MA30084

GENERALISED LINEAR MODELS

Lecture Notes

Evangelos Evangelou

Last Updated: 6th January 2025

Contents

Prolegomena

Statistical models are devised to explain and understand the variability in the observed values of a characteristic of interest from a sample of the population under study. A statistical model is simply an assumption about the behaviour of nature which incorporates randomness to account for this variability. In your first-year statistics course, you have learnt models for one variable, such as the normal distribution model, the Poisson distribution, etc. These models were extended in the second-year statistics course to incorporate linear relationships between the variable of interest (the response variable) and other measured variables (the explanatory variables).

On the other hand, the linear model assumption can be inappropriate in many interesting applications. For example, suppose that we wish to predict the outcome of an election. Using data from surveys and past elections, we can analyse a candidate's policies and predict what share of the vote this particular candidate will receive. However, the impact of each policy on the candidate's share is non-linear. To see why this is the case, suppose that a certain policy, let's call it Policy A, is favoured by the majority of the electorate but it is not among this candidate's policies. What would happen if our candidate ends up adopting Policy A? That would depend on the share of the vote the candidate had before adopting this policy. If the candidate's share was low, a significant portion of the electorate who previously opposed this candidate may now be convinced to vote for them. On the other hand, if the candidate's share was already high, the impact of adopting Policy A would be smaller. Therefore, the effect of a change in policy is not always the same, hence a non-linear relationship between policy and vote share.

This unit builds on the foundations of the linear model and extends it to allow for the modelling of non-linear relationships and non-normally-distributed observations, giving rise to a new class of models, the generalised linear model, having the linear model as a special case. We will see applications of this model in medicine, biology, insurance, among other fields. In this unit, you will learn the many applications of the general theory of maximum likelihood estimation in the context of statistical modelling. You will also learn how to assess the goodness-of-fit of a model, and techniques for improving the model fit. All methodology will be applied to real data examples using R.

Modelling flexibility comes with a price. The linear model is simple-enough for statisticians to be able to derive analytical results for parameter estimation, confidence intervals, and hypothesis tests. For generalised linear models, no such closed-form expressions exist in general, therefore we rely on asymptotic theory and numerical algorithms to be able to perform inference. The theory developed here is not restricted to generalised linear models but it is almost universal, therefore, the techniques that you will learn here will be useful in more challenging modelling problems.

These lecture notes is a product of many years of me teaching this unit as an undergraduate statistics course in generalised linear models. In producing these notes, I assimilated knowledge from various textbooks such as "An introduction to generalized linear models" by Dobson, "Generalized linear models" by McCullagh and Nelder, and "Extending the linear model with R" by Faraway, as well as from my many years of research in the field. The notes offer a comprehensive study of the theory and applications of generalised linear models, and include several optional sections for further study, which I hope will make an interesting read.

1. Likelihood Theory

This chapter summarises the results on likelihood functions and maximum likelihood estimation. The results from this chapter will be used later in the development of the methodology for inference in generalised linear models.

1.1 Log-likelihood function

In a typical statistical modelling exercise, we observe data $\boldsymbol{y} = (y_1, \ldots, y_n)$, which we want to analyse. To that end, we propose a (statistical) model, which is a mathematical expression of the data-generating process. The simplest model is the *independent and identically distributed* (iid) random sample, where the components of $\boldsymbol{y}$ are assumed to be pairwise independent, and from the same distribution. We write $f(\boldsymbol{y}; \boldsymbol{\theta})$ for the assumed joint probability density/mass function (pdf) of $\boldsymbol{y}$, which typically depends on unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. For the iid model, $f(\boldsymbol{y}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta})$, where $f(y_i; \boldsymbol{\theta})$ denotes the pdf of the ith observation. We can think of the value of $f(\boldsymbol{y}; \boldsymbol{\theta})$ as how "likely" it is to observe $\boldsymbol{y}$ for a given $\boldsymbol{\theta}$ value. In other words, we are given $\boldsymbol{\theta}$, and we are asked about $\boldsymbol{y}$. In statistical inference, however, we observe the data $\boldsymbol{y}$, and ask questions about $\boldsymbol{\theta}$. To emphasise that interest is in $\boldsymbol{\theta}$ instead of $\boldsymbol{y}$, we write the joint pdf as $L(\boldsymbol{\theta}|\boldsymbol{y})$, i.e., $L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$ and call it the *likelihood function*.

Example 1.1. The number of customers arriving at the queue of a supermarket till every minute for a ten-minute period are recorded and shown below.

$$y_1 = 2,\ y_2 = 1,\ y_3 = 2,\ y_4 = 3,\ y_5 = 1,\ y_6 = 2,\ y_7 = 1,\ y_8 = 1,\ y_9 = 0,\ y_{10} = 3.$$

These observations consist of our data. The whole data vector is denoted by $\boldsymbol{y} = (y_1, \ldots, y_{10}) = (2, 1, 2, 3, 1, 2, 1, 1, 0, 3)$.

A sensible model is to assume that the observations are independently distributed from the Poisson distribution with rate θ. Thus, each y_i, for $i = 1, \ldots, n$, where $n = 10$ denotes the sample size, is an iid observation from the Poisson distribution with rate θ. The model can be succinctly expressed as

$$y_i \overset{\text{iid}}{\sim} \text{Po}(\theta),\ i = 1, \ldots, n,$$

where the notation $\overset{\text{iid}}{\sim}$ is read as "are independently and identically distributed as". Alternatively, the model could be written as

$$y_i \overset{\text{ind}}{\sim} \text{Po}(\theta),\ i = 1, \ldots, n,$$

where the notation $\overset{\text{ind}}{\sim}$ is read as "are independently distributed as". The "identically" part is implied since the distribution on the right-hand size of $\overset{\text{ind}}{\sim}$ does not depend on i. Note that if we just write

$$y_i \sim \text{Po}(\theta),\ i = 1, \ldots, n,$$

without specifying ind (or iid, which implies ind), then our model would be unspecified, as we have only provided the marginal distributions, but not the joint distribution of $\boldsymbol{y}$.

Based on the assumed model, the probability mass function for y_i is given by

$$f(y_i; \theta) = \frac{\theta^{y_i}}{y_i!} e^{-\theta},$$

so the probability of observing the sample $\boldsymbol{y}$ is

$$f(\boldsymbol{y}; \theta) = \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta} = \theta^{n\bar{y}} e^{-n\theta} / (\prod_{i=1}^{n} y_i!), \tag{1.1}$$

where $\bar{y} = \sum y_i / n$. The likelihood function is (1.1) as a function of θ for given $\boldsymbol{y}$. The likelihood is denoted by $L(\theta|\boldsymbol{y})$. So $L(\theta|\boldsymbol{y}) = f(\boldsymbol{y}; \theta)$. ▶

Definition 1.1 (Likelihood and score functions)**.**
Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a random sample of size n from a joint distribution that depends on a parameter $\boldsymbol{\theta}$ ($\boldsymbol{\theta}$ may be a vector). The joint pdf of the n observations is written as

$$f(\boldsymbol{y}; \boldsymbol{\theta}) =: L(\boldsymbol{\theta}|\boldsymbol{y}). \tag{1.2}$$

When the expression in (1.2) is viewed as a function of the parameter $\boldsymbol{\theta}$, it is called the *likelihood function*.

The *log-likelihood* is simply the logarithm of the likelihood function.

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) := \log L(\boldsymbol{\theta}|\boldsymbol{y}).$$

The first derivative of the log-likelihood function is called the *score function* and is denoted by

$$\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) := \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\boldsymbol{y}).$$

If $\boldsymbol{\theta}$ is a vector, say $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ then the score function is vector-valued with elements

$$\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) = \left(\frac{\partial}{\partial \theta_1} \ell(\boldsymbol{\theta}|\boldsymbol{y}), \ldots, \frac{\partial}{\partial \theta_p} \ell(\boldsymbol{\theta}|\boldsymbol{y}) \right).$$

If we assume that the observations are independent, with the ith marginal pdf denoted by $f_i(y_i; \boldsymbol{\theta})$, then

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f_i(y_i; \boldsymbol{\theta}),$$
$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \log f_i(y_i; \boldsymbol{\theta}).$$

Example 1.2. Refer to Example 1.1. From equation (1.1) we have

$$L(\theta|\boldsymbol{y}) = \frac{\theta^{n\bar{y}}}{\prod y_i!} e^{-n\theta}$$

and the log-likelihood is

$$\ell(\theta|\boldsymbol{y}) = n\bar{y}\log\theta - n\theta - \log\prod y_i!.$$

The log-likelihood function is plotted in Figure 1.1 using the data from the example. Note that the last term is a constant in terms of θ, so it does not impact inference. Therefore, we could simplify the log-likelihood by writing

$$\ell(\theta|\boldsymbol{y}) = n\bar{y}\log\theta - n\theta + \text{constant}.$$

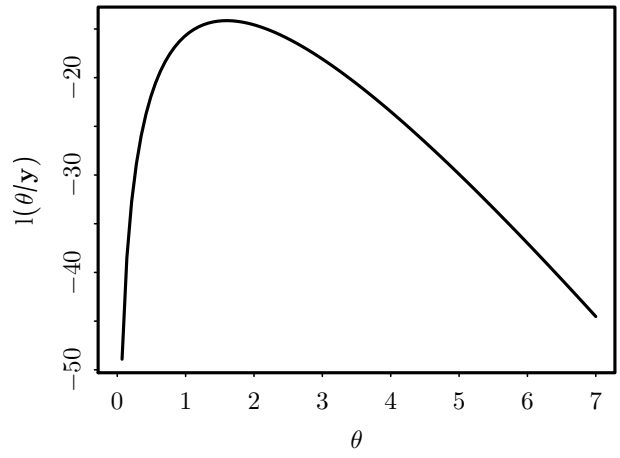


Figure 1.1: Log-likelihood function for Example 1.2.

The score function is

$$\boldsymbol{u}(\theta|\boldsymbol{y}) = \frac{n\bar{y}}{\theta} - n.$$

▶

Example 1.3. The normal distribution is probably the most ubiquitous distribution in statistics. It is a continuous distribution with two parameters, the mean μ and the variance σ^2. Its density function is given by

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}, \quad y \in \mathbb{R}.$$

In this example the parameter is a 2-dimensional vector, $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\mu, \sigma^2)$.

Suppose that we obtain an iid sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ from the normal distribution of size n. The log-likelihood function is

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\theta_2) - \frac{1}{2\theta_2}\sum_{i=1}^{n}(y_i - \theta_1)^2$$

Differentiating with respect to each parameter we have

$$\frac{\partial}{\partial\theta_1}\ell(\theta|\boldsymbol{y}) = \frac{1}{\theta_2}\sum_{i=1}^{n}(y_i - \theta_1)$$

$$\frac{\partial}{\partial\theta_2}\ell(\theta|\boldsymbol{y}) = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(y_i-\theta_1)^2$$

so the score function is

$$\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) = \left[\frac{1}{\theta_2}\sum_{i=1}^{n}(y_i-\theta_1)\ , \quad -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(y_i-\theta_1)^2\right]$$

▶

1.2 Fisher information matrix

The score function may also be viewed as a random vector, being a function of the random sample $\boldsymbol{y}$. The expected value of this random variable is

$$\mathsf{E}\,\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) = 0.$$

To see this write

$$\begin{aligned}
\mathsf{E}\,\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) &= \mathsf{E}\,\frac{\partial}{\partial\boldsymbol{\theta}}\log L(\boldsymbol{\theta}|\boldsymbol{y}) \\
&= \mathsf{E}\,\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{y};\boldsymbol{\theta}) \\
&= \int f(\boldsymbol{y};\boldsymbol{\theta})\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{y};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{y} \\
&= \int f(\boldsymbol{y};\boldsymbol{\theta})\frac{\frac{\partial}{\partial\boldsymbol{\theta}}f(\boldsymbol{y};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta})}\,\mathrm{d}\boldsymbol{y} \\
&= \int \frac{\partial}{\partial\boldsymbol{\theta}}f(\boldsymbol{y};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{y} \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\int f(\boldsymbol{y};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{y} \qquad \text{(by Leibniz rule)} \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}1 \\
&= 0.
\end{aligned}$$

The variance of this random variable is called the *Fisher information matrix*. It depends on the parameter $\boldsymbol{\theta}$ and it is denoted by $\mathcal{I}(\boldsymbol{\theta})$,

$$\mathcal{I}(\boldsymbol{\theta}) := \mathsf{Var}\,\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) = \mathsf{E}\left\{\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y})\boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y})^{\mathsf{T}}\right\}.$$

It can be shown that

$$\mathcal{I}(\boldsymbol{\theta}) = -\,\mathsf{E}\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}}\ell(\boldsymbol{\theta}|\boldsymbol{y})\right\}.$$

For a given sample $\boldsymbol{y}$, the matrix of negative observed second derivatives is called the *observed information matrix* and is denoted by

$$\mathcal{J}(\boldsymbol{\theta}|\boldsymbol{y}) := -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}}\ell(\boldsymbol{\theta}|\boldsymbol{y})$$

so $\mathcal{I}(\boldsymbol{\theta}) = \mathsf{E}\,\mathcal{J}(\boldsymbol{\theta}|\boldsymbol{y})$.

Example 1.4. For the Poisson distribution with mean θ and for sample $\boldsymbol{y}$ of size n,

$$\ell(\theta|\boldsymbol{y}) = \sum y_i \log\theta - n\theta - \log\prod y_i!$$

so

$$\boldsymbol{u}(\theta|\boldsymbol{y}) = \frac{\sum y_i}{\theta} - n.$$

Therefore

$$\begin{aligned}
\mathsf{E}\,\boldsymbol{u}(\theta|\boldsymbol{y}) &= \mathsf{E}\left\{\frac{\sum y_i}{\theta} - n\right\} \\
&= \frac{\sum \mathsf{E}\, y_i}{\theta} - n \\
&= \frac{\sum \theta}{\theta} - n \\
&= \frac{n\theta}{\theta} - n \\
&= 0,
\end{aligned}$$

as expected, and variance

$$\begin{aligned}
\mathcal{I}(\theta) &= \mathsf{Var}\,\boldsymbol{u}(\theta|\boldsymbol{y}) \\
&= \frac{\sum \mathsf{Var}\, y_i}{\theta^2} \\
&= \frac{n}{\theta}.
\end{aligned}$$

The observed information is

$$\mathcal{J}(\theta|\boldsymbol{y}) = \frac{\sum y_i}{\theta^2},$$

with expectation

$$\begin{aligned}
\mathsf{E}\,\mathcal{J}(\theta|\boldsymbol{y}) &= \mathsf{E}\,\frac{\sum y_i}{\theta^2} \\
&= \frac{n\,\theta}{\theta^2} \\
&= \frac{n}{\theta} \\
&= \mathcal{I}(\theta).
\end{aligned}$$

In this case the Fisher information and the observed information matrices are actually scalars because there is only one parameter. For the normal distribution of Example 1.3, they will be 2×2 symmetric matrices, i.e.

$$\mathcal{J}(\theta|\boldsymbol{y}) = \begin{pmatrix} -\dfrac{\partial^2\ell}{\partial\theta_1^2} & -\dfrac{\partial^2\ell}{\partial\theta_1\partial\theta_2} \\ -\dfrac{\partial^2\ell}{\partial\theta_2\partial\theta_1} & -\dfrac{\partial^2\ell}{\partial\theta_2^2} \end{pmatrix} \qquad \mathcal{I}(\theta) = \begin{pmatrix} -\mathsf{E}\,\dfrac{\partial^2\ell}{\partial\theta_1^2} & -\mathsf{E}\,\dfrac{\partial^2\ell}{\partial\theta_1\partial\theta_2} \\ -\mathsf{E}\,\dfrac{\partial^2\ell}{\partial\theta_2\partial\theta_1} & -\mathsf{E}\,\dfrac{\partial^2\ell}{\partial\theta_2^2} \end{pmatrix}$$

▶

It is clear from its definition that the Fisher information matrix is positive definite. The importance of the Fisher information matrix is that if $\boldsymbol{\theta}$ is unknown and $\tilde{\boldsymbol{\theta}}(\boldsymbol{y})$ is an unbiased estimator of $\boldsymbol{\theta}$ (i.e. $\mathsf{E}\,\tilde{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{\theta}$) based on the sample $\boldsymbol{y}$, then

$$\mathsf{Var}\,\tilde{\boldsymbol{\theta}}(\boldsymbol{y}) \geq \mathcal{I}(\boldsymbol{\theta})^{-1}. \tag{1.3}$$

(For $p \times p$ matrices A and B, $A \geq B$ indicates that the difference $A - B$ is a positive semi-definite matrix.) The inequality (1.3) is known as the *Cramer-Rao inequality*. Therefore if we can find some unbiased estimator $\hat{\boldsymbol{\theta}}$ such that $\mathsf{Var}\,\hat{\boldsymbol{\theta}}(\boldsymbol{y}) = \mathcal{I}(\boldsymbol{\theta})^{-1}$ then this is the best possible estimator we can have. Such estimator is called *efficient*.

1.3 Maximum likelihood estimation

If we compare the likelihood function at two parameter points, $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$, and find that

$$L(\boldsymbol{\theta}'|\boldsymbol{y}) > L(\boldsymbol{\theta}''|\boldsymbol{y}),$$

then the sample we actually observed is more likely to have occurred if $\boldsymbol{\theta} = \boldsymbol{\theta}'$ than if $\boldsymbol{\theta} = \boldsymbol{\theta}''$, which may be interpreted by saying that $\boldsymbol{\theta}'$ is a more plausible value for the parameter $\boldsymbol{\theta}$ than $\boldsymbol{\theta}''$. It therefore makes sense to search for the value of $\boldsymbol{\theta}$ that maximises the likelihood function (or equivalently the log-likelihood function).

Definition 1.2 (Maximum likelihood estimator)**.**
For given data $\boldsymbol{y}$ from a distribution depending on an unknown parameter $\boldsymbol{\theta}$, belonging to a parameter space Θ, the parameter value $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{y}) \in \Theta$ that maximises the log-likelihood is called the *maximum likelihood estimator* (MLE),

$$\hat{\boldsymbol{\theta}} := \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}|\boldsymbol{y}).$$

We note that the maximum likelihood estimator depends on the given sample. Its value will be different for different samples and is therefore a random variable.

Theorem 1.1 (Asymptotic distribution of the maximum likelihood estimator)**.**
Suppose that $y_1, \ldots, y_n$ is an iid sample from a distribution depending on the p-dimensional parameter $\boldsymbol{\theta}$ and let $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ denote the MLE of $\boldsymbol{\theta}$. Then, as $n \to \infty$,

$$\hat{\boldsymbol{\theta}}(\boldsymbol{y}) \mathrel{\dot{\sim}} \mathrm{N}_p(\boldsymbol{\theta}, \mathcal{I}(\boldsymbol{\theta})^{-1}),$$

i.e., the p-dimensional multivariate normal distribution with mean $\boldsymbol{\theta}$ and variance-covariance matrix $\mathcal{I}(\boldsymbol{\theta})^{-1}$. The conditions for the theorem are the following. The true parameter value must be in the interior of the parameter space, the log-likelihood function must be three times differentiable and its third derivatives must be bounded.

Here and subsequently the symbol $\mathrel{\dot{\sim}}$ will be used to denote the asymptotic distribution of a random variable as the sample size increases to infinity ($n \to \infty$). In practice this means that if the sample size is large enough, then the actual distribution of the random variable in question is close to its asymptotic distribution.

The importance of Theorem 1.1 is that it allows us to construct approximate confidence intervals and hypothesis tests for the unknown parameters based on the MLE. Furthermore it tells us that the MLE is asymptotically unbiased and efficient. In practice, however, the parameter $\boldsymbol{\theta}$ is unknown so the variance of the MLE, $\mathcal{I}(\boldsymbol{\theta})^{-1}$, is also unknown and a confidence

interval cannot be constructed. To overcome this, we replace it by its estimate $\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}$, i.e. we evaluate the Fisher information at the MLE.

If the log-likelihood is concave, we can find the maximum likelihood estimator by setting the score function to 0. In this case we find $\hat{\boldsymbol{\theta}}$ that satisfies $\boldsymbol{u}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = 0$. Sometimes solving for the MLE cannot be done analytically and an iterative procedure must be used.

All iterative procedures start with an initial guess $\boldsymbol{\theta}^{(0)}$ and perform a sequence of approximations of the form

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + \Delta^{(i)},\ i = 0, 1, 2, \ldots$$

until convergence is achieved.

To define the iterative procedure, perform a Taylor expansion to the score function evaluated at the MLE around a trial value $\boldsymbol{\theta}^*$ so that

$$\boldsymbol{u}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \approx \boldsymbol{u}(\boldsymbol{\theta}^*|\boldsymbol{y}) + \boldsymbol{u}'(\boldsymbol{\theta}^*|\boldsymbol{y})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*). \tag{1.4}$$

We can easily see that $\boldsymbol{u}'(\boldsymbol{\theta}^*|\boldsymbol{y}) = -\mathcal{J}(\boldsymbol{\theta}^*|\boldsymbol{y})$ and by definition $\boldsymbol{u}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = 0$. Therefore, solving for $\hat{\boldsymbol{\theta}}$ in (1.4),

$$\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}^* + \mathcal{J}(\boldsymbol{\theta}^*|\boldsymbol{y})^{-1}\boldsymbol{u}(\boldsymbol{\theta}^*|\boldsymbol{y}). \tag{1.5}$$

Since (1.5) is still an approximation, it will not give the correct value for $\hat{\boldsymbol{\theta}}$ but it can be used to define an iterative procedure for obtaining the MLE. The iterative procedure based on the updating equation (1.5) is called the *Newton-Raphson*. An alternative procedure known as the *Fisher scoring* is to replace $\mathcal{J}(\boldsymbol{\theta}^*|\boldsymbol{y})$ in (1.5) by its expectation $\mathcal{I}(\boldsymbol{\theta}^*)$.

Definition 1.3 (Fisher scoring algorithm)**.**
The Fisher scoring algorithm is an iterative procedure for finding the MLE of a parameter $\boldsymbol{\theta}$ given the sample $\boldsymbol{y}$. It starts with an initial guess $\boldsymbol{\theta}^{(0)}$ and updates the current value of $\boldsymbol{\theta}^{(i)}$ by

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + \mathcal{I}(\boldsymbol{\theta}^{(i)})^{-1}\boldsymbol{u}(\boldsymbol{\theta}^{(i)}|\boldsymbol{y}) \tag{1.6}$$

until convergence.

Example 1.5 (MLE for the mean of the Poisson distribution)**.** Recall that the score function of an i.i.d. sample of size n from the Poisson distribution is $\boldsymbol{u}(\theta|\boldsymbol{y}) = \frac{\sum y_i}{\theta} - n$. Solving for $\boldsymbol{u}(\hat{\theta}|\boldsymbol{y}) = 0$ we obtain

$$\hat{\theta} = n^{-1}\sum_{i=1}^{n} y_i = \bar{y}.$$

In Example 1.1 with $n = 10$ and $\sum y_i = 16$, we obtain $\hat{\theta} = 1.6$.

Recall also that $\mathcal{I}(\theta) = \frac{n}{\theta}$ so $\mathcal{I}(\hat{\theta}) = \frac{10}{1.6}$ and $\mathsf{Var}(\hat{\theta}) \approx 0.16$. Therefore, an approximate 95% confidence interval for θ is $1.6 \pm (1.96)\sqrt{0.16} = (0.816, 2.384)$.

Suppose we want to test the hypothesis $H_0 : \theta = 1$ v.s $H_1 : \theta \neq 1$ at level $\alpha = 5\%$. The z-statistic is $z = (1.6 - 1)/\sqrt{0.16} = 1.5$ which in absolute value is less than the critical value of 1.96. Therefore the conclusion is that there is not enough evidence to reject the null hypothesis.

What are the Newton-Raphson and Fisher scoring algorithms for the MLE? Recall that $\mathcal{J}(\theta|\boldsymbol{y}) = \frac{n\bar{y}}{\theta^2}$ so $\mathcal{J}(\theta|\boldsymbol{y})^{-1}\boldsymbol{u}(\theta|\boldsymbol{y}) = \theta - \frac{\theta^2}{\bar{y}}$. So the Newton-Raphson algorithm is

$$\theta^{(j+1)} = \theta^{(j)} + \theta^{(j)} - \frac{\theta^{(j)2}}{\bar{y}} = 2\theta^{(j)} - \frac{\theta^{(j)2}}{\bar{y}}$$

while the Fisher scoring algorithm is

$$\theta^{(j+1)} = \theta^{(j)} + \theta^{(j)}(\frac{\bar{y}}{\theta^{(j)}} - 1) = \bar{y}.$$

In this case the Fisher scoring will converge after one iteration. ▶

Example 1.6 (Geometric distribution)**.** Consider a sequence of Bernoulli experiments with probability of success $\theta \in (0, 1)$. The geometric distribution measures the number of successes in the sequence until the first failure. Its probability mass function is given by

$$f(y; \theta) = \theta^y(1 - \theta),\ y = 0, 1, 2, \ldots.$$

The expected value of the geometric distribution is

$$\begin{aligned}
\mathsf{E}\, Y &= \sum_{y=0}^{\infty} y\, \theta^y(1 - \theta) \\
&= (1 - \theta) \sum_{y=0}^{\infty} (y + 1)\, \theta^y - (1 - \theta) \sum_{y=0}^{\infty} \theta^y \\
&= (1 - \theta) \frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{y=0}^{\infty} \theta^{y+1} - (1 - \theta) \frac{1}{1 - \theta} \\
&= (1 - \theta) \frac{\mathrm{d}}{\mathrm{d}\theta} \frac{\theta}{1 - \theta} - 1 \\
&= \frac{\theta}{1 - \theta}.
\end{aligned}$$

Suppose an i.i.d. sample $\boldsymbol{y}$ of size n is obtained from the geometric distribution with probability of success θ. The likelihood function is given by

$$L(\theta|\boldsymbol{y}) = \theta^{\sum y_i}(1 - \theta)^n$$

and the log-likelihood is

$$\ell(\theta|\boldsymbol{y}) = n\, \bar{y} \log \theta + n \log(1 - \theta).$$

The score function is

$$\boldsymbol{u}(\theta|\boldsymbol{y}) = n \left\{ \frac{\bar{y}}{\theta} - \frac{1}{1 - \theta} \right\}$$

so the MLE for θ, found by setting the score equal to 0, is

$$\hat{\theta} = \frac{\bar{y}}{1 + \bar{y}}$$

The observed information is

$$\mathcal{J}(\theta|\boldsymbol{y}) = n \left\{ \frac{\bar{y}}{\theta^2} + \frac{1}{(1 - \theta)^2} \right\}$$

so the Fisher information is

$$\begin{aligned}
\mathcal{I}(\theta) &= \mathsf{E}\, \mathcal{J}(\theta|\boldsymbol{y}) \\
&= n \left\{ \frac{\mathsf{E}\, \bar{y}}{\theta^2} + \frac{1}{(1 - \theta)^2} \right\}
\end{aligned}$$

$$= n\left\{\frac{\theta}{(1-\theta)\theta^2} + \frac{1}{(1-\theta)^2}\right\}$$
$$= n\frac{1}{(1-\theta)^2\theta}$$

Suppose that a random sample of size $n = 4$ gave $\bar{y} = 1.5$. The MLE for θ is $\hat{\theta} = \frac{1.5}{1+1.5} = 0.6$. The variance of the MLE is estimated by inverting the Fisher information to be $\frac{(1-0.6)^2(0.6)}{4} = 0.024$ and a 95% confidence interval for θ is $0.6 \pm (1.96)\sqrt{0.024} = (0.3, 0.9)$.

Although the MLE can be obtained analytically, we demonstrate the use of the Fisher scoring algorithm for iteratively finding the MLE. The rule for updating the current estimate $\theta^{(i)}$ to a new value $\theta^{(i+1)}$ is, by (1.6),

$$\theta^{(i+1)} = \theta^{(i)} + (1-\theta^{(i)})^2\theta^{(i)}\left\{\frac{\bar{y}}{\theta^{(i)}} - \frac{1}{1-\theta^{(i)}}\right\}$$
$$= (1-\theta^{(i)})^2\bar{y} + (\theta^{(i)})^2.$$

With starting value $\theta^{(0)} = 0.40$ the Fisher scoring algorithm converges to the MLE in 6 iterations. The following R code illustrates the convergence.

```
#### Fisher scoring for the MLE of the geometric distribution
fs.geom <- function (theta, avg) {
  ## Input:
  ## theta: Initial guess for the MLE
  ## avg: The sample average
  maxit <- 100 # Maximum number of iterations
  epsilon <- 1e-5 # Criterion for convergence
  theta.trace <- array (NA, maxit+1, dimnames=list(0:maxit))
  theta.trace[1] <- theta # Store initial estimate
  for (i in 1:maxit) { # Begin iterative procedure
    theta0 <- theta # Previous value of theta
    theta <- (1-theta)^2*avg + theta^2 # Update current estimate
    theta.trace[i+1] <- theta # Store current estimate
    converge <- abs(theta - theta0)/theta0 < epsilon # Converge?
    if (converge) break # If yes, no more iterations
  }
  ## Output:
  ## theta: MLE for theta
  ## niter: Number of iterations until convergence
  ## theta.trace: Sequence of updates until convergence
  ## convergence: Indicator whether the alogrithm has converged
  list (theta=theta, niter=i, theta.trace=theta.trace[1:(1+i)],
        converge=converge)
}

> fs.geom (0.4, 1.5)
$theta
[1] 0.6

$niter
[1] 6

$theta.trace
       0        1        2        3        4        5        6
0.400000 0.700000 0.625000 0.601563 0.600006 0.600000 0.600000

$converge
```

```
[1] TRUE
```

▶

Proof of Theorem 1.1 (Optional). We will fist show that $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ (in probability), i.e. that the MLE is a *consistent* estimator of $\boldsymbol{\theta}$. To keep the proof simple, we will assume that $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_*$ (ignoring questions about convergence of $\hat{\boldsymbol{\theta}}$), and we will show that $\boldsymbol{\theta}_* = \boldsymbol{\theta}$, the true parameter value.

To that end, suppose $\boldsymbol{\theta}_* \neq \boldsymbol{\theta}$, and let $\ell(\boldsymbol{\theta}|\boldsymbol{y})$ denote the log-likelihood evaluated at the true value $\boldsymbol{\theta}$ and $\ell(\boldsymbol{\theta}_*|\boldsymbol{y})$ denote the log-likelihood evaluated at the limit $\boldsymbol{\theta}_*$. Then

$$\begin{aligned}
\frac{1}{n}(\ell(\boldsymbol{\theta}|\boldsymbol{y}) - \ell(\boldsymbol{\theta}_*|\boldsymbol{y})) &= \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n} \log f(y_i|\boldsymbol{\theta}_*) \\
&\to \mathsf{E} \log f(y|\boldsymbol{\theta}) - \mathsf{E} \log f(y|\boldsymbol{\theta}_*) \text{ (by the law of large numbers)} \\
&= -\mathsf{E} \log \left\{ \frac{f(y|\boldsymbol{\theta}_*)}{f(y|\boldsymbol{\theta})} \right\} \\
&> -\log \mathsf{E} \left\{ \frac{f(y|\boldsymbol{\theta}_*)}{f(y|\boldsymbol{\theta})} \right\} \text{ (by Jensen's inequality and the assumption } \boldsymbol{\theta}_* \neq \boldsymbol{\theta}) \\
&= -\log \int \frac{f(y|\boldsymbol{\theta}_*)}{f(y|\boldsymbol{\theta})} f(y|\boldsymbol{\theta}) \, \mathrm{d}y \\
&= -\log \int f(y|\boldsymbol{\theta}_*) \, \mathrm{d}y \\
&= -\log(1) = 0.
\end{aligned}$$

So, $\lim \ell(\boldsymbol{\theta}|\boldsymbol{y}) - \ell(\boldsymbol{\theta}_*|\boldsymbol{y}) > 0$. However, since $\hat{\boldsymbol{\theta}}$ is the MLE, $\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) > \ell(\boldsymbol{\theta}|\boldsymbol{y})$ so $\lim \ell(\boldsymbol{\theta}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \leq 0$ which is a contradiction. Therefore $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$.

The next step is to use the mean value theorem and the central limit theorem to derive the asymptotic normality of the MLE. Consider the score function evaluated at the MLE and apply the mean value theorem around the true value $\boldsymbol{\theta}$:

$$\begin{aligned}
0 = \boldsymbol{u}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) &= \boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) + \boldsymbol{u}'(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\
&\qquad \text{(for some } \tilde{\boldsymbol{\theta}} \text{ between } \hat{\boldsymbol{\theta}} \text{ and } \boldsymbol{\theta}) \\
\Rightarrow 0 &= \boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) - \mathcal{J}(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\
\Rightarrow \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} &= \mathcal{J}(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})^{-1} \boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) \\
&= \left\{ \frac{1}{n} \mathcal{J}(\tilde{\boldsymbol{\theta}}|\boldsymbol{y}) \right\}^{-1} \left\{ \frac{1}{n} \boldsymbol{u}(\boldsymbol{\theta}|\boldsymbol{y}) \right\}.
\end{aligned}$$

Note that the first term converges to $\mathcal{I}(\boldsymbol{\theta})^{-1}$ by the law of large numbers and the fact that $\tilde{\boldsymbol{\theta}} \to \boldsymbol{\theta}$, and the second term converges, by the central limit theorem, to a random variable Z with distribution $Z \sim \mathrm{N}_p(0, \mathcal{I}(\boldsymbol{\theta}))$. Putting these together, gives the desired result. □

1.4 Likelihood ratio test

Suppose that the parameter $\boldsymbol{\theta}$ is a p-dimensional vector, i.e. $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ and the parameter space is $\Theta \subseteq \mathbb{R}^p$ and we wish to perform a hypothesis test on a subset of the parameter vector $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_q)$ where $q \leq p$. In other words we wish to test

$$H_0 : \boldsymbol{\theta}' = \boldsymbol{\theta}'_0 \text{ vs } H_1 : \boldsymbol{\theta}' \neq \boldsymbol{\theta}'_0. \tag{1.7}$$

Under the null hypothesis the parameter space is restricted to say $\Theta' = \{\boldsymbol{\theta} \in \Theta : \boldsymbol{\theta}' = \boldsymbol{\theta}'_0\} \subset \Theta$ and we need to estimate the remaining $p' = p - q$ parameters. For example in the case of the normal distribution of Example 1.3, $\boldsymbol{\theta} = (\mu, \sigma^2)$, $p = 2$ and $\Theta = (-\infty, \infty) \times (0, \infty)$. Under the null hypothesis $H_0 : \mu = 0$, or equivalently $H_0 : \theta_1 = 0$, we have $\boldsymbol{\theta}' = (\theta_1)$, $q = 1$ (as μ is fixed under H_0), $p' = 1$ (as we only need to estimate σ^2 under H_0), and $\Theta' = \{0\} \times (0, \infty)$.

Now let $\hat{\boldsymbol{\theta}}'$ and $\hat{\boldsymbol{\theta}}$ be the MLE under the null and alternative hypotheses respectively. So

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}}\, L(\boldsymbol{\theta}|\boldsymbol{y})$$
$$\hat{\boldsymbol{\theta}}' = \underset{\boldsymbol{\theta} \in \Theta'}{\operatorname{argmax}}\, L(\boldsymbol{\theta}|\boldsymbol{y}).$$

It is clear that $L(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \geq L(\hat{\boldsymbol{\theta}}'|\boldsymbol{y})$, however if $L(\hat{\boldsymbol{\theta}}'|\boldsymbol{y})$ is relatively close to $L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ that would be evidence for the null hypothesis. The statistic that allows us to perform the hypothesis test (1.7) is called the *log-likelihood-ratio statistic* and is given by

$$\Lambda := 2 \log \frac{L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})}{L(\hat{\boldsymbol{\theta}}'|\boldsymbol{y})} = 2(\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}'|\boldsymbol{y})).$$

The asymptotic distribution of the log-likelihood-ratio statistic is given in the following theorem.

Theorem 1.2 (Asymptotic distribution of the log-likelihood-ratio statistic)**.**
For an independent sample $\boldsymbol{y}$ of size n and under the null hypothesis (1.7), assuming that $\hat{\boldsymbol{\theta}}'$ is in the interior of the parameter space Θ, the distribution of the log-likelihood-ratio statistic as $n \to \infty$ is approximately the chi-square distribution with $q = p - p'$ degrees of freedom, i.e.

$$\Lambda \dot{\sim} \mathcal{X}^2_{p-p'}.$$

The importance of Theorem 1.2 is that it allows us to quantify how large should the value of the log-likelihood-ratio test statistic be in order to reject the H_0. In other words we reject the H_0 at level α if

$$\Lambda > \mathcal{X}^2_{p-p';1-\alpha}.$$

Example 1.7. The following random sample is an i.i.d. sample from the normal distribution with unknown mean (μ) and variance (σ^2) parameters

$$y_1 = 0.93 \quad y_2 = 1.31 \quad y_3 = -0.45 \quad y_4 = 0.94 \quad y_5 = 0.88 \quad y_6 = 1.85$$

We wish to test the hypothesis

$$H_0 : \mu = 0 \text{ v.s. } H_1 : \mu \neq 0$$

at level $\alpha = 5\%$.

The MLE under the alternative hypothesis gives

$$\hat{\mu} = \bar{y} = 0.91$$
$$\hat{\sigma}^2 = (1/n) \sum (y_i - \hat{\mu})^2 = 0.48$$
$$L(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = (2\pi\hat{\sigma}^2)^{-n/2} \exp\left\{-\frac{1}{2\hat{\sigma}^2} \sum (y_i - \hat{\mu})^2\right\} = \exp(-6.33)$$

and under the null hypothesis

$$\begin{aligned}\hat{\mu} &= 0\\ \hat{\sigma}^2 &= (1/n)\sum(y_i-\hat{\mu})^2 = 1.31\\ L(\hat{\boldsymbol{\theta}}'|\boldsymbol{y}) &= (2\pi\hat{\sigma}^2)^{-n/2}\exp\left\{-\frac{1}{2\hat{\sigma}^2}\sum(y_i-\hat{\mu})^2\right\} = \exp(-9.33)\end{aligned}$$

The log-likelihood-ratio statistic is $\Lambda = 2\times(-6.33+9.33) = 6.00$. The critical value of the $\mathcal{X}_1^2$ distribution is 3.84 therefore the conclusion is that there is enough evidence to reject the null hypothesis. ▶

Proof of Theorem 1.2 (Optional). Apply the mean value theorem to the log-likelihood evaluated at $\hat{\boldsymbol{\theta}}'$ around $\hat{\boldsymbol{\theta}}$:

$$\begin{aligned}\ell(\hat{\boldsymbol{\theta}}'|\boldsymbol{y}) &= \ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + \boldsymbol{u}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})^\intercal(\hat{\boldsymbol{\theta}}'-\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\hat{\boldsymbol{\theta}}'-\hat{\boldsymbol{\theta}})^\intercal\mathcal{J}(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}'-\hat{\boldsymbol{\theta}})\\ &\qquad\text{(for some } \tilde{\boldsymbol{\theta}} \text{ between } \hat{\boldsymbol{\theta}}' \text{ and } \hat{\boldsymbol{\theta}})\\ &= \ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \frac{1}{2}(\hat{\boldsymbol{\theta}}'-\hat{\boldsymbol{\theta}})^\intercal\mathcal{J}(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}'-\hat{\boldsymbol{\theta}})\\ &\qquad\text{(because } \boldsymbol{u}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = 0)\\ \Rightarrow 2(\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}'|\boldsymbol{y})) &= (\hat{\boldsymbol{\theta}}'-\hat{\boldsymbol{\theta}})^\intercal\mathcal{J}(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}'-\hat{\boldsymbol{\theta}})\\ &= (\hat{\boldsymbol{\theta}}'_{1:q}-\hat{\boldsymbol{\theta}}_{1:q})^\intercal\mathcal{J}_{1:q,1:q}(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}'_{1:q}-\hat{\boldsymbol{\theta}}_{1:q})\\ &\quad+(\hat{\boldsymbol{\theta}}'_{1:q}-\hat{\boldsymbol{\theta}}_{1:q})^\intercal\mathcal{J}_{1:q,q+1:p}(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}'_{q+1:p}-\hat{\boldsymbol{\theta}}_{q+1:p})\\ &\quad+(\hat{\boldsymbol{\theta}}'_{q+1:p}-\hat{\boldsymbol{\theta}}_{q+1:p})^\intercal\mathcal{J}_{q+1:p,1:q}(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}'_{1:q}-\hat{\boldsymbol{\theta}}_{1:q})\\ &\quad+(\hat{\boldsymbol{\theta}}'_{q+1:p}-\hat{\boldsymbol{\theta}}_{q+1:p})^\intercal\mathcal{J}_{q+1:p,q+1:p}(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}'_{q+1:p}-\hat{\boldsymbol{\theta}}_{q+1:p}),\end{aligned}$$

where we partitioned each vector (and matrix) into two blocks, where the first block contains its 1 up to q elements (rows and columns), and the second the rest.

Because of the assumption that the null hypothesis is true, $\hat{\boldsymbol{\theta}}'_{1:q} = \boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}'_{q+1:p} - \hat{\boldsymbol{\theta}}_{q+1:p}$ is negligible compared to $\hat{\boldsymbol{\theta}}_{1:q} - \hat{\boldsymbol{\theta}}'_{1:q} = \hat{\boldsymbol{\theta}}_{1:q} - \boldsymbol{\theta}_0$. Therefore, the last three terms in the expansion become negligible. The first term can be shown that it converges to the $\mathcal{X}_q^2$ distribution using Theorem 1.1 and the fact that $\mathcal{J}_{1:q,1:q}(\tilde{\boldsymbol{\theta}}|\boldsymbol{y}) \to \mathcal{I}_{1:q,1:q}(\boldsymbol{\theta})$. □

1.5 Inference for the transformation of a parameter

Sometimes the parameter of interest is a function $\varphi := \varphi(\theta)$ of θ. If $\varphi(\cdot)$ is one-to-one then the invariance property of the MLE says that the MLE for φ is $\hat{\varphi} := \varphi(\hat{\theta})$.

1.5.1 The delta method (Optional)

Sometimes we wish to know the distribution of a transformation of a random variable. For example we know from Theorem 1.1 that for the MLE of θ, $\hat{\theta} \dot{\sim} N(\theta, \mathcal{I}(\theta)^{-1})$ but we wish to estimate a function $\varphi = \varphi(\theta)$ of θ. According to the invariance property of the MLE, the MLE of φ is $\hat{\varphi} = \varphi(\hat{\theta})$ but in order to be able to construct confidence intervals and perform hypothesis tests we need to know the distribution of $\hat{\varphi}$. The delta method provides a solution to this problem.

Theorem 1.3 (Delta method)**.**
Suppose that $\hat{\theta} \dot{\sim} N(\theta, \sigma^2)$. Then for a continuous differentiable function $\varphi(\hat{\theta})$, asymptotically,

$$\varphi(\hat{\theta}) \dot{\sim} N\left(\varphi(\theta), \sigma^2 \varphi'(\theta)^2\right).$$

Proof. By the mean value theorem,

$$\varphi(\hat{\theta}) = \varphi(\theta) + \varphi'(\theta^*)\,(\hat{\theta} - \theta),$$

for some θ^* between θ and $\hat{\theta}$. However, since $\hat{\theta} \to \theta \Rightarrow \theta^* \to \theta \Rightarrow \varphi'(\theta^*) \to \varphi'(\theta)$. Therefore $\varphi(\hat{\theta}) \to \varphi(\theta) + \varphi'(\theta)\, N(0, \sigma^2)$ which proves the theorem. □

Example 1.8. Refer to Example 1.5. Suppose that we are interested in estimating the parameter $\varphi = 1/\theta$ which is interpreted as how long we have to wait until the next customer enters the queue. In Example 1.5 we found that $\hat{\theta} \dot{\sim} N(\theta, 0.16)$ and $\hat{\theta} = 1.6$. Then $\hat{\varphi} = 1/1.6 = 0.625$, so a new customer arrives at the queue every 0.625 minutes and $\frac{\mathrm{d}}{\mathrm{d}\theta}\varphi(\hat{\theta}) = -1/(1.6)^2 = -0.39$. Therefore $\mathsf{Var}(\hat{\varphi}) = (0.16)(0.39)^2 = 0.024$ and a 95% confidence interval for φ is $0.625 \pm (1.96)\sqrt{0.024} = (0.469, 0.781)$. ▶

1.5.2 Transformed confidence intervals

The disadvantage of the delta method is that it might not be accurate if the sample size is not large and therefore the coverage of the confidence interval produced by it might be far from the desired level. If the transformation $\varphi(\cdot)$ is monotone then a better way of obtaining confidence intervals for φ is by transforming the limits of the confidence interval for θ. In other words if (θ_L, θ_U) is a level $1-\alpha$ confidence interval for θ and $\varphi(\cdot)$ is increasing then $(\varphi(\theta_L), \varphi(\theta_U))$ is a level $1-\alpha$ confidence interval for φ. This is based on the result that $\Pr(\theta_L < \theta < \theta_U) = 1 - \alpha \Leftrightarrow \Pr(\varphi(\theta_L) < \varphi(\theta) < \varphi(\theta_U)) = 1 - \alpha$.

Similarly if $\varphi(\cdot)$ is decreasing, the transformed confidence interval is $(\varphi(\theta_U), \varphi(\theta_L))$.

Example 1.9 (Example 1.8 continued)**.** In Example 1.5 we found that a 95% confidence interval for θ is $(0.816, 2.384)$. The transformation $\varphi = 1/\theta$ is monotonically decreasing so a 95% confidence interval for φ is $(1/2.384, 1/0.816) = (0.419, 1.225)$. ▶

1.6 A word of caution (Optional)

Although this chapter encourages the use of MLE for parameter estimation, we should be aware that it is not a panacea, and care should be taken when applied. In 1948, statisticians Jerzy Neyman and Elizabeth L. Scott published a famous paper in the journal *Econometrica*, titled "Consistent estimates based on partially consistent observations", where they showed that the MLE can be inconsistent in certain cases, i.e., it does not converge to the true parameter value as the sample size grows to infinity. The paper shocked many statisticians at the time, who used to think that the MLE is always well-behaved and "best", and inspired further research into the theory of maximum likelihood method. The following example is a simplified version of an example from the above paper, which later became known as the Neyman-Scott paradox.

Example 1.10. Suppose we have a sample of n individuals and we measure (subject to random noise) the same quantity twice for each individual. Let x_i, y_i denote the two measurements from the ith individual. We assume independence between the measurements and that

$$x_i, y_i \overset{\text{ind}}{\sim} \mathrm{N}(\mu_i, \sigma^2),\ i = 1, \ldots, n,$$

where $\mu_1, \ldots, \mu_n$ and σ^2 are unknown parameters. It is easy to see that the MLE for μ_i is $\hat{\mu}_i = (x_i + y_i)/2$, i.e., the average of the two corresponding measurements, which implies that

$$x_i - \hat{\mu}_i = \hat{\mu}_i - y_i = \frac{x_i - y_i}{2} = \frac{z_i}{2}, \tag{1.8}$$

where $z_i = x_i - y_i$. By the properties of the normal distribution, $z_i \overset{\text{ind}}{\sim} N(0, 2\sigma^2)$, $i = 1, \ldots, n$.

The log-likelihood is

$$\begin{aligned}\ell(\mu_1, \ldots, \mu_n, \sigma^2 | \boldsymbol{x}, \boldsymbol{y}) &= \sum_{i=1}^{n} \Big\{ \log f(x_i; \mu_i, \sigma^2) + \log f(y_i; \mu_i, \sigma^2) \Big\} \\ &= -n \log(2\pi) - n \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \Big\{ (x_i - \mu_i)^2 + (y_i - \mu_i)^2 \Big\}.\end{aligned}$$

After plugging-in the MLE for μ_i and applying (1.8) in the last term, we have,

$$\begin{aligned}\ell(\sigma^2 | \boldsymbol{x}, \boldsymbol{y}) &= -n \log(2\pi) - n \log(\sigma^2) - \frac{1}{4\sigma^2} \sum_{i=1}^{n} z_i^2. \\ \Rightarrow u(\sigma^2 | \boldsymbol{x}, \boldsymbol{y}) &= -\frac{n}{\sigma^2} + \frac{1}{4(\sigma^2)^2} \sum_{i=1}^{n} z_i^2. \\ \Rightarrow \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} \frac{z_i^2}{4}.\end{aligned}$$

The law of large numbers says that, as $n \to \infty$, $\hat{\sigma}^2$ converges to the common mean of $z_i^2/4$, i.e.,

$$\hat{\sigma}^2 \to \mathsf{E}\, \frac{z_i^2}{4} = \frac{\mathsf{Var}\, z_i}{4} = \frac{2\sigma^2}{4} = \frac{\sigma^2}{2} \neq \sigma^2.$$

Therefore, $\hat{\sigma}^2$ is inconsistent. ▶

The problem with Example 1.10 is that there are too many parameters, $n + 1$ to be precise. The number of parameters grows linearly with the sample size instead of staying fixed, so there aren't enough data to reduce the finite-sample bias of the MLE.

Biographies

Jerzy Neyman Neyman is one of the most prominent statisticians of the last century. Through his collaboration with Karl Pearson the 1930's, he has developed a new paradigm of statistical inference, culminated by the famous Neyman-Pearson lemma used in hypothesis testing. The use of confidence intervals for parameter estimation was also his idea. Neyman was also the founder of the Statistics Department at the University of California, Berkeley and was instrumental in establishing Berkeley as a major statistical centre. In 1963, he made extended visits to the southern USA, to give lectures at various universities and witnessed the segregation of black and white communities. On his return to Berkeley, he established a special scholarship programme to help prepare talented underprivileged young people for university. Among his accolades include being member of the US National Academy of Sciences (1963), the Polish National Academy (1966), and the Royal Society (1979), as well as receiving the Royal Statistical Society's Guy Medal in Gold (1966) and the US National Medal of Science (1969).

Elizabeth L. Scott Scott studied mathematics and astronomy and later worked as a teaching assistant of astronomy at a time where there were not many opportunities for women in the field. Indeed, women were often forbidden from using the telescopes themselves. Frustrated by this, she decided to switch fields and study for a PhD in mathematics under the supervision of Jerzy Neyman, which she obtained in 1949. She worked in the areas of experimental design, distribution theory, and medical statistics. She was a life-long active member of the Women's Faculty Club at Berkeley and developed an interest in women's issues that would encompass much of her later work. The Elizabeth L. Scott Award was established in her honour to recognise an individual's efforts to further the careers of women in academia. The award is granted to an individual, male or female, who has helped foster opportunities in statistics for women.

2. Linear Models

This chapter presents some useful material from the linear models theory. It also introduces the statistical package R and how to use it to define and fit models.

2.1 Linear predictors

Statistical models are used to explain the variability in the values of one variable, the *response* or *dependent* variable (symbolised by y), in terms of the values of p other variables, the *explanatory* or *regressor*, or *predictor* variables (symbolised by $\boldsymbol{x} = (x_1, \ldots, x_p)$).

The simple linear model with one explanatory variable is defined as

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

The term $\eta := \beta_0 + \beta_1 x$ is the linear combination of the explanatory variables. Let $\boldsymbol{x} = (1, x)$, $\boldsymbol{\beta} = (\beta_0, \beta_1)$. Then, $\eta = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$.

In general we consider p explanatory variables combined linearly in the form

$$\eta := \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} + \beta_p x_p, \tag{2.1}$$

where $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p)$ are unknown parameters called the *regressor coefficients*. The term (2.1), henceforth written in the more compact form $\eta = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$, is called the *linear predictor*.

If $x_p \equiv 1$ always then the last term in the linear predictor (2.1) becomes just $\beta_p x_p =: \beta_0$, in which case β_0 is called the *intercept*. In this special case the linear predictor is written as

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1}. \tag{2.2}$$

Note that there are still p explanatory variables in (2.2), which are the $p-1$ variables $x_1, \ldots, x_{p-1}$, plus the one "variable" that is in fact a constant equal to 1 which is multiplied by β_0.

In general β_i is interpreted as the effect of the variable x_i in the linear predictor η. If, for some i, $\beta_i = 0$ then no matter what x_i is, its value has no effect on the linear predictor. In this case we call the regressor x_i *not significant*.

In practice, a sample $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$ is collected and inference about the value of the parameter $\boldsymbol{\beta}$ is sought. Consider the $n \times p$ matrix $\boldsymbol{X}$ with ith row $\boldsymbol{x}_i$, called the *design matrix*. Let $\boldsymbol{\eta}$ denote the vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$. Then the following relationship holds

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta},$$

thus a linear predictor can be defined in terms of the design matrix. The design matrix plays an important role in inference for any model.

2.1.1 Components of the linear predictor

Numerical explanatory variables

Consider first the simple case where there is only one explanatory variable x and that this variable is numerical. The linear predictor, based on the sample (x_i, y_i), $i = 1, \ldots, n$ is

$$\eta_i = \beta_0 + \beta_1 x_i, \ i = 1, \ldots, n,$$

so the parameters consist of an intercept and a slope: $\boldsymbol{\beta} = (\beta_0, \beta_1)$. Writing these equations explicitly as

$$\begin{aligned} \eta_1 &= \beta_0 + \beta_1 x_1 \\ \eta_2 &= \beta_0 + \beta_1 x_2 \\ &\vdots \\ \eta_n &= \beta_0 + \beta_1 x_n \end{aligned}$$

makes it easier to see that the design matrix has the form

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

To generalise this, suppose we have $p-1$ numerical explanatory variables $x_1, \ldots, x_{p-1}$ and we obtain a sample $(x_{i,1}, \ldots, x_{i,p-1}, y_i), i = 1, \ldots, n$. The linear predictor is

$$\eta_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1}, \ i = 1, \ldots, n,$$

the set of parameters is $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ and the design matrix has the form

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ 1 & x_{2,1} & \cdots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{bmatrix}.$$

Notice that with this notation $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$, where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$.

In summary, every numerical explanatory variable enters in the design matrix in a single column as is. The intercept is represented by a column of ones.

Example 2.1. The Open University Energy Research Group ran a constant-heating experiment in an unoccupied test house in Great Linford, Milton Keynes. The house was well-insulated, faces south and has most of its glazing on the south side. For a ten-week period from February to May 1982, the house was heated to a constant 21°C using thermostatically controlled electric heaters. Among other things, the researchers measured the following three variables.

E The electrical energy (kWh per day) required to keep the house at 21°C.

S Energy from solar radiation (kWh per square metre per day) falling on a south-facing vertical surface at the house site.

TD Temperature difference (°C) between the inside and outside of the house.

The data are shown in Table 2.1 correspond to the weekly averages of the measured variables. How is E determined by the other two variables?

	S	TD	E		S	TD	E
1	1.36	15.1	74.5	6	2.00	12.4	47.7
2	2.77	16.8	60.5	7	2.67	15.7	52.8
3	2.79	16.1	63.2	8	3.43	13.5	31.1
4	2.70	16.0	55.6	9	2.30	12.0	33.4
5	3.27	15.6	45.0	10	2.68	14.4	48.9

Table 2.1: Data for Example 2.1.

The sample size of this data set is $n = 10$. The response variable is the energy required E and there are two explanatory variables S and TD. We use y_i to symbolise the ith response variable in the sample, so $y_1 = 74.5$, $y_2 = 60.5$, ..., $y_{10} = 48.9$. Let x_1 denote the energy from solar radiation, S, and x_2 denote the temperature difference, TD. The linear predictor may have the form

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

The parameters β_1 and β_2 denote the effects of the two explanatory variables while the parameter β_0 is the intercept. This notation follows (2.2) with $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. The vector of explanatory variables is then $\boldsymbol{x}_1 = (1, 1.36, 15.1)$, $\boldsymbol{x}_2 = (1, 2.77, 16.8)$, ..., $\boldsymbol{x}_{10} = (1, 2.68, 14.4)$. The design matrix is the matrix with rows $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{10}$, i.e.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1.36 & 15.1 \\ 1 & 2.77 & 16.8 \\ \vdots & \vdots & \vdots \\ 1 & 2.68 & 14.4 \end{bmatrix}$$

▶

Factors

Factors are categorical explanatory variables, e.g. sex, nationality, etc. The categories in each factor are called the *levels*. The analysis of variance (ANOVA) is concerned with modelling against factors. The simple ANOVA model for modelling observations $\{y_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, m\}$ against a factor of m levels is

$$y_{ij} = \mu_j + \varepsilon_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, m.$$

The parameter μ_j corresponds to the mean at the jth level. There are m such means that need to be estimated.

An alternative way to write the model is to set $\beta_0 = \mu_1$ and define, for $j = 1, \ldots, m$, $\alpha_j = \mu_j - \mu_1$. With this notation, β_0 corresponds to the mean at the first level while each α_j

corresponds to the average *increase* at level j compared to level 1. Obviously, $\alpha_1 = \mu_1 - \mu_1 = 0$ so the unknown parameters under this parametrisation are still m in size, $\boldsymbol{\beta} = (\beta_0, \alpha_2, \ldots, \alpha_m)$.

The linear predictor for this model is

$$\eta_{ij} = \begin{cases} \beta_0 & \text{if } j = 1 \\ \beta_0 + \alpha_j & \text{if } j = 2, \ldots, m \end{cases} \tag{2.3}$$

Now let $\boldsymbol{\eta}$ denote the vector constructed by stacking each component of the linear predictor one below the other, i.e. $\boldsymbol{\eta} = (\eta_{11}, \ldots, \eta_{n_1 1}, \eta_{12}, \ldots, \eta_{n_2 2}, \ldots, \eta_{1m}, \ldots, \eta_{n_m m})$. Then (2.3) may be written in the form

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$$

where

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \Big\} n_1 \text{ times}$$
$$\begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \end{bmatrix} \Big\} n_2 \text{ times}$$
$$\begin{bmatrix} 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \end{bmatrix} \Big\} n_3 \text{ times}$$
$$\begin{bmatrix} \vdots & \vdots & \vdots & & \vdots \end{bmatrix}$$
$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix} \Big\} n_m \text{ times}$$

Assuming an intercept term is present in the linear predictor, a factor of m levels introduces into the model $m - 1$ parameters and $m - 1$ columns in the design matrix (not counting the intercept). Each column indicates whether the corresponding level is present at each observation with the value 1 if present and the value 0 if not.

Example 2.2. To compare the effects of two fertilisers on the yield of tomatoes a gardener randomly selected tomato plants to treat with one of two types of fertiliser. The corresponding tomato yields are shown in the table below. The question is whether fertiliser B (a supposedly improved fertiliser) yields more than fertiliser A (the standard one).

Fertiliser A	Fertiliser B
29.9	26.6
11.4	23.7
25.3	28.5
16.5	14.2
21.1	17.9
	24.3

Table 2.2: Observed yields of tomato plants using two different fertilisers.

Let β_0 denote the average yield when fertiliser A is used. If fertiliser B is used instead, then the average yield is expected to increase by an amount α_B (a negative value indicates a

decrease). In other words α_B denotes the average effect of fertiliser B *in comparison* to the average effect of fertiliser A. Therefore the average yield for fertiliser B is given by $\beta_0 + \alpha_B$.

Note that the data in Table 2.2 may be presented alternatively by stacking the second column of Table 2.2 under the first. This is illustrated in Table 2.3.

Fertiliser	Yield
A	29.9
A	11.4
A	25.3
A	16.5
A	21.1
B	26.6
B	23.7
B	28.5
B	14.2
B	17.9
B	24.3

Table 2.3: Alternative representation of the data in Table 2.2.

Let $\boldsymbol{y} = (29.9, 11.4, \ldots, 24.3)$ denote the observed yields, in Table 2.3. The number of observations is $n = 11$ and the number of parameters is $p = 2$ with $\boldsymbol{\beta} = (\beta_0, \alpha_B)$. The linear predictor is

$$\eta_{ij} = \begin{cases} \beta_0 & \text{for } j = A,\ i = 1, \ldots, 5 \\ \beta_0 + \alpha_B & \text{for } j = B,\ i = 1, \ldots, 6 \end{cases}.$$

Therefore, to express the linear predictor in the form $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ we must choose

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_{1A} \\ \vdots \\ \eta_{5A} \\ \eta_{1B} \\ \vdots \\ \eta_{6B} \end{bmatrix} \quad \text{and} \quad \boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} 5 \text{ times} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} 6 \text{ times} \end{matrix}$$

▶

Interactions

When the effect of one explanatory variable depends on the value of another explanatory variables, we say that the two variables interact.

In the simplest model, the effect of an explanatory variable is independent of the values of the other variables. If that's not the case, we can consider adding interactions to the model. We can have an interaction between a numerical and a factor variable or between two factor variables. The interactions imply that the effect of the first variable (numerical or factor) depends on the level of the second variable.

• *Interaction between a numerical and a factor variable*

Suppose that the terms in the linear predictor are

$$\eta_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_j, \qquad \alpha_1 = 0,$$

for $i = 1, \ldots, n_j$, $j = 1, \ldots, m$. The parameter β_1 denotes the increase in the linear predictor when x is increased by one unit at a fixed level j, *no matter* the level of the factor variable. The interaction means that if β_1 is the effect when $j = 1$, the effect for $j = 2$ is something else, $\beta_1 + \beta_2^*$ say, and for $j = 3$ it is $\beta_1 + \beta_3^*$, and so on. In other words

$$\eta_{ij} = \begin{cases} \beta_0 + \beta_1 x_{ij} + \alpha_j & \text{if } j = 1, \\ \beta_0 + (\beta_1 + \beta_2^*) x_{ij} + \alpha_j & \text{if } j = 2, \\ \quad \vdots & \\ \beta_0 + (\beta_1 + \beta_m^*) x_{ij} + \alpha_j & \text{if } j = m, \end{cases}$$

which can be written in one equation as

$$\eta_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_j + \beta_j^* x_{ij},$$
$$\alpha_1 = \beta_1^* = 0.$$

The number of parameters associated with the interaction are therefore $m-1$: $\beta_2^*, \ldots, \beta_m^*$. Each β_j^* denotes the increase in the effect of the numerical variable for level j compared to the effect for level 1.

To create the design matrix, note that the linear predictor can be expanded to contain all parameters as

$$\begin{aligned} \eta_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_j \\ + \beta_2^* \times 0 + \ldots + \beta_{j-1}^* \times 0 + \beta_j^* \times x_{ij} + \beta_{j+1}^* \times 0 + \ldots + \beta_m^* \times 0. \end{aligned}$$

This suggests the following rule.

Assuming an intercept term is present in the linear predictor, the interaction between a numerical variable and a factor variable of m levels corresponds to $m-1$ parameters in the linear predictor (not counting the intercept). In the design matrix, it is represented by $m-1$ columns (not counting the intercept). Each column is created by multiplying the column corresponding to the numerical variable with each of the $m-1$ columns corresponding to the different levels of the factor variable.

• *Interaction between two factor variables*

Suppose that the terms in the linear predictor are

$$\eta_{ijk} = \beta_0 + \alpha_j^1 + \alpha_k^2, \qquad \alpha_1^1 = \alpha_1^2 = 0,$$

for $i = 1, \ldots, n_{jk}$, $j = 1, \ldots, m_1$, $k = 1, \ldots, m_2$. The parameter α_k^2 denotes the increase in the linear predictor for level k of the second factor compared to the first level of the second factor. This increase is the same *no matter* what the first factor level is. Including the interaction between the two factors removes this constraint. Therefore we need to estimate a separate effect for each level k. We can let α_k^2 be this effect when the first factor level is 1 and $\alpha_k^2 + \alpha_{jk}^{12}$ when the first factor level is $j = 2, \ldots, m_1$, and we have to do this for $k = 2, \ldots, m_2$. This brings $(m_1 - 1) \times (m_2 - 1)$ parameters α_{jk}^{12}.

The model for the linear predictor can be written as

$$\eta_{ij} = \begin{cases} \beta_0 & \text{if } j = 1 \text{ and } k = 1, \\ \beta_0 + \alpha_j^1 & \text{if } j = 2, \ldots, m_1 \text{ and } k = 1, \\ \beta_0 + \alpha_k^2 & \text{if } j = 1 \text{ and } k = 2, \ldots, m_2, \\ \beta_0 + \alpha_j^1 + \alpha_k^2 + \alpha_{jk}^{12} & \text{if } j = 2, \ldots, m_1 \text{ and } k = 2, \ldots, m_2, \end{cases}$$

or in one equation as

$$\eta_{ijk} = \beta_0 + \alpha_j^1 + \alpha_k^2 + \alpha_{jk}^{12}, \qquad \alpha_1^1 = \alpha_1^2 = \alpha_{1k}^{12} = \alpha_{j1}^{12} = 0,$$

which can be expanded to include all parameters as before. We can conclude the following.

Assuming an intercept term is present in the linear predictor, the interaction between one factor variable of m_1 levels and another factor variable of m_2 levels corresponds to $(m_1 - 1) \times (m_2 - 1)$ parameters in the linear predictor (not counting the intercept). In the design matrix, it is associated with $(m_1 - 1) \times (m_2 - 1)$ parameters (not counting the intercept). Each column is created by multiplying one column from the first variable with a column from the second variable.

Example 2.3. The times taken to cycle up a hill (in seconds), as well as the bicycle seat height (Low or High), use of dynamo, and tyre pressure (in psi) were recorded and shown in Table 2.4.

	Seat	Dynamo	Tyre	Time		Seat	Dynamo	Tyre	Time
1	Low	No	40	51	9	High	No	40	41
2	Low	No	45	54	10	High	No	45	43
3	Low	No	50	50	11	High	No	50	39
4	Low	No	55	48	12	High	No	55	39
5	Low	Yes	40	54	13	High	Yes	40	44
6	Low	Yes	45	60	14	High	Yes	45	43
7	Low	Yes	50	53	15	High	Yes	50	41
8	Low	Yes	55	51	16	High	Yes	55	44

Table 2.4: Data for example 2.3.

We are interested in modelling the time taken against the seat height, use of dynamo, and tyre pressure. In this example time is the response variable, observed $n = 16$ times with values $y_1 = 51, y_2 = 54, \ldots, y_{16} = 44$. The model parameters consist of an intercept, the effect of seat height, the effect of dynamo, and the effect of tyre pressure. Seat and Dynamo are factors while Tyre is numerical.

For Seat = Low, Dynamo = No, and Tyre = x the average time taken to cycle up the hill is modelled as $\eta = \beta_0 + \beta_1 x$. If Seat = High instead of Low the average time is shifted by an amount α_H^S, and if Dynamo = Yes instead of No the average time is shifted by an amount α_Y^D. This suggests a model for the linear predictor

$$\begin{gathered} \eta_{ijk} = \beta_0 + \beta_1 x_{ijk} + \alpha_j^S + \alpha_k^D, \\ \alpha_L^S = 0, \ \alpha_N^D = 0, \\ i = 1, 2, 3, 4, \ j = L, H, \ k = N, Y, \end{gathered} \tag{2.4}$$

where $x_{1LN} = 40$, $x_{2LN} = 45$, $x_{3LN} = 50$, and so on.

The unknown parameters of the model defined in (2.4) are $\boldsymbol{\beta} = (\beta_0, \beta_1, \alpha_H^S, \alpha_Y^D)$ and the

design matrix is

$$\boldsymbol{X} = \begin{bmatrix} 1 & 40 & 0 & 0 \\ 1 & 45 & 0 & 0 \\ 1 & 50 & 0 & 0 \\ 1 & 55 & 0 & 0 \\ 1 & 40 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 55 & 1 & 1 \end{bmatrix}$$

Suppose that we wish to test the claim that there is an interaction between Dynamo and Tyre. That is to say that if Dynamo = Yes the effect of Tyre is no longer β_1 but something else, $\beta_1+\beta_Y^D$ say. In other words, when Dynamo = Yes the linear predictor is $\eta = \beta_0+(\beta_1+\beta_Y^D)x+\alpha_j^S+\alpha_Y^D = \beta_0 + \beta_1 x + \alpha_j^S + \alpha_Y^D + \beta_Y^D x$. The new model is

$$\begin{gathered} \eta_{ijk} = \beta_0 + \beta_1 x_{ijk} + \alpha_j^S + \alpha_k^D + \beta_k^D x_{ijk}, \\ \alpha_L^S = 0,\ \alpha_N^D = 0,\ \beta_N^D = 0, \\ i = 1,2,3,4,\ j = L,H,\ k = N,Y, \end{gathered}$$

with unknown parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \alpha_H^S, \alpha_Y^D, \beta_Y^D)$ and design matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & 40 & 0 & 0 & 0 \\ 1 & 45 & 0 & 0 & 0 \\ 1 & 50 & 0 & 0 & 0 \\ 1 & 55 & 0 & 0 & 0 \\ 1 & 40 & 0 & 1 & 40 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 55 & 1 & 1 & 55 \end{bmatrix}$$

where the last column is derived as the product between the second and fourth columns.

Suppose instead, that we wish to account for the interaction between Dynamo and Seat. That is to say that if Dynamo = Yes, the effect of Seat is no longer α_j^S but something else, $\alpha_j^S + \alpha_{jY}^{SD}$ say. In other words, when Dynamo = Yes the linear predictor is $\eta = \beta_0+\beta_1 x+(\alpha_j^S+\alpha_{jY}^{SD})+\alpha_Y^D = \beta_0 + \beta_1 x + \alpha_j^S + \alpha_Y^D + \alpha_{jY}^{SD}$. The new model is

$$\begin{gathered} \eta_{ijk} = \beta_0 + \beta_1 x_{ijk} + \alpha_j^S + \alpha_k^D + + \alpha_{jk}^{SD}, \\ \alpha_L^S = 0,\ \alpha_N^D = 0,\ \alpha_{Lk}^{SD} = \alpha_{jN}^{SD} = 0, \\ i = 1,2,3,4,\ j = L,H,\ k = N,Y, \end{gathered}$$

with unknown parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \alpha_H^S, \alpha_Y^D, \alpha_{HY}^{SD})$ and design matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & 40 & 0 & 0 & 0 \\ 1 & 45 & 0 & 0 & 0 \\ 1 & 50 & 0 & 0 & 0 \\ 1 & 55 & 0 & 0 & 0 \\ 1 & 40 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 55 & 1 & 1 & 1 \end{bmatrix}$$

where the last column is derived as the product between the third and fourth columns. ▶

Offset variables

Sometimes we wish to include a numerical variable in the linear predictor where its coefficient is known to be 1, i.e. $\beta_1 = 1$. This often happens when our response variable is measured over different time lengths or different population sizes each time. Consider the following example.

Example 2.4. The sale price for different properties in Bath was recorded along with the property's type, number of bedrooms, and floor area, and is shown in Table 2.5. We wish to predict the property's price per square foot against its type and number of bedrooms.

Sale price (100,000's)	Type	Bedrooms	Floor Area (sq ft)
2.8	Detached	3	1121
3.1	Detached	4	1421
2.5	Semi-detached	3	1087
2.9	Semi-detached	4	1356
2.1	Semi-detached	2	887
2.0	Terraced	2	798
2.1	Terraced	2	870

Table 2.5: House price data for Example 2.4.

Let y denote the sale price, A denote the area, t the type, and x the number of bedrooms. Because price and area are always positive, one possible model is to use as a response variable the logarithmic transformation of price per square foot, i.e.

$$\begin{aligned} \log(y_{it}/A_{it}) = \beta_0 + \alpha_t + \beta_1 x_{it} + \varepsilon_{it}, \quad \alpha_D = 0, \\ i = 1, \ldots, n_t, \ t = D, S, T. \end{aligned} \tag{2.5}$$

Since $\log(y_{it}/A_{it}) = \log y_{it} - \log A_{it}$, the model in (2.5) is equivalent to

$$\begin{aligned} \log y_{it} = \beta_0 + \alpha_t + \beta_1 x_{it} + \log A_{it} + \varepsilon_{it}, \quad \alpha_D = 0, \\ i = 1, \ldots, n_t, \ t = D, S, T. \end{aligned} \tag{2.6}$$

Note that $\log A_{it}$ is a numerical variable appearing in the linear predictor in (2.6) without a coefficient. Such variable is called an *offset*.

Because the offset variable doesn't have any parameter associated with it, it doesn't appear in the design matrix. Therefore, the design matrix for the model in (2.6) is

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & 3 \\ 1 & 0 & 0 & 4 \\ 1 & 1 & 0 & 3 \\ 1 & 1 & 0 & 4 \\ 1 & 1 & 0 & 2 \\ 1 & 0 & 1 & 2 \\ 1 & 0 & 1 & 2 \end{bmatrix},$$

and the vector of parameters is $\boldsymbol{\beta} = (\beta_0, \alpha_S, \alpha_T, \beta_1)$. ▶

Definition 2.1 (Offset variable)**.**
An offset variable is the variable that appears in the right-hand side of the linear predictor with a known coefficient equal to 1.

Offset variables do not have any parameters, therefore they are not associated with any columns in the design matrix.

2.2 Modelling in R

2.2.1 Model formulation

A model formula in R has the form

```
response ~ linear_predictor
```

where the `response` corresponds to the response variable and `linear_predictor` specifies the explanatory variables that enter in the linear predictor (as a linear combination of the explanatory variables). The simple linear model with one explanatory variable, stored in the array `x` and one response variable, stored in the array `y` is simply expressed in the R language as

```
y ~ x
```

where the intercept is implied. The main expressions for constructing a formula in R are the following (this list is not exhaustive)

`-1` If we wish to fit a model without an intercept then the corresponding R expression becomes `y ~ x - 1` so the `-1` means that the default intercept term should not be included in the model.

`x1 + x2` To include more than one explanatory variables without interactions we separate them with a `+` sign. So the linear predictor $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ is expressed in the R language as `x1 + x2`. The variables may be numerical or categorical. To add more than two explanatory variables in the linear predictor we use more `+` signs: `x1 + x2 + x3 + x4`.

`x1:x2` The colon is used to denote the interaction between two variables. Note that the order in which the variables appear around the colon sign does not matter, i.e. `x1:x2` is equivalent to `x2:x1`. Interaction between more than two variables may be added, e.g. `x1:x2:x3` includes the interaction between three variables.

`x1*x2` The `*` is a shorthand notation for adding each variable separately into the linear predictor as well as their interaction. Therefore the expression `x1*x2` is equivalent to `x1 + x2 + x1:x2` and the expression `x1*x2*x3` is equivalent to `x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3`.

`(x1+x2+x3)^2` The `^k` where `k` is an integer denotes crossing to the `k`th degree. For example `(x1+x2+x3)^2` is equivalent to `x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3`. Note that if we want to include the power of a variable in the model, say x_1^2, we have to use the `I` function (see below).

`x1/x2` This is a shorthand notation for including the term before the `/` sign together with the interaction between the two terms. For example `x1/x2` is equivalent to `x1 + x1:x2`.

`-x1` In order to exclude variables from the model we use the - sign. This is useful when we want to exclude a term which was added automatically through the `*` expression or when we wish to update the model by excluding one variable. For example the expression `x1*x2 -x2` is equivalent to `x1 + x1:x2`.

offset(x1) Sometimes we wish to add a variable into the model without a parameter. This is done using the `offset` function. For example the linear predictor $\beta_0 + \beta_1 x_1 + x_2$ is represented in the R language by `x1 + offset(x2)`.

I(x1+x2) If we want to evaluate an R expression before it is included in the linear predictor, we use the function `I`. For example the linear predictor $\beta_0 + \beta_1 x_1 + \beta_2(x_2 + x_3)$ cannot be expressed as `x1 + (x2+x3)` because of the special meaning of the `+` sign in defining the model. The correct way to express this in the R language is by `x1 + I(x2+x3)`. The same applies for the `^` symbol. However, other arithmetic operations without special meaning in model definition need not be protected by the `I` function. For example the model $\beta_0 + \beta_1 x_1 + \beta_2 \log(x_2)$ can be simply expressed as `x1 + log(x2)`.

2.2.2 Model fitting

After we specify the model formula we can fit it to the data using one of R's functions. For example the function `lm` is used to fit a linear model via least squares. Another R function which we will use is `glm` which is used to fit a (generalised) linear model via Fisher scoring. A typical model fitting consists of the commands

```
mymodel <- response ~ linear_predictor
mylmfit <- lm(mymodel)   # For linear model fit
myglmfit <- glm(mymodel) # For generalised linear model fit
```

The first line specifies the model formula and stores it in the object `mymodel`. The second and third lines fit the formula using the data and store the results into the objects `mylmfit` and `myglmfit`. The model formula may be inputted directly into the fitting function, so one may simply write, e.g

```
mylmfit <- lm(response ~ linear_predictor) # For linear model fit
```

Once the model is fitted and stored into an R object, `myfit` say, the following functions are available for that object.

summary Produce and print a summary of the fit including parameter estimates along with their standard errors and p-values.

anova Produce and print the ANOVA table of the model.

formula The model formula in the form `response ~ linear_predictor`.

model.matrix The design matrix of the model. This function can be applied directly on a model formula as well.

family The assumed distribution of the response variable and the relationship between the left and right-hand sides of the model formula (i.e. the link function).

fitted or **fitted.values** Extracts the mean prediction for each observation of the response, i.e returns the estimate $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu} = \mathsf{E}\,\boldsymbol{y}$.

predict Obtains predictions for the fitted model. The second optional argument `newdata` should be a list with the new values of the explanatory variables. If omitted the current data are used.

coef or **coefficients** Extracts the parameter estimates from the model fit.

confint Confidence intervals for the model parameters. This has an optional argument `level` for the desired level of the confidence intervals which defaults to 0.95 if omitted.

vcov Computes the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$.

resid or **residuals** The residuals from the fit.

deviance Sum of squared residuals.

df.residuals Degrees of freedom of the sum of squared residuals.

logLik The maximum log-likelihood value along with its degrees of freedom.

nobs Extract the number of observations from the fitted model.

AIC and **BIC** Compute the AIC and BIC of the fitted model.

2.2.3 Updating and comparing models

Once a model is fitted we might wish to add new explanatory variables or remove some of the variables which were not significant and compare the new model with the old one. R provides commands for updating and comparing models. These are the following

update This command is used for updating and refitting a model. Its first input is the object containing the fit of the model we wish to update. The subsequent inputs indicate how the model is updated. The second input is usually the formula for the new model and that is used when we wish to add, remove or transform variables. When updating a formula, the dot (`.`), in either the left or right-hand sides of the `~` has a special meaning: it corresponds to the old model's left or right-hand side formula respectively.

For example the command

```
fit2 <- update(fit1, . ~ . - x1 + x3)
```

means "remove variable `x1` from the model stored in `fit1` and add the variable `x3`; refit the model and store it in the variable `fit2`".

Similarly the command

```
fit3 <- update(fit1, log(.) ~ .)
```

means "refit the model stored in `fit1` but with the log transformation of the response variable".

anova Produce and print the ANOVA table for comparing several models. The inputs to this command are the objects where the model fits are stored. Each subsequent model entered must be an extension of the previous in the sense that it contains the same explanatory variables as the previous model plus some more.

add1 Add one variable to the fitted model. The first input should be the object containing the fitted model and the second input a formula with the variables to be considered.

drop1 Similar to `add1` but drops instead of adding one variable to the fitted model.

step Repeatedly call `add1` and `drop1` to produce a better model.

Subject	Species	Temperature	Rest	Exercise
1	A	High	0.133	0.194
2	A	High	0.140	0.198
3	A	Low	0.107	0.152
4	A	Low	0.114	0.163
5	B	High	0.118	0.171
6	B	High	0.110	0.165
7	B	Low	0.098	0.144
8	B	Low	0.093	0.136
9	C	High	0.159	0.196
10	C	High	0.166	0.204
11	C	Low	0.126	0.182
12	C	Low	0.138	0.196
13	D	High	0.184	0.244
14	D	High	0.192	0.232
15	D	Low	0.154	0.207
16	D	Low	0.141	0.191

Table 2.6: Data for oxygen consumption of frogs

2.3 Example: Oxygen consumption of frogs

Frogs of four species had their oxygen consumption measured at two temperatures and two exercise levels. There were two frogs of each species at each temperature, and each of the two was measured both at rest and during forced exercise. The data are shown in Table 2.6.

There are four variables in the dataset. Variable *Species* is a factor with four levels, *Temperature* is also a factor with two levels, while *Rest* and *Exercise* are both numerical, measured at ml/g/hr.

We wish to predict oxygen consumption after exercise when the species, temperature and oxygen consumption at rest are given.

First we note a couple of things. Both *Rest* and *Exercise* are naturally positive quantities and evidently, *Rest* is smaller than *Exercise*. It probably makes more sense to model the logarithm of the relative increase in consumption after exercise from consumption at rest. Therefore, one may define the following variable

$$\text{RelativeIncrease} = \frac{\text{Exercise} - \text{Rest}}{\text{Rest}}$$

We let y denote the logarithm of the relative increase and let the index j correspond to the different species ($j = A, B, C, D$), the index k correspond to the temperature ($k = H, L$) and i correspond to the subject within each species and temperature ($i = 1, 2$). In other words for subject 5 we have $i = 1$, $j = B$, $k = H$, $y_{1BH} = \log\{(0.171 - 0.118)/0.118\} = -0.80$, while for subject 8 $i = 2$, $j = B$, $k = L$, $y_{2BL} = \log\{(0.136 - 0.093)/0.093\} = -0.77$. With this notation, a suitable linear model may be

$$\begin{aligned} y_{ijk} &= \beta_0 + \alpha_j^S + \alpha_k^T + \varepsilon_{ijk}, \\ \alpha_A^S &= \alpha_H^T = 0, \\ \varepsilon_{ijk} &\sim \mathrm{N}(0, \sigma^2) \end{aligned} \tag{M1}$$

where α_j^S denotes the average effect of the jth species and α_k^T the average effect of the kth temperature.

Fitting model (M1), we obtain the following estimates: $\hat{\beta}_0 = -0.95$, $\hat{\alpha}_B^S = 0.09$, $\hat{\alpha}_C^S = -0.31$, $\hat{\alpha}_D^S = -0.36$, $\hat{\alpha}_L^T = 0.22$. Here we take as a reference level to be *Species* = A and *Temperature* = High.

The parameter β_0 is interpreted as the average logarithm relative consumption at the reference level and each of the parameters α correspond to the average increase from the reference level when the corresponding species or temperature changes. To see this, let

$$\eta_{jk} = \beta_0 + \alpha_j^S + \alpha_k^T$$

be the average logarithm of the relative increase. Then, when $j = A$ and $k = H$, $\eta_{AH} = \beta_0$. Similarly, for any j and for $k = H$,

$$\eta_{jH} = \beta_0 + \alpha_j^S,$$

while if k changes to $k = L$,

$$\eta_{jL} = \beta_0 + \alpha_j^S + \alpha_L^T.$$

By subtracting one equation from the other we have $\eta_{jL} - \eta_{jH} = \alpha_L^T$ so α_L^T is the change in the average logarithm relative increase in consumption when the temperature changes from High to Low regardless of the species.

One may want to add the interaction between *Species* and *Temperature* into the model. If for example we believe that some species are more tolerant to changes in temperature than others, then the interaction term would make sense. To that end, an appropriate model would be

$$\begin{aligned} y_{ijk} &= \beta_0 + \alpha_j^S + \alpha_k^T + \alpha_{jk}^{ST} + \varepsilon_{ijk}, \\ \alpha_A^S = \alpha_H^T = \alpha_{Ak}^{ST} = \alpha_{jH}^{ST} &= 0, \\ \varepsilon_{ijk} &\sim \mathrm{N}(0, \sigma^2) \end{aligned} \tag{M2}$$

The interaction term adds $(4-1) \times (2-1) = 3$ new parameters in the model: α_{BL}^{ST}, α_{CL}^{ST} and α_{DL}^{ST}. The estimates for the parameters of (M2) are $\hat{\beta}_0 = -0.83$, $\hat{\alpha}_B^S = 0.08$, $\hat{\alpha}_C^S = -0.64$, $\hat{\alpha}_D^S = -0.51$, $\hat{\alpha}_L^T = -0.02$, $\hat{\alpha}_{BL}^{ST} = 0.01$, $\hat{\alpha}_{CL}^{ST} = 0.65$ and $\hat{\alpha}_{DL}^{ST} = 0.32$.

Suppose that we want to test for the significance of the interaction term. Then, testing for each of the three parameters individually would be incorrect since the tests do not necessarily hold simultaneously. The method is to use a likelihood-ratio test. In this case the distribution of the test statistic is not the chi-square but the F distribution with degrees of freedom 3 and 8 (this is something that we will discuss in Chapter 4). The value of the test statistic is 6.74 which corresponds to a p-value of 0.014. Therefore, we conclude that the interaction term is significant at the 1% level.

Let us now define a new variable

$$\text{ConsumptionDifference} = \text{Exercise} - \text{Rest}$$

which measures the change in consumption from rest to exercise. One may notice that the response of the original model (M1) may be written as

$$\begin{aligned} \log \text{RelativeIncrease} &= \log \frac{\text{Exercise} - \text{Rest}}{\text{Rest}} \\ &= \log(\text{Exercise} - \text{Rest}) - \log \text{Rest} \end{aligned}$$

$$= \log \text{ConsumptionDifference} - \log \text{Rest}$$

so if we let y^* be the logarithm of the difference in consumption and u^* be the logarithm of consumption at rest, then an equivalent definition of (M1) is

$$\begin{aligned} y^*_{ijk} &= \beta_0 + \alpha^S_j + \alpha^T_k + u^*_{ijk} + \varepsilon_{ijk}, \\ \alpha^S_A &= \alpha^T_H = 0, \\ \varepsilon_{ijk} &\sim \mathrm{N}(0, \sigma^2) \end{aligned} \tag{M1'}$$

The variable u^*_{ijk} appearing in the right-hand side of the model (M1') is an offset variable.

By fitting model (M1') indeed gives the same estimates for the parameters as expected.

2.3.1 R session

We store the data in a file called `frogs.dat` in the R working directory. The file is loaded using the R function `read.table` as follows

```
> ############################## Read data ################################
> frogs <- read.table("frogs.dat", header=TRUE)
> frogs
   Species Temperature  Rest Exercise
1        A         High 0.133    0.194
2        A         High 0.140    0.198
3        A          Low 0.107    0.152
4        A          Low 0.114    0.163
5        B         High 0.118    0.171
6        B         High 0.110    0.165
7        B          Low 0.098    0.144
8        B          Low 0.093    0.136
9        C         High 0.159    0.196
10       C         High 0.166    0.204
11       C          Low 0.126    0.182
12       C          Low 0.138    0.196
13       D         High 0.184    0.244
14       D         High 0.192    0.232
15       D          Low 0.154    0.207
16       D          Low 0.141    0.191
>
> attach(frogs) # Make variables in dataset available in workspace
```

Next we define and fit model (M1).

```
> ############################## Model M1 ################################
> RelativeIncrease <- (Exercise - Rest)/(Rest)
>
> M1 <- log(RelativeIncrease) ~ Species + Temperature
> fit1 <- lm(M1)
> summary(fit1)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.9526     0.1068   -8.92  2.3e-06
SpeciesB         0.0875     0.1350    0.65    0.530
SpeciesC        -0.3097     0.1350   -2.29    0.043
SpeciesD        -0.3553     0.1350   -2.63    0.023
TemperatureLow   0.2196     0.0955    2.30    0.042
```

```
Residual standard error: 0.191 on 11 degrees of freedom
Multiple R-squared:  0.66,      Adjusted R-squared:  0.536
F-statistic: 5.34 on 4 and 11 DF,  p-value: 0.0123
```

The summary table for the parameter estimates of the above model suggests that the *Temperature* variable is significant. The p-value associated with the *SpeciesB* parameter is not significant. This suggests that on average, species A and B don't differ in terms of the relative increase in consumption.

To add the interaction term (model (M2)) in the model we run the following. The last line uses the R function `anova` to perform the likelihood-ratio test between the model with and without interaction. The null hypothesis is that the model without interaction (the simpler model) is valid.

```
> ############################## Model M2 ################################
> M2 <- update(M1, . ~ . + Species*Temperature) # Add interaction
> fit2 <- lm(M2)
> summary(fit2)

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.83034    0.08433   -9.85  9.5e-06
SpeciesB                 0.08357    0.11926    0.70   0.5033
SpeciesC                -0.63586    0.11926   -5.33   0.0007
SpeciesD                -0.51427    0.11926   -4.31   0.0026
TemperatureLow          -0.02493    0.11926   -0.21   0.8396
SpeciesB:TemperatureLow  0.00784    0.16866    0.05   0.9641
SpeciesC:TemperatureLow  0.65226    0.16866    3.87   0.0048
SpeciesD:TemperatureLow  0.31784    0.16866    1.88   0.0962

Residual standard error: 0.119 on 8 degrees of freedom
Multiple R-squared:  0.904,     Adjusted R-squared:  0.819
F-statistic: 10.7 on 7 and 8 DF,  p-value: 0.00166

>
> anova(fit1, fit2) # Test if the interaction is significant using LR
Analysis of Variance Table

Model 1: log(RelativeIncrease) ~ Species + Temperature
Model 2: log(RelativeIncrease) ~ Species + Temperature + Species:Temperature
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     11 0.401
2      8 0.114  3     0.287 6.74  0.014
```

From the ANOVA table we conclude that the extended model (M2) is preferred at the 5% level. We can let R discover which is the best model (based on the AIC) using the function `step`. The input and output from the command is

```
> step(fit2) # Select the best model
Start:  AIC=-63.14
log(RelativeIncrease) ~ Species + Temperature + Species:Temperature

                       Df Sum of Sq   RSS   AIC
<none>                              0.114 -63.1
- Species:Temperature  3     0.287 0.401 -49.0
```

From the output above, the best model is the one currently fitted (`<none>`) and the second best is the one that excludes the interaction term (`- Species:Temperature`).

Finally, fitting (M1') using an `offset` term

```
> ############################## Model M1' ################################
> ConsumptionDifference <- Exercise - Rest
>
> M1b <- log(ConsumptionDifference) ~ Species + Temperature + offset(log(Rest))
> fit1b <- lm(M1b)
> summary(fit1b) # Note sames estimates as M1

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.9526     0.1068   -8.92  2.3e-06
SpeciesB         0.0875     0.1350    0.65    0.530
SpeciesC        -0.3097     0.1350   -2.29    0.043
SpeciesD        -0.3553     0.1350   -2.63    0.023
TemperatureLow   0.2196     0.0955    2.30    0.042

Residual standard error: 0.191 on 11 degrees of freedom
Multiple R-squared:  0.124,     Adjusted R-squared:  -0.195
F-statistic: 0.389 on 4 and 11 DF,  p-value: 0.813
```

Note here that the models defined by (M1) and (M1'), being equivalent, give the same estimates for the parameters.

3. Generalised Linear Models

In this chapter we introduce the generalised linear models as an extension of the linear models. We introduce the exponential family of distributions which can be used to model continuous, count, and binary data, and discuss some of its properties. We also discuss link functions that represent non-linear relationships between the mean of the distribution and the linear predictor.

3.1 The linear model

The linear model assumes that the relationship between the response y and the explanatory variables $\boldsymbol{x}$ is linear: $\boldsymbol{x}^\top\boldsymbol{\beta}$, for some p-dimensional vector $\boldsymbol{\beta}$, perturbed by some random error ε, i.e,

$$y = \boldsymbol{x}^\top\boldsymbol{\beta} + \varepsilon. \tag{3.1}$$

For a given sample $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$ we have

$$y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + \varepsilon_i,\ i = 1, \ldots, n$$

which is more conveniently written in vector form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$, $\boldsymbol{X}$ is a $n \times p$ matrix design matrix with ith row $\boldsymbol{x}_i$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$. If the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ is invertible the ordinary least squares estimator for $\boldsymbol{\beta}$ is

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

Then we have the following result about $\hat{\boldsymbol{\beta}}$

Theorem 3.1.
If $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d., then $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator for $\boldsymbol{\beta}$ and is the best unbiased estimator in the sense that the variance of any other unbiased estimator for $\boldsymbol{\beta}$ is larger than the variance of $\hat{\boldsymbol{\beta}}$, regardless of the value of $\boldsymbol{\beta}$.

On the other hand, the assumption of Theorem 3.1 is not appropriate for all types of data, e.g. binary, count or skewed data. Generalisation of this assumption leads to a new class of models called the *generalised linear model*.

3.2 Generalised linear models

Let $\mu_i := \mathsf{E}\, y_i$. The normality assumption of Theorem 3.1 is equivalent to

$$y_i \overset{\text{ind}}{\sim} \mathrm{N}(\mu_i, \sigma^2),\ \mu_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}. \tag{3.2}$$

Looking at (3.2) we can see that a linear model assumes that

1. the response is normally distributed,
2. the mean equals the linear predictor, and
3. the variance is not related to the mean.

These assumptions may not be valid in general. For example, when the response is binary, then the normal distribution is not valid and instead we want to model it using the Bernoulli distribution. The mean of the Bernoulli distribution is the probability of success. In this case, relating μ_i and x_i linearly is not reasonable since (i) the mean must be bounded between 0 and 1 and this constraint is difficult to impose on a linear function such as $\beta_0 + \beta_1 x_i$; and (ii) the effect of the regressor may not be linear on the scale of the mean, which is a probability but may be more likely on a transformed scale, for example the logit scale $\log\{\mu_i/(1-\mu_i)\}$. Furthermore, $\mathsf{Var}(y_i) = \mu_i(1-\mu_i)$. Thus knowing the mean of the data, we also know the variance. In other words the variance is a function of the mean.

Example 3.1 (Credit scoring)**.** Lenders, such as banks, use credit-scoring models to decide whether to grant a loan to an applicant, based on whether that applicant will be able to meet their financial obligation or default on it. These models are built using historical data from previous borrowers, such as their income, intended use of the funds, sex, marital status, etc, and whether they repaid their debt (a binary variable). A credit-scoring model is then able to estimate the probability that a new applicant will be able to repay their debt.

Here, we will consider a simple case where two applicants, Alice the actress and Bob the banker, each apply for a £50,000 loan from a bank. The bank would decide whether to grant the loan to each based on their annual salary only. Suppose that Alice's salary is $x_A = £20,000$ and Bob's salary is $x_B = £80,000$. Let π_A and π_B be the probabilities that Alice and Bob repay their debt respectively. Any sensible model would estimate $\pi_A < \pi_B$ because Bob's salary is higher, and $0 \leq \pi_A, \pi_B \leq 1$. These quantities are shown in Figure 3.1.

Now suppose that right before the bank decides on whether to grant the loan or not, Alice finds out that she was cast for a new play and her salary will increase by £10,000, i.e., her new salary now is $x'_A = £30,000$. Similarly, Bob just found out that, because his bank was making sensible investments, he is set to receive a £10,000 bonus, raising his annual salary to $x'_B = £90,000$. Let π'_A and π'_B be the new probabilities that Alice and Bob will repay their debts after their salary increase. The new probabilities will be higher than the old probabilities, i.e., $\pi_A < \pi'_A$ and $\pi_B < \pi'_B$, however, as Alice's salary was low, this salary increase is expected to have a bigger impact on her ability to repay her debt compared to Bob whose old probability of repayment was high anyway. In other words, the change $\pi'_A - \pi_A > \pi'_B - \pi_B$, even though $x'_A - x_A = x'_B - x_B$. Figure 3.1 demonstrates this. So if we were to model the relationship between salary, x, and probability of paying back a debt, $\pi(x)$, we would choose a non-linear relationship of the form illustrated in Figure 3.1.

Considering the uncertainty associated with each prediction, we would be fairly certain that individuals with relatively high salaries would repay their debt and individuals with relatively low salaries won't. We are less certain about individuals whose salaries fall in the middle,

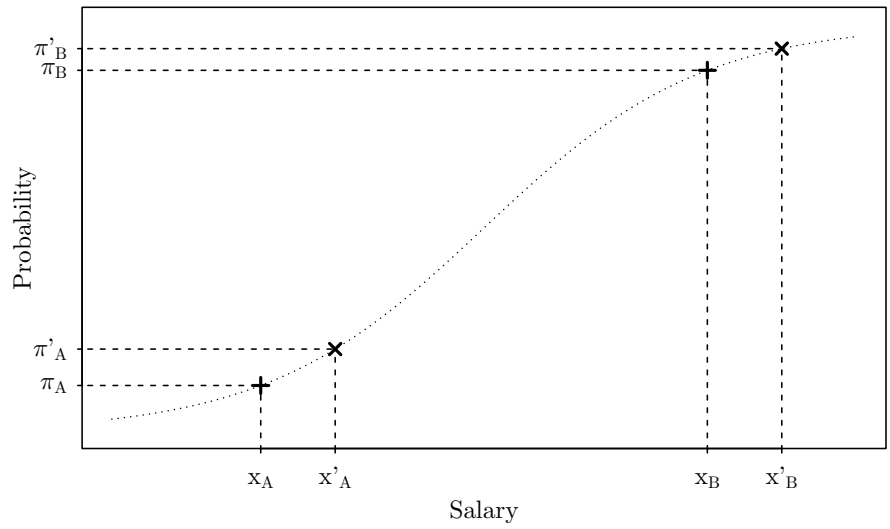


Figure 3.1: Probabilities that Alice and Bob repay their debt before and after they receive a salary increase. The dotted line shows the predicted probability for a given salary from the credit-scoring model.

because a change in their circumstances can easily affect their ability to repay their debt. Thus, the variance of our prediction should be low when the predicted probability is close to 0 or 1, but high when the probability is close to 0.5. ▶

A generalised linear model is a broader class of models which generalises the three assumptions above implied by the linear model.

Therefore, to specify a *generalised linear model* (GLM) we require three components:

- a distribution for the response variable,
- a linear predictor,
- a function that maps (links) the mean of the response to the linear predictor.

Specifically, a GLM may be written in the following way

$$y_i \overset{\text{ind}}{\sim} \text{EF}(\theta_i, \phi),\ g(\mu_i) = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}, \tag{3.3}$$

where

- $\mu_i = \mathsf{E}\, y_i$;
- $\text{EF}(\theta, \phi)$ denotes a member of the *exponential family* of distributions with *canonical parameter* $\theta \in \Theta$ and *dispersion parameter* $\phi > 0$ (see Definition 3.1 below);
- and g is a strictly monotone function called the *link function*.

Note that, instead of equating the mean with the linear predictor, we now assume that some monotonic function $g(\mu)$ equals the linear predictor. In (3.2), g is the identity function, $g(\mu) = \mu$. In Example 3.1 one may choose the logit function, $g(\mu) = \log\{\mu/(1-\mu)\}$, which maps the interval $(0, 1)$, i.e. the possible values of $\mu = \pi$, onto the real line, the possible values of the linear predictor $\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$. The link function enables us to model non-linear relationships between the data and the explanatory variables that often arise with binary, count, or skewed continuous data. The distribution of the data may therefore be discrete or skewed and is chosen from a

wide class of distributions, i.e., a family of distributions. The extension to (3.3) allows us to model data from a wider class of distributions including *inter alia* the binomial, Poisson, and gamma.

3.2.1 Exponential family of distributions

Definition 3.1 (Exponential family)**.**
We say that the random variable y has a distribution that belongs to the exponential family with *canonical parameter* $\theta \in \Theta$ and *dispersion parameter* $\phi > 0$, written $y \sim \mathrm{EF}(\theta, \phi)$, if its probability density/mass function can be written in the form

$$f(y; \theta, \phi) = \exp\{(y\theta - \psi(\theta))/\phi + c(y, \phi)\}. \tag{3.4}$$

The functions $\psi(\theta)$ and $c(y, \phi)$ are functions which are specific to each distribution from this family. The function $\psi(\theta)$ is called the *cumulant function* and has the property $\psi''(\theta) > 0$.

Example 3.2. The normal distribution, $\mathrm{N}(\mu, \sigma^2)$, is a member of the EF.

To see this express its pdf in the form (3.4).

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

which is of the form (3.4) with $\theta = \mu$, $\phi = \sigma^2$, $\psi(\theta) = \frac{1}{2}\theta^2$, and $c(y, \phi) = -\frac{1}{2}y^2/\phi - \frac{1}{2}\log(2\pi\phi)$. Thus, under the normality assumption and taking g to be the identity function, (3.3) reduces to (3.2). ▶

Variance function

It turns out (see Exercise 3.2) that $\mathsf{E}(y) = \psi'(\theta)$ and $\mathsf{Var}(y) = \phi\,\psi''(\theta)$. This relationship means that, unless ψ'' is constant, the mean and variance of the distribution depend on the same parameter, θ. In other words the variance is a function of the mean of the distribution, $v(\mu)$ say, called the *variance function* so that $\mathsf{Var}(y) = \phi\, v(\mu)$. The only case where the variance in not a function of the mean is the normal distribution and that is because indeed $\psi'' \equiv 1$ so that $v(\mu) \equiv 1$. Figure 3.2 shows the graphs of the variance functions for common distributions.

Remark 3.1. In model (3.3), the canonical parameter, which is related to the mean of the distribution, depends on the index i, while the dispersion parameter, which is related to the variance of the distribution, does not.

Exercise 3.1. Show that the following distributions belong to the EF by expressing their density (mass) function in the form (3.4).

1. Gamma(μ, ϑ) with pdf $f(y; \mu, \vartheta) = \frac{\vartheta^\vartheta}{\mu^\vartheta \Gamma(\vartheta)}\, y^{\vartheta-1} \exp(-\vartheta\, y/\mu),\ y > 0$; hence the exponential distribution also belongs to the EF.
2. Poisson with mean $\lambda > 0$.

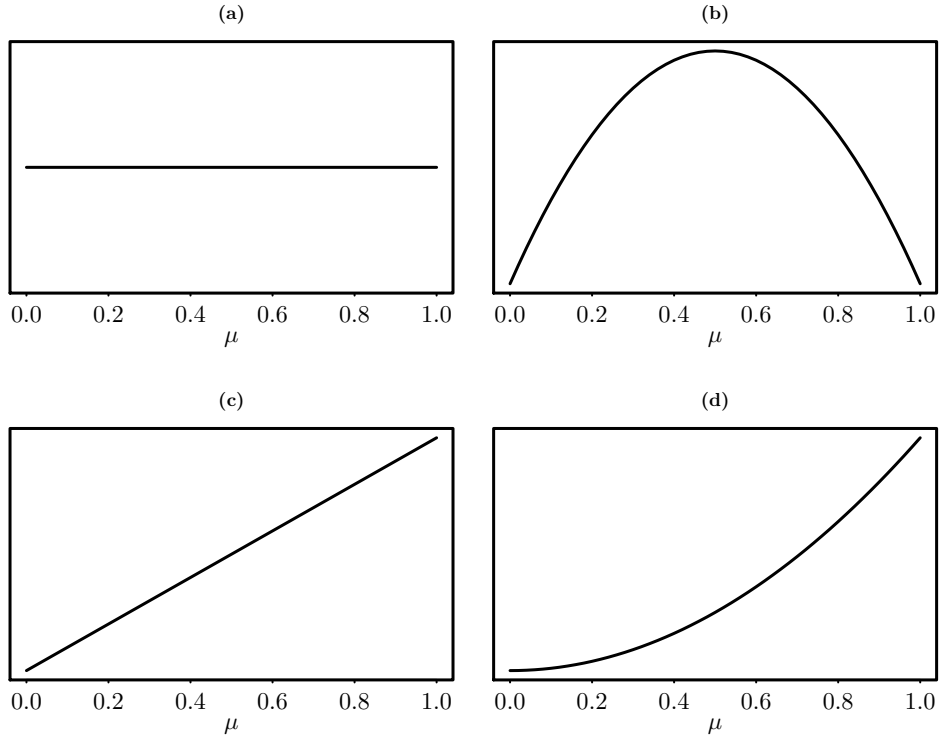


Figure 3.2: Plots of the variance function for four distributions. (a) Normal; (b) Bernoulli; (c) Poisson; (d) gamma.

3. Binomial(m, π) with m known.

4. Inverse-Gaussian IG(μ, λ) with pdf $f(y; \mu, \lambda) = \left(\frac{\lambda}{2\pi y^3}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{\lambda\,(y-\mu)^2}{\mu^2 y}\right\}$, $y > 0$, $\mu, \lambda > 0$.

5. Negative binomial with pmf $f(y; \kappa, \pi) = \frac{\Gamma(y+\kappa)}{y!\Gamma(\kappa)}(1-\pi)^\kappa \pi^y$, $y \in \{0, 1, 2, \ldots\}$, $\pi \in (0, 1)$, $\kappa > 0$ with κ known.

Exercise 3.2. Show that if $y \sim \mathrm{EF}(\theta, \phi)$, then $\mathsf{E}(y) = \psi'(\theta)$ and $\mathsf{Var}(y) = \phi\,\psi''(\theta)$. *Hint.* Use the properties of the mean and variance of the score function.

3.2.2 Link functions and the canonical link

The link function relates the mean μ of the datum y with the linear predictor $\eta = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$. In particular

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

The range of values for the mean of distributions from the exponential family is not always $(-\infty, \infty)$: for example the mean of a Bernoulli distributed random variable is restricted in $(0, 1)$ while the mean of the Poisson and gamma distributions is a number within $(0, \infty)$. On the other hand, the linear predictor, $\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}$, has the potential of attaining any value in $(-\infty, \infty)$ so for a wide range of the regressor variables it might give unreasonable estimates if set equal to the mean. Through the link function we transform the mean so that its values match $(-\infty, \infty)$. Additionally, to facilitate interpretation of the results from the GLM regression, the link function should be strictly monotone and continuous.

If $\mu \in (0, \infty)$, such as for the Poisson or gamma models, a popular choice for the link is the logarithm

$$\eta = \log(\mu) \Leftrightarrow \mu = \exp(\eta).$$

If $\mu \in (0, 1)$, such as for the Bernoulli model, the three most common choices are:

1. *logit*:
$$\eta = \text{logit}(\mu) = \log \frac{\mu}{1-\mu} \Leftrightarrow \mu = \frac{e^\eta}{1+e^\eta};$$

2. *probit*: (defined in terms of the normal CDF Φ)
$$\eta = \Phi^{-1}(\mu) \Leftrightarrow \mu = \Phi(\eta);$$

3. *complementary log-log*:
$$\eta = \log\{-\log(1-\mu)\} \Leftrightarrow \mu = 1 - \exp\{-\exp(\eta)\}.$$

Plots of these functions are shown in Figure 3.3. It is easy to see that all three inverse link functions have the same shape (a sigmoid), they are monotonically increasing and tend to 0 and 1 as η tends to $-\infty$ and $+\infty$ respectively.

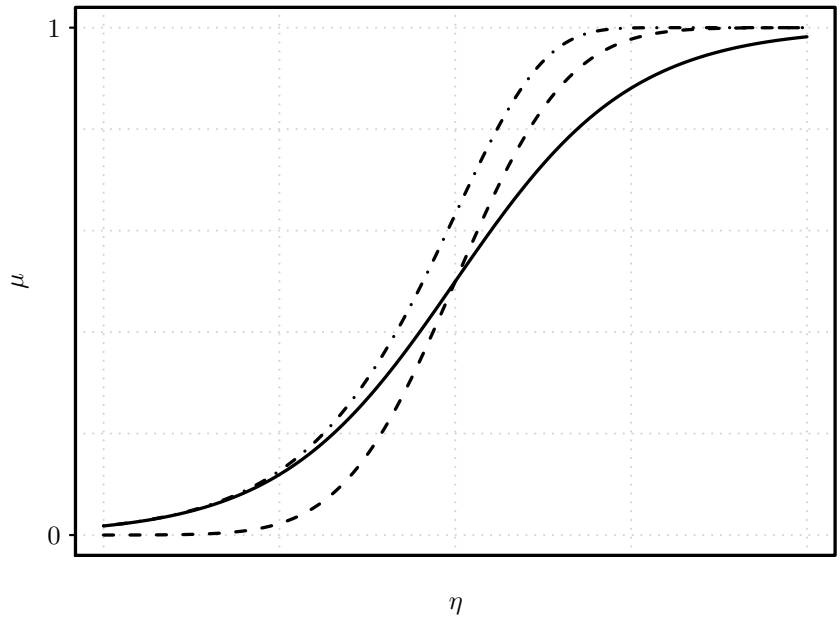


Figure 3.3: Plots of the inverse link functions for the Bernoulli GLM: logit (—); probit (– –); complementary log-log (– · –).

Example 3.3 (Interpretation of regressor coefficients)**.** Suppose that we are fitting a model with only one regressor variable, x, and an intercept. Then the linear predictor at x is $\eta_x = \beta_0 + \beta_1 x$, and the mean at x is $\mu_x = g^{-1}(\beta_0 + \beta_1 x)$. Therefore, the interpretation of β_0 and β_1 depends on the choice of link function. We can observe that, at $x = 0$, $\beta_0 = g(\mu_0)$. When x increases by one unit, we have $g(\mu_{x+1}) = \beta_0 + \beta_1(x+1) = \beta_0 + \beta_1 x + \beta_1 = g(\mu_x) + \beta_1 \Rightarrow \beta_1 = g(\mu_{x+1}) - g(\mu_x)$.

Identity link: If we choose the identity link, as is common for the classical linear model, then

$$\mu_x = \beta_0 + \beta_1 x$$

so the intercept β_0 is interpreted as the value of the mean when $x = 0$ and the coefficient β_1 is interpreted as the increase in the mean when the value of x is increased by 1 unit.

Logarithmic link: Suppose alternatively that we choose the logarithmic link function, as is common for Poisson regression. Then we have the expression

$$\log \mu_x = \beta_0 + \beta_1 x \Leftrightarrow \mu_x = e^{\beta_0+\beta_1 x}.$$

Then β_0 is the value of the $\log \mu_x$ when $x = 0$. Comparing the value of the mean at x, μ_x, and at $x + 1$, μ_{x+1}, we have that $\mu_{x+1} = e^{\beta_0+\beta_1(x+1)} = e^{\beta_0+\beta_1 x} e^{\beta_1} = \mu_x\, e^{\beta_1}$, so if the regressor variable x is increased by 1 unit then the mean is multiplied by a factor of e^{β_1}, and β_1 is the increase in the logarithm of the mean when x is increased by one unit. ▶

A special link function is the canonical link. On the one hand we have the relationship between the mean and the linear predictor: $g(\mu) = \boldsymbol{x}^\intercal \boldsymbol{\beta}$ and on the other hand we have the relationship between the mean and the canonical parameter $\mu = \psi'(\theta)$ so by combining the two we have $g(\psi'(\theta)) = \boldsymbol{x}^\intercal \boldsymbol{\beta}$ which defines the relationship between the canonical parameter and the linear predictor. It is then possible to choose g in such way so that $\theta = \boldsymbol{x}^\intercal \boldsymbol{\beta}$. This particular choice of g is the *canonical link*.

Definition 3.2 (Canonical link)**.**
The canonical link for a GLM is the link function g that satisfies the relationship $\theta = \boldsymbol{x}^\intercal \boldsymbol{\beta}$. In this case $g = \psi'^{-1}$.

Table 3.1 shows the characteristics of four common distributions belonging to the EF including the canonical link and variance function.

3.2.3 Some examples

Log-linear models for counts

In the early stages of an epidemic the mean number of infections in the population is increasing exponentially with time. Let x_i denote the time elapsed from the start of the epidemic and μ_i the mean number of infections after time x_i. The observed number of new infections y_i after time x_i is modelled as Poisson with mean (rate) μ_i. This suggests the following model

$$y_i \overset{\text{ind}}{\sim} \text{Po}(\mu_i),\ \mu_i = \mu_0 e^{\alpha x_i},\ i = 1, \ldots, n$$

Letting $\beta_0 = \log(\mu_0)$, $\beta_1 = \alpha$, the expression for the mean becomes

$$\mu_i = e^{\beta_0+\beta_1 x_i} \Leftrightarrow \log(\mu_i) = \beta_0 + \beta_1 x_i$$

Example 3.4 (AIDS cases in Belgium)**.** The number of new AIDS cases each year since 1980 in Belgium were recorded. We are interested to know how the disease evolves in time in order to predict the number of new cases for the next year.

The data were fitted using a Poisson log-linear GLM. Figure 3.4 shows the observed number of cases together with the prediction from the GLM fit. The prediction corresponds to an exponential curve. ▶

	Normal	Poisson	Bernoulli	Gamma
Notation	$\mathrm{N}(\mu, \sigma^2)$	$\mathrm{Po}(\mu)$	$\mathrm{Bin}(1, \mu)$	$\mathrm{G}(\mu, \vartheta)$
Density/mass	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$	$\frac{\mu^y}{y!}e^{-\mu}$	$\mu^y(1-\mu)^{1-y}$	$\frac{1}{\Gamma(\vartheta)}\left(\frac{\vartheta}{\mu}\right)^{\vartheta} y^{\vartheta-1}e^{-y\vartheta/\mu}$
Support of y	$(-\infty, \infty)$	$\{0, 1, 2, \ldots\}$	$\{0, 1\}$	$(0, \infty)$
θ	μ	$\log(\mu)$	$\mathrm{logit}(\mu)$	$-1/\mu$
ϕ	σ^2	1	1	ϑ^{-1}
$\psi(\theta)$	$\theta^2/2$	e^θ	$\log(1+e^\theta)$	$-\log(-\theta)$
$\mu = \psi'(\theta)$	θ	e^θ	$e^\theta/(1+e^\theta)$	$-1/\theta$
$v(\mu)$	1	μ	$\mu(1-\mu)$	μ^2
Canonical link	μ	$\log(\mu)$	$\mathrm{logit}(\mu)$	$-1/\mu$
R `family`	`gaussian`	`poisson`	`binomial`	`Gamma`
R default `link`	μ	$\log(\mu)$	$\mathrm{logit}(\mu)$	$1/\mu$

Table 3.1: Characteristics of some common distributions in the exponential family.

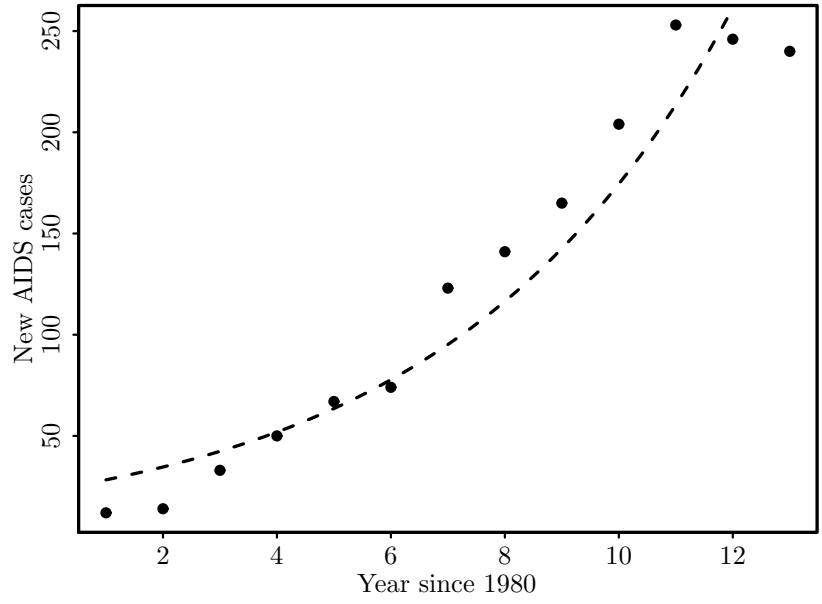


Figure 3.4: New AIDS cases in Belgium and GLM fit.

Logistic regression for proportions

Suppose we are interested in the probability of occurrence π of a certain event under conditions $\boldsymbol{x}$. For example π could denote the probability of a patient to recover from a disease following treatment given the patient's age, gender, weight, etc. Suppose data $(\boldsymbol{x}_i, y_i)$ are collected where

y_i is 1 or 0 according to whether the event has occurred or not. A common practice is to model the logarithm of the odds linearly with respect to the parameters:

$$y_i \sim \text{Bin}(1, \mu_i), \ \log \frac{\mu_i}{1-\mu_i} = \boldsymbol{x}_i^\intercal \boldsymbol{\beta}.$$

Example 3.5 (Heart attacks and creatine kinase)**.** Data were collected for examining the efficacy of blood creatine kinase (CK) levels as a diagnostic when patients present with symptoms that may indicate a heart attack. Figure 3.5 shows the observed proportion of heart attacks against the patient's CK level. In this example we would like to know what is the risk (i.e. the probability) of heart attack for a patient for a given CK level.

The logistic regression model was used to fit the data and the plotted curve shows the fit. The curve corresponds to the logit^{-1} function from where we can infer that a patient with CK level around 100 has 50% chance of having a heart attack and a patient with CK level around 150 has 90% chance of having a heart attack. ▶

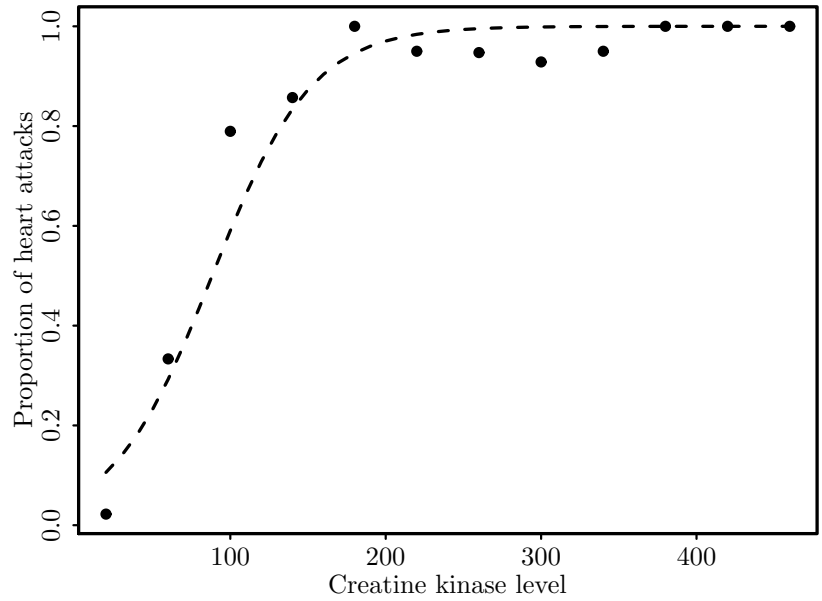


Figure 3.5: Proportion of heart attacks plotted against the creatine kinase level. The GLM fit is also shown.

Probit model

In toxicology experiments test animals are divided into n groups and each animal is subject to a known level x of toxin. The level x varies from group to group but within each group it's the same. For the ith set, the number y_i of survivors out of m_i is recorded. We would like to know the probability μ_x of surviving a dose x of the toxin. This probability is usually expressed in terms of $\Phi(\cdot)$, the normal cumulative distribution function, as

$$\mu_x = \Phi(\beta_0 + \beta_1 x)$$

which suggests the model

$$y_i \sim \text{Bin}(m_i, \mu_i), \ \mu_i = \Phi(\beta_0 + \beta_1 x_i).$$

In this case $g(\mu) = \Phi^{-1}(\mu)$.

Example 3.6 (Toxicity of the chemical diglyme)**.** Diglyme ($C_6H_{14}O_3$) is a component of industrial solvents, used widely in the manufacture of protective coatings such as lacquers, metal coatings, baking enamels, etc. Although to date, several attempts have proved inadequate to evaluate the potential of glycol ethers to produce human reproductive toxicity, several related compounds have been identified as reproductive toxicants.

In one study, data were collected to determine the toxicity of diglyme on mice. In the study a specified amount of diglyme was dosed daily to pregnant mice, and after ten days all foetuses were examined. The proportion of stillborn foetuses was recorded for different concentrations of the chemical. Figure 3.6 shows the proportion of stillborn foetuses observed under different concentrations (in mg/Kg/day). A binomial GLM with probit link was fitted to the data. The fit is shown by a dashed line in Figure 3.6. From the fit we can infer that a dose of 500 mg per day would kill about 50% of the foetuses.

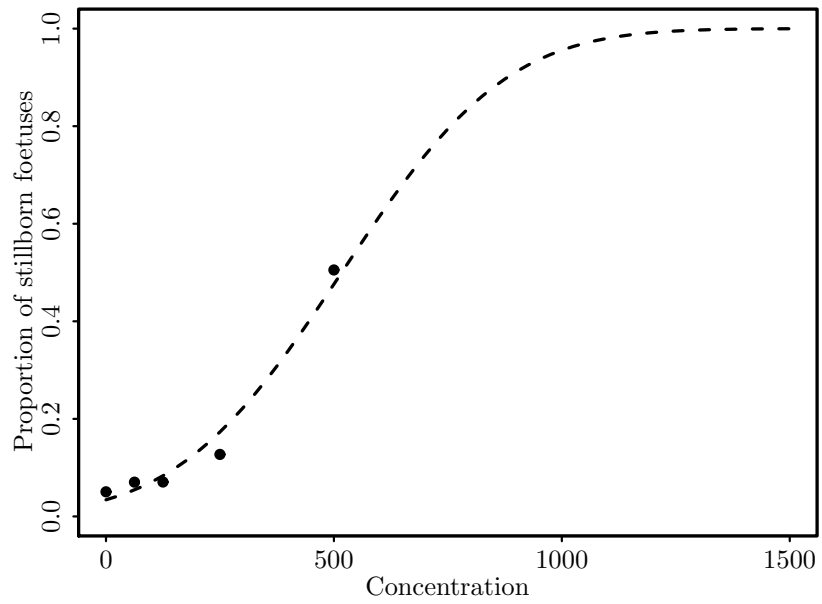


Figure 3.6: Proportion of stillborn foetuses plotted against the chemical concentration. The GLM fit is also shown.

3.3 Historical notes (optional)

Generalised linear models were first proposed by statisticians John A Nelder and Robert W M Wedderburn in a paper titled "Generalized Linear Models" published in the *Journal of the Royal Statistical Society, Series A* in the year 1972. Although these models existed previously as individual models, Nelder and Wedderburn were the first to propose this unifying framework, and also developed much of the theory for inference which will be presented in the next chapters. Wedderburn would sadly pass away unexpectedly three years later in 1975, but not before making another major contribution to the subject, that of quasi-likelihood inference, which we will come to in Chapter 5.

4. Inference for Generalised Linear Models

In this chapter we discuss parameter estimation and goodness-of-fit tests for GLM.

4.1 Estimation of the parameters of a GLM

The general form of the model we wish to fit is

$$y_i \overset{\text{ind}}{\sim} \text{EF}(\theta_i, \phi),\ g(\mu_i) = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta},\ i = 1, \ldots, n$$

where the parameter vector $\boldsymbol{\beta}$ is unknown and the parameter ϕ may be known (as in the Bernoulli and Poisson) or unknown (as in the normal and gamma). The parameters μ_i and θ_i are related to $\boldsymbol{\beta}$, so if we know $\boldsymbol{\beta}$, then we also know these parameters.

4.1.1 Maximum likelihood estimation for the regressor coefficients

Estimation for $\boldsymbol{\beta}$ is performed by the method of maximum likelihood. Using the exponential family form, the likelihood is given by

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \exp\left\{\frac{1}{\phi}\left(\sum y_i\theta_i - \sum \psi(\theta_i)\right) + \sum c(y_i, \phi)\right\}$$

where the dependence of L on $\boldsymbol{\beta}$ is only through θ_i by means of the relationship

$$\begin{aligned} \mu_i &= \psi'(\theta_i) \\ g(\mu_i) &= \eta_i \\ \eta_i &= \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} \end{aligned} \tag{4.1}$$

The log-likelihood is

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \frac{1}{\phi}\left(\sum y_i\theta_i - \sum \psi(\theta_i)\right) + \sum c(y_i,\phi)$$

which is a nonlinear function in $\boldsymbol{\beta}$. In fact, note that $\boldsymbol{\beta}$ appears only in the term in the brackets implying that the maximum likelihood estimates (MLE) for $\boldsymbol{\beta}$ don't depend on the, possibly unknown, parameter ϕ.

The Fisher scoring algorithm (Definition 1.3) is the iterative procedure used for finding the MLE. For the implementation of the Fisher scoring, the score vector and the Fisher information matrix are needed.

The score vector has jth element

$$\begin{aligned} u_j(\boldsymbol{\beta}|\boldsymbol{y}) &= \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} \\ &= \frac{1}{\phi}\left(\sum y_i \frac{\partial \theta_i}{\partial \beta_j} - \sum \psi'(\theta_i)\frac{\partial \theta_i}{\partial \beta_j}\right) \\ &= \frac{1}{\phi}\sum (y_i - \psi'(\theta_i))\frac{\partial \theta_i}{\partial \beta_j} \end{aligned}$$

where the summation is over $i = 1, \ldots, n$. For the moment we refrain giving an expression for $\partial\theta_i/\partial\beta_j$ but note that this can be computed for a given $\boldsymbol{\beta}$ using the chain rule and equations (4.1) and its value will depend on the particular distribution and link function assumed. In particular, under the canonical link, $\partial\theta_i/\partial\beta_j = x_{ij}$, the (i,j) element of the design matrix $\boldsymbol{X}$, because in that case $\theta_i = \eta_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}$.

The observed information matrix $\mathcal{J}$ is the negative of the second order derivatives of the log-likelihood and has (j,k) element given in a similar manner by

$$\begin{aligned} \mathcal{J}_{jk}(\boldsymbol{\beta}|\boldsymbol{y}) &= -\frac{\partial^2 \ell(\boldsymbol{\beta}|\boldsymbol{y})}{\partial\beta_j \partial\beta_k} \\ &= -\frac{1}{\phi}\sum \left(y_i - \psi'(\theta_i)\right)\frac{\partial^2\theta_i}{\partial\beta_j\partial\beta_k} + \frac{1}{\phi}\sum \psi''(\theta_i)\frac{\partial\theta_i}{\partial\beta_j}\frac{\partial\theta_i}{\partial\beta_k}. \end{aligned} \tag{4.2}$$

Noting that the first term of (4.2) has expectation 0 because $\mathsf{E}\, y_i = \psi'(\theta_i)$, while the second term is not random, the Fisher information matrix for $\boldsymbol{\beta}$, $\mathcal{I}(\boldsymbol{\beta})$, has (j,k) element given by

$$\begin{aligned} \mathcal{I}_{jk}(\boldsymbol{\beta}) &= \mathsf{E}\,\mathcal{J}_{jk}(\boldsymbol{\beta}|\boldsymbol{y}) \\ &= \frac{1}{\phi}\sum \psi''(\theta_i)\frac{\partial\theta_i}{\partial\beta_j}\frac{\partial\theta_i}{\partial\beta_k}. \end{aligned}$$

We also need a way of computing $\frac{\partial\theta_i}{\partial\beta_j}$. To that end, applying the chain rule on equations (4.1),

$$\begin{aligned} \frac{\partial\theta_i}{\partial\beta_j} &= \frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j} \\ &= \frac{1}{\psi''(\theta_i)}\frac{1}{g'(\mu_i)}x_{ij}. \end{aligned}$$

Substituting into the expression for the Fisher information matrix, we have (see Appendix B.1)

$$\mathcal{I}(\boldsymbol{\beta}) = \frac{1}{\phi}\boldsymbol{X}^\mathsf{T}V^{-1}\boldsymbol{X} \tag{4.3}$$

where V is diagonal with ith diagonal element given by $v_i = \psi''(\theta_i)(g'(\mu_i))^2 = v(\mu_i)(g'(\mu_i))^2$ and $v(\mu)$ is the variance function. Note also that

$$\frac{\partial \theta_i}{\partial \beta_j} = x_{ij} v_i^{-1} g'(\mu_i),$$

and define the n-dimensional vector $\boldsymbol{z}$ by $z_i := \eta_i + g'(\mu_i)(y_i - \mu_i)$. Then

$$\boldsymbol{u}(\boldsymbol{\beta}) = \frac{1}{\phi} \boldsymbol{X}^\top V^{-1}(\boldsymbol{z} - \boldsymbol{\eta}). \tag{4.4}$$

The Fisher scoring iterative procedure for the MLE proceeds by updating $\boldsymbol{\beta}^{(m)}$ to $\boldsymbol{\beta}^{(m+1)}$ for $m = 0, 1, 2, \ldots$ until convergence by

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathcal{I}(\boldsymbol{\beta}^{(m)}))^{-1} \boldsymbol{u}(\boldsymbol{\beta}^{(m)}) \tag{4.5}$$

Observe that ϕ is cancelled in the above equation, therefore the algorithm can in fact be implemented without knowing the true ϕ.

Example 4.1. Consider the AIDS data of Example 3.4. The fitted model is

$$y_i \overset{\text{ind}}{\sim} \text{Po}(\mu_i), \ i = 1, \ldots, 13$$
$$\log \mu_i = \beta_0 + \beta_1 x_i$$

where y_i denotes the number of cases and x_i the number of years since 1980. For this model, $g(\mu) = \log \mu$ and $v(\mu) = \mu$ so the diagonal matrix V has elements $v_i = v(\mu_i)(g'(\mu_i))^2 = 1/\mu_i$ and the vector $\boldsymbol{z}$ has elements $z_i = \eta_i + g'(\mu_i)(y_i - \mu_i) = \eta_i + (y_i - \mu_i)/\mu_i$.

The R code below illustrates the fitting.

```
> y <- c(12, 14, 33, 50, 67, 74, 123, 141, 165, 204, 253, 246, 240)
> x <- 1:13
> fit1 <- glm(y ~ x, family = poisson) # Using the glm function
> coef(fit1)
(Intercept)           x
  3.1405895   0.2021212
>
> ## Demonstration of the Fisher scoring algorithm
> X <- cbind(1,x)               # Design matrix X
> beta <- c(1, 0)               # Initial guess
> for (m in 1:200) {            # Fisher scoring iterations
>   eta <- beta[1] + beta[2]*x  # eta
>   mu <- exp(eta)              # mean
>   Vinv <- diag(mu)            # diagonal matrix V
>   I <- t(X) %*% Vinv %*% X    # Fisher information matrix (phi = 1)
>   z <- eta + (y - mu)/mu      # Vector z
>   u <- t(X) %*% Vinv %*% (z - eta) # Score
>   beta.old <- beta
>   beta <- beta + solve(I,u)   # Update: beta(m+1) = beta(m) + I^{-1} * u
>   convergence <- max(abs(beta-beta.old)) < 1e-6
>   if(convergence) break       # Stop iterations if convergence
> }
> beta
                      x
  3.1405895   0.2021212
```

▶

Equation (4.4) has an interesting interpretation. Consider the weighted linear model with response z_i on regressor variables $\boldsymbol{x}_i$ with weights v_i, $i = 1, \ldots, n$,

$$z_i = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i,\ \varepsilon_i \overset{\text{ind}}{\sim} N(0, v_i\phi),\ i = 1, \ldots, n,$$

where the weights v_i are known. Such models arise when the response is the total or average over strata of different sizes, e.g., z_i may represent the unemployment rate at city i of population P_i, in which case, $v_i = 1/P_i$. The weighted least squares estimate of $\boldsymbol{\beta}$ is obtained by solving $\boldsymbol{u}(\hat{\boldsymbol{\beta}}) = 0$, where $\boldsymbol{u}(\boldsymbol{\beta})$ is given by (4.4). However, in the case of GLM z_i and v_i depend on the response $\boldsymbol{y}$, so a GLM is not equivalent to a weighted linear model. Nevertheless, this similarity is helpful because it shows that, in a first approximation, the GLM resembles the weighted linear model, so the theory of linear models can be used to derive approximate results for GLM. In fact, this approach is used to justify the theory developed in this chapter.

Exercise 4.1. Show that under canonical link, the Newton-Raphson (1.5) reduces to the Fisher scoring algorithm.

Exercise 4.2. Give expressions for the vector $\boldsymbol{z}$ and the matrix V for the following GLM's

1. Poisson model with logarithmic link,
2. Bernoulli model with logit link,
3. Bernoulli model with probit link,
4. Bernoulli model with complementary log-log link.

4.1.2 Estimation of the dispersion parameter

For some GLMs such as the Bernoulli and Poisson the dispersion parameter ϕ is known but for others such as the normal and gamma it is not. However knowledge of the dispersion parameter is not necessary for model fitting since the likelihood for $\boldsymbol{\beta}$ does not depend on it. Therefore the dispersion parameter is estimated after the estimation of $\boldsymbol{\beta}$ using the method of moments estimator.

First note that

$$\phi v(\mu) = \mathsf{Var}\, y = \mathsf{E}(y-\mu)^2 \Rightarrow \phi = \mathsf{E}\,\frac{(y-\mu)^2}{v(\mu)}$$

which suggests the following estimator

$$\hat{\phi} = \frac{1}{n-p}\sum \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}, \tag{4.6}$$

where $\hat{\mu}_i$ is the estimated mean for the ith observation. The asymptotic distribution of $\hat{\phi}$ is, in analogy to the case of $\hat{\sigma}^2$ in classical linear models.

Theorem 4.1.
Asymptotically, as $n \to \infty$,

$$\frac{(n-p)\hat{\phi}}{\phi} \dot{\sim} \mathcal{X}^2_{n-p}. \tag{4.7}$$

A proof of this theorem is given in Section 4.4.

Example 4.2. Consider the gamma GLM with logarithmic link

$$y_i \overset{\text{ind}}{\sim} \text{Gamma}(\mu_i, \vartheta), \;\; \log \mu_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}, \;\; i = 1, \ldots, n,$$

where $\text{Gamma}(\mu, \vartheta)$ denotes the gamma distribution with mean μ and shape parameter ϑ, and $\boldsymbol{\beta} \in \mathbb{R}^p$. For this model, the dispersion parameter is unknown and is given by $\phi = 1/\vartheta$ (see Table 3.1). For this gamma distribution, the variance function is $v(\mu) = \mu^2$. Let $\hat{\mu}_i = \exp(\boldsymbol{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}})$. Then,

$$\begin{aligned}
\hat{\phi} &= \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} \\
&= \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2} \\
&= \frac{1}{n-p}\sum_{i=1}^{n}\left(\frac{y_i}{\hat{\mu}_i} - 1\right)^2,
\end{aligned}$$

is the method of moments estimator for ϕ. ▶

4.1.3 Fitting GLM in R

The R function `glm` is used to fit a GLM in R via the Fisher scoring algorithm. The format of the function is

```
glm(formula, family = gaussian, data, weights, subset,
    start = NULL, etastart, mustart, control = list(...), ...)
```

with input (see the R help documentation for more details by issuing the command `?glm`)

`formula` The model formula in the form of `response ~ terms`, i.e a symbolic description of the model to be fitted. In the case of the binomial GLM the response should be a two-column matrix with the first column indicating the number of successes and the second column the number of failures.

`family` The assumed distribution of the data in the form `distribution(link)` where `distribution` is one of `binomial`, `gaussian`, `Gamma`, `poisson` and others (see the R help documentation for `family`) and `link` is the name of the link function. If the `link` is not specified the canonical link is assumed. The `gaussian` family accepts the links (as names) `identity`, `log` and `inverse`; the `binomial` family the links `logit`, `probit`, and `cloglog` (complementary log-log) among others; the `Gamma` family the links `inverse`, `identity` and `log`; and the `poisson` family the links `log`, `identity`, and `sqrt`.

`data` An optional data frame, list or environment containing the variables in the model.

`weights` An optional vector of 'prior weights' to be used in the fitting process. Should be `NULL` or a numeric vector.

`subset` An optional vector specifying a subset of observations to be used in the fitting process.

`start` Optional starting values for the parameters in the linear predictor.

`etastart` Optional starting values for the linear predictor.

`mustart` Optional starting values for the vector of means.

`control` Optional list of parameters for controlling the fitting process. The list may have the following named variables. `epsilon`: convergence tolerance (default is 10^{-8}); `maxit`: an integer giving the maximum number of iterations (default is 25); `trace`: logical indicating if output should be produced for each iterations (default is `FALSE`).

`...` Optional arguments to be used to form the `control` argument if it is not supplied directly.

The R functions described in section 2.2 can be applied to the object created from using the `glm` function.

4.2 Confidence intervals and hypothesis tests

The natural step following estimation of the parameters is inference.

4.2.1 Inference for the regressor parameters

According to Theorem 1.1 and (4.3), the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} \dot{\sim} N_p \left(\boldsymbol{\beta}, \phi \, (\boldsymbol{X}^\intercal V^{-1} \boldsymbol{X})^{-1} \right) \tag{4.8}$$

with the variance obtained by inverting the Fisher information matrix in (4.3). Equation (4.8) is the basis for constructing confidence intervals and performing hypothesis tests on $\boldsymbol{\beta}$.

As is, equation (4.8) cannot be used in practice because the variance depends on μ (through the matrix V), which is unknown. In practice μ is replaced by $\hat{\mu}$ (see next section), essentially replacing V by $\hat{V} = V(\hat{\mu})$ in (4.8).

Equation (4.8) suggests that the marginal distribution of $\hat{\beta}_j$ is asymptotically normal with mean β_j and variance denoted by $\sigma_j^2 = \phi u_j$, where u_j is the (j, j) element of $(\boldsymbol{X}^\intercal \hat{V}^{-1} \boldsymbol{X})^{-1}$.

We distinguish two cases depending on whether the dispersion parameter ϕ is known (such as in the binomial and Poisson) or not (such as in the normal and gamma). In any case, the following result holds

$$z = \frac{\hat{\beta}_j - \beta_j}{\sigma_j} \dot{\sim} N(0, 1). \tag{4.9}$$

Case ϕ known:
Then, from (4.9), z is a pivot for β_j. Then a typical $(1 - \alpha)100\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm z^* \, \sigma_j$$

where z^* is the $1 - \alpha/2$ quantile of the standard normal distribution.

For a hypothesis test, we consider

$$H_0 : \beta_j = \beta_{(0)} \text{ vs } H_1 : \beta_j \neq \beta_{(0)} \tag{4.10}$$

where $\beta_{(0)}$ is some possible value for β_j, typically 0. Then, from (4.9),

$$z = \frac{\hat{\beta}_j - \beta_{(0)}}{\sigma_j} \dot{\sim} N(0, 1) \text{ under the } H_0$$

We would reject the H_0 if the value of z above is extreme compared to the $N(0, 1)$. Therefore, a level α test would reject H_0 if $|z| \geq z^*$. The p-value of the test is

$$\text{p-value} = 2 \Pr(N(0, 1) > |z|).$$

Case ϕ estimated:

In this case we cannot use σ_j because we don't know ϕ. We let $\hat{\sigma}_j^2 = \hat{\phi} u_j$ where u_j is as before the (j,j) element of $(\boldsymbol{X}^\intercal \hat{V}^{-1} \boldsymbol{X})^{-1}$ and $\hat{\phi}$ is the estimator from (4.6).

Then, combining (4.9), (4.7), and the fact that a t_ν distribution is defined as

$$t_\nu = \frac{N(0,1)}{\sqrt{\mathcal{X}_\nu^2/\nu}},$$

we have

$$t = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \dot{\sim} t_{n-p}. \tag{4.11}$$

In this case the level-α confidence interval for β_j becomes

$$\hat{\beta}_j \pm t^* \, \hat{\sigma}_j$$

where t^* is the $1-\alpha/2$ quantile of the Student's t_{n-p} distribution. For the hypothesis test (4.10) the null hypothesis is rejected if $|t| \geq t^*$ where

$$t = \frac{\hat{\beta}_j - \beta_{(0)}}{\hat{\sigma}_j} \dot{\sim} t_{n-p} \text{ under the } H_0 \tag{4.12}$$

with p-value

$$\text{p-value} = 2\Pr(t_{n-p} > |t|). \tag{4.13}$$

Implementation in R.

The R function `summary` provides a table with the parameter estimates $\hat{\boldsymbol{\beta}}$ along with their standard errors, z or t statistics for testing the significance of each parameter and corresponding p-values. The general call of the function is as follows.

```
summary(object, dispersion = NULL, correlation = FALSE)
```

where

`object` Is the object containing the output from the call to the function `glm`.

`dispersion` Is an optional input used to set the dispersion parameter ϕ to a known value. If it is `NULL` (the default) then it is inferred from the fitted model.

`correlation` If `TRUE` then the correlation matrix of $\hat{\boldsymbol{\beta}}$ is computed and returned as well.

Example 4.3. Continuing from Example 4.1 we demonstrate in the following R code the construction of the summary of a GLM fit.

```
> summary(fit1, correlation = TRUE)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.140590   0.078247   40.14   <2e-16
x           0.202121   0.007771   26.01   <2e-16

(Dispersion parameter for poisson family taken to be 1)

Correlation of Coefficients:
  (Intercept)
```

```
x -0.95

>
> X <- cbind(1, x) # Design matrix
> phi <- 1
> I <- (1/phi)*(t(X) %*% diag(1/V) %*% X) # I = (Xtranspose * V^{-1} * X)/phi
> Var.beta <- solve(I)          # Inverse of the Fisher information matrix
> sigma.beta <- sqrt(diag(Var.beta))
> zstat <- beta/sigma.beta      # Test statistic
> pval <- 2*pnorm(abs(zstat), lower.tail = FALSE) # p-value
>
> # Summary table
> cbind(Estimate = beta, SE = sigma.beta, z = zstat, p = pval)
             Estimate          SE        z             p
(Intercept) 3.1405895 0.078247000 40.13687  0.000000e+00
x           0.2021212 0.007771492 26.00803 4.017647e-149
>
> # Correlation
> Var.beta[1, 2]/(sigma.beta[1]*sigma.beta[2])
-0.9483162
```

Then, a 95% confidence interval for β_1 is $0.202 \pm (1.96)(0.0078) = (0.187, 0.217)$ which does not contain 0, indicating that the regressor x is significant.

Similarly, the p-value that corresponds to the hypothesis test H_0: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$ is very close to 0 so the conclusion is that the variable is significant. ▶

4.2.2 Estimation of the mean

The mean $\mu_{\boldsymbol{x}}$ for a given value of regressors $\boldsymbol{x}$ is sometimes of interest. For example, in the Poisson GLM the mean is the rate at which an event occurs while in the Bernoulli GLM the mean is associated with the probability of success. The mean is related to the linear predictor $\eta_{\boldsymbol{x}}$ through the link function and the asymptotic distribution of $\hat{\eta}_{\boldsymbol{x}} = \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}$ is, by (4.8),

$$\hat{\eta}_{\boldsymbol{x}} \dot{\sim} N(\boldsymbol{x}^\top \boldsymbol{\beta}, \phi\, \boldsymbol{x}^\top (\boldsymbol{X}^\top V^{-1} \boldsymbol{X})^{-1} \boldsymbol{x}) \tag{4.14}$$

using the properties for the mean and variance of the multivariate normal distribution.

The maximum likelihood estimator for the mean $\mu_{\boldsymbol{x}} := g^{-1}(\eta_{\boldsymbol{x}})$ is obtained by inverting the link function:

$$\hat{\mu}_{\boldsymbol{x}} = g^{-1}(\hat{\eta}_{\boldsymbol{x}}) = g^{-1}(\boldsymbol{x}^\top \hat{\boldsymbol{\beta}}).$$

Thus a confidence interval for the mean is obtained by transforming the corresponding confidence interval for the linear predictor based on (4.14). In other words, if $(\hat{\eta}^L_{\boldsymbol{x}}, \hat{\eta}^U_{\boldsymbol{x}})$ is a $(1-\alpha)100\%$ confidence interval for $\eta_{\boldsymbol{x}}$, derived from (4.14), then

$$\left(g^{-1}(\hat{\eta}^L_{\boldsymbol{x}}), g^{-1}(\hat{\eta}^U_{\boldsymbol{x}})\right)$$

is a $(1-\alpha)100\%$ confidence interval for $\mu_{\boldsymbol{x}}$.

Implementation in R.
Both $\hat{\eta}_{\boldsymbol{x}}$ and $\hat{\mu}_{\boldsymbol{x}}$ can be obtained in R using the function `predict`. The structure of the function predict is as follows

```
predict(object, newdata = NULL, type = c("link", "response"),
        se.fit = FALSE, dispersion = NULL)
```

where

object Is the object containing the output from the call to the function `glm`.

newdata Is a `list` with named elements containing the values of the regressors $\boldsymbol{x}$ to use for the prediction. If omitted, the design matrix $\boldsymbol{X}$ is used instead.

type Indicates the type of prediction required. The type `"link"` (the default) computes $\hat{\eta}_{\boldsymbol{x}}$. The type `"response"` computes $\hat{\mu}_{\boldsymbol{x}}$.

se.fit Indicates whether the standard error of the fit should be computed and returned as well.

dispersion The dispersion of the GLM fit to be assumed in computing the standard errors. If omitted, that returned by `summary` applied to the object is used.

Example 4.4. Continuing from Example 4.3, suppose that we wish to predict the expected number of cases for the next year. In this case $\boldsymbol{x} = (1, 14)$. Then $\hat{\eta}_{\boldsymbol{x}} = 3.141 + 0.202 \times 14 = 5.969$ and $\mathrm{Var}\,\hat{\eta}_{\boldsymbol{x}} = (0.0782)^2 + (0.0078)^2 \times 14^2 - 2 \times 14 \times (0.948)(0.0782)(0.0078) = (0.043)^2$. Therefore, a 95% confidence for $\eta_{\boldsymbol{x}}$ is $5.969 \pm (1.96)(0.043) = (5.88, 6.05)$. Since $\mu_{\boldsymbol{x}} = \exp(\eta_{\boldsymbol{x}})$, $\hat{\mu}_{\boldsymbol{x}} = e^{5.969} = 391$ and a 95% confidence interval for $\mu_{\boldsymbol{x}}$ is $(e^{5.88}, e^{6.05}) = (360, 426)$.

The above is verified with the following R code

```
> xx <- 14
>
> eta.x <- predict(fit1, newdata = list(x = xx), type = "link", se.fit = TRUE)
> eta.x
$fit
5.970286

$se.fit
[1] 0.04258572

> eta.ci <- eta.x$fit + c(-1, 1) * 1.96 * eta.x$se.fit
> eta.ci
[1] 5.886818 6.053754
> mu.ci1 <- exp(eta.ci)
> mu.ci1
[1] 360.2573 425.7083
```


4.3 Residuals

Residuals are important in performing diagnostics about the fit of the model and for identifying potential outliers. Some of the uses of residuals include the following.

- Examining them to identify poorly fitted observations;
- plotting them to examine effects of potential new explanatory variables or nonlinear effects of those already in the fitted model;
- combining them into overall goodness-of-fit tests.

There are several ways to define the residuals in GLM, each used for a different purpose.

Response residuals

The response residuals,

$$_{\mathrm{R}}e_i := y_i - \hat{\mu}_i,$$

are central to model checking for the normal linear model. For the GLM, these residuals are neither normally distributed, nor do they have constant variance, for any distribution other than the normal and their use in GLM is limited.

Pearson residuals

The Pearson residuals are a scaled version of the response residuals:

$$_{\mathrm{P}}e_i := \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}.$$

Notice the use of the Pearson residuals in the estimation of the dispersion parameter in (4.6): $\hat{\phi} = (n-p)^{-1} \sum {}_{\mathrm{P}}e_i^2$.

The scaling puts the residuals on similar scales of variance so that standard plots of Pearson residuals versus the regressors or versus the dependent variable will reveal dependencies of the variance on these factors (if such dependencies exist). However for these purposes it is better from a theoretical point of view to use the deviance residuals (defined below).

Deviance residuals

The deviance residual is defined in the next section. Intuitively, the deviance residual measures the square-root contribution of the ith observation to the discrepancy of the GLM fit, called the deviance. In general, the deviance residual is preferred to the Pearson residual for model checking since its distributional properties are closer to the residuals arising in standard linear regression models. Thus the distribution of the residuals should be symmetric around 0 and a normal q-q plot of the deviance residuals should show a straight line if the fit is good and the distribution assumptions are correct.

If a particular residual is large, then observation i is contributing "too much" to the deviance, indicating a departure from the model assumptions for that observation or an outlier. The asymptotic distribution of the deviance residuals is normal with variance ϕ, therefore a residual greater than $(3.5)\sqrt{\phi}$ in absolute value is an indication that that observation is not explained by the model.

Typically deviance residuals are examined by plotting them against the explanatory variables to identify transformations of the explanatory variables that should be entered in the model.

Implementation in R.

The R function `residuals` (or its alias `resid`) can be used to return the residuals of the GLM fit. The general call to the function is as follows.

```
residuals(object, type = c("deviance", "pearson", "response"))
```

where

`object` Is the object containing the output from the call to the function `glm`.

`type` Is the type of residual required. The default is to return the deviance residual.

Example 4.5. Figure 4.1 shows two plots using the deviance residuals for Example 4.1. The left is the normal quantile plot and the one on the right is a plot of the residuals against the linear predictor. Both plots show that the fit is not particularly good. Specifically, the right plot suggests that a quadratic term should be included in the model. It can also be seen that two observations have particularly large residuals, $|e_i| > 3.5$. Based on this information we decide to add a quadratic term in the linear predictor. Figure 4.2 shows the fit with the linear and quadratic terms. Evidently the model with the quadratic term is a better fit. ▶

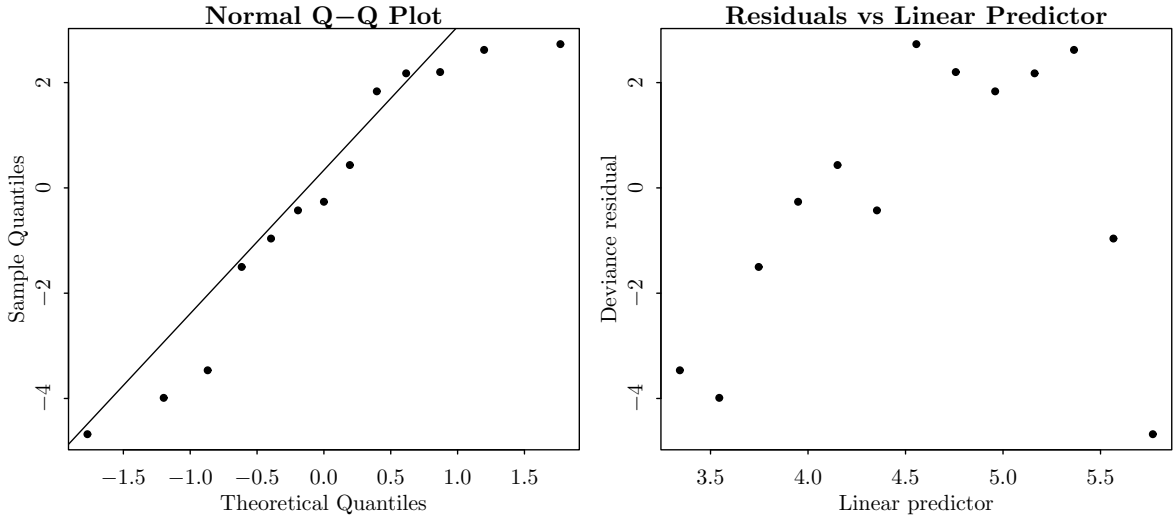


Figure 4.1: Residual plots for Example 4.5.

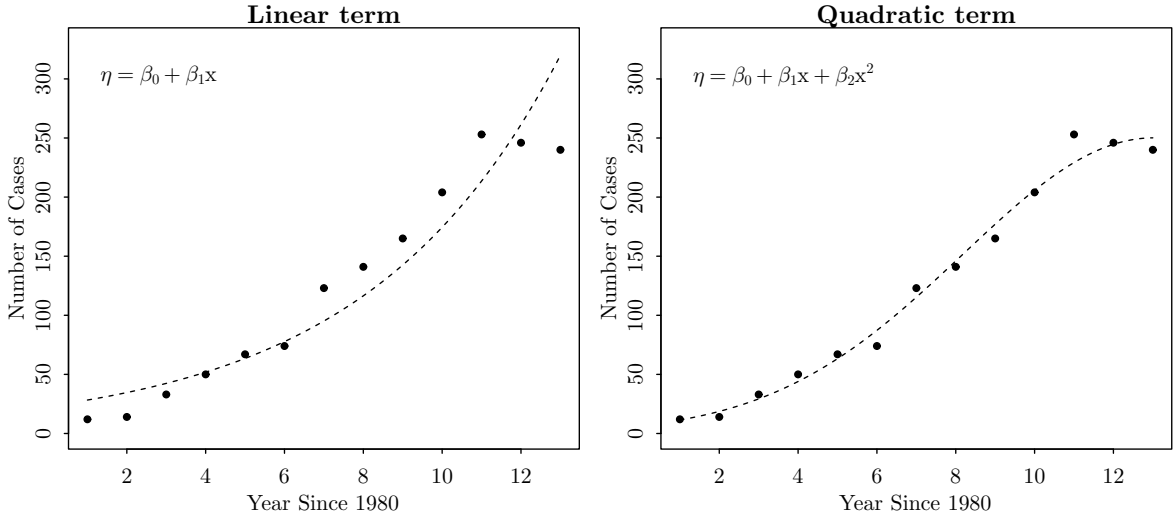


Figure 4.2: Fits for linear and quadratic terms for Example 4.5.

4.4 Goodness of fit tests and the deviance

The goodness of fit of a model to data is a natural question arising with all statistical modelling. Models are evaluated against two opposing criteria: model error and model complexity. A highly

complex model will have generally lower error than a simpler model. It is desirable to have a model with low error and low complexity.

One way of assessing the fit of a given model is to compare it against two extreme models: against the model with the lowest possible error and against the model with the lowest possible complexity. The lowest possible error will be obtained when there are as many parameters as observations ($p = n$): this is called a *saturated* or *full model*, however, this model also has the highest possible complexity and tends to overfit the data. On the other extreme, the model with the lowest complexity is called the *null model*: the model with no other explanatory variables but an intercept ($p = 1$), however this model has the highest error and tends to underfit. These models are represented in Figure 4.3. Figure 4.4 shows the null and full (saturated) models for the data of Example 4.1 along with the fitted model. Notice that the null model is represented by a horizontal line while the saturated model passes from every data point.

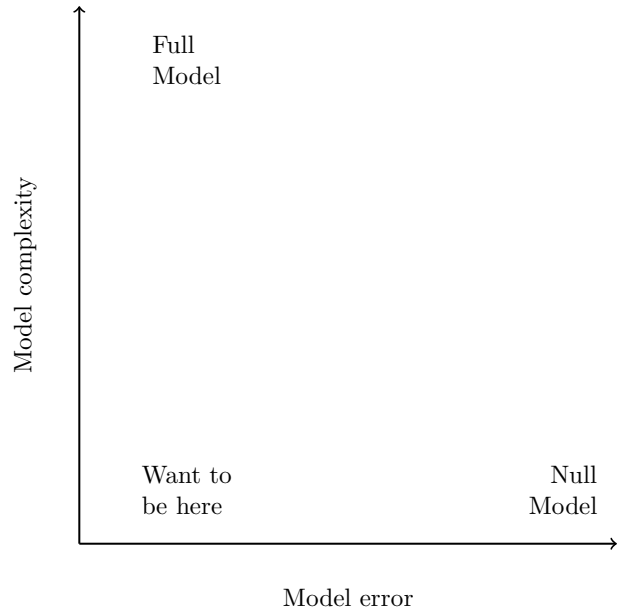


Figure 4.3: Illustration of model complexity and model error.

If the full model is significantly better than the fitted model, then there is still room for improving our model. On the other hand, if the fitted model is not significantly better than the null model, then the regressors do not contribute to the model fit so their overall significance is zero.

The value of the log-likelihood for the saturated model is (keeping only the terms that depend on θ)

$$\tilde{\ell} = \phi^{-1} \sum_{i=1}^{n} \left\{ y_i \tilde{\theta}_i - \psi(\tilde{\theta}_i) \right\},$$

which is the maximum possible likelihood attainable by any model. This value is compared to $\hat{\ell}$, the value of the log-likelihood based on the explanatory variables. Alternatively, let $\ell(\boldsymbol{\mu}|\boldsymbol{y})$ be the log-likelihood of the model expressed in terms of the mean $\boldsymbol{\mu}$ instead of $\boldsymbol{\beta}$. Then $\hat{\ell} = \ell(\hat{\boldsymbol{\mu}}|\boldsymbol{y})$ where $\hat{\boldsymbol{\mu}}$ is the predicted mean of $\boldsymbol{y}$ and $\tilde{\ell} = \ell(\boldsymbol{y}|\boldsymbol{y})$ since under the full model the predicted mean matches the data, i.e $\tilde{\mu}_i = y_i$.

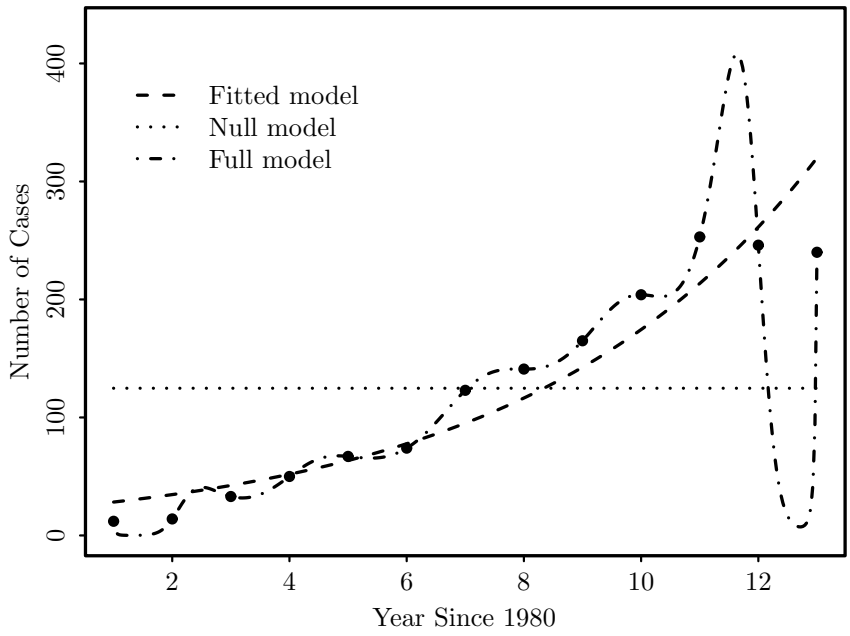


Figure 4.4: Fitted, null and saturated models for Example 4.1.

Definition 4.1 (Deviance)**.**
The *deviance*, denoted by Δ, is defined as the log-likelihood-ratio statistic (Section 1.4) between the full and fitted models, and is interpreted as a measure of the distance between these two models.

$$\Delta := 2(\tilde{\ell} - \hat{\ell}).$$

By Theorem 1.2

$$\Delta \dot{\sim} \mathcal{X}^2_{n-p}.$$

Table 4.1 shows the deviance for the common distributions. A direct calculation shows that

$$\Delta = \frac{1}{\phi} \sum_{i=1}^{n} 2 \left\{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - \psi(\tilde{\theta}_i) + \psi(\hat{\theta}_i) \right\}, \tag{4.15}$$

which, up to a potentially unknown constant factor $1/\phi$, is a sum of n terms. The square root of the ith term defines the ith deviance residual, e_i, such that

$$e_i = \text{sign}(y_i - \hat{\mu}_i)\sqrt{2} \left\{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - \psi(\tilde{\theta}_i) + \psi(\hat{\theta}_i) \right\}^{1/2},$$

so that $\Delta = \frac{1}{\phi} \sum e_i^2$. Note that we only omit unknown constant factors in the definition of the deviance residual. The term

$$D := \sum e_i^2$$

is called the *residual deviance* and is the analogue of the sum of squared residuals in classical linear models.

When the model provides a good fit, then $\hat{\ell}$ is expected to be close to (but not greater than) $\tilde{\ell}$. A large value of the deviance indicates a badly fitted model.

Normal	$(1/\sigma^2)\sum(y_i - \hat{\mu}_i)^2$
Poisson[a]	$\sum 2\{y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i\}$
Bernoulli[a]	$\sum 2\{y_i \log(y_i/\hat{\mu}_i) + (1-y_i)\log[(1-y_i)/(1-\hat{\mu}_i)]\}$
Gamma	$\vartheta \sum 2\{y_i/\hat{\mu}_i - 1 - \log(y_i/\hat{\mu}_i)\}$

Table 4.1: Deviance for common GLMs. [a]In the Poisson and Bernoulli cases, by convention $0 \log 0 = 0$.

In the case where the dispersion parameter ϕ is known, the deviance can be calculated and its value is compared against the $1-\alpha$ quantile of the $\mathcal{X}^2_{n-p}$ distribution. A deviance goodness-of-fit test would reject the fitted model at level α if

$$\Delta > \chi^2_{n-p;1-\alpha}.$$

Exercise 4.3. Derive the formulae for the deviance in Table 4.1. *Hint.* First write the log-likelihood in terms of y_i and μ_i. For $\hat{\ell}$ replace μ_i by $\hat{\mu}_i$ and for $\tilde{\ell}$ replace μ_i by y_i.

Implementation in R.
The R function `summary` outputs the residual deviance D and the residual null deviance D_0 along with their degrees of freedom as well as the value of the dispersion parameter ϕ. The functions `deviance` and `logLik` return the deviance and log-likelihood respectively of a fitted model.

Example 4.6. Suppose $y_i \overset{\text{ind}}{\sim} \text{Po}(\mu_i)$, $i = 1, \ldots, n$. Then

$$f(y_i; \mu_i) = \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i},$$

so the likelihood for $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ is

$$\begin{aligned} L(\boldsymbol{\mu}|\boldsymbol{y}) &= \prod_{i=1}^{n} \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i} \\ &= \prod_{i=1}^{n} \exp\{y_i \log \mu_i - \mu_i - \log y_i!\} \\ &= \exp\left\{\sum_{i=1}^{n} [y_i \log \mu_i - \mu_i] - \sum_{i=1}^{n} \log y_i!\right\} \\ \Rightarrow \ell(\boldsymbol{\mu}|\boldsymbol{y}) &= \sum_{i=1}^{n} [y_i \log \mu_i - \mu_i] - \sum_{i=1}^{n} \log y_i!. \end{aligned}$$

When the log-likelihood is evaluated at the estimated μ_i, it gives the log-likelihood of the fitted model, i.e.

$$\hat{\ell} = \ell(\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}|\boldsymbol{y}) = \sum_{i=1}^{n} [y_i \log \hat{\mu}_i - \hat{\mu}_i] - \sum_{i=1}^{n} \log y_i!.$$

For the full model we set $\tilde{\mu}_i = y_i$ because in this case the model has no error (see Figure 4.4), so

$$\tilde{\ell} = \ell(\boldsymbol{\mu} = \boldsymbol{y}|\boldsymbol{y}) = \sum_{i=1}^{n} [y_i \log y_i - y_i] - \sum_{i=1}^{n} \log y_i!.$$

Then,

$$\Delta = 2(\tilde{\ell} - \hat{\ell}) = 2\sum_{i=1}^{n} \left[y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i \right].$$

Continuing with Example 4.1, the fitted Poisson model with logarithmic link and a linear term $\eta_i = \beta_0 + \beta_1 x_i$ has residual deviance $D = 80.686$ so $\Delta = 80.686$. The degrees of freedom for the deviance are $n - p = 13 - 2 = 11$ and $\mathcal{X}^2_{11;0.95} = 19.68$. Because $\Delta > 19.68$, we reject this model as being significantly worse that the full model.

The model with the quadratic term $\eta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$ has residual deviance $D = 9.24$ so $\Delta = 9.24$ with degrees of freedom $n - p = 13 - 3 = 10$ and $\mathcal{X}^2_{10;0.95} = 18.31$. Because $\Delta < 18.31$, this model is not significantly worse than the full model. ▶

Proof of Theorem 4.1 (Optional). We will show that $(n-p)\hat{\phi}/\phi$ is approximately equal to the deviance. As the deviance is approximately $\mathcal{X}^2_{n-p}$ distributed, this will complete the proof.

From the relationship $y_i = \psi'(\tilde{\theta}_i)$, we have, by Taylor expansion around $\hat{\theta}_i$,

$$\begin{aligned} y_i &\approx \psi'(\hat{\theta}_i) + \psi''(\hat{\theta}_i)(\tilde{\theta}_i - \hat{\theta}_i) \\ &\approx \hat{\mu}_i + v(\hat{\mu}_i)(\tilde{\theta}_i - \hat{\theta}_i) \\ \Rightarrow \tilde{\theta}_i - \hat{\theta}_i &\approx \frac{y_i - \hat{\mu}_i}{v(\hat{\mu}_i)}. \end{aligned}$$

By another application of Taylor expansion around $\hat{\theta}_i$, we have

$$\begin{aligned} \psi(\tilde{\theta}_i) - \psi(\hat{\theta}_i) &\approx \psi'(\hat{\theta}_i)(\tilde{\theta}_i - \hat{\theta}_i) + \frac{1}{2}\psi''(\hat{\theta}_i)(\tilde{\theta}_i - \hat{\theta}_i)^2 \\ &\approx \hat{\mu}_i(\tilde{\theta}_i - \hat{\theta}_i) + \frac{1}{2}v(\hat{\mu}_i)(\tilde{\theta}_i - \hat{\theta}_i)^2. \end{aligned}$$

Substituting this into (4.15), we have

$$\begin{aligned} \Delta &\approx \frac{1}{\phi}\sum_{i=1}^{n} 2\left\{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - \hat{\mu}_i(\tilde{\theta}_i - \hat{\theta}_i) - \frac{1}{2}v(\hat{\mu}_i)(\tilde{\theta}_i - \hat{\theta}_i)^2 \right\} \\ &\approx \frac{1}{\phi}\sum_{i=1}^{n} 2\left\{ (y_i - \hat{\mu}_i)(\tilde{\theta}_i - \hat{\theta}_i) - \frac{1}{2}v(\hat{\mu}_i)(\tilde{\theta}_i - \hat{\theta}_i)^2 \right\} \\ &\approx \frac{1}{\phi}\sum_{i=1}^{n} 2\left\{ (y_i - \hat{\mu}_i)\frac{y_i - \hat{\mu}_i}{v(\hat{\mu}_i)} - \frac{1}{2}v(\hat{\mu}_i)\left(\frac{y_i - \hat{\mu}_i}{v(\hat{\mu}_i)}\right)^2 \right\} \\ &\approx \frac{1}{\phi}\sum_{i=1}^{n} 2\left\{ \frac{1}{2}\frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} \right\} \\ &= \frac{(n-p)\hat{\phi}}{\phi}, \end{aligned}$$

as required. □

4.5 Testing for the overall significance of the regressors

Comparing the fitted model against the null model is a way of assessing the significance of the regressor variables simultaneously. The deviance associated with the null model is called the *null deviance*, $\Delta_0 = \phi^{-1} D_0$ where D_0 is the corresponding residual null deviance. The null deviance is always larger than the model deviance Δ and their difference $\Delta_0 - \Delta$ is the likelihood ratio statistic for comparing the null and fitted models. To see this, let $\hat{\ell}_0$ be the maximum attainable log-likelihood under the null model. Then the likelihood ratio statistic is

$$2(\hat{\ell} - \hat{\ell}_0) = 2(\tilde{\ell} - \hat{\ell}_0) - 2(\tilde{\ell} - \hat{\ell}) = \Delta_0 - \Delta.$$

By the properties of the likelihood ratio statistic,

$$\Delta_0 - \Delta \dot{\sim} \mathcal{X}^2_{p-1}.$$

If ϕ is known, a level-α test for assessing the significance of the fitted model against the null model only would reject the null model if

$$\frac{D_0 - D}{\phi} > \chi^2_{p-1;1-\alpha}.$$

When ϕ is unknown, let $\hat{\phi}$ be its estimate under the fitted model. An F test can be constructed in this case as follows

$$F = \frac{\dfrac{D_0 - D}{\phi(p-1)}}{\dfrac{(n-p)\hat{\phi}}{\phi(n-p)}} = \frac{D_0 - D}{\hat{\phi}(p-1)} \dot{\sim} F_{p-1,n-p}$$

so in this case the null model is rejected at level α if

$$\frac{D_0 - D}{\hat{\phi}\,(p-1)} > F_{p-1,n-p;1-\alpha}. \tag{4.16}$$

4.6 Model selection

Let M denote the GLM with design matrix $\boldsymbol{X}$ of p regressor variables and M' be a nested model of M with design matrix $\boldsymbol{X}'$ of $p' < p$ regressor variables defined by deleting $p - p'$ columns from $\boldsymbol{X}$. We would like to know whether the reduced model M' is significantly different from M.

Another way to think about model selection is as follows. Suppose that the larger model M is fitted and let $\beta_{p'+1}, \ldots, \beta_p$ be the coefficients associated with the deleted variables in M. Then an equivalent hypothesis test is

$$H_0 : \beta_{p'+1} = \ldots = \beta_p = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0 \text{ for some } j \in \{p'+1, \ldots, p\}. \tag{4.17}$$

The above is a multiple hypothesis test about the regressor coefficients which can be performed using their asymptotic distribution (4.8). Formally this is called the deviance test, discussed below.

Deviance criterion

The deviance of the reduced model will be larger but if the increase in the residual deviance D' of the reduced model M' is not significant, then this is evidence in favour of M'. The statistic for testing the null hypothesis that model M' fits the data as well as model M is the difference between the two deviances: $\Delta' - \Delta = (D' - D)/\phi$ with asymptotic distribution

$$\frac{D' - D}{\phi} \dot{\sim} \mathcal{X}^2_{p-p'}$$

and the reduced model M' is rejected at level α if

$$\frac{D' - D}{\phi} > \mathcal{X}^2_{p-p';1-\alpha}.$$

If the dispersion parameter ϕ is unknown then by (4.7)

$$\frac{D' - D}{\hat{\phi}\,(p - p')} \dot{\sim} F_{p-p',n-p}$$

so a level-α test would reject the hypothesis that the reduced model M' is as good as model M if

$$\frac{D' - D}{\hat{\phi}\,(p - p')} > F_{p-p',n-p;1-\alpha}.$$

Remark: Note that the estimate $\hat{\phi}$ is the one derived from the larger model M. This is justified by considering the equivalent hypothesis test (4.17).

Akaike's Information Criterion (AIC)

The deviance criterion described earlier can be used to compare two nested models only, i.e. the smaller model is a special case of the larger model when some of the parameters of the larger model are fixed.

Akaike's information criterion (AIC) provides a measure for model selection that can be used more broadly. It can be considered as a aggregate measure of model error (measured by the deviance) and model complexity (measured by the number of parameters). Let d denote the number of estimated parameters, i.e., $d = p$ if ϕ is known, and $d = p + 1$ if ϕ is also estimated. The complexity measure is set to $2d$, so the aggregate measure of model error and model complexity is

$$\text{Aggregate measure} = \Delta + 2d = 2(\tilde{\ell} - \hat{\ell}) + 2d = -2\hat{\ell} + 2d + 2\tilde{\ell}.$$

The final term in the above equation does not change between models (as long as we use the same distribution), so we define AIC to be

$$\text{AIC} = -2\hat{\ell} + 2d.$$

The model with the smallest AIC is preferred.

Implementation in R.

The R function `anova` performs a deviance test to compare two or more nested models. The general input to the function is as follows.

```
anova(object, ..., dispersion = NULL, test = NULL)
```

where

object Is the object containing the output from the call to the function `glm`.

... Are objects containing additional models. Each subsequent model should be an extension of the previous model. If omitted then the regressors used in `object` are used to define the new models starting from the null model and sequentially adding one regressor at a time.

dispersion The dispersion parameter for the fitting family. By default it is obtained from the last entered object.

test Either `"Chisq"` or `"F"` depending on the type of test desired.

The functions `add1` and `drop1` are used for adding or dropping one term from the currently fitted model. The general call to these functions is as follows.

```
add1(object, scope, scale = 0, test = c("none", "Chisq", "F"))

drop1(object, scope, scale = 0, test = c("none", "Chisq", "F"))
```

where

object Is the object containing the output from the call to the function `glm`.

scope A formula giving the terms to be considered for adding or dropping.

scale Not used.

test As in the function `anova` above. If `"none"` then no test is performed but only the table containing the deviances and information criteria are returned.

The function `AIC` is used for computing the AIC criterion for one or several models. The general call to the functions is

```
AIC(object, ...)
```

where

object Is the object containing the output from the call to the function `glm`.

... Additional objects containing fitted models.

Example 4.7. Continuing from Example 4.5, we wish to test the goodness of fit of the model using the deviance. The residual deviance of the fitted model is found to be $D = 80.686$ and the null deviance $D_0 = 872.206$.

The test statistic for a deviance goodness of fit test is then $\Delta = D/\phi = 80.686$ at $n - p = 13 - 2 = 11$ degrees of freedom and its critical value at the 5% level is $\mathcal{X}^2_{0.95;11} = 19.68$. The conclusion then is to reject the fitted model.

Comparing the fitted model with the null model, we have $\dfrac{D_0 - D}{\phi} = 872.206 - 80.686 = 791.52$ at $p - 1 = 1$ degree of freedom. The critical value in this case is $\mathcal{X}^2_{0.95;1} = 3.84$ so the null model is rejected over the fitted model.

Suppose that we wish to add a quadratic term as the residual plots suggest. To that end, the residual deviance for the quadratic model is 9.240 and the test statistic for comparing our original model against the new model is $80.686 - 9.240 = 71.446$ at 1 degree of freedom (1 new parameter added to the model). The critical value is 3.84 which suggests that the model with the quadratic term is significantly better than the model without.

We may also observe that the model with the quadratic term passes the deviance goodness of fit test as the critical value in this case is $\mathcal{X}^2_{0.95;10} = 18.31$. ▶

4.7 Example: Oxygen consumption of frogs

We continue the example from section 2.3 where we have fitted a linear model to the data. In this section we will fit a GLM to the data with *RelativeIncrease* as a response and *Species* and *Temperature* as explanatory. Since the response variable is positive and continuous it makes sense to consider a distribution which has these properties such as the gamma or the inverse-Gaussian. The data are loaded into R using the following commands.

```
> frogs <- read.table("frogs.dat", header=TRUE)
> attach(frogs)
> RelativeIncrease <- (Exercise - Rest)/(Rest)
```

Let y_{ijk} denote the relative increase in consumption after exercise for the ith frog within the jth species and at kth temperature ($k = H$ for high temperature and $k = L$ for low temperature) for $i \in \{1, 2\}$, $j \in \{A, B, C, D\}$ and $k \in \{H, L\}$ and let μ_{ijk} be the corresponding mean of y_{ijk}. We choose to fit the model

$$y_{ijk} \overset{\text{ind}}{\sim} \mathrm{G}(\mu_{ijk}, \vartheta),\ \mu_{ijk} = \exp(\beta_0 + \alpha^S_j + \alpha^T_k),$$
$$\alpha^S_A = \alpha^T_H = 0$$

i.e. a gamma GLM with logarithmic link with α^S_j denoting the effect of the jth species and α^T_k the effect of the kth temperature. This model is defined and fitted in R using the following command.

```
> M1 <- RelativeIncrease ~ Species + Temperature
> glm1 <- glm(M1, family = Gamma(log))
> summary(glm1, correlation = TRUE)

Call:
glm(formula = M1, family = Gamma(log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2652  -0.1285   0.0278   0.1631   0.2219

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.9410     0.1050   -8.96  2.2e-06
SpeciesB         0.0871     0.1329    0.66    0.526
SpeciesC        -0.2958     0.1329   -2.23    0.048
SpeciesD        -0.3502     0.1329   -2.64    0.023
TemperatureLow   0.2118     0.0939    2.25    0.045

(Dispersion parameter for Gamma family taken to be 0.0353)

    Null deviance: 1.06560  on 15  degrees of freedom
Residual deviance: 0.39726  on 11  degrees of freedom
```

```
AIC: -33.25

Number of Fisher Scoring iterations: 5

Correlation of Coefficients:
               (Intercept) SpeciesB SpeciesC SpeciesD
SpeciesB       -0.63
SpeciesC       -0.63        0.50
SpeciesD       -0.63        0.50     0.50
TemperatureLow -0.45        0.00     0.00     0.00
```

In the first line of the output the R command executed to produce the model fit is displayed. R makes internal use of this command when it is requested to update a model.

In the next line it is displayed a five-number summary of the deviance residuals. Recall that if a particular residual is larger in absolute value than $(3.5)\sqrt{\phi}$ then the corresponding observation is not fitted well. In this example the dispersion parameter is estimated to be $\hat{\phi} = 0.0353$ so $(3.5)\sqrt{\hat{\phi}} = 0.658$ and the smallest and largest residuals are within this bound. Also notice that the five-number summary of the residuals is roughly symmetric as expected from a correct model. Therefore there is no indication of a bad fit using the residuals five-number summary.

Below the residual five-number summary the estimates of the parameters of the model are shown along with their standard errors, t-statistics and p-values for the significance of these parameters using the formulae in (4.12) and (4.13). From the table we conclude that both variables are significant at the 5% level.

The residual deviance D and the residual null deviance D_0 are displayed along with their degrees of freedom. This information can be used to assess the overall significance of the regressors. Comparing the fitted model against the null model we have, by (4.16), $F_{4,11;.95} = 3.357$ and the test statistic is $(1.0656 - 0.39726)/(0.0353 \times 4) = 4.733$ which suggests that the fitted model is an improvement over the null model. The model's AIC is also shown which is useful for model selection.

Next, the number of iterations of the Fisher scoring algorithm until convergence is displayed.

Finally, the correlation matrix for the parameter estimates is given. The matrix was requested by setting `correlation=TRUE` in the `summary` function. This is useful for inferring about linear combinations of the parameters, e.g. in the case of the linear predictor or the mean.

The significance of the fitted model is also verified by performing an F test on the models derived by removing each of the two explanatory variables:

```
> drop1(glm1, test = "F")
Single term deletions

Model:
RelativeIncrease ~ Species + Temperature
            Df Deviance   AIC F value Pr(>F)
<none>           0.397 -33.3
Species      3   0.948 -23.7    5.08  0.019
Temperature  1   0.573 -30.3    4.88  0.049
```

The table above shows that in terms of AIC and the F test, the model that contains both explanatory variables is significantly better than the models without one of these. On the other hand, *Temperature* is only narrowly significant in terms of the F test.

Two residual plots are shown next (Figure 4.5).

```
> e.d <- resid(glm1)
> eta <- predict(glm1)
> qqnorm(e.d)
> qqline(e.d)
> plot(eta, e.d, xlab = "Linear Predictor", ylab = "Deviance Residuals",
+      main = "Residuals vs Linear Predictor",
+      pch=ifelse(Temperature == "Low", tolower(Species), toupper(Species)))
> abline(h = c(-0.188, 0, 0.188), lty = 3)
```

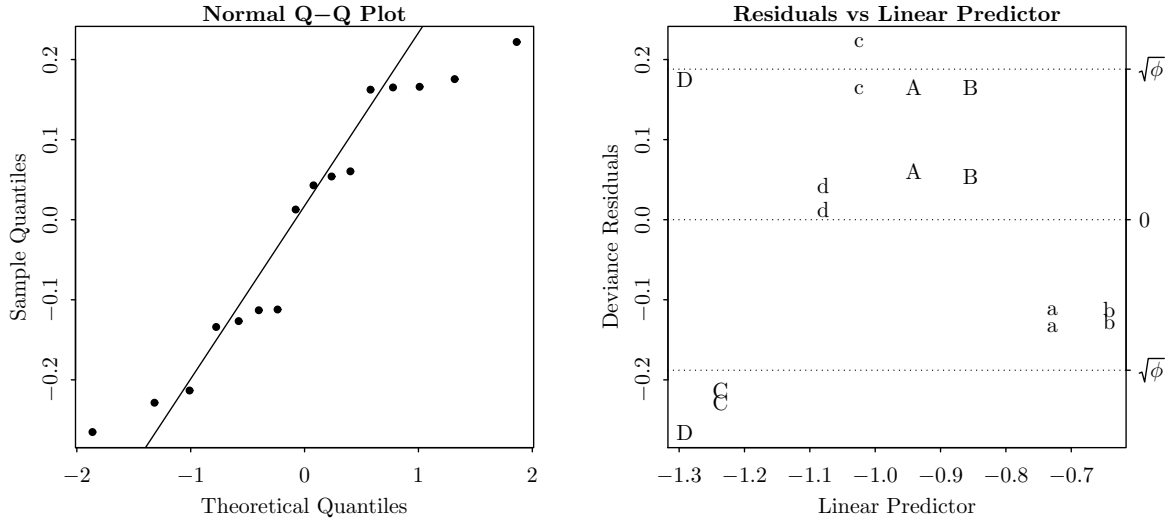


Figure 4.5: Residual plots for the frogs example. In the right-hand plot the letters denote the species and the upper/lower case denotes high/low temperature.

Both plots show that the model does not fit well. This indicates that perhaps a different distribution, other than the gamma may be more suitable.

With the currently fitted model, suppose that we wish to predict the oxygen consumption after exercise of a frog that belongs to species B at low temperature and with consumption at rest equal to 0.15ml/g/hr. The linear predictor has value $\hat{\eta}_{\boldsymbol{x}} = -0.9410 + 0.0871 + 0.2118 = -0.6421$. The variance of this estimate is $\mathsf{Var}\,\hat{\eta}_{\boldsymbol{x}} = (0.1050)^2 + (0.1329)^2 + (0.0939)^2 - 2(0.63)(0.1050)(0.1329) - 2(0.45)(0.1050)(0.0939) = (0.1051)^2$. The 97.5% quantile of the t_{11} distribution is 2.2 so a 95% confidence interval for $\eta_{\boldsymbol{x}}$ is $-0.6421 \pm (2.2)(0.1051) = (-0.87332, -0.41088)$

Since we are using the logarithmic link $\hat{\mu}_{\boldsymbol{x}} = \exp(\hat{\eta}_{\boldsymbol{x}}) = 0.5262$. So we estimate the frog's relative increase after exercise to be 0.5262. Since the frog's oxygen consumption at rest is 0.15ml/g/hr, this means its expected consumption after exercise is $(0.15)(1.5262) = 0.2289$ml/g/hr. A 95% confidence interval is obtained by transforming the confidence interval for $\eta_{\boldsymbol{x}}$: $((0.15) \times (1 + e^{-0.87332}), (0.15) \times (1 + e^{-0.41088})) = (0.213, 0.249)$.

Finally, we consider adding an interaction term to the model.

```
> glm2 <- update(glm1, . ~ . + Species:Temperature)
> deviance(glm2)
[1] 0.1129
> summary(glm2)$dispersion
```

```
[1] 0.01381
```

There are 3 more parameters (interaction terms) added to the model. The F statistic is $F = (0.3973 - 0.1129)/(3 \times 0.01381) = 0.6865$ with critical value 4.066 ($F_{3,8}$). The test fails to reject H_0 so the interaction term is not significant in this case. Note that the interaction was significant for the normal linear model which points to the differences between the two models.

4.7.1 Exponential model

The `summary` command accepts further input arguments. The input `dispersion` sets the dispersion parameter ϕ to a known value and is therefore not estimated. For example setting `dispersion=1` fits an exponential GLM since the exponential distribution is a special case of the gamma for $\phi = 1$. In this case the z test is used instead of the t-test.

```
> summary(glm1, correlation = TRUE, dispersion = 1)

Call:
glm(formula = M1, family = Gamma(log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.2652  -0.1285   0.0278   0.1631   0.2219

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.9410     0.5590   -1.68    0.092
SpeciesB         0.0871     0.7071    0.12    0.902
SpeciesC        -0.2958     0.7071   -0.42    0.676
SpeciesD        -0.3502     0.7071   -0.50    0.620
TemperatureLow   0.2118     0.5000    0.42    0.672

(Dispersion parameter for Gamma family taken to be 1)

    Null deviance: 1.06560  on 15  degrees of freedom
Residual deviance: 0.39726  on 11  degrees of freedom
AIC: -33.25

Number of Fisher Scoring iterations: 5

Correlation of Coefficients:
               (Intercept) SpeciesB SpeciesC SpeciesD
SpeciesB       -0.63
SpeciesC       -0.63        0.50
SpeciesD       -0.63        0.50     0.50
TemperatureLow -0.45        0.00     0.00     0.00
```

Notice the similarities and differences between the two models. The parameter estimates $\hat{\boldsymbol{\beta}}$, the correlations between components of $\hat{\boldsymbol{\beta}}$, residuals and therefore residual deviance are the same between the two models. This happens because these quantities don't depend on the value of ϕ. The standard errors for $\hat{\boldsymbol{\beta}}$ do depend on ϕ and in particular SE(Gamma Model) = $\sqrt{\hat{\phi}} \times$ SE(Exponential Model). Note that for the AIC printed above corresponds to the one of the gamma model and not the exponential model. Nevertheless, the AIC as it is may still be used for model comparison between different exponential models since their ordering is preserved.

We can perform a deviance goodness-of-fit test for the exponential model. The residual deviance is as before, $D = 0.397$ and $\phi = 1$ so $\Delta = 0.397$. Compare this against the 5% upper

quantile of a $\mathcal{X}^2_{11}$ which is 19.675. Therefore the deviance is significantly small to indicate good fit.

5. Models for Over-dispersed Data

In this chapter we discuss extensions of models with one-parameter distributions. The proposed models are more flexible and can better capture the variability in the data, making them suitable in circumstances where the variability of the response variable is different from the one assumed by the original one-parameter model. Recall that in a general exponential family distribution, there are two parameters, the canonical parameter θ, which is related to the mean of the distribution, through the relationship $\mathsf{E}\, y = \mu = \psi'(\theta)$, and the dispersion parameter, which is related to the variance of the distribution, through the relationship $\mathsf{Var}\, y = \phi\psi''(\theta) = \phi v(\mu)$. In one-parameter distributions, such as the Bernoulli and Poisson distributions, the dispersion parameter is set to 1 ($\phi = 1$), so $\mathsf{Var}\, y = v(\mu)$, therefore, the variance of the distribution is solely determined by its mean. These distributions lack flexibility in modelling the variance.

Extra variability can occur when the response variable is affected by variables that have not been considered or measured. Under idealised experimental conditions when successive events occur independently, the binomial and Poisson distributions are used to model the proportion of successes and the number of events observed respectively. However, even in well-conducted laboratory experiments, departures from these idealised models are to be expected for several reasons. For example Geiger counters experience a 'dead-time' following the arrival of a particle. During this short interval the apparatus is incapable of recording further particles. Consequently when the radioactive decay rate is high, the "dead-time" phenomenon leads to noticeable departures from the Poisson model. In mortality studies and insurance claims, person-specific variability or under-reporting leads to over-dispersion relative to the standard model.

5.1 Quasi likelihood model

In cases where inspections of the residual diagnostics show a good fit of the model to the data but the deviance goodness of fit test fails, a quasi likelihood model may be used.

A restriction of the binomial and Poisson models is that their variance is determined solely by their mean. For these distributions, the dispersion parameter has a known value $\phi = 1$, so $\mathsf{Var}(y) = v(\mu)$. This is sometimes a constraint of these models. However, as we have seen in Section 4.1.1, the value of the dispersion parameter does not affect the estimate for $\boldsymbol{\beta}$ so the MLE would be the same if we relax the constraint that $\phi = 1$ and assume that ϕ is unknown even though it is not. The dispersion parameter ϕ is subsequently estimated using the method of moments as discussed in Section 4.1.2.

The quasi-binomial and quasi-Poisson models are extensions of the standard binomial and Poisson models with an unknown dispersion. The MLE $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is computed first using the

Fisher scoring algorithm and the estimate $\hat{\phi}$ of the dispersion parameter ϕ is computed based on the estimate $\hat{\boldsymbol{\beta}}$. As we already mentioned, the estimate $\hat{\boldsymbol{\beta}}$ does not depend on the value of ϕ so one obtains the same $\hat{\boldsymbol{\beta}}$ regardless of whether the standard model or the quasi model is used, and therefore any quantity which depends only on $\hat{\boldsymbol{\beta}}$ is the same between the two models. However, the standard error of $\hat{\boldsymbol{\beta}}$ does depend on ϕ so the standard errors between the standard and quasi models differ.

On the other hand, by doing so we contradict our model because for values of ϕ other than the known $\phi = 1$ the pdf does not integrate to 1 (hence the term "quasi"). For this reason the model does not have a proper pdf. Consequently the AIC and other information criteria cannot be used for model selection. Fortunately, we can still use the likelihood ratio (deviance) tests since in the ratio the value of the integral of the pdf would cancel.

Table 5.1 shows which components stay the same and which change between the two models. In addition, the tests for the significance of the regressors coefficients which were compared against the $\mathrm{N}(0,1)$ distribution should consider the t_{n-p} distribution for the quasi model. Similarly, the deviance tests based on the $\mathcal{X}^2_k$ distribution, where k are the appropriate degrees of freedom, in the quasi version of the model they follow the $F_{k,n-p}$ distribution.

Same	**Change**
Estimates for regressor coefficients: $\hat{\boldsymbol{\beta}}_Q = \hat{\boldsymbol{\beta}}_S$	Standard errors: $\mathrm{SE}(\hat{\boldsymbol{\beta}}_Q) = \sqrt{\phi}\mathrm{SE}(\hat{\boldsymbol{\beta}}_S)$
Correlation: $\mathrm{Corr}(\hat{\boldsymbol{\beta}}_Q) = \mathrm{Corr}(\hat{\boldsymbol{\beta}}_S)$	
Response mean: $\mu_Q = \mu_S$	Variance of the linear predictor: $\mathsf{Var}(\hat{\eta}_Q) = \phi\,\mathsf{Var}(\hat{\eta}_S)$
Residuals (all types): $e_Q = e_S$	
Residual deviance: $D_Q = D_S$	Deviance: $\Delta_Q = \frac{1}{\phi}\Delta_S$
	AIC_Q does not exist

Table 5.1: Relationship between quantities associated with the standard and quasi models.

Since the value of ϕ does not affect the estimation of $\boldsymbol{\beta}$, we get the same $\hat{\boldsymbol{\beta}}$ regardless of whether we fit the standard or the quasi model. However, from (4.8), $\mathsf{Var}\,\hat{\boldsymbol{\beta}} = \phi(\boldsymbol{X}^\top V^{-1}\boldsymbol{X})^{-1}$ so for the standard model where $\phi = 1$, $\mathsf{Var}\,\hat{\boldsymbol{\beta}}_S = (\boldsymbol{X}^\top V^{-1}\boldsymbol{X})^{-1}$ and for the quasi model $\mathsf{Var}\,\hat{\boldsymbol{\beta}}_Q = \hat{\phi}(\boldsymbol{X}^\top V^{-1}\boldsymbol{X})^{-1}$. In other words the variance of $\hat{\boldsymbol{\beta}}$ for the quasi model differs from the variance of the standard model by a factor of $\hat{\phi}$, i.e.

$$\mathsf{Var}(\text{quasi model}) = \hat{\phi} \times \mathsf{Var}(\text{standard model}).$$

Implementation in R.
In R the families `quasibinomial` and `quasipoisson` can be used to fit the corresponding quasi model. These families can be used as the family argument of the function `glm`.

The `summary` function can be used to print the results of the GLM fit. In the case of the quasi model the output is almost identical to the output of the standard model with the only difference that the standard errors are now multiplied by a factor of $\sqrt{\phi}$. In addition to this the test statistics for the significance of the regressor coefficients and p-values correspond to the t_{n-p} distribution. Moreover, the value of AIC is not valid.

Example 5.1. The data in this example were obtained from the Ames Salmonella/microsome mutagenicity assay. The response variable (denoted by `m`) corresponds to the number of revertant

colonies of the salmonella bacteria after given a certain dosage of quinoline (denoted by `dose`). The data are shown in Table 5.2.

dose	m	dose	m	dose	m
0	15	33	16	333	33
0	21	33	26	333	38
0	29	33	33	333	41
10	16	100	27	1000	20
10	18	100	41	1000	27
10	21	100	60	1000	42

Table 5.2: The salmonella data set.

We observe that dose increases exponentially. To reduce the influence of the largest dosages, we consider the logarithmic transformation of dosage, so we let $x_i = \log(1 + \text{dose}_i)$. Also let y_i denote the number of revertant colonies. We fit the following model

$$y_i \overset{\text{ind}}{\sim} \text{Po}(\mu_i),\ i = 1, \ldots, n = 18$$
$$\log(\mu_i) = \beta_0 + \beta_1 x_i.$$

Examination of the residuals indicates that the 12th observation is an outlier, thus we refit the model with that observation deleted. The following code illustrates the model fit.

```
> fit2 <- glm(m ~ log(1+dose), poisson(log), salmonella, subset = -12)
> summary(fit2)

Call:
glm(formula = m ~ log(1 + dose), family = poisson(log), data = salmonella,
    subset = -12)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.9764     0.0999   29.80  < 2e-16
log(1 + dose)  0.0816     0.0208    3.93  8.6e-05

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 51.379  on 16  degrees of freedom
Residual deviance: 35.492  on 15  degrees of freedom
AIC: 126.1

Number of Fisher Scoring iterations: 4
```

We note that the deviance of the model is $\Delta = 35.492$, while the corresponding chi-squared quantile is $\mathcal{X}^2_{15;0.95} = 25$. Therefore, the fitted model fails the deviance goodness-of-fit test. We can examine the residuals to see if there are any issues. A normal q-q plot and a plot of the residuals against the linear predictor is shown in Figure 5.1. We can see from the plot that the residuals are approximately normally distributed, which indicates good fit of the model, and that the plot of the residuals against the linear predictor does not show any pattern, so we don't need to include any transformations in the model.

As the model fails the deviance goodness-of-fit test, we then decide to fit a quasi Poisson model to the same data.

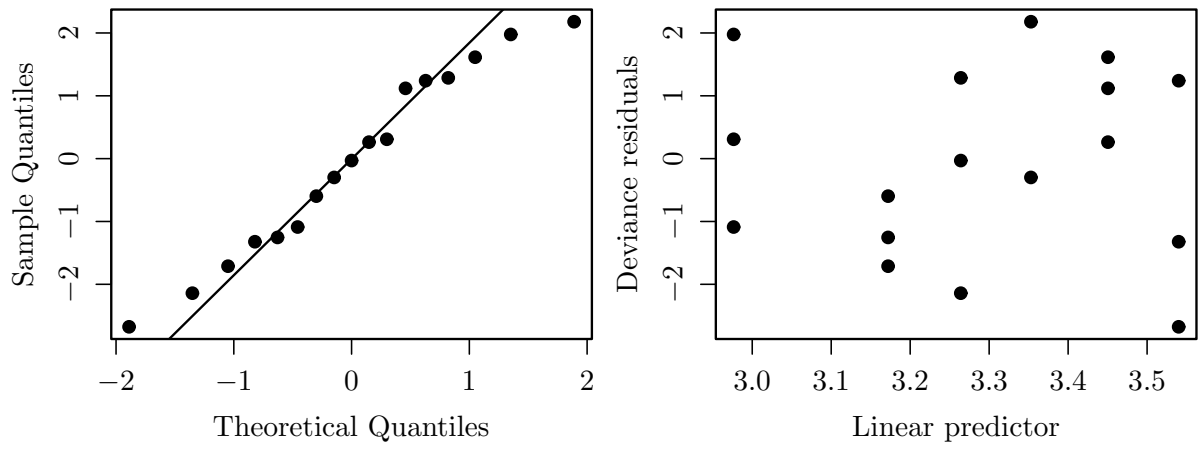


Figure 5.1: Residual plots for Example 5.1.

```
> fit3 <- glm(m ~ log(1+dose), quasipoisson(log), salmonella, subset = -12)
> summary(fit3)

Call:
glm(formula = m ~ log(1 + dose), family = quasipoisson(log),
    data = salmonella, subset = -12)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.9764     0.1522   19.55  4.4e-12
log(1 + dose)   0.0816     0.0317    2.58    0.021

(Dispersion parameter for quasipoisson family taken to be 2.323)

    Null deviance: 51.379  on 16  degrees of freedom
Residual deviance: 35.492  on 15  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

We observe that, in the quasi Poisson model, we obtain an estimate for the dispersion parameter $\hat{\phi} = 2.323$, while for the standard Poisson model $\phi = 1$. The estimates for the regressor parameters remain the same, however the standard errors for the quasi Poisson model are obtain by multiplying the standard errors of the standard Poisson model by $\sqrt{\hat{\phi}}$. The p-value of the explanatory variable is calculated based on the t distribution with 15 degrees of freedom and is equal to 0.021, so the explanatory variable is significant at the 5% level. ▶

Example 5.2. Recall the toxicity data of Example 3.6 where a binomial GLM with probit link was fitted. A quasi-binomial model is defined as

$$y_i \overset{\text{ind}}{\sim} \text{QBin}(m_i, \pi, \phi), \ i = 1, \ldots, 5$$
$$\pi_i = \Phi(\beta_0 + \beta_1 x_i)$$

where y_i denotes the number of stillborn foetuses, m_i the total number of mice, and x_i the dosage for the ith cohort. The function Φ denotes the CDF of the standard normal distribution implied by the probit link. The following R commands illustrate the fitting of this model.

```
> stillborn <- c(15,17,22,38,144)
> total <- c(297,242,312,299,285)
> concentration <- c(0,62.5,125,250,500)
> fit2 <- glm(cbind(stillborn,total-stillborn) ~ concentration,
+             family=quasibinomial(probit))
> summary(fit2)

Deviance Residuals:
     1       2       3       4       5
 1.456   1.039  -0.843  -2.189   0.987

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.823574   0.138449   -13.2  0.00095
concentration  0.003527   0.000425     8.3  0.00366

(Dispersion parameter for quasibinomial family taken to be 3.226)

    Null deviance: 259.1073  on 4  degrees of freedom
Residual deviance:   9.6742  on 3  degrees of freedom
AIC: NA
```

For this model $\hat{\phi} = 3.226$.

An F test for the significance of the fitted model against the null model gives $F = \frac{259.1073-9.6742}{(3.226)1} = 77.32$ with critical value $F_{1,3;0.95} = 10.13$ so the null model is rejected. ▶

5.2 Compound model

The quasi models generalise the mean-variance relationship by imposing a multiplicative factor ϕ to the relationship derived from the standard model, i.e. $\mathsf{Var}\, y = \phi v(\mu)$. A different way of introducing additional variability is by letting the mean be random instead of a fixed parameter. This introduces a different mean-variance relationship than the quasi model of the form $\mathsf{Var}\, y = w(\mu, \vartheta)$, where ϑ is a new parameter related to the distribution of the mean.

Let $\mathrm{EF}(\mu)$ denote the standard binomial or Poisson distribution with mean μ and known dispersion parameter. The compound model assumes that $\mu = z$ where z is a random variable with pdf $f(z)$. In other words, conditioned on the value of z, y is a random sample from the distribution $\mathrm{EF}(z)$. The joint distribution of (y, z) can be expressed hierarchically as follows

$$\begin{aligned} y|z &\sim \mathrm{EF}(z) \\ z &\sim f(z). \end{aligned} \tag{5.1}$$

In this case the marginal distribution of y is not necessarily in the exponential family and, as z is unobserved, in order to fit a distribution defined by (5.1) we need the marginal density $f(y)$.

Using the rule of conditional probability, the conditional distribution of $y|z$ is

$$f(y|z) = \frac{f(y,z)}{f(z)}, \tag{5.2}$$

in other words, the ratio of the joint pdf $f(y, z)$ over the marginal pdf of the conditioned variable $f(z)$. The marginal distribution is obtained from the joint distribution by integrating with respect to the variable we wish to eliminate. Thus,

$$f(y) = \int f(y,z)\,\mathrm{d}z$$

$$= \int f(y|z)\, f(z)\, \mathrm{d}z \tag{5.3}$$

from (5.2). The marginal distribution of y is said to be the distribution that results from compounding the conditional distribution of $y|z$ with the distribution of z. It is not generally possible to derive a closed-form expression for the integral in (5.3) so exact maximum likelihood inference is in most situations hampered by the need to evaluate (5.3) numerically. In two special cases, the so-called gamma-Poisson (aka negative binomial) and the beta-binomial models, we are able to write down the likelihood function in closed form and subsequently fit the model by MLE.

5.2.1 Negative binomial model

The negative binomial model is a generalisation of the Poisson model by assuming that the mean is a gamma distributed random variable. In other words we assume the following model

$$y|z \sim \mathrm{Po}(z), \tag{5.4}$$
$$z \sim \mathrm{G}(\mu, \vartheta). \tag{5.5}$$

Equation (5.4) says that the response is Poisson distributed with mean z and equation (5.5) says that z is gamma distributed with mean μ and shape parameter ϑ.

As we derive below the marginal distribution of y is the negative binomial distribution with success probability $\pi = \frac{\mu}{\mu+\vartheta}$ and shape $\kappa = \vartheta$. Recall from Appendix A.2.3, the negative binomial is a discrete distribution with pdf

$$f(x) = \frac{\Gamma(x+\kappa)}{x!\Gamma(\kappa)}(1-\pi)^{\kappa}\pi^{x};\ x = 0, 1, \ldots. \tag{5.6}$$

The mean of (5.6) is $\frac{\pi\,\kappa}{1-\pi}$ and its variance is $\frac{\pi\,\kappa}{(1-\pi)^2}$.

By (5.5), the pdf of z is given by (see section A.3.2)

$$f(z) = \frac{\vartheta^{\vartheta}}{\mu^{\vartheta}\Gamma(\vartheta)}\, z^{\vartheta-1}\, \exp\left(-\frac{\vartheta}{\mu}\, z\right).$$

As a side note, the fact that $\int f(z)\, \mathrm{d}z = 1$ suggests that

$$\int z^{\vartheta-1}\, \exp\left(-\frac{\vartheta}{\mu}\, z\right)\, \mathrm{d}z = \Gamma(\vartheta)\left(\frac{\vartheta}{\mu}\right)^{-\vartheta}. \tag{5.7}$$

The conditional pdf of $y|z$ is

$$f(y|z) = \frac{z^y}{y!}e^{-z}$$

so the joint pdf from (5.2) is

$$f(y, z) = \frac{\vartheta^{\vartheta}}{\mu^{\vartheta}\,\Gamma(\vartheta)\,y!}\, z^{\vartheta+y-1}\, \exp\left\{-\left(\frac{\vartheta}{\mu}+1\right)\, z\right\}$$

and the marginal distribution of y, from (5.3),

$$f(y) = \int \frac{\vartheta^{\vartheta}}{\mu^{\vartheta}\,\Gamma(\vartheta)\,y!}\, z^{\vartheta+y-1}\, \exp\left\{-\left(\frac{\vartheta}{\mu}+1\right)\, z\right\}\, \mathrm{d}z$$

$$
\begin{aligned}
&= \frac{\vartheta^\vartheta}{\mu^\vartheta\,\Gamma(\vartheta)\,y!}\int z^{\vartheta+y-1}\,\exp\left\{-\left(\frac{\vartheta}{\mu}+1\right)\,z\right\}\,\mathrm{d}z \\
&= \frac{\Gamma(\vartheta+y)}{\Gamma(\vartheta)\,y!}\frac{\vartheta^\vartheta}{\mu^\vartheta}\left(\frac{\vartheta}{\mu}+1\right)^{-\vartheta-y} \qquad \text{from (5.7)} \\
&= \frac{\Gamma(\vartheta+y)}{\Gamma(\vartheta)\,y!}\frac{\vartheta^\vartheta}{\mu^\vartheta}\left(\frac{\mu}{\mu+\vartheta}\right)^{\vartheta+y} \\
&= \frac{\Gamma(\vartheta+y)}{\Gamma(\vartheta)\,y!}\left(\frac{\mu}{\mu+\vartheta}\right)^{y}\left(\frac{\vartheta}{\mu+\vartheta}\right)^{\vartheta} \\
&= \frac{\Gamma(\vartheta+y)}{\Gamma(\vartheta)\,y!}\left(\frac{\mu}{\mu+\vartheta}\right)^{y}\left(1-\frac{\mu}{\mu+\vartheta}\right)^{\vartheta}. \qquad (5.8)
\end{aligned}
$$

The last expression is of the form of the negative binomial distribution (5.6) for $\kappa = \vartheta$ and $\pi = \frac{\mu}{\mu+\vartheta}$. Then, using the formula for the mean and variance of the negative binomial distribution,

$$
\begin{aligned}
\mathsf{E}\,y &= \frac{\vartheta\,\frac{\mu}{\mu+\vartheta}}{1-\frac{\mu}{\mu+\vartheta}} = \mu \\
\mathsf{Var}\,y &= \frac{\vartheta\,\frac{\mu}{\mu+\vartheta}}{\left(1-\frac{\mu}{\mu+\vartheta}\right)^2} = \mu + \vartheta^{-1}\,\mu^2
\end{aligned}
$$

so the mean is μ but the variance is now larger than μ. Note that when $\vartheta = \infty$, $\mathsf{Var}\,z = 0$ and z takes the constant value μ. In this case we fall back to the Poisson model with mean μ.

Negative binomial GLM

When ϑ is known, equation (5.8) can be written in the form of the exponential family. To that end, let $c(y) = \log\frac{\Gamma(\vartheta+y)}{\Gamma(\vartheta)\,y!}$ and $\theta = \log\frac{\mu}{\mu+\vartheta}$. Then

$$
f(y) = \exp\Big\{y\,\theta + \vartheta\,\log(1-e^\theta) + c(y)\Big\}
$$

so that $\psi(\theta) = -\vartheta\,\log(1-e^\theta)$ and $\phi = 1$. However, the case where ϑ is unknown is not a GLM.

The negative binomial GLM for known ϑ, link function $g(\cdot)$, and observations $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ is defined as follows

$$
\begin{aligned}
y_i &\sim \mathrm{NegBin}(\pi_i, \vartheta) \\
g(\mu_i) &= \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}
\end{aligned} \qquad (5.9)
$$

The link function g is as in the Poisson model. Estimation of the parameters $\boldsymbol{\beta}$ is performed as in the other GLMs, namely by the iterative Fisher scoring algorithm.

In the standard negative binomial GLM, the dispersion parameter is known, $\phi = 1$. We can extend (5.9) to a quasi-negative binomial GLM by assuming that ϕ is unknown as with the standard binomial and Poisson models, therefore allowing for two types of overdispersion relative to the Poisson model.

As we have already mentioned the case where ϑ is not known the above model is not a GLM. We can, nevertheless, estimate the parameter ϑ by MLE simultaneously with the estimation of $\boldsymbol{\beta}$.

Implementation in R.
The familiar `glm` function can be used for fitting the negative binomial and quasi-negative binomial models with known ϑ. The family is not available in R by default but only through the

MASS package.[a] The package is loaded into R by executing the command `library(MASS)`. Then the `negative.binomial` family can be used inside the `glm` function. The format of the `negative.binomial` family is as follows

```
negative.binomial(theta, link = "log")
```

where

`theta` The known value of the shape parameter, ϑ.

`link` The link function. Can be one of `log` (the default), `sqrt` or `identity`.

The `summary` function (see section 4.2.1) prints the results of the model fit. The output of the `summary` function corresponds to the quasi-negative-binomial model. For the standard negative binomial model we need to specify `dispersion=1` as an argument to the function.

The case where ϑ is also unknown is handled by a different function, called `glm.nb`. The format of the function is as follows

```
glm.nb(formula, data, weights, subset, start = NULL,
       etastart, mustart, control = glm.control(...), ...,
       init.theta, link = log)
```

where

`formula, data, weights, subset, start, etastart, mustart, control, ...` are the same arguments as in the `glm` function (see section 4.1.3).

`init.theta` Optional initial value for the theta parameter.

`link` The link function. Must be one of `log` (the default), `sqrt` or `identity`.

The function `summary` may be used to print the results of the fit of the function.

Residuals for Poisson v.s. negative binomial (Optional)

In this section we investigate what happens to the residuals if we wrongly fit a Poisson model to over-dispersed data. This will help us detect when a negative binomial model is appropriate after fitting a Poisson model.

First note the following property of the gamma $\mathrm{G}(\mu, \vartheta)$ distribution. If $X \sim \mathrm{G}(\mu, \vartheta)$ and a is a scalar, then $Y = a\,X$ is also gamma distributed and $Y \sim \mathrm{G}(a\,\mu, \vartheta)$. Thus we can write $z \sim \mathrm{G}(\mu, \vartheta)$ of the gamma-Poisson model as $z = \mu\,w$ where $w \sim \mathrm{G}(1, \vartheta)$. With this modification the model becomes

$$\begin{aligned} y &\sim \mathrm{Po}(\mu\,w) \\ w &\sim \mathrm{G}(1, \vartheta) \\ \log \mu &= \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}. \end{aligned} \tag{5.10}$$

Suppose we observe w and let $\mu' = \mu\,w$. The model in (5.10) can be written as

$$\begin{aligned} y &\sim \mathrm{Po}(\mu') \\ \log \mu' &= \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta} + \log w. \end{aligned} \tag{5.11}$$

[a] Packages are addons to the default R installation which provide additional functions and data. These are loaded into R using the command `library(`*`package_name`*`)`

so $\log w$ plays the role of an offset. The difference between the Poisson and negative binomial models is that the former ignores the term $\log w$ in (5.11).

Let e_P, e_{NB} be the deviance residual of the Poisson and negative binomial models respectively. Then

$$e_P^2 = 2\left\{y \log \frac{y}{\mu} - y + \mu\right\}$$

where $\log \mu = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$, and

$$\begin{aligned} e_{NB}^2 &= 2\left\{y \log \frac{y}{\mu'} - y + \mu'\right\} \\ &= 2\left\{y \log \frac{y}{\mu\, w} - y + \mu\, w\right\} \\ &= 2\left\{y \log \frac{y}{\mu} - y \log w - y + \mu\, w\right\} \end{aligned}$$

so

$$e_{NB}^2 - e_P^2 = 2\left\{-y \log w + \mu\,(w-1)\right\}. \tag{5.12}$$

Taking expectations on both sides of (5.12), and noting that for the correct negative binomial model $e_{NB} \dot{\sim} \mathrm{N}(0,1)$, $\mathsf{E}\, w = 1$ and $\mathsf{E}(y \log w) = \mu\, \mathsf{E}(w \log w) = \mu\, \xi$ where $\xi = \mathsf{E}(w \log w)$ does not depend on μ[b] we have

$$\mathsf{Var}\, e_P = 1 + 2\,\xi\,\mu = 1 + 2\,\xi\, \exp(\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}). \tag{5.13}$$

Equation (5.13) suggests that if we plot the deviance residuals of the Poisson model against the linear predictor we should observe an exponential increase in the variability of the residuals if the correct model is the negative binomial.

This behaviour is verified by the simulation study below. In this setting we simulate $n = 1001$ observations with mean $\mu = \exp(\beta_0 + \beta_1\, x)$ where $\beta_0 = -1$, $\beta_1 = 5$, and $x = 0, 0.002, 0.004, \ldots, 2$ as shown below.

```
> ### Create some data
> n <- 1001                              # Sample size
> b0 <- -1                               # Intercept
> b1 <- 5                                # Coefficient
> x <- seq(0, 2, length=n)               # Regressor
> mu <- exp(b0 + b1*x)                   # Mean
```

First we show how the deviance residuals behave if the correct model is used. To that end we simulate from the Poisson model and pretending we don't know β_0 and β_1 and try to estimate them. The plot of the residuals against the linear predictor (Figure 5.2, left) shows a fairly random pattern (the fact that there is some pattern at the low values of the linear predictor is due to the fact that the corresponding mean is very low and the response takes only the values 0, 1, and 2 but is not and indication of a bad fit) and the normal quantile plot (Figure 5.2, right) verifies the fact that the deviance residual follow approximately the standard normal distribution.

[b] In particular $\xi = \vartheta^{\vartheta-2}(1 - \gamma - \log \vartheta)/\Gamma(\vartheta)$ where $\gamma \approx 0.577$ is the Euler's constant.

```
> ### Poisson counts
> y1 <- rpois(n,mu)                    # Poisson counts
> glm1 <- glm(y1 ~ x, family=poisson) # Fit (correct model)
> e1 <- resid(glm1)                    # Deviance residuals
> eta1 <- predict(glm1)                # Linear predictor
> plot(eta1, e1)                       # Residuals vs Linear predictor
> qqnorm(e1)                           # Normal qq plot
```

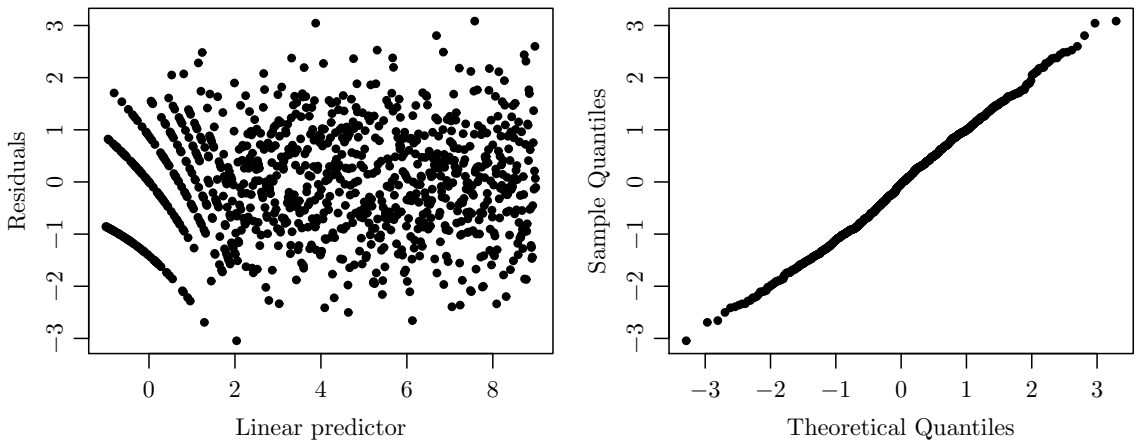


Figure 5.2: Residual plots for the correct Poisson model. Against linear predictor (left) and normal quantile plot (right).

Now suppose that the data follow the gamma-Poisson (negative binomial) model and we wrongly choose to fit a Poisson model. This is fitted using the commands below. In this case the plot of the residuals against the linear predictor (Figure 5.3, left) shows that the variability is increasing exponentially and the normal quantile plot shows that the normal distribution (Figure 5.3, right) is not valid.

```
> ### Negative binomial counts modelled as Poisson
> theta <- 1                           # Shape parameter
> w2 <- rgamma(n, shape=theta)         # Simulate from gamma
> z2 <- mu*w2                          # Mean
> y2 <- rpois(n, z2)                   # Negative binomial counts
> glm2 <- glm(y2 ~ x, family=poisson) # Fit (wrong model)
> e2 <- resid(glm2)                    # Deviance residuals
> eta2 <- predict(glm2)                # Linear predictor
> plot(eta2, e2)                       # Residuals vs Linear predictor
> qqnorm(e2)                           # Normal qq plot
```

5.2.2 Beta binomial model

The beta binomial model is the analog to the gamma-Poisson model for over-dispersed binomial data. In this model we assume that the probability of success is a random variable z having a beta distribution, i.e. for $\vartheta > 0$ and $0 < \pi < 1$,

$$y|z \sim \text{Bin}(m, z) \tag{5.14}$$

$$z \sim \text{Beta}(\vartheta\pi, \vartheta(1-\pi)) \tag{5.15}$$

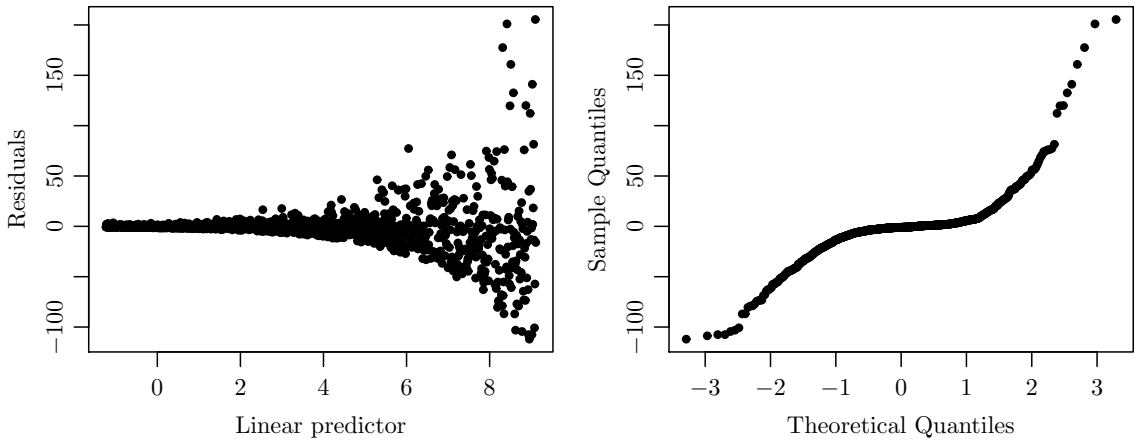


Figure 5.3: Residual plots for the wrong Poisson model. Against linear predictor (left) and normal quantile plot (right).

so that $\mathsf{E}\, z = \pi$ and $\mathsf{Var}\, z = \pi(1-\pi)/(\vartheta+1)$ where the probability density function of the beta distribution is given by

$$f(z;a,b) = \frac{1}{B(a,b)} z^{a-1}(1-z)^{b-1};\ z \in (0,1),\ a,b>0$$

and $B(a,b)$ denotes the *beta function*

$$B(a,b) := \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

It turns out that the marginal mean and variance of y are

$$\begin{aligned}\mathsf{E}\, y &= m\pi \\ \mathsf{Var}\, y &= m\pi(1-\pi)(1+(m-1)/(\vartheta+1)).\end{aligned}$$

The marginal distribution of y is called the *beta binomial distribution* and its probability mass function is given by

$$f(y;m,\vartheta,\pi) = \binom{m}{y} \frac{\mathrm{B}(y+\vartheta\pi, m-y+\vartheta(1-\pi))}{\mathrm{B}(\vartheta\pi,\vartheta(1-\pi))},\ y=0,1,\ldots,m$$

which is in general not a member of the exponential family. Parameter estimation for the beta binomial model is done using the maximum likelihood method.

Implementation in R (Optional)**.**
As the beta binomial model is not a GLM, we cannot use the `glm` function to fit such model. The function `aodml` provided by the package `aods3` may be used to fit the model. The same function may be used to fit a negative binomial model as well. The format of the function is as follows

```
aodml(formula, data, family = c("bb", "nb"),
      link = c("logit", "cloglog", "probit", "log"))
```

where

formula, data are the same as in the `glm` function (see section 4.1.3).

family The assumed distribution of the response which can be either beta binomial or negative binomial.

link The link function.

The `summary` function can be used to print the results from the model fit. The function defines a parameter $\varphi = 1/(\vartheta + 1)$ and estimates φ instead of ϑ. Then $\hat{\vartheta} = \hat{\varphi}^{-1} - 1$.

5.2.3 Testing for the significance of the compound model

When the variance of the assumed distribution of z is 0, the mean is not longer random but a constant so the compound model reduces to the standard GLM. Recall Chebyshev's inequality: $\forall\, \varepsilon > 0,\ \Pr(|z - \mathsf{E}\, z| > \varepsilon) \leq \varepsilon^{-2}\, \mathsf{Var}(z)$, so as $\mathsf{Var}(z) \to 0$, $z \to \mathsf{E}\, z$ almost surely.

In the case of the negative binomial model, the assumed $\mathrm{G}(\mu, \vartheta)$ distribution for z in (5.5) has mean μ and variance μ^2/ϑ so as $\vartheta \to \infty$, $z \to \mu$ almost surely. In this case (5.4) becomes the standard Poisson distribution.

Similarly, in (5.15), if $\vartheta \to \infty$, $z \to \pi$ almost surely and (5.14) becomes the standard binomial distribution.

In both cases, a very large estimated ϑ suggests that the standard model for the response would suffice. A formal test of H_0: $\vartheta = \infty$ v.s. H_1: $\vartheta < \infty$ is based on the likelihood ratio test. Since $\vartheta = \infty$ is at the boundary of the possible range $\vartheta \leq \infty$, the distribution of the test statistic is non-standard and requires care. If the data are generated from the Poisson model, then half of the time, i.e., with probability 0.5, the likelihood will be unbounded, and the MLE for ϑ is ∞. Therefore, under the null hypothesis, the likelihood ratio test statistic, $\Lambda = 2(\hat{\ell}_C - \hat{\ell}_S)$, where $\hat{\ell}_C$ and $\hat{\ell}_S$ are the maximum values of the log-likelihood under the compound and standard models respectively, will attain the value 0 with probability 0.5, and, otherwise, it will be distributed according to the $\mathcal{X}^2_1$ distribution. We say that distribution of the test statistic has a mass of 0.5 at zero, and a half-$\mathcal{X}^2_1$ distribution above zero. This distribution is shown in Figure 5.4. A test with a significance level α, requires a rejection region corresponding to the upper 2α point of the $\mathcal{X}^2_1$ distribution. For a 5% level test, the critical value corresponds to $\mathcal{X}^2_{1;0.90} = 2.71$.

5.2.4 Example: Car insurance claims

This data set contains records of the number of insurance claims to an insurance company in a twelve-month period in mid-1980's in each of 176 geographical areas in New South Wales, Australia. Other recorded variables are the number of accidents and population size for each area. Let μ_i denote the expected number of claims in the ith area and let μ'_i be as μ_i per unit in the population. That is $\mu'_i = \mu_i/N_i$ where N_i is the size of the population in the ith area.

We expect that the average number of claims per person would be affected by the number of accidents in each area. To that end, let x_i denote the logarithm of the number of accidents in the ith area. Then we model

$$\log \mu'_i = \beta_0 + \beta_1 x_i \Rightarrow \log(\mu_i/N_i) = \beta_0 + \beta_1 x_i \Rightarrow \log(\mu_i) = \beta_0 + \beta_1 x_i + \log N_i.$$

We therefore have the following model for the observed number of claims y_i at area i

$$y_i \overset{\mathsf{ind}}{\sim} \mathrm{Po}(\mu_i),\ \log(\mu_i) = \beta_0 + \beta_1 x_i + \log(N_i),\ i = 1, 2, \ldots, 176.$$

Notice that the variable $\log N_i$ appearing in the equation for the linear predictor has known coefficient equal to 1. Such variable is defined in Definition 2.1 and is called an *offset*.

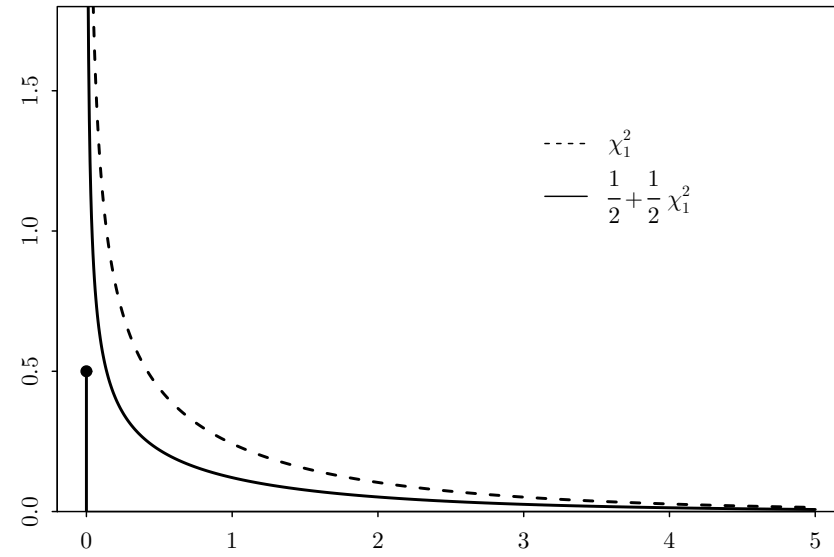


Figure 5.4: Illustration of the distribution of the likelihood ratio test statistic for testing the significance of the compound model. The solid line (—) represents the pdf of the distribution. For comparison, the $\mathcal{X}^2_1$ pdf is represented by a dashed line (– –).

We initially fit a Poisson GLM with logarithmic link to obtain estimates $\hat{\beta}_0 = -7.09$, $\hat{\beta}_1 = 0.26$, both significant. The deviance of the model is 15837, way above the 95% quantile of the $\mathcal{X}^2_{174}$ which indicates that the model does not fit well. An alternative quasi Poisson model is fitted which gave an estimate of $\hat{\phi} = 102$. A plot of the deviance residuals (Figure 5.5) shows that the model does not do a good job in explaining the variability in the data.

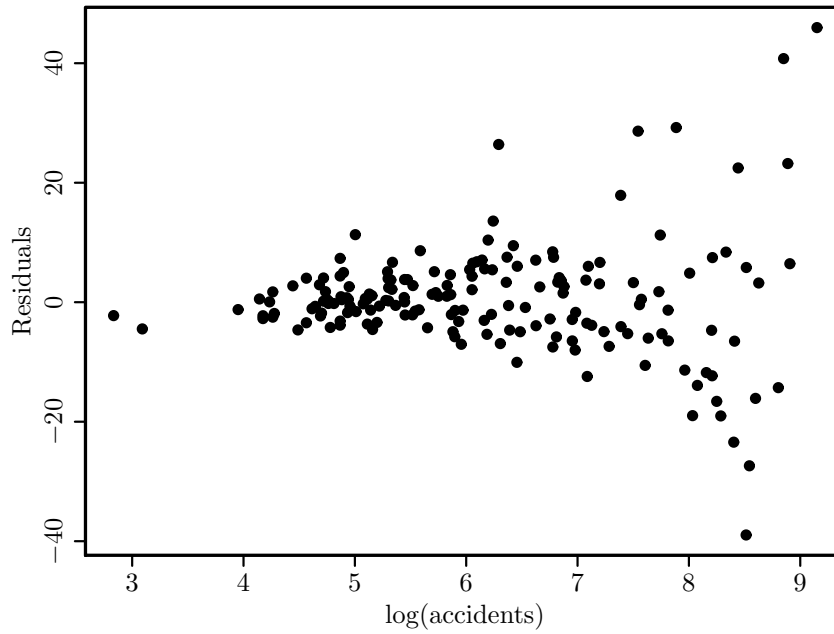


Figure 5.5: Residuals of the quasi-Poisson model for the insurance data.

Observe in Figure 5.5 that as the variability increases in a non-linear fashion as the regressor variable increases. The plot suggest that a negative binomial model would be more appropriate.

Suppose that we would like to fit the negative binomial model with known parameter ϑ. For the sake of example, let us assume that $\vartheta = 2.90$. Fitting the model we get $\hat{\beta}_0 = -6.98$ and $\hat{\beta}_1 = 0.26$. The model with unknown ϑ gives estimates $\hat{\beta}_0 = -6.95$, $\hat{\beta}_1 = 0.25$ and $\hat{\vartheta} = 5.83$. The estimates for $\hat{\boldsymbol{\beta}}$ for the two negative binomial models are very similar.

The plots of the residuals of the second fit in Figure 5.6 look much better however there is some evidence of outliers in the observations.

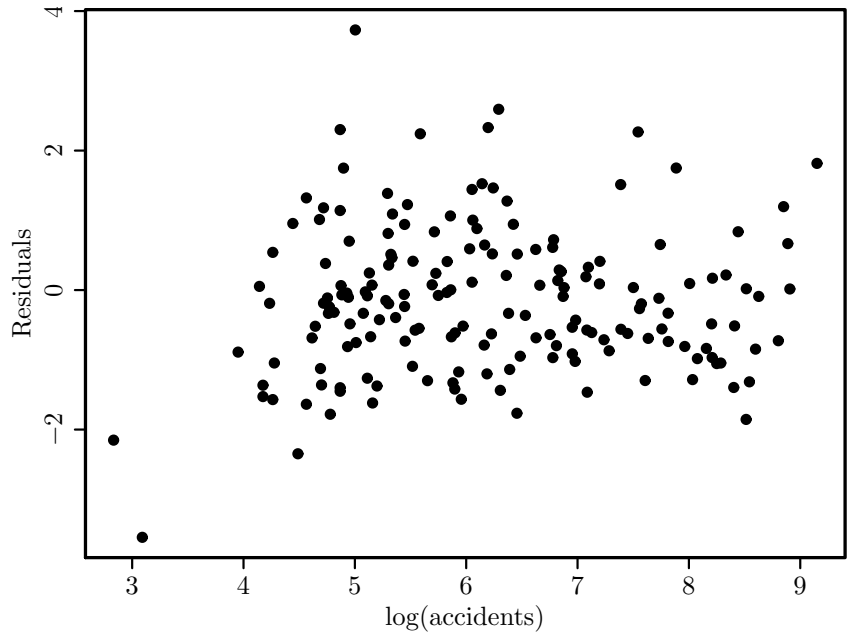


Figure 5.6: Residuals for the negative-binomial model for the insurance data

Testing for significance of the negative binomial model against the Poisson we have that the log-likelihood for the negative binomial model is -1017.627 while the log-likelihood for the Poisson model is -8531.218. The test statistic is therefore $2(-1017.627 + 8531.218) = 15027$ which is an indication that the negative binomial model is better than the Poisson model.

Figure 5.7 shows the fit between the three models. Notice that the fit between the two negative binomial models is about the same.

R code

Below is the R code used for the analysis. Notice the use of the function `offset` in the formula field of the function `glm` used to define an offset variable. Also note the use of `negative.binomial` for specifying a negative binomial fit with known parameter ϑ and the use of the function `glm.nb` to fit the model with the same parameter being unknown. The R function `logLik` was used to extract the log-likelihood of the model in order to construct the likelihood-ratio test statistic.

```
> library(MASS) # Needed for negative binomial GLM.
> insurance <- read.csv("insurance.csv",header=TRUE,row.names="area")
>
> y <- insurance$claims
> N <- insurance$population
```

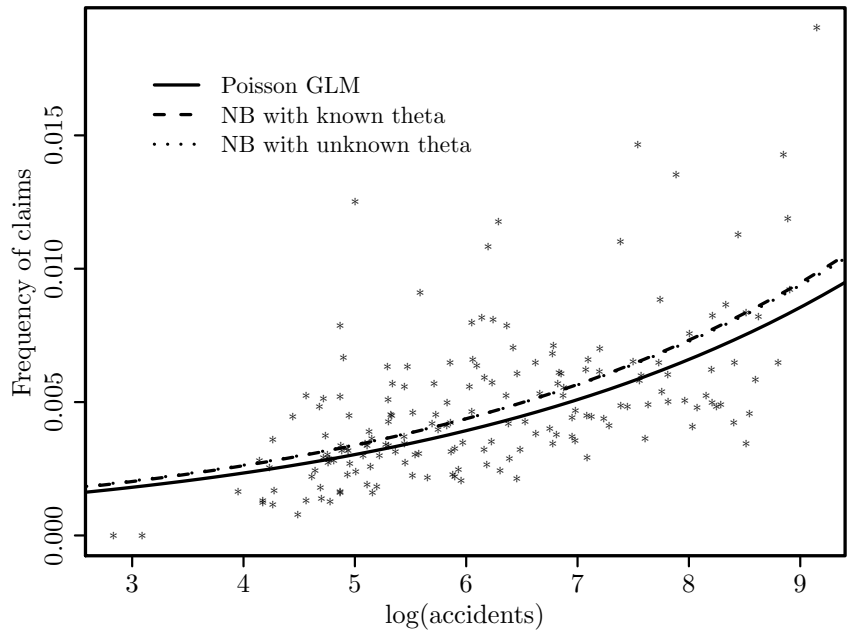


Figure 5.7: GLM fit for the insurance data

```
> x <- log(insurance$accidents)
>
> ################################ Poisson GLM ################################
> fitp <- glm(y ~ x + offset(log(N)), family=poisson)
> summary(fitp)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-38.96   -3.55    0.12    3.84   45.96

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.09381    0.02699  -262.8   <2e-16
x            0.25910    0.00338    76.8   <2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 22393  on 175  degrees of freedom
Residual deviance: 15837  on 174  degrees of freedom
AIC: 17066

Number of Fisher Scoring iterations: 4

> ### Residual plot
> plot(x, resid(fitp), xlab="log(accidents)", ylab="Residuals")
>
> ############################ Quasi-Poisson fit ##############################
> fitqp <- glm(y ~ x + offset(log(N)), family=quasipoisson)
> fitqp.smr <- summary(fitqp)
> fitqp.smr

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-38.96   -3.55    0.12    3.84   45.96
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.0938     0.2722  -26.06  < 2e-16
x             0.2591     0.0341    7.61  1.7e-12

(Dispersion parameter for quasipoisson family taken to be 101.7)

    Null deviance: 22393  on 175  degrees of freedom
Residual deviance: 15837  on 174  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4

>
> fitqp.smr$dispersion # Estimate of dispersion parameter
[1] 101.7
>
> ################## Negative binomial fit with known theta ##################
> theta.nb <- 2.90 # Fix theta for NB model
> fitnb1 <- glm(y ~ x + offset(log(N)),
+               family=negative.binomial(theta.nb))
> coef(fitnb1)
(Intercept)           x
    -6.9757      0.2574
>
> ################ Negative binomial fit with unknown theta #################
> fitnb2 <- glm.nb(y ~ x + offset(log(N)))
> coef(fitnb2)
(Intercept)           x
    -6.9544      0.2539
> fitnb2$theta # Estimate of theta
[1] 5.831
>
> plot(x, resid(fitnb2), xlab="log(accidents)", ylab="Residuals")
>
> ## Test for NB is significant:
> logLik(fitp)
'log Lik.' -8531 (df=2)
> logLik(fitnb2)
'log Lik.' -1018 (df=3)
> 2*(logLik(fitnb2) - logLik(fitp))
[1] 15027
>
> ###################### Plot of the different models ######################
> xnew <- list(x = seq(2,10,length=51), N = rep(1,51))
>
> plot(x, y/N, xlab="log(accidents)", ylab="Frequency of claims",
+      pch="*")
> lines(xnew$x, predict(fitp,xnew,"response"), lty=1)
> lines(xnew$x, predict(fitnb1,xnew,"response"), lty=2)
> lines(xnew$x, predict(fitnb2,xnew,"response"), lty=3)
> legend(3, .018, lty=1:3, bty="n",
+        legend=c("Poisson GLM","NB with known theta",
+                 "NB with unknown theta"))
```

Appendix

A. Some Probability Distributions

A.1 Useful functions

A.1.1 The gamma function

The *gamma function*, $\Gamma(x)$, is a function of a real positive argument, given by

$$\Gamma(x) := \int_0^\infty t^{x-1} e^{-t} \, \mathrm{d}t.$$

In particular $\Gamma(1) = 1$ and $\Gamma(0.5) = \sqrt{\pi}$.

The function satisfies the recursion

$$\Gamma(x+1) = x\,\Gamma(x)$$

so for integer values n we have

$$\Gamma(n) = (n-1) \times (n-2) \times \ldots \times 1 = (n-1)!$$
$$\Gamma(n+0.5) = \frac{(2\,n-1) \times (2\,n-3) \times \ldots \times 1}{2^n} \sqrt{\pi}$$

A.1.2 The beta function

The *beta function*, $B(x, y)$, is a function of two real positive arguments, given by

$$B(x, y) := \int_0^1 t^{x-1} (1-t)^{y-1} \, \mathrm{d}t.$$

It can be expressed in terms of the gamma function as

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

A.2 Discrete probability distributions

A.2.1 The binomial and Bernoulli distributions

The *binomial distribution* is denoted by $\mathrm{Bin}(m, \pi)$ where m, a positive integer, denotes the number of trials and $\pi \in (0, 1)$ the probability of success.

Its probability mass function given by

$$f(x) = \binom{m}{x} \pi^x (1-\pi)^{m-x}; \; x = 0, 1, \ldots, m.$$

The mean and variance are $m\,\pi$ and $m\,\pi\,(1-\pi)$ respectively.

In the special case where $m = 1$ the distribution is called the *Bernoulli distribution*, denoted by $\mathrm{Bernoulli}(\pi)$.

A.2.2 The Poisson distribution

The *Poisson distribution* is denoted by $\text{Po}(\mu)$ where $\mu > 0$ is called the rate.

Its probability mass function given by

$$f(x) = \frac{\mu^x}{x!} e^{-\mu};\ x = 0, 1, \ldots.$$

The mean and variance both equal μ.

A.2.3 The negative binomial and geometric distributions

The *negative binomial distribution* is denoted by $\text{NegBin}(\pi, \kappa)$ where $\pi \in (0, 1)$ is the success probability and $\kappa > 0$ is a shape parameter.

Its probability mass function given by

$$f(x) = \frac{\Gamma(x + \kappa)}{x! \Gamma(\kappa)} (1 - \pi)^\kappa \pi^x;\ x = 0, 1, \ldots,$$

where $\Gamma(\cdot)$ denotes the gamma function.

When κ is a positive integer the negative binomial distribution is interpreted as the number of successes in Bernoulli trials with probability of success π until κ failures occur.

The mean equals $\dfrac{\pi\,\kappa}{1 - \pi}$ and variance equals $\dfrac{\pi\,\kappa}{(1 - \pi)^2}$.

In the special case where $\kappa = 1$ the distribution is called the *geometric distribution*, denoted by $\text{Geometric}(\pi)$.

A.2.4 The beta-binomial distribution

The *beta-binomial distribution* is denoted by $\text{BetaBin}(m, a, b)$ where m is a positive integer and $a, b > 0$.

Its probability mass function given by

$$f(x) = \binom{m}{x} \frac{B(x + a, m - x + b)}{B(a, b)};\ x = 0, 1, \ldots, m,$$

where $B(\cdot, \cdot)$ denotes the beta function.

The mean equals $\dfrac{m\,a}{a + b}$ and the variance equals $\dfrac{m\,a\,b\,(m + a + b)}{(a + b)^2\,(a + b + 1)}$.

A.3 Continuous probability distributions

A.3.1 The normal distribution

The *normal* or *Gaussian distribution* is denoted by $\text{N}(\mu, \sigma^2)$ where μ is the mean and $\sigma^2 > 0$ is the variance.

Its probability density function is given by

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

A.3.2 The gamma, exponential, and chi-squared distributions

The *gamma distribution* is denoted by $\mathrm{G}(\mu, \vartheta)$ where $\mu > 0$ is the mean and $\vartheta > 0$ is the called the shape parameter.

Its probability density function is given by

$$f(x) = \frac{\vartheta^{\vartheta}}{\mu^{\vartheta}\Gamma(\vartheta)}\, x^{\vartheta-1}\, \exp(-\vartheta\, x/\mu);\ x > 0,$$

where $\Gamma(\cdot)$ denotes the gamma function mentioned above.

The variance equals μ^2/ϑ.

The case $\vartheta = 1$ reduces to the *exponential distribution* and the case $\vartheta = \mu/2$ to the *chi-squared distribution* with μ degrees of freedom.

A.3.3 The inverse-Gaussian distribution

The *inverse-Gaussian* distribution is denoted by $\mathrm{IG}(\mu, \lambda)$ for $\mu, \lambda > 0$. The parameter μ is the mean of the distribution and λ is called the shape parameter.

Its probability density function is given by

$$f(x) = \lambda^{\frac{1}{2}}(2\pi x^3)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\frac{\lambda\,(x-\mu)^2}{\mu^2 x}\right\};\ x > 0.$$

The variance equals μ^3/λ.

A.3.4 The beta distribution

The *beta distribution* is denoted by $\mathrm{Beta}(a, b)$ where $a, b > 0$ are two shape parameters. It is a continuous distribution in the interval (0,1).

Its probability density function is given by

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)};\ x \in (0,1),$$

where $B(\cdot,\cdot)$ denotes the beta function mentioned above.

Its mean is $\dfrac{a}{a+b}$ and its variance is $\dfrac{ab}{(a+b)^2(a+b+1)}$.

B. Vector Notation

B.1 Vector calculus

Consider the n-dimensional vectors $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_n)$. Let $A = (a_{ij})$ be an $n \times n$ matrix and $D = \mathrm{diag}(d_i)$ be an $n \times n$ diagonal matrix.

We have the following representations, where the indices i and j range from 1 to n.

1. $\sum_{i,j} u_i v_j = \mathbf{u}^\mathsf{T}\mathbf{v}$. In particular $\sum_i u_i^2 = \mathbf{u}^\mathsf{T}\mathbf{u}$.
2. $\sum_{i,j} a_{ij} u_i v_j = \mathbf{u}^\mathsf{T} A \mathbf{v}$.
3. $\sum_i d_i u_i v_i = \mathbf{u}^\mathsf{T} D \mathbf{v}$.

Now suppose $\mathbf{u}$ and $\mathbf{v}$ are vector-valued functions of a vector argument $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$. Their derivatives $\nabla\mathbf{u}$ and $\nabla\mathbf{v}$ are $n \times p$ matrices with (i,j) elements $\frac{\partial u_i}{\partial x_j}$ and $\frac{\partial v_i}{\partial x_j}$ respectively. Some of the properties of $\nabla\mathbf{u}$ and $\nabla\mathbf{v}$ are

1. $\nabla(A\mathbf{u}) = A(\nabla\mathbf{u})$
2. $\nabla(\mathbf{u} + \mathbf{v}) = (\nabla\mathbf{u}) + (\nabla\mathbf{v})$
3. $\nabla(\mathbf{u}^\mathsf{T}\mathbf{v}) = (\nabla\mathbf{u})^\mathsf{T}\mathbf{v} + (\nabla\mathbf{v})^\mathsf{T}\mathbf{u}$
4. $\nabla(\mathbf{u}^\mathsf{T} A \mathbf{v}) = (\nabla\mathbf{u})^\mathsf{T} A \mathbf{v} + (\nabla\mathbf{v})^\mathsf{T} A^\mathsf{T}\mathbf{u}$
5. $\nabla(\mathbf{u}^\mathsf{T} A \mathbf{u}) = (\nabla\mathbf{u})^\mathsf{T}(A + A^\mathsf{T})\mathbf{u}$
6. If A is symmetric, $\nabla(\mathbf{u}^\mathsf{T} A \mathbf{u}) = 2(\nabla\mathbf{u})^\mathsf{T} A \mathbf{u}$

B.2 Properties of random vectors

If $\mathbf{u}$ is a random vector, its mean $\boldsymbol{\mu}$ is an n-dimensional vector with components $\mu_i = \mathsf{E}\, u_i$ and its variance $\boldsymbol{\Sigma}$ is an $n \times n$ positive definite matrix defined by

$$\boldsymbol{\Sigma} := \mathsf{E}\{(\mathbf{u} - \boldsymbol{\mu})(\mathbf{u} - \boldsymbol{\mu})^\mathsf{T}\},$$

i.e. a matrix with (i,j) element $\sigma_{ij} = \mathsf{E}\{(u_i - \mu_i)(u_j - \mu_j)\}$. In particular the diagonal elements of $\boldsymbol{\Sigma}$ correspond to the variances of the elements of $\mathbf{u}$: $\mathsf{Var}(u_i) = \sigma_{ii} = \sigma_i^2$, and the off-diagonal elements to the covariances between the respective elements: $\mathsf{Cov}(u_i, u_j) = \sigma_{ij} = \sigma_i \sigma_j \rho_{ij}$ where ρ_{ij} is the correlation between u_i and u_j.

Theorem B.1.
We have the following properties.

1. $\boldsymbol{\Sigma} = \mathsf{E}(\mathbf{u}\mathbf{u}^\mathsf{T}) - \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T}$
2. $\mathsf{E}(A\mathbf{u}) = A\boldsymbol{\mu}$
3. $\mathsf{Var}(A\mathbf{u}) = A\boldsymbol{\Sigma}A^\mathsf{T}$
4. $\mathsf{E}(\mathbf{u}^\mathsf{T} A \mathbf{u}) = \boldsymbol{\mu}^\mathsf{T} A \boldsymbol{\mu} + \mathrm{tr}(A\boldsymbol{\Sigma})$

Proof of property 4. Note that $\mathbf{u}^\mathsf{T} A\mathbf{u}$ is a scalar therefore $\mathbf{u}^\mathsf{T} A\mathbf{u} = \mathrm{tr}(\mathbf{u}^\mathsf{T} A\mathbf{u}) = \mathrm{tr}(A\mathbf{u}\mathbf{u}^\mathsf{T})$ where the last equality makes use of the property for the trace of the product of two matrices: $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

Then

$$\begin{aligned}
\mathsf{E}(\mathbf{u}^\mathsf{T} A\mathbf{u}) &= \mathsf{E}\,\mathrm{tr}(A\mathbf{u}\mathbf{u}^\mathsf{T}) \\
&= \mathrm{tr}\,\mathsf{E}(A\mathbf{u}\mathbf{u}^\mathsf{T}) \\
&= \mathrm{tr}\{A\,\mathsf{E}(\mathbf{u}\mathbf{u}^\mathsf{T})\} \\
&= \mathrm{tr}\{A(\boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} + \boldsymbol{\Sigma})\} \\
&= \mathrm{tr}(A\boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T}) + \mathrm{tr}(A\boldsymbol{\Sigma}) \\
&= \mathrm{tr}(\boldsymbol{\mu}^\mathsf{T} A\boldsymbol{\mu}) + \mathrm{tr}(A\boldsymbol{\Sigma}) \\
&= \boldsymbol{\mu}^\mathsf{T} A\boldsymbol{\mu} + \mathrm{tr}(A\boldsymbol{\Sigma})
\end{aligned}$$

□

Making use of the above properties, let $a_1, a_2, \ldots, a_n$ be fixed numbers and define the random variable $v := \sum_i a_i u_i = a_1 u_1 + a_2 u_2 + \ldots + a_n u_n$. Then

- $\mathsf{E}\, v = \sum_i a_i \mu_i = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n$
- $\mathsf{Var}\, v = \sum_i \sum_j a_i a_j \sigma_{ij} = \sum_i a_i^2\sigma_i^2 + 2\sum_{i<j} a_i a_j \sigma_{ij}$
  $= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots + a_n^2\sigma_n^2 + 2\,a_1a_2\sigma_{1,2} + 2\,a_1a_3\sigma_{1,3} + \ldots + 2\,a_{n-1}a_n\sigma_{n-1,n}$

B.2.1 Multivariate normal distribution

Definition B.1 (Multivariate normal distribution)**.**
We say that the n-dimensional random vector $\mathbf{u}$ follows the *multivariate normal distribution* with mean an n-dimensional vector $\boldsymbol{\mu}$ and variance an $n \times n$-dimensional positive definite matrix $\boldsymbol{\Sigma}$, written as $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its probability density function (pdf) has the form

$$f(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\,\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu})\right\}.$$

Theorem B.2.
Let $\mathbf{u} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and B is an $m \times n$ constant matrix. Define the m-dimensional random vector $\mathbf{v} := B\mathbf{u}$. Then $\mathbf{v} \sim N_m(B\boldsymbol{\mu}, B\boldsymbol{\Sigma}B^\mathsf{T})$.

Theorem B.3.
Let $\mathbf{u} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $(\mathbf{u}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu}) \sim \chi^2_n$.

C. A Short Tutorial of R

C.1 Basics

C.1.1 Arithmetic

The usual arithmetic operations defined in R are

- `+` for addition,
- `-` for subtraction,
- `*` for multiplication,
- `/` for division, and
- `^` for raising to a power.

The order in which these expressions are evaluated is powers first, then multiplication and division and lastly addition and subtraction. So, for example

```
> -1 + 2^2 * 3
[1] 11
```

gives the answer 11 because powers are evaluated first `2^2`, then multiplications `4*3` and finally the addition `-1+12`. The parentheses `()` may be used to change the order of the operations, e.g.

```
> (-1 + 2)^2 * 3
[1] 3
```

C.1.2 Variables and assignments

The following characters may be used for variable names

- Lowercase and uppercase Latin letters;
- Numbers, as long as they are not the first character in the name;
- The dot "." as long as it is not the first character in the name followed by a number;
- The underscore _ as long as it is not the first character in the name.

Variables are assigned values using either `<-` or `=`. For example

```
> x <- 1.5
```

assigns the value 1.5 to the variable `x`. Equivalently one may write

```
> 1.5 -> x
```

or

```
> x = 1.5
```

Once a variable is assigned a value, it can be printed or used in arithmetic operations, e.g.

```
> x
[1] 1.5
> 2 * x + 1
[1] 4
```

To see the variables defined in the workspace use the `ls` function

```
> ls()
[1] "x"
```

Note that the output does not list any variables whose name begins with a dot "." as these are regarded as hidden variables.

A special variable is the `NULL` which contains nothing, i.e. it is empty. Other special variables are `NA` for "Not Available"; `NaN` for "Not a Number"; and `Inf`, `-Inf` for infinity.

C.2 Logical values

The two logical values are `TRUE` and `FALSE`. The operators to logical values are

- `&` the **and** operator,
- `|` the **or** operator, and
- `!` the **not** operator.

So

```
> !FALSE
[1] TRUE
> !(!FALSE)
[1] FALSE
> TRUE & FALSE
[1] FALSE
> TRUE | FALSE
[1] TRUE
> FALSE | FALSE
[1] FALSE
```

Logical values are the values of the comparison operators

- < and > for **less-than** and **greater-than**;
- <= and >= for **less-than-or-equal** and **greater-than-or-equal**;
- == for **exactly equal**;
- != for **not equal**.

So

```
> 3 < 4                  # Is 3 less than 4?
[1] TRUE
> 4 <= 4                 # Is 4 less than or equal to 4?
[1] TRUE
> x <- -1                # Assign to the variable x the value -1
> (x < -2) | (x > 2)     # Is x not in [-2,2]?
[1] FALSE
```

C.3 Brackets

There are three types of brackets in R: round `()`, square `[]`, and curly `{}`. More conventionally, these are referred to as parentheses, brackets, and braces, respectively. Their use is as follows

parentheses These are used in function calls as the function arguments. Furthermore they are used to set priorities in arithmetic operations.

brackets Brackets are used as subscripts to vectors or lists.

braces These are used for grouping together several R commands. They are mostly used in the construction of functions, in if-then-else and for constructs.

C.4 Vectors and arrays

The simplest method of creating vectors is using the function `c` (for *concatenate*). Thus

```
> y <- c(1, 2, 4)
```

assigns to the variable `y` the value $(1, 2, 4)$. The basic arithmetic operations may be applied to vectors as well. For example

```
> y - 1
[1] 0 1 3
> y^2
[1]  1  4 16
> y * x
[1] -1 -2 -4
> y + c(0, 1, 2)
[1] 1 3 6
> y < 3
[1]  TRUE  TRUE FALSE
```

Vectors can be used inside functions such as `sum`, `prod`, `exp`, etc

```
> sum(y)
[1] 7
> exp(y)
[1]  2.718282  7.389056 54.598150
> c(0, y, 5)
[1] 0 1 2 4 5
```

A way of constructing vectors in R which have a pattern is by using the functions `rep` (for *replicate*) and `seq` (for *sequence*). The use of these functions is illustrated in the examples below.

```
> rep(1, 3)
[1] 1 1 1
> rep(c(1, 2), 3)
[1] 1 2 1 2 1 2
> rep(c(1, 2), each = 3)
[1] 1 1 1 2 2 2
> seq(4)
[1] 1 2 3 4
> seq(2, 4)
[1] 2 3 4
> seq(2, 6, 2)
[1] 2 4 6
> seq(2, 7, 2)
[1] 2 4 6
> seq(2, 9, length = 3)
```

```
[1] 2.0 5.5 9.0
> c(rep(c(1, 2), 2), seq(2, 4), rep(1, 2))
[1] 1 2 1 2 2 3 4 1 1
```

A shorthand for the `seq` function is the colon ":" operator

```
> seq(4)
[1] 1 2 3 4
> 1:4
[1] 1 2 3 4
> seq(4, -2)
[1]  4  3  2  1  0 -1 -2
> 4:-2
[1]  4  3  2  1  0 -1 -2
```

You can get the size of the vector using the function `length`

```
> length(y)
[1] 3
> length(x)
[1] 1
```

however, vectors are dimensionless

```
> dim(y)
NULL
```

Arrays are extensions of vectors with a dimension attribute. They are often constructed using the function `array`. The typical call to the function is

```
array(data = NA, dim = length(data), dimnames = NULL)
```

where

data is a vector giving data to fill the array. By default these are taken as missing values `NA`. The data vector is replicated until all the elements of the array are filled.

dim the dimension of the array to be created, that is a vector of length one or more giving the maximal indices in each dimension. By default this creates a vector with dimension the length of the inputted data.

dimnames either `NULL` or the names for the dimensions. By default no names are given.

The following examples illustrate the construction of arrays.

```
> a <- array(1:6)
> a
[1] 1 2 3 4 5 6
> b <- array(1:6, c(2, 3))
> b
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> dim(b)
[1] 2 3
> length(b)
[1] 6
> c <- 1:8
> dim(c) <- c(2, 4)
```

```
> c
     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8
> d <- c(c)
> d
[1] 1 2 3 4 5 6 7 8
> two <- array(2, c(2, 4))
> two
     [,1] [,2] [,3] [,4]
[1,]    2    2    2    2
[2,]    2    2    2    2
> i3 <- array(c(1, rep(0, 3)), c(3, 3))
> i3
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

C.4.1 Subscripts

Elements of vectors and arrays are extracted and assigned using the brackets. For example

```
> y                           # Vector y
[1] 1 2 4
> y[1]                        # Extract first element
[1] 1
> y[1:2]                      # Extract first and second elements
[1] 1 2
> y[-3]                       # Negative subscripts remove elements
[1] 1 2
> y[c(1,3)]                   # Extract first and third elements
[1] 1 4
> y[3] <- 5                   # Overwrite the third element
> y[2] <- y[1] + y[2]         # Overwrite the second element
> y
[1] 1 3 5
> b                           # Array b
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> b[2,1]                      # Extract the (2,1) element
[1] 2
> b[3]                        # Extract the third element
[1] 3
> b[,2]                       # Extract the second column
[1] 3 4
> b[2,]                       # Extract the second row
[1] 2 4 6
```

Subscripts may also contain logical vectors. For example

```
> z <- c(2, 4, 7, 9)
> z < 6
[1]  TRUE  TRUE FALSE FALSE
> z[z < 6]
[1] 2 4
> sum(z[z < 6])
[1] 6
> z[z < 6] <- 6
> z
```

```
[1] 6 6 7 9
```

C.5 Lists

Lists are collections of R objects under one object. Anything can be included in a list using the function `list` as follows

```
> mylist <- list(one = 1, 2, c, d = d, TRUE)
> mylist
$one
[1] 1

[[2]]
[1] 2

[[3]]
     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8

$d
[1] 1 2 3 4 5 6 7 8

[[5]]
[1] TRUE
```

The function `read.table` which is used for loading data into R creates a list with named elements.

Notice that some of the elements in the list are named and some not. Elements of lists can be extracted and assigned using the double bracket indexing `[[]]`. The named elements may also be extracted and assigned using the `$` symbol.

```
> mylist[[1]]
[1] 1
> mylist$one
[1] 1
> mylist[[2]]
[1] 2
> mylist$one + mylist[[2]]
[1] 3
> length(mylist)
[1] 5
> mylist[[3]] <- 3^3
> mylist$happy <- ":)"
> mylist
$one
[1] 1

[[2]]
[1] 2

[[3]]
[1] 27

$d
[1] 1 2 3 4 5 6 7 8

[[5]]
```

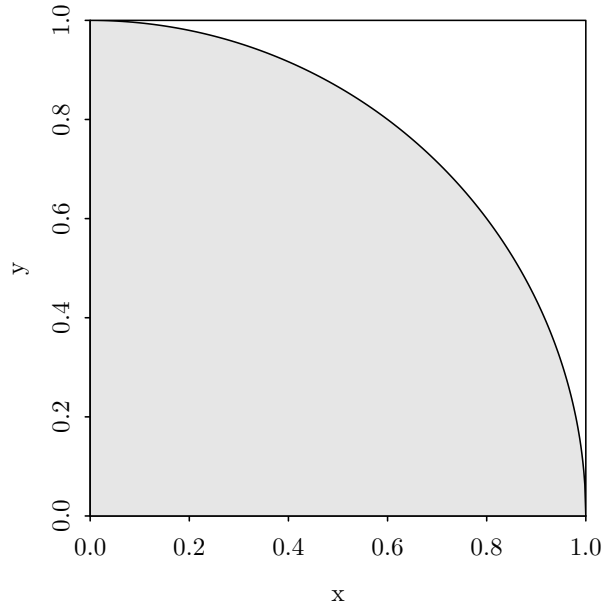
```
[1] TRUE

$happy
[1] ":)"
```

C.6 Examples

C.6.1 Approximating the number π

The area of a quarter of a circle of radius 1 is $\frac{\pi}{4}$. If a large number of points are generated uniformly in the unit square the proportion of these that will fall inside the first quadrant of a unit circle centred at 0 will then be approximately (Area of quadrant)/(Area of square) $= \frac{\pi}{4}$.



Below is the R code for approximating the value of π. In the code that follows notice the use of the function `runif` for generating random samples from the uniform distribution in $(0, 1)$ and the use of the function `mean` with logical array argument for computing the proportion of true cases in the array.

```
> N <- 1000000  # Random sample size
> x <- runif(N) # Random 1st coordinate from uniform dist'n
> y <- runif(N) # Random 2nd coordinate from uniform dist'n
> piq <- mean(x^2 + y^2 < 1) # The proportion of points that fall in the
>                            # unit circle
> 4*piq         # Approximate value of pi
[1] 3.1418
> pi            # Actual value of pi
[1] 3.1416
```

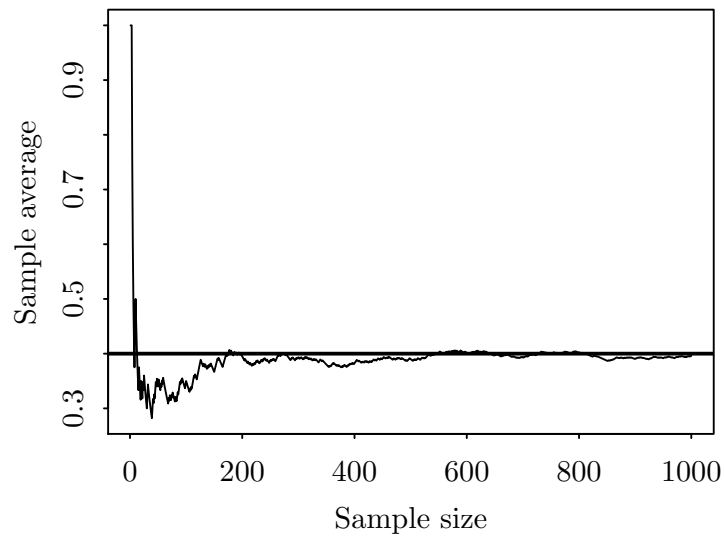
C.6.2 The law of large numbers

The law of large numbers roughly says that if $X_1, X_2, \ldots, X_n$ is an i.i.d. sample from a distribution with mean μ such that $\mathsf{E}|X_i| < \infty$, then as $n \to \infty$, $\bar{X} \to \mu$.

The law of large numbers for the case of a Bernoulli random sample with probability of success 0.4 is illustrated with the following R code. Notice the use of the construct `for` for

performing calculations iteratively. The function `abline` is used for drawing straight lines to a plot.

```
> N <- 1000
> p <- 0.4
> xx <- rbinom(N,1,p)
> xbar <- array(0,N)
> for (i in 1:N) {
+   xbar[i] <- mean(xx[1:i])
+ }
> plot(1:N,xbar,t='l',xlab='Sample size',ylab='Sample average')
> abline(h=p,lwd=2)
```



C.6.3 The Monty Hall problem

The *Monty Hall problem* is a probability puzzle loosely based on the American television game show *Let's Make a Deal* and named after the show's original host, Monty Hall. The problem goes as follows.

> Suppose you are on a game show, and you are given the choice of three doors: Behind one door is a prize; behind the others, nothing. You pick a door, say No. 1 (but the door is not opened), and the host, who knows which door contains the prize, opens another door, say No. 3, which has nothing. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

The answer is "yes". This is because in the long run the player initially selects the door with the prize 1/3 of the time. In the remaining 2/3 of the time the player will initially choose one of the doors with no prize. In this case the host will be forced to open the other door with no prize leaving the one with the prize closed. Then if the player switches to the remaining door there is a 2/3 chance that he or she will win.

The R code below illustrates the Monty Hall problem for a long number of plays. The player's strategy is to always switch from the initial choice. Roughly 2/3 of the time the player wins with this strategy. Note the use of the function `sample(x,n)` for randomly selecting n elements from the array x. The function `dimnames` is used for assigning names to the rows and columns

of an array. Also the construct `if ... else ...` is used for performing calculation subject to conditions.

```
> N <- 15 # Number of games
> tbl <- array(NA,c(N,5)) # Table storing the games
> dimnames(tbl) <- list(NULL,c("Prize","Player","Host","Switch","Win"))
> for (i in 1:N) {
+   prize <- sample(1:3,1) # Door with the prize
+   player1 <- sample(1:3,1) # Initial choice of player
+   if (prize == player1) { # This will happen approx 1/3 of the time
+     host <- sample((1:3)[-prize],1)
+   } else { # The remaining 2/3 of the time
+     host <- (1:3)[-c(prize,player1)]
+   }
+   player2 <- (1:3)[-c(host,player1)] # Player chooses other door
+   win <- player2 == prize # Did the player win?
+   tbl[i,] <- c(prize,player1,host,player2,win) # Store in table
+ }
> tbl
      Prize Player Host Switch Win
 [1,]     3      1    2      3   1
 [2,]     2      2    3      1   0
 [3,]     3      3    1      2   0
 [4,]     3      1    2      3   1
 [5,]     1      3    2      1   1
 [6,]     3      2    1      3   1
 [7,]     2      2    3      1   0
 [8,]     1      3    2      1   1
 [9,]     1      3    2      1   1
[10,]     2      2    3      1   0
[11,]     3      1    2      3   1
[12,]     2      2    1      3   0
[13,]     2      1    3      2   1
[14,]     3      1    2      3   1
[15,]     1      1    3      2   0
```