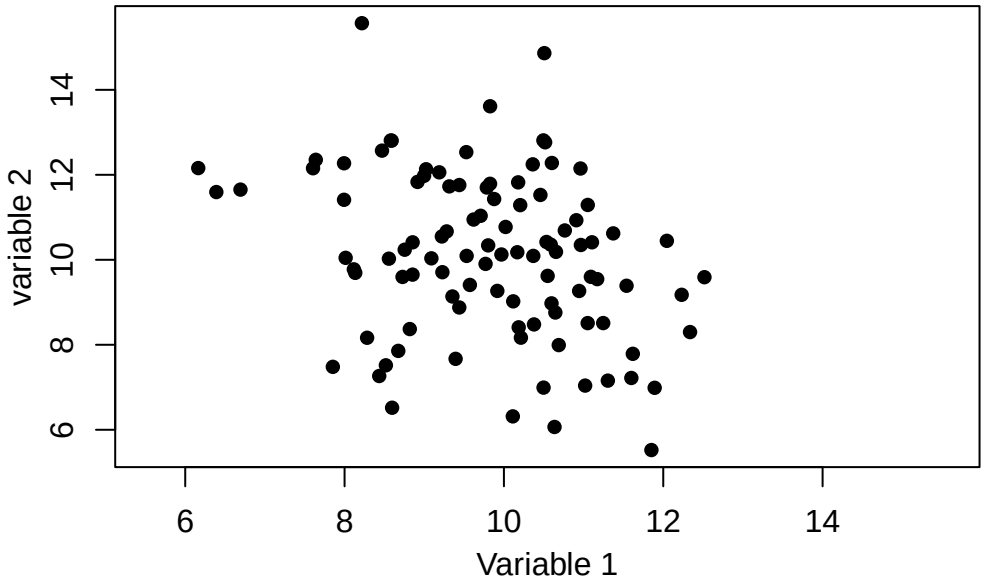
- (A) -1.0
- (B) -0.7
- (C) -0.3
- (D) 0.0
- (E) +0.3
- (F) +0.7
- (G) +1.0



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

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

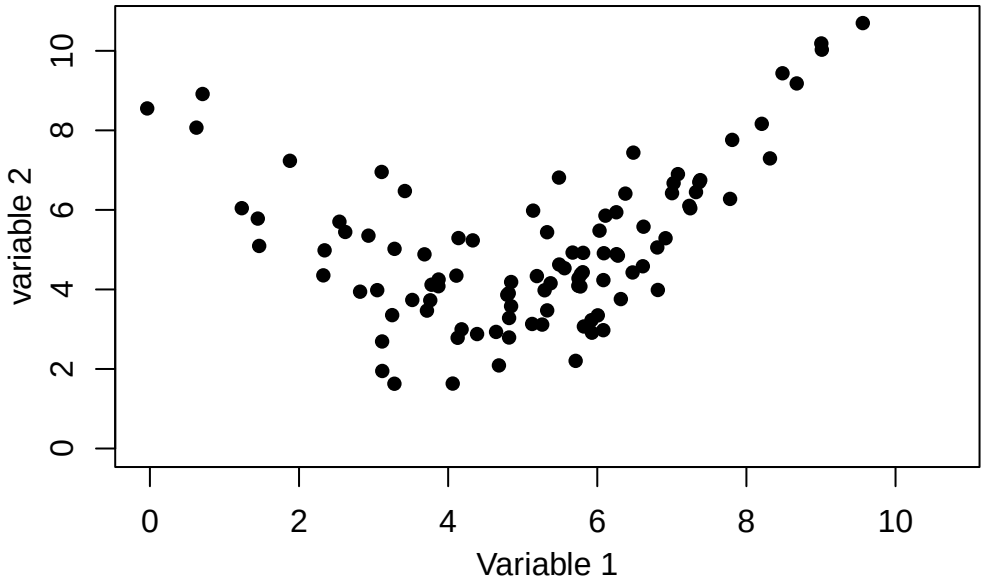
- (D) 0.0
- (E) +0.3
- (F) +0.7
- (G) +1.0



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

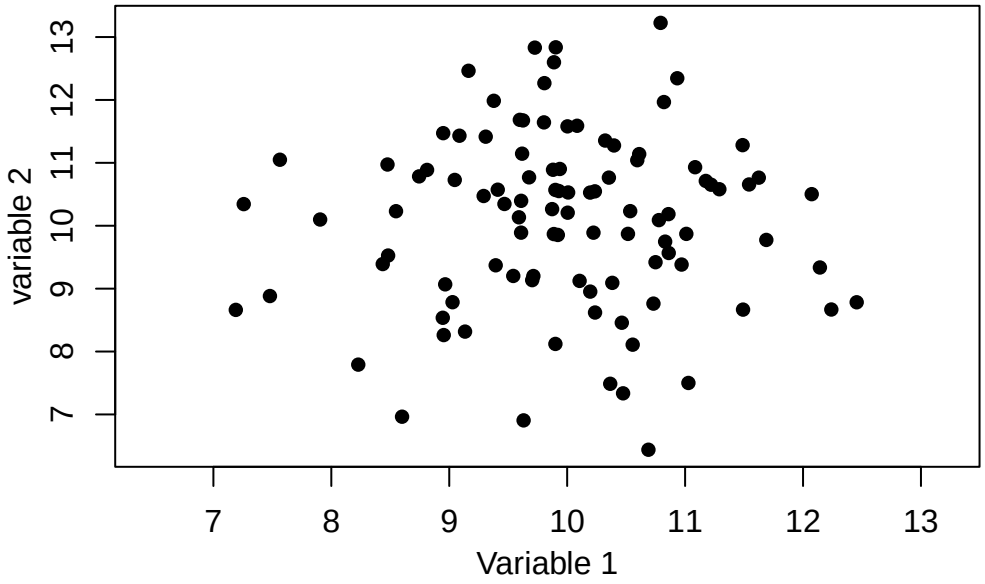
- (A) -1.0
- (B) -0.7
- (C) -0.3
- (D) 0.0
- (E) +0.3
- (F) +0.7

- (G) +1.0



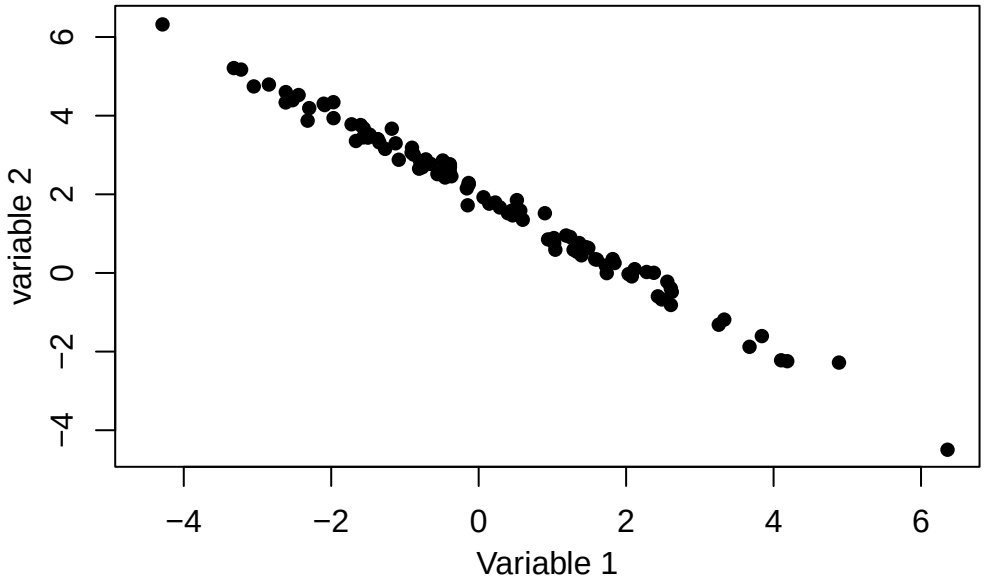
9. What is the most likely correlation for the data below?

- (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

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

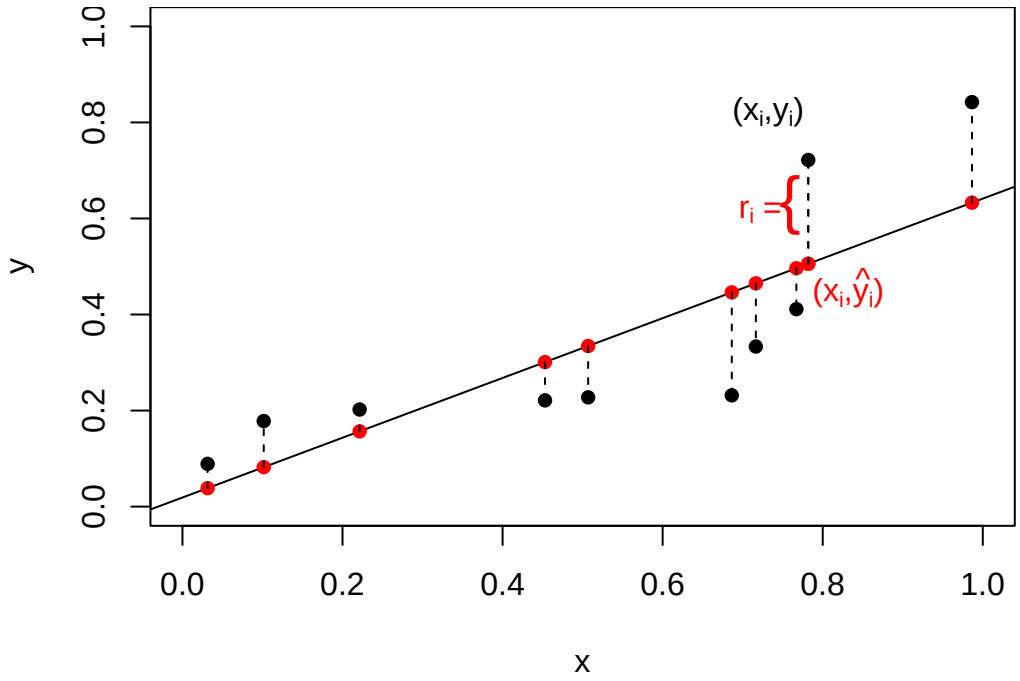


Figure 1.2: Diagram showing linear regression method

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.

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.

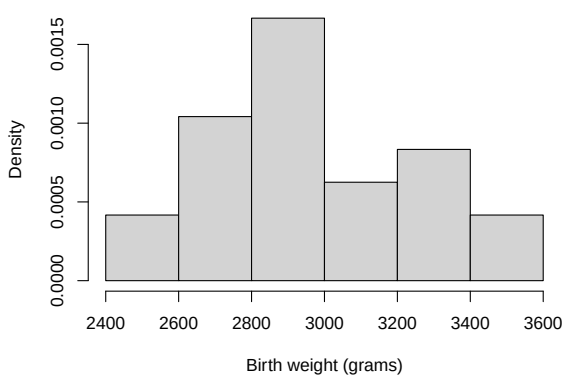
Focus on modelling quiz

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

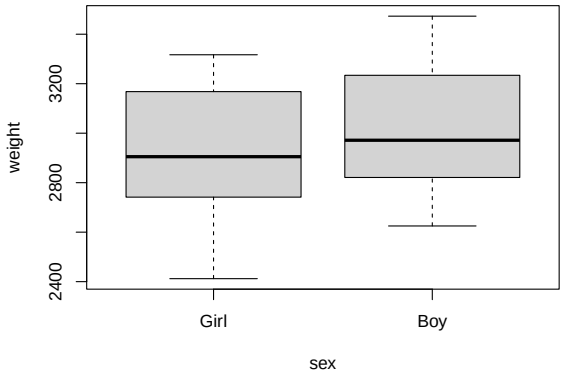
For each situation, choose one of the following statements which you think is **most likely** to apply.

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

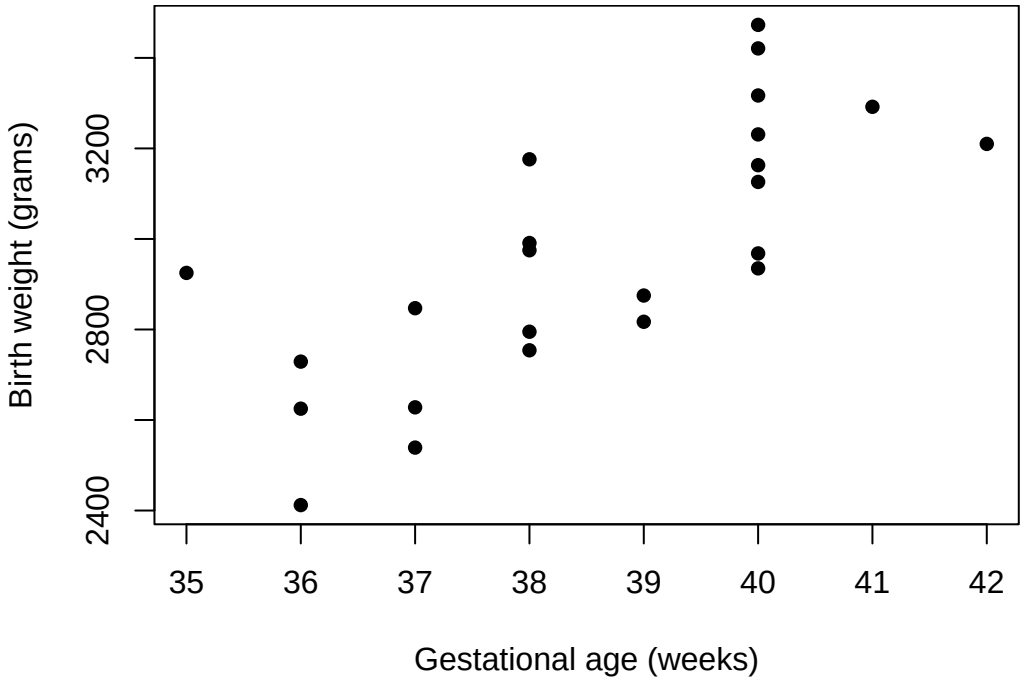
[1] Dobson and Barnett, 3rd edition, Table 2.3.



(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

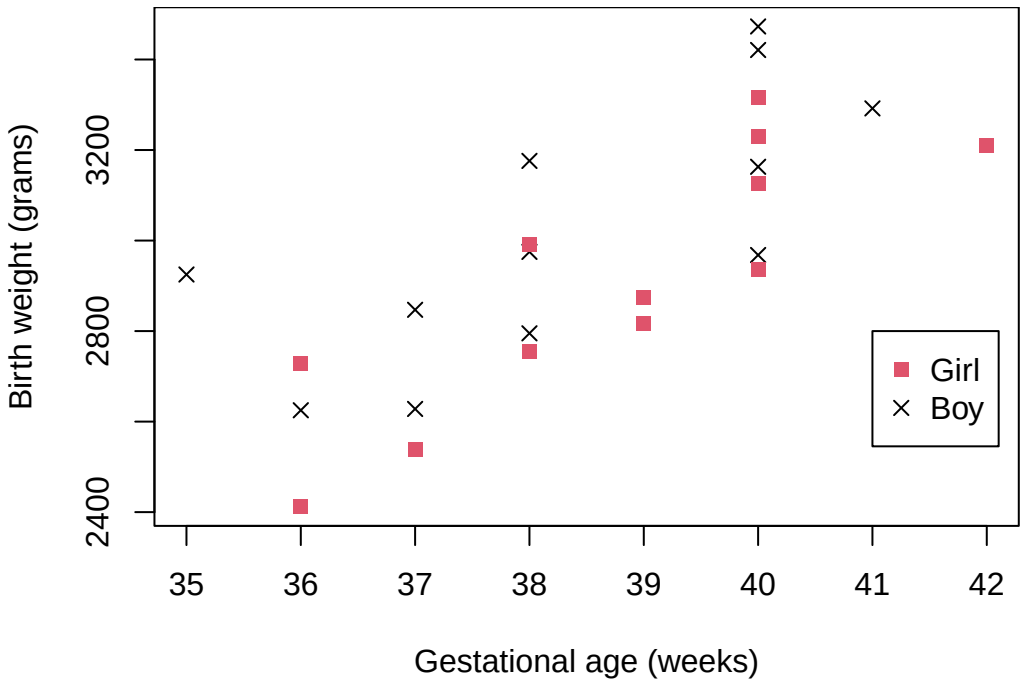


Figure 2.2: Birthweight and gestational age for 12 girls (red squares) and 12 boys (black crosses).

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

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

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

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

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

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

- (A) After fitting a regression model we should always consider residuals

- (B) A linear model can be fitted to any data set
- (C) A linear regression model will fit well if the correlation is close to zero
- (D) All regression models describe linear relationships
- (E) A linear regression can sometimes be used to approximate a non-linear relationship

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

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

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

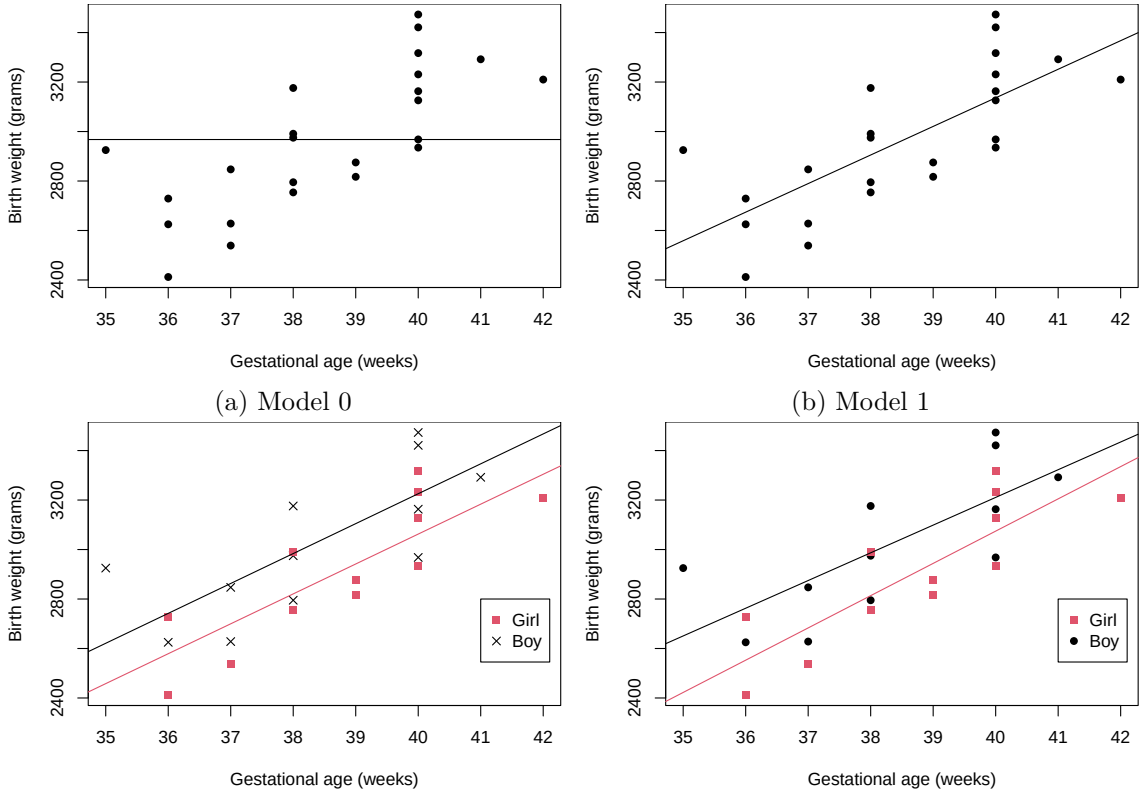
that `Model 2` is a special case of `Model 3` (consider the situation when $\delta = 0$) – such models are called *nested*.

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:

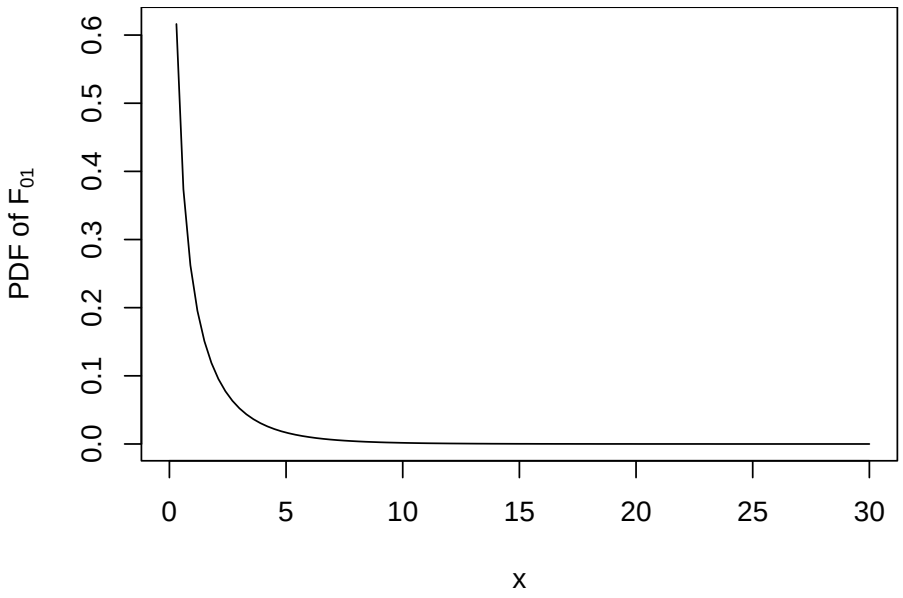


Figure 2.4: Probability density function of F_{01} distribution.

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

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

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

- (A) Can be used to judge how well the model fits the data

- (B) All should be zero if the model is a good fit
- (C) Can help to identify unusual data points
- (D) Should always be plotted as part of a data analysis
- (E) Will always sum to zero

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

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

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

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

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

- (A) Is the data reliable and have suitable data checks been performed
- (B) Do we need authorization from the person collecting the data before we can use it

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

- (D) Are the data representative of what we are aiming to investigate

- (E) Background information is not needed before analyzing a new data set – it's only the numbers that matter

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

2. What is the distribution of the error term?

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

3. Which quantity represents the model parameters?

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

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

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

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

- (A) X is singular
- (B) X is full rank
- (C) X can be zero
- (D) X is not square
- (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) factorial
- (B) as.factor
- (C) as.ordinal
- (D) as.nominal
- (E) as.numerical

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

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

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

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

Example of fitting a linear model in R

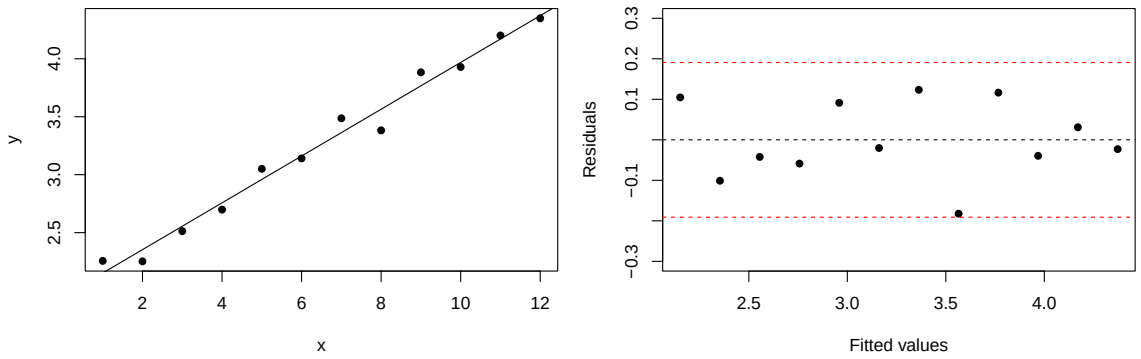
Here is a toy example of **R** commands for modelling a response in terms of one quantitative explanatory variable.

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

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.18236 -0.04632 -0.02162  0.09472  0.12353

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.949587   0.061636   31.63 2.35e-11 ***
x           0.201892   0.008375   24.11 3.43e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1001 on 10 degrees of freedom
Multiple R-squared:  0.9831,    Adjusted R-squared:  0.9814
F-statistic: 581.2 on 1 and 10 DF,  p-value: 3.433e-10
```



(a) Data and fitted model (b) Residual 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 resulting fitted linear regression describes the data well. A residual plot shows that all residuals are within two standard deviations of zero, again supporting that the model fits well.

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

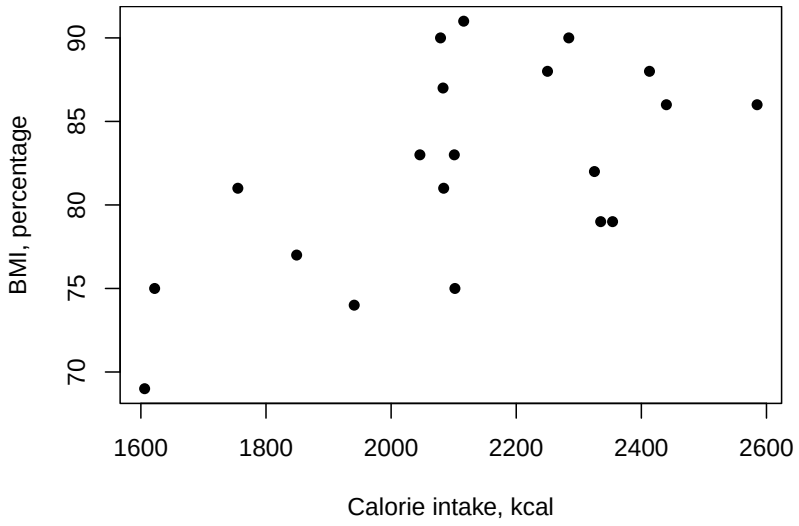
> **!** Important
>
> Unless otherwise stated, data files will be available online at: rgaykroyd.github.io/MATH3823/Datasets/*filename.ext*, where *filename.ext* is the stated filename with extension.

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.

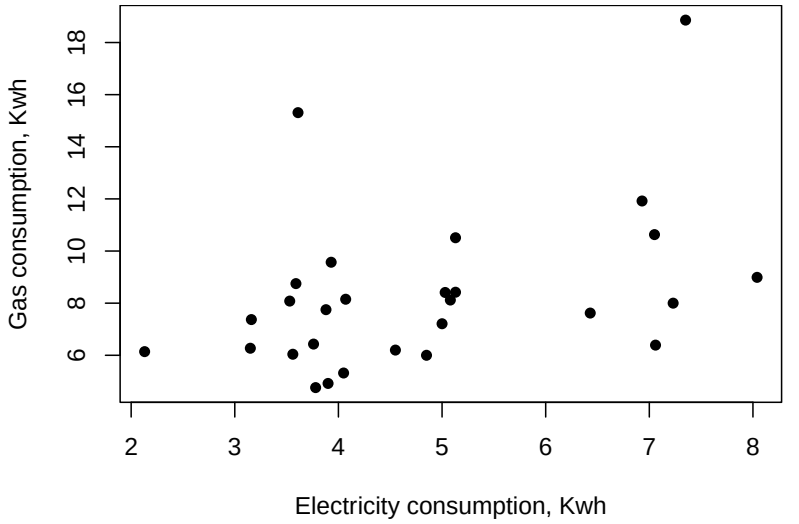
Click here to see hints.

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.

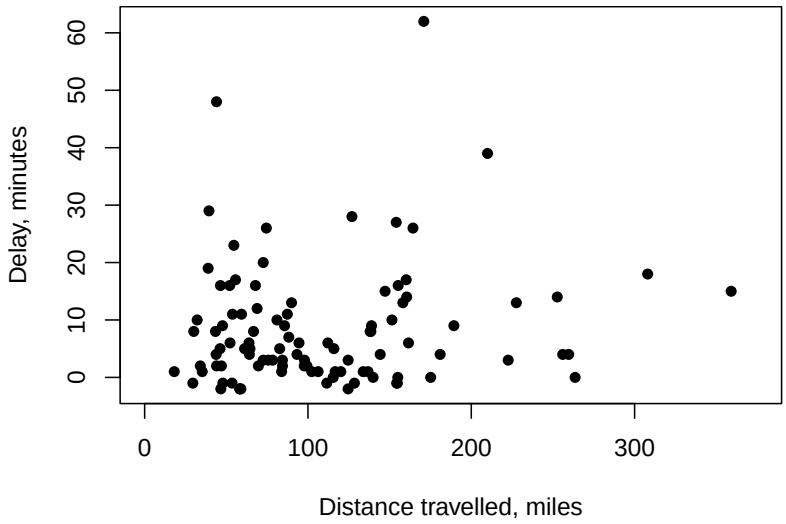
a. Childhood obesity is a serious medical condition which can lead to long-term health problems. A mixed-sex secondary school for ages 11-18 measures height and weight of all its new students (mostly aged 11 years old) on the first day of term and calculates their body mass index (BMI) and the average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



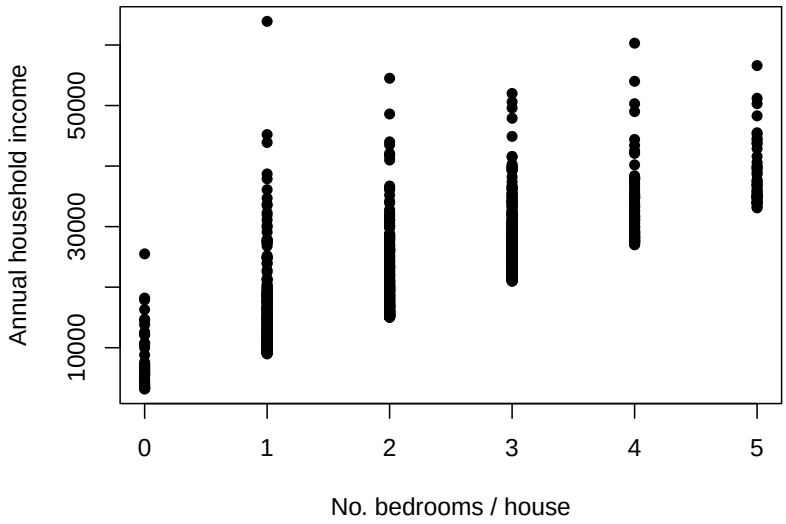
b. Domestic UK "smart meters" aim to record cumulative electricity and gas usage once per day. They use a wireless internet connection to transmit information to the energy providers, but they do not save previous values. The central system records readings as they are received and calculates the difference 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

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

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.

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.

[4]Source unknown.
[5]Source unknown.

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.