World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:3, No:3, 2009

Categorical Data Modeling: Logistic Regression Software

Abdellatif Tchantchane

Open Science Index, Computer and Information Engineering Vol:3, No:3, 2009 publications.waset.org/11499.pdf

Abstract**—**A Matlab based software for logistic regression is developed to enhance the process of teaching quantitative topics and assist researchers with analyzing wide area of applications where categorical data is involved.
The software offers an option of performing stepwise logistic regression to select the most significant predictors. The software includes a feature to detect influential observations in data, and investigates the effect of dropping or misclassifying an observation on a predictor variable. The input data may consist either as a set of individual responses (yes/no) with the predictor variables or as grouped records summarizing various categories for each unique set of predictor variables' values. Graphical displays are used to output various statistical results and to assess the goodness of fit of the logistic regression model. The software recognizes possible convergence constraints when present in data, and the user is notified accordingly.

Keywords**—**Logistic regression, Matlab, Categorical data, Influential observation.

I. INTRODUCTION

LOGISTIC regression is used in a wide range of applications leading to binary dependent data analysis. A Matlab based software for the analysis of logistic regression is developed. This work is motivated by the need of providing university-level students with statistical analysis tools that are interactive and easy to use. Other statistical tools have shortcoming that have been addressed by this software as follows: a) stepwise regression is performed in a simple and interactive manner where the regression model is built either by progressively adding variables, or removing variables from the original full model. b) detection of influential observations is done with ease in terms of the ability to examine the effect of their presence or absence on the significance of the overall model, or their impact on variables' significance. c) Graphical User Interface that is flexible, interactive and easy to use. Further more, the main window of the Matlab-based statistical package provides the user with a variety of functions to manipulate input data like, creating new variables from existing ones, filtering, selecting, removing observations, and selecting variables for study in the model.

In this paper, formulating the logistic regression and pinpointing its constraints and limitations is addressed. Next the interpretation of logistic regression results is given for single effect and the overall fit of the model. Two applications are treated to illustrate the various features provided by the software.

A. Tchantchane is an Assistant Professor with the University of Wollongong in Dubai, College of Computer Science and Engineering, Dubai, U.A.E. (phone:971-50-6731693; email: tchanlatif@uowdubai.ac.ae).

II. LOGISTIC REGRESSION FORMULATION

A. Deriving the logistic regression model

Logistic regression is used to analyze the dependence of a binary response variable y on a set of K independent explanatory variables:

$$\log(\frac{P_i}{1-P_i}) = \log(odds) = \beta_0 + \beta_1.X_{i,1} + ... + \beta_k.X_{i,K} = X_i.\beta \qquad (1)$$

P_i is the predicted probability of occurrence (y_i=1) for the ith observation (i=1..N). 1-P_i is the probability of non-occurrence (y_i =0). β is a (K+1) column vector of unknown parameters to be estimated including the intercept term. X_i is a (K+1) row vector of explanatory variables accounting for the ith observation. The explanatory variables may be continuous, categorical or both. The *odds* is defined as the ratio of the probability of occurrence over the probability of non occurrence.

Linear regression based on (1) can not be used for the following reasons [1]:

1. The response y_i is either 0 or 1 so the left hand size of (1) can not be evaluated.
2. The response variable is a discrete binary data and it can not be assumed to be normally distributed.
3. The predicted response may fall outside the (0-1) range, thus yielding meaningless results.

Equation (1) may be equivalently rewritten to yield the predicted probability of occurrence satisfying the constraint: 0 <P_i<1 [2]:

$$P_i(y_i = 1 \mid X_i) = \frac{e^{X_i.\beta}}{1 + e^{X_i.\beta}} = \frac{1}{1 + e^{-X_i.\beta}} \qquad (2)$$

The vector β is estimated by maximizing the likelihood function taking into account the contribution of the N independent observations:

$$\text{Maximize} \prod_{i=1..N} (\frac{1}{1+e^{-X_i.\beta}})^{y_i}.(1-\frac{1}{1+e^{-X_i.\beta}})^{1-y_i} \qquad (3)$$

Equation (3) may be transformed, by taking the natural logs, to yield the following maximization problem:

Open Science Index, Computer and Information Engineering Vol:3, No:3, 2009 publications.waset.org/11499.pdf

Maximize $f(\beta)=\sum_{i=1..N}\ln(\frac{1}{1+e^{-X_i.\beta}})^{y_i}+\ln(1-\frac{1}{1+e^{-X_i.\beta}})^{1-y_i}$ (4)

where $f(\beta)$ is the log-likelihood function and may be rewritten as:

$$f(\beta) = y'X.\beta - \sum_{i=1..N}\ln(1+e^{+X_i\beta}) \quad (5)$$

where $\boldsymbol{X}$ is (N)x($K+1$) matrix corresponding to the $\boldsymbol{N}$ observations of the K-explanatory variables including a column vector of ones for the intercept.

B. Log-Likelihood maximization

The Newton-Raphson iterative method, derived from the multivariate second order Taylor's expansion around β, is applied to estimate the parameters that maximize the log-likelihood function:

$$\beta_{t+1} = \beta_t - [\frac{\partial^2 f}{\partial\beta\partial\beta'}]^{-1}.\frac{\partial f(\beta)}{\partial\beta} \quad (6)$$

Where the $\frac{\partial f(\beta)}{\partial\beta}$ denotes a ($K+1$) vector of partial derivatives of the function $f(\beta)$ and is given by:

$$\frac{\partial f(\beta)}{\partial\beta}=\sum_{i=1..N}y_iX_i'-\frac{e^{X_i.\beta}}{1+e^{X_i\beta}}.X_i'=\sum_{i=1..N}(y_i-\frac{e^{X_i.\beta}}{1+e^{X_i\beta}}).X_i' \quad (7.a)$$

which is rewritten in matrix form as:

$$\frac{\partial f(\beta)}{\partial\beta} = X'.(y-\frac{e^{X.\beta}}{1+e^{X.\beta}}) \quad (7.b)$$

$\frac{\partial^2 f}{\partial\beta\partial\beta'}$ denotes a ($K+1$)x($K+1$) square symmetric matrix of second order derivatives known as the Hessian matrix of the function $f(\beta)$ and is given by:

$$\frac{\partial^2 f}{\partial\beta\partial\beta'}=\sum_{i=1..N}-\frac{(1+e^{X_i.\beta}).e^{X_i.\beta}-e^{X_i.\beta}(e^{X_i.\beta})}{(1+e^{X_i.\beta})^2}.X_i'.X_i \quad (8.a)$$

which is rewritten in matrix form as: $\frac{\partial^2 f}{\partial\beta\partial\beta}=-X'.D.X$ (8.b)

Where $\boldsymbol{D}$ is an (N)x(N) diagonal matrix:

$$D_{i,i}=\frac{(1+e^{X_i.\beta}).e^{X_i.\beta}-e^{X_i.\beta}(e^{X_i.\beta})}{(1+e^{X_i.\beta})^2}=P_i.(1-P_i) \quad (8.c)$$

The iteration procedure starts with an initial guess set to:

$$\beta_0=(X'X)^{-1}.X'.y \quad (9)$$

At each iteration, a new estimate of the vector β is obtained where (7) and (8) are used to evaluate respectively the derivative vector and the Hessian matrix. Convergence is reached if the norm of the derivative vector is sufficiently close to zero.

C. Convergence criteria

The iterative process to maximize the likelihood function fails to converge if [2],[3],[4]:

1. There exists a perfect multicollinearity among the explanatory variables which may be detected if the matrix $X'.X$ is singular.
2. There exists a perfect or quasicomplete separation of the response variable with respect to the explanatory variables. The perfect separation is detected if the response of each observation is predicted with probability 1 and the log-likelihood goes to zero. In the case of quasicomplete separation, the maximum likelihood estimate does not exist as the Hessian matrix becomes unbounded.
3. Presence of small and/or sparse data set.
4. A low percentage of values in the data set for which y_i=1 or for which y_i = 0.

III. Digesting Logistic Regression Results

A. Single effect of the explanatory variable

The coefficient β_j estimated by the logistic regression models the single effect of the ***j***-th explanatory variable on the response variable. Based on (1), a change of ΔX_j in the ***j***-th explanatory variable while holding the rest of the variables constant, would change the predicted *odds* to:

$$\begin{aligned}\Delta\log(odds)&=\log(odds_{after\_the\_change})-\log(odds_{before\_the\_change})\\&=\beta_0+...+\beta_j(X_{.j}+\Delta X_j)+...+\beta_k.X_{.K}-(\beta_0+...+\beta_jX_{.j}+...+\beta_k.X_{.K})\\&=\beta_j\Delta X_j\end{aligned} \quad (10.a)$$

and by a factor of

$$\Delta\ odds = e^{\beta_j.\Delta X_j} \quad (10.b)$$

B. Marginal effect of the explanatory variable

The marginal effect of the j-th explanatory variable on the response variable is derived from (2) [1]:

$$\frac{\partial P}{\partial X_{.j}} = P(1-P).\beta_j \quad (11)$$

Unlike in ordinary least squares, the marginal effect is not constant and depends not only on the coefficient of regression but also on the quadratic function of P. The marginal effect is maximal when P is close to 0.5.

C. Significance test of the individual regression coefficients

The negative inverse of the Hessian matrix calculated by (8.b) is used as an approximate to the variance-covariance matrix of the logistic regression. The standard deviation vector is estimated from the Newton-Raphson final iteration and given by:

$$s = Diag\{sqrt(X'.D.X)^{-1}\} \quad (12)$$

The Wald test χ^2 statistics, based on the ratio squared of the parameter value over its standard deviation, provides a significance test for the logistic regression coefficients β with

Open Science Index, Computer and Information Engineering Vol:3, No:3, 2009 publications.waset.org/11499.pdf

one degree of freedom. The significance p_{value} of each explanatory variable is determined using the Matlab *cdf* cumulative distribution function:

$$p_{value} = 1 - cdf\ ('chi\ 2', \left(\frac{\beta_j}{s_j}\right)^2, df = 1) \quad (13)$$

Low p_{value} indicates that the regression coefficient is significantly different from zero.

D. Goodness of fit test

The overall goodness of the model fit is assessed by the following χ^2 distributed term:

$$G^2 = -2\ [Max\ f(\beta_0) - Max\ f(\beta_0,...\beta_K)] \quad (14)$$

$Max\ f(\beta_0)$ corresponds to the likelihood function under the null hypothesized of the intercept-only model, and $Max\ f(\beta_0,...\beta_K)$ corresponds to the estimated *K*-variables full model: H_0:$\beta_1=\beta_2=...=\beta_K=0$, versus H_1: with at least one of the β_j is significantly different from zero.

Based on G^2 and using the chi-square cumulative distribution on a degree of freedom equal to the number of explanatory variables, the overall significance p_{value} is obtained:

$$p_{value} = 1 - cdf\ ('chi2', G^2, df = K) \quad (15)$$

A low p_{value} indicates that the model fits well and the null hypothesis should be rejected.

E. Naglelkerke and Cox-Snell R^2

Cox-Snell R^2 and Nagelkerke's R^2 are likelihood based indicators measuring the strength of the relationship between the dependent variable and the explanatory variables [5], [6]:

$$R^2_{Cox\ -\ Snell} = 1 - \left[\frac{e^{\max\ f(\beta_0)}}{e^{\max\ f(\beta_0,...,\beta_K)}}\right]^{2/N} \quad (16)$$

$$R^2_{Nagel\ ker\ ke} = \frac{R^2_{Cox\ \_\ Snell}}{1 - \left[e^{\max\ f(\beta_0)}\right]^{2/N}} \quad (17)$$

R^2 varies between 0 and 1; high values of R^2 (close to 1) mean that the observed and predicted responses correlate tightly. However, low values of these indicators are common even in the presence of strong relationship between the response and the explanatory variables.

IV. Application of the Logistic Regression

In order to illustrate the various software features we present the analysis of two applications selected from the literature and treated by the Matlab-based logistic regression software.

A. Application I: Investigating risky software

The first application investigates whether or not a software project is risky or not, based on four criteria of the project: *insufficient estimation for the requirements*, *lack of stakeholders' commitment for estimation*, *lack of breakdown of the work product* and the *insufficient planning of project monitoring*. All the explanatory variables are scaled from 0 to 3. The response variable y_i equals 1 for a risky software project and 0 for a not risky software project. The data consists of 32 observations and it is available in [3].

1) Interpreting results

Performing logistic regression on the data yields the results given in TABLE I. For each regression coefficient estimated, the standard deviation and the significance p_{value} based on the Wald test statistics are reported. The variation in the *odds* calculated by (10.b) for a one-unit change in the explanatory variable ($\Delta X_j = 1$) is given in column 6. The 95% confidence interval, given in column 7, is determined using the regression coefficient standard deviation:

$$CI = e^{\beta_j \pm 1.96\ s_j} \quad (18)$$

The fit for this model gives a deviance χ^2=-2[-39.75/2–14.64/2]=25.11 on four degrees of freedom and a corresponding p_{value}<0.000 indicating a good overall fit.

TABLE I
LOGISTIC REGRESSION RESULTS: SINGLE EFFECT (LEFT OF TABLE) AND OVERALL EFFECT (RIGHT OF TABLE)

Logistic Regression

Response Variable: Risky_Software

File: D:\pr-oct-2008\Project\logistic-regression\riskysoftware.txt

Var. Name	Beta	Sd.E.	Wald	p_value	Exp(Beta)	95% CI Exp(Beta)
Y_Intercept	-8.8352	4.091	4.665	0.0308	0.000	0.0 to 0.4
Insufficient_Estimation	1.5772	1.025	2.367	0.1239	4.841	0.6 to 36.1
Lack_Of_Commitment	0.9642	0.630	2.342	0.1259	2.623	0.8 to 9.0
Lack_Of_BreakDown	1.2283	0.688	3.185	0.0743	3.416	0.9 to 13.2
Insufficient_Planning	2.2222	1.156	3.697	0.0545	9.228	1.0 to 88.9

Overall Fit Results

Full Model:-2*LogLikeliHood	14.64
Null Model:-2*LogLikeliHood	39.750
Chi^2	25.11
p_value(df=4)	0.0000
N_ones	10
N_zeros	22
# of match (y=Yes)	70.0 %
# of match(y=No)	95.5 %
R^2(NagelKerke)	0.764
R^2(Cox_Snell)	0.544
Niterations	30

In Fig. 1, predicted responses are plotted versus observed responses. The observations 16 and 25 are poorly predicted. The four data display of the response variable with respect to 2-explanatory variables is illustrated by Fig. 2. The straight lines represent the regression contour lines described by (1) for different probabilities ***P*** (0.1, 0.5, 0.9) holding the remaining variables to their mean values.

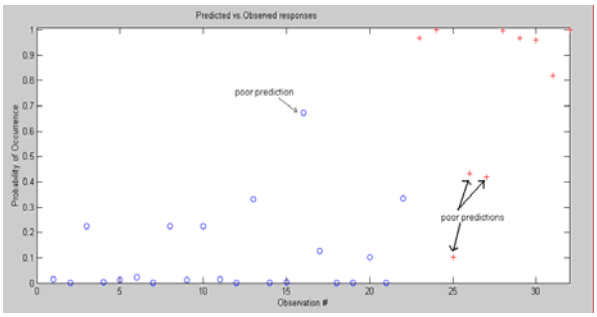


Fig. 1 Predicted probability versus the observed response: (+) denotes observation with y_i=1, (o) denotes a response with y_i=0.

Open Science Index, Computer and Information Engineering Vol:3, No:3, 2009 publications.waset.org/11499.pdf

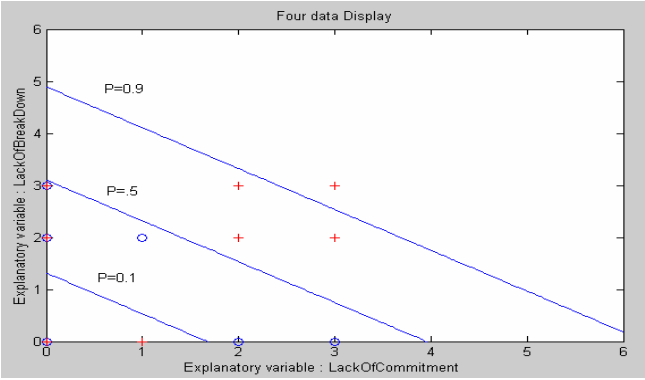


Fig. 2 Four data display: two numerical predictors "*lack of break down*", "*lack of commitment*" and the binary outcome. (+) denotes a response with y_i=1, (o) denotes a response with y_i=0. The contour lines are displayed as a function of the probability of occurrence.

2) Detecting influential observations

The software provides a utility for plotting both single effects and the overall effect due to the removal or the misclassification of an observation which may change a variable from non significant to significant [7].

The effect of dropping an observation from the data on the significance of the variable "*lack of break down*" is shown by Fig. 3. Dropping any of the observations 16, 25 or 27 would deteriorate the significance of the variable.

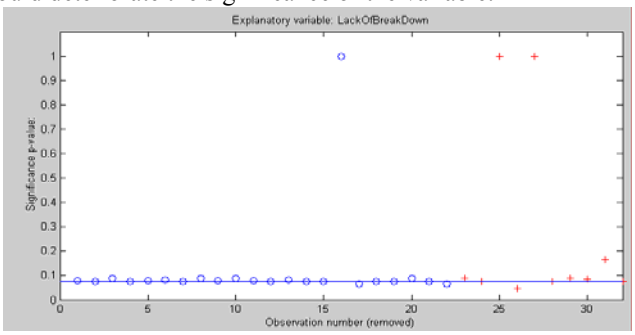


Fig. 3 The effect of dropping an observation on the p_{value} of the variable "*lack of break down*". The solid line is the original p_{value}. (+) denotes a response with y_i=1, the (o) denotes a response with y_i=0.

The single effect due to the misclassification of a response from 1/0 or vice versa is plotted in Fig. 4 for the variable "*Insufficient planning*".

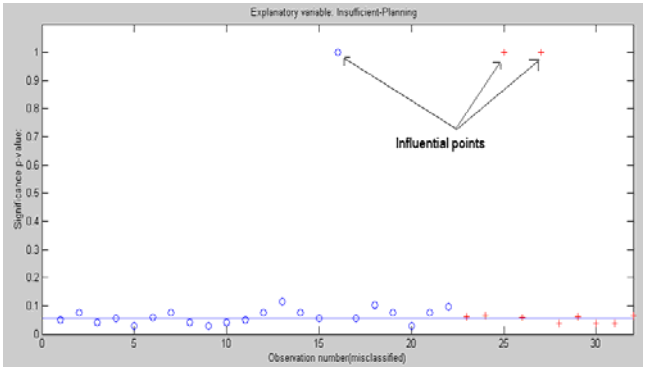


Fig. 4 The effect of misclassification of a single observation on the significance $\boldsymbol{p}_{value}$ of the variable "*insufficient planning*". The solid line is the original $\boldsymbol{p}_{value}$. (+) denotes a response with y_i=1 the (o) denotes a response with y_i=0.

However, unlike in ordinary least squares, the plots of the residuals versus explanatory variables are not provided in the software because in the case of logistic regression, any apparent trends of dependence of the residuals on the explanatory variables would not necessarily reveal a violation in the fit [8].

In addition the algorithm, detailed in [7] for detecting influential observations, is implemented in the software. The results are presented in Fig. 5 and appear to be in agreement with those reported in Fig. 3 and Fig. 4.

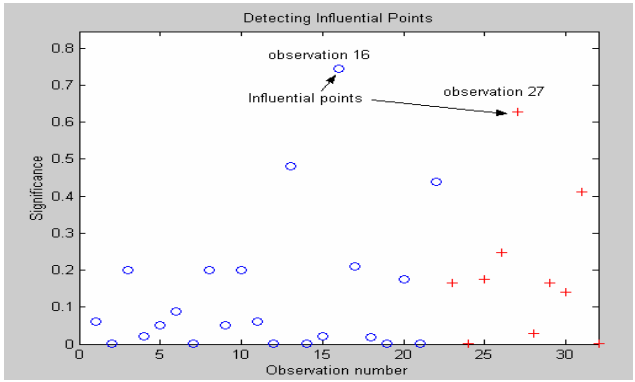


Fig. 5 Detecting influential observations in the data

3) Stepwise Regression

The software provides a feature for investigating the effect of adding or removing iteratively one explanatory variable from the model. Based on (14), a χ^2 deviance test on the difference between restricted and unrestricted model with one degree of freedom is used to test the significance of adding or removing a variable from the model. The effects of removing a single variable from the 4-variable unrestricted model are reported in TABLE II. These results are in agreement with those reported in TABLE I where the variable "*lack of commitment*" has the least effect on the dependent variable and the variable "*insufficient planning*" has the most significant effect.

TABLE II
EFFECT OF REMOVING A SINGLE VARIABLE FROM THE FULL MODEL

Effect of Removing one Variable from the original model

Variable. Name Removed	-2*maximum Likelihood	Chi_2	p_value	Rsquare NagelKerke	Rsquare Cox_Snell
Insufficient-Estimation	18.8	20.98	0.00011	0.6761	0.4809
Lack-Of-Commitment	17.3	22.45	0.00005	0.7088	0.5042
Lack-Of-BreakDown	19.4	20.32	0.00015	0.6609	0.4701
Insufficient-Planning	24.2	15.59	0.00137	0.5423	0.3857

4) Forecasting the probability of occurrence

The logistic regression software provides an interactive

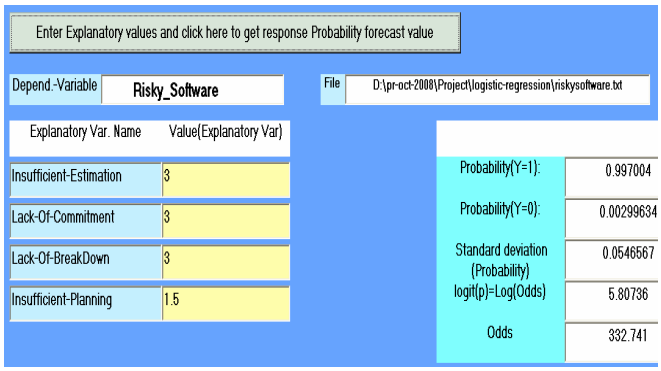
Open Science Index, Computer and Information Engineering Vol:3, No:3, 2009 publications.waset.org/11499.pdf

tool for predicting the probability of a project being risky for a given set of the explanatory variables. This tool is illustrated by TABLE III. The expected probability of success is obtained from (2) and the variance is obtained from the binomial distribution [7],[9]:

$$\text{var}(P_j) = E([(P_j - E(P_j))]^2) = (1 - \frac{e^{x_j\beta}}{1+e^{x_j\beta}})(0 - \frac{e^{x_j\beta}}{1+e^{x_j\beta}})^2 + (\frac{e^{x_j\beta}}{1+e^{x_j\beta}})(1 - \frac{e^{x_j\beta}}{1+e^{x_j\beta}})^2 \quad (19.a)$$

which is simplified to: $$\text{var}(P_j) = \frac{e^{x_j\beta}}{(1+e^{x_j\beta})^2} \quad (19.b)$$

TABLE III PREDICTED PROBABILITY $P(Y_i=1|X_i)$ OF A PROJECT BEING RISKY



B. *Application 2: The Titanic survivors*

The second application investigates the survival factors of the 2201 people who were on board of the Titanic ship according to their *economic status*, *gender* and *age*. As an alternative way of representing the data of the 2201 subjects, the data is grouped into 14 unique combinations of the explanatory variables (e.g. male/first-class/adult) [10]. This way of grouping the data is not suitable with continuous explanatory variables. The log-likelihood function for grouped data is given by:

$$f(\beta) = \sum_{j=1..J} s_j \ln(\frac{1}{1+e^{-X_j.\beta}}) + (m_j - s_j)\ln(1 - \frac{1}{1+e^{-X_j.\beta}}) \quad (20)$$

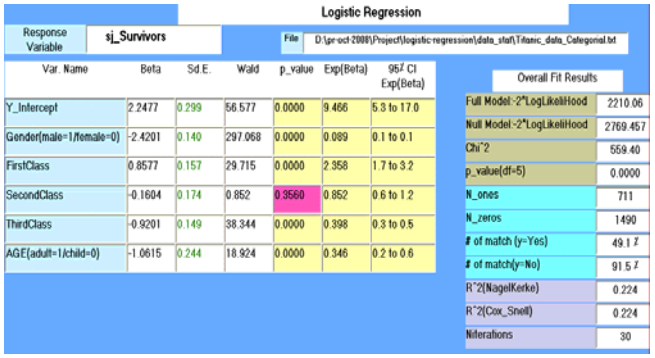
Where $\boldsymbol{X_j}$ is a unique set of values recording the economic status (first/ second/ third/ fourth class), gender (male=1/ female=0) and age (adult=1/ child=0). The economic class status is represented by 3 variables; the *fourth* class is coded by 0 in the three variables. The number of survivors ($\boldsymbol{s_j}$) and the total number of subjects ($\boldsymbol{m_j}$) for each category are recorded.

The log-likelihood in (20) may be simplified similarly as done previously to:

$$f(\beta) = s' X.\beta - m' \sum_{j=1..J} \ln(1 + e^{+X_j\beta}) \quad (21)$$

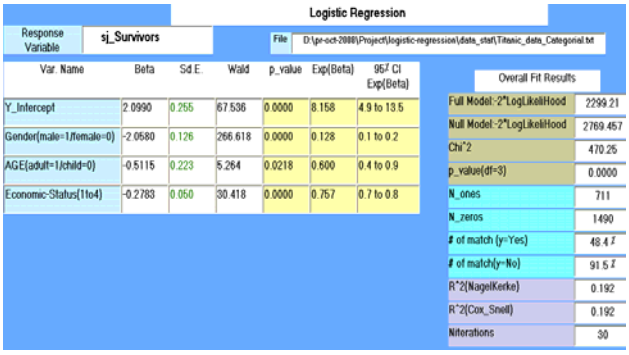
The logistic regression results are given in TABLE IV. The deviance statistics for the model is 559 on 5 degrees of freedom. The regression coefficient of the second class is not significant (p_{value}>0.35). The gender regression coefficient is 2.4201 indicating that males had a lower priority for survival than women with an *odds* of 1/0.089 ≈ 11 to 1.

TABLE IV RESULTS OF THE TITANIC APPLICATION (USING 3 EXPLANATORY VARIABLES FOR THE ECONOMIC STATUS). NUMBER OF CATEGORIES J=14 AND NUMBER OF SUBJECTS =2201



When the economic status is represented by only one variable and coded from 1 to 4 (for the four classes), the logistic regression of this model results in a deviance statistics of 470 on 3 degrees of freedom ($\boldsymbol{p}_{value}$ ≈0). As displayed in TABLE V, the individual effects of all the three explanatory variables are significant. The difference between these two models is not significant as the regression coefficients from the two designs are within the 95% confidence interval of each other.

TABLE V RESULTS OF THE TITANIC (USING ONE VARIABLE FOR THE ECONOMIC STATUS).NUMBER OF CATEGORIES J=14 AND NUMBER OF SUBJECTS =2201



V. CONCLUSION AND FUTURE WORK

In this paper, two applications were used to explain some of the features of the Matlab-based software. The statistical analysis shown reflects the flexibility of the software in terms of user interactivity, manipulation of complex functions, and the ease of use due to the graphical-oriented interface. It is hoped that this statistical tool would help university students, staff and researchers in aiding the process of learning statistics, applying its various tools to complex business applications, and advancing the state of research of automation in quantitative tools to support both educators and researchers.

In future work, the implementation details and interpretation of other dependent variable models, including ordinal, multinomial and panel data will be explored.

REFERENCES

[1] Mark A. Huselid and Nancy E. Day , "Organizational commitment, job involvement, and turnover: A substantive and methodological analysis", Journal of Applied Psychology 1991, vol. 76, N0 3 380-391).
[2] Elizabeth N. King and Thomas P. Ryan "A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression", The American Statistician, August 2002, vol. 56 No 3.
[3] Yasunari Takagi, Osamu Mizuno and Tohru Kikuno, Empirical Software Engineering, 10, 495-515, 2005.
[4] [Santner T.J. and Duffy, E.D. (1986), "A note on A. Albert and J.A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models." Biometrika, 73, pp. 755-758
[5] D. R. Cox and Nanny Wermuth, "A comment on the coefficient of determination for binary responses." The American Statistician, February 1992, vol. 46 N0 1.
[6] Marija J. Norusis, "SPSS 16.0 Advanced Statistical Procedures Companion", Prentice Hall Inc (2008). ISBN- 13:978-0-13-606140-3.
[7] Michael P.Fay, "Measuring a binary response's range of influential in logistic regression", The American Statistician, February 2002, vol. 56. N0 1.
[8] Iain Pardoe and R. Dennis Cook "A graphical method for assessing the fit of a logistic regression model", The American Statistician, November 2002, vol. 56, No 4.
[9] Nicholas J. Horton and Stuart R. Lipsitz , "Review of software to fit generalized estimation equation regression models", The American Statistician, May 1999, vol. 53.
[10] Jeffrey S. Simonoff, "Logistic regression, categorical predictors, and goodness-of-fit: It depends on who you ask", The American Statistician, February 1998, vol. 52 N0 1.

Open Science Index, Computer and Information Engineering Vol:3, No:3, 2009 publications.waset.org/11499.pdf