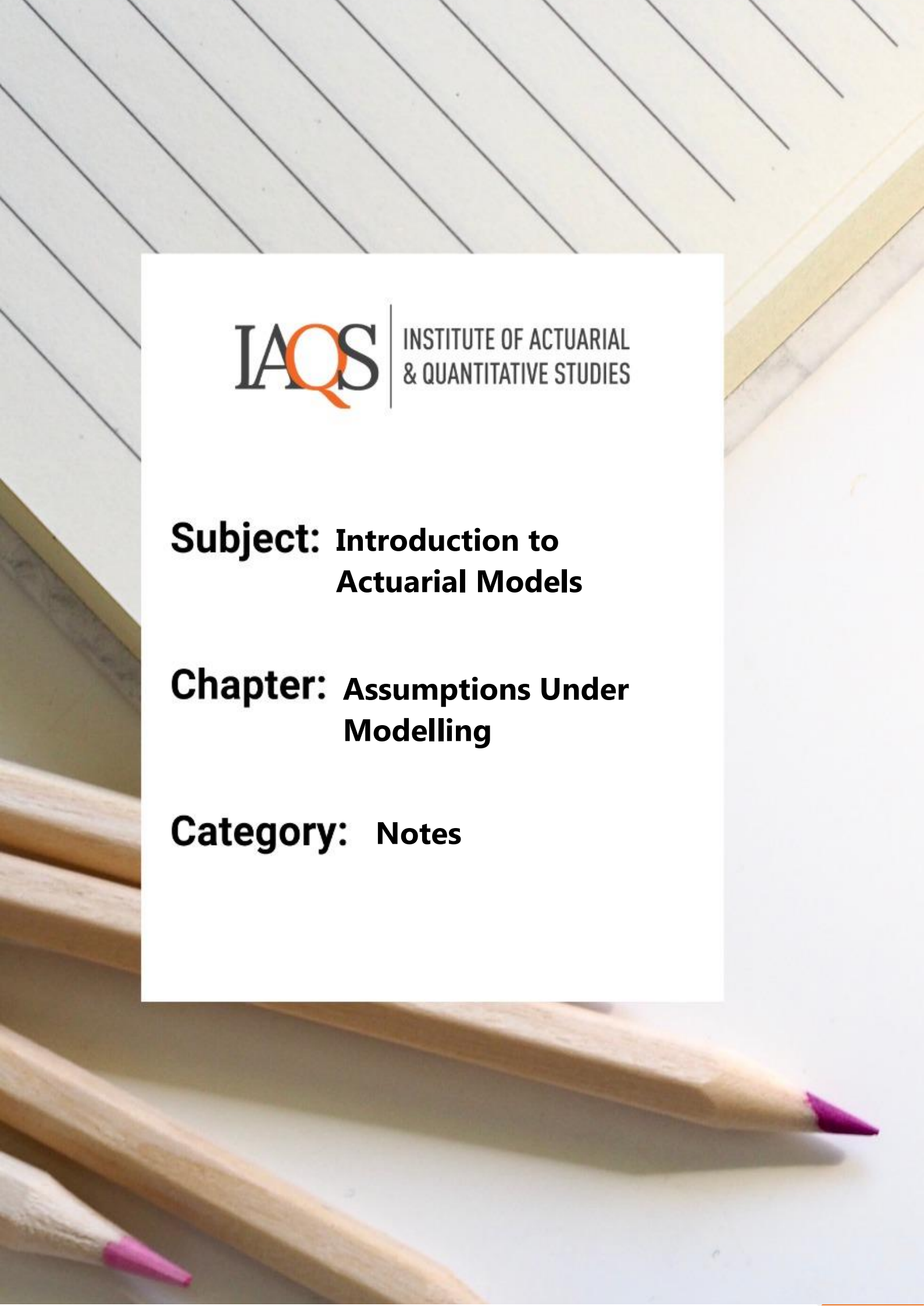


INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Subject: Introduction to Actuarial Models

Chapter: Assumptions Under Modelling

Category: Notes

Underlying Assumptions in Modelling

Assumptions underlie most actions.

Most, if not all, thoughtful actions that people take are based on ideas, or assumptions, about how those actions will affect the goals they want to achieve.
It is important to understand what the assumptions are for any modeling method because the validity of these assumptions affect whether or not the goals of the analysis will be met.

Checking Assumptions Provides Feedback on Actions

Checking model assumptions is essential in building a model that will be used for prediction. If assumptions are not met, the model may inaccurately reflect the data and will likely result in inaccurate predictions. Each model has different assumptions that must be met, so checking assumptions is important both in choosing a model and in verifying that it is the appropriate model to use.

Checking model assumptions may be relatively simple, but hugely important step in optimizing model performance and increasing model reliability. Prior to building your model, check to see if your data meets the specific assumptions that go with your chosen model. Start with a visual check. If your visualizations are even a bit unclear on whether or not your data meets the specific assumption you are checking for, use a more specific diagnostic tool to either confirm or deny your suspicions. This way, you can assure that you are using the most appropriate model for your data, which will lead to better prediction capabilities.

Overview

In this chapter we consider:

What are the typical underlying assumptions in modelling?

It includes:

1. The process is a *statistical* process.
2. The means of the random errors are zero.
3. The random errors have a constant standard deviation.
4. The random errors follow a normal distribution.
5. The data are randomly sampled from the process.
6. The explanatory variables are observed without error.

The Assumptions

Typical Assumptions under Modelling

Here we discuss the assumptions which are common to most of the models build under statistics.
However certain models may require few more specific assumptions to be made further, but this completely depends on the model and the data.

1]
The process is a ***statistical*** *process.*

The most basic assumption inherent to all statistical methods for modelling is that the process to be described is actually a statistical process.

"Statistical" implies Random Variation.

In order to successfully model using statistical methods, it must include random variation.
Random variation is what makes the process statistical rather than purely deterministic.

The random variation serves as a baseline for drawing conclusions about the nature of the deterministic part of the process. If there were no random noise in the data, then conclusions based on statistical methods would no longer make sense or be appropriate.

2]
The means of the random errors are zero.

The error term accounts for the variation in the dependent variable that the independent variables do not explain. For your model to be unbiased, the average value of the error term must equal zero.

Suppose the average error is +7. This non-zero average error indicates that our model systematically underpredicts the

observed values. Statisticians refer to systematic error like this as bias, and it signifies that our model is inadequate because it is not correct on average.

Stated another way, we want the expected value of the error to equal zero. If the expected value is +7 rather than zero, part of the error term is predictable, and we should add that information to the regression model itself. We want only random error left for the error term.

3]
The random errors have a constant standard deviation.

The assumption about the random error term is that its probability distribution remains the same for all observations of *X* and in particular that the variance of each error term is the same for all values of the explanatory variables, i.e the variance of errors is the same across all levels of the independent variables.

This assumption is known as the assumption of homoscedasticity or the assumption of constant variance of the error term.

If the assumption of homoscedastic disturbance (Constant Variance) is not fulfilled, following are the consequence:

1. We cannot apply the formula of the variance of the coefficient to conduct tests of significance and construct confidence intervals

2. The prediction would be inefficient, because of the variance of prediction includes the variance of error and of the parameter estimates which are not minimal due to the incidence of heteroscedasticity i.e. The prediction of *Y* for a given value of *X* based on the estimates β^'s from the original data, would have a high variance.

3. The estimates of the coefficients also would be inefficient.

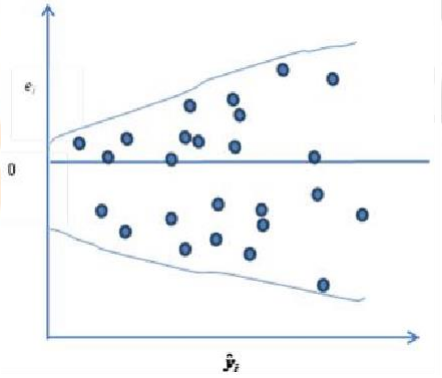
The presence of any drift, outliers or any other trends will distort the assumption of zero mean and constant variance.

The graphical analysis of residuals is a very effective way to investigate the adequacy of the fit of a model and to check the underlying assumptions. Various types of graphics can be examined for different assumptions.

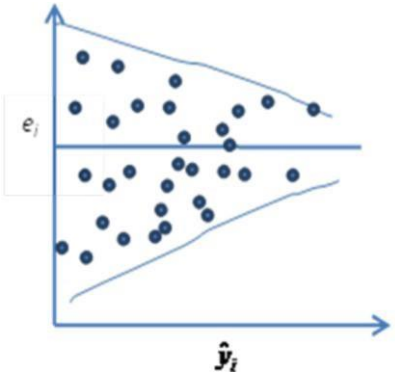
For instance,

- Residuals plotted against fitted values.

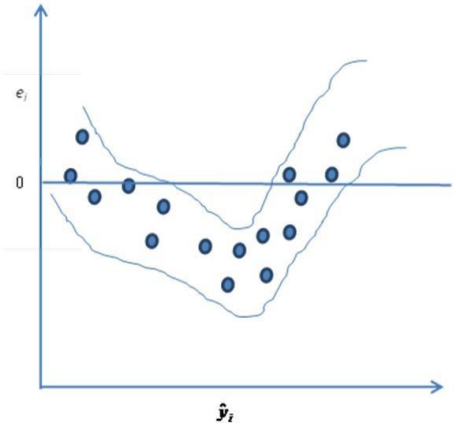
a) If plot is such that the residuals can be contained is an outward opening funnel then such pattern indicates that the variance of errors is not constant but it is an increasing function of y. Check in the graph below:



b) If the case is opposite, that is when we have inward opening funnel, then such pattern would indicate that variance of errors is not constant but it is a decreasing function of y.



c) If plot is such that the residuals are contained inside a curved plot, then it indicates nonlinearity.



Note: A plot of residuals against fitted values may also reveal one or more unusually large residuals. These points are potential outliers. Large residuals that occur at the extreme of y^ values could also indicate that either the variance is not constant or the true relationship between y and X is nonlinear. These possibilities should be investigated before the points are considered outliers.

- Residuals plotted against explanatory variables.

Plotting of residuals against the corresponding values of each explanatory variable can also be helpful.

It is also helpful to plot the residuals against explanatory variables that are not currently in the model, but which could potentially be included. Any structure in the plot of residuals versus an omitted variable indicates that incorporation of that variable could improve the model.

- Plots of residuals in time sequence

If the time sequence in which the data were collected is known, then the residuals can be plotted against the time order.

The time sequence plot of residuals may indicate that the errors at one time period are correlated with those at other time periods. The correlation between model errors at different time periods is called autocorrelation. This can be further examined.

Thus, various trends in the plot could indicate the model inadequacies and cope for improvement.

4]
The random errors follow a normal distribution.

After fitting a model to the data and validating it, questions are usually answered by computing statistical intervals for required quantities using the model. These intervals give the range of plausible values for the model parameters based on the data and the underlying assumptions about it.
In order for these intervals to truly have their specified probabilistic interpretations, the form of the distribution of the random errors must be known.

With most modelling methods, inferences about the data are based on the idea that the random errors are drawn from a normal distribution.

One reason this is done is because the normal distribution often describes the actual distribution of the random errors in real-world processes reasonably well.

The normal distribution is also used because the mathematical theory behind it is well-developed and supports a broad array of inferences on functions of the data relevant to different types of questions about the models.

The methods used for parameter estimation can also imply the assumption of normally distributed random errors.

The normal probability plots help in verifying the assumption of normal distribution. If errors come from thicker and heavier tails than normal, then the least squares fit may be sensitive to small data sets.

Hence, we have the assumption of normality coming into picture!

5]
The data are randomly sampled from the population. (random sampling)

Random sampling is a part of the sampling technique in which each sample has an equal probability of being chosen. A sample chosen randomly is meant to be an unbiased representation of the total population. An unbiased random sample is important for drawing conclusions
If for some reasons, the sample does not represent the population, the variation is called a sampling error.

Data best reflects the population Via unbiased sampling
Given that we can never determine what the actual random errors in a particular data set are, representative samples of data are best obtained by randomly sampling data from the population. Random sampling ensures that the act of data

collection does not leave behind any biases in the data, on average. This means that most of the time, over repeated samples, the data will be representative of the population.

Paying careful attention to data collection procedures and employing experimental design principles will yield a sample of data that is as close as possible to being perfectly randomly sampled from the population.

6] The explanatory variables are observed without error.

A fundamental assumption in all the statistical analysis is that all the observations are correctly measured. In the context of regression models, it is assumed that the observations on study and explanatory variables are observed without any error.

Satisfying this assumption is necessary for efficient estimation of parameters. If the magnitude of measurement errors is small, then they can be assumed to be merged in the disturbance (variance) term and they will not affect the statistical inferences much. On the other hand, if they are large in magnitude, then they will lead to incorrect and invalid statistical inferences.