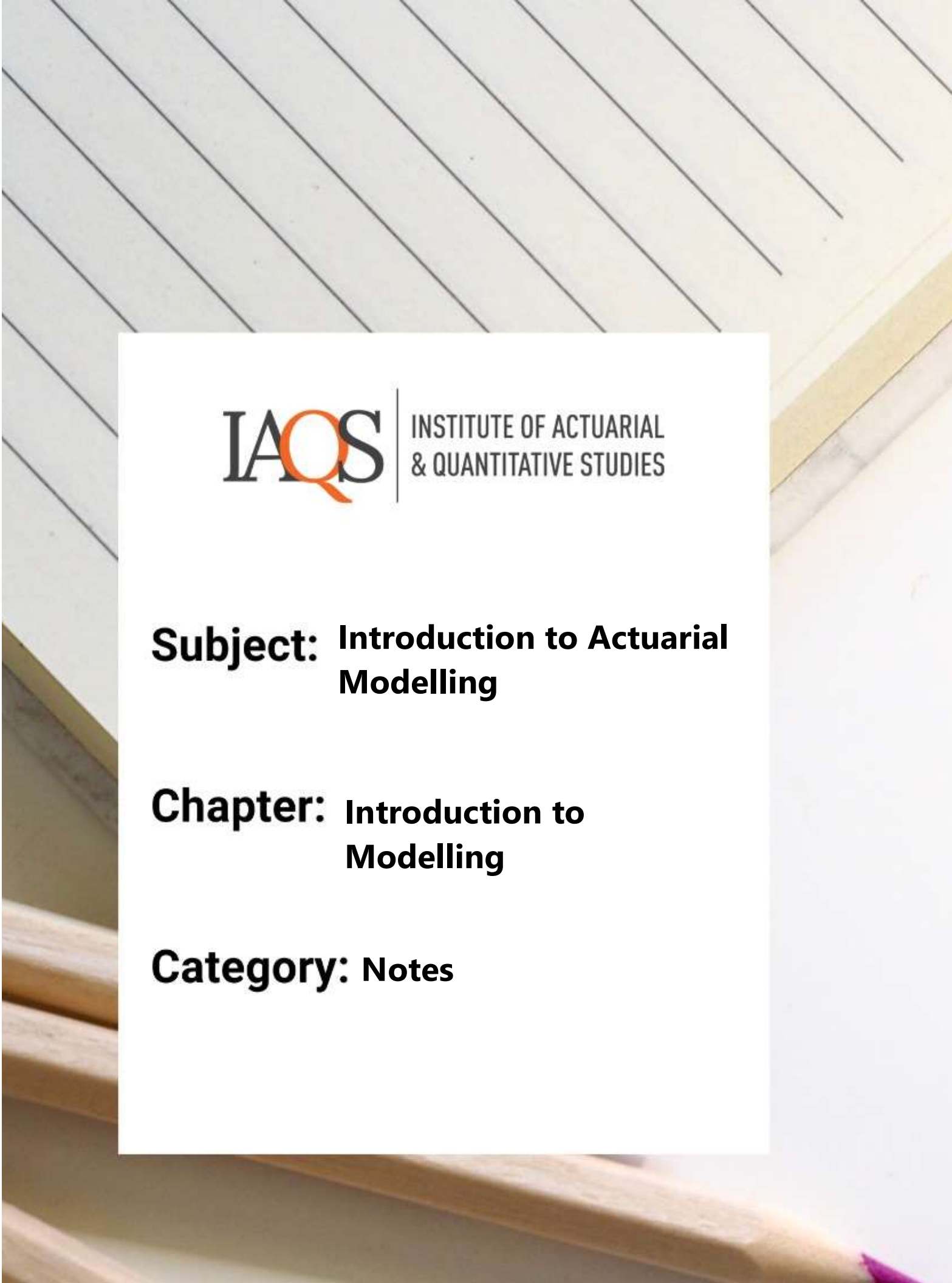


IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Subject: Introduction to Actuarial Modelling

Chapter: Introduction to Modelling

Category: Notes

Introduction

Overview

In modelling, the total variation of the measured data can be described by partitioning the variability into its deterministic part, which is a function of the data values, plus some left-over random error.

The data that vary deterministically can be explained by a certain relationship using mathematical functions.
For example, using line of best fit for linear relationships.

However, for the left-over random errors, the relationship will not be purely deterministic. The random errors cannot be characterized individually, but will follow some probability distribution that will describe the relative frequencies of occurrence of different-sized errors.
For example, a histogram can be fit to the random errors which shows the relative frequencies of observing different-sized random errors.
Then the relative frequencies of the random errors can be summarized by a (normal) probability distribution.

Basic Definition

In simple terms, **statistical modeling** is a simplified, mathematically-formalized way to approximate reality (i.e. what generates your data) and optionally to make predictions from this approximation.

The statistical model is the mathematical equation that is used.

Precise form

It is the concise description of the total variation in one quantity Y, ,by partitioning it into

1. **a deterministic component** given by a mathematical

function of one or more other quantities x_1, x_2 , …

plus

2. **a random component** that follows a particular probability distribution.

Example (Linear regression analysis)

Suppose that we have a population of school children, with the ages of the children distributed uniformly, in the population.

The height of a child will be stochastically related to the age: e.g. when we know that a child is of age 7, this influences the chance of the child being 1.5 meters tall.

We could formalize that relationship in a linear regression model, like this: $height_i = b_0 + b_1 age_i + \varepsilon_i$, where b_0 is the intercept, b_1 is a parameter that age is multiplied by to obtain a prediction of height, ε_i is the error term, and i identifies the child.

This implies that height is predicted by age, with some error.

Example (Multiple regression analysis)

When forecasting financial statements for a company it may be useful to do a multiple regression analysis to determine how changes in certain assumption or drivers of business will impact revenue or expenses in the future.

For instance; there may be very high correlation between the number of salespeople employed by a company, the number of stores they operate and the revenue the business generates.

What terminology do statisticians use to describe models?

Model Components

There are three main parts to every model. These are

1. the response variable, usually denoted by y,

2. the mathematical function, usually denoted as, $f(\vec{x};\vec{\beta})$,

and

3. the random errors, usually denoted by €.

Form of Model

The general or basic form is given as;

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

Most of the models have this general form.

The random errors that are included in the model make the relationship between the response variable and the predictor variables a "statistical" one, rather than a perfect deterministic one. This is because the functional relationship between the response and predictors holds only on average, not for each data point

Let's understand each component separately.

Response Variable

The **response variable** is the one we want to describe, to explain, to predict. As a rule of thumb, the dependent variable is often the one we represent on the Y axis in modelling charts. It is also known as the dependent variable as the value of this

variable is dependent on certain factors or certain other variables.

Explanatory Variable

Explanatory variables, also referred to as **independent** variables, are the ones we use to explain, to describe or to predict the dependent variable(s).

Explanatory variables are often represented on the X axis. They are also called as predictors or regressors.

Mathematical Function

The mathematical function consists of two parts. These parts are the predictor variables x1, x2,... and the parameters, $\beta_1 \beta_2$,...

The predictor variables are observed along with the response variable. They are the quantities described on the previous page as inputs to the mathematical function, $f(\vec{x}; \vec{\beta})$.

The collection of all of the predictor variables is denoted by

$\vec{x}$ $\vec{x} \equiv (x_1, x_2, \ldots)$

The parameters are the quantities that will be estimated during the modelling process. Their true values are unknown and unknowable, except in simulation experiments.
As for the predictor variables, the collection of all of the parameters is denoted by

$\vec{\beta}$ $\vec{\beta} \equiv (\beta_0, \beta_1, \ldots)$

The parameters and predictor variables are combined in different forms to give the function used to describe the deterministic variation in the response variable.

Random errors

Like the parameters in the mathematical function, the random errors are unknown.

They are simply the difference between the data and the mathematical function. They are assumed to follow a particular probability distribution, however, which is used to describe their aggregate behaviour.

The probability distribution that describes the errors has a mean of zero and an unknown standard deviation.

What are the models used for?

Main purposes

Models are used for four main purposes:
1. Estimation,
2. Prediction,
3. Calibration, and
4. Optimization.

More detailed explanations of the uses for models are given below;

Estimation

The goal of estimation is to determine the value of the regression function (i.e., the average value of the response variable), for a particular combination of the values of the predictor variables.

Regression function values can be estimated for any combination of predictor variable values, including values for which no data have been measured or observed.

Function values estimated for points within the observed space of predictor variable values are sometimes called interpolations.

Estimation of regression function values for points outside the observed space of predictor variable values, called extrapolations, are sometimes necessary, but require caution.

A critical part of estimation is an assessment of how much am estimated value will fluctuate due to the noise in the data

Prediction

The goal of prediction is to determine either
1. the value of a new observation of the response variable, or

2. the values of a specified proportion of all future observations of the response variable,

for a particular combination of the values of the predictor variables.

Predictions can be made for any combination of predictor variable values, including values for which no data have been measured or observed.

As in the case of estimation, predictions made outside the observed space of predictor variable values are sometimes necessary, but require caution.

Calibration

The goal of calibration is to quantitatively relate measurements made using one measurement system to those of another measurement system. This is done so that measurements can be compared in common units or to tie results from a relative measurement method to absolute units.

Optimization

Optimization is performed to determine the values of process inputs that should be used to obtain the desired process output.

Typical optimization goals might be to maximize the yield of a process, to minimize the processing time required to fabricate a product, or to hit a target product specification with minimum variation in order to maintain specified tolerances.

What are some of the different statistical methods for model building?

Selecting an Appropriate Stat Method for Modeling

In order to build a statistical model, we need to be very careful in selecting the method.

There are more general approaches and more competing techniques available for model building. There is often more than one statistical tool that can be effectively applied to a given modeling application.

The large menu of methods applicable to modeling problems means that there is both more opportunity for effective and efficient solutions and more potential to spend time doing different analyses, comparing different solutions and mastering the use of different tools.

In the process of developing the model we will often come across situations where we build a first basic model and then run the model, perform calculations and try to improvise.

Every model you run tells you a story. Stop and listen to it. Look at the coefficients. Look at R-squared. Did it change? How much do coefficients change from a model with control variables to one without?
When you pause to do this, you can make better decisions on the model to run next.

Now we discuss some of the most popular and well-established statistical techniques that are useful for different model building situations

Modelling methods that we will discuss

1. Linear Least Squares Regression

2. Nonlinear Least Squares Regression

3. Weighted Least Squares Regression

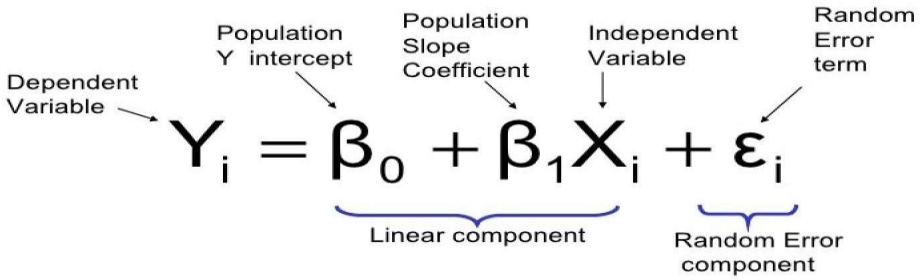
1] *Linear Least Squares Regression*

Least squares regression method is by far the most widely used modelling method.

The least squares method is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve.
Least squares regression is used to predict the behavior of dependent variables.

This method of regression analysis begins with a set of data points to be plotted on an x- and y-axis graph. An analyst using the least squares method will generate a line of best fit that explains the potential relationship between independent and dependent variables.

Linear least square regression equation example;



Linear least squares regression also gets its name from the way the estimates of the unknown parameters are computed – using the "method of least squares". In this method the unknown parameters are estimated by minimizing the sum of the squared deviations between the data and the model.

Example of the Least Squares Method
An example of the least squares method is an analyst who wishes to test the relationship between a company's stock returns, and the returns of the index for which the stock is a component.

In this example, the analyst seeks to test the dependence of the stock returns on the index returns. To achieve this, all of the returns are plotted on a chart. The index returns are then designated as the independent variable, and the stock returns are the dependent variable.

The line of best fit provides the analyst with coefficients explaining the level of dependence.

2] *Nonlinear Least Squares Regression*

Nonlinear least squares regression extends linear least squares regression for use with a much larger and more general class of functions. Almost any function that can be written in closed form can be incorporated in a nonlinear regression model.

Nonlinear regression is a regression in which the dependent variables are modelled as a non-linear function of model parameters and one or more independent variables.
The reason that these models are called nonlinear regression is because the relationships between the dependent and independent parameters are not linear.

As the name suggests, a nonlinear model is any model of the basic form,

$$Y = f(\vec{x};\vec{\beta})+\varepsilon,$$

in which
i) the functional part of the model is *not linear* with respect to the unknown parameters, β0,β1,..., and
ii) the method of least squares is used to estimate the values of the unknown parameters.

Some examples of nonlinear models include:

a) $f(x;\vec{\beta}) = (\beta_0+\beta_1 x)/(1+\beta_2 x)$

b) $f(x;\vec{\beta}) = \beta_1 x^{\beta 2}$

Advantages

The biggest advantage of nonlinear regression over other techniques is the broad range of functions that can be fit.

Disadvantages

The major cost of moving to nonlinear least squares from simpler modelling techniques like linear least squares is the need to use iterative optimization procedures to compute the parameter estimates.

3] *Weighted Least Squares Regression*

One of the common assumptions underlying most process modelling methods, including linear and nonlinear least squares regression, is that each data point provides equally precise information about the deterministic part of the total process variation.

Weighted Least Squares is an extension of simple regression. Non-negative constants (weights) are attached to data points. The values scattered close to each other in the centre certainly reflect more information and hence should be given more weightage than those scattered far away from each other and those which are outliers.

It is used when any of the following are true:
1. Your data violates the assumption of homoscedasticity.
2. You want to concentrate on certain areas
3. You have any other situation where data points should not be treated equally**.**

Advantages

Weighted least squares has several advantages over other methods, including:
It's well suited to extracting maximum information from small data sets.

It is the only method that can be used for data points of

varying quality.

Disadvantages

It requires that you know exactly what the weights are. Estimating weights can have unpredictable results, especially when dealing with small samples.
Therefore, the technique should only be used when your weight estimates are fairly precise. In practice, precision of weight estimates usually isn't possible.

Sensitivity to outlier is a problem. A rogue outlier given an inappropriate weight could dramatically skew your results.