A Survey of Statistical Modeling Tools

Madhuri Kulkarni (A survey paper written under the guidance of Prof. Raj Jain)

Abstract:
A plethora of statistical modeling tools are available in the market providing various features. Some of them are comprehensive while others suit a specific area of technology. The need for proper selection of tools is essential for proper modeling and analysis. This paper discusses the overview of modeling and the popular statistical tools used for modeling like R, Minitab, SPSS, Matlab and Mathematica.

Keywords: Statistical Modeling Tools, SPSS, R, Matlab, Minitab, Mathematica, statistics.

Table of Contents

1. Introduction
2. Overview of Modeling
3. Modeling Tools
   3.1 Generic Tools
      3.1.1 S-PLUS
      3.1.2 R
      3.1.3 SPSS
      3.1.4 Minitab
   3.2 Domain Specific Tools
      3.2.1 MedCalc
      3.2.2 Partek
      3.2.3 Primer-E
   3.3 Statistical Programming Tools
      3.3.1 Matlab
      3.3.2 Mathematica
4. Summary
References
List of Acronyms

1. Introduction

In simple terms a model is a relationship between variables. Models are often designed to estimate the behavior of the system based on the past performance of that system. It also helps estimate the probable errors that could be expected in the system.

Various tools are used today in the industry to model the behavior of their systems. Numerous modeling tools are available in the market. Some are free, open sourced while other popular tools are built to suit a specific area of interest.

Section 1 gives a brief introduction followed by section 2 which gives an overview of modeling. Section 3 gives an idea about the popular modeling tools. Finally, Section 4 describes the other modeling tools available.

2. Overview of Modeling

The main aim of modeling is to study or understand a system or process to predict its future behavior. Thus with the availability of observed data of a system, models can help infer various alternatives for the system.

Data is first collected and observed for patterns. Various factors, parameters are analyzed and then the choice of appropriate model is crucial. A model is developed depending on this analysis. Regression analysis is the most commonly used. Predictive and Uplift analysis are the newer practices observed today in the market. These methods help better predict the system and improve performance of the system. Hence the tools to execute this analysis are of great importance.

The next section describes the popular tools used in the market today, since listing of all the tools available is difficult given the time and space for this survey.

3. Modeling Tools

Following are some of the popular tools used today. They have been classified as generic, non generic and language based tools. Generic tools include S-PLUS, R, Minitab, SPSS (Statistical Package for the Social Sciences). Domain specific tools include Medcalc, Primer-e, Partek where as the statistical programming tools include the Matlab and Mathematica tools.
3.1 Generic Tools

Modeling tools that provide various analysis and modeling functions that can be used across domains can be termed as generic tools. Generic tools are popular and widely used in the industry. They provide user friendly interface and standard modeling and analysis features. Few of these tools like R, are open source and are available under the GNU license.

3.1.1 S-PLUS

S-PLUS is a much evolved version of S language which was developed at the lucent technologies [formerly Bell labs]. Object oriented S language was specially built to support various data analysis and modeling functions such as data visualization, data exploration, statistical modeling and data programming. S has since then been developed and enhanced to support a comprehensive library of statistical functions. Users can take advantage of this robust architecture to fit alternative models to find the best model suitable for your system.

S-PLUS support a wide range of such as microarray analysis, financial econometrics, large-scale constrained optimization, Clinical trial design and analysis, analysis of spatial data, wavelet and signal series analysis, integration module for, environmental statistics. Fig 1 shows the graphics created in S language for prediction of soybean future contracts. S-PLUS also provides an IDE [Integrated Development Environment], project management tools as well as validation, debugging and profiling tools.[ S PLUS]

S-PLUS has a Windows based and a Unix based version. Based on S language, R is another language which is open source and is discussed next.

![S-PLUS graphlet](S-PLUS fig)

Figure 1: S-PLUS single page graphlet with labels associated with each of the data points. SOURCE: [ S-PLUS fig]

3.1.2 R

R is an object oriented, free, open source software environment for statistical analysis and graphics. It is based on its predecessor language S and hence is sometimes termed as GNU S and has been freely available under the GNU license.

Although a GUI [Graphical User Interface] is available for R, it mainly uses the command line interface. R supports traditional statistical tests, time series analysis, linear & non linear modeling, classification, clustering. Since R is based on the S it has a comprehensive library of functions to support the above functions.[ R]

![R graphlet](R-Fig)

Figure 2: (C) R Foundation. Unix version of R Source: [ R-Fig]

Its different versions run on Windows, Mac OS X, Linux and Unix like systems. Fig 2 shows the snapshot of the software on Unix desktop. The next statistical package is SPSS which is widely used in the industry.

3.1.3 SPSS
SPSS [Statistical Package for the Social Sciences] was developed in the late 1960s by students of Stanford University. 37 years later it has been one of the widely used softwares in the industry. SPSS includes a variety of tools statistics and data mining, modeling, data collection, text mining and deployment.

SPSS provides automated functionality for various features thus making it convenient for users. It provides tools for entire life cycle of the statistical process from data preparation to , modeling, analysis, report generation and deployment. SPSS is windows based and support various versions of windows OS such as Xp, Vista and Windows 2003. Fig 3 shows the different graphs supported by SPSS. Next statistical software is Minitab which is one of the popular softwares.

![Different graphs in SPSS](image)

**3.1.4 Minitab**

Minitab contains easy to use features for both novice as well as expert users. It provides functionality for statistical process control, regression analysis, analysis of variance, design of experiments, data and file management, measurement systems analysis, graphs and graphs editing, quality tools, multivariate analysis, simulations and various other functions. It is also used to implement six sigma for the users process.

This concludes the generic tools available for statistical modeling and analysis that can be used across technologies. Other tools in this category include SAS (Statistical Analysis Software), Stata, JMP and Statistica. The next sections describes tools that are used in specific areas to make accurate decisions.

**3.2 Domain Specific Tools**

Domain based modeling tools are built to benefit a specific area of technology so that the decision makers can make decision with highest level of precision. These domains suffer wide loses if they are subjected to the smallest of errors. Due to the large number of such industries and nearly equal amount of products floating in the market, it is impossible to list all of them. Hence only a few below are discussed.

**3.1.1 Medcalc**

MedCalc is a statistical package for biomedical researchers. For comparison studies it supports functionality for Bland & Altman plot, Passing and Bablok and Deming regression. It also provides Receiver Operating Characteristic curve analysis.

Medcalc is a windows based and runs on most windows versions like xp, vista, windows 2000 and windows server 2003. Fig4 shows a mountain plot of two continuous variables projected by Medcalc.

![Mountain plot](image)

**3.1.2 Partek**

Partek is another statistical software build for genome researchers and is discussed below.

**3.1.2 Partek**

Partek is a statistical package for genomics. It supports statistics, data mining and visualization tools to identify correlations between various chemical and biological activities. It has been developed to support genomic studies with high dimensions. It also provides facilities to import the data from leading chip platforms for analysis and processing.
Primer-e is statistical software for ecologists and is discussed next.

### 3.1.3 Primer-e

Primer-e is a multivariate statistical package for ecologists. It supports complex and wide range of functionalities necessary for the research. Fig 5 shows some of the functionalities supported by Primer-e.

The following section is based on statistical softwares that allow us to build our own package using the language and library of functions provided by them.

![Figure 5: Ordination and visualization of fitted models. Source](PE)

### 3.3 Statistical Programming Tools

Matlab and Mathematica are few of those packages that help users program their own statistical software according to the needs of the systems or process they are working on. A sound knowledge of the language and the library functions is necessary for building statistical software.

#### 3.3.1 Matlab

Matlab provides functionalities for statistical data analysis and modeling. Functions include modeling, simulation, developing statistical algorithms, analyzing trends and developing multi dimensional non linear models.

Since these functions are written in Matlab language which is open source, users can inspect, or edit exiting functions or include new ones to enhance the functionality to suit its needs.

![Figure 6: Non classical multi dimensional scaling. Source](Mat Fig)

Matlab versions run on Windows, Linux, MAC OS X and Solaris. Fig 6 shows one of the functionalities supported by matlab software. Mathematica is another statistical package discussed below.

#### 3.3.2 Mathematica

Mathematica provides various functionalities such as analysis of variance, hypothesis testing, confidence interval estimation, data smoothing, univariate & multivariate statistics, linear & nonlinear regression and optimization techniques. It consists of a symbolic engine that let users create additional functions, commands using its programming environment which help solve isolated problems.

### 4. Summary

<table>
<thead>
<tr>
<th>Table 1: Summary of Modeling Tools</th>
<th></th>
</tr>
</thead>
</table>

---

---
<table>
<thead>
<tr>
<th>Modeling tool</th>
<th>Type</th>
<th>Free</th>
<th>OS</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-PLUS</td>
<td>Generic</td>
<td>No</td>
<td>Windows, Unix</td>
<td>Life Sciences, Finance, Marketing Analytics, Manufacturing, Telecommunications, Academia, Government</td>
</tr>
<tr>
<td>R</td>
<td>Generic</td>
<td>Yes</td>
<td>Windows, Mac OS X, Linux, Unix like systems</td>
<td>Using CRAN: Bayesian, Environmetrics, Finance, Genetics, Machine Learning, Social Sciences, Spatial, Time Series, etc.</td>
</tr>
<tr>
<td>SPSS</td>
<td>Generic</td>
<td>No</td>
<td>Windows XP, Vista, Windows 2003</td>
<td>Pharmaceutical, government, business operations, software etc</td>
</tr>
<tr>
<td>Minitab</td>
<td>Generic</td>
<td>No</td>
<td>Microsoft Windows 2000, XP, or Vista</td>
<td>Multi domain for: Six sigma, CMMI (Capability Maturity Model Integration), educational etc</td>
</tr>
<tr>
<td>MedCalc</td>
<td>Domain Specific</td>
<td>No</td>
<td>Microsoft Windows 2000, XP, or Vista, Windows 2003</td>
<td>Medical</td>
</tr>
<tr>
<td>Partek</td>
<td>Domain Specific</td>
<td>No</td>
<td>Windows, Linux, Macintosh</td>
<td>Genomics</td>
</tr>
<tr>
<td>Primer-e</td>
<td>Domain Specific</td>
<td>No</td>
<td>Windows</td>
<td>Environmental, Biodiversity</td>
</tr>
<tr>
<td>Matlab</td>
<td>Generic, Programming support</td>
<td>No</td>
<td>Linux, Mac OS X, Solaris, Windows, Windows</td>
<td>Multi domain</td>
</tr>
<tr>
<td>Mathematica</td>
<td>Generic, Programming support</td>
<td>No</td>
<td>Windows, Mac OS X, Linux, Solaris</td>
<td>Multi domain</td>
</tr>
</tbody>
</table>

Due to the plethora of tools available in the market a good research is necessary to buy the most beneficial software also taking into consideration the cost factor, and the number of user since license costs vary accordingly. The generic tools provide industry wide solutions by providing a wide range of functionality but at a considerably high cost. The domain specific tools are less expensive but may not always be available for all domains. The programming statistical tools are a good solution if users are not happy with the above statistical tools. Sound knowledge of these softwares and investment of time are necessary for proper implementation.

References

[R-Fig] [4] (C) R Foundation, from http://www.r-project.org
[CRAN] [16] http://cran.r-project.org/web/views/index.html

List of Acronyms

SPSS Statistical Package for the Social Sciences
GUI Graphical User Interface
IDE Integrated Development Environment
OS Operating Systems
SAS Statistical Analysis Software
CRAN Comprehensive R Archive Network
CMMI Capability Maturity Model Integration