LECTURE NOTES

for short course

ADVANCED ANALYTICAL METHODS FOR CLIMATE RESEARCH

given at

INSTITUTE OF ATMOSPHERIC PHYSICS
BEIJING
12TH–14TH JUNE 2001

by

Dr Richard E. Chandler
Department of Statistical Science
University College London
Gower Street
London WC1E 6BT
ENGLAND

Contents

APPENDICES

Acknowledgements

These lecture notes would not exist, unless I had been invited to give a short lecture series at the Institute of Atmospheric Physics in Beijing. First and foremost, therefore, thanks are due to Zhongwei Yan for instigating the visit. I am also grateful to the IAP for hosting the lectures; in particular, to Professor Renhe Zhang for the invitation, and to staff at the Temperate East Asia Regional Research Centre and the State Key Laboratory of Numerical Simulation for Atmospheric Sciences and Geophysical Fluid Dynamics. Special mention must be made of Xiaobo Ren, for dealing with much of the administration, and setting up the web site for the lectures.

Various people have helped with the preparation of these lectures, by providing data and climatological advice. The North-West Pacific tropical storm numbers, used in the first case study in Lecture 2, were provided Jenni Collins of the Mullard Space Science Laboratory, UCL. The Niño 3 indices used in this study were obtained from the IRI/LDEO Climate Data Library[1]. The windspeed data used in the fourth case study were obtained from the Royal Netherlands Meteorological Institute (KNMI). Many thanks are due to Job Verkaik and Ilja Smits, for all their hard work in preparing this dataset.

Finally, I would like to apologise to the audience, for the fact that these lecture notes are in English. Wǒ Zhōngguóhuà shuō de bù hǎo!

Richard Chandler
June 2001

[1] `http://ingrid.ldgo.columbia.edu/`

Lecture 1

Probability and statistical modelling

1.1 Motivation

Climate researchers face two fundamental problems in the course of their work. The first is the complexity of the climate system, and the second is the difficulty in obtaining reliable climate measurements. Consequently, all results in climate research have some degree of uncertainty attached to them. This has implications for all who use the results, whether they are decision makers developing new policies or scientists seeking to develop further their understanding of the climate. It is therefore useful to develop scientific methods that include some recognition of uncertainty.

The aim of these lectures is to introduce statistical methods that deal with uncertainty by using probability models. In this lecture, we provide some background material, and give an accessible overview of the relevant theory. Subsequent lectures introduce, and illustrate, the use of Generalized Linear Models (GLMs) in climate research, as well as briefly discussing other types of model. The reason for focusing on GLMs is that they are well established in statistics, and are extremely flexible and powerful. Most good statistical software packages have the facility for fitting GLMs. We will illustrate the application of GLMs using the free package R (for details on how to obtain this package, see Appendix A.1).

1.1.1 Examples of problems in climate research

Typically, the aims of any climate investigation fall into one or more of the following categories:

1. To seek an understanding of climatic processes, by studying the relationships between variables. For example, we may wish to determine whether the state of ENSO can be related to rainfall patterns in a particular part of the world. The presence or absence of such a relationship may suggest mechanisms that can be used, for example, to improve the performance of dynamical or physical climate models.

2. To examine the evidence for changes in climate regimes, and to determine the nature and

extent of any such changes. Of particular interest at present is the detection of signals that may be attributed to anthropogenic effects. However, such signals are difficult to detect and even harder to interpret. There are two basic reasons for this:

(a) Climate varies naturally on many different timescales. It is difficult to tell whether any apparent trends over, say, the last 30, 50 or 100 years are associated with human activity, or whether they can be explained in terms of natural variability.

(b) The climate system involves complex interactions. Typically, changes will be different in different locations, and at different times of year. For example, in North-Western Europe it is generally accepted that the late 20th century has seen a trend towards wetter winters and drier summers. Analyses of annual rainfall here, then, may not reveal any trends because the two effects cancel each other out. This is a simple example, which could be studied straightforwardly by analysing data from ‘winter’ and ‘summer’ separately. However, not all situations are as obvious as this, and there is a danger of misinterpretation if the potential complexity of a system is not recognised.

3. To make some useful statements about future climate. Under this heading, we include short-term and seasonal forecasting, climate change impacts studies and risk assessment from extreme events.

Later in the lectures, we will consider four specific case studies, each of which can be regarded as falling into one or more of these categories.

1.1.2 Uncertainty, and the need to confront it

In all of the situations outlined above, some uncertainty is involved. In an application such as weather forecasting, this arises because (i) our numerical models are incomplete (ii) our data are subject to measurement error. It may be argued that as we develop more powerful computers, more accurate models and better quality data, such uncertainty will be reduced and we will be able to generate better and better forecasts. This may be true, but ultimately there is a limit to what can be achieved — this has been known since the development of quantum theory in the mid-1920s. We may ask (i) is this important? and (ii) if so, is it possible to quantify the uncertainty in a forecast?

In response to the first of these, we may answer as follows: *it is important to know whether it is important*! If a forecast has a ‘small’ error then there is little uncertainty and we may choose to ignore it (exactly how we define ‘small’ here will depend on the application). If, on the other hand, the error in a forecast is ‘large’ then we may have to account for it. Of course, we do not know in advance what the forecast error will be, and therefore the best we can do is to give an indication of its likely magnitude.

This leads us to the second question above: is it possible to quantify the uncertainty in a forecast? To address this, consider first that a perfect forecast requires perfect measurements of all factors influencing the climate system (as well as a perfect representation of the system itself). In reality, it is not possible to observe all of the relevant factors — and those that are observed are subject to measurement error. As a result, a forecaster has only partial knowledge of the true state of the system when the forecast is issued, and is unable to distinguish between the many different 'true' states which are consistent with this partial knowledge. The forecast for each of these states will be the same, but the actual outcomes will be different. We can imagine collecting together all of these possible outcomes and studying them. If we did this, we could make some useful quantitative statements such as '90% of the actual values were within 5% of the forecast value'. Such a statement gives a useful indication of the accuracy of a forecast, or equivalently of the uncertainty associated with it. The point of this argument is that we have just made a probability statement. However, we have at no stage claimed that climate is in any way 'random'. We return to this in the next section.

Although the above example deals with weather forecasting, the fundamental points — that we need to know about uncertainty, and that simple quantitative statements can help — are valid for all applications. Unfortunately, in many areas of science the importance of these points is often not recognised. There are a number of reasons for this — largely due to the way in which the human mind works. For example, a scientist will often have an instinctive feeling that his or her results are 'more or less' right, and therefore that uncertainty can be neglected.

Example 1.1: Many complex models, such as GCMs, involve large numbers of parameters. In such models, individual parameters may be identified very accurately. However, there will always be some error. The cumulative effect of such errors can be much larger than expected. This is partly because of the psychological effect of breaking down complex models into smaller modules — usually, it is not possible to comprehend the workings of a complex model in sufficient detail to understand the way in which errors cumulate. Additionally, there is a tendency for the human mind to be optimistic, and to imagine (incorrectly) that errors will compensate for each other i.e. that overestimation in one part of a model will be balanced by underestimation in another.

In extreme cases, it may be that several complex models are joined together, and that the cumulative uncertainty is so large that the combined model is effectively useless. To consider a hypothetical example: a hydrologist may be interested in determining future river levels from GCM output. One approach to this may be to take the precipitation output from a GCM (model 1), apply a downscaling procedure (model 2), and input the downscaled precipitation to a rainfall runoff model (model 3) to determine future river levels. Each of these models may, in itself, provide reasonable output, but this does not guarantee that the combined modelling system will be useful.

It is tempting to think that such an attitude is unproductive. If we use the best available models for a system, surely their output represents our best attempt at understanding the system, and this can only be useful? Well, yes and no. It can be useful if we really do have some understanding of the system. However, if the cumulative uncertainty in a model is very large, we must accept that this understanding does not exist. In this case, there is a real risk of making costly wrong decisions if we act on the basis of model output. *It can be extremely useful to know that we don't know* *anything!* ■

Another common scenario is that a scientist recognises that a system is too complex to be modelled accurately; in this case there is a temptation to think that representation of uncertainty is ‘too difficult’, since all of the system’s complexities need to be taken into account when calculating the likely magnitude of an error.

A final problem that arises when considering uncertainty is that its effects are often unexpected, or at least counter-intuitive — often, people recognise that uncertainty is present, but do not correctly understand its implications (this presumably explains why so much money is spent across the world on gambling and lotteries!). A simple example of a ‘counter-intuitive’ result is the following:

Example 1.2: Suppose that, at some time t, a particle in a simple system is travelling at speed V_t metres per second. By considering the dominant forces acting on the particle, it is possible to write down its equations of motion. If all necessary observations are available at time t, these equations may be solved to to forecast the particle’s speed at some future time, $t + \ell$ say. Call this forecast $\hat{V}_{t+\ell}$. Typically, the actual speed at time $V_{t+\ell}$ will not be equal to $\hat{V}_{t+\ell}$, because of approximations in the equations. However, in a simple system we expect the error $V_{t+\ell} - \hat{V}_{t+\ell}$ to be ‘small’. Moreover, if we repeat the exercise many times, and compute the average of all the errors, we would expect the average error to be zero.

Now consider what happens if we wish to use our forecast of $V_{t+\ell}$ to obtain a forecast of the particle’s kinetic energy, $\frac{1}{2}V_{t+\ell}^2$. The natural forecast of this quantity is just $\frac{1}{2}\hat{V}_{t+\ell}^2$. Intuitively, it seems reasonable to expect that on average, the resulting forecast errors will be zero. In fact this is incorrect — it can be shown that such a scheme will, on average, overestimate the kinetic energy, even though the average error in the forecast speed is zero. The magnitude of the overestimation increases with the uncertainty in the speed forecasts.

This example is a special case of a more general phenomenon. The surprising result occurs because the initial error in $\hat{V}_{t+\ell}$ is transformed in a nonlinear way. The key point to note here, apart from the unexpectedness of the result, is that in order to forecast kinetic energy *on average*, we need to know about the uncertainty in the forecast of speed. ■

1.1.3 The role of probability

In the previous section, we expressed the uncertainty in a weather forecast using a statement of the form ‘90% of the time, the actual value will be within 5% of the forecast value’. This statement takes an EVENT (in this case, that the actual value will be within 5% of the forecast value) and assigns to it a number (90). This number is interpreted as the percentage of time that the event will occur, in the long run. Equivalently, we could allocate a number between 0 and 1 representing the *proportion* of time that the event will occur. For present purposes, this defines the PROBABILITY of the event. Here, the probability that the actual value will be within 5% of the forecast value is 0.9.

Probability statements are often interpreted as though they relate to 'random' phenomena. The exact meaning of the word 'random' in this context is not clear, although most people would agree that it implies a lack of ordered structure at some level. This perception is unfortunate, and incorrect. In the previous section, we made a probability statement for a deterministic system. The statement was meaningful, because the system could not be observed completely (this sounds like quantum theory again!). Probability statements *are* meaningful for such systems, and provide a simple way of expressing uncertainty. Indeed, probability may be thought of as the language of uncertainty.

In practice, of course, it is necessary to develop techniques that enable us to make probability statements on the basis of observations. The modern discipline of STATISTICAL SCIENCE (usually referred to just as STATISTICS) is largely concerned with the development of these techniques. Statistics in climate research is often perceived to be about 'analysing data'. To some extent, this is true, but it is more than that. Most professional statisticians would agree that statistics is about *interpreting information*. Since there is usually uncertainty associated with any such interpretation, it is inevitable that probability finds its way, in some form or another, into most modern statistical methods.

In order to appreciate some of these methods, and the way in which they can be used to interpret information, it is necessary to summarise a few theoretical results. These provide the background and justification for some of the material to be discussed in the remaining lectures. We do not intend to give too much technical detail; rather, to give a broad overview of some of the key ideas.

1.2 Overview of probability theory

1.2.1 Probability as a relative frequency

In probability theory, we speak of an EXPERIMENT as any process that can result in a number of possible OUTCOMES. An EVENT is just a collection of these outcomes.

Example 1.3: Suppose we wish to determine whether it will rain tomorrow in Beijing, on the basis of information available today. The evolution of the weather between today and tomorrow is, in probabilistic terms, an experiment, since it is a process under which many different possible scenarios, or outcomes, are possible. Some of these outcomes will result in rain tomorrow in Beijing; others will not. Therefore the event 'it rains tomorrow in Beijing' can be regarded as a collection of different outcomes of the experiment. ■

Formally, PROBABILITY is defined as an allocation of numbers, between 0 and 1, to events. The allocation must satisfy certain requirements; however, these do not concern us here. Of more interest is the interpretation of probabilities. The 'classical' interpretation is that the probability of an event is the proportion of times it would occur in a long sequence of repetitions of the experiment, under identical conditions. In climate research, it is not immediately clear that this is useful. In Example 1.3 above, it seems that the experiment (i.e. the evolution of the weather between today and tomorrow) cannot be repeated a large number of times. However, in this case 'repetitions' can be obtained from all days for which the synoptic conditions are identical — or, at least, very close — to today's. If the probability of rain tomorrow in Beijing is 0.9, this can be interpreted as saying '90% of days like today are followed by days during which it rains in Beijing'.

This way of understanding probabilities is called the RELATIVE FREQUENCY interpretation, and is the view that we will take for most of these lectures. However, it is not the only way in which probability statements may be interpreted. An alternative view is mentioned briefly in Lecture 3.

1.2.2 Some important results, and notation

It is convenient to introduce some mathematical notation in this section. We will denote events by $A_1, A_2, \ldots$. The probability of an event A_i is denoted by $P(A_i)$.

Often, we are interested in finding the probability of an event, A_1 say, when we know that another event A_2 has occurred. This is referred to as the CONDITIONAL PROBABILITY OF A_1 GIVEN A_2, and denoted by $P(A_1|A_2)$.

Example 1.4: Suppose that in Beijing, on average it rains on 20% of days. However, at the peak of the monsoon season it rains on 60% of days. We can express this in probability notation as follows: let A_1 denote the event 'it rains today in Beijing', and A_2 be the event 'today, we are in the peak of the monsoon season'. Then $P(A_1) = 0.2$, and $P(A_1|A_2) = 0.6$. ■

This example shows, very simply, how conditional probability may be used to to study relationships between events. Our understanding of the climate system tells us that rain is more likely during the monsoon. Because of this, the conditional probability of rain during the monsoon differs from the UNCONDITIONAL PROBABILITY (0.2 in this example).

Definition: Formally, $P(A_1|A_2)$ is defined as

$$\frac{P(A_1 \text{ and } A_2)}{P(A_2)} .$$

Here, when we write $P(A_1 \text{ and } A_2)$, we mean that the events A_1 and A_2 both occur.

A useful way to think of conditional probability is this: if we know A_2 has occurred, we have effectively changed the experiment. In Example 1.4 above, we can think in terms of an experiment 'pick any day, and see if it rains in Beijing'. In this case, the probability of rain is 0.2. If we define an alternative experiment: 'pick any day during the peak of the monsoon season, and see if it rains in Beijing', then the probability of rain is 0.6.

If the occurrence of A_2 tells us nothing about A_1, then $P\left(A_1|A_2\right) = P\left(A_1\right)$. In this case, events A_1 and A_2 are said to be *independent*. A rearrangement of the formal definition of conditional probability tells us that if A_1 and A_2 are independent, $P\left(A_1 \text{ and } A_2\right) = P\left(A_1\right) \times P\left(A_2\right)$.

Some key results concerning conditional probabilities are as follows:

1. **Bayes' Theorem:** Let A_1 and A_2 be events with $P(A_2) > 0$. Then

$$P\left(A_1|A_2\right) = \frac{P\left(A_2|A_1\right) P\left(A_1\right)}{P(A_2)} \; .$$

In fact, this is a simplified statement of the theorem, but it is adequate for our purposes. We will return to this result in Lecture 3.

2. **Generalised Multiplication Law:** For arbitrary events $A_1, \ldots, A_n$ such that $P\left(A_1 \text{ and } A_2, \ldots \text{and } A_{n-1}\right) > 0$,

$$\begin{aligned} P\left(A_1 \text{ and } A_2, \ldots \text{and } A_n\right) &= P\left(A_1\right) \times P\left(A_2|A_1\right) \times P\left(A_3|A_1 \text{ and } A_2\right) \times \ldots \\ &\quad \times P\left(A_n|A_1 \text{ and } A_2, \ldots \text{and } A_{n-1}\right) \; . \end{aligned}$$

This result is essentially an extension, and rearrangement, of the formal definition of conditional probability given above. It will be used during the discussion of likelihood theory in Section 1.4.3 below. For the moment a simple example will suffice:

Example 1.5: In the West of Ireland, the probability that any day will experience rainfall is 0.7. However, if rainfall was experienced on the preceding day this probability increases to 0.85. What is the probability that rain falls on two successive days in the West of Ireland?

Solution: Let A_1 be the event 'first day is wet', and A_2 be the event 'second day is wet'. Then $P(A_1) = 0.7$, and $P(A_2|A_1) = 0.85$. Hence

$$P\left(A_1 \text{ and } A_2\right) = P(A_1) \times P(A_2|A_1) = 0.7 \times 0.85 = 0.595 \; .$$

Hence, if we take all pairs of successive days in Ireland, approximately 60% of these pairs will experience rain on both days. ■

1.2.3 Random variables

Most of the time, we are not so much interested in answering questions such as ‘Will it rain today?’ as in studying numerical quantities (e.g. the amount of rainfall, or the number of typhoons in the North-West Pacific in a particular season). In the framework of probability theory, this may be dealt with by allocating numbers to the outcomes of an experiment. Any rule for allocating numbers to outcomes is called a RANDOM VARIABLE (note that ‘random’ here is used as a formal mathematical term, not in the sense discussed in Section 1.1.3). In these lectures, the letter Y denotes a random variable.

To summarise:

- An EXPERIMENT is a process that can result in a number of possible OUTCOMES. Typically, in climate studies these outcomes would be different scenarios that could occur over a given time period.
- An EVENT is a collection of outcomes. For example, there are many different scenarios that will give rise to 15 tropical cyclones in the North-West Pacific this year.
- A RANDOM VARIABLE is a rule for allocating numbers to outcomes. The number of tropical cyclones in the North-West Pacific this year is a random variable, since it provides a single-number summary of each possible climate scenario.

Discrete random variables

A DISCRETE RANDOM VARIABLE takes values in a countable set. In practice, discrete random variables usually arise as counts — for example, the number of tropical cyclones in a time period.

Suppose Y is a discrete random variable. Its entire probability structure can be summarised by making a list of all the values it can take, and allocating the appropriate probabilities: i.e. by specifying $P(Y = k)$ for all the possible values of k. Such a specification is called the PROBABILITY DISTRIBUTION of Y. Note that ‘$Y = k$’ is an event (it corresponds to a statement of the form ‘there are 15 tropical cyclones in the North-West Pacific’).

Often, it is useful to be able to summarize the behaviour of a random variable by giving a single ‘typical’ value. This leads to the notion of ‘averages’, and to some further definitions. The average of a set of observations is usually understood to be the result of adding them together, and dividing by the number of observations. The result represents, in some sense, the value that would have been observed if all of the observations had been equal. So the operation of ‘averaging’ provides us with a convenient single-number summary of an entire dataset.

This notion can be extended to random variables. The reason is that when we make probability statements about random variables, we are imagining many repetitions of an experiment under identical conditions. Each of these repetitions will give rise to a single value of the variable. The average of all these values is a single number which we may regard as representative.

Definition: The EXPECTED VALUE, or EXPECTATION, of a discrete random variable Y, denoted by $E(Y)$, is defined as

$$E(Y) = \sum_k kP\left(Y = k\right) \ ,$$

providing the sum (which is over all values taken by Y) is well-defined.

$E(Y)$ represents an idealised long-run average for the values of Y. It is also called the (POPULATION) MEAN of Y, and is usually denoted by μ. It need not necessarily be a value that Y can take (for example, we might find that the expected number of tropical cyclones in the North-West Pacific this year is 22.6).

If Y is a random variable, then so is any transformation of Y, say $g(Y)$; hence we can talk about $E\left(Y^2\right), E(\ln Y)$ etc. Note that, in general, $E\left(g(Y)\right)$ is *not* equal to $g\left(E(Y)\right)$.

The expectation, μ, of a random variable Y is intended to give a 'representative' value. It is natural to ask '*How* representative, exactly?'. Put another way: if we forecast that the value of Y will be μ, how big will our error be, on average?

Definition: The VARIANCE of Y is defined by

$$\text{Var}(Y) = E\left((Y-\mu)^2\right) \ ,$$

where it exists.

$\text{Var}(Y)$ is often denoted by σ^2. If we repeat an experiment many times, each time issuing a forecast 'The value of Y will be μ', σ^2 is the average squared forecast error. Therefore it is a direct measure of uncertainty. The square root of the variance, σ, is called the STANDARD DEVIATION of Y. The motivation for this is that σ is measured in the same units as Y, and hence has a direct interpretation. The standard deviation can be thought of as the likely magnitude of a forecast error (recall the discussion in Section 1.1.2).

For random variables that cannot take negative values, the ratio of standard deviation to mean (σ/μ) is a dimensionless quantity called the COEFFICIENT OF VARIATION.

Continuous random variables

Not all random variables are discrete. For example, temperature can take any value on a continuous scale. This is an example of a CONTINUOUS RANDOM VARIABLE. If Y is a continuous random variable, it is not obvious how we can calculate probabilities such as $P(Y = 18)$ — in the long run, how often will a value of exactly 18 (rather than $17.9999\ldots$ or $18.000\ldots001$) arise?

It is easier to think of continuous random variables by considering events of the form '$Y \leq y$', since probabilities can easily be allocated to such events (it is straightforward to make statements such as ‘75% of the time, the temperature will be less than 18°C’).

Definition: for any random variable Y, the function

$$F(y) = P\left(Y \le y\right)$$

is called the (CUMULATIVE) DISTRIBUTION FUNCTION (cdf) of Y.

Distribution functions are defined for both discrete and continuous random variables. We can use the distribution function to find the probability of obtaining a value in any interval: if $a < b$ then $P(a < Y \le b) = F(b) - F(a)$. Now, if $F(.)$ is continuous and differentiable, such that $dF/dy = f(y)$, we have

$$\int_a^b f(y)dy = F(b) - F(a) = P(a < Y \le b)\ ,$$

since integration is the opposite of differentiation.

The function $f(.)$ here is called the (PROBABILITY) DENSITY FUNCTION of Y. Note that $\int_a^b f(y)dy$ is the area under the graph of $f(.)$ between a and b. Hence, if Y has density $f(.)$, $P(a < Y \le b)$ can be obtained as the area under the graph of $f(.)$, between a and b (see Figure 1.1). Note that $f(y)$ is *not* the same as the probability that $Y = y$. In fact, for any continuous random variable, the above discussion forces us to accept that for any number y, $P(Y = y) = 0$ (since the area under the graph of $f(.)$, between y and y, is zero). This is not intuitively obvious, but the logic is correct! We have discovered that ‘possible’ events may have probability zero.

Notice that $F(.)$ can be obtained from $f(.)$ as $F(y) = \int_{-\infty}^{y} f(u)du$). Hence $f(.)$ and $F(.)$ each specify completely the probability distribution of a continuous random variable.

How can we define expectation for continuous random variables? Well, we could choose to approximate a such a variable by a discrete one which takes values on a very finely spaced grid of points (separated by intervals of length δy, say). In this case, if δy is small, we will find that $f(y + \delta y) \approx f(y)$, so that the area under the graph of $f(.)$ between y and $y + \delta y$ is approximately $f(y)\delta y$. If we say that the discrete approximation takes the value y when Y lies between y and $y + \delta y$, then the expected value of this discrete random variable is $\sum y f(y)\delta y$. The approximation will improve as δy becomes smaller and smaller; in the limit as δy tends to zero, the sum tends to an integral. Therefore we have

Definition: Let Y be a continuous random variable with density $f(.)$. The EXPECTED VALUE of Y is defined as

$$E(Y) = \int_{-\infty}^{\infty} y f(y)dy = \mu, \text{ say,}$$

providing this integral is well defined.

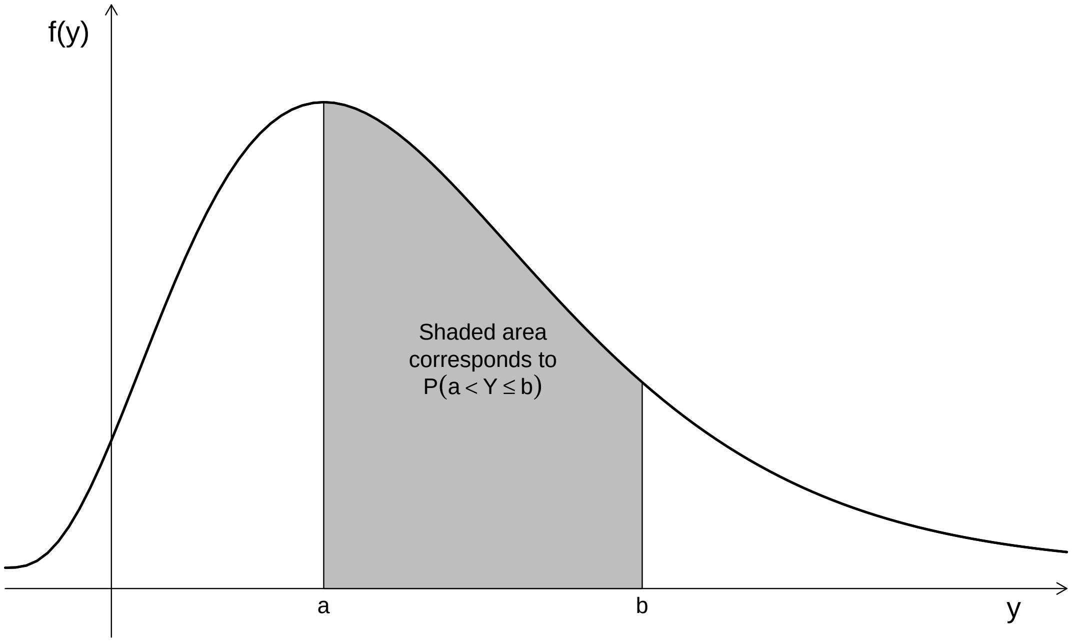

Figure 1.1: Interpretation of a probability density function: probabilities are represented by areas under the graph.

The definition of variance for a continuous random variable is the same as that for the discrete case i.e. $E\left((Y-\mu)^2\right)$. Interpretation of expectation and variance is the same as for the discrete case. In general, definitions and formulae are also the same, except that we replace $\sum$ by $\int$, and $P(Y=k)$ by $f(y)dy$.

Some random variables have distributions that are a mixture of discrete and continuous components. Daily rainfall is a good example — even in Ireland, a proportion of days experience exactly zero rainfall, but if the rainfall amount is non-zero it may be regarded as a continuous random variable.

Joint and conditional distributions

In Section 1.2.2, we introduced conditional probability as a way of studying relationships between events. In general, we will want to study relationships between random variables in a similar manner. If $Y_1,\ldots,Y_p$ are each discrete random variables, we can define their JOINT PROBABILITY DISTRIBUTION by considering

$$P\left(Y_1=k_1 \text{ and } Y_2\ =k_2\ \ldots \text{ and } Y_p=k_p\right)$$

for all the possible values of $k_1,\ldots,k_p$. When we study several random variables together, we will often assemble them into a vector for ease of notation:$\boldsymbol{Y}=(Y_1\ \ldots\ Y_p)'$ say. $\boldsymbol{Y}$ is called a

RANDOM VECTOR.

Occasionally, we will be interested in studying functions of several random variables and, in particular, finding their expectations. The expected value of $g\left(Y_1, \ldots, Y_p\right)$ is just

$$\sum_{k_1,\ldots,k_p} g\left(k_1, \ldots, k_p\right) \times P\left(Y_1 = k_1 \text{ and } Y_2 \ = k_2 \ \ldots \ \text{and } Y_p = k_p\right) \ .$$

The most important examples of this are the following:

Definition: Let Y_1 and Y_2 be random variables, with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2, respectively. The COVARIANCE between Y_1 and Y_2 is defined as

$$\text{Cov}\left(Y_1, Y_2\right) = E\left(\left(Y_1 - \mu_1\right)\left(Y_2 - \mu_2\right)\right) = E\left(Y_1 Y_2\right) - \mu_1 \mu_2 \ ,$$

and the CORRELATION between Y_1 and Y_2 is defined as

$$\text{Corr}\left(Y_1, Y_2\right) = \frac{\text{Cov}\left(Y_1, Y_2\right)}{\sigma_1 \sigma_2} \ .$$

It can be shown that Corr $\left(Y_1, Y_2\right)$ takes values between -1 and $+1$, and that it only achieves these values if Y_1 and Y_2 are linearly related: $Y_1 = a + bY_2$, where a and b are constants. If $b < 0$ then Corr $\left(Y_1, Y_2\right) = -1$; otherwise it is $+1$.

Note that Cov$(Y, Y) =$ Var(Y). The main reason for introducing covariance at this point is to present the following result:

Result: Let Y_1 and Y_2 be two random variables. Then

$$\text{Var}\left(Y_1 \pm Y_2\right) = Var\left(Y_1\right) + \text{Var}\left(Y_2\right) \pm 2\text{Cov}\left(Y_1, Y_2\right) \ .$$

This tells us how uncertainty is propagated when we combine information. We will return to this point below.

Now suppose we observe the values of all of the Ys except for Y_1. We may define the CONDITIONAL DISTRIBUTION OF Y_1 GIVEN $Y_2, \ldots, Y_p$, by considering

$$P\left(Y_1 = k_1 \middle|\ Y_2 \ = k_2 \text{ and } \ldots \text{ and } Y_p = k_p\right) = \frac{P\left(Y_1 = k_1 \text{ and } Y_2 = k_2 \text{ and } \ldots \text{ and } Y_p = k_p\right)}{P\left(Y_1 = k_1 \text{ and } Y_2 \ = k_2 \ \ldots \ \text{and } Y_p = k_p\right)} \ ,$$

according to the definition of conditional probability given in Section 1.2.2. We can, if we wish, calculate the CONDITIONAL MEAN and CONDITIONAL VARIANCE of this distribution, in exactly the same way as for any other discrete distribution.

If the Ys are continuous rather than discrete, then we can define their JOINT DISTRIBUTION FUNCTION

$$F\left(y_1, \ldots, y_p\right) = P\left(Y_1 \leq y_1 \text{ and } \ldots \text{ and } Y_p \leq y_p\right) \ ,$$

and the corresponding JOINT DENSITY

$$f\left(y_1,\ldots,y_p\right)=\frac{\partial^p F}{\partial y_1\ldots\partial y_p}\ .$$

As before, results and formulae for continuous distributions are equivalent to those for discrete distributions, replacing sums by integrals and probabilities by densities. For example, the conditional distribution of Y_1 given $Y_2,\ldots,Y_p$ has density

$$f\left(y_1|Y_2\ =y_2 \text{ and } \ldots \text{ and } Y_p=y_p\right)=\frac{f\left(y_1,\ldots,y_p\right)}{f\left(y_2,\ldots,y_p\right)}\ .$$

(the denominator here is the joint density of $Y_2,\ldots,Y_p$). Also, the expected value of a function of continuous random variables is

$$E\left(g\left(Y_1,\ldots,Y_p\right)\right)=\int g\left(y_1,\ldots,y_p\right)f\left(y_1,\ldots,y_p\right)dy_1\ldots dy_p\ .$$

From now on, we will discuss only continuous random variables. All of the results hold in the discrete case as well.

Independence of random variables

In the same way as we defined independent events in Section 1.2.2, we can now define independence of random variables. Specifically, we say that Y_1 and Y_2 are independent if *all* probability statements about Y_1 are unaffected by observing Y_2, and vice versa. If Y_1 and Y_2 are independent continuous random variables then their joint density is given by the product of their individual densities: $f\left(y_1,y_2\right)\ =\ f_1\left(y_1\right)f_2\left(y_2\right)$, say. In addition, $E\left(Y_1Y_2\right)\ =\ E\left(Y_1\right)E\left(Y_2\right)$ so that the covariance (and hence the correlation) between Y_1 and Y_2 is zero. Note, however, that a zero correlation between Y_1 and Y_2 does *not* imply that they are independent.

From the formula for Var $\left(Y_1\pm Y_2\right)$ given above, in the independent case the variance of a sum (or difference) is just the sum of the individual variances. Since variance is a direct measure of uncertainty (see page 13), this tells us that if we combine outputs from two or more unrelated models (as in Example 1.1), the uncertainty will accumulate — we cannot rely on errors in one model compensating for those in another.

The idea of independence can be extended to that of CONDITIONAL INDEPENDENCE. Suppose we have three random variables Y_1, Y_2 and Y_3, and that neither of these is independent of either of the other two. However, it may be that the conditional distribution of Y_1 given Y_2 is the same as that of Y_1 given *both* Y_2 and Y_3. This could be expressed, in terms of densities, as

$$f\left(y_1|Y_2=y_2\right)=f\left(y_1|Y_2=y_2 \text{ and } Y_3=y_3\right)\ .$$

The implication of this is that, once we know Y_2, Y_3 tells us nothing new about Y_1. This notion is extremely important, and is particularly relevant for the study of complex systems such as the climate. An example would probably help:

Example 1.6: Suppose Y_1 is a random variable taking the value 1 if it rains in the West of Ireland today, and 0 otherwise (such a variable, taking the value 1 if an event occurs and 0 otherwise, is called an INDICATOR VARIABLE). Let Y_2 be an indicator for the passage of a cold front over the area today, and let Y_3 be an indicator for rain yesterday.

In Example 1.5, we saw that $P(Y_1 = 1) = 0.7$, and that $P(Y_1 = 1|Y_3 = 1) = 0.85$. For the sake of argument, suppose that on 98% of days when a cold front passes over the region, precipitation occurs. Suppose also that a cold front passes over the region today. Knowing this, we can say that $P(Y_1 = 1|Y_2 = 1) = 0.98$. Now we learn, in addition, that it rained yesterday. Realistically, this information is now irrelevant — today, it will rain because a cold front is present, not because it rained yesterday. The probability of rain today is still 0.98 i.e. $P(Y_1 = 1|Y_2 = 1 \text{ and } Y_3 = 1) = P(Y_1 = 1|Y_2 = 1)$. ■

The calculation in this example is not enough to demonstrate conditional independence of Y_1 and Y_3 given Y_2. A complete analysis would require computation of the probabilities for all possible combinations of values for the three variables. However, it illustrates the general concept. The important point is that dependence between two variables may be completely explained by the effect of a third variable. In climate research this situation often arises when, as here, the third variable represents a genuine physical mechanism. In this case, Y_3 only tells us about Y_1 in the absence of information about Y_2.

1.3 Simple distributions

In this section, we summarise a few of the more commonly-used probability distributions. We give, without proof, the important properties of each distribution. We also give, where appropriate, an overview of situations in which each distribution might arise.

1.3.1 The Bernoulli distribution

This is the simplest possible probability distribution, for a random variable Y which takes either of the values 0 and 1. If $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$, we say that Y has a BERNOULLI DISTRIBUTION WITH PARAMETER p. In this case, $E(Y) = p$ and $\text{Var}(Y) = p(1 - p)$.

The Bernoulli distribution is not often mentioned by name, but it is commonly used. The random variables in Example 1.6 were all Bernoulli random variables. It is often useful to think of indicator variables in terms of the Bernoulli distribution.

1.3.2 The Binomial distribution

Suppose an experiment is repeated n times, under identical conditions. The probability of some event A occurring each time is p, independently of all other repetitions. Let Y be the total number of repetitions in which A occurs. Then Y takes values in $\{0, 1, \ldots, n\}$. We say that Y follows a

BINOMIAL DISTRIBUTION WITH PARAMETERS n AND p, and write $Y \sim Bin(n,p)$. The probabilities are given by

$$P(Y=k) = \binom{n}{k} p^k(1-p)^{n-k} \;=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}\;\; (k=0,1,\ldots,n).$$

The mean of the distribution is np, and the variance is $np(1-p)$.

The binomial distribution may not be directly applicable to many climate variables, since there will be few datasets where all of the observations have been obtained independently and under identical conditions. However, it may be that homogeneous subsets of the observations can be regarded as following different binomial distributions.

Notice that a Binomial random variable can be regarded as the sum of n independent Bernoulli random variables.

1.3.3 The Poisson distribution

A discrete random variable Y is said to have a POISSON DISTRIBUTION WITH PARAMETER μ if

$$P(Y=k) = \frac{\mu^k e^{-\mu}}{k!} \qquad (k=0,1,2,\ldots).$$

We write $Y \sim Poi(\mu)$. The distribution, incidentally, is named after Siméon Poisson (1781–1840). The mean and variance are both equal to μ. A further useful property is that if Y_1 and Y_2 are independent Poisson variables, with means μ_1 and μ_2, then $Y_1+Y_2 \sim Poi(\mu_1+\mu_2)$.

The Poisson distribution can be used as an approximation to the Binomial (n,p) when n is large and p is small. However, in climate and similar applications this particular aspect is rarely useful.

The Poisson Process

It is common to encounter variables that represent the number of occurrences of some phenomenon during a time period. For example, we may wish to count the number of typhoons in the North-West Pacific this year. A schematic diagram showing this situation is shown in Figure 1.2. A process that can be represented in this way is called a POINT PROCESS (because the occurrences can be regarded as points on a line).

Let Y be the number of occurrences in the time interval of interest. Denote the duration of this interval by t, and assume that (i) no more than 1 occurrence is associated with each time instant; (ii) the numbers of occurrences in non-overlapping time intervals are independent random variables; and (iii) the mean number of occurrences in any time interval of length 1 time unit is λ. Then it can be shown that Y has a Poisson distribution with mean λt.

A process of occurrences, which satisfies the assumptions in the previous paragraph, is called a HOMOGENEOUS POISSON PROCESS. The parameter λ is called the RATE of the process. At first

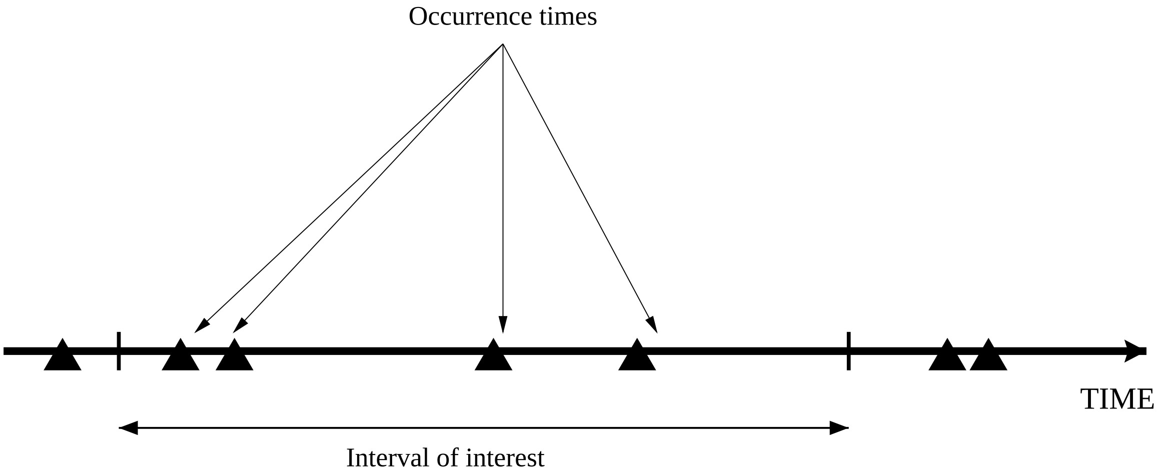

Figure 1.2: Schematic diagram of a point process.

sight, Assumptions (ii) and (iii) appear restrictive. Assumption (ii) requires that non-overlapping time intervals should be unrelated to each other: this seems unlikely to hold for many climate applications. Assumption (iii) requires that the rate of occurrences is the same for all time.

In fact, Assumption (iii) can be relaxed: suppose that there exists an INTENSITY FUNCTION, $\lambda(t)$ say, such that the probability of an occurrence in the interval $[t, t+\delta t)$ is approximately $\lambda(t)\delta t$. In this case, and under Assumptions (i) and (ii) above, it can be shown that the number of events in the interval $[a, b)$ has a Poisson distribution with mean $\int_a^b \lambda(t)dt$. Such a process, with time-varying intensity, is called an INHOMOGENEOUS POISSON PROCESS.

Assumption (ii) — that non-overlapping time intervals are independent of each other — remains. It seems that in most climate applications, this assumption will not hold. However, there are certain situations in which the Poisson process structure will hold approximately. Some examples of these are:

Superposition: If many point processes, none of which predominate, are superposed then the result is approximately a Poisson Process.

Thinning: If each occurrence in a point process is deleted with probability p, independently of all other occurrences, the resulting 'thinned' process is approximately a Poisson Process if p is large (i.e. close to 1).

Translation: If we take any point process, and randomly displace each occurrence independently of all the others,then under certain conditions on the displacement mechanism, the result is approximately a Poisson Process.

Rare events: In many situations where we are interested in studying 'rare events', a Poisson Process model may be appropriate. For example, we may be interested in studying days for which rainfall amounts exceed some threshold. If this threshold is high enough, the point process of exceedances is approximately a Poisson Process.

The presentation here is deliberately non-technical. In all of the situations above, certain conditions are necessary in order that the Poisson approximation is valid. However, they do suggest mechanisms under which it may be appropriate to use a Poisson distribution for counts.

Example 1.7: In the eastern parts of tropical oceans, *easterly waves* form roughly every 3 days. These are weak troughs in sea level pressure, that move westwards. As they move, some of them develop into hurricanes (i.e. develop windspeeds above 33ms^{-1}). In the Atlantic, there are typically 100 easterly waves per year, of which 8 reach tropical cyclone status.

On the basis of this, it may be that tropical cyclones occur in a Poisson process, since we start out with a point process of easterly waves, and delete each one with high probability so that we are left with just a few tropical cyclones. The rate of cyclone formation is seasonally varying; hence an inhomogeneous Poisson Process model may be appropriate. If this is the case, then the number of tropical cyclones in any one year has a Poisson distribution. ■

The Poisson process model is also applicable for counts of occurrences in regions of space, under similar assumptions as those given above.

1.3.4 The Normal distribution

A continuous random variable Y has a NORMAL (or GAUSSIAN) DISTRIBUTION WITH PARAMETERS μ AND σ^2 if its density is

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

We write $Y \sim N(\mu, \sigma^2)$. The mean of the distribution is μ, and the variance is σ^2.

The normal distribution is perhaps the most important model for continuous random variables, and much statistical theory is based upon it. The main reason for its importance is the CENTRAL LIMIT THEOREM. This states that if Y is the sum of a large number of independent random variables from *any* distribution then, under very general conditions, Y has approximately a normal distribution. Also, linear combinations of normal random variables are themselves normally distributed. Therefore, it is natural to use the normal distribution when studying quantities that are 'averages'.

Example 1.8: Suppose we wish to study monthly mean temperatures. Since each monthly mean is an average of around 30 daily values, it may be appropriate to consider that it is drawn from a normal distribution. Typically, however, each mean will be drawn from a *different* normal distribution, because of effects such as seasonality. ■

The normal distribution can be used as an approximation to many other distributions. For example, if $Y \sim Bin(n, p)$ then Y can be regarded as a sum of n independent Bernoulli random variables (see Section 1.3.2 above). Hence, if n is large, the distribution of Y can be approximated by a normal distribution. Similarly, the normal distribution can be used to approximate $Poi(\mu)$ when μ is large.

Another use for the normal distribution is to study 'errors of measurement' — in fact, this was one of its first uses (by Laplace in 1783, 35 years before Gauss used it — and 50 years after DeMoivre had used it as an approximation to the Binomial!).

The normal distribution with mean zero and variance 1 is referred to as the STANDARD NORMAL DISTRIBUTION. Standard normal random variables are usually denoted by Z. The density of the standard normal distribution is denoted by $\phi(.)$, and its distribution function by $\Phi(.)$. An important result is that if $Y \sim N(\mu, \sigma^2)$ and $Z = (Y - \mu)/\sigma$, then $Z \sim N(0, 1)$.

1.3.5 The Gamma, Exponential and Chi-squared distributions

A continuous random variable Y has a GAMMA DISTRIBUTION WITH PARAMETERS $\nu > 0$ AND $\lambda > 0$ if its density is

$$f(y) = \begin{cases} \lambda^\nu y^{\nu-1} e^{-\lambda y}/\Gamma(\nu) & y \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

where $\Gamma(.)$ denotes the GAMMA FUNCTION: $\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du = (\alpha - 1)\Gamma(\alpha - 1)$. The gamma function is sometimes called the GENERALISED FACTORIAL since, if α is an integer, $\Gamma(\alpha) = (\alpha - 1)!$.

If Y has this distribution, we write $Y \sim \Gamma(\nu, \lambda)$. The mean of the distribution is $\mu = \nu/\lambda$, and the variance is ν/λ^2. In Lectures 2 and 3, we will express the Gamma distribution via the parameters μ and ν, rather than λ and ν. In this case, the density for $y \geq 0$ is written as

$$\frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1} \exp\left[-\frac{\nu y}{\mu}\right] .$$

If $Y_1 \sim \Gamma(\nu_1, \lambda)$ and $Y_2 \sim \Gamma(\nu_2, \lambda)$ are independent then $Y_1 + Y_2 \sim \Gamma(\nu_1 + \nu_2, \lambda)$ (note that the parameter λ must be common to both distributions for this result to hold). It follows that, when ν is large, the $\Gamma(\nu, \lambda)$ distribution can be approximated by a normal distribution.

There are two situations in which the Gamma distribution arises naturally. These are as follows:

1. In a homogeneous Poisson Process of rate λ (see Section 1.3.3 above), the time until the kth event is distributed as $\Gamma(k, \lambda)$.

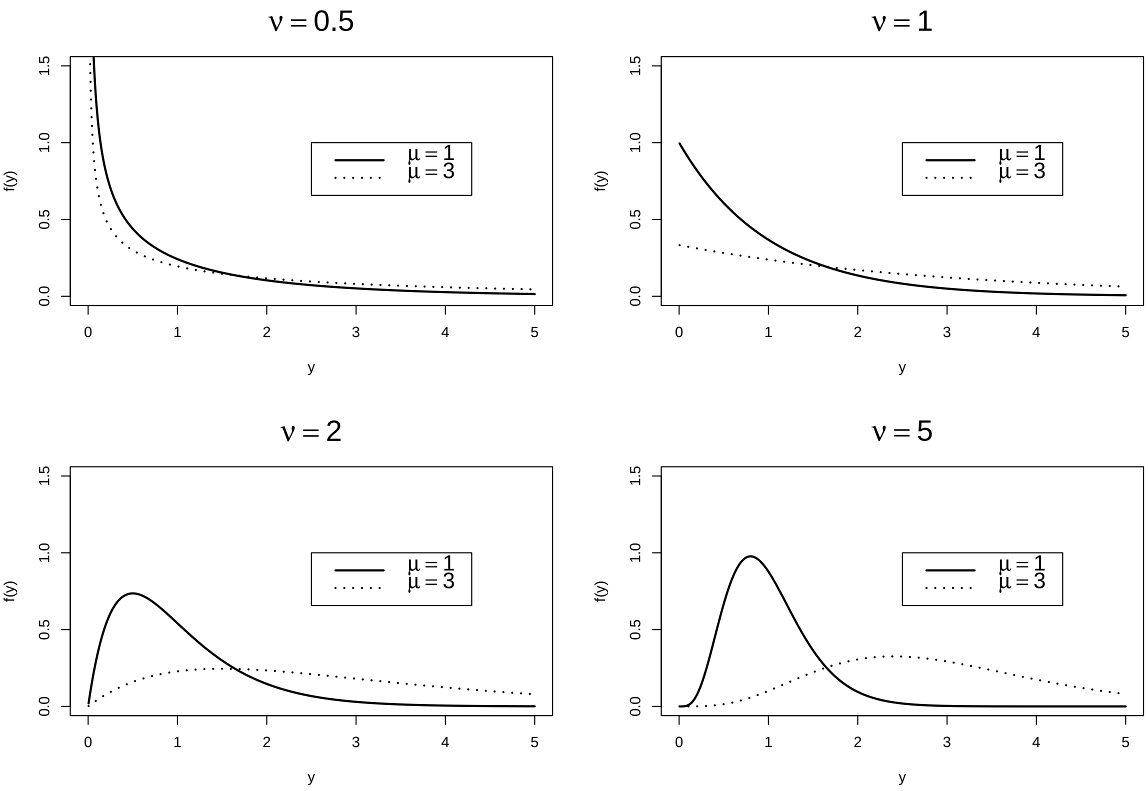

Figure 1.3: Examples of gamma densities.

2. If $Z_1, \ldots, Z_n$ are independent standard normal random variables and $Y = \sum_{i=1}^n Z_i^2$, then $Y \sim \Gamma(n/2, 1/2)$. In fact, this distribution is usually referred to as the CHI-SQUARED DISTRIBUTION WITH n DEGREES OF FREEDOM. We write $\sum_{i=1}^n Z_i^2 \sim \chi_n^2$.

More often, however, the gamma family of distributions provides a convenient model for any positive-valued random variable, simply because a wide variety of distributional shapes are available. Some of these are illustrated in Figure 1.3. We see that the parameter ν controls the shape of the distribution, and for this reason it is often called the SHAPE PARAMETER. For $\nu \leq 1$, the density has a maximum at $y = 0$. As ν increases, the distribution becomes more symmetric. Another interpretation for ν is in terms of the coefficient of variation (see page 13) of the distribution. This is equal to $\sqrt{\nu/\lambda^2}/\left(\nu/\lambda\right) = 1/\sqrt{\nu}$.

Figure 1.3 also shows that varying μ (or, equivalently, λ) does not affect the shape of the distribution: it merely scales the graph horizontally and vertically. For this reason, λ is usually referred to as the SCALE PARAMETER of the distribution.

When $\nu = 1$ the Gamma density is, for $y \geq 0$,

$$f(y) = \lambda e^{-\lambda x} \qquad \text{or} \qquad f(y) = \frac{1}{\mu} e^{-x/\mu} \,,$$

which is the EXPONENTIAL DISTRIBUTION — if Y has this distribution we write $Y \sim Exp(\lambda)$. This arises as the time to the first event in a Poisson process of rate λ. Note also, from the discussion above, that the χ^2_2 distribution is the same as $Exp\left(\frac{1}{2}\right)$.

1.3.6 The Weibull distribution

A random variable Y has a WEIBULL DISTRIBUTION WITH PARAMETERS α AND β if its density is

$$f(y) = \begin{cases} \frac{\alpha}{\beta}\left(y/\beta\right)^{\alpha-1}\exp\left[-\left(y/\beta\right)^{\alpha}\right] & y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

We write $Y \sim Wei(\alpha, \beta)$. α is the SHAPE PARAMETER of the distribution, and β is the SCALE PARAMETER. The mean is $\beta\Gamma\left(1+\alpha^{-1}\right)$, and the variance is $\beta^2\left[\Gamma\left(1+2\alpha^{-1}\right)-\Gamma^2\left(1+\alpha^{-1}\right)\right]$. A further property is that if $Y \sim Wei(\alpha, \beta)$ then $Y^k \sim Wei\left(\alpha/k, \beta^k\right)$. When $\alpha = 1$, we obtain an exponential distribution with parameter $\lambda = 1/\beta$.

The Weibull distribution does not arise naturally in many ‘obvious’ situations, at least in climatology. Two mechanisms that give rise to the Weibull distribution are as follows:

1. If $Y_1 \sim N(0, \sigma^2)$ and $Y_2 \sim N(0, \sigma^2)$ are independent, then $\sqrt{Y_1^2 + Y_2^2} \sim Wei\left(2, \sigma\sqrt{2}\right)$. This distribution is called the RAYLEIGH DISTRIBUTION. This particular result has motivated the use of the Weibull distribution to model windspeeds — if the u and v components of wind velocity are zero-mean normal random variables with the same variance, then the windspeed has a Rayleigh distribution.

2. Let $X_1, \ldots, X_n$ be independent continuous random variables, drawn from the same distribution and taking values in the range $[\tau, \infty)$ for some threshold τ. Define $Y = \min_i\left(X_i\right) - \tau$. Then, under certain conditions on the distribution of the Xs, the distribution of Y is approximately Weibull. As a consequence of this, the Weibull distribution is used in the study of extreme events. This will be discussed further in Lecture 3.

As with the Gamma distribution, the Weibull is often used simply because it provides a flexible class of distributions with different shapes. In practice it may be difficult to distinguish between the Gamma and Weibull distributions. Statistical inference (see Section 1.4.3 below) is more difficult for the Weibull than for the Gamma. The Weibull is useful, however, when we want to study the probability of observing values over some threshold, since

$$P(Y > y) = \exp\left[-\left(\frac{y}{\beta}\right)^{\alpha}\right],$$

for the Weibull modell. This has a particularly simple form, and is much easier to calculate than the corresponding expression for the Gamma distribution.

1.3.7 The Multivariate Normal distribution

The final distribution we study here is a joint distribution (see page 15). Suppose $Y_1, \ldots, Y_p$ are random variables: the mean of Y_i is μ_i, the variance is σ_{ii} and the covariance between Y_i and Y_j (see page 16) is σ_{ij}. The Ys can be assembled into a vector $\boldsymbol{Y} = (Y_1 \ldots Y_p)'$, and the μs into a corresponding vector $\boldsymbol{\mu}$. In addition, we can define a $p \times p$ matrix, $\boldsymbol{\Sigma}$ say, whose (i,j)th element is σ_{ij}. If the joint density of the Ys can be written as

$$f(\boldsymbol{y}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}-\boldsymbol{\mu})\right],$$

where $(\boldsymbol{y}-\boldsymbol{\mu})'$ denotes the transpose of the vector $(\boldsymbol{y}-\boldsymbol{\mu})$, then $\boldsymbol{Y}$ has a MULTIVARIATE NORMAL DISTRIBUTION WITH MEAN $\boldsymbol{\mu}$ AND VARIANCE-COVARIANCE MATRIX $\boldsymbol{\Sigma}$. We write $\boldsymbol{Y} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The multivariate normal distribution has many appealing properties. The most important are:

1. Any subset of the Ys also has a multivariate normal distribution, with mean vector and covariance matrix obtained from the corresponding elements of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

2. The distribution of Y_i is $N(\mu_i, \sigma_{ii})$.

3. If Cov $(Y_i, Y_j) = 0$ then Y_i and Y_j are independent. Apart from some trivial cases involving Bernoulli random variables (see Section 1.3.1), the multivariate normal distribution is the *only* distribution for which zero covariance (i.e. lack of correlation) implies independence.

4. If $\boldsymbol{A}$ is an $r \times p$ matrix, with $r \leq p$, then $\boldsymbol{AY} \sim MVN(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A'})$. In particular, if we choose a $p \times p$ matrix $\boldsymbol{A}$ in such a way that $\boldsymbol{A'A} = \boldsymbol{\Sigma}^{-1}$ (for example, via Cholesky decomposition), then $\boldsymbol{AY} \sim MVN(\boldsymbol{A\mu}, \boldsymbol{I})$ where $\boldsymbol{I}$ is the $p \times p$ identity matrix. This result can be used to transform $\boldsymbol{Y}$ into a vector of independent variables.

5. Suppose we split $\boldsymbol{Y}$ into two parts, with corresponding splits for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ — i.e. write

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \text{ and } \ \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

   Suppose next that we observe $\boldsymbol{Y}_2 = \boldsymbol{y}_2$. The conditional distribution of Y_1 is now

$$MVN\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right).$$

   The matrix $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ is called the matrix of (POPULATION) REGRESSION COEFFICIENTS. The effect of observing $\boldsymbol{Y}_2$ is to adjust the mean of $\boldsymbol{Y}_1$, and to reduce the uncertainty.

We might imagine that, if $Y_1, \ldots, Y_p$ all have normal distributions, then their joint distribution is multivariate normal. In fact is not necessarily the case.

1.4 Probability models and statistical methods

So far, we have compiled a list of useful results in probability theory. We now review some of the ideas that are needed to apply these results.

1.4.1 Overview of probability modelling

Usually, climate investigations involve analysis of information, in the form of data. There are two stages to any analysis: the first is to describe the structures in the data and the second is to interpret them. We are concerned here primarily with interpretation, although description is equally important (without a clear description of structure, interpretation is impossible).

We will assume that our information is a set of numerical observations $y_1, \ldots, y_n$. These are typically measurements of some variable of interest, and can be assembled into a single vector $\boldsymbol{y}$. Usually, each y is associated with corresponding values of other variables, say $x^{(1)}, \ldots, x^{(p)}$, and we want to learn how the xs affect y. Formally, we regard the $\boldsymbol{y}$s as having been produced by a DATA GENERATING MECHANISM, in which the xs play a part. We aim to discover this mechanism.

The fundamental idea behind the methods in these lectures is that $\boldsymbol{y}$ is the observed value of a vector, $\boldsymbol{Y}$, of random variables. In other words, for the particular set of xs that were observed, we could have observed many different $\boldsymbol{y}$s. This approach does not exclude the fact that the data generating mechanism may be purely deterministic — this was discussed earlier, in Section 1.1.2.

Since $\boldsymbol{Y}$ is a vector of random variables, it has a joint distribution (see page 15), with density $f(.)$, say. This can be specified by the values of one or more parameters. We distinguish between two classes of parameter: 'statistical' and 'physical' (although this distinction may not always be obvious). A statistical parameter merely describes probability structure; examples are μ or σ^2 in the parametrisation of the normal distribution. A physical parameter is used in a simplified representation of the mechanism generating the data — gravitational acceleration is an example. In our probability-based framework, we physical parameters will usually contribute to the mean vector of the joint distribution, since this is our 'expectation' given our understanding of the system.

The values of some parameters (such as gravitational acceleration) may be known in advance. The remaining unknown parameters can be assembled into a vector, $\boldsymbol{\theta}$ say. In this case, the dependence of the joint density $f(.)$ upon $\boldsymbol{\theta}$ can be emphasised by writing $f(.;\boldsymbol{\theta})$. Then our objective, of discovering the data generating mechanism, can be re-stated as follows:

Objective: Given data $\boldsymbol{y}$, to specify a suitable joint density $f(.;\boldsymbol{\theta})$ for the underlying joint distribution, and to learn as much as possible about $\boldsymbol{\theta}$.

In this context, $f(.;\boldsymbol{\theta})$ is a PROBABILITY MODEL for the data. Ideally, its specification will use an understanding of the underlying mechanism to specify a plausible mean structure. This can then be combined with a knowledge of probability theory to select an appropriate distribution. For example, if the ys are averages, we might consider using normal distributions to model them.

Learning about $\boldsymbol{\theta}$ may involve estimating its value from the available data, or testing whether the data are consistent with some prespecified value, $\boldsymbol{\theta}_0$ say. Several methods may be available for carrying out such tasks, each of which yields a slightly different answer. It is then natural to ask: which method should we prefer? We now consider some of the issues involved in answering this question, before giving an overview of some commonly-used methods in Section 1.4.3.

1.4.2 Estimation and inference — issues

Suppose, for convenience, that there is a single unknown parameter, θ. Hopefully, this can be estimated using some function of the observations $\boldsymbol{y}$. Any function of the random vector $\boldsymbol{Y}$ is called a STATISTIC.

Definition: A statistic $T(\boldsymbol{Y})$ is an ESTIMATOR of a parameter θ if its value $t = T(\boldsymbol{y})$ is used as an estimate of θ.

Since the Ys are random variables, so is any statistic. Thus any estimator T has a probability distribution. The properties of this distribution determine whether or not T is a 'good' estimator.

Definition: If $E(T) = \theta$ then T is an UNBIASED estimator of θ. The difference $b(T;\theta) = E(T) - \theta$ is the BIAS of T as an estimator of θ.

An unbiased estimator will give us the right value 'on average' i.e. in a long series of experiments. Of course, we only observe $\boldsymbol{y}$ once! We would like to be fairly sure that $T(\boldsymbol{y})$ is close to the actual value of θ. This will happen if T is unbiased and has a small standard deviation:

Definition: The distribution of an estimator is called its SAMPLING DISTRIBUTION. The STANDARD ERROR of an estimator is the standard deviation of this sampling distribution.

The terminology is potentially confusing, but is used to emphasise the fact that we are talking about the typical magnitude of an estimation error.

Unbiased estimators may not be unique. So, given two unbiased estimators, how do we choose between them? Obviously, we want the one which is expected to give the smallest error, in some sense. One way of achieving this is via the criterion of mean square error:

Definition: The MEAN SQUARE ERROR of an estimator T for a parameter θ is defined as

$$\text{MSE}(T) = E\left[(T-\theta)^2\right] .$$

It can be shown that $\mathrm{MSE}(T) = \mathrm{Var}(T;\theta) + b^2(T,\theta)$. There is a tradeoff, in terms of MSE, between bias and variance. If we concentrate on unbiased estimators then the smallest possible MSE is achieved by the estimator with minimum variance. However, by considering estimators with a small amount of bias, we may find something that has a much smaller variance than the best unbiased estimator, so that we can reduce the MSE.

Interval estimation

So far, we have only considered giving a single number as our estimate of θ i.e. a POINT ESTIMATE. Of course, since the estimate is a realised value of a random variable it is unlikely to be exactly equal to θ (in fact, if $T(\boldsymbol{Y})$ is a continuous random variable, $P(T(Y) = \theta) = 0!$). In this case, we might wish to give a *range* of θ values as our estimate. Such a range is called an INTERVAL ESTIMATE or a CONFIDENCE INTERVAL.

Confidence intervals are constructed in such a way that they have a specified probability of including the 'true' value of θ. For example, a 95% confidence interval will contain the true value with probability 0.95. A 99% interval will include the true value with probability 0.99, and therefore needs to be wider than a 95% interval.

Typically, if the variance of our estimator is small then we will be reasonably sure that our point estimate is close to θ, and will therefore our confidence interval will be narrow. On the other hand, if the variance of the estimator is large then our confidence interval will be wide. Thus the length of the interval tells us something about the precision of the estimator. Viewed in another way, it tells us about our uncertainty regarding θ.

There are various ways of calculating confidence intervals in practice. The 'obvious', and most common, way is to use the interval

$$t \pm (k \times \text{standard error}) \ ,$$

where t is the estimate and k is a constant, chosen appropriately so as to give the right probability of including the true value of θ. However, this method will only be correct if the standard error of an estimator does not depend on θ. An alternative method is discussed in the next section.

Hypothesis testing

In some ways related to the idea of interval estimation is that of HYPOTHESIS TESTING. This theory was developed for use in controlled experiments involving simple structures. The idea is to test whether or not the data are consistent with some prespecified value of θ, say θ_0. Typically, this is done by constructing a TEST STATISTIC which is expected to take 'small' values if the true value of θ is θ_0, and 'large' values otherwise.

If the observed value of the test statistic is less than some constant, c say, we conclude that the data are consistent with the NULL HYPOTHESIS $\theta = \theta_0$; otherwise, we REJECT the null hypothesis in favour of an ALTERNATIVE HYPOTHESIS. There are two types of error we can make in this procedure: either we reject the null hypothesis when it is true (a TYPE 1 ERROR), or we accept it when it is false (a TYPE 2 ERROR). The value of c is usually chosen so that the probability of making a type 1 error is equal to some prespecified value such as 0.05 or 0.01. This probability is called the SIGNIFICANCE LEVEL of the test. If the null hypothesis is false, the probability of making a type 2 error then depends upon factors such as the sample size and the magnitude of the difference between θ_0 and the true value of θ. If β is the probability of making a type 2 error, then the quantity $1 - \beta$ is called the POWER of the test. It measures the ability of the test to detect genuine departures from the null hypothesis.

An equivalent approach to testing is to calculate the probability that, under the null hypothesis, the test statistic will exceed its observed value. This probability is called the OBSERVED SIGNIFICANCE LEVEL or p-VALUE. For a test at the 5% level (i.e. where the type 1 error rate is 0.05), we will reject the null hypothesis if the p-value is less than 0.05.

Suppose now that two test procedures are available to us. How do we choose which one to use? From the discussion above, one approach would be to fix a significance level, and then choose the procedure that has the highest power. We return to this below.

Although hypothesis testing is widespread, in many applications it should be used with caution. The reason is that any testing procedure is designed to answer the question ‘Is θ equal to θ_0?’. The answer to this, when studying a complex system such as the climate, is almost certainly ‘No’. If we have a large dataset, any powerful testing procedure is very likely to reject a null hypothesis, even if the difference between θ_0 and the true value of θ is very small. The procedure also assumes that there *is* a true value of θ, which will only be the case if data generating mechanism follows the particular mathematical form that we are using!

1.4.3 Estimation and inference — techniques

Finally in this lecture, we introduce three methods that are commonly used to implement the ideas discussed above.

Method of moments

The natural way to estimate the parameter vector $\boldsymbol{\theta}$ in the probability model $f(.;\boldsymbol{\theta})$ is to match properties of the distribution with the corresponding properties of the data. For example, to estimate the mean of any distribution we could use $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ — i.e. use the mean of the data to estimate the mean of the distribution. In some situations, we may modify the idea slightly to obtain unbiased estimators. For example, the obvious estimator of the variance of a normal distribution is $n^{-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$, but this is biased so we usually use the estimator $s^2 = (n-1)^{-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$ instead. s^2 is called the SAMPLE VARIANCE.

The method of moments is usually easy to implement. However, the choice of properties to use is not always obvious. A disadvantage is that two scientists, analysing the same data, may choose different properties and reach different conclusions. In simple cases this is not usually a problem, but it can be a major drawback when dealing with complex models.

Least squares

In cases where we wish to model the mean structure of a distribution (i.e. $E(Y_i)$ depends on $\boldsymbol{\theta}$), we may wish to choose $\boldsymbol{\theta}$ so as to minimise the quantity $\sum_{i=1}^{n}(Y_i - E(Y_i))^2$. Estimators derived in this way are called LEAST SQUARES ESTIMATORS. In some sense they improve upon moment estimators, since there is no arbitrary choice of fitting properties. However, the approach can only be used to estimate parameters relating to the mean of a distribution. In its simplest form, it is appropriate only when all of the Ys are drawn from distributions with the same variance. However, modifications exist to deal with unequal variances.

If all of the Ys have the same mean μ, the least squares estimator of μ is $\bar{Y}$.

Maximum likelihood

Another possible approach to estimation relies on a very simple idea: we should choose the value of $\boldsymbol{\theta}$ which allocates the highest probability to the observations $\boldsymbol{y}$. Specifically, we define the LIKELIHOOD for $\boldsymbol{\theta}$ given $\boldsymbol{y}$, as

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y};\boldsymbol{\theta}) \ ,$$

and define the MAXIMUM LIKELIHOOD ESTIMATE OF $\boldsymbol{\theta}$ to be the value which maximises $L(\boldsymbol{\theta}|\boldsymbol{y})$. Equivalently, it is the value that maximises the LOG-LIKELIHOOD $\ln L(\boldsymbol{\theta}|\boldsymbol{y})$. In practice, maximising $\ln L(\boldsymbol{\theta}|\boldsymbol{y})$ is often easier than maximising $L(\boldsymbol{\theta}|\boldsymbol{y})$. If the Ys are all independent random variables such that the density of Y_i is $f_i(.;\boldsymbol{\theta})$, then their joint density is just $f(\boldsymbol{y};\boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(y_i;\boldsymbol{\theta})$, so that the log-likelihood is

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

We denote the maximum likelihood estimate (MLE) by $\hat{\boldsymbol{\theta}}$.

Now consider how we might use the likelihood to form a confidence interval for a parameter. Previously, we said that a confidence interval could be calculated as

$$t \pm (k \times \text{standard error}) \ .$$

Another way of defining a confidence interval is as the set of all values of $\boldsymbol{\theta}$ for which the likelihood (or, equivalently, the log-likelihood) exceeds some threshold. An example will be given in Section 3.2.5. Such an interval will only be similar to the 'obvious' one if the likelihood is fairly symmetric about its maximum point.

Finally, we consider hypothesis testing. In a likelihood-based framework, we can test whether the data are consistent with an underlying value of $\boldsymbol{\theta}_0$ by examining the LIKELIHOOD RATIO $\Lambda = L(\hat{\boldsymbol{\theta}}|\boldsymbol{y})/L(\boldsymbol{\theta}_0|\boldsymbol{y})$, or its logarithm. By definition of $\hat{\boldsymbol{\theta}}$, $L(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \geq L(\boldsymbol{\theta}_0|\boldsymbol{y})$. Values of Λ close to 1 (i.e. values of $\ln \Lambda$ close to zero) are consistent with the null hypothesis; larger values are not.

Likelihood-based procedures have a number of appealing properties. A precise statement of these is lengthy and theoretical. However, for practical purposes the most important ones can be summarised as follows:

1. In a wide class of useful models, maximimum likelihood estimators have the smallest Mean Squared Error of *any* estimator.

2. For such models, likelihood-based confidence intervals are generally the shortest that can be found, at a specified confidence level.

3. The most powerful test for distinguishing between two hypotheses is based on the likelihood ratio (the NEYMAN-PEARSON LEMMA). This means that if a weak signal is present in a noisy record, a likelihood ratio test may be able to detect it when other procedures cannot.

The only disadvantage to likelihood-based inference is that it requires the probability model $f(\boldsymbol{y};\boldsymbol{\theta})$ to be completely specified. Results and conclusions will depend on this specification, so we need to ensure that the model structure is realistic. This can be achieved by combining prior knowledge of the problem with an understanding of probability mechanisms such as those discussed in Section 1.3 above.

1.5 Further reading

This lecture has summarised an extremely wide range of material, and we will not attempt to give an exhaustive reference list. There is very little applied literature that deals with the general ideas of probability modelling. The material in Section 1.4 is quite technical, if treated thoroughly, and there are few accessible statistical texts that cover it. For this reason, to find out more about the general ideas discussed in this lecture the best approach may be to consult a good undergraduate-level statistics text. Rice (1995) and Wackerly *et al*. (1996) are both excellent examples. These both give a very clear account of basic probability theory, as well as a good overview of statistical methods (including relatively accessible accounts of the material in Section 1.4).

The theory of point processes (Section 1.3.3) is not covered in so much detail in the texts cited above. Cox and Isham (1980) give a good theoretical account. Diggle (1983) gives a more applied treatment, focusing primarily upon processes in space rather than time — however, the basic ideas are the same in each case.

The discussion of the multivariate normal distribution (Section 1.3.7) closely follows that in Krzanowski (1988). This text is an excellent introduction to multivariate techniques in general — including many methods which are commonly used in climatology.

Lecture 2

Generalized Linear Models

In the previous lecture we discussed the need for probability-based modelling in climate research and gave an overview of some of the issues involved, together with a collection of necessary background material. In this lecture, we introduce a specific technique that can be used to apply these ideas.

The basic problem we consider is to determine how some quantity of interest is affected by other quantities. The quantity of interest is called the DEPENDENT or RESPONSE VARIABLE, and the other quantities will be referred to as EXPLANATORY VARIABLES, PREDICTORS or COVARIATES. For example, we might want to know if, or how, El Niño affects tropical cyclone formation in the North-West Pacific. In this case the dependent variable may be the number of cyclones in a year, and the predictors are appropriately-chosen El Niño indices.

The approach described here is an extension of the familiar technique of linear regression. However, our view of regression may differ from that normally encountered in climate research. We therefore start by outlining this view.

2.1 Overview of linear regression

In the simplest case, linear regression can be described as follows: we have n pairs of observations $\{(x_i, y_i) : i = 1, \ldots, n\}$, and a plot shows that the points are more or less scattered about a straight line. Accordingly, we may decide that the data can be summarised by the n equations

$$y_i = b_0 + b_1 x_i + e_i \qquad (i = 1, \ldots, n) \ ,$$

where e_i is the prediction error for the ith point, and b_0 and b_1 are chosen to minimise the sum of squared prediction errors, $\sum_{i=1}^{n} e_i^2$.

This procedure seems sensible, if a little *ad hoc*. We can make things a little more rigorous if we try to embed the procedure within the framework of a probability model. In most situations, if we have two x values that are the same the associated y values will differ, as discussed in Lecture 1. Thus y_i can be regarded as the observed value of a random variable Y_i, whose distribution depends on x_i. An appropriate model for such a situation might be

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ ,$$

where the $\{\varepsilon_i\}$ are independent random variables with zero mean and common variance (σ^2, say). Equivalently, we can write

$$E(Y_i) = \beta_0 + \beta_1 x_i = \mu(x_i) \text{ , say.}$$

If we also regard the xs as realised values of random variables $\{X_i : i = 1, n\}$, we are effectively modelling the conditional distribution (see page 16) of Y_i given $X_i = x_i$. If we make the extra assumption that the $\{\varepsilon_i\}$ are normally distributed, then we are predicting a probability distribution for Y given a value of x — we are saying that $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

2.1.1 Parameter estimation

The discussion in Lecture 1 suggests that we should use Maximum Likelihood to estimate parameters wherever possible. Since the Ys are assumed to be independent of each other, the likelihood is just the product of the individual densities (see Section 1.4.3) i.e.

$$L(\beta_0, \beta_1, \sigma^2; \boldsymbol{y}) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right] \right) \ ,$$

from the definition of the normal density in Section 1.3.4. The log-likelihood is therefore

$$\ln L(\beta_0, \beta_1, \sigma^2; \boldsymbol{y}) = \sum_{i=1}^{n} \left(-\frac{1}{2} \ln \sigma^2 - \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right) + \text{constant.}$$

To maximise with respect to β_0 and β_1, we calculate $\partial \ln L / \partial \beta_0$ and $\partial \ln L / \partial \beta_1$, set to zero and solve to obtain estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, say. It is clear, from the log-likelihood, that these estimates do not depend on σ^2. In fact, the only part of the log-likelihood which contributes to estimation of β_0 and β_1 is $-\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$, which we are trying to maximise. This is equivalent to minimising $\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ i.e. to Least Squares estimation. So, if we assume an underlying normal probability model, the 'obvious' Least Squares procedure yields Maximum Likelihood estimates — and is therefore optimal from a variety of viewpoints, as discussed in Section 1.4.3. The quantity $\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$ is called the RESIDUAL SUM OF SQUARES (RSS).

A Maximum Likelihood estimator of σ^2 can be obtained similarly — it depends on RSS. In fact this estimator is biased, and is usually adjusted to account for this. We will not discuss estimation of σ^2 further here. In the next section we will assume its value is known, since this simplifies the discussion and does not substantially affect any of the theory, at least for large samples.

2.1.2 Hypothesis testing

Suppose now that, knowing the value of σ^2, we wish to test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$. The scientific question being asked here is: is there any evidence for a genuine linear association between the ys and the xs? The standard way of doing this is to examine the ratio of $\hat{\beta}_1$ to its standard error. If this ratio is smaller than some critical value, we conclude that the data are consistent with an underlying process in which there is no association. The critical value depends on the significance level of the test. For a test at the 5% level (i.e. such that the probability of rejecting H_0 when it is true is 0.05), it is 1.96. When σ^2 is unknown, the critical value also depends on the sample size, but it is usually around 2 for a test at the 5% level.

An alternative test procedure is based on the likelihood ratio (see Section 1.4.3). Under H_0, the log-likelihood is

$$\sum_{i=1}^{n}\left(-\frac{1}{2}\ln\sigma^2 - \frac{(y_i-\beta_0)^2}{2\sigma^2}\right) + \text{constant},$$

and in this case the maximum likelihood estimate of β_0 is just the sample mean $\bar{y}$. The log likelihood ratio statistic is then

$$\begin{aligned}\ln\Lambda &= \sum_{i=1}^{n}\left(-\frac{1}{2}\ln\sigma^2 - \frac{(Y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2}{2\sigma^2} + \frac{1}{2}\ln\sigma^2 + \frac{(Y_i-\bar{Y})^2}{2\sigma^2}\right)\\ &= \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(Y_i-\bar{Y})^2 - \text{RSS}\right) .\end{aligned}$$

There are two points to note here:

1. The first term in the observed value of this statistic is proportional to the sample variance of the ys (see page 29). The residual sum of squares is the amount of variation that is unexplained by the regression. The likelihood ratio statistic is therefore closely related to the proportion of variance explained (i.e. the COEFFICIENT OF DETERMINATION, usually denoted by R^2).

2. Consider the expression for $2\ln\Lambda$:

$$2\ln\Lambda = \sum_{i=1}^{n}\left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2 - \frac{\text{RSS}}{\sigma^2} .$$

   This looks like the difference between two sums of squares of normally-distributed random variables. We expect, from Section 1.3.5, that each sum of squares will have a chi-squared distribution. Under the null hypothesis, this is indeed the case — the first term is distributed as χ^2_{n-1}, the second as χ^2_{n-2}, and the difference between them as χ^2_1. We will therefore accept H_0 at the 95% level if $2\ln\Lambda$ is less than the 95% point of a χ^2_1 distribution (which is 3.841); and at the 99% level if $2\ln\Lambda < 5.991$.

Finally, in this section, we introduce the concept of DEVIANCE. This may be thought of as a measure of discrepancy between the fitted model and a 'perfect' model. By a perfect model, we mean one in which each y value is perfectly predicted (i.e. in which $E(Y_i) = y_i$). In this case, the log-likelihood is

$$\sum_{i=1}^{n}\left(-\frac{1}{2}\ln\sigma^2 - \frac{(y_i-y_i)^2}{2\sigma^2}\right) + \text{constant} = -\frac{n}{2}\ln\sigma^2 + \text{constant}.$$

The discrepancy between our fitted model and a perfect model can be measured, in the usual way, by examining the ratio of likelihoods — or twice its logarithm(!). The statistic is

$$\begin{aligned} & 2\left(-\frac{n}{2}\ln\sigma^2 + \text{constant}\right) - 2\sum_{i=1}^{n}\left(-\frac{1}{2}\ln\sigma^2 - \frac{(Y_i - \hat\beta_0 - \hat\beta_1 x_i)^2}{2\sigma^2} + \text{constant}\right) \\ = & \sum_{i=1}^{n}\frac{(Y_i - \hat\beta_0 - \hat\beta_1 x_i)^2}{\sigma^2} \quad = \quad \frac{\text{RSS}}{\sigma^2} \quad = \; D \text{ , say.} \end{aligned}$$

The statistic D is called the SCALED DEVIANCE. The quantity $\sigma^2 D$ (here equal to RSS) is called the DEVIANCE. If the fitted model is correct, the scaled deviance is distributed as χ^2_{n-2}.

The important message from this section is that common concepts from the common Least Squares approach to regression have been embedded within the framework of a probability model, and interpreted via a likelihood-based approach to inference. This suggests how the ideas of regression may be extended to situations where the Ys have distributions other than the normal.

2.2 The extension to Generalised Linear Models

Suppose now we have a vector of random variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$. Associated with Y_i is a vector of p predictor variables: $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$. Let $\mu_i = E(Y_i|\boldsymbol{x}_i)$. Then a GENERALIZED LINEAR MODEL (GLM) for $\boldsymbol{Y}$ can be specified by choosing a family of probability distributions (e.g. Poisson, normal or gamma) for the Ys, and setting

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \eta_i \text{ ,say,}$$

where $g(.)$ is a monotonic function called the LINK FUNCTION, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters. η_i is called the LINEAR PREDICTOR.

Example 2.1: The linear regression model of the previous section is a GLM. In this case, the Ys are all normally distributed, and $\eta_i = \mu_i = \beta_0 + \beta_1 x_i$, so that the link function is the identity. ■

Example 2.2: Suppose the Ys are Bernoulli random variables (see Section 1.3.1), with $\mu_i = E(Y_i) = P(Y_i = 1)$. In this case, it does not make sense to write a linear equation such as

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \text{ ,}$$

since we know that μ_i must lie between 0 and 1. Therefore, we usually apply a transformation which maps the interval $[0, 1]$ to the whole real line. Various transformations are possible: perhaps the most common is the logistic transform $g(\mu_i) = \ln(\mu_i/(1-\mu_i))$. This gives rise to the LOGISTIC REGRESSION MODEL

$$\ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \text{ .}$$

Distribution	**Canonical link**	**Nuisance parameter**
$N(\mu, \sigma^2)$	$g(\mu) = \mu$	$\psi = \sigma^2$
$Poi(\mu)$	$g(\mu) = \ln \mu$	$\psi = 1$
$\Gamma\left(\nu, \frac{\nu}{\mu}\right)$	$g(\mu) = \mu^{-1}$	$\psi = \nu^{-1}$
$Ber(\mu)$	$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$	$\psi = 1$

Table 2.1: Examples of canonical link functions and nuisance parameters. Refer to Section 1.3 for details of the parameterisations used here.

Here the link function is the logistic transform. Once we know η_i we can find $\mu_i = P(Y_i = 1) = e^{\eta_i} / (1 + e^{\eta_i})$; then the distribution of Y_i is $Ber(\mu_i)$. ■

In practice, the link function is usually chosen to ensure that any restrictions on values of the fitted means are automatically satisfied (as in the logistic regression example above). Apart from this, there may be physical reasons for choosing a particular link function; however, often the choice is made purely for convenience. Many software packages use default link functions (which can be changed by the user). Unfortunately, these defaults are chosen for their mathematical elegance, which is not always the same as practical usefulness! They are referred to as CANONICAL LINKS. For our purposes, the only problem case is the gamma distribution, where the canonical link is $g(\mu) = \mu^{-1}$. Use of this link function does not guarantee that the fitted means will all be positive, and in applications it is far more natural to use the link $g(\mu) = \ln \mu$ instead.

In addition to the βs, most GLMs require additional parameters to be estimated; these typically specify the variance structure, and are assumed to be constant for all cases. For example, in a normal distribution we need to estimate the variance σ^2 and in a gamma distribution (see Section 1.3.5) we need to estimate the shape parameter ν. We refer to such parameters as NUISANCE PARAMETERS, and denote them by ψ. In the literature, and in software packages, they are often called SCALE PARAMETERS (which is confusing for the gamma distribution!). Some distributions, such as the Poisson and Bernoulli distributions, do not have nuisance parameters. In such cases, we take $\psi = 1$. Table 2.1 gives nuisance parameters, and canonical links, for a few distributions.

2.2.1 Inference and likelihood theory

Given a data vector $\boldsymbol{y}$, the βs in a GLM may be estimated by Maximum Likelihood. This is usually done using ITERATIVE WEIGHTED LEAST SQUARES, which is an efficient numerical algorithm that works for a wide range of useful distributions (these are distributions in the EXPONENTIAL FAMILY). Likelihood theory can therefore be applied to problems such as hypothesis testing within the GLM framework.

Typically, software packages will output parameter estimates together with their standard errors. These can be used to obtain approximate confidence intervals for each of the parameters, and to test hypotheses in the same way as for linear regression (discussed in Section 2.1.2 above). However, the results of such a procedure should be interpreted with *extreme* caution when two or more predictors are highly associated. We will see an example of this in the first case study below.

In general, it is better to use likelihood ratio tests to assess the significance of predictors. The theory is a direct extension of the linear regression case discussed in Section 2.1.2 above. Specifically, we suppose that the linear predictor in our model has the form

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \ ,$$

and we wish to test the null hypothesis $H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$, for some $q < p$. The likelihood ratio test procedure in this case is:

1. Fit the REDUCED MODEL (i.e. the model containing the first q predictors) using Maximum Likelihood; denote the resulting log-likelihood by $\ln L_0$.

2. Fit the model containing all of the xs, and denote the resulting log-likelihood by $\ln L_1$. This will never be less than $\ln L_0$.

3. Calculate the likelihood ratio test statistic $2 \ln \Lambda = 2 \left(\ln L_1 - \ln L_0 \right)$. If this is larger than the appropriate percentage point of a χ^2 distribution with $(p - q)$ degrees of freedom, reject the null hypothesis; otherwise accept it.

Unless the Ys are normally distributed, the χ^2 distribution here is actually a large-sample approximation. However, in climatology, most datasets are so large that the result is almost exact.

To use likelihood ratio tests, we need to decide upon a sensible hierarchy of models, because two models can only be compared if one is a special case of the other. This requires the modeller to think rather carefully about a problem before starting to test hypotheses. In climatology, there is often a natural hierarchy of models, based on our understanding of climatic processes. For example, we may be interested in investigating the effect of El Niño upon rainfall in China. We know that rainfall varies with location and season: therefore, we should account for these effects before we begin to study the effect of El Niño. A sensible strategy in this case would be to build a ‘simple’ model that accounts for seasonality and regional variation; and then to compare this with an extended model that incorporates the effects of El Niño. When interpreting the results, however, we should bear in mind the comments made in Section 1.4.2 about the appropriateness of hypothesis testing in complex systems, especially where large datasets are involved. Usually, we will need to exercise some degree of scientific judgement in such situations.

Many software packages do not output log-likelihoods directly; rather, they output deviances. As defined in the previous section, the scaled deviance for a model is defined as

$$D = 2\left(\ln L_F - \ln L\right) \ ,$$

where L_F is the likelihood for a FULL MODEL in which we set $\mu_i = y_i$, and L is the likelihood for the model under consideration. Tests based on the scaled deviance are directly equivalent to those based on the likelihood ratio statistic. However, software packages usually report the *unscaled* deviance. This is defined as $\hat{\psi}D$, where $\hat{\psi}$ is an estimate of the nuisance parameter. The reason is that, for small samples, estimation of ψ can affect the χ^2 distribution theory upon which likelihood ratio tests are based. We do not discuss this further here: it is mentioned merely to aid understanding of software output! For distributions without a nuisance parameter, such as the Poisson and Bernoulli, the scaled and unscaled deviances are identical and χ^2 testing is appropriate.

The deviance is equivalent to the residual sum of squares (RSS) in a linear regression. For this reason, procedures such as Analysis of Variance (which describes how different predictors in a regression account for the variability in the Ys) are generalised to ‘Analysis of Deviance’ in GLMs.

2.3 Introduction to case studies

We have now covered all of the necessary theoretical background. To illustrate the ideas, we next introduce four climatological case studies, and consider how the GLM approach may be applied in each case. At present, we focus on general issues such as choice of distribution and predictors.

2.3.1 Case study 1: Tropical cyclones in the North-West Pacific

The first study is an extremely simple example. It has been chosen because it provides a nice illustration of the ideas, on a small dataset. Figure 2.1 shows numbers of tropical storms in the North-West Pacific Ocean, for each year between 1959 and 2000. ‘Tropical storms’ are classified as events with a maximum windspeed above 17 ms^{-1}; ‘typhoons’ as events with a maximum windspeed over 33 ms^{-1}; and ‘intense typhoons’ as events with a maximum windspeed over 49 ms^{-1}. The intense typhoon record starts in 1972, because the Dvorak technique for assessing maximum windspeeds was not available before this. The data in this figure can be downloaded from the web site for this lecture series[1].

It is well known that the state of El Niño in any year is related to tropical storm activity in the North-West Pacific the following year. This can be seen in Figure 2.2, in which the number of typhoons in each year is plotted against the Niño 3 anomaly for each month of the preceding year.

[1] http://www.tea.ac.cn/chinese/meeting/study1/study1.html.

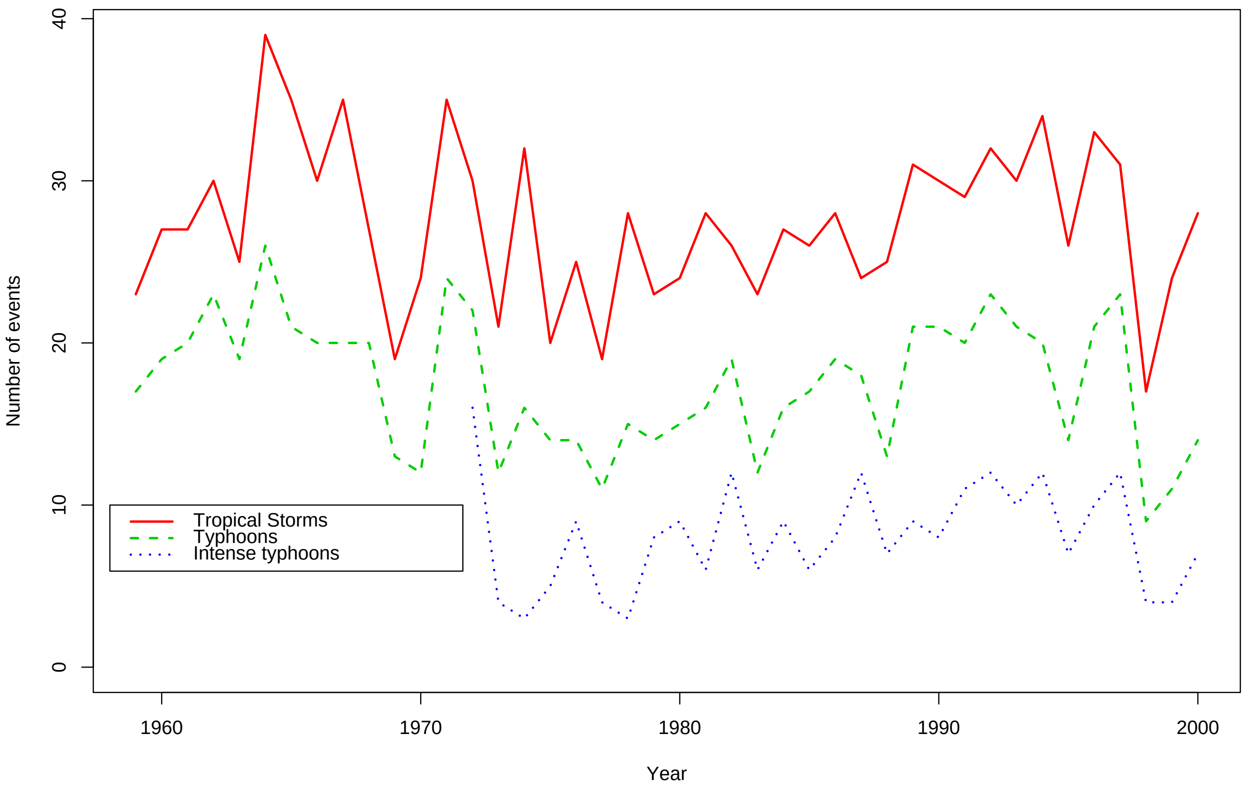

Figure 2.1: Annual numbers of tropical storms in the North-West Pacific, 1959–2000.

The sample correlation coefficients for each plot are also given. Within our probability modelling framework, these are estimates of ‘true’ underlying correlation coefficients as defined on page 16. For each plot, we have also tested the hypothesis that the ‘true’ underlying correlation is zero. The p-value on each plot is supposed to represent the probability of obtaining a correlation at least as large as the one observed, if there is no relationship. On this basis, it appears that significant relationships exist for all months between June and December. These relationships, together with others that are not considered here, may be exploited in seasonal forecasting models.

However an analysis via correlations is no more than a useful starting point. If we wish to construct seasonal forecasts, we might ask the following questions:

1. The relationship between Niño 3 values and storm numbers is clear but weak. Therefore, any seasonal forecasts that just use these relationships will be imperfect. How can uncertainty in the forecasts be recognised, while at the same time providing quantitative information?

2. Niño 3 values in successive months are highly correlated. Given this, how many months’ Niño 3 values should we include in our forecasting model?

3. Is there any evidence of systematic structure in storm numbers, that is attributable to factors

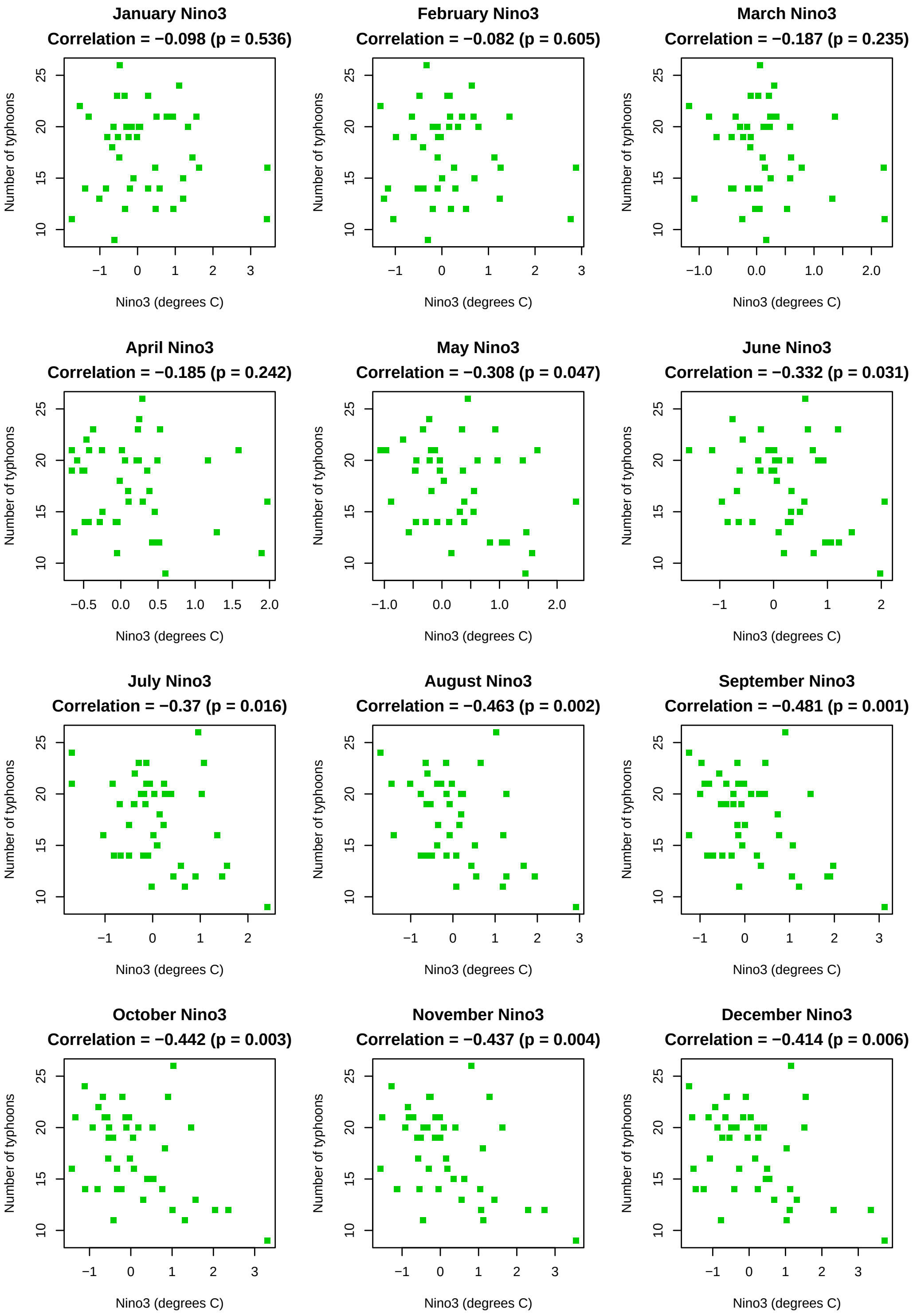

Figure 2.2: Annual typhoon numbers versus previous year’s monthly Niño 3 anomalies.

other than El Niño?

We will tackle this problem within the GLM framework. The response variable here is the annual number of typhoons, and the potential predictors are the twelve Niño 3 values from the previous year. We need to choose a suitable probability distribution for tropical storm numbers. The obvious candidate, as discussed in Example 1.7 on page 21, is the Poisson distribution, since tropical storms can be regarded as arising from a 'thinned' process of easterly waves. As a preliminary check that this is reasonable, we can plot out histograms of storm numbers. These are presented in Figure 2.3, for each of the storm categories. One very quick check on the Poisson assumption at this stage is to see whether the sample means and variances of storm numbers are approximately equal. For the 'tropical storms' category, in Figure 2.3, the variance is somewhat less than the mean; however, for the other two categories the mean and variance are similar.

In Section 2.7, we will return to this example and use the software package R to fit some GLMs.

2.3.2 Case study 2: Daily rainfall in Western Ireland

The Galway Bay area of Western Ireland (see Figure 2.4) experiences flooding every winter. However, this flooding was exceptionally severe in the winters of 1990, 1991, 1994 and 1995. Prior to the 1990s, flooding on this scale had occurred on average every 30 years. After the 1991 flood event, the Irish Government commissioned an extensive study whose aims were:

1. To assess the extent to which the flooding was caused by abnormal rainfall, rather than other factors such as changes in land use;

2. To determine whether or not rainfall patterns in the area are changing systematically; and

3. To explore a variety of engineering solutions to the flooding problem, and determine their likely effectiveness.

For this study, daily rainfall data were available from raingauges within the study area, for the period 1941–1997. To assess the likely effectiveness of engineering solutions, it was necessary to estimate the probability of large floods recurring, and to generate synthetic daily rainfall series for input to hydrological and hydrogeological models.

Figure 2.5 shows the time series of winter (December, January and February) rainfalls over the Galway Bay area. It is clear that winter rainfall was exceptionally high in the flooding years (especially 1994 and 1995), and that winter rainfalls seem to have increased in the 1990s. There is therefore systematic structure in the rainfall record, which is associated with severe flood events.

Unfortunately, this structure is difficult to detect in the daily rainfall record, which is very noisy (only 1.2% of the variability is associated with seasonality!). Ideally, any analysis of changing climate in this area would be based on data at monthly timescales or longer, to smooth out this noise. However, the need for synthetic daily rainfall series means that ultimately a study of daily rainfall

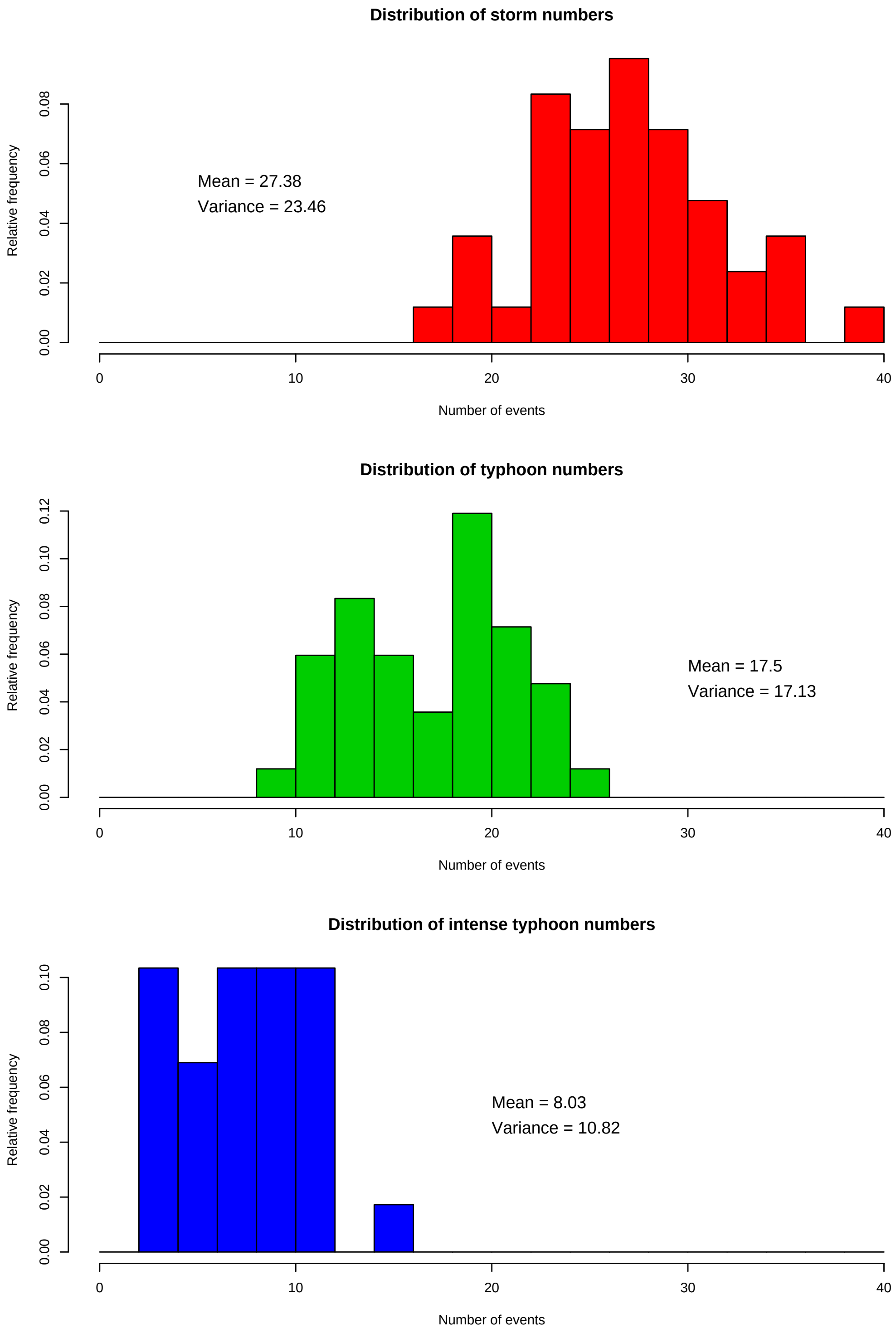

Figure 2.3: Distributions of annual numbers of tropical storms, typhoons and intense typhoons in the North-West Pacific, 1959–2000.

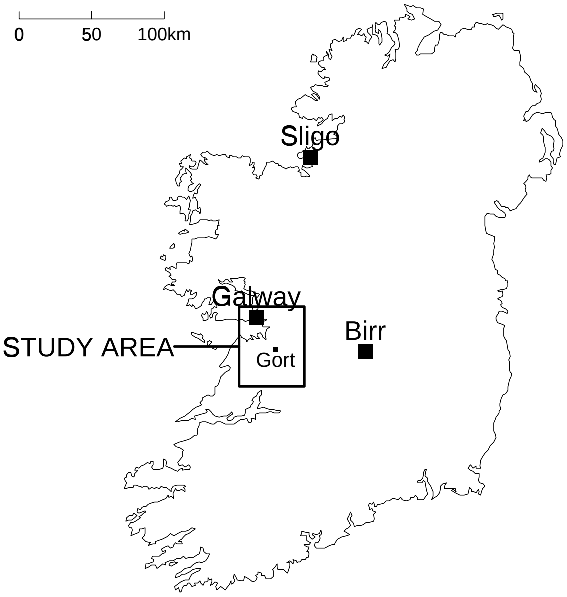

Figure 2.4: Location of the Galway Bay study area within Ireland.

is required. Generalized Linear Models are able to identify weak signals in noisy data (because they use likelihood methods in a probability-based framework), and are therefore particularly appropriate for this kind of problem.

The modelling of daily rainfall is complicated by the fact that the distribution has a mixture of discrete and continuous components. Within the GLM framework, this can be overcome by fitting *two* models. The first is used to predict the probability that rain will occur at a site on a given day, and the second to predict a distribution for the amount of rain if non-zero.

To model the probability of rain, it is natural to use logistic regression (see Example 2.2 above). It is harder to choose a suitable distribution for the amount of rain on wet days — there is no 'obvious' physical mechanism that suggests a suitable candidate. However, the gamma family of distributions provides a flexible class of models for positive-valued random variables (see Figure 1.3), and so it is natural to work with this family.

At this point, we should recall from Section 2.2 that when using GLMs we usually assume that the nuisance parameter ψ is common to all observations. For the gamma distribution, $\psi = \nu^{-1}$ (see Table 2.1). In Section 1.3.5, we saw that the coefficient of variation of a gamma distribution is $\nu^{-1/2}$. Hence, if we use a gamma GLM with constant ν, we are assuming a common coefficient of variation for all of the observations. Before we start, we should calculate the coefficient of variation for subsets of the data (e.g. for each month, or each site), to check that this assumption is reasonable. For this particular dataset, the coefficient of variation in daily rainfall does indeed

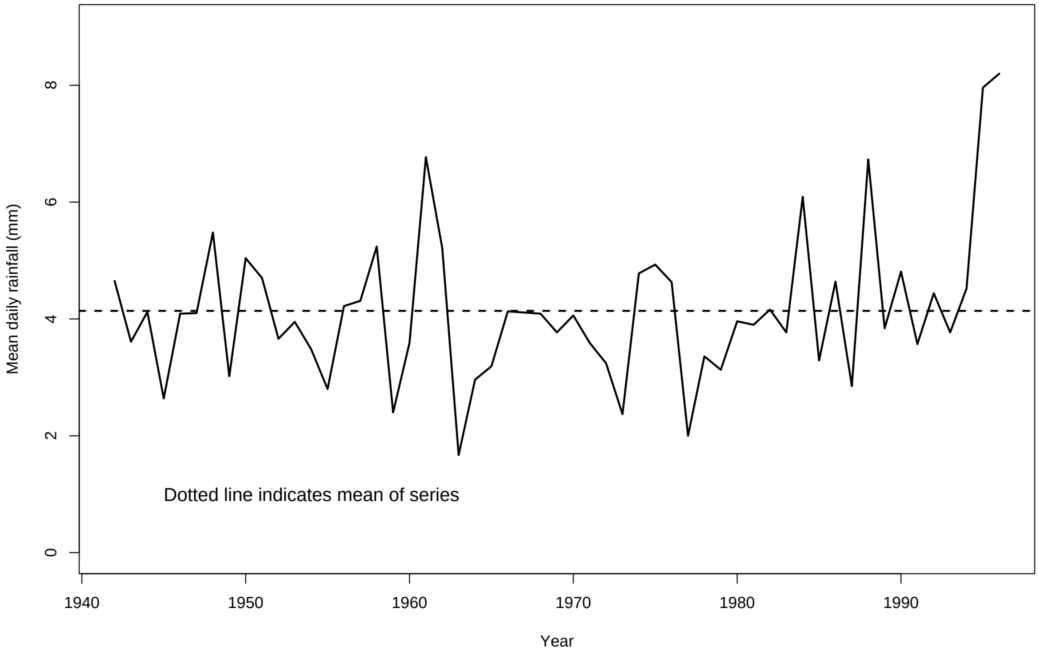

Figure 2.5: Mean daily winter rainfall amounts for Galway Bay area, 1941–1995.

appear constant over months and between sites. Indeed, this is a feature that is often observed in daily rainfall data, from all over the world (providing the area of interest is not too large). There is no obvious physical explanation for this, but it is very convenient!

Predictors which may affect daily rainfall include seasonality, altitude and ‘external’ factors such as the North Atlantic Oscillation (NAO). Some of these predictors may have variable effects — for example, the dominant impact of the NAO is known to be in the winter months. Finally, we mention that typical daily rainfall sequences are autocorrelated in time, so that the individual Y values in a GLM cannot be regarded as independent (this is usually required, in order to write down and maximise a likelihood). We return to these points in Section 2.4 below.

2.3.3 Case study 3: Mean monthly temperatures in the USA

Our third case study uses GLMs to study climate at a continental scale. The aim is to develop a model for monthly mean temperatures at any location in the USA. Temperature data for the period 1948–1997 are available from 2600 weather stations. The dataset has a total of 1,560,000 observations.

It is natural to use the normal distribution to model monthly mean temperatures, because of the Central Limit Theorem (see Section 1.3.4). A GLM using the normal distribution is just a standard multiple regression model. By incorporating predictors representing seasonal effects and regional variability, we can specify different normal distributions for each observation in the dataset.

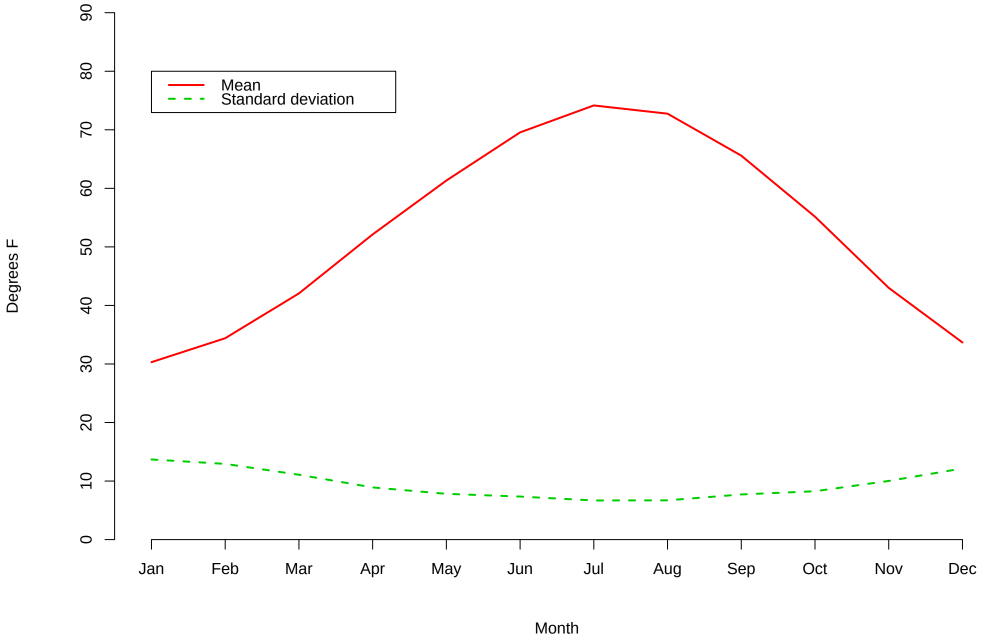

Figure 2.6: Seasonal variation of monthly temperature statistics in the USA.

There is a problem, however. As described in Section 2.2 above, in GLMs it is usually assumed that the nuisance parameter is constant for all observations. The validity of this assumption should be checked before starting to develop models. Simple plots are sufficient to demonstrate that the assumption does not hold. For example, Figure 2.6 shows that the standard deviation of temperatures is lower in the summer months than in winter. Similar plots show that the standard deviation also varies with factors such as latitude.

One approach to this problem would be to work with standardised temperature anomalies rather than actual temperatures. This technique is commonly used in climatology. In this particular case we might calculate, for each site, a separate mean and standard deviation for each month of the year; then transform each data value using these means and standard deviations, so that all systematic seasonal and regional variability has been removed.

From the viewpoint of probability modelling, the idea of working with anomalies is not altogether satisfactory, and may lead to under-representation of uncertainty. This is because the approach is effectively fitting a model with a separate mean and variance for every single site and month — in this case, when we have 2600 sites, this 'model' has $2600 \times 12 = 31200$ parameters representing mean temperature structure, and a further 31200 parameters representing variance structure. From a statistical viewpoint, this means that 62400 parameters have been estimated! Without a proper analysis, we do not know what impact this will have upon our conclusions, but it is an issue that should be investigated.

Fortunately, within the GLM framework it is possible to deal with changes in both mean and variance in an elegant manner, at least when we are working with normal distributions. The approach relies upon a number of results from Lecture 1: the variance of any distribution is the expected squared deviation from its mean (Section 1.2.3); any normal random variable can be transformed to a standard normal random variable by suitable scaling (Section 1.3.4); the square of a standard normal random variable is distributed as χ^2_1 by definition; and the χ^2_1 distribution is equivalent to the gamma distribution with shape parameter equal to $1/2$ (Section 1.3.5). To summarise: If $Y \sim N(\mu, \sigma^2)$ then $(Y - \mu)^2 \sim \Gamma\left(\frac{1}{2}, (2\sigma^2)^{-1}\right)$. As a result of this, we can model both the mean and the variance of monthly temperatures using a combination of models. We use a normal model to specify

$$\mu_i = E\left(Y_i\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} \ ,$$

and use a gamma model with known shape parameter to specify $\sigma_i^2 = E\left((Y_i - \mu_i)^2\right)$. In the GLM framework, σ_i^2 is allowed to depend on various other predictors, $\xi_{1i}, \ldots, \xi_{qi}$ say (which may or may not be the same as the xs used in the model for μ_i) in such a way that

$$h\left(\sigma_i^2\right) = \gamma_0 + \sum_{j=1}^{q} \gamma_j \xi_{ji} \ ,$$

where the γs are a further set of coefficients to be estimated, and $h(.)$ is a monotonic function. In this case, we have chosen $h\left(\sigma_i^2\right) = \sigma_i \Rightarrow h(x) = \sqrt{x}$, on the basis of other preliminary analyses which are not reported here.

This modelling approach has the disadvantage of being computationally intensive. This is because the Maximum Likelihood estimates of the βs in the model for μ_i depend on the values of σ_i^2 for all cases in the dataset (in fact, the βs are estimated by Weighted Least Squares in this case, where the weights are inversely proportional to the associated variances). However, in order to specify the value of σ_i^2 we need to know the values of the γs. These can only be estimated by fitting a model to the squared errors $\left\{(Y_i - \mu_i)^2 : i = 1, \ldots, n\right\}$, which are themselves unknown because we don't know the values of the μs! An iterative approach to model fitting is required. The algorithm, which yields Maximum Likelihood estimates of all parameters, is:

1. Start by assuming that all variances are equal.

2. Use Weighted Least Squares to estimate the βs in the mean part of the model. The weight for the ith case is $\hat{\sigma}_i^{-2}$, where $\hat{\sigma}_i^2$ is the current estimate of the variance.

3. Use the estimated βs to calculate the estimated means $\{\hat{\mu}_i : i = 1, \ldots, n\}$. Calculate the corresponding squared errors $\left\{(Y_i - \hat{\mu}_i)^2 : i = 1, \ldots, n\right\}$, and estimate the γs from these using a gamma GLM with shape parameter fixed at $\nu = \frac{1}{2}$.

4. If the change in all parameter estimates is 'small', stop; otherwise go back to step (2).

This procedure is non-standard, and some programming is required to implement it in any software package. However, the extra effort is worthwhile since all the usual benefits of GLMs are available to us — likelihood ratio tests for determining the significance of predictors, confidence intervals for every parameter in the model, and a completely specified probability distribution for each observation. The resulting model is far more parsimonious than that underlying the 'climate anomaly' approach (the model we discuss in Lecture 3 contains 139 parameters instead of 62400), and can be used to specify probability distributions for monthly temperatures at locations where no data have been observed.

2.3.4 Case study 4: Daily Maximum Windspeed in the Netherlands

The final case study is similar to Case Study 2 above. The aim is to provide data and some specimen analyses, for a realistic climatological example (in contrast to Case Study 1, which is rather simple).

The study is concerned with windspeeds in the Netherlands. Large areas of this country are below sea level, and are protected by a system of dikes. These dikes are continually attacked by waves, and it is necessary to continually monitor the risk of dike failure. The main risk of dike failure occurs during high winds, since these cause large waves. It is therefore of particular interest to study extreme windspeeds.

For this study, we use daily windspeed data from 9 sites, each of which have continuous records from 1961-1998. Each daily value is the largest of 4 instantaneous hourly values at times 0600, 1200, 1800 and 2400, and is referred to as the *Daily Maximum Windspeed* (DMWS). The sites are listed, together with some summary statistics, in Table 2.2. Bubble maps, showing the magnitude of the statistics for each location, are given in Figure 2.7 (the larger the circle, the larger the value being represented). The data have been strictly pre-processed by the Royal Netherlands Meteorological Observatory, to remove inhomogeneities, and are of extremely high quality. They can be downloaded from the web site for this lecture series[2].

The daily statistics in Table 2.2 (including the mean) are self-explanatory. The annual standard deviation is computed from the time series of annual mean DMWS at each site, and provides a background against which the values in the 'Decadal trend' column can be judged. These values are obtained by fitting a straight line through annual time series plots at each site using linear regression; they represent the decadal change in mean windspeed according to this fitted line.

Once again, such procedures are only a first step in a full analysis. Summary statistics can only give approximate indications of structure, since they ignore all of the other factors that affect windspeeds. However, they are useful for exploratory purposes. For example, Figure 2.7 shows a lot of spatial structure in all of the summary statistics — windspeeds near the coast are higher, and more variable, than those inland. Also, windspeeds at coastal sites appear to have increased between 1961–1998, while those at inland locations have decreased. This pattern is not due to

[2] `http://www.tea.ac.cn/chinese/meeting/study4/study4.html.`

Site name	Mean	Daily standard deviation	Daily coefficient of variation	Annual standard deviation	Decadal trend
Ijmuiden	8.44	3.36	0.40	0.38	0.28
Schiphol	7.24	3.08	0.43	0.35	-0.06
Soesterberg	6.05	2.50	0.41	0.33	-0.06
Eindhoven	6.37	2.74	0.43	0.43	-0.30
De Bilt	5.68	2.39	0.42	0.37	-0.09
Deelen	6.61	2.74	0.41	0.35	-0.24
Eelde	6.51	2.75	0.42	0.37	0.20
Vlissingen	7.23	3.00	0.41	0.34	0.09
Gilze Rijen	6.27	2.61	0.42	0.35	-0.15

Table 2.2: Summary statistics for DMWS series (ms^{-1}) from 9 sites in the Netherlands.

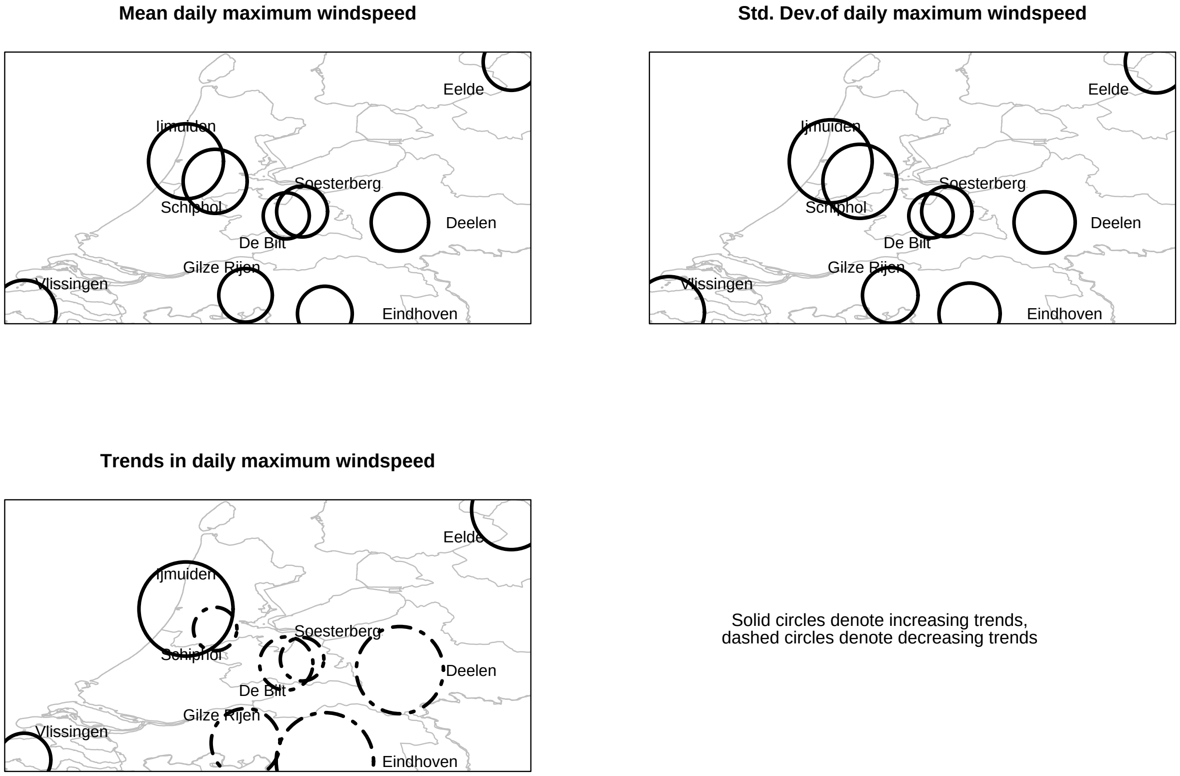
Mean daily maximum windspeed
Std. Dev.of daily maximum windspeed
Trends in daily maximum windspeed
Eelde
Ijmuiden
Soesterberg
Schiphol
Deelen
De Bilt
Gilze Rijen
Vlissingen
Eindhoven
Solid circles denote increasing trends,
dashed circles denote decreasing trends

Figure 2.7: Regional variation of summary statistics for DMWS in the Netherlands.

inhomogeneities in the station records (a comparison with NCEP windspeed data over Europe supports this). What is interesting here is the contrast in DMWS regimes over a small area.

If we wish to analyse this dataset within the GLM framework, we need to specify a suitable family of probability distributions for DMWS. Historically, windspeeds have often been modelled using the Weibull distribution, but the motivation for this is weak (see Section 1.3.6). Moreover, the Weibull is not a member of the exponential family of distributions (see Section 2.2.1 above), so that standard numerical techniques cannot be applied directly to maximise the likelihood in this case[3]. Among the distributions commonly used for Generalized Linear Modelling, the gamma is the most natural candidate here, since it offers a flexible range of shapes for dealing with positive-valued variables. To use gamma GLMs, we need to assume that the coefficient of variation is constant as in Case Study 2 above. Table 2.2 shows that this assumption appears reasonable, at least across different sites.

To predict a probability distribution for DMWS we might consider using predictors representing seasonal variability, location effects and long-term trends (possibly represented via 'external' factors such as the North Atlantic Oscillation, which are expected to control some aspects of European climate). However, we need to allow trends to vary systematically between sites.

2.4 Common features of climate-related problems

These case studies illustrate a variety of features commonly encountered in climatological problems. We now summarise these, and discuss how they may be dealt with in the GLM framework.

2.4.1 Autocorrelation in time

Inference for GLMs is carried out using Maximum Likelihood. We therefore need to be able to write down a realistic joint density for the observations, as defined in Section 1.4.3. If the observations are all independent, this is straightforward since the joint density is a product of individual terms. However, the assumption of independence does not hold for many climatological time series. This is especially true for daily series such as those in Case Studies 2 and 4. Suppose, then, that our data vector $\boldsymbol{y}$ arises as a time series. In this case, the Generalised Multiplication Law (page 11) tells us that the corresponding joint density can be written, in an obvious notation, as $f\left(\boldsymbol{y};\boldsymbol{\theta}\right) = f_1\left(y_1;\boldsymbol{\theta}\right) \times f_2\left(y_2|y_1;\boldsymbol{\theta}\right) \times \ldots \times f_n\left(y_n|y_1,y_2,\ldots,y_{n-1};\boldsymbol{\theta}\right)$ i.e. as a product of *conditional* densities. The log-likelihood is then

$$\ln L(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \ln f_i\left(y_i|y_1,\ldots,y_{i-1};\boldsymbol{\theta}\right) \ .$$

The point of this is that the log-likelihood can be written as a sum of terms. However, each term involves the density of the corresponding observation *given all previous observations*. Within the

[3] In fact, there are algorithms for fitting Weibull GLMs, but the theory is complex and will not be covered here.

GLM framework, this can be dealt with very straightforwardly. All we have to do is to include previous observations as predictors in the model.

Often, we will find that the inclusion of previous observations into a GLM has a dramatic effect on other terms in the model, and upon our assessment of their significance. Indeed, it is not unusual for previous observations to dominate a model completely, in terms of measures such as variance explained. In models for daily climate time series, this merely reflects the fact that variability is dominated by weather scale fluctuations. Usually, when we incorporate previous observations into a GLM, the result will be a reduction in the number of other factors that are deemed to be significant. This may lead to the suspicion that previous observations are somehow obscuring some genuine relationships. This is not the case — likelihood techniques are typically able to detect weak signals in noisy data, at least with the size of dataset usually encountered in climatology. We may be confident that by including previous observations into our GLMs, genuine relationships will be identified and spurious ones will be discarded. Such spurious relationships can arise very easily — for example, if two unrelated time series both have a lot of internal structure, they may appear to be related simply because they both show long runs of high or low values.

2.4.2 Interactions

In climate processes, it is common to find predictors whose effects vary with the values of some other variable (representing, for example, location or time of year). We have seen this, for example, in Case Study 4 — windspeeds have increased in some parts of the Netherlands over the last 40 years, but reduced in others.

To simplify the discussion, suppose that there are two predictors x_1 and x_2, and that these are related to $\mu = E(Y)$ in such a way that $g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. However, the value of β_2 itself may depend on the value of x_1, such that $\beta_2 = \gamma_0 + \gamma_1 x_1$. In this case, we have

$$g(\mu) = \beta_0 + \beta_1 x_1 + (\gamma_0 + \gamma_1 x_1)\, x_2 = \beta_0 + \beta_1 x_1 + \gamma_0 x_2 + \gamma_1 x_1 x_2 \ .$$

This looks exactly like our original model except that now we have three predictors (x_1, x_2 and $x_1 x_2$) instead of two. The predictors x_1 and x_2 are said to INTERACT. The extra predictor $x_1 x_2$ is an INTERACTION TERM in the model. The idea can be extended to deal with more complex interactions. For example, the regional difference between trends in the Netherlands may be more pronounced in the winter than the summer. This would be represented by 3-WAY INTERACTIONS between predictors representing trends, location effects and seasonality. To include interactions in a model, we just need to add terms corresponding to products of the relevant predictors.

2.4.3 Spatial dependence

Climate datasets often take the form of a number of time series from different spatial locations as in Case Studies 2, 3 and 4. In such cases, we need to address the issue of spatial dependence. We distinguish two forms of such dependence, as follows:

Systematic regional variation: In Case Study 3, average temperatures in the Northern USA will be cooler than those in the South. Similarly, in Case Study 4 we saw that windspeeds are highest in coastal regions of the Netherlands. These are systematic features of regional climate, and can be regarded as 'average' effects of location. They can be incorporated into a GLM by including predictors whose values depend upon spatial location. Typically, these predictors will be nonlinear functions of quantities such as latitude and longitude. The problem of how to specify such functions will be discussed in the next section.

In a sense, systematic regional variation is not really 'spatial dependence', although the term is often used to describe it (presumably because it refers to 'dependence upon spatial location').

Spatial autocorrelation: As a result of systematic regional variation, data from two nearby sites will tend to be similar on average, simply because they are close to each other. However, the fluctuations about the sites' average values will also tend to be similar, as a result of the mechanisms generating the data. For example, in Case Study 2 we might expect rainfall amounts to increase, on average, with altitude. However, on any given day we would expect rainfall amounts to be higher, or lower, than average across the whole area. This is because over an area of this size, daily rainfall is dominated by the effects of weather systems — an intense frontal system will produce a lot of rain everywhere, regardless of altitude. The effect is to induce *spatial autocorrelation* into a dataset.

One effect of spatial autocorrelation is that data from different sites at the same time point cannot be regarded as independent. Spatial autocorrelation is much more difficult to deal with than autocorrelation in time, however, since in this case it is not usually possible to write down a simple factorisation of the likelihood (the problem is that there is no natural ordering of sites). This issue is the subject of a lot of current research in statistics — most of the available procedures are too computationally expensive for use with large datasets such as those found in climatology . It can be shown, however, that if we ignore the spatial autocorrelation and treat sites as independent, our parameter estimates will be extremely close to the exact Maximum Likelihood estimates, so long as a sufficiently long record is available from each site. This is therefore the procedure we will use. We should bear in mind, however, that hypothesis tests and confidence intervals obtained under the assumption of independence will be incorrect. As a rough guide, for the levels of spatial autocorrelation in the datasets considered here, the 'independence' standard errors of parameters may be underestimated by a factor of between 1.5 and 2.5. Hence some informal judgement is required to decide whether the effect of a predictor is 'significant' in the presence of spatial autocorrelation.

From now on, when we talk about 'spatial dependence' we will be referring to spatial autocorrelation rather than systematic regional variation.

2.4.4 Nonlinearities

It is common to encounter situations where a response variable is associated with some predictor, but where the relationship is actually with a nonlinear transformation of the predictor. Examples include the investigation of possible long-term cycles in the climate of an area (where the fundamental predictor is a time index, but a cyclical pattern implies that the relationship is really with a sine wave derived from the time index), and the realistic modelling of systematic regional effects (where the underlying predictors might be latitude and longitude, but systematic variability cannot be represented by including these directly into a GLM — for example, in Case Study 3, it is unlikely that average temperatures in the USA vary as a linear function of longitude!).

Such nonlinear transformations may be divided into two categories, as follows:

Category 1: in this case, there is an obvious parametric form for the transformation. The example of fitting a cyclical trend function falls into this category. In such a case, given an underlying predictor x, we wish to include in the model a term of the form $h\left(x, \boldsymbol{\phi}\right)$, where $h(.)$ is a known function and $\boldsymbol{\phi}$ is a vector of parameters in the transformation. Usually these parameters are unknown and must be estimated along with all of the βs in the model. This can be achieved using an extension of the usual iterative weighted least squares algorithm. Unfortunately, this feature is not implemented in many software packages.

Category 2: in this case, there is no obvious way in which a nonlinear transformation may be parametrised. The modelling of systematic regional variation in temperatures across the USA is an example of this. We suggest adopting a nonparametric approach. Specifically, suppose we wish to represent the unknown transformation $h(.)$ over the interval (a, b). Let $\{\zeta_j : j = 0, 1, 2, \ldots\}$ be an ORTHOGONAL BASIS of functions i.e. a collection satisfying

$$\int_a^b \zeta_j(t)\zeta_k(t)dt = \begin{cases} 1 & j = k \\ 0 & \text{otherwise.} \end{cases}$$

Then $h(x)$ can be expressed, over the interval (a, b), as the infinite sum

$$h(x) = \sum_{j=0}^{\infty} A_j \zeta_j(x) \ .$$

for some set of coefficients $\{A_j : j = 0, 1, 2, \ldots\}$.

In practice, providing the ζs are chosen intelligently, most of the coefficients A_j will be very small and can be neglected, so that $h(.)$ can be represented to a very good degree of approximation using a small finite collection of ζs. This procedure reduces a highly nonlinear dependency into linear dependence upon a set of known functions: if we use ζs directly as predictors in the GLM, the coefficients $\{A_j\}$ will appear as βs. They can therefore be estimated, and the problem is reduced into linear form.

Orthogonality of the basis functions is not required for this approach to work. However, if the data points are scattered approximately uniformly over the range (a, b), then an orthogonal basis will produce predictors which are approximately uncorrelated. As a consequence, the coefficient of any of the ζs will not be greatly affected by the presence or absence of other terms in the model.

The disadvantage of orthogonal series representation is that it may be quite parameter-intensive, as several ζs may be needed to obtain an adequate representation of an effect. This problem can be minimised by careful selection of basis functions. For example, if a transformation is likely to be essentially monotonic, it might be represented efficiently using a polynomial basis such as Legendre polynomials. Effects which are more oscillatory may be represented more parsimoniously using Fourier series. Both of these bases have $\zeta_0 = 1$, so that the coefficient of ζ_0 is estimated as part of the constant term (β_0) in any GLM.

A straightforward extension of these arguments shows that we can represent the combined effect of two variables (such as latitude and longitude) in a similar way. To do this, we simply need to add ζs for latitude and longitude, and interactions between them.

There is one potential pitfall when using orthogonal series to model regional effects with few sites. If the total number of ζs and their interactions approaches the number of sites, there is a danger of severely overfitting the model to match exactly the observed means at each site. As a general rule, the total number of site effects in the model (including interactions) should be kept well below the number of sites available.

2.5 Model checking

Having fitted any model, we need to check it. This is an area which often does not receive the attention it deserves. However, for GLMs there are a few simple but informative checks that should be carried out routinely. All of the techniques rely on analyses of model residuals, so we start by discussing these.

2.5.1 Residuals

In a GLM, we regard each observation as coming from a different probability distribution. Potentially, this makes it difficult to check models directly. However, it is always possible to transform the data in such a way that, if the fitted model is correct, all of the transformed values come from distributions with the same properties. It is natural to consider transformations representing some measure of 'error'. A few of the more commonly-used measures are:

Pearson residuals: If, for the ith case in the dataset, we forecast some probability distribution with mean μ_i and standard deviation σ_i, then the PEARSON RESIDUAL for this case is

$$r_i^{(P)} = \frac{y_i - \mu_i}{\sigma_i} \ ,$$

where y_i is the observed value. If the fitted model is correct, all Pearson residuals come from distributions with mean 0 and variance 1. Pearson residuals are usually simple to interpret.

Variations on this theme are possible. For example, for a gamma distribution we have $\sigma_i = \mu_i/\sqrt{\nu}$ (see Section 1.3.5), so that $r_i^{(P)} = \sqrt{\nu}\,(y_i - \mu_i)\,/\mu_i$. However, if ν is the same for all cases in the dataset we may prefer to use the definition $(y_i - \mu_i)\,/\mu_i$ instead, since this is just the proportional error which is more directly interpretable.

Anscombe residuals: In some applications, it may be useful to define residuals that all come from the same *normal* distribution if the model is correct. Such measures are called ANSCOMBE RESIDUALS. They do not always exist (for example, if the Ys are discrete it is not possible to define residuals that have a continuous distribution).

Often, Anscombe residuals are defined to have an approximate, rather than exact, normal distribution. For example, for a gamma distribution with mean μ the Anscombe residual may be defined as $(y/\mu)^{1/3}$. The distribution of this quantity is not exactly normal, but it is usually extremely close. The mean and variance of the approximating normal distribution depend only on ν. Therefore, if ν is common to all observations, the Anscombe residuals all come from the same normal distribution.

Deviance residuals: In Section 2.2.1 above, we saw that deviance is equivalent to the residual sum of squares (RSS) in a linear regression model. Since RSS is just a sum of terms of the form $(y_i - \mu_i)^2$ (i.e. of squared errors), we might consider trying to write the deviance as a similar sum of squares: $\sum_{i=1}^{n}\left(r_i^{(D)}\right)^2$, say. Since the deviance is a difference of log-likelihoods, each of which is expressed as a sum of contributions from each observation, definitions of deviance residuals can be derived by inspection of the log-likelihood.

Deviance residuals are often difficult to interpret. However, they are often provided by statistical software packages. Since the scaled deviance is supposed to have a χ^2 distribution, we might expect deviance residuals to have an approximate normal distribution. However, this approximation is often quite poor.

We now discuss how residuals can be used to check GLMs.

2.5.2 Checks on forecast probability distributions

The GLM framework deals with uncertainty in a response variable by specifying a probability distribution conditional on the values of predictors. Since parameter estimates are obtained via Maximum Likelihood, we need to check that the chosen family of distributions is realistic. Correct specification of the forecast distributions is also important if the fitted models are subsequently to be used in simulations, particularly if extreme events are of interest.

For continuous response variables, the easiest way to check the form of the forecast distribution is via quantile-quantile plots of residuals. For instance, for distributions where Anscombe residuals are defined we could produce a normal probability plot — a straight line on such a plot indicates that the distributional assumptions are satisfied. Most statistical packages will produce normal probability plots very easily.

For discrete random variables, things may be more difficult. However, we can use the relative frequency interpretation of probability to construct some sensible test procedures. For example, let $\boldsymbol{Y} = (Y_1 \ldots Y_n)'$ be a vector of discrete random variables, each of which can take the value 0. Suppose we fit a GLM to $\boldsymbol{Y}$. We can use this calculate $P\left(Y_i = 0\right) = p_i$, say, for each i. We expect to observe $Y_i = 0$ on a total of $\sum_{i=1}^{n} p_i$ occasions. A comparison of the observed and expected numbers of zeroes provides a check on the probability structure of the model. Of course, this procedure should then be repeated for all other values in the dataset. We will illustrate this procedure with reference to Case Study 1, in Section 2.7.

For very simple distributions, we can be more thorough. Consider, for example, the case when the Ys are all Bernoulli random variables (see Section 1.3.1). Suppose we examine all cases for which the forecast probability $P\left(Y_i = 1\right)$ is close to some value p^*. We expect to observe a value of 1 in a proportion p^* of these cases. Unless the observed and expected proportions of 1s agree across the whole range of forecast probabilities, there is something wrong with the probability structure of the model. In practice, we implement this idea by dividing the forecast probability range into suitable intervals e.g. $(0.0, 0.1), \ldots, (0.9, 1.0)$. We will illustrate this when we check the logistic regression model for rainfall occurrence in Case Study 2.

2.5.3 Checks for unexplained patterns

As well as checking the probability structure of a GLM, we need to check that all of the relationships between variables in a dataset have been accounted for correctly. In statistics, such checks are usually carried out by plotting residuals against values of the linear predictor (see page 35), and against individual covariates. Such plots may be produced both for covariates which appear in the model, and for factors that have not been included. Any apparent structure in these plots indicates a problem with the model. Note that a plot of residuals against observed values is *not* informative, and will usually show some structure even if the model is correct.

A typical feature of climatological datasets is their large size. If there are many points on a residual plot, it can be difficult to interpret. In this case, it may be better to focus on summary statistics for residual measures over subgroups of observations. For example, to check that seasonality is well reproduced we can compute the mean and root mean squared error of Pearson residuals for each month of the year, and plot these: any pattern in the plot, or values which are 'significantly' different from zero, indicate seasonal structure which has not been captured.

To aid visual interpretation of such plots, it is helpful to include approximate confidence bands, indicating the range within which mean residuals are expected to lie if the model is correct. If mean residuals are computed from a network of sites, it may be necessary to adjust the width of these confidence bands to account for spatial dependence.

2.6 Interpreting models

Generalized Linear Modelling is, at its most basic level, a descriptive technique for summarising relationships between predictors and response variables. However, it is also possible to interpret the models, so as to draw meaningful conclusions about the mechanisms that generated the data.

The most obvious way in which we can interpret a GLM is by examining the coefficients of the predictors. Depending on the link function used, it may be possible to infer the effect of a particular predictor upon the mean of the response distribution. For example, if the link function is the identity (as in linear regression), β_i represents the average effect upon Y, of a unit increase in the ith predictor. For models with a log link, e^{β_i} is the average multiplicative effect of the ith predictor.

This idea can be extended, to build up pictures of nonlinear effects where these have been represented nonparametrically (as described in Section 2.4.4 above). Suppose, for example, that we have represented systematic regional variations using orthogonal basis functions of latitude and longitude. If we extract all of these predictors from a model, together with their associated βs, then we obtain a function that can be evaluated at any spatial location. A map of this function shows us the systematic regional variation in the mean of the response. A similar idea allows us to study the effects of large-scale climate indices by extracting, and mapping, all terms in a model that represent their interactions with location effects. This will be illustrated in Case Study 3, in the next lecture.

Finally on the subject of interpretation, we should not overlook the standard errors associated with each parameter in a model. If one or two parameters in a model have large standard errors, we know that they have not been estimated very precisely. This means that the available data do not contain much information about these parameters, which is potentially useful knowledge.

2.7 Worked example — Case Study 1

To conclude this lecture, we work through the first case study above, using the free statistical package `R` to perform the analysis. `R` is a computer language, designed for easy implementation of a wide range of advanced statistical and graphical procedures. The language is object-oriented, and can be used either by typing commands at a prompt or by running `R` programs. Appendix A.1 gives details of how to obtain the package.

2.7.1 Reading the data

The data for this case study can be obtained from the web site for this lecture series, as described in Section 2.3.1 above. There are two data files. File `nstorms.dat` contains annual counts of tropical storm numbers in the North-West Pacific, and file `nino3.dat` contains monthly Niño 3 index values. In addition to these files, the `R` program `TC_anal.r` may be downloaded. The simplest way to analyse these data is to download these three files to the same directory, start up R in this directory, type source("TC_anal.r") at the prompt, and then type q() to quit R . However, the only thing this will teach you is how to run R programs! We will start by looking at some of the simple commands in this program.

The first few lines begin with the "#" character. These are comments, and will be ignored by R. The first commands in the program are

```
storm.data <- read.table(file="nstorms.dat",header=T)
attach(storm.data)
```

The first line here reads the data file nstorms.dat, and stores the contents in an object called storm.data. To view the contents of any R object, just type its name at the prompt. If we just want to view the first 4 rows of storm.data, we type

```
storm.data[1:4,]
```

and obtain

```
  Year Storms Typhoons Intense
1 1959     23       17      NA
2 1960     27       19      NA
3 1961     27       20      NA
4 1962     30       23      NA
```

The columns are automatically named. R has picked up the column names from the data file as a result of the "header=T" part of the read.table command. The NA values in the Intense column are used by R to denote missing data (recall from Section 2.3.1 that we have not used intense typhoon numbers from any year before 1972).

The command attach(storm.data) is used to tell R to search the variable names of storm.data, as well as the index of R objects, when we refer to an object. For example, if we now type Intense at the R prompt, R will first look for an object called "Intense". If it does not find one, it will look at the column headings in storm.data, and identify the fourth column. Alternative ways of specifying this are storm.data$Intense, and storm.data[,4]. The attach() command provides a convenient way of referencing parts of objects.

2.7.2 Simple plots

The next few lines of the program demonstrate how to produce a simple plot (Figure 2.1) in R . Following these, we have the lines

```
custom.hist <- function(x,breaks,colour,name,statloc) {
                        .
                        .
                        .
}
```

This section of code is a *user-defined function*. Its purpose is to produce a customized histogram of the data in an object `x`, according to the values of various options specified in the arguments `breaks`,`colour`,`name` and `statloc`. The next four lines:

```
par(mfrow=c(3,1))
custom.hist(Storms,seq(0,40,2),2,"storm",c(5,0.05))
custom.hist(Typhoons,seq(0,40,2),3,"typhoon",c(30,0.05))
custom.hist(Intense,seq(0,40,2),4,"intense typhoon",c(20,0.05))
```

can then be used to generate Figure 2.3. The command `par(mfrow=c(3,1))` is used to put 3 plots on a single page.

Having produced the histograms, the program reads the Niño 3 data and merges this with the `storm.data` object. This can be done because the data files `nino3.dat` and `nstorms.dat` both have a column headed `Year`. To match the storm numbers to the *previous* year's Niño 3 values, we add 1 to the Niño 3 `Year` column before merging.

The next step is to produce Figure 2.2. Note the use of `par(mfrow=c(4,3))`, to produce a 4×3 array of plots on a page: also the use of a loop (`for (i in 1:12)` { `...` }) to produce plots for each month; and the automatic labelling of each plot using month names, which are defined in the object `monthlabs`.

2.7.3 Fitting a GLM

Now for the interesting part. In Section 2.3.1, we suggested that the Poisson distribution might be appropriate for modelling storm numbers. Here, we will only consider modelling typhoon numbers — analyses of tropical storms and intense typhoons can be carried out in a similar way. On the basis of Figure 2.2, typhoon numbers appear to be more strongly associated with Niño 3 values in September than in any other month. At this point in the program, the September Niño 3 values are held in the `N3.m09` column of `storm.data`. Let Y_i be the number of typhoons in year i, and let x_i be the September Niño 3 value from the previous year. To fit a Poisson GLM with a log link function (see Section 2.2), we type

```
glm(Typhoons ~ N3.m09,family=poisson(link="log"))
```

This fits the model $Y_i \sim Poi\left(\mu_i\right)$, where $\ln \mu_i = \beta_0 + \beta_1 x_i$. The output is

```
Call:  glm(formula = Typhoons ~ N3.m09, family = poisson(link = "log"))

Coefficients:
(Intercept)       N3.m09
     2.8740      -0.1262

Degrees of Freedom: 41 Total (i.e. Null);  40 Residual
Null Deviance:      41.3
Residual Deviance: 31.63        AIC: 232.2
```

The parameter estimates are $\hat{\beta}_0 = 2.8740$ and $\hat{\beta}_1 = -0.1262$. The deviance for this model (again, see Section 2.2) is 31.63. In general, R outputs the deviance rather than the scaled deviance, as discussed in Section 2.2.1 — however, for the Poisson distribution, these quantities are the same. The *null deviance* is the deviance for a model containing no predictors. Recall that, in a GLM the deviance is the equivalent of the residual sum of squares in a linear regression. In this case, the deviance is reduced from 41.3 to 31.63 — the ‘percentage of deviance explained’ is $100(41.3 - 31.63)/41.63 = 23.4$.

We may want to do rather more with this model — for example, we may want to obtain standard errors for the parameter estimates, and perform some model checks. This is straightforward in R . All we need to do is to store the model in an object, which we can then interrogate:

```
storm.model1 <- glm(Typhoons ~ N3.m09,family=poisson(link="log"))
```

The object `storm.model1` now holds all of the information about the fitted model. If we want some more detail, we can type `summary(storm.model1)`[4] to obtain

```
Call:
glm(formula = Typhoons ~ N3.m09, family = poisson(link = "log"))

Deviance Residuals:
      Min          1Q      Median          3Q        Max
-1.777846  -0.756295  -0.001514   0.641291   2.344787

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.87400    0.03692  77.842  < 2e-16 ***
N3.m09      -0.12624    0.04141  -3.048  0.00230 **
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1
```

[4] In file `TC_anal.r`, this command appears as `print(summary(storm.model1))`. This is necessary to force output to be written when the entire program is run using the command `source("TC_anal.r")`.

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 41.296  on 41  degrees of freedom
Residual deviance: 31.630  on 40  degrees of freedom
AIC: 232.22

Number of Fisher Scoring iterations: 3
```

The output is more detailed. We obtain some summary statistics for deviance residuals (see Section 2.5.1). If the model is a good one, we expect these to be approximately symmetrically distributed about zero. We also obtain assessments of the significance of the individual predictors in the model. The "z value" column is the ratio of each estimate to its standard error, and the "Pr($>$ $|$z$|$)" column gives a p-value for testing the hypothesis $H_0 : \beta_i = 0$ against the alternative $H_1 : \beta_i \neq 0$. We might informally accept evidence of a genuine relationship if the values are less than 0.05. In fact, R highlights 'significant' relationships with asterisks. The p-value associated with the September Niño 3 value is 0.00230, suggesting extremely strong evidence of association. Note, however, that this p-value is twice that given in Figure 2.2. In this case, our conclusions are not affected, but it does illustrate the potential sensivity of results to the analysis method chosen. The p-values in Figure 2.2 will be correct if typhoon numbers are normally distributed; those from the GLM will be correct if the underlying distribution is Poisson.

2.7.4 Comparing models

We now have a model which uses September Niño 3 values to foreast a Poisson distribution for the number of typhoons in the North-West Pacific the following year. Is this the best possible model? We could try fitting models which use Niño 3 values from other months in place of the September value. If we do this, we find that none of the other models gives a deviance as low as the September one. Equivalently, the 'September' model has the highest likelihood. However, we may want to ask: can we improve our model by including other months' Niño 3 values in addition to the September value? For example, suppose we consider using both the August and September Niño 3 values:

```
storm.model2 <- glm(Typhoons ~ N3.m09 + N3.m08,
                                family=poisson(link="log"))
summary(storm.model2)
```

The results are given in Table 2.3. The p-values here indicate that neither of the two Niño 3 values is significantly associated with typhoon numbers — this appears to contradict the findings of the previous section, where we found strong evidence of an association between September Niño 3 values and typhoon numbers.

The reason for the apparent problem is that August and September Niño 3 values are highly correlated, and tests based on standard errors of coefficients can be misleading in this situation (see

```
Call:
glm(formula = Typhoons ~ N3.m09 + N3.m08, family = poisson(link = "log"))

Deviance Residuals:
      Min          1Q     Median          3Q         Max
-1.768344  -0.766805   0.006982   0.650814   2.354857

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.873309   0.038393  74.840   <2e-16 ***
N3.m09      -0.118073   0.131130  -0.900    0.368
N3.m08      -0.008747   0.133295  -0.066    0.948
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 41.296  on 41  degrees of freedom
Residual deviance: 31.625  on 39  degrees of freedom
AIC: 234.21

Number of Fisher Scoring iterations: 3
```

Table 2.3: R output for Poisson GLM, fitting annual typhoon numbers to previous year's August and September Niño 3 values.

Section 2.2.1). Effectively, the hypotheses being tested here are of the form $H_0 : \beta_i = 0$ *with all other βs fixed at their current values*. Since September and August values are highly correlated, the August term is not likely to add much information once the September term is in the model, and vice versa.

The way around this problem is to use tests based upon the deviance or scaled deviance (which are the same for the Poisson distribution). We have two models: the ‘September only’ model and the ‘September plus August’ model. The deviance for the first (reduced) model is 31.63, and that for the second is 31.625 (from Table 2.3). To test whether the second model offers a significant improvement over the first, we compare the deviance reduction with the upper 5% point of a χ^2 distribution with 1 degree of freedom (since there is 1 extra parameter in the extended model). This is 3.84 (check this in R by typing `qchisq(0.95,1)`). Since the observed deviance reduction is only 0.005, we conclude that there is no improvement, and the model should not be extended.

The `anova()` command can be used to carry out this procedure automatically in `R` . Type `anova(storm.model2,test="Chi")`, to obtain the following Analysis of Deviance table:

```
Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      41     41.296
N3.m09  1    9.667         40     31.630     0.002
N3.m08  1    0.004         39     31.625     0.948
```

Each row of this table represents the effect of adding an extra term to the model. For example, by adding the September Niño 3 value to the null model (i.e. the model containing no predictors), the deviance is reduced by 9.667. The associated p-value is 0.002, in agreement with the result of the previous section. However, when we add the August Niño 3 value to the model, the deviance reduces by 0.004 (our manual calculation gave 0.005, but this is because `R` only outputs the deviances to 3 decimal places). The associated p-value is 0.948.

Of course, we could repeat this analysis the other way round — adding the August term to the model first. In this case we find that the September term does not improve significantly upon a model that just contains the August term. The conclusion is that only one of the two predictors is necessary. Since the 'September only' model gives a lower deviance (i.e. a higher likelihood) than the 'August only' model, this is the model we prefer.

Finally, we might ask whether we can improve the model by adding *any* Niño 3 values from the last 5 months of the year (since, according to Figure 2.2, this is the period for which the relationship with typhoon numbers is strongest). We can do this by comparing our September model with an extended model containing all 5 months' values:

```
storm.model3 <- glm(Typhoons ~ N3.m08 + N3.m09 + N3.m10 + N3.m11 +
                               N3.m12,family=poisson(link="log"))
anova(storm.model1,storm.model3,test="Chi")
```

We obtain the following Analysis of Deviance table:

```
Model 1: Typhoons ~ N3.m09
Model 2: Typhoons ~ N3.m08 + N3.m09 + N3.m10 + N3.m11 + N3.m12
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        40     31.630
2        36     30.773  4    0.857     0.931
```

Now, the deviance reduces by 0.857 for the addition of 4 extra parameters (in the 'Df' column). The associated p-value is 0.931 — again, we find no evidence that we can improve our model by adding extra terms.

Further notes

This example has been chosen for its simplicity rather than its realism. In practice, model choice is usually more complex than in this study. However, we have not covered all of the capabilities of R here (for a full list, type `help.start()` and see the online help!), and some of these may be of use in more complicated situations. A few additional features are as follows:

Interactions: these can be specified easily in R . For example, the command `glm(y ~ x1 + x2 + x1:x2, family = ...)` can be used to fit a model containing the predictors `x1`, `x2` and their interaction.

Stepwise fitting: various 'automatic' methods of model selection exist. These are similar to stepwise regression. The R command for this is `step`. For example, to select predictors from the last 5 months' Niño 3 values we could fit a model containing all of these values (i.e. `storm.model3` above), and then use the `step()` command to decide which terms should be kept in the model. In the above example, the command `step(storm.model3)` may be used. However, such automatic methods should be used with caution. Stepwise fitting is *not* guaranteed to find the 'best' model, and the results can depend upon the way in which the stepwise search is carried out. There are various options in the `step()` command. A safe strategy for researchers is: if you don't understand these options, don't use the command! In any case, in climate there is usually an 'obvious' hierarchy of models so that the use of automatic techniques is not necessary. This was discussed earlier, in Section 2.2.1.

GLMs with nuisance parameters: for the Poisson distribution, there is no nuisance parameter. Therefore, the deviance and scaled deviance are the same and deviance tests based on the χ^2 distribution are appropriate, even for small samples. However, for other distributions such as the gamma, it may be more appropriate to calculate Analysis of Deviance tables using '`test = "F"`' instead of '`test = "Chi"`' in the `anova()` command. In fact, '`test = "F"`' may be used for the Poisson case as well, and gives the same results as the χ^2 version.

2.7.5 Model checking

At this point, we have decided that typhoon numbers may be predicted using September Niño 3 values in a Poisson GLM. Our model is stored as an R object called `storm.model1`. We need to carry out some checks. R in fact provides a variety of diagnostic plots — we just type

```
par(mfrow=c(2,2))
plot(storm.model1)
```

Figure 2.8 shows the result. First, we have a plot of deviance residuals against linear predictors (see Sections 2.5.1 and 2.5.3). Since we are using a log link function, the linear predictors are the

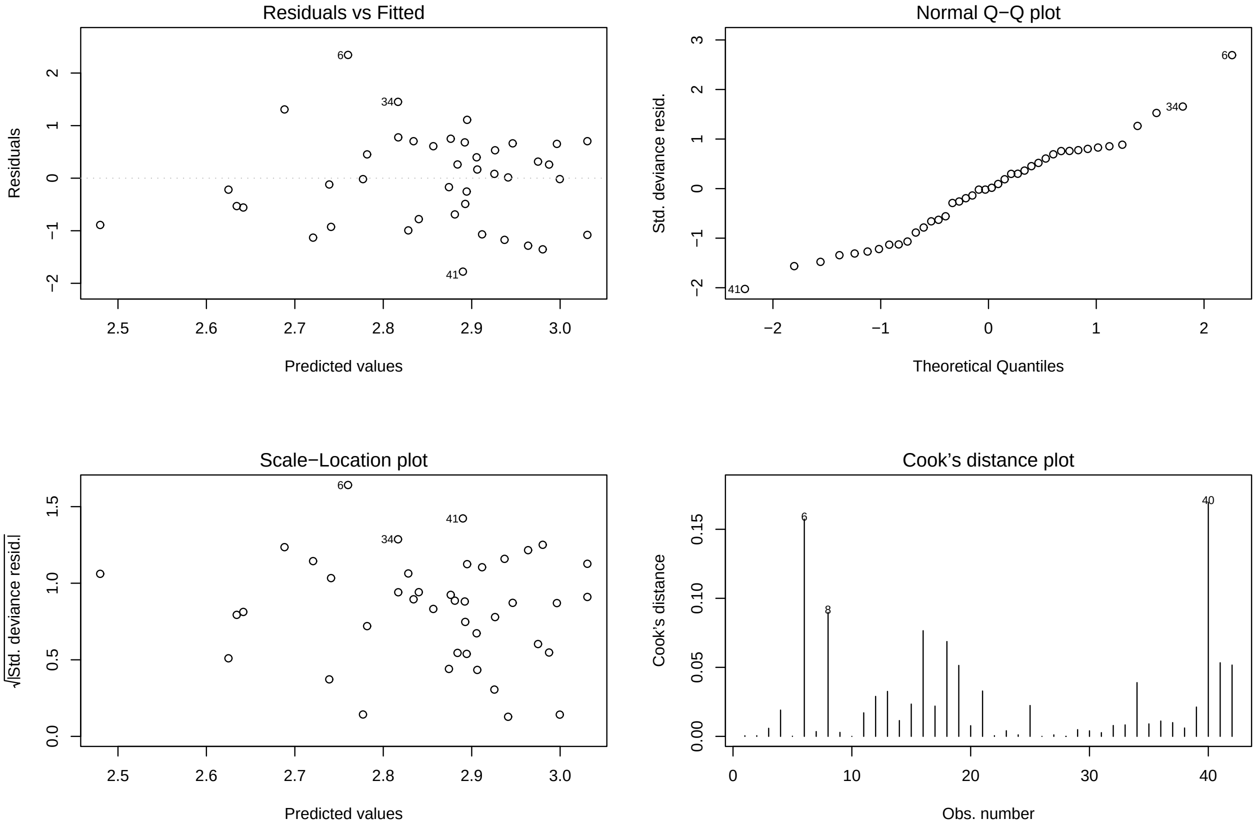

Figure 2.8: Diagnostic plots for Poisson GLM fitted to North-West Pacific typhoon numbers.

logarithms of the fitted means for each case. This plot provides a check of the link function in the model, as well as highlighting any other obvious anomalies. If the model is correct, there should be no structure here — this appears to be the case, which is a good sign.

The second plot is a normal quantile-quantile plot of deviance residuals. This is based on the idea that, as discussed in Section 2.5.1, deviance residuals may be expected to have a distribution close to normal. In this case, the quantile-quantile plot should appear close to a straight line — departures from this would suggest that the Poisson assumption is not valid. However, in practice the normal approximation for deviance residuals may be poor, so this plot may be moderately curved even if the model is correct. In this case, there is no cause for concern.

The scale-location plot is another check on distributional assumptions. It is designed to check the variances of the fitted distributions (for the Poisson distribution, the variance is equal to the mean). If the chosen family of distributions is correct, the points on this plot should be randomly scattered about the line $y = 1$. Again, there do not appear to be any problems.

Finally, the Cook's distance plot tells us which observations have the most influence upon the fitted model (in the sense that, if these observations were deleted, the parameter estimates would change a lot). Observations 6, 8 and 40 are all highlighted as influential. To identify these observations, we can type storm.data[c(6,8,40),c(1,3,13)] (this command selects rows 6, 8 and 40, and columns 1, 3 and 13, of storm.data). The corresponding years are 1964, 1966 and 1998. We may wish to investigate these years further — for example, to determine whether anything unusual occurred, or whether there is any data error.

The plots in Figure 2.8 all seem to indicate that the Poisson model fits well. However, as discussed in Section 2.5.3, we may also wish to find out whether there are other factors that have not been accounted for. An obvious question is whether typhoon numbers have changed over time. To answer this, we can plot a time series of model residuals for each year. We will use Pearson residuals, since they are more easily interpretable than deviance residuals. For the Poisson distribution, the Pearson residual for the ith case in a dataset is defined as $r_i^{(P)} = (y_i - \mu_i)/\sqrt{\mu_i}$. In R , the plot is generated using the commands

```
pearson.model1 <- resid(storm.model1,type="pearson")
par(mfrow=c(1,1))
plot(Year,pearson.model1,lwd=2,type="l",xlab="Year",ylab="Residual")
abline(0,0)
```

The result is shown in Figure 2.9. This plot does not look random, indicating that there is interannual structure that has not been captured by the model. In this study, we do not propose to try and find out the cause of this; however, it does show that simple plots can be informative.

We can also check that the mean and variance of Pearson residuals from this model agree with their expected values of 0 and 1 respectively. The commands for this may be found in file TC_anal.r. The mean is 0, and the variance is 0.799. This variance is lower than expected — this may indicate that the Poisson distribution does not fit the data. If this is the case then we have a very interesting result. We chose the Poisson distribution by considering the mechanism of cyclone formation. If the Poisson distribution is not a good fit, we must conclude that typhoons do not follow a Poisson process and therefore that our suggested mechanism for typhoon formation is incorrect. The most likely explanation is that easterly waves do not develop into typhoons independently of each other.

We conclude this section with a final check. In Section 2.5.2, we described a method for checking the probability structure of GLMs for discrete responses. To implement this method, we compare the observed and expected numbers of years in which 0, 1, 2 . . . typhoons occurred. The R code for doing this is included in file TC_anal.r. The result is shown in Figure 2.10. The observed distribution is bimodal, with two clearly-defined peaks at values around 14 and 20 typhoons per year. The expected distribution does not capture this pattern. This may indicate a problem with the model, or the pattern in the observations may be a chance effect due to the small sample size. This is something that should be investigated further.

We do not attempt to improve this model here. The main point of this case study has been to demonstrate that GLMs can be fitted, and checked, very easily in R , and that simple checks can highlight possible problems with the models. These problems suggest directions for further model development, and can potentially enhance our understanding of climate processes.

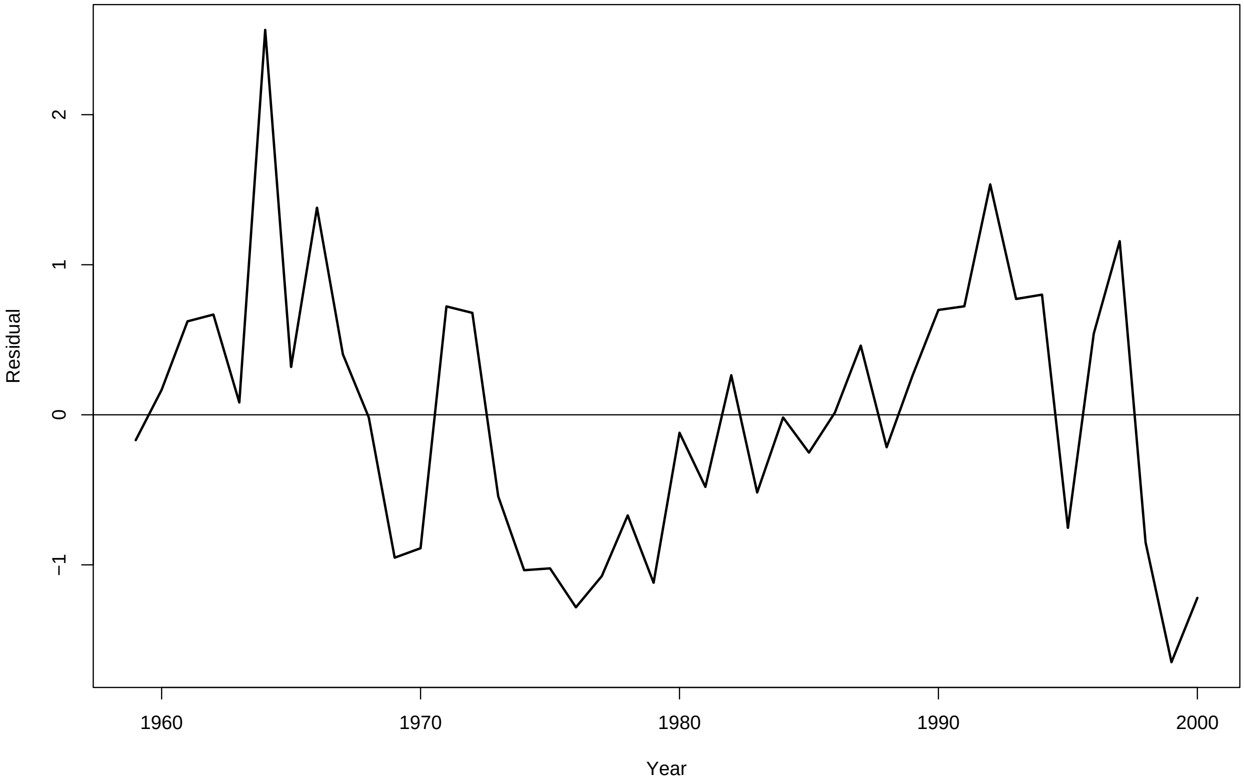

Figure 2.9: Annual Pearson residuals for Poisson GLM fitted to North-West Pacific typhoon numbers.

2.8 Further reading

There is an enormous amount of statistical literature on Generalised Linear Models. They were introduced by Nelder and Wedderburn (1972); the theory is covered quite comprehensively in the classic book by McCullagh and Nelder (1989). A less detailed and simpler treatment is given by Dobson (1990). Chandler (1998b) contains some discussion of the issues involved when predictors are correlated.

In the climatological literature, GLMs have not received much attention. Some fundamental papers are Coe and Stern (1982) and Stern and Coe (1984) — these authors used GLMs to model daily rainfall sequences. Further recent developments of GLMs, in the context of rainfall modelling, are given in Chapter 4 of Wheater *et al.* (2000b). This reference contains more detailed pointers to some of the relevant statistical literature, as well as a thorough analysis of the Irish data considered here in Case Study 2.

Poisson models have been used to model tropical cyclone numbers by Elsner and Schmertmann (1993) and Elsner *et al.* (2001) — the latter paper gives a nice introduction to the use of the

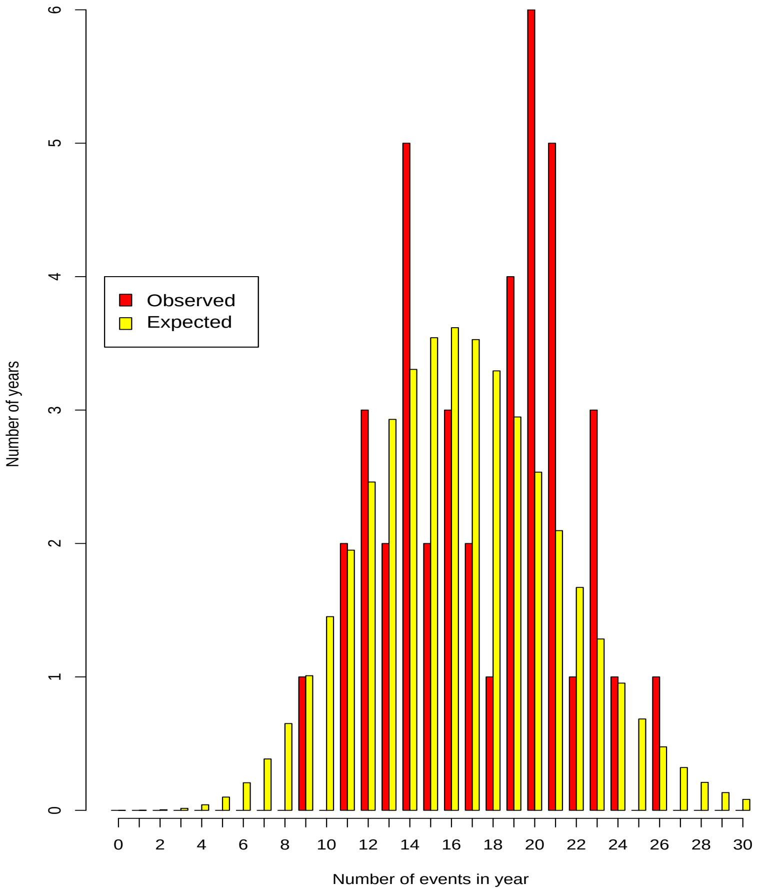

Figure 2.10: Observed distribution of annual typhoon numbers, compared with that expected under Poisson GLM.

deviance statistic in a climatological application.

The need to account for autocorrelation, when analysing relationships between time series, is very clearly demonstrated by Walther (1999). This paper should be compulsory reading for everybody in climate research!

An overview of the relevant issues in model checking is given by Chandler (1998*a*). For a more detailed treatment from a statistical perspective, see Atkinson (1985) — this book is a good reference for the various types of residual plot used by `R` .

The windspeed data in Case Study 4 have been provided by the Royal Netherlands Meteorological Institute (KNMI). Documentation for this dataset can be found on the KNMI website at `http://www.knmi.nl/samenw/hydra/`, and in Verkaik (2000*a*). The procedures used to remove inhomogeneities in station records are published in Verkaik (2000*b*).

Any graduate-level mathematical textbook will contain details of orthogonal basis functions. Abramowitz and Stegun (1965) and Press *et al*. (1992) are useful references for Legendre polynomials, in particular.

Lecture 3

Applications, and alternatives

3.1 Case studies

To begin this final lecture, we return to Case Studies 2, 3 and 4, which were introduced in Section 2.3. The aim is to illustrate what can be achieved using GLMs in a variety of different situations.

3.1.1 Case study 2

In the study of Irish rainfall, we have decided to use logistic regression to model the sequence of wet and dry days at each site, and then to fit gamma distributions to the amount of rain on wet days. These will be referred to as 'occurrence' and 'amounts' models respectively.

To develop a model, it is natural to start with 'obvious' predictors, and to add extra terms and interactions gradually. The value of extra terms can be assessed by examining the log-likelihood, predictive performance and residuals for each model. Recall, however, that when our data consist of time series from several sites, log-likelihoods should be treated informally because of the spatial dependence between sites.

One of the aims of this study was to investigate the evidence for changing rainfall patterns in the West of Ireland. Within the GLM framework, this can be investigated by fitting basic models corresponding to a stationary climate, and comparing these with 'extended' models incorporating nonstationary effects. In selecting predictors to represent trends over time, four basic alternatives have been considered here. The first three are deterministic functions of time:

$$f_1(t) = t\ , \qquad f_2(t) = \begin{cases} 0 & t \leq t_0 \\ t - t_0 & \text{otherwise.} \end{cases} \qquad \text{and} \qquad f_3(t) = \cos\left(\frac{2\pi(t-\phi)}{\omega}\right)\ .$$

It is implausible to extrapolate $f_1(t)$ indefinitely outside the range of the data, but it may provide a good approximation to any monotonic trend over the period of record. $f_2(t)$ is a crude representation of anthropogenic 'climate change' (t_0 being the year in which the change started to occur). $f_3(t)$ is included to investigate the possible presence of cycles in the area's climate. $f_2(t)$ and $f_3(t)$

Model number	Trend scenario	Number of parameters in model	Log-likelihood	RMSE (mm)
RAINFALL OCCURRENCE				
1	None	35	-67990.600	Not applicable
2	f_1	42	-67805.609	
3	f_2	43	-67805.609	
4	f_3	44	-67804.388	
5	NAO	41	-67583.472	
6	NAO plus f_1	49	-67470.000	
RAINFALL AMOUNTS				
1	None	30	-194086.831	5.580
2	f_1	45	-194023.122	5.579
3	f_2	46	-194023.122	5.579
4	f_3	43	-193995.569	5.578
5	2 cycles	48	-193941.679	5.577
6	NAO	38	-193862.769	5.567
7	NAO plus f_3	42	-193822.092	5.566
8	NAO plus 2 cycles	51	-193762.053	5.565

Table 3.1: Summary of models for the daily rainfall record in the Galway Bay area. For each trend scenario, the summary refers to the best model that was found. Log-likelihoods are calculated as though data from different sites are independent. The numbers of observations were 143,682 for the occurrence models and 101,448 for the amounts.

are both nonlinear transformations of time, involving unknown parameters (t_0, ϕ and ω) that must be estimated from the data as described in Section 2.4.4.

These trend functions are all essentially descriptive in nature. It is natural to ask whether there is a physical explanation for any apparent trends. Therefore we have investigated the impact of the North Atlantic Oscillation (NAO) in addition to the deterministic trend functions. The NAO is known to be associated with European precipitation patterns — see the reference list at the end of this lecture for further details.

Table 3.1 gives the numbers of parameters, and log-likelihoods, for a variety of models. Additionally, for each amounts model the root mean squared error (RMSE) is reported. For both occurrence and amounts, Model 1 contains predictors representing systematic regional effects, seasonal variability, previous days' rainfall (5 previous days for occurrence, and 4 for amounts), and interactions between previous days' rainfall and seasonal predictors. These interactions reflect seasonal variations in the strength of autocorrelation. Autocorrelation is higher in winter than in summer, because the area experiences more persistent frontal rainfall in the winter.

The log-likelihoods in Table 3.1 indicate that the best fits are obtained by occurrence model 6 and amounts model 8. For both occurrence and amounts, the log-likelihoods clearly identify the NAO as the dominant source of interannual variability. However, it does not account for all the trends in the data, since the likelihoods for occurrence model 5 and amounts model 6 are both significantly increased by adding extra terms corresponding to linear and cyclical trends respectively. The standard deviation of rainfall amounts on wet days is 5.758mm: hence amounts model 8 explains 6.6% of the variability. This is actually quite impressive given the level of noise in the data (see Section 2.3.2). The improvement is due to the incorporation of previous days' rainfalls (i.e. of 'weather variability'), and the NAO, into the models.

These results show that GLM methodology is able to distinguish between the 'genuine' NAO mechanism and the artifical deterministic trend scenarios. The methodology picks out significant interactions between the NAO and seasonal predictors (the dominant effect is in winter), and a 3-way interaction between the NAO, seasonality and the rainfall 1 day ago. We can use this to illustrate how interactions may be interpreted.

Example 3.1: We will consider the interpretation of the 3-way interaction in the amounts model. From a physical viewpoint we may be interested in the coefficient associated with rainfall 1 day ago, since this tells us about the strength of autocorrelation in the rainfall series, and hence about the types of weather system that affect the area. In the present discussion, let Y_t be the rainfall amount on day t when it is non-zero, and let μ_t denote the mean of the distribution of Y_t. We are using a gamma GLM with a log link function (see Section 2.2). In our model, we have chosen to use the predictor $\ln(1+Y_{t-1})$ to represent the effect of rainfall 1 day ago. According to the fitted model, the contribution of Y_{t-1} to $\ln \mu_t$ is

$$\begin{aligned}\ln(1+Y_{t-1})\Big[0.194 &+ 0.070\cos\frac{2\pi\times\text{day}}{365} + 0.030\sin\frac{2\pi\times\text{day}}{365} - (0.010\times\text{NAO})\\ &- \left(0.015\times\text{NAO}\times\cos\frac{2\pi\times\text{day}}{365}\right) + \left(0.002\times\text{NAO}\times\sin\frac{2\pi\times\text{day}}{365}\right)\Big]\ ,\end{aligned}$$

where 'day' is the day of the year (running from 1 to 365), and 'NAO' is the current value of the monthly NAO index. This index fluctuates about a zero value. Therefore, if we put NAO = 0 in this equation we will obtain an 'average' seasonal cycle for the coefficient of $\ln(1+Y_{t-1})$. If we put NAO = 1, we will obtain the corresponding cycle for a year in which NAO takes the value 1 in every month i.e. in which there is a reasonably strong, and persistent, positive anomaly.

The result is shown in Figure 3.1. As expected, autocorrelation is weaker in the summer than in the winter. The effect of the NAO is to decrease the autocorrelation in the winter months, but it has very little effect in the summer. This suggests that a positive NAO is associated with a reduction in the homogeneity of weather systems in winter. This is quite a complex structure to identify from noisy data; however, the GLM is able to detect this, and to represent it straightforwardly. ■

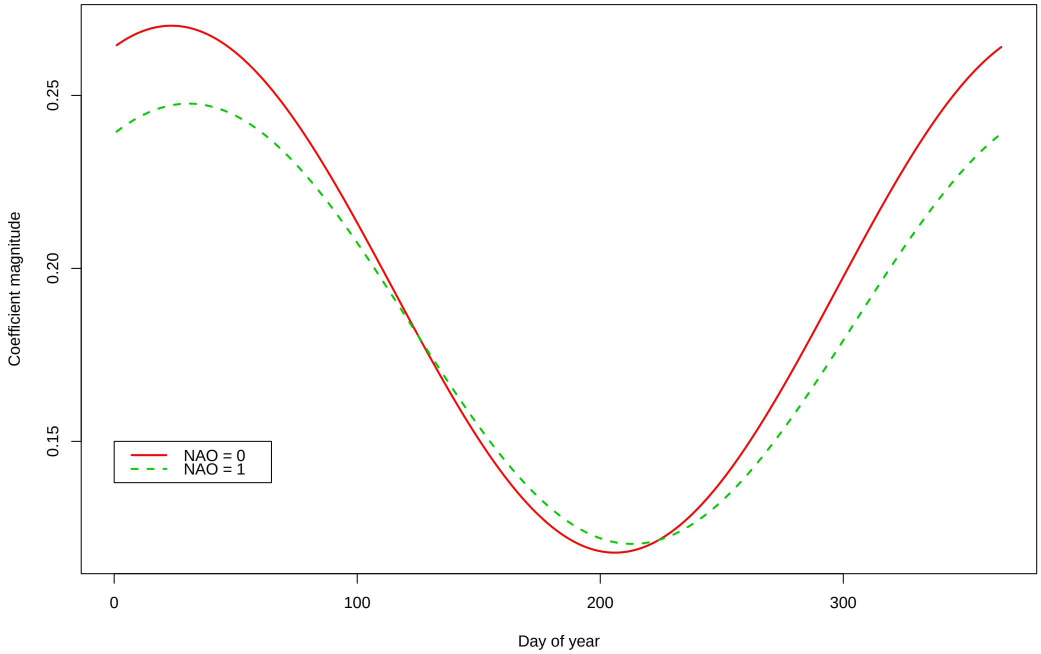

Figure 3.1: Seasonal variation of the coefficient of $\ln(1+Y_{t-1})$ in Galway rainfall amounts model, and the effect of the NAO upon this variation.

In serious climate studies, we would probably want to explore this dataset further. In particular, we may want to determine whether there are any large-scale climate indices that account for the deterministic trends in the models presented here. However, no new methods would be involved so we do not attempt this here.

To check the models, we start by studying Pearson residuals. Here however, because the dataset is large, we do not plot individual residuals, but summary statistics over subsets of data. For example, Figure 3.2 shows summary statistics for monthly and annual Pearson residuals from the best model for rainfall amounts. The amounts models use gamma distributions, for which we use the modified Pearson residuals which are just the proportional errors $(y_i - \mu_i)/\mu_i$ for each case (see Section 2.5.1). The plots of mean residuals include 95% confidence bands, approximately adjusted for spatial dependence. If the model is correct, approximately 95% of mean residuals should lie within these bands. There is little systematic structure in either of the mean residual plots. This indicates that the model gives a good representation of seasonal and interannual variability.

Figure 3.2 also shows the root mean squared Pearson residual for each month and year — i.e. $\sqrt{\sum_{i=1}^{n}\left(r_i^{(P)}\right)^2}$. These plots are designed to highlight any problems with the variance structure of the model (i.e. to check the assumption that the coefficient of variation is constant). The horizontal lines on the plots are drawn at $1/\sqrt{\hat{\nu}}$, where $\hat{\nu}$ is the estimated shape parameter of the gamma distributions. This is the expected value of the squared Pearson residuals if the model is correct.

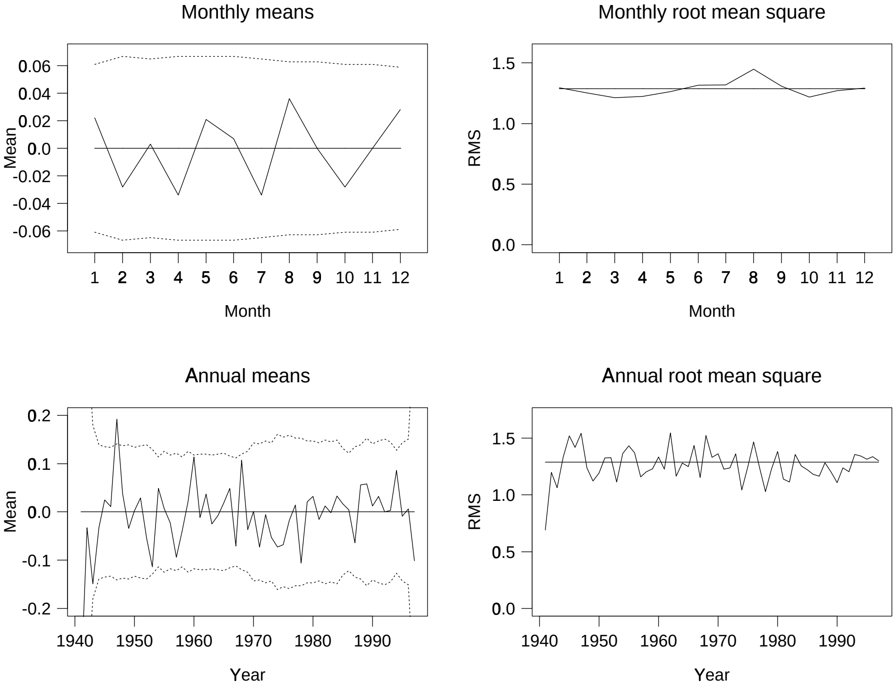

Figure 3.2: Seasonal and annual structure of Pearson residuals from rainfall amounts model 8 in Table 3.1.

We see that there is a very small amount of seasonal structure in the top plot, but little interannual structure. For practical purposes, there is nothing to worry about here.

In addition to checking seasonal and interannual structure, we could check that regional variability is well represented, by computing summary statistics for Pearson residuals at each site. We also need to produce similar plots for the occurrence model.

To check the probability structure of the amounts model, we can produce a normal probability plot of Anscombe residuals (see Section 2.5.1). If the gamma distribution fits the data well, this should appear as a straight line. The result is shown in Figure 3.3. The plot is linear, except in the lower tail where there are not enough very small residuals. This has been investigated. Almost all of the values in the lower tail correspond to rainfall amounts which were recorded as ‘trace’ (i.e. the value was recorded as ‘less than 0.1mm’). The lack of fit here is therefore to be expected, and will not cause any problems in practical applications.

In Table 3.1, the rainfall occurrence models use logistic regression, and provide an opportunity to demonstrate the checking of probability structure for Bernoulli random variables. The technique was described in Section 2.5.2. We split the dataset according to the modelled probability of rain, and compare the observed and expected numbers of rainy days in each subset of data. The result, for the best of the occurrence models, is shown in Table 3.2. There is a close agreement between observed and expected proportions of rainy days, across the entire range of forecast probabilities.

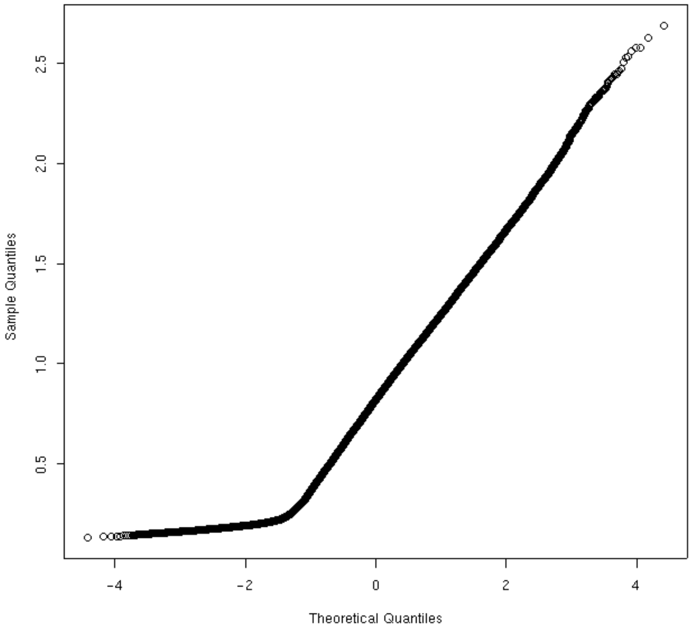

Figure 3.3: Normal probability plot of Anscombe residuals from rainfall amounts model 8 in Table 3.1.

The worst cases are for the forecast ranges $(0.5, 0.6)$ and $(0.9, 1.0)$. For example, of the 6,444 days when we forecast a probability between 0.5 and 0.6, we expect 54.6% to be wet, but only observe 53.2%. This difference is actually statistically significant, but is of little practical interest.

On the basis of these, and other similar analyses, we may conclude that the Irish daily rainfall record is well modelled using a combination of occurrence model 6 and amounts model 8, in Table 3.1. In many climatological applications, we might stop at this point. We have learned that the apparent nonstationarity in Irish rainfall is not a 'chance' effect, and that some of it is due to fluctuations in the NAO. We have also obtained some interesting, and interpretable, results showing how the NAO affects autocorrelation in rainfall sequences. We could interrogate the fitted models further to find out, for example, how NAO affects rainfall amounts (rather than autocorrelations) at different times of year.

However, for this particular study, we need to do more than this. Recall, from Section 2.3.2, that another aim of the study was to estimate the probabilty of large floods recurring, and to provide synthetic rainfall sequences for input to hydrological models. Within the GLM framework, this is straightforward. All we have to do is to simulate sequences from the fitted model. From a single sequence, we can derive any quantity of interest (such as the winter rainfall amount in a particular year). By simulating many sequences, we can obtain a simulated probability distribution for this quantity of interest — this tells us about our uncertainty. In fact, we have to allow for spatial

	Forecast probability of rain									
	0–0.1	**0.1–0.2**	**0.2–0.3**	**0.3–0.4**	**0.4–0.5**	**0.5–0.6**	**0.6–0.7**	**0.7–0.8**	**0.8–0.9**	**0.9–1.0**
N	0	4989	14449	10433	8912	6444	6591	18088	43776	30000
O	0.000	0.179	0.254	0.357	0.455	0.532	0.646	0.752	0.850	0.938
E	0.000	0.178	0.249	0.347	0.449	0.546	0.656	0.760	0.856	0.927

Table 3.2: Checking the probability structure of occurrence model 6 in Table 3.1. Row **N** gives the total number of days in each column. Rows **O** and **E** give the observed and expected proportions of these that were wet.

dependence when simulating over a network of sites. The theory behind this can be complex, and is not covered here. Details are given in the references the end of the lecture.

Figure 3.4 shows what can be achieved using simulation. Here, 1000 daily rainfall sequences have been generated over the period 1989–1997. These sequences were all initialised using observed data. From each simulated daily sequence, winter rainfalls were extracted for every year. The figure shows percentiles of the winter rainfall distributions obtained in this way. It also shows the observed winter rainfall amounts, with some uncertainty owing to missing data.

The simulated distributions in Figure 3.4 show some interranual variability, as a result of the NAO. Each day's probability distribution depends upon the value of the NAO, and the observed NAO sequence from 1989–1997 was used in all of the simulations. The high rainfall in 1995, and low rainfall in 1996, are both strongly associated with NAO activity, since this is the only factor in the model that could possibly produce the dramatic change in simulated distributions between these two years.

Figure 3.4 also shows that the observed winter rainfalls over this period fall within the simulated distributions. The most extreme observed rainfalls, in 1994 and 1995, appear to lie between the upper 5% and 1% point of the simulated distribution. We should not interpret these results too literally. However, they do indicate that floods at least as extreme as those in 1994 and 1995 may occur again under similar NAO conditions. Qualitative conclusions like this are of use to policymakers[1].

3.1.2 Case study 3

In Section 2.3.3, we introduced the problem of modelling temperatures in the USA, to illustrate the use of the GLM approach in a continental-scale study. We decide to use normal distributions, and to fit separate models for the mean and variance of each observation. In this section, we summarise

[1] Since this work was carried out, water levels in the area have again been very high. To some extent, this justifies the conclusions presented here!

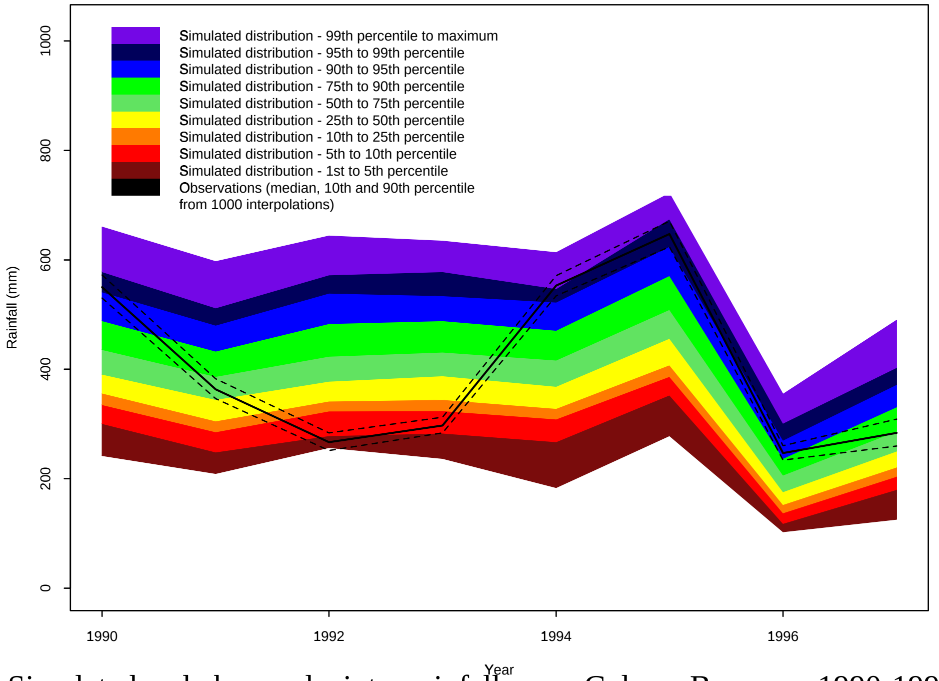

Figure 3.4: Simulated and observed winter rainfalls over Galway Bay area, 1990-1997. The distributions are based on 1000 realisations.

the procedure that was used to fit a model to these data, and illustrate the use of orthogonal series to represent systematic regional variations, as described in Section 2.4.4. We will not discuss model checking here: all of the issues have been covered in Case Studies 1 and 2.

As mentioned in Section 2.3.3, the joint modelling of mean and variance is computationally expensive. This is a particular problem in this case study, where the dataset is large. As a first step to reduce the computational demands, we split the dataset into 2 groups. Half of the 2,606 weather stations were sampled, and their data were used to fit the model. Data from the remaining stations were subsequently used in an independent model validation exercise.

The 'fitting' dataset contains 733,169 observations. This is still large. Within the GLM framework, typically we choose predictors by fitting different models and comparing them. However, because of the computational expense of fitting a joint mean/variance model to a dataset of this size, it is not feasible here to fit every possible model (typically it takes between 4 and 12 hours to fit any realistic model, using an extremely powerful Sun computer). In order to simplify matters, therefore, we first use standard multiple regression techniques (which are computationally cheap) to identify possible predictors for the mean part of the model. Next, we study the squared residuals from the resulting multiple regression model to identify predictors for the variance. We then fit a very large model, containing all of these possible predictors. Finally, we delete groups of similar terms from this large model, if they appear unimportant. When choosing predictors, we should rely on residual analyses and informal methods, rather than upon formal hypothesis tests. This is because the dataset is so large here that any formal test is likely to highlight very small effects as statistically significant, even if they are of no practical importance.

In this case study, orthogonal series have been used to represent systematic regional variability in temperatures. As mentioned in Section 2.4.4, a potential disadvantage of the orthogonal series approach is that it may require a lot of parameters to represent an effect. The number of parameters increases with the complexity of the function being represented. It is likely that the dominant effects of spatial location upon temperature are those of latitude and altitude. In order to keep model complexity to a minimum, we therefore incorporate site altitude as a predictor, and then use orthogonal functions of latitude and longitude to represent any remaining regional variability. Hopefully, this remaining variability will be reasonably smooth.

In order to use an orthogonal series representation, we need to specify an interval (a, b) over which this representation is valid. Orthogonal series will usually provide an accurate approximation to any function, except near the ends of the chosen interval. This suggests that we should choose a and b to be well outside the range of the available data. However, if we did this we could not consider the available values to be uniformly distributed over (a, b), which is necessary in order for the resulting predictors to be approximately uncorrelated. Therefore, in practice we should choose the interval to be slightly wider than the range of the data. For this study, we represent longitude effects over the range $-130°$ to $-60°$, and latitude effects over the range $20°$ to $50°$. The region is shown in Figure 3.5.

Over the range of latitudes considered here, temperatures increase on average from North to South. This suggests that the effect of latitude may be represented using a low-order polynomial. The Legendre polynomials form an orthogonal basis, and so we have used these. Polynomials up to degree 4 have been investigated. Degree 4 offered no improvement over degree 3, and accordingly latitude is represented via the 3 predictors

$$\begin{aligned} \zeta_1^{LAT} &= (\text{Latitude} - 35)/15 \\ \zeta_2^{LAT} &= \left(3\left(\zeta_1^{LAT}\right)^2 - 1\right)/2 \\ \text{and } \zeta_3^{LAT} &= \left(5\left(\zeta_1^{LAT}\right)^3 - 3\zeta_1^{LAT}\right)/2\ . \end{aligned}$$

The choice of a basis for representing longitude effects is not so obvious. The only thing we can do is to try different bases and see which produces the best model. In this study, we investigated the use of Legendre polynomials up to degree 8, and a Fourier representation using the first 5 Fourier frequencies. After dropping insignificant terms from the models, their performance was virtually indistinguishable. However, the Fourier-based model contained fewer terms. This is therefore the preferred basis, and longitude effects are represented using the functions

$$\zeta_j^{LONG} = \begin{cases} \cos\left(\frac{(j+1)\pi}{70} \times \text{longitude}\right) & j = 1, 3, 5, 7\ . \\ \sin\left(\frac{j\pi}{70} \times \text{longitude}\right) & j = 2, 4, 6, 8\ . \end{cases}$$

When using Fourier series to represent a function over an interval, we should remember that

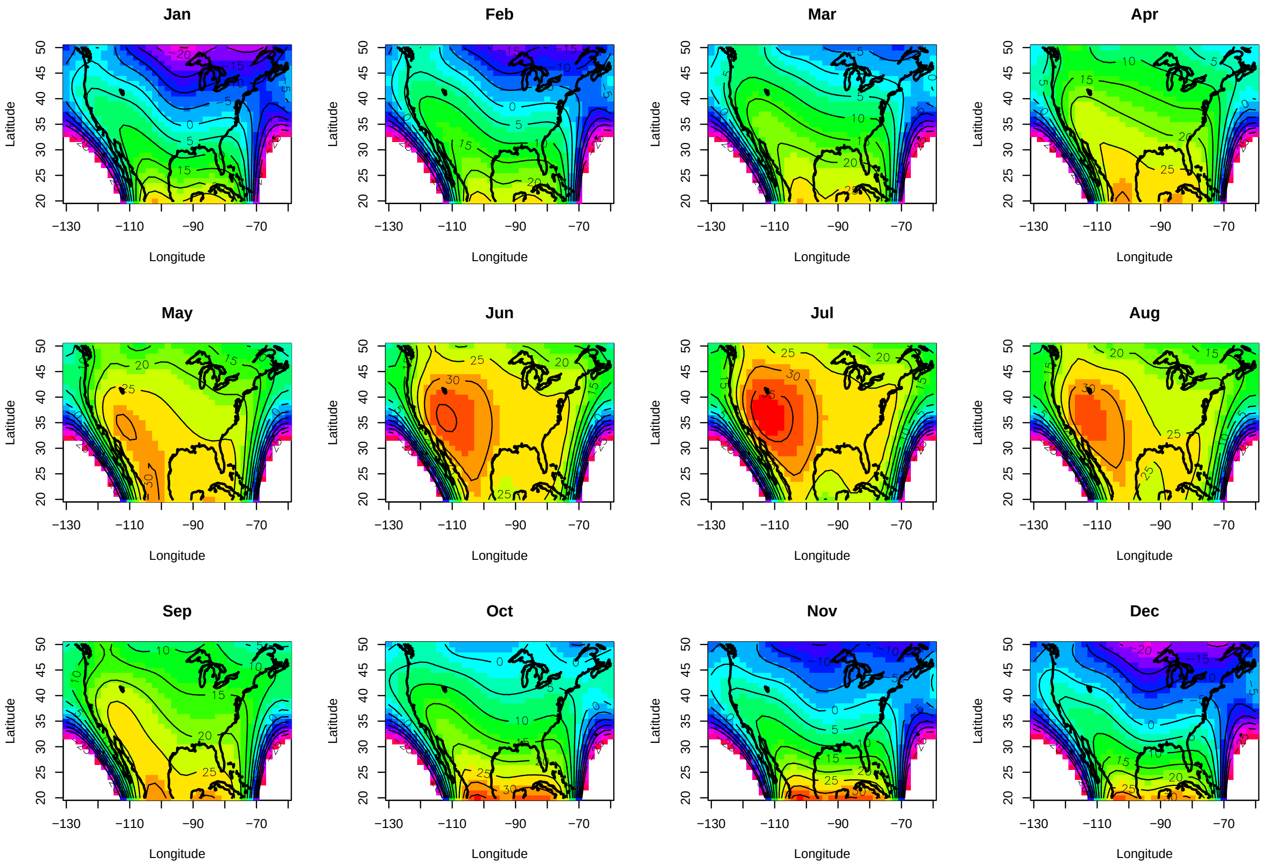

Figure 3.5: Systematic regional variation of temperatures over the USA, represented by orthogonal basis functions of latitude and longitude, and their interactions with seasonal predictors. Contour units are degrees Fahrenheit.

Fourier series are periodic. If $\hat{f}(.)$ is the Fourier series representation of a function $f(.)$ over the interval (a, b), then $\hat{f}(a) = \hat{f}(b)$. The Fourier basis should therefore only be used in situations where it is reasonable to assume that $f(a)$ and $f(b)$ are approximately equal.

In Section 2.4.4, we said that to represent the systematic variation of temperature with latitude and longitude, it is necessary to include the ζ^{LAT}s, ζ^{LONG}s and their interactions into the model. Usually, many of the interactions will be insignificant and can be deleted from the model. However, to start with we must consider all of them.

At a continental scale, regional temperature patterns will vary with season. Therefore, we need to include interactions between the ζs and the seasonal predictors in the model. As in Case Study 2, seasonality may be represented by sine and cosine functions. We can use the interactions in a model to show the systematic regional variation in US temperatures in each month. If we take any GLM, and extract the terms corresponding to seasonality, regional variation and their interactions, we obtain a function of latitude, longitude and time of year. This represents an average deviation from some baseline, whose level is determined by the remaining terms in the model. In each month, the seasonal predictors take different values, and hence the coefficients associated with each of the ζs for that month can be calculated. The resulting function of latitude and longitude can then be plotted. The functions defined in our model for μ_i are plotted, for each month, in Figure 3.5 (note that the effect of altitude is omitted from these maps). The idea is exactly the same as that used to produce Figure 3.1 in Case Study 2. The dominant feature in the winter is a North-South temperature gradient. In the summer, the desert regions in the Western USA become very hot. This result is not very exciting scientifically, but it does indicate the ability of the GLM approach to represent known climate patterns.

There are no large-scale climate indices in the temperature models considered here. The model for mean temperature contains 112 predictors, of which 99 are required to define the structure in Figure 3.5. The variance model contains 27 predictors, of which 23 represent systematic regional effects (the remaining 4 represent seasonal variation). In a climatological study, the resulting model could be treated as a base: predictors representing climatological variables could be added to assess their significance and impacts. Maps, such as those in Figure 3.5, could be used to display the impacts at different times of year.

3.1.3 Case study 4

We have now illustrated most of the important areas in which GLMs can be applied to climatology. The final case study gives you an opportunity to try out some of these ideas for yourself.

In Section 2.3.4, we decided that these windspeed data could be modelled using a gamma GLM. In principle, this could be done using R as in Case Study 1. However, the dataset is fairly large (it contains 125,181 observations), and fitting GLMs in R can be slow for large datasets. Here, we will use a suite of FORTRAN programs to fit models. These programs fit gamma distributions, and logistic regression models — originally, they were written for daily rainfall modelling and used to analyse Case Study 2. Details of how to obtain the programs may be found in Appendix A.2. The software will need to be installed and compiled before you can try this case study for yourself. On a Unix system, the compilation will produce an executable called `fit_gamm`, which is the program we will use to fit gamma GLMs.

The GLM fitting programs require a number of input files. These are all described in the software documentation. The first is a data file which may be downloaded from the website for this lecture series, as described in Section 2.3.4. The other required files are `siteinfo.def`, which contains details of site locations; `mn_preds.dat`, which is used to define monthly 'external' predictors (here, we just provide NAO data, as an example); and a model definition file. All of the files may be downloaded from the website. A number of model definition files are included. These illustrate how we may build up a GLM gradually from a few 'obvious' predictors.

The simplest possible model for this study is the one in which all observations come from the same gamma distribution. In this model, $E(Y_i) = \mu_i$, such that $\ln \mu_i = \beta_0$ for all i. The definition

```
     Results after   9 iterations:
     Log-likelihood -     -76773.038
     Number of observations - 125181
     No. of parameters estimated -  1
     Nuisance parameter (NU) -  5.321411 (ML estimate is

Final parameter estimates:

Main effect:                                      Coefficient  Std Err
------------                                      -----------  -------
Constant                                            1.899159   0.0012

Spatial dependence structure:
------------------------------
Structure used is Independence


                              RESIDUAL ANALYSIS
                              =================

     Mean of observations:   6.680
     Standard deviation of observations:   2.896

     Mean error (observed - predicted):   0.000
     Root mean squared error:   2.896
     Proportion of variance explained by model:  0.000

     Mean Pearson residual:   0.000
     Standard deviation of Pearson residuals:   0.433
     Expected std dev of Pearson residuals:   0.433

     Mean Anscombe residual:  0.9792 (expected:  0.9791)
     Std Dev of Anscombe residuals:  0.1441 (expected:  0.1445)

     ...
```

Figure 3.6: Example of output from gamma GLM fitting program (simple model with no predictors except a constant).

Model	File	Predictors in model
0	model_0.def	Constant only
1	model_1.def	Constant + autocorrelation: $\ln(1 + Y_{t-j})$ for $j = 1, 2, 3$
2	model_2.def	As model 1, + seasonal cycle
3	model_3.def	As model 2, + linear functions of latitude and longitude
4	model_4.def	As model 3, + autocorrelation/seasonal interactions
5	model_5.def	As model 4, with some insignificant interactions removed
6	model_6.def	As model 5, + site/seasonal interactions
7	model_7.def	As model 5, + site/autocorrelation interactions
8	model_8.def	As model 7, with some insignificant interactions removed
9	model_9.def	As model 8, + NAO
10	model_10.def	As model 9, + NAO/site interactions
11	model_11.def	As model 9, + NAO/seasonal interactions
12	model_12.def	As model 11, + NAO/site/seasonal interactions
13	model_13.def	As model 11, + linear trend
14	model_14.def	As model 13, + trend/site interactions
15	model_15.def	As model 14, + trend/seasonal interactions
16	model_16.def	As model 14, + trend/site/seasonal interactions
17	model_17.def	As model 14, including spatial dependence

Table 3.3: Description of models for which definition files are provided, for use with Case Study 4.

file for this model model_0.def. To fit this model, copy model_0.def to gammamdl.def and run the fitting program. Some important parts of the output are shown in Figure 3.6. These include the log-likelihood for the fitted model, the number of observations, the parameter estimates and some residual analyses. The estimate of β_0 is 1.899. This corresponds to a mean windspeed of $e^{1.899} = 6.679\text{ms}^{-1}$. The full output contains further residual analyses, which are omitted here for reasons of space. Use Case Study 2 as a guide when interpreting these analyses.

The model fitting program contains a number of output files. These are described in the software documentation. The important ones are gammamdl.res, which contains results, and gammamdl.de2, which is a model definition file corresponding to the fitted model[2]. If we now want to extend the model by adding extra predictors, we can copy gammamdl.de2 to gammamdl.def, add some extra lines corresponding to the extra predictors, and fit the new model.

[2] The file anscombe.ps may also be produced. This produces a normal probability plot of Anscombe residuals. However, this may be incorrect and should be ignored! Use R to produce normal probability plots.

To give some idea of a typical sequence of models for this dataset, 17 model definition files have been provided. The models are summarised in Table 3.3. Note the following:

1. The first predictors to be added are those representing autocorrelation. This is extremely important, because likelihoods can only be used to compare models that have been fitted to the same data values. In Model 1, we can only fit models to cases where the values of each of the 3 previous Y values are known. Since some data are missing, we therefore have to discard some days from the database to fit the model, and the sample size decreases (from 125,181 to 123,311). Therefore, we cannot compare the likelihood from Model 1 to that from Model 0. However, all remaining models will be fitted to the same data as Model 1, so comparisons can be made between these.

2. After accounting for autocorrelation, we add predictors representing seasonality and site effects. From the preliminary analysis in Section 2.3.4, these obviously affect windspeed. The linear representation of site effects seems reasonable, from Figure 2.7.

3. The general procedure for dealing with interactions is to add a group of them at a time, then delete the ones that appear insignificant.

4. Model 8 is a 'baseline' model, that is deemed to account for seasonality, regional variability and autocorrelation. We can investigate questions of climatological interest, such as the effect of the NAO, by comparing extended models with this baseline.

5. Model 17 is the same as Model 14, except that the fitting program will also estimate the correlations between Anscombe residuals at each pair of sites. These are stored in file `cor_gamm.dat`, and may be used subsequently to simulate correlated daily windspeed sequences at these 9 sites, if desired. The simulation program supplied with the software can be used to achieve this.

We will not discuss this case study any further. It is almost time for the statistician to stop, and for the climatologists to take over!

3.2 Other statistical methods

In these lectures, we have focused upon the use of GLMs to analyse climate data. There are, of course, many other statistical methods that are commonly used in climatology. In this section we summarise some of these, to place the GLM methodology in context.

3.2.1 Extreme value theory

In many applications, we are interested in studying 'extreme' events, since these often have a large impact upon society. In Case Study 2, for example, we studied the probability of severe flooding in Ireland by building a GLM for daily rainfall, and simulating this to estimate the probability of large floods recurring.

Usually, extreme events are studied using Extreme Value Theory. This may involve either of the following techniques:

1. Split a dataset into K time periods, each of which contains a 'large' number of observations. Extract the largest observation from each time period, and fit a Generalised Extreme Value (GEV) distribution to the resulting sample of K maxima. For daily data, we will typically take a year to be the basic time period so that we are fit distributions to the annual maximum values. From these distributions, we can deduce statements such as 'the probability that the maximum daily rainfall this year will exceed y is 0.01', for some threshold y.

2. Consider just those observations ($y_1, \ldots, y_m$ say) that exceed some large threshold τ, and fit a Generalised Pareto Distribution (GPD) to the threshold exceedances $y_1 - \tau, \ldots, y_m - \tau$. Again, the fitted distribution can be used to make probability statements about large values.

These techniques are both based upon large-sample theory (and, in fact, are equivalent from a theoretical point of view). Under very general conditions, the maximum of a large number of identically-distributed random variables has a GEV distribution, regardless of the distribution of the individual variables. A similar result justifies the use of the GPD in modelling threshold exceedances. These results are similar to the Central Limit Theorem, which suggests that we should use the normal distribution to model 'averages'.

Historically, the method of moments (see Section 1.4.3) has often been used to fit distributions in extreme value analyses. In modern statistical practice, however, maximum likelihood is usually used. Using maximum likelihood, it is possible to incorporate predictors into the fitted distributions, and to assess the significance of these predictors using likelihood ratio tests. This allows us to assess the effects of predictors upon extremes directly (in a GLM, we have to study extremes by simulating models that are fitted to all of the data), which can be very useful if we are only interested in extreme events.

The main advantage of Extreme Value Theory is that we do not have to choose a distribution for the variable of interest. It is known that the GEV distribution is the only possible distribution for the maximum of a large number of observations, and that the GPD is the only possible distribution for threshold exceedances. By fitting these distributions, we are can therefore be reasonably confident that our estimates of extreme event probabilities will be fairly accurate. By contrast, if we derive extreme event probabilities by simulating a GLM, we need to be extremely careful that the model structure is correct.

Extreme Value Theory does have some drawbacks. These are mostly associated with the need to fit distributions to a small subset of the available data, and with difficulties in applying the theory to data from more than one site. If we fit distributions to annual maxima, or to observations that exceed some high threshold, then our sample size will be greatly reduced. As a result, we may not be able to detect weak relationships among variables. Also, note that we can only use Extreme

Value Theory if the quantity of interest is an extreme event over a relatively small timescale. In Case Study 2, for example, we were interested in large winter rainfalls. To perform an Extreme Value Analysis of winter rainfall, we would need winter rainfall data for many years, so that we could extract the 'large' values and still have enough data to fit a distribution.

In summary: Extreme Value Theory is very useful for situations where extremes of direct interest. When its use is appropriate, it will probably estimate extreme event probabilities more accurately than a GLM. However, it is not always appropriate, and the need to discard most of the data before fitting distributions means that weak climatological relationships may not be detected using this technique. Ideally, any analysis of extremes would combine both an Extreme Value Analysis, and a GLM.

3.2.2 Multivariate techniques

It is probably fair to say that multivariate techniques are currently the most popular statistical methods in the climatological literature. Such techniques include Principal Components Analysis (PCA, also known as 'Empirical Orthogonal Functions', and effectively the same as a Singular Value Decomposition) and Canonical Correlation Analysis (CCA). All of them are designed for the analysis of high-dimensional datasets. In climatology, the high dimension usually arises from the simultaneous observation of a single variable at a network of sites. For ease of presentation, we will only discuss Principal Components Analysis here. The general comments apply to all similar methods, however.

Suppose we observe values of a climatological variable, at S sites. The observations at any time can be assembled into an S-dimensional vector $\boldsymbol{y}$, say, and we observe values of this vector over many time points. If S is large, it can be difficult to visualise all of the observations. However, in many applications, values from neighbouring sites at the same time will be very similar. This means that each $\boldsymbol{y}$ vector effectively contains far fewer than S pieces of information, and suggests that we might search for a way of representing the $\boldsymbol{y}$s in a small number of dimensions. PCA is a method for achieving this, by transforming each $\boldsymbol{y}$ vector into S new variables, which are linear combinations of its elements. These new variables are the principal components of the system. They are mutually orthogonal (effectively, this means that they are uncorrelated), and are chosen in such a way that, in some sense, the first j principal components give the best possible representation of the entire dataset in j dimensions. It may be possible to interpret a principal component, by examining the weights associated with each y value in the linear transformation. For example, we may conduct a PCA of gridded global sea surface temperatures, and find that the second principal component allocates high weight to all grid nodes in the equatorial East Pacific, and low weight everywhere else. In this case, we may be justified in interpreting this as an 'El Niño' component.

This approach is very different to the one used in Generalised Linear Modelling. In a GLM, each $\boldsymbol{y}$ would be regarded as the realised value of a random vector $\boldsymbol{Y}$. The corresponding mean vector $\boldsymbol{\mu}$ would be derived from predictors at each site, and the dependence between sites might be specified by studying some structure among suitably-defined residuals.

In comparing the approaches, notice that PCA is essentially a descriptive technique. Its primary aim is to reduce a high-dimensional dataset to a few variables. There is no notion of probability involved here (in fact, PCA can be embedded within the framework of a probability model, but the theory behind this is complex and the models involved are not particularly helpful for many practical applications). Note also that the definition of the principal components is an artificial one, made on purely mathematical grounds. It is tempting to try and ascribe a meaningful interpretation to each of the principal components of a system (as with the El Niño example above). However, the climate is a complex system, and nobody really believes that it can be regarded as a collection of uncorrelated variables! For this reason, it is potentially dangerous to seek interpretations of principal components.

The GLM approach has the advantage that it models climate variables directly, thereby avoiding the artificiality of the PCA representation. Moreover, uncertainty in a GLM is easily quantified because it is a probability-based framework. In any PCA, we ought to ascribe some uncertainty to the weights in each linear combination, since these will change with the time period used for analysis. However, this is rarely (if ever) done — the result is to underestimate uncertainty in any analysis based on PCA.

PCA can potentially be very helpful for the climatologist using GLMs. Suppose we wish to use an entire field of, say, sea surface temperatures (SSTs) to build a model for some variable of interest. We could build a GLM containing each of the individual SST values as predictors. However, this would lead to a huge model (and many predictors would appear insignificant, because the SST values at neighbouring sites would be highly correlated). In such a situation, it may be useful to carry out a PCA of the SST field, and use a few principal components as predictors. There is no guarantee that the first principal components will be the ones most strongly associated with the variable of interest, so it may be necessary to try a variety of different models using this procedure. However, it does offer the opportunity to reduce a complex problem to a manageable form. Alternatives based on canonical correlations are possible.

In summary: from a statistical perspective, the application of most multivariate techniques in climatology should be primarily descriptive, and over-interpretation of their output should be avoided where possible. In applications, it is more informative to investigate variables of interest directly, rather than via an artificial construction that is motivated by mathematical elegance rather than practical usefulness. The ability to reduce the number of dimensions in a large dataset is useful however, and may be applied to obtain a few predictor variables from a large spatial field.

3.2.3 Time series modelling

When data arise as sequences in time, it is common to analyse them using time series models. These are typically based upon the Autoregressive-Moving Average (ARMA) class of models and its extensions. An ARMA model, for a stationary sequence of random variables (Y_y), takes the form

$$Y_t = \mu + \sum_{j=1}^{p} \phi_j \left(Y_{t-j} - \mu\right) + \varepsilon_t - \sum_{k=1}^{q} \theta_k \varepsilon_{t-k}$$

for some parameters $\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$. (ε_t) here is a sequence of uncorrelated, identically distributed random variables with zero mean.

In an ARMA model, the values of $Y_{t-1}, \ldots, Y_{t-p}$ and $\varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$ will be known at time t. They can therefore be regarded as predictors, and we can write

$$E\left(Y_t | Y_{t-1}, \ldots, Y_{t-p}, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}\right) = \mu + \sum_{j=1}^{p} \phi_j \left(Y_{t-j} - \mu\right) - \sum_{k=1}^{q} \theta_k \varepsilon_{t-k} \; .$$

This takes the form of a GLM, with the previous Ys and εs as predictors. Therefore ARMA-type time series models can be regarded as special cases of GLMs. Indeed, the class of GLMs that includes previous Y values as predictors is sometimes called the class of GENERALIZED AUTOREGRESSIVE MODELS.

This description appears to trivialise the enormous amount of literature on time series modelling. However, there are many issues, mainly regarding the theoretical properties of models, that we have not mentioned in our discussion of GLMs. If we really want to understand the structure and implications of our models, we need to address these issues. For the wider class of GLMs, they are poorly understood. In climatology, the large size of datasets means that this lack of understanding may be unimportant — any sensible model, fitted and checked using a large dataset, is unlikely to have poor properties. We should be aware, however, that at present there is plenty of theoretical work to be done, when using GLMs to study time series!

3.2.4 Stochastic models

So far we have discussed purely statistical techniques. These may be contrasted with STOCHASTIC MODELS, whose aim is to give a simplified representation of a process in probabilistic terms. For example, we know that rain occurs in ‘storms’. Within storms, local areas of convection (‘rain cells’) produce enhanced rainfall. The complete mechanics of the rainfall process are extremely complex. However, a very crude model for a rainfall sequence at a site is as follows:

1. Storm origins follow a Poisson Process (see Section 1.3.3).

2. Each storm gives rise to a random number of rain cells, which arrive at the site in a sequence after the storm origin.

3. Each cell has a random duration and intensity.

In order to complete the description of such a model, we need to specify distributions for the numbers of cells per storm, and for the cell durations and intensities. The model’s parameters are physically interpretable quantities such as storm arrival rate, mean number of cells per storm and mean cell duration.

It might appear that such a model is far too simplistic to be useful. However, in practice it is found that this type of structure can do an extremely good job of reproducing many features of observed rainfall sequences, under a variety of different climate regimes. Stochastic models are particularly appropriate for the generation of high-resolution synthetic data, which is required in a variety of applications (notably in hydrology).

3.2.5 Bayesian methods

In our discussion of statistical modelling, we have adopted the view that models should be fitted by Maximum Likelihood whenever possible. This is based on the idea that, if we wish to use a data vector $\boldsymbol{y}$ to learn about the probability distribution that generated it, Maximum Likelihood estimators are often optimal in various senses.

However, in many situations we may be able to do better than this. The reason is that our knowledge of a system does not just come from $\boldsymbol{y}$. Past experience with similar data, and understanding of the processes within the system, may both give us some idea of realistic parameter values before we even see the data. A simple example illustrates this:

Example 3.2: Suppose we wish to find the probability that a coin comes down heads when we toss it. To do this, we toss the coin 100 times, independently. On the ith toss, we record $Y_i = 1$ if the coin shows a head, and 0 if it shows a tail. The Ys are therefore independent Bernoulli variables, with unknown parameter p.

When the experiment is carried out, we observe values $y_1, \ldots, y_{100}$, and assemble them into a vector $\boldsymbol{y}$. The log-likelihood function for $p \in [0, 1]$ is then

$$\ln L\left(p|\boldsymbol{y}\right) = \sum_{i=1}^{100} \ln P\left(Y_i = y_i\right) = \sum_{y_i=1} \ln p + \sum_{y_i=0} \ln(1-p) = x \ln p + (100-x)\ln(1-p) \ ,$$

where x is the total number of heads observed. To maximise the likelihood, we differentiate with respect to p and set to zero: the maximum likelihood estimate is $\hat{p} = x/100$, as expected.

To express our uncertainty regarding the value of p, we could estimate the standard error associated with $\hat{p}$. However, as mentioned in Section 1.4.3, it is more accurate to give a confidence interval based on the likelihood. A 95% confidence interval is the set of values of p_0 for which the null hypothesis $H_0 : p = p_0$ is not rejected at the 5% level. Using the usual theory for likelihood ratio tests, this is the set of values $\{p_0 : 2\left[\ln L(\hat{p}|\boldsymbol{y}) - \ln L\left(p_0|\boldsymbol{y}\right)\right] \leq 3.84\}$, since 3.84 is the upper 5% point of a χ^2_1 distribution. The endpoints of this interval are therefore the values of p_0 satisfying

$$\begin{aligned} x\ln\left(\frac{x}{100}\right) + (100-x)\ln\left(\frac{100-x}{100}\right) - x\ln p_0 - (100-x)\ln\left(1-p_0\right) &= 1.92 \\ \Rightarrow \quad \left(\frac{x}{100p_0}\right)^x \left(\frac{100-x}{100\left(1-p_0\right)}\right)^{100-x} &= e^{1.92} \ . \end{aligned}$$

The solutions to this equation can be found straightforwardly using numerical methods.

Suppose now that we carry out this experiment, and observe 62 heads. We therefore compute $\hat{p} = 62/100 = 0.62$, and the 95% likelihood-based confidence interval for p is $(0.523, 0.711)$. However, most people would not accept this. They would argue that the true value of p is 'obviously' 0.5, and that the results of this experiment are due to chance — in effect, we have made a Type I error (see Section 1.4.2) in testing the null hypothesis $H_0 : p = 0.5$ against the alternative $H_1 : p \neq 0.5$.

The reason for this reaction is that we have some prior understanding of the coin-tossing experiment. We have a very strong belief that a coin is equally likely to show heads or tails. If we record 620 heads in 1000 tosses, or 620,000 heads in 1,000,000 tosses, we may suspect that the coin is biased; however, we are unlikely to change our prior belief on the basis of the results considered here. The likelihood analysis takes no account of prior knowledge, since it uses only the observed data for the experiment. ■

How can we incorporate prior knowledge of a parameter vector $\boldsymbol{\theta}$ into a statistical analysis? Note first that we have some uncertainty about $\boldsymbol{\theta}$ (if we didn't, there would be no need for any analysis!). The natural way to express this uncertainty is via a probability distribution: we may consider that $\boldsymbol{\theta}$ is itself a random vector with density $\pi(\boldsymbol{\theta})$, say. This is called the PRIOR DISTRIBUTION of $\boldsymbol{\theta}$. Note, however, that probability statements about $\boldsymbol{\theta}$ cannot be interpreted in the 'classical' way. We cannot usually obtain many different values of $\boldsymbol{\theta}$ by repeating some experiment a large number of times. However, $\pi(\boldsymbol{\theta})$ does have an intuitive interpretation, since it conveys information about uncertainty. In Example 3.2 above, we may choose a prior distribution for p that is extremely concentrated about 0.5. An example is shown in the left panel of Figure 3.7.

The effect of observing $\boldsymbol{y}$ is to modify our prior belief about $\boldsymbol{\theta}$. Our modified belief can be expressed by the conditional probability distribution of $\boldsymbol{\theta}$ given $\boldsymbol{y}$. This conditional distribution is called the POSTERIOR DISTRIBUTION for $\boldsymbol{\theta}$, and is denoted by $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. By Bayes' Theorem (page 11), we have

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

Here, $f(\boldsymbol{y}|\boldsymbol{\theta})$ is the density of $\boldsymbol{y}$ given $\boldsymbol{\theta}$, and $f(\boldsymbol{y})$ is the unconditional density of $\boldsymbol{y}$ (which depends upon $\pi(\boldsymbol{\theta})$, rather than upon $\boldsymbol{\theta}$ itself). For the analysis of any dataset $\boldsymbol{y}$, $f(\boldsymbol{y})$ is fixed. It can be regarded as a constant, chosen to ensure that the posterior density integrates to 1 (since it represents a probability distribution). We therefore have

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) , \qquad \text{or equivalently} \qquad \pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto L(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta}) ,$$

where $L(\boldsymbol{\theta}|\boldsymbol{y})$ is the likelihood for $\boldsymbol{\theta}$ given $\boldsymbol{y}$ (see page 30).

The point of all this is that the posterior distribution allows us to modify our prior belief using the available data. As we gather more data, the contribution of the likelihood will tend to dominate that of the prior. As a result, for very large datasets, inference based on the posterior will often be very similar to that based on the likelihood. However, for small or moderately-sized datasets, the

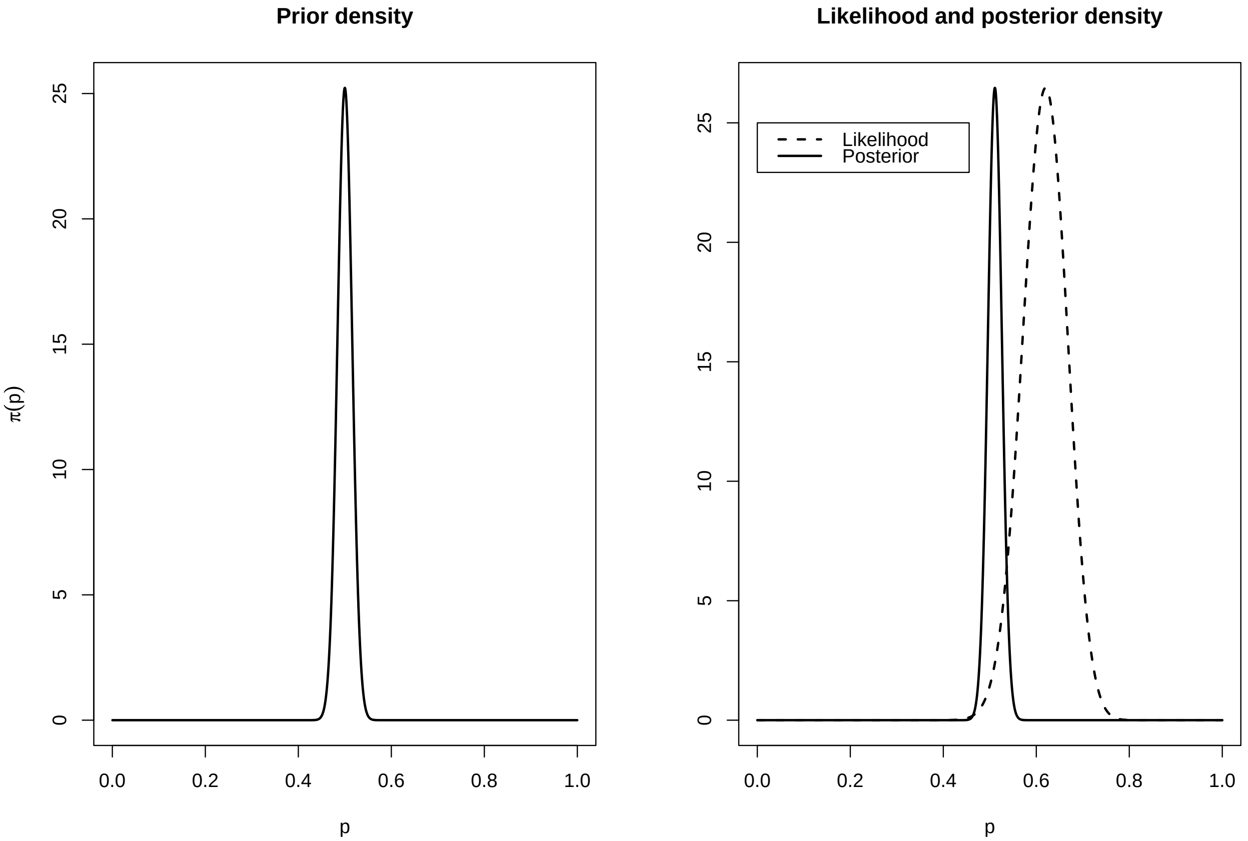

Figure 3.7: Left panel: a possible prior distribution for the unknown parameter p in a coin-tossing experiment (see Example 3.2). Right panel: the likelihood function for p when 62 heads are recorded out of 100 tosses, and the posterior density that results from combining this with the suggested prior. The likelihood here is defined up to a constant of proportionality, and is presented on the same scale as the posterior for ease of comparison.

prior can influence results to a considerable extent. This is illustrated in the right-hand panel of Figure 3.7, where the likelihood and posterior density are presented for the coin-tossing example discussed above. The strong prior information dominates the results, and the posterior is much more in agreement with our intuition than the likelihood.

Statistical inference based on posterior distributions is called *Bayesian inference*. This caused some controversy in the statistical community during the 20th century, mainly because the choice of prior distribution was seen as subjective. However, Bayesian methods are now generally accepted, since they can be very powerful if applied appropriately. The idea has tremendous potential in some areas of climate research, where a lot of prior knowledge can be assimilated into an analysis. There are, however, a number of difficult issues involved. In particular, calculation of posterior densities in real problems is usually difficult, and computationally demanding. Moreover, for large datasets such as those considered in Case Studies 2, 3 and 4 here, any Bayesian analysis is likely to be dominated by the contribution of the likelihood to the posterior density, so that results and conclusions would be very similar to those we have presented already.

3.3 Relationships with physical and dynamical modelling

The methods discussed in these lectures have all been based on probability models. The use of such models does not mean that we regard the climate as 'random' in the usual sense of the word. We regard our observations as realised values of random variables, but in this context 'random variable' is a formal mathematical concept, as defined in Section 1.2.3.

As an alternative to statistical modelling, we could study the climate as a system which is essentially deterministic. This is usually done by writing down equations that represent various laws governing the system's behaviour, and studying the evolution of the system according to these equations. We refer to such an approach as 'dynamical modelling'. Inevitably, there is some degree of approximation involved in this, but there can be little doubt that the more 'expert knowledge' can be embedded in a model, the better it is likely to perform. Since climatologists are experts on the climate, their climate models are likely to be better than those developed by a statistician! Nonetheless, no climate model is perfect. There will always be uncertainty, and hence the opportunity to incorporate probability modelling, in climate research.

The potential for incorporating probability into dynamical models is only just starting to be recognised. We may begin to think about how to achieve this, using the framework set out in Section 1.4. Essentially, a dynamical model tells us what to expect from a system. This expectation may be regarded as the mean of a probability distribution. There is a clear connection with GLMs here: our observations are regarded as drawn from some probability distribution, whose mean may be related to the values of various predictors. In a very simple case we could take a dynamical model for, say, rainfall, and use the model output as a single predictor in a GLM. The GLM could then be used to express the uncertainty in the system.

An interesting question that arises here is: what type of probability distribution should we use in such a scenario? To continue with the rainfall example: 'statistical' models, such as those considered in Case Study 2, typically use gamma distributions at a daily timescale. Would the gamma distribution still be appropriate if we were to use mesoscale model 24-hour forecasts as predictors in a GLM? If our dynamical model is accurate, then its errors over small time intervals can probably be modelled using a normal distribution (since this has applications to 'measurement error' problems — see Section 1.3.4). Perhaps, in the case of rainfall, this normal distribution may be regarded as a Gamma distribution with a large shape parameter (see Section 1.3.5). If this is the case, then we might expect that as we use our mesoscale model to forecast rainfall at longer and longer time intervals, the increasing uncertainty may result in a steady reduction in the shape parameter of the gamma distributions. This is speculation, however — and it makes no attempt to deal with the problem of forecasting 'dry' intervals in which there is no rain.

Questions like these can only be answered through collaboration between meteorologists, climatologists and statisticians. To answer them requires both familiarity with probability theory, and a good understanding of the climate system.

3.4 Possibilities for the future

In these lectures, we have tried to illustrate how probability models may be applied in climate research. From a climatological perspective, the examples here are probably rather simplistic. However, we hope that they do at least illustrate the potential of modern statistical methods. To conclude the lectures, we set out some possible areas where probability modelling may usefully be applied in climate research.

The first area is in combining statistical and dynamical modelling approaches, as outlined in the previous section. This would not necessarily affect the ‘average’ performance of climate models. However, the use of a probability-based framework would give us a realistic and structured representation of uncertainty. As we said in Lecture 1, we need to know how big our uncertainty is, in order that we know whether it is important!

The second area is in climate change studies. In Case Study 2, we illustrated the use of GLMs to study changes in the climate of an area. These changes are expressed as changes in probability distributions, rather than in mean climate. This gives us a far more powerful and flexible approach than more traditional analyses based on monthly or annual mean climate data.

The problem of downscaling GCM output is one which receives a lot of attention in the hydrological and climatological literature. This particular problem involves an enormous amount of uncertainty, even without accounting for errors in GCM output. Some currently-available methods do use probability in some form, but many do not. From the point of view of probability modelling, the downscaling problem is actually rather difficult, but also very interesting. This problem is presented as a challenge!

Finally, an application which may be of more relevance to meteorology than climatology is forecasting. There is scope both for the incorporation of probability into forecasting models (as already discussed), and for the development of methods for assessing the performance of probability forecasts. We have mentioned some of the issues involved here, in our discussions of model checking for GLMs. This area is the subject of much current research in the United States, in particular.

These thoughts are inevitably a random collection of items of personal interest; there are doubtless other areas where probability may usefully be applied in climate research. The word ‘random’ here *should* be interpreted in the usual sense!

3.5 Further reading

For further details of Case Study 2, see Wheater *et al*. (2000*b*). This includes a description of the way in which spatial dependence may be incorporated when simulating from a GLM at several sites. Further technical details, including calculation of Anscombe residual properties and methods for dealing with 'trace' values, may be found in Chandler and Wheater (1998*a*) and Chandler and Wheater (1998*b*). A useful reference, dealing with the impact of the North Atlantic Oscillation upon European precipitation, is Hurrell (1995). The NAO index used in this study is the normalised monthly pressure difference between stations in Iceland and Gibraltar, defined by Jones *et al*. (1997).

The use of probability plots to check distributions is a standard statistical procedure. The use of techniques such as the one presented in Table 3.2 is not, however. Such techniques, for assessing probability forecasts, were largely developed in the meteorological literature through the 1960s and 1970s, by Allan Murphy and co-workers. Dawid (1986) contains a good overview. Murphy and Epstein (1967) is an important reference.

We are not aware of other studies that use our approach to modelling temperatures in Case Study 3. However, the method for jointly modelling the mean and variance of a normal distribution is described in Chapter 10 of McCullagh and Nelder (1989).

A good modern text on Extreme Value Theory is Embrechts *et al*. (1997). Although this is predominantly theoretical, it is very accessible and gives a thorough overview of the subject. Smith (1989) illustrates the use of the theory to study trends in extreme ozone levels — this provides a nice example of how the modern statistical approach may be applied to environmental problems. Stuart Coles (currently at the University of Bristol, UK) has written some software for carrying out extreme value analysis. This can be downloaded from `http://www.stats.bris.ac.uk/~masgc/`, together with a set of lecture notes that illustrate its use. The software is written in SPlus, which is very similar to `R` . As provided, it does not run in `R` , but small modifications to the code should fix this.

The use of multivariate techniques in climatology is standard. We therefore give a single reference: Krzanowski (1988) gives an excellent, and very readable, account of the area from a statistical viewpoint. It gives a clear and balanced account of the advantages and disadvantages associated with a variety of methods.

For a good introduction to time series analysis, see Chatfield (1996). This book contains a brief outline of many different time series modelling techniques. Fahrmeir and Tutz (1994) give some theoretical details of Generalized Autoregressive Models.

To date, the only climatological variable for which stochastic models have been extensively developed is rainfall. Recent developments in this area are summarised by Wheater *et al*. (2000*a*), and by Wheater *et al*. (2000*b*). Both of these references contain extensive literature surveys.

A brief introduction to Bayesian methods is given in Chapter 15 of Rice (1995). A more detailed summary of the area is the book by Gelman *et al*. (1995), which is aimed primarily at a statistical audience but contains some useful material that is accessible to non-statisticians. In the climate literature, Chandler *et al*. (2000) illustrate the use of Bayesian methods to downscale GCM output. An alternative probabilistic approach to downscaling is presented by Hughes *et al*. (1999).

Appendix A

Useful software

A.1 The R project for statistical computing

The R package was introduced in Lecture 2, where it was used perform some simple analyses, and fit some GLMs, in Case Study 1. On the R project web site, the package is described as follows:

> “R is available as Free Software under the terms of the Free Software Foundation’s GNU General Public License in source code form. It compiles and runs out of the box on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux). It also compiles and runs on Windows 9x/NT/2000 and MacOS.”

For more details, and to download the package, see the R project homepage at `http://www.R-project.org/`.

In Section 2.7, a variety of R commands were presented. A couple of other pieces of code may be helpful. Firstly, to illustrate how to draw simple diagrams, here is the code used to generate Figure 1.1:

```
#
# Draw the graph of a plausible probability density function - use
# a gamma distribution for convenience, and put arrows on the axes
#
x <- seq(0,10,0.1)
fx <- dgamma(x,4,1) + 0.01
ylim <- c(-0.1*max(fx),1.2*max(fx))
plot(x,fx,type="l",axes=F,lwd=2,xlim=c(0,10),ylim=ylim,xlab="",ylab="")
text(9.5,ylim[1]/2,"y",cex=1.5)
text(0.6,0.95*ylim[2],"f(y)",cex=1.5)
arrows(0,0,10,0,length=0.1)
arrows(1,ylim[1],1,ylim[2],length=0.1)
#
```

```
# Now add a shaded polygon, with some text in it
#
ab <-c(3,6)
ba <- seq(ab[2],ab[1],-0.1)
absector <- dgamma(ba,4,1) + 0.01
abpoly <- list(x = c(ab,ba), y = c(0,0,absector))
polygon(abpoly,col="grey")
lines(x,fx,lwd=2)
text(ab[1],ylim[1]/2,"a",cex=1.2,adj=c(0.5,0))
text(ab[2],ylim[1]/2,"b",cex=1.2,adj=c(0.5,0))
text(mean(ab),max(absector)/2,"Shaded area\ncorresponds to",cex=1.2)
text(mean(ab),max(absector)/2.5,expression(P(a < Y <= b)),cex=1.2)
#
# And output to a Postscript file
#
dev.copy(postscript,"density.ps",horizontal=T,paper="a4")
dev.off()
```

Secondly, to illustrate how to add mathematical labels to plots, here is the code used to generate Figure 1.3.

```
#
# Initialise - set screen to 2x2, set up arrays of shape parameters
# and means, and set plotting ranges
#
par(mfrow=c(2,2))
nu <- c(0.5,1,2,5)
mu <- c(1,3)
xlim <- c(0,5)
ylim <- c(0,1.5)
xvals <- (1:1000)/200
#
# Now produce 1 plot for each shape parameter. Each plot contains 2
# lines - 1 for each value of mu. Also, plots get annotated with Greek
# letters.
#
for (k in nu) {
        lambda <- k/mu[1]
        gammden <- dgamma(xvals,k,1/lambda)
        plot(xvals,gammden,type="l",xlab="y",ylab="f(y)",
                                        xlim=xlim,ylim=ylim,lwd=2)
        lambda <- k/mu[2]
        gammden <- dgamma(xvals,k,1/lambda)
        lines(xvals,gammden,lty=3,lwd=2)
        title(substitute(nu == nuval,list(nuval=k)),cex.main=2)
        legend(2.5,1,c(expression(mu == 1),expression(mu == 3)),
                lwd=c(2,2),lty=c(1,3),cex=1.5)
}
#
# And print to postscript file.
#
dev.copy(postscript,"gammdens.ps",horizontal=T,paper="a4")
dev.off()
```

A.2 FORTRAN code for Generalised Linear Modelling

In Lecture 3, we used a suite of FORTRAN programs to fit gamma GLMs to daily data from a network of sites. The FORTRAN source code for these programs is available from `http://www.ucl.ac.uk/∼ucakarc/work/rain_glm.html`. The distribution is zipped into the single file `rain_glm.zip`. Download this file, unzip it, read the `README` file, and hopefully everything will be clear! The software also contains simulation routines, which were used to generate the synthetic data in Case Study 2.

Bibliography

Abramowitz, M. and Stegun, I.A. (1965). *Handbook of mathematical functions: with formulas, graphs and mathematical tables*. Dover, New York.

Atkinson, A.C. (1985). *Plots, transformations and regression*. Clarendon Press, Oxford.

Chandler, R.E. (1998*a*). Model checking. In *Encyclopedia of Biostatistics*, (ed. P. Armitage and T. Colton). Wiley, Chichester.

Chandler, R.E. (1998*b*). Orthogonality. In *Encyclopedia of Biostatistics*, (ed. P. Armitage and T. Colton). Wiley, Chichester.

Chandler, R.E., Mackay, N., Wheater, H.S., and Onof, C. (2000). Bayesian image analysis and the disaggregation of rainfall. *J. Atmos. and Oceanic Technol.*, **17**, 641–50.

Chandler, R.E. and Wheater, H.S. (1998*a*). Climate change detection using Generalized Linear Models for rainfall — a case study from the West of Ireland. I. Preliminary analysis and modelling of rainfall occurrence. Technical report, no. 194, Department of Statistical Science, University College London. http://www.ucl.ac.uk/Stats/research/abstracts.html.

Chandler, R.E. and Wheater, H.S. (1998*b*). Climate change detection using Generalized Linear Models for rainfall — a case study from the West of Ireland. II. Modelling of rainfall amounts on wet days. Technical report, no. 195, Department of Statistical Science, University College London. http://www.ucl.ac.uk/Stats/research/abstracts.html.

Chatfield, C. (1996). *The analysis of time series — an introduction (fifth edition)*. Chapman and Hall, London.

Coe, R. and Stern, R.D. (1982). Fitting models to daily rainfall. *J. Appl. Meteorol.*, **21**, 1024–31.

Cox, D.R. and Isham, V. (1980). *Point processes*. Chapman and Hall, London.

Dawid, A.P. (1986). Probability forecasting. In *Encyclopedia of Statistical Sciences*, (ed. S. Kotz and N. Johnson). Wiley, New York.

Diggle, P. J. (1983). *Statistical analysis of spatial point patterns*. Academic Press.

Dobson, A.J. (1990). *An introduction to generalized linear models*. Chapman and Hall, London.

Elsner, J.B., Bossak, B.H., and Niu, X. (2001). Multidecadal variability of the ENSO-hurricane teleconnection. Available from http://garnet.acns.fsu.edu/˜jelsner/html/Research.htm. In review.

Elsner, J.B. and Schmertmann, C.P. (1993). Improving extended-range seasonal predictions of intense Atlantic hurricane activity. *Weather and Forecasting*, **8**, 345–51.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events for insurance and finance*. Springer, Berlin.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate statistical modelling based on Generalized Linear Models*. Springer-Verlag, New York.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian data analysis*. Chapman and Hall, London.

Hughes, J.P., Guttorp, P., and Charles, S. (1999). A Nonhomogeneous Hidden Markov Model for precipitation. *Appl. Statist.*, **48**, 15–30.

Hurrell, J.W. (1995). Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science*, **269**, 676–9.

Jones, P.D., Jónsson, T., and D.Wheeler (1997). Extension to the North Atlantic Oscillation using early instrumental pressure observations from Gibraltar and South-West Iceland. *Int. J. Climatol.*, **17**, 1433–50.

Krzanowski, W.J. (1988). *Principles of multivariate analysis*. Oxford University Press.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models (second edition)*. Chapman and Hall, London.

Murphy, A.H. and Epstein, E.S. (1967). Verification of probabilistic predictions: a brief review. *J. Appl. Meteorol*, **6**, 748–55.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *J. R. Statist. Soc., Series A*, **135**, 370–84.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992). *Numerical recipes in FORTRAN (second edition)*. Cambridge University Press.

Rice, J.A. (1995). *Mathematical statistics and data analysis (second edition)*. Duxbury Press.

Smith, R.L. (1989). Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. *Statistical Science*, **4**, 367–93.

Stern, R.D. and Coe, R. (1984). A model fitting analysis of rainfall data (with discussion). *J. Roy. Stat. Soc.*, **A147**, 1–34.

Verkaik, J.W. (2000*a*). Documentatie windmetingen in Nederland (documentation on wind speed measurements in the Netherlands). Technical report, Koninklijk Nederlands Meteorologisch Instituut. Available from http://www.knmi.nl/samenw/hydra/documents/docum0.htm. In Dutch.

Verkaik, J.W. (2000*b*). Evaluation of two gustiness models for exposure correction calculations. *J. Appl. Meteorol.*, **39**, 1613–26.

Wackerly, D.D., Mendenhall, W., and Scheaffer, R.L. (1996). *Mathematical statistics with applications (fifth edition)*. Duxbury Press.

Walther, G. (1999). On the solar-cycle modulation of the Homestake solar neutrino capture rate and the shuffle test. *The Astrophysical Journal*, **513**, 990–6.

Wheater, H.S., Isham, V.S., Cox, D.R., Chandler, R.E., Kakou, A., Northrop, P.J., Oh, L., Onof, C., and Rodriguez-Iturbe, I. (2000*a*). Spatial-temporal rainfall fields: modelling and statistical aspects. *Hydrological and Earth Systems Science*, **4**, 581–601.

Wheater, H.S., Isham, V.S., Onof, C., Chandler, R.E., Northrop, P.J., Guiblin, P., Bate, S.M., Cox, D.R., and Koutsoyiannis, D. (2000*b*). Generation of spatially consistent rainfall data. Report to the Ministry of Agriculture, Fisheries and Food (2 volumes). Also available as Research Report no. 204, Department of Statistical Science, University College London (http://www.ucl.ac.uk/Stats/research/abstracts.html).