MTH5120 Statistical Modelling 1

Lecture Notes, Semester B 2024

Chris Sutton FIA FHEA, Senior Lecturer in Actuarial Science
School of Mathematical Sciences, Queen Mary University of London

1. Principles of Statistical Modelling

1.1 Why and how models are used

A model is an imitation of a real-world system or process. Models of many activities can be developed, for example, in economics, medicine and business. Suppose we wished to 'predict' the effect that a real-world change would have. In some cases, it might be too risky, or too expensive or too slow, to try a proposed change in the real-world even on a sample basis. Trying out the change first without the benefit of a model could have serious consequences. A model enables the possible consequences to be investigated. The effect of changing certain input parameters can be studied before a decision is made to implement the plans in the real-world.

To build a model of a system or process, a set of mathematical or logical assumptions about how it works needs to be developed. The complexity of a model is determined by the complexity of the relationships between the various model parameters. For example, in modelling the profitability of a business, consideration must be given to issues such as regulations, taxation and sales terms. Future events affecting interest rates, inflation, new business and expenses also affect these relationships.

In order to produce the model and determine suitable parameters, data is needed, and judgements need to be made as to the relevance of the observed data to the future environment. Such data may result from past observations, from current observations or from expectations of future changes.

Where observed data is considered to be suitable for producing the parameters for a chosen model, statistical methods can be used to fit the data.

Before finalising the choice of model and parameters, it is important to consider the objectives for creation and use of the model. For example, in many cases there may not be a desire to create the most accurate model, but instead to create a model that will not understate costs or other risks that may be involved.

While in reality a modelling process does not follow a rigid pattern of prescribed steps, it is helpful in introducing the topic to imagine a set of key steps. In practice, statisticians who build and use models move back and forth between these key steps continuously to improve the model.

The key steps in a modelling process can be described as follows:

i. Develop a well-defined set of objectives which need to be met by the modelling process.
ii. Plan the modelling process and how the model will be validated.
iii. Collect and analyse the necessary data for the model.

iv. Define the parameters for the model and consider appropriate parameter values.
v. Define the model initially by capturing the essence of the real-world system. Refining the level of detail in the model can come at a later stage.
vi. Involve experts on the real-world system you are trying to imitate to get feedback on the validity of the conceptual model.
vii. Write the computer program for the model.
viii. Test the reasonableness of the output from the model.
ix. Review and carefully consider the appropriateness of the model in the light of small changes in input parameters.
x. Analyse the output from the model.
xi. Ensure that any relevant professional guidance has been complied with.
xii. Communicate and document the results and the model.

1.2 Modelling the benefits and limitations

In many areas of work, one of the most important benefits of modelling is that systems with long time frames can be studied in compressed time.

Other benefits include:

• Complex systems with stochastic elements, such as the operation of a company can be studied.

• Different future policies or possible actions can be compared to see which best suits the requirements or constraints of a user.

• In a model of a complex system we can usually get control over the experimental conditions so that we can reduce the variance of the results output from the model without upsetting their mean values.

However, models are not the simple solution to all problems – they have drawbacks that must be understood when interpreting the output from a model and communicating the results.

The drawbacks include:

• Model development requires a considerable investment of time, and expertise. The financial costs of development can be quite large given the need to check the validity of the model's assumptions, the computer code, the reasonableness of results and the way in which results can be interpreted in plain language by the target audience.

• In a stochastic model, for any given set of inputs each run gives only estimates of a model's outputs. So, to study the outputs for any given set of inputs, several independent runs of the model are needed. As a rule, models are more useful for comparing the results of input variations than for optimising outputs.

• Models can look impressive when run on a computer so that there is a danger that one gets lulled into a false sense of confidence. If a model has not passed the tests of validity and verification, its impressive output is a poor substitute for its ability to imitate its corresponding real-world system.

• Models rely heavily on the data input. If the data quality is poor or lacks credibility, then the output from the model is likely to be flawed.

• It is important that the users of the model understand the model and the uses to which it can be safely put. There is a danger of using a model from which it is assumed that all results are valid without considering the appropriateness of using that model for the data input and the output expected.

• It is not possible to include all future events in a model. For example, a change in legislation could invalidate the results of a model, but may be impossible to predict when the model is constructed.

• It may be difficult to interpret some of the outputs of the model. They may only be valid in relative rather than absolute terms, as when, for example, comparing the level of risk of the outputs associated with different inputs.

1.3 Stochastic and deterministic models

If it is desired to represent reality as accurately as possible, the model needs to imitate the random nature of the variables. A stochastic model is one that recognises the random nature of the input components. A model that does not contain any random component is deterministic in nature.

In a deterministic model, the output is determined once the set of fixed inputs and the relationships between them have been defined. By contrast, in a stochastic model the output is random in nature – like the inputs, which are random variables. The output is only a snapshot or an estimate of the characteristics of the model for a given set of inputs. Several independent runs are required for each set of inputs so that statistical theory can be used to help in the study of the implications of the set of inputs.

A deterministic model is really just a special (simplified) case of a stochastic model.

Whether to use a deterministic or a stochastic model depends on whether you are interested in the results of a single 'scenario' or in the distribution of results of possible 'scenarios'. A deterministic model will give one the results of the relevant calculations for a single scenario; a stochastic model gives distributions of the relevant results for a distribution of scenarios.

1.4 Discrete and continuous states and time

The state of a model is the set of variables that describe the system at a particular point in time taking into account the goals of the study.

Discrete states are where the variables exhibit step function changes in time. For example, from a state of alive to dead, or an increase in the number of cars manufactured in a factory. Continuous states are where the variables change continuously with respect to time. For example, real time changes in values of investments.

The decision to use a discrete or continuous state model for a particular system is driven by the objectives of the study, rather than whether or not the system itself is of a discrete or continuous nature.

A model may also consider time in a discrete or a continuous way.

1.5 Suitability of a model

In assessing the suitability of a model for a particular exercise it is important to consider the following:

- The objectives of the modelling exercise.
- The validity of the model for the purpose to which it is to be put.
- The validity of the data to be used.
- The validity of the assumptions.
- The possible errors associated with the model or parameters used not being a perfect representation of the real-world situation being modelled.
- The impact of correlations between the random variables that 'drive' the model.
- The extent of correlations between the various results produced from the model.
- The current relevance of models written and used in the past.
- The credibility of the data input.
- The credibility of the results output.
- The dangers of spurious accuracy.
- The ease with which the model and its results can be communicated.
- Regulatory requirements.

1.6 Short-term and long-term properties of a model

The stability of the relationships incorporated in the model may not be realistic in the longer term. For example, exponential growth can appear linear if surveyed over a short period of time. If changes can be predicted, they can be incorporated in the model, but often it must be accepted that longer term models are suspect.

Models are by definition, simplified versions of the real-world. They may, therefore, ignore 'higher order' relationships which are of little importance in the short term, but which may accumulate in the longer term.

1.7 Analysing the output of a model

Statistical sampling techniques are needed to analyse the output of a model, as a simulation is just a computer-aided statistical sampling project. The statistician must exercise great care and judgement at this stage of the modelling process as the observations in the process are correlated with each other and the distributions of the successive observations change over time. Therefore we need to be particularly careful before making any assumptions that rely on independence or identical distributions.

1.8 Sensitivity testing

Where possible, it is important to test the reasonableness of the output from the model against the real-world. To do this, an examination of the sensitivity of the outputs to small changes in the inputs or their statistical distributions should be carried out. The appropriateness of the model should then be reviewed, particularly if small changes in inputs or their statistical distributions give rise to large changes in the outputs. In this way, the key inputs and relationships to which particular attention should be given in designing and using the model can be determined.

1.9 Communication of the results

The final step in the modelling process is the communication and documentation of the results and the model itself to others. The communication must be such that it takes account of the knowledge of the target audience and their viewpoint. A key issue here is to make sure that the audience accepts the model as being valid and a useful tool in decision making. It is important to ensure that any limitations on the use and validity of the model are fully appreciated.

2. The Simple Linear Regression Model

2.1. The Model

Let us begin with a simple situation where we have

- one response variable, *Y*
- one explanatory variable, *X*

In many situations,

- *X* can be controlled and is known
- *Y* is unknown but can be observed
- we have *n* pairs of observations { (x_1,y_1), (x_2,y_2), …, (x_n,y_n) }

We would like to use these observations to estimate (or predict) the mean value of *Y* for some given values of *X*. A good place to start exploring the relationship between *X* and *Y* is often to plot them using the *n* pairs of observations. This plot might begin to show the nature of the relationship between *X* and *Y.*

It is good practice to seek the simplest model that describes the relationship well. This idea is called the *principle of parsimony*. At this stage we have not defined what "describes well" means and we will return to this issue a number of times through the module. A *linear* relationship (which would be indicated by something close to a straight-line plot) is the obvious place to start when looking for a simple model.

Given observation data (x_i, y_i) for $i = 1, 2, \ldots n$ we can fit a straight line to describe the response variable *Y* in terms of the explanatory variable *X* where

$$Y = \beta_0 + \beta_1 X$$

where,

- β_0 denotes the intercept
- β_1 is the slope of the line

However this is a *deterministic* model, meaning it does not allow for any randomness. As such it is unlikely that this model properly describes the data which will usually include some random elements.

We introduce randomness by having a *probabilistic* or *stochastic* element to the model where the model for *Y* has two parts:

- the value for Y we expect to observe for a given value of *X*
- an additional uncontrolled random value

This model can be written either as

$$Y_i = E[Y_i|X = x_i] + \varepsilon_i$$

or as

$$Y_i = \beta_0 + \beta_1 X + \varepsilon_i$$

for i = 1, 2, … n

Here ε_i is the *random error.*

It is usual to make the following three standard assumptions about the random error:

(1) $E[\varepsilon_i] = 0$ for all i

(2) $var[\varepsilon_i] = \sigma^2$ for all i

(3) $cov[\varepsilon_i, \varepsilon_j] = 0$ for all $i \neq j$

Because ε_i is a random variable, Y_i is also a random variable. We can re-write the three assumptions above in terms of $Y_i|X = x_i$ rather than ε_i

(1) $E[\,Y_i|X = x_i] = \mu_i = \beta_0 + \beta_1 x_i$ for all i

(2) $var[\,Y_i|X = x_i] = \sigma^2$ for all i

(3) $cov[Y_i|X = x_i,\ Y_j|X = x_j] = 0$ for all $i \neq j$

Putting these three assumptions into words we might say that

(1) the dependence of Y on X is linear

(2) the variance of Y at each value of X is constant and does not depend on x_i

(3) Y_i and Y_j are uncorrelated

Rather than keep writing $Y_i|X = x_i$ we often use $y_i = (Y_i|X = x_i)$ and then the simple linear model can be written as

$$E[y_i] = \beta_0 + \beta_1 x_i$$

and

$$var[y_i] = \sigma^2$$

It is often convenient to make a further assumption, that the conditional distribution of Y_i is Normal. This is the *Normal Simple Linear Regression Model* which can be written in one of three (equivalent) ways:

(A) $y_i \sim N(\mu_i, \sigma^2)$ where $\mu_i = \beta_0 + \beta_1 x_i$

(B) $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

(C) $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where the ε_i are iid $\varepsilon_i \sim N(0, \sigma^2)$

It can be convenient to redefine the parameters of the simple linear model into a *centred* form. This expresses the response variable y_i in terms of both the explanatory variable x_i and the mean level of that explanatory variable $\overline{x}$ where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

if we set

$\alpha = \beta_0 + \beta_1 \overline{x}$ and $\beta = \beta_1$

then the centred form of the model is

$$y_i = \alpha + \beta(x_i - \overline{x}) + \varepsilon_i$$

This model is mathematically identical to the previous (non-centred) form of the model but with new parameters (α and β instead of β_0, β_1) which allows a new interpretation:

- the slope β is the same as that in the previous model β_1
- the new intercept α is the mean response at the mean level of the explanatory variable

2.2. Least Squares Estimation

The model parameters (β_0, β_1 in the simple linear regression model above) are unknown. With a data set we can *estimate* these parameters – that is find values for the parameters that best explain the data we have observed. There are various ways in which parameters can be estimated. Here we consider *least squares* estimation. In later statistics modules you will see other methods e.g. *maximum likelihood estimation*.

The least squares estimators of the model parameters β_0 and β_1 are the parameter values that minimise the sum of the squares of the errors $S(\beta_0, \beta_1)$.

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

To find the minimum we need to differentiate $S(\beta_0, \beta_1)$ with respect to both β_0 and β_1 and set each differential to zero, then solve the two simultaneous equations in β_0 and β_1. The values of β_0 and β_1 that satisfy these simultaneous equations are $\widehat{\beta_0}$ and $\widehat{\beta_1}$.

$$\frac{dS}{d\beta_0} = -2\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] = 0 \qquad (A)$$

and

$$\frac{dS}{d\beta_1} = -2\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]x_i = 0 \qquad (B)$$

if we divide by -2 and separate the items in the brackets in (A) and (B) above we get

$$n\widehat{\beta_0} + \widehat{\beta_1}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad (C)$$

and

$$\widehat{\beta_0}\sum_{i=1}^{n} x_i + \widehat{\beta_1}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad (D)$$

where (C) and (D) are sometimes called the *normal equations*.

If we divide (C) by *n* we have

$$\widehat{\beta_0} = \frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{\beta_1}\frac{1}{n}\sum_{i=1}^{n} x_i$$

or

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$$

and from (D) substituting for $\widehat{\beta_0}$ from above and rearranging gives

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} x_i\, y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

or

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

which can be written in shorthand as

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$$

Now in calculus, to check that this is indeed a minimum not a maximum for $S(\beta_0, \beta_1)$ we need to find all the second derivatives $\frac{d^2S}{d\beta_0^2}$, $\frac{d^2S}{d\beta_1^2}$, $\frac{dS}{d\beta_0\beta_1}$ and $\frac{dS}{d\beta_1\beta_0}$ to check that all are > 0.

Note that the equations for the least squares estimators $\widehat{\beta_0}$ and $\widehat{\beta_1}$ above are functions of *Y* as well as of *X.* Now *Y* is a random variable and is generally unknown. This means that $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are also random variables. Because the response variable *Y* is not known, all that we can do is calculate values for $\widehat{\beta_0}$ and $\widehat{\beta_1}$ given a particular set of observations for (x_i, y_i). These values for $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are called *least squares estimates.* The *estimator* is the algebraic form depending on the variables X_i and Y_i whilst the *estimate* is that form evaluated for a certain set of observations (x_i, y_i). If we use a different set of observations, we should expect to get different values for the estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$.