Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 5, Issue, 03, pp.439-444, March, 2013

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

DEVELOPMENT OF PREDICTIVE REGRESSION MODEL FOR STUDY OF THERAPEUTIC PROPERTIES OF CHEMICAL COMPOUNDS

***,[1]Patle, M. R. and [2]Katre M. G.**

[1]Department of Chemistry, D. B. Science College, Gondia
[2]Department of Computer Science, D. B. Science College, Gondia

ARTICLE INFO

Article History:
Received 29th December, 2012
Received in revised form
16th January, 2013
Accepted 23rd February, 2013
Published online 19th March, 2013

Key words:
Regression analysis,
Therapeutic properties,
Data mining,
VISUAL BASIC,

ABSTRACT

Regression analysis is an important tool in the computer based drug design practitioner's toolbox for a number of reasons. First, this method saves lot of time and by using this model, lot of compound can be screened for their therapeutic properties. Secondly, varieties of regression analysis methods are available depending on the nature of the problem being studied. In current project, a regression model "Regression Analysis Software Package (RASTP)" is developed as a measure for Study of Therapeutic Properties of Chemical Compounds. This model will compare the unknown molecule with the set of known molecules with respect to their structural properties and select or reject the given molecule or set of molecule on the basis of correlation and Regression Coefficient parameters. In this study, the known set of molecules, which are anticancer compounds with known biological activity, will be retrieved from the databank. These set of molecules will act as training set during the model building and will also used for validation purpose of the model. From the data obtained from model, it is clear that the important regression analysis parameters for the predictor such as coefficient estimate, standard error; mean R squared error; adjusted R square, etc. are in good accord with the respective parameters for the known set. Hence, "Regression Analysis Software Package (RASTP)" model is a good tool for the calculation of biological activity which in turn is used to predict the therapeutic importance of the chemical compound.

INTRODUCTION

In current project, we promote the use of regression model as a measure for Study of therapeutic Properties of Chemical Compounds. We then propose an efficient algorithm for tuning a regression model to further increase its efficiency. In contrast with previous statistical methods, which are customized to particular functions, this algorithm can deal with any functions without modifying the underlying regression methods. We have evaluated the algorithm by using the known data set. Our results show that the proposed algorithm significantly gives the values which are in good agreement. The asks of regression are at the heart of data mining. Collectively, what we are doing here can be termed in data mining as Predictive Analytics (Sridhar *et al.*, ? and David Hand *et al.*, 2001). Most of what we learn from a traditional data mining course focuses on the algorithms from machine learning and statistics that build regression models. These models can then be used to evaluate new entities. The actual structure of the model also gives us insight into the relationships between the variables that are important in differentiating the classes (Apte, 1997 and Prabhu, ?). This model will compare the unknown molecule with the set of known molecules with respect to their structural properties and select or reject the given molecule or set of molecule on the basis of correlation and Regression Coefficient parameters (Mogull, Robert 2004; Yule Udny, 1897; Fisher, 1922).

In current study, the known set of molecules, which are anticancer compounds with known biological activity, will be retrieved from the databank. These set of molecules will act as training set during the model building and will also used for validation purpose of the model. In statistics, regression analysis includes any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables.

Corresponding author: manojpatle14@gmail.com

More specifically, regression analysis helps us understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables -that is, the average value of the dependent variable when the independent variables are held fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function, which can be described by a probability distribution (Mogull, Robert 2004; Francis Galton 1885; Yule, Udny 1897; Fisher, 1922 Ronald; Fisher, 1954; Aldrich, John 2005; David Freedman, 2005; Dennis Cook, 1982; Anon, 2000). Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. (Anon, 2000)

METHODOLOGY

This study used the "Regression Analysis Software Package (RASTP)" to calculate biological activity of chemical compounds on the basis of topological descriptors by using regression analysis methodology. A brief overview of the aspects of the methodology specific to this study is discussed below and further details are presented.

The first step was to divide the compounds into the three sets

- the training set,
- cross-validation set,
- Prediction set.

This resulted in a training set containing 63 compounds, a cross-validation set containing 15 compounds and a prediction set containing 9 compounds. Next, this descriptor pool was reduced using objective feature selection. Setting correlation and identical cutoffs, this procedure resulted in a reduced pool of descriptors. This descriptor pool was used to generate the linear model. After objective feature selection, predictive models were generated by using a simulated annealing or genetic algorithm to search the descriptor space for optimal subsets of descriptors. The optimization routines were coupled with a multiple linear regression routine to find the best predictive models.

Model Development

Once the descriptors are selected, reduced the original pool to a more manageable size can then proceed to build a set of models and choose the best one. The methodology for model development involves following steps. First a set of regression models are developed using descriptor subsets selected by using matrix calculation method. The best model is selected based on R^2 and RMSE value. Once a number of descriptor subsets have been obtained, the final architecture is obtained as described above. The result of this procedure is to create a set of models to investigate different aspects of the structure-property relationship and to understand the trends present in the dataset as well as provide good predictive ability for new observations.

Model Building

Overview of the research approach

The aim of the current project is to develop a model "Regression Analysis Software Package (RASTP)" for Study of therapeutic Properties of Chemical Compounds on the basis of regression analysis methodology. To achieve the objectives of the project, programming technology are used. The model consists of various modules to achieve the concerned task. Each module is defined according to the tasks which are being researched with an emphasis on the assistive technology. The Fig. 2.2 shows flowchart for showing the steps involved in predicting molecular properties or activities from molecular structure. And Table 2.1 shows the overall process for "Regression Analysis Software Package (RASTP)" model.

Model Building parameters

Visual Basic 6.0

VISUAL BASIC (VB) is a high level programming language which evolved from the earlier DOS version called BASIC. BASIC means Beginners' All-purpose Symbolic Instruction Code. It is a very easy programming language to learn. The code looks a lot like English Language. Different software companies produced different versions of BASIC, such as Microsoft QBASIC, QUICKBASIC, GWBASIC, IBM BASICA and so on. However, people prefer to use Microsoft Visual Basic today, as it is a well developed programming language and supporting resources are available everywhere. Now, there are many versions of VB exist in the market, the most popular one and still widely used by many VB programmers is none other than

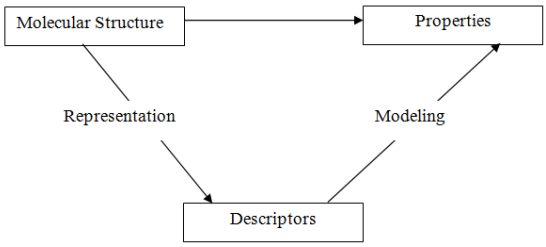
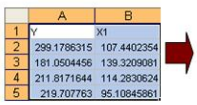
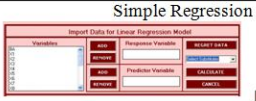
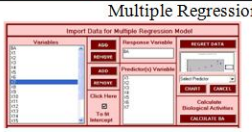
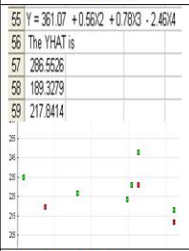


Fig. 2.1: A flowchart showing the steps involved in predicting molecular properties or activities from molecular structure

Table 2.1: Overview of the research approach

Data	Processing	Result
	Simple Regression  Multiple Regression 	
Data Source: Excel Sheet	Regression Analysis: Selection of response Variables, predictor variables and intercept if any.	Regretted Data & Charts

Prediction, Validation and Interpretation

After a model has been developed the next step is to investigate its predictive ability. The simplest method is to test the model on a subset of the dataset that has not been used during the model development process (the prediction set). The statistics obtained from the results of the prediction set can give us some indication of the model's predictive ability. The most common statistics for linear models are R^2 and RMSE, though the former is not always a very reliable indicator of the goodness of fit.

Visual Basic 6. We also have VB.net, VB2005 and the latest VB2008, which is a fully object oriented programming (OOP) language. It is more powerful than VB6 but looks more complicated to master (Evangelos Petroutsos, ?). VISUAL BASIC is a VISUAL and events driven Programming Language. These are the main divergence from the old BASIC. In BASIC, programming is done in a text-only environment and the program is executed sequentially. In VB, programming is done in a graphical environment. In the old BASIC, you have to write program code for each graphical object you wish to display it on screen, including its position and its color. However, In

VB, you just need to drag and drop any graphical object anywhere on the form, and you can change its color any time using the properties windows. [16] On the other hand, because the user may click on certain object randomly, so each object has to be programmed independently to be able to response to those actions (events). Therefore, a VB Program is made up of many subprograms, each has its own program code, and each can be executed independently and at the same time each can be linked together in one way or another. (Evangelos Petroutsos, ? and Peter Norton's, ?)

shows the selected data for known compounds which acts as a training set for the current model.

Step III: Predictive Model

The total number of data (i.e. the total number of rows) used in a regression (named as N)
Number of Data Points: 9
N=9

3.1 Step I: Data Selection

Table 3.1 Parameters of Known chemical compound

Values	BA	Symbol						
		GGI1	GGI2	GGI3	GGI4	GGI5	GGI6	GGI7
		X1	X2	X3	X4	X5	X6	X7
4a	3.81	4	4	3.847	1.917	1.421	0.825	0.612
5a	3.2	3.5	2.889	3.153	1.494	1.008	0.55	0.411
5b	3.9	4	2.667	3.153	1.449	1.279	0.469	0.308
5c	4	3	1.778	1.757	0.836	0.549	0.31	0.128
5d	4	3	1.778	1.917	0.957	0.563	0.395	0.15
5e	3.3	3.5	1.778	1.639	0.907	0.723	0.367	0.209
5f	3.6	3.5	2.667	3.09	1.441	1.286	0.641	0.352
5g	2.9	2.5	2	1.917	1.042	0.653	0.279	0.035
5h	3.4	3	2.444	2.354	1.623	1.049	0.687	0.347

On the basis of above data selection the "Regression Analysis Software Package (RASTP)" gives the output as given below

3.2 Step II: Variable Selection in Which Response Variables, Predictor Variables and Intercept

Table 3.2 The "Y" vector (dependent variable) and "X" matrix (constant and independent variables)

BA	Intercept	X1	X2	X3	X4	X5	X6	X7
3.81	1	4	4	3.847	1.917	1.421	0.825	0.612
3.2	1	3.5	2.889	3.153	1.494	1.008	0.55	0.411
3.9	1	4	2.667	3.153	1.449	1.279	0.469	0.308
4	1	3	1.778	1.757	0.836	0.549	0.31	0.128
4	1	3	1.778	1.917	0.957	0.563	0.395	0.15
3.3	1	3.5	1.778	1.639	0.907	0.723	0.367	0.209
3.6	1	3.5	2.667	3.09	1.441	1.286	0.641	0.352
2.9	1	2.5	2	1.917	1.042	0.653	0.279	0.035
3.4	1	3	2.444	2.354	1.623	1.049	0.687	0.347

Procedure for building the regression model

An entirely VB 6.0 based approach is only the way to build a regression utility. If the entire programming is written to the visual basic and some built in functionality of visual basic 6.0 are employed, large data sets can be calculated rather quickly. The purpose of the following section is to illustrate visual basic programming methods by building a regression model completely on the visual basic environment.

In order to calculate a regression model on the worksheet, the following elements are needed.

1. Data cleansed of missing values, and assembled from range reference; (Known and unknown parameters of chemical compounds)
2. A column of 1's added to the data matrix if an intercept is specified; and
3. Various matrices calculations.
4. A Chart-space technology for designing a Scattered Graphs.

Data Analysis and Interpretation

The "Regression Analysis Software Package (RASTP)" methodology involves several steps. The first step is to select data of calculate molecular structure descriptors for the dataset. The next step involved variable selection in which response variables, predictor variables and intercept are included. After objective feature selection, in step three, predictive models were generated by using regression analysis methodology. The optimization routines were coupled with multiple linear regression routine to find the best predictive models. Table 3.1

The total number of data (i.e. the total number of column except the vector "Y") used in a regression (named P)
Number of coefficients to estimate: 8
P=8

The degree of freedom (DF) is calculated by subtracting the Number of data points (N) and total number of coefficients to estimates (P).
DF=N-P = 9-8 =1 Number of DEGREE OF FREEDOM: 1

The "Y" vector (dependent variable) ("BA" is named as "Y") and its deviations from mean (Y-AVG(Y)) and squared deviation from mean (Y-AVG(Y) ^ 2). Total Average of Y = 3.56777777777778

Table 3.3 The "Y" vector (dependent variable) and its deviations from mean (Y-AVG(Y)) and squared deviation from mean (Y-AVG(Y) ^ 2)

Y	Y-AVG(Y)	Y-AVG(Y) ^ 2
3.81	0.242222	0.058672
3.2	-0.367778	0.13526
3.9	0.332222	0.110372
4	0.432222	0.186816
4	0.432222	0.186816
3.3	-0.267778	0.071705
3.6	0.032222	0.001038
2.9	-0.667778	0.445927
3.4	-0.167778	0.028149

From above beta hat the equation for regression model is as follows.
BA = -4.04 + 2.18X1 + 0.09X2 + 0.84X3 + 0.17X4 - 3.67X5 + 8.82X6 - 11.18X7 The Sum of Squares for Regression Model is: 0.900158768668191

Table 3.4 Shows the selected data for unknown compounds

Table 3.4 Parameters of Unknown chemical compound

Values	X1	X2	X3	X4	X5	X6	X7
mol a	3.5	2	1.693	0.712	0.702	0.488	0.195
mol b	3	2.222	1.639	0.877	0.681	0.332	0.222
mol c	3	2	1.451	0.356	0.368	0.275	0.151
mol d	3.5	2.222	1.701	0.846	0.424	0.275	0.214
mol e	5	2.667	1.882	1.131	0.479	0.316	0.276
mol f	3	2.222	1.639	0.877	0.681	0.332	0.222
mol g	3.5	2.222	1.514	0.712	0.594	0.51	0.311
mol h	3.5	2.667	2.326	1.228	1.105	0.78	0.315
mol i	5	3.5556	2.826	1.771	1.285	0.819	0.484

Parameters of Unknown Chemical Compound The Regression Equation is: BA = -4.04 + 2.18X1 + 0.09X2 + 0.84X3 + 0.17X4 - 3.67X5 + 8.82X6 - 11.18X7 On putting the values of X1, X2, X3, X4, X5, X6, and X7 in the above regression equation we get a Calculated Biological Activities. Here is the Output for Calculated Biological Activities for Unknown Parameters of Known chemical compound

Table 3.5: Output for Calculated Biological Activities

Calculated Biological Activities	X1	X2	X3	X4	X5	X6	X7
4.86	3.5	2	1.693	0.712	0.702	0.488	0.195
2.17	3	2.222	1.639	0.877	0.681	0.332	0.222
3.35	3	2	1.451	0.356	0.368	0.275	0.151
3.84	3.5	2.222	1.701	0.846	0.424	0.275	0.214
6.82	5	2.667	1.882	1.131	0.479	0.316	0.276
2.17	3	2.222	1.639	0.877	0.681	0.332	0.222
4.02	3.5	2.222	1.514	0.712	0.594	0.51	0.311
5.3	3.5	2.667	2.326	1.228	1.105	0.78	0.315
6.95	5	3.5556	2.826	1.771	1.285	0.819	0.484

Table 4.1: Calculated Biological activity for Unknown compounds

Calculated Biological Activities	X1	X2	X3	X4	X5	X6	X7
4.86	3.5	2	1.693	0.712	0.702	0.488	0.195
2.17	3	2.222	1.639	0.877	0.681	0.332	0.222
3.35	3	2	1.451	0.356	0.368	0.275	0.151
3.84	3.5	2.222	1.701	0.846	0.424	0.275	0.214
6.82	5	2.667	1.882	1.131	0.479	0.316	0.276
2.17	3	2.222	1.639	0.877	0.681	0.332	0.222
4.02	3.5	2.222	1.514	0.712	0.594	0.51	0.311
5.3	3.5	2.667	2.326	1.228	1.105	0.78	0.315
6.95	5	3.5556	2.826	1.771	1.285	0.819	0.484

Table 4.2: Analysis of Variance for Known Parameters

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Regression Model	0.9	7	0.129
Error	0.325	1	0.325
Total	1.225	8	

Table 4.3: Analysis of Variance for Unknown Parameters

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Regression Model	24.885	7	3.555
Error	9.85E-23 = 0	1	9.85E-23 = 0
Total	24.885	8	

Table 4.4: Regression Statistics for Known Parameters

Observations	9
Variance	1.225
Population Variance	0.136
Sample Variance	0.153
Sample Average of Squared ERRORS	0.036
R Squared Error	0.735
Mean R Squared Error	0.325
Adjusted R Squared	-1.12
Standard Error of Estimate	0.57

Table 4.5: Regression Statistics for Unknown Parameters

Observations	9
Variance	24.885
Population Variance	2.765
Sample Variance	3.111
Sample Average of Squared ERRORS	1.1E-23 = 0
R Squared Error	1
Mean R Squared Error	9.85E-23 = 0
Adjusted R Squared	1
Standard Error of Estimate	9.93E-12 = 0

The Fig. 3.1, 3.2, 3.3, 3.4, 3.5, 3.6 and 3.7 shows that the XY Scatter graph for variable X1, X2, X3, X4, X5, X6 and X7

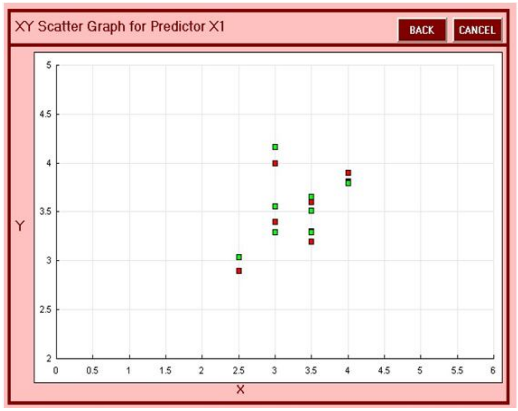


Fig. 3.1: XY Scatter graph for predictor X1

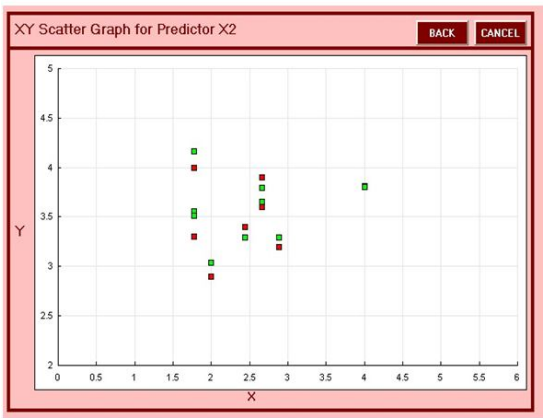


Fig. 3.2: XY Scatter graph for predictor X2

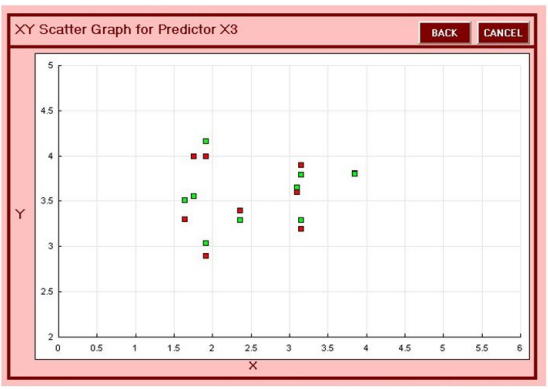


Fig. 3.3: XY Scatter graph for predictor X3

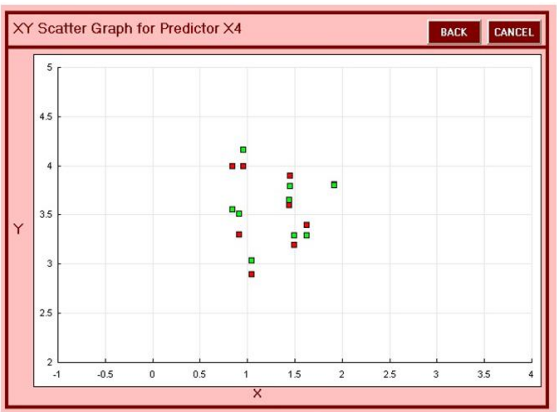


Fig. 3.4: XY Scatter graph for predictor X4

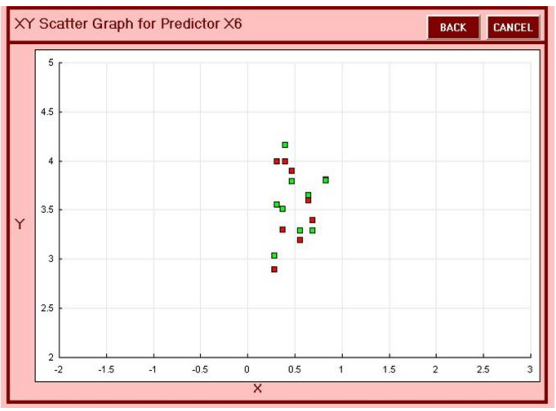


Fig. 3.6: XY Scatter graph for predictor X6

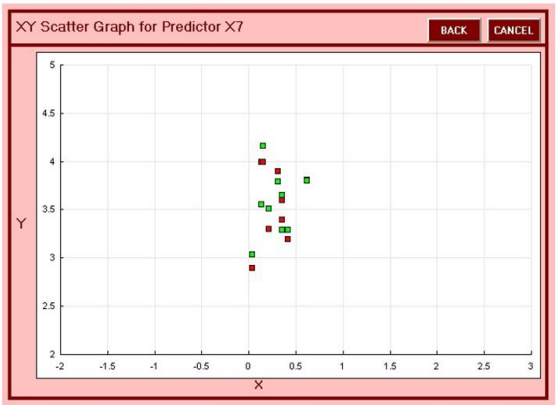


Fig. 3.7: XY Scatter graph for predictor X7

Conclusion

This study involves the development of a set of linear models to predict and interpret the biological activity of a set of unknown compounds. The dataset consist of compounds with the biological activity. However, this study is restricted to linear regression models using topological descriptors only. The current model presented concentrate on the biological activity values. The wide varieties of descriptors are used rather than restricting to any single class for the prediction model. From the regression analysis by using the "Regression Analysis Software Package (RASTP)" model, the biological activity of unknown compound is calculated which is given in Table 4.1 on the basis of their topological descriptors.

Statistics for Regression Analysis

The calculated biological activity for unknown compound obtained by using the current model is with good confidence limit. This fact is evident from the following regression analysis parameters which are in good agreement for both training set and for unknown compound. From above data, it is clear that the important regression analysis parameters for the predictor such as coefficient estimate, standard error; mean R squared error; adjusted R square, etc. are in good accord with the respective parameters for the known set. Hence, "Regression Analysis Software Package (RASTP)" model is a good tool for the calculation of biological activity which in turn is used to predict the therapeutic importance of the chemical compound.

REFERENCES

Aldrich, John (2005). "Fisher and Regression". Statistical Science 20 (4): 401-417. doi:10.1214/088342305000000331

Anon 2000 Regression Analysis http://www.en.wikipedia. org/wiki/ Regression_analysis

Apte, C. Data Mining: An Industrial Research Perspective. IEEE Computational Science and Engineering. v 4, 1997.

David A. Freedman, Statistical Models: Theory and Practice, Cambridge University Press (2005)

David Hand, Heikki Mannila and Padhraic Smyth, Principles of Data Mining, MIT Press, 2001.

Dennis Cook R.; Sanford Weisberg Criticism and Influence Analysis in Regression, Sociological Methodology, Vol. 13. (1982), pp. 313-361

Evangelos Petroutsos, Maastering Vsual Baseic 6.0, BPB, Publication

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., From Data Mining to Knowledge Discovery in Databases. AI Magazine 17(3): Fall 1996, 37-54

Fisher, R.A. (1922). "The goodness of fit of regression formulae, and the distribution of regression coefficients". J. Royal Statist. Soc. 85: 597–612. doi:10.2307/2341124.

Francis Galton. Presidential address, Section H, Anthropology. (1885) (Galton uses the term "regression" in this paper, which discusses the height of humans.)

Mogull, Robert G. (2004). Second-Semester Applied Statistics. Kendall/Hunt Publishing Company. pp. 59. ISBN 0-7575-1181-3.

Peter Norton's, Guide to Visual basic 6.0, SAMS Tec Media.

Prabhu C. S. R., Data Warehousing – Concepts, Techniques, Products and Applications, New Delhi, Prentice-Hall of India Pvt. Ltd., ISBN-81-203-2068-9

Ronald A. Fisher (1954). Statistical Methods for Research Workers (Twelfth ed.). Oliver and Boyd.

Sridhar and Dunham Margaret H., Data Mining, Introduction and Advanced Topics, Pearson Education ISBN 81-7758-785-4

Yule, G. Udny (1897). "On the Theory of Correlation". J. Royal Statist. Soc.: 812–54.

****