Theory of Statistical Inference

Statistical inference is about figuring out what *data* are telling us about the world. We gather data related to some question of interest and construct a *model* of how the data could have been generated. The model is incomplete, being formulated in terms of *parameters* whose values are unknown (or perhaps having some other unknown component). Knowing the values of these unknowns would tell us something useful about the world. The theory of statistical inference addresses how we can use data to make inferences about the values of the unknowns. Since models and data are central to this enterprise, we should start with some examples.

1 Some example models of data

A statistical model is a simplified mathematical description of how a set of data might have been generated: a sort of mathematical cartoon. A key feature is that it describes the random variability in the data generating process, recognising that gathering a replicate set of data under similar circumstances would not give an identical dataset, but rather a stochastically similar dataset.

1. Consider the following 60-year record of mean annual temperatures in New Haven, Connecticut (in °F, and available as `nhtemp` in R).

   ```
   49.9 52.3 49.4 51.1 49.4 47.9 49.8 50.9 49.3 51.9 50.8 49.6 49.3 50.6 48.4 50.7 50.9 50.6 51.5 52.8
   51.8 51.1 49.8 50.2 50.4 51.6 51.8 50.9 48.8 51.7 51.0 50.6 51.7 51.5 52.1 51.3 51.0 54.0 51.4 52.7
   53.1 54.6 52.0 52.0 50.9 52.6 50.2 52.6 51.6 51.9 50.5 50.9 51.7 51.4 51.7 50.8 51.9 51.8 51.9 53.0
   ```

   A simple model would treat these data as independent observations from an $N(\mu, \sigma^2)$ distribution, where μ and σ^2 are unknown parameters. Then the p.d.f. for the random variable corresponding to a single measurement, y_i, is

   $$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y_i-\mu)^2}{2\sigma^2}}.$$

   The joint p.d.f. for the vector $\mathbf{y}$ is the product of the p.d.f.s for the individual random variables, because the model specifies independence of the y_i, i.e.

   $$f(\mathbf{y}) = \prod_{i=1}^{60} f(y_i).$$

2. The New Haven temperature data seem to be 'heavy tailed' relative to a normal: that is, there are more extreme values than are implied by a normal with the observed standard deviation. A better model might be

   $$\frac{y_i - \mu}{\sigma} \sim t_\alpha,$$

   where μ, σ and α are unknown parameters. Denoting the p.d.f. of a t_α distribution as f_{t_α}, standard transformation theory, combined with independence of the y_i, implies that the p.d.f. of $\mathbf{y}$ is

   $$f(\mathbf{y}) = \prod_{i=1}^{60} \frac{1}{\sigma} f_{t_\alpha}\{(y_i - \mu)/\sigma\}.$$

3. Air temperature, a_i, is measured at times t_i (in hours) spaced half an hour apart for a week. The temperature is believed to follow a daily cycle, with a long-term drift over the course of the week, and to be subject to random autocorrelated departures from this overall pattern. A suitable model might then be

   $$a_i = \theta_0 + \theta_1 t_i + \theta_2 \sin(2\pi t_i/24) + \theta_3 \cos(2\pi t_i/24) + e_i,$$

   where $e_i = \rho e_{i-1} + \epsilon_i$ and the ϵ_i are i.i.d. $N(0, \sigma^2)$. This model implicitly defines the p.d.f. of $\mathbf{a}$, but as specified we have to do a little work to actually find it. Writing $\mu_i = \theta_0 + \theta_1 t_i + \theta_2 \sin(2\pi t_i/24) + \theta_3 \cos(2\pi t_i/24)$, we have $a_i = \mu_i + e_i$. Because e_i is a weighted sum of zero mean normal random variables, it is itself a zero mean normal random variable, with covariance matrix $\mathbf{\Sigma}$ such that $\Sigma_{i,j} = \rho^{|i-j|}\sigma^2/(1-\rho^2)$. So the p.d.f. of $\mathbf{a}$,[1] the vector of temperatures, must be multivariate normal,

$$f_{\mathbf{a}}(\mathbf{a}) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{a}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{a}-\boldsymbol{\mu})},$$

where $\mathbf{\Sigma}$ depends on parameter ρ and σ, while $\boldsymbol{\mu}$ depends on parameters $\boldsymbol{\theta}$ and covariate $\mathbf{t}$.

4. Data were collected at the Ohio State University Bone Marrow Transplant Unit to compare two methods of bone marrow transplant for 23 patients suffering from non-Hodgkin's lymphoma. Each patient was randomly allocated to one of two treatments. The *allogenic* treatment consisted of a transplant from a matched sibling donor. The *autogenic* treatment consisted of removing the patient's marrow, 'cleaning it' and returning it after a high dose of chemotherapy. For each patient the time of death, relapse or last follow up (still healthy) is recorded. The 'right-censored' last follow up times are marked with an over-bar.

	Time (Days)											
Allo	28	32	49	84	357	$\overline{933}$	$\overline{1078}$	$\overline{1183}$	$\overline{1560}$	$\overline{2114}$	$\overline{2144}$	
Auto	42	53	57	63	81	140	176	$\overline{210}$	252	$\overline{476}$	524	$\overline{1037}$

The data are from Klein and Moeschberger (2003). A reasonable model is that the death or relapse times are observations of independent random variables having exponential distributions with parameters θ_l and θ_u respectively (mean survival times are $\theta_{u/l}^{-1}$). Medically the interesting question is whether the data are consistent with $\theta_l = \theta_u$.

For the allogenic group, denote the time of death, relapse or censoring by t_i. So we have

$$f_l(t_i) = \begin{cases} \theta_l e^{-\theta_l t_i} & \text{uncensored} \\ \int_{t_i}^{\infty} \theta_l e^{-\theta_l t} dt = e^{-\theta_l t_i} & \text{censored} \end{cases}$$

where f_l is a density for an uncensored t_i (death) or a probability of dying after t_i for a censored observation. A similar model applies for the autogenic sample. For the whole dataset we then have

$$f(\mathbf{t}) = \prod_{i=1}^{11} f_l(t_i) \prod_{i=12}^{23} f_u(t_i).$$

The key point with all these models is that if they are correct, then given appropriate values for their parameters, they should be able to generate data that is similar to the original data (while almost never being identical to it).

One feature of these example is that the only unknowns they contain are parameters: hence they are known as *parametric* models, and the theory for making inferences about them is known as *parametric statistical inference*. There other possibilities. For example models sometimes contain unknown functions as well as unknown parameter (e.g. a model assumes that the relationship between death rate from respiratory disease and time is a smooth function). Such models are often known as *semi-parametric*, in part because we can often turn them back into something parametric with a lot of parameters, by finding appropriate ways of representing the functions. Finally there are truly *nonparametric models* which try to make minimal assumptions about how the data were generated, beyond those absolutely necessary to answer the question of interest. For example we might have a model: 'these 2 samples are from the same population' which we want to compare with the model 'these two samples come from populations with different means'. Clearly these are not models from which we could simulate data if we only knew the values of some unknown parameters, and hence have a rather different character to the (semi) parametric models. In this course we will focus on parametric inference.

[1] For aesthetic reasons I will use phrases such as 'the p.d.f. of $\mathbf{y}$' to mean 'the p.d.f. of the random vector of which $\mathbf{y}$ is an observation'.

2 Inferential questions and approaches

Statistical inference is about inferring the model unknowns (usually parameters) from the data that the model is supposed to have generated. This should be done in a way that accounts for the random variability in the data, so that our conclusions will hold generally, and will not overly reflect the random peculiarities of the particular data set being analysed. Basically we are interested in answering three questions:

1. What range of parameter values are consistent with the data?
2. Are some pre-specified values (or restrictions) for the parameters consistent with the data?
3. Could our model have generated the data at all?

Question 2 is often generalized to:

> Which of several alternative models could most plausibly have generated the data?

This reflects the fact that you can view models with and without some restriction on the parameters as alternative models to be compared.

There is another closely related question.

> How could we better arrange the data gathering process to improve the answers to the preceding questions?

This is really a question of *statistical design*, but it is an important one to appreciate, as it will have a bearing on what we can conclude from statistical inference.

2.1 Basic approaches to inference: Bayesian and frequentist

There are two main approaches to statistical inference: *Bayesian* and *frequentist*. Both are useful, but imperfect.

In the *frequentist* approach we treat model unknowns (usually parameters) as fixed states of nature, which we want to estimate. Random variability is present in the data generating process, and inherited by our estimates, but the parameters themselves are unknown fixed quantities and are not random. Probability is used to describe the variability that would occur under replication of the data gathering and estimation process.

In the *Bayesian* approach we use probability to describe our knowledge of the model parameters. Specifically, we augment our model for generating data given parameter values, with a model describing our initial uncertainty about the parameter values. This initial uncertainty model treats the model parameters as random variables and takes the form of a probability distribution for these parameters, known as the *prior distribution*. Inference consists of updating the prior distribution to a *posterior distribution*, which is simply the probability distribution of the parameters *given the data*.

In this course we will be interested in developing general inferential methods applicable to classes of models that are as wide as possible. Usually this generality will come at the cost of approximation. In the frequentist approach we will often use 'asymptotic' results which become exact only as the sample size tends to infinity. In the Bayesian case we will typically have to use simulation methods, which are exact only as the simulation sample size tends to infinity. But to fix ideas consider application of the frequentist and Bayesian approaches to a simple (and unusually tractable) example model with one parameter.

2.1.1 A simple example

Suppose that we observe data $x_1, x_2, \ldots, x_n$ which can be modelled as observations of independent $N(\mu, 1)$ random variables, where μ is an unknown parameter. We want to make inferences about μ, focussing on the question of the range of μ values consistent with the data.

Frequentist approach. We need an estimate of μ. Later we will look at general principles for finding estimates, but for now let's use the obvious estimate $\hat{\mu} = \bar{x}$ ($\bar{x} = \sum_i x_i/n$). The x_i are observations of random variables. $\hat{\mu}$

considered as a function of those random variables is the *estimator* of μ, and is obviously itself a random variable. Hence we can consider its expectation, variance and distribution. We have

$$E(\hat{\mu}) = E(\bar{x}) = \frac{1}{n}\sum_i E(x_i) = \mu \text{ and } \text{var}(\hat{\mu}) = \frac{1}{n^2}\sum_i \text{var}(x_i) = \frac{1}{n}.$$

Furthermore, since $\hat{\mu}$ is a linear transformation of normal random variables it must also be a normal random variable, so

$$\hat{\mu} \sim N(\mu, n^{-1}).$$

So we have an estimate of μ, and know that the standard deviation associated with that estimate is $n^{-1/2}$. We might also want to make probability statements about the range of values of μ that are plausible. To do so requires a little work, as the distribution of $\hat{\mu}$ depends on the unknown parameter μ. The usual way to address such a problem is to find a *pivotal quantity*, with a distribution with no unknown parameters, and work from this. For example $(\hat{\mu} - \mu)n^{1/2} \sim N(0, 1)$, so

$$\text{Pr}(-1.96 < (\hat{\mu} - \mu)n^{1/2} < 1.96) = 0.95 \Rightarrow \text{Pr}(\hat{\mu} - 1.96n^{-1/2} < \mu < \hat{\mu} + 1.96n^{-1/2}) = 0.95,$$

i.e. the *confidence interval* $\bar{x} \pm 1.96n^{-1/2}$ has a 95% chance of including the true value of μ (notice that it is the interval that is random here, not μ).

Bayesian approach. Now we are going to treat μ as a random variable, whose distribution describes our beliefs about the value of μ. Therefore we need to augment our model with a *prior* distribution for μ. Let's use $\mu \sim N(0, \sigma_\mu^2)$ (where σ_μ^2 is some fixed constant of our choosing). Bayes rule follows from the observation that the joint density of $\mathbf{x}$ and μ is obviously equal to the joint density of μ and $\mathbf{x}$, so

$$f(\mu|\mathbf{x})f(\mathbf{x}) = f(\mathbf{x}|\mu)f(\mu) \Rightarrow f(\mu|\mathbf{x}) = f(\mathbf{x}|\mu)f(\mu)/f(\mathbf{x})$$

(f with different arguments denote different distributions here). We are interested in finding $f(\mu|\mathbf{x})$ (with the values of $\mathbf{x}$ plugged in), which describes our beliefs about μ after observing the data, and is hence known as the *posterior distribution* of μ. When $f(\mu|\mathbf{x})$ is tractable at all, there is a trick to finding it, which avoids the difficult process of evaluating $f(\mathbf{x}) = \int f(\mathbf{x}|\mu)f(\mu)d\mu$.

The trick is to start from $f(\mu|\mathbf{x}) \propto f(\mathbf{x}|\mu)f(\mu)$, and to realise that any factor of $f(\mu|\mathbf{x})$ which only involves $\mathbf{x}$, but not μ, must be part of the normalizing constant of $f(\mu|\mathbf{x})$. Hence it suffices to find the factor of $f(\mu|\mathbf{x})$ containing the dependence on μ, and to identify the distribution to which it corresponds. So in the current case

$$\begin{aligned} f(\mu|\mathbf{x}) &\propto \exp\left\{-\left(\textstyle\sum_i (x_i - \mu)^2 + \mu^2/\sigma_\mu^2\right)/2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(-2n\bar{x}\mu + \frac{n\sigma_\mu^2 + 1}{\sigma_\mu^2}\mu^2\right)\right\} = \exp\left\{-\frac{1}{2}\left(\mu^2 - 2\bar{x}\mu\frac{n\sigma_\mu^2}{n\sigma_\mu^2 + 1}\right)\frac{n\sigma_\mu^2 + 1}{\sigma_\mu^2}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\mu - \bar{x}\frac{n\sigma_\mu^2}{n\sigma_\mu^2 + 1}\right)^2\frac{n\sigma_\mu^2 + 1}{\sigma_\mu^2}\right\}, \end{aligned}$$

from which we can recognise that

$$\mu|\mathbf{x} \sim N\left(\bar{x}\frac{n\sigma_\mu^2}{n\sigma_\mu^2 + 1}, \frac{\sigma_\mu^2}{n\sigma_\mu^2 + 1}\right). \tag{1}$$

The fact that the posterior distribution is the same type of distribution as the prior (i.e. Gaussian) is a rather special property known as *conjugacy*. If we had chosen a different distribution for the prior then this would not have happened. For example a gamma prior for the mean of a Gaussian distribution will not yield a gamma posterior.

There are no unknowns in (1) so producing a 95% *credible interval* for μ is easy:

$$\bar{x}\frac{n\sigma_\mu^2}{n\sigma_\mu^2 + 1} \pm 1.96\frac{\sigma_\mu}{(n\sigma_\mu^2 + 1)^{1/2}}.$$

This fixed interval has a 95% chance of including the random variable μ. Compare this to the interpretation of the frequentist confidence interval, produced earlier.

Despite the difference in interpretation it is worth comparing the Bayesian credible interval and frequentist confidence interval when either n or σ^2_μ tend to infinity. $n \to \infty$ corresponds to the sample size tending to infinity (more data), while $\sigma^2_\mu \to \infty$ corresponds to the prior becoming vaguer (less prior information). In either case the Bayesian credible interval tends to the frequentist confidence interval, something which applies more widely than just this simple example. It's also interesting to compare the distributional results in this large sample/vague prior limit: $\hat\mu \sim N(\mu, n^{-1})$ versus $\mu|\mathbf{x} \sim N(\hat\mu, n^{-1})$.

2.2 Inference by resampling

The examples of Bayesian and frequentist inference given above used mathematics to make inferences about parameter values and their uncertainty. An alternative approach substitutes computation for mathematical manipulation. The idea is that uncertainty about parameter values is inherited from the uncertainty in the data sampling process. When we calculate the uncertainties associated with parameters we are essentially computing the uncertainty that would result from repeating the data gathering and analysis process a large number of times. Continuing the simple example from above, we could imagine gathering replicate data sets again and again and again, computing $\bar x$ for each, and thereby building up an empirical distribution for $\hat\mu = \bar x$.

In most cases such replication of the data is not possible, and even if we could do it, we would then prefer to combine all the data to get a larger sample size. However we can *simulate* repeated replication of the data gathering process, by simply repeatedly *resampling* from the one data set that we have. In particular, we can randomly resample the original data with replacement, to create simulated replicate datasets of the same size as the original. Unknown parameters can then be estimated from each of these datasets, in the same way as was done for the original dataset, in order to assess the uncertainty of the parameter estimators. This process is known as *bootstrapping*. Here it is applied to the simple example from the previous section…

```
set.seed(1);n <- 50
x <- rnorm(n) + 4 ## simulate data

n.rep <- 1000
mubs <- rep(0,n.rep);mubs[1] <- mean(x)
for (i in 2:n.rep) {
  xb <- sample(x,n,replace=TRUE) ## bootstrap resample data
  mubs[i] <- mean(xb)            ## compute mu.hat for this resample
}
quantile(mubs,c(.025,.975)) ## quantiles of bootstrap distribution give CI
    2.5%    97.5%
3.867025 4.315665
```

For comparison here is the original frequentist interval using the same data.

```
mu.hat <- mean(x)
mu.sd <- 1/sqrt(n)
c(mu.hat - 1.96*mu.sd,mu.hat + 1.96*mu.sd)
[1] 3.823262 4.377634
```

Finally the Bayesian interval, assumming $\sigma_\mu = 3$.

```
sig.mu <- 3 ## prior sd
mub <- mean(x)*(n*sig.mu^2)/(n*sig.mu^2+1)
sdb <- sqrt(sig.mu^2/(n*sig.mu^2+1))
c(mub - 1.96*sdb,mub + 1.96*sdb)
[1] 3.814478 4.368235
```

Notice how the Bayesian interval is shifted slightly towards the prior mean of zero, relative to the frequentist interval. The bootstrap interval is slightly narrower than the others. In this case it is slightly *too* narrow and its real coverage probability is slightly under the nominal 95%. This would improve with increasing sample size. Increasing the sample size causes all the intervals to converge on each other.

Careful consideration of the bootstrapping process might lead you to worry about the way these bootstrap percentile intervals have been computed - we've done it as if the bootstrap $\hat{\mu}$ values were samples from the posterior $\mu|\mathbf{x}$, rather than being replicates of $\hat{\boldsymbol{\mu}}$. In fact various approaches can be justified here, but for the moment, let's move on. The key point to take from this section is the notion of quantifying uncertainty by resampling from your data to simulate replication of the data gathering process.

3 Inference for linear models

We are interested in developing fairly general theory for statistical inference, and since statistical models are central to statistical inference it is therefore helpful to consider some quite general classes of statistical model. The class of *linear models* is a good starting point, due to its wide applicability in applied statistics and machine learning, the elegant theory that it possesses and the wider applicability of many of the ideas that are required to use these models.

3.1 Linear models

A linear model is one in which a *response* vector $\mathbf{y}$ is linear in some parameters $\boldsymbol{\beta}$ and some zero mean random errors $\boldsymbol{\epsilon}$, so that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

The *model matrix* $\mathbf{X}$ is determined by some known *predictor* variables (also known as *covariates*[2]), observed along with each response observation y_i. Usually the elements of $\boldsymbol{\epsilon}$ are assumed to be mutually independent with constant variance σ^2. For the purposes of finding confidence intervals and testing hypotheses for $\boldsymbol{\beta}$, the ϵ_i are also assumed to have a normal distribution.

Two types of predictor variable form the basic ingredients of $\mathbf{X}$.

1. *Metric* predictor variables are measurements of some quantity that may help to predict the value of the response. For example, if the response is the blood pressure of patients in a clinical trial, then age, fat mass and height are potential metric predictor variables.

2. *Factor* variables are labels that serve to categorize the response measurements into groups, which may have different expected values. Continuing the blood pressure example, factor variables might be sex and drug treatment received (drug A, drug B or placebo, for example). Somewhat confusingly, the groups of a factor variable are referred to as *levels* although the groups generally have no natural ordering, and even if they do, the model structure ignores it.

To understand the construction of $\mathbf{X}$ it helps to consider an example. Suppose that along with y_i we have metric predictor variables x_i and z_i and factor variable g_i, which contains labels dividing y_i into three groups. Suppose further that we believe the following model to be appropriate:

$$y_i = \gamma_{g_i} + \alpha_1 x_i + \alpha_2 z_i + \alpha_3 z_i^2 + \alpha_4 z_i x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where there is a different γ parameter for each of the three levels of g_i. Collecting the γ and α parameters into one

[2] Response and predictor variables are sometimes known as 'dependent' and 'independent' variables, a particularly confusing terminology, not used here.

vector, $\boldsymbol{\beta}$, we can rewrite the model in matrix-vector form as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ . \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & x_1 & z_1 & z_1^2 & z_1x_1 \\ 1 & 0 & 0 & x_2 & z_2 & z_2^2 & z_2x_2 \\ 1 & 0 & 0 & x_3 & z_3 & z_3^2 & z_3x_3 \\ 0 & 1 & 0 & x_4 & z_4 & z_4^2 & z_4x_4 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 0 & 0 & 1 & x_n & z_n & z_n^2 & z_nx_n \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ . \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

where y_1 - y_3 are in group 1, y_4 is in group 2, and y_n is in group 3 of the factor g. Notice how the factor levels/-groups each get a dummy indicator column in the model matrix, with elements showing whether the corresponding y_i belongs to the group or not. Notice also how the metric variables can enter the model nonlinearly: the model is linear in the parameters and error term, but not necessarily in the predictors.

3.1.1 Some simple examples

1. The `cars` dataframe in R contains data relating stopping `distance` of cars to their `speed` at the point at which the signal to stop is given. Physically we expect stopping distance to be made up of two components — a component relating to reaction time, which is therefore proportional to `speed` and a component proportional to the kinetic energy of the car, i.e. to $\texttt{speed}^2$ (work done by the brakes is braking force times distance which has to equal initial kinetic energy). This suggests a linear model

   $$\texttt{distance}_i = \beta_1\texttt{speed}_i + \beta_2\texttt{speed}_i^2 + \epsilon_i$$

   where the ϵ_i are independent zero mean random variables with constant variance. i.e. $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$ for all i. We might go a little further and also assume $\epsilon_i \sim N(0, \sigma^2)$.

2. As statisticians we need to be open to the possibility that our model is wrong. In the `cars` example, we might mitigate against this by allowing the model a bit more flexibility than the physics suggests. For example, we might use

   $$\texttt{distance}_i = \beta_0 + \beta_1\texttt{speed}_i + \beta_2\texttt{speed}_i^2 + \beta_3\texttt{speed}_i^3 + \epsilon_i,$$

   and then test statistically whether $\beta_0 = 0$ and $\beta_3 = 0$, as expected.

3. The `PlantGrowth` dataframe in R contains data on the `weight` of plant seedlings grown under three alternative experimental conditions.

   ```
   > PlantGrowth
      weight group
   1     4.17  ctrl
   2     5.58  ctrl
   .      .     .
   11    4.81  trt1
   12    4.17  trt1
   .      .      .
   21    6.31  trt2
   22    5.12  trt2
   .      .      .
   ```

   A model for these data is that each treatment `group` has a separate mean. i.e.

   $$\texttt{weight}_i = \beta_{g(i)} + \epsilon_i$$

   where $g(i)$ indicates which of the 3 groups the i^{th} observation belongs to. The assumptions on ϵ_i are as above.

3.2 Linear model estimation and checking

We will start with a frequentist treatment of the linear model. That is we will assume that the parameters $\boldsymbol{\beta}$ and σ^2 are fixed states of nature, but unknown to us, and that all the uncertainty is in our estimates, inherited from the random variability in the data. We will derive results that apply to any linear model, and can be programmed into a computer to perform the actual calculations for any specific model.

Point estimates of the linear model parameters, $\boldsymbol{\beta}$, can be obtained by the method of least squares; that is, by minimising the residual sum of squares

$$\mathcal{S} = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \mu_i)^2,$$

with respect to $\boldsymbol{\beta}$, where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. Fitting a model so that we minimise the squared error terms, or equivalently try to make the μ_i as close as possible to the y_i, has intuitive appeal. Later we will see that it also has some theoretical justification.

To use least squares with a linear model, written in general matrix-vector form, first recall the link between the Euclidean length of a vector and the sum of squares of its elements. If $\mathbf{v}$ is any vector of dimension, n, then $\|\mathbf{v}\|^2 \equiv \mathbf{v}^{\mathrm{T}}\mathbf{v} \equiv \sum_{i=1}^{n} v_i^2$. Hence

$$\mathcal{S} = \|\mathbf{y} - \boldsymbol{\mu}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Since $\mathcal{S}$ is simply the squared (Euclidian) length of the vector $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, its value will be unchanged if $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ is rotated or reflected. This observation is the basis for a practical method for finding $\hat{\boldsymbol{\beta}}$ and for developing the distributional results required to use linear models.

Specifically, as with any real matrix, $\mathbf{X}$ can always be decomposed

$$\mathbf{X} = \boldsymbol{\mathcal{Q}} \left[\begin{array}{c} \mathbf{R} \\ \mathbf{0} \end{array} \right] = \mathbf{Q}\mathbf{R}, \tag{2}$$

where $\mathbf{R}$ is a $p \times p$ upper triangular matrix,[3] and $\boldsymbol{\mathcal{Q}}$ is an $n \times n$ orthogonal matrix, the first p columns of which form $\mathbf{Q}$. Recall that orthogonal matrices rotate/reflect vectors, but do not change their length. Orthogonality also means that $\boldsymbol{\mathcal{Q}}\boldsymbol{\mathcal{Q}}^{\mathrm{T}} = \boldsymbol{\mathcal{Q}}^{\mathrm{T}}\boldsymbol{\mathcal{Q}} = \mathbf{I}_n$. Multiplying $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ by $\boldsymbol{\mathcal{Q}}^{\mathrm{T}}$ implies that

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \|\boldsymbol{\mathcal{Q}}^{\mathrm{T}}\mathbf{y} - \boldsymbol{\mathcal{Q}}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}\|^2 = \left\| \boldsymbol{\mathcal{Q}}^{\mathrm{T}}\mathbf{y} - \left[\begin{array}{c} \mathbf{R} \\ \mathbf{0} \end{array} \right] \boldsymbol{\beta} \right\|^2 .$$

Defining p vector $\mathbf{f}$ and $n - p$ vector $\mathbf{r}$ so that $\left[\begin{array}{c} \mathbf{f} \\ \mathbf{r} \end{array} \right] \equiv \boldsymbol{\mathcal{Q}}^{\mathrm{T}}\mathbf{y}$, yields[4]

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \left\| \left[\begin{array}{c} \mathbf{f} \\ \mathbf{r} \end{array} \right] - \left[\begin{array}{c} \mathbf{R} \\ \mathbf{0} \end{array} \right] \boldsymbol{\beta} \right\|^2 = \|\mathbf{f} - \mathbf{R}\boldsymbol{\beta}\|^2 + \|\mathbf{r}\|^2.$$

The length of $\mathbf{r}$ does not depend on $\boldsymbol{\beta}$ and $\|\mathbf{f} - \mathbf{R}\boldsymbol{\beta}\|^2$ can be reduced to zero by choosing $\boldsymbol{\beta}$ so that $\mathbf{R}\boldsymbol{\beta}$ equals $\mathbf{f}$. Hence, provided that $\mathbf{X}$ and therefore $\mathbf{R}$ have full column rank,

$$\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{f} \tag{3}$$

is the least squares estimator of $\boldsymbol{\beta}$. Notice that $\|\mathbf{r}\|^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$, the *residual sum of squares* for the model fit.

3.2.1 Unbiasedness and variance

It is easy to show that $\hat{\boldsymbol{\beta}}$ is unbiased (has expectation equal to the true $\boldsymbol{\beta}$).

$$E(\hat{\boldsymbol{\beta}}) = E(\mathbf{R}^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{y}) = \mathbf{R}^{-1}\mathbf{Q}^{\mathrm{T}}E(\mathbf{y}) = \mathbf{R}^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \mathbf{R}^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\mathbf{R}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

The covariance matrix for $\hat{\boldsymbol{\beta}}$ is also straightforward. The covariance matrix of $\mathbf{y}$ is $\mathbf{I}\sigma^2$ and so the covariance matrix of $\mathbf{f} = \mathbf{Q}^{\mathrm{T}}\mathbf{y}$ is $\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\sigma^2 = \mathbf{I}\sigma^2$. Hence the covariance matrix of $\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{f}$ is

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \mathbf{R}^{-1}\mathbf{R}^{-\mathrm{T}}\sigma^2. \tag{4}$$

[3] That is, $R_{i,j} = 0$ if $i > j$

[4] If the final equality is not obvious recall that $\|\mathbf{x}\|^2 = \sum_i x_i^2$, so if $\mathbf{x} = \left[\begin{array}{c} \mathbf{v} \\ \mathbf{w} \end{array} \right]$, $\|\mathbf{x}\|^2 = \sum_i v_i^2 + \sum_i w_i^2 = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2$.

3.2.2 Checking

Further inference, beyond simply estimating $\boldsymbol{\beta}$, will require that the assumptions of independence and constant variance of the ϵ_i hold. This means that we need some means for checking these assumptions, if our inference is to be sound. Formally we will also assume normality of the ϵ_i, but in reality there is a certain robustness to this assumption, as a result of the central limit theorem.

The most useful checks of model assumptions are often graphical, as these tend to indicate not just that an assumption is violated, but also *how* it is violated. This in turn may allow us to fix the problem. Once we have estimated a linear model, all the information about model fit is contained in the model *residuals*

$$\hat{\epsilon}_i = y_i - \hat{\mu}_i$$

(where $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, of course). To check the model assumptions the $\hat{\epsilon}_i$ are usually plotted against the model predictor variables and the *fitted values*, $\hat{\mu}_i$. Systematic pattern in the mean of the residuals will violate the independence assumption, and often results from something wrong with how the dependence on a predictor variable has been specified. Systematic pattern in the variability of the residuals indicates violation of the constant variance assumption. This is most often manifested in the plot of $\hat{\epsilon}_i$ against $\hat{\mu}_i$.

3.3 Is $\hat{\boldsymbol{\beta}}$ a good estimator? The Gauss Markov Theorem

As presented above, the choice of minimizing the residual sum of squares to find $\hat{\boldsymbol{\beta}}$ was essentially arbitrary. Why not minimise $\sum_i |\epsilon_i|$ or $\sum_i \epsilon_i^4$ or $\sum \epsilon_i^2/\mu^2$ or something completely different? One of the central questions of frequentist statistical inference is whether an estimator is in some sense 'optimal'. To address this we need to define what makes a good estimator.

An obvious approach is to decide that we want $\hat{\boldsymbol{\beta}}$ to be 'as close as possible' to $\boldsymbol{\beta}$. For example we might want to favour estimators that have a low value for $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$, or its expected value. That is we might favour estimators that have low *mean square error*. However $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$, and similar criteria, are somewhat problematic. What if our data contain lots of information for estimating β_i and very little information for estimating β_j: does it then make sense to give equal weight to (squared) errors in β_j and β_i? Quite perverse outcomes are possible, especially in high dimensions, if we are not rather careful about such issues.

A somewhat less obvious, but theoretically easier, approach is to require estimators to be *unbiased*, meaning that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. Among unbiased estimators we then favour those with minimum variance. In the case of vector parameters *minimum variance* means smallest covariance matrix, in a sense that takes care of some of the issues around how to weight individual parameters relative to each other.

Unbiasedness has the attraction of a sort of honesty in the estimation process. It means that if we replicated our data gathering and estimation process repeatedly, then eventually the long run average of our replicate $\hat{\boldsymbol{\beta}}$s would settle down to $\boldsymbol{\beta}$. i.e. our $\hat{\boldsymbol{\beta}}$ has something of the property of fair dice. Given unbaisedness then minimising the estimator variance means that we are getting as close as possible to $\boldsymbol{\beta}$.

In the context of linear models these ideas lead us to an optimality result for least squares estimation, the *Gauss Markov Theorem*.

Theorem 1. Suppose that $\boldsymbol{\mu} \equiv E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{V}_y = \sigma^2\mathbf{I}$, and let $\tilde{\phi} = \mathbf{c}^{\mathrm{T}}\mathbf{Y}$ be any *unbiased* linear estimator of $\phi = \mathbf{t}^{\mathrm{T}}\boldsymbol{\beta}$, where $\mathbf{t}$ is an arbitrary vector. Then:

$$\mathrm{var}(\tilde{\phi}) \geq \mathrm{var}(\hat{\phi})$$

where $\hat{\phi} = \mathbf{t}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{Y}$ is the least squares estimator of $\boldsymbol{\beta}$ ($\mathbf{X} = \mathbf{QR}$). Notice that, since $\mathbf{t}$ is arbitrary, this theorem implies that each element of $\hat{\boldsymbol{\beta}}$ is a minimum variance unbiased estimator.

Proof. Since $\tilde{\phi}$ is a linear transformation of $\mathbf{Y}$, $\mathrm{var}(\tilde{\phi}) = \mathbf{c}^{\mathrm{T}}\mathbf{c}\sigma^2$. To compare variances of $\hat{\phi}$ and $\tilde{\phi}$ it is also useful to express $\mathrm{var}(\hat{\phi})$ in terms of $\mathbf{c}$. Because $\tilde{\phi}$ is unbiased we have

$$E(\mathbf{c}^{\mathrm{T}}\mathbf{Y}) = \mathbf{t}^{\mathrm{T}}\boldsymbol{\beta} \Rightarrow \mathbf{c}^{\mathrm{T}}E(\mathbf{Y}) = \mathbf{t}^{\mathrm{T}}\boldsymbol{\beta} \Rightarrow \mathbf{c}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \mathbf{t}^{\mathrm{T}}\boldsymbol{\beta} \Rightarrow \mathbf{c}^{\mathrm{T}}\mathbf{X} = \mathbf{t}^{\mathrm{T}}.$$

So the variance of $\hat{\phi}$ can be written as

$$\mathrm{var}(\hat{\phi}) = \mathrm{var}(\mathbf{t}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) = \mathrm{var}(\mathbf{c}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathrm{var}(\mathbf{c}^{\mathrm{T}}\mathbf{QR}\hat{\boldsymbol{\beta}}).$$

This is the variance of a linear transformation of $\hat{\boldsymbol{\beta}}$, and the covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\mathbf{R}^{-1}\mathbf{R}^{-\mathrm{T}}\sigma^2$, so

$$\mathrm{var}(\hat{\phi}) = \mathrm{var}(\mathbf{c}^{\mathrm{T}}\mathbf{Q}\mathbf{R}\hat{\boldsymbol{\beta}}) = \mathbf{c}^{\mathrm{T}}\mathbf{Q}\mathbf{R}\mathbf{R}^{-1}\mathbf{R}^{-\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{c}\sigma^2 = \mathbf{c}^{\mathrm{T}}\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{c}\sigma^2.$$

Hence,

$$\mathrm{var}(\tilde{\phi}) - \mathrm{var}(\hat{\phi}) = \mathbf{c}^{\mathrm{T}}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}})\mathbf{c}\sigma^2.$$

Because the columns of $\mathbf{Q}$ are orthogonal, $\mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\mathbf{Q}^{\mathrm{T}}$, and it follows that

$$\mathbf{c}^{\mathrm{T}}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}})\mathbf{c} = \{(\mathbf{I} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}})\mathbf{c}\}^{\mathrm{T}}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}})\mathbf{c} \geq 0,$$

since this is just the sum of squares of the elements of the vector $(\mathbf{I} - \mathbf{Q}\mathbf{Q}^{\mathrm{T}})\mathbf{c}$. Hence the result is proved. □

So amongst estimators that are unbiased and linear in $\mathbf{Y}$, least squares estimators have minimum variance. This leaves open the possibility that some non-linear estimator might sometimes be even better, and it could be that we could achieve lower overall error (variance + squared bias) by allowing our estimators to be biased, but we have at least got some theoretical justification for using least squares.

3.4 Further inference: intervals and testing

Having covered estimation and checking, we are now in a position to consider how to answer the inferential questions:

1. What range of parameter values are consistent with the data?
2. Are some pre-specified values (or restrictions) for the parameters consistent with the data?

In the specific case of linear models, we are want to find confidence intervals for the β_i and to *test hypotheses* about the β_i. For example, we often want to test the null hypothesis that one or several elements of $\boldsymbol{\beta}$ are zero, corresponding to terms being omitted from the linear model.

Confidence intervals and hypothesis tests are probabilistic concepts, so at this stage we need to turn our linear model into a fully probabilistic model, by specifying a full distribution for the ϵ_i. We will simply assume that independently $\epsilon_i \sim N(0, \sigma^2)$, which implies that $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2)$, and hence, upon recycling results from above

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathbf{V}_{\hat{\boldsymbol{\beta}}}) \text{ where } \mathbf{V}_{\hat{\boldsymbol{\beta}}} = \mathbf{R}^{-1}\mathbf{R}^{-\mathrm{T}}\sigma^2.$$

This is not yet in the most directly usable form: it gives the distribution of $\hat{\boldsymbol{\beta}}$ in terms of the unknowns $\boldsymbol{\beta}$ and σ^2. As we saw in section 2.1.1, if we want to make inferences about β_i it is helpful if we can start from a *pivotal* statistic, which has a distribution with no unknowns.

3.4.1 $(\hat{\beta}_i - \beta_i)/\hat{\sigma}_{\hat{\beta}_i} \sim t_{n-p}$

First, consider the distribution of $\boldsymbol{\mathcal{Q}}^{\mathrm{T}}\mathbf{y}$. It is a linear transform of a normal random vector, so must also be a normal random vector, and from standard theory on transformation matrices its covariance matrix is

$$\mathbf{V}_{\boldsymbol{\mathcal{Q}}^{\mathrm{T}}\mathbf{y}} = \boldsymbol{\mathcal{Q}}^{\mathrm{T}}\mathbf{I}\boldsymbol{\mathcal{Q}}\sigma^2 = \mathbf{I}\sigma^2.$$

Since, for the normal distribution only, zero covariance implies independence, the elements of $\boldsymbol{\mathcal{Q}}^{\mathrm{T}}\mathbf{y}$ are mutually independent. Furthermore,

$$E\begin{bmatrix} \mathbf{f} \\ \mathbf{r} \end{bmatrix} = E(\boldsymbol{\mathcal{Q}}^{\mathrm{T}}\mathbf{y}) = \boldsymbol{\mathcal{Q}}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}\boldsymbol{\beta}$$

$$\Rightarrow E(\mathbf{f}) = \mathbf{R}\boldsymbol{\beta} \text{ and } E(\mathbf{r}) = \mathbf{0}.$$

So we have that

$$\mathbf{f} \sim N(\mathbf{R}\boldsymbol{\beta}, \mathbf{I}_p\sigma^2) \;\text{ and }\; \mathbf{r} \sim N(\mathbf{0}, \mathbf{I}_{n-p}\sigma^2)$$

with both vectors independent of each other.

Now, because the $n-p$ elements of $\mathbf{r}$ are i.i.d. $N(0,\sigma^2)$ random variables,

$$\frac{1}{\sigma^2}\|\mathbf{r}\|^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n-p} r_i^2 \sim \chi^2_{n-p}.$$

The mean of a χ^2_{n-p} r.v. is $n-p$, so this result is sufficient (but not necessary) to imply that

$$\hat{\sigma}^2 = \|\mathbf{r}\|^2/(n-p) \tag{5}$$

is an unbiased estimator of σ^2. The independence of the elements of $\mathbf{r}$ and $\mathbf{f}$ also implies that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent.[5]

Now consider a single-parameter estimator, $\hat{\beta}_i$, with standard deviation, $\sigma_{\hat{\beta}_i}$, given by the square root of element i,i of $\mathbf{V}_{\hat{\beta}}$. An unbiased estimator of $\mathbf{V}_{\hat{\beta}}$ is $\hat{\mathbf{V}}_{\hat{\beta}} = \mathbf{V}_{\hat{\beta}}\hat{\sigma}^2/\sigma^2 = \mathbf{R}^{-1}\mathbf{R}^{-\mathrm{T}}\hat{\sigma}^2$, so an estimator, $\hat{\sigma}_{\hat{\beta}_i}$, is given by the square root of element i,i of $\hat{\mathbf{V}}_{\hat{\beta}}$, and it is clear that $\hat{\sigma}_{\hat{\beta}_i} = \sigma_{\hat{\beta}_i}\hat{\sigma}/\sigma$. Hence,

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} = \frac{\hat{\beta}_i - \beta_i}{\sigma_{\hat{\beta}_i}\hat{\sigma}/\sigma} = \frac{(\hat{\beta}_i - \beta_i)/\sigma_{\hat{\beta}_i}}{\sqrt{\frac{1}{\sigma^2}\|\mathbf{r}\|^2/(n-p)}} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-p}/(n-p)}} \sim t_{n-p} \tag{6}$$

(where the independence of $\hat{\beta}_i$ and $\hat{\sigma}^2$ has been used).

3.4.2 Confidence intervals for β_i

Suppose we want a $(1-2\alpha)100\%$ CI for β_i. Let $t_{n-p}(\alpha)$ denote the point above which a t_{n-p} r.v. lies with probability α. Then

$$\Pr\left(-t_{n-p}(\alpha) < \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} < t_{n-p}(\alpha)\right) = \Pr\left(\hat{\beta}_i - t_{n-p}(\alpha)\sigma_{\hat{\beta}_i} < \beta_i < \hat{\beta}_i + t_{n-p}(\alpha)\sigma_{\hat{\beta}_i}\right) = 1-2\alpha.$$

i.e. $\hat{\beta}_i \pm t_{n-p}(\alpha)\sigma_{\hat{\beta}_i}$ is the required interval.

3.4.3 Hypothesis tests about β_i

Another possibility is that we want to test $H_0 : \beta_i = \beta_{i0}$, for some given value β_{i0}, versus an alternative, such as $H_1 : \beta_i \neq \beta_{i0}$. If H_0 is true then

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{\hat{\sigma}_{\hat{\beta}_i}} \sim t_{n-p},$$

but otherwise we would expect the magnitude of T to usually be too large for consistency with a t_{n-p} distribution. Suppose that t_{obs} is our observed value of T. We can judge its compatibility with H_0 by assessing the probability of getting a t_{obs} value at least this large, *if H_0 is true*. Such a probability is known as *p-value* for H_0. A small value indicates that t_{obs}, and by extension the data, are improbable under the null hypothesis. This is evidence against H_0. Conversely if the p-value is large then t_{obs}, and hence the data, are quite probable under H_0, and we have no reason to doubt it.

The actual calculation is just a matter of assessing $p = \Pr(|T| \geq |t_{\text{obs}}|)$ where $T \sim t_{n-p}$. This can easily be done using the cumulative distribution function of the t_{n-p} distribution, built in to R, for example. Note two points:

1. From its definition, the p-value itself is an observation of a $U(0,1)$ r.v. if H_0 is true.

2. A $(1-2\alpha)100\%$ confidence interval for β_i is also the range of β_{i0} values that would result in a p-value $\geq 2\alpha$.

[5] Recall that $\|\mathbf{r}\|^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$.

These days it is usual to compute the p-value and use its magnitude to judge the evidence against the null hypothesis, H_0. However in the days before widespread availability of computers, calculating the p-value was difficult, as statisticians relied on printed tables of to assess the probability of test statistics. For this reason it was common practice to make a binary decision of whether to reject or accept H_0 on the basis of whether the p-value was above or below some threshold *significance* level, such as .05 or 0.01. Operating this way you could work out the range of test statistic values that would cause you to reject the null hypothesis, and see if your observed test statistic was within that range.

Since we are no longer reliant on tables of probabilities, binary reject/accept decisions are generally avoided in practice, in favour of quoting the p-value itself. This is because the the actual p-value contains substantially more information about the evidence against H_0 than is contained in a statement like 'H_0 was rejected at the 5% level'. The latter fails to differentiate between p-values of 0.049 and 10^{-10}, for example.

However the notion of accepting and rejecting tests is still useful in two circumstances. The first is in the theory of testing, when we want to consider what makes a good test. In that case the probability of rejecting a false H_0 is the *power* of a test, and higher is obviously better. Power depends on the significance level and the true value of β_i, of course. The second case is when we are performing large numbers of tests on the same dataset. From the definition of p-values, you can see that if we test at a significance level of 5%, for example, then we will have a 5% probability of rejecting H_0 when it is actually true. If we perform a large number of tests, then we can expect a substantial proportion of our rejected null hypotheses to be this sort of 'false discovery'. This is a problem that requires some work to control.

Later on we will return to the theory of testing in more generality. For now let us move on to a slightly more general linear model testing problem.

3.4.4 Testing $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$

Recall that factor variables serve to classify observations into different groups. In a linear model each factor variable will usually have several associated parameters - in principle, one for each group. If we want to test whether the factor variable should be in the model or not, we therefore need to test whether all the factor's parameters could simultaneously be zero. This is a special case of the general null hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$. Note that approaching such a testing problem with a separate test for each parameter of the factor is a bad idea: how would you go about combining the different p-values, for example?

Suppose that $\mathbf{C}$ is a $q \times p$ matrix and $\mathbf{d}$ a q vector, where $q < p$. Under H_0 we have $\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d} \sim N(\mathbf{0}, \mathbf{C}\mathbf{V}_{\hat{\beta}}\mathbf{C}^{\mathrm{T}})$, from basic properties of the transformation of normal random vectors. Any positive definite matrix can be factored into the crossproduct of a triangular matrix with itself: a *Cholesky decomposition*. Forming the Cholesky decomposition $\mathbf{L}^{\mathrm{T}}\mathbf{L} = \mathbf{C}\mathbf{V}_{\hat{\beta}}\mathbf{C}^{\mathrm{T}}$, it is then easy to show that $\mathbf{L}^{-\mathrm{T}}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \sim N(\mathbf{0}, \mathbf{I})$, so,

$$(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^{\mathrm{T}}(\mathbf{C}\mathbf{V}_{\hat{\beta}}\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) = (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{L}^{-\mathrm{T}}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \sim \sum_{i=1}^{q} N(0,1)^2 \sim \chi^2_q.$$

As in the previous section, plugging in $\hat{\mathbf{V}}_{\hat{\beta}} = \mathbf{V}_{\hat{\beta}}\hat{\sigma}^2/\sigma^2$ gives the computable test statistic and its distribution under H_0:

$$\begin{aligned} F = \frac{1}{q}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^{\mathrm{T}}(\mathbf{C}\hat{V}_{\hat{\beta}}\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) &= \frac{\sigma^2}{q\hat{\sigma}^2}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^{\mathrm{T}}(\mathbf{C}V_{\hat{\beta}}\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \\ &= \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^{\mathrm{T}}(\mathbf{C}V_{\hat{\beta}}\mathbf{C}^{\mathrm{T}})^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})/q}{\frac{1}{\sigma^2}\|\mathbf{r}\|^2/(n-p)} \sim \frac{\chi^2_q/q}{\chi^2_{n-p}/(n-p)} \sim F_{q,n-p}. \quad (7) \end{aligned}$$

This result can be used to compute a p-value for the test.

When our interest is in testing whether several elements of $\boldsymbol{\beta}$ could simultaneously be zero, then $\mathbf{d} = \mathbf{0}$ while $\mathbf{C}$ is made up of a selection of rows of the $p \times p$ identity matrix. In that case let RSS_1 denote the residual sum of squares for the full model, and RSS_0 denote the residual sum of squares for the reduced model in which the H_0 is true. It can be shown that the above test statistic then reduces to

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/q}{\mathrm{RSS}_1/(n-p)}.$$

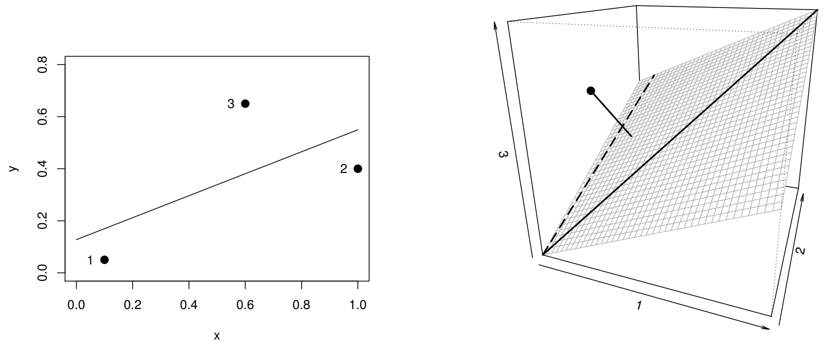

Figure 1: Illustration of the geometry of least squares. Left: a straight line fit to three x, y data. Right: the space in which the y coordinates of the data define a single point, while the columns of the model matrix (solid and dashed line) span the subspace shown in grey. The least squares estimate of $E(\mathbf{y})$ is the orthogonal projection of the data point onto the model subspace.

3.5 The geometry of linear models

Least squares estimation of linear models amounts to finding the orthogonal projection of the n dimensional response data $\mathbf{y}$ onto the p dimensional linear subspace spanned by the columns of $\mathbf{X}$. The linear model states that $E(\mathbf{y})$ lies in the space spanned by all possible linear combinations of the columns of the $n \times p$ model matrix $\mathbf{X}$, and least squares seeks the point in that space that is closest to $\mathbf{y}$ in Euclidean distance. Figure 1 illustrates this geometry for the model,

$$\begin{bmatrix} .05 \\ .40 \\ .65 \end{bmatrix} = \begin{bmatrix} 1 & .1 \\ 1 & 1 \\ 1 & .6 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}.$$

Given this view of least squares fitting as orthogonal projection it is interesting to look at the projection matrix that maps the data $\mathbf{y}$ to the fitted values $\hat{\boldsymbol{\mu}}$. We have

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Q}\mathbf{R}\mathbf{R}^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{y} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{y}.$$

So the required projection matrix is $\mathbf{A} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}}$, which is often known as the *influence matrix*, or *hat matrix* of the linear model. (Don't forget that $\mathbf{Q}$ is an $n \times p$ matrix with orthogonal columns, but non-orthogonal rows, so that $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$ but $\mathbf{Q}\mathbf{Q}^{\mathrm{T}} \neq \mathbf{I}$.)

One interesting property of $\mathbf{A}$ is that it is idempotent: $\mathbf{A}\mathbf{A} = \mathbf{A}$. In a way this is obvious, as the orthogonal projection of $\hat{\boldsymbol{\mu}}$ onto the column space of $\mathbf{X}$ must be $\hat{\boldsymbol{\mu}}$.

3.6 Results in terms of X

The presentation so far has been in terms of the QR based method actually used to fit linear models in practice. By taking this approach, results (6) and (7) can be derived relatively easily. However, for historical reasons, these results are more usually presented in terms of the model matrix, $\mathbf{X}$, rather than the components of its QR decomposition.

For example the covariance matrix of $\hat{\boldsymbol{\beta}}$ becomes $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2$, which is easily seen to be equivalent to (4):

$$\begin{aligned} \mathbf{V}_{\hat{\boldsymbol{\beta}}} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2 = \left(\mathbf{R}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\mathbf{R}\right)^{-1}\sigma^2 = \left(\mathbf{R}^{\mathrm{T}}\mathbf{R}\right)^{-1}\sigma^2 \\ &= \mathbf{R}^{-1}\mathbf{R}^{-\mathrm{T}}\sigma^2. \end{aligned}$$

The expression for the least squares estimates is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$, which is equivalent to (3):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{R}^{-1}\mathbf{R}^{-\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{y} = \mathbf{R}^{-1}\mathbf{Q}^{\mathrm{T}}\mathbf{y} = \mathbf{R}^{-1}\mathbf{f}.$$

It follows that the influence/hat matrix can be written as $\mathbf{A} = \mathbf{X}(\mathbf{X}^{\rm T}\mathbf{X})^{-1}\mathbf{X}^{\rm T}$. These results are of theoretical interest, but should not usually be used for computational purposes.

3.7 Bayesian analysis

For a fully Bayesian analysis of the linear model we need to supply prior distributions for $\boldsymbol{\beta}$ and σ^2. To obtain analytical tractability we need to use conjugate priors. For the regression coefficients use $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{\psi}^{-1})$ where $\boldsymbol{\beta}_0$ and $\boldsymbol{\psi}$ are given. Defining the precision $\tau = 1/\sigma^2$ then $\tau \sim G(a, b)$ is conjugate, where $G(a, b)$ denotes a gamma random variable with p.d.f. $f(\tau) = b^a\tau^{a-1}e^{-b\tau}/\Gamma(a)$. So a, b, $\boldsymbol{\beta}_0$ and $\boldsymbol{\psi}$ are quantities to which we would have to give actual values, for a practical analysis. Now the joint distribution of data and parameters is

$$f(\mathbf{y}, \boldsymbol{\beta}, \tau) \propto \tau^{n/2}e^{-\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2\tau/2}e^{-(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\rm T}\boldsymbol{\psi}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)/2}e^{-b\tau}\tau^{a-1}.$$

The distribution of $\tau|\boldsymbol{\beta}, \mathbf{y}$ is proportional to this, and can be identified by reading off the factors dependent on τ (everything else being part of the normalising constant)...

$$f(\tau|\boldsymbol{\beta}, \mathbf{y}) \propto \tau^{n/2+a-1}e^{-\tau(b+\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2/2)}.$$

This is immediately identifiable as a $G(n/2 + a, b + \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2/2)$ distribution. Similarly the $\boldsymbol{\beta}$ dependent factor from the joint distribution lets us identify the distribution of $\boldsymbol{\beta}|\tau, \mathbf{y}$.

$$\begin{aligned} f(\boldsymbol{\beta}|\tau, \mathbf{y}) &\propto e^{-\frac{1}{2}(\boldsymbol{\beta}^{\rm T}\mathbf{X}^{\rm T}\mathbf{X}\boldsymbol{\beta}\tau - 2\boldsymbol{\beta}\mathbf{X}^{\rm T}\mathbf{y}\tau + \boldsymbol{\beta}^{\rm T}\boldsymbol{\psi}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\rm T}\boldsymbol{\psi}\boldsymbol{\beta}_0)} \\ &\propto e^{-\frac{1}{2}\{\boldsymbol{\beta}-(\mathbf{X}^{\rm T}\mathbf{X}\tau+\boldsymbol{\psi})^{-1}(\tau\mathbf{X}^{\rm T}\mathbf{y}+\boldsymbol{\psi}\boldsymbol{\beta}_0)\}^{\rm T}(\mathbf{X}^{\rm T}\mathbf{X}\tau+\boldsymbol{\psi})\{\boldsymbol{\beta}-(\mathbf{X}^{\rm T}\mathbf{X}\tau+\boldsymbol{\psi})^{-1}(\tau\mathbf{X}^{\rm T}\mathbf{y}+\boldsymbol{\psi}\boldsymbol{\beta}_0)\}}, \end{aligned}$$

which is recognisable as a $N((\mathbf{X}^{\rm T}\mathbf{X}\tau + \boldsymbol{\psi})^{-1}(\tau\mathbf{X}^{\rm T}\mathbf{y} + \boldsymbol{\psi}\boldsymbol{\beta}_0), (\mathbf{X}^{\rm T}\mathbf{X}\tau + \boldsymbol{\psi})^{-1})$ distribution.

If the sample size tends to infinity so that $\mathbf{X}^{\rm T}\mathbf{X}\tau$ dominates $\boldsymbol{\psi}$, or if the prior precision matrix tends to the zero matrix, then $f(\boldsymbol{\beta}|\tau, \mathbf{y})$ tends to a $N(\hat{\boldsymbol{\beta}}, (\mathbf{X}^{\rm T}\mathbf{X})^{-1}\sigma^2)$ density, so that once again Bayesian and frequentist intervals for $\boldsymbol{\beta}$ will coincide.

Notice that we have not quite produced the full joint posterior density of $\boldsymbol{\beta}, \tau|\mathbf{y}$, but rather the conditional densities of $\boldsymbol{\beta}|\tau, \mathbf{y}$ and $\tau|\boldsymbol{\beta}, \mathbf{y}$. In practice there are several ways to proceed using just these. One option is to iteratively find the posterior modes of $\boldsymbol{\beta}$ given the estimated mode of τ and the posterior mode of τ given the estimated modes of $\boldsymbol{\beta}$, until the mode of τ converges, and then just plug this into the conditional density of $\boldsymbol{\beta}$. An alternative is to integrate $\boldsymbol{\beta}$ out of $f(\tau|\boldsymbol{\beta}, \mathbf{y})$ to obtain $f(\tau|\mathbf{y})$. This *marginal likelihood* can then be maximized to find $\hat{\tau}$, which is plugged into $f(\boldsymbol{\beta}|\tau, \mathbf{y})$ (this procedure is known as 'empirical Bayes'). Finally we can simply alternate simulation of $\boldsymbol{\beta}$ from $f(\boldsymbol{\beta}|\tau, \mathbf{y})$, given (simulated) τ with simulation from $f(\tau|\boldsymbol{\beta}, \mathbf{y})$ given the last simulated $\boldsymbol{\beta}$, to generate joint draws $\tau, \boldsymbol{\beta}$ from the posterior $f(\boldsymbol{\beta}, \tau|\mathbf{y})$. This approach is known as *Gibbs sampling*. The generated $\tau, \boldsymbol{\beta}$ vectors have some serial correlation, but otherwise are a random sample from the posterior, which can be used as the basis for inference. We will look in more detail at this sort of Bayesian stochastic simulation later.

4 Causality, confounding and randomization

There are several different purposes for which we may wish to use the methods of statistical inference.

1. For prediction. For example we build a model of patients' probability of heart disease based on lifestyle factors such as diet and exercise, estimate the model from a representative group of subjects whose heart disease status is known, and use it to predict the status of new subjects.

2. To establish association between variables. For example we suspect that being educated in a single sex school may be negatively associated with length of marriage, and use statistical modelling to assess whether this putative association is real, or could just be due to chance variability in the data.

3. To establish causation. Often, especially in science, we want to know whether something is an actual cause of something else. Famous examples are: does smoking cause cancer? and does the MMR vaccine cause autism?

The methods we have covered so far are quite sufficient for addressing 1 and 2, but the *causal inference* required in 3 is more difficult, and the sort of methods covered so far are not sufficient to establish causality *on their own*.

To understand the issues, it is important to really have grasped the difference between correlation (association) and causality. An oft quoted example is the correlation between the population of storks and the birth rate, in Europe. The correlation is real, but there is obviously no causation. Rather, industrialization raised health care, living and educational standards in Europe leading to reduced birth rates, but that same industrialization led to a reduction in suitable habitat for storks. That is, the association is caused by other factors which may be causative to both variables of interest.

Confusing correlation with causation is a famously good way of selling newspapers. A study from the university of Bath was reported as revealing that men who preferred women with a smaller than usual hip to waist ratio were more likely to have autistic children. The association was real, but it was the interpretation in terms of the preference being somehow causative that made the story interesting. In fact, the likely explanation of the association is that higher than average testosterone in the womb seems to be a causative factor in autism, and women with higher than average testosterone tend to have smaller than usual hip to waist ratio. For obvious reasons they also tend to have children with men who find them attractive. This explanation makes for a duller press release, of course. Again the key is a hidden variable related to the variables of direct interest.

4.1 Confounding: some simulated examples

To appreciate the effect of correlated predictors on interpretation and attempts at causal inference, it is worth looking at some simulations, where we are in complete control of the correlation structure between variables. The following code simulates two correlated predictor variables, x and h:

```
require(mgcv) ## supplies rmvn
V <- matrix(c(1,.95,.95,1),2,2) ## cov matrix
n <- 1000 ## number of data
set.seed(7) ## just for reproducibility
X <- rmvn(n,rep(0,2),V)
x <- X[,1];h <- X[,2]
```

Now consider the case when $y_i = x_i + \epsilon_i$, but we fit the model $y_i = \beta_0 + \beta_1 x_i + \beta_2 h_i + \epsilon_i$.

```
> y <- x + rnorm(n)
> summary(lm(y ~ x + h))
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.003268   0.031170  -0.105   0.9165
x            0.812568   0.098924   8.214 6.59e-16 ***
h            0.209603   0.099475   2.107   0.0354 *
...
> summary(lm(y ~ x ))
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.002117   0.031219  -0.068    0.946
x            1.010417   0.031191  32.395   <2e-16 ***
...
```

Because of the high correlation between x and h it is difficult to distinguish their effects, with the result that when both are included in the model the ‘explanation’ of the response tends to get shared out between them, leading to rather variable estimates of the real effect of x. In contrast when the spurious h term is omitted from the model, the estimate of β_1 is much more accurate. Actually, for most replicates of the data simulation, β_2 would not have appeared to be significantly different from zero, so we would have been likely to drop the term, and arrive at the correct model using just the inferential tools already developed.

However, it is worth also looking at what would have happened if we had only observe h but not x:

```
> summary(lm(y ~ h))
...
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.007193   0.032187  -0.223    0.823
h            0.985161   0.032336  30.466   <2e-16 ***
...
```

The high correlation between h and x means that we find a strongly significant association between h and y, despite the fact that h played no part at all in the simulation of y! In this case x is an example of a hidden *confounding variable*: it is correlated with both our predictor, h, and our response y, and therefore messes up our inference about the true causal influence of h on y (which is zero in this case).

Here is a second example of confounding. Now both h and x have a causal effect on y, and if we only observe x, we will overestimate its true effect (overestimate, because the correlation is positive).

```
> y <- x + h + rnorm(n)
> summary(lm(y ~ x ))
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01714    0.03249   0.528    0.598
x            1.92020    0.03246  59.155   <2e-16 ***
...
> summary(lm(y ~ x + h))
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01257    0.03145   0.400    0.689
x            1.13516    0.09980  11.374  < 2e-16 ***
h            0.83168    0.10036   8.287 3.71e-16 ***
...
```

i.e. when h is hidden from us, our estimate of β_1 is almost twice as large as it should be in this case.

Note that this tendency of the model fitting to try and compensate for missing confounding variables, is a profound difficulty for causal inference, but a blessing for prediction. The fact that the model has adjusted for the missing confounders in this way actually improves prediction.

4.2 Controlled experiments and randomization

The gold standard for establishing causality is to analyse data from a properly designed experiment, in which we use the process of *randomization* to break any possible association between unmeasured confounders and the predictor whose effect we want to measure. The idea is basically to turn the systematic variation in the response caused by confounders into something that we can model as independent random variability between experimental units.

An example is the most useful way to convey the idea. Suppose that we want to examine the relationship between exercise and fat mass in people aged 40 to 50. There are a huge number of factors contributing to people's fat mass, and if we simply look at a sample of people who already exercise by different amounts it will be very difficult to isolate the effect of exercise separate from all the other confounders (diet, working hours, profession, whether they have children, drinking habits, smoking, wealth etc.). Suppose instead that we set up a study in which participants enrol and are then assigned to one of several 'treatments', consisting of exercise programs of differing intensity (or a control of none). Now if we assign subjects to the different treatments randomly then there can be no possible association between all the other variables contributing to fat mass, and the one that we are interested in: amount of exercise. Indeed all the variability in fat mass attributable to the confounders can now be treated as random variability within each treatment. Any effect of exercise that we then find is causal: the thing we controlled having an effect on the variable we are interested in. Of course this approach does not stop us from including measured covariates in the analysis model, in order to reduce the amount of variability being modelled as random, and thereby increase precision, but it does mean that we no longer have to worry about the effect of hidden confounders.

To appreciate the issue mathematically, consider a linear model for which the full model matrix is $(\mathbf{X}, \mathbf{H})$ (w.l.o.g. assume that the columns of $\mathbf{H}$ are centred). If we had not observed the variables in $\mathbf{H}$, then our estimates of the effects of $\mathbf{X}$ would be $\tilde{\boldsymbol{\beta}}_x = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$, whereas if we observed all the data then it would be

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_x \\ \hat{\boldsymbol{\beta}}_h \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{H} \\ \mathbf{H}^{\mathrm{T}}\mathbf{X} & \mathbf{H}^{\mathrm{T}}\mathbf{H} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^{\mathrm{T}} \\ \mathbf{H}^{\mathrm{T}} \end{bmatrix} \mathbf{y}.$$

Clearly if $\mathbf{X}^{\mathrm{T}}\mathbf{H} \neq \mathbf{0}$ then $\tilde{\boldsymbol{\beta}}_x \neq \hat{\boldsymbol{\beta}}_x$, that is the missing confounders in $\mathbf{H}$ interfere with the estimation of $\boldsymbol{\beta}_x$. However if the confounders are independent of $\mathbf{X}$, then at least in the large sample limit, $\mathbf{X}^{\mathrm{T}}\mathbf{H} = \mathbf{0}$, so that

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_x \\ \hat{\boldsymbol{\beta}}_h \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}^{\mathrm{T}}\mathbf{H} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^{\mathrm{T}} \\ \mathbf{H}^{\mathrm{T}} \end{bmatrix} \mathbf{y} = \begin{bmatrix} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{X}^{\mathrm{T}} \\ \mathbf{H}^{\mathrm{T}} \end{bmatrix} \mathbf{y}.$$

i.e. when the confounders are independent of the response we get the same estimate of $\boldsymbol{\beta}_x$ whether or not they are included in the model (and its expected value is correct). Randomization ensures that this occurs.

4.3 Instrumental variables

So properly designed controlled experiments with random allocation of experimental units to treatments are the best way of establishing causal effects. However there are many cases in which experimental manipulation is impractical, unethical and/or illegal. For example, if we are interested in establishing the causative effect of alcohol consumption on heart disease, it is simply unethical to conduct a study in which we randomly allocate subjects to a heavy drinking treatment (the same goes for any treatment where we have prior cause to suspect it may be harmful). Economics is another field beset with such problems: for example it is impractical to conduct a controlled experiment to establish the causative factors controlling a company's share price, and anything that came close would likely count as illegal market manipulation. This sort of problem is so ubiquitous in economics that it is from the field of *econometrics* that much of the work on causal inference from observational data comes.

In this section we will look at one method: instrumental variables. The approach can be applied more widely than just to linear models, but the ideas are easiest to grasp in the linear model context. Let's again use the setup in which $\mathbf{X}$ is the model matrix for the observed effects, $\mathbf{H}$ is the (column centred) model matrix for the hidden confounders and the response data are generated by,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_x + \mathbf{H}\boldsymbol{\beta}_h + \boldsymbol{\epsilon}.$$

Since we don't have access to $\mathbf{H}$ we fit $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_x + \mathbf{e}$, and therein lies the problem, since in reality $\mathbf{e} = \mathbf{H}\boldsymbol{\beta}_h + \boldsymbol{\epsilon}$, which is unlikely to meet the linear model assumptions about the residual vector. Indeed,

$$E(\hat{\boldsymbol{\beta}}_x) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}, \mathbf{H})\boldsymbol{\beta} = \boldsymbol{\beta}_x + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{H}\boldsymbol{\beta}_h \neq \boldsymbol{\beta}_x.$$

Now the problem here is that the space spanned by $\mathbf{X}$ is not orthogonal to $\mathbf{e}$, because of the interdependence of $\mathbf{X}$ and $\mathbf{H}$. This would not be a problem if we could somehow orthogonalize $\mathbf{X}$ and $\mathbf{e}$, for example by projecting $\mathbf{X}$ onto a space that is orthogonal to $\mathbf{e}$.

Suppose then, that we have available some *instrumental variables* giving rise to a (column centred) model matrix $\mathbf{Z}$, where $\mathbf{Z}$ is not part of the true model for $\mathbf{y}$, but is correlated with $\mathbf{X}$ *and* is independent of $\mathbf{H}$ (and hence $\mathbf{e}$). We'll assume that $\mathbf{Z}$ has at least as high a rank as $\mathbf{X}$, and w.l.o.g. its columns are centred to have zero mean. We could then project $\mathbf{X}$ onto the column space of $\mathbf{Z}$. That is replace $\mathbf{X}$ with $\mathbf{A}_z\mathbf{X}$ where $\mathbf{A}_z = \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}$, and regress $\mathbf{y}$ on $\mathbf{A}_z\mathbf{X}$. In that case

$$\hat{\boldsymbol{\beta}}_x = (\mathbf{X}^{\mathrm{T}}\mathbf{A}_z\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{A}_z\mathbf{y}$$

and we now have

$$E(\hat{\boldsymbol{\beta}}_x) = (\mathbf{X}^{\mathrm{T}}\mathbf{A}_z\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{A}_z\mathbf{X}\boldsymbol{\beta}_x + (\mathbf{X}^{\mathrm{T}}\mathbf{A}_z\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{A}_z\mathbf{H}\boldsymbol{\beta}_h \simeq \boldsymbol{\beta}_x,$$

since $\mathbf{A}_z\mathbf{H} \simeq \mathbf{0}$ as a result of $\mathbf{Z}^{\mathrm{T}}\mathbf{H} \simeq \mathbf{0}$ by the independence of $\mathbf{Z}$ and $\mathbf{H}$ (alternatively we could also take expectations over the distribution of the predictors to obtain equality).

Here is a simulated illustration. First simulate observed x, confounder h and instrument z, by simulating multivariate random variables with appropriate correlation structure, using x and h to simulate response y.

```
V <- matrix(c(1,.7,0.9,.7,1,0,.9,0,1),3,3)
library(mgcv); n <- 1000; set.seed(0)
X <- rmvn(n,rep(0,3),V)
x <- X[,1]; h <- X[,2]; z <- X[,3]
y <- x + h + rnorm(n)*.3
```

Now fit the model naively, and then again using the instrumental variable.

```
> summary(lm(y ~ x)) ## naive fit
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01760    0.02426   0.726    0.468
x            1.70146    0.02425  70.173   <2e-16 ***
...
> zx <- fitted(lm(x~z-1)) ## project x onto z
> summary(lm(y ~ zx)) ## fit iv model
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.08042    0.05434    1.48    0.139
zx           1.01547    0.07557   13.44   <2e-16 ***
...
```

Clearly the IV model gives an estimate that is much closer to correct. The price paid is precision: the standard error is substantially larger than we would get if h had been available for inclusion in the model (and that is without correcting for the uncertainty in the regression of x on z). This approach is not the only way of using instrumental variables, and their use is not restricted to linear models, but this method is sufficient for introducing the idea.

The success of such methods rests on being able to find good instrumental variables, which are independent of the potential confounders, but strongly correlated with the observed variables (weak correlation will inflate the standard errors of the estimates). Another way of stating this is that the instrumental variables must be correlated with the response variable *only via the instrument's correlation with the observed predictor variables*.

Finding such variables is obviously not easy, and one might question how often such a magical combination of correlation with $\mathbf{X}$ but independence of $\mathbf{H}$ can really apply, and be knowable when the confounders are not. However the technique of 'Mendelian randomization', provides an application of instrumental variable use that does seem to be practical. Take the example of alcohol and heart disease. There are genetic polymorphisms that predict the propensity to consume alcohol, but are completely unrelated to any other plausible predictors of heart disease. Hence a subject's genotype (with respect to this polymorphism) can be used as an instrumental variable. The method is called 'Mendelian randomization' after the genetic pioneer Gregor Mendel, and the fact that the genotype is assigned randomly by meiosis when passed from parents to offspring, thereby avoiding correlation with things that it does not actually cause.

'Mendelian randomization' is important in epidemiology when trying to uncover health effects of environmental exposures that can not (ethically) be manipulated experimentally, however finding appropriate polymorphisms and establishing that they are uncorrelated with any unobserved confounders is obviously challenging.

5 Beyond linear models

Suitably chastened by the difficulties associated with ascertaining causation, let's move on with some more straightforwardly mathematical theory. In particular, it is time to move on from the linear model, to consider statistical models in greater generality. Here are some motivating examples.

1. The linear models covered so far admit a rich range of possibilities for modelling the expected value of the response variable, while permitting only the simplest possible model of variability about that mean. Many situations demand a richer structure to model the randomness in the response, and linear mixed models provide this by allowing a linear structure for the randomness that is similar to that allowed for the mean. Specifically,
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon} \text{ where } \mathbf{b} \sim N(0, \boldsymbol{\psi}(\boldsymbol{\theta})) \text{ and } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2).$$

$\mathbf{Z}$ is a model matrix constructed in a similar way to $\mathbf{X}$, but $\mathbf{b}$ is a vector of *random effects*. The random effects covariance matrix $\boldsymbol{\psi}(\boldsymbol{\theta})$ typically has some relatively simple structure and is dependent on a parameter vector $\boldsymbol{\theta}$ which is the target of inference, alongside $\boldsymbol{\beta}$ and σ^2. For example, suppose we want to model the time course of repeated blood pressure measurements of patients starting on different blood pressure treatments. As well as the dependence of blood pressure on treatment and time, included in $\mathbf{X}\boldsymbol{\beta}$ we should also allow for the differences between individual patients, and might model these by including a random effect for each patient, and perhaps also some random slope with respect to time for each patient: these random components are included in $\mathbf{Zb}$. Our existing linear model theory is insufficient for inference with these *linear mixed models*.

2. In many situations a Gaussian distribution is inappropriate as a model for a response variable of interest, although a linear model structure might be appropriate for the expected value of the response. For example a discrete response variable might follow a Poisson or binomial distribution, or a positive continuous response variable might be better described by a gamma distribution. Our existing theory is then insufficient: for a start all of these distributions break the ordinary linear model constant variance assumption.

3. The bone marrow survival model, in section 1, is similarly different to anything we can express as a linear model.

So without actually straying very far from the linear model (these models all still have a univariate 'response variable', for example), we have several obvious motivations to develop inference methods that apply beyond the linear model.

Whether we take a Bayesian or frequentist approach to these models, the *likelihood* will be central to the inference methods we will discuss here. That is, if $\boldsymbol{\theta}$ denotes the model parameters and $\mathbf{y}$ the data that we are treating as random, everything will depend on $f(\mathbf{y}|\boldsymbol{\theta})$, the probability density[6] of $\mathbf{y}$ given $\boldsymbol{\theta}$, evaluated at the observed data. For the frequentist approach the theory will revolve around using the likelihood to estimate parameters, and involves developing some general methods for doing this. In the Bayesian case, there are no new principles involved: we use Bayes' theorem as before, however to make this practical in general we have to develop new methods for either approximating, or simulating from, the posterior.

The methods we will cover here are not the only approaches to frequentist or Bayesian inference. In the frequentist case there are alternatives and generalizations, such as the method of moments and M-estimation, which my be useful when we require robustness to getting some aspect of the model specification wrong. Similarly *Approximate Bayesian Computation* is a method that is useful when it is too difficult to work directly with the likelihood, but simulation from the model is relatively easy.

6 Maximum likelihood estimation

We'll start with the frequentist approach: parameters are fixed states of nature and the uncertainty is all in our estimates of those parameters. The key idea of *maximum likelihood estimation* is simply this:

> Parameter values that make the observed data appear relatively probable are more *likely* to be correct than parameter values that make the observed data appear relatively improbable.

For example, we would much prefer an estimate of $\boldsymbol{\theta}$ that assigned a probability density of 0.1 to our observed $\mathbf{y}$, according to the model, to an estimate for which the density was 0.00001.

So the idea is to judge the *likelihood* of parameter values using $f_\theta(\mathbf{y})$, the model p.d.f. according to the given value of $\boldsymbol{\theta}$, evaluated at the observed data. Because $\mathbf{y}$ is now fixed and we are considering the likelihood as a function of $\boldsymbol{\theta}$, it is usual to write the likelihood as $L(\boldsymbol{\theta}) \equiv f_\theta(\mathbf{y})$. In fact, for theoretical and practical purposes it is usual to work with the log likelihood $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$. The *maximum likelihood estimate* (MLE) of $\boldsymbol{\theta}$ is then

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}}\, l(\boldsymbol{\theta}).$$

[6] Unless noted otherwise, everything discussed for probability densities applies equally to probability functions, provided we replace integration with summation where appropriate.

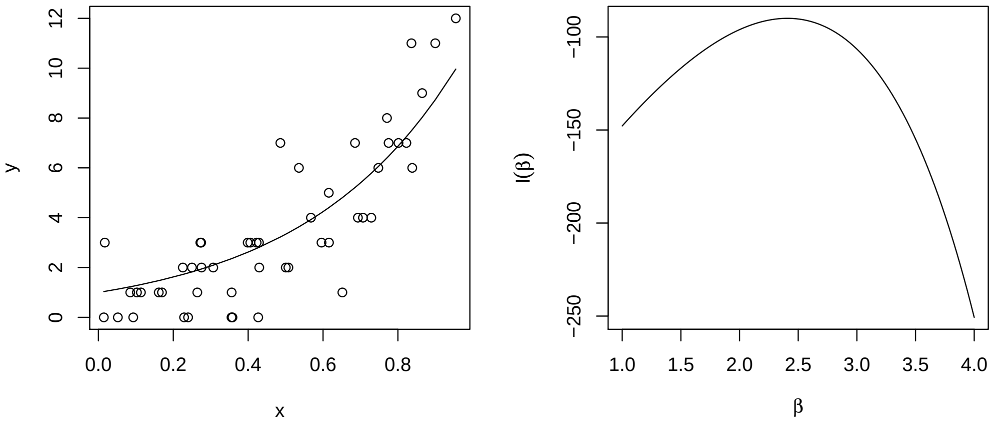

Figure 2: Left: 50 Poisson data, y, dependent on predictor data, x, and described by the model in section 6.1. The black curve is the maximum likelihood estimate of the Poisson mean as a function of x. Right: the log likelihood function for the model parameter β, computed using the data in the left panel and as described in the text.

In general, numerical methods will be needed to actually find $\hat{\boldsymbol{\theta}}$, but reliable and general methods are available.

There is more to maximum likelihood estimation than just its intuitive appeal. As we will see shortly, under some regularity conditions and in the large sample limit as $n \to \infty$,

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \boldsymbol{\mathcal{I}}^{-1}), \tag{8}$$

where $\boldsymbol{\mathcal{I}} = -E(\partial^2 l / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}})$ — the expected second derivative matrix of the negative log likelihood (also known as the *information matrix*). This turns out to be the best that can be achieved for an unbiased estimator (although MLEs are only unbiased in the large sample limit). In fact a similar result applies in terms of the observed (rather than expected) second derivative matrix. If $\hat{\boldsymbol{\mathcal{I}}} = -\partial^2 l / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}$ then in the $n \to \infty$ limit,

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \hat{\boldsymbol{\mathcal{I}}}^{-1}).$$

Another large sample result is also useful. Suppose that we want to compare two *nested* models, meaning that one model can be written as the other model subject to r restrictions on its parameters, $\boldsymbol{\theta}$, that can be written as $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$. Let $\hat{\boldsymbol{\theta}}_0$ denote the MLEs under the restriction. Then we can test $H_0 : \mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$ (i.e. test the null hypothesis that the restricted model is correct), by using the large sample result that

$$2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_0)\} \sim \chi^2_r \tag{9}$$

if the null hypothesis is correct. If the null hypothesis is not correct then the log likelihood ratio statistic on the left hand side will tend to be too large for consistency with a χ^2_r distribution. To actually judge compatibility of data and null hypothesis we compute a p-value — the probability of a χ^2_r random variable being at least as large as the observed test statistic. As usual a small p-value is evidence against the null hypothesis.

6.1 A simple example

Consider the one parameter Poisson model:

$$y_i \underset{\text{ind}}{\sim} \text{Poisson}\{\exp(\beta x_i)\}$$

and suppose that we have n pairs of x_i, y_i data. Generally the probability function for a Poisson random variable with mean λ_i is $f(y_i) = \lambda_i^{y_i} \exp(-\lambda_i)/y_i!$. So for our model

$$f(y_i) = \exp(\beta x_i)^{y_i} \exp\{-\exp(\beta x_i)\}/y_i!$$

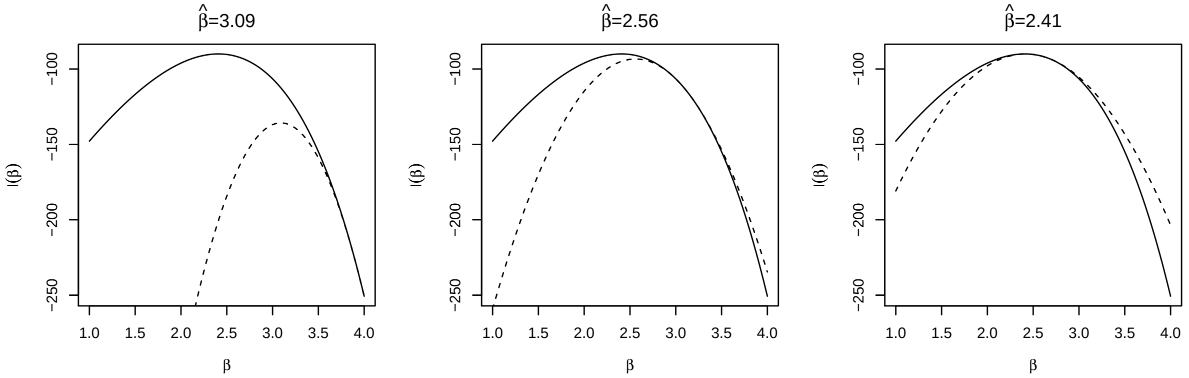

Figure 3: Three steps of Newton's method applied to the log likelihood for the Poisson example of section 6.1, starting on the left from $\beta = 4$. The dashed curve shows the approximating quadratic matching the value and first two derivatives of the log likelihood at the previous best estimate of β. The panel headings give the value, $\hat{\beta}$ maximising the approximating quadratic at each step. In this case no step length control is needed and the second derivative is always negative.

To compute the likelihood we need to compute the joint probability of all the data. Because our model says that the data are independent, this joint probability is just the product of the individual probabilities for each y_i, i.e. the likelihood function is just

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp(\beta x_i)^{y_i} \exp\{-\exp(\beta x_i)\}}{y_i!}.$$

and repeatedly using the facts that $\log(a^b) = b\log(a)$ and $\log(ab) = \log(a) + \log(b)$, the log likelihood is therefore

$$l(\beta) = \sum_{i=1}^{n} y_i \beta x_i - \exp(\beta x_i) - \log y_i!$$

Figure 2 shows some x, y data (left) for which this model is appropriate, and plots the log-likelihood function for a range of β values (right). $\hat{\beta} = 2.41$ is the MLE, and the continuous curve overlaid on the left panel is the Poisson mean that this implies.

The key idea then, is that the log likelihood function provides us with a generic way of measuring how well a model fits data, and which parameter values do the best job at explaining the data. The simple Poisson example only considered a single parameter, but in principle we can compute the log likelihood for a model with as many parameters as we like, although we do need to be sure that we have sufficient data if we are intending to estimate large numbers of parameters.

6.1.1 Practical likelihood maximization

It is rarely possible to find explicit expressions for the maximum likelihood estimates, and instead we must maximize the log likelihood numerically. Newton's method is the gold standard for doing this, and works my optimizing successive quadratic approximations to the log likelihood. We start with an initial parameter value guess, and try to obtain a quadratic approximation to the log-likelihood that behaves like the log-likelihood itself in the vicinity of that guess. To do this we can use a Taylor approximation to the log-likelihood, discarding the terms above second order. That is we evaluate the log likelihood and its first two derivatives[7] at the parameter value, and then find the unique quadratic function which also has this value and derivatives at the parameter value. Hopefully the approximation has a maximum close to the maximum of the log-likelihood itself, and it is easy to write down an explicit

[7] first derivative vector and second derivative matrix in usual case of a parameter vector.

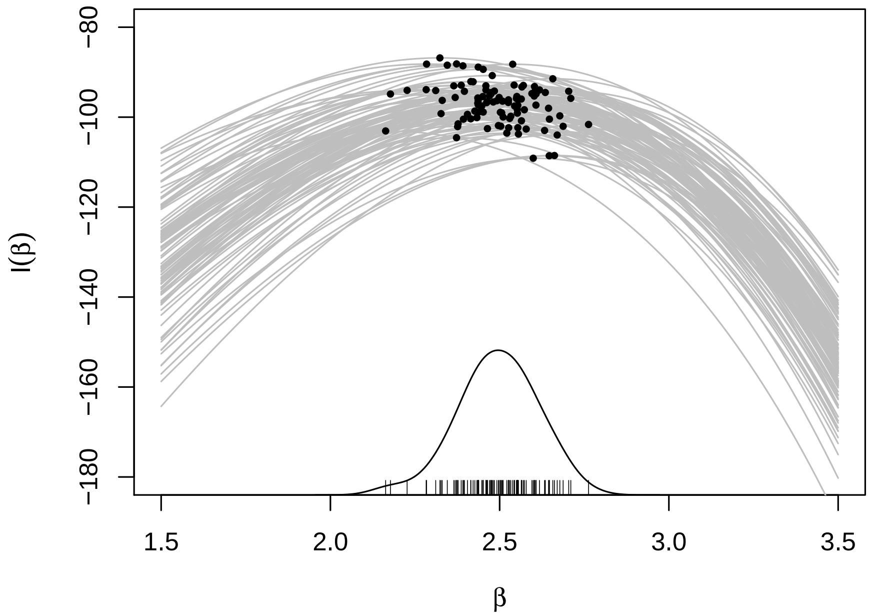

Figure 4: Log likelihood functions (grey) for 100 replicates of the section 6.1 Poisson dataset, when $\beta = 2.5$. The black dots show the maximum of each curve. The corresponding MLEs, $\hat{\beta}$, are shown as vertical bars on the x-axis (a rug plot). The continuous black curve is a (rescaled and vertically shifted) kernel density estimate of the distribution of $\hat{\beta}$ based on the shown 100 estimates. Notice how the distribution of $\hat{\beta}$ is approximately normal, with mean at the true value, $\beta = 2.5$, although there is zero probability that $\hat{\beta} = 2.5$ exactly.

form for the maximizer of the approximating quadratic, using this as our updated parameter guess. Iterating this procedure we eventually reach the maximum of the log likelihood itself.

There are two modifications required to guarantee that Newton's method converges to the MLE. Firstly we need to ensure that the approximating quadratic is one that actually has a maximum (as opposed to a minimum, or even, in more than one dimension, a saddlepoint). In one dimension this is a matter of checking that the second derivative is negative and replacing it with a negative number if it isn't. In more dimensions there is also an equivalent fix (check that the negative of the second derivative matrix is positive definite, and if necessary modify it to force this condition to hold). The second modification is to check that the proposed change in parameter values actually increases the log likelihood itself. If it doesn't we just move the parameter back towards the previous parameter guess, until the log likelihood is increased at the new values. With these modifications Newton's method will find a maximum of the log likelihood, provided it has one.

Figure 3 illustrates Newton's method for the case of the single parameter Poisson regression model. In this case the maximum is reached in 3 steps (to graphical accuracy). Notice how the final quadratic matches the log likelihood in the vicinity of the MLE.

6.1.2 Illustrating $\hat{\theta}$ uncertainty

Figure 4 shows log likelihood curves for 100 replicates of the data shown in the left panel of figure 2, when the true β is 2.5, along with the corresponding MLEs of $\hat{\beta}$, and the corresponding non-parametric estimate of the distribution of $\hat{\beta}$. Notice how the likelihood function and $\hat{\beta}$ vary from replicate to replicate, but how the distribution of $\hat{\beta}$ is approximately normal and centred on the true parameter value of 2.5. This is exactly what (8) implies will happen.

7 MLE via numerical optimization

To use MLE theory for all but the simplest practical problems requires numerical optimization of the likelihood. So if we are interested in practical statistical inference we need some idea how this works. We will look only at the Newton type methods already introduced, since these have the appeal of being based on exactly the second derivative matrix of the log likelihood that occurs in the distributional result for $\hat{\boldsymbol{\theta}}$.

7.1 Numerical optimisation

Most optimisation literature and software, including in R, concentrates on the minimisation of functions. This section follows this convention, bearing in mind that our goal of maximising log likelihoods can always be achieved by minimising negative log likelihoods. Generically, then, we are interested in automatic methods for finding

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmin}}\, f(\boldsymbol{\theta}). \tag{10}$$

There are some very difficult problems in this class, so some restrictions are needed. Specifically, assume that the *objective function*, f, is a sufficiently smooth function, bounded below, and that the elements of $\boldsymbol{\theta}$ are unrestricted real parameters. So f might be a negative log likelihood, for example. f may also depend on other known parameters and data, but there is no need to clutter up the notation with these. The assumption that $\boldsymbol{\theta}$ is unrestricted means that, if we want to put restrictions on $\boldsymbol{\theta}$, we need to be able to implement them by writing $\boldsymbol{\theta} = \mathbf{r}(\boldsymbol{\theta}_r)$, where $\mathbf{r}$ is a known function and $\boldsymbol{\theta}_r$ is a set of unrestricted parameters. Then the problem becomes $\min_{\theta_r} f\{\mathbf{r}(\boldsymbol{\theta}_r)\}$.

Even given these assumptions, it is not possible to guarantee finding a solution to (10) unless we know that f is convex, which is generally an assumption too far. Pragmatically, the best we can hope for is to develop methods to find a *local minimum*; that is, a point $\hat{\boldsymbol{\theta}}$ such that $f(\hat{\boldsymbol{\theta}} + \boldsymbol{\Delta}) \geq f(\hat{\boldsymbol{\theta}})$, for any *sufficiently small* perturbation $\boldsymbol{\Delta}$. The resulting methods are adequate for many statistical problems.

7.2 Newton's method

A very successful optimisation method is based on iteratively approximating f by a truncated Taylor expansion and seeking the minimum of the approximation at each step. With a little care, this can be made into a method that is guaranteed[8] to converge to a local minimum. Taylor's theorem states that if f is a twice continuously differentiable function of $\boldsymbol{\theta}$, and $\boldsymbol{\Delta}$ is of the same dimension as $\boldsymbol{\theta}$, then for some $t \in (0, 1)$,

$$f(\boldsymbol{\theta} + \boldsymbol{\Delta}) = f(\boldsymbol{\theta}) + \nabla f(\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{\Delta} + \frac{1}{2}\boldsymbol{\Delta}^{\mathrm{T}}\nabla^2 f(\boldsymbol{\theta} + t\boldsymbol{\Delta})\boldsymbol{\Delta} \tag{11}$$

$$\text{where } \nabla f(\boldsymbol{\theta}^*) = \left.\frac{\partial f}{\partial \boldsymbol{\theta}}\right|_{\theta^*} \text{ and } \nabla^2 f(\boldsymbol{\theta}^*) = \left.\frac{\partial^2 f}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}}\right|_{\theta^*}.$$

From (11), the condition $f(\hat{\boldsymbol{\theta}} + \boldsymbol{\Delta}) \geq f(\hat{\boldsymbol{\theta}})$, for any *sufficiently small* perturbation $\boldsymbol{\Delta}$, is equivalent to

$$\nabla f(\hat{\boldsymbol{\theta}}) = \mathbf{0} \text{ and } \nabla^2 f(\hat{\boldsymbol{\theta}}) \text{ positive semi-definite,} \tag{12}$$

which are the useful conditions for a minimum.

A second consequence of (11) is that for sufficiently small α, a $\boldsymbol{\theta}$ that is not a turning point, and *any* positive definite matrix $\mathbf{H}$ of appropriate dimension, then $f\{\boldsymbol{\theta} - \alpha\mathbf{H}\nabla f(\boldsymbol{\theta})\} < f(\boldsymbol{\theta})$. Under the given conditions we can approximate f by a first-order Taylor approximation. Hence, in the small α limit, $f\{\boldsymbol{\theta} - \alpha\mathbf{H}\nabla f(\boldsymbol{\theta})\} = f(\boldsymbol{\theta}) - \alpha\nabla f(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{H}\nabla f(\boldsymbol{\theta}) < f(\boldsymbol{\theta})$, where the inequality follows from the fact that $\nabla f(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{H}\nabla f(\boldsymbol{\theta}) > 0$ since $\mathbf{H}$ is positive definite and $\nabla f(\boldsymbol{\theta}) \neq \mathbf{0}$. In short,

$$\Delta = -\mathbf{H}\nabla f(\boldsymbol{\theta}) \tag{13}$$

is a *descent direction* if $\mathbf{H}$ is any positive definite matrix. After Taylor's theorem, this is probably the second most important fact in the optimisation of smooth functions.

[8] The statistical literature contains many statements about the possibility of Newton's method diverging, and the consequent difficulty of guaranteeing convergence. These statements are usually outdated, as a look at any decent textbook on optimisation shows.

Now consider Newton's method itself. Suppose that we have a guess, $\boldsymbol{\theta}'$, at the parameters minimising $f(\boldsymbol{\theta})$. Taylor's theorem implies that

$$f(\boldsymbol{\theta}' + \boldsymbol{\Delta}) \simeq f(\boldsymbol{\theta}') + \nabla f(\boldsymbol{\theta}')^{\mathrm{T}}\boldsymbol{\Delta} + \frac{1}{2}\boldsymbol{\Delta}^{\mathrm{T}}\nabla^2 f(\boldsymbol{\theta}')\boldsymbol{\Delta}.$$

Provided that $\nabla^2 f(\boldsymbol{\theta}')$ is positive semi-definite, the right-hand-side of this expression can be minimised by differentiating with respect to $\boldsymbol{\Delta}$ and setting the result to zero, which implies that

$$\nabla^2 f(\boldsymbol{\theta}')\boldsymbol{\Delta} = -\nabla f(\boldsymbol{\theta}'). \tag{14}$$

So, in principle, we simply solve for $\boldsymbol{\Delta}$ given $\boldsymbol{\theta}'$ and update $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta}' + \boldsymbol{\Delta}$ repeatedly until the conditions (12) are met. By Taylor's theorem itself, this process must converge if we start out close enough to the minimising $\hat{\boldsymbol{\theta}}$. But if we knew how to do that we perhaps would not need to be using Newton's method in the first place.

The method should converge when started from parameter guesses that are a long way from $\hat{\boldsymbol{\theta}}$, requiring two modifications of the basic iteration:

1. $\nabla^2 f(\boldsymbol{\theta}')$ is only guaranteed to be positive (semi) definite close to $\hat{\boldsymbol{\theta}}$. So $\nabla^2 f(\boldsymbol{\theta}')$ must be modified to make it positive definite, if it is not. The obvious alternatives are (i) to replace $\nabla^2 f(\boldsymbol{\theta}')$ by $\nabla^2 f(\boldsymbol{\theta}') + \delta\mathbf{I}$, where δ is chosen to be just large enough to achieve positive definiteness;[9] or (ii) take the symmetric eigen-decomposition $\nabla^2 f(\boldsymbol{\theta}') = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$, where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues, and replace $\nabla^2 f(\boldsymbol{\theta}')$ by $\mathbf{U}\tilde{\boldsymbol{\Lambda}}\mathbf{U}^{\mathrm{T}}$, where $\tilde{\boldsymbol{\Lambda}}$ is $\boldsymbol{\Lambda}$ with all nonpositive eigenvalues replaced by positive entries (e.g. $|\Lambda_{ii}|$). Using the perturbed version in (14), results in a step of the form (13), so if we are not at a turning point, then a sufficiently small step in the direction $\boldsymbol{\Delta}$ is guaranteed to reduce f.

2. The second-order Taylor approximation about a point far from $\hat{\boldsymbol{\theta}}$ could be poor at $\hat{\boldsymbol{\theta}}$, so that there is no guarantee that stepping to its minimum will lead to a reduction in f. However, given the previous modification, we know that the Newton step is a descent direction. A small enough step in direction $\boldsymbol{\Delta}$ must reduce f. Therefore, if $f(\boldsymbol{\theta}' + \boldsymbol{\Delta}) > f(\boldsymbol{\theta}')$, repeatedly set $\boldsymbol{\Delta} \leftarrow \boldsymbol{\Delta}/2$, until a reduction in f is achieved.

With these two modifications each step of Newton's method must reduce f until a turning point is reached.
In summary, starting with $k = 0$ and a guesstimate $\boldsymbol{\theta}^{[0]}$, iterate these steps:

1. Evaluate $f(\boldsymbol{\theta}^{[k]})$, $\nabla f(\boldsymbol{\theta}^{[k]})$ and $\nabla^2 f(\boldsymbol{\theta}^{[k]})$.

2. Test whether $\boldsymbol{\theta}^{[k]}$ is a minimum using (12), and terminate if it is.[10]

3. If $\mathbf{H} = \nabla^2 f(\boldsymbol{\theta}^{[k]})$ is not positive definite, perturb it so that it is.

4. Solve $\mathbf{H}\boldsymbol{\Delta} = -\nabla f(\boldsymbol{\theta}^{[k]})$ for the search direction $\boldsymbol{\Delta}$.

5. If $f(\boldsymbol{\theta}^{[k]} + \boldsymbol{\Delta})$ is not $< f(\boldsymbol{\theta}^{[k]})$, repeatedly halve $\boldsymbol{\Delta}$ until it is.

6. Set $\boldsymbol{\theta}^{[k+1]} = \boldsymbol{\theta}^{[k]} + \boldsymbol{\Delta}$, increment k by one and return to step 1.

In practice ∇f is not tested for exact equality to zero at 2, and we instead test whether $\|\nabla f(\boldsymbol{\theta}^{[k]})\| < |f(\boldsymbol{\theta}^{[k]})|\epsilon_r + \epsilon_a$, for small constants ϵ_r and ϵ_a.

7.2.1 Newton's method examples

As a single-parameter example, consider an experiment on antibiotic efficacy. A 1-litre culture of 5×10^5 cells is set up and dosed with antibiotic. After 2 hours, and then every subsequent hour up to 14 hours after dosing, 0.1ml

[9] Positive definiteness can be tested by attempting a Choleski decomposition of the matrix concerned: it will succeed if the matrix is positive definite and fail otherwise. Use a pivoted Choleski decomposition to test for positive semi-definiteness. Alternatively, simply examine the eigenvalues returned by any symmetric eigen routine.

[10] If the objective function contains a saddlepoint, then theoretically Newton's method might find it, in which case the gradient would be zero and the Hessian indefinite: in this rare case further progress can only be made by perturbing the Newton step directly.

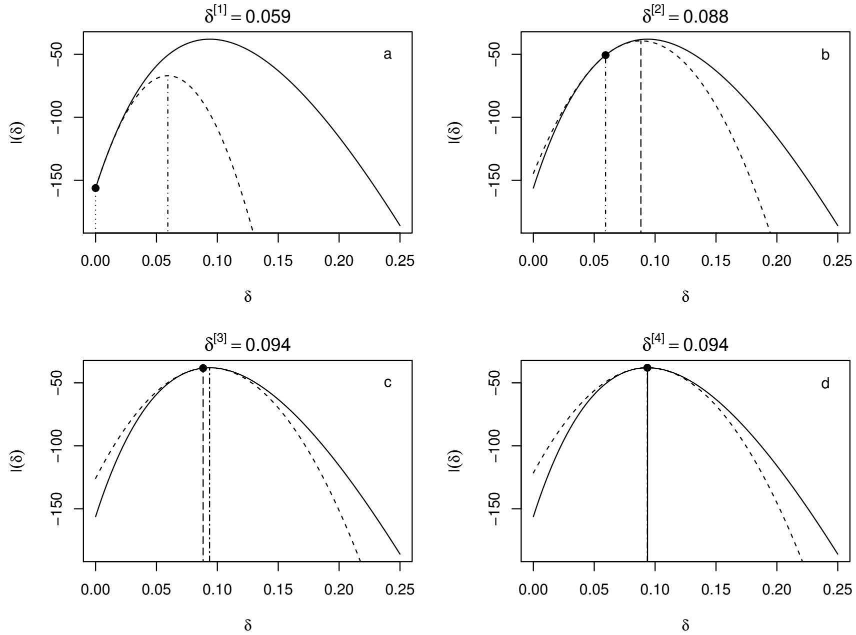

Figure 5: Newton's method for the antibiotic example of Section 7.2. Each panel shows one step of Newton's method, with the log likelihood (black) and the second-order Taylor approximation about ● (dashed). Vertical lines show the estimated value of δ at the start and end of the step. Each panel title gives the end estimate. **a** starts from $\delta^{[0]} = 0$. **b** to **d** show subsequent iterations until convergence.

of the culture is removed and the live bacteria in this sample counted under a microscope, giving counts, y_i, and times, t_i (hours). The data are

t_i	2	3	4	5	6	7	8	9	10	11	12	13	14
y_i	35	33	33	39	24	25	18	20	23	13	14	20	18

A simple model for the sample counts, y_i, is that their expected value is $E(Y_i) = \mu_i = 50e^{-\delta t_i}$, where δ is an unknown 'death rate' parameter (per hour) and t_i is the sample time in hours. Given the sampling protocol, it is reasonable to assume that the counts are observations of independent Poi(μ_i) random variables, with probability function $f(y_i) = \mu_i^{y_i} e^{-\mu_i}/y_i!$ So the log likelihood is

$$l(\delta) = \sum_{i=1}^{n} \{y_i \log(\mu_i) - \mu_i - \log(y_i!)\} = \sum_{i=1}^{n} y_i\{\log(50) - \delta t_i\} - \sum_{i=1}^{n} 50e^{-\delta t_i} - \sum_{i=1}^{n} \log(y_i!),$$

where $n = 13$. Differentiating w.r.t. δ,

$$\frac{\partial l}{\partial \delta} = -\sum_{i=1}^{n} y_i t_i + \sum_{i=1}^{n} 50 t_i e^{-\delta t_i} \text{ and } \frac{\partial^2 l}{\partial \delta^2} = -50 \sum_{i=1}^{n} t_i^2 e^{-\delta t_i}.$$

The presence of the t_i term in $e^{-\delta t_i}$ precludes a closed-form solution for $\partial l / \partial \delta = 0$, and Newton's method can be applied instead. Figure 5 illustrates the method's progression. In this case the second derivatives do not require perturbation, and no step-length halving is needed.

Now consider an example with a vector parameter. The following data are reported AIDS cases in Belgium, in the early stages of the epidemic.

Year (19–)	81	82	83	84	85	86	87	88	89	90	91	92	93
Cases	12	14	33	50	67	74	123	141	165	204	253	246	240

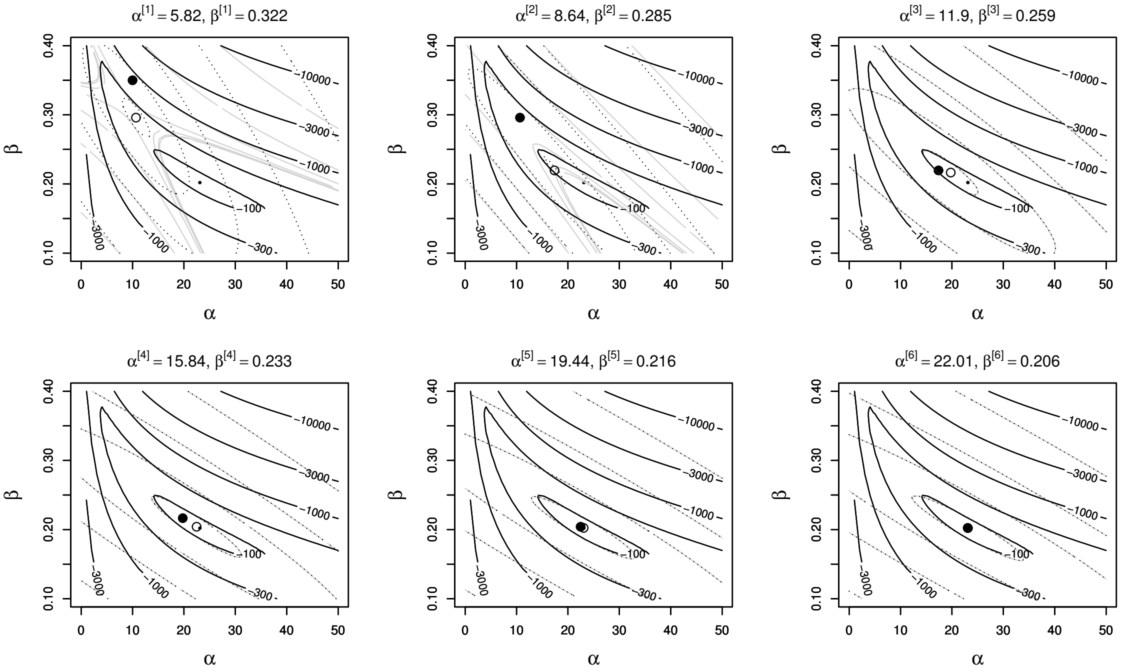

Figure 6: Newton's method for the AIDS example of Section 7.2. Each panel shows one step, with the log likelihood contoured in black, the second order Taylor approximation about • in grey, and the quadratic given by the positive definite corrected Hessian as dotted. The panel caption and ◦ give the Newton method proposed parameter values at the end of each step. The iteration starts at the top left and has converged by the lower right.

One important question, early in such epidemics, is whether control measures are beginning to have an impact or whether the disease is continuing to spread essentially unchecked. A simple model for unchecked growth leads to an 'exponential increase' model. The model says that the number of cases, y_i, is an observation of an independent Poisson r.v., with expected value $\mu_i = \alpha e^{\beta t_i}$ where t_i is the number of years since 1980. So the log likelihood is

$$l(\alpha, \beta) = \sum_{i=1}^{n} y_i \{\log(\alpha) + \beta t_i\} - \sum_{i=1}^{n} (\alpha e^{\beta t_i} - y_i!),$$

and hence,

$$\nabla l = \left[\begin{array}{c} \sum y_i/\alpha - \sum \exp(\beta t_i) \\ \sum y_i t_i - \alpha \sum t_i \exp(\beta t_i) \end{array} \right] \text{ and } \nabla^2 l = \left[\begin{array}{cc} -\sum y_i/\alpha^2 & -\sum t_i e^{\beta t_i} \\ -\sum t_i e^{\beta t_i} & -\alpha \sum t_i^2 e^{\beta t_i}. \end{array} \right].$$

A swift glance at the expression for the gradients should be enough to convince you that numerical methods will be required to find the MLEs of the parameters. Starting from an initial guess $\alpha^{[0]} = 4$, $\beta^{[0]} = .35$, here is the first Newton iteration:

$$\left[\begin{array}{c} \alpha^{[0]} \\ \beta^{[0]} \end{array} \right] = \left[\begin{array}{c} 4 \\ .35 \end{array} \right] \Rightarrow \nabla l = \left[\begin{array}{c} 88.4372 \\ 1850.02 \end{array} \right], \nabla^2 l = \left[\begin{array}{cc} -101.375 & -3409.25 \\ -3409.25 & -154567 \end{array} \right]$$

$$\Rightarrow (\nabla^2 l)^{-1} \nabla l = \left[\begin{array}{c} -1.820 \\ 0.028 \end{array} \right] \Rightarrow \left[\begin{array}{c} \alpha^{[1]} \\ \beta^{[1]} \end{array} \right] = \left[\begin{array}{c} \alpha^{[0]} \\ \beta^{[0]} \end{array} \right] - (\nabla^2 l)^{-1} \nabla l = \left[\begin{array}{c} 5.82 \\ 0.322 \end{array} \right].$$

After eight more steps the likelihood is maximised at $\hat{\alpha} = 23.1$, $\hat{\beta} = 0.202$. Figure 6 illustrates six Newton steps, starting from the more interesting point $\hat{\alpha}_0 = 4$, $\hat{\beta}_0 = 0.35$. Perturbation to positive definiteness is required in the first two steps, but the method converges in six steps.

7.2.2 Newton variations: avoiding f evaluation and the expected Hessian

Occasionally we have access to ∇f and $\nabla^2 f$, but f itself is either unavailable or difficult to compute in a stable way. Newton's method only requires evaluation of f in order to check that the Newton step has led to a reduction in f. It is usually sufficient to replace the condition $f(\boldsymbol{\theta}+\boldsymbol{\Delta}) \le f(\boldsymbol{\theta})$ with the condition that f must be non-increasing in the direction $\boldsymbol{\Delta}$ at $\boldsymbol{\theta}'+\boldsymbol{\Delta}$. That is, $\nabla f(\boldsymbol{\theta}'+\boldsymbol{\Delta})^{\mathrm{T}}\boldsymbol{\Delta} \le 0$. In many circumstances, step-length control based on this condition ensures convergence in cases where the iteration would otherwise have diverged, but unlike the function-value based control, pathological cases can easily be dreamt up to defeat it. Such step-length reduction should only be applied after testing that the step has not already met the convergence criteria.

Another common variation on the method, used in maximum likelihood estimation, is to replace $-\nabla^2 l(\boldsymbol{\theta})$ by $-E\{\nabla^2 l(\boldsymbol{\theta})\}$ (so-called *Fisher scoring*). Because the replacement is always positive (semi-)definite, perturbation to positive definiteness is not required and by the arguments surrounding (13) the method converges when used with simple step-length control.

7.3 Other optimization methods

Optimization is a huge subject, but two alternative methods are worth a brief mention here. Quasi-Newton methods are Newton type methods in which an approximation to the Hessian matrix, or its inverse, is built up iteratively from the first derivative information computed at each trial set of parameter values. Not requiring the computation of second derivative information makes implementation simpler and sometimes more efficient, at the cost of requiring somewhat more effort with step length control than is required for Newton's method. In R quasi-Newton is available using `optim(...,method="BFGS")`.

A word of warning. Truncating the Taylor expansion of the objective at first order leads to the 'obvious' first derivative based method, of steepest descent, where the trial step is set to a multiple of the negative of the gradient of the objective w.r.t. $\boldsymbol{\theta}$. Step length control is always needed in this case, as the first order Taylor expansion tells you nothing about how far to step. Worse than that, the first order term in the Taylor expansion *always* becomes smaller than the higher order terms as we approach $\hat{\boldsymbol{\theta}}$, so that it is a useless approximation to the objective function at exactly the point we are most interested in. The result is that steepest descent is often absurdly slow to converge, and is rarely worth using.

There are also methods that optimize based only on function values. Of these, the most useful in a likelihood context is the 'Nelder Meade simplex (or polytope) method'. We will not cover how this works here, but it the default method for the R `optim` function.

8 MLE theory in more detail

Having resolved the issue of how to actually compute MLEs, it is time to consider where the large sample inferential results come from. This section briefly covers the derivation of the key results at a level suitable for ensuring their reliable application. See Cox and Hinkley (1974), Silvey (1970) and Davison (2003) for more detail.

8.1 Some properties of the expected log likelihood

Large sample theory for maximum likelihood estimators relies on some results for the expected log likelihood and on the observed likelihood tending to its expected value as the sample size tends to infinity. The results for the expected log likelihood are derived here. Recall that $l(\boldsymbol{\theta}) = \log f_\theta(\mathbf{y})$, and let $\boldsymbol{\theta}_t$ be the vector of true parameter values.

1.
$$E\left(\left.\frac{\partial l}{\partial \boldsymbol{\theta}}\right|_{\theta_t}\right) = \mathbf{0}, \tag{15}$$

where the expectation is taken at $\boldsymbol{\theta}_t$. Proof is straightforward provided that there is sufficient regularity to allow the order of differentiation and integration to be exchanged:

$$E\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f_\theta(\mathbf{y})\right\} = \int \frac{1}{f_\theta(\mathbf{y})}\frac{\partial f_\theta}{\partial \boldsymbol{\theta}} f_\theta(\mathbf{y})d\mathbf{y} = \int \frac{\partial f_\theta}{\partial\boldsymbol{\theta}}d\mathbf{y} = \frac{\partial}{\partial\boldsymbol{\theta}}\int f_\theta(\mathbf{y})d\mathbf{y} = \frac{\partial \mathbf{1}}{\partial\boldsymbol{\theta}} = \mathbf{0}.$$

2.

$$\text{cov}\left(\left.\frac{\partial l}{\partial \boldsymbol{\theta}}\right|_{\theta_t}\right) = E\left(\left.\frac{\partial l}{\partial \boldsymbol{\theta}}\right|_{\theta_t}\left.\frac{\partial l}{\partial \boldsymbol{\theta}^{\rm T}}\right|_{\theta_t}\right), \tag{16}$$

which follows directly from the previous result and the definition of a covariance matrix. Recall here that $\partial l/\partial\boldsymbol{\theta}$ is a column vector and $\partial l/\partial\boldsymbol{\theta}^{\rm T}$ a row vector.

3.

$$\boldsymbol{\mathcal{I}} = E\left(\left.\frac{\partial l}{\partial \boldsymbol{\theta}}\right|_{\theta_t}\left.\frac{\partial l}{\partial \boldsymbol{\theta}^{\rm T}}\right|_{\theta_t}\right) = -E\left(\left.\frac{\partial^2 l}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\rm T}}\right|_{\theta_t}\right). \tag{17}$$

$\boldsymbol{\mathcal{I}}$, is known as the *Fisher information matrix*. The terminology relates to the fact that a likelihood containing lots of information about $\boldsymbol{\theta}$ will be sharply peaked ($\boldsymbol{\mathcal{I}}$ will have large magnitude eigenvalues), whereas a less informative likelihood will be less sharply peaked.

Proof is straightforward. From (15) we have

$$\int \frac{\partial \log f_\theta}{\partial \boldsymbol{\theta}} f_\theta(\mathbf{y})d\mathbf{y} = \mathbf{0} \Rightarrow \int \frac{\partial^2 \log f_\theta}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\rm T}} f_\theta(\mathbf{y}) + \frac{\partial \log f_\theta}{\partial \boldsymbol{\theta}}\frac{\partial f_\theta}{\partial \boldsymbol{\theta}^{\rm T}} d\mathbf{y} = \mathbf{0},$$

$$\text{but } \frac{\partial \log f_\theta}{\partial \boldsymbol{\theta}^{\rm T}} = \frac{1}{f_\theta}\frac{\partial f_\theta}{\partial \boldsymbol{\theta}^{\rm T}}, \text{ so } \int \frac{\partial^2 \log f_\theta}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\rm T}} f_\theta(\mathbf{y})d\mathbf{y} = -\int \frac{\partial \log f_\theta}{\partial \boldsymbol{\theta}}\frac{\partial \log f_\theta}{\partial \boldsymbol{\theta}^{\rm T}} f_\theta(\mathbf{y})d\mathbf{y},$$

and the result is proven.

4. The expected log likelihood has a global maximum at $\boldsymbol{\theta}_t$. i.e.

$$E\{l(\boldsymbol{\theta}_t)\} \geq E\{l(\boldsymbol{\theta})\}\ \forall\ \boldsymbol{\theta}. \tag{18}$$

Since log is a concave function, Jensen's inequality[11] implies that

$$E\left[\log\left\{\frac{f_\theta(\mathbf{y})}{f_{\theta_t}(\mathbf{y})}\right\}\right] \leq \log\left[E\left\{\frac{f_\theta(\mathbf{y})}{f_{\theta_t}(\mathbf{y})}\right\}\right] = \log\int\frac{f_\theta(\mathbf{y})}{f_{\theta_t}(\mathbf{y})}f_{\theta_t}(\mathbf{y})d\mathbf{y} = \log\int f_\theta(\mathbf{y})d\mathbf{y} = \log(1) = 0,$$

and the result is proven.

8.2 The Cramér-Rao lower bound

In the context of linear models, the discussion of the properties of a good estimator came to the conclusion that minimum variance unbiased estimation was a reasonable goal to aim for. The Gauss-Markov theorem then provided the justification for least squares estimation of linear models. But in general how do we know when we have achieved minimum variance? The following result is part of the answer.

The Cramér-Rao lower bound. $\boldsymbol{\mathcal{I}}^{-1}$ provides a lower bound on the variance matrix of any unbiased estimator $\tilde{\boldsymbol{\theta}}$, in the sense that $\text{cov}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{\mathcal{I}}^{-1}$ is positive semi-definite.

Proof: Since $f\partial \log f/\partial\theta = \partial f/\partial\theta$, $\partial\boldsymbol{\theta}_t/\partial\boldsymbol{\theta}_t^{\rm T} = \mathbf{I}$ and $\tilde{\boldsymbol{\theta}}$ is unbiased,

$$\int \tilde{\boldsymbol{\theta}} f_{\theta_t}(\mathbf{y})d\mathbf{y} = \boldsymbol{\theta}_t \Rightarrow \int \tilde{\boldsymbol{\theta}}\left.\frac{\partial \log f_{\theta_t}}{\partial \boldsymbol{\theta}_t^{\rm T}}\right|_{\theta_t} f_{\theta_t}(\mathbf{y})d\mathbf{y} = \mathbf{I}.$$

Hence, given (15), the matrix of covariances of elements of $\tilde{\boldsymbol{\theta}}_t$ with elements of $\partial\log f_{\theta_t}/\partial\boldsymbol{\theta}_t$ can be obtained:

$$\text{cov}\left(\tilde{\boldsymbol{\theta}}, \left.\frac{\partial \log f_{\theta_t}}{\partial \boldsymbol{\theta}_t}\right|_{\theta_t}\right) = E\left(\tilde{\boldsymbol{\theta}}\left.\frac{\partial \log f_{\theta_t}}{\partial \boldsymbol{\theta}_t^{\rm T}}\right|_{\theta_t}\right) - E(\tilde{\boldsymbol{\theta}})E\left(\left.\frac{\partial \log f_{\theta_t}}{\partial \boldsymbol{\theta}_t^{\rm T}}\right|_{\theta_t}\right) = \mathbf{I}.$$

[11] For any random variable X and concave function c, $c\{E(X)\} \geq E\{c(X)\}$.

Combining this with (16) we obtain the variance-covariance matrix,

$$\text{cov}\left[\begin{array}{c}\tilde{\boldsymbol{\theta}}\\ \frac{\partial \log f_{\theta_t}}{\partial \boldsymbol{\theta}_t}\Big|_{\theta_t}\end{array}\right] = \left[\begin{array}{cc}\text{cov}(\tilde{\boldsymbol{\theta}}) & \mathbf{I}\\ \mathbf{I} & \boldsymbol{\mathcal{I}}\end{array}\right],$$

which is positive semi-definite by virtue of being a variance-covariance matrix. It follows that

$$\left[\begin{array}{cc}\mathbf{I} & -\boldsymbol{\mathcal{I}}^{-1}\end{array}\right]\left[\begin{array}{cc}\text{cov}(\tilde{\boldsymbol{\theta}}) & \mathbf{I}\\ \mathbf{I} & \boldsymbol{\mathcal{I}}\end{array}\right]\left[\begin{array}{c}\mathbf{I}\\ -\boldsymbol{\mathcal{I}}^{-1}\end{array}\right] = \text{cov}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{\mathcal{I}}^{-1}$$

is positive semi-definite, and the result is proven.

If the sense in which $\boldsymbol{\mathcal{I}}^{-1}$ is a lower bound is unclear, consider the variance of any linear transformation of the form $\mathbf{a}^{\text{T}}\tilde{\boldsymbol{\theta}}$. By the result just proven, and the definition of positive semi-definiteness,

$$0 \le \mathbf{a}^{\text{T}}\{\text{cov}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{\mathcal{I}}^{-1}\}\mathbf{a} = \text{var}(\mathbf{a}^{\text{T}}\tilde{\boldsymbol{\theta}}) - \mathbf{a}^{\text{T}}\boldsymbol{\mathcal{I}}^{-1}\mathbf{a},$$

$\Rightarrow$ $\text{var}(\mathbf{a}^{\text{T}}\tilde{\boldsymbol{\theta}}) \ge \mathbf{a}^{\text{T}}\boldsymbol{\mathcal{I}}^{-1}\mathbf{a}$. For example, the lower bound on $\text{var}(\tilde{\theta}_i)$ is given by the i^{th} element on the leading diagonal of $\boldsymbol{\mathcal{I}}^{-1}$.

The importance of this result in the current context is that, in the large sample limit subject to regularity conditions, maximum likelihood estimators are unbiased and achieve the CR bound.

8.3 Consistency of MLE

Maximum likelihood estimators are usually consistent, meaning that as the sample size tends to infinity, $\hat{\boldsymbol{\theta}}$ tends to $\boldsymbol{\theta}_t$ (provided that the likelihood is informative about the parameters). This occurs because in regular situations $l(\boldsymbol{\theta})/n \to E\{l(\boldsymbol{\theta})\}/n$ as the sample size, n, tends to infinity, so that eventually the maximum of $l(\boldsymbol{\theta})$ and $E\{l(\boldsymbol{\theta})\}$ must coincide at $\boldsymbol{\theta}_t$ by (18). The result is easy to prove if the log likelihood can be broken down into a sum of independent components (usually one per observation), so that the law of large numbers implies convergence of the log likelihood to its expectation. Consistency can fail when the number of parameters is growing alongside the sample size in such a way that, for at least some parameters, the information per parameter is not increasing with sample size.

8.4 Large sample distribution of MLE

Taylor's theorem implies that

$$\frac{\partial l}{\partial \boldsymbol{\theta}}\Big|_{\hat{\theta}} \simeq \frac{\partial l}{\partial \boldsymbol{\theta}}\Big|_{\theta_t} + \frac{\partial^2 l}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\text{T}}}\Big|_{\theta_t}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t)$$

with equality in the large sample limit, for which $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t \to 0$. From the definition of $\hat{\boldsymbol{\theta}}$, the left-hand side is $\mathbf{0}$. So assuming $\boldsymbol{\mathcal{I}}/n$ is constant (at least in the $n \to \infty$ limit), then as the sample size tends to infinity,

$$\frac{1}{n}\frac{\partial^2 l}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\text{T}}}\Big|_{\theta_t} \to -\frac{\boldsymbol{\mathcal{I}}}{n}, \text{ while } \frac{\partial l}{\partial \boldsymbol{\theta}}\Big|_{\theta_t}$$

is a random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\mathcal{I}}$ by (16) and (15).[12] Therefore in the large sample limit,

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t \sim \boldsymbol{\mathcal{I}}^{-1}\frac{\partial l}{\partial \boldsymbol{\theta}}\Big|_{\theta_t},$$

implying that $E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t) = 0$ and $\text{var}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t) = \boldsymbol{\mathcal{I}}^{-1}$. Hence in regular situations in the large sample limit, maximum likelihood estimators are unbiased and achieve the Cramér-Rao lower bound. This partly accounts for their popularity.

[12] In the limit the random deviation of $n^{-1}\partial^2 l/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\text{T}}$ from its expected value, $n^{-1}\boldsymbol{\mathcal{I}}$, is negligible relative to $n^{-1}\boldsymbol{\mathcal{I}}$ itself (provided it is positive definite). This is never the case for the $\partial l/\partial\boldsymbol{\theta}$, because its expected value is zero.

It remains to establish the large sample distribution of $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_t$. In the case in which the likelihood is based on independent observations, then $l(\boldsymbol{\theta}) = \sum_i l_i(\boldsymbol{\theta})$, where l_i denotes the contribution to the log likelihood from the i^{th} observation. In that case $\partial l/\partial\boldsymbol{\theta} = \sum_i \partial l_i/\partial\boldsymbol{\theta}$, so that $\partial l/\partial\boldsymbol{\theta}$ is a sum of independent random variables. Hence, under mild conditions, the central limit theorem applies, and in the large sample limit

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_t, \boldsymbol{\mathcal{I}}^{-1}). \tag{19}$$

In any circumstance in which (19) holds, it is also valid to use $-\partial^2 l/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}$ in place of $\boldsymbol{\mathcal{I}}$ itself. When the likelihood is not based on independent observations, it is very often the case that $\partial l/\partial\boldsymbol{\theta}$ has a limiting normal distribution so that (19) holds anyway. The key is that the expected information increases without limit with increasing sample size. In any case, achievement of the Cramér-Rao lower bound does not depend on normality.

8.5 Hypothesis testing

When introducing inference in the context of the linear model, we derived hypothesis testing procedures based on p-values, without much consideration of the basis of the testing process. Now that we are considering statistical inference in greater generality, it is time to examine the process of hypothesis testing in more generality. The basic question of interest is whether some defined restriction on $\boldsymbol{\theta}$ is consistent with $\mathbf{y}$?

8.5.1 p-values: the fundamental idea

Suppose that we have a model defining a p.d.f., $f_\theta(\mathbf{y})$, for data vector $\mathbf{y}$ and that we want to test the *null hypothesis*, $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is some specified value. That is, we want to establish whether the data could reasonably be generated from $f_{\theta_0}(\mathbf{y})$. An obvious approach is to ask how probable data like $\mathbf{y}$ are under H_0. It is tempting to simply evaluate $f_{\theta_0}(\mathbf{y})$ for the observed $\mathbf{y}$, but then deciding what counts as 'probable' and 'improbable' is difficult to do in a generally applicable way.

A better approach is to assess the probability, p_0 say, of obtaining data at least as improbable as $\mathbf{y}$ under H_0 (better read that sentence twice). For example, if only one dataset in a million would be as improbable as $\mathbf{y}$, according to H_0, then assuming we believe our data, we ought to seriously doubt H_0. Conversely, if half of all datasets would be expected to be at least as improbable as $\mathbf{y}$, according to H_0, then there is no reason to doubt it.

A quantity like p_0 makes good sense in the context of goodness of fit testing, where we simply want to assess the plausibility of f_{θ_0} as a model without viewing it as being a restricted form of a larger model. But when we are really testing $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against the alternative H_1 : '$\boldsymbol{\theta}$ unrestricted' then p_0 is not satisfactory, because it makes no distinction between $\mathbf{y}$ being improbable under H_0 but probable under H_1, and $\mathbf{y}$ being improbable under both.

A very simple example illustrates the problem. Consider independent observations y_1, y_2 from $N(\mu, 1)$, and the test $\mathrm{H}_0 : \mu = 0$ versus $\mathrm{H}_0 : \mu \neq 0$. Figure 7 shows the p.d.f. under the null, and, in grey, the region over which the p.d.f. has to be integrated to find p_0 for the data point marked by •. Now consider two alternative values for y_1, y_2 that yield equal $p_0 = 0.1$. In one case (black triangle) $y_1 = -y_2$, so that the best estimate of μ is 0, corresponding exactly to H_0. In the other case (black circle) the data are much more probable under the alternative than under the null hypothesis. So because we include points that are more compatible with the null than with the alternative in the calculation of p_0, we have only weak discriminatory power between the hypotheses.

Recognizing the problems with p_0, a possible solution is to standardise $f_{\theta_0}(\mathbf{y})$ by the highest value that $f_\theta(\mathbf{y})$ could have taken for the given $\mathbf{y}$. That is, to judge the *relative plausibility* of $\mathbf{y}$ under H_0 on the basis of $f_{\theta_0}(\mathbf{y})/f_{\hat{\theta}}(\mathbf{y})$ where $\hat{\boldsymbol{\theta}}$ is the value maximising $f_\theta(\mathbf{y})$ for a given $\mathbf{y}$. In the context of the example in Figure 7 this approach is much better. The black triangle now has relative plausibility 1, reflecting its compatibility with the H_0, whereas the black circle has much lower plausibility, reflecting the fact that it would be much more probable under a model with a mean greater than zero. So we could now seek a revised measure of consistency of the data and null hypothesis:

> p is the probability, under the null hypothesis, of obtaining data at least as relatively implausible as that observed.

Actually the reciprocal of this relative plausibility is generally known as the *likelihood ratio* $f_{\hat{\theta}}(\mathbf{y})/f_{\theta_0}(\mathbf{y})$ of the two hypotheses, because it is a measure of how likely the alternative hypothesis is relative to the null, given the data. So we have the more usual equivalent definition:

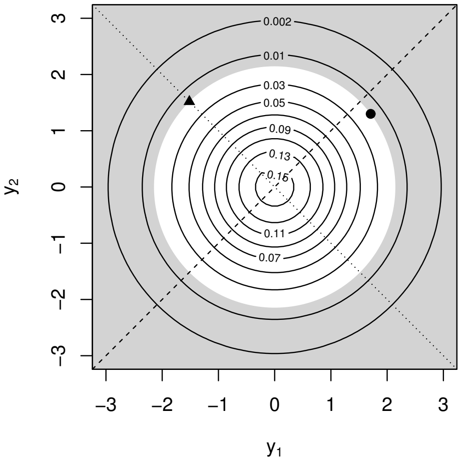

Figure 7: Why the alternative hypothesis is needed in defining the p-value, and p_0 will not do. Contour plot of the joint p.d.f. for the null model that y_1, y_2 are independent $N(0,1)$. The dashed line illustrates the possible values for the expected values of y_1 and y_2, under the alternative model that they are independent $N(\mu,1)$. The black triangle and black circle show two possible values for y_1, y_2, while the grey region shows the region of at least as improbable y_1, y_2 pairs, corresponding to $p_0 = 0.1$. The problem is that although the black circle is much more probable under the alternative model, it has the same p_0 value as the black triangle, for which $y_1 = -y_2$ and the estimated μ would be exactly the null model value of zero. The dotted line is $y_1 = -y_2$.

> p is the probability, under the null hypothesis, of obtaining a likelihood ratio at least as large as that observed.

p is generally referred to as the *p-value* associated with a test. If the null hypothesis is true, then from its definition, the p-value should have a uniform distribution on $[0,1]$ (assuming its distribution is continuous). By convention p-values in the ranges $0.1 \geq p > 0.05$, $0.05 \geq p > 0.01$, $0.01 \geq p > 0.001$ and $p \leq 0.001$ are sometimes described as providing, respectively, 'marginal evidence', 'evidence', 'strong evidence' and 'very strong evidence' against the null model, although the interpretation should really be sensitive to the context.

8.5.2 Generalisations

For the purposes of motivating p-values, the previous subsection considered only the case where the null hypothesis is a *simple* hypothesis, specifying a value for every parameter of f, while the alternative is a *composite* hypothesis, in which a range of parameter values are consistent with the alternative. Unsurprisingly, there are many situations in which we are interested in comparing two composite hypotheses, so that H_0 specifies some restrictions of $\boldsymbol{\theta}$, without fully constraining it to one point. Less commonly, we may also wish to compare two simple hypotheses, so that the alternative also supplies one value for each element of $\boldsymbol{\theta}$. This latter case is of theoretical interest, but because the hypotheses are not nested it is somewhat conceptually different from most cases of interest.

All test variants can be dealt with by a slight generalisation of the likelihood ratio statistic to $f_{\hat{\theta}}(\mathbf{y})/f_{\hat{\theta}_0}(\mathbf{y})$ where $f_{\hat{\theta}_0}(\mathbf{y})$ now denotes the maximum possible value for the density of $\mathbf{y}$ under the null hypothesis. If the null hypothesis is simple, then this is just $f_{\theta_0}(\mathbf{y})$, as before, but if not then it is obtained by finding the parameter vector that maximises $f_\theta(\mathbf{y})$ subject to the restrictions on $\boldsymbol{\theta}$ imposed by H_0.

In some cases the p-value can be calculated exactly from its definition, and the relevant likelihood ratio. When this is not possible the large sample result (9) from section 6 can be employed to compute an approximate p-value.

$f_{\hat{\theta}}(\mathbf{y})/f_{\hat{\theta}_0}(\mathbf{y})$ is an example of a *test statistic*, which takes low values when the H_0 is true, and higher values when H_1 is true. Other test statistics can be devised in which case the definition of the p-value generalises to:

p is the probability of obtaining a test statistic at least as favourable to H_1 as that observed, if H_0 is true.

This generalisation immediately raises the question: what makes a good test statistic? The answer is that we would like the resulting p-values to be as small as possible when the null hypothesis is not true (for a test statistic with a continuous distribution, the p-values should have a $U(0,1)$ distribution when the null is true). That is, we would like the test statistic to have *high power* to discriminate between null and alternative hypotheses.

8.5.3 The Neyman-Pearson lemma

The Neyman-Pearson lemma provides some support for using the likelihood ratio as a test statistic, in that it shows that doing so provides the best chance of rejecting a false null hypothesis, albeit in the restricted context of a simple null versus a simple alternative. The result can be seen as formalizing some of the considerations that originally led us to the likelihood ratio statistic and associated p-value.

Formally, consider testing $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $\mathrm{H}_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$. Suppose that we decide to reject H_0 if the p-value is less than or equal to some value α. Let $\beta(\boldsymbol{\theta})$ be the probability of rejection if the true parameter value is $\boldsymbol{\theta}$ — the test's *power*.

In this accept/reject setup the likelihood ratio test rejects H_0 if $\mathbf{y} \in R = \{\mathbf{y} : f_{\theta_1}(\mathbf{y})/f_{\theta_0}(\mathbf{y}) > k\}$ and k is such that $\Pr_{\theta_0}(\mathbf{y} \in R) = \alpha$. It is useful to define the function $\phi(\mathbf{y}) = 1$ if $\mathbf{y} \in R$ and 0 otherwise. Then $\beta(\boldsymbol{\theta}) = \int \phi(\mathbf{y}) f_\theta(\mathbf{y}) d\mathbf{y}$. Note that $\beta(\boldsymbol{\theta}_0) = \alpha$.

Now consider using an alternative test statistic and again rejecting if the p-value is $\leq \alpha$. Suppose that the test procedure rejects if

$$\mathbf{y} \in R^* \text{ where } \Pr_{\theta_0}(\mathbf{y} \in R^*) \leq \alpha.$$

Let $\phi^*(\mathbf{y})$ and $\beta^*(\boldsymbol{\theta})$ be the equivalent of $\phi(\mathbf{y})$ and $\beta(\boldsymbol{\theta})$ for this test. Here $\beta^*(\boldsymbol{\theta}_0) = \Pr_{\theta_0}(\mathbf{y} \in R^*) \leq \alpha$.

The *Neyman-Pearson Lemma* then states that $\beta(\boldsymbol{\theta}_1) \geq \beta^*(\boldsymbol{\theta}_1)$ (i.e. the likelihood ratio test is the most powerful test possible).

Proof follows from the fact that

$$\{\phi(\mathbf{y}) - \phi^*(\mathbf{y})\}\{f_{\theta_1}(\mathbf{y}) - k f_{\theta_0}(\mathbf{y})\} \geq 0,$$

since from the definition of R, the first bracket is non-negative whenever the second bracket is non-negative, and it is non-positive whenever the second bracket is negative. In consequence,

$$\begin{aligned} 0 &\leq \int \{\phi(\mathbf{y}) - \phi^*(\mathbf{y})\}\{f_{\theta_1}(\mathbf{y}) - k f_{\theta_0}(\mathbf{y})\} d\mathbf{y} \\ &= \beta(\boldsymbol{\theta}_1) - \beta^*(\boldsymbol{\theta}_1) - k\{\beta(\boldsymbol{\theta}_0) - \beta^*(\boldsymbol{\theta}_0)\} \leq \beta(\boldsymbol{\theta}_1) - \beta^*(\boldsymbol{\theta}_1), \end{aligned}$$

since $\{\beta(\boldsymbol{\theta}_0) - \beta^*(\boldsymbol{\theta}_0)\} \geq 0$. So the result is proven. Casella and Berger (1990) give a fuller version of the lemma, on which this proof is based.

8.6 Distribution of the generalised likelihood ratio statistic

Having examined the motivation for p-values, and the justification for using the likelihood ratio as a test statistic, it is time to derive the general result (9). Consider testing:

$$H_0 : \mathbf{R}(\boldsymbol{\theta}) = \mathbf{0} \quad \text{vs.} \quad H_1 : \mathbf{R}(\boldsymbol{\theta}) \neq \mathbf{0},$$

where $\mathbf{R}$ is a vector-valued function of $\boldsymbol{\theta}$, such that H_0 imposes r restrictions on the parameter vector. If H_0 is true, then in the limit as $n \to \infty$,

$$2\lambda = 2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_0)\} \sim \chi^2_r, \tag{20}$$

where l is the log-likelihood function and $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$. $\hat{\boldsymbol{\theta}}_0$ is the value of $\boldsymbol{\theta}$ maximising the likelihood subject to the constraint $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$. This result is used to calculate approximate p-values for the test.

To derive (20), first re-parameterise so that $\boldsymbol{\theta}^{\mathrm{T}} = (\boldsymbol{\psi}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})$, where $\boldsymbol{\psi}$ is r dimensional and the null hypothesis can be rewritten $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$. Such re-parameterisation is always possible, but is only necessary for deriving (20), not for its use.

Let the unrestricted MLE be $(\hat{\boldsymbol{\psi}}^{\rm T}, \hat{\boldsymbol{\gamma}}^{\rm T})$, and let $(\boldsymbol{\psi}_0^{\rm T}, \hat{\boldsymbol{\gamma}}_0^{\rm T})$ be the MLE under the restrictions defining the null hypothesis. To make progress, $\hat{\boldsymbol{\gamma}}_0$ must be expressed in terms of $\hat{\boldsymbol{\psi}}$, $\hat{\boldsymbol{\gamma}}$ and $\boldsymbol{\psi}_0$. Taking a Taylor expansion of l around the unrestricted MLE, $\hat{\boldsymbol{\theta}}$, yields

$$l(\boldsymbol{\theta}) \simeq l(\hat{\boldsymbol{\theta}}) - \frac{1}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^{\rm T} \mathbf{H} \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right), \tag{21}$$

where $H_{i,j} = - \left. \partial^2 l / \partial\theta_i \partial\theta_j \right|_{\hat{\boldsymbol{\theta}}}$. Exponentiating produces

$$L(\boldsymbol{\theta}) \simeq L(\hat{\boldsymbol{\theta}}) \exp\left[- \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^{\rm T} \mathbf{H} \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)/2 \right]$$

(i.e. the likelihood can be approximated by a function proportional to the p.d.f. of an $N(\hat{\boldsymbol{\theta}}, \mathbf{H}^{-1})$ random vector). So, in the large sample limit and defining $\boldsymbol{\Sigma} = \mathbf{H}^{-1}$, the likelihood is proportional to the p.d.f. of

$$N\left(\left[\begin{array}{c} \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{\gamma}} \end{array} \right], \left[\begin{array}{cc} \boldsymbol{\Sigma}_{\psi\psi} & \boldsymbol{\Sigma}_{\psi\gamma} \\ \boldsymbol{\Sigma}_{\gamma\psi} & \boldsymbol{\Sigma}_{\gamma\gamma} \end{array} \right] \right).$$

If $\boldsymbol{\psi} = \boldsymbol{\psi}_0$ then this p.d.f. will be maximised by $\hat{\boldsymbol{\gamma}}_0 = E(\boldsymbol{\gamma}|\boldsymbol{\psi}_0)$, which, from general properties of MVN densities[13], is

$$\hat{\boldsymbol{\gamma}}_0 = \hat{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}_{\gamma\psi}\boldsymbol{\Sigma}_{\psi\psi}^{-1}(\boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}}). \tag{22}$$

If the null hypothesis is true, then in the large sample limit $\hat{\boldsymbol{\psi}} \to \boldsymbol{\psi}_0$ (in probability) so that the approximate likelihood tends to the true likelihood, and we can expect (22) to hold for the maximisers of the true likelihood.

It helps to express (22) in terms of the partitioned version of $\mathbf{H}$. Writing $\boldsymbol{\Sigma}\mathbf{H} = \mathbf{I}$ in partitioned form

$$\left[\begin{array}{cc} \boldsymbol{\Sigma}_{\psi\psi} & \boldsymbol{\Sigma}_{\psi\gamma} \\ \boldsymbol{\Sigma}_{\gamma\psi} & \boldsymbol{\Sigma}_{\gamma\gamma} \end{array} \right] \left[\begin{array}{cc} \mathbf{H}_{\psi\psi} & \mathbf{H}_{\psi\gamma} \\ \mathbf{H}_{\gamma\psi} & \mathbf{H}_{\gamma\gamma} \end{array} \right] = \left[\begin{array}{cc} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{array} \right],$$

and multiplying out, results in two useful equations:

$$\boldsymbol{\Sigma}_{\psi\psi}\mathbf{H}_{\psi\psi} + \boldsymbol{\Sigma}_{\psi\gamma}\mathbf{H}_{\gamma\psi} = \mathbf{I} \text{ and } \boldsymbol{\Sigma}_{\psi\psi}\mathbf{H}_{\psi\gamma} + \boldsymbol{\Sigma}_{\psi\gamma}\mathbf{H}_{\gamma\gamma} = \mathbf{0}. \tag{23}$$

Rearranging (23) while noting that, by symmetry, $\mathbf{H}_{\psi\gamma}^{\rm T} = \mathbf{H}_{\gamma\psi}$ and $\boldsymbol{\Sigma}_{\psi\gamma}^{\rm T} = \boldsymbol{\Sigma}_{\gamma\psi}$, yields

$$\boldsymbol{\Sigma}_{\psi\psi}^{-1} = \mathbf{H}_{\psi\psi} - \mathbf{H}_{\psi\gamma}\mathbf{H}_{\gamma\gamma}^{-1}\mathbf{H}_{\gamma\psi} \tag{24}$$

and $-\mathbf{H}_{\gamma\gamma}^{-1}\mathbf{H}_{\gamma\psi} = \boldsymbol{\Sigma}_{\gamma\psi}\boldsymbol{\Sigma}_{\psi\psi}^{-1}$. Substituting the latter into (22), we obtain

$$\hat{\boldsymbol{\gamma}}_0 = \hat{\boldsymbol{\gamma}} + \mathbf{H}_{\gamma\gamma}^{-1}\mathbf{H}_{\gamma\psi}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0). \tag{25}$$

Now provided that the null hypothesis is true so that $\hat{\boldsymbol{\psi}}$ is close to $\boldsymbol{\psi}_0$, we can reuse the expansion (21) and write the log likelihood at the restricted MLE as

$$l(\boldsymbol{\psi}_0, \hat{\boldsymbol{\gamma}}_0) \simeq l(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\gamma}}) - \frac{1}{2} \left[\begin{array}{c} \boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\gamma}} \end{array} \right]^{\rm T} \mathbf{H} \left[\begin{array}{c} \boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\gamma}} \end{array} \right].$$

Hence

$$2\lambda = 2\{l(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\gamma}}) - l(\boldsymbol{\psi}_0, \hat{\boldsymbol{\gamma}}_0)\} \simeq \left[\begin{array}{c} \boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\gamma}} \end{array} \right]^{\rm T} \mathbf{H} \left[\begin{array}{c} \boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}} \\ \hat{\boldsymbol{\gamma}}_0 - \hat{\boldsymbol{\gamma}} \end{array} \right].$$

[13] Suppose that $\mathbf{Z}$ and $\mathbf{X}$ are random vectors with a multivariate normal joint distribution. Partitioning their joint covariance matrix

$$\boldsymbol{\Sigma} = \left[\begin{array}{cc} \boldsymbol{\Sigma}_z & \boldsymbol{\Sigma}_{zx} \\ \boldsymbol{\Sigma}_{xz} & \boldsymbol{\Sigma}_x \end{array} \right] \text{ then } \mathbf{X}|\mathbf{z} \sim N(\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xz}\boldsymbol{\Sigma}_z^{-1}(\mathbf{z} - \boldsymbol{\mu}_z), \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_{xz}\boldsymbol{\Sigma}_z^{-1}\boldsymbol{\Sigma}_{zx}).$$

Substituting for $\hat{\gamma}_0$ from (25) and writing out $\mathbf{H}$ in partitioned form gives

$$\begin{aligned}2\lambda &\simeq \begin{bmatrix} \boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}} \\ \mathbf{H}_{\gamma\gamma}^{-1}\mathbf{H}_{\gamma\psi}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathbf{H}_{\psi\psi} & \mathbf{H}_{\psi\gamma} \\ \mathbf{H}_{\gamma\psi} & \mathbf{H}_{\gamma\gamma} \end{bmatrix} \begin{bmatrix} \boldsymbol{\psi}_0 - \hat{\boldsymbol{\psi}} \\ \mathbf{H}_{\gamma\gamma}^{-1}\mathbf{H}_{\gamma\psi}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \end{bmatrix} \\ &= (\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^{\mathrm{T}}\left[\mathbf{H}_{\psi\psi} - \mathbf{H}_{\psi\gamma}\mathbf{H}_{\gamma\gamma}^{-1}\mathbf{H}_{\gamma\psi}\right](\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \\ &= (\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^{\mathrm{T}}\boldsymbol{\Sigma}_{\psi\psi}^{-1}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0),\end{aligned}$$

where the final equality follows from (24). If H_0 is true, then as $n \to \infty$ this expression will tend towards exactness as $\hat{\boldsymbol{\psi}} \to \boldsymbol{\psi}_0$. Furthermore, provided $\mathbf{H} \to \boldsymbol{\mathcal{I}}$ as $n \to \infty$, then $\boldsymbol{\Sigma}$ tends to $\boldsymbol{\mathcal{I}}^{-1}$, and hence $\boldsymbol{\Sigma}_{\psi\psi}$ tends to the covariance matrix of $\hat{\boldsymbol{\psi}}$, by (19). Hence, by the asymptotic normality of the MLE $\hat{\boldsymbol{\psi}}$, $2\lambda \sim \chi^2_r$, under H_0.

8.7 Confidence intervals: what range of hypotheses would we accept?

Everyone who has done any statistics is familiar with the notion of testing a hypothesis by looking to see whether a hypothetical value lies within a confidence interval. For example, to test $H_0 : \theta = .5$ at the 5% level, we just need to see whether .5 lies within the 95% CI for θ, or not. We can also reverse this construction, and use a hypothesis test to create a confidence interval.

A frequentist $100(1-\alpha)\%$ confidence set is a random set with a probability of $1-\alpha$ of including the true parameter value. When hypothesis testing, if we choose to accept a null hypothesis whenever the p-value is $>\geq \alpha$ (and reject otherwise), then we will accept the null hypothesis with probability $1-\alpha$ if it is true.

So suppose we find the *set* containing the values that we would have accepted in a hypothesis test. Over multiple replicates of the data gathering process such a set must have a probability of $1-\alpha$ of including the true parameter value (since with probability $1-\alpha$ the test will accept the true parameter value, and with probability α will reject it). Hence our set of acceptable values is a $100(1-\alpha)\%$ confidence set. For a scalar parameter the set is often continuous within some interval, and we have a confidence interval.

Finding confidence intervals/sets by finding the range of values that would be accepted as the null hypothesis in a test is known as *test inversion*. For example we can look at the range of θ_0 values that would be accepted when testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ using a GLRT. This usually has to be done numerically (e.g. produce a fine grid of possible θ_0 values, and produce a p-value for each: the interval is then all the values for which the p-value is $\geq$ the threshold α).

A nice property of such *Wilks* (approximate) confidence intervals is that every parameter value inside the interval has a higher likelihood than every value outside the interval. This contrasts with intervals based on approximation (19), which are symmetric about $\hat{\boldsymbol{\theta}}$, and hence may even include impossible values (e.g. negative values for a variance parameter).

8.8 Regularity conditions

The preceding results depend on some assumptions.

1. The densities defined by distinct values of $\boldsymbol{\theta}$ are distinct. If this is not the case the parameters need not be *identifiable*, and there is no guarantee of consistency.

2. $\boldsymbol{\theta}_t$ is interior to the space of possible parameter values. This is necessary in order to be able to approximate the log likelihood by a Taylor expansion in the vicinity of $\boldsymbol{\theta}_t$.

3. Within some neighbourhood of $\boldsymbol{\theta}_t$, the first three derivatives of the log likelihood exist and are bounded, while the Fisher information matrix satisfies (17) and is positive definite and finite. The various Taylor expansions and the arguments leading to (19) depend on this.

When these assumptions are met the results of this section are very general and apply in many situations well beyond the i.i.d setting. When they are not met, some or all of the results of Sections 8.3 to 8.6 will fail.

8.9 AIC: Akaike's information criterion

Suppose that we want to decide between two models that are not nested, so that the GLRT is inapplicable. Equally we might not want to follow the hypothesis testing approach to model selection, which is essentially *stick with the simplest model not obviously inconsistent with the data*. We might want to simply choose the model that is 'best' in some sense. Obviously we can not do this by choosing the model that just has the highest likelihood, since more complicated models always tend to have higher likelihood, whether or not the extra complication really reflects something in the true data generating process.

One approach is to select models that are as close as possible to the true model in the probabilistic average sense of minimising the *Kullback-Leibler divergence*

$$K(f_\theta, f_t) = \int \{\log f_t(\mathbf{y}) - \log f_\theta(\mathbf{y})\} f_t(\mathbf{y}) d\mathbf{y}, \tag{26}$$

where f_t is the true density of $\mathbf{y}$ and f_θ is the model approximation to it. To make this aspiration practical, we need to choose some version of K that can be estimated, and it turns out that the expected value of $K(f_{\hat\theta}, f_t)$ is tractable, where $\hat{\boldsymbol\theta}$ is the MLE.

Although we cannot compute it, consider the value of $\boldsymbol\theta$ that would minimise (26), and denote it by $\boldsymbol\theta_K$. Now consider the Taylor expansion

$$\log f_{\hat\theta}(\mathbf{y}) \simeq \log f_{\theta_K}(\mathbf{y}) + (\hat{\boldsymbol\theta} - \boldsymbol\theta_K)^{\mathrm{T}} \left.\frac{\partial \log f_\theta}{\partial \boldsymbol\theta}\right|_{\theta_K} + \frac{1}{2}(\hat{\boldsymbol\theta} - \boldsymbol\theta_K)^{\mathrm{T}} \left.\frac{\partial^2 \log f_\theta}{\partial \boldsymbol\theta \partial \boldsymbol\theta^{\mathrm{T}}}\right|_{\theta_K} (\hat{\boldsymbol\theta} - \boldsymbol\theta_K). \tag{27}$$

If $\boldsymbol\theta_K$ minimises K then $\int \partial \log f_\theta / \partial\theta|_{\theta_K} f_t d\mathbf{y} = 0$, so substituting (27) into $K(f_{\hat\theta}, f_t)$, while *treating $\hat{\boldsymbol\theta}$ as fixed*[14] results in

$$K(f_{\hat\theta}, f_t) \simeq K(f_{\theta_K}, f_t) + \frac{1}{2}(\hat{\boldsymbol\theta} - \boldsymbol\theta_K)^{\mathrm{T}} \boldsymbol{\mathcal{I}}_{\theta_K} (\hat{\boldsymbol\theta} - \boldsymbol\theta_K), \tag{28}$$

where $\boldsymbol{\mathcal{I}}_{\theta_K}$ is the information matrix at $\boldsymbol\theta_K$. Now assume that the model is sufficiently correct that $E(\hat{\boldsymbol\theta}) \simeq \boldsymbol\theta_K$ and cov$(\hat{\boldsymbol\theta}) \simeq \boldsymbol{\mathcal{I}}_{\theta_K}$, at least for large samples. In this case, and reusing results from the end of Section 8.6,

$$E\{l(\hat{\boldsymbol\theta}) - l(\boldsymbol\theta_K)\} \simeq E\left\{\frac{1}{2}(\hat{\boldsymbol\theta} - \boldsymbol\theta_K)^{\mathrm{T}} \boldsymbol{\mathcal{I}}_{\theta_K} (\hat{\boldsymbol\theta} - \boldsymbol\theta_K)\right\} \simeq p/2 \tag{29}$$

where p is the dimension of $\boldsymbol\theta$. So taking expectations of (28) and substituting an approximation from (29),

$$EK(f_{\hat\theta}, f_t) \simeq K(f_{\theta_K}, f_t) + p/2. \tag{30}$$

Since this still involves the unknownable f_t, consider

$$\begin{aligned} E\{-l(\hat{\boldsymbol\theta})\} &= E[-l(\boldsymbol\theta_K) - \{l(\hat{\boldsymbol\theta}) - l(\boldsymbol\theta_K)\}] \\ &\simeq -\int \log\{f_{\theta_K}(\mathbf{y})\} f_t(\mathbf{y}) d\mathbf{y} - p/2 \text{ by } (29) \\ &= K(f_{\theta_K}, f_t) - p/2 - \int \log\{f_t(\mathbf{y})\} f_t(\mathbf{y}) d\mathbf{y}. \end{aligned}$$

Using this result to eliminate $K(f_{\theta_K}, f_t)$ from (30) suggests the estimate

$$\widehat{EK(f_{\hat\theta}, f_t)} = -l(\hat{\boldsymbol\theta}) + p + \int \log\{f_t(\mathbf{y})\} f_t(\mathbf{y}) d\mathbf{y}.$$

Since the last term on the right-hand side only involves the truth, this last estimate is minimised by whichever model minimises

$$\text{AIC} = -2l(\hat{\boldsymbol\theta}) + 2p,$$

[14] By treating $\hat{\boldsymbol\theta}$ as fixed we are effectively assessing the expected likelihood ratio between model and truth for new data.

where the factor of 2 is by convention, to put AIC on the same scale as 2λ from Section 8.6. A possible concern here is that (29) is not justified if the model is oversimplified and hence poor, but in practice this is unproblematic, because the log likelihood decreases sharply as the approximation deteriorates. See Davison (2003) for a fuller derivation.

An objection to AIC is that it is not consistent: as $n \to \infty$, the probability of selecting the correct model does not tend to 1. For nested models (20) states that the difference in $-2l(\hat{\theta})$ between the true model and an overly complex model follows a χ^2_r distribution, where r is the number of spurious parameters. Neither χ^2_r nor $2p$ depends on n, so the probability of selecting the overly complex model by AIC is nonzero and independent of n (for n large). The same objection could also be made about hypothesis testing, unless we allow the accept/reject threshold to change with n.[15]

9 Bayesian Inference

When we applied Bayes' theorem to the linear model, it was fairly straightforward to compute the posterior distribution, but in general such direct calculations are not possible, and to apply the Bayesian approach we must do one of two things: either approximate the posterior, or posterior expectations that are of more direct interest, or find a way of simulating from the posterior that avoids the difficulties inherent in trying to compute posterior densities directly. You can appreciate why working with the posterior directly is difficult by looking at the form of Bayes theorem:

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{y})}.$$

The top of the r.h.s. is readily computable directly from the model specification, but what about $f(\mathbf{y})$? To evaluate it we have to evaluate

$$f(\mathbf{y}) = \int f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}$$

but the integral is only tractable in special cases, and high dimensional numerical integration is usually very difficult.

Since this course is of finite length, we will only cover the simulation approach here. The key is to use the fact that

$$f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\boldsymbol{\theta}, \mathbf{y}) \ \ (= f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}))\,,$$

to devise methods for simulating from $f(\boldsymbol{\theta}|\mathbf{y})$, which only require evaluation of $f(\boldsymbol{\theta}, \mathbf{y})$ with the observed data values plugged in for $\mathbf{y}$. The resulting Markov chain Monte Carlo (MCMC) methods simulate (correlated) samples from the distribution of the model unknowns (parameters and any random effects), given the data. Based on the unknowns at one step, a new set of unknowns is generated in such a way that the stable distribution of the resulting Markov chain is the distribution of interest.

9.1 Markov chains

To use MCMC we do not require much theoretical background on Markov chains, but some basic concepts are needed. A sequence of random vectors, $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots$, constitutes a *Markov chain* if, for any j,

$$f(\mathbf{x}_j|\mathbf{x}_{j-1}, \mathbf{x}_{j-2}, \ldots, \mathbf{x}_1) = f(\mathbf{x}_j|\mathbf{x}_{j-1}).$$

For notational convenience let us rewrite the density of $\mathbf{x}_j$ given $\mathbf{x}_{j-1}$ as $P(\mathbf{x}_j|\mathbf{x}_{j-1})$, the *transition kernel* of the Markov chain. If there exists a density f_x such that

$$f_x(\mathbf{x}_j) = \int P(\mathbf{x}_j|\mathbf{x}_{j-1})f_x(\mathbf{x}_{j-1})d\mathbf{x}_{j-1}$$

[15] The inconsistency of AIC is not in itself the reason for the empirical observation that AIC tends to select increasingly complex models as n increases. If the true model is among those considered then (20) does not imply that the probability of rejecting it increases with sample size. However, if all the models under consideration are wrong, then we will tend to select increasingly complex approximations as the sample size increases and the predictive disadvantages of complexity diminish.

(where f_x denotes the same density on both sides), then this is the *stationary distribution* of the chain. Existence of a stationary distribution depends on P being *irreducible*, meaning that wherever we start the chain, there is a positive probability of visiting all possible values of $\mathbf{X}$. If the chain is also *recurrent*, meaning that if its length tends to infinity it will revisit any non-negligible set of values an infinite number of times, then its stationary distribution is also its *limiting distribution*. This means that the chain can be started from any possible value of $\mathbf{X}$, and its marginal distribution will eventually converge on f_x. In consequence as the simulation length, J, tends to infinity,

$$\frac{1}{J}\sum_{j=1}^{J}\phi(\mathbf{X}_j) \to E_{f_x}\{\phi(\mathbf{X})\}$$

(also known as *ergodicity*). This extension of the law of large numbers to this particular sort of correlated sequence is what makes MCMC methods useful, so the methods discussed in this section will be set up to produce chains with the required properties.

9.2 Reversibility

Now we turn to the issue of constructing Markov chains to generate sequences $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ from $f(\boldsymbol{\theta}|\mathbf{y})$. An MCMC scheme will generate samples from $f(\boldsymbol{\theta}|\mathbf{y})$, if it satisfies the *detailed balance* condition (also termed *reversibility*). Let $P(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j)$ be the p.d.f. of $\boldsymbol{\theta}_i$ given $\boldsymbol{\theta}_j$, according to the chain. We require

$$P(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{j-1})f(\boldsymbol{\theta}_{j-1}|\mathbf{y}) = P(\boldsymbol{\theta}_{j-1}|\boldsymbol{\theta}_j)f(\boldsymbol{\theta}_j|\mathbf{y}). \tag{31}$$

The left hand side of (31) is the joint p.d.f. of $\boldsymbol{\theta}_j, \boldsymbol{\theta}_{j-1}$ from the chain, if $\boldsymbol{\theta}_{j-1}$ is from $f(\boldsymbol{\theta}|\mathbf{y})$. Integrating w.r.t. $\boldsymbol{\theta}_{j-1}$ gives the corresponding marginal density of $\boldsymbol{\theta}_j$,

$$\begin{aligned}\int P(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{j-1})f(\boldsymbol{\theta}_{j-1}|\mathbf{y})d\boldsymbol{\theta}_{j-1} &= \int P(\boldsymbol{\theta}_{j-1}|\boldsymbol{\theta}_j)f(\boldsymbol{\theta}_j|\mathbf{y})d\boldsymbol{\theta}_{j-1} \\ &= f(\boldsymbol{\theta}_j|\mathbf{y}).\end{aligned}$$

That is, given $\boldsymbol{\theta}_{j-1}$ from $f(\boldsymbol{\theta}|\mathbf{y})$, the chain generates $\boldsymbol{\theta}_j$ also from $f(\boldsymbol{\theta}|\mathbf{y})$ as result of (31). So provided that we start with a $\boldsymbol{\theta}_1$ that is not impossible according to $f(\boldsymbol{\theta}|\mathbf{y})$, then the chain will generate from the target distribution. How quickly it will converge to the high-probability region of $f(\boldsymbol{\theta}|\mathbf{y})$ is another matter.

9.3 Metropolis Hastings

The Metropolis-Hastings method constructs a chain with an appropriate P. It works as follows:

1. Pick a *proposal distribution* $q(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{j-1})$ (e.g. a normal centred on $\boldsymbol{\theta}_{j-1}$). Then pick a value $\boldsymbol{\theta}_0$, set $j = 1$ and iterate steps 2 and 3:

2. Generate $\boldsymbol{\theta}'_j$ from $q(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{j-1})$.

3. Set $\boldsymbol{\theta}_j = \boldsymbol{\theta}'_j$ with probability

$$\alpha = \min\left\{1, \frac{f(\mathbf{y}|\boldsymbol{\theta}'_j)f(\boldsymbol{\theta}'_j)q(\boldsymbol{\theta}_{j-1}|\boldsymbol{\theta}'_j)}{f(\mathbf{y}|\boldsymbol{\theta}_{j-1})f(\boldsymbol{\theta}_{j-1})q(\boldsymbol{\theta}'_j|\boldsymbol{\theta}_{j-1})}\right\}, \tag{32}$$

   otherwise setting $\boldsymbol{\theta}_j = \boldsymbol{\theta}_{j-1}$. Increment j.

Note that the q terms cancel if q depends only on the magnitude of $\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1}$ (e.g. if q is a normal centred on $\boldsymbol{\theta}_{j-1}$). The same goes for the prior densities, $f(\boldsymbol{\theta})$, if they are improper uniform. If both of these simplifications hold then we have $\alpha = \min\left\{1, L(\boldsymbol{\theta}'_j)/L(\boldsymbol{\theta}_{j-1})\right\}$, so that we are accepting or rejecting on the basis of the likelihood ratio.

An important consideration is that $\boldsymbol{\theta}_1$ may be very improbable so that the chain may take many iterations to reach the high-probability region of $f(\boldsymbol{\theta}|\mathbf{y})$. For this reason we usually need to discard a *burn-in period* consisting of the first few hundred or thousand $\boldsymbol{\theta}_j$ vectors simulated.

9.4 Why Metropolis Hastings works

As we saw in Section 9.2, the Metropolis-Hastings (MH) method will work if it satisfies detailed balance. It does, and proof is easy. To simplify notation let $\pi(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$, so that the MH acceptance probability from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ is

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}.$$

We need to show that $\pi(\boldsymbol{\theta})P(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}')P(\boldsymbol{\theta}|\boldsymbol{\theta}')$. This is trivial if $\boldsymbol{\theta}' = \boldsymbol{\theta}$. Otherwise we know that $P(\boldsymbol{\theta}'|\boldsymbol{\theta}) = q(\boldsymbol{\theta}'|\boldsymbol{\theta})\alpha(\boldsymbol{\theta}', \boldsymbol{\theta})$, from which it follows that

$$\begin{aligned}\pi(\boldsymbol{\theta})P(\boldsymbol{\theta}'|\boldsymbol{\theta}) &= \pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})\min\left\{1, \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\} \\ &= \min\left\{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta}), \pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')\right\} = \pi(\boldsymbol{\theta}')P(\boldsymbol{\theta}|\boldsymbol{\theta}'),\end{aligned}$$

where the final equality is by symmetry of the third term above.

9.5 A toy example with Metropolis Hastings

To illustrate the basic simplicity of the approach, consider an example for which simulation is certainly not required. Suppose we have 20 independent observations x_i, that can each be modelled as $N(\mu, \sigma^2)$ random variables, and we are interested in inference about μ and σ. In the absence of real prior knowledge about these parameters, suppose that we decide on prior independence and improper prior densities, so that $f(\mu) \propto k$ and $f(\log \sigma) \propto c$ where k and c are constants (values immaterial). We work with $\log \sigma$, because σ is inherently positive.[16]

This specification sets up the Bayesian model. To simulate from the corresponding posterior for the parameters using MH we also need a proposal distribution. In this case let us choose independent scaled t_3 distributions centred on the current parameter values for both parameters. This means that we will propose new parameter values by simply adding a multiple of t_3 random deviates to the current parameter values: this is an example of a *random walk* proposal. The proposed values are then accepted or rejected using the MH mechanism.

Here is some R code to implement this example, using simulated x data from $N(1, 2)$. The parameters are assumed to be in vectors $\boldsymbol{\theta} = (\mu, \log \sigma)^{\mathrm{T}}$ to be stored columnwise in a matrix `theta`.

```
set.seed(1);x <- rnorm(20)*2+1 ## simulated data
n.rep <- 10000; n.accept <- 0
theta <- matrix(0,2,n.rep) ## storage for sim. values
ll0 <- sum(dnorm(x,mean=theta[1,1],
             sd=exp(theta[2,1]),log=TRUE))
for (i in 2:n.rep) { ## The MH loop
  theta[,i] <- theta[,i-1] + rt(2,df=3)*.5 ## proposal
  ll1 <- sum(dnorm(x,mean=theta[1,i],
               sd=exp(theta[2,i]),log=TRUE))
  if (exp(ll1-ll0)>runif(1)) { ## MH accept/reject
    ll0 <- ll1; n.accept <- n.accept + 1 ## accept
  } else theta[,i] <- theta[,i-1] ## reject
}
n.accept/n.rep ## proportion of proposals accepted
```

Working on the log probability scale is a sensible precaution against probabilities underflowing to zero (i.e. being evaluated as zero, merely because they are smaller than the smallest number the computer can represent). The acceptance rate of the chain is monitored, to try to ensure that it is neither too high, nor too low. Too low an acceptance rate is obviously a problem, because the chain then stays in the same state for long periods, resulting in very high correlation and the need for very long runs in order to obtain a sufficiently representative sample. A low acceptance rate may result from a proposal that tries to make very large steps, which are almost always rejected. Less obviously, very high acceptance rates are also a problem because they only tend to occur when the

[16] Note that a uniform prior on $\log \sigma$ puts a great deal of weight on $\sigma \approx 0$, which can cause problems in cases where the data contain little information on σ.

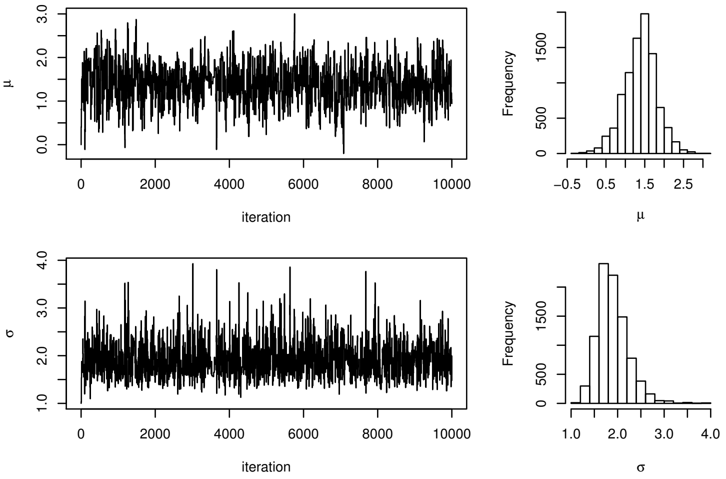

Figure 8: Results of the Metropolis-Hastings method applied to the toy model in Section 9.5. The left panels show the simulated values for the parameters at each step of the chain, joined by lines. The chains converge very rapidly and mix well in this case. The right panels show histograms of the simulated values of the parameters after discarding the first 1000 steps as burn-in.

proposal is making very small steps, relative to the scale of variability suggested by the posterior. This again leads to excessively autocorrelated chains and the need for very long runs. It turns out that in many circumstances it is near optimal to accept about a quarter of steps (Roberts et al., 1997). Here we can control the acceptance rate through the standard deviation of the proposal distribution: some experimentation was needed to find that setting this to 0.5 gave an acceptance rate of about 23%.

We need to look at the output. The following code nicely arranges plots of the chain components against iteration, and histograms of the chain components after discarding a burn-in period of 1000 iterations:

```
layout(matrix(c(1,2,1,2,3,4),2,3))
plot(1:n.rep,theta[1,],type="l",xlab="iteration",
                        ylab=expression(mu))
plot(1:n.rep,exp(theta[2,]),type="l",xlab="iteration",
                        ylab=expression(sigma))
hist(theta[1,-(1:1000)],main="",xlab=expression(mu))
hist(exp(theta[2,-(1:1000)]),main="",
                        xlab=expression(sigma))
```

The results are shown in Figure 8. The left-hand plots show that the chains appear to have reached a stable state very quickly (rapid convergence) and then move rapidly around that distribution (good mixing). The right-hand histograms illustrate the shape of the marginal distributions of the parameters according to the posterior.

9.6 Designing proposal distributions

The catch with Metropolis-Hastings is the proposal distribution. To get the chain to mix well we have to get it right, and for complex models it is seldom the case that we can get away with updating all elements of the parameter vector with independent random steps, all with the same variance, as in the toy example from the last section. In most practical applications, several pilot runs of the MH sampler will be needed to ‘tune’ the proposal distribution, along with some analysis of model structure. In particular:

1. With simple independent random walk proposals, different standard deviations are likely to be required for different parameters.

2. As its dimension increases it often becomes increasingly difficult to update all elements of $\boldsymbol{\theta}$ simultaneously, unless uselessly tiny steps are proposed. The difficulty is that a purely random step is increasingly unlikely

to land in a place where the posterior is non-negligible as dimension increases. In addition it is hard to tune componentwise standard deviations if all elements are proposed together. A solution is to break the proposal down into smaller parts and to only update small mutually exclusive subsets of the parameter vector at each step. The subset to update can be chosen randomly, or we can systematically work through all subsets in some order. This approach only affects the computation of the proposal; the computation of the acceptance ratio is unchanged. But notice that we increase the work required to achieve an update of the whole vector, because the computations required for the accept/reject decision have to be repeated for each subset of parameters.

3. It may be necessary to use correlated proposals, rather than updating each element of $\boldsymbol{\theta}$ independently. Bearing in mind the impractical fact that the perfect proposal would be the posterior itself, it is tempting to base the proposal on a Gaussian approximation to the posterior, obtained analytically, or from a pilot run.

9.7 Gibbs sampling

When considering the design of proposal distributions, two facts were important. First, it is often necessary to update parameters in blocks. Second, the perfect proposal is the posterior itself: substituting it for q in (32), we find that $\alpha = 1$, so that such proposals would always be accepted. On its own the second fact is impractical, but applied blockwise it can result in a very efficient scheme known as Gibbs sampling.[17]

The basic idea is this. Suppose we have a random draw from the joint posterior distribution of $\boldsymbol{\theta}^{[-1]} = (\theta_2, \theta_3, \ldots \theta_q)^{\mathrm{T}}$, and would like a draw from the joint posterior distribution of the whole of $\boldsymbol{\theta}$. This is easy given that $f(\boldsymbol{\theta}|\mathbf{y}) = f(\theta_1|\boldsymbol{\theta}^{[-1]}, \mathbf{y}) f(\boldsymbol{\theta}^{[-1]}|\mathbf{y})$: simulate θ_1 from $f(\theta_1|\boldsymbol{\theta}^{[-1]}, \mathbf{y})$, append the result to $\boldsymbol{\theta}^{[-1]}$ and we are done. There is nothing special about θ_1 in this process. The same thing would have worked for any other θ_i, or indeed for several θ_i simultaneously. In fact, if we have access to the conditional distributions for all elements of $\boldsymbol{\theta}$ then we could simply cycle through the θ_i updating each in turn and thereby generating a (correlated) sequence of draws from $f(\boldsymbol{\theta}|\mathbf{y})$.

In general then, suppose that the parameter row vector is partitioned into subvectors $\boldsymbol{\theta} = (\boldsymbol{\theta}^{[1]}, \boldsymbol{\theta}^{[2]}, \ldots, \boldsymbol{\theta}^{[K]})$. Further define

$$\tilde{\boldsymbol{\theta}}_j^{[-k]} = (\boldsymbol{\theta}_{j+1}^{[1]}, \boldsymbol{\theta}_{j+1}^{[2]}, \ldots, \boldsymbol{\theta}_{j+1}^{[k-1]}, \boldsymbol{\theta}_j^{[k+1]}, \ldots, \boldsymbol{\theta}_j^{[K]}).$$

Then, given an initial $\boldsymbol{\theta}_1$, J steps of the Gibbs sampler proceed as follows

1. For $j = 1, \ldots, J$ repeat...
2. For $k = 1, \ldots, K$ simulate $\boldsymbol{\theta}_{j+1}^{[k]} \sim f(\boldsymbol{\theta}^{[k]}|\tilde{\boldsymbol{\theta}}_j^{[-k]}, \mathbf{y})$.

Notice from the definition of $\tilde{\boldsymbol{\theta}}_j^{[-k]}$ that we always condition on the most recently simulated values.

At this point, the obvious question is how all those conditional distributions are to be found. It is often natural to specify a model in terms of a hierarchy of conditional dependencies, but these dependencies all run in one direction, leaving the problem of working out the conditional dependencies in the other direction. Alternatively, if we attempt to specify the model directly in terms of all its conditional distributions, we will have the no less tricky problem of checking that our specification actually corresponds to a properly defined joint distribution.

Actually, the problem of identifying the conditionals is less daunting than it first seems, and even if we cannot recognise the conditionals as belonging to some standard distribution, it is always possible to devise some way of simulating from them, as the last resort simply using a Metropolis Hastings step for the component. The main trick for recognising conditionals is to use the fact that, for any p.d.f., multiplicative factors that do not involve the argument of the p.d.f. must be part of the normalising constant. To identify a p.d.f. it therefore suffices to recognise its form, to within a normalising constant. The following example helps to clarify this.

9.8 Toy Gibbs sampling example

Consider again the toy example from Section 9.5, but this time with proper priors on the parameters of the normal model. So we have $n = 20$ observations of a $N(\mu, \phi)$ random variable, where $1/\phi \sim G(a, b)$ (a gamma random

[17] In honour of the physical model to which it was first applied.

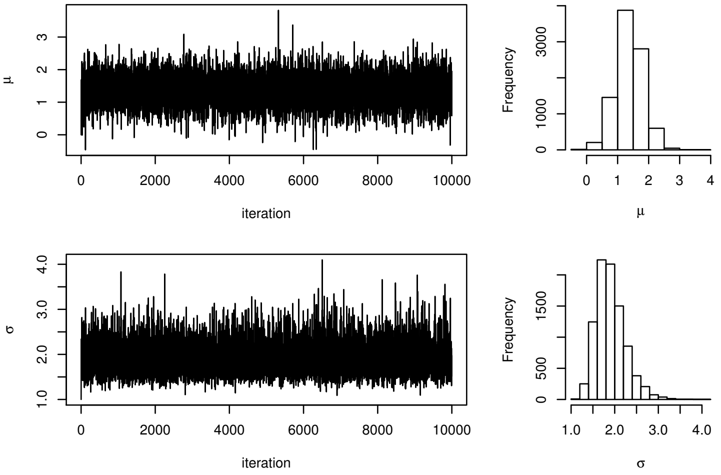

Figure 9: Results of Gibbs sampling applied to the toy model in Section 9.8. The left panels show the simulated values for the parameters at each step of the chain, joined by lines. The right panels show histograms of the simulated values of the parameters after discarding the first 1000 steps.

variable, with p.d.f. $f(y) = b^a y^{a-1} e^{-by}/\Gamma(a)$) and (independently) $\mu \sim N(c, d)$. a, b, c and d are constants to be specified. The joint density is given by the product of the three densities involved:

$$f(\mathbf{x}, \mu, \phi) \propto \frac{1}{\phi^{n/2}} e^{-\sum_i (x_i-\mu)^2/(2\phi)} e^{-(\mu-c)^2/(2d)} \frac{1}{\phi^{a-1}} e^{-b/\phi}$$

where factors not involving $\mathbf{x}$, ϕ or μ have been omitted because they only contribute to the normalising constant. As we saw in section 2.1.1 and 9, the conditional densities are proportional to the joint density, at the conditioning values. So we can read off the conditional for $1/\phi$, again ignoring factors that do not contain ϕ (and hence contribute only to the normalising constant):

$$f(1/\phi|\mathbf{x}, \mu) \propto \frac{1}{\phi^{n/2+a-1}} e^{-\sum_i (x_i-\mu)^2/(2\phi) - b/\phi}.$$

If this is to be a p.d.f., then it is recognisable as a $G(n/2 + a, \sum_i (x_i - \mu)^2/2 + b)$ p.d.f.

The conditional for μ is more tedious,

$$\begin{aligned} f(\mu|\mathbf{x}, \phi) &\propto e^{-\sum_i (x_i-\mu)^2/(2\phi) - (\mu-c)^2/(2d)} \\ &\propto e^{-(n\mu^2 - 2\bar{x}n\mu)/(2\phi) - (\mu^2 - 2\mu c)/(2d)} \\ &= e^{-\frac{1}{2\phi d}(dn\mu^2 - 2\bar{x}dn\mu + \phi\mu^2 - 2\mu\phi c)} = e^{-\frac{dn+\phi}{2\phi d}\left(\mu^2 - 2\mu\frac{dn\bar{x}+\phi c}{dn+\phi}\right)} \\ &\propto e^{-\frac{dn+\phi}{2\phi d}\left(\mu - \frac{dn\bar{x}+\phi c}{dn+\phi}\right)^2}, \end{aligned}$$

where terms involving only $\sum_i x_i^2$ and c^2 were absorbed into the normalising constant at the second '$\propto$', and the constant required to complete the square at the final '$\propto$' has been taken from the normalising constant. From the final line, we see that

$$\mu|\mathbf{x}, \phi \sim N\left(\frac{dn\bar{x} + \phi c}{dn + \phi}, \frac{\phi d}{dn + \phi}\right).$$

Now it is easy to code up a Gibbs sampler:

```
n <- 20;set.seed(1);x <- rnorm(n)*2+1 ## simulated data

n.rep <- 10000;
thetag <- matrix(0,2,n.rep)

a <- 1; b <- .1; c <- 0; d <- 100 ## prior constants
```

```
xbar <- mean(x)            ## store mean
thetag[,1] <- c(mu <- 0,phi <- 1) ## initial guesses
for (j in 2:n.rep) { ## the Gibbs sampling loop
  mu <- rnorm(1,mean=(d*n*xbar+phi*c)/(d*n+phi),
                 sd=sqrt(phi*d/(d*n+phi)))
  phi <- 1/rgamma(1,n/2+a,sum((x-mu)^2)/2+b)
  thetag[,j] <- c(mu,phi) ## store results
}
```

The equivalent of Figure 8 is shown in Figure 9. Notice the rather limited effect of the change in prior between the two figures. This is because even the proper priors used for the Gibbs sampler are very vague, providing very limited information on the probable parameter values (plot the $\Gamma(1, .1)$ density to see this), while the data are informative.

9.9 Limitations of Gibbs sampling

Gibbs sampling largely eliminates the difficulty of choosing a good proposal that complicates Metropolis Hastings, but this is not quite the free lunch that it might appear. The catch is that Gibbs sampling produces slowly moving chains if parameters have high posterior correlation, because sampling from the conditionals then produces very small steps. Sometimes updating parameters in blocks or re-parameterising to reduce posterior dependence can then help to improve mixing. The other practical consideration is that if improper priors are used with Gibbs sampling then it is important to check that the posterior is actually proper: it is not always possible to detect impropriety from the output of the sampler.

9.10 Checking for convergence

The great appeal of MCMC methods is their impressive generality. In principle we can work with almost any model, and if we are prepared to simulate for enough iterations, then we will generate a sample from its posterior. The difficulty is in identifying what is enough iterations. Well-designed samplers for relatively benign posteriors may require only several thousand or several hundred thousand iterations. In other situations all the computer power on earth running for the age of the universe might be required to adequately sample from the posterior: for example, when the posterior consists of several well-separated and compact modes in a high-dimensional space, where proposing a move that will take us from one mode to another is all but impossible, let alone doing it often enough to sample from the modes in the correct proportion.

So it is important to check for apparent convergence of MCMC chains. Obvious checks are the sort of plots produced in the left-hand panels of Figures 8 and 9, which give us some visual indication of convergence and how well the chain is mixing. If there is any suspicion that the posterior could be multimodal, then it is sensible to run multiple chains from radically different starting points to check that they appear to be converging to the same distribution. If interest is actually in some scalar valued function of the parameters $h(\boldsymbol{\theta})$, then it makes sense to produce the plots and other diagnostics directly for this quantity.

One step up in sophistication from the simple trace plots is to examine how specific quantiles of the sample, up to iteration j, behave when plotted against j. For example, the following code produces plots that overlay the 0.025, 0.5 and 0.975 quantiles of the sample so far, on top of a simple line plot of the chain's progress:

```
qtplot <- function(theta,n.plot=100,ylab="") { ## simple MCMC chain diagnostic plot
  cuq <- Vectorize(function(n,x) ## cumul. quantile func.
       as.numeric(quantile(x[1:n],c(.025,.5,.975))),
       vectorize.args="n")
  n.rep <- length(theta)
  plot(1:n.rep,theta,col="lightgrey",xlab="iter",
                                   ylab=ylab,type="l")
  iter <- round(seq(1,n.rep,length=n.plot+1)[-1])
  tq <- cuq(iter,theta)
  lines(iter,tq[2,])
  lines(iter,tq[1,],lty=2);lines(iter,tq[3,],lty=2)
}
```

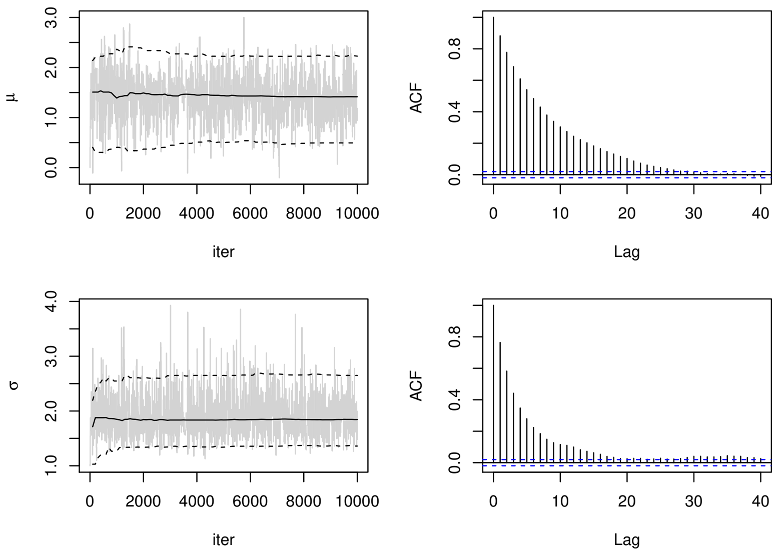

Figure 10: Basic MCMC checking plots, relating to the example from Section 9.5 as discussed in Section 9.10. The left panels show trace plots for the two chain components (grey) overlaid with the evolution of the chain median (continuous), 0.025 and 0.975 quantiles (dashed). Everything is stabilising nicely here. The right-hand panels show the equivalent autocorrelation function plots for the chains, illustrating the substantial degree of autocorrelation when using this sampler, which is higher for the μ component than for the σ component.

A call to `qtplot(theta[1,],ylab=expression(mu))` produces the upper left plot in Figure 10, with a slightly modified call producing the lower left plot. In both cases it appears that the median and both the other quantiles stabilise rapidly.

For slightly more formal checking it helps to have some idea of the *effective sample size* of the chains. Roughly speaking, what size of independent sample from $f(\boldsymbol{\theta}|\mathbf{y})$ would be equivalent to the correlated sample from our MCMC scheme? Or, if we were to retain only every k^{th} sample from the chain, how large would k have to be before we could reasonably treat the resulting thinned sample as approximately independent? To answer these questions it is helpful to examine the autocorrelation function (ACF) of the chain components, as shown in the plots on the right hand side of Figure 10, which are produced by, for example, `acf(theta[1,])` in R. Apparently retaining every 25th μ value and every 20th σ value from the chain would give us almost independent samples.

Actually the `acf` function also returns the estimated correlations at each lag (silently). For example,

```
> mu.ac <- acf(theta[1,])[[1]][,,1];mu.ac
 [1] 1.0000000 0.8831924 0.7777484 0.6863556 0.6096274
 [6] 0.5406225 0.4834136 0.4303171 0.3796966 0.3389512
  .    .          .          .          .          .
```

The *autocorrelation length* associated with a chain is defined as twice the sum of the correlations minus 1. The summation is strictly over all lags up to infinity, but in practice we can sum over the lags up to the point at which autocorrelation appears to have vanished (see R package `coda` for a better method). The corresponding effective sample size is then defined as the sequence length divided by the autocorrelation length. For example,

```
> acl <- 2*sum(mu.ac)-1; acl
[1] 16.39729
> n.rep/acl ## effective sample size
[1] 609.8569
```

So the effective sample size for the μ component is about 600 (although usually we would discard the burn-in period before computing this). Repeating the exercise for σ gives an autocorrelation length of around 10 and an effective sample size of about 1000. For the Gibbs sampler in this simple case the autocorrelation length for μ drops to close to 1.

Armed with this information, more formal tests are possible. For example, given multiple chains, we can subsample to obtain approximately independent samples between each chain and then apply ANOVA methods to see if there appear to be variance components associated with the difference between chains. With a single chain, we might want to divide the apparently converged chain into two parts and formally test whether the two samples appear to come from the same distribution. A two-sample Kolmogorov-Smirnov test is appropriate here, provided that the two samples are of independent draws from the two distributions, so again sub-sampling is needed. Here is an example, for the simple μ chain:

```
> th0 <- theta[1,1001:5500]
> th1 <- theta[1,5501:10000]
> ind <- seq(1,4500,by=16) ## subsampling index
> ks.test(th0[ind],th1[ind])

      Two-sample Kolmogorov-Smirnov test

data: th0[ind] and th1[ind]
D = 0.0745, p-value = 0.4148
alternative hypothesis: two-sided
```

With a p-value of 0.4 there is no evidence for a difference in distribution between the two halves of the chain, so this result provides no reason to doubt convergence. The exact choice of sampling interval is not important: increasing the sampling interval to 25, as might be implied by simply examining the ACF, leads to the same conclusion. However, using a much lower sampling rate is a disaster: failure to subsample at all completely violates the independence assumption of the test, resulting in a computed p-value around 10^{-13}, and even a sampling interval of five results in an erroneously low p-value of 0.016.

This section only scratches the surface of convergence checking. See Robert and Casella (2009, Ch. 8) for more information, and the `coda` package in R for an extensive set of checking functions (Plummer et al., 2006).

9.11 Interval estimation

Given reliable posterior simulations from a chain, interval estimates and quantities for model comparison can be computed. The former is straightforward, because intervals can be based directly on the observed quantiles of the simulated parameters. For example, with the simple toy model of Section 9.8, it is easy to produce 95% *credible intervals* (CIs) for μ and σ, as follows (discarding the first 1000 samples as burn-in):

```
quantile(thetag[1,-(1:1000)],c(0.025,0.975)) ## CI mu
quantile(thetag[2,-(1:1000)]^.5,c(0.025,0.975)) # CI sig
```

which yields $0.52 < \mu < 2.22$ and $1.39 < \sigma < 2.67$.

10 Graphical models and automatic Gibbs sampling

The implementation of MCMC samplers is obviously rather time consuming, and the question of automating the construction of samplers arises. It turns out that automatic Gibbs sampling can be very successfully implemented for *Bayesian graphical models* in which the dependency structure between variables in the model can be represented by a *directed acyclic graph* (DAG). The basic trick is to break down simulation from a high-dimensional posterior into a sequence of Gibbs sampling steps of intrinsically low dimension.

The automation process is bound up with the model's DAG structure, so we need to explore the concepts here. A directed graph consists of a set of *nodes* connected by directed edges: arrows. These arrows run from *parents* to *children*. Every variable in a graphical model is a node, and the key feature of such models is that the distribution of a variable/node is completely known if you know the values of all its parent nodes. The fact that the graphs are *acyclic* means that no node is its own ancestor: you cannot find a path through the graph that follows edges in the direction of the arrows and arrives back at the node you started from.

It is helpful to distinguish three types of node.

1. Stochastic nodes are variables with a distribution that depends stochastically on other nodes. They may be observed (i.e. correspond to data), or unobserved (parameters or random effects).

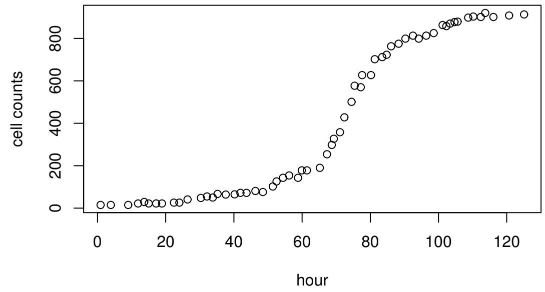

Figure 11: Algal cell counts in samples taken from a laboratory chemostat experiment, against hour.

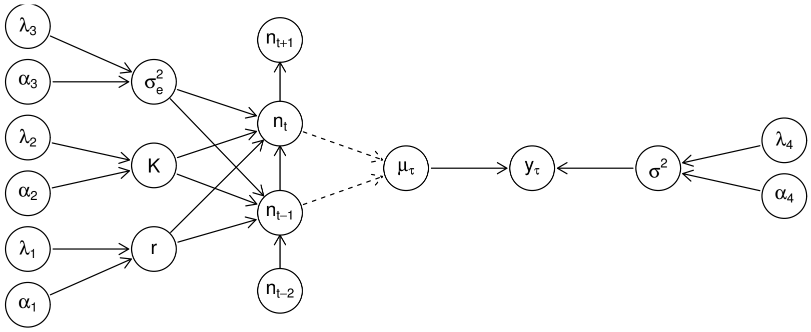

Figure 12: Part of the DAG for the algal model introduced in Section 10.1. The figure focuses on a section of the graph relating to an observation at time τ lying between discrete simulation times t and $t+1$. Most of the edges connected to nodes n_{t-2} and n_{t+1} are not shown. Notice how the left and rightmost nodes, which are fixed constants defining priors, have no parents. Conversely, the observed data node y_t has no children, but this is a feature of this model, rather than being a requirement of the formalism.

2. Deterministic nodes are nodes that are deterministic (logical) functions of other nodes. They cannot be observed.

3. Constants are fixed numbers and have no parents.

There are two types of arrows. Deterministic/logical relationships between nodes are usually shown as dashed arrows, whereas stochastic relationships are shown as solid arrows.

10.1 A non-trivial example

Figure 11 shows counts of algal cells in samples drawn from a laboratory chemostat experiment. A possible model for the population growth in the chemostat is that it follows a self-damping growth model, such as

$$N_{t+1} = e^r N_t e^{-N_t/K + e_t}, \quad e_t \sim N(0, \sigma_e^2), \tag{33}$$

where the independent e_t terms reflect the fact that the population growth will not be fully deterministic. This model would usually operate with a fixed timestep (e.g. t might index hour, or two hour period). If we want to estimate N between N_t and N_{t+1} then we might use linear interpolation. The cell population, y, is then modelled as being a noisy observation of the underlying population N, perhaps Gaussian, with unknown variance, σ^2. The data plotted in Figure 11 are not evenly spaced in time. The spacing ranges from 0.4 to 5.1 hours. Hence we will need to interpolate the solution to (33) in order to use this model for the data.

Figure 12 illustrates a portion of the DAG for this model, assuming that proper gamma(α_j, λ_j) priors have been specified for r, K, $1/\sigma_e^2$ and $1/\sigma^2$. The parentless nodes at the far left and right of the figure are the constants specifying the various gamma priors: actual numbers would have to be supplied here. The portion of the graph shown surrounds a data node y_τ whose time of observation lies between discrete update times t and $t+1$, so that its expected value is obtained by linear interpolation between nodes n_t and n_{t+1}. The fact that this linear interpolation is purely deterministic is the reason that the arrows from n_t and n_{t-1} to deterministic node μ_τ are dashed.

10.2 Exploiting the DAG

Now consider what makes graphical models convenient for automatic Gibbs sampling. Generically, let x_i denote the variable corresponding to the i^{th} node of the graph. From the dependencies encoded in the graph it follows that the joint density of the (non constant) nodes is

$$f(\mathbf{x}) = \prod_i f(x_i|\text{parent}\{x_i\}), \tag{34}$$

where the product is over the non constant nodes. If this is not obvious, start with the childless (terminal) nodes and work back using the basic relations between conditional and joint densities.

Gibbs sampling involves simulating from the full conditionals of all stochastic nodes other than those corresponding to data, which are fixed at their observed values. It turns out that these conditionals usually involve far fewer terms than the full joint density. From the definition of a conditional p.d.f.,

$$f(x_j|\mathbf{x}_{-j}) = \frac{f(\mathbf{x})}{\int f(\mathbf{x})dx_j} = \frac{\prod_i f(x_i|\text{parent}\{x_i\})}{\int \prod_i f(x_i|\text{parent}\{x_i\})dx_j},$$

but the only terms in the product that have to stay inside the integral are those that involve x_j: the conditional density of x_j given its parents, and the conditional densities of each child of x_j, given that child's parents. All other terms in (34) can be taken out of the integral and therefore cancel between the top and bottom of $f(x_j|\mathbf{x}_{-j})$. In short,

$$\begin{aligned} f(x_j|\mathbf{x}_{-j}) &= \frac{f(x_j|\text{parent}\{x_j\})\prod_{i\in\text{child\{j\}}} f(x_i|\text{parent}\{x_i\})}{\int f(x_j|\text{parent}\{x_j\})\prod_{i\in\text{child\{j\}}} f(x_i|\text{parent}\{x_i\})dx_j} \\ &\propto f(x_j|\text{parent}\{x_j\})\prod_{i\in\text{child\{j\}}} f(x_i|\text{parent}\{x_i\}), \end{aligned}$$

so that, however complicated the model and corresponding DAG, $f(x_j|\mathbf{x}_{-j})$, required for the Gibbs update of x_j, depends only on the parent-conditional-densities of x_j and its children.

10.3 Building the samplers

The preceding discussion forms the basis for the automatic construction of Gibbs samplers. The model's DAG structure is used to identify the relatively small number of terms that play a part in each $f(x_j|\mathbf{x}_{-j})$, an attempt is made to identify the exact distribution for $f(x_j|\mathbf{x}_{-j})$, and when this is not possible a more costly general-purpose sampler is constructed. Identification of the exact distributions rests on known *conjugacy* relationships between distributions, while the ingenious method of *slice sampling* is often useful otherwise.

10.3.1 Conjugate distributions

Again consider,

$$f(x_j|\mathbf{x}_{-j}) \propto f(x_j|\text{parent}\{x_j\})\prod_{i\in\text{child\{j\}}} f(x_i|\text{parent}\{x_i\}).$$

The right hand side has exactly the structure of a prior, $f(x_j|\text{parent}\{x_j\})$, for x_j, multiplied by a likelihood term for x_j, in which the children x_i play the role of data. This fact allows what is known about *conjugacy* of distributions to be exploited in automatically ascertaining the density that gives $f(x_j|\mathbf{x}_{-j})$.

If the prior and posterior distribution for some quantity are from the same family[18], for a given likelihood, then that distribution is said to be *conjugate* for that likelihood. We have already seen an example of this when

[18] For example, a normal prior yields a normal posterior, or a gamma prior yields a gamma posterior. Of course, the parameters of the distributions change from prior to posterior.

constructing the simple Gibbs sampler in Section 9.8: there the normal distribution was shown to be conjugate for the mean of a normal likelihood term, while the gamma distribution was conjugate for the precision (reciprocal variance) of a normal. A whole library of such standard conjugacy results is known and can therefore be exploited in the automatic construction of Gibbs samplers.

10.3.2 Slice sampling

When no convenient analytic form for a density $f(x_j|\mathbf{x}_{-j})$ can be obtained, then some other stochastic simulation method must be employed for that density. A beautifully simple approach is *slice sampling* (Neal, 2003). The basic observation is this: if we plot $kf(x)$ against x for any finite non-zero k, and then generate a coordinate x, y from a uniform density over the region bounded by $kf(x)$ and the x axis, then the resulting x value will be a draw from $f(x)$.

The problem, of course, is that generating directly from the required uniform density of x, y is no easier than generating from $f(x)$ itself. The simplicity arises when we consider a Gibbs update of x and y. Trivially

$$f(y|x) \sim U(0, kf(x))$$

while

$$f(x|y) \sim U(x : kf(x) \geq y),$$

so a Gibbs update would draw a y uniformly from the interval $(0, kf(x))$ and then draw x uniformly from the set of x values for which $kf(x) \geq y$ (the 'slice' of the technique's name). The only problem now is identifying the required set of x values. For a unimodal distribution, this set will constitute a single interval, which may be easy to locate, but for multimodal distributions several intervals may need to be identified. Of course, in practice it is only necessary to identify an interval or a set of intervals that bracket the required set: then we can generate uniformly on the bracketing interval(s) until we obtain an x such that $kf(x) \geq y$. If the bracketing interval(s) are too wide this will be inefficient, of course.

10.4 BUGS and JAGS

The most widely used software for statistics via automatically constructed Gibbs samplers is BUGS (Bayesian Updating via Gibbs Sampling), now followed by openBUGS, which has an R package interface `brugs`. BUGS established a simple language for the specification of graphical models, leading to other implementations including JAGS (Just Another Gibbs Sampler), with the R interface package `rjags`, which is covered here.

JAGS has to be installed as a standalone program and can be used as such, but it is very convenient to use it via `rjags`, which is the method covered here. `rjags` also provides easy integration with the `coda` package for convergence diagnostics. JAGS models are specified in a text file using a dialect of the BUGS language. The name of this file, together with a list providing the corresponding data, is supplied to the `jags.model` function, which calls JAGS itself to automatically generate a sampler, returned as an object of class `"jags"`. This object can then be used to generate samples using calls to `jags.samples` or the closely related `coda.samples` (depending on exactly what format you would like the data returned in).

10.4.1 Toy example

Recall the toy normal model example of Section 9.8 in which we have $n = 20$ observations $y_i \sim N(\mu, \phi)$, where $1/\phi \sim G(a, b)$ (a gamma random variable) and (independently) $\mu \sim N(c, d)$. a, b, c and d are constants. In graphical model terms each y_i, μ and $\tau = 1/\phi$ are stochastic nodes, whereas a, b, c and d are constant nodes: the graph has 26 nodes in total. The BUGS language is set up for convenient specification of each node, and their relationships (directed edges). Here is the contents of the file `norm.jags` coding up our toy model

```
model {
 for (i in 1:N) {
   y[i] ~ dnorm(mu,tau)
 }
 mu ~ dnorm(0.0, 0.01)
 tau ~ dgamma(0.05,0.005)
}
```

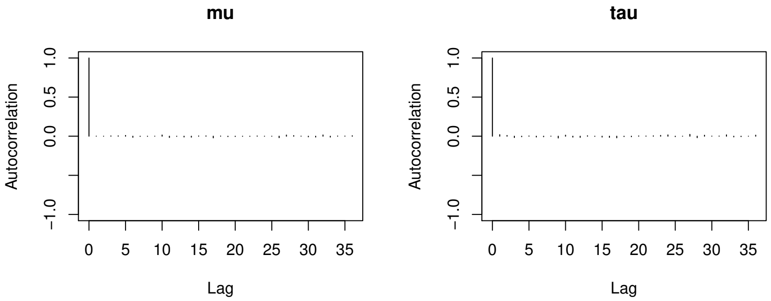

Figure 13: ACFs computed by `autocorr.plot` from the `coda` package for the toy model simulated using JAGS in Section 10.4.1. There is little autocorrelation here: to get more detail on the low correlations the `acfplot` function could be used.

For anyone who has read this far, the language is very intuitive. The symbol ~ specifies a stochastic dependence (i.e. a → in the graph), with the BUGS/JAGS statement `y[i] ~ dnorm(mu,tau)` being exactly equivalent to the mathematical statement $y_i \sim N(\mu, 1/\tau)$. By default the normal distribution is parameterised in terms of its precision, rather than its variance. Notice the use of loops to deal with vectors. R programmers are used to avoiding loops for populating vectors, but here there is no problem: this code is going to be compiled by JAGS to produce a sampler, and the R efficiency concerns do not apply.

Here is the R code to get JAGS to build a sampler from R, given 20 observations in a vector, `y`:

```
library(rjags)
setwd("some/directory/somewhere")
jan <- jags.model("norm.jags",data=list(y=y,N=20))
```

Function `setwd` sets R's working directory to the location of `norm.jags`. `jags.model` then creates the sampler, setting the nodes identified in `data` to their observed values. JAGS counts `N` as a constant node in the model, so it reports that the model has 27 nodes, rather than the 26 counted before.

The `jags.model` function also runs a number of adaptation iterations, tuning those component samplers that can be tuned to try to optimise their performance. The `n.adapt` argument controls the number of adaptation iterations and has a default value of 1000. At this stage `jan` contains a JAGS model object, which is ready to use to generate samples. Once built and initialized the sampler can be used:

```
> um <- jags.samples(jan,c("mu","tau"),n.iter=10000)
  |**************************************************| 100%
```

The second argument, `c("mu","tau")`, specifies the nodes that should be monitored at each step, and `n.iter` gives the number of steps (by setting argument `thin` to an integer greater than 1, we could monitor every `thin` steps). The results are stored in a two element list `um`, which could be used to produce a plot almost identical to Figure 9.

`rjags` can also create output in a manner convenient for use with the MCMC diagnostics package `coda`. Here is some code to do this, plot the chain ACFs shown in Figure 13 and compute effective sample sizes:

```
> er <- coda.samples(jan,c("mu","tau"),n.iter=10000)
  |**************************************************| 100%
> autocorr.plot(er)
> effectiveSize(er)
       mu       tau
 9616.071 10000.000
```

JAGS has a built-in function `dic.samples` for obtaining the *Deviance Information Criterion* (DIC) for a model, which is an AIC like model selection criterion, computable by MCMC (see section 11.1.4). Relative to the standard computation given in section 11.1.4, JAGS uses a slightly different computation for p_D which requires samples from two independent chains (see argument `n.chains` of `jags.model`). For example,

```
jan <- jags.model("norm.jags",data=list(y=y,N=20),
                              n.chains=2)
dic.samples(jan,n.iter=10000)
```

10.5 JAGS algal population example

The algal population growth model of Section 10.1 can easily be implemented in JAGS. The graph in this case has 874 nodes, and a portion of it is shown in Figure 12. The contents of the model specification file are as follows:

```
model {
  n[1] ~ dnorm(n0,tau)
  for (i in 2:M) {
    n[i] ~ dnorm(n[i-1] + r - exp(n[i-1])/K,tau)
  }
  for (i in 1:N) {
    mu[i] <- wm[i]*n[im[i]] + wp[i]*n[ip[i]]
    y[i] ~ dnorm(exp(mu[i]),tau0)
  }
  K ~ dgamma(1.0,.001)
  tau ~ dgamma(1.0,.1)
  r ~ dgamma(1.0,.1)
  n0 ~ dgamma(1.0,.1)
  tau0 ~ dgamma(1.0,.1)
}
```

Notice that there are two loops now. The first iterates the dynamic model for the log population for M steps. The second loop works through the N observed population nodes `y[i]`, relating them to the `n[i]`. The required linear interpolation is implemented via the deterministic nodes `mu[i]`, using pre-computed interpolation indices and weights as generated by the function `lint`, below. The notation '`<-`' is equivalent to the dashed arrows in Figure 12.

The interpolation weight routine `lint` returns vectors `im`, `ip`, `wm` and `wp`, each of the same length as `t`, such that if `N` is the m vector of evenly spaced N values, `N[im] *wm + N[ip] *wp` gives the vector of interpolated N estimates, corresponding to `t`. It also returns `dt` and an appropriate value for m:

```
lint <- function(t,t0=0,dt=1) {
## produce interpolation indices and weights.
  n <- length(t)
  ts <- seq(t0,max(t),by=dt)
  ts <- c(ts,max(ts)+dt)
  m <- length(ts)
  im <- floor((t-t0)/dt)+1;ip <- im+1;ip[ip>m] <- m
  list(im=im,ip=ip,wm=(ts[ip] - t)/dt,
          wp=(t - ts[im])/dt,m=m,dt=dt)
}
```

The R code to use this model is a little more involved for this example, because we need to produce interpolation weights and an initial value for the state vector `n`. Without a reasonable initial value JAGS is unsuccessful at initialising this model.

```
library(rjags)
setwd("~location/of/model/file")
li <- lint(alg$hour,t0=0,dt=4)
dat <- list(y=alg$cell.pop,N=length(alg$cell.pop),M=li$m,
         ip=li$ip,im=li$im,wp=li$wp,wm=li$wm)
## initial state for n by linear interpolation of data...
ni <- log(c(alg$cell.pop[1],approx(alg$hour,alg$cell.pop,
         1:(li$m-2)*li$dt)$y,max(alg$cell.pop)))
jal <- jags.model("algae.jags",data=dat,inits=list(n=ni),
              n.adapt=10000)
```

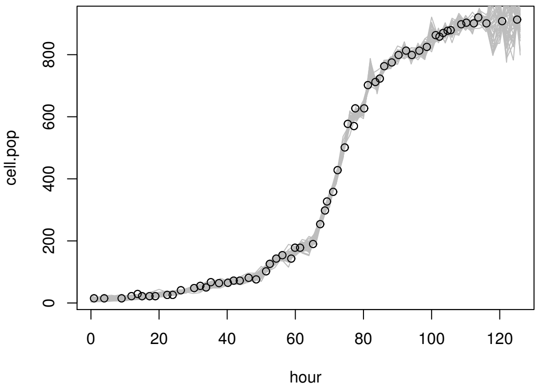

Figure 14: 40 replicates of the underlying state $\exp(n)$ of the algal population model of section 10.5, as simulated using JAGS, shown as grey lines. The observed data are shown as open circles. Clearly the state is too variable towards the end of the data.

```
ug <- coda.samples(jal,c("n0","r","K","tau0"),
              n.iter=40000,thin=10)
```

Here only every 10th sample has been stored in `ug`. According to `coda` the effective sample sizes are 2600, 615, 1514 and 4000 for K, n_0, r and τ_0, respectively, and other diagnostic plots look reasonable. Let us look at the underlying state n, by having JAGS monitor it every 1000 iterations:

```
pop <- jags.samples(jal,c("n"),n.iter=40000,thin=1000)
plot(alg)
ts <- 0:(li$m-1)*li$dt
for (i in 1:40) lines(ts,exp(pop$n[,i,1]),col="grey")
with(alg,points(hour,cell.pop))
```

The results are shown in Figure 14: clearly the state is too variable at high population sizes, and the model would benefit from some modification.

11 More Bayesian Inference

We have already seen how to obtain interval estimates under the Bayesian approach. Under the Bayesian paradigm we do not estimate parameters: rather we compute a whole distribution for the parameters given the data. Even so, we can still pose the question of which parameters are most consistent with the data. A reasonable answer is that it is the most probable value of $\boldsymbol{\theta}$ according to the posterior: the *posterior mode*,

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}} f(\boldsymbol{\theta}|\mathbf{y}).$$

More formally, we might specify a loss function quantifying the loss associated with a particular $\hat{\boldsymbol{\theta}}$ and use the minimiser of this over the posterior distribution as the estimate. This *decision theory* based approach has considerable philosophical appeal, and leads to some very elegant theory, but it can also lead us to parameter 'estimates' that are in the far tails of the posterior probability distribution, if we are not very careful about how the loss is specified (especially in high dimensions).

If we specify an improper uniform prior $f(\boldsymbol{\theta}) = k$, then $f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})$ and the posterior modes are exactly the maximum likelihood estimates (given that $f(\mathbf{y})$ does not depend on $\boldsymbol{\theta}$). In fact, for data that are informative about a fixed dimension parameter vector $\boldsymbol{\theta}$, then as the sample size tends to infinity the posterior modes tend to the maximum likelihood estimates in any case, because the prior is then dominated by the likelihood.

The connection to MLE goes further. In many circumstances in the large sample limit when the likelihood dominates the prior,

$$\boldsymbol{\theta}|\mathbf{y} \sim N(\hat{\boldsymbol{\theta}}, \boldsymbol{\mathcal{I}}^{-1}),$$

where $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\mathcal{I}}$ are as in (8), as we saw in the case of the linear model. However there are real differences when it comes to model selection.

11.1 Model comparison, Bayes factors, prior sensitivity, BIC, DIC

Hypothesis testing does not fit easily with the Bayesian approach, and a criterion somehow similar to AIC is also not straightforward. The obvious approach to Bayesian model selection is to include all possible models in the analysis and then to compute the marginal posterior probability for each model (e.g. Green, 1995). This sounds clean, but it turns out that those probabilities are sensitive to the priors put on the model parameters, which is problematic when these are 'priors of convenience' rather than well-founded representations of prior knowledge. Computing such probabilities is also not easy. This section examines the issue of sensitivity to priors and then covers two of the attempts to come up with a Bayesian equivalent to AIC.

11.1.1 Marginal likelihood, the Bayes factor and sensitivity to priors

In the Bayesian framework the goal of summarising the evidence for or against two alternative models can be achieved by the *Bayes factor* (which therefore plays a somewhat similar role to frequentist p-values). A natural way to compare two models, M_1 and M_0, is via the ratio of their probabilities.[19] As a consequence of Bayes theorem, the prior probability ratio transforms to the posterior probability ratio as

$$\frac{\Pr(M_1|\mathbf{y})}{\Pr(M_0|\mathbf{y})} = \frac{f(\mathbf{y}|M_1)\Pr(M_1)}{f(\mathbf{y}|M_0)\Pr(M_0)} = B_{10}\frac{\Pr(M_1)}{\Pr(M_0)},$$

by definition of B_{10}, which is known as the Bayes factor for comparing M_1 with M_0. So B_{10} measures the amount by which the data have shifted the prior probability ratio in favour of M_1. As with p-values, there are conventions about how to describe the degree of evidence that different magnitudes of Bayes factor represent (Kass and Raftery, 1995). Working with $2 \log B_{10}$ (for comparability with the log likelihood ratio) we have

$2 \log B_{10}$	Evidence against M_0
$0 - 2$	Barely worth mentioning
$2 - 6$	Positive
$6 - 10$	Strong
> 10	Very strong

To actually compute B_{10} we need to obtain the marginal density of $\mathbf{y}$ given each model, also known as the *marginal likelihood*. For example,

$$f(\mathbf{y}|M_1) = \int f(\mathbf{y}|\boldsymbol{\theta}_1)f(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1, \tag{35}$$

where $\boldsymbol{\theta}_1$ denotes the parameters of M_1. The need to integrate over all possible parameter values is a major difference between the Bayes factor and the likelihood ratio statistic for model comparison, but integration gives the Bayes factor some advantage. The likelihood ratio statistic is a ratio of maximised likelihoods evaluated at both models' best fit parameters, giving the larger model an inevitable advantage, which we then have to allow for in interpreting the ratio; hence the need for p-values or AIC. By integrating over all possible parameter values, the marginal likelihood does not suffer from this bias towards large models — irrelevant flexibility can decrease the marginal likelihood. Computing (35) is generally not straightforward. Since part of the motivation for MCMC is to avoid having to evaluate the marginal likelihood, it is unsurprising that it requires substantially more effort to get marginal likelihood estimates by simulation. The alternative is to use approximations to the posterior. However, there is also a more fundamental problem to consider.

[19] If M_1 and M_0 are the only two possibilities then the probability ratio is also the *odds* of M_1; that is, the probability of M_1 over the probability of not M_1.

Examination of (35) indicates an immediate problem with the use of vague, uninformative or improper priors (i.e. with any prior that is chosen to represent a broad statement of ignorance, rather than a precise characterisation of prior knowledge). The difficulty is that the value of (35) is very sensitive to the prior. It is easy to see the problem by example. Suppose the likelihood indicates that a single parameter θ is almost certain to lie in the interval $(0, 1)$, but because we had no real prior information on θ, we used a $U(-100, 100)$ prior, obtaining a value for the marginal likelihood of k. Now suppose that we replace the prior with $U(-1000, 1000)$. This produces a negligible change in the posterior for θ, but reduces the marginal likelihood to approximately $k/10$ (corresponding to a 'positive' change in the Bayes factor in the earlier table). If we had used an improper prior then it would only have been possible to compute the marginal likelihood to within an arbitrary multiplicative constant, rendering the Bayes factor completely useless unless the same improper priors apply to both models.

Another way of seeing the problem with the marginal likelihood is to recognise that it is the likelihood for the fixed parameters of the prior (that is, for those parameter values we chose during model specification), all other parameters having been integrated out. Choosing between models on the basis of the relative likelihood of the fixed parameters of the prior is not always a natural approach. Indeed it is completely arbitrary if those values were selected merely to be as uninformative as possible about $\boldsymbol{\theta}$.

In summary, because the prior is inescapably part of the model in the Bayesian approach, marginal likelihoods, Bayes factors and posterior model probabilities are inescapably sensitive to the choice of prior. In consequence, it is only when those priors that differ between alternative models are really precise and meaningful representations of prior knowledge that we can justify using Bayes factors and posterior model probabilities for model selection. Even then, the computation of the marginal likelihood is often difficult. These difficulties are part of the motivation for attempting to produce AIC-like model selection criteria in the Bayesian setting. But before looking at these, let us consider fixing the Bayes Factor.

11.1.2 Intrinsic, fractional and partial Bayes factors

Given that Bayes factors require meaningfully informative priors, which are often not available at the model formulation stage, it is worth considering the alternative of using part of the data to generate priors. The basic idea is to split the data $\mathbf{y}$ into two parts $\mathbf{x}$ and $\mathbf{z}$, and to use $f(\boldsymbol{\theta}|\mathbf{x})$ as the prior for computing the marginal likelihood based on $\mathbf{z}$. That is, marginal likelihoods of the form

$$f(\mathbf{z}|M_i, \mathbf{x}) = \int f(\mathbf{z}|\boldsymbol{\theta}_i, \mathbf{x}) f(\boldsymbol{\theta}_i|\mathbf{x}) d\boldsymbol{\theta}_i$$

are used to form a sort of partial Bayes factor. Substitute $f(\boldsymbol{\theta}_i|\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\theta}_i)f(\boldsymbol{\theta}_i)/\int f(\mathbf{x}|\boldsymbol{\theta}_i)f(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$ into the preceding expression to see why this improves matters,

$$\begin{aligned} f(\mathbf{z}|M_i, \mathbf{x}) &= \frac{\int f(\mathbf{z}|\boldsymbol{\theta}_i, \mathbf{x}) f(\mathbf{x}|\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i}{\int f(\mathbf{x}|\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i} \\ &= \frac{\int f(\mathbf{y}|\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i}{\int f(\mathbf{x}|\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i}. \qquad (36) \end{aligned}$$

Since we have the same prior on the top and bottom of this last expression, the sensitivity to priors seen in the full marginal likelihood is much reduced.

Two variants of this basic idea are *intrinsic* and *fractional* Bayes factors. Intrinsic Bayes factors (Berger and Pericchi, 1996) use a subset $\mathbf{x}$ just large enough to ensure that $f(\boldsymbol{\theta}_i|\mathbf{x})$ is proper, and then average the resulting partial Bayes factors over all such subsets to remove the arbitrariness attendant on any particular choice of $\mathbf{x}$. The required averaging can be somewhat computationally costly. Hence *fractional* Bayes factors (O'Hagan, 1995) use the fact that if $b = \dim(\mathbf{x})/\dim(\mathbf{y})$, then $f(\mathbf{x}|\boldsymbol{\theta}_i) \approx f(\mathbf{y}|\boldsymbol{\theta}_i)^b$ (at least for large dimensions and exchangeable observations) and this approximation can be plugged into (36). Note that if the fractional Bayes factors are to select the right model in the large sample limit, then $b \to 0$ as the sample size tends to infinity. This consistency is automatic for intrinsic Bayes factors.

11.1.3 BIC: the Bayesian information criterion

An older approach, avoiding the difficulties of sensitivity to priors, is due to Schwarz (1978). Dropping the notation relating to particular models in the interests of clarity, the computation of the Bayes factor requires the evaluation of the marginal likelihood,

$$P = \int f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta},$$

for each model. Let n be the dimension of $\mathbf{y}$ and p be the dimension of $\boldsymbol{\theta}$, and define $f_p(\boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$ ($\mathbf{y}$ is the observed data vector here). Let $\tilde{\boldsymbol{\theta}}$ be the value of $\boldsymbol{\theta}$ maximising f_p. A Taylor expansion gives

$$\begin{aligned}
\log f_p(\boldsymbol{\theta}) &\simeq \log f_p(\tilde{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathrm{T}} \left(-\frac{\partial^2 \log f_p}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \right) (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \\
\Rightarrow f_p(\boldsymbol{\theta}) &\simeq f_p(\tilde{\boldsymbol{\theta}}) \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathrm{T}} \left(-\frac{\partial^2 \log f_p}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \right) (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \right\}.
\end{aligned}$$

Recognising the term in $\{\}$ from the multivariate normal p.d.f., we have

$$P = \int f_p(\boldsymbol{\theta})d\boldsymbol{\theta} \simeq f_p(\tilde{\boldsymbol{\theta}})(2\pi)^{p/2} \left| -\frac{\partial^2 \log f_p}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \right|^{-1/2}.$$

Now assume, at least in the $n \to \infty$ limit, that $-\partial^2 \log f_p / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}} = n\boldsymbol{\mathcal{I}}_0$, where $\boldsymbol{\mathcal{I}}_0$ is a matrix such that $|\boldsymbol{\mathcal{I}}_0|$, is bounded above and below by finite positive constants, independent of n (and ideally close to 1). In the case of i.i.d. data, then $\boldsymbol{\mathcal{I}}_0$ is the (fixed) information matrix for a single observation. Under this assumption we have

$$\left| -\frac{\partial^2 \log f_p}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \right| = n^p |\boldsymbol{\mathcal{I}}_0|$$

and so

$$\log P \simeq \log f(\mathbf{y}|\tilde{\boldsymbol{\theta}}) + \log f(\tilde{\boldsymbol{\theta}}) + \frac{p}{2}\log(2\pi) - \frac{p}{2}\log n - \frac{1}{2}\log|\boldsymbol{\mathcal{I}}_0|.$$

Now as $n \to \infty$, $\tilde{\boldsymbol{\theta}} \to \hat{\boldsymbol{\theta}}$ (the MLE) while the terms that do not depend on n become negligible compared to those that do. So we arrive at

$$\mathrm{BIC} = -2\log f(\mathbf{y}|\hat{\boldsymbol{\theta}}) + p\log n \ \ (\approx -2\log P).$$

Hence the difference in BIC between two models is a crude approximation to twice the log Bayes factor, and all other things being equal, we would select the model with the lowest BIC. Notice some arbitrariness here: there is really nothing in the preceding derivation to stop us from multiplying n in BIC by the finite positive constant of our choice. One interesting feature of BIC is that because it drops the prior, it is not susceptible to the problem with sensitivity to priors that affects the Bayes factor itself, but on the other hand the justification for dropping the prior seems somewhat artificial.

11.1.4 DIC: the deviance information criterion

In complex Bayesian models it is not always clear how to count the number of free parameters in the model. For example, the distinction between random effects and parameters in models is really terminological rather than formal in the Bayesian setting. This makes application of BIC problematic, as does BIC's dependence on knowledge of the posterior modes, which are not directly available via simulation.

In essence, the problem with counting free parameters is the introduction of priors. A prior restricts the freedom of a parameter to vary. In the limit in which the prior was a Dirac delta function, then the corresponding parameter would behave as a fixed constant, insensitive to the data, and should clearly not count in the tally of free parameters at all. Moving smoothly from this extreme to the other extreme of a fully uninformative prior, it seems reasonable that the contribution of the corresponding parameter to the free parameter count should increase smoothly from 0 to 1. This idea leads us to the notion of *effective degrees of freedom*.

Spiegelhalter et al. (2002) suggest a measure of the effective degrees of freedom of a Bayesian model that is readily computed from simulation output. They first define the *deviance* as

$$D(\boldsymbol{\theta}) = -2\log f(\mathbf{y}|\boldsymbol{\theta}) + c,$$

where c is a neglectable constant depending only on the $\mathbf{y}$ (and hence not varying between models of a given $\mathbf{y}$). Using the notation $\bar{x}$ for 'mean of x', the proposed definition of the effective degrees of freedom is

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}).$$

The definition is appealing because in the large sample limit in which the likelihood dominates the prior and the posteriors are approximately Gaussian, then $D(\boldsymbol{\theta}) - D\{E(\boldsymbol{\theta})\} \sim \chi^2_r$, by the same argument that leads to (9). But p_D is a direct estimate of $E[D(\boldsymbol{\theta}) - D\{E(\boldsymbol{\theta})\}]$, and $E(\chi^2_r) = r$. The deviance information criterion is then

$$\text{DIC} = D(\bar{\boldsymbol{\theta}}) + 2p_D,$$

which clearly has a somewhat similar form to AIC. Derivation of DIC in the context of approximately Gaussian posteriors is relatively straightforward, but it is applied much more widely, with some associated controversy. In any case the DIC cannot be justified if p_D is not small relative to the number of data in $\mathbf{y}$. It has the pragmatic appeal of being readily computable from simulation output, while being much less sensitive to the choice of vague priors than the marginal likelihood.

References

Berger, J. O. and L. R. Pericchi (1996). The intrinsic Bayes factor for model selection and prediction. *Journal Of The American Statistical Association 91*(433), 109–122.

Casella, G. and R. Berger (1990). *Statistical inference*. Belmont, CA: Duxbury Press.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. London: Chapman & Hall.

Davison, A. C. (2003). *Statistical Models*. Cambridge: Cambridge University Press.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*(4), 711–732.

Kass, R. and A. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Klein, J. and M. Moeschberger (2003). *Survival Analysis: Techniques for Censored and Truncated Data* (2nd ed.). New York: Springer.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics 31*, 705–767.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 99–138.

Plummer, M., N. Best, K. Cowles, and K. Vines (2006). coda: Convergence diagnosis and output analysis for MCMC. *R News 6*(1), 7–11.

Robert, C. and G. Casella (2009). *Introducing Monte Carlo methods with R*. New York: Springer.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability 7*(1), 110–120.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

Silvey, S. D. (1970). *Statistical Inference*. London: Chapman & Hall.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B 64*(4), 583–639.