1

Introduction

Mike Christie[1], Andrew Cliffe[2], Philip Dawid[3] and Stephen Senn[4]

[1]*Institute of Petroleum Engineering, Heriot Watt University, Edinburgh, UK*
[2]*School of Mathematical Sciences, University of Nottingham, UK*
[3]*Centre for Mathematical Sciences, University of Cambridge, UK*
[4]*School of Mathematics and Statistics, University of Glasgow, UK*

In this introductory chapter we make some brief remarks about this book, what its purpose is, how it relates to the Simplicity Complexity and Modelling (SCAM) project and also more widely about what the purpose of modelling is and what various traditions in modelling there are.

1.1 The origins of the SCAM project

In January 2006 the Engineering and Physical Research Council (EPSRC) organized a 'sandpit' or 'ideas factory' at Shrigley Park under the directorship of Peter Grindrod with the title 'Scientific Uncertainty and Decision Making for Regulatory and Risk Assessment Purposes' in which scientists from a wide variety of disciplines participated. At the ideas factory there were frequent informal and formal meetings to discuss issues relevant to uncertainty in modelling. As the week progressed various themes emerged, projects were mooted and teams coalesced. These teams then competed with each other for funding from the EPSRC. Among those that were successful was a project which had the following specific objectives:

- First, given that data are finite, what is the appropriate balance between simplicity and complexity required in modelling complex data?

Simplicity, Complexity and Modelling, First Edition. Edited by Mike Christie, Andrew Cliffe, Philip Dawid and Stephen Senn.

- Second, where more than one plausible candidate model is used, how should forecasts be combined?
- Third, where model uncertainty exists, how should this uncertainty be propagated into predictions?

However, the project also had the more general and wider purposes of making modellers in different traditions mutually aware of what they were doing and also of making the different terminology that they employed intelligible to each other.

Funding for the project was agreed and the name *Simplicity, Complexity and Modelling* (SCAM) was chosen. This is the book of the SCAM project.

1.2 The scope of modelling in the modern world

Scientists working in many diverse areas are engaged in modelling the world. Obviously, the various fields in which the models they create are applied vary considerably and this is reflected in the approaches they adopt to build, fit, test and use the models they devise. Consider, for example, credit scoring and climate modelling. In the former case the data consist of billions of transactions every day. The field is data-rich and the opportunities to test the ability of the fitted models to predict (say) good and bad debts abundant. A model that is fitted today can be tested tomorrow and again the day after and so on. On the other hand, climate modellers are trying to predict a unique future. If current trends in human activity persist, will this lead to global warming and what will be the consequences? If the models suggest that the consequences of current activity are serious and if mankind acts on the warning and mends its ways then the prediction will never be validated. Climate modellers are thus cast in the role of Cassandras: if heeded they will ultimately be doubted because what they predict will not come to pass and only disaster will reveal them to have spoken the truth. This may seem somewhat fanciful, yet consider the case of the so-called *millennium bug*. Huge sums of money were invested in fixing computer code. The world computing network survived the arrival of the year 2000, and now some are convinced that it was all a fuss about nothing while others believe that it was only foresight and action that prevented disaster.

Yet, if one looks a little deeper even in these very different fields there are points in common. For example, in the wake of the global financial crisis of 2008 many financial analysts are no doubt pondering how well the current approach to forecasting the credit weather will serve if the credit climate is changing.

Nevertheless, some things are very different as one moves from one field to another, and it is the belief that knowledge of such differences is valuable that is one of the justifications for this book. On the other hand, some things that appear different are in fact the same or similar, and it is the vocabulary that differs from field to field and sometimes within a field, rather than the concept. For example, the terms *random effects model*, *hierarchical model* and *mixed model* used within the discipline of statistics are either synonyms or so readily interchangeable that they might be applied, depending on author, to exactly the same algebraic construct. However, those who work in pharmacometrics use machinery that is identical to random effects models but are likely to refer to such as population models (Sheiner et al. 1977). This reflects, of course, the fact that even within the same discipline different individuals responding to different perceived needs have stumbled across the same solution, and that as one switches discipline the scope for this phenomenon is even greater.

It is the object of this book and of the SCAM project, to represent various modelling traditions and application areas with a view to making researchers aware of a rich diversity but also that there are many concerns they share in common.

1.3 The different professions and traditions engaged in modelling

However, it would be foolish of us to claim that the team members cover all disciplines and hence that our book encompasses the whole field. We are, in fact, three statisticians (APD, JO and SS), an applied mathematician (AC), a climate modeller (PC), a geographer (SD) and three engineers (MC, ZK and JH). Not included in the team, for example, are any computer scientists. Also absent, to name but a few scientific professions, are any econometricians, financial analysts or pharmacometricians (although SS has some interests in the latter field). The bias towards the physical sciences in the team is thus clear. In fact the application areas covered by us include topics from the physical sciences such as climate, oil exploration, flood prevention, nuclear waste disposal, water distribution networks, and simpler approximations of complex computer programs. The modelling of treatment effects in drug development is perhaps the only exception to this theme.

We do not claim that the breadth of the book is great enough to cover all fields or even all lessons that might be learned from study of such fields, but hope that it is great enough to be interesting and valuable and that it will serve to make the strange familiar by drawing parallels where they can be found and to make the familiar strange by alerting modellers in a given field to the fact that others do not necessarily do things the same way and hence that what they take for granted may be far from obvious.

1.4 Different types of models

Cox (1990) identifies two major types of model: substantive and empirical. Models of the former type arise as a result of careful consideration of some well-established or at least plausible background scientific theory. Careful thought concerning processes involved suggests a relationship between quantities of interest. The theory thus embodied may suggest some difficult or intricate mathematical work, and this receives expression in a model. We give a simple example of the thinking that might go into such a model from the field of pharmacokinetics.

Various physiological considerations may suggest that a particular pharmaceutical given by injection will be eliminated at a rate that is proportional to its concentration in the blood. Suppose we have an experiment in which a healthy volunteer is given a pharmaceutical by intravenous injection and then blood samples are drawn at regular and frequent intervals. A differential equation suggests that the concentration–time relationship can then be modelled with concentration on the log scale as a linear function of time. Of course nothing is measured perfectly, so that some random variation should be allowed for. It may thus be valuable to think in terms of data which have a signal plus some noise. The signal part of the model can then be modelled as

$$\mu_t = \mu_0 e^{-kt}, \tag{1.1}$$

where μ_t is the 'true' concentration at time t after dosing, μ_0 is the concentration in the blood at time 0 and k is a so-called *elimination constant*. One could regard such a

model as being a simple (incomplete) example of a substantive model. Making it realistic using purely theory-based considerations may be difficult, however. A log transformation is particularly appealing and we can then write

$$\log(\mu_t) = \log(\mu_0) - kt. \tag{1.2}$$

(Here we follow the usual statistician's convention of writing natural logarithms as log.) We do not, however, observe μ_t directly but (say) a quantity Y_t. The model given in (1.1) may then be extended to represent observable quantities by proposing some simple relationship between a given observed concentration Y_i taken at time t_i and the true unobserved concentration μ_{t_i} that involves an unobserved random variable ϵ_i. One possible relationship is

$$\log(Y_i) = \log(\mu_{t_i}) + \epsilon_i = \log(\mu_0) - kt_i + \epsilon_i. \tag{1.3}$$

However, this model is itself not complete until we specify how the ϵ_i are distributed. If we can assume that they are identically, independently distributed with unknown variance σ^2 which does not vary with time (and hence with concentration) then a rather good way to estimate the unknown parameter seems to be via ordinary least squares on the log concentration scale.

So far, some limited subject-matter theory (to do with plausible models for drug elimination) has been used for developing the model for the signal. The model for the noise, however, is rather 'off the peg' but it can be refined by further considerations. For instance, the theory of ordinary least squares tells us that where such a model applies and n blood samples have been taken, the variance of the estimate k, $\hat{k}$, is given by

$$\text{var}(\hat{k}) = \frac{\sigma^2}{\sum_{t=1}^{n}(t - \bar{t})^2}. \tag{1.4}$$

This raises the question, given that a fixed number of samples should be taken, when should we choose to take them. If formula (1.4) is correct the answer is half at baseline and half at infinity, since this is the arrangement that maximizes the denominator in (1.4) for given n and hence minimizes (1.4) for given n and σ^2. This is, however, absurd and its absurdity can be traced to two inappropriate assumptions in the error model: first, that on the log scale the error variance is constant; and second, that the error terms are independent. Recognizing that the variance (on this log scale) is likely to increase with time makes it less reasonable to measure at high values of t. Allowing the ϵ_i to have a correlation that decays with time will indicate that, other things being equal, measurements taken more closely together provide less information.

Many models employed, however, are not the result of these sorts of consideration. These are models of the type Cox calls empirical. For example, in a clinical trial in adults suffering from asthma (Senn 1993) we may be measuring forced expiratory volume in one second (FEV_1). We will of course have treatment given as an explanatory factor in the model. However, we know that, other things being equal, women have lower FEV_1 than men and older adults have lower FEV_1 than younger ones. As a first attempt at a model we might include a dummy variable for sex, taking on the value 0 for females and 1 for men, say. We could have a simple linear term for age but might consider also adding age squared and age cubed. Or perhaps we could use some other polynomial scheme such as that of so-called *fractional polynomials* (Royston and Altman 1994; Royston and Sauerbrei 2004). The general point here, however, is that the model we use is governed

much more by what has been observed to work in the past and some general modelling habits we have, rather than by some considerations based on the physiology of the lung and (say) some biological model of how it deteriorates with age.

The choice of a suitable model may depend on context as well as purpose. Does one need to make predictions under conditions that are physically different from ones in which any of the observations have been made? To take an example from flood modelling, one may wish to predict how high the flood waters will be after construction of a dam. If one was just interested in predicting water levels next week, by which time the dam would *not* have been constructed, one could use a Kalman filter or a machine learning algorithm or some such, preferably rather parsimonious, empirical model. But if one wants to predict in changed circumstances one may have to go to the trouble of setting up a hydraulic model, estimating roughness parameters, and then changing the geometry to represent the future and unobserved conditions.

Of course, the distinction between these two types of model is not absolute. For instance, to return to pharmacokinetics, a modern approach builds up models of drug elimination from more fundamental models of various organ classes of the human body – liver, gut, skin, blood and so on – as well as biochemical models of the pharmaceutical (Krippendorff et al. 2009) to predict what sort of model of serum concentration in the blood will be adequate. From the perspective of this approach, adopting a model such as (1.1) directly without such background modelling is rather empirical.

One can also give examples tending in the other direction. A common approach to comparing generic formulations of a pharmaceutical to the innovator product for the purpose of obtaining a licence is to use a so-called bioequivalence study (Patterson and Jones 2006; Senn 2001). This compares the concentration–time profile in the blood of both formulations given on different occasions (the sequence being random) to healthy volunteers. Commonly these curves are compared using summary statistics such as area under the curve (AUC) and concentration maximum (Cmax) and a model is built relating AUC (say) to formulation, subject and period. From the perspective of someone who builds a model like (1.1) this is also very *ad hoc* and empirical. However, theoretical considerations can be produced based on a model like (1.1) to show that AUC is in fact a good measure to use to compare two concentration–time profiles.

The various examples of modelling in this book cover this spectrum pretty widely. Examples will be found of empirical modelling but also of complex models that are built up from more fundamental scientific considerations.

1.5 Different purposes for modelling

Different sciences have developed their own modelling traditions and approaches. Some use entirely deterministic models, others allow for uncertainty and random variation. Some attempt to model finely detailed structure, others a coarser ‘big picture’. The ‘fitness for purpose’ of a model will depend on many considerations. One important aspect is complexity: while incorporating more detail may allow a more accurate description, an over-complex model will be hard to identify from observations, and this can lead to poor predictions. Note, however, that a poorly identified model is not necessarily bad at prediction. For example, the parameter estimates may have high standard errors but be strongly negatively correlated. The variance of a prediction may then include a contribution not only from large variances of individual parameters but also from important negative covariance terms. For example, to return to the case of a clinical trial in asthma,

any model that includes height, sex, age and baseline FEV_1 in the model may find that the estimates have large standard errors since height, sex and age are all strongly predictive of FEV_1. The problem is, however, that the collinearity makes it difficult to establish the separate contribution of each precisely. However, for a prediction for any given patient it is the joint effect of them all that is needed, and this may be measured quite well.

Nevertheless, it is important to strike the right balance between too much simplicity (which may miss important patterns in the world and signals in the data) and too much complexity (which may lose the signal in a halo of noise). A variety of methods has been developed to tackle this subtle but vital issue.

However, whatever the science, two purposes of models are commonly encountered. One is to increase understanding of a particular field. In the field of statistics this is very much associated with causal analysis (Pearl 2000). In the hard sciences it is to use models as a means of establishing and understanding 'laws'. A further purpose, however, is for prediction. In the hard sciences the analogy would be to work out the consequences of the laws established.

1.6 The purpose of the book

The primary purpose of this book is to make it easier for modellers in different disciplines to interact and understand each other's concerns and approaches. This is largely achieved, we hope, through the subject-specific contributions (Chapters 3–10) which provide an introduction to modelling in various fields. We hope that the reader will emerge from perusing these chapters with the same sense of surprise that we experienced through our interactions with each other throughout the course of the project, namely that there is much more to modelling than we originally thought.

What the book is *not* is a basic introduction to linear models, generalized linear models or statistical modelling generally. For the reader who is in search of such, excellent texts that fulfil this purpose that we can recommend are the classics on linear models by Draper et al. (1998) and Seber and Lee (1977), that on generalized linear models by McCullagh and Nelder (1999) and three more general texts on statistical modelling, with very different but valuable perspectives, by Harrell (2001), Davison (2003) and Freedman (2005). For a Bayesian approach we recommend Gelman et al. (2004).

Nevertheless, a brief technical introduction to modelling is provided in Chapter 2, and in Chapter 11 we try and draw some threads together. We also provide a glossary, which we hope will help modellers to understand each other's vocabulary.

1.7 Overview of the chapters

The book contains ten further chapters after this one, two of which are general in scope and eight of which cover specific application areas reflecting the interests of the members of the team.

Chapter 2, by Philip Dawid and Stephen Senn, is a general purpose methodological one on model selection but also including some remarks on a matter that goes to the heart of the SCAM project. A model that is finally chosen may be a clear winner in that it seems to be the only model among many that adequately describes the data. On the other hand, it might simply be the best by a narrow margin among a wide set of candidate models. It would seem plausible that in the first case the true uncertainty in prediction is better captured by a within-model analysis than in the second. In the second case some

consideration of the road or roads not taken would seem to be necessary in order to express uncertainty honestly. Yet if model selection and fitting proceeds, as it often has in practice, through a first stage of selection and then a second stage of prediction using the model selected as if one knew it were true, the true uncertainty is underestimated.

Chapter 3 is the first of the subject-matter chapters. In it Stephen Senn considers the field of drug development and, in particular, the analysis of so-called phase III trials. This is interesting not because the modelling is complex – in fact it is frequently very simple, although increasingly complex models are being used to deal, for instance, with the vexed problem of missing data (Molenberghs and Kenward 2007) – but rather because progress can often be made without complex modelling, albeit at a price.

The price is a reduction in precision. Under best conditions, randomized clinical trials yield unbiased estimates of the effect of treatments. However, including covariates in the model can often make these estimates more precise. Thus, simplicity has a price in the form of the need for larger sample sizes. On the other hand, it seems to be a psychological fact that simpler models (rightly or wrongly) are often trusted more than complex ones. Thus the reduction in statistical uncertainty is bought at some apparent increase in epistemic certainty.

In Chapter 4 Jeremy Oakley considers statistical issues in the use of deterministic substantive models. Such models are often described as 'computer models', in the sense that they are implemented using computers. These models may be of such complexity that to run them for all the combinations of interest of the various parameter values would be far too costly in terms of time to be practical. A standard technique for dealing with this problem is to construct an emulator: a statistical model of the computer model, which can be used as a fast surrogate. The emulator is then a simplified fitted model that, it is hoped, will yield very similar outputs to the more complex one given the same inputs. Of course, there will inevitably be some loss in the quality of the output at a given parameter combination. On the other hand, it becomes much more feasible to study many combinations. Thus there is clearly a trade-off in moving from sparsely run complex models to abundantly run simple models. Unlike Chapter 3, where Senn discusses modelling in the frequentist framework, which predominates in drug regulation, the statistical framework in which Oakley operates is Bayesian and this is extensively illustrated in this chapter.

The emulator is perhaps an example of the ultimate black-box model. The required inputs are known, a set of input and outputs are available, and it is simply required to produce outputs reliably in future without too much concern about how this is achieved. This field also has an intriguing potential that most modelling fields do not. Technological developments and the operation of Moore's law may mean that the performance of an emulator may in the fullness of time be given a perfect assessment.

A very different situation occurs in the field covered by Chapters 5 and 6. Here the theme is climate modelling. Extremely complex models, based on physical theory, are created to predict a unique future that mankind may take action to avoid. SCAM team member Peter Challenor, together with co-author Robert Tomakin, considers not only the various physical anthropogenic processes that may lead to global warming but also the various types of uncertainty attendant on any modelling of this process. As a particular example of a problem in climate modelling they take the possible collapse of the North Atlantic thermohaline circulation.

In a further chapter, SCAM team member Suraje Dessai and co-author Jeroen van der Sluijs examine the modelling of climate change impacts. This chapter illustrates the numerous professions and traditions involved in modelling, as climate change impacts

are the result of linking a variety of different models. Such models include integrated energy–economy–environment assessment models, global and regional climate models, and hydrological models. The chapter shows that computational constraints, pragmatism and scientific traditions have led to multiple routes of uncertainty assessment in this field.

The next three chapters are examples of modelling in engineering dealing with rather different aspects of fluids and their large-scale management, all with extremely important implications for human activity. Zoran Kapelan looks at modelling of water distribution systems, Jim Hall looks at flood prediction, prevention and management, and Mike Christie at oil reservoir modelling. In all these applications well-known physical laws are included as part of the model building. Again there is a difference in statistical frameworks. Christie's approach is Bayesian, and this is perhaps particularly suited to a situation in which many of the factors one would like to know about are hidden and must be estimated but also where this uncertain knowledge must be synthesized. The situations that Kapelan and Hall face are somewhat different. Many key elements of the problem required for the model, for example the structure of coastal defences or the details of a distribution grid, are known in great detail but system complexity makes exact forecasting difficult. Even in the structural model, however, probability plays a part. For instance, a coastal defence system may have a large number of fallible components. Hence, probability of failure of various components becomes a key element of any model.

Chapter 10, by Andrew Cliffe, considers modelling in radioactive waste disposal. Many features of the problem are related to physics that is well understood. Nevertheless there are many aspects of the problem to which uncertainty applies, and Cliffe considers these in this chapter.

Finally, after these various subject-matter chapters have been considered, we try and bring the lessons learned together and in a final chapter offer some general advice on modelling.

References

Cox, D.R. (1990) Role of models in statistical analysis. *Statistical Science*, **5**(2), 169–174.

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

Draper, N.R., Smith, H. and Pownell, E. (1998) *Applied Regression Analysis*. New York: John Wiley & Sons, Inc.

Freedman, D.A. (2005) *Statistical Models: Theory and Practice*. Cambridge: Cambridge University Press.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004) *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Harrell, F.E. (2001) *Regression Modeling Strategies*. New York: Springer.

Krippendorff, B.F., Kuester, K., Kloft, C. and Huisinga, W. (2009) Nonlinear pharmacokinetics of therapeutic proteins resulting from receptor mediated endocytosis. *Journal of Pharmacokinetics and Pharmacodynamics*, **36**(3), 239–260.

McCullagh, P. and Nelder, J.A. (1999) *Generalized Linear Models*. Boca Raton, FL: Chapman & Hall/CRC.

Molenberghs, G. and Kenward, M.G. (2007) *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons, Ltd.

Patterson, S.D. and Jones, B. (2006) *Bioequivalence and Statistics in Clinical Pharmacology*. Boca Raton, FL: Chapman & Hall/CRC.

Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Royston, P. and Altman D.G. (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, **43**(3), 429–467.

Royston, P. and Sauerbrei, W. (2004) A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine*, **23**(16), 2509–2525.

Seber, G.A.F. and Lee, A.J. (1977) *Linear Regression Analysis*. New York: John Wiley & Sons, Inc.

Senn, S. (1993) Statistical issues in short-term trials in asthma. *Drug Information Journal*, **27**, 779–791.

Senn, S. (2001) Statistical issues in bioequivalance. *Statistics in Medicine*, **20**(17–18), 2785–2799.

Sheiner, L.B., Rosenberg, B and Marathe V.V. (1977) Estimation of population characteristics of pharmacokinetic parameters from routine clinical data. *Journal of Pharmacokinetics and Pharmacodynamics*, **5**(5), 445–479.