Data and Systems

Guojun Gan*

February 28, 2018

This chapter covers the learning areas on data and systems outlined in the IAA (International Actuarial Association) Education Syllabus published in September 2015.

1 Data

1.1 Data Types and Sources

In terms of how data are collected, data can be divided into two types (Hox and Boeije, 2005): primary data and secondary data. Primary data are original data that are collected for a specific research problem. Secondary data are data originally collected for a different purpose and reused for another research problem. A major advantage of using primary data is that the theoretical constructs, the research design, and the data collection strategy can be tailored to the underlying research question to ensure that the data collected indeed help to solve the problem. A disadvantage of using primary data is that data collection can be costly and time-consuming. Using secondary data has the advantage of lower cost and faster access to relevant information. However, using secondary data may not be optimal for the research question under consideration.

In terms of the degree of organization of the data, data can be also divided into two types (Inmon and Linstedt, 2014; O'Leary, 2013; Hashem et al., 2015; Abdullah and Ahmad, 2013; Pries and Dunnigan, 2015): structured data and unstructured data. Structured data have a predictable and regularly occurring format. In contrast, unstructured data are unpredictable and have no structure that is recognizable to a computer. Structured data consists of records, attributes, keys, and indices and are typically managed by a database management system (DBMS) such as IBM DB2, Oracle, MySQL, and Microsoft SQL Server. As a result, most units of structured data can be located quickly and easily. Unstructured data have many different forms and variations. One

*Department of Mathematics, University of Connecticut, 341 Mansfield Road, Storrs, CT, 06268-1009, USA. Email: guojun.gan@uconn.edu.

common form of unstructured data is text. Accessing unstructured data is clumsy. To find a given unit of data in a long text, for example, sequentially search is usually performed.

In terms of how the data are measured, data can be classified as qualitative or quantitative. Qualitative data is data about qualities, which cannot be actually measured. As a result, qualitative data is extremely varied in nature and includes interviews, documents, and artifacts (Miles et al., 2014). Quantitative data is data about quantities, which can be measured numerically with numbers. In terms of the level of measurement, quantitative data can be further classified as nominal, ordinal, interval, or ratio (Gan, 2011). Nominal data, also called categorical data, are discrete data without a natural ordering. Ordinal data are discrete data with a natural order. Interval data are continuous data with a specific order and equal intervals. Ratio data are interval data with a natural zero.

There exist a number of data sources. First, data can be obtained from university-based researchers who collect primary data. Second, data can be obtained from organizations that are set up for the purpose of releasing secondary data for general research community. Third, data can be obtained from national and regional statistical institutes that collect data. Finally, companies have corporate data that can be obtained for research purpose.

While it might be difficult to obtain data to address a specific research problem or answer a business question, it is relatively easy to obtain data to test a model or an algorithm for data analysis. In nowadays, readers can obtain datasets from the Internet easily. The following is a list of some websites to obtain real-world data:

UCI Machine Learning Repository This website (url: `http://archive.ics.uci.edu/ml/index.php`) maintains more than 400 datasets that can be used to test machine learning algorithms.

Kaggle The Kaggle website (url: `https://www.kaggle.com/`) include real-world datasets used for data science competition. Readers can download data from Kaggle by registering an account.

DrivenData DrivenData aims at bringing cutting-edge practices in data science to solve some of the world's biggest social challenges. In its website (url: `https://www.drivendata.org/`), readers can participate data science competitions and download datasets.

Analytics Vidhya This website (url: `https://datahack.analyticsvidhya.com/contest/all/`) allows you to participate and download datasets from practice problems and hackathon problems.

KDD Cup KDD Cup is the annual Data Mining and Knowledge Discovery competition organized by ACM Special Interest Group on Knowledge

Discovery and Data Mining. This website (url: `http://www.kdd.org/kdd-cup`) contains the datasets used in past KDD Cup competitions since 1997.

U.S. Government's open data This website (url: `https://www.data.gov/`) contains about 200,000 datasets covering a wide range of areas including climate, education, energy, and finance.

AWS Public Datasets In this website (url: `https://aws.amazon.com/datasets/`), Amazon provides a centralized repository of public datasets, including some huge datasets.

1.2 Data Structures and Storage

As mentioned in the previous subsection, there are structured data as well as unstructured data. Structured data are highly organized data and usually have the following tabular format:

	V_1	V_2	$\cdots$	V_d
$\mathbf{x}_1$	x_{11}	x_{12}	$\cdots$	x_{1d}
$\mathbf{x}_2$	x_{21}	x_{22}	$\cdots$	x_{2d}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$\mathbf{x}_n$	x_{n1}	x_{n2}	$\cdots$	x_{nd}

In other words, structured data can be organized into a table consists of rows and columns. Typically, each row represents a record and each column represents an attribute. A table can be decomposed into several tables that can be stored in a relational database such as the Microsoft SQL Server. The SQL (Structured Query Language) can be used to access and modify the data easily and efficiently.

Unstructured data do not follow a regular format (Abdullah and Ahmad, 2013). Examples of unstructured data include documents, videos, and audio files. Most of the data we encounter are unstructured data. In fact, the term "big data" was coined to reflect this fact. Traditional relational databases cannot meet the challenges on the varieties and scales brought by massive unstructured data nowadays. NoSQL databases have been used to store massive unstructured data.

There are three main NoSQL databases (Chen et al., 2014): key-value databases, column-oriented databases, and document-oriented databases. Key-value databases use a simple data model and store data according to key-values. Modern key-value databases have higher expandability and smaller query response time than relational databases. Examples of key-value databases include Dynamo used by Amazon and Voldemort used by LinkedIn. Column-oriented databases store and process data according to columns rather than

rows. The columns and rows are segmented in multiple nodes to achieve expandability. Examples of column-oriented databases include BigTable developed by Google and Cassandra developed by FaceBook. Document databases are designed to support more complex data forms than those stored in key-value databases. Examples of document databases include MongoDB, SimpleDB, and CouchDB. MongoDB is an open-source document-oriented database that stores documents as binary objects. SimpleDB is a distributed NoSQL database used by Amazon. CouchDB is an another open-source document-oriented database.

1.3 Data Quality

Accurate data are essential to useful data analysis. The lack of accurate data may lead to significant costs to organizations in areas such as correction activities, lost customers, missed opportunities, and incorrect decisions (Olson, 2003).

Data has quality if it satisfies its intended use, that is, the data is accurate, timely, relevant, complete, understood, and trusted (Olson, 2003). As a result, we first need to know the specification of the intended uses and then judge the suitability for those uses in order to assess the quality of the data. Unintended uses of data can arise from a variety of reasons and lead to serious problems.

Accuracy is the single most important component of high-quality data. Accurate data have the following properties (Olson, 2003):

- The data elements are not missing and have valid values.
- The values of the data elements are in the right ranges and have the right representations.

Inaccurate data arise from different sources. In particular, the following areas are common areas where inaccurate data occur:

- Initial data entry. Mistakes (including deliberate errors) and system errors can occur during the initial data entry. Flawed data entry processes can result in inaccurate data.
- Data decay. Data decay, also known as data degradation, refers to the gradual corruption of computer data due to an accumulation of non-critical failures in a storage device.
- Data moving and restructuring. Inaccurate data can also arise from data extracting, cleaning, transforming, loading, or integrating.
- Data using. Faulty reporting and lack of understanding can lead to inaccurate data.

Reverification and analysis are two approaches to find inaccurate data elements. To ensure that the data elements are 100% accurate, we must use reverification. However, reverification can be time-consuming and may not be possible for some data. Analytical techniques can also be used to identify inaccurate data elements. There are five types of analysis that can be used to identify inaccurate data (Olson, 2003): data element analysis, structural analysis, value correlation, aggregation correlation, and value inspection

Companies can create a data quality assurance program to create high-quality databases. For more information about data quality issues management and data profiling techniques, readers are referred to (Olson, 2003).

1.4 Data Cleaning

Raw data usually need to be cleaned before useful analysis can be conducted. In particular, the following areas need attention when preparing data for analysis (Janert, 2010):

Missing values It is common to have missing values in raw data. Depending on the situations, we can discard the record, discard the variable, or impute the missing values.

Outliers Raw data may contain unusual data points such as outliers. We need to handle outliers carefully. We cannot just remove outliers without knowing the reason for their existence. Sometimes the outliers are caused by clerical errors. Sometimes outliers are the effect we are looking for.

Junk Raw data may contain junks such as nonprintable characters. Junks are typically rare and not easy to get noticed. However, junks can cause serious problems in downstream applications.

Format Raw data may be formated in a way that is inconvenient for subsequent analysis. For example, components of a record may be split into multiple lines in a text file. In such cases, lines corresponding to a single record should be merged before loading to a data analysis software such as R.

Duplicate records Raw data may contain duplicate records. Duplicate records should be recognized and removed. This task may not be trivial depending on what you consider "duplicate."

Merging datasets Raw data may come from different sources. In such cases, we need to merge the data from different sources to ensure compatibility.

For more information about how to handle data in R, readers are referred to (Forte, 2015) and (Buttrey and Whitaker, 2017).

2 Data Analysis Preliminary

Data analysis involves inspecting, cleansing, transforming, and modeling data to discover useful information to suggest conclusions and make decisions. Data analysis has a long history. In 1962, statistician John Tukey defined data analysis as (Tukey, 1962):

> procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

Recently, Judd and coauthors defined data analysis as the following equation (Judd et al., 2017):

$$\text{Data} = \text{Model} + \text{Error}, \tag{1}$$

where Data represents a set of basic scores or observations to be analyzed, Model is a compact representation of the data, and Error is simply the amount the model fails to represent accurately. Using the above equation for data analysis, an analyst must resolve the following two conflicting goals:

- to add more parameters to the model so that the model represents the data better.
- to remove parameters from the model so that the model is simple and parsimonious.

In this section, we give a high-level introduction to data analysis, including different types of methods.

2.1 Data Analysis Process

Data analysis is part of an overall study. For example, Figure 1 shows the process of a typical study in behavioral and social sciences as described in (Albers, 2017). The data analysis part consists of the following steps:

Exploratory analysis The purpose of this step is to get a feel of the relationships with the data and figure out what type of analysis for the data makes sense.

Statistical analysis This step performs statistical analysis such as determining statistical significance and effect size.

Make sense of the results This step interprets the statistical results in the context of the overall study.

Determine implications This step interprets the data by connecting it to the study goals and the larger field of this study.

The goal of the data analysis as described above focuses on explaining some phenomenon (See Section 2.5).

Figure 1: The process of a typical study in behavioral and social sciences.

Shmueli (2010) described a general process for statistical modeling, which is shown in Figure 2. Depending on the goal of the analysis, the steps differ in terms of the choice of methods, criteria, data, and information.

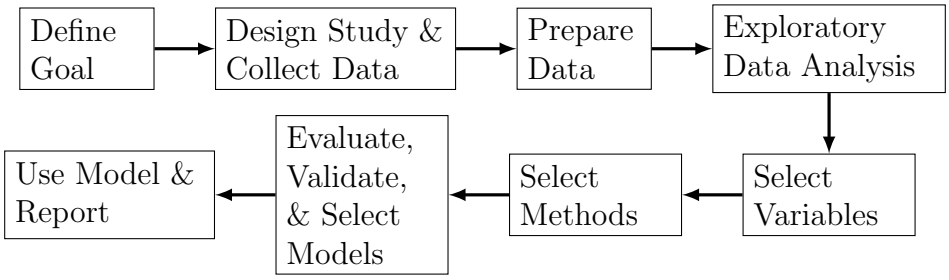

Figure 2: The process of statistical modeling.

2.2 Exploratory versus Confirmatory

There are two phases of data analysis (Good, 1983): exploratory data analysis (EDA) and confirmatory data analysis (CDA). Table 1 summarizes some differences between EDA and CDA. EDA is usually applied to observational data with the goal of looking for patterns and formulating hypotheses. In contrast, CDA is often applied to experimental data (i.e., data obtained by means of a formal design of experiments) with the goal of quantifying the extent to which discrepancies between the model and the data could be expected to occur by chance (Gelman, 2004).

Techniques for EDA include descriptive statistics (e.g., mean, median, standard deviation, quantiles), distributions, histograms, correlation analysis, dimension reduction, and cluster analysis. Techniques for CDA include the traditional statistical tools of inference, significance, and confidence.

2.3 Supervised versus Unsupervised

Methods for data analysis can be divided into two types (Abbott, 2014; Igual and Segu, 2017): supervised learning methods and unsupervised learning

Table 1: Comparison of exploratory data analysis and confirmatory data analysis.

	EDA	**CDA**
Data	Observational data	Experimental data
Goal	Pattern recognition, formulate hypotheses	Hypothesis testing, estimation, prediction
Techniques	Descriptive statistics, visualization, clustering	Traditional statistical tools of inference, significance, and confidence

methods. Supervised learning methods work with labeled data, which include a target variable. Mathematically, supervised learning methods try to approximate the following function:

$$Y = f(X_1, X_2, \ldots, X_p),$$

where Y is a target variable and X_1, X_2, ..., X_p are explanatory variables. Other terms are also used to mean a target variable. Table 2 gives a list of common names for different types of variables (Frees, 2009). When the target variable is a categorical variable, supervised learning methods are called classification methods. When the target variable is continuous, supervised learning methods are called regression methods.

Target Variable	**Explanatory Variable**
Dependent variable	Independent variable
Response	Treatment
Output	Input
Endogenous variable	Exogenous variable
Predicted variable	Predictor variable
Regressand	Regressor

Table 2: Common names of different variables.

Unsupervised learning methods work with unlabeled data, which include explanatory variables only. In other words, unsupervised learning methods do not use target variables. As a result, unsupervised learning methods are also called descriptive modeling methods.

2.4 Parametric versus Nonparametric

Methods for data analysis can be parametric or nonparametric (Abbott, 2014). Parametric methods assume that the data follow a certain distribution. Nonparametric methods do not assume distributions for the data and therefore are called distribution-free methods.

Parametric methods have the advantage that if the distribution of the data is known, properties of the data and properties of the method (e.g., errors, convergence, coefficients) can be derived. A disadvantage of parametric methods is that analysts need to spend considerable time on figuring out the distribution. For example, analysts may try different transformation methods to transform the data so that it follows a certain distribution.

Since nonparametric methods make fewer assumptions, nonparametric methods have the advantage that they are more flexible, more robust, and applicable to non-quantitative data. However, a drawback of nonparametric methods is that the conclusions drawn from nonparametric methods are not as powerful as those drawn from parametric methods.

2.5 Explanation versus Prediction

There are two goals in data analysis (Breiman, 2001; Shmueli, 2010): explanation and prediction. In some scientific areas such as economics, psychology, and environmental science, the focus of data analysis is to explain the causal relationships between the input variables and the response variable. In other scientific areas such as natural language processing and bioinformatics, the focus of data analysis is to predict what the responses are going to be given the input variables.

Shmueli (2010) discussed in detail the distinction between explanatory modeling and predictive modeling, which reflect the process of using data and methods for explaining or predicting, respectively. Explanatory modeling is commonly used for theory building and testing. However, predictive modeling is rarely used in many scientific fields as a tool for developing theory.

Explanatory modeling is typically done as follows:

- State the prevailing theory.
- State causal hypotheses, which are given in terms of theoretical constructs rather than measurable variables. A causal diagram is usually included to illustrate the hypothesized causal relationship between the theoretical constructs.
- Operationalize constructs. In this step, previous literature and theoretical justification are used to build a bridge between theoretical constructs and observable measurements.

- Collect data and build models alongside the statistical hypotheses, which are operationalized from the research hypotheses.
- Reach research conclusions and recommend policy. The statistical conclusions are converted into research conclusions. Policy recommendations are often accompanied.

Shmueli (2010) defined predictive modeling as the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations. Predictions include point predictions, interval predictions, regions, distributions, and rankings of new observations. Predictive model can be any method that produces predictions.

2.6 Data Modeling versus Algorithmic Modeling

Breiman (2001) discussed two cultures in the use of statistical modeling to reach conclusions from data: the data modeling culture and the algorithmic modeling culture. In the data modeling culture, the data are assumed to be generated by a given stochastic data model. In the algorithmic modeling culture, the data mechanism is treated as unknown and algorithmic models are used.

Data modeling gives the statistics field many successes in analyzing data and getting information about the data mechanisms. However, Breiman (2001) argued that the focus on data models in the statistical community has led to some side effects such as

- Produced irrelevant theory and questionable scientific conclusions.
- Kept statisticians from using algorithmic models that might be more suitable.
- Restricted the ability of statisticians to deal with a wide range of problems.

Algorithmic modeling was used by industrial statisticians long time ago. However, the development of algorithmic methods was taken up by a community outside statistics (Breiman, 2001). The goal of algorithmic modeling is predictive accuracy. For some complex prediction problems, data models are not suitable. These prediction problems include speech recognition, image recognition, handwriting recognition, nonlinear time series prediction, and financial market prediction. The theory in algorithmic modeling focuses on the properties of algorithms, such as convergence and predictive accuracy.

2.7 Big Data Analysis

Unlike traditional data analysis, big data analysis employs additional methods and tools that can extract information rapidly from massive data. In particular, big data analysis uses the following processing methods (Chen et al., 2014):

Bloom filter A bloom filter is a space-efficient probabilistic data structure that is used to determine whether an element belongs to a set. It has the advantages of high space efficiency and high query speed. A drawback of using bloom filter is that there is a certain misrecognition rate.

Hashing Hashing is a method that transforms data into fixed-length numerical values through a hash function. It has the advantages of rapid reading and writing. However, sound hash functions are difficult to find.

Indexing Indexing refers to a process of partitioning data in order to speed up reading. Hashing is a special case of indexing.

Tries A trie, also called digital tree, is a method to improve query efficiency by using common prefixes of character strings to reduce comparison on character strings to the greatest extent.

Parallel computing Parallel computing uses multiple computing resources to complete a computation task. Parallel computing tools include MPI (Message Passing Interface), MapReduce, and Dryad.

Big data analysis can be conducted in the following levels (Chen et al., 2014): memory-level, business intelligence (BI) level, and massive level. Memory-level analysis is conducted when the data can be loaded to the memory of a cluster of computers. Current hardware can handle hundreds of gigabytes (GB) of data in memory. BI level analysis can be conducted when the data surpass the memory level. It is common for BI level analysis products to support data over terabytes (TB). Massive level analysis is conducted when the data surpass the capabilities of products for BI level analysis. Usually Hadoop and MapReduce are used in massive level analysis.

2.8 Reproducible Analysis

As mentioned in Section 2.1, a typical data analysis workflow includes collecting data, analyzing data, and reporting results. The data collected are saved in a database or files. The data are then analyzed by one or more scripts, which may save some intermediate results or always work on the raw data. Finally a report is produced to describe the results, which include relevant plots, tables, and summaries of the data. The workflow may subject to the following potential issues (Mailund, 2017, Chapter 2):

- The data are separated from the analysis scripts.
- The documentation of the analysis is separated from the analysis itself.

If the analysis is done on the raw data with a single script, then the first issue is not a major problem. If the analysis consists of multiple scripts and a script saves intermediate results that are read by the next script, then the scripts describe a workflow of data analysis. To reproduce an analysis, the scripts have to be executed in the right order. The workflow may cause major problems if the order of the scripts is not documented or the documentation is not updated or lost. One way to address the first issue is to write the scripts so that any part of the workflow can be run completely automatically at any time.

If the documentation of the analysis is synchronized with the analysis, then the second issue is not a major problem. However, the documentation may become completely useless if the scripts are changed but the documentation is not updated.

Literate programming is an approach to address the two issues mentioned above. In literate programming, the documentation of a program and the code of the program are written together. To do literate programming in R, one way is to use the R Markdown and the `knitr` package.

2.9 Ethical Issues

Analysts may face ethical issues and dilemmas during the data analysis process. In some fields, for example, ethical issues and dilemmas include participant consent, benefits, risk, confidentiality, and data ownership (Miles et al., 2014). For data analysis in actuarial science and insurance in particular, we face the following ethical matters and issues (Miles et al., 2014):

Worthness of the project Is the project worth doing? Will the project contribute in some significant way to a domain broader than my career? If a project is only opportunistic and does not have a larger significance, then it might be pursued with less care. The result may be looked good but not right.

Competence Do I or the whole team have the expertise to carry out the project? Incompetence may lead to weakness in the analytics such as collecting large amounts of data poorly and drawing superficial conclusions.

Benefits, costs, and reciprocity Will each stakeholder gain from the project? Are the benefit and the cost equitable? A project will likely to fail if the benefit and the cost for a stakeholder do not match.

Privacy and confidentiality How do we make sure that the information is kept confidentially? Where raw data and analysis results are stored and how will have access to them should be documented in explicit confidentiality agreements.

3 Data Analysis Techniques

Techniques for data analysis are drawn from different but overlapping fields such as statistics, machine learning, pattern recognition, and data mining. Statistics is a field that addresses reliable ways of gathering data and making inferences based on them (Bandyopadhyay and Forster, 2011; Bluman, 2012). The term machine learning was coined by Samuel in 1959 (Samuel, 1959). Originally, machine learning refers to the field of study where computers have the ability to learn without being explicitly programmed. Nowadays, machine learning has evolved to the broad field of study where computational methods use experience (i.e., the past information available for analysis) to improve performance or to make accurate predictions (Bishop, 2007; Clarke et al., 2009; Mohri et al., 2012; Kubat, 2017). There are four types of machine learning algorithms (See Table 3) depending on the type of the data and the type of the learning tasks.

Table 3: Types of machine learning algorithms.

	Supervised	**Unsupervised**
Discrete Label	Classification	Clustering
Continuous Label	Regression	Dimension reduction

Originating in engineering, pattern recognition is a field that is closely related to machine learning, which grew out of computer science. In fact, pattern recognition and machine learning can be considered to be two facets of the same field (Bishop, 2007). Data mining is a field that concerns collecting, cleaning, processing, analyzing, and gaining useful insights from data (Aggarwal, 2015).

3.1 Exploratory Techniques

Exploratory data analysis techniques include descriptive statistics as well as many unsupervised learning techniques such as data clustering and principal component analysis.

3.1.1 Descriptive Statistics

In the mass noun sense, descriptive statistics is an area of statistics that concerns the collection, organization, summarization, and presentation of data (Bluman, 2012). In the count noun sense, descriptive statistics are summary statistics that quantitatively describe or summarize data.

Table 4: Some commonly used descriptive statistics.

	Descriptive Statistics
Measures of central tendency	Mean, median, mode, midrange
Measures of variation	Range, variance, standard deviation
Measures of position	Quantile

Table 4 lists some commonly used descriptive statistics. In R, we can use the function `summary` to calculate some of the descriptive statistics. For numeric data, we can visualize the descriptive statistics using a boxplot.

In addition to these quantitative descriptive statistics, we can also qualitatively describe shapes of the distributions (Bluman, 2012). For example, we can say that a distribution is positively skewed, symmetric, or negatively skewed. To visualize the distribution of a variable, we can draw a histogram.

3.1.2 Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that transforms a dataset described by possibly correlated variables into a dataset described by linearly uncorrelated variables, which are called principal components and are ordered according to their variances. PCA is a technique for dimension reduction. If the original variables are highly correlated, then the first few principal components can account for most of the variation of the original data.

To describe PCA, let $X_1, X_2, \ldots, X_d$ be a set of variables. The first principal component is defined to be the normalized linear combination of the variables that has the largest variance, that is, the first principal component is defined as

$$Z_1 = w_{11}X_1 + w_{12}X_2 + \cdots + w_{1d}X_d, \tag{2}$$

where $\mathbf{w}_1 = (w_{11}, w_{12}, \ldots, w_{1d})'$ is a vector of loadings such that $\mathrm{Var}\,(Z_1)$ is maximized subject to the following constraint:

$$\mathbf{w}_1'\mathbf{w}_1 = \sum_{j=1}^{d} w_{1j}^2 = 1. \tag{3}$$

For $i = 2, 3, \ldots, d$, the ith principal component is defined as

$$Z_i = w_{i1}X_1 + w_{i2}X_2 + \cdots + w_{id}X_d, \tag{4}$$

where $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{id})'$ is a vector of loadings such that $\text{Var}(Z_i)$ is maximized subject to the following constraints:

$$\mathbf{w}_i'\mathbf{w}_i = \sum_{j=1}^{d} w_{ij}^2 = 1, \tag{5a}$$

$$\text{Cov}(Z_i, Z_j) = 0, \quad j = 1, 2, \ldots, i-1. \tag{5b}$$

The principal components of the variables are related to the eigenvectors and eigenvectors of the covariance matrix of the variables. For $i = 1, 2, \ldots, d$, let $(\lambda_i, \mathbf{e}_i)$ be the ith eigenvalue-eigenvector pair of the covariance matrix Σ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$ and the eigenvectors are normalized. Then the ith principal component is given by

$$Z_i = \mathbf{e}_i'\mathbf{X} = \sum_{j=1}^{d} e_{ij} X_j, \tag{6}$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_d)'$. It can be shown that $\text{Var}(Z_i) = \lambda_i$. As a result, the proportion of variance explained by the ith principal component is calculated as

$$\frac{\text{Var}(Z_i)}{\sum_{j=1}^{d} \text{Var}(Z_j)} = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_d}. \tag{7}$$

For more information about PCA, readers are referred to (Mirkin, 2011).

3.1.3 Cluster Analysis

Cluster analysis (aka data clustering) refers to the process of dividing a dataset into homogeneous groups or clusters such that points in the same cluster are similar and points from different clusters are quite distinct (Gan et al., 2007; Gan, 2011). Data clustering is one of the most popular tools for exploratory data analysis and has found applications in many scientific areas.

During the past several decades, many clustering algorithms have been proposed. Among these clustering algorithms, the k-means algorithm is perhaps the most well-known algorithm due to its simplicity. To describe the k-means algorithm, let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a dataset containing n points, each of which is described by d numerical features. Given a desired number of clusters k, the k-means algorithm aims at minimizing the following objective function:

$$P(U, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il} \|\mathbf{x}_i - \mathbf{z}_l\|^2, \tag{8}$$

where $U = (u_{il})_{n \times k}$ is an $n \times k$ partition matrix, $Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k\}$ is a set of cluster centers, and $\|\cdot\|$ is the L^2 norm or Euclidean distance. The partition matrix U satisfies the following conditions:

$$u_{il} \in \{0, 1\}, \quad i = 1, 2, \ldots, n, \; l = 1, 2, \ldots, k, \tag{9a}$$

$$\sum_{l=1}^{k} u_{il} = 1, \quad i = 1, 2, \ldots, n. \tag{9b}$$

The k-means algorithm employs an iterative procedure to minimize the objective function. It repeatedly updates the partition matrix U and the cluster centers Z alternately until some stop criterion is met. When the cluster centers Z are fixed, the partition matrix U is updated as follows:

$$u_{il} = \begin{cases} 1, & \text{if } \|\mathbf{x}_i - \mathbf{z}_l\| = \min_{1 \le j \le k} \|\mathbf{x}_i - \mathbf{z}_j\|; \\ 0, & \text{if otherwise,} \end{cases} \tag{10}$$

When the partition matrix U is fixed, the cluster centers are updated as follows:

$$z_{lj} = \frac{\sum_{i=1}^{n} u_{il} x_{ij}}{\sum_{i=1}^{n} u_{il}}, \quad l = 1, 2, \ldots, k, \ j = 1, 2, \ldots, d, \tag{11}$$

where z_{lj} is the jth component of $\mathbf{z}_l$ and x_{ij} is the jth component of $\mathbf{x}_i$.

For more information about k-means, readers are referred to (Gan et al., 2007) and (Mirkin, 2011).

3.2 Confirmatory Techniques

Confirmatory data analysis techniques include the traditional statistical tools of inference, significance, and confidence.

3.2.1 Linear Models

Linear models, also called linear regression models, aim at using a linear function to approximate the relationship between the dependent variable and independent variables. A linear regression model is called a simple linear regression model if there is only one independent variable. When more than one independent variables are involved, a linear regression model is called a multiple linear regression model.

Let X and Y denote the independent and the dependent variables, respectively. For $i = 1, 2, \ldots, n$, let (x_i, y_i) be the observed values of (X, Y) in the ith case. Then the simple linear regression model is specified as follows (Frees, 2009):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{12}$$

where β_0 and β_1 are parameters and ϵ_i is a random variable representing the error for the ith case.

When there are multiple independent variables, the following multiple linear regression model is used:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \tag{13}$$

where β_0, β_1, ..., β_k are unknown parameters to be estimated.

Linear regression models usually make the following assumptions:

(a) $x_{i1}, x_{i2}, \ldots, x_{ik}$ are nonstochastic variables.

(b) $\text{Var}(y_i) = \sigma^2$, where $\text{Var}(y_i)$ denotes the variance of y_i.

(c) $y_1, y_2, \ldots, y_n$ are independent random variables.

For the purpose of obtaining tests and confidence statements with small samples, the following strong normality assumption is also made:

(d) $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are normally distributed.

3.2.2 Generalized Linear Models

The generalized linear model (GLM) is a wide family of regression models that include linear regression models as special cases. In a GLM, the mean of the response (i.e., the dependent variable) is assumed to be a function of linear combinations of the explanatory variables, i.e.,

$$\mu_i = E[y_i], \tag{14a}$$

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta} = g(\mu_i), \tag{14b}$$

where $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})'$ is a vector of regressor values, μ_i is the mean response for the ith case, and η_i is a systematic component of the GLM. The function $g(\cdot)$ is known and is called the link function. The mean response can vary by observations by allowing some parameters to change. However, the regression parameters $\boldsymbol{\beta}$ are assumed to be the same among different observations.

GLMs make the following assumptions:

(a) $x_{i1}, x_{i2}, \ldots, x_{in}$ are nonstochastic variables.

(b) $y_1, y_2, \ldots, y_n$ are independent.

(c) The dependent variable is assumed to follow a distribution from the linear exponential family.

(d) The variance of the dependent variable is not assumed to be constant but is a function of the mean, i.e.,

$$\text{Var}\,(y_i) = \phi\nu(\mu_i), \tag{15}$$

where ϕ denotes the dispersion parameter and $\nu(\cdot)$ is a function.

As we can see from the above specification, the GLM provides a unifying framework to handle different types of dependent variables, including discrete and continuous variables. For more information about GLMs, readers are referred to (de Jong and Heller, 2008) and (Frees, 2009).

3.2.3 Tree-based Models

Decision trees, also known as tree-based models, involve dividing the predictor space (i.e., the space formed by independent variables) into a number of simple regions and using the mean or the mode of the region for prediction (Breiman et al., 1984). There are two types of tree-based models: classification trees and regression trees. When the dependent variable is categorical, the resulting tree models are called classification trees. When the dependent variable is continuous, the resulting tree models are called regression trees.

The process of building classification trees is similar to that of building regression trees. Here we only briefly describe how to build a regression tree. To do that, the predictor space is divided into non-overlapping regions such that the following objective function

$$f(R_1, R_2, \ldots, R_J) = \sum_{j=1}^{J} \sum_{i=1}^{n} I_{R_j}(\mathbf{x}_i)(y_i - \mu_j)^2 \tag{16}$$

is minimized, where I is an indicator function, R_j denotes the set of indices of the observations that belong to the jth box, μ_j is the mean response of the observations in the jth box, $\mathbf{x}_i$ is the vector of predictor values for the ith observation, and y_i is the response value for the ith observation.

In terms of predictive accuracy, decision trees generally do not perform to the level of other regression and classification models. However, tree-based models may outperform linear models when the relationship between the response and the predictors is nonlinear. For more information about decision trees, readers are referred to (Breiman et al., 1984) and (Mitchell, 1997).

4 Some R Functions

R is an open-source software for statistical computing and graphics. The R software can be downloaded from the R project website at `https://www.r-project.org/`. In this section, we give some R function for data analysis, especially the data analysis tasks mentioned in previous sections.

Table 5 lists a few R functions for different data analysis tasks. Readers can read the R documentation for examples of using these functions. There are also other R functions from other packages to do similar things. However, the functions listed in this table provide good start points for readers to conduct data analysis in R. For analyzing large datasets in R in an efficient way, readers are referred to (Daroczi, 2015).

Table 5: Some R functions for data analysis.

Data Analysis Task	R package	R Function
Descriptive Statistics	`base`	`summary`
Principal Component Analysis	`stats`	`prcomp`
Data Clustering	`stats`	`kmeans`, `hclust`
Fitting Distributions	`MASS`	`fitdistr`
Linear Regression Models	`stats`	`lm`
Generalized Linear Models	`stats`	`glm`
Regression Trees	`rpart`	`rpart`
Survival Analysis	`survival`	`survfit`

5 Summary

In this chapter, we gave a high-level overview of data analysis. The overview is divided into three major parts: data, data analysis, and data analysis techniques. In the first part, we introduced data types, data structures, data storages, and data sources. In particular, we provided several websites where readers can obtain real-world datasets to horn their data analysis skills. In the second part, we introduced the process of data analysis and various aspects of data analysis. In the third part, we introduced some commonly used techniques for data analysis. In addition, we listed some R packages and functions that can be used to perform various data analysis tasks.

References

Abbott, D. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. Wiley, Hoboken, NJ.

Abdullah, M. F. and Ahmad, K. (2013). The mapping process of unstructured data to structured data. In *2013 International Conference on Research and Innovation in Information Systems (ICRIIS)*, pages 151–155.

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer, New York, NY.

Albers, M. J. (2017). *Introduction to Quantitative Data Analysis in the Behavioral and Social Sciences*. John Wiley & Sons, Inc., Hoboken, NJ.

Bandyopadhyay, P. S. and Forster, M. R., editors (2011). *Philosophy of Statistics*. Handbook of the Philosophy of Science 7. North Holland.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer, New York, NY.

Bluman, A. (2012). *Elementary Statistics: A Step By Step Approach*. McGraw-Hill, New York, NY.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC, Raton Boca, FL.

Buttrey, S. E. and Whitaker, L. R. (2017). *A Data Scientists Guide to Acquiring, Cleaning, and Managing Data in R*. Wiley, Hoboken, NJ.

Chen, M., Mao, S., Zhang, Y., and Leung, V. C. (2014). *Big Data: Related Technologies, Challenges and Future Prospects*. Springer, New York, NY.

Clarke, B., Fokoue, E., and Zhang, H. H. (2009). *Principles and theory for data mining and machine learning*. Springer-Verlag, New York, NY.

Daroczi, G. (2015). *Mastering Data Analysis with R*. Packt Publishing, Birmingham, UK.

de Jong, P. and Heller, G. Z. (2008). *Generalized linear models for insurance data*. Cambridge University Press, Cambridge, UK.

Forte, R. M. (2015). *Mastering Predictive Analytics with R*. Packt Publishing, Birmingham, UK.

Frees, E. W. (2009). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press.

Gan, G. (2011). *Data Clustering in C++: An Object-Oriented Approach*. Data Mining and Knowledge Discovery Series. Chapman & Hall/CRC Press, Boca Raton, FL, USA.

Gan, G., Ma, C., and Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. SIAM Press, Philadelphia, PA.

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, 13(4):755–779.

Good, I. J. (1983). The philosophy of exploratory data analysis. *Philosophy of Science*, 50(2):283–295.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., and Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47:98 – 115.

Hox, J. J. and Boeije, H. R. (2005). Data collection, primary versus secondary. In *Encyclopedia of social measurement*, pages 593 – 599. Elsevier.

Igual, L. and Segu, S. (2017). *Introduction to Data Science. A Python Approach to Concepts, Techniques and Applications*. Springer, New York, NY.

Inmon, W. and Linstedt, D. (2014). *Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault*. Morgan Kaufmann, Cambridge, MA.

Janert, P. K. (2010). *Data Analysis with Open Source Tools*. O'Reilly Media, Sebastopol, CA.

Judd, C. M., McClelland, G. H., and Ryan, C. S. (2017). *Data Analysis. A Model Comparison Approach to Regression, ANOVA and beyond*. Routledge, New York, NY, 3rd edition.

Kubat, M. (2017). *An Introduction to Machine Learning*. Springer, New York, NY, 2nd edition.

Mailund, T. (2017). *Beginning Data Science in R: Data Analysis, Visualization, and Modelling for the Data Scientist*. Apress.

Miles, M., Hberman, M., and Sdana, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook*. Sage, Thousand Oaks, CA, 3rd edition.

Mirkin, B. (2011). *Core Concepts in Data Analysis: Summarization, Correlation and Visualization*. Springer, London, UK.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT Press, Cambridge, MA.

O'Leary, D. E. (2013). Artificial intelligence and big data. *IEEE Intelligent Systems*, 28(2):96–99.

Olson, J. E. (2003). *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, San Francisco, CA.

Pries, K. H. and Dunnigan, R. (2015). *Big Data Analytics: A Practical Guide for Managers*. CRC Press, Boca Raton, FL.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67.