MATH3823 Generalized Linear Models

Robert G Aykroyd

3/2/23

Table of contents

Weekly schedule **3**

Overview **5**
Preface . . . 5

Official Module Description **6**
Module summary . . . 6
Objectives . . . 6
Syllabus . . . 6
University Module Catalogue . . . 6

1 Introduction **7**
1.1 Overview . . . 7
1.2 Motivating example . . . 8
1.3 Revision of least-squares estimation . . . 9
1.4 Types of variables . . . 11
1.5 Exercises . . . 11

2 Essentials of Normal Linear Models **13**
2.1 Overview . . . 13
2.2 Types of normal linear model . . . 20
2.3 Matrix representation of linear models . . . 21
2.4 Construction of the design matrix . . . 22
2.4.1 Example: Simple linear regression . . . 23
2.4.2 Example: One-way ANOVA . . . 23
2.5 Model shorthand notation . . . 24
2.6 Exercises . . . 25

3 GLM Theory **28**
3.1 Motivating examples . . . 28
3.2 The GLM structure . . . 29
3.3 The random part of a GLM . . . 30
3.4 Moments of exponential-family distributions . . . 32
3.5 The systematic part of the model . . . 34
3.6 The link function . . . 34
3.7 Exercises . . . 39

4 GLM Estimation **41**
4.1 The identically distributed case . . . 41
4.1.1 Maximum likelihood estimation . . . 41
4.1.2 Estimation accuracy . . . 42
4.2 The general case . . . 43
4.2.1 MLE Estimation . . . 43
4.2.2 The score function and Fisher information . . . 44
4.2.3 The saturated case . . . 46
4.3 Model deviance . . . 46
4.4 Model residuals . . . 47
4.5 Fitting generalized linear models in **R** . . . 48
4.5.1 GLM-related **R** commands . . . 48
4.5.2 Example of fitting Poisson GLM in **R** . . . 49
4.6 Exercises . . . 51

Weekly schedule

Week 5 (27 February - 3 March)

- **Before next Lecture:** Re-read *Chapter 4: Section 4.1.*
- **Lecture on Tuesday:** Cancelled due to illness. Please read *Chapter 4: Section 4.2.*
- **Lecture on Thursday:** *Chapter 4: Sections 4.3, 4.4 & 4.5.*
- **Weekly feedback:** Start Exercises in *Chapter 4.*

Week 4 (20 - 24 February)

- **Before next Lecture:** Be confident with all material up to, and including, *Section 3.4 Moments of exponential-family distributions.*
- **Lecture on Tuesday:** *Chapter 3: Sections 3.5 & 3.6*
- **Lecture on Thursday:** Start Chapter 4 by covering *Section 4.1.*
- **Weekly feedback:** Complete Exercises in *Chapter 3.*

Week 3 (13 - 17 February)

- **Before next Lecture:** Be confident with material in *Chapter 2: Essentials of Normal Linear Models.*
- **Lecture on Tuesday:** Cancelled due to UCU strike. Instead, self-study *Chapter 3: Sections 3.1 & 3.2.*
- **Lecture on Thursday:** Cancelled due to UCU strike. Instead, self-study *Chapter 3: Sections 3.3 & 3.4.*
- **Before next Lecture:** Complete questions and check solutions, including video(s), for all Exercises in *Chapters 1 and 2.* Start Exercises in *Chapter 3.*

Week 2 (6 - 10 February)

- **Before next Lecture:** Please re-read *Section 2.1: Overview* and read *Section 2.2: Types of normal linear model*.
- **Lecture on Tuesday:** We will briefly cover all remaining material in *Chapter 2: Essentials of Normal Linear Models*.
- **Before next Lecture:** Please re-read *Chapter 2* carefully.
- **Lecture on Thursday:** Cancelled due to UCU strike.
- **Weekly feedback:** Self-study the Exercises in *Section 2.6* – solutions to be posted during Week 3.

Week 1 (30 January - 3 February)

- **Before next Lecture:** Please read the *Overview*.
- **Lecture on Tuesday:** We will briefly cover all material in *Chapter 1: Introduction*.
- **Before next Lecture:** Please re-read *Chapter 1* carefully.
- **Lecture on Thursday:** Start *Chapter 2: Essentials of Normal Linear Models* with *Section 2.1: Overview*.
- **Weekly feedback:** Self-study the Exercises in *Section 1.5* – solutions to be posted during Week 1.

Coursework Practical Sessions (20 - 24 March)

- Coursework for this module involves a single written report worth 20% of the module grade. This will mainly involve investigating different models using R and interpreting the results.
- Tasks are expected to be handed out before 16 March with hand-in deadline expect to be after 28 March. Further details to follow in early March.

Overview

Preface

These lecture notes are produced for the University of Leeds module *MATH3823 - Generalized Linear Models* for the academic year 2022-23. Please note that this material also forms part of the module *MATH5824 - Generalized Linear and Additive Models*. They are based on those 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 is the first year, however, that they have been produced in accessible format and hence some errors might occur during this conversion process. For information, 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 have not previously used Quarto/RMarkdown nor Github and hence please be patient if there are hitches along the way.

RG Aykroyd, Leeds, November 22, 2022

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

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 in these notes 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*.

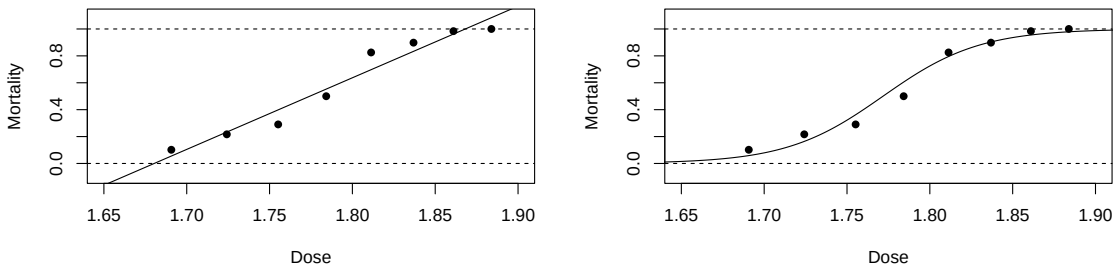
1.2 Motivating example

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.

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)

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

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.

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

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

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.

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. *[Hint: 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.

1.3 Starting from Equation 1.1, derive the estimation equations given in Equation 1.2. Further, show that $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of α and β. *[Hint: Check your MATH1712 lecture notes.]*

What can be said about $\hat{\sigma}^2$ as an estimator of σ^2? *[Hint: There is a careful theoretical proof, 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`. *[Hint: You already met this example in MATH1712.]*

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.

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.

Read the data file 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 and add the corresponding fitted linear models. Which variable do you think helps explain Reynolds number the best? Why do you think this?

Here are an infinite number of further numerical examples from **maths e.g.** (thanks to https://www.mathcentre.ac.uk/):
Finding the intersercept
Finding the slope - Part 1
Finding the slope - Part 2

[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.

2 Essentials of Normal Linear Models

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.

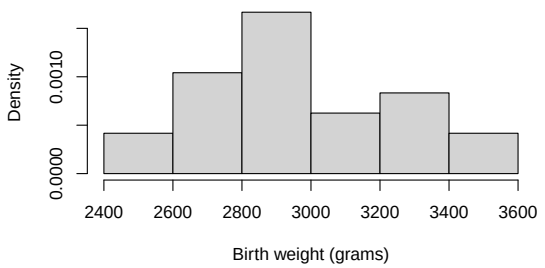
Before considering possible models, Figure 2.2 again shows the relationship between weight and age but this time with the points coloured according to the baby's sex. This, perhaps, shows the boys to have generally higher weights across the age range than girls.

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:`	$\texttt{Weight} = \alpha + \beta.\texttt{Age} + \gamma.\texttt{Sex} + \delta.\texttt{Age.Sex}$

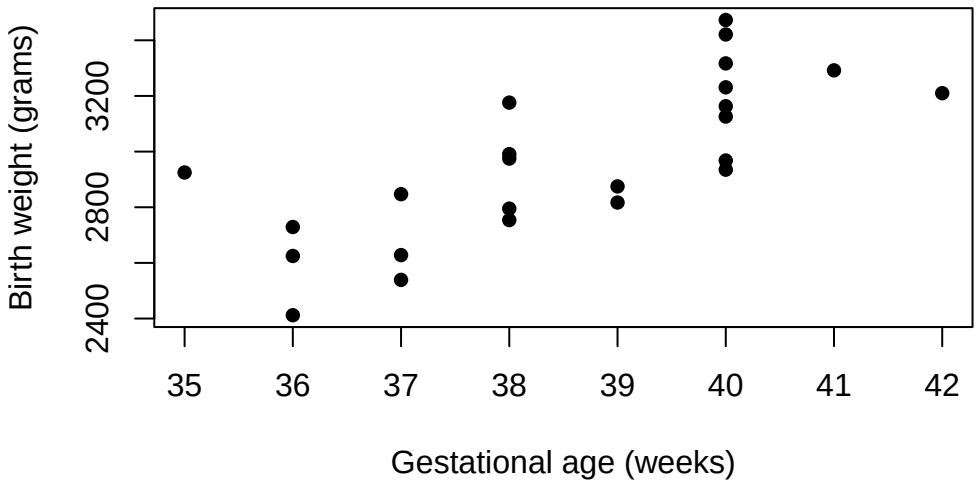
In these models, `Weight` is called the *response* variable (sometimes called the *dependent* variable) and `Age` and `Sex` are called the *covariates* or *explanatory* variables (sometimes called the *predictor* or *independent* variables). Here, `Age` is a continuous variable whereas `Sex` is coded 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.

[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

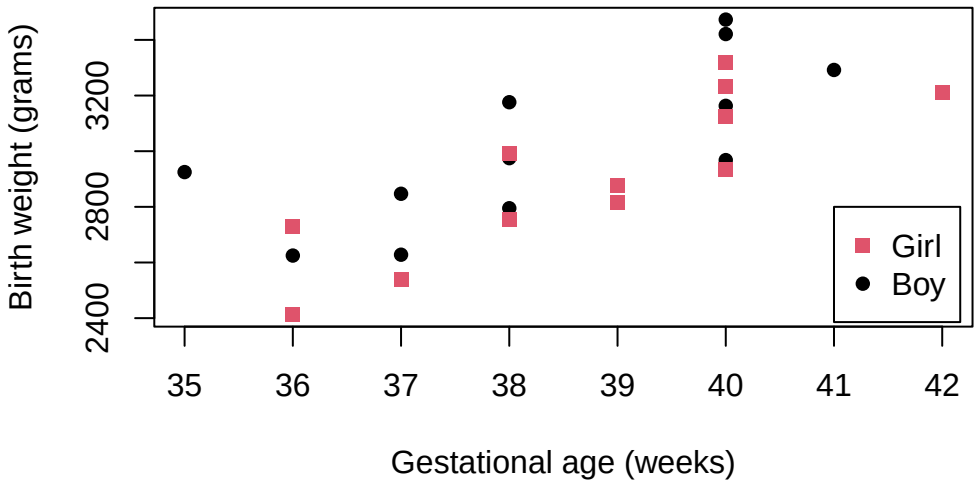


Figure 2.2: Birthweight and gestational age for 12 girls (red squares) and 12 boys (black dots).

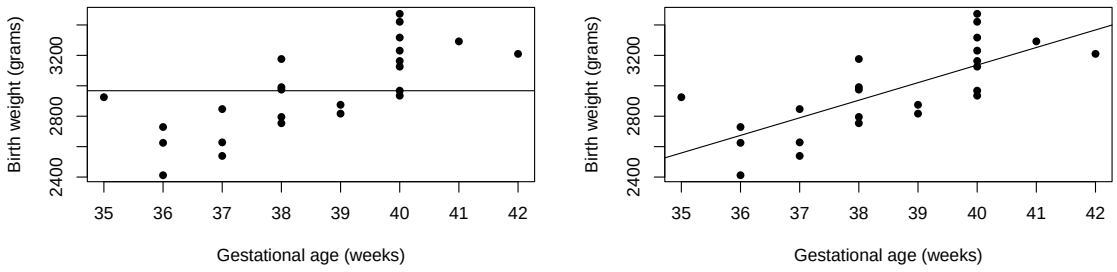
Note that `Model 0` is a special case of `Model 1` (consider the situation when $\beta = 0$) and that `Model 1` is a special case of `Model 2` (consider the situation when $\gamma = 0$) and finally that `Model 2` is a special case of `Model 3` (consider the situation when $\delta = 0$) – such models 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.

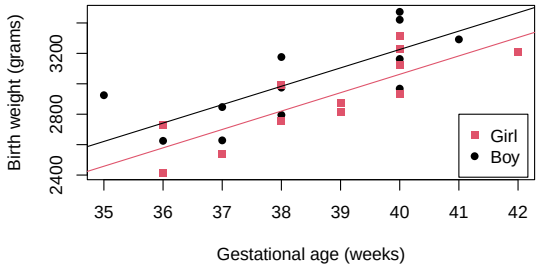
These models can be fitted to the data using (Ordinary) *Least Squares* to produce the results presented in Figure 2.3.

Which model should we use?

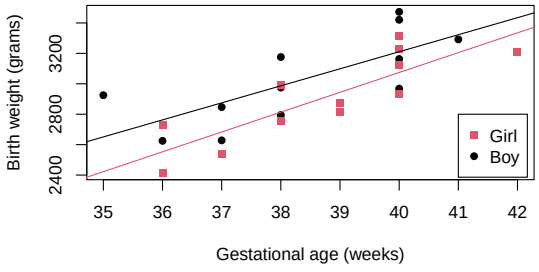


(a) Model 0

(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

$$F_{01} = \frac{113799/1}{816074/22} = 27.33,$$

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

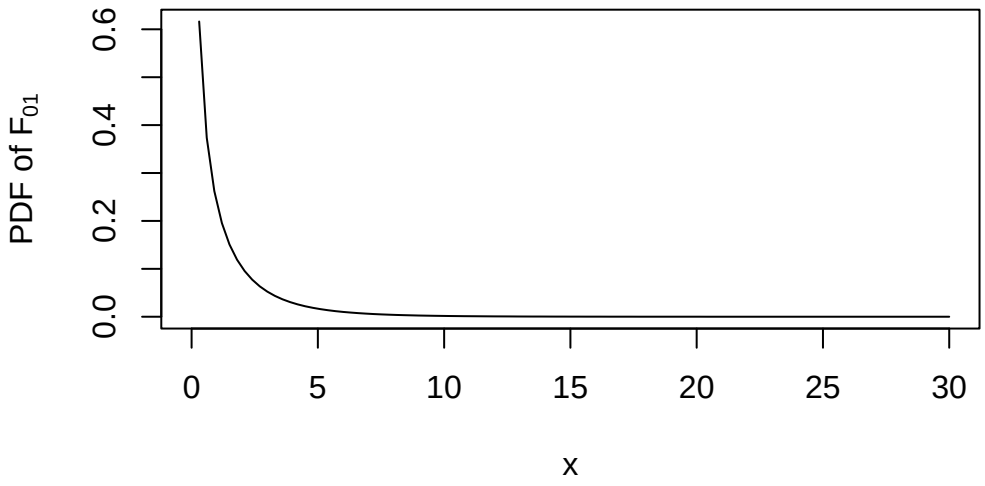


Figure 2.4: Probability density function of F_{01} distribution.

mean-square R_2/r_2. Thus our test statistic for comparing `Model 2` to the simpler `Model 1` is:

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

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: $\hat{\alpha} = -1773.3$, $\hat{\beta} = 120.9$ and $\hat{\gamma} = 163.0$, 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.

2.2 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 $x = (x_1, x_2, \ldots, x_p)$ 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$y = \alpha + \sum_{j=1}^{k} \delta_j\ I(x = j) + \epsilon$

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.

>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.

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.

2.3 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.

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

- 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.

2.4 Construction of the design matrix

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.

2.4.1 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:

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

2.4.2 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).

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. $\text{poly}(E;\ell)$: 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).

2.6 Exercises

> **!** Important
>
> Please note that these questions are currently unchecked. I aim to look at them very soon but a apologize is anything is incomplete or there are errors.

2.1. 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.

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?

2.2. 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?

2.3 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.

2.4 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

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

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.

3 GLM Theory

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

$$\Pr(y_i;\ p_i, m_i) = \binom{m_i}{y_i} p_i^{y_i}(1-p_i)^{m_i-y_i} \tag{3.1}$$

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

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 $\{0, 1, 2, \dots\}$.

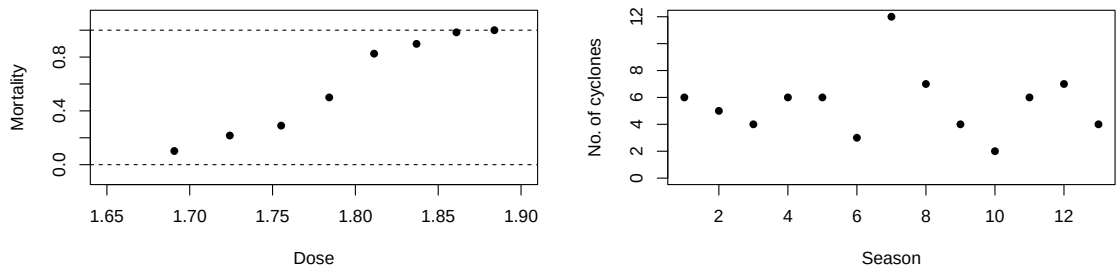
Table 3.1: Numbers of tropical cyclones in $n = 13$ successive seasons[1]

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 x_i or i.

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.

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



(a) Binomial model

(b) Poisson model

Figure 3.1: Examples of non-normally distributed data.

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*.

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.

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 $Y \sim \mathrm{Po}(\lambda)$, then Y takes values in $\{0, 1, 2, \dots\}$ 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 $Y_i \sim \mathrm{Po}(\lambda)$. The parameter would be simply estimated by the sample mean, $\widehat{\lambda} = \bar{y} = 5.5384615$.

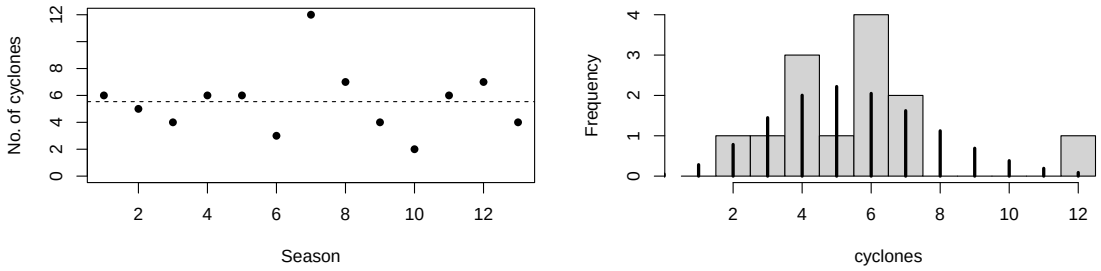
Example: Binomial distribution

Let Y have a Binomial distribution, that is $Y \sim \mathrm{Bin}(m, p)$, with m fixed. Then Y is discrete, taking values in $\{0, 1, \dots, m\}$, 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

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



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

Figure 3.2: Poisson model fitted to cyclone data.

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.

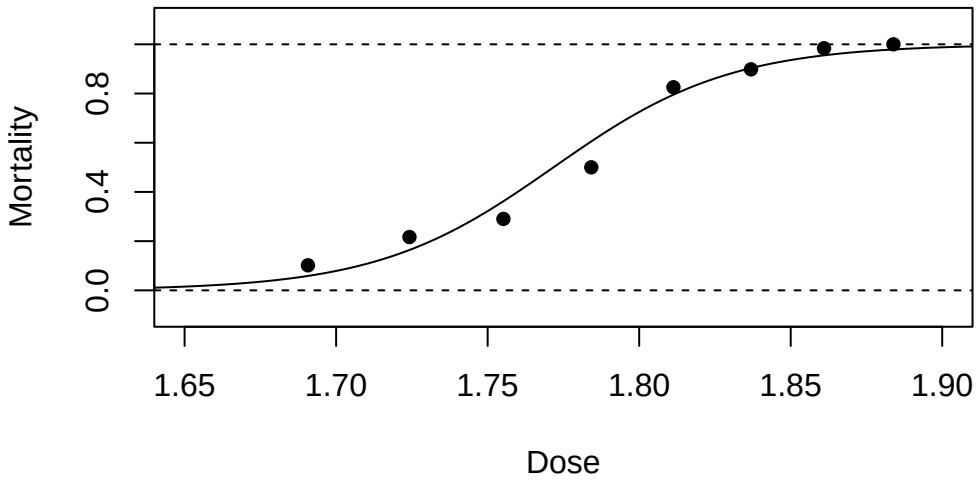


Figure 3.3: Binomial model fitted to beetle data.

Example: Normal distribution 32

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 2.2 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:

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

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,

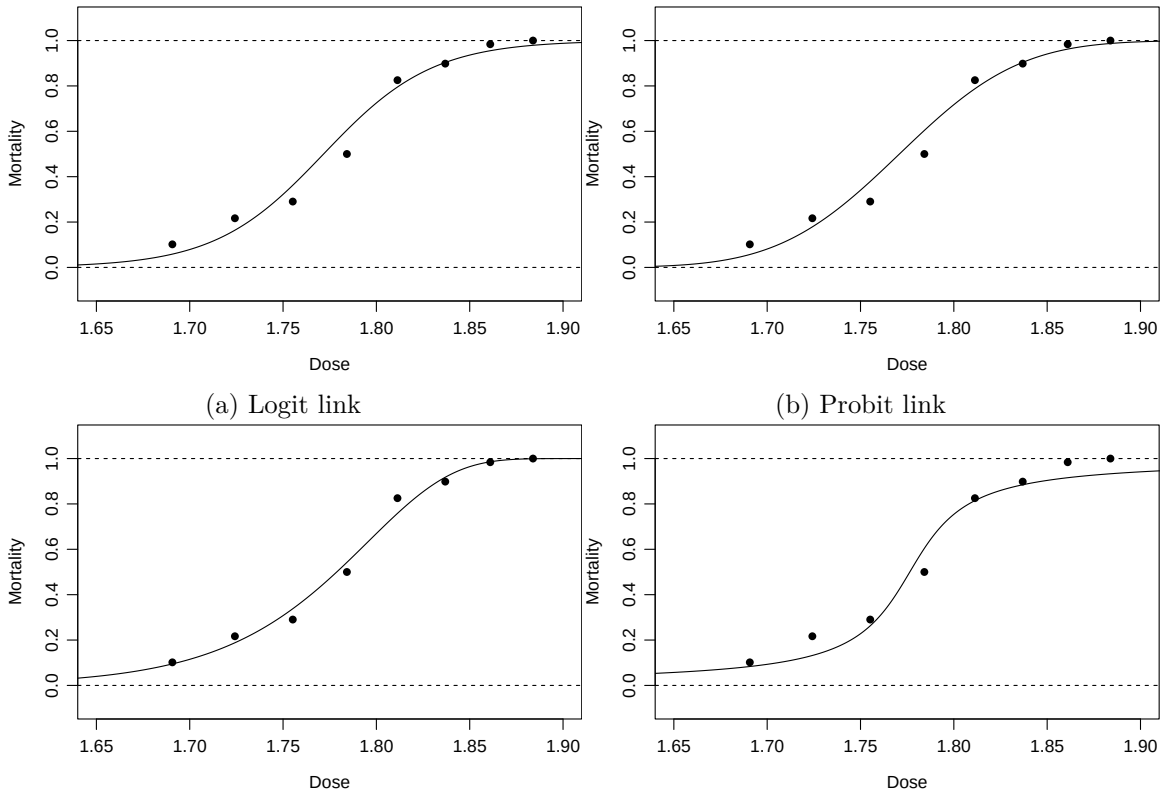
$$\mu = b'(\theta).$$

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:

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



(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.

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.

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. [*Hint: 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 $\theta = \text{logit } p = \log(p/(1 - p))$.

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

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

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

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, possibly with a scale parameter.

3.5 Prove Proposition 3.1 assuming that Y follows a discrete distribution. Verify the results for the Poisson in Table 3.5 and then derive similar results for the binomial, $Y \sim \text{Bin}(n, p)$. From these results, what are the mean and variance of Y.

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

[*Hint for the gamma, treat* $1/\alpha$ as a scale parameter and let θ be a suitable function of λ and α.]

3.7 Using Proposition 3.2, verify the canonical link function, $g(\mu)$, for the binomial and gamma distributions shown in Table 3.6.

4 GLM Estimation

> **Warning**
>
> Please note that to make these lecture notes available on time, there has not been sufficient time to check the following sections. This, in particular, may mean that some of the cross-referencing is not accurate, as well as small typos. Further, the final few sections are not yet complete.

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 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.

By independence, the joint distribution of all the observations $\mathbf{y} = \{y_i,\ i = 1, \ldots, n\}$ 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 them) 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 θ of interest 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 $y_i,\ i = 1, \ldots, n$, 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, and

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

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

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

[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.

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

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

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, Po(λ), 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

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

4.2 The general case

4.2.1 MLE Estimation

Suppose that now the n independent observations $\{y_i,\ i = 1, \ldots, n\}$ 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 $y_i, i = 1, \ldots, n$, given the parameters β, the likelihood function is:

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

and the log-likelihood by

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

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 the Newton–Raphson algorithm is usually adopted.

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

$$\widehat{\beta} = \left(X^T X\right)^{-1} X^T \mathbf{y}. \tag{4.11}$$

4.2.2 The score function and Fisher information

We define the *score function*:

$$U(\beta) = \frac{\partial l(\beta; \mathbf{y}, \phi)}{\partial \beta}, \tag{4.12}$$

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

$$I(\beta) = -\frac{\partial U(\beta)}{\partial \beta^T} = -\frac{\partial^2 l(\beta; \mathbf{y}, \phi)}{\partial \beta\, \partial \beta^T}, \tag{4.13}$$

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*:

$$J(\beta) = \mathrm{E}\left[-\frac{\partial^2 l(\beta; \mathbf{y}, \phi)}{\partial \beta\, \partial \beta^T}\right], \tag{4.14}$$

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

Proposition 4.1. *With definitions Equation 4.12, Equation 4.13, Equation 4.14 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

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

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

$$\hat{\beta} \sim N_p(\beta, J^{-1}(\beta)). \qquad (4.15)$$

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

From Equation 4.15, variances $\mathrm{Var}(\hat{\beta}_k)$, standard errors $\mathrm{se}(\hat{\beta}_k)$ and correlations between parameter estimates $\mathrm{Corr}(\hat{\beta}_k, \hat{\beta}_h)$ 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, $\mathrm{E}[Y_i] = \mu_i = b'(\theta_i)$ 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*.

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,

$$D = 2\phi\left\{l(\tilde{\theta}; \mathbf{y}, \phi) - l(\hat{\theta}; \mathbf{y}, \phi)\right\} \tag{4.16}$$

$$= \begin{cases} \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \text{Residual sum of squares} & \text{Normal} \\ 2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right\} & \text{Binomial} \\ 2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i + \hat{\mu}_i\right\} & \text{Poisson} \end{cases}$$

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:

$$\hat{\phi} = \frac{D_3}{n - r_3}. \tag{4.17}$$

Then test M_1 against M_2 using

$$\mathrm{F} = \frac{(D_1 - D_2)/(r_2 - r_1)}{\hat{\phi}} = \frac{(D_1 - D_2)/(r_2 - r_1)}{D_3/(n - r_3)} \tag{4.18}$$

with

$$\mathrm{F} \sim F_{r_2 - r_1, n - r_3}. \tag{4.19}$$

So if the observed value of the statistic Equation 4.18 was within the upper (say 5%) tail of the F-distribution Equation 4.19, 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

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

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

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

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

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 ~ 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 ~ 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 ~ 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.

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)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-1.8740  -0.6834   0.1287   0.7151   1.5956

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
```

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

Are these estimators unbiased for θ? What is the variance of the estimator in each case?

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 $\text{Var}[\hat{\mu}] = \sigma^2/n$?

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 4.11.

Further questions to be added later