MATH3091: Statistical Modelling II

Dr Helen Ogden

Contents

Preface

The pre-requisite module MATH2010: Statistical Modelling I covered in detail the theory of linear regression models, where explanatory variables are used to explain the variation in a response variable, which is assumed to be normally distributed.

However, in many practical situations the data are not appropriate for such analysis. For example, the response variable may be **binary**, and interest may be focused on assessing the dependence of the probability of 'success' on potential explanatory variables. Alternatively, the response variable may be a **count** of events, and we may wish to infer how the rate at which events occur depends on explanatory variables. Such techniques are important in many disciplines such as finance, biology, social sciences and medicine.

The aim of this module is to cover the theory and application of what are known as **generalised linear models** (GLMs). This is an extremely broad class of statistical models, which incorporates the linear regression models studied in MATH2010, but also allows binary and count response data to be modelled coherently.

Chapter 1

Introduction

1.1 Elements of statistical modelling

Probability and statistics can be characterised as the study of variability. In particular, statistical inference is the science of analysing statistical data, viewed as the outcome of some random process, in order to draw conclusions about that random process.

Statistical models help us to *understand* the random process by which observed data have been generated. This may be of interest in itself, but also allows us to make *predictions* and perhaps most importantly *decisions* contingent on our inferences concerning the process.

It is also important, as part of the modelling process, to acknowledge that our conclusions are only based on a (potentially small) sample of possible observations of the process and are therefore subject to error. The science of statistical inference therefore involves assessment of the uncertainties associated with the conclusions we draw.

Probability theory is the mathematics associated with randomness and uncertainty. We usually try to describe random processes using probability models. Then, statistical inference may involve estimating any unspecified features of a model, comparing competing models, and assessing the appropriateness of a model; all in the light of observed data.

In order to identify 'good' statistical models, we require some principles on which to base our modelling procedures. In general, we have three requirements of a statistical model

- Plausibility
- Parsimony
- Goodness of fit

The first of these is not a statistical consideration, and a subject-matter expert usually needs to be consulted about this. For some objectives, like prediction, it might be considered unimportant. Parsimony and goodness of fit are statistical issues. Indeed, there is usually a trade-off between the two and our statistical modelling strategies will take account of this.

1.2 Regression models

Many statistical models, and all the ones we shall deal with in MATH3091, can be formulated as *regression* models.

In practical applications, we often distinguish between a *response* variable and a group of *explanatory* variables. The aim is to determine the pattern of dependence of the response variable on the explanatory variables. A regression model has the general form

response = function(structure and randomness)

The structural part of the model describes how the response depends on the explanatory variables and the random part defines the probability distribution of the response. Together, they produce the response and the statistical modeller's task is to 'separate' these out.

1.3 Example data to be analysed

1.3.1 `nitric`: Nitric acid

This data set relates to 21 successive days of operation of a plant oxidising ammonia to nitric acid. The response `yield` is ten times the percentage of

ingoing ammonia that is lost as unabsorbed nitric acid (an indirect measure of the yield). The aim here is to study how the yield depends on flow of air to the plant (`flow`), temperature of the cooling water entering the absorption tower (`temp`) and concentration of nitric acid in the absorbing liquid (`conc`). These data will be analysed in worksheet 2 using multiple linear regression models.

1.3.2 `birth`: Weight of newborn babies

This data set contains weights of 24 newborn babies. There are two explanatory variables, sex (`Sex`) and gestational age in weeks (`Age`) together with the response variable, birth weight in grams (`Weight`). The aim here is to study how birth weight depends on sex and gestational age. This data set will be analysed in worksheet 3 by using multiple linear regression models including both categorical and continuous explanatory variables.

1.3.3 `survival`: Time to death

This data set, analysed in worksheet 4, contains survival times in 10 hour units (`time`) of 48 rats each allocated to one of 12 combinations of 3 poisons (`poison`) and 4 treatments (`treatment`). The aim here is to study how survival time depends on the poison and the treatment, and to determine whether there is an interaction between these two categorical variables.

1.3.4 `beetle`: Mortality from carbon disulphide

This data set represents the number of beetles exposed (`exposed`) and number killed (`killed`) in eight groups exposed to different doses (`dose`) of a particular insecticide. Interest is focussed on how mortality is related to dose. It seems sensible to model the number of beetles killed in each group as the binomial random variable with probability of death depending on dose. This will be discussed in worksheet 5.

1.3.5 `shuttle`: Challenger disaster

This data set concerns the 23 space shuttle flights before the Challenger disaster. The disaster is thought to have been caused by the failure of a number of O-rings, of which there were six in total. The data consist of four variables, the number of damaged O-rings for each pre-Challenger flight (`n_damaged`), together with the launch temperature in degrees Fahrenheit (`temp`), the pressure at which the pre-launch test of O-ring leakage was carried out (`pressure`) and the name of the orbiter (`orbiter`). The Challenger launch temperature on 20th January 1986 was 31F. The aim is to predict the probability of O-ring damage at the Challenger launch. This will be discussed in worksheet 6.

1.3.6 `heart`: Treatment for heart attack

This data set represents the results of a clinical trial to assess the effectiveness of a thrombolytic (clot-busting) treatment for patients who have suffered an acute myocardial infarction (heart attack). There are four categorical explanatory variables, representing

- the `site` of infarction: anterior, inferior or other
- the `time` between infarction and treatment: ≤ 12 or > 12 hours
- whether the patient was already taking Beta-blocker medication prior to the infarction, `blocker`: yes or no
- the `treatment` the patient was given: active or placebo.

For each combination of these categorical variables, the dataset gives the total number of patients (`n_patients`), and the number who survived for for 35 days (`n_survived`). The aim is to find out how these categorical variables affect a patient's chance of survival. These data will be analysed in worksheet 7.

1.3.7 `accident`: Road traffic accidents

This example concerns the number of road accidents (`number`) and the volume of traffic (`volume`), on each of two roads in Cambridge (`road`), at various

times of day (`time`, taking values `morning`, `midday` or `afternoon`). We should be able to answer questions like:

1. Is Mill Road more dangerous than Trumpington Road?
2. How does time of day affect the rate of road accident?

These issues will be considered in worksheet 8.

1.3.8 `lymphoma`: Lymphoma patients

The `lymphoma` data set represents 30 lymphoma patients classified by sex (`Sex`), cell type of lymphoma (`Cell`) and response to treatment (`Remis`). It is an example of data which may be represented as a three-way ($2 \times 2 \times 2$) contingency table. The aim here is to study the complex dependence structures between the three classifying factors. This is taken up in worksheet 9.

Chapter 2

Parametric statistical inference

2.1 Introduction

Probability distributions like the binomial, Poisson and normal, enable us to calculate probabilities, and other quantities of interest (e.g. expectations) for a probability model of a random process. Therefore, given the model, we can make statements about possible outcomes of the process.

Statistical inference is concerned with the inverse problem. Given outcomes of a random process (observed data), what conclusions (inferences) can we draw about the process itself?

We assume that the n observations of the response $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ are observations of random variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, which have joint p.d.f. $f_{\boldsymbol{Y}}$ (joint p.f. for discrete variables). We use the observed data $\boldsymbol{y}$ to make inferences about $f_{\boldsymbol{Y}}$.

We usually make certain assumptions about $f_{\boldsymbol{Y}}$. In particular, we often assume that $y_1, \ldots, y_n$ are observations of *independent* random variables. Hence

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{Y_1}(y_1) f_{Y_2}(y_2) \cdots f_{Y_n}(y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i).$$

In parametric statistical inference, we specify a joint distribution $f_{\boldsymbol{Y}}$, for $\boldsymbol{Y}$, which is known, except for the values of parameters $\theta_1, \theta_2, \ldots, \theta_p$ (sometimes

denoted by $\boldsymbol{\theta}$). Then we use the observed data $\boldsymbol{y}$ to make inferences about $\theta_1, \theta_2, \ldots, \theta_p$. In this case, we usually write $f_{\boldsymbol{Y}}$ as $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})$, to make explicit the dependence on the unknown $\boldsymbol{\theta}$.

2.2 The likelihood function

We often think of the joint density $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})$ as a function of $\boldsymbol{y}$ for fixed $\boldsymbol{\theta}$, which describes the relative probabilities of different possible values of $\boldsymbol{y}$, given a particular set of parameters $\boldsymbol{\theta}$. However, in statistical inference, we have observed $y_1, \ldots, y_n$ (values of $Y_1, \ldots, Y_n$). Knowledge of the probability of alternative possible realisations of $\boldsymbol{Y}$ is largely irrelevant. What we want to know about is $\boldsymbol{\theta}$.

Our only link between the observed data $y_1, \ldots, y_n$ and $\boldsymbol{\theta}$ is through the function $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})$. Therefore, it seems sensible that parametric statistical inference should be based on this function. We can think of $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ for fixed $\boldsymbol{y}$, which describes the relative *likelihoods* of different possible (sets of) $\boldsymbol{\theta}$, given observed data $y_1, \ldots, y_n$. We write

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})$$

for this *likelihood*, which is a function of the unknown parameter $\boldsymbol{\theta}$. For convenience, we often drop $\boldsymbol{y}$ from the notation, and write $L(\boldsymbol{\theta})$.

The likelihood function is of central importance in parametric statistical inference. It provides a means for comparing different possible values of $\boldsymbol{\theta}$, based on the probabilities (or probability densities) that they assign to the observed data $y_1, \ldots, y_n$.

Notes

1. Frequently it is more convenient to consider the *log-likelihood* function $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$.
2. Nothing in the definition of the likelihood requires $y_1, \ldots, y_n$ to be observations of independent random variables, although we shall frequently make this assumption.

3. Any factors which depend on $y_1, \ldots, y_n$ alone (and not on $\boldsymbol{\theta}$) can be ignored when writing down the likelihood. Such factors give no information about the relative likelihoods of different possible values of $\boldsymbol{\theta}$.

Example 2.1 (Bernoulli)**.** $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, independent identically distributed (i.i.d.) Bernoulli(p) random variables. Here $\theta = (p)$ and the likelihood is

$$L(p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i}.$$

The log-likelihood is

$$\ell(p) = \log L(p) = n\bar{y}\log p + n(1-\bar{y})\log(1-p).$$

Example 2.2 (Normal)**.** $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. $N(\mu, \sigma^2)$ random variables. Here $\boldsymbol{\theta} = (\mu, \sigma^2)$ and the likelihood is

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i-\mu)^2\right) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2\right) \\ &\propto (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2\right). \end{aligned}$$

The log-likelihood is

$$\ell(\mu, \sigma^2) = \log L(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i-\mu)^2.$$

2.3 Maximum likelihood estimation

One of the primary tasks of parametric statistical inference is *estimation* of the unknown parameters $\theta_1, \ldots, \theta_p$. Consider the value of $\boldsymbol{\theta}$ which maximises the likelihood function. This is the 'most likely' value of $\boldsymbol{\theta}$, the one which makes the observed data 'most probable'. When we are searching for an estimate of $\boldsymbol{\theta}$, this would seem to be a good candidate.

We call the value of $\boldsymbol{\theta}$ which maximises the likelihood $L(\theta)$ the *maximum likelihood estimate* (MLE) of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$. $\hat{\boldsymbol{\theta}}$ depends on $\boldsymbol{y}$, as different observed data samples lead to different likelihood functions. The corresponding function of $\boldsymbol{Y}$ is called the *maximum likelihood estimator* and is also denoted by $\hat{\boldsymbol{\theta}}$.

Note that as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, the MLE for any component of $\boldsymbol{\theta}$ is given by the corresponding component of $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^T$. Similarly, the MLE for any function of parameters $g(\boldsymbol{\theta})$ is given by $g(\hat{\boldsymbol{\theta}})$.

As log is a strictly increasing function, the value of $\boldsymbol{\theta}$ which maximises $L(\boldsymbol{\theta})$ also maximises $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$. It is almost always easier to maximise $\ell(\boldsymbol{\theta})$. This is achieved in the usual way; finding a stationary point by differentiating $\ell(\boldsymbol{\theta})$ with respect to $\theta_1, \ldots, \theta_p$, and solving the resulting p simultaneous equations. It should also be checked that the stationary point is a maximum.

Example 2.3 (Bernoulli)**.** $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables. Here $\boldsymbol{\theta} = (p)$ and the log-likelihood is

$$\ell(p) = n\bar{y} \log p + n(1 - \bar{y}) \log(1 - p).$$

Differentiating with respect to p,

$$\frac{\partial}{\partial p}\ell(p) = \frac{n\bar{y}}{p} - \frac{n(1 - \bar{y})}{1 - p}$$

so the MLE $\hat{p}$ solves

$$\frac{n\bar{y}}{\hat{p}} - \frac{n(1 - \bar{y})}{1 - \hat{p}} = 0.$$

Solving this for $\hat{p}$ gives $\hat{p} = \bar{y}$. Note that

$$\frac{\partial^2}{\partial p^2}\ell(p) = -n\bar{y}/p^2 - n(1 - \bar{y})/(1 - p)^2 < 0$$

everywhere, so the stationary point is clearly a maximum.

Example 2.4 (Normal)**.** $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. $N(\mu, \sigma^2)$ random variables. Here $\boldsymbol{\theta} = (\mu, \sigma^2)$ and and the log-likelihood is

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum (y_i - \mu)^2.$$

Differentiating with respect to μ

$$\frac{\partial}{\partial \mu}\ell(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum(y_i - \mu) = \frac{n(\bar{y} - \mu)}{\sigma^2}$$

so $(\hat{\mu}, \hat{\sigma}^2)$ solve

$$\frac{n(\bar{y} - \hat{\mu})}{\hat{\sigma}^2} = 0. \tag{2.1}$$

Differentiating with respect to σ^2

$$\frac{\partial}{\partial \sigma^2}\ell(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum(y_i - \mu)^2,$$

so

$$-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum(y_i - \hat{\mu})^2 = 0 \tag{2.2}$$

Solving (2.1) and (2.2), we obtain $\hat{\mu} = \bar{y}$ and

$$\hat{\sigma}^2 = \frac{1}{n}\sum(y_i - \hat{\mu})^2 = \frac{1}{n}\sum(y_i - \bar{y})^2.$$

Strictly, to show that this stationary point is a maximum, we need to show that the Hessian matrix (the matrix of second derivatives with elements $[\boldsymbol{H}(\boldsymbol{\theta})]_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j}\ell(\theta)$) is negative definite at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, that is $\boldsymbol{a}^T\boldsymbol{H}(\hat{\boldsymbol{\theta}})\boldsymbol{a} < 0$ for every $\boldsymbol{a} \neq \boldsymbol{0}$. Here

$$\boldsymbol{H}(\hat{\mu}, \hat{\sigma}^2) = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2(\hat{\sigma}^2)^2} \end{pmatrix}$$

which is clearly negative definite.

2.4 Score

Let

$$u_i(\boldsymbol{\theta}) \equiv \frac{\partial}{\partial \theta_i}\ell(\theta), \quad i = 1, \ldots, p$$

and $\boldsymbol{u}(\boldsymbol{\theta}) \equiv [u_1(\boldsymbol{\theta}), \ldots, u_p(\boldsymbol{\theta})]^T$. Then we call $\boldsymbol{u}(\boldsymbol{\theta})$ the *vector of scores* or *score vector*. Where $p = 1$ and $\boldsymbol{\theta} = (\theta)$, the *score* is the scalar defined as

$$u(\theta) \equiv \frac{\partial}{\partial \theta} \ell(\theta).$$

The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ satisfies

$$u(\hat{\boldsymbol{\theta}}) = \mathbf{0},$$

that is,

$$u_i(\hat{\boldsymbol{\theta}}) = 0, \quad i = 1, \ldots, p.$$

Note that $u(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$ for fixed (observed) $\boldsymbol{y}$. However, if we replace $y_1, \ldots, y_n$ in $u(\boldsymbol{\theta})$, by the corresponding random variables $Y_1, \ldots, Y_n$ then we obtain a vector of random variables $U(\boldsymbol{\theta}) \equiv [U_1(\boldsymbol{\theta}), \ldots, U_p(\boldsymbol{\theta})]^T$.

An important result in likelihood theory is that the expected score at the true (but unknown) value of $\boldsymbol{\theta}$ is zero:

Theorem 2.1. *We have* $E[U(\boldsymbol{\theta})] = \mathbf{0}$, i.e. $E[U_i(\boldsymbol{\theta})] = 0$, $i = 1, \ldots, p$, *provided that*

1. *The expectation exists.*
2. *The sample space for* $\boldsymbol{Y}$ *does not depend on* $\boldsymbol{\theta}$.

Proof. Our proof is for continuous $\boldsymbol{y}$ – in the discrete case, replace $\int$ by $\sum$. For each $i = 1, \ldots, n$

$$\begin{aligned}
E[U_i(\boldsymbol{\theta})] &= \int U_i(\boldsymbol{\theta}) f_{\boldsymbol{Y}}(\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{y} \\
&= \int \frac{\partial}{\partial \theta_i} \ell(\theta) f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} \\
&= \int \frac{\partial}{\partial \theta_i} \log f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} \\
&= \int \frac{\frac{\partial}{\partial \theta_i} f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})}{f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})} f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} \\
&= \int \frac{\partial}{\partial \theta_i} f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} \\
&= \frac{\partial}{\partial \theta_i} \int f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{y} \\
&= \frac{\partial}{\partial \theta_i} 1 = 0,
\end{aligned}$$

as required. □

Example 2.5 (Bernoulli)**.** $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables. Here $\boldsymbol{\theta} = (p)$ and

$$u(p) = n\bar{y}/p - n(1-\bar{y})/(1-p).$$

Since $E[U(p)] = 0$, we must have $E[\bar{Y}] = p$ (which we already know is correct).

Example 2.6 (Normal)**.** $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. $N(\mu, \sigma^2)$ random variables. Here $\boldsymbol{\theta} = (\mu, \sigma^2)$ and

$$\begin{aligned}
u_1(\mu, \sigma^2) &= n(\bar{y} - \mu)/\sigma^2 \\
u_2(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (y_i - \mu)^2
\end{aligned}$$

Since $E[\boldsymbol{U}(\mu, \sigma^2)] = \mathbf{0}$, we must have $E[\bar{Y}] = \mu$ and $E[\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu)^2] = \sigma^2$.

2.5 Information

Suppose that $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, whose joint p.d.f. $L(\theta)$ is completely specified except for the values of p unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$. Previously, we defined the Hessian matrix $H(\boldsymbol{\theta})$ to be the matrix with components

$$[H(\boldsymbol{\theta})]_{ij} \equiv \frac{\partial^2}{\partial\theta_i \partial\theta_j}\ell(\theta) \qquad i = 1, \ldots, p;\ j = 1, \ldots, p.$$

We call the matrix $-H(\boldsymbol{\theta})$ the *observed information matrix*. Where $p = 1$ and $\boldsymbol{\theta} = (\theta)$, the *observed information* is a scalar defined as

$$-H(\theta) \equiv -\frac{\partial}{\partial\theta^2}\ell(\theta).$$

As with the score, if we replace $y_1, \ldots, y_n$ in $H(\boldsymbol{\theta})$, by the corresponding random variables $Y_1, \ldots, Y_n$, we obtain a matrix of random variables. Then, we define the *expected information matrix* or *Fisher information matrix*

$$[\mathcal{I}(\boldsymbol{\theta})]_{ij} = E(-[H(\boldsymbol{\theta})]_{ij}) \qquad i = 1, \ldots, p;\ j = 1, \ldots, p.$$

An important result in likelihood theory is that the variance-covariance matrix of the score vector is equal to the expected information matrix:

Theorem 2.2. *We have* $Var[U(\boldsymbol{\theta})] = \mathcal{I}(\boldsymbol{\theta})$, i.e.

$$Var[U(\boldsymbol{\theta})]_{ij} = [\mathcal{I}(\boldsymbol{\theta})]_{ij}, \quad i = 1, \ldots, p, \quad j = 1, \ldots, p$$

provided that

1. *The variance exists.*
2. *The sample space for* $\boldsymbol{Y}$ *does not depend on* $\boldsymbol{\theta}$.

Proof. Our proof is for continuous $\boldsymbol{y}$ – in the discrete case, replace $\int$ by $\sum$.

For each $i = 1, \ldots, p$ and $j = 1, \ldots, p$,

$$\begin{aligned}
\mathrm{Var}[U(\boldsymbol{\theta})]_{ij} &= E[U_i(\boldsymbol{\theta})U_j(\boldsymbol{\theta})] \\
&= \int \frac{\partial}{\partial\theta_i}\ell(\theta)\frac{\partial}{\partial\theta_j}\ell(\theta) f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\
&= \int \frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}\log f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta}) f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\
&= \int \frac{\frac{\partial}{\partial\theta_i}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}\frac{\frac{\partial}{\partial\theta_j}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})} f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\
&= \int \frac{1}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}\frac{\partial}{\partial\theta_i}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y}.
\end{aligned}$$

Now

$$\begin{aligned}
[\mathcal{I}(\boldsymbol{\theta})]_{ij} &= E\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta)\right] \\
&= \int -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta}) f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\
&= \int -\frac{\partial}{\partial\theta_i}\left[\frac{\frac{\partial}{\partial\theta_j}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}\right] f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\
&= \int \left[-\frac{\frac{\partial^2}{\partial\theta_i\partial\theta_j}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})} + \frac{\frac{\partial}{\partial\theta_i}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})^2}\right] f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\
&= -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\int f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} + \int \frac{1}{f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})}\frac{\partial}{\partial\theta_i}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})\frac{\partial}{\partial\theta_j}f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{y} \\
&= \mathrm{Var}[U(\boldsymbol{\theta})]_{ij},
\end{aligned}$$

as required. □

Example 2.7 (Bernoulli)**.** $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables. Here $\boldsymbol{\theta} = (p)$ and

$$
\begin{aligned}
u(p) &= \frac{n\bar{y}}{p} - \frac{n(1-\bar{y})}{(1-p)} \\
-H(p) &= \frac{n\bar{y}}{p^2} + \frac{n(1-\bar{y})}{(1-p)^2} \\
\mathcal{I}(p) &= \frac{n}{p} + \frac{n}{(1-p)} = \frac{n}{p(1-p)}.
\end{aligned}
$$

Example 2.8 (Normal)**.** $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. $N(\mu, \sigma^2)$ random variables. Here $\boldsymbol{\theta} = (\mu, \sigma^2)$ and

$$
\begin{aligned}
u_1(\mu, \sigma^2) &= \frac{n(\bar{y} - \mu)}{\sigma^2} \\
u_2(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum (y_i - \mu)^2.
\end{aligned}
$$

Therefore

$$
-\boldsymbol{H}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{n(\bar{y}-\mu)}{(\sigma^2)^2} \\ \frac{n(\bar{y}-\mu)}{(\sigma^2)^2} & \frac{1}{(\sigma^2)^3} \sum (y_i - \mu)^2 - \frac{n}{2(\sigma^2)^2} \end{pmatrix}
$$

$$
\mathcal{I}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2(\sigma^2)^2} \end{pmatrix}.
$$

2.6 Asymptotic distribution of the MLE

Maximum likelihood estimation is an attractive method of estimation for a number of reasons. It is intuitively sensible and usually reasonably straightforward to carry out. Even when the simultaneous equations we obtain by differentiating the log-likelihood function are impossible to solve directly, solution by numerical methods is usually feasible.

Perhaps the most compelling reason for considering maximum likelihood estimation is the asymptotic behaviour of maximum likelihood estimators.

Suppose that $y_1, \ldots, y_n$ are observations of independent random variables $Y_1, \ldots, Y_n$, whose joint p.d.f. $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_{Y_i}(y_i; \boldsymbol{\theta})$ is completely specified except for the values of an unknown parameter vector $\boldsymbol{\theta}$, and that $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$.

Then, as $n \to \infty$, the distribution of $\hat{\boldsymbol{\theta}}$ tends to a multivariate normal distribution with mean vector $\boldsymbol{\theta}$ and variance covariance matrix $\mathcal{I}(\boldsymbol{\theta})^{-1}$.

Where $p = 1$ and $\boldsymbol{\theta} = (\theta)$, the distribution of the MLE $\hat{\theta}$ tends to $N[\theta, 1/\mathcal{I}(\theta)]$.

For 'large enough n', we can treat the asymptotic distribution of the MLE as an approximation. The fact that $E(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{\theta}$ means that the maximum likelihood estimator is *approximately unbiased* for large samples. The variance of $\hat{\boldsymbol{\theta}}$ is approximately $\mathcal{I}(\boldsymbol{\theta})^{-1}$. It is possible to show that this is the smallest possible variance of any unbiased estimator of $\boldsymbol{\theta}$ (this result is called the Cramér–Rao lower bound, which we do not prove here). Therefore the MLE is the 'best possible' estimator in large samples (and therefore we hope also reasonable in small samples, though we should investigate this case by case).

2.7 Quantifying uncertainty in parameter estimates

The usefulness of an estimate is always enhanced if some kind of measure of its precision can also be provided. Usually, this will be a *standard error*, an estimate of the standard deviation of the associated estimator. For the maximum likelihood estimator $\hat{\theta}$, a standard error is given by

$$s.e.(\hat{\theta}) = \frac{1}{\mathcal{I}(\hat{\theta})^{\frac{1}{2}}},$$

and for a vector parameter $\boldsymbol{\theta}$

$$s.e.(\hat{\theta}_i) = [\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}]_{ii}^{\frac{1}{2}}, \quad i = 1, \ldots, p.$$

An alternative summary of the information provided by the observed data about the location of a parameter θ and the associated precision is a *confidence interval*.

The asymptotic distribution of the maximum likelihood estimator can be used to provide approximate large sample confidence intervals. Asymptotically, $\hat{\theta}_i$ has a $N(\theta_i, [\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii})$ distribution and we can find $z_{1-\frac{\alpha}{2}}$ such that

$$P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{\theta}_i - \theta_i}{[\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii}^{\frac{1}{2}}} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Therefore

$$P\left(\hat{\theta}_i - z_{1-\frac{\alpha}{2}}[\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii}^{\frac{1}{2}} \leq \theta_i \leq \hat{\theta}_i + z_{1-\frac{\alpha}{2}}[\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii}^{\frac{1}{2}}\right) = 1 - \alpha.$$

The endpoints of this interval cannot be evaluated because they also depend on the unknown parameter vector $\boldsymbol{\theta}$. However, if we replace $\mathcal{I}(\boldsymbol{\theta})$ by its MLE $\mathcal{I}(\hat{\boldsymbol{\theta}})$ we obtain the approximate large sample $100(1-\alpha)\%$ confidence interval

$$[\hat{\theta}_i - z_{1-\frac{\alpha}{2}}[\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}]_{ii}^{\frac{1}{2}}, \hat{\theta}_i + z_{1-\frac{\alpha}{2}}[\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}]_{ii}^{\frac{1}{2}}].$$

For $\alpha = 0.1, 0.05, 0.01$, $z_{1-\frac{\alpha}{2}} = 1.64, 1.96, 2.58$.

Example 2.9 (Bernoulli)**.** If $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables then asymptotically $\hat{p} = \bar{y}$ has a $N(p, p(1-p)/n)$ distribution, and a large sample 95% confidence interval for p is

$$\begin{aligned}
&[\hat{p} - 1.96[\mathcal{I}(\hat{p})^{-1}]^{\frac{1}{2}}, \hat{p} + 1.96[\mathcal{I}(\hat{p})^{-1}]^{\frac{1}{2}}] \\
&= [\hat{p} - 1.96[\hat{p}(1-\hat{p})/n]^{\frac{1}{2}}, \hat{p} + 1.96[\hat{p}(1-\hat{p})/n]^{\frac{1}{2}}] \\
&= [\bar{y} - 1.96[\bar{y}(1-\bar{y})/n]^{\frac{1}{2}}, \bar{y} + 1.96[\bar{y}(1-\bar{y})/n]^{\frac{1}{2}}].
\end{aligned}$$

Chapter 3

Comparing statistical models

3.1 Introduction

If we have a set of competing probability models which might have generated the observed data, we may want to determine which of the models is most appropriate. In practice, we proceed by comparing models pairwise. Suppose that we have two competing alternatives, $f_{\boldsymbol{Y}}^{(0)}$ (model M_0) and $f_{\boldsymbol{Y}}^{(1)}$ (model M_1) for $f_{\boldsymbol{Y}}$, the joint distribution of $Y_1, \ldots, Y_n$. Often H_0 and H_1 both take the same parametric form, $f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})$ but with $\boldsymbol{\theta} \in \Theta^{(0)}$ for H_0 and $\boldsymbol{\theta} \in \Theta^{(1)}$ for H_1, where $\Theta^{(0)}$ and $\Theta^{(1)}$ are alternative sets of possible values for $\boldsymbol{\theta}$. In the regression setting, we are often interested in determining which of a set of explanatory variables have an impact on the distribution of the response.

3.2 Hypothesis testing

A hypothesis test provides one mechanism for comparing two competing statistical models. A hypothesis test does not treat the two hypotheses (models) symmetrically. One hypothesis,

$$H_0\text{: the data were generated from model } M_0,$$

is accorded special status, and referred to as the *null hypothesis*. The null hypothesis is the reference model, and will be assumed to be appropriate

unless the observed data strongly indicate that H_0 is inappropriate, and that

$$H_1\text{: the data were generated from model } M_1,$$

(the *alternative* hypothesis) should be preferred. The fact that a hypothesis test does not reject H_0 should not be taken as evidence that H_0 is true and H_1 is not, or that H_0 is better supported by the data than H_1, merely that the data does not provide sufficient evidence to reject H_0 in favour of H_1.

A hypothesis test is defined by its *critical region* or *rejection region*, which we shall denote by C. C is a subset of $\mathbb{R}^n$ and is the set of possible $\boldsymbol{y}$ which would lead to rejection of H_0 in favour of H_1, *i.e.*

- If $\boldsymbol{y} \in C$, H_0 is rejected in favour of H_1;
- If $\boldsymbol{y} \notin C$, H_0 is not rejected.

As $\boldsymbol{Y}$ is a random variable, there remains the possibility that a hypothesis test will produce an erroneous result. We define the *size* (or *significance level*) of the test

$$\alpha = \max_{\boldsymbol{\theta} \in \Theta^{(0)}} P(\boldsymbol{Y} \in C; \boldsymbol{\theta})$$

This is the maximum probability of erroneously rejecting H_0, over all possible distributions for $\boldsymbol{Y}$ implied by H_0. We also define the power function

$$\omega(\boldsymbol{\theta}) = P(\boldsymbol{Y} \in C; \boldsymbol{\theta})$$

It represents the probability of rejecting H_0 for a particular value of $\boldsymbol{\theta}$. Clearly we would like to find a test with where $\omega(\boldsymbol{\theta})$ is large for every $\boldsymbol{\theta} \in \Theta^{(1)} \setminus \Theta^{(0)}$, while at the same time avoiding erroneous rejection of H_0. In other words, a good test will have small size, but large power.

The general hypothesis testing procedure is to fix α to be some small value (often 0.05), so that the probability of erroneous rejection of H_0 is limited. In doing this, we are giving H_0 precedence over H_1. Given our specified α, we try to choose a test, defined by its rejection region C, to make $\omega(\boldsymbol{\theta})$ as large as possible for $\boldsymbol{\theta} \in \Theta^{(1)} \setminus \Theta^{(0)}$.

3.3 Likelihood ratio tests for nested hypotheses

Suppose that H_0 and H_1 both take the same parametric form, $f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta^{(0)}$ for H_0 and $\boldsymbol{\theta} \in \Theta^{(1)}$ for H_1, where $\Theta^{(0)}$ and $\Theta^{(1)}$ are alternative sets of possible values for $\boldsymbol{\theta}$. A *likelihood ratio test* of H_0 against H_1 has a critical region of the form

$$C = \left\{ \boldsymbol{y} : \frac{\max_{\boldsymbol{\theta}\in\Theta^{(1)}} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\in\Theta^{(0)}} L(\boldsymbol{\theta})} > k \right\} \tag{3.1}$$

where k is determined by α, the size of the test, so

$$\max_{\boldsymbol{\theta}\in\Theta^{(0)}} P(\boldsymbol{y} \in C; \boldsymbol{\theta}) = \alpha.$$

Therefore, we will only reject H_0 if H_1 offers a distribution for $Y_1, \ldots, Y_n$ which makes the observed data much more probable than any distribution under H_0. This is intuitively appealing and tends to produce good tests (large power) across a wide range of examples.

In order to determine k in (3.1), we need to know the distribution of the likelihood ratio, or an equivalent statistic, under H_0. In general, this will not be available to us. However, we can make use of an important asymptotic result.

First we notice that, as log is a strictly increasing function, the rejection region is equivalent to

$$C = \left\{ \boldsymbol{y} : 2\log\left(\frac{\max_{\boldsymbol{\theta}\in\Theta^{(1)}} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\in\Theta^{(0)}} L(\boldsymbol{\theta})}\right) > k' \right\}$$

where

$$\max_{\boldsymbol{\theta}\in\Theta^{(0)}} P(\boldsymbol{y} \in C; \boldsymbol{\theta}) = \alpha.$$

Write

$$L_{01} \equiv 2\log\left(\frac{\max_{\boldsymbol{\theta}\in\Theta^{(1)}} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\in\Theta^{(0)}} L(\boldsymbol{\theta})}\right)$$

for the *log-likelihood ratio* test statistic. Provided that H_0 is *nested within* H_1, the following result provides a useful large-n approximation to the distribution of L_{01}.

Theorem 3.1. *Suppose that* H_0*:* $\boldsymbol{\theta} \in \Theta^{(0)}$ *and* H_1*:* $\boldsymbol{\theta} \in \Theta^{(1)}$*, where* $\Theta^{(0)} \subset \Theta^{(1)}$*. Let* $d_0 = \dim(\Theta^{(0)})$ *and* $d_1 = \dim(\Theta^{(1)})$*. Under* H_0*, the distribution of* L_{01} *tends towards* $\chi^2_{d_1-d_0}$ *as* $n \to \infty$*.*

Proof. First we note that in the case where $\boldsymbol{\theta}$ is one-dimensional and $\boldsymbol{\theta} = (\theta)$, a Taylor series expansion of $\ell(\theta)$ around the MLE $\hat{\theta}$ gives

$$\ell(\theta) = \ell(\hat{\theta}) + (\theta - \hat{\theta})U(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 U'(\hat{\theta}) + \ \ldots$$

Now, $U(\hat{\theta}) = 0$, and if we approximate $U'(\hat{\theta}) \equiv H(\hat{\theta})$ by $E[H(\theta)] \equiv -\mathcal{I}(\theta)$, and also ignore higher order terms, we obtain

$$2[\ell(\hat{\theta}) - \ell(\theta)] = (\theta - \hat{\theta})^2 \mathcal{I}(\theta)$$

As $\hat{\theta}$ is asymptotically $N[\theta, \mathcal{I}(\theta)^{-1}]$, $(\theta - \hat{\theta})^2\mathcal{I}(\theta)$ is asymptotically χ^2_1, and hence so is $2[\ell(\hat{\theta}) - \ell(\theta)]$.

Similarly it can be shown that when $\boldsymbol{\theta} \in \Theta$, a multidimensional space, $2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta})]$ is asymptotically χ^2_p, where p is the dimension of Θ.

Now, suppose that H_0 is true and $\boldsymbol{\theta} \in \Theta^{(0)}$ and therefore $\boldsymbol{\theta} \in \Theta^{(1)}$. Furthermore, suppose that $\ell(\boldsymbol{\theta})$ is maximised in $\Theta^{(0)}$ by $\hat{\boldsymbol{\theta}}^{(0)}$ and is maximised in $\Theta^{(1)}$ by $\hat{\boldsymbol{\theta}}^{(1)}$. Then

$$\begin{aligned} L_{01} &\equiv 2\log\left(\frac{\max_{\boldsymbol{\theta}\in\Theta^{(1)}} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\in\Theta^{(0)}} L(\boldsymbol{\theta})}\right) \\ &= 2\log L(\hat{\boldsymbol{\theta}}^{(1)}) - 2\log L(\hat{\boldsymbol{\theta}}^{(0)}) \\ &= 2[\log L(\hat{\boldsymbol{\theta}}^{(1)}) - \log L(\boldsymbol{\theta})] - 2[\log L(\hat{\boldsymbol{\theta}}^{(0)}) - \log L(\boldsymbol{\theta})] \\ &= L_1 - L_0. \end{aligned}$$

Therefore $L_1 = L_{01} + L_0$ and we know that, under H_0, L_1 has a $\chi^2_{d_1}$ distribution and L_0 has a $\chi^2_{d_0}$ distribution. Furthermore, it is possible to show (although we will not do so here) that under H_0, L_{01} and L_0 are independent. It can also be shown that under H_0 the difference $L_1 - L_0$ can be expressed as a quadratic form of normal random variables. Therefore, it follows that under H_0, the log likelihood ratio statistic L_{01} has a $\chi^2_{d_1-d_0}$ distribution. □

Example 3.1 (Bernoulli)**.** $y_1, \ldots, y_n$ are observations of $Y_1, \ldots, Y_n$, i.i.d. Bernoulli(p) random variables. Suppose that we require a size α test of the hypothesis H_0: $p = p_0$ against the general alternative H_1: 'p is unrestricted' where α and p_0 are specified.

Here $\boldsymbol{\theta} = (p)$, $\Theta^{(0)} = \{p_0\}$ and $\Theta^{(1)} = (0, 1)$ and the log likelihood ratio statistic is

$$L_{01} = 2n\bar{y} \log\left(\frac{\bar{y}}{p_0}\right) + 2n(1 - \bar{y}) \log\left(\frac{1 - \bar{y}}{1 - p_0}\right).$$

As $d_1 = 1$ and $d_0 = 0$, under H_0, the log likelihood ratio statistic has an asymptotic χ^2_1 distribution. For a log likelihood ratio test, we only reject H_0 in favour of H_1 when the test statistic is too large (observed data are much more probable under model H_1 than under model H_0), so in this case we reject H_0 when the observed value of the test statistic above is 'too large' to have come from a χ^2_1 distribution. What we mean by 'too large' depends on the significance level α of the test. For example, if $\alpha = 0.05$, a common choice, then we should reject H_0 if the test statistic is greater than the 3.84, the 95% point of the χ^2_1 distribution.

3.4 Information criteria for model comparison

It is more difficult to use the likelihood ratio test of Section 3.3 to compare two models if those models are not nested. An alternative approach is to record some criterion measuring the quality of the model for each of a candidate set of models, then choose the model which is the best according to this criterion.

When we were estimating the unknown parameters θ of a model, we chose the value which maximised the likelihood: that is, the value of θ that maximises the probability of observing the data we actually saw. It is tempting to use a similar system for choosing between two models, and to choose the model which has the greater likelihood, under which the probability of seeing the data we actually observed is maximised. However, if we do this we will always end up choosing complicated models, which fit the observed data very closely, but do not meet our requirement of parsimony.

For a given model depending on parameters $\theta \in \mathbb{R}^p$, let $\hat{\ell} = \ell(\hat{\theta})$ be the log-likelihood function for that model evaluated at the MLE $\hat{\theta}$. It is not sensible to choose between models by maximising $\hat{\ell}$ directly, and instead it is common to choose a model to maximise a criteria of the form

$$\hat{\ell} - \text{penalty},$$

where the penalty term will be large for complex models, and small for simple models.

Equivalently, we may choose between models by minimising a criteria of the form

$$-2\hat{\ell} + \text{penalty}.$$

By convention, many commonly-used criteria for model comparison take this form. For instance, the Akaike information criterion (AIC) is

$$\text{AIC} = -2\hat{\ell} + 2p,$$

where p is the dimension of the unknown parameter in the candidate model, and the Bayesian information criterion (BIC) is

$$\text{BIC} = -2\hat{\ell} + \log(n)p,$$

where n is the number of observations.

Chapter 4

Linear Models

4.1 The linear model

In practical applications, we often distinguish between a *response* variable and a group of *explanatory* variables. The aim is to determine the pattern of dependence of the response variable on the explanatory variables. We denote the n observations of the response variable by $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$. These are assumed to be observations of *random variables* $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$. Associated with each y_i is a vector $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ of values of p explanatory variables.

In a linear model, we assume that

$$\begin{aligned} Y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \\ &= \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i \\ &= \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i \\ &= [\boldsymbol{X}\boldsymbol{\beta}]_i + \epsilon_i, \qquad i = 1, \ldots, n \end{aligned} \tag{4.1}$$

where $\epsilon_i \sim N(0, \sigma^2)$ independently,

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of fixed but unknown parameters describing the dependence of Y_i on $\boldsymbol{x}_i$. The four ways of describing the linear model in (4.1) are equivalent, but the most economical is the matrix form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{4.2}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$.

The $n \times p$ matrix $\boldsymbol{X}$ consists of known (observed) constants and is called the *design matrix*. The ith row of $\boldsymbol{X}$ is $\boldsymbol{x}_i^T$, the explanatory data corresponding to the ith observation of the response. The jth column of $\boldsymbol{X}$ contains the n observations of the jth explanatory variable.

The error vector $\boldsymbol{\epsilon}$ has a multivariate normal distribution with mean vector $\boldsymbol{0}$ and variance covariance matrix $\sigma^2\boldsymbol{I}$, since $\mathrm{Var}(\epsilon_i) = \sigma^2$, and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, as $\epsilon_1, \ldots, \epsilon_n$ are independent of one another. It follows from (4.2) that the distribution of $\boldsymbol{Y}$ is multivariate normal with mean vector $\boldsymbol{X}\boldsymbol{\beta}$ and variance covariance matrix $\sigma^2\boldsymbol{I}$, *i.e.* $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$.

4.2 Examples of linear model structure

Example 4.1 (The null model)**.** If we do not include any variables x_i in the model, we have

$$Y_i = \beta_0 + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad i = 1, \ldots, n,$$

so

$$\boldsymbol{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = (\beta_0).$$

This is one (dummy) explanatory variable. In practice, this variable is present in all models.

Example 4.2 (Simple linear regression)**.** If we include a single variable x_i in the model, we might have

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2) \qquad i = 1, \ldots, n$$

so

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

There are two explanatory variables: the dummy variable and one 'real' variable.

Example 4.3 (Polynomial regression)**.** If we want to allow for a non-linear impact of x_i on the mean of Y_i, we might model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_{p-1} x_i^{p-1} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad i = 1, \ldots, n,$$

so

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{p-1} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}.$$

There are p explanatory variables: the dummy variable and one 'real' variable, transformed to $p-1$ variables.

Example 4.4 (Multiple regression)**.** To include multiple explanatory variables, we might model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i\,p-1} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad i = 1, \ldots, n,$$

so

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1\,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2\,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n\,p-1} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}.$$

There are p explanatory variables: the dummy variable and $p-1$ 'real' variables.

Example 4.5 (One categorical explanatory variable)**.** Suppose x_i is a categorical variable, taking values in a set of k possible categories. For simplicity of notation, we will give each category a number, and write $x_i \in \{1, \ldots, k\}$. We wish to model

$$Y_i = \mu_{x_i} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad i = 1, \ldots, n,$$

so that the mean of Y_i is the same for all observations in the same category, but differs for different categories.

We could rewrite this model to include an intercept, as

$$Y_i = \beta_0 + \beta_{x_i} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad i = 1, \ldots, n,$$

so that $\mu_j = \beta_0 + \beta_j$, for $j = 1, \ldots, k$. It is not possible to estimate all of the $\boldsymbol{\beta}$ parameters separately, as they only affect the distribution through the combination $\beta_0 + \beta_j$. Instead, we choose a **reference category** l, and set $\beta_l = 0$. The intercept term β_0 then gives the mean for the reference category, with β_j giving the difference in mean between category j and the reference category. In `R`, categorical variables are called `factor`s, and by default the reference category will be the first category when the names of the categories (the `levels` of the `factor`) are sorted alphabetically.

We can rewrite the model as a form of multiple regression by first defining a new explanatory variable $\boldsymbol{z}_i$

$$\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ik})^T,$$

where

$$z_{ij} = \begin{cases} 1 & \text{if } x_i = j \\ 0 & \text{otherwise.} \end{cases}$$

$\boldsymbol{z}_i$ is sometimes called the **one-hot encoding** of x_i, as it contains precisely one 1 (corresponding to the category x_i), and is 0 everywhere else. We then have

$$Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_k z_{ik} + \epsilon_i,$$

so

$$\boldsymbol{X} = \begin{pmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1k} \\ 1 & z_{21} & z_{22} & \cdots & z_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & z_{n2} & \cdots & z_{nk} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},$$

where each row of $\boldsymbol{X}$ will have two ones, and the remaining entries will be zero.

Example 4.6 (Two categorical explanatory variables)**.** Suppose we have two categorical variables $x_{i1} \in \{1, \ldots, k_1\}$ and $x_{i2} \in \{1, \ldots, k_2\}$. We might consider a model

$$Y_i = \beta_0 + \beta^{(1)}_{x_{i1}} + \beta^{(2)}_{x_{i2}} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad i = 1, \ldots, n,$$

where

$$\boldsymbol{\beta} = \left(\beta_0, \beta^{(1)}_1, \ldots, \beta^{(1)}_{k_1}, \beta^{(2)}_1, \ldots, \beta^{(2)}_{k_2}\right)^T,$$

where as in Example 4.5 we choose reference categories l_1 and l_2 for each categorical variable, and set $\beta^{(1)}_{l_1} = \beta^{(2)}_{l_2} = 0$. The terms $\beta^{(1)}_j$ are called the **main effects** for the categorical variables x_{i1}, and $\beta^{(2)}_j$ are the main effects for x_{i2}.

We might also want to allow an **interaction** between x_{i1} and x_{i2}, letting

$$Y_i = \beta_0 + \beta^{(1)}_{x_{i1}} + \beta^{(2)}_{x_{i2}} + \beta^{(1,2)}_{x_{i1},x_{i2}} + \epsilon_i,$$

where

$$\boldsymbol{\beta} = \left(\beta_0, \beta^{(1)}_1, \ldots, \beta^{(1)}_{k_1}, \beta^{(2)}_1, \ldots, \beta^{(2)}_{k_2}, \beta^{(1,2)}_{11}, \beta^{(1,2)}_{1,2}, \ldots, \beta^{(1,2)}_{k_1,k_2}\right)^T.$$

The terms $\beta^{(1,2)}_{j_1,j_2}$ are called the **interaction effects**. This model is equivalent to

$$Y_i = \mu_{x_{i1},x_{i2}} + \epsilon_i,$$

allowing a different mean for each possible combination of categories. To allow us to estimate the parameters, given reference categories l_1 and l_2, we set

$$\beta^{(1)}_{l_1} = \beta^{(2)}_{l_2} = 0; \qquad \beta^{(1,2)}_{l_1,j} = 0, \quad j = 1, \ldots, k_2; \qquad \beta^{(1,2)}_{j,l_2} = 0, \quad j = 1, \ldots, k_1.$$

As in Example 4.5, it is possible to rewrite the model with a design matrix $\boldsymbol{X}$, by using one-hot encoding of x_{i1} and x_{i2}.

4.3 Maximum likelihood estimation

The regression coefficients $\beta_1, \ldots, \beta_p$ describe the pattern by which the response depends on the explanatory variables. We use the observed data $y_1, \ldots, y_n$ to *estimate* this pattern of dependence.

The likelihood for a linear model is

$$L(\boldsymbol{\beta}, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right). \tag{4.3}$$

This is maximised with respect to $(\boldsymbol{\beta}, \sigma^2)$ at

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)^2.$$

The corresponding fitted values are

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

or

$$\hat{y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}, \quad i = 1, \ldots, n.$$

The residuals $\boldsymbol{r} = (r_1, \ldots, r_n)$ are $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ or $r_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ for $i = 1, \ldots, n$.. These residuals describe the variability in the observed responses $y_1, \ldots, y_n$ which has not been explained by the linear model. We call

$$D = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)^2$$

the *residual sum of squares* or *deviance* for the linear model.

4.4 Properties of the MLE

As $\boldsymbol{Y}$ is normally distributed, and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ is a linear function of $\boldsymbol{Y}$, then $\hat{\boldsymbol{\beta}}$ must also be normally distributed. We have $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, so

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}).$$

It is possible to prove (although we shall not do so here) that

$$\frac{D}{\sigma^2} \sim \chi^2_{n-p}$$

which implies that

$$E(\hat{\sigma}^2) = \frac{n-p}{n}\sigma^2,$$

so the maximum likelihood estimator is biased for σ^2 (although still asymptotically unbiased as $\frac{n-p}{n} \to 1$ as $n \to \infty$). We often use the unbiased estimator of σ^2

$$\tilde{\sigma}^2 = \frac{D}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2.$$

The denominator $n - p$, the number of observations minus the number of linear coefficients in the model is called the *degrees of freedom* of the model. Therefore, we estimate the residual variance by the deviance divided by the degrees of freedom.

4.5 Comparing linear models

If we have a set of competing linear models which might explain the dependence of the response on the explanatory variables, we will want to determine which of the models is most appropriate.

As described previously, we proceed by comparing models pairwise using a likelihood ratio test. For linear models this kind of comparison is restricted to situations where one of the models, H_0, is *nested* in the other, H_1. This usually means that the explanatory variables present in H_0 are a subset of those present in H_1. In this case model H_0 is a special case of model H_1,

where certain coefficients are set equal to zero. We let $\boldsymbol{\theta}$ represent the collection of linear parameters for model H_1, together with the residual variance σ^2, and let $\Theta^{(1)}$ be the unrestricted parameter space for $\boldsymbol{\theta}$. Then $\Theta^{(0)}$ is the parameter space corresponding to model H_0, *i.e.* with the appropriate coefficients constrained to zero.

We will assume that model H_1 contains p linear parameters and model H_0 a subset of $q < p$ of these. Without loss of generality, we can think of H_1 as the model

$$Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \quad i = 1, \ldots, n$$

and H_0 being the same model with

$$\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0.$$

Now, a likelihood ratio test of H_0 against H_1 has a critical region of the form

$$C = \left\{ \boldsymbol{y} : \frac{\max_{(\boldsymbol{\beta},\sigma^2)\in\Theta^{(1)}} L(\boldsymbol{\beta},\sigma^2)}{\max_{(\boldsymbol{\beta},\sigma^2)\in\Theta^{(0)}} L(\boldsymbol{\beta},\sigma^2)} > k \right\}$$

where k is determined by α, the size of the test, so

$$\max_{\boldsymbol{\theta}\in\Theta^{(0)}} P(\boldsymbol{y} \in C; \boldsymbol{\beta}, \sigma^2) = \alpha.$$

For a linear model,

$$L(\boldsymbol{\beta},\sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right).$$

This is maximised with respect to $(\boldsymbol{\beta},\sigma^2)$ at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\sigma^2 = \hat{\sigma}^2 = D/n$. Therefore

$$\begin{aligned}\max_{\boldsymbol{\beta},\sigma^2} L(\boldsymbol{\beta},\sigma^2) &= (2\pi D/n)^{-\frac{n}{2}} \exp\left(-\frac{n}{2D}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2\right) \\ &= (2\pi D/n)^{-\frac{n}{2}} \exp\left(-\frac{n}{2}\right).\end{aligned}$$

This form applies for both $\boldsymbol{\theta} \in \Theta^{(0)}$ and $\boldsymbol{\theta} \in \Theta^{(1)}$, with only the model changing. Let the deviances under models H_0 and H_1 be denoted by D_0 and D_1 respectively. Then the critical region for the likelihood ratio test is of the form

$$\frac{(2\pi D_1/n)^{-\frac{n}{2}}}{(2\pi D_0/n)^{-\frac{n}{2}}} > k$$

so

$$\left(\frac{D_0}{D_1}\right)^{\frac{n}{2}} > k,$$

and

$$\left(\frac{D_0}{D_1} - 1\right)\frac{n-p}{p-q} > k'$$

for some k'. Rearranging,

$$\frac{(D_0 - D_1)/(p-q)}{D_1/(n-p)} > k'.$$

We refer to the left hand side of this inequality as the F-statistic. We reject the simpler model H_0 in favour of the more complex model H_1 if F is 'too large'.

As we have required H_0 to be nested in H_1, $F \sim F_{p-q,\,n-p}$ when H_0 is true. To see this, note that

$$\frac{D_0}{\sigma^2} = \frac{D_0 - D_1}{\sigma^2} + \frac{D_1}{\sigma^2}.$$

Furthermore, under H_0, $D_1/\sigma^2 \sim \chi^2_{n-p}$ and $D_0/\sigma^2 \sim \chi^2_{n-q}$. It is possible to show (although we will not do so here) that under H_0, $(D_0 - D_1)/\sigma^2$ and D_0/σ^2 are independent. Therefore, from the properties of the chi-squared distribution, it follows that under H_0, $(D_0 - D_1)/\sigma^2 \sim \chi^2_{p-q}$, and $F \sim F_{p-q,\,n-p}$ distribution.

Therefore, the precise critical region can be evaluated given the size, α, of the test. We reject H_0 in favour of H_1 when

$$\frac{(D_0 - D_1)/(p-q)}{D_1/(n-p)} > k$$

where k is the $100(1-\alpha)\%$ point of the $F_{p-q,\,n-p}$ distribution.

Chapter 5

Generalised Linear Models

5.1 Regression models for non-normal data

The linear model of Chapter 4 assumes each response $Y_i \sim N(\mu_i, \sigma^2)$, where the mean μ_i depends on explanatory variables through $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. For many types of data, this assumption of normality of the response may not be justified. For instance, we might have

- a binary response ($Y_i \in \{0, 1\}$), for instance representing whether or not a patient recovers from a disease. A natural model is that $Y_i \sim$ Bernoulli(p_i), and we might want to model how the 'success' probability p_i depends on explanatory variables $\boldsymbol{x}_i$.
- a count response ($Y_i \in \{0, 1, 2, 3, \ldots\}$), for instance representing the number of customers arrive at a shop. A natural model is that $Y_i \sim$ Poisson(λ_i), and we might want to model how the rate λ_i depends on explanatory variables.

In Section 5.2, we define the exponential family, which includes the Bernoulli and Poisson distributions as special cases. In a generalised linear model, the response distribution is assumed to be a member of the exponential family.

To complete the specification of a generalised linear model, we will need to model how the parameters of the response distribution (e.g. the success probability p_i or the rate λ_i) depend on explanatory variables $\boldsymbol{x}_i$. We need to do this in a way which respects constraints on the possible values which

these parameters may take: for instance, we **should not** model $p_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ directly, as we need to enforce $p_i \in [0, 1]$.

5.2 The exponential family

A probability distribution is said to be a member of the exponential family if its probability density function (or probability function, if discrete) can be written in the form

$$f_Y(y; \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right). \tag{5.1}$$

The parameter θ is called the *natural* or *canonical* parameter. The parameter ϕ is usually assumed known. If it is unknown then it is often called the *nuisance* parameter.

The density (5.1) can be thought of as a likelihood resulting from a single observation y. Then the log-likelihood is

$$\ell(\theta, \phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

and the score is

$$u(\theta) = \frac{\partial}{\partial\theta}\ell(\theta, \phi) = \frac{y - \frac{\partial}{\partial\theta}b(\theta)}{a(\phi)} = \frac{y - b'(\theta)}{a(\phi)}.$$

The Hessian is

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

so the expected information is

$$\mathcal{I}(\theta) = E[-H(\theta)] = \frac{b''(\theta)}{a(\phi)}.$$

From the properties of the score function in Section 2.4, we know that $E[U(\theta)] = 0$. Therefore

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

so $E[Y] = b'(\theta)$. We often denote the mean by μ, so $\mu = b'(\theta)$.

Furthermore,

$$\text{Var}[U(\theta)] = \text{Var}\left[\frac{Y - b'(\theta)}{a(\phi)}\right] = \frac{\text{Var}[Y]}{a(\phi)^2},$$

as $b'(\theta)$ and $a(\phi)$ are constants (not random variables). We also know from Section 2.5 that $\text{Var}[U(\theta)] = \mathcal{I}(\theta)$. Therefore

$$\text{Var}[Y] = a(\phi)^2\text{Var}[U(\theta)] = a(\phi)^2\mathcal{I}(\theta) = a(\phi)b''(\theta).$$

and hence the mean and variance of a random variable with probability density function (or probability function) of the form (5.1) are $b'(\theta)$ and $a(\phi)b''(\theta)$ respectively.

The variance is the product of two functions; $b''(\theta)$ depends on the canonical parameter θ (and hence μ) only and is called the *variance function* ($V(\mu) \equiv b''(\theta)$); $a(\phi)$ is sometimes of the form $a(\phi) = \sigma^2/w$ where w is a known *weight* and σ^2 is called the *dispersion parameter* or *scale parameter*.

Example 5.1 (Normal distribution)**.** Suppose $Y \sim N(\mu, \sigma^2)$. Then

$$\begin{aligned} f_Y(y; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \qquad y \in \mathbb{R};\ \ \mu \in \mathbb{R} \\ &= \exp\left(\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]\right). \end{aligned}$$

This is in the form (5.1), with $\theta = \mu$, $b(\theta) = \frac{1}{2}\theta^2$, $a(\phi) = \sigma^2$ and

$$c(y, \phi) = -\frac{1}{2}\left[\frac{y^2}{a(\phi)} + \log(2\pi a[\phi])\right].$$

Therefore

$$E(Y) = b'(\theta) = \theta = \mu,$$

$$\text{Var}(Y) = a(\phi)b''(\theta) = \sigma^2$$

and the variance function is

$$V(\mu) = 1.$$

Example 5.2 (Poisson distribution)**.** Suppose $Y \sim \text{Poisson}(\lambda)$. Then

$$\begin{aligned} f_Y(y;\lambda) &= \frac{\exp(-\lambda)\lambda^y}{y!} \qquad y \in \{0,1,\ldots\}; \quad \lambda \in \mathcal{R}_+ \\ &= \exp\left(y\log\lambda - \lambda - \log y!\right). \end{aligned}$$

This is in the form (5.1), with $\theta = \log\lambda$, $b(\theta) = \exp\theta$, $a(\phi) = 1$ and $c(y,\phi) = -\log y!$. Therefore

$$E(Y) = b'(\theta) = \exp\theta = \lambda,$$

$$\text{Var}(Y) = a(\phi)b''(\theta) = \exp\theta = \lambda$$

and the variance function is

$$V(\mu) = \mu.$$

Example 5.3 (Bernoulli distribution)**.** Suppose $Y \sim \text{Bernoulli}(p)$. Then

$$\begin{aligned} f_Y(y;p) &= p^y(1-p)^{1-y} \qquad y \in \{0,1\}; \quad p \in (0,1) \\ &= \exp\left(y\log\frac{p}{1-p} + \log(1-p)\right) \end{aligned}$$

This is in the form (5.1), with $\theta = \log\frac{p}{1-p}$, $b(\theta) = \log(1+\exp\theta)$, $a(\phi) = 1$ and $c(y,\phi) = 0$. Therefore

$$E(Y) = b'(\theta) = \frac{\exp\theta}{1+\exp\theta} = p,$$

$$\text{Var}(Y) = a(\phi)b''(\theta) = \frac{\exp\theta}{(1+\exp\theta)^2} = p(1-p)$$

and the variance function is

$$V(\mu) = \mu(1-\mu).$$

Example 5.4 (Binomial distribution)**.** Suppose $Y^* \sim \text{Binomial}(n,p)$. Here, n is assumed known (as usual) and the random variable $Y = Y^*/n$ is taken as the *proportion* of successes, so

$$f_Y(y;p) = \binom{n}{ny} p^{ny}(1-p)^{n(1-y)} \qquad y \in \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\right\}; \quad p \in (0,1)$$
$$= \exp\left(\frac{y \log \frac{p}{1-p} + \log(1-p)}{\frac{1}{n}} + \log\binom{n}{ny}\right).$$

This is in the form (5.1), with $\theta = \log \frac{p}{1-p}$, $b(\theta) = \log(1+\exp\theta)$, $a(\phi) = \frac{1}{n}$ and $c(y,\phi) = \log\binom{n}{ny}$. Therefore

$$E(Y) = b'(\theta) = \frac{\exp\theta}{1+\exp\theta} = p,$$

$$\mathrm{Var}(Y) = a(\phi)b''(\theta) = \frac{1}{n}\frac{\exp\theta}{(1+\exp\theta)^2} = \frac{p(1-p)}{n}$$

and the variance function is

$$V(\mu) = \mu(1-\mu).$$

Here, we can write $a(\phi) \equiv \sigma^2/w$ where the scale parameter $\sigma^2 = 1$ and the weight w is n, the binomial denominator.

5.3 Components of a generalised linear model

5.3.1 The random component

As in a linear model, the aim is to determine the pattern of dependence of a response variable on explanatory variables. We denote the n observations of the response by $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$. In a generalised linear model (GLM), these are assumed to be observations of *independent* random variables $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$, which take the same distribution from the exponential family. In other words, the functions a, b and c and usually the scale parameter ϕ are the same for all observations, but the canonical parameter θ may differ. Therefore, we write

$$f_{Y_i}(y_i;\theta_i,\phi_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i,\phi_i)\right)$$

and the joint density for $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ is

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\phi}) &= \prod_{i=1}^{n} f_{Y_i}(y_i; \theta_i, \phi_i) \\ &= \exp\left(\sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i, \phi_i) \right) \end{aligned} \tag{5.2}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ is the collection of canonical parameters and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$ is the collection of nuisance parameters (where they exist).

Note that for a particular sample of observed responses, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$, (5.2) is the likelihood function $L(\boldsymbol{\theta}, \boldsymbol{\phi})$ for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

5.3.2 The systematic (or structural) component

Associated with each y_i is a vector $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ of p explanatory variables. In a generalised linear model, the distribution of the response variable Y_i depends on $\boldsymbol{x}_i$ through the *linear predictor* η_i where

$$\begin{aligned} \eta_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \\ &= \sum_{j=1}^{p} x_{ij} \beta_j \\ &= \boldsymbol{x}_i^T \boldsymbol{\beta} \\ &= [\boldsymbol{X}\boldsymbol{\beta}]_i, \qquad i = 1, \ldots, n, \end{aligned} \tag{5.3}$$

where, as with a linear model,

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of fixed but unknown parameters describing the dependence of Y_i on $\boldsymbol{x}_i$. The four ways of describing the linear predictor in (5.3) are equivalent, but the most economical is the matrix form

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}. \tag{5.4}$$

Again, we call the $n \times p$ matrix $\boldsymbol{X}$ the *design matrix*. The ith row of $\boldsymbol{X}$ is $\boldsymbol{x}_i^T$, the explanatory data corresponding to the ith observation of the response. The jth column of $\boldsymbol{X}$ contains the n observations of the jth explanatory variable.

As for the linear model in Section 4.2, this structure allows quite general dependence of the linear predictor on explanatory variables. For instance, we can allow non-linear dependence of η_i on a variable x_i through polynomial regression (as in Example 4.3), or include categorical explanatory variables (as in Examples 4.5 and 4.6).

5.3.3 The link function

For specifying the pattern of dependence of the response variable on the explanatory variables, the canonical parameters $\theta_1, \ldots, \theta_n$ in (5.2) are not of direct interest. Furthermore, we have already specified that the distribution of Y_i should depend on $\boldsymbol{x}_i$ through the linear predictor η_i. It is the parameters $\beta_1, \ldots, \beta_p$ of the linear predictor which are of primary interest.

The link between the distribution of $\boldsymbol{Y}$ and the linear predictor $\boldsymbol{\eta}$ is provided by the *link function* g,

$$\eta_i = g(\mu_i), \quad i = 1, \ldots, n,$$

where $\mu_i \equiv E(Y_i),\ i = 1, \ldots, n$. Hence, the dependence of the distribution of the response on the explanatory variables is established as

$$g(E[Y_i]) = g(\mu_i) = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}, \quad i = 1, \ldots, n.$$

In principle, the link function g can be any one-to-one differentiable function. However, we note that η_i can in principle take any value in $\mathbb{R}$ (as we make no restriction on possible values taken by explanatory variables or model parameters). However, for some exponential family distributions μ_i is restricted. For example, for the Poisson distribution $\mu_i \in \mathbb{R}_+$; for the Bernoulli distribution $\mu_i \in (0, 1)$. If g is not chosen carefully, then there may exist a possible $\boldsymbol{x}_i$

and $\boldsymbol{\beta}$ such that $\eta_i \neq g(\mu_i)$ for any possible value of μ_i. Therefore, 'sensible' choices of link function map the set of allowed values for μ_i onto $\mathbb{R}$.

Recall that for a random variable Y with a distribution from the exponential family, $E(Y) = b'(\theta)$. Hence, for a generalised linear model

$$\mu_i = E(Y_i) = b'(\theta_i), \quad i = 1, \ldots, n.$$

Therefore

$$\theta_i = b'^{-1}(\mu_i), \quad i = 1, \ldots, n$$

and as $g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$, then

$$\theta_i = b'^{-1}(g^{-1}[\boldsymbol{x}_i^T \boldsymbol{\beta}]), \quad i = 1, \ldots, n. \tag{5.5}$$

Hence, we can express the joint density (5.2) in terms of the coefficients $\boldsymbol{\beta}$, and for observed data $\boldsymbol{y}$, this is the likelihood $L(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$. As $\boldsymbol{\beta}$ is our parameter of real interest (describing the dependence of the response on the explanatory variables) this likelihood will play a crucial role.

Note that considerable simplification is obtained in (5.5) if the functions g and b'^{-1} are identical. Then

$$\theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} \qquad i = 1, \ldots, n$$

and the resulting likelihood is

$$L(\boldsymbol{\beta}) = \exp\left(\sum_{i=1}^{n} \frac{y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - b(\boldsymbol{x}_i^T \boldsymbol{\beta})}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i, \phi_i) \right).$$

The link function

$$g(\mu) \equiv b'^{-1}(\mu)$$

is called the *canonical* link function. Under the canonical link, the canonical parameter is equal to the linear predictor.

The canonical link functions are:

Distribution	$b(\theta)$	$b'(\theta) \equiv \mu$	$b'^{-1}(\mu) \equiv \theta$	Link	Name
Normal	$\frac{1}{2}\theta^2$	θ	μ	$g(\mu) = \mu$	Identity
Poisson	$\exp\theta$	$\exp\theta$	$\log\mu$	$g(\mu) = \log\mu$	Log
Binomial	$\log(1 + \exp\theta)$	$\frac{\exp\theta}{1+\exp\theta}$	$\log\frac{\mu}{1-\mu}$	$g(\mu) = \log\frac{\mu}{1-\mu}$	Logit

5.4 Examples of generalised linear models

5.4.1 The linear model

The linear model considered in Chapter 4 is also a generalised linear model. We assume $Y_1, \ldots, Y_n$ are independent normally distributed random variables, so that $Y_i \sim N(\mu_i, \sigma^2)$. We have seen in Example 5.1 that the normal distribution is a member of the exponential family.

The explanatory variables enter a linear model through the linear predictor

$$\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad i = 1, \ldots, n.$$

The link between $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and the linear predictor $\boldsymbol{\eta}$ is through the (canonical) identity link function

$$\mu_i = \eta_i, \quad i = 1, \ldots, n.$$

5.4.2 Models for binary data

In binary regression, we assume either $Y_i \sim$ Bernoulli(p_i), or $Y_i \sim$ binomial(n_i, p_i), where n_i are known. The objective is to model the success probability p_i as a function of the explanatory variables $\boldsymbol{x}_i$. We have seen in Examples 5.3 and 5.4 that the Bernoulli and binomial distributions are members of the exponential family.

When the canonical (logit) link is used, we have

$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i} = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

This implies

$$p_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} = \frac{1}{1 + \exp(-\eta_i)}.$$

The function $F(\eta) = \frac{1}{1+\exp(-\eta)}$ is the cumulative distribution function (cdf) of a distribution called the logistic distribution.

The cumulative distribution functions of other distributions are also commonly used to generate link functions for binary regression. For example, if we let

$$p_i = \Phi(\boldsymbol{x}_i^T \boldsymbol{\beta}) = \Phi(\eta_i),$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution, then we get the link function

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

which is called the **probit** link.

5.4.3 Models for count data

If Y_i represent counts of the number of times an event occurs in a fixed time (or a fixed region of space), we might model $Y_i \sim \text{Poisson}(\lambda_i)$. We have seen in Example 5.2 that the Poisson distribution is a member of the exponential family.

With the canonical (log) link, we have

$$\log \lambda_i = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

or

$$\lambda_i = \exp\{\eta_i\} = \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\}.$$

This model is often called a log-linear model.

Now suppose that Y_i represents a count of the number of events which occur in a given region i, for instance the number of times a particular drug is prescribed on a given day, in a district i of a country. We might want to model the prescription rate **per patient** in the district λ_i^*. Write N_i is the number of patients registered in district i, often called the **exposure** of observation i. We model $Y_i \sim \text{Poisson}(N_i \lambda_i^*)$, where

$$\log \lambda_i^* = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

Equivalently, we may write the model as $Y_i \sim \text{Poisson}(\lambda_i)$, where

$$\log \lambda_i = \log N_i + \boldsymbol{x}_i^T \boldsymbol{\beta},$$

(since $\lambda_i = N_i \lambda_i^*$, so $\log \lambda_i = \log N_i + \log \lambda_i^*$). The log-exposure $\log N_i$ appears as a fixed term in the linear predictor, without any associated parameter. Such a fixed term is called an **offset**.

5.5 Maximum likelihood estimation

The regression coefficients $\beta_1, \ldots, \beta_p$ describe the pattern by which the response depends on the explanatory variables. We use the observed data $y_1, \ldots, y_n$ to *estimate* this pattern of dependence.

As usual, we maximise the log-likelihood function which, from (5.2), can be written

$$\ell(\boldsymbol{\beta}, \boldsymbol{\phi}) = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i, \phi_i) \tag{5.6}$$

and depends on $\boldsymbol{\beta}$ through

$$\begin{aligned} \theta_i &= (b')^{-1}(\mu_i), \\ \mu_i &= g^{-1}(\eta_i), \\ \eta_i &= \boldsymbol{x}_i^T \boldsymbol{\beta} = \sum_{i=1}^{p} x_{ij} \beta_j, \quad i = 1, \ldots, n. \end{aligned}$$

To find $\hat{\boldsymbol{\beta}}$, we consider the scores

$$u_k(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_k} \ell(\boldsymbol{\beta}, \boldsymbol{\phi}) \qquad k = 1, \ldots, p$$

and then find $\hat{\boldsymbol{\beta}}$ to solve $u_k(\hat{\boldsymbol{\beta}}) = 0$ for $k = 1, \ldots, p$.

From (5.6)

$$\begin{aligned}
u_k(\boldsymbol{\beta}) &= \frac{\partial}{\partial \beta_k}\ell(\boldsymbol{\beta}, \boldsymbol{\phi}) \\
&= \frac{\partial}{\partial \beta_k}\sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + \frac{\partial}{\partial \beta_k}\sum_{i=1}^{n} c(y_i, \phi_i) \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial \beta_k}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)}\right] \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial \theta_i}\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)}\right]\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_k} \\
&= \sum_{i=1}^{n}\frac{y_i - b'(\theta_i)}{a(\phi_i)}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_k}, \quad k = 1, \ldots, p,
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial \theta_i}{\partial \mu_i} &= \left[\frac{\partial \mu_i}{\partial \theta_i}\right]^{-1} = \frac{1}{b''(\theta_i)} \\
\frac{\partial \mu_i}{\partial \eta_i} &= \left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1} = \frac{1}{g'(\mu_i)} \\
\frac{\partial \eta_i}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k}\sum_{j=1}^{p} x_{ij}\beta_j = x_{ik}.
\end{aligned}$$

Therefore

$$u_k(\boldsymbol{\beta}) = \sum_{i=1}^{n}\frac{y_i - b'(\theta_i)}{a(\phi_i)}\frac{x_{ik}}{b''(\theta_i)g'(\mu_i)} = \sum_{i=1}^{n}\frac{y_i - \mu_i}{\mathrm{Var}(Y_i)}\frac{x_{ik}}{g'(\mu_i)}, \quad k = 1, \ldots, p, \tag{5.7}$$

which depends on $\boldsymbol{\beta}$ through $\mu_i \equiv E(Y_i)$ and $\mathrm{Var}(Y_i)$, $i = 1, \ldots, n$.

In theory, we solve the p simultaneous equations $u_k(\hat{\boldsymbol{\beta}}) = 0,\ k = 1, \ldots, p$ to evaluate $\hat{\boldsymbol{\beta}}$. In practice, these equations are usually non-linear and have no analytic solution. Therefore, we rely on numerical methods to solve them.

First, we note that the Hessian and Fisher information matrices can be derived directly from (5.7).

$$[\boldsymbol{H}(\boldsymbol{\beta})]_{jk} = \frac{\partial^2}{\partial\beta_j\partial\beta_k}\ell(\boldsymbol{\beta},\boldsymbol{\phi}) = \frac{\partial}{\partial\beta_j}u_k(\boldsymbol{\beta}).$$

Therefore

$$\begin{aligned}[\boldsymbol{H}(\boldsymbol{\beta})]_{jk} &= \frac{\partial}{\partial\beta_j}\sum_{i=1}^{n}\frac{y_i-\mu_i}{\mathrm{Var}(Y_i)}\frac{x_{ik}}{g'(\mu_i)} \\ &= \sum_{i=1}^{n}\frac{-\frac{\partial\mu_i}{\partial\beta_j}}{\mathrm{Var}(Y_i)}\frac{x_{ik}}{g'(\mu_i)} + \sum_{i=1}^{n}(y_i-\mu_i)\frac{\partial}{\partial\beta_j}\left[\frac{x_{ik}}{\mathrm{Var}(Y_i)g'(\mu_i)}\right]\end{aligned}$$

and

$$\begin{aligned}[\mathcal{I}(\boldsymbol{\beta})]_{jk} &= \sum_{i=1}^{n}\frac{\frac{\partial\mu_i}{\partial\beta_j}}{\mathrm{Var}(Y_i)}\frac{x_{ik}}{g'(\mu_i)} - \sum_{i=1}^{n}(E[Y_i]-\mu_i)\frac{\partial}{\partial\beta_j}\left[\frac{x_{ik}}{\mathrm{Var}(Y_i)g'(\mu_i)}\right] \\ &= \sum_{i=1}^{n}\frac{\frac{\partial\mu_i}{\partial\beta_j}}{\mathrm{Var}(Y_i)}\frac{x_{ik}}{g'(\mu_i)} \\ &= \sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{\mathrm{Var}(Y_i)g'(\mu_i)^2}.\end{aligned}$$

Hence we can write

$$\mathcal{I}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} \tag{5.8}$$

where

$$\boldsymbol{X} = \begin{pmatrix}\boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T\end{pmatrix} = \begin{pmatrix}x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np}\end{pmatrix},$$

$$\boldsymbol{W} = \mathrm{diag}(\boldsymbol{w}) = \begin{pmatrix}w_1 & 0 & \cdots & 0 \\ 0 & w_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & w_n\end{pmatrix}$$

and

$$w_i = \frac{1}{\mathrm{Var}(Y_i)g'(\mu_i)^2}, \quad i = 1, \ldots, n.$$

The Fisher information matrix $\mathcal{I}(\boldsymbol{\beta})$ depends on $\boldsymbol{\beta}$ through $\boldsymbol{\mu}$ and $\mathrm{Var}(Y_i),\ i = 1, \ldots, n$.

We notice that the score in (5.7) may now be written as

$$u_k(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mu_i)x_{ik}w_i g'(\mu_i) = \sum_{i=1}^{n} x_{ik}w_i z_i, \quad k = 1, \ldots, p,$$

where

$$z_i = (y_i - \mu_i)g'(\mu_i), \quad i = 1, \ldots, n.$$

Therefore

$$\boldsymbol{u}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{z}. \tag{5.9}$$

One possible method to solve the p simultaneous equations $\boldsymbol{u}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$ that give $\hat{\boldsymbol{\beta}}$ is the (multivariate) Newton-Raphson method.

If $\boldsymbol{\beta}^{(m)}$ is the current estimate of $\hat{\boldsymbol{\beta}}$ then the next estimate is

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \boldsymbol{H}(\boldsymbol{\beta}^{(m)})^{-1}\boldsymbol{u}(\boldsymbol{\beta}^{(m)}). \tag{5.10}$$

In practice, an alternative to Newton-Raphson replaces $\boldsymbol{H}(\boldsymbol{\theta})$ in (5.10) with $E[\boldsymbol{H}(\boldsymbol{\theta})] \equiv -\mathcal{I}(\boldsymbol{\beta})$. Therefore, if $\boldsymbol{\beta}^{(m)}$ is the current estimate of $\hat{\boldsymbol{\beta}}$ then the next estimate is

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \mathcal{I}(\boldsymbol{\beta}^{(m)})^{-1}\boldsymbol{u}(\boldsymbol{\beta}^{(m)}). \tag{5.11}$$

The resulting iterative algorithm is called *Fisher scoring*. Notice that if we substitute (5.8) and (5.9) into (5.11) we get

$$\begin{aligned}
\boldsymbol{\beta}^{(m+1)} &= \boldsymbol{\beta}^{(m)} + [\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{z}^{(m)} \\
&= [\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}]^{-1}[\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}\boldsymbol{\beta}^{(m)} + \boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{z}^{(m)}] \\
&= [\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}[\boldsymbol{X}\boldsymbol{\beta}^{(m)} + \boldsymbol{z}^{(m)}] \\
&= [\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}[\boldsymbol{\eta}^{(m)} + \boldsymbol{z}^{(m)}],
\end{aligned}$$

where $\boldsymbol{\eta}^{(m)}$, $\boldsymbol{W}^{(m)}$ and $\boldsymbol{z}^{(m)}$ are all functions of $\boldsymbol{\beta}^{(m)}$.

Note that this is a weighted least squares equation, that is $\boldsymbol{\beta}^{(m+1)}$ minimises the weighted sum of squares

$$(\boldsymbol{\eta}+\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}(\boldsymbol{\eta}+\boldsymbol{z}-\boldsymbol{X}\boldsymbol{\beta})=\sum_{i=1}^{n} w_i\left(\eta_i+z_i-\boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2$$

as a function of $\boldsymbol{\beta}$ where $w_1,\ldots,w_n$ are the weights and $\boldsymbol{\eta}+\boldsymbol{z}$ is called the *adjusted dependent variable*. Therefore, the Fisher scoring algorithm proceeds as follows.

1. Choose an initial estimate $\boldsymbol{\beta}^{(m)}$ for $\hat{\boldsymbol{\beta}}$ at $m=0$.
2. Evaluate $\boldsymbol{\eta}^{(m)}$, $\boldsymbol{W}^{(m)}$ and $\boldsymbol{z}^{(m)}$ at $\boldsymbol{\beta}^{(m)}$.
3. Calculate

$$\boldsymbol{\beta}^{(m+1)}=[\boldsymbol{X}^T\boldsymbol{W}^{(m)}\boldsymbol{X}]^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(m)}[\boldsymbol{\eta}^{(m)}+\boldsymbol{z}^{(m)}].$$

4. If $||\boldsymbol{\beta}^{(m+1)}-\boldsymbol{\beta}^{(m)}||>\epsilon$, for some prespecified (small) tolerance ϵ then set $m\to m+1$ and go to 2.
5. Use $\boldsymbol{\beta}^{(m+1)}$ as the solution for $\hat{\boldsymbol{\beta}}$.

As this algorithm involves iteratively minimising a weighted sum of squares, it is sometimes known as *iteratively (re)weighted least squares*.

Notes

1. Recall that the canonical link function is $g(\mu)=b'^{-1}(\mu)$ and with this link $\eta_i=g(\mu_i)=\theta_i$. Then

$$\frac{1}{g'(\mu_i)}=\frac{\partial\mu_i}{\partial\eta_i}=\frac{\partial\mu_i}{\partial\theta_i}=b''(\theta_i),\quad i=1,\ldots,n.$$

   Therefore $\mathrm{Var}(Y_i)g'(\mu_i)=a(\phi_i)$ which does not depend on $\boldsymbol{\beta}$, and hence

$$\frac{\partial}{\partial\beta_j}\left[\frac{x_{ik}}{\mathrm{Var}(Y_i)g'(\mu_i)}\right]=0$$

   for all $j=1,\ldots,p$. It follows that $\boldsymbol{H}(\boldsymbol{\theta})=-\mathcal{I}(\boldsymbol{\beta})$ and, for the canonical link, Newton-Raphson and Fisher scoring are equivalent.
2. The linear model is a generalised linear model with identity link, $\eta_i=g(\mu_i)=\mu_i$ and $\mathrm{Var}(Y_i)=\sigma^2$ for all $i=1,\ldots,n$. Therefore $w_i=$

$[\mathrm{Var}(Y_i)g'(\mu_i)^2]^{-1} = \sigma^{-2}$ and $z_i = (y_i - \mu_i)g'(\mu_i) = y_i - \eta_i$ for $i = 1, \ldots, n$. Hence $\boldsymbol{z} + \boldsymbol{\eta} = \boldsymbol{y}$ and $\boldsymbol{W} = \sigma^{-2}\boldsymbol{I}$, neither of which depend on $\boldsymbol{\beta}$. So the Fisher scoring algorithm converges in a single iteration to the usual least squares estimate.

3. Estimation of an unknown scale parameter σ^2 is discussed later. A common (to all i) σ^2 has no effect on $\hat{\boldsymbol{\beta}}$.

5.6 Confidence intervals

Recall from Section 2.6 that the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with mean $\boldsymbol{\beta}$ (it is unbiased) and variance covariance matrix $\mathcal{I}(\boldsymbol{\beta})^{-1}$. For 'large enough n' we treat this distribution as an approximation.

Therefore, standard errors (estimated standard deviations) are given by

$$s.e.(\hat{\beta}_i) = [\mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}]_{ii}^{\frac{1}{2}} = [(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}]_{ii}^{\frac{1}{2}} \qquad i = 1, \ldots, p.$$

where the diagonal matrix $\hat{\boldsymbol{W}} = \mathrm{diag}(\hat{\boldsymbol{w}})$ is evaluated at $\hat{\boldsymbol{\beta}}$, that is $\hat{w}_i = (\hat{\mathrm{Var}}(Y_i)g'(\hat{\mu}_i)^2)^{-1}$ where $\hat{\mu}_i$ and $\hat{\mathrm{Var}}(Y_i)$ are evaluated at $\hat{\boldsymbol{\beta}}$ for $i = 1, \ldots, n$. Furthermore, if $\mathrm{Var}(Y_i)$ depends on an unknown scale parameter, then this too must be estimated in the standard error.

The asymptotic distribution of the maximum likelihood estimator can be used to provide approximate large sample confidence intervals. For given α we can find $z_{1-\frac{\alpha}{2}}$ such that

$$P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{\beta}_i - \beta_i}{[\mathcal{I}(\boldsymbol{\beta})^{-1}]_{ii}^{\frac{1}{2}}} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Therefore

$$P\left(\hat{\beta}_i - z_{1-\frac{\alpha}{2}}[\mathcal{I}(\boldsymbol{\beta})^{-1}]_{ii}^{\frac{1}{2}} \leq \beta_i \leq \hat{\beta}_i + z_{1-\frac{\alpha}{2}}[\mathcal{I}(\boldsymbol{\beta})^{-1}]_{ii}^{\frac{1}{2}}\right) = 1 - \alpha.$$

The endpoints of this interval cannot be evaluated because they also depend on the unknown parameter vector $\boldsymbol{\beta}$. However, if we replace $\mathcal{I}(\boldsymbol{\beta})$ by its

MLE $\mathcal{I}(\hat{\boldsymbol{\beta}})$ we obtain the approximate large sample $100(1-\alpha)\%$ confidence interval

$$[\hat{\beta}_i - s.e.(\hat{\beta}_i)z_{1-\frac{\alpha}{2}}\ ,\ \hat{\beta}_i + s.e.(\hat{\beta}_i)z_{1-\frac{\alpha}{2}}].$$

For $\alpha = 0.10, 0.05, 0.01,\ z_{1-\frac{\alpha}{2}} = 1.64, 1.96, 2.58,$ respectively.

5.7 Comparing generalised linear models

5.7.1 The likelihood ratio test

If we have a set of competing generalised linear models which might explain the dependence of the response on the explanatory variables, we will want to determine which of the models is most appropriate. Recall that we have three main requirements of a statistical model; plausibility, parsimony and goodness of fit, of which parsimony and goodness of fit are statistical issues.

As with linear models, we proceed by comparing models pairwise using a likelihood ratio test. This kind of comparison is restricted to situations where one of the models, H_0, is *nested* in the other, H_1. Then the asymptotic distribution of the log likelihood ratio statistic under H_0 is a chi-squared distribution with known degrees of freedom.

For generalised linear models, 'nested' means that H_0 and H_1 are

1. based on the same exponential family distribution, and
2. have the same link function, but
3. the explanatory variables present in H_0 are a subset of those present in H_1.

We will assume that model H_1 contains p linear parameters and model H_0 a subset of $q < p$ of these. Without loss of generality, we can think of H_1 as the model

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j \qquad i = 1, \ldots, n$$

and H_0 is the same model with

$$\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0.$$

Then model H_0 is a special case of model H_1, where certain coefficients are set equal to zero, and therefore $\Theta^{(0)}$, the set of values of the canonical parameter $\boldsymbol{\theta}$ allowed by H_0, is a subset of $\Theta^{(1)}$, the set of values allowed by H_1.

Now, the log likelihood ratio statistic for a test of H_0 against H_1 is

$$\begin{aligned} L_{01} &\equiv 2 \log \left(\frac{\max_{\boldsymbol{\theta} \in \Theta^{(1)}} L(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \Theta^{(0)}} L(\boldsymbol{\theta})} \right) \\ &= 2 \log L(\hat{\boldsymbol{\theta}}^{(1)}) - 2 \log L(\hat{\boldsymbol{\theta}}^{(0)}), \end{aligned} \tag{5.12}$$

where $\hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\theta}}^{(0)}$ follow from $b'(\hat{\theta}_i) = \hat{\mu}_i$, $g(\hat{\mu}_i) = \hat{\eta}_i$, $i = 1, \ldots, n$ where $\hat{\boldsymbol{\eta}}$ for each model is the linear predictor evaluated at the corresponding maximum likelihood estimate for $\boldsymbol{\beta}$. Here, we assume that $a(\phi_i)$, $i = 1, \ldots, n$ are known; unknown $a(\phi)$ is discussed in Section 5.9.

Recall that we reject H_0 in favour of H_1 when L_{01} is 'too large' (the observed data are much more probable under H_1 than H_0). To determine a threshold value k for L_{01}, beyond which we reject H_0, we set the size of the test α and use the result of Section 3.3 that, because H_0 is nested in H_1, L_{01} has an asymptotic chi-squared distribution with $p - q$ degrees of freedom. For example, if $\alpha = 0.05$, we reject H_0 in favour of H_1 when L_{01} is greater than the 95% point of the χ^2_{p-q} distribution.

Note that setting up our model selection procedure in this way is consistent with our desire for parsimony. The simpler model is H_0, and we do not reject H_0 in favour of the more complex model H_1 unless the data provide convincing evidence for H_1 over H_0, that is unless H_1 fits the data significantly better.

5.8 Scaled deviance and the saturated model

Consider a model where $\boldsymbol{\beta}$ is n-dimensional, and therefore $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$. Assuming that $\boldsymbol{X}$ is invertible, then this model places no constraints on the linear predictor $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$. It can take any value in $\mathbb{R}^n$. Correspondingly the means $\boldsymbol{\mu}$ and the canonical parameters $\boldsymbol{\theta}$ are unconstrained. The model is of

dimension n and can be parameterised equivalently using $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\mu}$ or $\boldsymbol{\theta}$. Such a model is called the *saturated* model.

As the canonical parameters $\boldsymbol{\theta}$ are unconstrained, we can calculate their maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ directly from their likelihood (5.2) (without first having to calculate $\hat{\boldsymbol{\beta}}$)

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i, \phi_i). \tag{5.13}$$

We obtain $\hat{\boldsymbol{\theta}}$ by first differentiating with respect to $\theta_1, \ldots, \theta_n$ to give

$$\frac{\partial}{\partial\theta_k}\ell(\boldsymbol{\theta}) = \frac{y_k - b'(\theta_k)}{a(\phi_k)} \qquad k = 1, \ldots, n.$$

Therefore $b'(\hat{\theta}_k) = y_k,\ k = 1, \ldots, n$, and it follows immediately that $\hat{\mu}_k = y_k,\ k = 1, \ldots, n$. Hence the saturated model fits the data perfectly, as the *fitted values* $\hat{\mu}_k$ and observed values y_k are the same for every observation $k = 1, \ldots, n$.

The saturated model is rarely of any scientific interest in its own right. It is highly parameterised, having as many parameters as there are observations. This goes against our desire for parsimony in a model. However, every other model is necessarily nested in the saturated model, and a test comparing a model H_0 against the saturated model H_S can be interpreted as a goodness of fit test. If the saturated model, which fits the observed data perfectly, does not provide a significantly better fit than model H_0, we can conclude that H_0 is an acceptable fit to the data.

The log likelihood ratio statistic for a test of H_0 against H_S is, from (5.12)

$$L_{0s} = 2\log L(\hat{\boldsymbol{\theta}}^{(s)}) - 2\log L(\hat{\boldsymbol{\theta}}^{(0)}),$$

where $\hat{\boldsymbol{\theta}}^{(s)}$ follows from $b'(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\mu}} = \boldsymbol{y}$ and $\hat{\boldsymbol{\theta}}^{(0)}$ is a function of the corresponding maximum likelihood estimate for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$. Under H_0, L_{0s} has an asymptotic chi-squared distribution with $n - q$ degrees of freedom. Therefore, if L_{0s} is 'too large' (for example, larger than the 95% point of the χ^2_{n-q} distribution) then we reject H_0 as a plausible model for the data, as it does not fit the data adequately.

The *degrees of freedom* of model H_0 is defined to be the degrees of freedom for this test, $n - q$, the number of observations minus the number of linear parameters of H_0. We call L_{0s} the *scaled deviance* (R calls it the *residual deviance*) of model H_0.

From (5.12) and (5.13) we can write the deviance of model H_0 as

$$L_{0s} = 2\sum_{i=1}^{n} \frac{y_i[\hat{\theta}_i^{(s)} - \hat{\theta}_i^{(0)}] - [b(\hat{\theta}_i^{(s)}) - b(\hat{\theta}_i^{(0)})]}{a(\phi_i)}, \tag{5.14}$$

which can be calculated using the observed data, provided that $a(\phi_i),\ i = 1, \ldots, n$ is known.

Notes

1. The log likelihood ratio statistic (5.12) for testing H_0 against a non-saturated alternative H_1 can be written as

$$\begin{aligned} L_{01} &= 2\log L(\hat{\boldsymbol{\theta}}^{(1)}) - 2\log L(\hat{\boldsymbol{\theta}}^{(0)}) \\ &= [2\log L(\hat{\boldsymbol{\theta}}^{(s)}) - 2\log L(\hat{\boldsymbol{\theta}}^{(0)})] - [2\log L(\hat{\boldsymbol{\theta}}^{(s)}) - 2\log L(\hat{\boldsymbol{\theta}}^{(1)})] \\ &= L_{0s} - L_{1s}. \end{aligned} \tag{5.15}$$

   Therefore the log likelihood ratio statistic for comparing two nested models is the difference of their deviances. Furthermore, as $p - q = (n - q) - (n - p)$, the degrees of freedom for the test is the difference in degrees of freedom of the two models.
2. The asymptotic theory used to derive the distribution of the log likelihood ratio statistic under H_0 does not really apply to the goodness of fit test (comparison with the saturated model). However, for binomial or Poisson data, we can proceed as long as the relevant binomial or Poisson distributions are likely to be reasonably approximated by normal distributions (*i.e.* for binomials with large denominators or Poissons with large means). However, for Bernoulli data, we cannot use the scaled deviance as a goodness of fit statistic in this way.
3. An alternative goodness of fit statistic for a model H_0 is Pearson's X^2 given by

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i^{(0)})^2}{\hat{\text{Var}}(Y_i)}. \tag{5.16}$$

X^2 is small when the squared differences between observed and fitted values (scaled by variance) is small. Hence, large values of X^2 correspond to poor fitting models. In fact, X^2 and L_{0s} are asymptotically equivalent and under H_0, X^2, like L_{0s}, has an asymptotic chi-squared distribution with $n - q$ degrees of freedom. However, the asymptotics associated with X^2 are often more reliable for small samples, so if there is a discrepancy between X^2 and L_{0s}, it is usually safer to base a test of goodness of fit on X^2.

4. Although the deviance for a model is expressed in (5.14) in terms of the maximum likelihood estimates of the canonical parameters, it is more usual to express it in terms of the maximum likelihood estimates $\hat{\mu}_i,\ i = 1, \ldots, n$ of the mean parameters. For the saturated model, these are just the observed values $y_i,\ i = 1, \ldots, n$, and for the model of interest, H_0, we call them the *fitted values*. Hence, for a particular generalised linear model, the scaled deviance function describes how discrepancies between the observed and fitted values are penalised.

Example 5.5 (Poisson)**.** Suppose $Y_i \sim \text{Poisson}(\lambda_i),\ i = 1, \ldots, n$. Recall from Section 5.2 that $\theta = \log \lambda$, $b(\theta) = \exp \theta$, $\mu = b'(\theta) = \exp \theta$ and $\text{Var}(Y) = a(\phi)V(\mu) = 1 \cdot \mu$. Therefore, by (5.14) and (5.16)

$$\begin{aligned} L_{0s} &= 2 \sum_{i=1}^{n} y_i [\log \hat{\mu}_i^{(s)} - \log \hat{\mu}_i^{(0)}] - [\hat{\mu}_i^{(s)} - \hat{\mu}_i^{(0)}] \\ &= 2 \sum_{i=1}^{n} y_i \log \left(\frac{y_i}{\hat{\mu}_i^{(0)}} \right) - y_i + \hat{\mu}_i^{(0)} \end{aligned}$$

and

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i^{(0)})^2}{\hat{\mu}_i^{(0)}}.$$

Example 5.6 (Binomial)**.** Suppose $n_i Y_i \sim \text{Binomial}(n_i, p_i),\ i = 1, \ldots, n$. Recall from Section 5.2 that $\theta = \log \frac{p}{1-p}$, $b(\theta) = \log(1 + \exp \theta)$, $\mu = b'(\theta) = \frac{\exp \theta}{1+\exp \theta}$ and $\text{Var}(Y) = a(\phi)V(\mu) = \frac{1}{n} \cdot \mu(1 - \mu)$. Therefore, by (5.14) and (5.16)

$$L_{0s} = 2\sum_{i=1}^{n} n_i y_i \left[\log\frac{\hat{\mu}_i^{(s)}}{1-\hat{\mu}_i^{(s)}} - \log\frac{\hat{\mu}_i^{(0)}}{1-\hat{\mu}_i^{(0)}}\right] + 2\sum_{i=1}^{n} n_i \left[\log(1-\hat{\mu}_i^{(s)}) - \log(1-\hat{\mu}_i^{(0)})\right]$$
$$= 2\sum_{i=1}^{n}\left[n_i y_i \log\left(\frac{y_i}{\hat{\mu}_i^{(0)}}\right) + n_i(1-y_i)\log\left(\frac{1-y_i}{1-\hat{\mu}_i^{(0)}}\right)\right]$$

and

$$X^2 = \sum_{i=1}^{n} \frac{n_i(y_i - \hat{\mu}_i^{(0)})^2}{\hat{\mu}_i^{(0)}(1-\hat{\mu}_i^{(0)})}.$$

Bernoulli data are binomial with $n_i = 1,\ i = 1, \ldots, n$.

5.9 Models with unknown $a(\phi)$

The theory of Section 5.7 has assumed that $a(\phi)$ is known. This is the case for both the Poisson distribution ($a(\phi) = 1$) and the binomial distribution ($a(\phi) = 1/n$). Neither the scaled deviance (5.14) nor Pearson X^2 statistic (5.16) can be evaluated unless $a(\phi)$ is known. Therefore, when $a(\phi)$ is not known, we cannot use the scaled deviance as a measure of goodness of fit, or to compare models using (5.15). For such models, there is no equivalent goodness of fit test, but we can develop a test for comparing nested models.

Here we assume that $a(\phi_i) = \sigma^2/m_i,\ i = 1, \ldots, n$ where σ^2 is a common unknown scale parameter and $m_1, \ldots, m_n$ are known weights. (A linear model takes this form, as $\mathrm{Var}(Y_i) = \sigma^2,\ i = 1, \ldots, n$, so $m_i = 1,\ i = 1, \ldots, n$.) Under this assumption

$$L_{0s} = \frac{2}{\sigma^2}\sum_{i=1}^{n} m_i y_i[\hat{\theta}_i^{(s)} - \hat{\theta}_i^{(0)}] - m_i[b(\hat{\theta}_i^{(s)}) - b(\hat{\theta}_i^{(0)})] = \frac{1}{\sigma^2} D_{0s}, \qquad (5.17)$$

where D_{0s} is defined to be twice the sum above, which can be calculated using the observed data. We call D_{0s} the *deviance* of the model.

In order to test nested models H_0 and H_1 as set up in Section 5.7.1, we calculate the test statistic

$$F = \frac{L_{01}/(p-q)}{L_{1s}/(n-p)} = \frac{(L_{0s}-L_{1s})/(p-q)}{L_{1s}/(n-p)}$$
$$= \frac{\left(\frac{1}{\sigma^2}D_{0s} - \frac{1}{\sigma^2}D_{1s}\right)/(p-q)}{\frac{1}{\sigma^2}D_{1s}/(n-p)} = \frac{(D_{0s}-D_{1s})/(p-q)}{D_{1s}/(n-p)}. \tag{5.18}$$

This statistic does not depend on the unknown scale parameter σ^2, so can be calculated using the observed data. Asymptotically, if H_0 is true, we know that $L_{01} \sim \chi^2_{p-q}$ and $L_{1s} \sim \chi^2_{n-p}$. Furthermore, L_{01} and L_{1s} are independent (not proved here) so F has an asymptotic $F_{p-q,n-p}$ distribution. Hence, we compare nested generalised linear models by calculating F and rejecting H_0 in favour of H_1 if F is too large (for example, greater than the 95% point of the relevant F distribution).

The dependence of the maximum likelihood equations $\boldsymbol{u}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ on σ^2 (where $\boldsymbol{u}$ is given by (5.7)) can be eliminated by multiplying through by σ^2. However, inference based on the maximum likelihood estimates, as described in Section 5.6, does require knowledge of σ^2. This is because asymptotically $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ is the inverse of the Fisher information matrix $\mathcal{I}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$, and this depends on $w_i = \frac{1}{\mathrm{Var}(Y_i)g'(\mu_i)^2}$, where $\mathrm{Var}(Y_i) = a(\phi_i)b''(\theta_i) = \sigma^2 b''(\theta_i)/m_i$ here.

Therefore, to calculate standard errors and confidence intervals, we need to supply an estimate $\hat{\sigma}^2$ of σ^2. Generally, we do not use the maximum likelihood estimate. Instead, we notice that, from (5.17), $L_{0s} = D_{0s}/\sigma^2$, and we know that asymptotically, if model H_0 is an adequate fit, L_{0s} has a χ^2_{n-q} distribution. Hence

$$E(L_{0s}) = E\left(\frac{1}{\sigma^2}D_{0s}\right) = n-q \quad \Rightarrow \quad E\left(\frac{1}{n-q}D_{0s}\right) = \sigma^2.$$

Therefore the deviance of a model divided by its degrees of freedom is an asymptotically unbiased estimator of the scale parameter σ^2. Hence $\hat{\sigma}^2 = D_{0s}/(n-q)$.

An alternative estimator of σ^2 is based on the Pearson X^2 statistic. As $\mathrm{Var}(Y) = a(\phi)V(\mu) = \sigma^2 V(\mu)/m$ here, then from (5.16)

$$X^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{m_i (y_i - \hat{\mu}_i^{(0)})^2}{V(\hat{\mu}_i^{(0)})}. \tag{5.19}$$

Again, if H_0 is an adequate fit, X^2 has an chi-squared distribution with $n-q$ degrees of freedom, so

$$\hat{\sigma}^2 = \frac{1}{n-q} \sum_{i=1}^{n} \frac{m_i (y_i - \hat{\mu}_i^{(0)})^2}{V(\hat{\mu}_i^{(0)})}$$

is an alternative unbiased estimator of σ^2. This estimator tends to be more reliable in small samples.

Example 5.7 (Normal)**.** Suppose $Y_i \sim N(\mu_i, \sigma^2),\ i = 1, \ldots, n$. Recall from Section 5.2 that $\theta = \mu$, $b(\theta) = \theta^2/2$, $\mu = b'(\theta) = \theta$ and $\mathrm{Var}(Y) = a(\phi)V(\mu) = \sigma^2 \cdot 1$, so $m_i = 1,\ i = 1, \ldots, n$. Therefore, by (5.17),

$$D_{0s} = 2 \sum_{i=1}^{n} y_i [\hat{\mu}_i^{(s)} - \hat{\mu}_i^{(0)}] - [\frac{1}{2} \hat{\mu}_i^{(s)^2} - \frac{1}{2} \hat{\mu}_i^{(0)^2}] = \sum_{i=1}^{n} [y_i - \hat{\mu}_i^{(0)}]^2, \tag{5.20}$$

which is just the residual sum of squares for model H_0. Therefore, we estimate σ^2 for a normal GLM by its residual sum of squares for the model divided by its degrees of freedom. From (5.19), the estimate for σ^2 based on X^2 is identical.

5.10 Residuals

Recall that for linear models, we define the residuals to be the differences between the observed and fitted values $y_i - \hat{\mu}_i^{(0)},\ i = 1, \ldots, n$. From (5.20) we notice that both the scaled deviance and Pearson X^2 statistic for a normal GLM are the sum of the squared residuals divided by σ^2. We can generalise this to define residuals for other generalised linear models in a natural way.

For any GLM we define the *Pearson residuals* to be

$$r_i^P = \frac{y_i - \hat{\mu}_i^{(0)}}{\hat{\mathrm{Var}}(Y_i)^{\frac{1}{2}}} \qquad i = 1, \ldots, n.$$

Then, from (5.16), X^2 is the sum of the squared Pearson residuals.

For any GLM we define the *deviance residuals* to be

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i^{(0)}) \left[2 \frac{y_i[\hat{\theta}_i^{(s)} - \hat{\theta}_i^{(0)}] - [b(\hat{\theta}_i^{(s)}) - b(\hat{\theta}_i^{(0)})]}{a(\phi_i)} \right]^{\frac{1}{2}}, \quad i = 1, \ldots, n,$$

where $\text{sign}(x) = 1$ if $x > 0$ and -1 if $x < 0$. Then, from (5.14), the scaled deviance, L_{0s}, is the sum of the squared deviance residuals.

When $a(\phi) = \sigma^2/m$ and σ^2 is unknown, as in Section 5.9, the residuals are based on (5.17) and (5.19), and the expressions above need to be multiplied through by σ^2 to eliminate dependence on the unknown scale parameter. Therefore, for a normal GLM the Pearson and deviance residuals are both equal to the usual residuals, $y_i - \hat{\mu}_i^{(0)}$, $i = 1, \ldots, n$.

Residual plots are most commonly of use in normal linear models, where they provide an essential check of the model assumptions. This kind of check is less important for a model without an unknown scale parameter as the scaled deviance provides a useful overall assessment of fit which takes into account most aspects of the model.

However, when data have been collected in serial order, a plot of the deviance or Pearson residuals against the order may again be used as a check for potential serial correlation.

Otherwise, residual plots are most useful when a model fails to fit (scaled deviance is too high). Then, examining the residuals may give an indication of the reason(s) for lack of fit. For example, there may be a small number of outlying observations.

A plot of deviance or Pearson residuals against the linear predictor should produce something that looks like a random scatter. If not, then this may be due to incorrect link function, wrong scale for an explanatory variable, or perhaps a missing polynomial term in an explanatory variable.

Chapter 6

Models for categorical data

6.1 Contingency tables

A particularly important application of generalised linear models is the analysis of categorical data. Here, the data are observations of one or more categorical variables on each of a number of units (often individuals). Therefore, each of the units are *cross-classified* by the categorical variables.

For example, the `job` dataset gives the job satisfaction and income band of 901 individuals from the 1984 General Social Survey, which is summarised in Table 6.1.

	Job Satisfaction			
Income ($)	Very Dissat.	A Little Dissat.	Moderately Sat.	Very Sat.
<6000	20	24	80	82
6000-15000	22	38	104	125
15000-25000	13	28	81	113
>25000	7	18	54	92

Table 6.1: A contingency table of the `job` dataset.

A cross-classification table like this is called a *contingency table*. This is a *two-way table*, as there are two classifying variables. It might also be describe

	Job Satisfaction				
Income ($)	Very Dissat.	A Little Dissat.	Moderately Sat.	Very Sat.	**Sum**
<6000	20	24	80	82	**206**
6000-15000	22	38	104	125	**289**
15000-25000	13	28	81	113	**235**
>25000	7	18	54	92	**171**
Sum	**62**	**108**	**319**	**412**	**901**

Table 6.2: A contingency table of the `job` dataset, including one-way margins.

as a 4×4 contingency table (as each of the classifying variables takes one of four possible levels).

Each position in a contingency table is called a *cell* and the number of individuals in a particular cell is the *cell count*.

Partial classifications derived from the table are called *margins*. For a two-way table these are often displayed in the margins of the table, as in Table 6.2. These are one-way margins as they represent the classification of items by a single variable; income group and job satisfaction respectively.

The `lymphoma` dataset gives information about 30 patients, classified by cell type of lymphoma, sex, and response to treatment, as shown in Table 6.3. This is an example of a three-way contingency table. It is a $2 \times 2 \times 2$ or 2^3 table.

		Remission	
Cell Type	Sex	No	Yes
Diffuse	Female	3	1
	Male	12	1
Nodular	Female	2	6
	Male	1	4

Table 6.3: A contingency table of the `lymphoma` dataset.

For *multiway* tables, higher order margins may be calculated. For example,

for `lymphoma`, the two-way Cell type/Sex margin is shown in Table 6.4.

	Sex	
Cell Type	Female	Male
Diffuse	4	13
Nodular	8	5

Table 6.4: The two-way Cell type/Sex margin for the `lymphoma` dataset.

6.2 Log-linear models

We can model contingency table data using generalised linear models. To do this, we assume that the cell counts are observations of independent Poisson random variables. This is intuitively sensible as the cell counts are non-negative integers (the sample space for the Poisson distribution). Therefore, if the table has n cells, which we label $1, \ldots, n$, then the observed cell counts $y_1, \ldots, y_n$ are assumed to be observations of independent Poisson random variables $Y_1, \ldots, Y_n$. We denote the means of these Poisson random variables by $\mu_1, \ldots, \mu_n$. The canonical link function for the Poisson distribution is the log function, and we assume this link function throughout. A generalised linear model for Poisson data using the log link function is called a *log-linear model*, as we have already seen in Section 5.4.3.

The explanatory variables in a log-linear model for contingency table data are the cross-classifying variables, or *factors*. As usual with categorical variables, we can include interactions in the model as well as just main effects (see Example 4.6). Such a model will describe how the expected count in each cell depends on the classifying variables, and any interactions between them. Interpretation of these models will be discussed further in Section 6.5.

Table 6.5 shows the original data structure of the `job` dataset. It provides exactly the same data as the contingency table in Table 6.1, but in a different format, sometimes called *long* format. A log-linear model is just a Poisson GLM, where the response variable is `Count`, and `Income` and `Satisfaction` are explanatory variables.

Income	Satisfaction	Count
<6000	Very Dissatisfied	20
<6000	A Little Dissatisfied	24
<6000	Moderately Satisfied	80
<6000	Very Satisfied	82
6000-15000	Very Dissatisfied	22
6000-15000	A Little Dissatisfied	38
6000-15000	Moderately Satisfied	104
6000-15000	Very Satisfied	125
15000-25000	Very Dissatisfied	13
15000-25000	A Little Dissatisfied	28
15000-25000	Moderately Satisfied	81
15000-25000	Very Satisfied	113
>25000	Very Dissatisfied	7
>25000	A Little Dissatisfied	18
>25000	Moderately Satisfied	54
>25000	Very Satisfied	92

Table 6.5: The `job` dataset.

Cell	Sex	Remis	Count
Nodular	Male	No	1
Nodular	Male	Yes	4
Nodular	Female	No	2
Nodular	Female	Yes	6
Diffuse	Male	No	12
Diffuse	Male	Yes	1
Diffuse	Female	No	3
Diffuse	Female	Yes	1

Table 6.6: The `lymphoma` dataset.

Table 6.6 shows the `lymphoma` dataset in long format. Again, a log-linear model for the contingency table (Table 6.3) is just a Poisson GLM for this data, where in this case the response variable is `Cell`, and `Sex` and `Remis` are explanatory variables.

6.3 Multinomial sampling

Although the assumption of Poisson distributed observations is convenient for the purposes of modelling, it might not be a realistic assumption, because of the way in which the data have been collected. Frequently, when contingency table data are obtained, the total number of observations (the *grand total*, the sum of all the cell counts) is fixed in advance. In this case, no individual cell count can exceed the prespecified fixed total, so the assumption of Poisson sampling is invalid as the sample space is bounded. Furthermore, with a fixed total, the observations can no longer be observations of independent random variables.

For example, consider the `lymphoma` contingency table from Table 6.3. It may be that these data were collected over a fixed period of time, and that in that time there happened to be 30 patients. In this case, the Poisson assumption is perfectly valid. However, it may have been decided at the outset to collect data on 30 patients, in which case the grand total is fixed at 30, and the Poisson assumption is not valid.

When the grand total is fixed, a more appropriate distribution for the cell counts is the *multinomial* distribution. The multinomial distribution is the distribution of cell counts arising when a prespecified total of N items are each independently assigned to one of n cells, where the probability of being classified into cell i is p_i, $i = 1, \ldots, n$, so $\sum_{i=1}^{n} p_i = 1$. The probability function for the multinomial distribution is

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}; \mathbf{p}) &= P(Y_1 = y_1, \ldots, Y_n = y_n) \\ &= \begin{cases} N! \prod_{i=1}^{n} \frac{p_i^{y_i}}{y_i!} & \text{if } \sum_{i=1}^{n} y_i = N \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{6.1}$$

The binomial is the special case of the multinomial with two cells ($n = 2$).

We can still use a log-linear model for contingency table data when the data have been obtained by multinomial sampling. We model $\log \mu_i = \log(Np_i)$, $i = 1, \ldots, n$ as a linear function of explanatory variables. However, such a model must preserve $\sum_{i=1}^{n} \mu_i = N$, the grand total which is fixed in advance.

From (6.1), the log-likelihood for a multinomial log-linear model is

$$\ell(\boldsymbol{\mu}) = \sum_{i=1}^{n} y_i \log \mu_i - N \log N - \sum_{i=1}^{n} \log y_i! + \log N!.$$

Therefore, the maximum likelihood estimates $\hat{\boldsymbol{\mu}}$ maximise $\sum_{i=1}^{n} y_i \log \mu_i$ subject to the constraints $\sum_{i=1}^{n} \mu_i = N = \sum_{i=1}^{n} y_i$ (multinomial sampling) and $\log \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ (imposed by the model).

For a Poisson log-linear model,

$$L(\boldsymbol{\mu}) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}.$$

Therefore,

$$\ell(\boldsymbol{\mu}) = -\sum_{i=1}^{n} \mu_i + \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \log y_i! \tag{6.2}$$

$$= -\sum_{i=1}^{n} \exp(\log \mu_i) + \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \log y_i!. \tag{6.3}$$

Now any Poisson log-linear model with an intercept can be expressed as

$$\log \mu_i = \alpha + \text{other terms depending on } i, \qquad i = 1, \ldots, n$$

so that

$$\frac{\partial}{\partial \alpha} \ell(\boldsymbol{\mu}) = -\sum_{i=1}^{n} \exp(\log \mu_i) + \sum_{i=1}^{n} y_i. \tag{6.4}$$

$$\Rightarrow \qquad \sum_{i=1}^{n} \hat{\mu}_i = \sum_{i=1}^{n} y_i. \tag{6.5}$$

From (6.2), we notice that, at $\alpha = \hat{\alpha}$ the log-likelihood takes the form

$$\ell(\boldsymbol{\mu}) = -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \log y_i!.$$

Hence, when we maximise the log-likelihood, for a Poisson log-linear model with intercept, with respect to the other log-linear parameters, we are maximising $\sum_{i=1}^{n} y_i \log \mu_i$ subject to the constraints $\sum_{i=1}^{n} \mu_i = \sum_{i=1}^{n} y_i$ from (6.5) and $\log \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ (imposed by the model).

Therefore, the maximum likelihood estimates for multinomial log-linear parameters are identical to those for Poisson log-linear parameters. Furthermore, the maximised log-likelihoods for both Poisson and multinomial models take the form $\sum_{i=1}^{n} y_i \log \hat{\mu}_i$ as functions of the log-linear parameter estimates. Therefore any inferences based on maximised log-likelihoods (such as likelihood ratio tests) will be the same.

Therefore, we can analyse contingency table data using Poisson log-linear models, even when the data has been obtained by multinomial sampling. All that is required is that we ensure that the Poisson model contains an intercept term.

6.4 Product multinomial sampling

Sometimes margins other than just the grand total may be prespecified. For example, consider the `lymphoma` contingency table in Table 6.3. It may have

been decided at the outset to collect data on 18 male patients and 12 female patients. Alternatively, perhaps the distribution of both the Sex and Cell type of the patients was fixed in advance as in Table 6.4. In cases where a margin is fixed by design, the data consist of a number of fixed total subgroups, defined by the fixed margin. Neither Poisson nor multinomial sampling assumptions are valid. The appropriate distribution combines a separate, independent multinomial for each subgroup. For example, if just the Sex margin is fixed as above, then the appropriate distribution for modelling the data is two independent multinomials, one for males with $N = 18$ and one for females with $N = 12$. Each of these multinomials has four cells, representing the cross-classification of the relevant patients by Cell Type and Remission. Alternatively, if it is the Cell type/Sex margin which has been fixed, then the appropriate distribution is four independent two-cell multinomials (binomials) representing the remission classification for each of the four fixed-total patient subgroups.

When the data are modelled using independent multinomials, then the joint distribution of the cell counts $Y_1, \ldots, Y_n$ is the product of terms of the same form as (6.1), one for each fixed-total subgroup. We call this a distribution a *product multinomial*. Each subgroup has its own fixed total. The full joint density is a product of n terms, as before, with each cell count appearing exactly once.

For example, if the Sex margin is fixed for `lymphoma`, then the product multinomial distribution has the form

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \mathbf{p}) = \begin{cases} N_m! \prod_{i=1}^4 \frac{p_{mi}^{y_{mi}}}{y_{mi}!} N_f! \prod_{i=1}^4 \frac{p_{fi}^{y_{fi}}}{y_{fi}!} & \text{if } \sum_{i=1}^4 y_{mi} = N_m \text{ and } \sum_{i=1}^4 y_{fi} = N_f \\ 0 & \text{otherwise,} \end{cases}$$

where N_m and N_f are the two fixed marginal totals (18 and 12 respectively), $y_{m1}, \ldots, y_{m4}$ are the cell counts for the Cell type/Remission cross-classification for males and $y_{f1}, \ldots, y_{f4}$ are the corresponding cell counts for females. Here $\sum_{i=1}^4 p_{mi} = \sum_{i=1}^4 p_{fi} = 1$, $E(Y_{mi}) = N_m p_{mi}$, $i = 1, \ldots, 4$, and $E(Y_{fi}) = N_f p_{fi}$, $i = 1, \ldots, 4$.

Using similar results to those used in Section 6.3 (but not proved here), we can analyse contingency table data using Poisson log-linear models, even when the data has been obtained by product multinomial sampling. However, we must ensure that the Poisson model contains a term corresponding to the

fixed margin (and all marginal terms). Then, the estimated means for the specified margin are equal to the corresponding fixed totals.

For example, for the `lymphoma` dataset, for inferences obtained using a Poisson model to be valid when the Sex margin is fixed in advance, the Poisson model must contain the Sex main effect (and the intercept). For inferences obtained using a Poisson model to be valid when the Cell type/Sex margin is fixed in advance, the Poisson model must contain the Cell type/Sex interaction, and all marginal terms (the Cell type main effect, the Sex main effect and the intercept).

Therefore, when analysing product multinomial data using a Poisson log-linear model, certain terms **must** be present in any model we fit. If they are removed, the inferences would no longer be valid.

6.5 Interpreting log-linear models for two-way tables

Log-linear models for contingency tables enable us to determine important properties concerning the joint distribution of the classifying variables. In particular, the form of our preferred log linear model for a table will have implications for how the variables are associated.

Each combination of the classifying variables occurs exactly once in a contingency table. Therefore, the model with the highest order interaction (between all the variables) and all marginal terms (all other interactions) is the saturated model. The implication of this model is that every combination of levels of the variables has its own mean (probability) and that there are no relationships between these means (no structure). The variables are highly dependent.

To consider the implications of simpler models, we first consider a two-way $r \times c$ table where the two classifying variables R and C have r and c levels respectively. The saturated model $R * C$ implies that the two variables are associated. If we remove the RC interaction, we have the model $R + C$,

$$\log \mu_i = \alpha + \beta_R(r_i) + \beta_C(c_i), \qquad i = 1, \ldots, n$$

where $n = rc$ is the total number of cells in the table. Because of the equivalence of Poisson and multinomial sampling, we can think of each cell mean μ_i as equal to Np_i where N is the total number of observations in the table, and p_i is a cell probability. As each combination of levels of R and C is represented in exactly one cell, it is also convenient to replace the cell label i by the pair of labels j and k representing the corresponding levels of R and C respectively. Hence

$$\log p_{jk} = \alpha + \beta_R(j) + \beta_C(k) - \log N, \quad j = 1, \ldots, r, \quad k = 1, \ldots, c.$$

Therefore

$$P(R = j, C = k) = \exp[\alpha + \beta_R(j) + \beta_C(k) - \log N], \quad j = 1, \ldots, r, \quad k = 1, \ldots, c,$$

so

$$\begin{aligned} 1 &= \sum_{j=1}^{r} \sum_{k=1}^{c} \exp[\alpha + \beta_R(j) + \beta_C(k) - \log N] \\ &= \frac{1}{N} \exp[\alpha] \sum_{j=1}^{r} \exp[\beta_R(j)] \sum_{k=1}^{c} \exp[\beta_C(k)]. \end{aligned}$$

Furthermore

$$\begin{aligned} P(R = j) &= \sum_{k=1}^{c} \exp[\alpha + \beta_R(j) + \beta_C(k) - \log N] \\ &= \frac{1}{N} \exp[\alpha] \exp[\beta_R(j)] \sum_{k=1}^{c} \exp[\beta_C(k)], \quad j = 1, \ldots, r, \end{aligned}$$

and

$$\begin{aligned} P(C = k) &= \sum_{j=1}^{r} \exp[\alpha + \beta_R(j) + \beta_C(k) - \log N] \\ &= \frac{1}{N} \exp[\alpha] \exp[\beta_C(k)] \sum_{j=1}^{r} \exp[\beta_R(j)], \quad k = 1, \ldots, c. \end{aligned}$$

Therefore

$$\begin{aligned} P(R=j)P(C=k) &= \frac{1}{N}\exp[\alpha]\exp[\beta_C(k)]\exp[\beta_R(j)] \times 1 \\ &= P(R=j, C=k), \quad j=1,\ldots,r, \quad k=1,\ldots,c. \end{aligned}$$

Absence of the interaction $R*C$ in a log-linear model implies that R and C are independent variables. Absence of main effects is generally less interesting, and main effects are typically not removed from a log-linear model.

6.6 Interpreting log-linear models for multiway tables

In multiway tables, absence of a two-factor interaction does not necessarily mean that the two variables are independent. For example, consider the `lymphoma` dataset, with 3 binary classifying variables Sex (S), Cell type (C) and Remission (R). After comparing the fit of several possible models, we find that a reasonable log-linear model for these data is $R * C + C * S$. Hence the interaction between remission and sex is absent. The fitted cell probabilities from this log-linear model are shown in Table 6.7.

		Remission	
Cell Type	Sex	No	Yes
Diffuse	Female	0.1176	0.0157
	Male	0.3824	0.0510
Nodular	Female	0.0615	0.2051
	Male	0.0385	0.1282

Table 6.7: Fitted probabilities of each cell in the `lymphoma` dataset.

The estimated probabilities for the two-way Sex/Remission margin (together with the corresponding one-way margins) are shown in Table 6.8.

	Remission		
Sex	No	Yes	**Sum**
Female	0.1792	0.2208	**0.4**
Male	0.4208	0.1792	**0.6**
Sum	**0.6**	**0.4**	**1.0**

Table 6.8: Fitted marginal probabilities for the `lymphoma` dataset.

It can immediately be seen that this model does not imply independence of R and S, as $\hat{P}(R,S) \neq \hat{P}(R)\hat{P}(S)$. What the model $R*C + C*S$ implies is that R is independent of S *conditional on* C, that is

$$P(R,S|C) = P(R|C)P(S|C).$$

Another way of expressing this is

$$P(R|S,C) = P(R|C),$$

that is, the probability of each level of R given a particular combination of S and C, does not depend on which level S takes. Table 6.9 shows the estimated conditional probabilities for the `lymphoma` data. The probability of remission depends only on a patient's cell type, and not on their sex.

		Remission		
Cell Type	Sex	No	Yes	$\hat{P}(R\|S,C)$
Diffuse	Female	0.1176	0.0157	0.12
	Male	0.3824	0.0510	0.12
Nodular	Female	0.0615	0.2051	0.77
	Male	0.0385	0.1282	0.77

Table 6.9: Fitted probabilities of each cell and conditional probability of remission in the `lymphoma` dataset.

In general, if we have an r-way contingency table with classifying variables $X_1, \ldots, X_r$, then a log linear model which does not contain the $X_1 * X_2$ interaction (and therefore by the principle of marginality contains no interaction involving both X_1 and X_2) implies that X_1 and X_2 are *conditionally*

independent given $X_3, \ldots, X_r$, that is

$$P(X_1, X_2|X_3, \ldots, X_r) = P(X_1|X_3, \ldots, X_r)P(X_2|X_3, \ldots, X_r).$$

The proof of this is similar to the proof in the two-way case. Again, an alternative way of expressing conditional independence is

$$P(X_1|X_2, X_3, \ldots, X_r) = P(X_1|X_3, \ldots, X_r)$$

or

$$P(X_2|X_1, X_3, \ldots, X_r) = P(X_2|X_3, \ldots, X_r).$$

Although for the `lymphoma` dataset R and S are conditionally independent given C, they are not marginally independent. Using the marginal cell probabilities from Table 6.8, we find that the probability of remission is 0.30 for men and 0.55 for women. Male patients have a much lower probability of remission. The reason for this is that, although R and S are not directly associated, they are both associated with C. Observing the estimated values it is clear that patients with nodular cell type have a greater probability of remission, and furthermore, that female patients are more likely to have this cell type than males. Hence women are more likely to have remission than men.

Suppose the factors for a three-way tables are X_1, X_2 and X_3. We can list all possible models and the implications for the conditional independence structure:

1. Model 1 containing the terms X_1, X_2, X_3. All factors are mutually independent.
2. Model 2 containing the terms $X_1 * X_2, X_3$. The factor X_3 is jointly independent of X_1 and X_2.
3. Model 3 containing the terms $X_1 * X_2, X_2 * X_3$. The factors X_1 and X_3 are conditionally independent given X_2.
4. Model 4 containing the terms $X_1 * X_2, X_2 * X_3, X_1 * X_3$. There is no conditional independence structure. This is the model without the highest order interaction term.
5. Model 5 containing $X_1 * X_2 * X_3$. This is the saturated model. No more simplification of dependence structure is possible.

6.7 Simpson's paradox

Conditional and marginal association of two variables can therefore often appear somewhat different. Sometimes, the association can be 'reversed' so that what looks like a positive association marginally, becomes a negative association conditionally. This is known as *Simpson's paradox*.

In 1972-74, a survey of women was carried out in an area of Newcastle. A follow-up survey was carried out 20 years later. Among the variables observed in the initial survey was whether or not the individual was a smoker and among those in the follow-up survey was whether the individual was still alive, or had died in the intervening period. A summary of the responses is shown in Table 6.10.

Smoker	Dead	Alive	$\hat{P}$(Dead|Smoker)
Yes	139	443	0.24
No	230	502	0.31

Table 6.10: Number of respondents dead or alive at follow up, by smoking status

Looking at this table, it appears that the non-smokers had a greater probability of dying. However, there is an important extra variable to be considered, related to both smoking habit and mortality – age (at the time of the initial survey). When we consider this variable, we get Table 6.11. Conditional on every age at outset, it is now the smokers who have a higher probability of dying. The marginal association is reversed in the table conditional on age, because mortality (obviously) and smoking are associated with age. There are proportionally many fewer smokers in the older age-groups (where the probability of death is greater).

When making inferences about associations between variables, it is important that all other variables which are relevant are considered. Marginal inferences may lead to misleading conclusions.

Age	Smoker	Dead	Alive	$\hat{P}(\text{Dead}\|\text{Age}, \text{Smoker})$
18–34	Yes	5	174	0.03
	No	6	213	0.03
35–44	Yes	14	95	0.13
	No	7	114	0.06
45–54	Yes	27	103	0.21
	No	12	66	0.15
55–64	Yes	51	64	0.44
	No	40	81	0.33
65–74	Yes	29	7	0.81
	No	101	28	0.78
75–	Yes	13	0	1
	No	64	0	1

Table 6.11: Number of respondents dead or alive at follow up, by smoking status and age.