Subject : Introduction to Actuarial Models

Data Analysis

Data Analysis

Data analysis is the process by which data is gathered in its raw state and analysed or processed into information which can be used for specific purposes.

We look into

- Different forms of data analysis
- Processing steps
- Practical problems with data analytics

Aim of a data analysis

Three key forms of data analysis

- Descriptive
- Inferential
- Predictive

Descriptive

Data presented in its raw state can be difficult to manage and draw meaningful conclusions from, particularly where there is a large volume of data to work with. A descriptive analysis solves this problem by presenting the data in a simpler format, more easily understood and interpreted by the user.

Simply put, this might involve summarizing the data or presenting it in a format which highlights any patterns or trends. A descriptive analysis is not intended to enable the user to draw any specific conclusions. Rather, it describes the data actually presented.

Aim of a data analysis

Descriptive (Contd..)

Data Two key measures, or parameters, used in a descriptive analysis are the measure of central tendency and the dispersion. The most common measurements of central tendency are the mean, the median and the mode. Typical measurements of the dispersion are the standard deviation and ranges such as the interquartile range.

It can also be important to describe other aspects of the shape of the (empirical) distribution of the data, for example by calculating measures of skewness and kurtosis

E.g. Presenting health insurance claims raw data received from IIB to your Actuary using the above mentioned statistics rather than the raw data itself which can be quite large.

Aim of a data analysis

Inferential

Often it is not feasible or practical to collect data in respect of the whole population, particularly when that population is very large. For example, when conducting an opinion poll in a large country, it may not be cost effective to survey every citizen. A practical solution to this problem might be to gather data in respect of a sample, which is used to represent the wider population. The analysis of the data from this sample is called inferential analysis.

The sample analysis involves estimating the parameters as described in Section above and testing hypotheses. It is generally accepted that if the sample is large and taken at random (selected without prejudice), then it quite accurately represents the statistics of the population, such as distribution, probability, mean, standard deviation, However, this is also contingent upon the user making reasonably correct hypothesis about the population in order to perform the inferential analysis.

E.g. Taking several exit polls before and on the day of the elections.

Aim of a data analysis

Predictive

Predictive analysis extends the principles behind inferential analysis in order for the user to analyze past data and make predictions about future events.

It achieves this by using an existing set of data with known attributes (also known as features), known as the training set in order to discover potentially predictive relationships. Those relationships are tested using a different set of data, known as the test set, to assess the strength of those relationships.

A typical example of a predictive analysis is regression analysis. The simplest form of this is linear regression where the relationship between a scalar dependent variable and an explanatory or independent variable is assumed to be linear and the training set is used to determine the slope and intercept of the line. A practical example might be the relationship between a car's braking distance against speed.

The Data Analysis Process

The key steps in the data analysis process are:

- Develop a well-defined set of objectives which need to be met by the results of the data analysis.
- Identify the data items required for the analysis.
- Collection of the data from appropriate sources.
- Processing and formatting data for analysis, e.g. inputting into a spreadsheet, database or other model.
- Cleaning data, eg addressing unusual, missing or inconsistent values.
- Exploratory data analysis, which may include descriptive analysis, inferential analysis or predictive analysis.
- Modelling the data.
- Communicating the results.
- Monitoring the process; updating the data and repeating the process if required.

The Data Analysis Process

The key steps in the data analysis process are:

- Develop a well-defined set of objectives which need to be met by the results of the data analysis.
- Identify the data items required for the analysis.
- Collection of the data from appropriate sources.
- Processing and formatting data for analysis, e.g. inputting into a spreadsheet, database or other model.
- Cleaning data, e.g. addressing unusual, missing or inconsistent values.
- Exploratory data analysis, which may include descriptive analysis, inferential analysis or predictive analysis.
- Modelling the data.
- Communicating the results.
- Monitoring the process; updating the data and repeating the process if required.

Need to ensure the relevant professional guidance is followed and the modelling teams comply with any legal requirements.

Data Sources

- Step 3 of the process described in Section 2 above refers to collection of the data needed to meet the objectives of the analysis from appropriate sources. As consideration of Steps 3, 4, and 5 makes clear, getting data into a form ready for analysis is a process, not a single event. Consequently, what is seen as the source of data can depend on your viewpoint.
- Suppose you are conducting an analysis which involves collecting survey data from a sample of people in the hope of drawing inferences about a wider population. If you are in charge of the whole process, including collecting the primary data from your selected sample, you would probably view the 'source' of the data as being the people in your sample. Having collected, cleaned and possibly summarised the data you might make it available to other investigators in JavaScript object notation (JSON) format via a web Application programming interface (API). You will then have created a secondary 'source' for others to use.
- In this section we discuss how the characteristics of the data are determined both by the primary source and the steps carried out to prepare it for analysis – which may include the steps on the journey from primary to secondary source.

Data Sources (Contd…)

- Details of particular data formats (such as JSON), or of the mechanisms for getting data from an external source into a local data structure suitable for analysis, are not covered.
- Primary data can be gathered as the outcome of a designed experiment or from an observational study (which could include a survey of responses to specific questions). In all cases, knowledge of the details of the collection process is important for a complete understanding of the data, including possible sources of bias or inaccuracy.
- Issues that the analyst should be aware of include:
  - whether the process was manual or automated;
  - limitations on the precision of the data recorded;
  - whether there was any validation at source; and
  - if data wasn't collected automatically, how was it converted to an electronic form.

Data Sources (Contd…)

- Where randomization has been used to reduce the effect of bias or confounding variables it is important to know the sampling scheme used:
  - simple random sampling e.g. selecting 10 cars from a lot of 100 cars of different brands. Each car has an equal chance of being selected. Hence we could end up selecting 10 Honda cars.
  - stratified sampling e.g. where we split the 10 brands into groups and then select one from each group. Hence the sample reflects the split of the brands as seen in the group.
  - another sampling method.
- Data may have undergone some form of pre-processing. A common example is grouping (eg by geographical area or age band). In the past, this was often done to reduce the amount of storage required and to make the number of calculations manageable. The scale of computing power available now means that this is less often an issue, but data may still be grouped: perhaps to anonymize it, or to remove the possibility of extracting sensitive (or perhaps commercially sensitive) details.

Data Sources (Contd…)

- Other aspects of the data which are determined by the collection process, and which affect the way it is analysed include the following:
  - Cross-sectional data involves recording values of the variables of interest for each case in the sample at a single moment in time. E.g. Time spent by each Facebook users this week on the website.
  - Longitudinal data involves recording values at intervals over time. E.g. Time spent by the same user on the Facebook website each week.
  - Censored data occurs when the value of a variable is only partially known, for example, if a subject in a survival study withdraws, or survives beyond the end of the study: here a lower bound for the survival period is known but the exact value isn't. E.g. Patient not showing up for follow up checks post cancer surgery.
  - Truncated data occurs when measurements on some variables are not recorded so are completely unknown. E.g. When during a sleep study the disruption in the power grid resulted in 1 hour of the study not being recorded.

Big Data

- The term big data is not well defined but has come to be used to describe data with characteristics that make it impossible to apply traditional methods of analysis (for example, those which rely on a single, well-structured data set which can be manipulated and analysed on a single computer). Typically, this means automatically collected data with characteristics that have to be inferred from the data itself rather than known in advance from the design of an experiment.
- Few examples include
  - Data held by Apple from the Apple watch users collected multiple times (heart rate, oxygen levels etc.)
  - PMI 2.5 levels captured each day across the cities in India
  - Data held by e-commerce websites like Flipkart of the items viewed by each customers but not purchased

Big Data (Contd…)

- Given the description above, the properties that can lead data to be classified as ‘big’ include:
  - size, not only does big data include a very large number of individual cases, but each might include very many variables, a high proportion of which might have empty (or null) values – leading to sparse data;
  - speed, the data to be analysed might be arriving in real time at a very fast rate – for example, from an array of sensors taking measurements thousands of time every second;
  - variety, big data is often composed of elements from many different sources which could have very different structures – or is often largely unstructured;
  - reliability, given the above three characteristics we can see that the reliability of individual data elements might be difficult to ascertain and could vary over time (for example, an internet connected sensor could go offline for a period).

Big Data (Contd…)

- Although the four points above (size, speed, variety, reliability) have been presented in the context of big data, they are characteristics that should be considered for any data source.
- For example, an actuary may need to decide if it is advisable to increase the volume of data available for a given investigation by combining an internal data set with data available externally.
- In this case, the extra processing complexity required to handle a variety of data, plus any issues of reliability of the external data, will need to be considered.

Data security, privacy and regulations

- In the design of any investigation, consideration of issues related to data security, privacy and complying with relevant regulations should be paramount.
- It is especially important to be aware that combining different data from different ‘anonymized’ sources can mean that individual cases become identifiable.
- Another point to be aware of is that just because data has been made available on the internet, doesn’t mean that that others are free to use it as they wish. This is a very complex area and laws vary between jurisdictions.

Reproducible Research

Meaning

- Reproducibility refers to the idea that when the results of a statistical analysis are reported, sufficient information is provided so that an independent third party can repeat the analysis and arrive at the same results.
- In science, reproducibility is linked to the concept of replication which refers to someone repeating an experiment and obtaining the same (or at least consistent) results. Replication can be hard, or expensive or impossible, for example if:
  - the study is big;
  - the study relies on data collected at great expense or over many years; or
  - the study is of a unique occurrence (the standards of healthcare in the aftermath of a particular event).
- Due to the possible difficulties of replication, reproducibility of the statistical analysis is often a reasonably alternative standard.

Reproducible Research

Elements Required

- Typically, reproducibility requires the original data and the computer code to be made available (or fully specified) so that other people can repeat the analysis and verify the results. In all but the most trivial cases, it will be necessary to include full documentation (eg description of each data variable, an audit trail describing the decisions made when cleaning and processing the data, and full documented code).
- Full documented code can be achieved through literate statistical programming (as defined by Knuth, 1992) where the program includes an explanation of the program in plain language, interspersed with code snippets. Within the R environment, a tool which allows this is R-markdown.
- Although not strictly required to meet the definition of reproducibility, a good version control process can ensure evolving drafts of code, documentation and reports are kept in alignment between the various stages of development and review, and changes are reversible if necessary. There are many tools that are used for version control. A popular tool used for version control is git.

Reproducible Research

Elements Required

- Typically, reproducibility requires the original data and the computer code to be made available (or fully specified) so that other people can repeat the analysis and verify the results.
- In all but the most trivial cases, it will be necessary to include full documentation (eg description of each data variable, an audit trail describing the decisions made when cleaning and processing the data, and full documented code).
- Full documented code can be achieved through literate statistical programming (as defined by Knuth, 1992) where the program includes an explanation of the program in plain language, interspersed with code snippets.
- Within the R environment, a tool which allows this is R-markdown.

Reproducible Research

Elements Required (contd…)

- Although not strictly required to meet the definition of reproducibility, a good version control process can ensure evolving drafts of code, documentation and reports are kept in alignment between the various stages of development and review, and changes are reversible if necessary.
- There are many tools that are used for version control. A popular tool used for version control is git.
- In addition to version control, documenting the software environment, the computing architecture, the operating system, the software toolchain, external dependencies and version numbers can all be important in ensuring reproducibility.

Reproducible Research

Elements Required (contd…)

- Although not strictly required to meet the definition of reproducibility, a good version control process can ensure evolving drafts of code, documentation and reports are kept in alignment between the various stages of development and review, and changes are reversible if necessary.
- There are many tools that are used for version control. A popular tool used for version control is git.
- In addition to version control, documenting the software environment, the computing architecture, the operating system, the software toolchain, external dependencies and version numbers can all be important in ensuring reproducibility.
- Where there is randomness in the statistical or machine learning techniques being used (for example random forests or neural networks) or where simulation is used, replication will require the random seed to be set.

Reproducible Research

Elements Required (contd…)

- Doing things ‘by hand’ is very likely to create problems in reproducing the work. Examples of doing things by hand are:
  - manually editing spreadsheets (rather than reading the raw data into a programming environment and making the changes there);
  - editing tables and figures (rather than ensuring that the programming environment creates them exactly as needed);
  - downloading data manually from a website (rather than doing it programmatically); and
  - pointing and clicking (unless the software used creates an audit trail of what has been clicked).

Reproducible Research

Value of Reproducibility

- Many actuarial analyses are undertaken for commercial, not scientific, reasons and are not published, but reproducibility is still valuable:
  - reproducibility is necessary for a complete technical work review (which in many cases will be a professional requirement) to ensure the analysis has been correctly carried out and the conclusions are justified by the data and analysis;
  - reproducibility may be required by external regulators and auditors;
  - reproducible research is more easily extended to investigate the effect of changes to the analysis, or to incorporate new data;
  - it is often desirable to compare the results of an investigation with a similar one carried out in the past; if the earlier investigation was reported reproducibly an analysis of the differences between the two can be carried out with confidence;
  - the discipline of reproducible research, with its emphasis on good documentation of processes and data storage, can lead to fewer errors that need correcting in the original work and, hence, greater efficiency.

Reproducible Research

Issues not addressed by Reproducibility

- Reproducibility does not mean that the analysis is correct. For example, if an incorrect distribution is assumed, the results may be wrong – even though they can be reproduced by making the same incorrect assumption about the distribution.
- However, by making clear how the results are achieved, it does allow transparency so that incorrect analysis can be appropriately challenged.
- If activities involved in reproducibility happen only at the end of an analysis, this may be too late for resulting challenges to be dealt with. For example, resources may have been moved on to other projects.

Thank You