

Running Head: LOGISTIC REGRESSION: AN OVERVIEW

Logistic Regression: An Overview

Lawrence M. Healy

Eastern Michigan University, College of Technology

COT 711

March 20, 2006

Introduction

Regression Analysis is a multivariate statistical methodology to investigate relationships and predict outcomes. One type of regression analysis is known as logistic regression. Logistic regression is appropriate when the predicted outcome is binary (on/off, pass/fail, infected/not infected, etc.). Logistic regression techniques resolve inconsistencies associated with dichotomous dependent data and the assumptions of ordinary sum of squares regression methods. The independent variables that are used for outcome prediction may be dichotomous, categorical or continuous. Logistic regression is commonly used in manufacturing and health related studies. It can be used for any application where binary outcomes can be predicted. Logistic regression is based on the logit transformation of the dependent variable. The logit transformation generates a continuous logarithmic curve from non-continuous data so that a regression model can be developed. The outcome probabilities for each dependent variable value are the basis for the model. The logit transformation is necessary since dichotomous dependent data violates ordinary least squares assumptions. Another issue with dichotomous data is that the error terms are not normally distributed, thus ordinary sum of squares regression and all normality tests are invalid.

Logistic regression is less restrictive than ordinary sum of squares regression. It does not require normally distributed dependent data or homogeneity of variance. Predictions made by ordinary sum of squares regression are based on the observed changes in the independent data itself. Logistic regression is based on the log of the odds of a particular event occurring with a given set of observations. Logistic regression's underlying principles are based on probabilities and the nature of the log curve. The only assumptions of logistic regression are that the resulting logit transformation is linear, the dependent variable is dichotomous and that the resultant

logarithmic curve doesn't include outliers. Discriminant analysis and logistic regression will produce similar results with dichotomous dependent data except discriminant analysis is more restrictive and complex. Unlike discriminant analysis, logistic regression does not restrict the nature of the independent variable. In contrast with discriminant analysis, logistic regression doesn't restrict categorical independent variables. Discriminant analysis relies on strict adherence to normality and the equal variance assumptions while logistic regression does not have this requirement. Press and Wilson (1978) found in their comparison of discriminant analysis techniques with logistic regression that the predictions formed by these methods were consistent. They recommend that logistic regression be used whenever possible, especially when normality assumptions are violated or when there are a large number of qualitative variables. Logistic regression techniques are appropriate when dummy variables are required during analysis. Logistic regression is preferred by many researchers in the analytical fields due to its robust practical nature, intuitive assumptions and its ability to produce a predictive representation of real world situations.

Logistic regression's conceptual foundation is based on the benefits of the logarithmic curves flexibility. Figure 1(Pampel, 2000) illustrates the contrast between predictive values using the transformation to continuous data versus a plot for the original non-transformed dichotomous values. Without the logit transformation, dichotomous value prediction is not appropriate for ordinary multiple regression techniques. The similarities of logistic regression with other common multiple regression tools provides the researcher with a relatively easy to use, practical and less restrictive tool to analyze dichotomous data. Table 4 (Babin, et.al, 2006) depicts a summary of the comparison between multiple regression and logistic regression techniques.

This paper will present an overview of the logistic regression methodology and illustrate its usefulness with a real world case study.

Logistic Regression Method

As with any research the study objectives must first be defined. The researcher should then establish the best research design to address the objective. After stating all assumptions, the researcher should estimate the logistic regression model using the logit transformation and assess overall model fit and predictive accuracy. The results should then be interpreted and validated. Once the logistic regression model is estimated there are 3 primary considerations for diagnostic analysis: 1) How well does the model predict outcome? Does a relationship exist between the independent variables as a group and dependent variable such that the independent variables within a given level of confidence actually predict the outcome and that outcome is not random chance? 2) If the model works well, what is the relative predictive strength of each independent variable? 3) Are the assumptions of the model completely satisfied? The intent of this paper is to present a practical overview of logistic regression techniques and considerations. Theoretical derivations of logistic regression methods are outside the scope of this paper.

Generate the Logistic Regression Model

The first task in model estimation is to transform the independent variable and determine the coefficients of the independent variables. The basic logistic regression analysis begins with logit transformation of the dependent variable through utilization of maximum likelihood estimation. This is done using the odds ratio. The odds ratio for an event is represented as the probability of the event outcome / (1 - probability of event outcome).

The odds ratio can be described as

$$\text{Odds}_i = \left[\frac{p_i}{1 - p_i} \right] = e^{b_0 + b_1 X_1 + \dots + b_n X_n} \quad (1)$$

where

p_i is the probability of an event i ,

$b_0 + b_1 X_1 + \dots$ represents the regression model.

It represents all event probabilities, relationships and their exponential nature. The odds ratio has numerous advantageous properties. It clearly portrays the increased or decreased likelihood of an event outcome occurrence. If the odds ratio is less than one there is a decreased likelihood of an event occurring and if the odds ratio is greater than one then there will be an increased likelihood of the event occurring. The odds ratio provides an intuitive foundation for any sensitivity analysis of interest between the dependent and independent variable. The odds ratio is based on the probabilities that a specific binary outcome will occur when using particular model estimation. It is converted to a continuous function through the logit transformation. The new plot of the transformation of the independent data into probabilities versus the dichotomous dependent data will be continuous ranging from infinity to negative infinity. The log of the odds ratio is known as the logit.

For each data point, logit_i is represented by

$$\text{logit}_i = \ln \left[\frac{p_i}{1 - p_i} \right] \quad (2)$$

The maximum likelihood estimation (MLE) is now used to estimate the coefficients $(\beta_0, \beta_1, \dots, \beta_p)$ from the logit transformation. MLE is similar to the ordinary least squares used in multiple regression analysis. The likelihood is the probability that the observed values of the

dependent variable will be predicted by the observed independent variable data. The log likelihood (LL) is the log of that likelihood and is in the range of infinity to negative infinity. The logistic curve simplifies the coefficient estimation. The maximum likelihood estimate seeks to maximize the LL value and estimate the coefficient found at that maximum point. It is determined through an iterative process that is normally handled by computer software such as SAS or Minitab. One point worth mentioning is that MLE is extremely accurate for large sample sizes. Since the LL is the log probability that the dependent variables will be predicted by the observed independent variables, we should seek to maximize that probability. The coefficient estimate where the log likelihood is maximized will represent the best probability that the observed dependent variable is predicted by the observed independent variables. At this juncture SAS or some other statistical package has computed the log likelihood and logit transformations to estimate the coefficients for the initial model. This paper will now address logistic regression diagnostics to validate the proposed model derived through the logit transformation.

Logistic Regression Diagnostics

The first concern is the consistency of the model. How well does the model predict the outcome? The researcher should ascertain the predictive error occurrence and the model's sensitivity to that prediction error. The predictive accuracy of the model must be determined. This is accomplished through goodness-of-fit measures such as log-likelihood and the coefficient of determination (R^2). This section will address these two measures and then briefly discuss predictive accuracy considerations.

Goodness-of-Fit for the Estimate - Chi-square / Maximum Likelihood

Due to the non linear nature of dichotomous dependent data ordinary sum of squares methods will not be appropriate. The log likelihood method will be used. Manipulation of the log likelihood value leads to a test statistic known as the -2LL value. The -2LL value has an approximate chi-square distribution and therefore can be used to evaluate the significance of the logistic regression model. This is similar to the sum of squares error analysis used in multiple regression. The -2LL statistic is known as the likelihood ratio (goodness-of-fit). A perfect fit will yield a -2LL equal to zero (minimum value). As the -2LL value decreases the model is interpreted as having a better fit and predictive estimation. The log likelihood test is an alternative to another test statistic that is often used in logistic regression, the Wald test statistic. The likelihood ratio should now be used to compare a reduced model against the proposed logistic regression model. The log likelihood test can be used to test the overall model goodness-of-fit as well as the significance of each individual parameter.

The test of the overall model consists of comparing the -2LL statistic of a baseline null model (no independent variables, just a constant) against the full model with all proposed independent variables. The null hypothesis implies that the researcher should accept the baseline model without any beta coefficients ($\text{logit}(p) = \text{constant}$). The alternative hypothesis is that the full model is significant. The baseline model -2LL statistic is compared with the full model (all independent variables included). The difference between the baseline -2LL value and the full model -2LL value is known as the model chi-square test statistic. If the (model chi-square test statistic) is $\leq .05$ then we should reject the null hypothesis. If we reject the null hypothesis that knowing the independent variables makes no statistical difference in the prediction of the

dependent variable, then the overall model with all independent variables at this point is statistical significant.

The next step is to evaluate the significance of each individual variable. If removing a variable from the full model yields no statistically significant change in the -2LL test statistic then the variable is not significant and can be omitted from the full model. This process is repeated for all independent variables to determine whether any independent variables can be eliminated from the full model.

Goodness-of-Fit for the Estimate – R_L^2 , Coefficient of Determination

Another method to evaluate the overall model is by using a coefficient of determination approach. This method fits the estimation in a similar fashion that the coefficient of determination is used in multiple regression analysis.

The R^2 for the logit model is represented as

$$R^2_{\text{LOGIT}} = \frac{-2LL_{null} - (-2LL_{model})}{-2LL_{null}} ; 0 \leq R^2_{\text{LOGIT}} \leq 1 \quad (3)$$

It implies the degree of relative negative impact the independent variables has on the model versus not having any independent variables at all. It reflects how much the badness-of-fit is reduced as well as a proportionate reduction in the absolute value of log likelihood. The model fit improves as R^2_{LOGIT} increases from 0 to 1, and is a perfect fit at $R^2_{\text{LOGIT}} = 1$. After determining the significance of the overall model as well as the individual independent coefficient significance we should examine the predictive accuracy of the model.

Predictive Accuracy

In logistic regression, the predictive accuracy of a model is an issue that has received considerable discussion in the literature. It appears that many researchers are confident with model predictions using goodness-of-fit without utilizing predictive classification techniques. The literature indicates that there are varied ways to estimate predictive accuracy as well as many logistic regression statistic packages that will analyze the classification and predict the accuracy of the model. Therefore this paper will only briefly cover the basics of classification and predictive accuracy. This paper will now present one way to determine predictive accuracy through classification table assignment. In its most basic form, the classification tables can be computed by observing the predicted values for each set of independent variables and then comparing those values with the average predictive values for all observations. If the individual fitted logit predictive value is less than the average of the predicted values, then it is classified in the 0 group, else it is classified in the 1 group. To highlight errors in the model, the actual Y values for each data point are compared to the assigned group (i.e. 0 or 1). For instance, if a data point was assigned to group 1 and the observed Y for that set of independent variables was 0, this would indicate an error in the model. This comparison is performed for all data point sets and the overall model's predictive accuracy is determined.

The predictive accuracy is stated as:

$$\frac{\text{Correct_predictions} - \text{Number_of_incorrect_predictions}}{\text{Number_of_total_observations}} \quad (4)$$

The acceptable accuracy level is a practical consideration and dependent upon the sensitivity of the data. Table 1 (Ali, 2000) illustrates a simplistic example of a predictive accuracy computation. As shown in the example, the model correctly predicted rows 1, 2, 3 and 66

because the assigned group was equal to the observed Y. The model incorrectly predicted in rows 9 and 52. The average predicted accuracy for this example was 0.50.

The next question to address is: what is the relative predictive strength of each independent variable? Assuming the overall model is significant then the next step is to determine the relative strength of each independent variable. To determine substantive significance as related to the dependent variable we can observe the un-standardized regression coefficients and ascertain the strength of any causal relationships between the dependent and independent variables. The acceptable measure of change caused by the un-standardized regression coefficient must be determined for benchmark purposes. Another way to approach the determination of independent variable strength is to utilize the standardized regression coefficient. This represents the number of standard deviations a dependent variable changes as a response to one standard deviation change in the independent variable.

Testing Statistical Significance of the Coefficients

To determine the significance of the independent variables we can use either the Wald statistic or the likelihood ratio test. The likelihood ratio test requires extensive repetitive computations but these can be handled by most statistical software packages. The Wald statistic is a method to test whether the coefficients are significantly different from 0. A more common process to test the significance of the coefficients is to evaluate the exponentiated logistic expression of the coefficients. It is a transformation of the original logistic coefficient. Due to the logarithmic nature of the logistic coefficient, it can be difficult to interpret, but this difficulty can be overcome through many statistical software packages. There are two approaches to estimating the logistic coefficient significance. The original logistic coefficient can be used to

interpret changes in the logit function caused by the coefficient analysis or the exponentiated logistic expression can be used as a means to interpret changes in the odds. The direction of the relationship (positive or negative) of the original coefficients should be determined. The directional interpretation of the exponentiated coefficients is as follows: if the value of the exponentiated coefficient is greater than 1 then there is a positive relationship and if the value it is less than one, then there is a negative relationship. For example, if the original coefficient is positive, the transformed exponentiated logistic expression will be greater than one, this means that the odds will increase for any positive change in the independent variable. This results in a higher probability of occurrence. If the coefficient is zero then the exponentiated logistic expression will be equal to one and there will be no change in odds for a change in the independent variable.

The magnitude of the change will be reflected as follows

$$\text{Percent change} = (\text{exponentiated coefficient} - 1.0) \times 100 \quad (5)$$

Relative Strength of Relationship

The standardized coefficient will assist in the determination the relative strength of each variable. The standardized coefficient is an indicator of the number of standard deviations of change in a dependent variable associated with a 1 standard deviation change in the independent variable. Menard (1995) depicts a step by step process to compute the standardized coefficients using SPSS. It is summarized as follows:

1. b : Find the un-standardized logistic regression coefficient
2. R^2 : Use the predicted value of Y to calculate R^2

3. Use the predicted value of Y to calculate the predicted value of logit(Y), using the

$$\text{equation: } \text{logit}(\hat{Y}) = \ln[\hat{Y}/(1 - \hat{Y})]$$

4. $S_{\text{logit}(\hat{Y})}$: Calculate descriptive statistics for logit (\hat{Y}), including the standard deviation.

5. S_x : Calculate the standard deviation of all independent variables.

6. Compute the following for each independent variable:

$$b_{YX}^* = (b_{YX})(S_x) / \sqrt{S_{\text{logit}(\hat{Y})}^2 / R^2}$$

The final value computed in number 6 above can be interpreted as: for an increase of 1 standard deviation in independent variable x there will be a b_{YX}^* standard deviation increase (+)/decrease (-) in the dependent variable Y. So for example if $b_{YX}^* = .591$ then for every 1 standard deviation of x there will be an associated increase of .591 standard deviation in Y.

Menard (1995) suggests the following relative strengths for b_{YX}^* values:

Weak: 0 to .3

Moderate: .3 to .6

Strong, .7 to 1

Menard (1995) states the standardized coefficient will produce a more accurate picture than the un-standardized logistic regression coefficients. He postulates that the un-standardized logistic regression coefficients are more reliable for categorical variables. In essence, the standardized coefficient should be viewed as a relative ranking, not taken as an absolute value. If a variable has a $b_{YX}^* = .7$ and another has a $b_{YX}^* = .1$ then we know that one variable has a much stronger significance relative to the other, a smaller range between the two values might not as

conclusive. The standardized coefficient provides a measure of magnitude of the effects of the independent variables. The preceding sections have examined the significance, strength and predictive accuracy of the proposed logistic regression model and its variables. The final question to consider is: Are the assumptions of the model completely satisfied?

The final logistic regression diagnostics phase is to determine whether the model exhibits any significant violations of logistic regression assumptions. This includes issues related to biased coefficients, inefficient estimates and invalid statistical inferences. Specification error can lead to biased logistic regression coefficients; the model may be utilizing coefficients that are overestimated or underestimated. This can be caused by omitting material variables or including immaterial variables during the analysis. Immaterial variables can result in an increase in the model's standard error of the parameter estimates. Omitting material variables from the model will lead to false importance and relative strength of the remaining variables in the model. Another specification error that can lead to an erroneous model is observed when the model's $\text{logit}(Y)$ function exhibits non-linearity. This occurs when the change of $\text{logit}(Y)$ for a one unit change in the independent variable is not constant; it is strongly associated to the value of the dependent variable. The logit function requires a linear relationship between the dependent and independent variables. Co linearity occurs when the independent variables exhibit correlation amongst each other. This interaction effect is another potential issue in logistic regression analysis. If the change in the dependent variable associated with a one unit change in the independent variable is related to the value of another independent variable then this behavior is non-additive and it violates the assumptions of the model. The regression coefficient estimates and their standard error values will be affected. While Co linearity may indicate a biased coefficient, the degree of correlation between the variables will be the primary consideration

with regards to its impact. Numerical problems such as zero cell count for categorical data can also cause model assumption violations. When no data exists for a cell in a categorical independent variable there can be issues with the model. This is not a problem for continuous variables. The logistic regression analysis must yield residuals for the logi transformation that have an average equal to zero, be randomly distributed around the mean without pattern, and be normally distributed. This paper will now present an overview of a portion of a case study to illustrate the usefulness of logistic regression.

Step by Step Example – Dengue Fever Study

Dengue fever is an infectious disease that researchers wish to ascertain a model that predicts the risk of being infected using relevant predictors. All SAS computational results for this portion of the study are shown in Figures 2 and 3. A two-stage stratified random sample of 196 persons was selected from an area known to have a recent epidemic of the fever (Kleinbaum, 1998). 57 people of the sample were determined to already have the disease. The objective of this regression analysis is to identify risk factors associated with Dengue fever and create a regression model of risk prediction factors. The statistical software used for the study was SAS. The study used Dengue fever status (DENGUE) as the dependent variable for the predicted outcome with values of 1 for yes and 0 for no. Subject ID, AGE, MOSNET and SECTOR were the independent variables. MOSNET was defined as an indicator of whether or not mosquito netting was used by the subject (0=yes; 1=no) and SECTOR was the geographic sector the subject resided in. They divided the overall region into 5 sectors (1-5). SECTOR was treated as categorical variable and 4 dummy variables were created using sector 5 as the reference group. The 4 dummy variables were SECTOR1, SECTOR2, SECTOR3 and SECTOR4.

Each sector was defined by study researcher as

$$\text{If sector X then SECTORX} = 1, \text{ else } 0 \quad (6)$$

where

X is the sector number (1,2,3,4),

Y = dependent variable = DENGUE.

They determined the base logit function as follows

$$\begin{aligned} \log it[Y = 1] = & \beta_0 + \beta_1(AGE) + \beta_2(MOSNET) + \\ & \beta_3(SECTOR1) + \beta_4(SECTOR2) + \\ & \beta_5(SECTOR3) + \beta_6(SECTOR4) \end{aligned} \quad (7)$$

where

Y is the dependent variable (1 is yes, 2 is no)

The study author computed the coefficients using the SAS statistical software ($\hat{\beta}_0$ through $\hat{\beta}_6$).

The final logit model was as

$$\begin{aligned} \log it[Y = 1] = & -1.9001 + 0.243(AGE) + 0.3335(MOSNET) - \\ & 2.2200(SECTOR1) - 0.6589(SECTOR2) + \\ & 0.8121(SECTOR3) + 0.5310(SECTOR4) \end{aligned} \quad (8)$$

MOSNET Significance

The study author then estimated the odds of someone contracting Dengue fever when using mosquito netting versus those who don't use netting.

The odds ratio for the MOSNET variable was represented as

$$\hat{OR}_{(MOSNET=1 \text{ vs. } MOSNET=0 | age, sector)} \neq \beta_2. \quad (9)$$

The adjusted odds ratio (exponentiated) was

$$\hat{OR}_{(MOSNET=1 \text{ vs. } MOSNET=0|age,sector)} = e^{0.3335} = 1.396. \quad (10)$$

Using the computed standard error computed by the SAS statistical software, Kleinbaum (1998)

found the standard error for β_2 is 1.2718.

They were then able to compute the 95% confidence interval for e^{β_2} as

$$\exp[\beta_2 \pm 1.96(\text{Std Error of } \beta_2)] = \exp[0.3335 \pm 1.96(1.2718)] \quad (11)$$

Equation 11 results indicated that the upper confidence limit was 16.89 and the lower confidence limit was 0.115.

The null hypothesis was stated as

$$H_0: (\text{adjusted odds ratio}), e^{0.3335} = 1, \quad (12)$$

$$H_a: e^{0.3335} \neq 1 \quad (13)$$

where

Wald (chi-square) statistic for testing null hypothesis for β_2 was 0.0688,

p-value at 95% confidence was 0.7931.

At this point, the analysis indicated that the odds of contracting the fever were about 1.4 times higher for someone who does not use mosquito netting. The Wald statistic is not statistically significant. The wide range for the 95% confidence interval as well as the inclusion of the null value of 1 within the range indicates they should reject the null hypothesis. It can therefore be concluded that the subject's usage of mosquito netting does not present a statistically significant affect on the probability (risk) of contracting the fever.

To further strengthen the study's conclusion about the statistical insignificance of the mosquito netting usage by the subjects (MOSNET, β_2), the author utilized the log likelihood statistic (-2LL).

The null hypothesis was stated as

$$H_0: \beta_2 = 0, \quad (14)$$

$$H_a: \beta_2 \neq 0 \quad (15)$$

The -2LL for the full model was computed by SAS to be 203.706 and 203.778 without the MOSNET variable.

The likelihood ratio for model comparison with and without MOSNET was

$$\begin{aligned} &(-2LL \text{ with MOSNET}) \text{ minus } (-2LL \text{ without MOSNET}) \quad (16) \\ &= 203.778 - 203.706 = 0.072 \end{aligned}$$

By using the same significance values as in equations 14 and 15, with one degree of freedom, chi square indicates that this is not a statistically significant difference. The subject's usage of mosquito netting (MOSNET) does not present a statistically significant affect on the probability (risk) of contracting the fever (dependent variable, DENGUE).

AGE Significance

This study also investigated the affect of age as it related to risk of contracting Dengue fever controlling for MOSNET and SECTOR. Since AGE is continuous they needed to compare a larger range of age difference between 2 people so they compared 2 people with 5 yr age difference. One year difference in age isn't useful, but a 5 year difference will is more descriptive.

The adjusted odds ratio was found to be

$$\hat{OR}_{(AGE1-AGE0=5|MOSNET,sector)} \neq e^{5(0.0243)} = 1.13. \quad (17)$$

Kleinbaum (1998) found the standard error for β_1 was 0.0091.

Further computation for the 95% confidence interval for e^{β_1} found

$$\exp[5\beta_1 \pm 1.96(5(\text{Std Error of } \beta_1))] = \exp[5(0.0243) \pm 1.96(5)(0.0091)] \quad (18)$$

Equation 18 results indicated that the upper confidence limit was 1.03 and the lower confidence limit was 1.23.

The null hypothesis was stated as

$$H_0: (\text{adjusted odds ratio}), e^{5(0.0243)} = 1, \quad (19)$$

$$H_a: e^{5(0.0243)} \neq 1 \quad (20)$$

where

Wald (chi-square) statistic for testing null hypothesis for β_1 was 7.1778,

p-value at 95% confidence was 0.0074

Since the confidence interval does not contain the null tested value of 1, a 5 years AGE difference between 2 people is statistically significant with regards to being infected with the fever. The odds ratio of 1.13 indicated that the significance was small. The log likelihood analysis for age coefficient was not presented in the study. The author indicated that the results of the log likelihood for AGE supported the conclusions about age referenced above. Table 2 reflects the study's of risk of contracting the fever versus age significance. They found that as age difference increases, the odds ratio increases, therefore the significance of the AGE variable becomes has a more statistical significance.

Significance of Interaction of MOSNET and AGE(MSA)

The study also considered the interaction of 2 variables, in this case MOSNET and AGE.

The logit model was adjusted and computed as

$$\begin{aligned} \log it[Y = 1] = & \beta_0 + \beta_1(AGE) + \beta_2(MOSNET) + \\ & \beta_3(SECTOR1) + \beta_4(SECTOR2) + \\ & \beta_5(SECTOR3) + \beta_6(SECTOR4) + \beta_7(MSA) \end{aligned} \quad (21)$$

where

Y is the dependent variable (1 is yes, 0 is no),

MSA = MOSNET * AGE

The study author computed the coefficients using the SAS statistical software ($\hat{\beta}_0$ through $\hat{\beta}_7$).

The final logit model with the new variable (MSA) considered was found to be

$$\begin{aligned} \log it[Y = 1] = & -0.8080 - 0.00434(AGE) - 0.8043(MOSNET) - \\ & 2.2929(SECTOR1) - 0.6813(SECTOR2) + \\ & 0.8153(SECTOR3) + 0.5115(SECTOR4) + 0.0306(MSA) \end{aligned} \quad (22)$$

The author then presented the affect of MSA on the odds ratio for the MOSNET (mosquito netting usage) variable.

The new adjusted odds ratio for MOSNET considering interaction with AGE was

$$\hat{OR}_{(MOSNET=1vs.MOSNET=0)_{age,sec\ tor}} = \exp [\beta_2 + \beta_7(AGE)] \quad (23)$$

$$\hat{OR}_{(MOSNET=1vs.MOSNET=0)_{age,sec\ tor}} = e^{[-.8043 + 0.0306(age)]} \quad (24)$$

The value of the odds ratio when considering the MSA interaction variable will be dependent upon the AGE difference selected as depicted previously during the analysis of the age affect on fever contraction risk. The author highlights has concluded that AGE is an effect modifier of the relationship of AGE with MOSNET variables.

In Table 3, the age modifying effect is illustrated through comparative analysis of the age and the MSA interaction variable odds ratio. It depicts that the odds (risk) of getting the fever at age 20 for someone who does not use mosquito netting is .83 times that of someone at age 20 who uses mosquito netting. He postulates that the likelihood of contracting the fever when mosquito netting is not used (MOSNET=1) increases with age. For example, at age 40 a person not using mosquito netting is 1.52 times more likely to contract the fever than someone at age 40 that does use mosquito netting. The SAS computation as stated by Kleinbaum (1998) found that the standard error for β_2 was 1.2718.

They computed the 95% confidence interval for $e^{\beta_2 + \beta_7(AGE)}$ and found to be

$$95\% \text{ confidence interval} = e^{(\hat{L} \pm 1.96 S_{\hat{L}})} \quad (25)$$

$$95\% \text{ confidence} = e^{[\beta_2 + \beta_7(AGE) \pm 1.96(\text{Std Error of } \beta_2 + \beta_7(AGE))]} \quad (26)$$

where

$$S_{\hat{L}} = \sqrt{\widehat{VAR}(\hat{L})},$$

$$\hat{L} = \hat{\beta}_2 + \hat{\beta}_7(age),$$

$$\widehat{VAR}(\hat{L}) = \widehat{VAR}(\hat{\beta}_2) + (age)^2 \widehat{VAR}(\hat{\beta}_7) + 2(age) \widehat{COV}(\hat{\beta}_2, \hat{\beta}_7),$$

$$\widehat{VAR}(\hat{\beta}_2) = 2.7004, \widehat{VAR}(\hat{\beta}_7) = 0.001399, \widehat{COV}(\hat{\beta}_2, \hat{\beta}_7) = -0.0435.$$

The study then reflected on the equations 20 and 21 with an example.

At age 40 they found

$$\hat{L} = -0.8043 + 0.0306 + 0.0306(40) = 0.4197 \quad (27)$$

$$\widehat{VAR}(\hat{L}) = 2.7004 + (40)^2 (0.001399) + 2(40)(-0.0435) = 1.4588 \quad (28)$$

$$S_L = \text{SQRT}(1.4588) = 1.2078 \quad (29)$$

Confidence Level computations were conducted for the example.

95% confidence interval for the adjusted odds ratio

$$\text{Adjusted odds ratio} = e^{\beta_2 + \beta_7(AGE)} \quad (30)$$

$$\text{Adjusted odds} = \exp(\hat{L} \pm 1.96 S_{\hat{L}}) = \exp[0.4197 \pm (1.96)(1.2078)] \quad (31)$$

Equation 31 results indicated that the upper confidence limit was 0.14 and the lower confidence limit was 16.23. The author concluded that the confidence interval was wide and thus at age 40 the risk estimate for contracting the disease was not reliable. They performed analysis at various selected age groups and found outcomes consistent with the results mentioned in this example (age = 40). The data and computational analysis was not fully noted in the author's paper. Finally the author presented the Log likelihood ratio testing for the MSA interaction model against the model without the interaction.

Interaction Model Significance

The study concludes with a comparison of the interaction model with the non-interaction model.

The non-interaction model (less MSA) was re-stated

$$\begin{aligned} \log it[Y = 1] = & \beta_0 + \beta_1(AGE) + \beta_2(MOSNET) + \\ & \beta_3(SECTOR1) + \beta_4(SECTOR2) + \\ & \beta_5(SECTOR3) + \beta_6(SECTOR4) \end{aligned} \quad (32)$$

The model with interaction (less MSA) was re-stated

$$\begin{aligned} \log it[Y = 1] = & \beta_0 + \beta_1(AGE) + \beta_2(MOSNET) + \\ & \beta_3(SECTOR1) + \beta_4(SECTOR2) + \\ & \beta_5(SECTOR3) + \beta_6(SECTOR4) + \beta_7(MSA) \end{aligned} \quad (33)$$

Equations 32 and 33 were used to determine the -2LL test statistic was examined to determine whether the difference between the two models was statistically significant.

The computation result was summarized as

$$\text{-2LL difference} = (\text{-2LL without Interaction}) - (\text{-2LL MSA Interaction}) \quad (34)$$

$$\text{-2LL difference} = 203.706 - 202.995 = 0.711 \quad (35)$$

Given the Null Hypothesis as

$$H_0: \beta_7 = 0, \text{ (non- interaction without MSA variable)} \quad (36)$$

$$H_a: \beta_7 \neq 0, \text{ (interaction with MSA variable)} \quad (37)$$

The study researcher assumed a chi-square degree of freedom of 1. The difference of 0.711 is not significant at 95% confidence and therefore they failed to reject the null hypothesis. The non-interaction model should be used as the best alternative.

Conclusions

This paper has presented an overview of the logistic regression multivariate statistical analysis methods, its primary guiding principles and a research study that utilizes a few of the techniques. The Dengue study presented a reasonable overview of logistic regression methods with a few useful examples. This author concludes that logistic regression is quite robust, practical and relatively simple to use. The methodology is ideally suited to manufacturing, the clinical sciences as well as many analytical fields that require rigid pass/fail results. Logistic regression's relaxation of the more rigid normality and variance assumptions of OLS were found

to be advantageous. Often, typical real world data does not exhibit a normal distribution. This leads to complex and time consuming transformations to normalize the data. Overall, logistic regression is an excellent tool for a researcher who is attempting to utilize independent data (continuous, interval or dichotomous) in practical situations requiring a dichotomous dependent variable to appropriately predict the results.

References

- Chatterjee, S., Hadi, A. S., & Price, B. (2000). *Regression analysis by example* (3rd ed.). New York: Wiley-Interscience Publication.
- DeMaris, A. (1992). *Logit modeling: Practical applications* (Vol. 07-086). Newbury Park: Sage Publications.
- Dielman, T. (2001). *Applied regression analysis for business and economics* (3rd ed.). New York, NY: Duxbury Press.
- Greenhouse, J. B., Bromberg, J. A., & Fromm, D. (1995). An introduction to logistic regression with an application to the analysis of language recovery following a stroke. *Journal of Communication Disorders*, 28(3), 229-246.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis*. Upper Saddle River, NJ: Pearson, Prentice Hall.
- Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and other multivariable methods* (3rd ed.). New York, NY: Duxbury Press.
- Menard, S. (1995). *Applied logistic regression analysis* (Vol. 07-106). Thousand Oaks, CA: Sage Publications.
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47, 90-100.
- Pampel, F. C. (2000). *Logistic regression: A primer* (Vol. 07-132). Thousand Oaks, CA: Sage Publications.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73(364), pp. 699-705.

Table 1

Predictive Accuracy Example

Row (x_1, x_2, x_3)	Observed Y	Fitted Logit	Assigned Group
1	0	0.00	0
2	0	0.48	0
3	0	-0.12	0
.....			
9	0	0.52	1
....			
52	1	0.48	0
.....			
66	1	0.59	1

Table 2

Odds Ratio

$AGE_1 - AGE_0$	Odds Ratio	95% Confidence interval for Odds Ratio
5	1.13	(1.03, 1.23)
10	1.28	(1.07, 1.52)
15	1.44	(1.10, 1.88)
20	1.63	(1.14, 2.32)
25	1.84	(1.18, 2.87)
30	2.07	(1.21, 3.54)
35	2.34	(1.25, 4.37)
40	2.64	(1.30, 5.39)

Table 3

Odds Ratio and Age

AGE	Odds Ratio: $\exp[\beta_2 + \beta_7(AGE)]$
10	0.61
20	0.83
30	1.12
40	1.52
50	2.07
60	2.81

Table 4

Comparison of Regression Methods

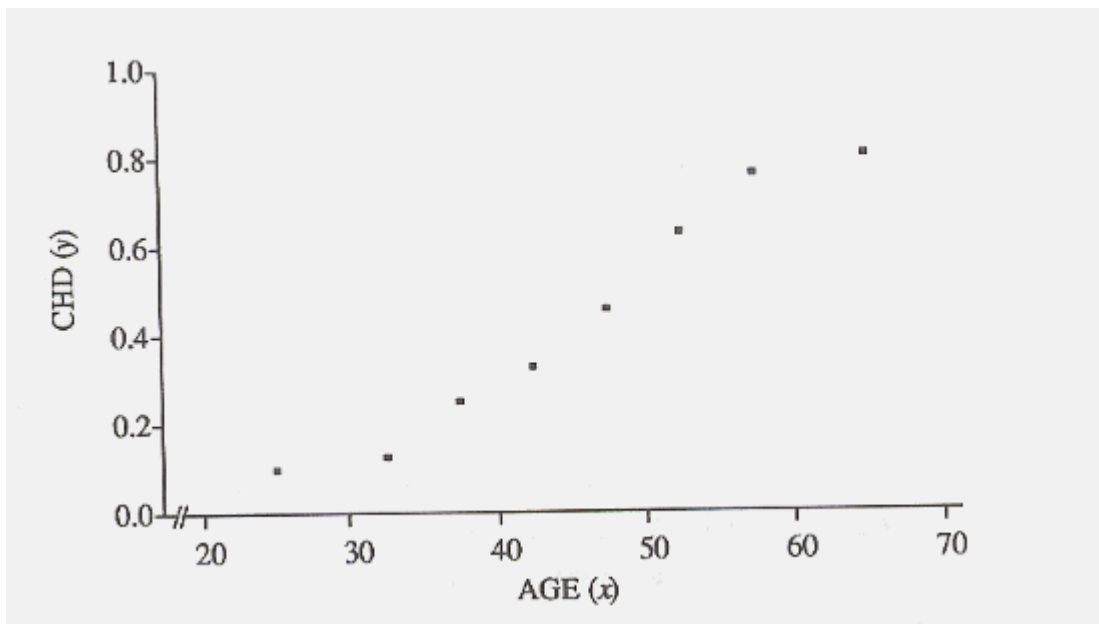
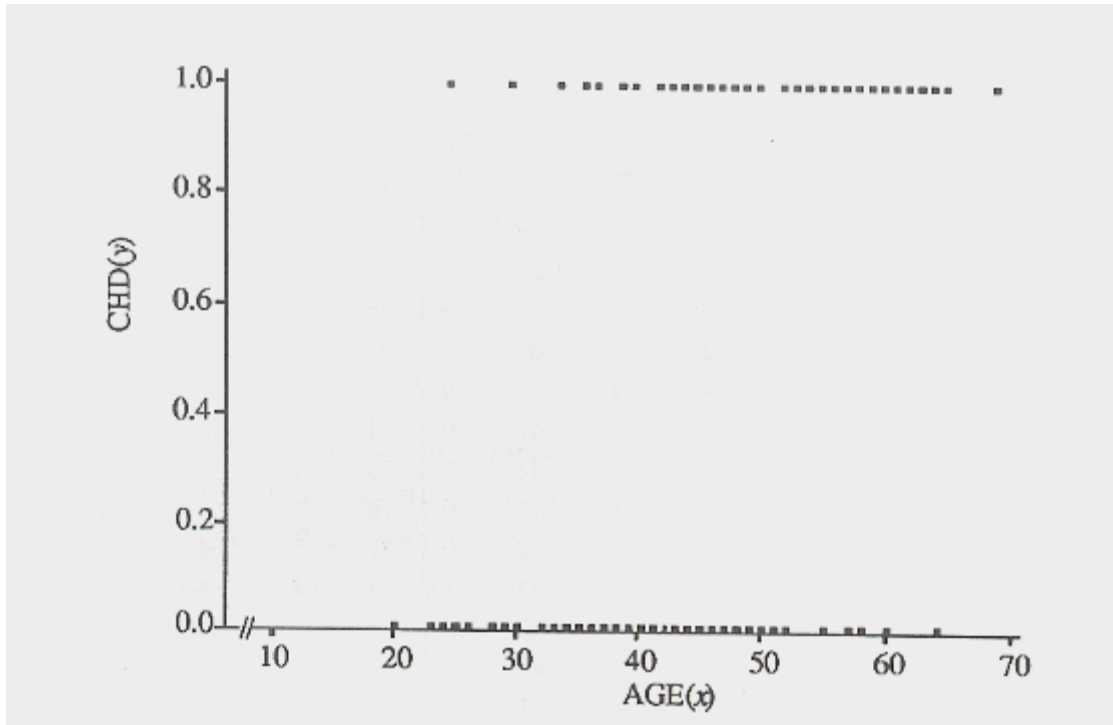
Multiple Regression	Logistic Regression
Total sum of squares	-2LL of base model
Error sum of squares	-2LL of proposed model
Regression sum of squares	Difference of -LL for base and proposed models
F test of model fit	Chi-square test of -2LL difference
Coefficient of Determination (R^2)	Pseudo R^2 measures

Figure Caption(s)

Figure 1. Dichotomous data without logit transformation (top), dichotomous data with logit transformation (bottom).

Figure 2. Dengue fever study, SAS output analysis response profile without MSA

Figure 3. Dengue fever study, SAS output analysis response profile with MSA



Response Profile						
Ordered Value		DENGUE		Count		
1		1		57		
2		2		139		

Criteria for Assessing Model Fit			
Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	238.329	217.706	.
SC	241.607	240.653	.
-2 LOG L	236.329	203.706	32.623 with 6 DF (p=0.0001)
Score	.	.	28.775 with 6 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates							
Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-1.9001	1.3254	2.0551	0.1517		0.150
AGE	1	0.0243	0.00906	7.1778	0.0074	0.252890	1.025
MOSNET	1	0.3335	1.2718	0.0688	0.7931	0.034212	1.396
SECTOR1	1	-2.2200	1.0723	4.2861	0.0384	-0.441811	0.109
SECTOR2	1	-0.6589	0.5536	1.4164	0.2340	-0.142513	0.517
SECTOR3	1	0.8121	0.4750	2.9235	0.0873	0.173824	2.253
SECTOR4	1	0.5310	0.4502	1.3911	0.2382	0.121456	1.701

The LOGISTIC Procedure
Response Profile

Ordered Value	DENGUE	Count
1	1	57
2	2	139

Criteria for Assessing Model Fit

Criterion	Intercept		Chi-Square for Covariates
	Intercept Only	Intercept and Covariates	
AIC	238.329	218.995	.
SC	241.607	245.220	.
-2 LOG L Score	236.329	202.995	33.334 with 7 DF (p=0.0001) 29.492 with 7 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-0.8080	1.6311	0.2454	0.6203		0.446
AGE	1	-0.00434	0.0362	0.0143	0.9048	-0.045185	0.996
MOSNET	1	-0.8043	1.6433	0.2396	0.6245	-0.082505	0.447
SECTOR1	1	-2.2929	1.0804	4.5042	0.0338	-0.456309	0.101
SECTOR2	1	-0.6813	0.5541	1.5118	0.2189	-0.147362	0.506
SECTOR3	1	0.8153	0.4756	2.9388	0.0865	0.174497	2.260
SECTOR4	1	0.5115	0.4515	1.2830	0.2573	0.116992	1.668
MSA	1	0.0306	0.0374	0.6689	0.4134	0.316912	1.031