2013 IEEE 10th International Conference on e-Business Engineering

SLinRA^{2}S: A Simple Linear Regression Analysis Assisting System

Chien-Ho Wu, Jung-Bin Li
Dept. of Statistics and Information Science
Fu Jen Catholic University
New Taipei City, TAIWAN
052845@mail.fju.edu.tw, 071635@mail.fju.edu.tw

Tsair-Yuan Chang
Dept. of Information Management
Ming Chuan University
Taoyuan County, TAIWAN
tychang@mail.mcu.edu.tw

Abstract—**Acquiring quality Information is critical to decision makings. In general the production of information for a particular decision scenario involves a process of analyzing data from various sources using some statistical methods. The proper application of the chosen statistical methods for analysis in turn relates, in a large portion, to the quality of the information generated for the decision scenario. To ensure a consistent and sound application of statistical methods to data analysis we followed the idea of active support and designed a tentative data analysis assistor, SLinRA^{2}S, which can guide a data analyst through the process of applying simple linear regression on data sets from external files or databases. SLinRA^{2}S is implemented in Java and it invokes R for statistical functions. The assistor not only promises the relief of a data analyst from computation errands but also contributes to the correct application of statistical methods.**

Keywords-linear regression; R; Java; active support

I. Introduction

The survival and prosperity of a business relies very much on the effectiveness of its decision making process, which requires prompt availability of quality business intelligence [11]. It is inevitable that the production of decision information very often involves selecting proper statistical methods for data analysis and ensuring the selected methods are properly conducted.

It is no argument that learning of Statistics is never an easy experience lest the proper application of statistic methods to analyzing data for decision support. In general the selection of statistical methods for data analysis relies basically on three factors:

- The nature of the decision problem at hand.
- The properties of data related to the decision problem.
- Assumptions of statistical methods.

The constraint of computational resources is not much of a problem due to the high performance/cost of ICT and availability of statistical software. From the perspective of providing high quality business intelligence for decision support, the issue of applying correctly the chosen statistical method to data analysis is of our major concern.

Most statistical methods are subject to some assumptions such as normality distribution of samples, equal variances of error terms, independence of error terms or free from multicollinearity et cetera [1, 2, 3, 4]. An experienced data analyst will test the assumptions for the chosen method and take appropriate remedies to correct possible biases. As such, the test of assumptions further complicates the applications of statistical methods.

In this research, we have proposed a systems approach to address the aforementioned difficulties in applying statistical methods to data analysis. In particular, we have developed a tentative system that can guide a data analyst through the process of applying simple linear regression technique. The tentative system was then evaluated for validity against outputs in SPSS. In a way the proposed systems approach can also ensure, to a certain degree, the quality of decision information generated from statistical data analysis.

II. Brief Review of Regression Model

Regression technology is one of the forecasting methods in Statistics. In regression technology the variable that is being predicted is called the dependent variable. The variables being used to predict the value of the dependent variable are called the independent variables.

The equation that describes how the dependent variable y is related to the independent variables $x_1,x_2,\ldots,x_n$ and an error term ϵ is called the regression model [1, 2, 4, 12]. The regression model has the following general form:

$$\mathrm{y}=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \epsilon \qquad (1)$$

The random variable ϵ accounts for the variability in y which cannot be explained by the linear effect of the n independent variables. While a simple linear regression model involves one independent variable, a multiple regression model is a regression model with two or more independent variables.

A. Assumptions about Model Error Term ϵ

The tests of significance in regression analysis are based on the following assumptions about the model error term:

- The error term ϵ is a normally distributed random variable. As a result the dependent variable y is also a normally distributed random variable.
- $\mathrm{E}(\epsilon)=0$, which implies that for given values of $x_1,x_2,\ldots,x_n$, the expected value of y is given by

$$\mathrm{E(y)}=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n \qquad (2)$$

978-0-7695-5111-1/13 $26.00 © 2013 IEEE
DOI 10.1109/ICEBE.2013.33

- The variance of ϵ, denoted σ^2, is the same for all values of x.
- The values of ϵ are independent.

B. Validation of Model Assumptions

An analysis of residuals will help determine whether the assumptions made about the regression model are appropriate. If the assumptions about the error term ϵ are questionable, the hypothesis tests about the significance of the regression relationship and the interval estimation results may not be valid.

Besides tests of normality, tests of independence and tests of constant variance, much of residual analysis is based on an examination of graphical plots. Plots of interest may include residual plot against $\hat{y}$, residual plot against x, normal Q–Q plot of residuals et cetera.

C. Outliers and Influential Observations

Outliers represent observations that are suspect and warrant careful examination. They may represent erroneous data; they may signal a violation of model assumptions; they may simply unusual values that should be retained. Influential observations are those that are valid and have high leverages on the estimated regression model.

Scatter diagrams and box plots of observations can help detect the existence of outliers or influential observations.

D. Procedure for Linear Regression Analysis

Regression analysis is an iterative process and involves some degree of experience and judgment. It may not be practical to prescribe a standard procedure for doing regression analysis. However, after the objective related to a particular regression analysis is defined, the analysis can follow the steps listed below.

1) *Detect missing values and outliers.*
   Replace or remove missing values and outliers as appropriate.
2) *Examine correlations between the dependent variable and independent variables.*
   Also look for possible multicollinearity between independent variables. In general, multicollinearity can cause problems if the absolute value of sample correlation coefficient exceeds 0.7 for any two of the independent variables [1, 2].
3) *Select an initial set of model building variables.*
4) *Choose a variable-selection procedure.*
   The four widely acknowledged procedures are stepwise procedure, forward-selection procedure, backward-elimination procedure and best-subsets procedure. The first three procedures select the regression model by adding or deleting independent variables one at a time. The best-subsets procedure aims at finding the best regression model given a specified number of independent variables.
5) *Detect influential observations.*
   Observations that have a strong influence on the resulting regress model are referred to as influential observations. Influential observations can be identified from a scatter plot when only one independent variable is present. For multiple regression models, Leverage and Cook's distance are two measures that can be used to identify influential observations [1, 2].
   Existence of influential observations suggests more intermediate values of independent variables are required to better understand the relationship between dependent variable and independent variables.
6) *Evaluate model significance.*
   First determine the overall model significance with F-test. Then if the F-test shows an overall significance, use t–test to determine whether each of the individual independent variable is significant.
7) *Test model assumptions.*
   A careful analysis of residuals should be conducted. The residual plot for the chosen model shall resemble approximately a horizontal band. Besides residual plots, Durbin-Watson test [6] can be used to test whether or not the error terms are independent. Kolmogorov–Smirnov test or Shapiro test can test the normality of residuals [8,9]. Equality of variances for residuals can be tested by a statistical procedure called the Levene test [7].
8) *Make the final choice.*
   Be sure that the functional form of the model is appropriate. The chosen model should be significant, parsimonious confoming to model assumptions and can serve the objective of the analysis.

Qualitative variables in regression analysis are represented by dummy variables. If a qualitative variable has *k* levels, *k-1* dummy variables are required, with each dummy variable being coded as 0 or 1 [1, 2, 4].

III. System Design

Simple linear regression technique is fundamental to the understanding of other more complicated regression models. To demonstrate the aforementioned analysis process, we have developed an assisting system for simple linear regression analysis. The system is implemented in the Java programming language and it invokes R functions for plotting and statistical computations.

A. The Use Case Diagram

Fig. 1 is the use case diagram [10] of SLinRA^{2}S. For the moment, SLinRA^{2}S provides users with three major functions, i.e. plotting function, correlation function and model building function. The main purpose of the plotting and correlation functions is to help users explore the nature of the data set. The model building function guides users through the process of building a simple linear regression model.

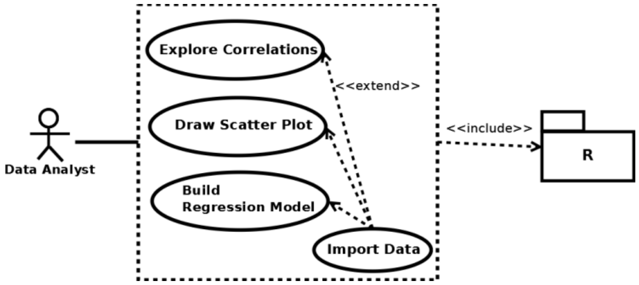


Figure 1. Use case diagram of SLinRA^{2}S.

B. The System Architecture

As shown in Fig. 2, SLinRA^{2}S consists of three control modules: interface module, database module and computation module.

The interface module interacts with and guides an analyst through the model building process. Essential explanations regarding the model built, such as implications of test results, are shown as appropriate through the interface module.

The database module connects to and accesses data stored in MySQL database. The database module assumes for now that the data to be fed to the computation module are preprocessed and the maintenance of the data is the responsibility of users.

The interface module may require functions of the database module to retrieve the data set in database. The computation module activates an R engine through the rJava/jri middleware [5]. The activated R engine may read data in database through the RODBC package. The data read, through either the database module or the RODBC package, are then submitted to R functions for model building.

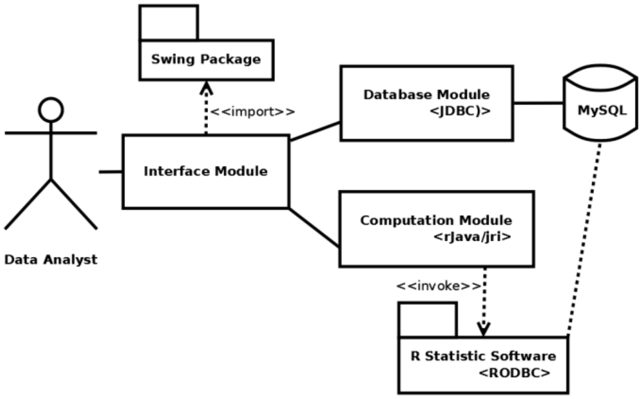


Figure 2. System architecture of SLinRA^{2}S.

C. The Activity Flow Diagram

The activity flow of the system is shown is Fig. 3 In general, the flow resembles the procedure laid out in section II. A user shall be clear about the objective of the analysis at hand. Make sure that linear regression analysis is the intended analysis method. Then, with the help of SLinRA^{2}S, read in the data set and examine the characteristics of the data set. If, for example, outliers, null values or multicollinearity are present, take corrective actions before proceeding to the model building stage.

After making sure that the data set is ready for analysis, submit the data set to the computation module for model building. SLinRA^{2}S can display on the interface the results of the model built and provides convenient designs to test model assumptions. Preliminary explanations on the test results for the model assumptions are also given. Users can then determine possible actions to follow if any of the model assumptions does not hold.

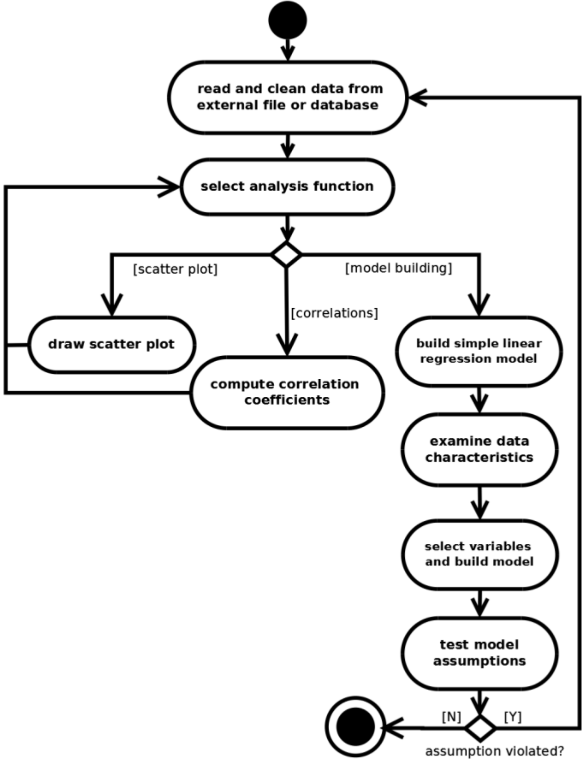


Figure 3. The activity flow of SLinRA^{2}S.

IV. SYTEM IMPLEMENTATION AND EVALUATION

SLinRA^{2}S is not complete in any way. It is developed to test our idea of systematically supporting the process of data analysis.

A. System Requirements

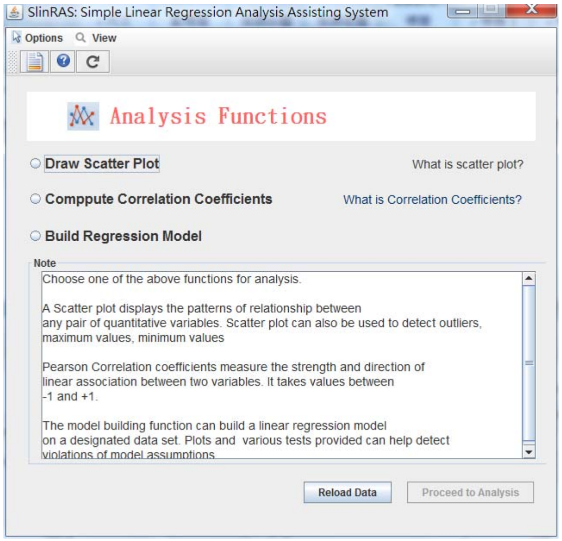
SLinRA^{2}S is implemented in Java with NetBeans IDE. Any platform conforming to the following requirements can run SLinRA^{2}S.

- Java runtime environment (SE. 1.6v or later)
- R statistical software (rJava/jri, RODBC incl.)
- ODBC manager,
- MySQL(5.0v or later) with JDBC connector and MySQL ODBC driver

SLinRA^{2}S does not include any proprietary code in its implementation. It has been tested in the Windows environment for convenience of evaluation against SPSS outputs.

B. Sample Screen Captures

Some screen captures of SLinRA^{2}S are shown from Fig. 4 to Fig. 9.


SlinRAS: Simple Linear Regression Analysis Assisting System
Options
View
Analysis Functions
○ Draw Scatter Plot
What is scatter plot?
○ Comppute Correlation Coefficients
What is Correlation Coefficients?
○ Build Regression Model
Note
Reload Data
Proceed to Analysis

Choose one of the above functions for analysis.

A Scatter plot displays the patterns of relationship between
any pair of quantitative variables. Scatter plot can also be used to detect outliers,
maximum values, minimum values

Pearson Correlation coefficients measure the strength and direction of
linear association between two variables. It takes values between
-1 and +1.

The model building function can build a linear regression model
on a designated data set. Plots and various tests provided can help detect
violations of model assumptions

Figure 4. SLinRA^{2}S analysis functions.

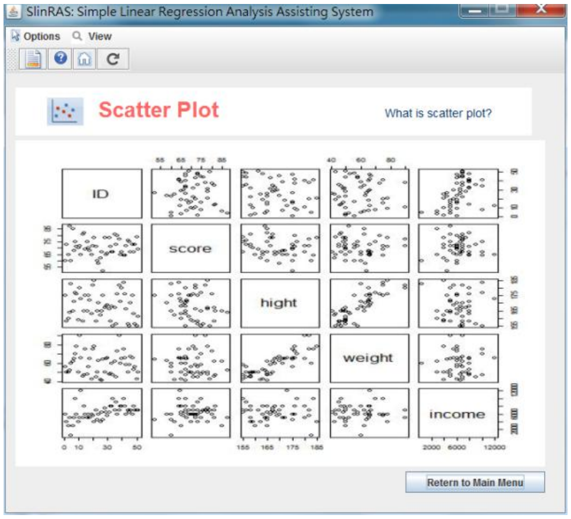
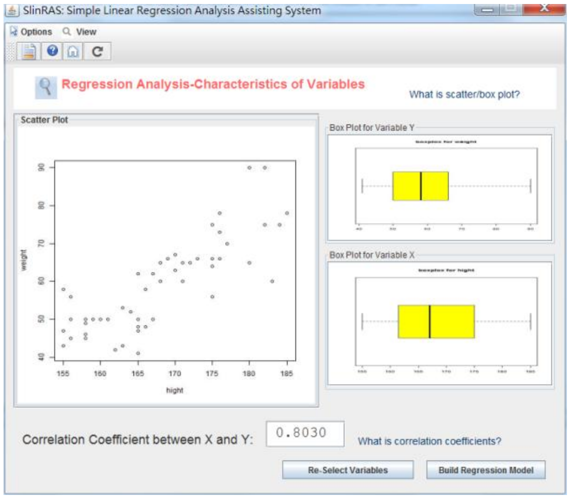
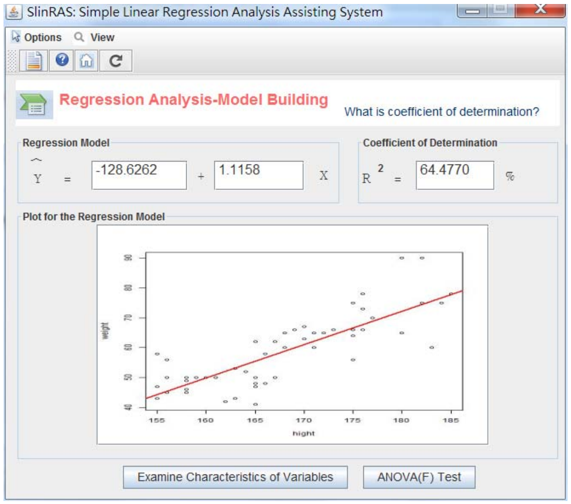


Figure 5. Scatter plot for sample data.


SlinRAS: Simple Linear Regression Analysis Assisting System
Options
View
Regression Analysis-Characteristics of Variables
What is scatter/box plot?
Scatter Plot
Box Plot for Variable Y
Box Plot for Variable X
weight
hight
Correlation Coefficient between X and Y:
What is correlation coefficients?
Re-Select Variables
Build Regression Model

Correlation Coefficient between X and Y: 0.8030

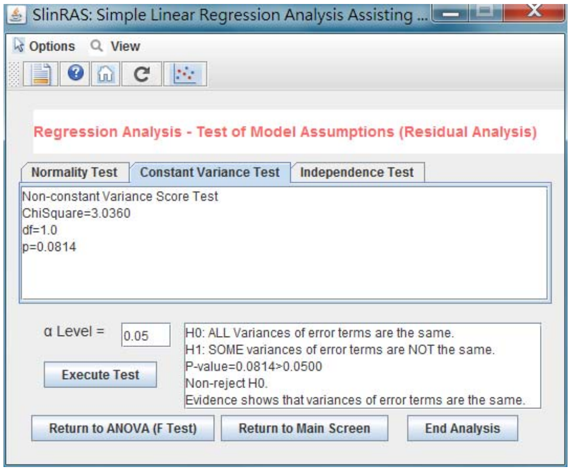
Figure 6. Box plots and correlations.


SlinRAS: Simple Linear Regression Analysis Assisting System
Options
View
Regression Analysis-Model Building
What is coefficient of determination?
Regression Model
Coefficient of Determination
Plot for the Regression Model
weight
hight
Examine Characteristics of Variables
ANOVA(F) Test

$\hat{Y}$ = -128.6262 + 1.1158 X

R^2 = 64.4770 %

Figure 7. Simple linear regression model.



```
Non-constant Variance Score Test
ChiSquare=3.0360
df=1.0
p=0.0814
```

α Level = 0.05

```
H0: ALL Variances of error terms are the same.
H1: SOME variances of error terms are NOT the same.
P-value=0.0814>0.0500
Non-reject H0.
Evidence shows that variances of error terms are the same.
```

Figure 8. Model assumption test for constant variance.

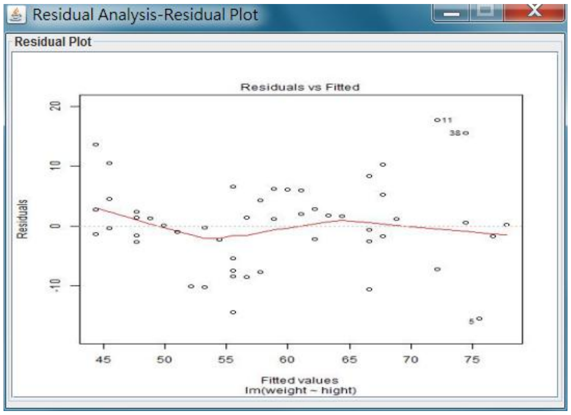
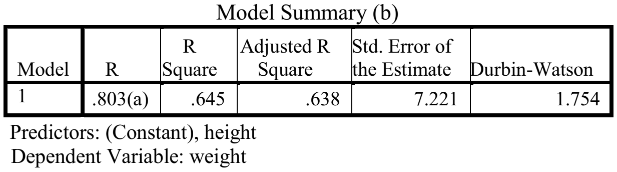


Figure 9. Residual plot.

C. *Evaluation against SPSS*

SPSS is one of the most widely used commercial statistical software available on the market. SPSS comes with an abundant of statistical and plotting functions.

However users rely very much on their own knowledge and experience to exploit those features. Fig. 10 gives an example output of a linear regression analysis in SPSS. Users must be equipped with due understandings of Statistics to interpret the meaning of the output. In general, while SPSS provides little explanation on the outputs, SLinRA^{2}S tries to display information that can help users interpret those outputs.



Model Summary (b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.803(a)	.645	.638	7.221	1.754

a Predictors: (Constant), height

b Dependent Variable: weight

Figure 10. An example output from SPSS

We believe that providing users with some degree of preliminary explanation on the outputs would prevent users from misinterpreting the abstruse yet important statistics.

V. Final Remarks

The provision of decision information inevitably involves some sort of statistical analysis on data. As such, the correct application of the chosen statistical analysis technique is crucial to the quality of the information generated from the analysis process. In this research we have developed an assisting system, i.e. SLinRA^{2}S, to test and evaluate our idea of guiding a data analyst through the process of simple linear regression analysis.

The assisting system not only imposes the steps for simple regression analysis but also provides preliminary interpretations of some important statistics. Furthermore it provides context-based hyper-links to related pages that can help users of the system understand important concepts related to regression analysis. SLinRA^{2}S is convenient to users and promises an active support of the analysis process.

Since SLinRA^{2}S invokes R for regression functions and is implemented in Java, it can be easily ported to various computing platforms. The experience gained from the implementation of SLinRA^{2}S also concurs that the post processing of outputs returned by an R Engine dictates much of programming efforts.

In the immediate future, this research can be extended in the following directions:

1) *Extension of analysis support.* Based on the implementation results and experiences from this research, this research can be extended to cover the support of multiple regression analysis and other statistical methods.
2) Provision of more intensive and contex-based documentation. By making an assisting system more *active* and *intelligent* in aiding the analysis process, we would have a better chance of ensuring the quality of decision information generated from the process.
3) *Dynamic building of regression models.* The contents in the original data set may be outdated or new data may be added to the orginal data set. It is often the case that the data set used for model building is time-variant and thus is not static. As such, there is a need to build revised models when data set is updated.

In a way, this research can be considered complementary to existing commercial statistical software.

Acknowledgment

Thanks to Yi-Hsin Li and Hsuan-Yu Chen for participating in the development of SLinRA^{2}S.

References

[1] D. R. Anderson, D. J. Sweeney, and T. A. Williams, Statistics for Business and Economics, 11th ed., OH: South-Western Cengage Learning, 2011.

[2] M. Kutner, C. Nachtsheim, and J. Neter, Applied Linear Regression Models, 4th ed., Europe: McGraw-Hill, 2004.

[3] P. Hoel, Introduction to Mathematical Statistics, 5th ed., NY: John Wiley & Sons, 1985.

[4] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to Linear Regression Analysis, 5th ed. NY: Wiley, 2012.

[5] W. N. Venables, D. M. Smith, and the R Core Team, An Introduction to R, The R Project, 2013.

[6] J. Durbin and G. S. Watson, "Testing for serial correlation in least squares regression. III". Biometrika, vol. 58, no. 1, Apr. 1971, pp. 1–19.

[7] H. Levene, "Robust Tests for Equality of Variances," in Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin et al. eds., CA: Stanford University Press, 1960, pp. 278–292.

[8] H. Lilliefors, "On the Kolmogorov–Smirnov test for normality with mean and variance unknown," Journal of the American Statistical Association, vol. 62, no. 318, Jun. 1967, pp. 399–402.

[9] S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," Biometrika, vol. 52, no. 3 and 4, Dec. 1965, pp. 591–611.

[10] G. Booch, J. Rumbaugh, and I. Jacobson, The Unified Modeling Language User Guide, 2th ed., IN: Addison-Wesley, 2005.

[11] C.H. Wu, J.B. Li, and T.Y. Chang, "Implementing Statistical Agents on JADE Platform," Applied Mathematics & Information Sciences, vol. 6. no. 2S, 2012, pp. 379S–385S.

[12] R. F. Nau, Lecture Notes on Forcasting, Duke University: The Fuqua School of Business, 2005.
(http://people.duke.edu/~rnau/411home.htm)