1 Introduction: An epidemiological problem

The following example is not completely realistic (and also a bit morbid), but it is a simple example of a dynamic linear model (DLM) in action.

Suppose we are part of a team of epidemiologists, who have been dispatched to an outbreak of a virulent disease in some large and inaccessible region. It's day 14 of the outbreak, and the doctor on the scene has compiled data on the number of people dying on each of the last nine days:

$$\text{Daily number dying} = \square, \square, \square, \square, \square, 3, 4, 4, 4, 5, 5, 5, 6, 5$$

(as is common in this situation, some initial data are not available). These data are shown as filled circles in Figure 1.1.

The doctor would like to know: (1) The number of people likely to die on each of the next seven days, and the total number of people likely to die next week; (2) The initial progress of the disease, corresponding to the period for which data are not available.

Formulate a simple model. We start by encoding the doctor's beliefs in a simple model. Suppose this disease passes through three stages, each stage taking one day. On the first day a person is infectious but shows no symptoms. On the second day a person is infectious and shows symptoms, and on the third day the person either recovers, or dies. The doctor believes (on average): (a) each stage I person infects 1.05 others, and 0.05 of stage I people die without passing through stage II; (b) each stage II person infects 0.1 others, and 0.2 of stage II people die.

We can collect the number of people in the different stages together into the vector θ_t:

$$\theta_t = \begin{pmatrix} \text{Number of stage I people in time } t \\ \text{Number of stage II people in time } t \\ \text{Number of people dying in time } t \end{pmatrix}.$$

Our supposition about the disease's behaviour can then be written in the form

$$E(\theta_t \mid \theta_{t-1}) = \begin{pmatrix} 1.05 & 0.1 & 0 \\ 0.95 & 0 & 0 \\ 0.05 & 0.2 & 0 \end{pmatrix} \theta_{t-1}.$$

If we denote by y_t the number of people dying in time t then our data comprise

$$E(y_t \mid \theta_t) = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \theta_t.$$

Quantify the uncertainties. Our models for $E(y_t \mid \theta_t)$ and for $E(\theta_t \mid \theta_{t-1})$ are only right 'on average'. Therefore we need to elicit from the doctor measures of how much the actual outcomes can vary around these averages. The doctor believes that the number of deaths observed is exactly right. The number going into stage I given θ_{t-1} is accurate to ± 4, the number going into stage II is accurate to ± 2, and the number dying is accurate to ± 1. From these quantities we can write

$$Var(y_t \mid \theta_t) = 0 \quad \text{and} \quad Var(\theta_t \mid \theta_{t-1}) = \begin{pmatrix} 2^2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.5^2 \end{pmatrix}$$

taking each range to represent ± 2 standard deviations.

We also need a starting point, representing beliefs about θ_0. The doctor is pretty sure that the disease started on the arrival of stage I people from another area, but she is quite uncertain about the number that arrived: she guesses about 5. On this basis we might choose

$$E(\theta_0) = \begin{pmatrix} 5 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad Var(\theta_0) = \begin{pmatrix} 2.5^2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

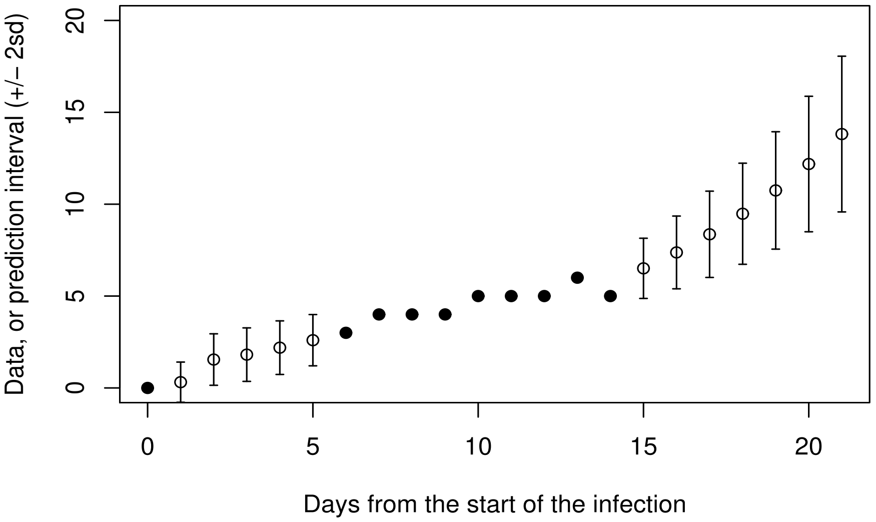

Figure 1.1: Number of deaths, actual or predicted

Forecasting. Using these quantities we can use the framework of a DLM to make predictions about mortality in the coming week. The precise calculations will be revealed in the next few lectures.

We denote the data collectively as D_{14}, being the number of deaths up to and including day 14 (including noting where the data are not available). Our prediction for deaths on day 15 comprises the mean and variance of $\theta_{15,3}$ (i.e. the third component of θ_{15}) conditional upon the data D_{14}:

$$E(\theta_{15,3} \mid D_{14}) = 6.5 \quad \text{and} \quad \mathit{Var}(\theta_{15,3} \mid D_{14}) = 0.7\,.$$

We can similarly predict the number of deaths at any point in the future, providing, of course, that we are satisfied that our original model holds. For example,

$$E(\theta_{21,3} \mid D_{14}) = 13.8 \quad \text{and} \quad \mathit{Var}(\theta_{21,3} \mid D_{14}) = 4.5\,.$$

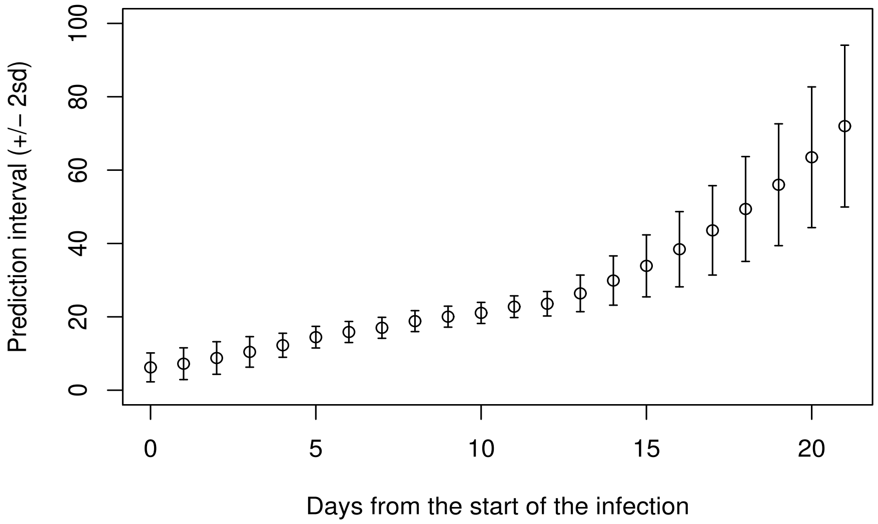

Figure 1.2: Number of stage I infectives, predicted

We can also predict linear combinations of components of θ. For example, the total number of deaths over the coming week is the sum $S := \theta_{15,3} + \cdots + \theta_{21,3}$:

$$E(S \mid D_{14}) = 68.5 \quad \text{and} \quad \mathit{Var}(S \mid D_{14}) = 8.8\,.$$

Filtering of 'backcasting'. The DLM framework also allows us to go backwards. We can use the current data D_{14} to make predictions about the number of deaths in the days before we collected any data. Our prediction for the number of deaths on day 5 (one day before we started collecting data) is

$$E(\theta_{5,3} \mid D_{14}) = 2.6 \quad \text{and} \quad \mathit{Var}(\theta_{5,3} \mid D_{14}) = 0.5\,.$$

Likewise, our prediction for the total number of deaths over the five missing days, $S' := \theta_{1,3} + \cdots + \theta_{5,3}$ is

$$E(S' \mid D_{14}) = 8.5 \quad \text{and} \quad Var(S' \mid D_{14}) = 2.1 \,.$$

Figure 1.1 shows the number of deaths per day, as data where it is available, and as a prediction based upon D_{14} where it is not, the latter with ± 2 standard deviation error bars.

If the doctor is interested in the number of stage I infectives, we can compute that too. Figure 1.2 shows the DLM prediction for the three weeks based upon the data D_{14}. Likewise we could compute the number of stage II infectives, or, in general, any linear combination of the components of θ_t for any t.

Exercises

1. Our DLM has several weakness in this application (some of which we will put right in later lectures). What do you think the main ones are?