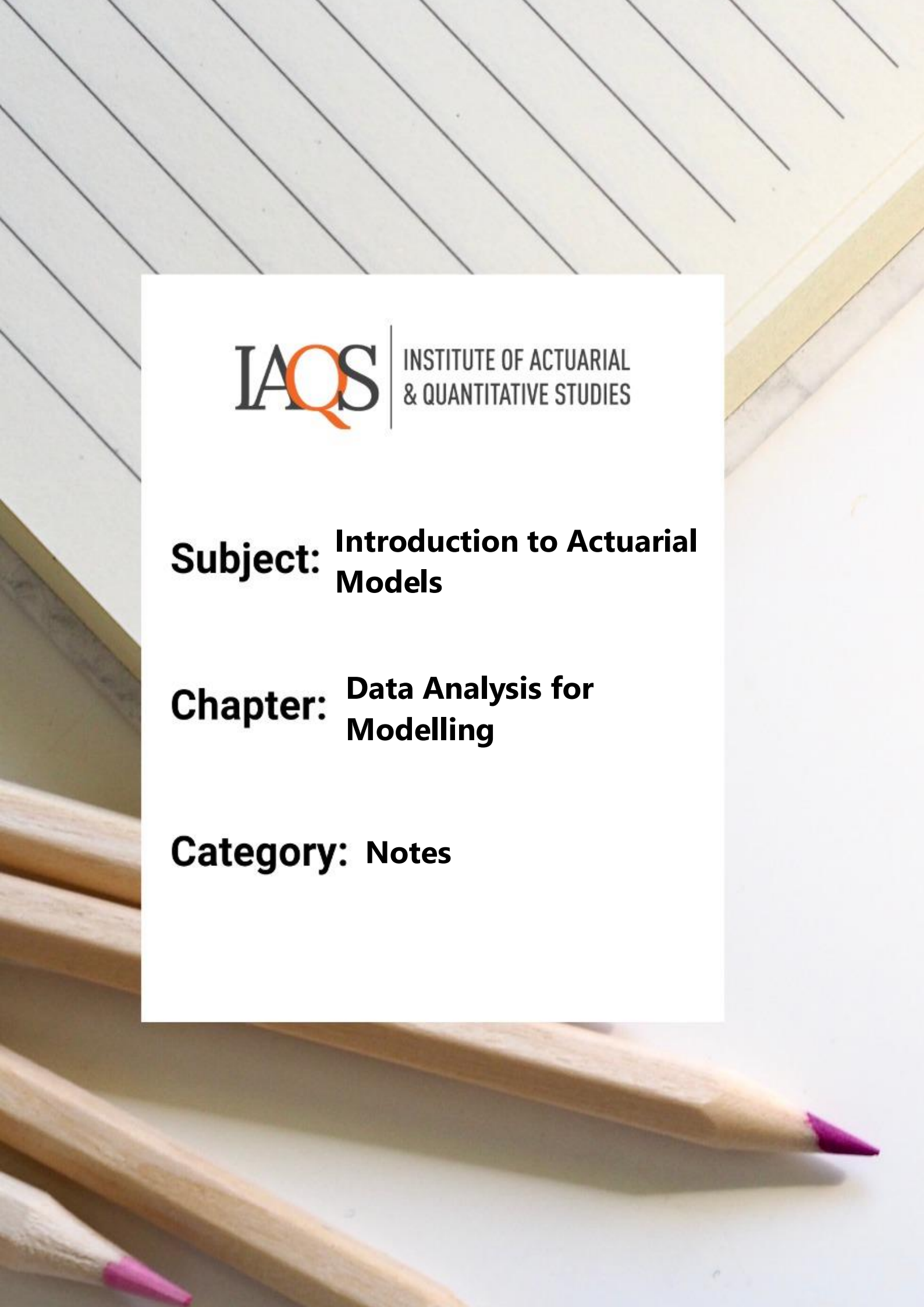

Subject: Introduction to Actuarial Models

Chapter: Data Analysis for Modelling

Category: Notes

Why is experimental design important for modelling?

Output from Process Model is Fitted Mathematical Function

The output from modelling is a fitted mathematical function with estimated coefficients. For example, in modelling, y, as a function of, x, an analyst may suggest the function

$$y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$$

in which the coefficients to be estimated are $\beta_0, \beta_1 \, and \, \beta_{11}$. Even for a given functional form, there is an infinite number of potential coefficient values that potentially may be used. Each of these coefficient values will in turn yield predicted values.

What are Good Coefficient Values?

Poor values of the coefficients are those for which the resulting predicted values are considerably different from the observed raw data "y".

Good values of the coefficients are those for which the resulting predicted values are close to the observed raw data "y".

The best values of the coefficients are those for which the resulting predicted values are close to the observed raw data "y", and the statistical uncertainty connected with each coefficient is small.

There are two considerations that are useful for the generation of "best" coefficients:

1. Least squares criterion
2. Design of experiment principles

For a given data set (e.g., 10 (x,y) pairs), the most common procedure for obtaining the coefficients for

Least Squares Criterion

$$y = f(\vec{x}, \vec{\beta}) + \epsilon$$

is the least squares estimation criterion. This criterion yields coefficients with predicted values that are closest to the raw data in the sense that the sum of the squared differences between the raw data and the predicted values is as small as possible.

Least squares estimates are popular because

1. the estimators are statistically optimal (BLUEs: Best Linear Unbiased Estimators);
2. the estimation algorithm is mathematically tractable, in closed form, and therefore easily programmable.

How then can this be improved? For a given set of "x" values it cannot be; but frequently the choice of the "x" values is under our control. If we can select the "x" values, the coefficients will have less variability than if the "x" are not controlled.

Design of Experiment Principles

As to what values should be used for the "x's", we look to established experimental design principles for guidance.

The first principle of experimental design is to control the values within the "x" vector such that after the "y" data are collected, the subsequent model coefficients are as good, in the sense of having the smallest variation, as possible.

The key underlying point with respect to design of experiments and modelling is that even though (for simple (x,y) fitting, for example) the least squares criterion may yield optimal (minimal

variation) estimators for a given distribution of "x" values, some distributions of data in the "x" vector may yield better (smaller variation) coefficient estimates than other "x" vectors. If the analyst can specify the values in the "x" vector, then he or she may be able to drastically change and reduce the noisiness of the subsequent least squares coefficient estimates.

To see the effect of experimental design on process modelling, consider the following simplest case of fitting a line:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Suppose the analyst can afford 10 observations (that is, 10 (x,y) pairs) for the purpose of determining optimal (that is, minimal variation) estimators of β_0 and β_1. What 10 "x" values should be used for the purpose of collecting the corresponding 10 "y" values? Colloquially, where should the 10 "x" values be sprinkled along the horizontal axis so as to minimize the variation of the least squares estimated coefficients for β_0 and β_1? Should the 10 values be:

1. ten equi-spaced values across the range of interest?
2. five replicated equi-spaced values across the range of interest?
3. five values at the minimum of the "x" range and five values at the maximum of the "x" range?
4. one value at the minimum, eight values at the midrange, and one value at the maximum?
5. four values at the minimum, two values at mid-range, and four values at the maximum?

or (in terms of "quality" of the resulting estimates for β_0 and β_1) perhaps it doesn't make any difference?

For each of the above five experimental designs, there will of course be "y" data collected, followed by the generation of least squares estimates for β_0 and β_1, and so each design will in turn yield a fitted line.

What are the basic steps for developing an effective model?

Basic Steps of Model Building

The basic steps used for model-building are the same across all modelling methods. The details vary somewhat from method to method, but an understanding of the common steps, combined with the typical underlying assumptions needed for the analysis, provides a framework in which the results from almost any method can be interpreted and understood.

The basic steps of the model-building process are:

1. model selection
2. model fitting, and
3. model validation.

These three basic steps are used iteratively until an appropriate model for the data has been developed. In the model selection step, plots of the data, and assumptions about the model are used to determine the form of the model to be fit to the data. Then, using the selected model and possibly information about the data, an appropriate model-fitting method is used to estimate the unknown parameters in the model. When the parameter estimates have been made, the model is then carefully assessed to see if the underlying assumptions of the analysis appear plausible. If the assumptions seem valid, the model can be used to answer the scientific questions that prompted the modelling effort. If the model validation identifies problems with

the current model, however, then the modelling process is repeated using information from the model validation step to select and/or fit an improved model.

A Variation on the Basic Steps.

The three basic steps of process modelling described in the paragraph above assume that the data have already been collected and that the same data set can be used to fit all of the candidate models. Although this is often the case in model-building situations, one variation on the basic model-building sequence comes up when additional data are needed to fit a newly hypothesized model based on a model fit to the initial data. In this case two additional steps, experimental design and data collection, can be added to the basic sequence between model selection and model-fitting.

Design of Initial Experiment

Of course, considering the model selection and fitting before collecting the initial data is also a good idea. Without data in hand, a hypothesis about what the data will look like is needed in order to guess what the initial model should be. Hypothesizing the outcome of an experiment is not always possible, of course, but efforts made in the earliest stages of a project often maximize the efficiency of the whole model-building process and result in the best possible models for the process.

How do I select a function to describe my modelling?

Selecting a function to Model

Selecting a model of the right form to fit a set of data usually requires the use of empirical evidence in the data, knowledge of the process and some trial-and-error experimentation. As mentioned previously, model building is always an iterative process. Much of the need to iterate stems from the difficulty in initially selecting a function that describes the data well. Details about the data are often not easily visible in the data as originally observed. The fine structure in the data can usually only be elicited by use of model-building tools such as residual plots and repeated refinement of the model form. As a result, it is important not to overlook any of the sources of information that indicate what the form of the model should be.

Sometimes the different sources of information that need to be integrated to find an effective model will be contradictory. An open mind and a willingness to think about what the data are saying is important. Maintaining balance and looking for alternate sources for unusual effects found in the data are also important.

Another helpful ingredient in model selection is a wide knowledge of the shapes that different mathematical functions can assume. Knowing something about the models that have been found to work well in the past for different application types also helps. A menu of different functions provides one way to learn about the function shapes and flexibility.

How are estimates of the unknown parameters obtained?

Parameter Estimation in General

After selecting the basic form of the functional part of the model, the next step in the model-building process is estimation of the unknown parameters in the function. In general, this is accomplished by solving an optimization problem in which the objective function (the function being minimized or maximized) relates the response variable and the functional part of the model containing the unknown parameters in a way that will produce parameter estimates that will be close to the true, unknown parameter values.

The unknown parameters are, loosely speaking, treated as variables to be solved for in the optimization, and the data serve as known coefficients of the objective function in this stage of the modelling process.

In theory, there are as many different ways of estimating parameters as there are objective functions to be minimized or maximized. However, a few principles have dominated because they result in parameter estimators that have good statistical properties. The two major methods of parameter estimation for process models are **maximum likelihood** and **least squares**. Both of these methods provide parameter estimators that have many good properties. Both maximum likelihood and least squares are sensitive to the presence of outliers, however. There are also many newer methods of parameter estimation, called robust methods, that try to balance the efficiency and desirable properties of least squares and maximum likelihood with a lower sensitivity to outliers.

Although robust techniques are valuable, they are not as well developed as the more traditional methods and often require

specialized software that is not readily available. Maximum likelihood also requires specialized algorithms, in general.

How can I tell if a model fits my data?

R^2 is Not Enough!

Model validation is possibly the most important step in the model building sequence. It is also one of the most overlooked. Often the validation of a model seems to consist of nothing more than quoting the R^2 statistic from the fit (which measures the fraction of the total variability in the response that is accounted for by the model). Unfortunately, a high R^2 value does not guarantee that the model fits the data well. Use of a model that does not fit the data well cannot provide good answers to the underlying scientific questions under investigation.

Main Tool: Graphical Residual Analysis

There are many statistical tools for model validation, but the primary tool for most process modelling applications is graphical residual analysis. Different types of plots of the residuals from a fitted model provide information on the adequacy of different aspects of the model. Numerical methods for model validation, such as the R^2 statistic, are also useful, but usually to a lesser degree than graphical methods. Graphical methods have an advantage over numerical methods for model validation because they readily illustrate a broad range of complex aspects of the relationship between the model and the data. Numerical methods for model validation tend to be narrowly focused on a particular aspect of the relationship between the model and the data and often try to compress that information into a single descriptive number or test result.

Numerical Methods'

Numerical methods do play an important role as

Forte

confirmatory methods for graphical techniques, however. There are also a few modelling situations in which graphical methods cannot easily be used. In these cases, numerical methods provide a fallback position for model validation. One common situation when numerical validation methods take precedence over graphical methods is when the number of parameters being estimated is relatively close to the size of the data set. In this situation residual plots are often difficult to interpret due to constraints on the residuals imposed by the estimation of the unknown parameters.

Residuals

The residuals from a fitted model are the differences between the responses observed at each combination values of the explanatory variables and the corresponding prediction of the response computed using the regression function. Mathematically, the definition of the residual for the *ith* observation in the data set is written

$$e_i = y_i - \hat{y} \text{ where } \hat{y} = f(\vec{x_i}, \vec{\beta})$$

with y_i denoting the *ith* response in the data set and $\vec{x_i}$ represents the list of explanatory variables, each set at the corresponding values found in the *ith* observation in the data set.

If my current model does not fit the data well, how can I improve it?

What to do next?

Validating a model using residual plots, formal hypothesis tests and descriptive statistics would be quite frustrating if discovery of a problem meant restarting the modelling process back at square one. Fortunately, however, there are also, techniques and tools to remedy many of the problems uncovered using residual analysis. In some cases, the model validation methods themselves suggest appropriate changes to a model at the same time problems are uncovered. This is especially true of the graphical tools for model validation, though tests on the parameters in the regression function also offer insight into model refinement.

Methods for Model Improvement

1. Updating the Function Based on Residual Plots
2. Accounting for Non-Constant Variation Across the Data
3. Accounting for Errors with a Non-Normal Distribution