MA50260 Lecture notes

Dr Ilaria Bussoli[*]

A.Y. 2023/2024, Semester II

Contents

[*] These notes are heavily based on notes from previous years by Drs Christian Rohrbeck and Deborshee Sen.

Overview of Statistical modelling

Content & Learning outcomes

Content:

This unit introduces the necessary theoretical background to understand and apply linear, generalised linear and mixed effect models. In details, we will talk about:

- Multiple linear regression models (LMs) – Inference techniques for the general linear regression model, diagnostics, transformation and variable selection.
- Generalised linear models (GLMs) – Exponential family of distributions and inference procedures. Focus on logistic regression and log-linear models.
- Generalised linear mixed effect models (GLMMs) – Hierarchical and grouped data, nested and crossed designs.

Learning outcomes:

After taking this unit, you should be able to:

- Choose an appropriate generalised linear mixed model for a given set of data.
- Fit this model, select terms for inclusion in the model and assess the adequacy of a selected model.
- Make inferences on the basis of a fitted model and recognise the assumptions underlying these inferences and possible limitations to their accuracy.

Organisation

Lecture notes

A set of comprehensive lecture notes are available through the unit Moodle page during the course of the semester. The notes are available in different formats (HTML, pdf, word and Epub) with identical content but different layouts. I will refer primarily to the HTML version.

In-person sessions

Lectures:

There are two weekly lectures on Fridays in CB 3.16, at 10:15am and 2:15pm.

Problem Classes:

There is only one problem class per week. It is on Tuesdays, at 9:15am in CB 5.6.

Recordings

Recordings of all sessions will be made available on Panopto and on the Moodle unit page.

Assessment

Summative assessment

Exam: 100% of unit mark (in May/June).

Resources

Q&As:

For any question,

- Feel free to put your hands up during sessions.
- Send me an email at ib641@bath.ac.uk, I usually answer

within the day.

- Use office hours, every Wednesdays from 3pm to 5pm, 2 South building, room 1.04A (how to reach 2 South). The library card is necessary to enter the building.
- Make use of the Moodle forum.

Extra textbooks:

This unit is self-contained in the sense that you will not strictly need to consult other textbooks. However, you may wish to consult extra books, available at the Library, to support your learning and understanding. A complete list of suggested works can be found on the Library List on Moodle, or in the References at the end of these lecture notes.

What if I have found an error in the lecture notes?

You can use the form below in case you find an error somewhere in the notes (it can happen!), by clicking on the link "Correcting typos and other errors".

Note: This form is visible only on the HTML version of the lecture notes.

A warning on question 4 - This question may be compiled in case of formulas, in the sense that, if explaining in which part a particular formula is wrong becomes difficult in words, you can take a screenshot of it and upload it there. In that case, the form is not anonymous any more.

Part I

Linear Regression

1 Introduction

In "Statistics for Data Science", you probably saw that statistical data science is based on the concepts of collecting, analysing and drawing inference from a sample of data. In "Statistical Modelling", we will focus on the basics of modelling data and introduce a wide range of statistical models.

We will start off with simple linear models and gradually introduce different forms of data that can be modelled, together with aspects of analysis and inference for more complex (and realistic) models. At the end of the course, you will be able to choose appropriate models for a wide range of applications, and be able to critically analyse their performance and validity.

1.1 Statistical Models and Parameter Estimation

A mathematical model is a mathematical representation of a real-world process, often taking the form of an equation (or set of equations) describing the relationships among several variables. Mathematical models are deterministic and do not allow for any uncertainty, for example, due to measurement errors.

On the other hand, a statistical model incorporates random variation in at least one of the quantities. In any statistical modelling problem, there is usually a variable of interest for which we want to build a model, pertaining to a subject-matter question. This variable of interest is known as the response variable (or the dependent variable). The statistical models introduced in this course seek to explain the variation in the response variable by a statistical model using explanatory variables which adequately describe the response.

A parametric statistical model for a set of data, $y_1, \dots, y_n$, consists of a common probability distribution (with cumulative distribution function F) with unknown parameters. First we must select an appropriate probability distribution; this will have one or more unknown parameters, denoted by θ, which we have to estimate. We denote an estimate using the 'hat' notation, i.e., $\hat{\theta}$ denotes the estimate for θ.

In addition to the definition of a statistical model and the estimation of the unknown parameters, we also consider aspects of:

- Model checking

  It is one thing to fit a model, but in order to trust any inferences that we might make from that model, we should ensure that it describes the observed data well, i.e., we should assess model fit: does the observed sample look like an independent random sample from the distribution F with parameter $\hat{\theta}$? A range of tools called diagnostics are available for this purpose. Which diagnostic(s) we choose will depend on the model fitted. Only then can we sensibly use the model to describe the behaviour of the population, test hypotheses or make predictions.

- Simplicity

  We want a model that is as simple as possible. The simpler it is, the easier it is to understand and draw inferences from. This is often known as parsimony. A parsimonious model gives better (less uncertain) predictions than one that is unnecessarily complicated.

Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.

— George Box

In summary, this course is about the skill and art of building models, or simplifying models, and of interpreting what they mean. The different aspects of the analysis are illustrated in the following pipeline in Figure 1.1:

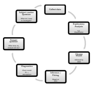

Figure 1.1: Statistical modeling pipeline.

1.2 Types of Response Variable

When we collect data, response/explanatory variables may be of different types. In this course, we use the following categorisation:

Table 1.1: Broad classification of variables.

Type	Description
Continuous	The variables can take any decimal value.
Count	The variable takes a positive whole number and represents a count of some phenomenon.
Categorical	The variable takes a positive whole number and represents quality or preference.
Binary	These are usually represented by 0 and 1 and could correspond to a yes/no response, or the presence or absence of some condition.

We start by modelling continuous data, and later we will consider categorical and count data. Different forms of models are needed for each of these types of response variable.

1.3 Statistical Background

1.3.1 Estimators and Estimates

Recall from “Statistics for Data Science” the definitions of estimator and estimate:

Definition 1.1 (Estimator).
An estimator $\hat{\theta}(\mathbf{Y})$ of θ, abbreviated as $\hat{\theta}_n$, is a function of the random variables $\mathbf{Y} = \{Y_1, \dots, Y_n\}$.

Definition 1.2 (Estimate).
An estimate $\hat{\theta}$ of θ is a function of the observed sample $y_1, \cdots, y_n$.

Remarks:

1. An estimator is itself a random variable.

2. An estimate is a number and a realisation of the estimator.

Sensible estimators for a population mean μ and variance σ^2 (e.g. using the method of moments) are, repsectively, the sample mean

$$\hat{\mu}_n \;=\; \hat{\mu}(Y_1, \ldots, Y_n) \;=\; \hat{\mu}(\mathbf{Y}) \;=\; \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

and the (unbiased[1]) sample variance

$$\hat{\sigma}^2_n \;=\; \hat{\sigma}^2(Y_1, \ldots, Y_n) \;=\; \hat{\sigma}^2(\mathbf{Y}) \;=\; \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2.$$

In “Statistics for Data Science”, you saw why these estimators were sensible. In particular, the following properties of estimators are important when assessing their usefulness.

[1] see Section 1.3.2

1.3.2 Properties of Estimators

The definition of an estimator makes no mention of a correspondence between the estimator and the parameter of interest. Of course, in reality we would like our estimators to be "good" in the sense that they generally give estimates that are "close" to the parameter we are interested in.

There are three properties of estimators that we are generally focused on:

1. Expectation, denoted as $\mathbb{E}\left[\widehat{\theta}(\mathbf{Y})\right]$ or $\mathbb{E}\left(\widehat{\theta}(\mathbf{Y})\right)$.

2. Variance, denoted as $\mathrm{Var}\left[\widehat{\theta}(\mathbf{Y})\right]$ or $\mathrm{Var}\left(\widehat{\theta}(\mathbf{Y})\right)$, and in particular we are interested in its behaviour as the sample size increases, i.e., as $n \to \infty$;

3. Sampling distribution of $\widehat{\theta}(\mathbf{Y})$.

1.3.2.1 Bias The main reason for our interest in the expectation of an estimator is to show that it is unbiased, i.e., it is accurate.

Definition 1.3 (Unbiased estimator).

An estimator is unbiased for θ if

$$\mathbb{E}\left[\hat{\theta}(\mathbf{Y})\right] = \theta,$$

where θ is the unknown true value of the parameter.

Thus, under the repeated sampling principle[2], an unbiased estimator will correctly estimate the true value of the parameter "on the average".

1.3.2.2 Variance In the same way that we can look at the expected value of an estimator, we can also examine its variance. We assume that $Y_1, \dots, Y_n$ constitutes an independent and identically distributed (i.i.d.) random sample from a population with mean μ and variance σ^2.

For a good estimator, its variance will decrease as the sample size n increases (i.e., tends to zero). In other words, as the sample size increases the between-sample variability in the estimates gets smaller and smaller. This is seen to be a desirable property, and is

[2] A statistical procedure should be evaluated on the basis of its behavior in hypothetical repetitions of the experiment that generated the original data.

related to the concept of consistency, i.e., the estimator should be precise.

Definition 1.4 (Consistent estimator).
An estimator of a parameter θ is consistent if it tends to the true value of the parameter as the sample size tends to infinity, i.e., as $n \to \infty$.

More formally, suppose we have a sample $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ and let θ be the true value of a parameter. Then we term $\widehat{\theta}(\mathbf{Y})$ a consistent estimator for θ if, for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\widehat{\theta}(\mathbf{Y}) - \theta\right| > \epsilon\right] = 0.$$

As a consequence of Chebyshev's inequality, an unbiased estimator is consistent if its variance tends to zero as $n \to \infty$. For example, $\widehat{\mu}(\mathbf{Y}) = \overline{Y}$ is a consistent estimator for μ.

Exercise 1.1.

Show that $\hat{\mu}_1(\mathbf{Y}) = \overline{Y}$ is an unbiased and consistent estimator for μ. Assume without loss of generality that $\overline{Y} = \{Y_1, \dots, Y_n\}$ is an i.i.d. random sample from a population with mean μ and variance σ^2.

Exercise 1.2.

Assume the same setting as Exercise 1.1 Show that $\hat{\mu}_2(\mathbf{Y}) = Y_1$ is an unbiased estimator for μ, but is not a consistent estimator for μ. Hint: How does the variance of $\hat{\mu}_1(\mathbf{Y})$ act as the sample size n increases?

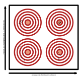

Figure 1.2: Relationships between accuracy and precision of an estimator.

1.3.2.3 Sampling Distribution So far, we have looked at the expectation and the variance of an estimator. In fact, it is useful to characterise the whole distribution of the estimator, over repeated samples, in order to quantify the uncertainty in $\hat{\theta}$. Sometimes this can be done exactly, but often it is stated approximately as the distribution can only be obtained as $n \rightarrow \infty$. Since an infinite sample size is never available, we have to assume that the sampling distribution is approximately true for large, but finite, sample sizes.

In general:

- Uncertainty occurs due to repeated sampling.

- We are using a sample of size n to tell us about a much larger population.

- Assuming that the sample was taken at random, there are many different samples of size n that we could have taken.

- How would our estimate change if we had a different sample?

- Suppose that we took 200 samples of size 10, what would the histograms of the 200 estimates for $\hat{\mu}(\mathbf{Y})$ and $\hat{\sigma}^2(\mathbf{Y})$ look like? In other words, what are the repeated sampling distributions of the estimators $\hat{\mu}(\mathbf{Y}) = \overline{Y}$ and $\hat{\sigma}^2(\mathbf{Y})$?

- What would happen to these distributions for $\hat{\mu}(\mathbf{Y})$ and

$\hat{\sigma}^2(\mathbf{Y})$ if the sample size was increased?

1.3.3 Maximum Likelihood Estimation

In "Statistics for Data Science" you were introduced to the method of moments estimator (MOME). Whilst it is often easy to compute, the estimators it yields can often be improved upon. For this reason, it has mostly been superseded by the method of maximum likelihood, which we will predominantly use in this course for estimating the parameters θ.

Definition 1.5 (Likelihood function).
Suppose we have data $y_1, \dots, y_n$, that arise from a population with probability mass (density) function (abbr., pmf or pdf) $f(\cdot)$, with an unknown parameter vector θ. Then the likelihood function of θ is the probability (density) of the observed data for given values of θ. If the data are i.i.d., the likelihood function is defined as

$$L(\theta \mid y_1, \dots, y_n) = \prod_{i=1}^{n} f(y_i \mid \theta).$$

The product arises because we assume independence, and joint probabilities (or densities) obey a product law under independence ("Statistics for Data Science").

Definition 1.6 (Maximum likelihood estimate).

For given data $y_1, \dots, y_n$, the maximum likelihood estimate (MLE) $\hat{\theta}$ is the value of θ that maximises the likelihood function $L(\theta \mid y_1, \dots, y_n)$.

The maximum likelihood estimator of θ based on the random variables $\mathbf{Y} = (Y_1, \dots, Y_n)$ will be denoted by $\hat{\theta}(\mathbf{Y})$.

So the MLE, $\hat{\theta}$, is the value of θ where the likelihood is the largest. This is intuitively sensible.

Note that it is often computationally more convenient to take logs of the likelihood function (in particular, the natural logarithm).

Definition 1.7 (Log-likelihood function).

The log-likelihood function is the natural logarithm of the likelihood function and, for i.i.d. data $y_1, \dots, y_n$, is denoted by

$$\ell(\theta \mid y_1, \dots, y_n) = \ln\left[L(\theta \mid y_1, \dots, y_n)\right] = \ln\left[\prod_{i=1}^{n} f(y_i \mid \theta)\right] = \sum_{i=1}^{n} \ln[f(y_i \mid$$

Remark: Since " ln " is a monotone increasing function, $\widehat{\theta}$ maximises the log-likelihood if and only if it maximises the likelihood; if we want to find a MLE we can choose which of these functions to try and maximise.

The reason why we will use maximum likelihood in this course is that the maximum likelihood estimator is asymptotically unbiased and consistent for a range of different models.

More specifically,

Theorem 1.1 (Asymptotic distribution of the MLE).
Let θ be a p-dimensional parameter vector. Then (subject to the likelihood function being smooth), as $n \to \infty$,

$$\widehat{\theta}(\mathbf{Y}) \sim \mathrm{MVN}_p\left(\theta, \mathcal{I}^{-1}(\theta)\right),$$

where

$$\mathcal{I}(\theta) \;=\; -\left[\mathbb{E}\left\{\frac{\partial^2 \ell(\theta \mid \mathbf{Y})}{\partial\theta_j\,\partial\theta_k}\right\}\right]_{j,k=1,\ldots,p}$$

is the Fisher information matrix.

This result gives us an idea about parameter uncertainty when using the MLE.

2 Linear Regression: Definition

2.1 Introduction

A linear regression model is a statistical model for linear relationships between variables. More precisely, it is a model for how the expected value (mean) of a response variable changes linearly with the value of one, or more, explanatory variables. It is useful for description, prediction or to improve understanding of a physical process.

As with all statistical models, it involves assumptions about the stochastic (or random) behaviour of the variables that we are modelling. Two models are often discussed in text books: simple linear regression and multiple linear regression. As we shall see, the former is a special case of the latter.

We start with an example to motivate the model.

Example 2.1 (Birth weights).
In a study on conception and foetus development, a group of newborn babies were weighed, and their (gestational) age recorded.

Figure 2.1 below shows a scatter plot of birth weight (grams) against gestational age (in weeks) at birth for 24 newborn babies. Gestational age shows time since conception.

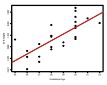

Figure 2.1: Birthweight (grams) against gestational age (weeks) for 24 children.

The straight line shows the estimated linear relationship between the two variables. We will look at the estimation of a linear regression model in Chapter 3. Such a linear regression model will also be useful to address various subject-matter questions, such as:

1. Is there evidence of a positive relationship between birth weight and gestational age? If so, what is this relationship?

2. Can we predict the birth weight for a child born at 34 weeks? What is a 95 % confidence interval for this prediction?

The answers to these questions relate to the intercept and slope of the fitted line, as well as the random variation in birth weights for a given gestation period.

2.2 Simple Linear Regression

In linear models we assume that some of the variability in the response variable could be explained by a linear relationship between this variable and a second variable. Simple linear regression provides a model for how the mean level of a response variable changes in response to changes in one explanatory variable.

Definition 2.1 (Response variable).
A response variable Y is a random variable whose distribution depends on the value of another variable.

Definition 2.2 (Explanatory variable).

An explanatory variable x is considered to be non-random and to influence the outcome of the response variable.

Definition 2.3 (Simple linear regression).

For observation (or individual) i, denote the response variable as Y_i and the explanatory variable as x_i. Then a simple linear regression model can be expressed as

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i, \qquad i = 1, \dots, n,$$

where n is the number of individuals in the study. The terms $\epsilon_1, \dots, \epsilon_n$ are termed regression residuals, and these are random variables.

We assume that

1. The residuals are mutually independent.

2. The residuals have zero mean and common variance σ^2, i.e., for each $i = 1, \dots, n$,

- $\mathbb{E}(\epsilon_i) = 0$,

- $\text{Var}(\epsilon_i) = \sigma^2$.

3. Each residual follows a normal distribution, i.e.,

$$\epsilon_i \sim \text{Normal}(0, \sigma^2), \qquad i = 1, \ldots, n.$$

Remarks:

1. The assumptions in Definition 2.3 can be combined to provide the following alternative definition of the simple linear regression model:

   a. $Y_i \sim \text{Normal}(\beta_1 + \beta_2 x_i,\, \sigma^2)$, for $i = 1, \ldots, n$,

   b. $Y_1, \ldots, Y_n$ are mutually independent.

2. In other words, the simple linear regression model allows the mean to vary across observations according to the value of

some explanatory variable x in the linear relationship

$$\mu_i = \mathbb{E}(Y_i) = \beta_1 + \beta_2 x_i.$$

Comments:

1. For a given data set, we observe values $y_1, \ldots, y_n$ of the response variable and associated values $x_1, \ldots, x_n$ of the explanatory variable. However, the following components of the model are unknown:

    a. The regression parameters, or coefficients, β_1 and β_2 ;

    b. The residual variance σ^2 ;

    c. The residuals $\epsilon_1, \ldots, \epsilon_n$.

2. The regression coefficients are interpreted as follows:

    a. The intercept, β_1 , is the expected value of the response variable when the explanatory variable is zero, i.e., $x = 0$.

b. The coefficient or slope β_2 is the expected change in the response variable for every unit increase in the explanatory variable.

Example 2.2 (Birth weights continued).
The estimated regression line for birth weight Y_i , conditional on gestational age x_i , is

$$\mathbb{E}(Y_i) = -1485 + 115.5x_i.$$

In other words, for every additional week of gestational age, the expected birth weight increases by about 116 grams.

Care should be taken when using a regression model to make predictions for Y outside the observed range of the explanatory variable x .

2.3 Multiple Linear Regression

Multiple linear regression follows exactly the same concept as simple linear regression, except that the expected value of the response variable may depend on more than one explanatory variable.

Example 2.3 (Birth weights continued).

As well as information on birth weight and gestational age, we know the sex-at-birth of each child. We can thus fit a separate linear relationship between birth weight and gestational age male and female newborns, which is illustrated in the following plot in Figure 2.2.

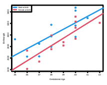

Figure 2.2: Birth weight (grams) against gestational age (weeks), split by sex-at-birth of the newborns. The red line shows the fitted linear regression when the infant is female, and the blue line correspond to the fitted linear regression model in case of a male infant.

Then, we can consider the following questions:

1. Do males and females gain weight at different rates?

2. Do we need both gestational age and sex-at-birth to explain variability in birth weights, or is one of these sufficient?

The multiple linear regression model has a very similar definition to the simple linear regression model, except that

1. Each individual has a single response variable Y_i, but a vector of explanatory variables ($x_{i,1}, \dots, x_{i,p}$). The number of explanatory variables is denoted by p.

2. There are p regression coefficients $\beta_1, \beta_2, \dots, \beta_p$, where β_j describes the effect of the j-th explanatory variable on the expected value of the response.

Definition 2.4 (Multiple linear regression model).
For $i = 1, \dots, n$, the multiple linear regression model takes the form

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_p x_{i,p} + \epsilon_i.$$

The residuals $\epsilon_1, \dots, \epsilon_n$ satisfy exactly the same assumptions as for the simple linear regression model. Hence, we assume that the residuals are independent and identically distributed, with a normal distribution, i.e., for $i = 1, \dots, n$,

$$\epsilon_i \sim \text{Normal}(0, \sigma^2).$$

An informal definition of the multiple linear model is

1. $Y_i \sim \text{Normal}\left(\sum_{j=1}^{p} \beta_j x_{i,j}\,,\, \sigma^2\right)$, for $i = 1, \dots, n$;

2. $Y_1, \dots, Y_n$ are independent.

Remark: To include an intercept term in the model, we set $x_{i,1} = 1$ for all $i = 1, \dots, n$, which gives

$$\mathbb{E}(Y_i) = \beta_1 + \sum_{j=2}^{p} \beta_j x_{i,j}.$$

The parameter β_1 is then termed the intercept.

2.3.1 Matrix Notation

Later in the course it will be useful to write our linear regression models using the following matrix notation.

Definition 2.5 (Response vector).

The response vector is $\mathbf{Y} = (Y_1, \dots, Y_n)$.

Definition 2.6 (Design matrix).

The design matrix $\mathbf{X}$ is an $n \times p$ matrix whose columns correspond to explanatory variables and whose rows correspond to subjects. That is, $X_{i,j}$ denotes the value of the j -th explanatory variable for individual i . If an intercept term is included, then the first column of X is a column of 1 's.

Definition 2.7 (Residual vector).

The residual vector is defined as $\underline{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$.

Definition 2.8 (Vector of coefficients).

The $p \times 1$ vector of coefficients is $\underline{\beta} = (\beta_1, \dots, \beta_p)$.

Then we can write the multiple linear regression model as follows, either

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \underline{\epsilon},$$

where $\underline{\epsilon} \sim \text{MVN}_n\left(0, \sigma^2 \mathbf{I}_n\right)$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix, or

$$\mathbf{Y} \sim \text{MVN}_n\left(\mathbf{X}\underline{\beta}\,,\, \sigma^2 \mathbf{I}_n\right).$$

Remark: Because of the normality assumption on the residuals, the full title of the model is the normal linear regression model. For the remainder of the course we will not distinguish between simple and multiple linear regression, since the former is a special case of the latter. Instead we refer to it as linear regression.

2.3.2 Response Variables

From the informal definition of the linear regression model, the response variable Y_i should be continuous and follow a normal distribution. It only makes sense to apply a linear regression model to data for which the response variable satisfies these criteria.

The assumption of normality can be verified with a normal QQ plot. If the response is continuous and non-normal, a transformation may be used to transform to normality. For example, we might take the logarithm or square root of the responses, as we will see in Chapter 5. Part 2 of this course will introduce the concept of the generalized linear model, in which the normality assumption is relaxed to cover a wider family of distributions, including the Poisson and binomial distributions.

2.3.3 Factors

Explanatory variables may be continuous or discrete, qualitative or quantitative.

Definition 2.9.

In this course we discuss two types of explanatory variable.

1. A covariate is a quantitative explanatory variable.

2. A factor is a qualitative explanatory variable.

The possible values for the factor are called levels. For example, sex-at-birth is a two-level factor with two levels: male and female.

Factors are represented by indicator variables in a linear regression model. For a p-level factor, p indicator variables are created.

For observation i, the indicator variable for level j takes the value 1 if that observation has level j of the factor; otherwise it takes the value zero.

To include sex-at-birth as an explanatory variable for example, we create two indicator variables $x_{i,1}$, to show whether individual i is male, and $x_{i,2}$, to show whether individual i is female. Then

$$x_{i,1} = \begin{cases} 1 & \text{if individual } i \text{ is male} \\ 0 & \text{if individual } i \text{ is female} \end{cases}$$

and

$$x_{i,2} = \begin{cases} 1 & \text{if individual } i \text{ is female} \\ 0 & \text{if individual } i \text{ is male} \end{cases}$$

2.4 Examples

Example 2.4 (Birth weights continued).

The response variable Y_i is birth weight. There are two explanatory variables, sex-at-birth (factor) and gestational age (continuous). Let $x_{i,1}$ and $x_{i,2}$ be indicator variables for male and female newborns respectively, and $x_{i,3}$ be gestational age.

One possible model is

$$\text{Model 1:} \quad \mathbb{E}(Y_i) = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}.$$

This assumes a different intercept for males and females, but a common slope for gestational age. It does not include an overall intercept term - we will see later why this is. The design matrix for this model has three columns; the first is the indicator for male newborns, the second is the indicator column for female newborns and the third contains the gestational age.

Example 2.5 (Birth weights continued).
A second possible model has a common intercept, but allows for separate slopes for male and female newborns; this is an interaction between sex-at-birth and gestational age of mothers.

$$\text{Model 2:} \quad \mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1} x_{i,3} + \beta_3 x_{i,2} x_{i,3}.$$

The design matrix for this model has three columns. The first is a column of 1's for the intercept. The second is the product of the indicator variable for being born male and gestational age of the mother. The third is the product of the indicator variable for being born female and gestational age of the mother.

Example 2.6 (Birth weights continued).
A third possible model, combining the first two, includes separate intercepts and separate slopes for the two sexes at birth,

$$\text{Model 3}: \quad \mathbb{E}(Y_i) = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1} x_{i,3} + \beta_4 x_{i,2} x_{i,3}.$$

All three model fits are shown below in Figure 2.3.

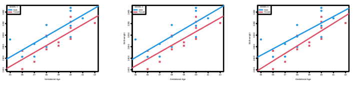

Figure 2.3: Birth weight (grams) against gestational age (weeks), split by sex-at-birth: three different models.

Remark: How can we choose which of models fits the data best? Intuitively the second model seems sensible - all babies start at the same weight, but sex-at-birth may affect the rate of growth. However, since our data only covers births from 35 weeks gestation onward, we should only think about the model which best reflects growth during this period. We will look at issues of model selection later.

Example 2.4 shows one way of including factors in linear models. In Example 2.4, indicator variables for all factors are included, but there is no intercept. Alternatively, we could include an intercept term, but indicator variables for only one of the two levels of the factor. In general, we include an intercept term and indicator variables for $p-1$ levels of a p-level factor. This ensures that the columns of the design matrix $\mathbf{X}$ are linearly independent - even if we include two or more factors in the model. For interpretation, one level of the factor is set as a 'baseline' and the regression coefficients for the remaining levels of the factor can be used to report the additional effect of the remaining levels on top of the baseline.

3 Linear Regression: Estimation

The linear regression model has three unknowns:

- The regression coefficients $\underline{\beta} = (\beta_1, \dots, \beta_p)$.
- The residual variance σ^2 .

- The residuals $\underline{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$.

In this section we will look at how each of these components can be estimated from a sample of data.

3.1 Estimation of Regression Coefficients

We shall use the method of least squares to estimate the vector of regression coefficients $\underline{\beta}$.

For a linear regression model, this approach to parameter estimation gives the same parameter estimates as the method of maximum likelihood, which will be discussed later when we look at generalized linear models.

The basic idea of least squares estimation is to find the estimate $\hat{\underline{\beta}}$ which minimises the sum of squares function

$$S\left(\underline{\beta}\right) = \sum_{i=1}^{n} \left(y_i - \beta_1 x_{i,1} - \quad - \beta_p x_{i,p}\right)^2 . \tag{3.1}$$

We can rewrite the linear regression model in terms of the residuals as

$$\epsilon_i = Y_i - \beta_1 x_{i,1} - \quad - \beta_p x_{i,p}.$$

By replacing Y_i with y_i , $S(\underline{\beta})$ can be interpreted as the sum of squares of the observed residuals.

In general, the sum of squares function $S(\underline{\beta})$ is a function of p unknown parameters, $\beta_1, \ldots, \beta_p$. To find the parameter values which minimise the function, we calculate all p first-order derivatives, set these derivatives equal to zero and solve simultaneously.

Using the expression for $S(\underline{\beta})$ above, the j -th first-order derivative is

$$\frac{\partial S\left(\underline{\beta}\right)}{\partial \beta_j} = -2 \sum_{i=1}^{n} x_{i,j} \left(y_i - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p}\right). \qquad (3.2)$$

We could solve the resulting system of p equations by hand, e.g., using substitution. Since this is time consuming we instead rewrite our equations using matrix notation. The j -th first-order derivative corresponds to the j -th element of the vector

$$-2\mathbf{X}^T\left(\mathbf{y}-\mathbf{X}\underline{\beta}\right).$$

Thus to find $\hat{\underline{\beta}}$ we must solve the equation,

$$-2\mathbf{X}^T\left(\mathbf{y}-\mathbf{X}\hat{\underline{\beta}}\right)=\mathbf{0}.$$

Multiplying out the brackets gives

$$-2\mathbf{X}^T\mathbf{y}+2\mathbf{X}^T\mathbf{X}\hat{\underline{\beta}}=\mathbf{0},$$

which can be rearranged to

$$\mathbf{X}^T\mathbf{X}\hat{\underline{\beta}}=\mathbf{X}^T\mathbf{y}.$$

Multiplying both sides by $(\mathbf{X}^T\mathbf{X})^{-1}$ gives the least squares estimate $\hat{\underline{\beta}}$ for $\underline{\beta}$,

$$\hat{\underline{\beta}}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

This is one of the most important results of the course!

3.2 Remarks on the Least Square Estimate

3.2.1 Dependence of Explanatory Variables

In order for the least squares estimate to exist, $(\mathbf{X}^T\mathbf{X})^{-1}$ must exist. In other words, the $p \times p$ matrix $\mathbf{X}^T\mathbf{X}$ must be non-singular:

- $\mathbf{X}^T\mathbf{X}$ is non-singular if and only if it has linearly independent columns (i.e., orthogonal columns).
- This occurs if and only if $\mathbf{X}$ has linearly independent columns.
- Consequently, explanatory variables must be linearly independent.

This relates back to the discussion on factors in Section 2.3.

Linear dependence occurs if

- An intercept term and the indicator variables for all levels of a factor are included in the model, since the columns representing the indicator variables sum to the column of 1's.

- The indicator variables for all levels of two or more factors are included in a model, since the columns representing the indicator variables sum to the column of 1's for each factor.

Consequently it is safest to include an intercept term and indicator variables for $p-1$ levels of each p-level factor.

3.2.2 Matrix Notation

If you want to bypass completely the summation notation used in the derivation of the least squares estimate, the sum of squares function $S(\underline{\beta})$ can be written as

$$S(\underline{\beta}) = \left(\mathbf{y} - \mathbf{X}\underline{\beta}\right)^T \left(\mathbf{y} - \mathbf{X}\underline{\beta}\right) = \mathbf{y}^T\mathbf{y} - \underline{\beta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\underline{\beta} + \underline{\beta}^T\mathbf{X}^T\mathbf{X}\underline{\beta}. \tag{3.3}$$

Now $\underline{\beta}^T\mathbf{X}^T\mathbf{y} = (\mathbf{y}^T\mathbf{X}\underline{\beta})^T$ and since both $\underline{\beta}^T\mathbf{X}^T\mathbf{y}$ and $\mathbf{y}^T\mathbf{X}\underline{\beta}$ are scalars, we have that

$$\underline{\beta}^T\mathbf{X}^T\mathbf{y} = \mathbf{y}^T\mathbf{X}\underline{\beta}.$$

Hence,

$$S\left(\underline{\beta}\right) = \mathbf{y}^T\mathbf{y} - 2\underline{\beta}^T\mathbf{X}^T\mathbf{y} + \underline{\beta}^T\mathbf{X}^T\mathbf{X}\underline{\beta}.$$

Differentiating with respect to $\underline{\beta}$ gives the vector of first-order derivatives

$$-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\underline{\beta} = -2\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\underline{\beta}\right)$$

as before.

3.2.3 Checking the Second-order Conditions

To prove that $\hat{\underline{\beta}}$ minimises the sum of squares function we must check that the matrix of second derivatives is positive definite at $\hat{\underline{\beta}}$. This is the multi-dimensional analogue to checking that the second derivative is positive at the minimum of a function in one unknown.

Returning once more to summation notation, we have

$$\frac{\partial^2 S(\underline{\beta})}{\partial \beta_k \, \partial \beta_j} = 2\sum_{i=1}^{n} x_{i,j}x_{i,k}.$$

This is the (j,k)-th element of the matrix $\mathbf{X}^T\mathbf{X}$. Thus the second derivative of $S(\underline{\beta})$ is $\mathbf{X}^T\mathbf{X}$.

To prove that $\mathbf{X}^T\mathbf{X}$ is positive definite, we must show that $\mathbf{z}^T\mathbf{X}^T\mathbf{X}\mathbf{z} > 0$ for all non-zero vectors $\mathbf{z}$.

Since $\mathbf{z}^T\mathbf{X}^T\mathbf{X}\mathbf{z}$ can be written as the product of a vector and its transpose, $(\mathbf{X}\mathbf{z})^T\mathbf{X}\mathbf{z}$, the result follows immediately.

3.3 Examples

3.3.1 Birth weight

We return to the birth weight data in Chapter 2. We will fit the simple linear regression for birth weight Y_i with gestational age x_i as explanatory variable,

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_i.$$

The response vector and design matrix are

$$\mathbf{y} = \begin{bmatrix} 2968 \\ 2795 \\ 3163 \\ 2925 \\ \vdots \\ 2875 \\ 3231 \end{bmatrix} \qquad \text{and} \qquad \mathbf{X} = \begin{bmatrix} 1 & 40 \\ 1 & 38 \\ 1 & 40 \\ 1 & 35 \\ \vdots & \\ 1 & 39 \\ 1 & 40 \end{bmatrix}.$$

To find the estimate $\underline{\hat{\beta}}$ we use the formula

$$\underline{\hat{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

From above, we calculate

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 24 & 925 \\ 925 & 35727 \end{bmatrix},$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 19.6 & -0.507 \\ & \\ -0.507 & 0.0132 \end{bmatrix},$$

and

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 71224 \\ \\ 2753867 \end{bmatrix}.$$

Therefore,

$$\underline{\widehat{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} -1485 \\ \\ 115.5 \end{bmatrix}.$$

The fitted model for birth weight, given gestational age at birth is thus

$$\mathbb{E}(Y_i) = -1485 + 115.5x_i$$

We can interpret this as follows:

- For every additional week of gestation, expected birth weight increases by about 116 grams.

- If a child was born at zero weeks of gestation, their birth weight would be -1485 grams.

Because the matrices involved can be quite large, whether due to a large sample size n, a large number p of explanatory variables, or both, it is useful to be able to calculate parameter estimates using computer software. In **R**, we can do this "by hand" (treating **R** as a calculator), or we can make use of the function **lm** which will carry out the entire model fit – See MA50258 Applied Statistics.

Let's illustrate how to calculate $\underline{\hat{\beta}}$ "by hand" for the birth weight example. First we load the data set **bwt** into **R**,

```r
load("images/bwt0.Rdata")
## dim(bwt)
```

This tells us that there are 24 individuals and 3 variables. The variables are,

```r
names(bwt)
```

```
## [1] "Age"          "Weight"      "SexAtBirth"
```

To fit the simple linear regression of the previous example "by hand",

1. Set up the design matrix:

```
n = dim(bwt)[1] # or n = nrow(bwt)
X = matrix(cbind(rep(1, n), bwt$Age), ncol = 2)
```

2. Calculate $\underline{\hat{\beta}}$:

```
beta = solve( t(X) %*% X ) %*% t(X) %*% bwt$Weight
```

3. View results:

```
beta
```

```
##          [,1]
## [1,] -1485.0
## [2,]   115.5
```

3.3.2 Gas Consumption

The **gas** data set investigates the relationship between gas consumption and external temperature. To measure the effect of changes in the external temperature on gas consumption, we fit a multiple linear regression model. We will allow a different relationship between gas consumption and outside temperature before and after the installation of cavity wall insulation.

The model has four regression coefficients

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1} + \beta_3 x_{i,2} + \beta_4 x_{i,1} x_{i,2} \tag{3.4}$$

Here $x_{i,1}$ is the outside temperature and $x_{i,2}$ is an indicator variable taking the value 1 after installation.

To estimate the parameters by hand, we first set up the response vector and design matrix,

$$\mathbf{y} = \begin{bmatrix} 7.2 \\ 6.9 \\ 6.4 \\ \vdots \\ 2.6 \\ 4.8 \\ \vdots \\ 3.5 \\ 3.4 \end{bmatrix} \qquad \text{and} \qquad \mathbf{X} = \begin{bmatrix} 1 & -0.8 & 0 & 0 \\ 1 & -0.7 & 0 & 0 \\ 1 & 0.4 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 10.2 & 0 & 0 \\ 1 & -0.7 & 1 & -0.7 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 4.7 & 1 & 4.7 \\ 1 & 4.9 & 1 & 4.9 \end{bmatrix}.$$

Since $\mathbf{X}^T\mathbf{X}$ will be a 4×4 matrix, it is easier to do our calculations in **R**. First load the data set **gas**:

```
## load("gas.Rdata")
names(gas)
```

```
## [1] "Insulate"  "Temp"      "Gas"       "Insulate2"
```

- **`Insulate`** contains After or Before to indicate whether or not cavity wall insulation has taken place.

- **`Temp`** contains outside temperature.

- **`Gas`** contains gas consumption.

- **`Insulate2`** contains a 0 or 1 to indicate before (0) or after (1) cavity wall insulation.

To set up the design matrix $\mathbf{X}$, we use

```
n = nrow(gas)
X = matrix(cbind(rep(1, n), gas$Temp, gas$Insulate2, gas$In
```

Then to obtain $\underline{\hat{\beta}}$,

```
beta = solve( t(X) %*% X ) %*% t(X) %*% gas$Gas
beta
```

```
##           [,1]
## [1,]  6.8538
## [2,] -0.3932
## [3,] -2.2632
## [4,]  0.1436
```

Thus the fitted model is

$$\mathbb{E}(Y_i) = 6.85 - 0.393x_{i,1} - 2.26x_{i,2} + 0.144x_{i,1}x_{i,2}.$$

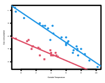

Figure 3.1: Gas consumption (1000's cubic feet) against outside temperature (` C), before (blue) and after (red) cavity wall insulation: separate straight line relationships added.

- Before cavity wall insulation, when the outside temperature is 0 ` C, the expected gas consumption is 6.85 1000's cubic feet.

- Before cavity wall insulation, for every increase in temperature of 1 ` C, the expected gas consumption decreases by 0.393 1000's cubic feet.

- After cavity wall insulation, for every increase in temperature of 1 ` C, the expected gas consumption decreases by 0.249 1000's cubic feet.

- Where does the figure 0.249 come from?

  Substitute $x_{i,2} = 1$ into the fitted model; $-0.393 + 0.144$ is the overall rate of change of gas consumption with temperature.

- What is the expected gas consumption after cavity wall insulation, when the outside temperature is 0 degree Celsius?

  We get $6.85 - 2.26 = 4.59$ thousand cubic feet.

3.4 Predicted Values

Once we have estimated the regression coefficients $\underline{\hat{\beta}}$, we can estimate predicted values of the response variable. The predicted value for individual i is defined as

$$\hat{\mu}_i = \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + ... + \hat{\beta}_p x_{i,p}. \tag{3.5}$$

This equation can also be used to obtain predicted values for combinations of explanatory variables unobserved in the sample. However, care should be taken not to extrapolate too far outside of the observed ranges of the explanatory variables.

The predicted value is interpreted as the expected value of the response variable for a given set of explanatory variable values. Predicted values are useful for checking model fit, calculating residuals and as model output.

Recall the simple linear regression example on birth weights,

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_i,$$

where x_i is gestational age at birth. We obtained $\underline{\hat{\beta}} = (-1485, 115.5)$.

FOOD FOR THOUGHTS

Can you predict the birth weight of a child at 37.5 weeks?

It's ___________ grams.

Click here to see the solution

Using the formula above, $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \times 37.5 = -1485 + 115.5 \times 37.5 = 2846.25$ grams.

3.5 Estimation of the Residual Variance σ^2

From the definition of the linear regression model, there is one other parameter to be estimated: the residual variance σ^2 . We estimate this parameter using the variance of the estimated residuals.

The estimated residuals are defined as,

$$\hat{\epsilon}_i = y_i - \hat{\mu}_i = y_i - \hat{\beta}_1 x_{i,1} - \ldots - \hat{\beta}_p x_{i,p},$$

and we estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

The heuristic reason for dividing by $n-p$, rather than n , is that, although the sum is over n residuals, these are not independent since each is a function of the p parameter estimates $\hat{\beta}_1, \ldots, \hat{\beta}_p$. Dividing by $n-p$ then gives an unbiased estimate of the residual variance. This is the same reason that we divide by $n-1$, rather than n, to get the sample variance. The square root of the residual variance, σ, is referred to as the residual standard error.

Returning to the simple linear regression on birth weight, to calculate the residuals we subtract the fitted birth weights from the observed birth weights.

The birth weights are

$$\begin{aligned}
\hat{\mu}_1 &= -1485 + 116 \times 40 = 3155, \\
\hat{\mu}_2 &= -1485 + 116 \times 38 = 2923, \\
&\vdots
\end{aligned}$$

Then the estimated residuals are

$$\begin{aligned}
\hat{\epsilon}_1 &= y_1 - \hat{\mu}_1 = 2968 - 3155 = -187, \\
\hat{\epsilon}_2 &= y_2 - \hat{\mu}_2 = 2795 - 2923 = -128,
\end{aligned}$$

$$\vdots$$

and the estimate of the residual variance is then

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} \hat{\epsilon}_i^2 = \frac{1}{24-2}[(-187)^2 + (-128)^2 + \ldots] \approx 37096,$$

since $n = 24$ and $p = 2$.

4 Sampling Distribution of Estimators

So far, we have focused on the estimation and interpretation of the regression coefficients. In practice, it is never sufficient just to report parameter estimates, without also reporting either a standard error or confidence interval. These measures of uncertainty can also be used to decide whether or not the relationships represented by the regression models are significant.

As for any estimators discussed thus far, $\underline{\hat{\beta}}(\mathbf{Y})$ and $\hat{\sigma}^2(\mathbf{Y})$ are random variables, since they are both functions of the response vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$. Consequently, they each have a sampling distribution. This is our starting point for deriving standard errors and confidence intervals.

4.1 Least Squares Estimator

We can write the least square estimator for the regression coefficients as

$$\underline{\hat{\beta}}(\mathbf{Y}) = \mathbf{A}\mathbf{Y},$$

where $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Since $\mathbf{A}$ is considered to be fixed, $\underline{\hat{\beta}}(\mathbf{Y})$ is a linear combination of the random variables $Y_1, \dots, Y_n$. By the definition of the linear model, $Y_1, \dots, Y_n$ are normal random variables, and so any linear combination of $Y_1, \dots, Y_n$ is also a normal random variable (by the linearity property of the normal distribution, see MA50215).

4.1.1 Expectation

Let's derive the expectation of the estimator $\underline{\hat{\beta}}(\mathbf{Y})$ in terms of $\mathbf{A}$, $\mathbf{X}$ and $\underline{\beta}$.

By linearity of the expectation, we have

$$\mathbb{E}\left[\underline{\hat{\beta}}(\mathbf{Y})\right] = \mathbb{E}\left[\mathbf{A}\mathbf{Y}\right] = \mathbf{A}\mathbb{E}[\mathbf{Y}].$$

Hence, $\mathbb{E}\left[\underline{\hat{\beta}}(\mathbf{Y})\right] = \mathbf{A}\mathbf{X}\underline{\beta}$ by definition of the linear model.

Now

$$\mathbf{A}\mathbf{X} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{I}_p,$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix. Combining these calculations, we find that

$$\mathbb{E}\left[\hat{\underline{\beta}}(\mathbf{Y})\right] = \mathbf{A}\mathbf{X}\underline{\beta} = \mathbf{I}_p\underline{\beta} = \underline{\beta}.$$

Consequently, the least square estimator is unbiased.

4.1.2 Variance

To find the variance, we first observe that

$$\text{Var}\left[\hat{\underline{\beta}}(\mathbf{Y})\right] = \text{Var}(\mathbf{A}\mathbf{Y}) = \mathbf{A}\text{Var}(\mathbf{Y})\mathbf{A}^T$$

by properties of the variance seen in MA50215.

By definition of the linear model,

$$\mathbf{A}\text{Var}(\mathbf{Y})\mathbf{A}^T = \mathbf{A}\left(\sigma^2\mathbf{I}_n\right)\mathbf{A}^T = \sigma^2\mathbf{A}\mathbf{A}^T.$$

Now

$$\mathbf{A}\mathbf{A}^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T$$

$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{I}_p$$

$$= (\mathbf{X}^T\mathbf{X})^{-1}.$$

Consequently,

$$\mathrm{Var}\left[\underline{\hat{\beta}}(\mathbf{Y})\right] = \sigma^2\mathbf{A}\mathbf{A}^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

To summarize, the sampling distribution for the estimator of the regression coefficient is

$$\underline{\hat{\beta}}(\mathbf{Y}) \sim \mathrm{MVN}_p\left(\underline{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right).$$

In practice, the residual variance σ^2 is usually unknown and must be replaced by its estimate $\hat{\sigma}^2$.

4.2 Residual Error

The sampling distribution of the estimator $\hat{\sigma}^2(\mathbf{Y})$ of the residual error follows a χ^2_{n-p} distribution. We do not give a formal proof of this here, but the intuition is that the estimator $\hat{\sigma}^2(\mathbf{Y})$ is the sum of squares of Normal random variables (the estimated residuals), and hence has a χ^2 distribution.

The degrees of freedom $n - p$ comes from the fact that the estimated residuals are not independent (each is a function of the estimated regression coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_p$).

Additionally, in the same way that the sample mean and variance are independent, so too are the estimators of the regression coefficients $\underline{\hat{\beta}}(\mathbf{Y})$ and the residual variance $\hat{\sigma}^2(\mathbf{Y})$. Although we do not prove this result, it is used below to justify a hypothesis test for the regression coefficient β_j.

4.3 Hypothesis Tests for the Regression Coefficients

The question that is typically asked of a regression model is "Is there evidence of a significant relationship between an explanatory variable and a response variable?". For example, "Is there evidence that domestic gas consumption increases as outside temperatures decrease?".

An equivalent way to ask this is "Is there evidence that the regression coefficient β_j associated with the explanatory variable x_j of interest is significantly different to zero?". This can be answered by testing

$$H_0 : \beta_j = 0 \qquad \text{vs.} \qquad H_1 : \beta_j \neq 0. \tag{4.1}$$

More generally we can test

$$H_0 : \beta_j = b \qquad \text{vs.} \qquad H_1 : \beta_j \neq b. \tag{4.2}$$

In analogy with the tests in MA50215, the test statistic required for the hypothesis test is

$$t = \frac{\hat{\beta}_j - b}{\sqrt{\hat{\sigma}^2 (\mathbf{X}^T\mathbf{X})^{-1}_{j,j}}} \tag{4.3}$$

where $(\mathbf{X}^T\mathbf{X})^{-1}_{j,j}$ is the j-th diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$.

Since

- $\hat{\beta}_j(\mathbf{Y})$ follows a Normal distribution,

- $\hat{\sigma}^2(\mathbf{Y})$ follows a χ^2_{n-p} distribution, and

- $\hat{\beta}_j(\mathbf{Y})$ is independent of $\hat{\sigma}^2(\mathbf{Y})$,

the test statistic follows a t-distribution with $n-p$ degrees of freedom under the null hypothesis H_0. Consequently, we decide whether to reject, or fail to reject, H_0 by comparing the test statistic t to the critical value determined by the t-distribution with $n-p$ degrees of freedom (or by using p-values).

Note that the standard error of $\hat{\beta}_j(\mathbf{Y})$ is $\sqrt{\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{j,j}}$.

Example 4.1 (Birth weight continued).
Recall the simple linear regression model relating birth weight to gestational age at birth,

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_i.$$

We want to test whether gestational age has a significant positive effect on birth weight, that is,

$$H_0 : \beta_2 = 0 \qquad \text{vs.} \qquad H_1 : \beta_2 > 0.$$

First, we calculate $(\mathbf{X}^T\mathbf{X})^{-1}$. From previous calculations in Section 3.3.1, we have

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 19.6 & -0.507 \\ -0.507 & 0.0132 \end{bmatrix}.$$

The test statistic is then

$$t = \frac{\hat{\beta}_2 - b}{\sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{2,2}}} = \frac{116 - 0}{\sqrt{37455 \times 0.0132}} = \frac{115.5}{22.2} = 5.22.$$

Compare $t = 5.22$ to the t_{22} distribution. From **R**, $t_{22}(0.95)$ is 1.72 . Since $5.22 > 1.72$, we conclude that there is evidence to reject H_0 at the 5% level and say that gestational age at birth does affect positively birth weight.

Example 4.2 (Gas consumption continued).

Recall the multiple linear regression model in Section 3.3.2 relating gas consumption to outside temperature and whether or not cavity wall insulation has been installed.

We now wish to address if, before cavity wall insulation was installed, there was a significant relationship between outside temperature and gas consumption.

To answer this question, we need to test

$$H_0 : \beta_2 = 0 \qquad \text{vs.} \qquad H_1 : \beta_2 \neq 0.$$

We firstly calculate the estimated residual variance as $\hat{\sigma}^2 = 0.0728$ by deriving the fitted values, then the residuals and finally using the formula in Section 3.5.

Since we are testing β_2, we need $(\mathbf{X}^T\mathbf{X})^{-1}_{2,2}$. We calculate that

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.17719 & -0.02593 & -0.17719 & 0.02593 \\ -0.02593 & 0.00485 & 0.02593 & -0.00485 \\ -0.17719 & 0.02593 & 0.40970 & -0.08888 \\ 0.02593 & -0.00485 & -0.08888 & 0.02724 \end{bmatrix}$$

The test statistic is thus

$$t = \frac{\widehat{\beta}_2 - 0}{\sqrt{\widehat{\sigma}^2 (X^T X)^{-1}_{2,2}}} = \frac{-0.393}{\sqrt{0.0728 \times 0.00485}} = -20.9.$$

Since $n = 44$ and $p = 4$, we compare to $t_{40}(0.975) = 2.021$. Clearly $|-20.9| = 20.9 > 2.021$, so we conclude that there is evidence to reject H_0 at the 5% level, i.e., there is evidence of a relationship between outside temperature and gas consumption.

4.4 Confidence Intervals for the Regression Coefficients

We can also use the sampling distributions of $\widehat{\beta}_j(\mathbf{Y})$ and $\widehat{\sigma}^2(\mathbf{Y})$ to create a $100(1-\alpha)\%$ confidence interval for β_j,

$$\widehat{\beta}_j \pm t_{n-p}(1-\alpha/2) \times \sqrt{\widehat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}_{j,j}}.$$

As discussed in MA50215, the confidence interval can be used to test $H_0 : \beta_j = b$ against $H_1 : \beta_j \neq b$. The null hypothesis is rejected at the $\alpha\%$ significance level if b does not lie in the $100(1-\alpha)\%$ confidence interval (in other words, if the confidence interval does not contain b, b is not a plausible value for the coefficient β_j under study).

To test against the one-tailed alternatives,

- if $H_1 : \beta_j > b$: calculate the $100(1-2\alpha)\%$ confidence interval and reject H_0 at the $\alpha\%$ level if b lies below the lower bound of the confidence interval;

- if $H_1 : \beta_j < b$: calculate the $100(1-2\alpha)\%$ confidence interval and reject H_0 at the $\alpha\%$ level if b lies above the upper bound of the confidence interval.

Example 4.3 (Birth weights: confidence interval and two-tailed test).
Let's derive a 95% confidence interval for the regression coefficient β_2 representing this relationship between weight and gestational age at birth.

We have all the information to do this from the previous Example 4.1, therefore:

1. $\hat{\beta}_2 = 115.5$, $\text{se}(\hat{\beta}_2) = \sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{2,2}} = 22.2$ and

$t_{22}(0.975) = 2.074\,.$

2. Then the 95% confidence interval for β_2 is

$$\hat{\beta}_2 \pm t_{22}(0.975) \times \sqrt{\hat{\sigma}^2 (\mathbf{X}^T\mathbf{X})^{-1}_{2,2}} = 115.5 \pm 2.074 \times 22.2 = (69.5, 161.$$

Suppose that we now want to test

$$H_0 : \beta_2 = 0 \qquad \text{vs.} \qquad H_1 : \beta_2 \neq 0.$$

Since zero (i.e., value of β_2 under H_0) lies outside the interval $(69.5, 161.5)$, there is evidence at the 5% level to reject H_0, i.e., there is evidence of a relationship between gestational age and weight at birth.

Example 4.4 (Birth weights: confidence interval and one-tailed test).
Suppose that we now want to test

$$H_0 : \beta_2 = 0 \qquad \text{vs.} \qquad H_1 : \beta_2 > 0$$

as in Example 4.1. To test at the 5% level, we have to use $t_{22}(0.90) = 1.717$ in order to calculate a 90% confidence interval. As above,

$$\hat{\beta}_2 \pm t_{22}(0.90) \times \sqrt{\hat{\sigma}^2 (\mathbf{X}^T\mathbf{X})^{-1}_{2,2}} = 116 \pm 1.717 \times 22.2 = (77.9, 154.1)$$

Since $0 < 77.9$, we conclude that there is evidence for a positive relationship between gestational age and weight at birth.

4.5 Linear Combinations of Regression Coefficients

4.5.1 Sampling Distribution

Recall the model for birth weight in Chapter 2 that included separate intercepts for male (β_1) and female (β_2) newborns. We might be interested in the difference between male and female birth weights, $\beta_1 - \beta_2$, estimated by $\hat{\beta}_1 - \hat{\beta}_2$. In particular, we might be interested in testing whether or not there is a difference between β_1 and β_2,

$$H_0 : \beta_1 - \beta_2 = 0 \qquad \text{vs.} \qquad H_1 : \beta_1 - \beta_2 \neq 0.$$

Core questions are:

- What is an appropriate test statistic for this test?

- What sampling distribution should we use to obtain the critical region, the p-value or confidence interval?

Since $\beta_1 - \beta_2$ is a linear combination of the regression coefficients, we can find the distribution of their estimators, and hence a test statistic for this test.

In general for the linear combination

$$\mathbf{a}^T \underline{\beta} = a_1\beta_1 + ... + a_p\beta_p = \sum_{j=1}^{p} a_j\beta_j,$$

we have

$$\mathbb{E}\left[\mathbf{a}^T \underline{\widehat{\beta}}(\mathbf{Y})\right] = \mathbf{a}^T \mathbb{E}\left[\underline{\widehat{\beta}}(\mathbf{Y})\right] = \mathbf{a}^T \underline{\beta}.$$

and

$$\begin{aligned}\text{Var}\left[\mathbf{a}^T \underline{\widehat{\beta}}(\mathbf{Y})\right] &= \mathbf{a}^T \text{Var}\left[\underline{\widehat{\beta}}(\mathbf{Y})\right]\mathbf{a} \\ &= \mathbf{a}^T \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}\end{aligned}$$

$$= \sigma^2 \mathbf{a}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{a}.$$

Further, because $\underline{\widehat{\beta}}(\mathbf{Y})$ follows a multivariate normal distribution, we have that $\mathbf{a}^T \underline{\widehat{\beta}}(\mathbf{Y})$ follows too a normal distribution,

$$\mathbf{a}^T \underline{\widehat{\beta}}(\mathbf{Y}) \sim \text{Normal}\left(\mathbf{a}^T \underline{\beta}\,,\, \sigma^2 \mathbf{a}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{a}\right).$$

Example 4.5 (Birth weight continued).

Let's revisit the model for the birth weight data with

$$\mathbb{E}(Y_i) = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3},$$

where $x_{i,1}$ and $x_{i,2}$ are indicators for sex-at-birth being male and female respectively, and $x_{i,3}$ is gestational age (see Example 2.4).

Using the data in the file **bwt**, we have

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 40 \\ 1 & 0 & 38 \\ 1 & 0 & 40 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 38 \\ 0 & 1 & 40 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 40 \end{bmatrix}, \qquad (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 19.7 & 19.8 & -0.512 \\ 19.8 & 20.1 & -0.517 \\ -0.512 & -0.517 & 0.0133 \end{bmatrix}$$

The expectation and variance of $\widehat{\beta}_1(\mathbf{Y}) - \widehat{\beta}_2(\mathbf{Y})$ are computed as follows.

First write $\widehat{\beta}_1(\mathbf{Y}) - \widehat{\beta}_2(\mathbf{Y}) = \mathbf{a}^T\underline{\widehat{\beta}}(\mathbf{Y})$ for some $\mathbf{a}$, e.g.,

$$\widehat{\beta}_1(\mathbf{Y}) - \widehat{\beta}_2(\mathbf{Y}) = 1 \times \widehat{\beta}_1(\mathbf{Y}) + (-1) \times \widehat{\beta}_2(\mathbf{Y}) + 0 \times \widehat{\beta}_3(\mathbf{Y}) = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix}$$

So $\mathbf{a} = (1, -1, 0)^T$. Consequently,

$$\mathbb{E}\left[\hat{\beta}_1(\mathbf{Y}) - \hat{\beta}_2(\mathbf{Y})\right] = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \underline{\beta}$$

$$= \beta_1 - \beta_2$$

and

$$\text{Var}\left[\hat{\beta}_1(\mathbf{Y}) - \hat{\beta}_2(\mathbf{Y})\right] = 31370 \times \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \left[\begin{matrix} 19.7 & 19.8 \\ 19.8 & 20.1 \\ -0.512 & -0.517 \end{matrix}\right.$$

$$= 31370 \times 0.169$$

$$= 5301.$$

4.5.2 Hypothesis Testing

A similar approach as in Section 4.3 can be taken for linear combinations of regression coefficients. We know that $\mathbf{a}^T\underline{\hat{\beta}}(\mathbf{Y})$ follows a normal distribution with mean $\mathbf{a}^T\underline{\beta}$ and variance

$\sigma^2 \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}\,.$

To test

$$H_0 : \mathbf{a}^T \underline{\beta} = b \qquad \text{vs.} \qquad H_1 : \mathbf{a}^T \underline{\beta} \neq b,$$

we consider the test statistic

$$t = \frac{\mathbf{a}^T \underline{\hat{\beta}} - b}{\sqrt{\hat{\sigma}^2 \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}}$$

which, under the null hypothesis, is t_{n-p} distributed.

Example 4.6 (Gas consumption continued).
To model gas consumption, we considered

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1} + \beta_3 x_{i,2} + \beta_4 x_{i,1} x_{i,2},$$

where $x_{i,1}$ represents whether wall insulation was installed and $x_{i,2}$ is the outside temperature.

Let's consider the following question: After cavity wall insulation was installed, is there a significant relationship between outside temperature and gas consumption?

We have seen that the relationship between gas consumption and outside temperature after insulation is given by $\beta_2 + \beta_4$. Hence, to answer the question above, we need to test

$$H_0 : \beta_2 + \beta_4 = 0 \qquad \text{vs.} \qquad H_1 : \beta_2 + \beta_4 \neq 0.$$

First, we calculate the variance of $\widehat{\beta}_2(\mathbf{Y}) + \widehat{\beta}_4(\mathbf{Y})$. Using results from MA50215, we get

$$\begin{aligned}\text{Var}\left[\widehat{\beta}_2(\mathbf{Y}) + \widehat{\beta}_4(\mathbf{Y})\right] &= \text{Cov}\left[\widehat{\beta}_2(\mathbf{Y}) + \widehat{\beta}_4(\mathbf{Y}), \widehat{\beta}_2(\mathbf{Y}) + \widehat{\beta}_4(\mathbf{Y})\right] \\ &= \text{Var}\left[\widehat{\beta}_2(\mathbf{Y})\right] + 2\text{Cov}\left[\widehat{\beta}_2(\mathbf{Y}), \widehat{\beta}_4(\mathbf{Y})\right] + \text{Var}\Big[\widehat{\beta}_4(\\ &= \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{2,2} + 2\sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{2,4} + \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{4,4} \\ &= \sigma^2 \left[(\mathbf{X}^T\mathbf{X})^{-1}_{2,2} + 2(\mathbf{X}^T\mathbf{X})^{-1}_{2,4} + (\mathbf{X}^T\mathbf{X})^{-1}_{4,4}\right].\end{aligned}$$

Next, we obtain the required elements from $(\mathbf{X}^T\mathbf{X})^{-1}$ and calculate the test statistic with $\mathbf{a} = (0, 1, 0, 1)^T$,

$$t = \frac{\widehat{\beta}_2 + \widehat{\beta}_4}{\sqrt{\text{Var}\left[\widehat{\beta}_2(\mathbf{Y}) + \widehat{\beta}_4(\mathbf{Y})\right]}}$$

$$= \frac{-0.393 + 0.144}{\sqrt{0.0728 \times (0.00485 + 2 \times (-0.00485) + 0.273)}}$$

$$= \frac{-0.250}{\sqrt{0.0728 \times 0.0272}}$$

$$= -6.18.$$

Finally, compare $t = |-6.18| = 6.18$ to $t_{40}(0.975) = 2.021$. Since $6.18 > 2.021$ we conclude that, at the 5% level, there is evidence of a relationship between gas consumption and outside temperature, once insulation has been installed. So, the insulation has not entirely isolated the house from the effects of external temperature, but it does appear to have weakened this relationship.

5 Linear Regression: Collinearity, Interactions and Transformations

We have introduced the linear model, and discussed some properties of its estimators. Over the remaining three sections on linear regression of this course, we discuss further modelling issues that can arise when fitting a linear regression model:

- Collinearity and interactions between explanatory variables.
- Covariate selection (sometimes referred to as model selection).
- Diagnostics (assessing goodness of model fit).

5.1 Collinearity

Collinearity arises when there is linear dependence (strong correlation) between two or more explanatory variables. We say that two explanatory variables x_i and x_j are

- Orthogonal if $\text{Corr}(x_i, x_j)$ is close to zero, and
- Collinear if $\text{Corr}(x_i, x_j)$ is close to 1.

Collinearity is undesirable because it means that the matrix $\mathbf{X}^T\mathbf{X}$ is ill-conditioned, and inversion of $(\mathbf{X}^T\mathbf{X})$ is numerically unstable. It can also make results difficult to interpret. Note, this issue is quite similar to the case of linearly dependent columns discussed in Chapter 3.

Example 5.1 (Cereal Prices).

We investigate global commodity price forecasts for various cereals from 1995–2015. Annual prices (dollars per tonne) are available for maize, barley and wheat, made available by the Economist Intelligence Unit and downloaded from http://datamarket.com/.

We aim to relate annual maize prices, Y_i, to annual prices of barley, $x_{i,1}$, and wheat, $x_{i,2}$.

Let's start by plotting the explanatory variables against the response variable.

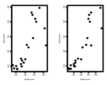

Figure 5.1: Annual price of barley against annual price of maize (left) and annual price of wheat against annual price of maize (right) in dollars per tonne.

The plots indicate that a linear relationship may exist for both Y_i and $x_{i,1}$ and Y_i and $x_{i,2}$. We thus decide to consider the following three models:

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1},$$

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,2},$$

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1} + \beta_3 x_{i,2}.$$

We fit the models using **R**:

```
lm(Maize ~ Barley, data = cereal)$coefficients
lm(Maize ~ Wheat, data = cereal)$coefficients
lm(Maize ~ Barley + Wheat, data = cereal)$coefficients
```

```
## (Intercept)      Barley
##      -9.485       1.086
## (Intercept)       Wheat
##    -30.8255      0.9491
## (Intercept)      Barley       Wheat
##    -25.6646     -0.5096      1.3208
```

We see that in the model with two covariates, the coefficients for barley and wheat are considerably different to the equivalent estimates obtained for the models with one covariate. In particular the relationship with barley is positive in the first model and negative in the third model.

What is going on here? To investigate, we check which of the covariates has a significant relationship with maize prices in each of the models. We will use the confidence interval method. For this we need the standard errors of the regression coefficients, which can be found by hand or using the **summary** function.

For instance, for the first model, we obtain

```
summary(lm(Maize ~ Barley, data = cereal))$coefficients
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   -9.485    32.2742 -0.2939 7.720e-01
## Barley         1.086     0.1919  5.6566 1.875e-05
```

The **R** output shows that the standard error for β_2 in the first model is 0.1919. The 95% confidence interval for β_2 (barley) is then

$$\hat{\beta}_2 \pm t_{21-2}(0.975) \times \mathrm{se}(\hat{\beta}_2) = 1.09 \pm 2.093 \times 0.1919 = (0.684, 1.487).$$

Using the same approach, the 95% confidence interval for β_2 (wheat) in the second model is

$$0.949 \pm 2.093 \times 0.1109 = (0.717, 1.18).$$

Finally, for the model with two covariates, the 95% confidence interval for β_2 (barley) is

$$\hat{\beta}_2 \pm t_{21-3}(0.975) \times \mathrm{se}(\hat{\beta}_2) = -0.510 \pm 2.101 \times 0.4076 = (-1.366, 0.347),$$

and the 95% confidence interval for β_3 (wheat) is

$$1.321 \pm 2.101 \times 0.3167 = (0.655, 1.99).$$

Overall,

Characteristic	Beta	95% CI[1]	p-value
Only barley			
(Intercept)	-9.485	-77.04, 58.07	0.772
Barley	1.086	0.6840, 1.487	<0.001

Only wheat			
(Intercept)	-30.83	-80.96, 19.31	0.214
Wheat	0.9491	0.7170, 1.181	<0.001
Both			
(Intercept)	-25.66	-76.01, 24.68	0.298
Barley	-0.5096	-1.366, 0.3467	0.227
Wheat	1.321	0.6554, 1.986	<0.001

[1] CI = Confidence Interval

We can conclude that

- If barley alone is included, then it has a significant relationship with maize price (at the 5% level).

- If wheat alone is included, then it has a significant relationship with maize price (at the 5% level).

- If both barley and wheat are included, then the relationship with barley is no longer significant (at the 5% level).

Why is that?

The answer comes if we look at the relationship between barley and wheat prices in the plot below.

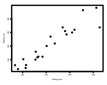

Figure 5.2: Annual prices for wheat against barley (dollars per tonne).

The sample correlation between these variables is 0.939, indicating a very strong linear relationship. Since their behaviour is so closely related, we do not need both as covariates in the model.

If we do include both, then it is impossible for the model to accurately identify the individual relationships. We should thus use either the first or second model. However, there is no statistical way to compare these two models; but one possibility is to select the one which has smallest p-value associated with β_2, i.e., the one with the strongest relationship between the covariate and the response.

5.2 Interactions

We already touched on the concept of an interaction in the gas consumption model given in Example 3.3.2. In this example, the relationship between gas consumption and outside temperature was altered by the installation of cavity wall insulation. This is an interaction between a factor (insulated or not) and a covariate (temperature).

Suppose that the response variable is Y_i and there are two explanatory variables $x_{i,1}$ and $x_{i,2}$. We could either

1. Model the main effects only,

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1} + \beta_3 x_{i,2},$$

2. Or include an interaction as well

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1} + \beta_3 x_{i,2} + \beta_4 x_{i,1} x_{i,2}.$$

Note, the interaction term is sometimes written as $x_{i,1} \times x_{i,2}$.

5.2.1 Interaction between two factors

We illustrate the idea of an interaction between two factors using a thought experiment.

A clinical trial is to be carried out to investigate the effect of the doses of two drugs A and B on a medical condition. Both drugs are available at two dose levels. All four combinations of drug-dose levels will be investigated. For the trial, N patients are randomly assigned to each of the possible combinations of drug-dose levels, so that $N/4$ patients receive each combination. The response variable is the increase from pre- to post-treatment red blood cell count. The average increase is calculated for each drug-dose level combination.

In all three outcomes, the level of both drugs affects cell count.

1. In outcome 1, cell count increases with dose level of both A and B. Since the size and direction of the effect of the dose level of drug A on the cell count is unchanged by changing the dose level of drug B there is no interaction.

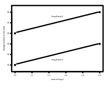

Figure 5.3: Average increase in cell counts with dose levels of both A and B in outcome 1.

2. In outcome 2, there is an interaction: at level 1 of drug B, the cell count is lower for drug A level 2, than for drug A level 1. Conversely, at level 2 of drug B, the cell count is lower for drug A level 1, than for drug A level 2. The direction of the effect of the dose levels of drug A is altered by changing the dose of drug B.

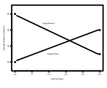

Figure 5.4: Average increase in cell counts with dose levels of both A and B in outcome 2.

3. In outcome 3, there is also an interaction. In this case increasing the dose level of drug A increases cell count, regardless of the level of drug B. But the difference in the response

for levels 1 and 2 of drug A is much greater for level 1 of drug B than it is for level 2 of drug B. The size of the effect of the dose levels of drug A is altered by changing the dose of drug B.

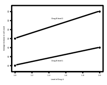

Figure 5.5: Average increase in cell counts with dose levels of both A and B in outcome 3.

5.2.2 Interaction between a factor and a covariate

In the gas consumption example in Chapter 3, we were interested in the relationship between outside temperature and gas consumption. We saw that the size of this relationship depends on whether or not the house has cavity wall insulation and we wrote this model formally as

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1} + \beta_3 x_{i,2} + \beta_4 x_{i,1} x_{i,2},$$

where

- The coefficient β_2 is the size of the main effect of outside temperature on gas consumption.

- The coefficient β_4 is the size of the interaction between the effect of outside temperature and whether or not insulation is installed.

To test whether or not there is an interaction, i.e., whether or not installing insulation has a significant effect on the relationship between outside temperature and gas consumption, we can test

$$H_0 : \beta_4 = 0 \qquad \text{vs} \qquad H_1 : \beta_4 \neq 0.$$

We have previously fitted this model in **R** and we will use the output from this model to speed up our testing procedure,

```
gaslm <- lm(gas$Gas ~ gas$Temp * gas$Insulate2)
summary(gaslm)
```

```
##
## Call:
## lm(formula = gas$Gas ~ gas$Temp * gas$Insulate2)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -0.620 -0.180  0.034  0.164  0.598
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|
## (Intercept)               6.8538     0.1136   60.32  < 2e
## gas$Temp                 -0.3932     0.0188  -20.93  < 2e
## gas$Insulate2            -2.2632     0.1728  -13.10  4.7e
## gas$Temp:gas$Insulate2   0.1436     0.0446    3.22   0.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 0.27 on 40 degrees of freedom
## Multiple R-squared:  0.936,  Adjusted R-squared:  0.931
## F-statistic:  195 on 3 and 40 DF,  p-value: <2e-16
```

Reading from the second column in the Coefficients table, we find that the standard error for $\hat{\beta}_4$ is 0.04455. The test statistic is then

$$t = \frac{\hat{\beta}_4 - 0}{\text{se}(\hat{\beta}_4)} = \frac{0.144}{0.04455} = 3.22,$$

which also appears in the third column of the Coefficients table.

To decide whether to reject H_0 or not, we have to compare the test statistic to the critical value $t_{40}(0.975) = 2.021$. Since $3.22 > 2.021$ there is evidence at the 5% level to reject H_0, i.e. there was a significant change in the relationship between outside temperature and gas consumption following insulation.

5.3 Transformations

We have seen in Question 2 on Problem Sheet 2 that linear transformations (centring) of an explanatory variable can, for instance, improve interpretation of the intercept.

Other transformations that may be considered in the linear regression setting include:

- Logarithmic transformation of the response or predictor

  It might make sense to take logs of the response variable, especially if the data are positive; we will consider such an example in the next chapter. We may additionally (separately) choose to take logs of the predictor, to mitigate for lack of linearity / additivity of variables. An alternative is the square root transformation, which is less severe than the log at the extremes, so may be more suitable.

- Box-Cox transformation of the response

  If the residuals appear to have a skewed distribution, then the normality assumption may not be valid. In this case, transforming the response to be more normally distributed may be appropriate. A commonly used such function is the Box-Cox family of transformations, defined by

$$y^* = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0, \\ \log(y) & \text{for } \lambda = 0. \end{cases} \tag{5.1}$$

  For $\lambda = 1$, this leaves the response unchanged (apart from a location shift); $\lambda = 0$ corresponds to the log transformation as above; $\lambda = \frac{1}{2}$ corresponds to the square root transformation, which is less severe than the log transform. These transformations can also have a linearising effect on the response-predictor relationship. A disadvantage of using such a transformation is that model coefficients become less interpretable.

  The transformation is available in **`R`** with the **`boxcox`** function in the **`MASS`** package.

6 Covariate Selection

Covariate selection refers to the process of deciding which of a number of explanatory variables best explain the variability in the response variable. You can think of it as finding the subset of explanatory variables which have the strongest relationships with the response variable.

6.1 Nested Models

We will only look at comparing nested models. Consider two models, the first has p_1 explanatory variables and the second has $p_2 > p_1$ explanatory variables. We refer to the model with fewer covariates as the simpler model.

An example of a pair of nested models is when the more complicated model contains all the explanatory variables in the simpler model, and an additional $p_2 - p_1$ explanatory variables.

For example, given a response Y_i and explanatory variables $x_{i,1}$, $x_{i,2}$ and $x_{i,3}$, we could create three possible models:

1. $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i,1}$,

2. $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}$,

3. $\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}$.

Which model(s) is nested inside model 3?

- (A) Both models 1 and 2 are nested inside model 3.

- (B) Only model 1 is nested inside model 3.

- (C) Only model 2 is nested inside model 3.

TRUE OR FALSE?

Are either of models 1 or 3 nested inside model 2? TRUE / FALSE

Click here to see the “TRUE OR FALSE?” solution

Model 1 is, since model 2 is model 1 with an additional covariate. Note that neither model 2 nor model 3 is nested in model 1.

Exercise 6.1.

Write down another model that is nested in model 3.

Definition 6.1 (Nested model).

Define model 1 as $\mathbb{E}(Y) = \mathbf{X}\underline{\beta}$, and model 2 as $\mathbb{E}(Y) = \tilde{\mathbf{X}}\underline{\gamma}$, where $\mathbf{X}$ is an $n \times p_1$ matrix and $\tilde{\mathbf{X}}$ is an $n \times p_2$ matrix, with $p_1 < p_2$. Assume $\mathbf{X}$ and $\tilde{\mathbf{X}}$ are both of full rank, i.e., neither has linearly dependent columns.

Then model 1 is nested in model 2 if $\mathbf{X}$ is a (strict) subspace of $\tilde{\mathbf{X}}$.

Given a pair of nested models, we will focus on deciding whether there is enough evidence in the data in favour of the more complicated model, or whether we are justified in staying with the simpler model.

The null hypothesis H_0 in this test is always that the simpler model is the best fit.

We start with an example.

Example 6.1 (Brain weights).
In this example we consider whether the body weight of a mammal could be used to predict its brain weight. In addition, we have the average number of hours of sleep per day for each species in the study.

Let Y_i denote brain weight, $x_{i,1}$ denote body weight and $x_{i,2}$ denote number of hours asleep per day of species i. We will consider the log of both brain and body weight in our model.

Which of the following models L1-L3 fits the data best?

$$\mathbb{E}(\log Y_i) \quad = \quad \beta_1 + \beta_2 \log x_{i,1},$$

$$\mathbb{E}(\log Y_i) \quad = \quad \beta_1 + \beta_2 x_{i,2},$$

$$\mathbb{E}(\log Y_i) \quad = \quad \beta_1 + \beta_2 \log x_{i,1} + \beta_3 x_{i,2}.$$

There are four species for which sleep time is unknown. For a fair comparison between models, we remove these species from the following study completely, leaving $n = 58$ observations.

We can fit each of the models in **R** as follows,

```
notnasleep = which(!is.na(sleep$TotalSleep))
L1 = lm(log(sleep$BrainWt) ~ log(sleep$BodyWt), subset = no
L2 = lm(log(sleep$BrainWt) ~ sleep$TotalSleep, subset = not
L3 = lm(log(sleep$BrainWt) ~ log(sleep$BodyWt) + sleep$Tota
```

Note that models L1 and L2 are both nested in model L3.

The estimated regression coefficients and standard errors are summarised in the following table.

Model	β_1	β_2	β_3
L1	2.15 (0.0991)	0.759 (0.0303)	NA
L2	6.17 (0.675)	-0.299 (0.0588)	NA
L3	2.60 (0.288)	0.728 (0.0352)	-0.0386 (0.0237)

For each model, we can test to see which of the explanatory variables is significant.

For model L1, we test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ by calculating

$$t = \frac{\hat{\beta}_2}{\text{se}(\hat{\beta}_2)} = \frac{0.759}{0.0303} = 25.09.$$

Comparing this to $t_{56}(0.975) = 2.00$, we see that β_2 is significantly different to zero at the 5% level. We conclude that there is evidence of a significant relationship between (log) brain weight and (log) body weight.

For model L2, to test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$, we calculate

$$t = \frac{\hat{\beta}_2}{\text{se}(\hat{\beta}_2)} = \frac{-0.299}{0.0588} = -5.092.$$

Again, the critical value is $t_{56}(0.975) = 2.00$. Since $|-5.092| >$

2.00 we conclude that there is evidence of a relationship between hours of sleep per day and (log) brain weight. This is a negative relationship: the more hours sleep per day, the lighter the brain. We cannot say that this is a causal relationship!

For model L3, we first test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$, using

$$t = \frac{\hat{\beta}_2}{\text{se}(\hat{\beta}_2)} = \frac{0.728}{0.352} = 20.67.$$

Next we test $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 \neq 0$, using

$$t = \frac{\hat{\beta}_3}{\text{se}(\hat{\beta}_3)} = \frac{-0.0386}{0.0237} = -1.632.$$

In both cases, the critical value is $t_{55}(0.975) = 2.00$; so, at the 5% level, there is evidence of a relationship between (log) brain weight and (log) body weight, but there is no evidence of a relationship between (log) brain weight and hours of sleep per day.

To summarise, individually, both explanatory variables appear to be significant. However, when we include both in the model, only one is significant. This appears to be a contradiction.

So which is the best model to explain variability amongst brain weights in mammals?

In general, we want to select the simplest possible model that explains the most variation (in machine learning jargon, the complexity of the model has an impact on the bias-variance trade-off).

Including additional explanatory variables will always increase the amount of variability explained - but is the increase sufficient to justify the additional parameter that must then be estimated?

6.2 The F-test

The F-test gives a formal statistical test to choose between two nested models. It is based on a comparison between the sum of squares for each of the two models.

Suppose that model 1 has p_1 explanatory variables, model 2 has $p_2 > p_1$ explanatory variables and model 1 is nested in model 2. Let model 1 have design matrix $\mathbf{X}$ and parameters $\underline{\beta}$; model 2 has design matrix $\tilde{\mathbf{X}}$ and parameters $\underline{\gamma}$.

First we show formally that adding additional explanatory variables will always improve model fit, by decreasing the residual sum of squares for the fitted model.

If $SS_1 = (\mathbf{y} - \mathbf{X}\underline{\beta})^T(\mathbf{y} - \mathbf{X}\underline{\beta})$ and $SS_2 = (\mathbf{y} - \tilde{\mathbf{X}}\underline{\gamma})^T(\mathbf{y} - \tilde{\mathbf{X}}\underline{\gamma})$ are the residual sums of squares for models 1 and 2 respectively. Then

$$SS_2 \leq SS_1.$$

Why does this last inequality hold?

Because of the nesting, we can always find a value $\underline{\tilde{\gamma}}$ such that

$$\mathbf{X}\underline{\hat{\beta}} = \tilde{\mathbf{X}}\underline{\tilde{\gamma}}.$$

Recalling the definition of the sum of squares and the least square estimate,

$$SS_2 = (\mathbf{y} - \tilde{\mathbf{X}}\underline{\gamma})^T(\mathbf{y} - \tilde{\mathbf{X}}\underline{\gamma})$$

$$\leq (\mathbf{y} - \tilde{\mathbf{X}}\underline{\tilde{\gamma}})^T(\mathbf{y} - \tilde{\mathbf{X}}\underline{\tilde{\gamma}})$$

$$= (\mathbf{y} - \mathbf{X}\underline{\beta})^T(\mathbf{y} - \mathbf{X}\underline{\beta})$$

$$= SS_1.$$

To carry out the F-test we must decide whether the difference between SS_1 and SS_2 is sufficiently large to merit the inclusion of the additional explanatory variables in model 2.

Consider the following hypothesis test,

H_0 : Model 1 is the best fit vs. H_1 : Model 2 is the best fit.

To test H_0 against H_1, first calculate the test statistic

$$F = \frac{(SS_1 - SS_2)/(p_2 - p_1)}{SS_2/(n - p_2)}. \tag{6.1}$$

Now compare the test statistic to the $F_{p_2-p_1,n-p_2}$ distribution, and reject H_0 if the test statistic exceeds the critical value (equivalently if the p-value is too small).

The critical value from the $F_{p_2-p_1,n-p_2}$ distribution can either be evaluated in **R**, or obtained from statistical tables.

Remark: We do not say that “Model 1 is the true model” or “Model 2 is the true model”. All models, be they probabilistic or deterministic, are a simplification of real life. No model can exactly describe a real life process. But some models can describe the truth “better” than others.

6.3 Example: Brain Weights

We proposed three models for (the logarithm of) brain weight with the following explanatory variables:

- in L1 , logarithm of body weight,
- in L2 , hours sleep per day,
- in L3 , logarithm of body weight and hours sleep per day.

Which of these models can we use the F-test to decide between?

The F-test does not allow us to choose between models L1 and L2 , since these are not nested. However, it does give us a way to choose between either the pair L1 and L3 , or the pair L2 and L3 .

To choose between L1 and L2 , we use a more ad-hoc approach by looking to see which of the explanatory variables is "more significant" than the other when we test

$$H_0 : \beta_2 = 0 \qquad \text{vs.} \qquad H_1 : \beta_2 \neq 0.$$

Using **`summary(L1)$coefficients`** and **`summary(L2)$coefficients`**, we see that the p-value for β_2 in L1 is $< 2e^{-16}$ and for β_2 in L2 is $4.30e^{-06}$ (both almost 0). As we saw earlier, both of these indicate highly significant relationships between the response and the explanatory variable in question.

Which of the single covariate models is preferable?

Since the p-value for logarithm of body weight in model L1 is lower, L1 is our preferred model.

We can now use the F-test to choose between our preferred single-covariate model L1 and the model L3 with two covariates,

$$H_0 : \text{L1 is the best fit} \qquad \text{vs.} \qquad H_1 : \text{L3 is the best fit.}$$

We first find the sum of squares for both models. For L1 , using the definition of the least squares,

$$SS(L1) = \sum_{i=1}^{58} \hat{\epsilon}_i^2 = \sum_{i=1}^{58} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i,1})^2 = \sum_{i=1}^{58} (y_i - 2.15 - 0.759 x_{i,1})^2.$$

To calculate this in **R**, we use

```r
deviance(L1)
# sum(L1$residuals^2)   #-- as alternative
```

```
## [1] 28
```

For L3 ,

$$SS(L3) = \sum_{i=1}^{58} \hat{\epsilon}_i^2 = \sum_{i=1}^{58} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i,1} - \hat{\beta}_3 x_{i,2})^2$$

$$= \sum_{i=1}^{58} (y_i - 2.60 - 0.728 x_{i,1} - (-0.0386) x_{i,2})^2.$$

To calculate this in **R**, we again use

```
deviance(L3)
# sum(L3$residuals^2)   #-- as alternative
```

```
## [1] 26.71
```

Next, we find the degrees of freedom for the two models. Since $n = 58$,

- L1 has $p_1 = 2$ regression coefficients, so the degrees of freedom are $n - p_1 = 58 - 2 = 56$.

- L3 has $p_2 = 3$ regression coefficients, so the degrees of freedom are $n - p_2 = 58 - 3 = 55$.

Finally we calculate the F -statistic,

$$F = \frac{[SS(L1) - SS(L3)]/(p_2 - p_1)}{SS(L3)/(n - p_2)}$$

$$= \frac{(28.0 - 26.7)/(3-2)}{26.7/(58-3)}$$

$$= 1.29/0.486$$

$$= 2.67.$$

The test statistic $F = 2.67$ is then compared to the F-distribution with $(p_2 - p_1, n - p_2) = (1, 55)$ degrees of freedom. From tables, the critical value is just above 4.00; from **R** (using **qf(.95, 1,55)**) it is 4.02.

Since $2.67 < 4.00$, we conclude that there is no evidence to reject H_0. Consequently, there is no evidence to choose the more complicated model and so the best fitting model is L1.

We can obtain the p-value associated to the test by

```
1 - pf(2.67, 1, 55)
```

```
## [1] 0.108
```

Remember that the F-distribution is positive and so in effect we examine upper tail of the distribution.

Remark: We should not be too surprised by this result, since we have already seen that the coefficient for total sleep time is not significantly different to zero in model L3. Once we have accounted for body weight, there is no extra information in total sleep time to explain any remaining variability in brain weights.

6.4 Where does the F-test come from?

From Section 3.5, the sum of squares, divided by the degrees of freedom, is an unbiased estimator of the residual variance,

$$\mathbb{E}\left[\frac{SS}{(n-p)}\right] = \sigma^2.$$

Alternatively,

$$\mathbb{E}(SS) = (n-p)\sigma^2.$$

So if both model 1 and model 2 fit the data, both of their normalised sums of squares are unbiased estimates of σ^2, and the expected difference in their sums of squares is

$$\mathbb{E}(SS_1 - SS_2) = \mathbb{E}(SS_1) - \mathbb{E}(SS_2) = (n-p_1)\sigma^2 - (n-p_2)\sigma^2 = (p_2-p_1)\sigma^2.$$

and $(SS_1 - SS_2)/(p_2 - p_1)$ is also an unbiased estimator of the residual variance σ^2 .

But if model 1 is not a sufficiently good model for the data

$$\mathbb{E}\left[\frac{SS_1 - SS_2}{p_2 - p_1}\right] > \sigma^2,$$

since the expected sum of squares for model 1 will be greater than σ^2 as the model does not account for enough of the variability in the response.

It follows that the F -statistic

$$F = \frac{(SS_1 - SS_2)/(p_2 - p_1)}{SS_2/(n - p_2)}.$$

is simply the ratio of two estimates of σ^2 . If model 1 is a sufficient fit, this ratio will be close to 1, otherwise it will be greater than 1.

To see how far the F -statistic must be from 1 for the result not to have occurred by chance, we need its sampling distribution. It turns out that the appropriate distribution is the $F_{(p_2-p_1),(n-p_2)}$ distribution. The proof of this is too long to be covered here.

7 Model Diagnostics

Even if a valid statistical method, such as the F-test, has been used to select our preferred linear model, checks should still be made to ensure that this model fits the data well. After all, it could be that all the models that we tried to fit actually described the data poorly, so that we have just made the best of a bad job. If this is the case, we need to go back and think about the underlying physical processes generating the data to suggest a better model.

Diagnostics refers to a set of tools which can be used to check how well the model describes (or fits) the data. We will use diagnostics to check that

1. The estimated residuals $\hat{\epsilon}_1, \dots, \hat{\epsilon}_n$ follow a Normal$(0, \sigma^2)$ distribution.

2. The estimated residuals are independent of the covariates used in the model.

3. The estimated residuals are independent of the fitted values

$\hat{\mu}_1, \ldots, \hat{\mu}_n$.

4. None of the observations is an outlier (perhaps due to measurement error).

5. No observation has undue influence on the estimated model parameters.

7.1 Normality of Residuals

One of the key underlying assumptions of the linear regression model is that the residuals $\epsilon_1, \ldots, \epsilon_n$ follow a normal distribution. In reality, we do not know the residuals and these are replaced with their estimates, $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$. We compare these estimated residuals to their model distribution using graphical diagnostics.

PP (probability-probability) and QQ (quantile-quantile) plots can be used to check whether or not a sample of data can be considered to be a sample from a statistical model (usually a probability distribution). In the case of the normal linear regression model, they can be used to check whether or not the estimated residuals are a sample from a $\text{Normal}(0, \sigma^2)$ distribution. PP plots show the same information as QQ plots, but on a different scale.

The PP plot is most useful for checking that values around the average (the body) fit the proposed distribution. It compares the percentiles of the sample of data, predicted under the proposed model, to the percentiles obtained for a sample of the same size, predicted from the empirical distribution.

The QQ plot is most useful for checking whether the largest and smallest values (the tails) fit the proposed distribution. It compares the ordered sample of data to the quantiles obtained for a sample of the same size from the proposed model.

First define the Pearson (normalized) residuals to be

$$\hat{r}_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}}.$$

From MA50215, normalizing by $\hat{\sigma}$ means that these should be a sample from a $\text{Normal}(0,1)$ distribution (approximately).

Denote by $\hat{r}^{(i)}$ the ordered Pearson residuals, so that $\hat{r}^{(1)}$ is the smallest residual, and $\hat{r}^{(n)}$ the largest. We compare the Pearson residuals to the standard normal distribution, using

- A PP plot,
  $$\left\{\Phi\left(\hat{r}^{(i)}\right), \frac{i}{n+1}\right\}$$
  for $i = 1, \dots, n$. Here $\Phi(\cdot)$ is the standard normal cumulative distribution function.

- A QQ plot,
  $$\left\{\hat{r}^{(i)}, \Phi^{-1}\left(\frac{i}{n+1}\right)\right\}$$
  for $i = 1, \dots, n$, where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cumulative distribution function.

If the normalized residuals follow a $\mathrm{Normal}(0, 1)$ distribution perfectly, both plots lie on the line $y = x$. Because of random variation, even if the model is a good fit, the points won't lie exactly on this line.

Example 7.1 (Brain weight continued).

In Example 6.1 we fitted the following linear regression model to try to explain variability in (log) brain weight (Y_i) using (log) body weight ($x_{i,1}$),

$$\mathbb{E}(\log Y_i) = 2.15 + 0.759 \log x_{i,1}.$$

We use **R** to create PP and QQ plots for the Pearson residuals.

First we will refit the model in **R** to obtain the required residuals,

```
notnasleep = which( !is.na(sleep$TotalSleep))
L1 = lm(log(BrainWt) ~ log(BodyWt), data = sleep, subset =
```

Next we need the residual variance,

```
sigmasq = sum(L1$residuals^2) / 56
```

and we can use this to get the Pearson residuals:

```
stdresid = L1$residuals / sqrt(sigmasq)
```

R does not have an inbuilt function for creating a PP plot, but we can create one using the function **qqplot**,

```r
qqplot(c(1:58) / 59, pnorm(stdresid), pch = 19,
       xlab = "Theoretical probabilities", ylab = "Sample p
abline(a = 0, b = 1, col = "firebrick")
```

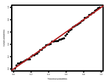

Figure 7.1: PP plot for the L1 model for brain weights of mammals.

Since we are comparing the normalized residuals to the standard Normal distribution, we can use the function **qqnorm** for the QQ plot,

```r
qqnorm(stdresid, pch = 19)
abline(a = 0, b = 1, col = 2)
```

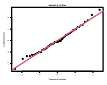

Figure 7.2: QQ plot for the L1 model for brain weights of mammals.

Both plots suggests that the Pearson residuals do follow the standard Normal distribution closely (straight lines indicate exact agreement between the residuals and a Normal $(0, 1)$ distribution).

Remarks:

1. In general, a QQ plot is more useful that a PP plot, as it tells us about the more “unusual” values , i.e., the very high and very low residuals. It is the behaviour of these values which is most likely to highlight a lack of model fit.

2. If the PP and QQ plots suggest that the residuals differ from the $\text{Normal}(0, 1)$ distribution in a systematic way, for example the points curve up (or down) and away from the 45` line at either (or both) of the tails, it may be more appropriate to

   - Transform your response, e.g., use the log or square root functions, before fitting the model.

   - Use a different residual distribution. This is discussed later in this module.

3. A lack of normality might also be due to the residuals having non-constant variance, referred to as heteroscedasticity.

This can be assessed by plotting the residuals against the explanatory variables included in the model to see whether there is evidence of variability increasing, or decreasing, with the value of the explanatory variable.

7.2 Residuals vs. Fitted Values

A further implication of the assumptions made in defining the linear regression model is that the residuals $\epsilon_1, \dots, \epsilon_n$ are independent of the fitted (predicted) values $\hat{\mu}_1, \dots, \hat{\mu}_n$, as we prove in the following.

Recall the model assumption that

$$\mathbf{Y} \sim \mathrm{MVN}_n \left(\mathbf{X}\underline{\beta}, \sigma^2 \mathbf{I}_n \right).$$

Then the fitted values and estimated residuals are defined as

$$\underline{\hat{\mu}}(\mathbf{Y}) = \mathbf{X}\underline{\hat{\beta}}(\mathbf{Y}) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y} \tag{7.1}$$

and

$$\underline{\hat{\epsilon}}(\mathbf{Y}) = \mathbf{Y} - \underline{\hat{\mu}}(\mathbf{Y}) = \mathbf{Y} - \mathbf{H}\mathbf{Y}, \tag{7.2}$$

where we define

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.$$

Remark: Since both $\underline{\hat{\mu}}(\mathbf{Y})$ and $\underline{\hat{\epsilon}}(\mathbf{Y})$ are both functions of $\mathbf{Y}$, they are themselves random variables. This means that they have sampling distributions. We focus on the joint behaviour of the two random variables.

To show that the fitted values and estimated residuals are independent, we show that the vectors $\underline{\hat{\mu}}(\mathbf{Y})$ and $\underline{\hat{\epsilon}}(\mathbf{Y})$ are orthogonal, i.e., that they have a product of zero.

By definition of $\underline{\hat{\mu}}(\mathbf{Y})$ and $\underline{\hat{\epsilon}}(\mathbf{Y})$,

$$\begin{aligned}\underline{\hat{\mu}}(\mathbf{Y})^T\underline{\hat{\epsilon}}(\mathbf{Y}) &= (\mathbf{HY})^T(\mathbf{Y} - \mathbf{HY})\\ &= \mathbf{Y}^T\mathbf{H}^T(\mathbf{Y} - \mathbf{HY})\\ &= \mathbf{Y}^T\mathbf{H}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{H}^T\mathbf{HY}\\ &= \mathbf{Y}^T\mathbf{H}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{H}^T\mathbf{Y},\end{aligned}$$

since $\mathbf{H}^T\mathbf{H} = \mathbf{H} = \mathbf{H}^T$. So $\underline{\hat{\mu}}(\mathbf{Y})^T\underline{\hat{\epsilon}}(\mathbf{Y}) = 0$. This result uses the identities $\mathbf{H}^T\mathbf{H} = \mathbf{H}$, i.e., $\mathbf{H}$ is idempotent, and $\mathbf{H}^T = \mathbf{H}$.

Exercise 7.1.

Can you prove these identities above? In other words, can you prove that $\mathbf{H}$ is idempotent and $\mathbf{H}^T = \mathbf{H}$, for $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$?

Since $\mathbf{H}$ is idempotent, when applied to its image, the image remains unchanged. In other words, $\mathbf{H}$ maps $\underline{\hat{\mu}}$ to itself. Mathematically, $\mathbf{H}$ can also be thought of as a projection. The matrix $\mathbf{H}$ is often referred to as the hat matrix, since it transforms the observations $\mathbf{y}$ to the fitted values $\underline{\hat{\mu}}$.

A sensible diagnostic to check the model fit is to plot the residuals against the fitted values and check that these appear to be independent:

$$\{(\hat{\mu}_i, \hat{\epsilon}_i) \ : \ i = 1, \dots, n\} .$$

Example 7.2 (Brain weights continued).

For the fitted brain weight regression model described in Example 7.1, a plot of the residuals against the fitted values is shown below.

```
plot(L1$fitted.values, L1$residuals, pch = 19,
     xlab = "Fitted Values", ylab = "Estimated Residuals")
Resid.lm = lm(L1$residuals ~ L1$fitted.values)
abline(a = Resid.lm$coefficients[1], b = Resid.lm$coefficie
```

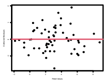

Figure 7.3: Residual vs. fitted values for the brain weight model. Straight lines show linear relationships, which is negligible.

The horizontal line indicates the line of best fit through the scatter plot. The correlation between the fitted values and residuals is

```
cor(L1$fitted.values, L1$residuals)
```

```
## [1] -5.191e-17
```

In this case, there is clearly no linear relationship between the residuals and fitted values and so, by this criterion, the model is a good fit.

7.3 Residuals vs. Explanatory Variables

For a well fitting model the residuals and the explanatory variables should also be independent. We can again prove this by showing that the vector of estimated residuals is independent of each of the explanatory variables. In other words, each column of the design matrix $\mathbf{X}$ is orthogonal to the vector of estimated residuals $\underline{\hat{\epsilon}}$.

Therefore, we need to show that

$$\mathbf{X}^T\underline{\hat{\epsilon}}(\mathbf{Y}) = 0.$$

Using the definition of the vector of estimated residuals in equation (7.2),

$$\begin{aligned}\mathbf{X}^T\underline{\hat{\epsilon}}(\mathbf{Y}) &= \mathbf{X}^T(\mathbf{Y} - \mathbf{H}\mathbf{Y})\\ &= \mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{H}\mathbf{Y}\\ &= \mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{Y}\\ &= 0.\end{aligned}$$

The penultimate step uses the result $\mathbf{X}^T\mathbf{H} = \mathbf{X}^T$, since, on substitution of the definition of $\mathbf{H}$,

$$\mathbf{X}^T\mathbf{H} = \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}^T.$$

Example 7.3 (Brain weights continued).

Let's plot the estimated residuals from the fitted brain weight regression model in Example 7.2 against the explanatory variable, the log of body weight.

```
plot(log(sleep$BodyWt), L1$residuals, xlab = "log(Body Weig
Resid.lm = lm(L1$residuals ~ log(sleep$BodyWt))
abline(Resid.lm$coefficients, col = 2)
```

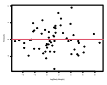

Figure 7.4: Residual vs. explanatory variable for the brain weight model. Straight lines show linear relationships, which is negligible.

The horizontal line is the line of best fit through the scatter plot, again indicating no linear relationship between the explanatory variable and the residuals. This is verified by the correlation

```r
cor(log(sleep$BodyWt), L1$residuals)
```

```
## [1] -6.755e-18
```

7.4 Outliers

An outlier is an observed response which does not seem to fit in with the general pattern of the other responses. Outliers may be identified using

- A simple plot of the response against the explanatory variable.
- Looking for unusually large residuals.
- Calculating standardized / studentized residuals.

The standardized (internally studentized) Pearson residual for observation i is defined as

$$s_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - \mathbf{H}_{i,i}}},$$

where $\mathbf{H}_{i,i}$ is the i-th element on the diagonal of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. The term $\hat{\sigma}\sqrt{1-\mathbf{H}_{i,i}}$ comes from the sampling distribution of the estimated residuals.

Remark: The diagonal terms $\mathbf{H}_{i,i}$ are referred to as the leverages. This name comes about since, as $\mathbf{H}_{i,i}$ gets closer to one, so the fitted value $\hat{\mu}_i$ gets closer to the observed value y_i. That is an observation with a large leverage will have a considerable influence on its fitted value, and consequently on the model fit.

We then like to test

$$H_0 : \text{Observation } i \text{ is not an outlier} \qquad \text{vs.} \qquad H_1 : \text{Observation } i \text{ is an out}$$

The test statistic for this test is the (externally) studentized Pearson residual

$$t_i = s_i\sqrt{\left(\frac{n-p-1}{n-p-s_i^2}\right)}.$$

These residuals are compared to the t-distribution with $n-p-1$ degrees of freedom. We test assuming a two-tailed alternative. If the test is significant, there is evidence that observation i is an outlier.

An alternative definition of t_i is based on fitting the regression model, without using observation i. This model is then used to predict the observation y_i, and the difference between the observed and predicted values is calculated. If this difference is small, the observation is unlikely to be an outlier as it can be predicted well using only information from the model and the remaining data.

The above discussions focus on identifying outliers, but don't specify what should be done with them. In practice, we should attempt to find out why the observation is an outlier. This reason will indicate whether the observation can safely be ignored (e.g. it occurred due to measurement error) or whether some additional term should be included in the model to explain it.

Example 7.4 (Atmospheric pressure).

Weisberg (2005)[3], p.4, presents data from an experiment by the physicist James D. Forbes (1857) on the relationship between atmospheric pressure and the temperature at which water boils.

[3] Weisberg, S. (2005), Applied Linear Regression, third edn, John Wiley and Sons, Inc., New York.

The 17 observations, and fitted linear regression model, are plotted below in Figure 7.5.

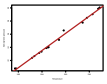

Figure 7.5: Atmospheric pressure against the boiling point of water, with fitted regression line.

Are any of the observations outliers?

Let's plot the estimated residuals against the recorded temperature.

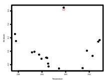

Figure 7.6: Residuals from the fitted model against temperature.

The plot suggests that observation $\mathbf{12}$ might be an outlier, since its residual is much larger than the rest (e.g., $\hat{\epsilon}_{12} = \mathbf{0.65}$).

To calculate the standardized residuals, we first set up the design matrix $\mathbf{X}$ and calculate the hat matrix $\mathbf{H}$,

```
presslm = lm(Pressure ~ Temp, data = pressure)
n = length(pressure$Temp)
X = cbind(1, pressure$Temp)
H = X %*% solve(t(X) %*% X) %*% t(X)
H[12, 12]
```

```
## [1] 0.06393
```

We also need the residual standard error:

```
summary(presslm)$sigma
```

```
## [1] 0.2328
```

From the **summary** command we see that the estimated residual standard error $\hat{\sigma}$ is 0.2328 and

```
presslm$residuals[12]
```

```
##     12
## 0.6499
```

gives the residual $\hat{\epsilon}_{12} = 0.65$.

Combining these results, the standardized residual is

$$s_{12} = \frac{\hat{\epsilon}_{12}}{\hat{\sigma}\sqrt{1 - H_{12,12}}}$$

$$= \frac{0.65}{0.2328 \times \sqrt{1 - 0.0639}}$$

$$= 2.89.$$

Since $n = 17$ and $p = 2$, the test statistic (studentized residual) is

$$t_{12} = 2.89\sqrt{\left(\frac{17 - 2 - 1}{17 - 2 - (2.89^2)}\right)} = 4.18.$$

The p-value to test whether or not observation 12 is an outlier is then

```
2 * (1 - pt(q = 4.18, df = 14))
```

```
## [1] 0.0009259
```

which is 9.25×10^{-4} . Since this is extremely small, we conclude that there is evidence that observation 12 is an outlier.

Note that a simpler way of obtaining the standardized (respectively studentized) residuals is by using the **`rstandard`** or **`rstudent`** commands on the regression model object.

7.5 Influence

Outliers can have an unduly large influence on the model fit, but this is not necessarily the case. Conversely, some points which are not outliers may actually have a disproportionate influence on the model fit. One way to measure the influence of an observation on the overall model fit is to refit this model without the observation.

Cook's distance summarises the difference between the parameter vector $\underline{\widehat{\beta}}$ estimated using the full data set and the parameter vector $\underline{\widehat{\beta}}_{(i)}$ obtained using all the data except observation i.

The formula for calculating Cook's distance for observation i is

$$D_i = \frac{s_i^2}{p} \frac{\mathbf{H}_{i,i}}{1 - \mathbf{H}_{i,i}}$$

where s_i is the previously defined standardized residual.

It is not straightforward to derive the sampling distribution for this test statistic. Instead it is common practice to follow the following guidelines:

1. First, look for observations with large D_i , since if these observations are removed, the estimates of the model parameters will change considerably.

2. If D_i is considerably less than 1 for all observations, none of the cases have an unduly large influence on the parameter estimates.

3. For every influential observation identified, the model should be refitted without this observation and the changes to the model noted.

Example 7.5 (Atmospheric pressure continued).
We calculate Cook's distance for the outlying observation (number

12). From the previous example $s_{12} = 2.89$, $H_{12,12} = 0.0639$ and $p = 2$. Therefore,

$$D_{12} = \frac{2.89^2 \times 0.0639}{2 \times (1 - 0.0639)} = 0.285.$$

Since this is reasonably far from 1 , we conclude that whilst observation 12 is an outlier, it does not appear to have an unduly large influence on the parameter estimates.

Remark: We can alternatively obtain the cooks distances by using the **`cooks.distance`** function:

```r
cooks.distance(presslm)
```

Observation \$i\$	Cook's Distance
1	0.0627
2	0.1883
3	0.0003
4	0.0001
5	0.0024
6	0.0104
7	0.0064
8	0.0069
9	0.0343
10	0.0229
11	0.0411
12	0.2843
13	0.0001
14	0.0894
15	0.0112
16	0.0631
17	0.0918

Part II

Generalized Linear Models

8 Generalized Linear Models: Introduction

8.1 Limitations of the Linear Model

So far, we have focused on (normal) linear regression to model continuous response data. In other words, we have assumed that the residuals after fitting a linear model are mutually independent and follow a normal distribution. Recall from the specification of the linear model, we have

$$Y_i = \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} + \epsilon_i, \quad \epsilon_i \sim \text{Normal}(0, \sigma^2), \qquad i = 1, \ldots, n,$$

which we could rewrite as

$$Y_i \sim \text{Normal}(\mu_i, \sigma^2), \qquad \mu_i = \mathbb{E}(Y_i) = \sum_{j=1}^{p} \beta_j x_{i,j} = \mathbf{x}_i^{\mathrm{T}} \underline{\beta}.$$

But what do we do if the response variables, $Y_1, \ldots, Y_n$, are not continuous?

More generally, when doing statistical modelling, there are some restrictions in the normal linear model that we may want to relax, or additional restrictions we may want to include, as in

- Non-normality of the residuals.
- Response variable is bounded by nature.
- Residual variance changes across observations.
- Non-linearity of the relationship between response and explanatory variables.

In this part of the course, we will generalize the (normal) linear model into a broader class of models that allows more flexibility in the type of response data which can be analysed and more flexibility in the relationship between the explanatory variables and response variable.

Recall that apart from continuous, we introduced three other types of response variables in Chapter 1:

- Count: The variable takes a positive whole number and represents a count of some phenomenon.

- Categorical: The variable takes a positive whole number and represents quality or preference.

- Binary: These are usually represented by 0 and 1 and could correspond to the presence or absence of some condition.

We can refine this classification further. For instance, counts may be considered bounded or unbounded. One particular subclass of continuous random variables are time-to-event responses, which represent the time to a particular event happening. These times are not usually normally distributed as they can only take positive values. There are many more cases we could consider, but we will focus on the types and distributions listed below:

Type of Response	Distribution
Continuous	Normal
Count (unbounded)	Poisson
Count (bounded)	Binomial
Categorical	Categorical
Binary	Bernoulli
Time-to-Event	Exponential, Gamma

8.2 Motivating Example: Beetle Mortality

In an early insecticide study, different groups of beetles were exposed to varying doses of insecticide for a fixed period of time. Within each group the total number of beetles was recorded and the number of beetles of those which had died due to the insecticide.

Let y_i denote the number of beetles killed in group i with m_i beetles. We are interested in the effectiveness of the insecticide and thus derive the proportion $p_i = y_i/m_i$ of beetles killed in each group.

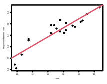

Figure 8.1: Dose against estimated probabilities of beetle mortality. The red line is the estimated linear normal model.

Examining the data and linear model fit, there might be evidence of a non-linear relationship between dose and the probability of death. A linear (normal) model is not suitable for the data because:

- The outcomes $\{p_i\}_{i=1}^{n}$ in the plot above represent proportions. Thus, any estimates of the regression coefficients must constrain the fitted values to lie between zero and one for the model to make sense (i.e., they are bounded, $p_i \in [0,1]$). However, some of the fitted values potentially lie outside of this range and hence this is not a satisfactory model.

- In the original (count) data, there are different sample sizes for each group and this needs to be taken into account.

- The observed proportions are unlikely to be normally distributed with constant variance. In fact, the count data are likely to be binomially distributed. Recall that for a random variable $Y \sim \text{Binomial}(m, p)$, we have $\text{Var}(Y) = mp(1-p)$. Hence the variance of Y_i, the number of dead beetles, depends on the number of beetles in the i-th sample.

There is also a contextual issue. In many settings it seems implausible that the expected outcome would change strictly linearly over the entire range of a continuous explanatory variable. For the beetle mortality example, the effect should be zero in the absence of any insecticide and increase to a maximum level, possibly corresponding to the proportion of the sample susceptible to the toxic effect, with increasing dose.

8.2.1 Transformation of the Response Variable

Given the non-linear relationship, we can consider a data transformation to capture this relationship. In particular, we use the logit transformation, the inverse of the logistic function (in machine learning jargon, an example of sigmoid function), which is widely used to map probabilities to the interval $(-\infty, \infty)$.

Formally,

$$
\begin{aligned}
\operatorname{logistic}(x) &= \frac{\exp(x)}{1+\exp(x)}, \qquad x \in (-\infty, \infty), \\
\operatorname{logit}(x) &= \log\left(\frac{x}{1-x}\right), \qquad x \in (0,1).
\end{aligned}
$$

The following plots illustrate that these functions are S-shaped.

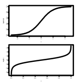

Figure 8.2: Comparison of logistic and logit transformations.

We apply the logit transformation to the beetle mortality data and obtain the following data plot (see Figure 8.3).

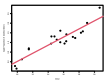

Figure 8.3: Dose against estimated probabilities of beetle mortality. The red line is the estimated linear normal model.

While the plot looks similar to the original data plot, we no longer have to consider that predicted values may lie outside a certain interval, as the estimated proportion is derived by applying the logistic transformation to the predicted value. However, the other issues are not resolved and thus we should not use a normal linear regression model.

IMPORTANT: We should include the transformation in the statistical model.

Remark: Another popular choice for mapping probabilities to the $(-\infty, \infty)$ interval is the probit transformation

$$\text{probit}(x) \;=\; \Phi^{-1}(x), \qquad x \in (0,1),$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function (quantile function) of the standard normal distribution (see Figure 8.4).

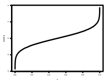

Figure 8.4: Probit transformations.

8.2.2 Logistic Regression

The plot of the transformed proportions indicates a linear relationship between $\text{logit}(y_i)$ and the dose x_i. We include the logit transformation in our model and define

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_1 + \beta_2 x_i.$$

We further noted that the number of beetles killed is likely to be binomially distributed. Combining these aspects results in a logistic regression model that can be split up into three different components:

- The linear predictor $\eta_i = \beta_1 + \beta_2 x_i$.

- The link function $\log\left(\frac{p_i}{1-p_i}\right) = \eta_i$ between the mean and the predictor.

- The distribution of the observations, $Y_i \sim \text{Binomial}(m_i, p_i)$.

We observe Y_i and m_i and so we have an observed probability for each group; the fitted p_i is estimated from the model.

8.3 The Generalized Linear Model (GLM)

We now extend the ideas above to introduce the generalized linear model (GLM) framework.

A GLM is generally defined by three components:

- $\eta_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \quad + \beta_p x_{i,p} = \mathbf{x}_i^T \underline{\beta}$

  A linear predictor consisting of regression coefficients and explanatory variables. If $x_{i,1} = 1$ for all $i = 1, \ldots, n$, the predictor includes an intercept term β_1. The linear predictor is also known as the systematic component.

- $g(\mu_i) = \eta_i$

  A link function mapping the linear predictor to the mean of the distribution.

- $Y_i \sim F(\mu_i)$

  A probability distribution from the exponential family, describing the observed data. The distribution may also have an unknown scale parameter ϕ. This is termed the random component.

The main idea is that changes in the explanatory variables in $\mathbf{x}_i^T\underline{\beta}$ lead to changes in the response Y_i . The relationship depends on fixed values of the parameter $\underline{\beta}$, the regression coefficients (common to all observations). The relationship is mediated through the linear predictor, η , which is one dimensional, and “aggregates” the separate variables in $\mathbf{x}_i^T$. The linear predictor has a one-to-one mapping with the mean response μ_i , through the link function, so that changes in $\mathbf{x}_i^T\underline{\beta}$ affect Y_i only through changing the μ_i parameter. The specific relationship between μ_i and Y_i depends on the particular distribution, F . This dependence is shown in the Figure 8.5 below.

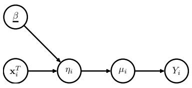

Figure 8.5: Schematic illustration of the dependence between the components of a GLM.

All of the types of response variables described at the end of Section 8.1 fall into a class of distributions which we can model using the generalized linear model framework. Specifically, all of the distributions listed in Section 8.1 are examples from a family of distributions known as the exponential family, which we consider in Section 8.3.2.

8.3.1 Link functions

The link function can be any smooth monotonic function and be selected based on a range of criteria, such as theory, experience, the data or the chosen model. In particular, there are a wide variety of link functions, each suitable for different contexts and data:

Name	Form
‘identity‘	$\mu_i = \eta_i$
‘logarithmic‘	$\log(\mu_i) = \eta_i$
‘reciprocal‘	$1/\mu_i = \eta_i$
‘square‘	$\mu_i^2 = \eta_i$
‘logit‘	$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i$
‘probit‘	$\Phi^{-1}(\mu_i) = \eta_i$
‘complementary log-log‘	$\log[-\log(1-\mu_i)] = \eta_i$

As above, $\eta_i = \beta_1 x_{i,1} + \beta_2 x_{2,1} + \quad + \beta_p x_{i,p}$ is the linear predictor, μ_i is the mean of observation i and Φ is the cumulative distribution function of the standard normal distribution.

We now explore examples of some common GLMs and link functions.

Example 8.1 (Normal distribution).
Suppose $Y_i \sim \text{Normal}(\mathbf{x}_i^T \underline{\beta}, \sigma^2)$, $i = 1, \ldots, n$ independently. Here, the link function is $g(\cdot) = \mathbf{1}(\cdot)$ (the identity), i.e., $\mu_i = \eta_i$.

Example 8.2 (Binomial distribution).
Suppose $Y_i \sim \text{Binomial}(m_i, p_i)$, $m_i \in \mathbb{N}$, $i = 1, \ldots, n$ independently. In other words, data from this model $\{y_i\}$ are observed counts, each having an observation-specific total, m_i . Here, p_i is a function of the linear predictor, $p_i = h(\eta_i) = h(\mathbf{x}_i^T \underline{\beta})$.

The link function is

$$g(\mu_i) = h^{-1}\left(\frac{\mu_i}{m_i}\right) = h^{-1}(p_i),$$

since we have

$$\mu_i = \mathbb{E}(Y_i) = m_i p_i = m_i h(\mathbf{x}_i^T \underline{\beta}) = m_i h(\eta_i).$$

Note, whilst of the same form, the link functions are actually different for each i (since the m_i are potentially different). This is not an issue. Examples of h commonly used are:

a. $h(\eta) = \Phi(\eta)$, where Φ is the distribution function of the standard normal distribution. The link function is the probit link function.

b. $h(\eta) = 1 - \exp[-\exp(\eta)]$, and so $\eta = h^{-1}(p) = \log[-\log(1-p)]$. The link function is the complementary log-log link function.

c. $h(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$, so that $h^{-1}\left(\frac{\mu}{m}\right) = \log\left(\frac{\mu}{m-\mu}\right)$. This is the logit link function.

We have met the binomial model with the logit link function before, when we looked at the beetle data. The logistic model allows for a smooth change in risk throughout the range of x, and has the property that risk increases slowly up to a threshold range of x, followed by a more rapid increase and a subsequent levelling off of risk (see the plots of the logistic and logit functions in Figure 8.2).

This particular choice of model and link is sometimes called the proportional odds model. This is because

$$\eta = h^{-1}(p) = \log\left(\frac{p}{1-p}\right)$$

means that additive increments to each explanatory variable x_j giving additive increments to $\eta_i = \mathbf{x}_i^T \underline{\beta}$ result in equal multiplicative increments to the odds, $\frac{p}{1-p}$. More specifically, we can write

$$\frac{p(x)}{1-p(x)} = \exp(\eta) = \exp(\beta_1 + \beta_2 x). \tag{8.1}$$

If we were then to compare two groups for example (with a factor variable s, e.g., gender), we would include this in the model through $\eta = \eta(s, x) = \gamma_s + \beta_1 + \beta_2 x$ and thus have

$$\frac{p(s,x)}{1-p(s,x)} \div \frac{p(s',x)}{1-p(s',x)} = \frac{\exp(\eta(s,x))}{\exp(\eta(s',x))} = \exp(\gamma_s - \gamma_{s'}) \quad \text{for all } x,$$

i.e., we have proportional odds – the odds ratio does not depend on x.

8.3.2 The Exponential Family

In a generalized linear mode, we generally assume that the distribution F (the random component) belongs to the exponential family.

Definition 8.1 (Exponential family).

The random variable Y has a distribution in the exponential family if its (probability) density function can be written in the canonical form

$$f(y \mid \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}, \tag{8.2}$$

where θ and ϕ are parameters (the canonical and scale parameters respectively), $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are specified functions, and we assume that the range of Y doesn't depend on the parameters.

Example 8.3.

Suppose $Y \sim \text{Normal}(\theta, \sigma^2)$ with density

$$f(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}.$$

We can rewrite the density as

$$f(y \mid \mu, \sigma^2) = \exp\left\{-\frac{y^2 - 2y\mu + \mu^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}\right\}.$$

Consequently, the density is of the form required in Definition 8.1 with parameters $\theta = \mu$ and $\phi = \sigma^2$, $b(\theta) = \frac{\theta^2}{2} = \frac{\mu^2}{2}$, $a(\phi) = \phi = \sigma^2$ and $c(y, \phi) = -\frac{y^2}{2\phi} - \log\sqrt{2\pi\phi}$.

It is difficult to see the interpretation of the functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. However recalling that (when viewed as a random variable) for a log-likelihood $\ell(\theta \mid Y) = \log L(\theta \mid Y)$, we have

$$\mathbb{E}_\theta\left(\frac{\partial \ell(\theta \mid Y)}{\partial \theta}\right) = 0$$

and

$$\mathrm{Var}_\theta\left(\frac{\partial \ell(\theta \mid Y)}{\partial \theta}\right) = \mathbb{E}_\theta\left(-\frac{\partial^2 \ell(\theta \mid Y)}{\partial \theta^2}\right),$$

i.e., these are moments of the score function, $U(\theta) = \frac{\partial \ell(\theta|Y)}{\partial \theta}$. Using the form of the exponential family of densities

$$\frac{\partial \ell(\theta \mid Y)}{\partial \theta} = \frac{\partial}{\partial \theta}\left\{\frac{Y\theta - b(\theta)}{a(\phi)} + c(Y, \phi)\right\} = \frac{Y - b'(\theta)}{a(\phi)}$$

and

$$\frac{\partial^2 \ell(\theta|Y)}{\partial \theta^2} = \frac{-b''(\theta)}{a(\phi)},$$

where we have treated ϕ as a constant.

Hence, taking expectations and solving, the mean and variance of the distribution can be derived as

$$\mathbb{E}(Y) = b'(\theta) \qquad \text{and} \qquad \text{Var}(Y) = a(\phi)b''(\theta).$$

Note that the variance depends on both the mean $b'(\theta)$ and the scale parameter ϕ through $a(\phi)$. We can choose $V(\mu)$ such that $\text{Var}(Y) = a(\phi)V(\mu)$; this function $V(\mu)$ is called the variance function and it describes each distribution by its mean-variance relationship (it represents the part of the variance which depends on θ and therefore μ). For the examples we consider in this course, $a(\phi)$ has the form $a(\phi) = \phi/w$, so that $\text{Var}(Y) = \phi b''(\theta)/w$, and the variance function is $V(\mu) = b''(\theta)/w$.

The function of μ_i obtained by inverting $\mu_i = \mathbb{E}(Y_i) = b'(\theta_i)$ is a commonly-used choice of g. Such a g is called the canonical link function. Notice that with the choice of the canonical link function, the canonical parameter θ_i is equal to the linear component η_i, that is, $\theta_i = g(\mu_i) = \eta_i$, $i = 1, \dots, n$, so is mathematically convenient and results in some nice properties for parameter estimation. The following table summarizes the canonical link functions for a range of distributions.

Notation	Normal(μ, σ^2)	Poisson(μ)	Bernoulli(μ)	Gamma(μ, ϑ)
pdf/pmf:	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$	$\frac{\mu^y}{y!}\exp(-\mu)$	$\mu^y(1-\mu)^{1-y}$	$\frac{1}{\Gamma(\vartheta)}\left(\frac{\vartheta}{\mu}\right)^{\vartheta} y^{\vartheta-1}e^{-y\vartheta/\mu}$
support:	$\mathbb{R}$	$\{0, 1, ...\}$	$\{0, 1\}$	$\mathbb{R}_+ = (0, \infty)$
θ:	μ	$\log(\mu)$	$\mathrm{logit}(\mu)$	$-1/\mu$
ϕ:	σ^2	1	1	$1/\vartheta$
$b(\theta)$:	$\theta^2/2$	$\exp(\theta)$	$\log[1+\exp(\theta)]$	$-\log(-\theta)$
$\mu = b'(\theta)$:	θ	$\exp(\theta)$	$\exp(\theta)/[1+\exp(\theta)]$	$-1/\theta$
$V(\mu)$:	1	μ	$\mu(1-\mu)$	μ^2
canonical link:	$\eta = \mu$	$\eta = \log(\mu)$	$\eta = \mathrm{logit}(\mu)$	$\eta = -1/\mu$
‘R‘ link name:	‘gaussian‘	‘poisson‘	‘binomial‘	‘gamma‘

9 Generalized Linear Models: Estimation

In Chapter 3, we used least squares to estimate the regression coefficients $\underline{\beta}$ of a (normal) linear models, i.e., the estimate $\underline{\beta}$ minimizes the sum of squares $S(\underline{\beta})$. However, in the GLM setting, the observation y_i and the predicted value $\mathbf{x}_i^T\underline{\beta}$ are on different scales – we have not included the link function $g(\cdot)$, which maps $\mu = \mathbb{E}(Y)$ onto the scale of the linear predictor η. Additionally, $Y_1, \dots, Y_n$ may have different variances, so we need a different approach to estimate $\underline{\beta}$. For all practical purposes, we will henceforth assume that $a_i(\phi) = \phi/w_i$, i.e., the scale parameter is the same across observations, up to proportionality (the constants w_i are assumed known).

Because of these issues, we will base model fitting in GLMs on the principle of maximum likelihood (recall that for the linear model, the method of least squares fitting is equivalent to maximum likelihood). Since the probability distribution $F(\cdot)$ in the GLM framework belongs to the exponential family, we have

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

Due to independence between observations, the log-likelihood is

$$\ell(\underline{\theta}, \phi \mid y_1, \dots, y_n) = \sum_{i=1}^{n} \log f(y_i \mid \theta_i, \phi)$$

where for convenience of notation, we will sometimes denote the contribution to the log-likelihood of individual i by ℓ_i, e.g., $\ell(\underline{\theta}, \phi \mid y_1, \dots, y_n) = \sum_{i=1}^{n} \ell_i$.

We can write the log-likelihood function as

$$\ell(\underline{\theta}, \phi \mid y_1, \dots, y_n) = \sum_{i=1}^{n} \left\{ w_i \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}$$

$$= \frac{1}{\phi} \left\{ \sum_{i=1}^{n} w_i [y_i\theta_i - b(\theta_i)] \right\} + \sum_{i=1}^{n} c(y_i, \phi). \tag{9.1}$$

Recall from Section 8.3 that $\mathbb{E}(Y_i) = b'(\theta_i)$. Since $\mathbb{E}(Y_i)$ is described via the linear predictor η_i, the log-likelihood $\ell(\underline{\theta}, \phi \mid y_1, \dots, y_n)$ can also be expressed in terms of the regression coefficient $\underline{\beta}$, and we denote this log-likelihood function as $\ell(\underline{\beta}, \phi \mid y_1, \dots, y_n)$.

We would like to maximize this likelihood to obtain estimates for $\underline{\beta}$. However, except in special cases (e.g., the Gaussian density), the likelihood cannot be maximized analytically. Hence, we will have to resort to numerical optimization methods, specifically iterative procedures to find a solution. The Iteratively Re-Weighted Least Squares (IRWLS) algorithm was developed to maximize the likelihood in GLMs. Note however, that the dispersion parameter ϕ appears outside of the bracket involving $\underline{\beta}$, and so maximization does not depend on knowing ϕ.

9.1 Newton-Raphson Numerical Method

The algorithm we will use to maximize the log-likelihood function above will be based on the Newton-Raphson numerical method, which attempts to find a solution to the equation $f(x) = 0$ by iteratively using gradient computations to move towards the optimal solution.

After an initial guess, $x^{(1)}$, at each iteration the Newton-Raphson method proposes an updated estimate of the solution $x^{(m+1)}$ with

$$x^{(m+1)} = x^{(m)} - \frac{f(x^{(m)})}{f'(x^{(m)})}$$

until convergence up to some tolerance (e.g. $|f(x)| < \varepsilon$). The following figure illustrates this update of $x^{(m)}$.

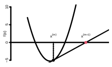

Figure 9.1: An illustration of the Newton-Raphson procedure.

In our case, to maximize the likelihood (e.g., for a single parameter) we would need to solve the equation $U(\theta) = 0$, where $U(\cdot)$ is the score function (the first derivative of the log-likelihood), and so the update step takes the form

$$\theta^{(m+1)} = \theta^{(m)} - \frac{U(\theta^{(m)})}{U'(\theta^{(m)})}.$$

To take into account that $U'(\cdot)$ may vary with data, it is common to replace $U'(\theta)$ by its expectation $\mathbb{E}[U'(\theta)]$, quantified through the information as

$$\mathcal{I} := \mathbb{E}[-U'(\theta)] = \mathbb{E}\left[-\sum_{i=1}^{n} U_i'(\theta)\right] = -\sum_{i=1}^{n} \mathbb{E}\left[U_i'(\theta)\right],$$

where $U_i(\theta)$ is the score of the i th observation, giving

$$\theta^{(m+1)} = \theta^{(m)} + \frac{U(\theta^{(m)})}{\mathcal{I}^{(m)}}. \tag{9.2}$$

This form is called Fisher scoring (since it involves the Fisher information, $\mathcal{I}$).

9.2 Fisher Scoring for GLMs

In reality, we are interested in obtaining estimates (MLEs) for the regression coefficients $\underline{\beta}$. Consider the log-likelihood in (9.1). Taking the derivative of the log-likelihood $\ell(\underline{\beta}, \phi \mid y_1, \dots, y_n)$ with respect to the j -th regression coefficient β_j ,

$$\frac{\partial \ell(\underline{\theta}, \phi \mid y_1, \dots, y_n)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

Since

$$\frac{\partial \ell_i}{\partial \theta_i} = w_i \frac{y_i - \mu_i}{\phi},$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = w_i V(\mu_i),$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij},$$

the derivative simplifies to

$$\frac{\partial \ell(\underline{\theta}, \phi \mid y_1, \dots, y_n)}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} w_i (y_i - \mu_i) \left[w_i V(\mu_i)\right]^{-1} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

$$= \frac{1}{\phi} \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}.$$

Setting these quantities to zero, this results in the likelihood equations

$$U(\underline{\beta}) := \left(\frac{\partial \ell(\underline{\beta}, \phi \mid y_1, \dots, y_n)}{\partial \beta_1}, \dots, \frac{\partial \ell(\underline{\beta}, \phi \mid y_1, \dots, y_n)}{\partial \beta_p} \right) = \frac{1}{\phi} \sum_{i=1}^{n} \frac{(y_i - \mu}{V(\mu_i)}$$

Note that this can be written

$$\frac{1}{\phi} \sum_{i=1}^{n} (y_i - \mu_i) W_{ii} \frac{\partial \eta_i}{\partial \mu_i} \mathbf{x}_i = 0, \tag{9.3}$$

with

$$W_{ii} = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \Big/ \, V(\mu_i) = \frac{1}{V(\mu_i) g'(\mu_i)^2}.$$

Similarly, differentiating $U(\underline{\beta})$ again, the Fisher information matrix $\mathcal{I} = \left[\mathcal{I}_{jk}\right]_{j,k=1,\dots,p}$ is

$$\left[-\mathbb{E}\left(\frac{\partial^2 \ell(\underline{\beta}, \phi \mid y_1, \dots, y_n)}{\partial \beta_j \partial \beta_k}\right)\right]_{j,k=1,\dots,p} = \left[\frac{1}{\phi}\sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right] = \frac{1}{\phi}$$

where $\mathbf{W}$ is a diagonal matrix with diagonal entries $\{W_{ii}\}_{i=1}^n$ and so for a multivariate parameter vector, $\underline{\beta}$, the Fisher scoring equation generalizes to

$$\underline{\beta}^{(m+1)} = \underline{\beta}^{(m)} + \left[\mathcal{I}^{(m)}\right]^{-1} \mathbf{U}\left(\underline{\beta}^{(m)}\right). \qquad (9.4)$$

In this equation, $\left[\mathcal{I}^{(m)}\right]^{-1}$ is the inverse of the Fisher information matrix. Multiplying by $\mathcal{I}^{(m)}$, we get

$$\mathcal{I}^{(m)} \underline{\beta}^{(m+1)} = \mathcal{I}^{(m)} \underline{\beta}^{(m)} + \mathbf{U}\left(\underline{\beta}^{(m)}\right),$$

which can be written as

$$\underline{\beta}^{(m+1)} = \underline{\beta}^{(m)} + (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{z}^*, \qquad (9.5)$$

with $z_i^* = (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) = (y_i - \mu_i) g'(\mu_i)$.

Alternatively, we can write

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\underline{\beta}^{(m+1)} = \mathbf{X}^T\mathbf{W}\mathbf{z}, \tag{9.6}$$

where $\mathbf{z}$ is a vector with entries

$$z_i = (y_i - \mu_i)g'(\mu_i) + \mathbf{x}_i^T\underline{\beta}^{(m)}.$$

The algorithm then takes the form of a Iteratively Re-Weighted Least Squares (IRWLS) (with diagonal weight matrix $\mathbf{W}$). In other words, we begin with an initial estimate $\underline{\beta}^{(0)}$ of $\underline{\beta}$, which is used (via the linear predictor η_i) to evaluate $\mathbf{z}^{(0)}$ and $\mathbf{W}$, and then solve the equation above to give $\underline{\beta}^{(1)}$. We see that $\underline{\beta}^{(1)}$ is given by

$$\underline{\beta}^{(m+1)} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z},$$

which correspond to a weighted least square estimate, which we can compute in **R** using the **lm** function. The process then iterates, each step obtaining $\underline{\beta}^{(m+1)}$ from $\underline{\beta}^{(m)}$. Note that we do not need to know the dispersion parameter ϕ to implement the procedure, since it is independent of the sum.

A suitable starting vector $\underline{\beta}^{(0)}$ can be derived by regressing $g(\mathbf{Y})$ on the explanatory variables (i.e., taking $\mu_i = y_i$ for all i in the linear predictor).

The algorithm can be summarized as follows:

1. Set initial estimates $\underline{\hat{\eta}}^{(0)}$ and $\underline{\hat{\mu}}^{(0)}$.

2. Compute the adjusted variable $\mathbf{z}^{(0)} = \underline{\hat{\eta}}^{(0)} + \left(\hat{\mathbf{y}} - \underline{\hat{\mu}}^{(0)}\right) \left.\frac{d\underline{\hat{\eta}}}{d\underline{\hat{\mu}}}\right|_{\underline{\hat{\eta}}^{(0)}}$.

3. Compute the weights $w_0^{-1} = \left.\left(\frac{d\underline{\hat{\eta}}}{d\underline{\hat{\mu}}}\right)^2\right|_{\underline{\hat{\eta}}^{(0)}} V\left(\underline{\hat{\mu}}^{(0)}\right)$.

4. Estimate $\underline{\hat{\beta}}^{(1)}$ using the weights to get $\underline{\hat{\eta}}^{(1)}$.

5. Iterate steps 2-4 until convergence (subject to some tolerance).

Remarks:

- The IRWLS algorithm is Fisher's scoring method and the resulting solutions correspond to maximum likelihood estimation.

- The algorithm consists of a number of simple steps using weighted least squares which are iterated to produce a sequence of parameter values.

- The iterations terminate when this sequence converges to a specified accuracy.

9.3 Illustration: Beetle Mortality

Recall the beetle data set in Section 8.2. We were interested in modelling the relationship between dose of an insecticide and beetle mortality:

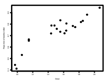

Figure 9.2: Dose against estimated probabilities of beetle mortality. The red line is the estimated linear normal model.

The data is of the form

```
##   dead alive  dose
## 1   33     2 20.68
## 2   24     7 17.51
## 3   30     4 19.05
## 4   26     7 14.42
## 5   25     5 15.58
## 6   27     6 18.25
```

The **glm** function in **R** gives

```
glm(cbind(dead,alive) ~ dose, family = binomial, data = bee
```

```
##
## Call:  glm(formula = cbind(dead, alive) ~ dose, family =
##
## Coefficients:
## (Intercept)         dose
##      -2.059        0.207
##
## Degrees of Freedom: 19 Total (i.e. Null);  18 Residual
## Null Deviance:       59
## Residual Deviance: 8.43  AIC: 83.9
```

We now want to illustrate that the algorithm in Section 9.2 gives the same result. For a binomial response, we have

$$\eta = \log\left(\frac{\mu}{1-\mu}\right), \qquad \frac{\partial \eta}{\partial \mu} = \frac{1}{\mu(1-\mu)}, \qquad V(\mu) = \mu(1-\mu), \qquad w = m$$

Here, y_i is the proportion of beetles killed. We initialise the algorithm with $\mu_i = y_i$ and the first update step is

```
m = beetles$dead + beetles$alive
y = beetles$dead / m
mu = y
eta = log( mu / (1-mu) )
z = eta + (y-mu) / ( mu * (1-mu) )
w = m * mu * ( 1 - mu )
lmod = lm( z ~ dose, weights=w, data=beetles )
coef( lmod )
```

```
## (Intercept)        dose
##     -2.0215      0.2035
```

We are already close to the estimates provided by the **`glm`** function. Let's perform five more iterations:

```r
for( i in 1:5 ){
  eta = lmod$fit
  mu = exp(eta) / ( 1 + exp(eta) )
  z = eta + (y-mu) / ( mu * (1-mu) )
  w = m * mu * (1 - mu)
  lmod = lm(z ~ dose, weights = w, data = beetles)
  cat(">> Iteration", i, "\n", names(coef(lmod)), "\n", coe
}
```

```
## >> Iteration 1
##  (Intercept) dose
##  -2.058 0.2071
## >> Iteration 2
##  (Intercept) dose
##  -2.059 0.2072
## >> Iteration 3
##  (Intercept) dose
##  -2.059 0.2072
## >> Iteration 4
##  (Intercept) dose
##  -2.059 0.2072
## >> Iteration 5
##  (Intercept) dose
##  -2.059 0.2072
```

We see that the algorithm converges rapidly to the estimates obtained by the **glm** function.

9.4 Estimation of ϕ

Recall that we can, conditional on ϕ, obtain maximum likelihood estimates for $\underset{\sim}{\beta}$ using the IRWLS algorithm. We would like to estimate ϕ, similar to estimating the residual variance in linear models.

Given that $\text{Var}(Y_i) = \phi V(\mu_i)$, we have that

$$\frac{1}{\phi}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)} \sim \chi^2_{n-p} \text{ approximately,}$$

since this is a sum of squared independent, zero mean, unit variance random variables. Due to the expectation of a χ^2_{n-p} random variable being $(n-p)$, this suggests an estimator for ϕ as

$$\hat{\phi}_P = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)}. \tag{9.7}$$

This is the Pearson's chi-square statistic $\mathcal{X}^2$, scaled by the degrees of freedom. This is essentially the estimator obtained using the method of moments.

For the Poisson model, we have that $\phi = 1 = w_i$ and also that $V(\mu_i) = \mu_i$. Thus, the approximation is

$$\frac{1}{\phi}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{V(\mu_i)} = \sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i} = \sum_{i=1}^{n}\frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i},$$

which takes the more familiar form of the Pearson's chi-square goodness of fit statistic.

10 Inference and Model Selection for GLMs

10.1 Inference for Model Parameters

Recall from Chapter 1 that the maximum-likelihood estimator is asymptotically normal, in particular,

$$\hat{\underline{\theta}}\,(\mathbf{Y}) \sim \text{MVN}\left(\underline{\theta}\,, \mathcal{I}^{-1}(\underline{\theta})\right).$$

Using the form of the Fisher information matrix derived in Chapter 9, we have

$$\hat{\underline{\beta}}(\mathbf{Y}) \sim \mathrm{MVN}_p\left(\underline{\beta}, \phi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\right),$$

where $\mathbf{W}$ is a diagonal matrix as defined in Chapter 9 with $W_{ii} = \frac{1}{V(\mu_i)g'(\mu_i)^2}$. Similar to the linear regression model, we can use this distribution to form (approximate) confidence intervals or to perform hypothesis tests for model parameters. In particular,

- When the dispersion parameter ϕ is known (e.g, binomial and Poisson models), then the normal distribution can be used.

- When we have to estimate ϕ, then (similar in spirit to other confidence intervals in Chapter 4), we use the t-distribution.

Note that the variance matrix of the MLE vector depends on the unknown mean $\underline{\mu}$ (via the weight matrix $\mathbf{W}$), and so in practice we estimate $\underline{\mu}$ and use $V(\hat{\mu}_i)$ in place of $V(\mu_i)$ in $\mathbf{W}$.

In what follows we denote by σ_j^2 the j-th diagonal entry of the variance matrix, Σ, i.e., $\sigma_j = \sqrt{\phi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}_{j,j}}$.

10.2 Hypothesis Tests for Regression Coefficients

We want to test

$$H_0 : \beta_j = b \qquad \text{vs.} \qquad H_1 : \beta_j \neq b.$$

at a given significance level α.

If ϕ is known, we consider the test statistic

$$z = \frac{\hat{\beta}_j - b}{\hat{\sigma}_j}$$

and compare $|z|$ to $z_{1-\alpha/2}$, the $(1-\alpha/2)\times 100\%$ quantile of the standard normal distribution. The corresponding p-value is calculated as $p = 2\,\mathbb{P}(Z > |z|)$, where $Z \sim \text{Normal}(0,1)$, and we reject H_0 if $p < \alpha$.

The hypothesis can be alternatively tested via the confidence interval

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\hat{\sigma}_j^2} \tag{10.1}$$

Equivalently to the (normal) linear regression framework, H_0 is rejected at the $\alpha\%$ significance level if b does not lie in the $100(1-\alpha)\%$ confidence interval.

In the case where ϕ is unknown, we first estimate ϕ using the Pearson's estimator in Section 9.4. Substituting the estimate $\hat{\phi}_P$ into the equation for the variance matrix, we obtain the estimate $\hat{\Sigma} = \hat{\phi}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$ (recall that we used $\hat{\sigma}^2$ in Chapter 4 for hypothesis testing). The test statistic is then

$$t = \frac{\hat{\beta}_j - b}{\hat{\sigma}_j}$$

and we compare with appropriate quantiles of the t_{n-p} distribution. The p-value is $p = 2\,\mathbb{P}(T > |t|)$, where T is t_{n-p}-distributed.

The $(1-\alpha)\times 100\%$ confidence interval to test the hypothesis would be

$$\hat{\beta}_j \pm t_{n-p}(1-\alpha/2)\sqrt{\hat{\sigma}_j^2},$$

and we reject H_0 if b lies outside the confidence interval.

Example 10.1 (Beetle Mortality continued).

We used a logistic regression model in Chapters 8 and 9 to analyse the between dose and number beetles killed. Let's consider the outputs provided by the **summary** function in **R**.

```
beetle_glm = glm(cbind(dead,alive) ~ dose, family = binomia
summary(beetle_glm)
```

```
##
## Call:
## glm(formula = cbind(dead, alive) ~ dose, family = binomi
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.0589     0.4432   -4.65  3.4e-06 ***
## dose          0.2072     0.0301    6.89  5.6e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 59.0188  on 19  degrees of freedom
## Residual deviance:  8.4295  on 18  degrees of freedom
## AIC: 83.88
##
## Number of Fisher Scoring iterations: 4
```

Since the dispersion parameter $\phi = 1$ is known for the binomial model, we are provided with a z value (instead of a t value as in the linear regression model). We see that the p-value for β_2 is close to zero, and we thus deduce that there is enough evidence to conclude that dose is significant at the 5% significance level. A 95% confidence interval for β_2 is given by

$$\hat{\beta}_2 \pm z_{1-\alpha/2}\sqrt{\sigma_2^2} = 0.207 \pm (1.96 \times 0.03) = (0.148, 0.266).$$

10.3 A Measure of Fit for GLMs: the Residual Deviance

For model fitting, instead of minimizing a sum of squares, we have been maximizing a log-likelihood function $\ell(\underline{\theta}, \phi)$, where $g[b'(\theta_i)] = \mathbf{x}_i^T \underline{\beta}$. Recall that in linear models, we use the residual sum of squares function $S(\underline{\hat{\beta}})$ – the amount of variation which is not explained by the model – as a numerical summary of how well a linear model fitted; as more terms are added to the model, the residual sum of squares decreases. The case $S(\underline{\hat{\beta}}) = 0$ represented a perfect fit.

For GLMs we will use a different measure.

What is the best we can do for fitting a GLM?

Recall that for an exponential family model, an observation-specific contribution to the log-likelihood (i.e., the summand) was

$$\ell_i = w_i \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi),$$

and so differentiating with respect to the parameter θ_i and setting to zero, we have

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{\partial}{\partial \theta_i}\left\{ w_i \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\} = w_i \frac{y_i - b'(\theta_i)}{\phi} = 0$$

i.e., $b'(\theta_i) = y_i$. A model that achieves this equality for all $i = 1, \dots, n$ is called the saturated model: we will need one parameter for each observation. Denote this particular value of the parameter by $\tilde{\theta}_i$ for each i ($\tilde{\theta}_i = b'^{-1}(y_i)$). Typically the full (saturated) model is not particularly useful as it gives us no more information than the data itself.

At the other extreme of the modelling choice problem, if we set $\mu_i = \mu$ for all $i = 1, \dots, n$, we have the null model: there is only one parameter for all observations, and the model represents the data as random variation.

Exercise 10.1.
Write down the null model for the Gaussian GLM, and think about what the parameter estimates will look like.

Now for a particular i, consider the quantity

$$D_i^* = 2\left[\ell_i\left(\tilde{\theta}_i\right) - \ell_i\left(\hat{\theta}_i\right)\right] = \frac{2w_i}{\phi}\left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right) + b\left(\hat{\theta}_i\right)\right],$$

where $\hat{\theta}_i$ is the MLE for θ_i. Maximizing the likelihood is equivalent to minimizing this quantity. The quantity is zero only if there is a perfect fit to the i-th observation. In other words for the observed data $y_1, \dots, y_n$, $\ell\left(\underline{\tilde{\theta}}\right)$ provides a baseline value for the log-likelihood.

Definition 10.1 (Deviance).

Suppose we have a data model f with associated log-likelihood ℓ. Then the deviance of f is defined as

$$D = \sum_{i=1}^{n} D_i = 2\left[\ell\left(\underline{\tilde{\theta}}\right) - \ell\left(\underline{\hat{\theta}}\right)\right]\phi,$$

which, if f is from the exponential family of distributions, is

$$D = \sum_{i=1}^{n} 2w_i\left[y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right) + b\left(\hat{\theta}_i\right)\right].$$

Dividing through by the scale parameter ϕ, we come to another measure:

Definition 10.2 (Scaled Deviance).

Suppose we have a data model f with associated log-likelihood ℓ. Then the scaled deviance of f is defined as

$$D^* = D/\phi = 2\left[\ell\left(\underline{\tilde{\theta}}\right) - \ell\left(\underline{\hat{\theta}}\right)\right].$$

The deviance above is an example of a more general likelihood ratio statistic, which is useful for comparing (nested) models. We will return to nested models and the deviance for comparing two candidate data models later on.

Remarks:

- The scaled deviance depends on the unknown scale/dispersion parameter ϕ, whereas the deviance is defined to be independent of ϕ.

- The closer a GLM fits the data, the more similar $\hat{\mu}_i$ is to y_i and so the smaller the deviance.

- Since $\ell\left(\underline{\tilde{\theta}}, \phi\right)$ is always greater than or equal to $\ell\left(\underline{\hat{\theta}}, \phi\right)$ (question for you: why?), the scaled deviance is always non-negative. However, the deviance can be negative depending on the sign of ϕ.

- For the Binomial and Poisson distributions, $\phi = 1$, and so the deviance and scaled deviance are the same, but this is not the case more generally.

Example 10.2 (Scaled deviance for the Normal GLM).

The likelihood for the normal density is

$$\ell(\underline{\mu}, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2 - \frac{n}{2}\log(2\pi\sigma^2),$$

and so the scaled deviance is

$$D^* = 2\left[\ell\left(\underline{\tilde{\mu}}\right) - \ell\left(\underline{\hat{\mu}}\right)\right]$$

$$= 2\Bigg[\Bigg(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - y_i)^2 - \frac{n}{2}\log(2\pi\sigma^2)\Bigg) - \Bigg(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2,$$

where as usual for a standard GLM, $\hat{\mu}_i = g^{-1}(\mathbf{x}_i^T\underline{\beta})$. In other words, the scaled deviance is the scaled residual sum of squares (RSS).

Under certain conditions (and where ϕ is known), the scaled deviance D^* is distributed χ^2_{n-p} asymptotically, where p is the number of parameters in $\underline{\beta}$. When the χ^2_{n-p} approximation is good, then this can form a goodness of fit test (a model with a large deviance does not fit the data well). In practice, we can use the goodness of fit test with the binomial and Poisson distributions since we know $\phi = 1$, provided we have a large dataset. The approximation is not accurate at all for binary data.

A large deviance will indicate a poor fit of the model to the data. Usually, this is due to one of two reasons (or both):

- The model (for the mean) is not appropriate for the data. One could consider for example, whether other (or additional) covariates should be included in the model.

- The assumption on the scale parameter, e.g. $\phi = 1$ is not suitable. The data will probably show evidence of overdispersion. We will discuss this later in detail.

In **R**, the (residual) deviance can be obtained for the `beetle` binomial GLM using

```
beetle_glm$deviance
```

```
## [1] 8.43
```

10.4 Model Comparison in a GLM framework

10.4.1 Comparing models with the Deviance: Analysis of Deviance

As we saw in the linear model part of the course, an important question when doing statistical modelling is to decide whether all the covariates are necessary, or whether some can be ignored without materially changing the model fit to the data.

In the (multiple) regression framework, this amounted to forming hypotheses for whether a subset of regression coefficients were zero or not, calculating a sum of squares measure of fit for both models, and assessing whether this measure was significantly reduced by including the extra variables, via the F-test.

For GLMs, since we allow non-identically distributed responses (and indeed any distribution within the exponential family), we will instead use the generalized likelihood ratio test.

Suppose we have two candidate GLMs with a density from the exponential family, the first has p_1 explanatory variables and the second has $p_2 < n$ explanatory variables, with $p_2 > p_1$ and $\mathbf{X}_1 \subset \mathbf{X}_2$ (i.e. Model 1 is nested in Model 2).

Our hypotheses take a similar form to before:

$H_0:$ The simpler model $\mathcal{M}_1$ adequately describes $\mathbf{y}$ $(\beta_{p_1+1} = ... = \beta_{p_2} =$

$H_1:$ The more complex model $\mathcal{M}_2$ is required.

Definition 10.3 (Deviance for model comparison).
Suppose we have the nested model situation above. The deviance (likelihood ratio test) statistic for model comparison is

$$D(\mathcal{M}_2, \mathcal{M}_1) = 2\left[\ell\left(\underline{\hat{\beta}}^{(2)}\right) - \ell\left(\underline{\hat{\beta}}^{(1)}\right)\right],$$

where $\underline{\hat{\beta}}^{(1)}$ and $\underline{\hat{\beta}}^{(2)}$ are the MLEs for $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively.

Sometimes $\mathcal{M}_2$ and $\mathcal{M}_1$ are referred to as the full and reduced model, respectively.

Under H_0, asymptotically as $n \to \infty$,

$$D(\mathcal{M}_2, \mathcal{M}_1) \sim \chi^2_{p_2 - p_1}.$$

Remarks:

- It is always the case that $\ell\left(\hat{\underline{\beta}}^{(2)}\right) \geq \ell\left(\hat{\underline{\beta}}^{(1)}\right)$. To see this, remember that $\ell\left(\hat{\underline{\beta}}^{(1)}\right)$ is a valid parameter choice for model $\mathcal{M}_2$, and $\ell\left(\hat{\underline{\beta}}^{(2)}\right)$ must have at least as high a likelihood under $\mathcal{M}_2$ by virtue of it being the MLE.

- Let D_1^*, D_2^* denote the scaled deviances for Models 1 and 2 respectively. Then

$$D(\mathcal{M}_2, \mathcal{M}_1) = 2\left[\ell\left(\hat{\underline{\beta}}^{(2)}\right) - \ell\left(\hat{\underline{\beta}}^{(1)}\right)\right]$$

$$= 2\left[\ell\left(\hat{\underline{\beta}}^{(2)}\right) - \ell\left(\tilde{\underline{\beta}}\right) + \ell\left(\tilde{\underline{\beta}}\right) - \ell\left(\hat{\underline{\beta}}^{(1)}\right)\right]$$

$$= D_1^* - D_2^* = \frac{D_1 - D_2}{\phi}.$$

Hence the model selection criterion can be calculated from the differences in two unscaled model deviances, provided that the deviances can be calculated (ϕ is known).

In this case, we should evaluate $D(\mathcal{M}_2, \mathcal{M}_1)$ and compare to the corresponding critical value, z_c^2 from the $\chi^2_{p_2-p_1}$ distribution. If $D(\mathcal{M}_2, \mathcal{M}_1) < z_c^2$ we do not reject H_0 and take the simpler model $\mathcal{M}_1$ as adequate.

10.4.2 Comparing (nested) models with the deviance when ϕ is unknown

The argument above can be used for performing model comparison in some cases, for example the Binomial and Poisson models. However, for many situations we do not know ϕ. What can we do in this situation?

Recall that asymptotically, $D_2^* \sim \chi^2_{n-p_2}$, i.e., we can estimate ϕ

by $D_2/(n-p_2)$. Thus we can eliminate ϕ by using the ratio

$$\frac{(D_1^* - D_2^*)/(p_2-p_1)}{D_2^*/(n-p_2)} = \frac{(D_1-D_2)/(p_2-p_1)}{D_2/(n-p_2)} \sim F_{p_2-p_1,n-p_2} \text{ for large sa}$$

and using that the ratio of two χ^2 distributions is an F distribution, specified by two degrees of freedom.

Note that this is still an approximation for all GLMs apart from the Gaussian model.

It is also useful to note here that the deviance is additive, i.e., for $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \mathcal{M}_3$, the statistic for comparing $\mathcal{M}_1$ and $\mathcal{M}_3$ is the sum of those for comparing $\mathcal{M}_1$ and $\mathcal{M}_2$ and comparing $\mathcal{M}_2$ and $\mathcal{M}_3$.

10.4.3 Other Measures of Model Fit

There are other quantities with which one may use to perform model selection. In particular, Akaike's Information Criterion (AIC) and Schwarz's Information Criterion (BIC) are well-used measures which take into account the model complexity (number of model parameters).

Akaike's Information Criterion is defined as

$$\mathrm{AIC} = -2\ell\left(\hat{\underline{\beta}}\right) + 2p,$$

and the Schwarz's Information Criterion (BIC) is

$$\mathrm{BIC} = -2\ell\left(\hat{\underline{\beta}}\right) + 2\log p,$$

where p is the number of explanatory variables in the linear predictor. For a better fitting model, we want a lower AIC or BIC. These measures are particularly useful for comparing non-nested models.

The introduced techniques also provide pathways to select the "best" set of explanatory variables. Suppose we started with the simplest model. We could then try adding single explanatory variables sequentially to see if they improve the model; this is known as forward model selection. In contrast, one could start with the "full" model, and try sequentially removing variables – this is known as backward model selection. Ideally we would like to sequentially cycle through dropping all variables.

In the next example, we look at different ways of including / excluding covariates and their effect on a model using **R**.

Example 10.3 (Contraceptive use).
The **cuse** data describes information on $n = 1607$ women from the Fiji Fertility Study, with variables indicating their age, whether they want more children, and whether they were currently using contraceptives.

```
## load("cuse.rda")
head(cuse)
```

```
##     age education wantsMore notUsing using
## 1   <25       low       yes       53     6
## 2   <25       low        no       10     4
## 3   <25      high       yes      212    52
## 4   <25      high        no       50    10
## 5 25-29       low       yes       60    14
## 6 25-29       low        no       19    10
```

We can fit a binomial GLM using the factor **age**:

```
cusefit1 = glm(cbind(using, notUsing) ~ age, family = binom
cusefit1$deviance
```

```
## [1] 86.58
```

The deviance is quite high. This suggests that more of the variability could be explained by adding other variables (try looking at the fitted values as for the beetle data set).

```
add1(cusefit1, ~. + education + wantsMore, test = "Chisq")
```

```
## Single term additions
##
## Model:
## cbind(using, notUsing) ~ age
##           Df Deviance AIC  LRT Pr(>Chi)
## <none>          86.6 166
## education  1     80.4 162  6.2    0.013 *
## wantsMore  1     36.9 118 49.7  1.8e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Both covariates are individually significant using a χ^2 (nested) model test, so we create a model with these included (see also AIC values),

```
cusefit2 = update(cusefit1, ~. + education + wantsMore)
```

and (checking), a nested model comparison shows the second is more appropriate:

```
anova(cusefit1, cusefit2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(using, notUsing) ~ age
## Model 2: cbind(using, notUsing) ~ age + education + want
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        12       86.6
## 2        10       29.9  2     56.7    5e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Recall that the logit link is used for the binomial as a default.

Similar to the beetle data set, the odds for the **`cusefit`** model can be expressed as

$$\frac{p_i}{1-p_i} = \exp(\eta(\mathbf{x}_i^T \underline{\beta})) = \exp(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}),$$

where $x_{i,1}$ corresponds to age, $x_{i,2}$ corresponds to education and $x_{i,3}$ corresponds to whether the individual wants more children or not. Note that in this model, the baseline corresponds to women under 25 years, who don't want more children and with high education:

```
cusefit2$coefficients
```

```
##  (Intercept)     age25-29     age30-39     age40-49 educ
##      -0.8082       0.3894       0.9086       1.1892
```

Looking at the coefficients of the model on the exponential scale, we have

```
##                beta exp(beta)
## (Intercept)   -0.81      0.45
## age25-29       0.39      1.48
## age30-39      0.91       2.48
## age40-49      1.19       3.28
## educationlow -0.32       0.72
## wantsMoreyes -0.83       0.43
```

In other words, compared to the baseline, the odds of contraception use trebles for the highest age group (**age40-49**), whereas the odds of contraceptive use more than halves if the individual reports wanting children (as $\exp(\widehat{\beta}_6) = 0.43$).

For this model, the deviance is also statistically unlikely to have been observed by random chance compared to a χ^2_{10} distribution with $n - p = 16 - 6$ degrees of freedom:

```
cusefit2$deviance
1 - pchisq(cusefit2$deviance, df = 10)
```

```
## [1] 29.92
## [1] 0.0008838
```

Let us now try a new model with interactions

```
cusefit3 = update(cusefit2, ~. + age * education + age * wa
cusefit3$deviance
```

```
## [1] 2.441
```

We can perform stepwise selection using the AIC, which cycles through all possible simpler models:

```
step(cusefit3, trace = 0)
```

```
##
## Call:  glm(formula = cbind(using, notUsing) ~ age + educ
##     age:education + age:wantsMore + education:wantsMore,
##     data = cuse)
##
## Coefficients:
##                (Intercept)                     age25-29
##                    -1.4516                       0.5979
##     age40-49:educationlow       age25-29:wantsMoreyes
##                    -0.9864                      -0.2254
##
```

```
## Degrees of Freedom: 15 Total (i.e. Null);  3 Residual
## Null Deviance:       166
## Residual Deviance: 2.44  AIC: 99.9
```

It turns out that the "full" model with the interactions isn't simplified using the AIC. Note that the AIC and deviance are low in comparison to both the original model **`cusefit1`** and the main effects model **`cusefit2`** (no interactions).

Remarks:

- In this model, the effect over the baseline for the 40-49 age group is 2.474 on the log-odds scale, meaning that the odds are $\exp(2.474) = 11.87$, i.e., nearly a 12-fold increase over the under 25 age group.

- However, if the women want children, this effect is reduced to $\exp(2.474 + 0.013 - 1.19) = 3.66$, because of the interaction. In other words, even though the women want children, there is still a 3.66 fold increase in odds that they will use contraception (over the under 25 age group who don't want children).

11 Modelling Aspects and Diagnostics

For the linear regression model, we highlighted that collinearity has to be taken into account, as it may lead to misleading model estimates. The same issue also arises in a GLM framework. Beyond collinearity, there are other aspects that have to be considered, and which we explore in this chapter. At the end of the chapter, we will describe diagnostics for GLMs.

11.1 Non-linear Predictors and Varying Exposure

Example 11.1 (Doctor Deaths).

In a cohort study in 1951, a number of (male) doctors were followed to investigate the effect of smoking on age of death from coronary disease. The data shows the doctors' age group, together with the total person-years of the individuals within the group (the number of years in the study at the time of analysis).

```
data(doctors, package = "dobson")
doctors
```

```
## # A tibble: 10 x 4
##    age      smoking    deaths `person-years`
##    <chr>    <chr>       <dbl>          <dbl>
##  1 35 to 44 smoker         32          52407
##  2 45 to 54 smoker        104          43248
##  3 55 to 64 smoker        206          28612
##  4 65 to 74 smoker        186          12663
##  5 75 to 84 smoker        102           5317
##  6 35 to 44 non-smoker      2          18790
##  7 45 to 54 non-smoker     12          10673
##  8 55 to 64 non-smoker     28           5710
##  9 65 to 74 non-smoker     28           2585
## 10 75 to 84 non-smoker     31           1462
```

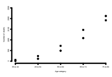

Figure 11.1: Number of deaths amongst doctors of a certain age group plotted by smoking status. The diamonds correspond to the non-smokers, while the crosses corresponds to the smokers.

Since these are (potentially unlimited) counts, a sensible choice of model for the deaths is a Poisson distribution. One possibility would be to use a rate $\mu_i = \exp(\eta_i) = \exp(\mathbf{x}_i^T \underline{\beta})$.

The plot suggests that there is a quadratic relationship between the rate of deaths and the age category. We demonstrate the lack of linear fit with the first model, and the improvement with the quadratic term.

```r
docglma = glm(deaths ~ agecat + smoking, family = poisson()
deviance(docglma)
docglmb = glm(deaths ~ agecat + I(agecat^2) + smoking, fami
deviance(docglmb)
```

```
## [1] 157.6
## [1] 14.65
```

The deviance of the first model is 157.59 whereas the second is 14.65 .

Since there seems to be a difference in the rate of increase between smoker and non-smoker, we should also consider including an interaction term.

However, these models do not take into account the varying exposure for each covariate pattern induced from the age groups: the doctors in each age group were "at risk" for differing times (or in this case, person-years). This is common in many modelling settings where the counts could be interpreted relative to some observed quantity affecting the rate θ_i . We have

$$Y_i \sim \text{Poisson}(\mu_i) \;\text{ with } \mu_i = u_i \exp(\eta_i) = u_i \exp(\mathbf{x}_i^T \underline{\beta}).$$

Using the (canonical) log link function on the logarithmic scale, the model can be written

$$\log(\mu_i) = \log(u_i) + \mathbf{x}_i^T \underline{\beta}.$$

The extra model term, the log of the exposure $\log(u_i)$, is called the offset. This can be included explicitly in the design matrix $\mathbf{X}$ by including a predictor with a known coefficient of one. This is normally included in **R** on the scale of the linear predictor.

Including an offset in the doctor mortality example, we find that the deviance is

```
docglmc = glm(deaths ~ agecat + I(agecat^2) + smoking + smo
                family = poisson(),
                  data = doctors)
deviance(docglmc)
```

```
## [1] 1.635
```

This new model fits quite well:

```
docglmcfits = fitted(docglmc)
obsdeaths = doctors$deaths
rbind(docglmcfits, obsdeaths)
```

```
##                 1     2     3     4     5     6     7
## docglmcfits 29.58 106.8 208.2 182.8 102.6 3.415 11.54 24
## obsdeaths   32.00 104.0 206.0 186.0 102.0 2.000 12.00 28
```

11.2 Overdispersed Models

In the case of the Poisson distribution, the mean is equal to the variance, specified by the single parameter μ. However, sometimes the variance of observations are much higher than their individual means (despite accounting for varying exposure). This can also happen in binomial GLMs. In this case, the observations are termed overdispersed. A high deviance despite a good model fit can signify overdispersion.

We can investigate overdispersion by looking at standardized residuals

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

to see if they are greater than 1. In fact the overdispersion measure

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} (r_i^P)^2$$

can be compared to a χ^2_{n-p} distribution to test for overdispersion.

To adjust for overdispersion, we can fit an overdispersed Poisson (known as the quasi-Poisson) model. This essentially relaxes the $\phi = 1$ condition of these distributions to an unknown dispersion parameter ϕ, so that (in the Poisson case) $\mathrm{Var}(Y_i) = \phi V(\mu_i) = \phi\mu$, i.e., the variance is proportional to the mean.

The name quasi- originates from the fact that with a non-unit ϕ, the distribution of the model does not sum/integrate to one. This means that the model does not have a “proper” likelihood. Despite this, since the maximization to estimate $\underline{\beta}$ does not involve ϕ , the estimates of $\underline{\beta}$ are the same for quasi-Poisson and quasi-Binomial models as their usual counterparts (as if assuming $\phi = 1$). The dispersion parameter can then be estimated as with other models using the method of moments (Pearson's estimator).

Note that due to the likelihood not being valid, inference measures and summaries based on the likelihood cannot be used for overdispersed models (for example the AIC).

However, the deviance can still be used since the ratio in its definition cancels the integral/sum of the distribution. For these models, we perform hypothesis tests and model comparison with an estimated dispersion, and so hypotheses on regression coefficients are tested against the t_{n-p} distribution (see Section 10.2). This is because we have to estimate ϕ ; the standard errors of the regression coefficients are then multiplied by $\sqrt{\phi}$ according to the asymptotic normal distribution.

For model comparison, we must use the F -distribution instead of the χ^2 distribution.

In **R**, the overdispersed models can be fit by setting **`family = quasipoisson`** (or **`family = quasibinomial`**).

Alternatively, a specified distribution with a different mean-variance relationship for count data is the negative binomial distribution, with probability mass function

$$f(y;\mu,\vartheta) = \frac{\Gamma(y+\vartheta)}{\Gamma(\vartheta)y!}\left(\frac{\mu}{mu+\vartheta}\right)^{y}\left(\frac{\vartheta}{mu+\vartheta}\right)^{\vartheta},$$

with $\mathbb{E}(Y) = \mu$ and $\operatorname{Var}(Y) = \mu(\mu+\vartheta^{-1})$, i.e., $\operatorname{Var}(Y) > \mathbb{E}(Y)$. Such a model is used when the overdispersion could be attributed to a specific source or mechanism. If ϑ is known, the negative binomial can be seen as an exponential family model with the same link choices as the Poisson GLM (we have seen that in Question 3 on Problem Sheet 5).

Example 11.2 (Disease Incidents).
In this example, we examine a small dataset of disease cases with a single predictor, the time at which the count was recorded. We fit models with and without overdispersion.

```r
## load("citydisease.rda")
fitp = glm(Incidents ~ Month, family = "poisson", data = ci

# Estimate the dispersion, which seems bigger than 1:
sum( (citydisease$Incidents - fitted(fitp) )^2 / fitted(fit
```

```
## [1] 2.609
```

```r
# Fit a quasi-Poisson model:
fitqp = glm(Incidents ~ Month, family = quasipoisson(), dat
summary(fitqp)
```

```
##
## Call:
## glm(formula = Incidents ~ Month, family = quasipoisson()
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5484     0.5520   -0.99  0.33361
## Month         0.1706     0.0346    4.93  0.00011 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## (Dispersion parameter for quasipoisson family taken to b
##
##     Null deviance: 126.891  on 19  degrees of freedom
## Residual deviance:  45.703  on 18  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
# Fit a negative binomial model with unknown theta:
library("MASS")
fitnb2 = glm.nb(Incidents ~ Month, data = citydisease)
summary(fitnb2)
```

```
##
## Call:
## glm.nb(formula = Incidents ~ Month, data = citydisease,
##     link = log)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.672      0.456   -1.47     0.14
## Month          0.180      0.032    5.62  1.9e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## (Dispersion parameter for Negative Binomial(4.046) famil
##
##     Null deviance: 64.298  on 19  degrees of freedom
## Residual deviance: 26.991  on 18  degrees of freedom
## AIC: 97.91
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  4.05
##           Std. Err.:  3.30
##
##  2 x log-likelihood:  -91.91
```

If we fit a Poisson GLM, we get an estimate of the dispersion parameter to be $\hat{\phi} = \mathbf{2.61}$, which seems to be greater than one.

The deviance for the model is 45.07; fitting a negative binomial GLM gives us a deviance of 26.9.

11.3 Diagnostics for GLMs

Recall that in normal linear models, the square roots of the individual terms in the (residual) sum of squares function $S(\underline{\beta})$ are the residuals, i.e., $\hat{\epsilon}_i = y_i - \mathbf{x}_i^T \underline{\beta}$. Since these quantities contain all the information (variability) not explained by the systematic part of the model, they can be examined to investigate model fit (e.g., unusually high residuals).

For GLMs, we cannot use the usual residuals directly, since they do not account for the mean-variance relationship of the distribution for the observations. Hence we standardize the residuals using the variance from the model so that (as far as possible) these standardized residuals act like residuals from the Gaussian linear model.

We will focus on two types of standardized residuals.

11.3.1 Pearson's Residuals

We can standardize the residuals by a quantity proportional to their standard deviation from the fitted model.

Definition 11.1 (Pearson Residuals).
For a GLM, the Pearson residuals are defined as

$$r_i^P = \frac{(y_i - \widehat{\mu}_i)}{\sqrt{V(\widehat{\mu}_i)}},$$

so that the residuals will be zero mean, with variance ϕ.

These are essentially the square roots of the individual terms in the definition of Pearson's chi square statistic (the sum of squares of the Pearson residuals gives the Pearson statistic).

11.3.2 Deviance Residuals

For some GLMs, Pearson's residuals aren't as symmetric as one would like (if we want them to behave like residuals from linear models). To counteract this, we can define alternative residuals for GLMs as follows.

Recall that the (unscaled) deviance acts like a measure of model fit, similar to the sum of squares function in normal linear models. Hence, in a similar way, we can examine the square roots of the individual terms in the (residual) deviance.

Definition 11.2 (Deviance Residuals).
For a GLM, the deviance residuals are defined as

$$r_i^D = \sqrt{D_i}\text{sign}(y_i - \hat{\mu}_i).$$

Note that the $\text{sign}(\cdot)$ is positive if y_i is greater than the fitted value, and negative otherwise. We can then compare the deviances of the observations to see which particular cases deviate from the fitted model. The sum of squares of these deviance residuals gives the deviance itself.

As discussed before, if the deviance were calculated for a model where all the parameters were known, then the scaled deviance would be $D^* \sim \chi^2_n$, and so we might expect that $d_i/\phi \sim \chi^2_1$, and so $r_i^D \sim \text{Normal}(0, \phi)$. Whilst the chi-squared distribution for the deviance is asymptotic, this is not the case, but nevertheless this argument suggests that we might expect the deviance residuals to behave something like $\text{Normal}(0, \phi)$ random variables, for a well fitting model, especially in cases for which the χ^2 approximation is reasonable.

Remarks:

- In the case of normal GLM, both deviance and Pearson residuals equal $y_i - \hat{\mu}_i$, the ordinary residual.

- In the case of a Poisson GLM, Pearson residuals equal $(y_i - \hat{\mu}_i)/\sqrt{\hat{\mu}_i}$, the usual standardized residual.

- For non-normal GLMs, the deviance residuals as a set are more nearly normal than the Pearson's residuals and are naturally more preferred. This is true, for example, in the case of logistic regression.

- The residuals should not display any trend in mean or variance (pattern) when plotted against the fitted values, or any model covariates.

11.3.3 Leverage and Influence

Leverages and influence can be defined similarly to the Normal linear model case. However, the weights involved in the IRWLS fitting algorithm affect the leverages in a GLM. Consequently, the hat matrix for a GLM is

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{W}^{1/2}$$

where $\mathbf{W}$ is the diagonal matrix with IRWLS weights W_{ii} as defined before.

The leverages can be extracted as before in **`R`** with the function **`hatvalues`** or **`influence(docglmc)$hat`**. Note that we can also examine the sensitivity of the model via Cook's distance as before.

11.3.4 Diagnostic plots

In the linear model case, we used fitted values against residuals to examine orthogonality. For GLMs, we have to be careful to use the correct scale to examine model fit. In particular, we usually plot the linear predictor $\underline{\hat{\eta}}$ against residuals:

```
plot(residuals(docglmc) ~ predict(docglmc, type = "link"),
     xlab = "Linear Predictor", ylab = "Deviance residuals"
abline(h = 0, col = "firebrick")
```

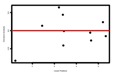

Figure 11.2: Plot to examine residuals against fitted of the model on the link (predictor) scale.

We should look at this plot to indicate any sign of a non-linear relationship or trend, in which case we might look at choice of model predictors.

Secondly, we can look if there is any evidence of heteroscedasticity: this might point to an alternative model, or a model with overdispersion.

In addition, we can look at half-normal quantile plots. These plots aren't used to look for departures from normality, but can be used to look for outliers. These diagnostics examine a sorted set of (positive) model quantities against the quantiles of the half-normal distribution: $\Phi^{-1}\left(\frac{n+i}{2n+1}\right)$, for $i = 1, \ldots, n$.

For example, we can examine the (studentized) residuals from a model fit:

```
library(daewr)
halfnorm(rstudent(docglmc))
```

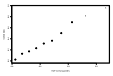

Figure 11.3: Half-normal residual plot to examine outliers as departures from the model fit.

The final plot in Figure 11.4 investigate whether there are any overly influential data points.

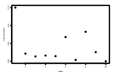

Figure 11.4: Examine influence of data points.

12 Ordinal Regression

12.1 Modelling Categorical Variables

In the generalized linear models in Chapters 8 to 11, we considered modelling a count, a continuous variable or a binary response. In this part of the course, we will focus on regression for categorical responses. Specifically, the categories are assumed to have a natural ordering; for instance, ratings for websites or movies fall into this class. This type of regression is termed ordinal regression.

The following plot shows one simulated example with 5 possible outcomes:

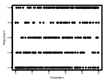

Figure 12.1: Plot of categorical response against covariate.

To model a categorical variable Y, we use a categorical distribution. If Y has K levels, the categorical distribution function has the form

$$\mathbb{P}(Y = k) = p_k \qquad (k = 1, \dots, K),$$

with $p_1 + \quad + p_K = 1$. We thus have $K - 1$ parameters, $p_1, \dots, p_{K-1}$, and the last probability is $p_K = 1 - p_1 - \quad - p_{K-1}$.

A regression framework thus has to model $p_1, \dots, p_{K-1}$ conditional on the explanatory variables $\mathbf{x}$.

12.2 The Ordinal Logistic Regression Model

Let F_k correspond to the probability that Y has a value of k or below, $F_k = p_1 + p_2 + \quad + p_k = \mathbb{P}(Y \leq k)$. Then we could use a logistic regression model for F_k with

$$\log\left(\frac{F_k}{1 - F_k}\right) = \eta,$$

where $\eta = \mathbf{x}^T \underline{\beta}$ is the linear predictor. In other words, the logistic regression model considers a binary random variable Z which takes value $Z = 1$ with probability F_k, and the value $Z = 0$ with probability $1 - F_k$.

We could consider each level $k = 1, \dots, K - 1$ individually and estimate separate logistic regression models for $F_1, \dots, F_{k-1}$. The probabilities $p_1, \dots, p_K$ are then given by

$$p_k = \mathbb{P}(Y \leq k) - \mathbb{P}(Y \leq k-1) = F_k - F_{k-1} \qquad (k = 1, \dots, K),$$

where $F_0 = 0$ and $F_K = 1$. The drawback of such an approach is that the fitted regression lines for $F_1, \dots, F_{K-1}$ may cross, which would give the contradiction $F_k > F_{k+1}$.

Instead of considering $F_1, \dots, F_{K-1}$ individually, the ordinal logistic regression model defines

$$\log\left(\frac{F_k}{1 - F_k}\right) = \alpha_k + \mathbf{x}^T\underline{\beta},$$

where $\alpha_1 < \alpha_2 < \quad < \alpha_{K-1}$. We see that the regression lines for $F_1, \dots, F_K$ are parallel and thus the case $F_k > F_{k+1}$ is prevented.

The ordinal logistic regression assumes that the coefficients that describe the relationship between, say, the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and all higher categories, etc. This is called the proportional odds assumption. Because the relationship between all pairs of groups is the same, there is only one set of coefficients.

If this is not the case, we would need different sets of coefficients in the model to describe the relationship between each pair of outcomes. As mentioned above, this provides additional constraints on the set of possible parameter values. An alternative is the application of nonparametric regression models[4] (which we don't cover

[4] Examples are kriging in spatial statistics, or regression trees and multivariate adaptive

in this course).

12.3 Ordinal Regression in **R**

In **R**, we can fit the ordinal logistic regression model using the **polr** function in the **MASS** library. Note that the estimates provided by the **summary** function consider the model

$$\log\left(\frac{F_k}{1-F_k}\right) = \alpha_k - \mathbf{x}^T\underline{\beta}.$$

Let's apply this function to the simulated data we plotted above.

```r
library(MASS)
m = polr(y ~ x, Hess = TRUE)
summary(m)
```

```
## Call:
## polr(formula = y ~ x, Hess = TRUE)
##
## Coefficients:
##   Value Std. Error t value
## x 0.583     0.0613    9.52
```

regression splines in machine learning, to cite few.

```
##
## Intercepts:
##     Value  Std. Error t value
## 1|2  0.366  0.168      2.178
## 2|3  1.190  0.175      6.790
## 3|4  1.974  0.189     10.437
## 4|5  2.500  0.200     12.472
##
## Residual Deviance: 1442.98
## AIC: 1452.98
```

We find $\alpha_1 = 0.37\,,\ \alpha_2 = 1.19\,,\ \alpha_3 = 1.98\,,\ \alpha_4 = 2.50$ and $\beta = -0.58$ for the model described in Section 12.2.

Inference for the model parameters is equivalent to the GLM framework considered in Chapter 10.

Part III

Generalized Mixed Effects Models

13 Linear Mixed Effects Models: Introduction

13.1 Motivation

In the models we have seen up to now, we assume that the observations $y_1, \dots, y_n$ are independent (but not necessarily identically distributed). In the considered linear and generalized models, the regression coefficients $\underline{\beta}$ were considered unknown but fixed, and estimated from the data. Such models are called fixed effects models.

However, in many situations we will observe data which are grouped in nature. This could be, for instance, due to repeated measurements. Some examples are:

- Observations in medical studies. Data in drug trials are often collected from the same individuals over time (longitudinal studies). It may be reasonable to assume that correlations exist among the observations from the same individual.

- Ecological / agricultural data. Yields of animal products or crops associated with the same field may be correlated; birth-weights of offspring in animal litters will be clustered by litter.

Note that these settings are examples of models with one level of grouping. There are also many situations where more complex multilevel groupings are required, i.e., a hierarchical structure is more appropriate. For example,

- Student exam scores. We often record data on exam marks over a number of years. A set of exam scores in a particular year correspond to an individual student, but are also associated to a teaching / tutorial class. These scores may further be viewed on a per school basis, or at a regional level.

The argument above suggests that we should take these groupings into account within our model. We should, for example, expect that an individual's set of exam scores are correlated, but also that there is potential correlation between individuals in a particular class, or students at a particular school.

One way of including such structure into (generalized) linear models is to assume that these coefficients are random. As such, these quantities cannot be estimated directly from the data, but we can estimate parameters associated with their distribution. For example, we cannot estimate the random coefficient of an unobserved student, but we can estimate its distribution. By accounting for the correlation structure in the data, the inclusion of these random coefficients allows us to effectively partition overall variation of the response variable into components corresponding to different levels of data hierarchy.

We refer to these random coefficient as random effects. Random effects can be thought of as grouping factors for which we are trying to control.

A mixed effects model contains both fixed and random effects.

13.2 Illustrative Example: Analysis of Estrone Levels

We introduce the mixed effects modelling framework using the following example.

Sixteen measurements of estrone[5] (each on a logarithmic scale) were taken from five menopausal women. We are primarily interested in quantifying the variability between measurements for each woman individually, as well as between women. The data for the five women are shown as boxplots below.

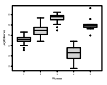

Figure 13.1: Boxplots of estrone measurements on logarithmic scale for five women.

We see that the women differ in their average estrone level. Let $Y_{i,j}$ denote the estrone level of woman i and measurement j. We now define a random effects model for $Y_{i,j}$. Since the women

[5] For the curious minds: Wikipedia page and the PubChem page on estrone.

are random with respect to the population, we model the estrone measurements using

$$Y_{i,j} = \mu + b_i + \epsilon_{i,j} \qquad (i = 1, \dots, 5;\ j = 1, \dots, 16),$$

where μ is the average measurement across the five women (fixed effect) and b_i is the person effect (random effect) and $\epsilon_{i,j}$ is the residual. In this model, we assume $\mathbb{E}(b_i) = \mathbb{E}(\epsilon_{i,j}) = 0$, $\mathrm{Var}(b_i) = \sigma_b^2$ and $\mathrm{Var}(\epsilon_{i,j}) = \sigma_\epsilon^2$.

We are generally not interested in how the estrone level changes from woman to woman, but in the variability between women. In other words, the focus does not lie on the random effects $b_1 \dots, b_I$ but their variance σ_b^2 (that's similar to σ^2 and $\epsilon_1, \dots, \epsilon_n$ in a linear regression framework).

For random effects models, we are generally interested in two questions arising from the data:

1. Is there evidence for variability in estrone between women?

2. If so, how large is this variability in relation to the variability of measurements for an individual woman?

The total variance of observations in this setting is

$$\mathrm{Var}(Y_{i,j}) = \sigma^2_{total} = \sigma^2_b + \sigma^2_\epsilon.$$

Hence we can express the first question mathematically as

$$H_0 : \sigma^2_b = 0 \qquad \text{vs} \qquad H_1 : \sigma^2_b > 0.$$

To address the second question, we quantify the correlation between observations from the same woman using the intraclass correlation

$$\rho = \frac{\sigma^2_b}{\sigma^2_b + \sigma^2_\epsilon}.$$

The case $\rho = 0$ implies no correlation between observations in the same group, i.e., $\sigma^2_b = 0$. Conversely, values close to $\rho = 1$ indicate that the interclass ("between groups") variability is larger than the intraclass ("within class") variability.

13.2.1 Connection to ANOVA

With a normality assumption, the framework above looks similar to a one-way ANOVA. We have data with m levels (groups) of a factor variable and n observations per group; this is called a balanced design. The key difference is the inclusion of random effects $(b_1, \ldots, b_m)$; you can think of this as a random effects ANOVA model.

Let $\overline{Y_i}$ denote the mean of the i-th group. The random effects model gives that

$$\overline{Y_i} = \mu + b_i + \frac{1}{J}\sum_{j=1}^{J} \epsilon_{i,j}.$$

The distribution of the random component is

$$b_i + \frac{1}{J}\sum_{j=1}^{J} \epsilon_{i,j} \sim \text{Normal}\left(0, \sigma_b^2 + \frac{\sigma_\epsilon^2}{J}\right).$$

13.3 Extending the Initial Model

As mentioned before, a mixed effects model has both fixed and random effects. In this setting, the fixed effects (as the name suggests) are fixed properties of the response variable, whereas the random effects account for the variability in the environment in which they are measured. Let's consider a more general mixed effects model to the one in Section 13.2.

Suppose our one explanatory variable x is a factor with I levels, and we want to model $Y_{i,j,k}$, corresponding to the k-th observation for the i-th level of x and the j-th group ($k = 1, \dots, n$; $j = 1 \dots, J$). A model for $Y_{i,j,k}$ would be

$$Y_{i,j,k} = \mu + \beta_i + b_j + \gamma_{i,j} + \epsilon_{i,j,k},$$

where $b_j \sim \text{Normal}(0, \sigma_b^2)$, $\gamma_{i,j} \sim \text{Normal}(0, \sigma_\gamma^2)$ and $\epsilon_{i,j,k} \sim \text{Normal}(0, \sigma_\epsilon^2)$, and as before, we assume mutual independence between all random variables.

Here, β_i is the effect for level i of x (given by the regression coefficients $\underline{\beta}$), and b_j is the j-th random effect. The notation $\gamma_{i,j}$ represents the $I \times J$ interaction terms between the factor x and the grouping.

We can assess the variability in the model (represented by b_j and $\gamma_{i,j}$) in a similar fashion to before. For example, the significance of the interaction terms can be considered using the hypothesis $H_0 : \sigma^2_\gamma = 0$, and comparing the models under the null and alternative hypotheses using the ANOVA / F-test procedure as before, aggregating the data as appropriate for the hypothesis.

13.4 General Notation for the Linear Mixed Effects Model

Recall that the normal linear (fixed effects) model can be written as

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \underline{\epsilon},$$

where $\epsilon \sim \mathrm{MVN}_n(0, \sigma^2\mathbf{I}_n)$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix. We can also write this model as

$$\mathbf{Y} \sim \mathrm{MVN}_n(\mathbf{X}\underline{\beta}, \sigma^2\mathbf{I}_n).$$

Suppose that the data we wish to analyse are grouped according to a single level with q groups. A mixed effects (normal) linear model can be written, conditional on the random effects $\mathbf{b} = (b_1, \dots, b_q)$, as

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \mathbf{Z}\mathbf{b} + \underline{\epsilon}$$

or

$$\mathbf{Y} \mid \mathbf{b} \sim \mathrm{MVN}_n(\mathbf{X}\underline{\beta} + \mathbf{Z}\mathbf{b}\,,\, \sigma^2\mathbf{I}_n),$$

where $\mathbf{Y}$, $\mathbf{X}$, $\underline{\beta}$ and $\underline{\epsilon}$ are defined as before, and $\mathbf{Z}$ is the matrix of covariates associated to the q (unknown) random effects.

Note that this model does not include any variability in the random effects. This variability is usually included by assuming

$$\mathbf{b} \sim \mathrm{MVN}_q\left(0, \mathcal{D}\right),$$

leading to the fully unconditional or marginal model

$$\mathbf{Y} \sim \mathrm{MVN}_n\left(\mathbf{X}\underline{\beta}\,,\, \Sigma + \mathbf{Z}\mathcal{D}\mathbf{Z}^T\right). \tag{13.1}$$

The variance matrix $\mathcal{D}$ is usually parametrized by a parameter vector $\underline{\theta}$, which will be the main quantity of interest in respect of the random effects. The residual error covariance matrix Σ usually takes a simple structure; in its simplest form it is $\Sigma = \sigma_\epsilon^2\mathbf{I}_n$, but could include dependence, such as an autoregressive structure.

For the estrone example in Section 13.2, the random intercept model can be written in the general mixed model form with the model quantities being:

- $\mathbf{Y} \in \mathcal{M}_{80\times 1}(\mathbb{R})$ (most intuitively with observations grouped according to woman);

- $\mathbf{X} \in \mathcal{M}_{80\times 1}(\mathbb{R})$ with all values equal to 1;

- $\mathbf{Z} \in \mathcal{M}_{80\times 5}(\mathbb{R})$ made up of indicator columns, one for each woman, with 16 ones per column (in an order corresponding to $\mathbf{Y}$);

- $\Sigma = \sigma_\epsilon^2 \mathbf{I}_{80} \in \mathcal{M}_{80\times 80}(\mathbb{R})$;

- $\mathcal{D} = \sigma_b^2 \mathbf{I}_5 \in \mathcal{M}_{5\times 5}(\mathbb{R})$;

- The vectors involved in the model are $\underline{\beta} = \mu$, $\mathbf{b} = (b_1, \dots, b_5)^T$ and $\underline{\theta} = \left(\sigma_b^2, \sigma_\epsilon^2\right)$.

In the next chapters we will explore some of the following remarks in detail:

- Testing for random effects proceeds as in the fixed effects case, by comparing models with and without the effect.

- We can test for fixed effects, whether or not there are significant random effects.

- The key step is to "average out" the data at each level of the random effect, producing a simpler model (for the aggregated data): the random effect is absorbed into the error term.

- Models for different aggregations will enable inferences to be made about different fixed and random effects.

- The variances of the random effects can be estimated from combinations of the usual variance estimates from the (aggregated) models.

14 Linear Mixed Effects Models: Estimation and Inference

Interest lies in estimating the parameters $\underline{\beta}$ and $\underline{\theta} = (\sigma_b^2, \sigma_\epsilon^2)$ of the linear mixed effects model

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \mathbf{Z}\mathbf{b} + \underline{\epsilon} \sim \mathrm{MVN}_n \left(\mathbf{X}\underline{\beta}\,,\, \Sigma + \mathbf{Z}\mathcal{D}\mathbf{Z}^T \right),$$

where $\mathcal{D} = \sigma_b^2 \mathbf{I}_q$ and $\Sigma = \sigma_\epsilon^2 \mathbf{I}_n$. Due to the correlation structure, inference in mixed effects models is different to that of fixed effects models. Additionally, we often have the situation where there are fewer observations than possible parameters, so estimation is not as straightforward as for those with only fixed effects.

In a simple random intercept model, such as the estrone levels example in Section 13.2, we can perform estimation using aggregated data and fitting fixed effect linear models (ANOVA), which is essentially based on least squares estimation. However, for more general mixed models, estimation is more complex and so we will use (variants of) maximum likelihood estimation.

14.1 Estimation of the Fixed Effects

By writing the variance-covariance matrix as $\mathbf{V}(\underline{\theta}) = \Sigma + \mathbf{Z}\mathcal{D}\mathbf{Z}^T$, the joint density of $y_1, \dots, y_n$ can be expressed as

$$f(\mathbf{y} \mid \underline{\beta}, \underline{\theta}) = \frac{1}{(2\pi)^{n/2}|V(\underline{\theta})|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y} - \mathbf{X}\underline{\beta}\right)^T \mathbf{V}(\underline{\theta})^{-1}\left(\mathbf{y} - \mathbf{X}\underline{\beta}\right)\right\},$$

where $|\mathbf{V}(\underline{\theta})|$ denotes the determinant of $\mathbf{V}(\underline{\theta})$. In other words, the log-likelihood of the parameters $\underline{\beta}$ and $\underline{\theta}$ is up to a proportionality constant

$$\ell(\underline{\beta}, \underline{\theta}) = -\frac{1}{2}\left\{\log|\mathbf{V}(\underline{\theta})| + \left(\mathbf{y} - \mathbf{X}\underline{\beta}\right)^T \mathbf{V}(\underline{\theta})^{-1}\left(\mathbf{y} - \mathbf{X}\underline{\beta}\right)\right\}$$

$$= -\frac{1}{2}\left\{\log|\mathbf{V}(\underline{\theta})| + \mathbf{y}^T\mathbf{V}(\underline{\theta})^{-1}\mathbf{y} - \underline{\beta}^T\mathbf{X}^T\mathbf{V}(\underline{\theta})^{-1}\mathbf{y} - \mathbf{y}^T\mathbf{V}(\underline{\theta})^{-1}\mathbf{X}\right.$$

Similar to the multiple linear regression case, we collect terms and obtain

$$\ell(\underline{\beta}, \underline{\theta}) = -\frac{1}{2}\left\{\log|\mathbf{V}(\underline{\theta})| + \mathbf{y}^T\mathbf{V}(\underline{\theta})^{-1}\mathbf{y} - 2\underline{\beta}^T\mathbf{X}^T\mathbf{V}(\underline{\theta})^{-1}\mathbf{y} + \underline{\beta}^T\mathbf{X}^T\mathbf{V}(\underline{\theta})\right.$$

We can maximize this log-likelihood to find the parameter estimates.

Suppose we knew the variance parameters $\underline{\theta}$. Then conditional on $\underline{\theta}$, the maximization of the likelihood via the score function involves solving the equation

$$U\left(\underline{\beta}_{\theta}, \underline{\theta}\right) = -\frac{1}{2}\left\{-2\mathbf{X}^T\mathbf{V}(\underline{\theta})^{-1}\mathbf{y} + \mathbf{X}^T\mathbf{V}(\underline{\theta})^{-1}\mathbf{X}\underline{\beta}_{\theta} + \underline{\beta}_{\theta}^T\mathbf{X}^T\mathbf{V}(\underline{\theta})^{-1}\mathbf{X}\right\}$$

$$= \mathbf{X}^T\mathbf{V}(\underline{\theta})^{-1}\mathbf{y} - \mathbf{X}^T\mathbf{V}(\underline{\theta})^{-1}\mathbf{X}\underline{\beta}_{\theta} = 0$$

i.e.,

$$\hat{\underline{\beta}}_{\theta} = \left(\mathbf{X}^T\mathbf{V}(\underline{\theta})^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}(\underline{\theta})^{-1}\mathbf{y}.$$

This is the same as the usual generalized least squares solution (for non-identically distributed observations). Then the variance parameters can be estimated by maximizing the profile likelihood

$$\ell\left(\underline{\theta} \mid \hat{\underline{\beta}}_{\theta}\right) = -\frac{1}{2}\left\{\log|\mathbf{V}(\underline{\theta})| + \left(\mathbf{y} - \mathbf{X}\hat{\underline{\beta}}_{\theta}\right)^T\mathbf{V}(\underline{\theta})^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\underline{\beta}}_{\theta}\right)\right\}$$

i.e., the likelihood with $\underline{\beta}$ replaced with $\hat{\underline{\beta}}_{\theta}$.

14.2 Estimation of the Random Effects

Using the distributional remarks earlier, the joint density of the observations and the random effects is given by

$$f(\mathbf{y}, \mathbf{b}) = f(\mathbf{y} \mid \mathbf{b})f(\mathbf{b}).$$

This leads to the log-likelihood being written as

$$\ell\left(\underline{\beta},\underline{\theta},\mathbf{b}\right)=-\frac{1}{2}\left\{\log|\Sigma|+\left(\mathbf{y}-\mathbf{X}\underline{\beta}-\mathbf{Zb}\right)^{T}\Sigma^{-1}\left(\mathbf{y}-\mathbf{X}\underline{\beta}-\mathbf{Zb}\right)\right\}-\frac{1}{2}\{$$

To estimate the random effects, we maximize this likelihood by taking the derivative with respect to $\mathbf{b}$:

$$\frac{\partial\ell(\underline{\beta},\underline{\theta},\mathbf{b})}{\partial\mathbf{b}}=\mathbf{Z}^{T}\Sigma^{-1}\left(\mathbf{y}-\mathbf{X}\underline{\beta}-\mathbf{Zb}\right)-\mathcal{D}^{-1}\mathbf{b}.$$

In other words, the MLE for $\mathbf{b}$ solves

$$\left(\mathbf{Z}^{T}\Sigma^{-1}\mathbf{Z}+\mathcal{D}^{-1}\right)\hat{\mathbf{b}}=\mathbf{Z}^{T}\Sigma^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\underline{\beta}}\right),$$

where we have substituted estimates for the fixed parameters $\hat{\underline{\beta}}$ in the equation above. It can be shown that $\hat{\mathbf{b}}=\mathcal{D}\mathbf{Z}^{T}\mathbf{V}(\underline{\theta})^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\underline{\beta}}\right)$, which is also known as the best linear unbiased predictor (BLUP).

14.3 Estimation of the Variance Components

In Section 13.2, we modelled estrone level using the model

$$Y_{i,j}=\mu+b_i+\epsilon_{i,j}\qquad(i=1,\dots,I;\ j=1,\dots,J),$$

with

$$\overline{Y}_i \sim \text{Normal}\left(\mu, \sigma_b^2 + \frac{\sigma_\epsilon^2}{J}\right), \qquad i = 1, \ldots, I.$$

We first note that we can decompose the residual sum of squares (the total variation) as follows:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{i,j} - \bar{y})^2 = \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{i,j} - \bar{y}_i)^2}_{\text{RSS}_1} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_i - \bar{y})^2}_{\text{RSS}_2}.$$

Using the results for variance estimators, the expectation of RSS_1 and RSS_2 is

$$\mathbb{E}(\text{RSS}_1) = I(J{-}1)\sigma_\epsilon^2 \qquad \text{and} \qquad \mathbb{E}(\text{RSS}_2) = J(I{-}1)\left(\sigma_b^2 + \sigma_\epsilon^2/J\right).$$

So, we get the estimates

$$\hat{\sigma}_\epsilon^2 = \frac{\text{RSS}_1}{I(J-1)} \qquad \text{and} \qquad \hat{\sigma}_b^2 = \frac{\text{RSS}_2}{J(I-1)} - \frac{\hat{\sigma}_\epsilon^2}{J}.$$

For more general mixed effects models, we could use maximum likelihood estimation. However, maximum likelihood estimation has drawbacks in the context of mixed effect models. One particular drawback is that variance estimates are biased. For example, for a sample of i.i.d. normally distributed data $y_1, \dots, y_n$, the MLE is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

while we saw that a denominator of $n-1$ is needed for an unbiased estimator. Given that the number of levels may not be large, the bias may be quite large.

Therefore, the principle of restricted maximum likelihood estimation is used instead of the usual maximization, in practice. In essence, this involves maximizing the likelihood of $\mathbf{K}^T\mathbf{Y}$ under the restriction that $k^T\mathbf{X} = 0$ for all columns k of the matrix $\mathbf{K}$. In this setting, the observations have the distribution

$$\mathbf{K}^T\mathbf{Y} \sim \mathrm{MVN}_n\left(0\,,\,\mathbf{K}^T\mathbf{V}(\underline{\theta})\mathbf{K}\right),$$

thus removing all fixed effects parameters.

Hence we can proceed by estimating the random effects first, and then estimate the fixed effects parameters. This approach is equivalent (not shown here) to doing estimation via profile likelihood.

14.4 Hypothesis Testing, Model Comparison and Diagnostics

14.4.1 Fixed Effects

Since the estimates for $\underline{\beta}$ are often heavily dependent on $\underline{\theta}$, we usually perform inference about fixed effects by conditioning on $\underline{\hat{\theta}}$ (i.e., treating $\underline{\theta}$ as if fixed at its estimated value). We can also then perform hypothesis tests for the fixed effects, as done in the classical normal linear model.

The reason that this conditional approach is usually preferable to an approach based on asymptotic likelihood theory, is that the approximations involved in conditioning on $\underline{\hat{\theta}}$ are usually better than those involved in using the large sample likelihood results at finite sample sizes.

Now suppose that the linear mixed models $\mathcal{M}_1$ and $\mathcal{M}_2$ only differ in terms of their fixed effects, with $\mathcal{M}_1$ being nested in $\mathcal{M}_2$. Then, we can compare the models using the likelihood ratio test statistic (deviance statistic)

$$2\left[\ell\left(\underline{\hat{\beta}}^{(2)}, \hat{\mathbf{b}}^{(2)}, \underline{\hat{\theta}}^{(2)}\right) - \ell\left(\underline{\hat{\beta}}^{(1)}, \hat{\mathbf{b}}^{(1)}, \underline{\hat{\theta}}^{(1)}\right)\right].$$

As in the GLM case, the test statistic is chi-square distributed with degrees of freedom equal to the difference in the dimensions of the two parameter spaces.

Note, we cannot use the REML estimates when using this approach for model comparison. The reason is that these estimates are obtained by considering linear combinations of the data that remove fixed effects. So, we have to derive estimates using the classical maximum likelihood approach if we wish to do a likelihood ratio test.

Similar to the GLM setting, we can also perform model comparison using the AIC.

14.4.2 Random Effects

Models differing in their random effects structure can be compared using likelihood ratio tests. Specifically, if ℓ_2 is the maximized log-likelihood of a model with p_2 parameters, and ℓ_1 is the maximized log-likelihood of a reduced version of the model (i.e. one with a simplified random effects structure) with p_1 parameters, then if the reduced model is correct

$$2(\ell_2 - \ell_1) \sim \chi^2_{p_2 - p_1}.$$

An important case we wish to test is $H_0 : \sigma_b^2 = 0$. However, the validity of comparing the likelihood ratio test statistic to a chi square distribution relies on the null hypothesis lying in the interior of the parameter space. Since $\sigma_b^2 = 0$ is the lower bound of the parameter space, this assumption does not hold, and the distribution of the test statistics is in general unknown.

When using the chi-square distribution, the test tends to be conservative - the p-values will tend to be larger than they should be. So, care must be taken when using this result for borderline results (i.e., p-values close to the significance level) since large sample results are very approximate – in this case bootstrapping may be a better approach. This also applies more generally in mixed model settings, where the χ^2 distribution can be very approximate.

A parametric bootstrapping approach can be performed to get better estimates of the variance of the parameters.

14.4.3 Prediction

Suppose we aim to predict the next observation for one of the $I = 5$ women in Section 13.2. Then, the predicted value is given by

$$\hat{\mu}_i = \hat{\mu} + \hat{b}_i,$$

where $\hat{\mu}$ is the estimated fixed effect and $\hat{b}_i$ is the best linear unbiased predictor for the person effect b_i .

Now consider that another woman is added to study. Then, our estimates for her estrone levels are

$$\hat{\mu}_{I+1} = \hat{\mu}.$$

14.4.4 Diagnostics

As for the normal linear model, we have to check whether the residuals are normally distributed. The estimated residuals are defined as

$$\underline{\hat{\epsilon}} = \mathbf{y} - \mathbf{X}\underline{\hat{\beta}} - \mathbf{Z}\hat{\mathbf{b}}.$$

So, we then derive the standardized residuals

$$\hat{r}_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_\epsilon}, \qquad i = 1 \ldots, N.$$

and use the PP plot

$$\left\{ \Phi(\hat{r}^{(i)}), \frac{i}{n+1} \right\}$$

or the QQ plot

$$\left\{ \hat{r}^{(i)}, \Phi^{-1}\left(\frac{i}{n+1}\right) \right\}$$

for $i = 1, \ldots, N$, where Φ^{-1} is the inverse of the standard normal cumulative distribution function.

We should also check whether the estimated residuals and fitted values are independent, as well as the estimated residuals and the observed explanatory variables; see Chapter 7. We can also use Cook's Distance to measure the influence of the individual observations.

15 Mixed Effects Models: Nested and Crossed Designs

15.1 Introduction

In this chapter, we describe the specification of linear mixed models with crossed random effects and compare it to the specification of linear mixed models with nested random effects.

Essentially,

- Nested effects are such that the variation of levels of one factor are completely contained within the levels of another factor.

- If the levels of a factor vary across levels of another, then the effects are said to be crossed.

To make the difference between the nested and crossed random effects more concrete, consider the following example.

Example 15.1.

Suppose we wish to assess the precision of machines in a manufacturing plant, which produce engineering components from a range of moulds. In the experiments, each machine, from a randomly selected group of N machines, produces n components out from each of P moulds, from which we measure the precision.

Let us denote by $Y_{i,j,s}$ the precision measurement obtained for the s-th component ($s = 1, \dots, n$) from the j-th mould ($j = 1, \dots, P$) for the i-th machine ($i = 1, \dots, N$). In other words, for each machine, we obtain in total $P \cdot n$ measurements.

We may be interested in assessing the influence of the effect of machine and mould on the measurements (i.e., variability between machines, components and moulds). Given that the machines and moulds are selected at random, we treat their effects as random.

15.2 A Model with Nested Random Effects

Referring to Example 15.1, suppose that the experiment is run so that each of the P series of n components for each machine is obtained from a different mould. We could then model the measurements from the experiment as

$$Y_{i,j,s} = \mu + b_{1,i} + b_{12,ij} + \epsilon_{i,j,s},$$

where

- $b_{1,i} \sim \text{Normal}\left(0, \sigma_M^2\right)$ is the random effect corresponding to machine i,

- $b_{12,ij} \sim \text{Normal}\left(0, \sigma_P^2\right)$ is the random effect corresponding to mould j specific to machine i (independent of $b_{1,i}$),

- $\epsilon_{i,j,s} \sim \text{Normal}(0, \sigma^2)$ is the residual (measurement) error, independent of both $b_{1,i}$ and $b_{2,ij}$.

Note that the mould effects are specific to (i.e., nested with) each machine. As a result, the model includes $N \cdot P$ mould effects. To indicate the nesting, we use the index ij in the symbolic representation of the random mould effect $b_{2,ij}$.

To summarize: the mould factor appears ONLY within a particular level of the machine factor (each mould belongs to a particular machine, and only that machine).

15.3 A Model with Crossed Random Effects

Alternatively, let us now assume that the experiment is run with only P moulds, so that the first n components for each machine are obtained from (the same) mould 1, the second n components from (the same) mould 2, and so on.

The measurements from the experiments can then be modelled as

$$Y_{i,j,s} = \mu + b_{1,i} + b_{2,j} + \epsilon_{i,j,s},$$

where

- $b_{1,i} \sim \text{Normal}\left(0, \sigma_M^2\right)$ is the random effect corresponding to machine i ,

- $b_{2,j} \sim \text{Normal}\left(0, \sigma_P^2\right)$ is the random effect corresponding to mould j (independent of $b_{1,i}$),

- $\epsilon_{i,j,s} \sim \text{Normal}\left(0, \sigma_{\epsilon}^{2}\right)$ is the residual (measurement) error independent of both $b_{1,i}$ and $b_{2,j}$.

Note that, compared to the nested setup in Section 15.2, the mould effects are no longer specific to machines, but each mould effect b_{j} remains the same for all machine effects b_{i} . We can say that the effects are crossed. As a result, the model with crossed effects includes only m mould effects. To indicate the crossing of the random effects, we use the index j in the symbolic representation of the random mould effect $b_{2,j}$.

To summarize: for crossed effects a given factor appears IN MORE THAN ONE level of another factor.

15.4 Notation of Nested and Crossed Random Effects Models

Note that in the nested model specification above, we used compound index notation to indicate that the (second level) mould effects were specific to machine i .

Following this notation, for all data the model formulation

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \mathbf{Z}_1\mathbf{b}_1 + \mathbf{Z}_{12}\mathbf{b}_{12} + \underline{\epsilon},$$

where $\mathbf{b}_1$ and $\mathbf{b}_{12}$ are vectors of random machine effects and mould effects, respectively.

By using the single subscript in $\mathbf{b}_1$, we indicate that the random effects contained in the vector $\mathbf{b}_\mathbf{1}$ are related to the levels of the first grouping factor, while by using the double subscript in $\mathbf{b}_{12}$, we indicate that the random effects contained in the vector $\mathbf{b}_{12}$ are related to the levels of the second grouping factor and are nested within the levels of the first factor. Note that $\mathbf{b}_1$ and $\mathbf{b}_{12}$ are constructed by "stacking", for all machines, the vectors of random machine and mould effects, respectively.

On the other hand, for model with a crossed design, the all-data specification would notationally be

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \mathbf{Z}_1\mathbf{b}_1 + \mathbf{Z}_2\mathbf{b}_2 + \underline{\epsilon},$$

where the notation is such that the random machine effects, contained in the vector $\mathbf{b_1}$, are related to the levels of the first grouping factor, while the random mould effects, contained in the vector $\mathbf{b_2}$, are related to the levels of the second grouping factor and are crossed with the random effects $\mathbf{b_1}$.

We can continue with this notation. In a similar manner, a model equation written as

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \mathbf{Z}_1\mathbf{b}_1 + \mathbf{Z}_{12}\mathbf{b}_{12} + \mathbf{Z}_{13}\mathbf{b}_{13} + \underline{\epsilon},$$

would be interpreted as specifying a linear mixed model with three sets of random effects:

1. The first set, represented by the vector $\mathbf{b}_1$, related to the levels of the grouping factor indexed by the first index.

2. The second set, represented by the vector $\mathbf{b}_{12}$, related to the levels of the grouping factor indexed by the second index, with the effects nested within the levels of the first factor.

3. The third set, represented by the vector $\mathbf{b}_{13}$, related to the

levels of the grouping factor indexed by the third index, with the effects nested within the levels of the first factor.

Note that the random effects contained in the vectors $\mathbf{b}_{12}$ and $\mathbf{b}_{13}$ are crossed within the levels of the first grouping factor.

16 Generalized Linear Mixed Models

16.1 Introduction

In Chapter 8, we extended the linear model to consider non-continuous response data, such as counts or true/false observations. Similarly, we now extend the linear mixed model in Chapter 13 to non-normal responses in order to form generalized linear mixed models (GLMMs).

Suppose that the data we wish to analyse are grouped, with I denoting the number of groups, and that we have J observations per group. We can specify a GLMM by assuming that the observations, conditional on the random effects $\mathbf{b}$, are independently distributed according to an exponential family distribution from

Definition 8.1, i.e., the probability density can be written as

$$f(y \mid \mathbf{b}, \theta, \phi) = \exp\left\{\frac{y\theta - d(\theta)}{a(\phi)} + c(y, \phi)\right\}.$$

We use $d(\cdot)$ here instead of $b(\cdot)$, as in Definition 8.1, to avoid confusion with the random effects $\mathbf{b}$.

We can think of the model as having a predictor of the form $\eta_{i,j} = \mathbf{x}_i^T \underline{\beta} + \mathbf{z}_i^T \mathbf{b}$ with an associated link function $g(\cdot)$ dependent on the exponential family distribution. Consequently, similar to the GLM framework, a GLMM is specified by three components:

1. The linear predictor $\eta_{i,j} = \mathbf{x}_i^T \underline{\beta} + \mathbf{z}_i^T \mathbf{b}$.

2. The link function $g(\mu_{i,j}) = \eta_{i,j}$ mapping the linear predictor to the mean of the distribution.

3. The probability distribution from the exponential family describing the observed data, $Y_{i,j} \sim F(\mu_{i,j})$.

Conditional on the random effects $\mathbf{b}$, $Y_{i,j}$ is independently distributed with mean $\mu_{i,j}$ and variance $\phi V(\mu_{i,j})$, i.e., $g(\mu_{i,j}) = g\left[\mathbb{E}(Y_{i,j})\right] = \eta_{i,j}$ (with the canonical link function, we can set $\theta_{i,j} = \eta_{i,j}$ as usual).

The marginal variance of the observations is

$$\mathrm{Var}(Y_{i,j}) = \mathbb{E}\{\mathrm{Var}(Y_{i,j} \mid \mathbf{b})\} + \mathrm{Var}\{\mathbb{E}(Y_{i,j} \mid \mathbf{b})\}$$

$$= \phi\,\mathbb{E}\{V(\mu_{i,j})\} + \mathrm{Var}(\mu_{i,j}),$$

thus separating the contributions to the variance from the exponential family model and that from the random effects. The distribution of the random effects is usually assumed to be the Gaussian distribution for computational tractability. Note that overdispersion can come from both the exponential family (in ϕ) or from the specification of random effects $\mathbf{b}$.

Remember that as before with GLMs, (random) contributions to the predictor no longer play a strictly additive role and instead can have a multiplicative effect.

16.2 Example: Seed Germination

In an experiment to assess the variability in seed germination, 20 seeds were planted on each of 10 experimental plates, and the number of seeds that germinated after a fixed period of time were counted. Our interest lies in accounting for plate-to-plate variability.

What are the values of I and J?

We have $I =$ ___ and $J =$ ___.

The data is of the following form:

```
## load("seeds.rda")
seeds
```

```
##    Plate Germinated Total NotGerminated
## 1      1          6    20            14
## 2      2          3    20            17
## 3      3         10    20            10
## 4      4         11    20             9
## 5        5          16     20             4
## 6        6           5     20            15
## 7        7           9     20            11
## 8        8           9     20            11
## 9        9           4     20            16
## 10      10          10     20            10
```

Let Y_i denote the number of seeds which germinated on plate i ($i = 1, \dots, 10$). Due to the fixed number of potentially germinated seeds on each plate, we model Y_i as

$$Y_i \sim \text{Binomial}\,(20, p_i)\,.$$

Note, this is equivalent to modelling $Y_{i,j} \sim \text{Binomial}(1, p_i)$ ($j = 1, \dots, 20$).

If we assume p_i to be the same across the $I = 10$ plates, i.e., $p_i = p$, then the maximum likelihood estimate is $n\hat{p} = 8.3$ (a total of 83 seeds germinated from $I = 10$ plates):

```
mean(seeds$Germinated)
```

```
## [1] 8.3
```

This estimate suggests that the variance should be about $n\hat{p}(1-\hat{p}) = 4.9$.

Let's compare this estimated variance to the empirical variance:

```
var(seeds$Germinated)
```

```
## [1] 15.12
```

We find that the empirical variance is `15.12` , which suggests that the model assuming a common probability across plates is not suitable.

Since we wish to assess the variability between plates, we would model these as random effects. More specifically, a plausible suggestion could be based on the logit link function with

$$\text{logit } p_i = \beta_1 + b_i \qquad (i = 1, \dots, I),$$

where β_1 is a fixed parameter and b_i represents the random plate effect. To complete the model, we could define

$$\mathbf{b} = (b_1, \dots, b_{10}) \sim \text{MVN}_I\left(0\,,\, \sigma_b^2\right).$$

16.3 Parameter Estimation

Suppose that the random effects $\mathbf{b}$ have a density $h(\mathbf{b} \mid \underline{\gamma})$ with parameters $\underline{\gamma}$. Then, with the canonical link, we can write the likelihood function as

$$L\left(\underline{\beta}, \phi, \underline{\gamma}\right) = \prod_{i=1}^{n} \int f\left(y_i \mid \underline{\beta}, \phi, \mathbf{b}\right) h\left(\mathbf{b} \mid \underline{\gamma}\right) \, \mathrm{d}\mathbf{b},$$

where we usually (as with the linear mixed model case) assume the random effects are such that $\mathbf{b} \sim \text{MVN}_I(0, \mathcal{D})$.

In the case where $f(\cdot)$ and $h(\cdot)$ are normal densities (the Normal-Normal model), the likelihood can be written explicitly in a nicer form. However, due to the complexity of the interplay between fixed and random effects in non-normal mixed models, inference is in general not straightforward. We outline two approaches in the following:

1. Approximation of the probability distribution (Penalized Quasi-Likelihood).

2. Approximation of the integral (Gauss-Hermite Quadrature).

16.3.1 Penalized Quasi-Likelihood (PQL)

The PQL method takes a iterative numerical algorithm, similar to the IRWLS for estimating a GLM. More specifically, we can view the generalized mixed model in the likelihood framework and for some fixed iteration k , define an adjusted variable

$$\mathbf{v}_k = \underline{\hat{\eta}_k} + (\mathbf{y} - \underline{\hat{\mu}_k})^T \left. \frac{\mathrm{d}\underline{\eta}}{\mathrm{d}\underline{\mu}} \right|_{\underline{\eta_k}} .$$

This is (as in the non-mixed case) a linearised version of the response variable, where the linearisation stems from a Taylor expansion of the likelihood (ignoring higher order terms). Hence we are using an approximation to the likelihood, the method outlined in this section is deemed a quasi-likelihood approach.

The approximation to the likelihood means that we can assume a (linear) mixed model for the adjusted variable of the form

$$\mathbf{v} \mid \mathbf{b} \sim \mathrm{MVN}_n\left(\mathbf{X}\underline{\beta} + \mathbf{Z}\mathbf{b}\,,\, \mathbf{W}^{-1}\phi\right), \qquad \mathbf{b} \sim \mathrm{MVN}_I\left(0\,,\, \mathcal{D}\right),$$

where $\mathbf{W}$ is the same diagonal matrix as in the GLM case. Thus we can use a iterative method in which we fit this linear mixed model at a particular iteration of the algorithm.

In what follows, let $\underline{\theta}$ be the vector of variance parameters for both fixed and random effects. To derive parameter estimates via the Penalized Quasi-Likelihood (PQL) approach, the following steps are performed

1. Set initial estimates $\hat{\mathbf{b}}_0$ and $\underline{\hat{\beta}_0}$. This could be done by setting $\hat{\mathbf{b}}_0 = 0$ and solving for $\underline{\hat{\beta}_0}$.

2. Compute the adjusted variable $\mathbf{v}_0$.

3. Compute the weight matrix $\mathbf{W}_0$.

4. Fit the linear mixed model

$$\mathbf{v}_0 = \mathbf{X}\underline{\beta} + \mathbf{Z}\mathbf{b} + \underline{\epsilon},$$

with $\underline{\epsilon} \sim \text{MVN}_n\left(0, \mathbf{W}^{-1}\phi\right)$ and $\mathbf{b} \sim \text{MVN}_I(0, \mathcal{D})$ to obtain estimates $\underline{\hat{\beta}}_1$, $\underline{\hat{\theta}}_1$, $\hat{\phi}_1$ and $\hat{\mathbf{b}}_1$.

5. Iterate steps 2-4 until convergence (subject to some tolerance).

The advantage of PQL estimation is that it is easy to implement. However, since the linearisation is essentially using an approximation to the likelihood, inference based on PQL will be approximate.

In **R**, we can estimate a GLMM via the PQL approach using the **`glmmPQL`** function in the **`MASS`** package.

16.3.2 Gauss-Hermite Quadrature (numerical approximation)

Whilst the PQL is an attractive method for finding parameters in non-normal mixed models, we can also approximate the (true) likelihood via numerical methods. In particular, for integrals of the form

$$\int_{-\infty}^{\infty} f(x)\exp(-x^2)\,\mathrm{d}x$$

(i.e., the likelihood in GLMMs), we can approximate the integral using a so-called Gauss-Hermite numerical approach; the function in the integral is evaluated at a number of points – the approximation error decreases as more function evaluations are used.

Such numerical methods are usually better than the PQL approach, because the inference based on such approximations is more reliable. However, these methods are in general computationally expensive.

In general we will prefer to use this numerical integration of the likelihood rather than PQL.

In **R**, we can estimate a GLMM via the Gauss-Hermite quadrature approach using the **`glmer`** function in the **`lme4`** package.

16.4 Example: Prevalence of Wheezing amongst Children in Ohio

Data were recorded for $I = 536$ children Ohio. Each child was studied for four years from age seven to ten, and we observed whether they wheezed, $y_{i,j} = 1$, or not, $y_{i,j} = 0$ ($i = 1, \ldots, I;\ j = 7, 8, 9, 10$). The explanatory variables are the child's age and whether the parents smoked when the child was seven years old. In the data set, age is shifted to values -2, -1, 0 and 1, with a value of 0 corresponding to the child being nine years old.

```r
library(faraway, warn.conflicts = F)
data("ohio")
head(ohio)
```

```
##   resp id age smoke
## 1    0  0  -2     0
## 2    0  0  -1     0
## 3    0  0   0     0
## 4    0  0   1     0
## 5    0  1  -2     0
## 6    0  1  -1      0
```

Suppose that there is no subject-level effect. We would then define a GLM

$$Y_{i,j} \sim \text{Bernoulli}(\mu_{i,j})$$

$$\text{logit}\,(\mu_{i,j}) = \mathbf{x}_{i,j}^T \underline{\beta}.$$

Let's estimate this model in **R**:

```
estim = glm(resp ~ age + smoke, data = ohio, family = binom
summary(estim)
```

```
##
## Call:
## glm(formula = resp ~ age + smoke, family = binomial, dat
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.8837     0.0838   -22.5   <2e-16 ***
## age          -0.1134     0.0541    -2.1    0.036 *
## smoke         0.2721     0.1235     2.2    0.028 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1829.1  on 2147  degrees of freedom
## Residual deviance: 1819.9  on 2145  degrees of freedom
## AIC: 1826
##
## Number of Fisher Scoring iterations: 4
```

The estimates for the regression coefficients are $\underline{\hat{\beta}} = (-1.88, -0.11, 0.27)^T$, and all effects are significant at the 5% significance level.

Let's now consider the inclusion of subject-level effects. The GLMM is then

$$Y_{i,j} \sim \text{Bernoulli}(\mu_{i,j})$$

$$\text{logit}\,\mu_{i,j} = \mathbf{x}_{i,j}^T \underline{\beta} + z_{i,j}^T \mathbf{b}$$

$$\mathbf{b} \sim \text{MVN}_I(0, \sigma_b^2 \mathcal{I}_I)$$

We now estimate this model using the Gauss-Hermite quadrature approach:

```r
library(lme4)
estimGH = glmer(resp ~ age + smoke + (1|id), data = ohio, n
summary(estimGH)
```

```
## Generalized linear mixed model fit by maximum likelihood
##  Family: binomial  ( logit )
## Formula: resp ~ age + smoke + (1 | id)
##    Data: ohio
##
##      AIC      BIC   logLik deviance df.resid
##   1603.3   1626.0   -797.6   1595.3     2144
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -1.373 -0.201 -0.177 -0.149  2.508
```

```
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 4.69     2.16
## Number of obs: 2148, groups:  id, 537
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.1015     0.2191  -14.16   <2e-16 ***
## age          -0.1756     0.0677   -2.60   0.0095 **
## smoke         0.3986     0.2731    1.46   0.1444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Correlation of Fixed Effects:
##       (Intr) age
## age    0.244
## smoke -0.493 -0.008
```

We obtain $\underline{\widehat{\beta}} = (-3.10, -0.18, 0.40)^T$. Note, the smoke effect is no longer significant at the 5% significance level, even though the regression coefficient is larger than for the GLM.

Let's compare the results to the estimates obtained via PQL:

```
library(MASS)
estimPQL = glmmPQL(resp ~ age + smoke, random = ~1|id, data
summary(estimPQL)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: ohio
##   AIC BIC logLik
##    NA  NA     NA
##
## Random effects:
##  Formula: ~1 | id
##         (Intercept) Residual
## StdDev:       2.057   0.6356
##
## Variance function:
##  Structure: fixed weights
##  Formula: ~invwt
## Fixed effects:  resp ~ age + smoke
##                Value Std.Error   DF t-value p-value
## (Intercept) -2.7658   0.14218 1610 -19.453  0.0000
## age         -0.1816   0.04365 1610  -4.160  0.0000
## smoke        0.3252   0.23132  535   1.406  0.1604
##  Correlation:
##       (Intr) age
## age    0.197
## smoke -0.591 -0.003
##
## Standardized Within-Group Residuals:
##     Min      Q1      Med      Q3     Max
## -2.6145 -0.2829 -0.2584 -0.2155  3.4444
##
## Number of Observations: 2148
## Number of Groups: 537
```

We find $\underline{\hat{\beta}} = (-2.77, -0.18, 0.33)^T$. Again, the smoke effect does not appear to be significant. Furthermore, note that the estimates obtained by PQL and Gauss-Hermite quadrature are similar, but not identical.

16.5 Inference and Diagnostics

16.5.1 Distributional Results and Model Selection

Note that the approximation to the likelihood justifies some distributional results which we can use (with care!) for inference. In particular, the result for the MLE (using IRWLS) from GLMs becomes

$$\underline{\widehat{\beta}}(\mathbf{Y}) \sim \mathrm{MVN}_p\left(\underline{\beta}, (\mathbf{X}^T(\mathbf{Z}\mathcal{D}\mathbf{Z}^T + \mathbf{W}^{-1}\phi)^{-1}\mathbf{X})^{-1}\right),$$

for large samples. This asymptotic normal distribution can be used for hypothesis tests and confidence intervals, where we may substitute in an estimate of the dispersion parameter ϕ.

As before, the AIC can still be used for model comparison.

Example 16.1 (Ohio wheeze data).
The GLM considered in Section 16.4 yields an AIC of 1825.9, while the AIC for the GLMM is 1597.9. Consequently, we would select the GLMM because it provides a lower AIC.

We can also still consider models with offset terms and overdispersion in a similar way to GLMs.

16.5.2 Diagnostics

In the generalized linear mixed model case, as with GLMs, we can

- Examine (deviance) residuals for patterns and outliers (e.g. with half-normal plots), or

- Examine the residuals vs the linear predictor.

Additional aspects to examine are the random effects and covariance structure. The difference between the fixed effects models and the mixed models in this chapter is that we have random effects which change the estimation and considerations of variability in the model.

We can examine the random effects using plots too. For example, we can consider:

- Dot plots of random effects to see if any post-analysis would be of interest.

- Check normality of random effects (PP plots or QQ plots).

- Check whether autocorrelation exists in the residuals.

Example 16.2.

Let's revisit the seed germination data in Section 16.2. We fit a GLMM with $\eta_i = \beta_1 + b_i$, logit link function and binomial distribution function:

```
library(lattice, warn.conflicts = FALSE)
estimGS = glmer(cbind( Germinated, NotGerminated) ~ (1|Plat
```

To investigate the difference between plates, we create a dot plot:

```
dotplot(ranef(estimGS))
```

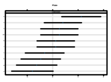

Figure 16.1: Dot plot for the seed germination data which shows the estimated random effects for the different plates.

```
## $Plate
```

The results show that the random effects ranges from less than -1 (Plate 2) to about 1 (Plate 5).

A Common distributions

A.1 Discrete random variables

A.1.1 Binomial & Bernoulli random variables

	Basic Information
Notation:	$X \sim \text{Binomial}(n, p)$, with <br> + n number of independent trials, and <br> + p probability of success
Sample space/support:	$x \in S = \{0, 1, \dots, n\}$
Parametric space:	$n \geq 1$, $p \in (0, 1)$
Probability function:	For $x \in S$, $f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$, and $f_X(x) = 0$ otherwise
Cumulative distribution function:	For $0 \leq x \leq n-1$, $F_X(x) = \sum_{y=0}^{x} \binom{n}{y} p^y (1-p)^{n-y}$. If $x < 0$, $F_X(x) = 0$, and

Properties:

- For $n = 1$ (a single trial), $X \sim \text{Binomial}(1, p) \equiv \text{Bernoulli}(p)$.

- If $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Binomial}(m, p)$, with X and Y independent, then $X + Y \sim \text{Binomial}(n + m, p)$ (therefore, the sum of n independent Bernoulli random variables of parameter p is equivalent to a Binomial random variable of parameters n and p).

A.1.2 Geometric random variables

	Basic Information
Notation:	$X \sim \text{Geometric}(p)$, with p probability of first success in a sequence of (possible infinite) trials.
Sample space/support:	$x \in S = \{1, 2, 3, ...\}$
Parametric space:	$p \in (0, 1)$
Probability function:	For $x \in S$, $$f_X(x) = p(1-p)^{x-1},$$ and $f_X(x) = 0$ otherwise
Cumulative distribution function:	For $x \in S$, $$F_X(x) = 1 - (1-p)^x.$$ If $x < 1$, $F_X(x) = 0$.
Moment generating function:	$M_X(t) = \mathbb{E}(e^{tX}) = \frac{pe^t}{1-e^t(1-p)}$ for $t < -\ln(1-p)$
Expectation:	$\mathbb{E}(X) = \frac{1}{}$

Properties:

- Memory absence – If $X \sim \text{Geometric}(p)$, then $\mathbb{P}(X > u + t \mid X > u) = \mathbb{P}(X > t)$, with $u, t \in S$.

- Negative Binomial connection – If each $X_1, \ldots, X_k$ are independent and follows the same $\text{Geometric}(p)$ distribution, then $Y = \sum_{i=1}^{k} X_i \sim \text{Negative Binomial}(k, p)$, expressing the k-th success with probability p in a series of (possible infinite) trials (expected value and variance are therefore $\mathbb{E}(Y) = \frac{k}{p}$ and $\text{Var}(Y) = \frac{k(1-p)}{p^2}$, respectively).

A.1.3 Poisson random variables

	Basic Information
Notation:	$X \sim \text{Poisson}(\lambda)$, with λ representing the expected number of counts per units of interest (e.g., average number of children per woman, or average yield per year and/or region, etc.)
Sample space/support:	$x \in S = \{0, 1, 2, ...\}$
Parametric space:	$\lambda > 0$
Probability function:	For $x \in S$, $f_X(x) = e^{-\lambda}\frac{\lambda^x}{x!}$, and $f_X(x) = 0$ otherwise
Cumulative distribution function:	For $x \in S$, $F_X(x) = \sum_{y=0}^{x} e^{-\lambda}\frac{\lambda^y}{y!}$. If $x < 0$, $F_X(x) = 0$.
Moment generating function:	$M_X(t) = \mathbb{E}(e^{tX}) =$

Properties:

- If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$, with X and Y independent, then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

- Binomial tends to Poisson – Let $\{X_n\}_{n \geq 1}$ be a sequence of random variables, each $X_n \sim \text{Binomial}(n, p_n)$ with $p_n = \frac{\lambda}{n}$ for $\lambda < n$ fixed. Then X_n converges in distribution to $X \sim \text{Poisson}(\lambda)$ (i.e, $\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$ for all $x \in \mathbb{R}$ at which F is continuous.)

- Modelling choices – In applications, whenever the response variable is count data, possible high-demand choices for modelling it are binomial[6], Poisson or negative binomial distributions. How to choose which one to use?

  - If $\frac{\text{variance}}{\text{expected value}} < 1$, the binomial model might be the more recommended.

[6] If the outcome is not binary (i.e., success and failure are not the only possible two outcomes), usually the multinomial distribution is used.

- If $\frac{\text{variance}}{\text{expected value}} = 1$, the Poisson model might be the more recommended.

- If $\frac{\text{variance}}{\text{expected value}} > 1$, the negative binomial model might be the more recommended (overdispersion - see Section 11.2).

A.2 Continuous random variables

A.2.1 Uniform random variables

	Basic Information
Notation:	$X \sim \text{Uniform}(a, b)$
Sample space/support:	$x \in S = [a, b]$
Parametric space:	$a, b \in \mathbb{R}$ with $b > a$
Probability function:	For $x \in S$, $f_X(x) = \dfrac{1}{b-a},$ and $f_X(x) = 0$ otherwise
Cumulative distribution function:	For $x \in S$, $F_X(x) = \dfrac{x-a}{b-a}.$ If $x < a$, $F_X(x) = 0$, and if $x > b$, $F_X(x) = 1$
Moment generating function:	For $t \neq 0$, $M_X(t) = \mathbb{E}(e^{tX}) = \dfrac{e^{tb} - e^{ta}}{t(b-a)}.$ For $t = 0$, $M_X(t) = 1.$

Properties:

- If $X \sim \text{Uniform}(a, b)$ and $Y = \mu + \sigma X$, with $\mu \in \mathbb{R}$ and $\sigma > 0$, then $Y \sim \text{Uniform}(\mu + \sigma a, \mu + \sigma b)$.

A.2.2 Gamma & Exponential random variables

	Basic Information
Notation:	$X \sim \text{Gamma}(\alpha, \lambda)$, where α is the shape parameter and λ is the rate parameter
Sample space/support:	$x > 0$
Parametric space:	$\alpha, \lambda > 0$
Probability function:	For $x > 0$, $f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x},$ and $f_X(x) = 0$ otherwise
Cumulative distribution function:	For $x > 0$, $F_X(x) = \int_0^x \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y} \, dy.$ If $x < 0$, $F_X(x) = 0$
Moment generating function:	For $t < \lambda$, $M_X(t) = \mathbb{E}(e^{tX}) = \left(1 - \frac{t}{\lambda}\right)^{-\alpha}.$

Properties:

- The failure rate function is defined as

$$r_X(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{f_X(x)}{1-F_X(x)} & \text{if } x > 0 \end{cases},$$

  and, in particular,

  - $\alpha > 1$ implies an increasing failure rate,
  - $\alpha = 1$ implies a constant failure rate,
  - $\alpha < 1$ implies a decreasing failure rate.

- If $X \sim \text{Gamma}(\alpha, \lambda)$ and $Y = cX$, $c > 0$, then $Y \sim \text{Gamma}(\alpha, \lambda/c)$.

- If $X \sim \text{Gamma}(\alpha, \lambda)$ and $Y \sim \text{Gamma}(\beta, \lambda)$, with X and Y independent and $\alpha, \beta, \lambda > 0$, then $X + Y \sim \text{Gamma}(\alpha + \beta, \lambda)$.

- For $\alpha = 1$ (constant failure rate), $X \sim \text{Gamma}(1, \lambda) \equiv$ Exponential(λ) (therefore, the sum of n independent Exponential random variables of parameter λ is equivalent to a Gamma random variable of parameters n and λ).

A.2.3 Normal & related random variables

Basic Information

Notation: $X \sim \text{Normal}(\mu, \sigma^2)$

Sample space/support: $x \in \mathbb{R}$

Parametric space: $\mu \in \mathbb{R}$ and $\sigma > 0$

Probability function:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Cumulative distribution function:

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} \, dy$$

Moment generating function:

$$M_X(t) = \mathbb{E}(e^{tX}) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}.$$

Expectation: $\mathbb{E}(X) = \mu$

Variance: $\text{Var}(X) = \sigma^2$

Properties:

- If $X \sim \text{Normal}(\mu, \sigma^2)$ and $Y = a + bX$, with $a \in \mathbb{R}$ and $b > 0$, then $Y \sim \text{Normal}(a + b\mu, b^2\sigma^2)$.

- Special case if $\mu = 0$ and $\sigma = 1$, denoted as $Z \sim \text{Normal}(0, 1)$ and termed standard normal distribution.

- If $Z \sim \text{Normal}(0, 1)$ and $Y = Z^2$, then Y follows a chi-squared distribution with 1 degree of freedom, with notation $Y \sim \chi^2_1$.

  Note that:

  - $Y \sim \chi^2_1 \equiv \text{Gamma}(1/2, 1/2)$.

  - More broadly, given $\nu > 0$,

    * $Y \sim \chi^2_\nu \equiv \text{Gamma}(\nu/2, 1/2)$, with $\mathbb{E}(Y) = \nu$ and $\text{Var}(Y) = 2\nu$.

    * If $Y \sim \chi^2_{\nu_1}$ and $W \sim \chi^2_{\nu_2}$, with Y independent

to W and $\nu_1, \nu_2 > 0$, then $Y + W \sim \chi^2_{\nu_1+\nu_2}$ (therefore, the sum of n independent squared standard normal random variables is equivalent to a chi-squared random variable with n degrees of freedom).

* If $Y \sim \chi^2_{\nu_1}$ and $W \sim \chi^2_{\nu_2}$, with Y independent to W and $\nu_1, \nu_2 > 0$, then the ratio between Y and W, weighted by their degrees of freedom, follows a F distribution with ν_1 and ν_2 degrees of freedom. In terms of notation, we have $F = \left(\frac{Y}{\nu_1}\right) / \left(\frac{W}{\nu_2}\right) \sim F_{\nu_1,\nu_2}$.

- If $Z \sim \text{Normal}(0,1)$ and $Y \sim \chi^2_\nu$, with Z independent of Y, then $T = \frac{Z}{\sqrt{Y/\nu}}$ follows a (Student's) t distribution with ν degrees of freedom, and, in terms of notation, we have $T \sim t_\nu$.

- Asymptotic results:

  – Given $X \sim \text{Binomial}(n, p)$, with n sufficiently large and $p \in (0,1)$, then it is possible to approximate the distribution of X with $X \approx \text{Normal}(np, np(1-p))$.

- Given $X \sim \text{Poisson}(\lambda)$, with $\lambda > 10$, then it is possible to approximate the distribution of X with $X \approx \text{Normal}(\lambda, \lambda)$.

- Given $X \sim \text{Negative binomial}(k, p)$, with k sufficiently large and $p \in (0,1)$, then it is possible to approximate the distribution of X with $X \approx \text{Normal}(k/p, k(1-p)/p^2)$.

- Given $X \sim \text{Gamma}(\alpha, \lambda)$, with $\alpha > 15$, then it is possible to approximate the distribution of X with $X \approx \text{Normal}(\alpha/\lambda, \alpha/\lambda^2)$.

- Given $X \sim \chi^2_\nu$, with $\nu > 30$, then it is possible to approximate the distribution of X with $X \approx \text{Normal}(\nu, 2\nu)$.

B Hypothesis testing procedures

For the rest of this section, we denote with $\theta \in \Theta$ the parameter which characterises a population of interest and that we want to test, where Θ is the parametric space (e.g., $\theta = \mu$ and $\Theta = \mathbb{R}$ for normal data, where μ represents the average value of such data).

The main idea behind a hypothesis testing procedure is to collect and judge evidence provided by data collected or measured. If such data were very unlikely to have occurred by chance if the null hypothesis was true, then the evidence raises more than a reasonable doubt in our minds about the null hypothesis.

There are 4 core steps to consider:

STEP 1 - Defining the null and alternative hypotheses.

STEP 2 - One or two tailed test.

STEP 3 - Test statistic and critical value.

STEP 4 - Final decision.

B.1 STEP 1 – Null and alternative hypotheses

We define two contrasting hypotheses:

- the null hypothesis, denoted as H_0, which expresses the status quo, i.e., no changes or no differences in the population of between populations (e.g., no differences between the effects of two drugs, null value of a regression coefficient of a variable, etc.), and

- the alternative hypothesis, denoted as H_1 or H_A, which expresses a change (positive or negative) in the population or between populations (e.g., drug A is more effective than drug B, regression coefficient of a variable is positive, etc.). Usually, the alternative hypothesis is the primary focus of the researcher, i.e., the hypothesis for which the researcher collects evidence.

Writing H_1, and therefore H_0, will depend on the problem of interest.

- If interest lies in testing if θ is different from a specific value θ_0 (provided by logic, literature, stakeholders, or other), then the hypotheses are

  $$H_0 : \theta = \theta_0 \qquad \text{against} \qquad H_1 : \theta \neq \theta_0.$$

  An example of such framework is testing for non-zero regression coefficients in a linear model.

- If interest lies in testing if θ is greater than a specific value θ_0 , then the hypotheses are

  $$H_0 : \theta \leq \theta_0 \qquad \text{against} \qquad H_1 : \theta > \theta_0.$$

  An example of such framework is testing for rising temperatures in the recent years.

- If interest lies in testing if θ is lesser than a specific value θ_0 , then the hypotheses are

  $$H_0 : \theta \geq \theta_0 \qquad \text{against} \qquad H_1 : \theta < \theta_0.$$

  An example of such framework is testing for drug A more effective than drug B for reducing the risk of a particular disease.

B.2 STEP 2 – One or two-tailed test

When $H_1 : \theta \neq \theta_0$, the test is called two-tailed (sometimes, two-sided). If data are very unlikely to have occurred by chance under the null hypothesis, then they end up in the tails of the distribution of the test statistic and, as no direction is given in H_1, both tails are to be considered.

For example, let $H_1 : \mu \neq 5$ (therefore, $H_0 : \mu = 5$) for normal data. Then, if the data are very unlikely to have been generated by a normal process centred in $\mu = 5$, we will find these data values in the extreme tails of said normal distribution, implying that it is more likely that such data have been generated by a normal process with $\mu \neq 5$ (either μ is much smaller or much greater than 5). See Figure B.1 for a representation of such example.

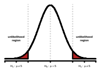

Figure B.1: Two-tailed test.

When $H_1 : \theta > \theta_0$, the test is called right one-tailed (sometimes, right one-sided). If data are very unlikely to have occurred by chance under the null hypothesis, then they end up in the right tail of the distribution of the test statistic.

For example, let $H_1 : \mu > 5$ (therefore, $H_0 : \mu \leq 5$) for normal data. Then, if the data are very unlikely to have been generated by a normal process centred in $\mu \leq 5$, we will find these data values in the extreme right tail of said normal distribution, implying that it is more likely that such data have been generated by a normal process with $\mu > 5$ (μ is much greater than 5). See Figure B.2 for a representation of such example.

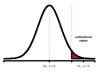

Figure B.2: Right one-tailed test.

When $H_1 : \theta < \theta_0$, the test is called left one-tailed (sometimes, left one-sided). If data are very unlikely to have occurred by chance under the null hypothesis, then they end up in the left tail of the distribution of the test statistic.

For example, let $H_1 : \mu < 5$ (therefore, $H_0 : \mu \geq 5$) for normal data. Then, if the data are very unlikely to have been generated by a normal process centred in $\mu \geq 5$, we will find these data values in the extreme left tail of said normal distribution, implying that it is more likely that such data have been generated by a normal process with $\mu < 5$ (μ is much smaller than 5). See Figure B.3 for a representation of such example.

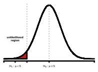

Figure B.3: Left one-tailed test.

B.3 STEP 3 – Test statistic and critical value

Evidence from data is obtained as transformation of the original random variables supposedly at the base of the process, and typically such transformation reflects the information that the parameter to be tested represents (e.g., if μ represents the mean of a population to be tested, then a related data transformation is the sample mean $\overline{X}$). Such transformation is called test statistic.

The core property of the test statistic is that, under the null hypothesis, its distribution is known and parameter free, e.g., Normal$(0,1)$, χ^2_ν with ν degrees of freedom, or others. For example, let $H_1 : \mu \neq 5$ (therefore, $H_0 : \mu = 5$) for normal data; then the test statistic is defined as $t_{stat} = (\overline{x} - \mu_{\text{under } H_0})/(s/\sqrt{n}) = (\overline{x} - 5)/(s/\sqrt{n})$, where s is the sample standard deviation and n is the sample size. Such value is a realization of the random variable

$$T = \frac{\overline{X} - 5}{S/\sqrt{n}} \sim t_{n-1}.$$

The distribution of the test statistic under the null hypothesis is fundamental in obtaining the critical value c_{value} , i.e., the quantile of order $1 - \alpha$ of such distribution that defines where the unlikelihood region starts, with α being the significance level of the test (i.e., how much unlikely is unlikely). In the previous example, the distribution is t_{n-1} which generates the critical value $c_{value} = t_{n-1}(1 - \alpha)$ (e.g., if $\alpha = 0.05$ and $n = 10$, then $t_9(0.95) = 1.833$ via **R** command **`qt(p = 0.95, df = 9)`**).

A clear distinction here is between two-tailed or one-tailed tests, as in the former case the unlikelihood region is composed by separated sets, and therefore requires more than one critical value.

B.4 STEP 4 – Final decision

If the absolute value of the test statistic is greater than the critical value, then the null hypothesis is rejected, as the evidence provided by data it is more unlikely to occur than if it would be under the null hypothesis. In other words, if the absolute value of the test statistic is more extreme than the supposed data in the null hypothesis, then the null hypothesis is rejected.

Sometimes, instead of comparing directly the observed test statistic with the critical value, their probabilities are calculated and juxtaposed. The probability of seeing a value more extreme than the critical value is α , the significance level of the test. Similarly, the probability of seeing a value more extreme than the observed test statistic is called the p-value the test, and the null hypothesis is rejected if the p-value is smaller than α (see Figure B.4).

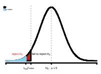

Figure B.4: Comparisons between test statistic and critical value, and between *p*-value and significance level of the test.

References

Casella, G., and Berger, R. (2001), Statistical inference, Cengage Learning, Inc.

D. C. Montgomery, E. A. P., and Vining, G. G. (2021), Introduction to linear regression analysis, John Wiley & Sons, Inc.

Dobson, A. J., and Barnett, A. G. (2018), An introduction to generalized linear models, CRC Press Taylor & Francis Group.

Draper, N. R., and Smith, H. (1998), Applied regression analysis, John Wiley & Sons, Inc.

Faraway, J. J. (2015), Linear models with r, CRC Press Taylor & Francis Group.

Faraway, J. J. (2016), Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models, CRC Press Taylor & Francis Group.

Fox, J. (2013), Applied regression analysis and generalized linear models, SAGE.

Pawitan, Y. (2013), In all likelihood: Statistical modelling and inference using likelihood, Clarendon Press.

Weisberg, S. (2014), Applied linear regression, John Wiley & Sons, Inc.