Overview of Modelling

Requirements

For this course, it is assumed that you have a good working knowledge of graphical methods for presenting data, basic probability theory, and statistical inference (particularly frequentist inference techniques). That is, if given a data set, you should know how to graphically display it, calculate summary statistics and their standard errors, and test hypotheses about (unknown, but interesting) parameters.

This course is concerned with statistical modelling of data: what we mean by a model, which type of models should be used in various situations, and how to analyse the results of applying a model to a data set. However, before discussing the various aspects of modelling, we will briefly examine experimental and observational studies since many data sets that we seek to model come from such studies.

Research Studies

Types of studies

Typically, the main objective of a study is to investigate the effect of various treatments on a response (including, perhaps, how this relationship is affected by various other factors) or, alternatively, to study the factors which are associated with the response. In order to achieve these objectives, the researcher will perform one of two types of studies, depending on what sort of questions he/she hopes to answer:

1. ***Experiment***: This is a study where the investigator has (more or less) total control over all parts of the study: for instance, he/she decides which units are selected for the study and, if applicable, which treatment group each selected unit is assigned to. Everything in this type of study will be done for a specific reason known to the investigator.

2. ***Observational Study***: The investigator may have control over the choice of units for inclusion in the study and the variables to be measured, but will often not have control over some of the essential features of the system (e.g., the treatment group to which each selected unit is assigned).

In an attempt to try to understand the differences between these two types of studies, consider the comparison of responses to two treatments, say *A* and *B*. In an experiment, we know why one individual had *A* and another had *B* (typically because randomisation, which will be discussed later, was used). In a comparable observational study, there is always the suspicion that whether an individual has treatment A or B partially depends on some unobserved variable (that may also affect the response variable of interest).

Many observational studies can be further classified as one of two types:

1. ***Prospective***: Units for study are selected based on their treatment status (and typically followed over time).
   Advantages: can look at several different kinds of responses
   Disadvantages: may require a large number of units and a long time

2. ***Retrospective***: Units for study are selected based on their response variable values, and then the researcher looks backwards through time for possible explanatory variables for the observed response.
   Advantages: can study rare responses without using as many units
   faster: background information can be collected quickly
   Disadvantages: typically, can only examine one response

From a statistical point of view, it is preferable to perform an experiment, rather than an observational study, to answer research questions because more potential sources of bias (discussed below) are eliminated in the former. However, it is often impossible to conduct the former type of study for a variety of reasons (e.g., ethical), particularly outside the scientific and clinical trials environments. For example, if you are going to look at the effects of radon gas exposure during childhood and the associated risks of developing childhood leukaemia, you obviously would not be able to perform an experiment. Instead, you might perform a prospective study in which you select a number of children and then monitor their exposure pattern to radon gas and their leukaemia status over a period of time, all the while measuring other variables that you may feel influence the likelihood of developing leukaemia. In another example, if you were interested in studying the effects of a particular fertiliser on crop yield, you would probably perform an experiment: you might grow a number of crops, giving them different varieties of fertiliser, and then compare the results after the crops have grown. In most research situations, it will be obvious whether an experiment can be used to answer to the questions of interest.

Controlling for Error

In designing a study, be it experimental or observational, two major objectives should be kept in mind: avoiding bias (systematic error) and minimizing random error. In addition, the researcher may want to design the study to be as economical as possible; this is particularly relevant in industrial studies and medical studies, where it may not be possible to sample as many units as is necessary to study every possible combination of treatment effects. However, the issues surrounding this objective will not be discussed as they are beyond the scope of the course.

When we attempt to avoid bias and minimize random error, we are essentially trying to reduce the effect of confounding variables. ***Confounding variables*** are background variables that alter the apparent associations between the main variables of interest, such as the treatment and response variables. For example, suppose we are conducting an investigation into the relationship between smoking and blood pressure. It is known that blood pressure increases with age, and, for this example, we can assume that smokers also tend to be older. Now, the increase in mean BP of a random sample of smokers over the mean BP of a random sample of non-smokers will tend to be an overestimate due to the confounding relationship between age and smoking.

There are two general approaches to dealing with the problem of confounding variables (and error): by altering the design of the study and/or by using a more complex model in the statistical analysis phase of the study. As an instance of the first approach, we might select our samples of smokers and non-smokers so that they are matched in terms of age. As an example of the second approach, we might use a statistical model that looks at differences in mean BP within various age groups.

Our efforts to establish ***causality*** (i.e., to show that the treatments do affect the response) can be hurt by confounding variables, which tend to play a much greater role when we do not attempt to control for error in our study. Further, the presence of error can hurt the ***generalisability*** of our study to the larger population of interest. Thus, it is important that we attempt to control for error in the study design or in the subsequent statistical analysis.

Types of Systematic Error

Before examining methods for controlling for the two types of error, let us first consider potential sources of bias. Simply put, ***bias*** is anything that makes the study sample unrepresentative of the target population. Different types of bias include:

1. *Sampling bias*: Can occur when the selection mechanism for study subjects is such that not all the units in the target population may be selected. For example, if a phonebook is used as the list from which to sample individuals, then unlisted individuals cannot not be selected for the study.

2. *Non-response bias*: Can occur when only those units who respond (to a mailed questionnaire, etc . . . .) can be included in the study. Typically, those people who respond have different characteristics from those who do not respond or refuse to take part in a study.

3. *Volunteer bias*: Can occur when a study is comprised of only those individuals who volunteer to be included (e.g., CNN online polls). Classically, those people who volunteer to take part in a study have significantly different characteristics from

those who do not.

4. *Group membership bias*: Can occur when only members of particular groups can be selected for study. Often, members of particular groups of people (e.g. radiation workers) tend to have different characteristics than the general population.

5. *Non-contemporaneous control bias*: If the study considers ideas or characteristics that change with time and place, then you need to ensure that the people in your study have the same changing characteristics as population of interest.

6. *Observer bias*: Observers may subconsciously (or deliberately!) try to verify their prior expectations using your study.

7. *Withdrawal bias*: Can occur when units selected for study do not remain in the study for its duration. Subjects who withdraw partway through a study may differ systematically from those who participate in the study until its conclusion.

8. *Recall bias*: Quite often, people who are knowledgeable about or have experienced the particular events being examined by the study will provide fuller and more detailed answers than those who have not.

9. *Self-selection bias*: Can occur when the treatments are not assigned but rather selected by the units under study. Often, the choice of treatment will be affected by a particular factor that also affects the response of interest.

10. *Treatment bias*: Sometimes, the *knowledge* of having a particular treatment will produce an effect on the response variable (aside from the effect produced by the actual treatment itself). This is sometimes referred to as the *placebo effect*.

Control of Systematic Error

There are two main methods of controlling for systematic error:

1. *Randomisation and Other Study Design Strategies*: Ideally, potential sources of bias are eliminated in the study design phase using techniques such as randomisation and various other strategies.

   Essentially, ***randomisation*** means that we randomly (i.e., mathematically) *select units* and *allocate treatments* to them. The randomisation of unit selection enables us to select a study sample that is (hopefully) representative of the overall target population, thereby furthering the goal of generalisability. The randomisation of treatment allocation (hopefully) results in treatment groups that are roughly the same in all characteristics except for the treatment they receive, thereby making it easier to establish a causal relationship between the treatments and the response of interest. For example, if we are studying the effect of a particular treatment against no treatment, we want to ensure that, in each half of the study, people are of a similar age group, similar gender ratio, similar social class, etc . . . . Randomisation will hopefully allow us to avoid any type of bias. When randomisation is used, the statistical models required to analyse the resulting data are typically fairly basic.

Lastly, there are other study design strategies that can be employed to help eliminate sources of bias that may enter during the study. For example, in a study with human units, one might *blind* the units to the type of treatment they are receiving, if possible, thereby eliminating treatment bias or the placebo effect. Further, one might also blind the person measuring the outcome variable so that the measurer does not know the treatment group to which the unit was assigned, thereby eliminating observer bias. Studies that employ both of these techniques are referred to as ***double-blind***.

2. *Retrospective Adjustment*: Sometimes, when it is not possible to eliminate possible sources of bias in the study's design, an attempt to do so can be made during the statistical analysis phase. It sometimes happens that, as a result of an oversight, there is a potentially important lack of balance, even after randomisation. Other times, randomisation was not possible, and the researcher was neither able to select the units used in his/her study nor to assign them to treatment groups, as is often the case in observational studies. Studies conducted under these conditions are prone to suffer from various types of bias. However, one way to deal with these problems is to measure various other characteristics of the units, and then to use these measurements in the subsequent statistical analysis to adjust for the (non-treatment) differences between units. As a result, the statistical models employed to analyse the data are more complex.

Control of Random Error

Unlike systematic error, random error cannot be eliminated by the study designer, even in a randomised and double-blind experiment. However, there are a number of techniques that might help reduce its effect on the study:

1. Use more uniform units (i.e., try to get the most standard material possible), use better measurement techniques, etc.
2. Use more units in the hope that sheer numbers will protect against random error.
3. Use a technique called ***blocking***, which basically divides up the units into blocks of units and allocates treatments randomly to units within each block (under the assumption that there is treatment balance within blocks, i.e., that each block has the same combinations of treatments). This technique then allows you to get more information by obtaining the results for each block, and using these to obtain the overall results of the experiment.
4. A key concept in the avoidance of error is that of ***matching***. The basic idea of matching is that if subjects receiving different treatments are matched for as many other factors that may influence their response as is possible, then you will be able to obtain a less biased estimate of the effect of the treatment. There are a number of

different ways that you can think about matching, but they basically result in the following two ideas: (a.) Match each subject in your study on a per subject basis (i.e., find a specific match for each of your cases – in the previous example this would involve finding a person of similar age to each case.) This can be naturally extended to the situation where instead of finding one person who matches your subject you find *k* people. (b.) Match groups of your subjects on the basis of key variables. For example, you can stratify the distributions of the confounding variables and then choose controls such that they have the same distribution. Alternatively, you can select your controls so that the means of the confounding variables are as alike as possible between the groups.

5. As in 2. in the previous section, use a more elaborate model in the statistical analysis phase, particularly a model that includes various (non-treatment) characteristics of the units in addition to their treatments and responses. This may be a desirable strategy even if the study is a randomised experiment since including covariates *can* improve the precision of the estimates of interest (the estimate of the effect of treatment on response).

Strategies 1.-5. all entail more elaborate planning in the design phase of the study, although strategies 3.-5. will also affect and complicate the model used in the statistical analysis phase.

Components of the Study

Next, we will assume that the study, whether observational or experimental, has been designed so as to eliminate potential sources of systematic error and to minimise random error. The resulting study will produce data set(s) that can typically be broken down into four basic components:

1. *Units of Study*: A unit is the smallest subdivision of material such that two different units might receive different treatments. Examples include individual patients, plots, batches of material, etc . . . .

2. *Treatments*: These are often chosen so that the research questions of interest can be formulated as a comparison of treatments. Examples include medical procedures, fertiliser combinations, etc . . . . In experiments, the treatments are assigned (typically via randomisation) to units, but in observational studies, the treatments are often self-selected by the units.

3. *Background Information:* Whether an experiment or an observational study is used to answer particular research questions, it is often useful to record certain background information for each unit. Examples include age, gender, etc . . . . Measurement of these variables and their inclusion in the subsequent statistical model can be particularly important in the case of observational studies.

4. *Observations*: For each unit, the researcher will want to observe the values of one or more particular outcomes that are thought to affected by the treatments and potentially by the background variables. Examples include crop yield, leukaemia status, blood pressure, etc . . . . The outcomes(s) might be observed at a number of different times: ***baseline observations*** may be made before the allocation of treatments, ***intermediate observations*** may be taken during the study, and, finally, ***responses*** will be measured at the end of the study.

In this course, we will assume that *n* ***units*** are included in the study. Further, we will use the term ***covariates*** to refer to variables that record background information or baseline or intermediate observations for the units. We will refer to variables that indicate the treatments (whether assigned or self-selected) received by each unit as ***design variables***. Lastly, we will refer to the variables the measure the response or responses for each unit as the ***response variables*.** Note that the response variable(s) may be ***univariate*** (i.e., there is only one response variable) or ***multivariate*** (there is more than one response variable). The latter type of response variable may arise because the same variable is measured at a number of different times, because multiple variables are measured at the same time, or both.

In some studies, the above variables will all be measured at the same time, in which case the study (and the resulting data set(s)) would be referred to as ***cross-sectional***. In other studies, the response variable(s) may be measured at a number of different times (as may some of the covariates), in which case the study would be referred to as ***longitudinal***. In longitudinal studies, the responses are multivariate.

Modelling

Having discussed the design and components of studies, which are often the source of the data sets that we are hoping to model, we now address models and the modelling process.

Overview of Models and the Modelling Process

A ***statistical model*** is used to *summarise* the relationship between two or more variables. For instance, we have previously discussed an example in which we are interested in the relationship between two variables—blood pressure and smoking status; we would like to find a statistical model that succinctly expresses the relationship between these two variables (and possibly between them and age as well).

In some models, all of the variables of interest are treated equally in the summary of their relationship. However, in this course, we will focus on statistical models that treat the variables asymmetrically. The variables in these asymmetric models can be divided into two groups:

1. ***Response or Dependent Variable(s)***: These are the variable(s) that we believe can be predicted / explained by the other variables which have been measured. That is to say, we suspect that some combination of the other recorded variables is "correlated" with the response variable(s). The response variable(s) are typically denoted by ***Y***.

2. ***Explanatory or Independent Variable(s)***: These are the variables that we believe can be used to give some form of explanation for the differences in the response variable observations. The explanatory variable(s) are typically denoted by ***X***.

Thus, the models examined in this course will summarise the way in which the explanatory variable(s) affect the response variable(s). This is a very general statement, as we have not said anything about the specific form of the relationship between the response and explanatory variables. We might use the following equation to state the relationship between the response and explanatory variables:

$$Y = f(X;\beta) + \varepsilon\,,$$

where ε denotes an error term, *f* is a function describing how *Y* changes in relation to *X*, and *β* represents the parameters of the model. The error term, ε, includes both random measurement error and systematic error, or possible aspects of *Y* that are not adequately explained by the model (i.e., by the function of the *X*s). It is usually too ambitious to model all aspects of the response variable, and thus it is common practice to analyse just how the *mean* of *Y* changes in relation to changes in *X*. Thus, a typical statistical model can be expressed as an equation that equates the mean(s) of the response variable(s) to some function of a linear combination of the explanatory variables:

$$E\big[Y\big|X = x\big] = \eta(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p) = \eta[LC(X;\beta)]\,, \qquad (1)$$

In the above model, the form of the function *η*( ) is known, as are *Y* and *X* (the latter for a particular choice of explanatory variables). However, the ***parameters*** of the model, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, are not known and must be estimated. These parameters can be used to make various statements about the mean of the response variable, in particular how it changes as each of the explanatory variables changes. The process of estimating these parameters (and calculating their errors, as well as various goodness of fit statistics) from the data will be referred to as *fitting* the model.

In the above, LC(*X; β*) = $\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ is a ***linear combination*** of the explanatory variables. Here, it should be noted that the above expression is linear in terms of the parameters (the *β*s), but not necessarily in terms of the covariates and design variables (in their original form): some of the explanatory variables may themselves be functions of the original covariates and design variables. For instance, in our blood pressure and smoking example, Age^2 or log(Age) might be used as an explanatory variable. In general, higher powers (such as squares, cubes, etc....) of original variables are often referred to as ***higher order terms***. Additionally, the model might include the product variable Age*Smoking Status; explanatory variables such as this one that incorporate more than one variable are referred to as ***interaction terms.*** In this course, we will usually consider only ***hierarchical models*** since their interpretation is typically much more straightforward. In hierarchical models, if a higher order power term is included in the model, then the lower order power terms must also be included (e.g., if Age^3 is included, then Age^2 and Age must also be included) and, if an interaction term is included, all component variables must be included (e.g., if Age*Smoking Status is included, then both Age and Smoking Status must be included).

One of the key components of the modelling process is determining which variables and which functions of those variables should be used as explanatory variables (i.e, which should be included in the LC($X; \beta$) part of the model). This is often thought of as trying to find the *best*[1] explanatory variables. Occasionally, we will also try to find the best functional form for $\eta(\,)$. These parts of the modelling process will be referred to as ***model selection***. Essentially, model selection involves trying to find a set of explanatory variables that, when used in a particular model, result in an expression that adequately summarises the relationship between the response variable(s) and those factors that affect it. Model selection is a balancing act because an adequate model should summarise the relationship enough to be useful (for explanation or prediction—see below) but not so much that the model is no longer a good representation of reality: in other words, our model must be a balance between the two goals of parsimony and goodness of fit.

The other key component of the modelling process, which we will refer to as ***model checking,*** involves investigating whether, for a particular choice of functional form and explanatory variables, the model's assumptions are satisfied for the particular data at hand. Further, we will want to investigate whether the model's summary of the relationship between the response and explanatory variables is an adequate representation of reality. Before proceeding, it is important to note that the model checking process is not actually separate from the model selection process, since we want to select explanatory variables (and possibly a form of $\eta(\,)$) such that the model's assumptions are satisfied and the model is an adequate representation of reality.

When it comes to model selection and checking, there is never a right or wrong answer, but only the best possible explanation based on the data you have observed. Further, even for a particular data set, it may be unclear which variables to use as there are a number of different ways to define *best*. It is often the case that although you have a good relationship between your explanatory and response variables, you actually have a model that is *subjectively* nonsense. You must always make sure that the models that you fit agree in some way with current subjective knowledge or prior research hypotheses.

[1] How we define *best* differs based on the particular model used and will be addressed in the following lectures.

Goals of Modelling

Exactly how we perform model selection and model checking will depend on the reasons why we are statistically modelling the data. These reasons can be divided into two loose groups: prediction and explanation.

If *prediction* is our goal, that simply means that we want to form an equation that can be used to calculate predictions of the response variable(s) based on future observations of the explanatory variables. With prediction, we are typically more interested in estimating and finding errors and confidence intervals for Y rather than for β. When prediction is our goal, model selection will typically focus less on parsimony and more on goodness of fit; this said, however, it is usually not desirable to include all possible (non-response) variables and all possible functions of them as explanatory variables since doing so may result in ***overfitting*** (this occurs when the resulting model is too specific to the data on which it was fit and, as a result, does not fit future data well). When prediction is the goal, model testing/checking will entail seeing how well the Y values predicted from the model match the true Y values for particular combinations of explanatory variables. Preferably, this model checking/testing is performed on a different data set (i.e., a ***validation*** data set) than was used to fit the model. However, if it is not possible to use a separate validation data set, then the original data set with which the model was fitted can be used to test/check the model (possibly via cross-validation procedures).

Alternatively, if *explanation* is our goal, we will be interested in determining which explanatory variables are associated with the response variable and how they are associated with it. In a scenario that we will term *confirmatory explanation*, we have some hypothesis that a particular variable has an association with the response, as is the case when we perform an experimental or observational study to test the effect of treatments on a particular response. In this scenario, we could use statistical modelling to test this hypothesis and investigate the extent and form of the association, possibly taking into account other factors that we may not be interested in but that may influence the response variable. In this scenario, the model selection process will often be used to determine which covariates should be included in the model (in addition to the design variables). In a scenario that we will term *exploratory explanation*, we do not have a particular hypothesis in mind but are interested in determining which variables affect the response variable and how. Here, the model selection process will be used on the data itself to find the answer to these questions. When our goal is explanation, whether confirmatory or exploratory, model selection will be based on a balance between parsimony and goodness of fit that is shifted slightly more towards the former, particularly in the second scenario. Further, with explanation, our focus will be on finding estimates, errors, and confidence intervals for the model parameters (the βs) since they can be used to quantify the relationship between the explanatory variables and the response variables, which is what we are interested in.

Before proceeding, it should be noted that the goal of exploratory explanation is somewhat less statistically sound than the goal of confirmatory explanation because the former goal can result in ***data dredging***. For instance, when there is a large number of potential explanatory variables none of which truly affect the response variable, a model may still suggest that a particular explanatory variable has an effect on the response variable because of the perils of multiple tests (as discussed in "Descriptive Statistics"). This is a particular danger with large data sets with many units.

Classes of Models

Equation (1) can be used to describe models belonging to a number of different classes. The particular model class that should be employed to answer questions about the data set at hand depends primarily on the form of the responses, especially on whether they are univariate or multivariate but also on their *type* (i.e., continuous, discrete, ordinal, or nominal). Further, the choice of model class also depends on the particular questions of interest (e.g., whether we are interested in the particular units in the data set or in the population of units) and on the study design (e.g., whether blocking was used, whether units in the sample belong to larger subgroups, etc . . . .). In addition, the choice of model class may be affected by the nature of the explanatory variable(s), although this aspect of the data may affect only the interpretation of the parameters in the chosen model.

In regards to the form of the response variable, the choice of model class depends first on whether the response variable is multivariate (or univariate) and then on the *type* (i.e., continuous, discrete, ordinal, or nominal) of the response variable(s). If there is more than one response variable, then there a number of large classes of models that can be considered: MANOVA or repeated measures ANOVA, GEE (Generalised Estimating Equation models), and random effects models such as GLMM (Generalised Linear Mixed Models), LME (Linear Mixed Effects models), and NLME (Non-Linear Mixed Effects models). For instance, if the response variable consists of longitudinal measurements for each unit, then it may be desirable to use a mixed effects (i.e., fixed effects and random effects) model or a generalised estimating equation model. (It should be noted that considerations other than the multivariate nature of the responses, such as a study design in which units belong to different subgroups, may also lead to the use of these model classes). If, on the other hand, there is only one response variable, then the choice of model class depends primarily on that variable's type: if it is continuous, then linear models or non-linear models are often considered (or, if the explanatory variables are all categorical, ANOVA, which is basically a special linear model); if it is continuous but always positive, a gamma model could be used; if it is discrete (i.e., a count variable), then a Poisson model may be used; if it is dichotomous or records proportions, then a logistic regression model might be used; if it is polytomous and ordinal, then a proportional odds model might be considered; and if the variable is polytomous and nominal, then a multiple logistic regression (or softmax) model might be used.

A Final Note on Models and Fitting Them

In this course, only ***parametric*** statistical models will be presented (with the exception of loess smoothing, which will be briefly mentioned in the lecture on Linear Models). In a parametric statistical model, the form of $\eta(\ )$ is known in the sense that a particular form for $\eta(\ )$ is specified before the model is fit to the data (instead of its form being estimated from the data)—this does not preclude the possibility that the analyst may fit the model using several different forms of $\eta(\ )$ and then decide between them. Also, in parametric models, there is a finite (and usually small, relative to the number of units) number of explanatory variables.

For the parametric models presented in this course, we will discuss only *frequentist* model selection and checking methods and estimation and inference techniques (for either β or Y). However, you should be aware that, for many of the models we will discuss, there are *Bayesian* alternatives to these methods and techniques that may be preferred in certain situations (and by certain individuals!).

A. Roddam (2000), K. Javaras (2002), and W. Vos (2002)