MT1007. Climate Change.

This first case study will examine the science behind climate change and the evidence for it. Since this, like the rest of the course, will involve mathematical and statistical models, it's a good idea to start by considering what a model is, what a mathematical model is and why we need statistical models.

1 Mathematical models

Science is about figuring out how nature works, so that we can predict what it will do in the future. Models play an important part in that process. Physical models of things have been used in science and engineering for a long time. For example, to understand and predict how bridges will respond to high winds, it is common practice to build model bridges and to measure how they behave in a wind tunnel. This helps understanding of airflow around the structure, and helps to predict which designs will perform well in practice and which will not. An important point about such models, is that they are not *exact* miniature replicas of the bridges of interest (we can't, for example, build a 1/200th scale model using 1/200th scale atoms), rather they are models designed to capture the features of a bridge that are important for air flow.

Similarly, if we want to produce a model suitable for investigating and explaining the workings of a steam engine, it would be foolish to take a full size working engine and scale it down: the walls of the boiler would be too thin for one thing, but also the full engine has many detailed bits of engineering designed to maximise its efficiency and optimize its practical usefulness. A good model steam engine really doesn't need these and should be considerably simplified.

Orreries are another example of physical models: these are clockwork models of the solar system, designed for showing and predicting the positions of the planets in the solar system. There is a particularly splendid orrery on display in the Physics building, just up the stairs from the PC classroom. Again, although often quite elaborate, orreries do not attempt to represent the system they describe exactly, but rather to represent some aspects of the system accurately enough to serve their purpose.

So, the key feature of models is that they are simplified representations of the thing that they represent, designed to describe some aspects of it quite accurately, but omitting many less relevant or important details in the interests of clarity and simplicity. The idea of mathematical modelling is to keep this basic approach, but to do away with the physical representation part, describing the modelled system by some equations rather than by a physical object.

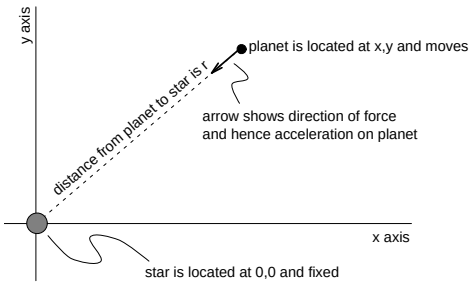
1.1 A mathematical model of planetary motion

As an example let's consider constructing a mathematical model to take the place of an orrery. That is, let's build a mathematical model of the motion of a smallish planet around a star.[1]. To start with we'll make some simplifying assumptions, to strip the model of un-neccessary or minor detail.

- Let's only consider the motion of the planet in 2 dimensions (actually just a matter of choosing the right co-ordinate system), and assume that the star is fixed at position $x = 0$, $y = 0$, while the planet moves around with its position given by co-ordinates x and y.
- We'll assume that the star's position is unaffected by the planet (a good enough approximation if the planet is much smaller than the star).
- Finally we'll assume that the gravity acts according to Newton's model of gravity and that the planets movement is governed by Newtons "laws" of motion (these are very accurate provided the planet isn't moving too fast, and are a much simpler model than Einstein's).

So the basic set up is:

[1] This is one of those examples that you should certainly not learn for the exam - it is here purely to help you understand the important concept of a model.



where the distance r from the star to the planet is given by $r = \sqrt{x^2 + y^2}$. Now, Newton's model of motion says that the acceleration[2] a experienced by the planet will be related to the force f exerted on it by

$$f = ma$$

where m is the mass of the planet. His model of gravity says that the force f exerted on the planet will be related to the distance r between the star and the planet according to:

$$f = \frac{k}{r^2}$$

where k is a constant involving the mass of the star and the planet. So the model says that the acceleration of the planet is given by:

$$a = \frac{k}{m}\frac{1}{r^2}$$

The constants m and k only involve masses and distances, so if we choose the units in which we measure length and mass carefully enough we can re-write this as:

$$a = \frac{1}{r^2}$$

without losing any important details. The acceleration is always in the direction towards the star. To get a model that we can solve to find the motion of the planet, we need to split the acceleration into the part a_x in the direction of the x axis and the part a_y in the direction of the y axis. Basic theory of vectors makes this quite easy:

$$a_x = -\frac{x}{r}a \qquad a_y = -\frac{y}{r}a$$

(The minus signs ensure that the star is attracting the planet and not repelling it!). Now let's define:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = x' \quad \text{and} \quad \frac{\mathrm{d}y}{\mathrm{d}t} = y' \tag{1}$$

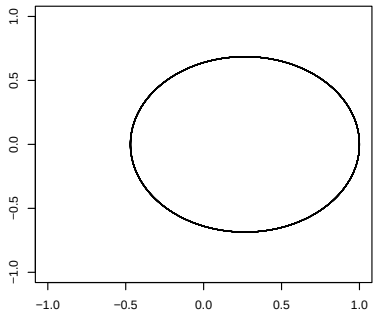
...the speed of the planet in the x and y directions. Acceleration is just rate of change of speed with respect to time, so the equations for a_x and a_y become:

$$\frac{\mathrm{d}x'}{\mathrm{d}t} = -\frac{x}{r^3} \quad \text{and} \quad \frac{\mathrm{d}y'}{\mathrm{d}t} = -\frac{y}{r^3} \tag{2}$$

Equations 1 and 2 constitute our mathematical model for the motion of a planet around a massive star. Given an initial position for the planet x, y, and an initial velocity x', y', the equations can be solved in order to find out where the planet will go.

There are excellent numerical methods available for solving systems of ordinary differential equations like this model. One such method is implemented in `R` package `mt1007`. Using it to solve the above system of equations with initial position (1,0) and initial velocity (0,0.8) gave the following picture for the path of the planet (obtained by plotting y against x):

[2] rate of change of speed



The planet traces out an ellipse according to the model, and this confirms astonishingly well with observation. That planets in reality have elliptical orbits was discovered by Johannes Keplar and published in 1609 - Keplar took 9 years to establish this using the lifetime of observations gathered by Tycho Brahe, a Danish astronomer. Keplar didn't have Newton's model to guide him of course, and Brahe gathered his data before the invention of the telescope.

So, this very simple mathematical model is able to predict the motion of the planets. Furthermore it is easily modified to include all the planets and even the influence of the planets on each other and the sun.

1.2 Determinism's limitations

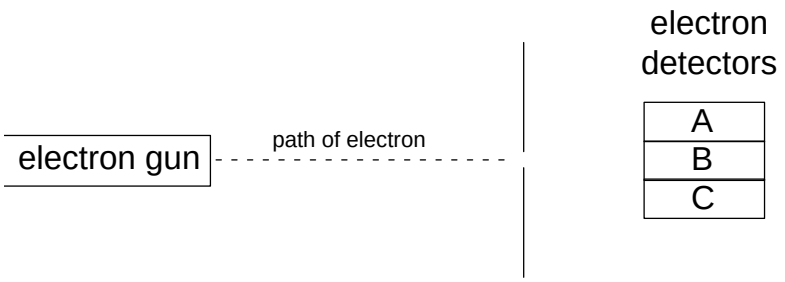
Models of planetary motion based on Newton's models of gravity and motion, were astonishingly successful, and this had a profound impact on the way in which people viewed mathematical models and in the way that we still view them. The 18th century French scientist/mathematician Laplace, extended the basic Newtonian model, given above, to model the motion of all the planets in the solar system, including their influences on each other. This model accounted for the motions of the planets as perfectly as they could be measured (including all the small deviations from elliptical orbits caused by the planets gravitational affects on each other). This seemed to be such a triumph that the belief grew up that everything in the universe could be described by such models, and that in principle the future could be predicted perfectly given such models and accurate measurements of the state of the universe now. Perhaps not surprisingly, Laplace was a vigorous promoter of this idea, which gained the name **determinism**.

These ideas lent their name to the concept of a **deterministic model** - that is a model which if solved from identical starting conditions always has the same solution (the planet model of the last section is an example of one of these - later we will meet statistical models, which do not have this property). For much of the nineteenth century determinism reined, and resulted in a great deal of heart searching about free-will and the like.

In the twentieth century three areas of science comprehensively overturned the idea that everything could be described by deterministic models. These were:

- Quantum mechanics in physics and chemistry.
- Genetics and cell biology.
- Chaos theory in meteorology.

The deterministic mathematical models of Newton, Laplace and later Einstein worked beautifully for describing fairly large simple objects, but physicists got into trouble when they tried to understand how very small things worked: e.g. the constituents of atoms. Models of atoms as miniature solar systems were a failure. In the early part of the 20th century quantum mechanics emerged as a theory which did work, but it was not a deterministic theory. For example, suppose that you fire electrons at an array of detectors, through a very small slit, as illustrated below:



Determinists would expect all the electrons to end up in detector B, but if the slit is narrow enough this is not what happens. Some electrons end up in B, but some also end up in A or C, and according to quantum mechanics there is nothing that can be done or measured to help predict which detector a particular electron will end up in. On the other hand if you fire a large number of electrons at the slit and count up the numbers ending up in A, B or C then there is a pattern - more electrons end up in B than in C or A, and roughly equal numbers end up in B and C. Hence you can not say precisely where a particular electron will end up, but you can give the probability that it ends up in any one of the three detectors. In fact quantum mechanics enables you to work out the probabilities. This sort of uncertainty pervades the physics of the very small: it is fundamentally impossible to make precise predictions about the behaviour of sub-atomic particles, all we can do is predict probabilities. This is perhaps best summed up by the famous Heisenberg uncertainty principle, which states that it is impossible to simultaneously measure the position and momentum (mass $\times$ speed) of a particle (e.g. an electron) accurately. Indeed (considering a particle moving in one dimension only) if the innaccuracy in knowledge of momentum is Δp and the innaccuracy in knowledge of position is Δx then:

$$\Delta p \Delta x \geq \frac{h}{4\pi} \qquad (h \text{ is Planck}'\text{s constant})$$

So quantum mechanical models are not deterministic. For example, each realisation of the quantum mechanical model for the electron passing through a slit will give a different result, only the probabilities of those results are constant. Quantum mechanics is the best description of the very small that we have, and randomness is fundamental to it.

Less fundamental, but of equal practical importance in overturning the sumpremacy of determinism were discoveries in biology. Biologists have long wanted to be able to produce mathematical models that predict and explain the behaviour of biological things, from cells to populations of plants or animals. One of the major problems in doing this is that individuals differ so much, even if they share parents and developed under near identical conditions. As the mechanisms underlying development, evolution and inheritance were uncovered, it became clear why all this variability exists. Essentially, at conception each of us gets a randomly selected half of our mother's genetic material and a randomly selected half of our father's. The mechanism by which the random selection is done allows a staggering number of possible ways of splitting the genetic material in half, and there is no way of predicting which of these ways will actually occur. Hence there is a randomizing mechanism right at the heart of biological systems, and the variability it produces ensures that biology is not deterministic in any useful practical sense.

Now, biological systems and sub atomic particles are respectively very complicated and very small. Deterministic models can't provide all the answers for these, but what about for simpler, middle sized things? For a while scientists tended to think that deterministic models would still always provide the best models in these cases. They thought, for example, that deterministic models would be completely adequate for describing the Earth's atmosphere, which is basically just a layer of gas subject to external heating. As we shall see, the phenomenon of chaos means that this is not so. But we will cover chaos in more detail having introduced the background to the scientific case study on climate.

So, in many cases, for quite fundamental reasons, deterministic mathematical models do not provide adequate models. Statistical models are a solution to this problem. Statistical models are mathematical models, each replicate realisation of which will be different from other realisations from the same model, even under identical conditions. Statistical models are the major means of making sense of real data and are the main focus of this course.

2 Climate change and carbon dioxide

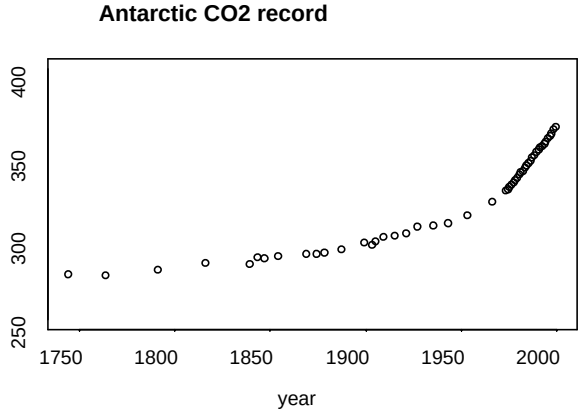
Climate change is one of the most important environmental, scientific and political issues facing people today. The scientific consensus is that human activities in the form of deforestation and the burning of fossil fuels has lead to an increase in the amount of carbon dioxide in the atmosphere. This carbon dioxide traps heat at the bottom of the earth's atmosphere, which changes the earth's climate[3]. However, there are some scientists and a large number of non-scientists (including US president George W. Bush), who are sceptical that increased carbon dioxide levels are changing the climate. In this case study we will use deterministic and statistical models to examine the issue of climate change.

Let's start with carbon dioxide. Life is carbon based. That is, the complicated organic compounds from which living things are composed all contain large amounts of carbon. Hence when organic matter burns or decays it releases carbon, much of it as carbon dioxide. In the absence of human activities the amount of CO_2 in the atmosphere tends to stay roughly constant: the CO_2 given off by decaying/burning plants or animals that have died is eventually just re-absorbed by new living plants. What upsets this is when humans remove and burn large areas of forest, without replacing it, or burn fossil fuels. Fossil fuels, like oil, gas and coal are all made up from the bodies of organisms that lived and died many millions of years ago. The bodies of these animals formed layers of sediment at the bottom of ancient seas and over time were buried. Eventually this carbon rich organic material was buried deep under layers of sedimentary rock, and over the millenia ended up forming, oil, gas and coal deposits. When humans extract this fossil fuel and burn it, a large proportion of the carbon content of the fuel ends up as CO_2 in the atmosphere. There is no way around this problem since it is the carbon that is the energy source in the fuel. The way in which carbon flows through the earth's ecosystem is fairly complicated: huge amounts of carbon circulate through the system all the time. In fact human production of CO_2 only accounts for 3-4% of the CO_2 entering the atmosphere at any given time. The problem arises, because this is 2-3% more than can be removed by natural processes, so the human contribution to atmospheric CO_2 just keeps on growing. For a while people hoped that all this extra carbon dioxide would just get mopped up naturally. It's worth looking at some data to see what is actually happening.

Long term records of carbon dioxide are a little hard to come by. It was not measured until fairly recently. Even recent records can be a little difficult to interpret, as CO_2 levels can vary locally as a result of local biological processes. Antarctica is quite a good place to measure CO_2, because it's fairly lifeless and hence likely to give a fairly reliable picture of CO_2 changes without local biases. Furthermore, as antarctic ice is deposited, air gets trapped within it. Ice deep within the ice cap is old and contains old air, and if this air is extracted then its CO_2 content can be measured. The `R` package `mt1007` includes a dataset `siple.co2` containing measured CO_2 concentrations dating back to 1744, while more recent trends are covered by dataset `south.co2`, which contains direct CO_2 measurements at the South Pole. These are easily plotted together using `R`:

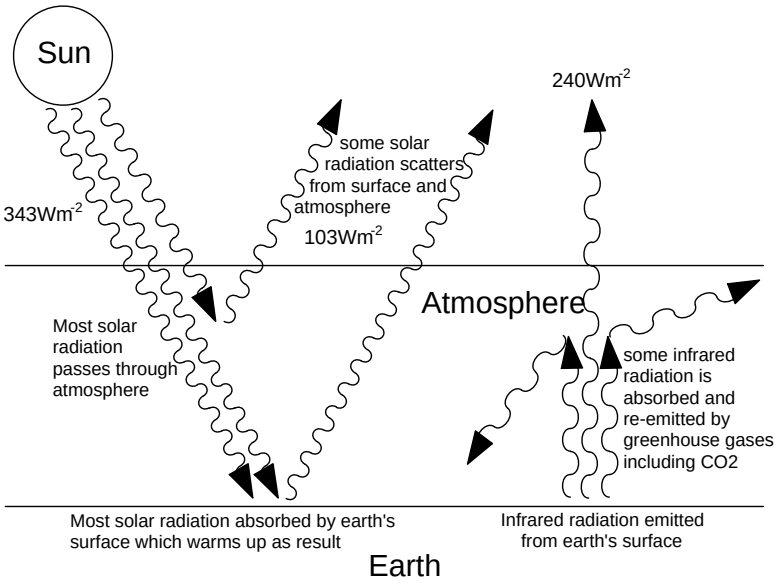
```
> library(mt1007)
> data(siple.co2)
> attach(siple.co2)
> data(south.co2)
> attach(south.co2)
> plot(year,co2,xlim=c(1744,2000),ylim=c(250,400),main="Antarctic co2 record")
> points(Year,Annual)
```

[3]The story is not quite as simple as the one given here, human activities have also increased the amount of methane in the atmosphere, which is also a greenhouse gas, and ozone depletion also contributes to the problem, by allowing more of the energy from the sun through the upper atmosphere. Then there is the contribution of water vapour the problem of reflective ice, snow and clouds, etc., etc.



Clearly the atmospheric CO_2 concentration has increased steadily since the start of the industrial revolution. All other measurements are in quite close agreement with the picture presented here, and the fact that CO_2 concentrations in the atmosphere have increased massively is not really disputed. Incidentally the increase is only about half what would be expected if all the carbon dioxide produced by humans had ended up in the atmosphere. This is partly because the oceans absorb a large amount. Notice that the increase from 280ppm to 360ppm means that about 22% of CO_2 in the atmosphere now probably results from human activity (various people and organisations, including the AA, seem to be unable to understand how adding an extra 3% $C0_2$ to the atmosphere year on year will eventually lead to such a high proportion of CO_2 being the result of human activity. I don't know why.)

CO_2 is a cause for concern because it absorbs infra-red radiation of wavelengths in the range $13 - 19 \times 10^{-6} m$. Infra-red radiation is just light with a very long wavelength. It's what is picked up by thermal imaging equipment, of the sort used by police to hunt for people at night. Warm objects emit it all the time. If there were no greenhouse gases, like CO_2, then infra-red radiation emitted from the earth's surface would simply pass through the atmosphere and out into space (thereby cooling the earth). But greenhouses gases get in the way, absorbing infra red photons (light particles) and then emitting them again. This means that the energy carried by these photons is trapped low down in the atmosphere, which warms it up. Furthermore, because the warmth carrying infra red radiation does not reach the upper atmosphere, this cools down. Schematically it works like this:

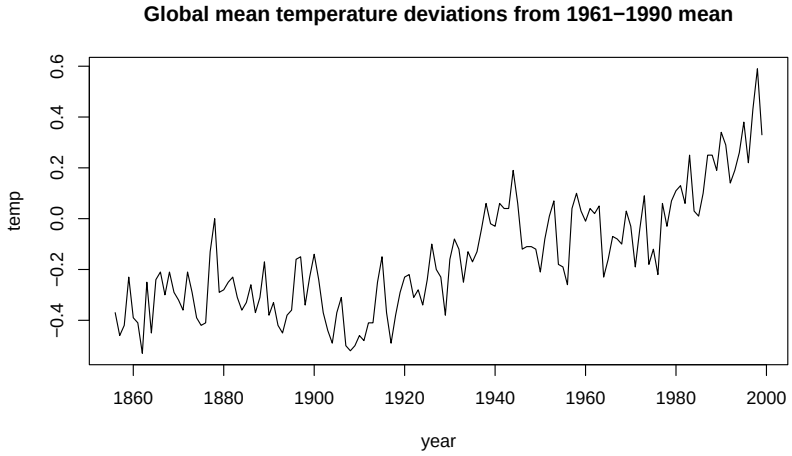


The figures give the average amount of energy per square metre carried by the different sorts of radiation when the whole system is in balance. In fact, without greenhouse gases the surface of the earth

would be very chilly, on average probably around 30C cooler than now, although CO_2 is only one of the contributers to this.

So, CO_2 has increased in the atmosphere as a result of human activities, and there are good scientific reasons to believe that this will increase temperatures in the lower atmosphere and on the earth's surface. Has this happened? The dataset `CT150`, contains average global temperatures since 1856, which is as far back as reliable direct measurements of temperature extend:

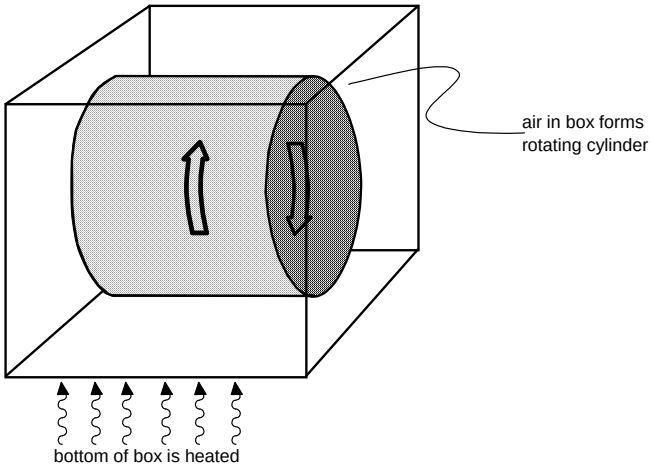
```
> data(GT150)
> attach(GT150)
> plot(year,temp,type="l",main="Global mean temperature deviations from 1961-1990 mean")
```



This plot does appear to show an increase in global mean temperature over the period of most rapid increase in CO_2 concentrations, but it's much less clear cut that the plot for CO_2 increase. In particular what is striking is how variable the mean temperature appears to be. It is this that causes sceptics to doubt that climate change is really happening. They argue that since climate is highly variable anyway, the apparent recent upswing in the data is just one of those chance events that is likely to happen from time to time. We'll examine this claim shortly using statistical models, but first let's have a look at how deterministic mathematical models cope with modelling weather and climate, and what they can tell us. This will also help explain why weather forecasting is so difficult and short term, and why chaos is yet another phenomenon that undermines determinism.

3 A simple mathematical weather model: the Lorenz model.

As part of ongoing attempts to understand the atmosphere, in order to be able to predict its behaviour, Lorenz suggested an extremely simple mathematical model of a hugely simplified atmosphere. At its most basic, the atmosphere is a volume of gas which is warmer at the bottom than at the top. This gross simplification is the basis of the Lorenz model. Specifically, imagine a closed rectangular box filled with air, and suppose that you apply a heat source uniformly to the bottom of the box. The physics of fluid flow tells you what will happen to the air in the box: if the bottom of the box is hot enough the air will start to circulate as illustrated here...



The warm air at the bottom starts to rise and cool air from the top has to sink to replace it - this leads to a cylindrical pattern of airflow with warm air rising up one side of the box and cold air sinking down the other side. It turns out that this situation can be described quite accurately by a simple model consisting of 3 ordinary differential equations[4]:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \delta(y - x) \tag{3}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = rx - y - xz \tag{4}$$

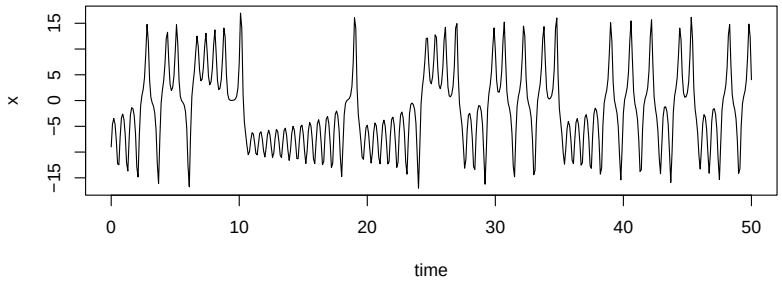
$$\frac{\mathrm{d}z}{\mathrm{d}t} = xy - bz \tag{5}$$

The model variables have the following meanings: x is the rate of rotation of the cylinder of air - essentially this gives the wind speed in the model atmosphere; y is the difference in temperature between the vertical sides of the cylinder, in effect between the rising warm air and the falling cold air; z measures how different the top-to-bottom temperature profile is from a straight line. The *parameters* of the model are as follows: δ is the ratio of the viscosity of the air to its thermal conductivity (!?); r is the difference in temperature between bottom and top of the box; b is the ratio of the width of the box to its height. The most interesting parameter for this case study is r, since we expect increased CO_2 concentrations to increase the temperature difference between the bottom and top of the atmosphere. $\delta = 10$, $r = 28$ and $b = 8/3$ are sensible parameter values.

Now let's solve the model numerically (it isn't possible to find a solution analytically). The `R` package `mt1007` includes a function to efficiently and accurately solve this model.

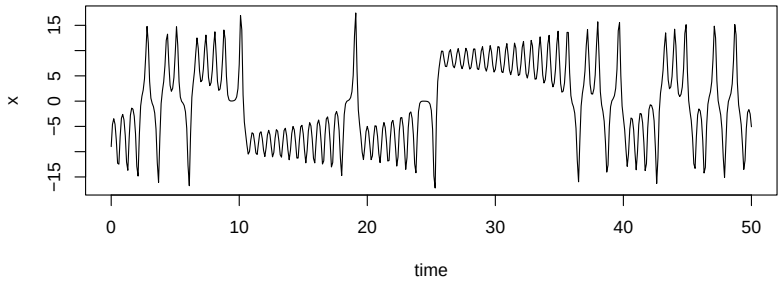
```
> library(mt1007)
> s<-c(-9,-3,32);
> co<-c(10,28,8/3)
> l<-lorenz(0,50,dt=0.1,s,co,tol=1e-7)
> plot(l$t,l$x,type="l")
```

[4]This will be the last differential equation model used!



Note that the alternation between positive and negative values of x corresponds to the direction of circulation changing. The irregular nature of the model solution is quite surprising, given the simplicity of the model that generates it. Even odder is what happens when a change of about 1 part in a 10 million is made to the starting value of x...

```
> s<-c(-9.000001,-3,32);
> l2<-lorenz(0,50,dt=0.1,s,co,tol=1e-7)
> plot(l2$t,l2$x,type="l",xlab="time",ylab="x")
```



The model solution starts out the same as before, but at some point it diverges completely. This property is not some sort of error introduced by solving the equations numerically: it can be proven to be intrinsic to the equations themselves. If we run the model twice with different starting conditions the two solutions will always diverge like this eventually, no matter how close the different starting values. Furthermore the phenomenon does not appear to be some sort of pathological property of the equations: the weather really does display this sort of sensitivity. Such extreme sensitivity to perturbations is known as chaos: many non-linear systems are chaotic in some circumstances.

Chaos is important because it further undermines the idea that nature is deterministic. The existence of chaos means that even if we can perfectly describe the workings of some natural system with a deterministic mathematical model, this does not necessarily mean that we can predict its behaviour far into the future. If the system is chaotic, then no matter how accurately we measure its state, the error left in the measurement will always mean that the error in our predictions will eventually grow so large that they become useless. This is exactly the problem that faces weather forecasters. They never know the state of the atmosphere perfectly, and in any case cannot produce a model that is a perfect description of the atmosphere. Because the atmosphere and models of it often behave in a chaotic way, these errors mean that forecasts can only ever be short term.

Chaos is another reason for considering alternatives to the often hopeless goal of producing models to predict exactly what a system will do: that is to work with statistical models that deal with probabilities of particular system behaviours.

4 Is the recent warming more than "random variation"?

Sceptics argue that the case for climate change actually happening is weak. They argue that there are so many complicating factors in the way that climate is controlled that we really cannot tell whether increasing $C0_2$ concentration will have a serious impact on the climate, or whether all sorts of natural control mechanisms (decreased water vapour in the atmosphere, increased cloud cover or whatever) will mean that the changes will be barely noticeable. After all, they argue, the Earth's climate has stayed within quite tight limits (give or take the odd ice age) for hundreds of millions of years, so there must be some natural control mechanisms in place.

In addition, as we have seen by examining the behaviour of the Lorenz model, there are considerable difficulties inherent in using deterministic models to help understand climate - we can never expect nature to support our models in the uniquivocal way that the motion of the planets supported Newton, Laplace and later Einstein. This isn't to say that nothing useful can be learned from deterministic climate models, just that we will need to use alternative approaches as well.

One of the main planks of the argument against climate change being real is that the data just don't provide enough support for it. The basic reasoning goes as follows:

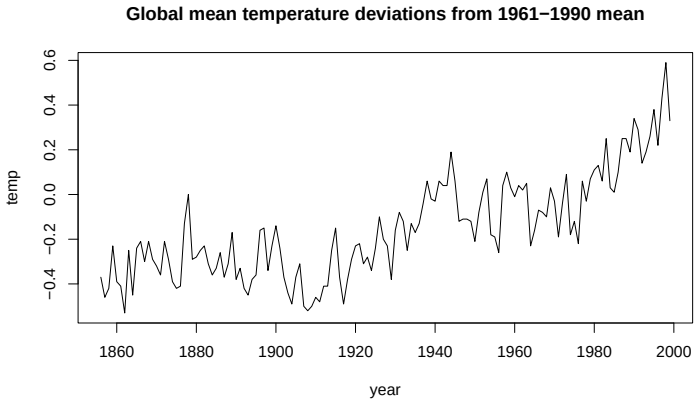
> The climate varies widely and randomly from year to year. This means that, even if the climate is not changing in the long run, every now and then we are bound to get a chance sequence of increasingly warm years like the sequence in the real record.

— this is an argument that we can investigate quite rigorously using statistical models. If the argument has substance, then we ought to be able to produce a statistical model that produces temperature trajectories that:

1. look like the real one,
2. have a reasonable probability of every now and then going on an excursion like the one seen at the end of the real data, but
3. do not display any long term trend.

If we can not do this, then it is reasonable to conclude that the recent increases are more than mere chance (or at least that the sceptics really need to be able to come up with a reasonable statistical model that *can* behave in the way that they are claiming the real system behaves).

Here, again, are the real data:



To build a sceptics statistical model of these data we need a mathematical way of building randomness into the model. The basic idea is that if the climate had started from a slightly different initial state in 1856, we would have got a different series - some of its properties would be like the original series, but the actual temperature in any given year would bear little or no relation to the temperature that we actually observed given the actual state in 1856. How can we model this sort of non-repeatability mathematically? The answer is to introduce the concept of a **random variable**.

4.1 Random variables

Before introducing random variables recall the definitions of two characteristics of a set of data $x_1, x_2, \ldots, x_n$.

1. The **mean** (or average) of the set of data is[5]
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$
2. The **variance** of the set of data is
$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
and is a measure of how spread out the data are around their mean (don't worry about the $n-1$ at this stage - it will be explained in the theory section).

Random variables will be given a fairly rigorous mathematical definition in the theory section of the course, for the moment they will be defined by their behaviour. Essentially a random variable is a variable which has a different value every time you make an observation of it. The defining property of a random variable is its distribution: this defines the probability that the variable takes any of its possible values. Here are some examples.

- Let X be a random variable which can take the values 0, 1 or 2, and suppose that the following table summarises its distribution.

x	0	1	2
$\Pr[X = x]$	0.3	0.5	0.2

What this means is that if we take an observation of X we can not say in advance exactly what value it will have. All we can say is that it has a probability of 0.3 of taking the value 0, a probability of 0.5 of taking the value of 1 and probability 0.2, of taking the value 2. In other words, if we took a very large number of observations of X we would expect about three tenths of them to be 0's, about half of them to be 1's and about a fifth of them to be 2's.

- Let Y be a random variable which can take any value in the interval $(0, 1)$ with equal probability. This means that before we take an observation of Y all we can say is that any value between zero and one is equally likely to occur, although the observation will not be outside that interval. Y is said to have a uniform distribution on $(0, 1)$: the mathematical notation for writing this compactly is $Y \sim U(0, 1)$. R has many built in functions for generating observations of random variables for you. The following generates 12 observation of Y:

```
> runif(12,0,1)
 [1] 0.37199711 0.73802148 0.57553934 0.22207002 0.70358750 0.87645082
 [7] 0.86158794 0.52360403 0.08799612 0.90849577 0.12890279 0.48354116
```

As you would expect, calling function `runif()` again yields a different set of numbers:

```
> runif(12,0,1)
 [1] 0.76815408 0.69212806 0.54376955 0.73546762 0.46736725 0.30032586
 [7] 0.92261148 0.04319168 0.92960260 0.25693702 0.87677743 0.45579485
```

- A rather useful model for all sorts of situations is provided by random variables that have a so called **normal distribution**. The normal distribution has two controlling parameters μ and σ^2. Writing $Z \sim \mathbf{N}(\mu, \sigma^2)$ means "Z has a normal distribution with parameters μ and σ^2". R has a function `rnorm()` for simulating normal r.v.s. For example let's simulate 12 normal random deviates with $\mu = 1$ and $\sigma^2 = 16$:

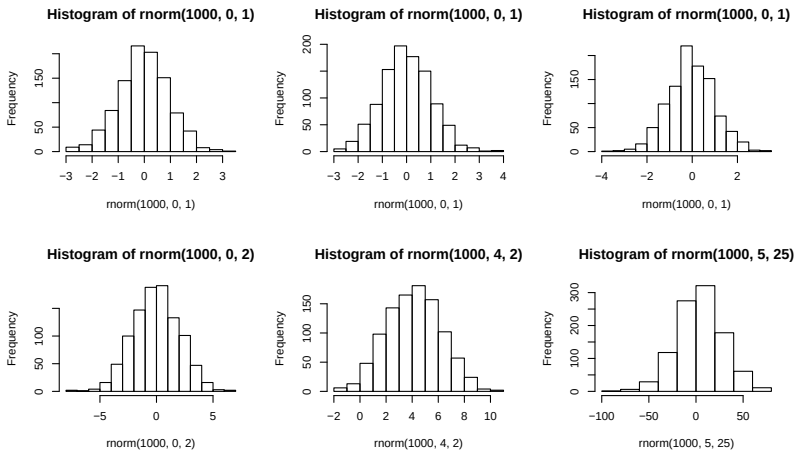
[5] Remember that $\sum_{i=1}^{n} z_i = z_1 + z_2 + z_3 + \ldots + z_n$, i.e. it is the sum of z_1, z_2, z_3 and so on up to z_n.

```
> rnorm(12,1,4) # note that rnorm expects sigma NOT sigma^2
 [1]  5.54235869 -5.08229221  7.98561034  2.25316833  3.23439155  2.50812412
 [7] -1.43311807 -0.92278691  7.22904658 -0.07715744  1.33164917 10.73981884
```

It is important to get a feel for the shape of the normal distribution: i.e. to get a feel for which values the r.v. is likely to take, which values are unlikely and which are practically impossible. To do this let's use R to simulate some normal r.v.s and plot histograms[6] of these. These commands:

```
> par(mfrow=c(2,3))
> hist(rnorm(1000,0,1))
> hist(rnorm(1000,0,1))
> hist(rnorm(1000,0,1))
> hist(rnorm(1000,0,2))
> hist(rnorm(1000,4,2))
> hist(rnorm(1000,5,25))
```

produced this output:



The first three plots are histograms of three different size 1000 samples from $\mathrm{N}(0,1)$. You can see that the random variable is concentrated around zero, with probability of occurance tailing away above and below zero. Because we are looking at finite samples of observations of random variables, the 3 histograms are not identical, but they do all show broadly the same pattern. Notice, for example, that the probability of being outside the interval (-2,2) is small in all cases (about 0.05 in fact). The fourth plot (lower left) shows what happens when σ^2 is increased to 4, while μ is left at 0: the range of probable values more or less doubles. Shifting μ up to 4 shifts the whole distribution up by 4, so that it is now centred on 4 rather than 0. Finally μ is set to 5 while σ^2 is set to 625. The distribution is now centred on 5 but has become much more spread out with few values lying outside (-45, 55).

The parameter μ is actually the mean of the Normal distribution. This means that if $Z \sim \mathrm{N}(\mu, \sigma^2)$ and we take n independent observations of Z, $z_1, z_2, \ldots z_1$, then as $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n} z_i \to \mu$.

The parameter σ^2 is the **variance** of the Normal distribution. The variance of a random variable is the average squared difference between the r.v. and its mean. In this case that means that as $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n}(z_i - \mu)^2 \to \sigma^2$.

[6] Remember that the defining feature of a histogram is that the area of each histogram bar is proportional to the number of datapoints that fall within the interval covered by the width of the bar.

4.2 Building a (sceptics) simple statistical model of the temperature anomaly series

Having introduced the notion of random variables, it is now possible to try constructing a suitable model of the temperature series, *that will embody the sceptics ideas about what is happening to the climate*. A sensible strategy to follow is to start simple, and only add just enough complexity to do an adequate job of modelling the series.

The simplest statistical model for the series is probably to assume that the temperature is a constant plus a random error term. Formally, our model is that global mean temperature in year i is:

$$T_i = \alpha + E_i$$

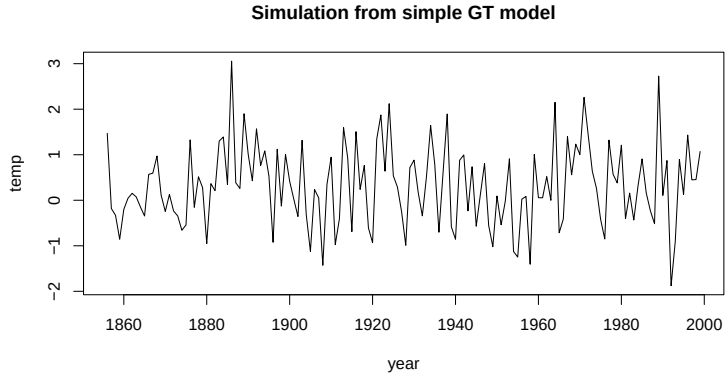
where E_i is a random variable and α is the average temperature. To complete the specification of the model we need to give E_i a distribution. Let's choose a Normal with mean 0 and variance σ^2. i.e.

$$E_i \sim N(0, \sigma^2)^7$$

Given values of α and σ^2, the model provides a recipe for generating data. The idea is that if the model is a good one, then it provides a recipe by which the original data could have been generated.

For example if $\alpha = 0.5$ and $\sigma = 0.9$ then we could simulate a set of data in R as follows:

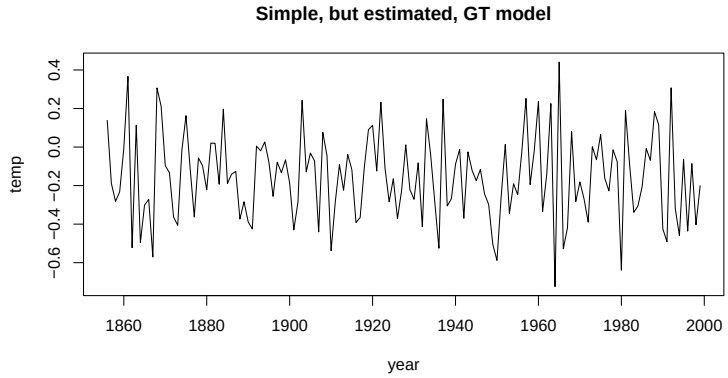
```
> alpha<-0.5
> e<-rnorm(144,0,0.9)
> temp<-alpha+e
> year<-1856:1999
> plot(year,t,type="l",main="Simulation from simple GT model" )
```



If you compare this with the real data it doesn't look great, but of course one problem is that α and σ^2 were just plucked out of the air, and it would make more sense to choose values that are representative of the data. Let's estimate α and σ^2 by setting them equal to the mean and variance of the data, respectively. This is easily done:

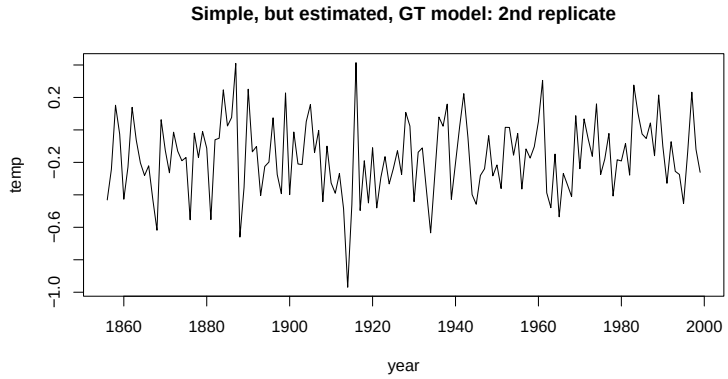
```
> data(GT150)
> alpha<-mean(GT150$temp)
> sigma<-var(GT150$temp)^0.5
> temp <- alpha + rnorm(144,0,sigma)
> plot(year,temp,type="l",main="Simple, but estimated, GT model" )
```

[7] We could have written down this model more simply as $T_i \sim \mathrm{N}(\alpha, \sigma^2)$, but this would not have been so easy to extend.



This is better, but still doesn't really look anything like the original series: the variations from one year to the next are too pronounced, the model never produces runs of warm or cool years in the way that the real data seem to. Clearly we wouldn't convince anybody that this model was a suitable description of the climate, at any level. But before moving on to try other models it's worth re-emphasising one key point about statistical models. Re-simulating the model under *identical* circumstances, gives a different trajectory:

```
> temp <- alpha + rnorm(144,0,sigma)
> plot(year,temp,type="l",main="Simple, but estimated, GT model: 2nd replicate")
```



The probabilities of the model being in various states haven't changed, but the details of what actually happened are quite different.

4.3 An alternative simple model

The first attempt at modelling the global temperature series produced model trajectories that looked nothing like the data, and hence can't be used as the basis for drawing any conclusions about the climate. Part of the problem was that we never got the runs of warm or cool years that the real data seem to show. As an alternative, let's try a model in which the temperature this year is given by the temperature last year plus a random term. i.e. the model for the temperature in year $i+1$ is:

$$T_{i+1} = T_i + E_i \quad \text{where} \quad E_i \sim \mathrm{N}(0, \sigma^2)$$

This sort of model is known as a "Random walk" (it actually works very well as a model of the way in which a molecule of gas moves in still air). In a way, this is an even simpler model than the previous one, in that it only has one parameter to estimate: σ^2. To see how to estimate σ^2, which is just the variance of E_i, note that:

$$E_i = T_{i+1} - T_i.$$

So we could estimate σ^2 from the data by finding the variance of the year to year *differences* in temperature:

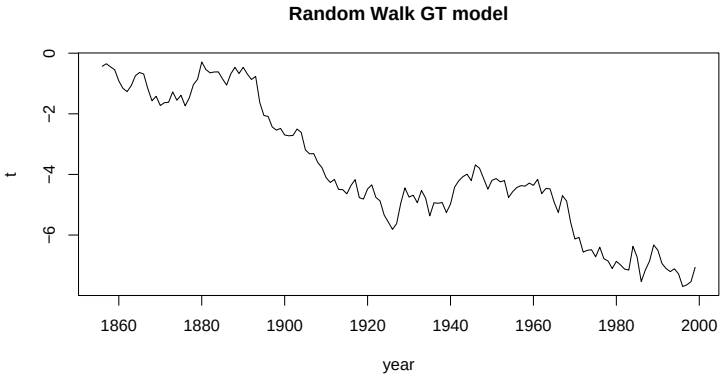
```
> attach(GT150)
> diff<-temp[2:144]-temp[1:143]
> sigma<-var(diff)^0.5
```

(The estimate of σ is 0.1220706.). Simulating from this model requires that we start with the first year's temperature and generate the next year's temperature by adding a normal random deviate to it; to get the third year's temperature we add another normal random deviate to the second year's temperature, and so on. We need to introduce a new feature of R to do this: a loop:

```
> t<-array(0,144)         # set up an array to hold temperatures
> t[1]<-temp[1]           # set first element to observed temp in 1856
> e<-rnorm(143,0,sigma)   # simulate random temperature changes
> for (i in 1:143) t[i+1]<-t[i]+e[i]  # simulate temperature record
```

The `for` loop repeats the instruction `t[i+1]<-t[i]+e[i]` for all integer values of `i` from 1 to 143. i.e. first the computer does `t[2]<-t[1]+e[1]` then it does `t[3]<-t[2]+e[2]`, and so on. Hence `t` now contains the simulated model temperatures.

```
> plot(year,t,type="l",main="Random Walk GT model")
```



Now, we seem to have the opposite problem to the previous model — there is too much correlation between one year and the next. More seriously, this model is completely inappropriate for representing the sceptics position, since it always produces long term trends: in fact under this model the temperatures are completely unbounded, they could "randomly walk" to any value at all given long enough (even a non-sensical negative one), and we know that the earths climate isn't *that* variable. Clearly we still need something a bit better.

4.4 "Auto-regressive" models

The problem with the first model was that it produced too much year to year variation and no longer term fluctuations. The second model produced too little year to year variation, and had the unfortunate ability to wander off towards $\pm\infty$, given enough time. A model in between these two extremes might be a good idea. Auto-regressive models provide just such middle models.

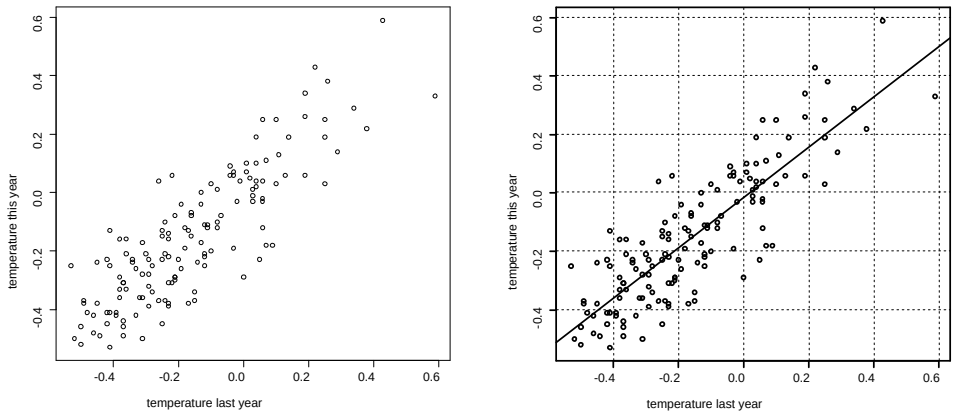
The basic idea is to set up a model in which this years temperature anomaly is a constant plus some proportion of last years temperature anomaly plus a random variable. i.e.

$$T_{i+1} = \alpha + \beta T_i + E_i \quad \text{where} \quad E_i \sim \mathrm{N}(0, \sigma^2)$$

Provided $\beta < 1$, this model should be bounded, unlike the simple random walk. How can α , β and σ^2 be estimated? The trick is to recognise that our model states that there is a straight line relationship

between T_{i+1} and T_i, corrupted by some random "error". Hence, if we plot the temperature anomalies for each year against the temperature anomalies the previous year, we should see a straight line relationship, from which α and β can be estimated. To do this, create an array of temperatures from 1857-1999 and an array of temperatures from 1856-1998, and plot one against the other:

```
> attach(GT150)
> t0<-temp[2:144]
> t1<-temp[1:143]
> plot(t1,t0,xlab="temperature last year",ylab="temperature this year")
```



The plot on the left is the raw plot, while the plot on the right superimposes an approximating straight line and a grid to help estimate its slope and intercept. The intercept term appears to be about -0.02, while the line goes up 3.5 squares while going across 4 squares, making a slope of about $3.5/4 \approx 0.88$. Hence -0.02 and 0.88 are reasonable estimates for α and β respectively.

What about σ^2? σ^2 is the variance of the E_i terms. Re-arranging the model gives:

$$E_i = T_{i+1} - \alpha - \beta T_i$$

Hence if we find the variance of the observed values of:

$$T_{i+1} + 0.02 - 0.88 T_i$$

we should get a reasonable estimate of σ^2 (recall that observed T_{i+1} are in `t0` while observed T_i are in `t1`):

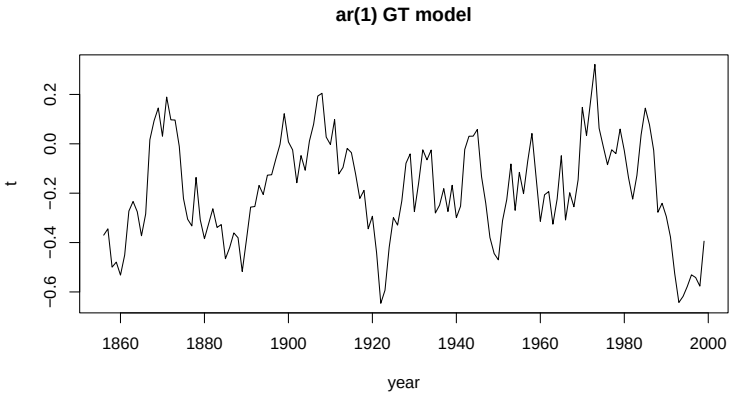
```
> var(t0+0.02-0.88*t1)
[1] 0.01398725
```

The R commands to simulate from this model and plot the results are straightforward. Complete them yourself:

```
> t<-array(0,144)
> t[1]<-temp[1]
> e<-rnorm(144,0,0.014^0.5)

> for (i in 1:143)

> plot(year,
```



. . . clearly we are now well on the way towards having a reasonable model. Given how much better this model is than the previous efforts, it is worth thinking more carefully about estimating its parameters: drawing a line by eye is not very objective, and it is possible to come up with much better mathematical methods.

4.4.1 Least squares estimation

Drawing a line on the T_{i-1} vs T_i plot was a matter of trying to find a line that on average is not too far from any of the data points. This is a sensible aim to keep in mind when constructing a mathematical method for estimating the slope and intercept of the model. So, to start constructing a mathematical method for estimating α and β, we need first to find some way of measuring how well the line fits the data. i.e. how well does the line $\alpha + \beta T_i$ fit the observations T_{i+1}? Consider the differences between T_{i+1} and $\alpha + \beta T_i$. Some sort of total of these differences would be a good measure of how well the model fits. A straight sum of the differences would be one possibility:

$$\sum_{i=1}^{n-1} (T_{i+1} - \alpha - \beta T_i)$$

. . . but the problem with this is that big positive terms in the sum cancel big negative terms, which means that the line can be a long way from some T_{i+1}'s without this measure of fit noticing. An obvious alternative would be to use:

$$\sum_{i=1}^{n-1} |T_{i+1} - \alpha - \beta T_i|$$

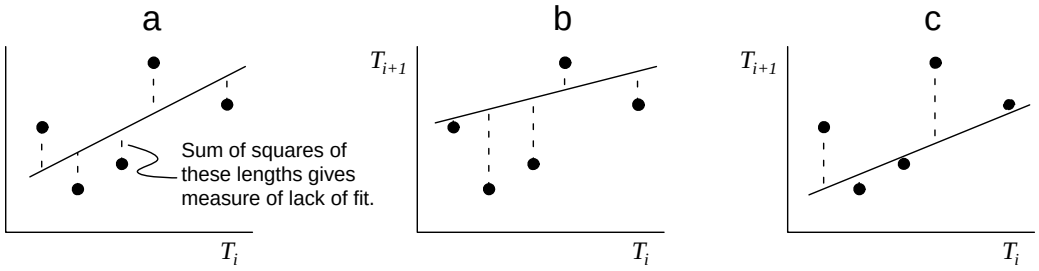
which would eliminate the problem with the straight summation, but suffers from the disadvantage of being mathematically quite awkward to work with because of having to take magnitudes. A mathematically easier quantity to work with is the sum of squared differences:

$$\sum_{i=1}^{n-1} (T_{i+1} - \alpha - \beta T_i)^2$$

. . . and methods based on this measure of model fit have some very "nice" properties.[8]

The following figure schematically illustrates what is being done:

[8] This measure of fit is also a very natural measure if the differences between the line and the data T_{i+1} can be modelled as observations of normal r.v.s, but more theoretical apparatus is needed in order to pursue this point.



a) shows a reasonably well fitting line, for which the sum of squares should be fairly low. b) and c) show poorly fitting lines where some of the deviations are large, leading to very large squared deviations and high sums of squares.

So, an objective way of finding the best fitting line would be to find the slope and intercept which minimise the mean square difference between the line and the T_{i+1}'s. That is, find the values of α and β that minimise the mean square difference. Estimating parameters in this way is known as the **method of least squares**. It turns out to be fairly straightforward to find the parameters minimising the mean square difference between line and data: the mathematical details will be given in the theory section. Quite general formulae can be found for the least squares estimates and R has these formulae built in. The appropriate R function is `lm()`. Recalling that `t0` contains a set of T_{i+1} values and `t1` the corresponding T_i values, finding the parameters α and β using `lm()` is simple:

```
> model<-lm(t0~t1)
> coef(model)
(Intercept)          t1
-0.01655453  0.86331118
```

...but needs some explanation! `lm()` requires a "model formula" to tell it what to do. The formula in this case is `t0~t1`. `lm()` interprets this as "find the parameters α and β that make the straight line $\alpha+\beta$`t1` best fit the data in `t0` in the least squares sense" or more concisely " fit a straight line to the x, y data in `t1`, `t0`". `lm()` returns a "fitted model object" which I have chosen to call `model`. `model` contains all sorts of useful information, but for now we only require parameter estimates. These can be extracted with the `coef()` function as shown (`model$coef` could also have been used). α, the intercept of the line, is estimated to be -0.0166, while β, the parameter multiplying the `t1` values, is estimated to be 0.8633.

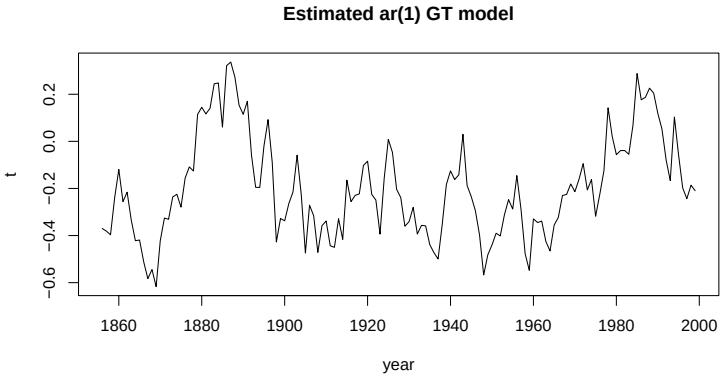
σ^2, the variance of the E_i's, still has to be estimated. Again this can be achieved by looking at the variance of the estimated values of $T_{i+1} - \alpha - \beta T_i$, but the `model` object already contains all the information for R to do this automatically:

```
> summary(model)$sigma^2
[1] 0.01407252
```

So all the model parameters have now been estimated objectively, rather than "by eye". Simulating the model with properly estimated parameters:

```
> sigma<-summary(model)$sigma
> alpha<-coef(model)[1]; beta<-coef(model)[2]
> e<-rnorm(144,0,sigma)
> t<-array(0,144); year<-1856:1999; t[1]<-GT150$temp[1]
> for (i in 1:143) t[i+1] <- alpha + beta*t[i] + e[i]
> plot(year,t,type="l",main="Estimated ar(1) GT model")
```

gives the result ...



4.4.2 Covariance, correlation and the acf

Now that we have a model that isn't obviously wrong, we need to devote a little more effort to finding ways of assessing just how close to reality it is. We are interested in knowing whether the randomness in the temperature series, coupled with the way in which one year's temperature is related to previous years' temperatures, is sufficient to produce runs of increasing temperature like the one that has happened, or whether we'll have to look for another explanation of the temperature rise (i.e. CO_2 driven climate change). The dependence structure of one year's temperature on previous year's temperatures is crucial here: what is needed is a direct way of representing this, in order to allow comparison of model output and the real data. To see how to do this we need a new concept: covariance.

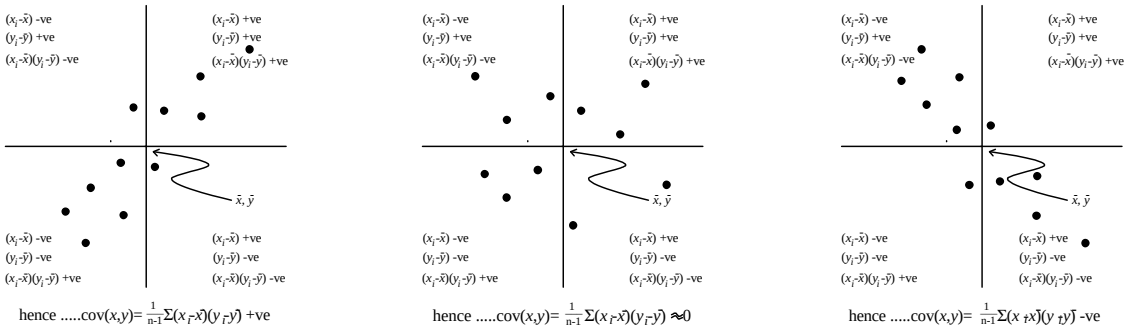
Recall that the variance of a set of data $x_1, x_2, \ldots, x_n$ is:

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The **covariance** of one set of data $x_1, x_2, \ldots, x_n$ with another set $y_1, y_2, \ldots, y_n$ is

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

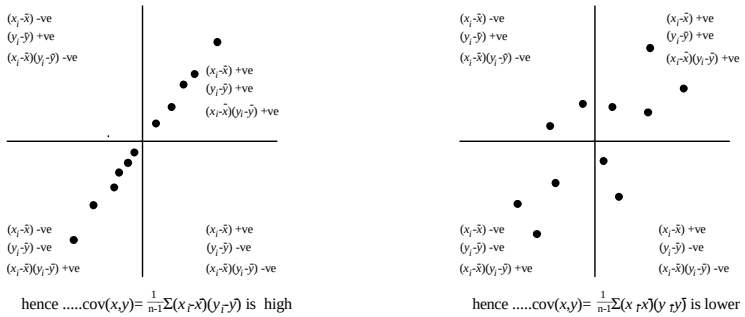
Notice how the covariance of the x_i's with themselves is just the variance of the x_i's. The following figure attempts to illustrate what it is that covariance measures:



The figure on the left shows a *positively* correlated set of x and y data (i.e. large x values correspond to large y values, small x values to small y values). The figure is split into 4 quadrants: $y > \bar{y}$ and $x > \bar{x}$; $y < \bar{y}$ and $x < \bar{x}$; $y < \bar{y}$ and $x > \bar{x}$; $y > \bar{y}$ and $x < \bar{x}$. Clearly, from the way that the points fall into the quadrants, this picture corresponds to positive covariance. The middle panel shows covariance quite close to zero (at least when compared to the *variances* of x and y) ...there are roughly equal numbers of points in all 4 quadrants, so that positive terms in $\sum(x_i - \bar{x})(y_i - \bar{y})$ will tend to balance the negative

terms, leading to a covariance near zero. The final panel on the right shows *negatively* correlated x and y data. In this case most points fall into the quadrants making a negative contribution to $\sum(x_i - \bar{x})(y_i - \bar{y})$, so the covariance is negative.

Notice how data lying very close to a straight line will tend to have mostly terms of the same sign in the covariance, which tends to increase the magnitude of the covariance. More scattered data will have more points falling in all 4 quadrants, so that there is always some cancelling of negative and positive terms, leading to a reduction in covariance:



The only problem with covariance, as a measure of how variables are related to each other, is that it mixes up the variability of the variables with their relatedness. For example, if x is very variable then $(x_i - \bar{x})$ will typically be large so that $\sum(x_i - \bar{x})(y_i - \bar{y})$ will tend to be large, even if x and y show little correlation. This can make the interpretation of covariances a bit difficult. A more interpretable measure of how related two variables are is the **correlation coefficient**, r which is obtained by dividing the covariance of x and y by their standard deviations:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

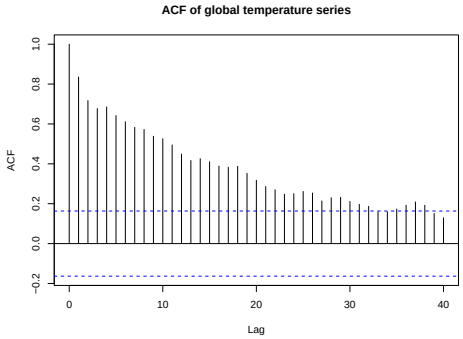
r measures the relatedness of x and y independent of how variable they are individually. It is positive, negative or near zero, in the same circumstances as the covariance, but is bounded between -1 and +1. $r = 1$ (-1), correspond to points lying exactly of a straight line, with positive (negative) slope.

Looking at the correlation between temperatures in neighbouring years is a very useful way of examining the patterns of relatedness between different years. For example, here is how you could work out the correlation between one year's temperature and the next:

```
> data(GT150)
> t0<-GT150$temp[2:144]
> t1<-GT150$temp[1:143]
> cor(t0,t1)
[1] 0.8520122
```

If we work out the correlations between year i and year $i - m$ for a range of m's and then plot these correlations against m, we'll get what is known as an **a**uto **c**orrelation **f**unction (acf). Production of acf plots is built in to R in package `ts`...

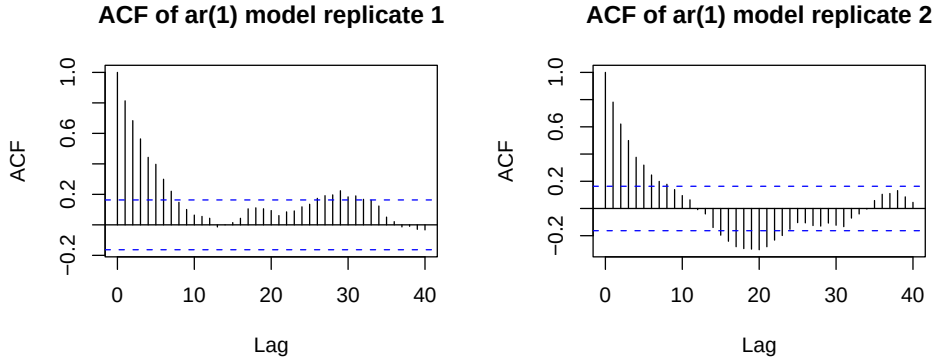
```
> library(ts)
> acf(GT150$temp,lag.max=40,main="ACF of global temperature series")
```



The height of the bars in this plot give the correlation coefficient between years separated by the `Lag` given on the x axis. Obviously the correlation at lag 0 is 1, since this is the correlation between the temperature in a year and itself. Between the horizontal dotted lines is the region where 95% of correlation coefficients will lie **if** there is really no relationship between the variables: these lines help you judge which correlation coefficients you should treat as important (i.e. ones outside the region between the dashed lines).

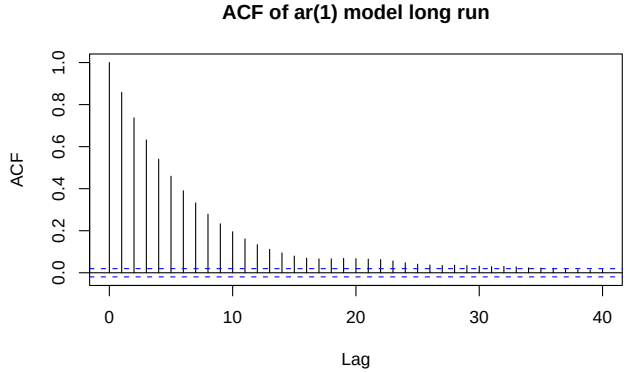
So, the acf gives a good summary of the relationship between temperatures in neighbouring years. If we have a good model, we would expect the acf's from model runs to look like the real acf. Let's see if they do for a couple of realisations of the current model:

```
> acf(t,lag.max=40,main="ACF of ar(1) model replicate 1")
> e<-rnorm(143,0,sigma)
> for (i in 1:143) t[i+1] <- alpha + beta*t[i] + e[i]
> acf(t,lag.max=40,main="ACF of ar(1) model replicate 2")
```



These plots indicate two things: firstly the models acf tails off rather more quickly than the original data, and secondly, the model acf's are quite variable between runs. An easy way to avoid the latter problem is to run the model for more years - this will much reduce the variability in the acf, leading to one less source of variability to obscure the comparison of model and data:

```
> t<-GT150$temp[1]
> for (i in 1:9999) t[i+1] <- alpha + beta*t[i] + e[i]
> par(mfrow=c(1,1))
> acf(t,lag.max=40,main="ACF of ar(1) model long run")
```



. . . a second replicate of this would be indistinguishable. Clearly the model correlation is tailing off too quickly, but one further round of model revisions will sort this problem out.

4.4.3 Higher order auto- regressive models

The simple auto- regressive model seems to display too little correlation over the longer term. This suggests that the temperature this year should depend not just on last year's temperature, but temperature over several previous years. For example, limiting the model to a 5 year lag we might try:

$$T_i = \alpha + \beta_1 T_{i-1} + \beta_2 T_{i-2} + \beta_3 T_{i-3} + \beta_4 T_{i-4} + \beta_5 T_{i-5} + E_i \quad \text{where} \quad E_i \sim N(0, \sigma^2)$$

We can estimate this model using the least squares method, just as we did for the previous model. This time we seek the parameters that minimise the sum of squares of differences between $\alpha+\beta_1 T_{i-1}+\beta_2 T_{i-2}+\beta_3 T_{i-3}+\beta_4 T_{i-4}+\beta_5 T_{i-5}$ and T_i. First create arrays containing all the lagged data:

```
> t0<-GT150$temp[6:144];t1<-GT150$temp[5:143];t2<-GT150$temp[4:142]
> t3<-GT150$temp[3:141];t4<-GT150$temp[2:140];t5<-GT150$temp[1:139]
```

Then fit the model and examine the parameter estimates:

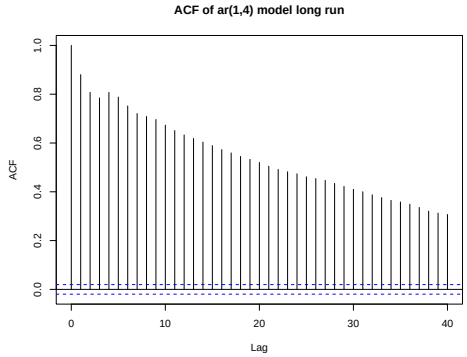
```
> mod<-lm(t0~t1+t2+t3+t4+t5)
> coef(mod)
 (Intercept)           t1           t2           t3           t4           t5
 0.008559785  0.599864792 -0.015875963  0.091810501  0.232809845  0.070571630
```

Because `t1 - t5` are all measurements of the same thing, it is valid to compare the magnitude of their associated parameter estimates directly: β_1 and β_4 are quite a bit larger than all the others, and will tend to dominate what the model produces. So, in the interests of simplicity, it's worth seeing if we can get away with a model featuring only β_1 and β_4:

```
> mod<-lm(t0~t1+t4-1)
> coef(mod)
       t1        t4
0.6347583 0.3154124
```

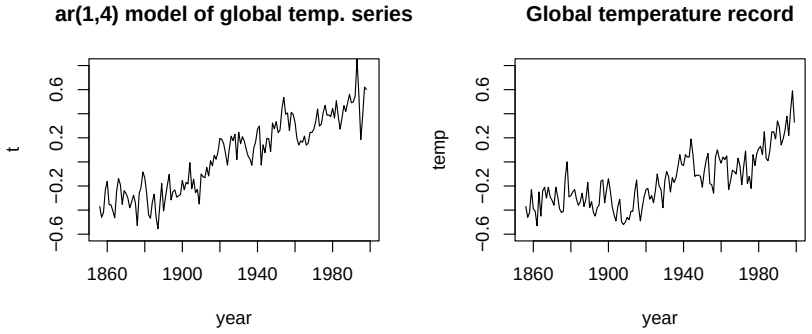
(the `-1` term in the call to `lm()` tells it not to include a constant α.) Notice that the parameter estimates for β_4 and β_1 change when the other terms are left out of the model - this is why it was important to re-estimate them. It's now straightforward to simulate from the new model to get an acf:

```
> sigma<-summary(mod)$sigma
> beta<-coef(mod)
> rm(t);t<-GT150$temp[1:5]
> e<-rnorm(9999,0,sigma)
> for (i in 5:9999) t[i]<-beta[1]*t[i-1]+beta[2]*t[i-4]+e[i]
> acf(t,main="ACF of r(1,4) model long run")
```

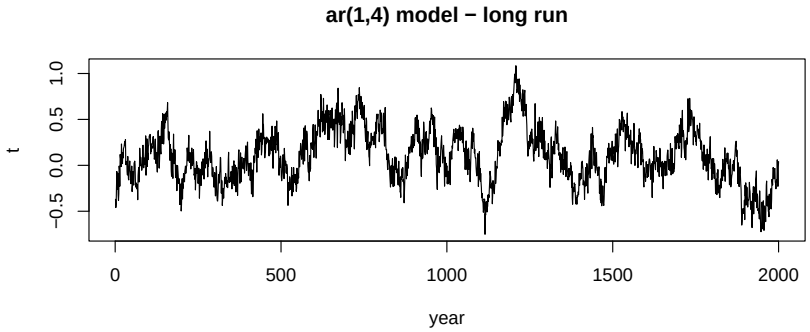


This is a much better match than the simpler model, although it has a slightly higher correlation at high lags than the data show. However, the difference is not substantial, bearing in mind the amount of variability that we expect in the acf for short series. Hence it is reasonable to decide that this model is good enough. Here is a simulation from it, next to the real data:

```
> t<-GT150$temp[1:5]
> year<-1856:1999
> e<-rnorm(143,0,sigma)
> for (i in 5:144) t[i]<-beta[1]*t[i-1]+beta[2]*t[i-4]+e[i]
> plot(year,t,type="l",main="ar(1,4) model of global temp. series",ylim=c(-0.6,0.8))
> plot(year,GT150$temp,ylab="temp",main="Global temperature record",type="l",ylim=c(-0.6,0.8))
```



. . . the simulation shown has a bigger temperature rise than the real data, purely as a result of random variation and the simple year to year correlation built into the model. It took me 7 simulations to find this example, and this seems to be quite typical - although the majority of model runs do not show as big a fluctuation as the real data, a sizeable minority do show such changes. It is worth looking at a longer run, to see how often the model produces this size of fluctuation and to check that the model indeed has no long term trend (as required by the sceptics hypothesis):



So, this exercise based on the directly measured record shows that the climate change sceptics could have a point. It is not hard to come up with simple statistical models, capable of producing series that look much like the real data, whose long term behaviour is not steady increase, but which can show quite marked runs of increasing temperature in the short term. The model produced here is not especially unreasonable, either. The Earth's climate does show multi-year cycles of various sorts (for example the Southern Oscillation, and the North Atlantic Oscillation). Correlation between climate in different years is also to be expected as a result of features of the system like the fact that oceans have enormous thermal inertia, so that their temperatures change very slowly.

On the other hand, our analysis has at least one major flaw. The record that we used to estimate the model parameters is rather short in relation to the timescale over which the recent change has happened. It is also the period of time over which large amounts of CO_2 have been released into the atmosphere. Hence, if climate change is happening, our analysis is a bit circular - we have set up a model of an atmosphere that is in "steady state" in the long term, but estimated its parameters from an atmosphere that might be changing dramatically - in this case perhaps it's not surprising that our model had no trouble producing the sort of rise actually seen in the data. For this sort of analysis to be conclusive we would have to examine a longer temperature record, which does not have the problem that it may be dominated by the very effect that we want to investigate.

4.5 Longer temperature records: climate reconstruction

As we have seen, to really figure out whether the recent rise is something more than the result of the random wanderings of the climate, we need a rather longer temperature record, less dominated by the period in which we suspect changes may be taking place. The difficulty with getting longer records is that prior to the middle of the nineteenth century people were not measuring temperature accurately enough, often enough, or over enough of the globe. The answer to this impasse is to use "climate reconstructions".

Average atmospheric temperature affects quite a large number of biological and physical processes on the Earth's surface. Some of these processes leave records that persist for centuries. For example, tree growth is quite sensitive to temperature. Furthermore, trees display clear annual growth rings. By studying these you can get some idea of the climatic conditions each year of the tree's life. Ancient trees contain a record of past climate, if you know how to read it! Another example comes from ice-caps. The deuterium content of snow turns out to depend on the temperature of the air when the snow fell. Hence the deuterium content of ice depends on the air temperature at the time at which the ice was deposited. So, deeply buried ancient ice contains a record of the atmospheric temperature in the past, which can be extracted by taking ice cores. Elsewhere, records of ice-melt give an indication of temperature. Coral growth is another example. Records of these "climate- proxies" are available from quite a large number of sites in the northern hemisphere.

All these proxies can be put together into a statistical model for predicting atmospheric temperature. Estimation and checking of the model is done using the period of time for which we have direct temperature measurements. Assuming that the *relationship* between the proxies and the temperature has not changed over time, the proxies can be fed into the model to estimate what the temperature would have been a long way back in to the past. A grossly simplified version of what is done might work like this:

Suppose we have tree ring width data w_i and ice core deuterium data d_i stretching back to 1400, and temperatures T_i from 1900-2000. We might believe that:

$$T = \alpha + \beta w + \gamma d + E \quad \text{where} \quad E \sim N(0, \sigma^2)$$

The parameters α, β and γ could be estimated by least squares using the w_i's, d_i's and T_i's from 1900-2000. We could then feed w_i and d_i values for 1400-1900 into the equation:

$$\hat{T}_i = \hat{\alpha} + \hat{\beta} w_i + \hat{\gamma} d_i$$

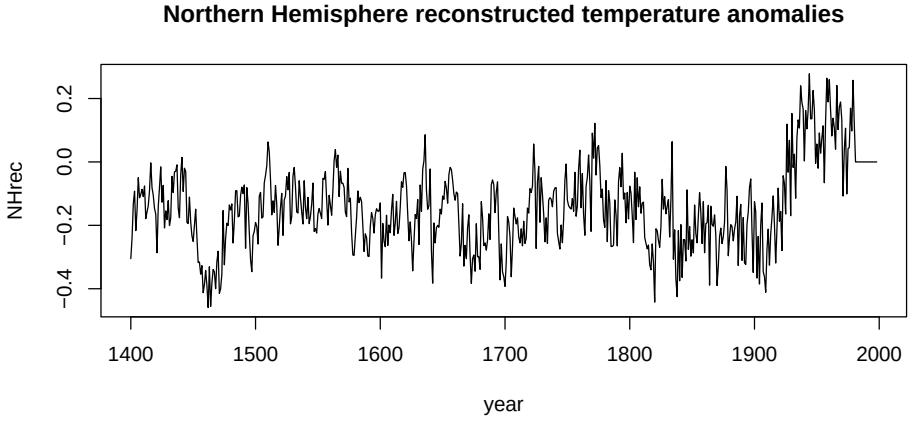
To get the estimated values of T for all years from 1400-1900. Note that I have used a $\hat{}$ to denote "estimated value of". In this case the predictions are in a sense "average predictions", since no simulated E_i terms have been added, rather the average value of the E_i's has been used, which is zero!

The real reconstruction uses substantially more complex methods and models than this, but the basic principle is exactly as described above. The dataframe `nh600` in package `mt1007` contains the annual mean reconstructed temperature anomalies for the northern hemisphere from 1400 onwards. You can find far more extensive details and further references at:

`http://www.ngdc.noaa.gov/paleo/pubs/mann1998/frames.htm`

`nh600` has the reconstructed temperature anomalies[9] for the northern hemisphere in array `NHrec`. Let's plot them against year:

```
>plot(year,NHrec,type="l",main="Northern Hemisphere reconstructed temperature anomalies")
```



Note that `NHrec` contains no reconstructions after 1980, but contains zeros instead. You should build an appropriate auto-regressive model for these data (don't go beyond lag 4) check that its acf looks ok and that it produces reasonable looking trajectories when you simulate from it. Use this model to see if the model appropriate for the climate from 1400 onwards is capable of producing the sort of excursion seen in the real data. Write yourself up a page of notes recording what you did and reporting what you find. As you do this keep the basic approach in mind: the idea is to build a model that embodies the sceptics position that the temperature record represents a process varying randomly around some long term average. If you can produce such a model that looks like the bulk of the data, and can reasonably often produce the sort of increase seen in the record, then the sceptics position seems plausible. If models embodying the sceptics hypothesis and looking like the data rarely or never produce such upswings then the sceptics position looks seems implausible.

Obviously this kind of simple modelling of the temperature series is only one part of the evidence relating to climate change: in addition to physical modelling results, there are also other data available and the next section looks at one such dataset.

[9] temperature less the 1901-1980 mean, in this case

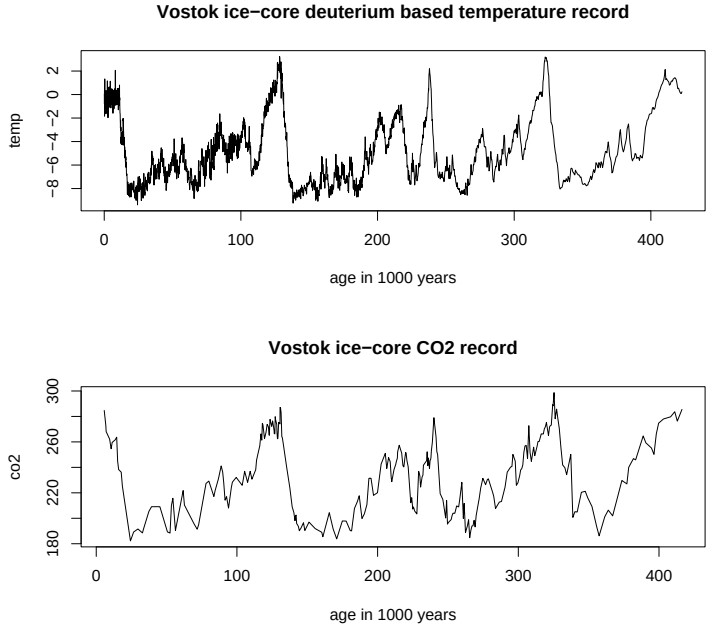
5 CO_2, temperature and the Vostok ice-core

In the previous lectures and practicals you have examined global temperature time-series, to investigate whether the recent warming is anomalous or could just be the result of "natural variability" in the climate system. To finish off this case study, we'll look at some of the evidence that is available relating directly to the link between temperature and CO_2 concentrations.

The evidence about the link between CO_2 and temperature over the long term comes from ice-core data. As has been mentioned before, large ice-sheets build up over time, with older ice occurring deeper below the surface than newer ice. Bubbles of air are trapped within the ice and these are roughly the same age as the ice (actually a bit younger as the bubbles are not sealed in until the ice is a little way below the surface). These bubbles can be analysed to find the CO_2 content of the atmosphere at the time the ice was formed. The ice also contains a record of air temperature when the snow that made up the ice fell. Deuterium occurs naturally in the atmosphere, and it turns out that the deuterium content of snow depends on air temperature. Since the snow eventually becomes the ice of the ice shelf, this means that the deuterium content of the ice gives the temperature of the atmosphere when the ice was formed.

It is possible to age ice-cores drilled out of the ice shelf. It is therefore possible to reconstruct records of past CO_2 concentrations and past atmospheric temperatures from the ice-cores. At Vostok station in Antarctica an ice-core over 4km deep has been drilled out. From this it is possible to reconstruct temperatures in the Antarctic atmosphere and atmospheric CO_2 concentrations over the last 414,000 years! The data are available in package `mt1007` as `vostok.co2` and `vostok.temp`:

```
> data(vostok.temp,vostok.co2)
> attach(vostok.temp)
> attach(vostok.co2)
> par(mfrow=c(2,1))
> plot(age,temp,type="l",main="Vostok ice-core deuterium based temperature record",
+ xlab="age in 1000 years")
> plot(ice.age,co2,type="l",main="Vostok ice-core CO2 record",xlab="age in 1000 years")
```



Notice that the CO_2 level has not been as high as it is now (over 360 ppm), ever in the last 414,000 years! The relationship between CO_2 and temperature appears to be fairly close from these plots. To get a clearer picture, let's try and plot the temperature data against the CO_2 data. There is a slight problem here (of the sort that often occurs with real data): the CO_2 measurements and temperature estimates

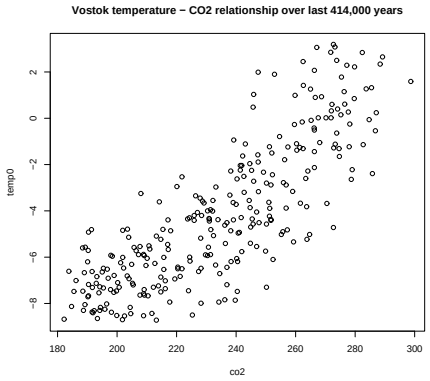
don't actually occur at exactly the same times. In fact there are far more temperature estimates than there are CO_2 measurements:

```
> length(vostok.temp$temp)
[1] 3311
> length(vostok.co2$co2)
[1] 283
```

It would therefore make sense to plot the CO_2 measurements against the temperature estimate that is closest to it in time. The following, more or less does this (it actually takes the closest younger value):

```
> a0<-age[1:3310]
> a1<-age[2:3311]
> for (i in 1:283) temp0[i]<-temp[a0<=ice.age[i]&a1>ice.age[i]]
```

This is picking out the records of `temp` such that the age of the record (stored in `a0`) is less than or equal to the `ice.age[i]` while the age of the next record (stored in `a1`) would be greater than `ice.age[i]`. This approach ensures that we get a set of temperature records that correspond closely in time to the CO_2 records. Plotting `temp0` against `co2` gives:



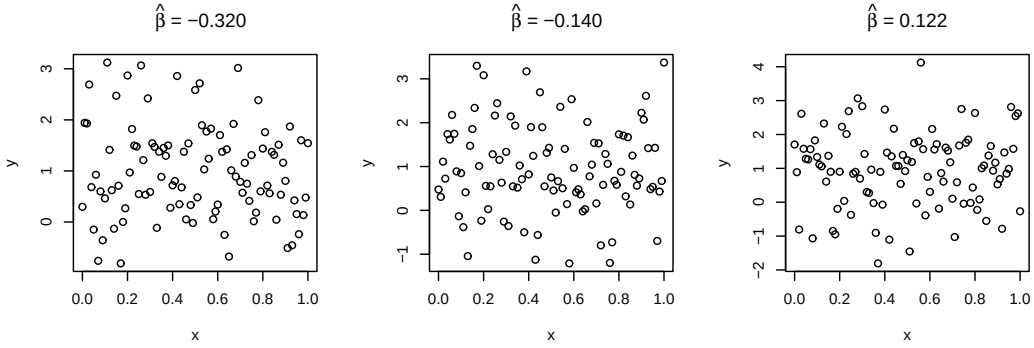
The plot shows fairly clearly that temperature and CO_2 seem to be positively correlated. A reasonable model for this would be:

$$T_i = \alpha + \beta C_i + E_i \qquad E_i \sim N(0, \sigma^2)$$

Where T_i is the i^{th} temperature record and C_i the corresponding CO_2 record. α and β can be estimated by a least squares fit to the data:

```
> mod<-lm(temp0~co2)
> coef(mod)
 (Intercept)          co2
-24.75740280   0.08797142
```

Obviously $\hat{\beta} = 0.088$ is quite small, and the data are quite scattered: which raises the question of how small $\hat{\beta}$ would have to be before we decided that there was insufficient evidence to conclude that there was a relationship between temperature and CO_2? This is an issue because we are estimating β. Even if the data were really generated from a model where $\beta = 0$ our *estimate* of β would (almost) never be exactly zero. By way of illustration of this, the following figures show replicate x, y data generated from a model $y_i = 1 + E_i$ where $E_i \sim N(0, 1)$: i.e. from the model above with $\alpha = 1$, $\beta = 0$ and $\sigma = 1$. Above each plot is the estimated value of β for the data:



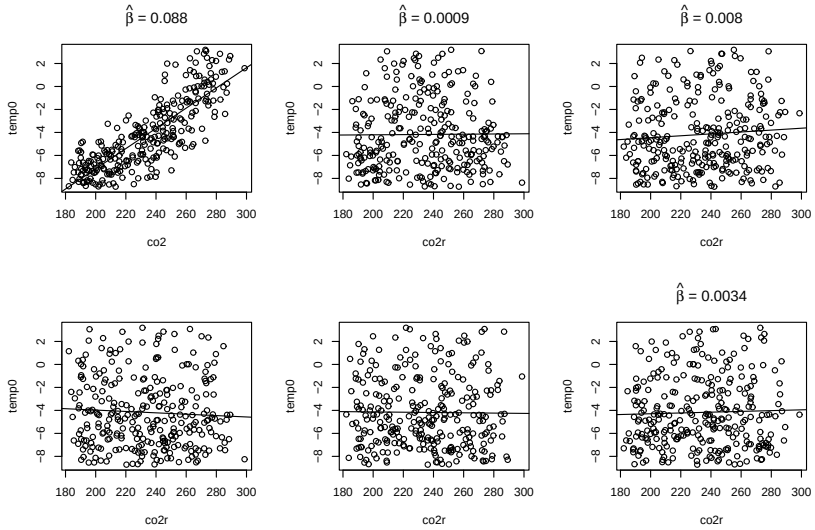
So the question is: how large does $\hat{\beta}$ have to be before we can be confident enough that β is not zero to conclude that CO_2 and temperature are definitely related? A theoretical approach to this question will be given in the theory section of the course. Now it will be addressed using a modern computer-intensive statistical method: randomization.

The idea is to simulate a great many replicate datasets, which share the characteristics of the real data *except* that we ensure that there is no real correlation between the temperature measurements and the CO_2 measurements. The easiest way to break the correlation in the original data is to randomly shuffle one of the variables, so that any relationship that was originally present is destroyed (we are effectively setting $\beta = 0$). Now, if we fit the model to each replicate simulated dataset, we will be able to generate the expected distribution of $\hat{\beta}$, *assuming that* $\beta = 0$. This immediately yields a way of evaluating whether the value of $\hat{\beta}$ obtained from the original data is consistent with β being zero. If the original $\hat{\beta}$ is improbable according to the distribution of $\hat{\beta}$'s that we expect if $\beta = 0$, then it's safe to conclude that $\beta \neq 0$.

To randomly re-shuffle the CO_2 data, we can use the `sample()` function:

```
> co2r<-sample(co2,283)
```

This tells R to pick 283 values from `co2`, at random, and put the resulting array in `co2r`. Since there are only 283 elements of `co2`, this amounts to randomly shuffling the CO_2 data. The following figure illustrates the randomization method:

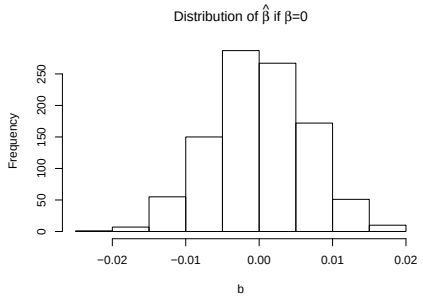


The top left panel shows the original data, the best fit straight line model and its slope, $\hat{\beta}$. The remaining 5 panels show replicate data sets where the situation of $\beta = 0$ has been simulated by randomly shuffling the CO_2 data (so that CO_2 and temperature are uncorrelated). For each replicate the best fitting straight line model and corresponding $\hat{\beta}$ are shown. From the picture it looks like 0.088 is well above the values

of $\hat{\beta}$ that would be consistent with $\beta = 0$. To check this more thoroughly R can be programmed to run a really large number of replicates, say 1000:

```
> b<-1   # need to initialize b to something for next line to work
> for (i in 1:1000) {co2r<-sample(co2,283);b[i]<-coef(lm(temp0~co2r))[2]}
```

What this does is to re-shuffle `co2`, and then fit a straight line model to the re-shuffled CO_2 data and the temperature data - $\hat{\beta}$ is extracted from this fitted model and stored in an element of `b`. When the loop finishes 1000 replicate $\hat{\beta}$ values are available, all generated under the assumption that $\beta = 0$. A histogram of `b` gives:



Clearly the original value $\hat{\beta} = 0.088$ is way outside this range. Hence, we can be **very** confident that the relationship apparent in the Vostok ice core data is real: the data are in no way consistent with an underlying model in which $\beta = 0$. Temperature is definitely correlated with atmospheric CO_2 concentration.