International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 4, Issue 6 (Nov-Dec 2016), PP. 47-52

Logistic Regression Analysis as a Future Predictor

[1]Jay Torasakar, [2]Rakesh Prabhu, [3]Pranay Rambade, [4]Manoj Kumar Shukla
Department of Computer Engineering, Atharva College of Engineering , Mumbai, India
[1]jaytorasakar8@gmail.com

***Abstract*—This paper presents a basic idea on Logistic Regression Analysis as a future predictor system. Logistic Regression is a mathematical model based approach which is widely applied in various important field like Data Mining, Prediction Mining, Machine Learning, etc. The main objective of this paper is to give a brief explanation about logistic model in regression analysis. It describes all the necessary and basic steps involved in determination of possible outcome using available past data ie. Training Data. This concept can be implemented easily to any real world applications where consideration of future outcome plays a vital role and some of them are listed viz. placement prediction of engineering students, predicting the cancer symptoms of patient, possibilities of obesity in childhood, etc.**

***Keywords*—Logistic Regression, Logit Function, Regression Analysis, Machine Learning, Data Mining, Prediction, Accuracy.**

I. INTRODUCTION

This paper is aimed to provide a basic but complete introduction about logistic regression. Going through this paper, reader will be able to understand and answer such tricky questions like "What is logistic regression?", "Why it is better than linear model?" and one of the most important one, "How it is predicting future outcome?"

Many researchers and mathematician has shown interest in calculating the logistic model for prediction of future outcome. It is done in order to determination of any future event or more precisely one should say that it does the predictions of events whether the event will succeed or fail, some of the examples of such application are whether the patient for a critical disease will survive till its cure or not, whether the student passes the university exam or not, whether students gets placed in campus placements or not, whether the system crashes or not on extra memory/user load, etc. Actually speaking, at least somewhere, the whole world is trying to foresee the future. So the scope of the logistic is very large and wide.

Logistic model is also very handy for developing supervised machine learning models where it is trained with the huge set of data specially when the data set is not linear and highly unordered. The outcome of logistic is in binary form, and hence is called as Dependent Variable (DV). The input applied in this model is nothing but the training data which is known as Independent Variable (IV).

The outcome of the logistic model is always a number and is probability indicating the chances of occurrence of an event. We always get the predicted number ranging between 0 and 1.

II. WORKING OF LOGISTIC REGRESSION

A. Methodology

Logistic regression is a statistical method used for analysing a training dataset. In this, there are one or more independent variables. These independent variables are denoted by Xb , where value of b ranges from 0 to (N-1), that is a total of N predictors, which determine the final prediction. The final result is measured in a variable which will have only two possible outcomes.

In logistic regression, the dependent variable is binary and the two values it exhibits are 0 and 1. The binary values we use are have a specific meaning. The value 1 represents Success while 0 represents Failure.

The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (β), standard errors (€) and significance levels of a formula to predict a logit transformation(π) of the probability(p).

Logit (p) $= \pi = \beta 0 + \beta 1\ X1 + \beta 2\ X2 + \beta 3\ X3 + \ldots\ldots\ldots + \beta 0XN\text{-}1$. [1]

Where "p" is probability of success.

The logit transformation is given using logged odds, which is described as: [1]

Odds = p/1-p = (Probability Success)/ (Probability Failure)

Logit (p) = π = log (p/1-p).

1) Dependent variable(Y)

The variable whose values we want to predict is called as dependent variable, with the condition that it must be binary. So it means that, dependent variable should only contain values coded as 0 or 1.

In our Example we will consider Student Placement Outcome in terms of whether a student would be placed or not. Here 0 represents "Not placed" while 1 represents "Placed". [2]

2) Independent variables(X)

Independent variables are the variables which influence the dependent variables. The independent variables are also called as Predictors which are responsible for final prediction [3].

In our example we will consider student details such as SSC percentage, HSC percentage, SGPA average of six

semesters and Aptitude marks in terms of rating or average out of 10.

B. Overall Model Fit

The null model -2 Log Likelihood is given by -2 * ln(L0) where L0 is the likelihood of obtaining the observations if the independent variables had no effect on the outcome.

The full model -2 Log Likelihood is given by -2 * ln(L) where L is the likelihood of obtaining the observations with all independent variables incorporated in the model.

The difference of these two yields a Chi-Squared statistic which is a measure of how well the independent variables affect the outcome or dependent variable.

If the P-value for the overall model fit statistic is less than the conventional 0.05 then there is evidence that at least one of the independent variables contributes to the prediction of the outcome.

Cox & Snell R2 and Nagelkerke R2 are other goodness of fit measures known as pseudo R-squareds. Note that Cox & Snell's pseudo R-squared has a maximum value that is not 1. Nagelkerke R2 adjusts Cox & Snell's so that the range of possible values extends to 1. [8]

C. Regression Coefficients

The logistic regression coefficients are the coefficients b0, b1, b2,... bk of the regression equation:

Logit (p) = π = β0 + β1 X1 + β2 X2 + β3 X3 +…….. + β0XN-1[8]

An independent variable with a regression coefficient not significantly different from 0 (P>0.05) can be removed from the regression model. If P<0.05 then the variable contributes significantly to the prediction of the outcome variable.

The logistic regression coefficients show the change (increase when β>0, decrease when β<0) in the predicted logged odds of having the characteristic of interest for a one-unit change in the independent variables.

The Wald statistic is the regression coefficient divided by its standard error squared: (β/SE)2.[8]

D. Odds Ratio with 95% Confidence Level.

By taking the exponential of both sides of the regression equation as given above, the equation can be rewritten as:

Odds= p/1-p = eβ0X0 * eβ1X1 * eβ2X2 *…….. eβN-1XN-1

It is clear that when a variable Xi increases by 1 unit, with all other factors remaining unchanged, then the odds will increase by a factor eβIXI.

This factor eβ is the odds ratio (O.R.) for the independent variable Xi and it gives the relative amount by which the odds of the outcome increase (O.R. greater than 1) or decrease (O.R. less than 1) when the value of the independent variable is increased by 1 units.

III. Our Own Case Study

The data used in our case study is a placement data taken from 20 arbitrary students who got placement in there campus interview. It consists of the marks obtained by students in secondary (SSC) and higher secondary school (HSC) examinations, there six semester SGPA average and aptitude score. The gathered data are analyzed and used to train the logistic regression model. The Table (I) gives the list of all variables and descriptions that explains their data types and dependency.

Table :-(I) List of Predictors used as Training Data

Name	Range	Type	ID/DP
SSC marks	0-100	Numeric/Continuous	ID
HSC marks	0-100	Numeric/Continuous	ID
SGPA AVG	0-100	Numeric/Continuous	ID
Aptitude score	0-100	Numeric/Continuous	ID

ID means Independent Variable. DP means Dependent Variable. The attributes selected as predictor variables Ranges from 0-100 but while applying on Logistic Model will treat it as Rating or average out of 10. Eg: SSC percentage 67% considered as 6.7.

The logistic regression equation is:

Logit (p) = π = β_0 + β_1SSC + β_2HSC + β_3SGPA + β_4Aptitude. [7]

$$p = \frac{1}{1 + e^{-logit(p)}}$$

Equation-:- (1)

The sample data given below in Diagram (II) is used to trained the Logistic Regression Model.

βi is Regression Coefficients used in calculation of Odds Ratio, Which increased the corresponding Predictor Variables value by βi percentage.

Logit (p) is the regression model/equation obtained from Training Data. The Graph between logit(p) and Probability (p) generates a Sigmoid curve (S-shape). This Sigmoid curve bounded between 0—1 and gives infinite probabilistic values ranges from 0 to 1. The Equation (1) represent sigmoid curve.

Diagram (II) Sample Data as Training Data for Logistic Regression Model.

NO	SSC Marks(X_0)	HSC Marks(X_1)	SGPA Average(X_2)	Aptitude Score(X_3)	Placement Outcome(Y)
1	9	7.9	7.9	7.5	1
2	9.4	8.3	7.9	9	1
3	8.5	7	7.9	9.5	1
4	8	6	6.8	7	0
5	8	7.5	7.5	7.6	1
6	7	6.7	8.6	6	0
7	7.8	6.5	7.8	4	0
8	8	7	7.8	7	0
9	9	8.4	9	7	1
10	7	6.4	7.5	7.4	0
11	8.5	7.5	7.5	8	1
12	8.56	7.3	7.556	3	0
13	8.5	6.4	5.8	4	0
14	9	8.6	9.33	9	1
15	8.3	7	8.4	8.5	1
16	7.5	6.4	6.4	6	0
17	7.54	6	6.57	5.5	0
18	7.6	6.2	7.5	6	0
19	8.67	6.8	8.45	8.5	1
20	8.9	7.2	7.9	8	1

Enter the Number Of Data Point :- 20
Enter the Number of Predictor Variables:-4
Enter the Confidence Level :- 95%

10 cases have Y=0; 10 cases have Y=1.

Variable	Average	Standard Deviation
1	8.2385	0.6670
2	7.0550	0.7645
3	7.7253	0.8350
4	7.0250	1.5614

Iteration History...

-2 Log Likelihood = 27.7259 (Null Model)
-2 Log Likelihood = 24.6059
-2 Log Likelihood = 19.7363
-2 Log Likelihood = 14.7663
-2 Log Likelihood = 10.5070
-2 Log Likelihood = 7.1298
-2 Log Likelihood = 4.5833
-2 Log Likelihood = 2.7626
-2 Log Likelihood = 1.5255
-2 Log Likelihood = 0.7306
-2 Log Likelihood = 0.2917
-2 Log Likelihood = 0.1103
-2 Log Likelihood = 0.0412
-2 Log Likelihood = 0.0153
-2 Log Likelihood = 0.0057
-2 Log Likelihood = 0.0021
-2 Log Likelihood = 0.0008
-2 Log Likelihood = 0.0003
-2 Log Likelihood = 0.0001
-2 Log Likelihood = 0.0000
-2 Log Likelihood = 0.0000
-2 Log Likelihood = 0.0000
-2 Log Likelihood = 0.0000
-2 Log Likelihood = 0.0000
-2 Log Likelihood = 0.0000
-2 Log Likelihood = 0.0000
-2 Log Likelihood = 0.0000 (Converged)

Overall Model Fit...
Chi Square= 27.7259;
Df =4; p= 0.0000

Coefficients, Standard Errors, Odds Ratios, and 95% Confidence Limits...

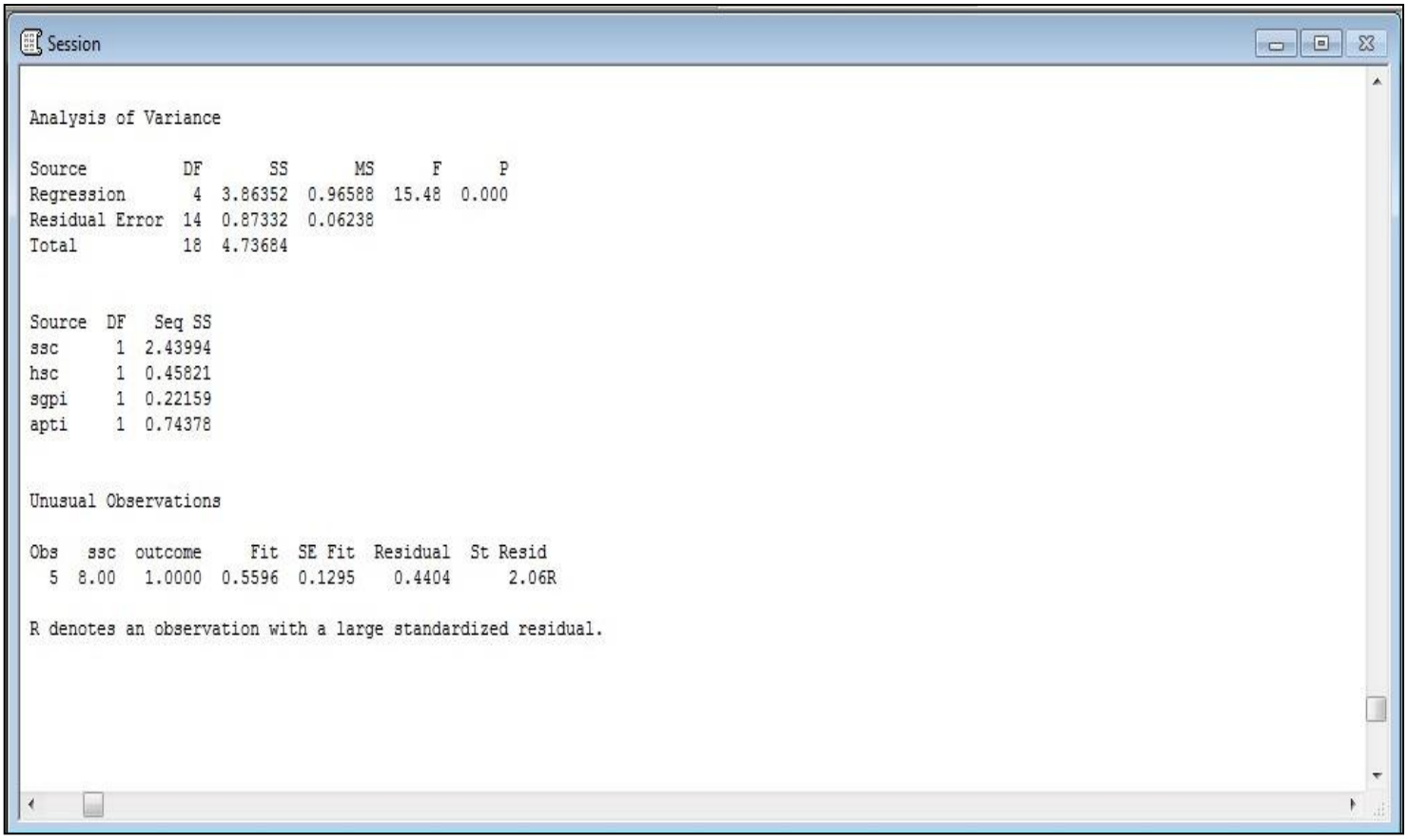
SR	Variable Coeff. (β)	Std.Err (σ)	p	O.R
1	0.2543	0.0578	0.9998	06337.7178
2	0.1078	6.4725	0.9992	235933.207
3	0.0480	3.9225	0.9997	0.0000
4	0.1585	2.6963	0.9984	07496731.3
Intercept	-3.8484	9.6518	0.9985	----------

Predicted Probability of Outcome, with 95% Confidence Limits...

Therefore final Logistic Regression Equation is

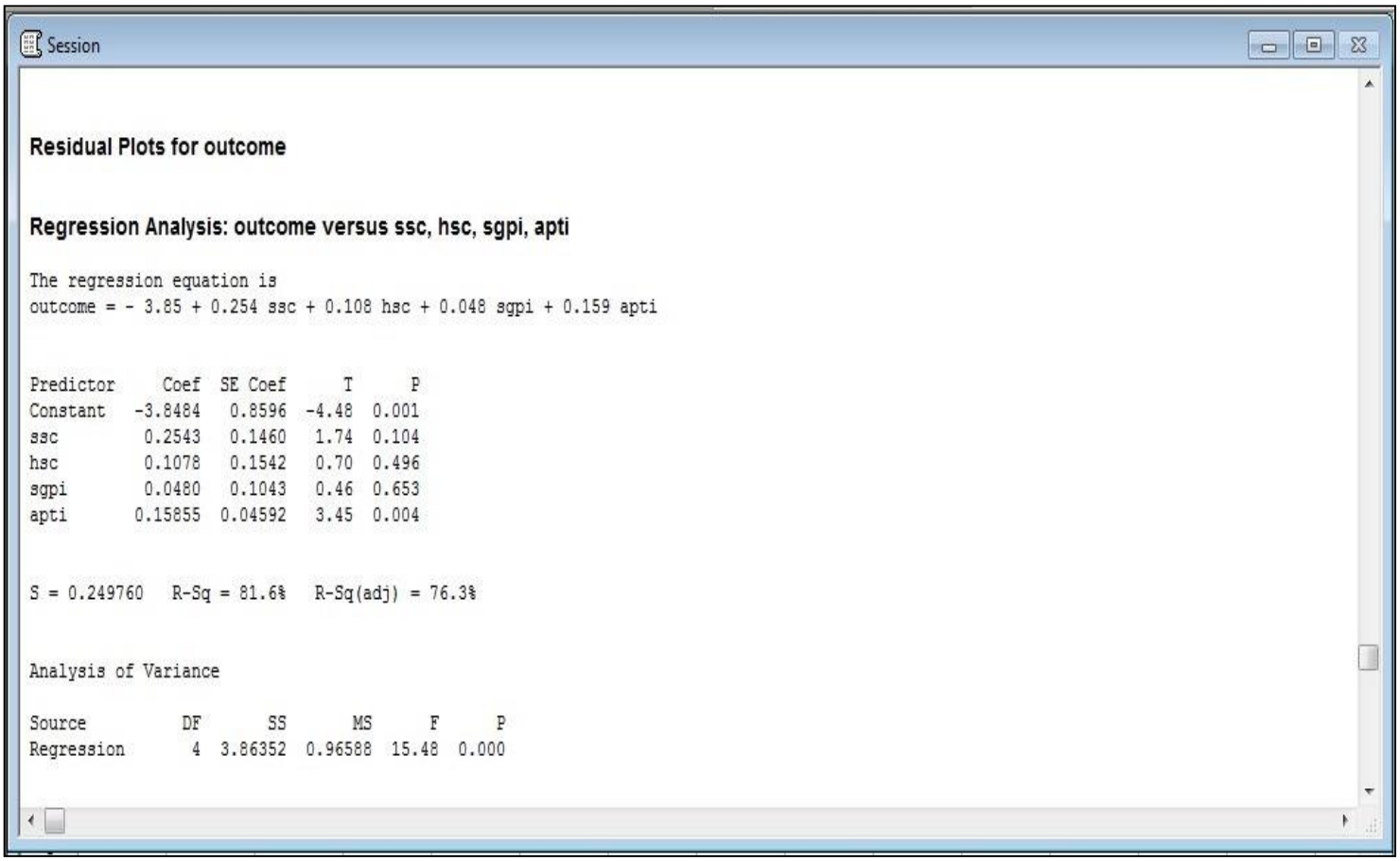
Probability (p) =

$$1/(1+e^{-(-3.8484+0.2543*SSC+0.1078*HSC+0.0480*SGPA+0.15855*Aptitude)})$$



Session

```
Analysis of Variance

Source          DF       SS       MS      F      P
Regression       4  3.86352  0.96588  15.48  0.000
Residual Error  14  0.87332  0.06238
Total           18  4.73684


Source  DF   Seq SS
ssc      1  2.43994
hsc      1  0.45821
sgpi     1  0.22159
apti     1  0.74378


Unusual Observations

Obs   ssc  outcome     Fit  SE Fit  Residual  St Resid
  5  8.00   1.0000  0.5596  0.1295    0.4404      2.06R

R denotes an observation with a large standardized residual.
```

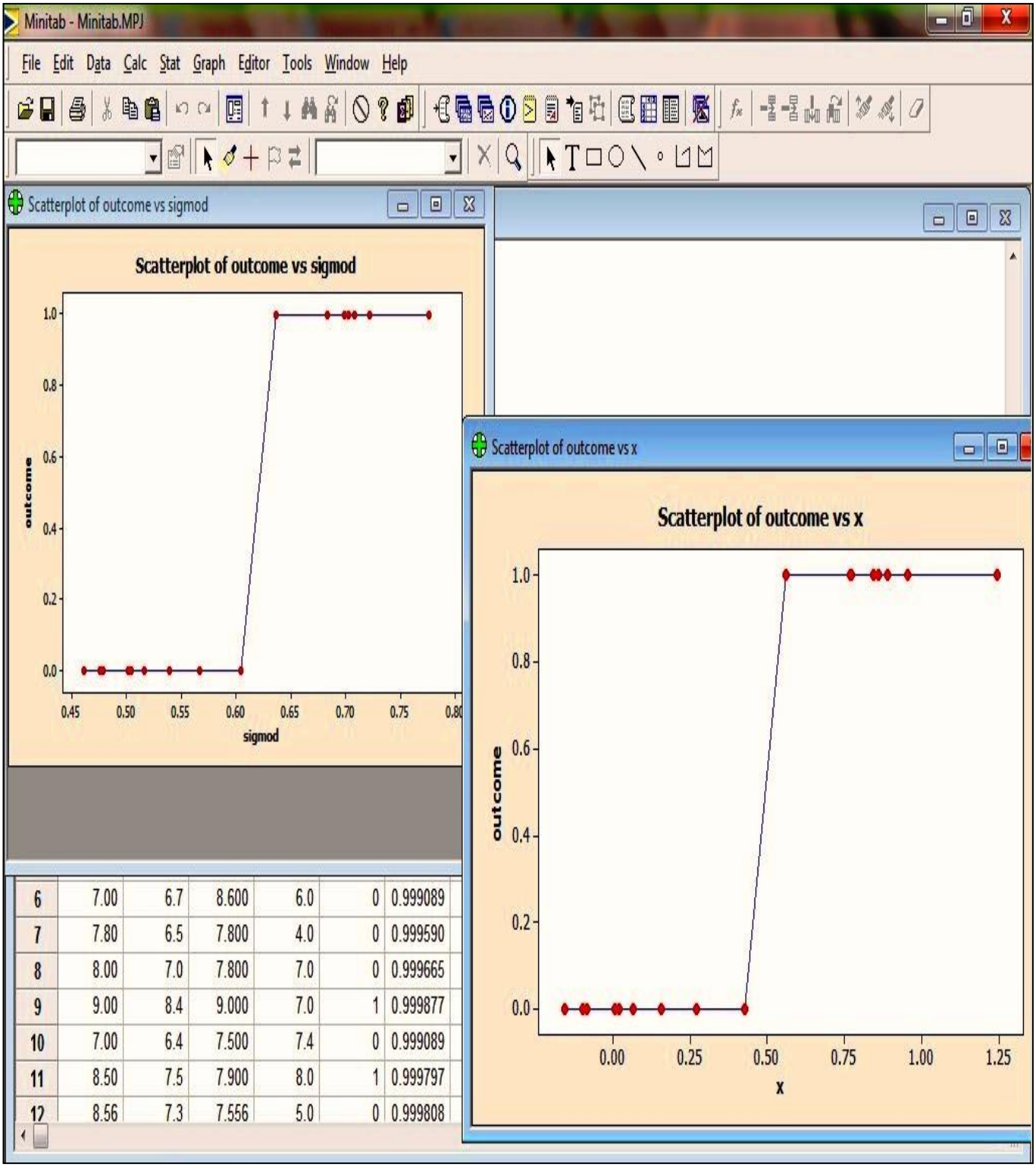


Session

Residual Plots for outcome

Regression Analysis: outcome versus ssc, hsc, sgpi, apti

```
The regression equation is
outcome = - 3.85 + 0.254 ssc + 0.108 hsc + 0.048 sgpi + 0.159 apti


Predictor     Coef  SE Coef      T      P
Constant   -3.8484   0.8596  -4.48  0.001
ssc         0.2543   0.1460   1.74  0.104
hsc         0.1078   0.1542   0.70  0.496
sgpi        0.0480   0.1043   0.46  0.653
apti       0.15855  0.04592   3.45  0.004


S = 0.249760   R-Sq = 81.6%   R-Sq(adj) = 76.3%


Analysis of Variance

Source          DF       SS       MS      F      P
Regression       4  3.86352  0.96588  15.48  0.000
```



IV. CONCLUSION

The final Logistic Model/Equation Obtained From Training Data is used to predict the Placement Chances for other students. The Sigmoid Function is designed using Data Sample from 20 arbitrary students, but the equation can become more

accurate if it is trained with huge data, so that the final curve will be more smooth and the prediction outcome will be more accurate. The Outcome is Probability based which predicts the chances of success.

ACKNOWLEDGMENT

The authors acknowledge Students from Atharva College Of Engineering, Mumbai for providing data support to carry out this work. The authors are thankful for the valuable suggestions made by Mrs. Deepali Maste, Professor, Computer Engineering Department.

REFERENCES

[1] David w.Hosmer, Stanley Lemeshow "Applied Logistic Regression" Second edition.

[2] Kabakchieva, Dorina. "Predicting student performance by using data mining methods for classification." Cybernetics and Information Technologies 13, no. 1 (2013): 61-72

[3] Scott Menard. "Applied Logistic Regression Analysis" , Issue 106.

[4] S.Philip Morgan, Jay D.Teachman , Vol. 50, No. 4 (Nov., 1988), pp.929-936. "Logistic Regression: Description, Example, Comparision "

[5] Viv Bewick, Liz Cheek and Jonathan Ball "Logistic Regression". Published Online 2005, Jan 13.

[6] Surjeet Kumar Yadav, Saurabh pal, World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012.

[7] Ajay Shiv Sharma, Keshav Kumar, 2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE), "PPS- Placement Prediction System Using Logistic Regression"

[8] Website:-" https://www.medcalc.org/manual/logistic_regression.php"