PROJECT REPORT
ON

"Model building using Logistic Regression & Comparison with regression tree model

For Partial fulfilment of
B.Tech Degree in Computer Science& Engineering

Submitted by
Arnab Banerjee
111cs0118
Department of Computer Science& Engineering
NIT Rourkela

Under the guidance of
Mrs. Vandita Srivastava
Scientist 'SE'

Submitted At:

Geoinformatics Department
Indian Institute of Remote Sensing (IIRS)
Indian Space Research Organization (ISRO)
Dehradun 248001, Uttarakhand, India
May 2014-July 2014

Abstract

Logistic Regression is one of the most fundamental methods used to fit a given set of data. Logistic Regression can be applied in a wide view of real-life applications such as predictions of climatic conditions such as temperature, humidity and others. Logistic Regression plays a major role in Machine Learning which is one of the recent and interesting arenas of Computer Science.

The aim of the report is to provide in-depth study and analysis of the model building strategies employed while applying logistic regression for model fit. Logistic regression demands a huge theoretical knowledge prior to model building. A good knowledge on probability is also required for efficient analysis. So the report concentrates on the literature involved in logistic regression along with the know-how of how to use Matlab in order to perform logistic regression. With that knowledge, we perform logistic regression modelling on two given datasets and also on the combined dataset. Testing and validation of the model is an important task before selecting any final model and is done after model building. The report finally compares the results of statistical Logistic regression model with the algorithmic Regression Tree model.

Acknowledgement

I am heartily thankful to Indian Institute of Remote Sensing, Dehradun and National Institute of Technology, Rourkela for providing me this opportunity.

I would like to express my deep gratitude to **Mrs Vandita Srivastava**, Scientist, SE, Geo-informatics Department, Indian Institute of Remote Sensing, for her guidance to complete the project successfully. Without her able guidance, I would not have been able to concentrate on the challenging project and complete it successfully.

I would like to convey my sincere gratitude and appreciation to my parents and my brother for their understanding and support throughout the course of the project.

Signature

Date:17/07/2014

Table of Contents

Page No.

1.0 Regression: A Brief Introduction

Regression is a technique which is used to model a set of data so that the information contained in the model could be further extrapolated for more information about the nature of data.

For example let X= {1,2,3} and Y={3,4,5}. Now suppose we need to model the data given above. If we observe carefully the above two sets satisfy the relation Y=X+2. After a relation is established we can predict value of Y given any value of X.

So, given any arbitrary dataset, regression can be efficiently used to predict relationship between the set of independent variables and a dependent variable.

Two of the most widely used models are discussed below:

1.1 Linear Model

Linear regression is the most widely used of all statistical techniques. It is the study of linear relationships between variables usually under the assumption that the errors are normally distributed(Hosmer,Lemeshow&Sturdivant,3rd ed).

A linear regression model takes the following form,

$$E[Y] = \beta_0+\beta_1*X_1+ \beta_2*X_2+\ldots+ \beta_n*X_n$$

Here Y is the dependent variable whereas X_1, $X_2\ldots$, X_n are the set of independent variables. E[Y] denotes the expected value of Y given X_1, $X_2\ldots$, X_n.

Modelling involves estimating the values of β_0, $\beta_1\ldots$ β_n so that the model fits the data best. That means the square of the error $Y_{observed} - E[Y]$ is minimized.

1.2 Generalized Regression Model

Logistic regression is one of the special models of Generalized Regression Models. If a non-linear relationship exists between the independent and the set of dependent variables then Generalized Regression Model is adopted. In GLM, a transformation on the independent variable is applied so that now the transformed version on the response variable has a linear relationship with the independent variables. Mathematically, the above statement can be represented as

$$F(y) = \beta_0+\beta_1*X_1+ \beta_2*X_2+\ldots+ \beta_n*X_n$$

The transformation F is called the link function. The following link functions are widely used for data modelling:

a) Logit function used in Logistic Regression
b) Probit function
c) Log function

In this report Logit function shall be covered in detail.

1.3 Non-Linear Model-Regression Trees

A tree is a non-linear data structure. A Regression tree is a decision tree which is constructed by divide and conquer greedy algorithm. Regression tree helps us to predict the value of independent variable, either continuous or categorical, given the set of predictor variables.

The basic algorithm[Chapter 3: Tree Based Regression] for constructing a regression tree is as follows

Algorithm Binary_Regression_Tree

Input: Set of n-data points

Output: A regression tree

```
If termination criterion THEN
      Create Leaf Node and assign it a Constant Value
      Return Leaf Node
ELSE
         Find Best Splitting Test s*
         Create node t with s*
         Left_Branch(t)= Binary_Regression_Tree(all data points belonging to s*)
         Right_Branch(t)=Binary_Regression_Tree(all data points not belonging to s*)
         Return t
ENDIF
```

The constant value which must be assigned to each leaf node is the **average of the outcome values** of the cases within the leaf node l. The previous statement is a result of a theorem which states that "The constant k which minimizes the expected value of the squared error is the mean value of the target variables". The proof of the above statement is given in the appendix.

Thus value assigned to each leaf node l is

$$k_l = \frac{\sum_{D_l} y_i}{n_l}$$

where D_l is the set of observations within the leaf l and n_l is the cardinality of D_l.

Each of the inner nodes will have two branches. The test data is split into two subsets depending upon some test on one of the predictor variable. The cases satisfying the test follow the left branch and the cases not satisfying the test follow the right branch.

Now, splitting of the data into two subsets is based on minimizing the error of the tree. The goal of the split is that it should maximize the decrease in the error of the tree resulting from this split.

Before we move further, let us understand error of a node t and thereby error of an entire tree T.

Error of node t is the average of the square differences between the Y-values of the data within node t and the node constant k_t.

Error of node t= $\frac{\sum_{D_t}(y_i - k_t)^2}{n_t}$

where n_t is the cardinality of D_t.

Error of tree T is the weighted average of the error in its leaves:

$$\text{Error of T}= \sum_{l \in T'} P(l) * Err(l) = \sum_{l \in T'} \frac{n_l}{n} * \frac{1}{n_l} * \sum_{D_l} (y_i - k_l)^2 = \frac{1}{n} \sum_{l \in T'} \sum_{D_l} (y_i - k_l)^2$$

where P(l) is the probability of a case falling into leaf l.
n is the total number of cases.
nl is the number of cases in leaf f
and T' is the set of leaves of the tree T.

Now, we define **error of a split s on a node t** as

Error(s,t)= $\frac{n_{t_l}}{n_t} * Err(t_l) + \frac{n_{t_r}}{n_t} * Err(t_r)$

where t_l is the left child of t defining a partition D_{tl} that contains the set of cases {<xi,yi>εD_t : xi->s} and n_{tl} is the cardinality of this set.
and t_r is the right child node of t defining a partition D_{tr} that contains the set of cases {<xi,yi>εD_t : xi $\not\rightarrow$ s) and n_{tr} is the cardinality of this set.

The best split s* is the split that **maximizes**

ΔErr(s,t)= Err(t)- Err(s,t)

The above is a greedy criterion based on which choices are made.

A frequently used criterion for stopping the recursion is to impose a minimum number of cases that once reached forces the termination of the algorithm. Another example of stopping criteria is to create a leaf if the error in the current node is below a fraction of the error in the root node.

2.0 Introduction to Logistic Regression

Suppose there is a company ‘X’ which deals in online shopping. A company normally maintains a database of its customers. So does ‘X’. Now, ‘X’ is in the process of selling 100 new items online which will replace previous items of the same category. For example the company sells books online.

Previously the books present in the children’s section were Book A, Book B, Book C. Now the new books that are going to replace those are Book A1, Book A2 and Book A3.

Now the company wants to know what is the probability that a customer ‘C ‘will buy a particular book?

Several factors may be taken into consideration such as 1) Gender of the customer 2) Age of the customer 3) Previously bought products of ‘X’ of particular type 4) Economic status of family 5) Number of websites visited where ‘X’ had given an advertisement.

Now, we consider a mathematical variable Y which denotes that customer bought a particular book. Suppose we code the variable as if the customer buys the book we say Y=1 and if the customer does not buy a book we say Y=0. Note that the variable Y is a dichotomous variable.

Now Y depends on the following factors or parameters:

a) Gender
b) Age
c) Previously bought products of ‘X’ of a particular type.
d) Economic status of family
e) Number of websites visited where ‘X’ had given an advertisement.

In the language of logistic regression, these parameters are called co-variates. Mathematically we can say that these 5 co-variates or independent variables will help in determining Y.

Now let us consider a situation in which the company asks its statisticians to list out the names of all those people whose probability of buying a new book would be more than 60%. Mathematically P(Y=1) should be greater than equal to 0.6. Then only the company would be sending e-mails informing those customers about the new products.

In order to perform such a task we perform logistic regression.

In general, Logistic Regression is applied to describe a relationship between a dependent variable Y and a set of independent variables. The independent variables can be continuous (Age, Economic status of family), dichotomous (Gender=0 denotes male while Gender=1 denotes female). The independent variable may be categorical as well.

Applications of Logistic Regression

a) Logistic Regression can be used to predict the videos one is most likely to watch in YouTube. The factors which may be taken into consideration is age, gender, country of residence, previous videos watched, titles of videos most frequently watched by the viewer.

b) Logistic Regression can be applied to predict temperature, humidity or other climatic factors. Given a set of data on temperature, logistic regression fits the data into a model. This model can be used for predicting temperature for future days.

c) Logistic Regression is applied in remote sensing based applications. The following are the research papers which use Logistic Regression models for prediction.

a) Forest Cover Dynamics Analysis and prediction modelling using Logistic Regression model by Rakesh Kumar, S. Nandy, Reshu Agarwal, S.P.S. Kushwaha.

b) Remote sensing and GIS-based landslide hazard analysis and cross-validation using multivariate logistic regression model on three test areas in Malaysia by Biswajeet Pradhan.

c) Logistic regression modelling of rock glacier and glacier distribution: Topographic and climatic controls in the semi-arid Andes by Alexander Brenning , Dario Trombotto.

3.0 Linear vs Logistic Regression

3.1 Differences

There are several differences between Linear Regression and Logistic Regression as mentioned below(Hosmer,Lemeshow,Sturdivant,3rd ed):

1) In linear regression the outcome variable is continuous while in logistic regression the outcome variable is dichotomous.

2) In linear regression the relationship between the dependent variable and the set of independent variables is linear while the same is not true for logistic Regression.

For example linear regression can be applied for the first relationship but will fail for the second relationship.

$$Y=\beta_0+\beta_1*X_1+\beta_2*X_2+\ldots+\beta_n*X_n\ldots.1)$$

$$Y=e^{\beta 0 + \beta 1*X1 + \beta 2*X2+\ldots+\beta n*Xn}\ldots..2)$$

In the first function we can see that the dependent variable is a straight line function of each of Xi's keeping other variables constant. We cannot say so for the next function.

3) In linear regression the error distribution is normal while in logistic regression is binomial. In linear regression the error distribution is normal with mean=0 and variance=σ2. In logistic regression the error distribution is binomial.

3.2 Assumptions of Logistic Regression

1) All the observations must be independent of each other i.e. a co-variate (predictor variable) pattern ={X1', X2', X3', X4'} and another co-variate pattern={X1'', X2'', X3'', X4''} do not have any relationship between them. All the cases must be independent. This assumption must be met while collecting data.

2) No important variables are excluded and no extraneous variables are included.

3) The independent variables should not be a linear combination of each other i.e. they should not be linearly correlated.

4.0 Mathematical Model

After reading the introduction to logistic regression, it is understandable that we are more interested in finding out probabilities of outcome having a specified value. To model such a situation, a function which ranges between 0 to 1 is required. One of the better choices(Wikipedia: http://en.wikipedia.org/wiki/Logistic_regression) is

$$F(t)= \frac{e^t}{1+e^t}$$

Domain of the function is $(-\infty, +\infty)$ and the range is (0,1). Thus this function can fit a probability function well.

Let us suppose $t= \beta_0+\beta_1*X_1+\beta_2*X_2+\ldots+\beta_n*X_n$

Replacing t in F(t) we obtain

$$F(X_1,X_2,X_3,\ldots,X_n) = \frac{e^{\beta 0+\beta 1*X1+\beta 2*X2+..+\beta n*Xn}}{1+e^{\beta 0+\beta 1*X1+\beta 2*X2+..+\beta n*Xn}}$$

We consider the left side of the above equation as the probability of Y, the dependent variable, to be equal to 1.

Thus,

$$P(Y=1) = \frac{e^{\beta 0+\beta 1*X1+\beta 2*X2+..+\beta n*Xn}}{1+e^{\beta 0+\beta 1*X1+\beta 2*X2+..+\beta n*Xn}}$$

The most important function in a logistic regression is the logit function which is defined as logit(x)=ln(f(x)/1-f(x)).

So, the mathematical model now becomes,

$$\text{logit}(Y=1) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta 0+\beta 1*X1+\beta 2*X2+\ldots+\beta n*Xn$$

Now let us analyse the second and the third term in the above equation.

Since any of the Xi's may be a continuous variable the third term may range from $-\infty$ to $+\infty$.

So the second term must also have the same range. On analysing the second term we conclude the following,

Since probability varies between 0 and 1, the ratio of probabilities may vary from 0 to ∞. On applying the natural logarithm the range now varies from -∞ to +∞.Thus the mathematical model is correct and can be used for logistic regression.

5.0 Type of Variables

The dependent variable, termed as outcome variable, must be dichotomous(Hosmer, Lemeshow,Sturdivant,3rd ed).

The set of independent variables, termed as predictor variables, could be of the following types:

5.1 Continuous variable

A continuous variable can assume any value between (--∞,+∞). For example the continuous variable AGE or WEIGHT can vary from (0, ∞).

5.2 Dichotomous variable

A dichotomous variable can only assume values 0 and 1. For example GENDER can be coded as 0 for males and 1 for females.

5.3 Categorical variable

In statistics, a categorical variable is a variable that can take one only value among a fixed set of values. Commonly, each of the possible values of a categorical variable is referred to as a level. In logistic regression also we refer to the different possible values as levels.

For example Colour may be a categorical variable with level 0 for red, 1 for green and 2 for blue.

In logistic regression a categorical variable is coded using design variables. To code k different levels of a categorical variable we need k-1 design variables.

For example the Colour variable may be coded with two design variables as k=3 here. Let these two variables be D1 and D2.

Colour	D1	D2
Red	0	0
Green	1	0
Blue	0	1

In order to represent colour in the logit equation, we will use D1 and D2 along with their coefficients.

6.0 Simple Logistic Regression Model

This section is a summary of Chapter 1 of the book *Applied Logistic Regression* by Hosmer, Lemeshow.
In simple logistic Regression we have only one independent variable and one dependent variable. Thus the logit equation of simple logistic regression is;

$$\textbf{logit(Y=1)} = \beta_0+\beta_1*X$$

or

$$\pi(X)=\frac{e^{\beta_0+\beta_1*X}}{1+e^{\beta_0+\beta_1*X}}$$

Here $\pi(x)$ represents the probability that Y=1 for a given X.

Now X can be a continuous variable or a dichotomous variable or a categorical variable with more than two categories.

6.1 Fitting of the logistic regression model:

This term means to find the values of β_0 and β_1 such that it maximizes the probability of obtaining the observed data. In order to find the best fitting model, we need to apply the concept of **maximum likelihood** so that we obtain the coefficient that maximizes the probability.

Likelihood function (lβ) is a function which expresses the probability of the observed data as a function of the unknown parameters β_0 and β_1. Since the value of the likelihood function must be maximum (probability must be maximum) for the values of β_0 and β_1, we can find the values by differentiating the likelihood function with respect to β_0 and β_1.

The maximum likelihood function on a set of observed values$((x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n))$ is

$$l(\beta)=\prod_{i=1}^{n}\pi(x_i)^{y_i}*(1-\pi(x_i))^{1-y_i}$$

Here $\pi(xi)$ represents the probability that $y_i=1$ given x_i.

On differentiating the above function with respect to β_0 and β_1 we obtain the following two equations,

$\sum [y_i-\pi(x_i)]= 0$

$\sum x_i [y_i - \pi(x_i)] = 0$

The above two equations are non-linear in the unknown parameters and thus require iterative solutions for it. **The coefficients can be obtained using Matlab.**

6.2 Testing the significance of the model

After obtaining unknown coefficients we test for the significance for it in the model. Now in case of simple logistic regression the model is

$$\textbf{logit(Y=1)} = \beta_0 + \beta_1 * X$$

We note that β_0 and β_1 are already known.

In order to test the significance of the model, we need to devise a statistic which would help us predict the significance. We need to ask a question that whether the variable included in the model tell us more about the outcome variable than when the variable is not included.

Introduction of the statistical variable **G= Deviance (without the concerned variable in model) – Deviance(with the concerned variable in model)** can help in this regard.

Now Deviance of a model is defined as

$$\text{Deviance} = -2 * \ln\left(\frac{likelihood\_of\_fitted\_model}{likelihood\_of\_saturated\_model}\right)$$

Such a test is known as the likelihood ratio test.

We note that likelihood of saturated model will be equals to 1 since conceptually a saturated model is a model which contains as many variables as required so as to correctly predict the outcome for each and every outcome.

Thus deviance becomes

Deviance= -2*ln(likelihood of the fitted model)

Thus replacing the deviance value in the equation for G,

G= 2*(log-likelihood of the fitted model with the concerned variables - log-likelihood of the fitted model without the concerned variables)

G will follow a **chi-square statistic** with the degree of freedom depending upon the hypothesis employed.

Matlab returns the value of deviance when logistic regression is performed using a model and thus the value of G can be found out quite easily.

In the case of simple logistic regression, we consider the hypothesis that $\beta_1=0$. In other words we are saying that the variable X does not provide a better model as compared to a model without X. Now since we have only one variable X, a model without X will only contain the constant term β_0.

To find the value of G in such a case, a formula exists for calculating G.

$$\mathbf{G= 2*(log\text{-}likelihood\ of\ model\ with\ X - [n_1 *ln(n_1)+n_0*ln(n_0) - n*ln(n)])}$$

n_1 represents the number of outcomes with Y=1.
n_0 represents the number of outcomes with Y=0.
n represents the total number of outcomes.

Log likelihood of the model with X can be obtained from Matlab. When we perform **glmfit()** on the given model, it returns **deviance as an output** argument and thus

log likelihood = -deviance/2

After calculating the value of G, we need to find the p-value for the corresponding G value using chi-square statistics. The degree of freedom in this case will be 1. **In general, the degree of freedom involved in the chi square distribution is equal to number of variables excluded.**

If the p-value is less than or equal to 0.05 then we reject the hypothesis that $\beta_1=0$ is rejected and we say that the model involving X is a better model than the one not involving X.

Matlab also provides the p-values if we perform logistic regression using glmfit(). This p-value can also be used to determine whether X is a significant variable or not. A value less than equal to 0.05 results in rejection of null hypothesis that $\beta_1=0$ and thus X is considered to be significant.

We can use any of the above two methods to assess the significance of the model in case of simple logistic regression.

6.3 Confidence Interval (CI) Estimation:

The basis of construction of the confidence interval estimators follows the same statistical theory that was used to formulate the test of the significance of the model. The confidence interval of the slope β_1 and the intercept β_0 can also be calculated. The confidence interval estimation is based on the respective Wald's test and are sometimes referred as **Wald-Based Confidence Intervals**.

The end-points of a 100(1-α) % confidence interval for the slope coefficient β_1 are

$$\text{Range}= \beta_1 \pm z_{1-\alpha/2} * SE(\beta_1)$$

The end-points of a 100(1-α) % confidence interval for the intercept β_0 are

Range= $\beta_0 \pm z_{1-\alpha/2} * SE(\beta_0)$

where $z_{1-\alpha/2}$ is the upper $100(1-\alpha/2)$ % point from the standard normal distribution and SE (β) represents the standard error of the parameter estimation.

Usually **α=5%** i.e. we want to predict **with 95% confidence** where the actual value of the parameter shall line. For α=5%, **$z_{1-\alpha/2}$=1.96** and the values of SE for each parameter can be obtained using glmfit() function in Matlab. Thus using the above two formulas we can formulate the interval between which we can say that the actual value of the parameters will lie with 95% confidence.

7.0 Multiple Logistic Regression Analysis

This section is a summary of Chapter 2 of the book *Applied Logistic Regression* by Hosmer, Lemeshow.
In case of multiple logistic regression, more than one independent variable(co-variate) are present in the model and each may be continuous, dichotomous or categorical.

Thus the logit equation can thus be extended to the following equation,

$$\textbf{logit}(Y=1) = \beta_0 + \beta_1 * X1 + \beta_2 * X2 + \ldots + \beta_p * X_p$$

So as we can see that we have **p** predictor variables (independent variables).

The equation may undergo some minor changes depending upon the type of variable:

a) Continuous: If a variable X is continuous just simply include it in the equation, multiplied by a beta term, without changing anything.

b) Dichotomous: If a variable X is dichotomous just simply include it in the equation, multiplied by a beta term, without changing anything.

c) Categorical with k levels: If a variable X is categorical we need to replace the variable X with k-1 design variables in the model. The method of assigning values to design variables is already discussed.

Suppose we have an equation $\text{logit}(Y=1) = \beta_0 + \beta_1 * X1 + \beta_2 * X2$.

Suppose X2 is a categorical variable with levels 3. That means we need 3-1=2 design variables. Suppose they are D_1 and D_2. Now we code the design variables as is explained earlier and after we have coded D_1 and D_2 simply replace X2 with those two variables.

Now the equation becomes $\text{logit}(Y=1) = \beta_0 + \beta_1 * X1 + \beta_2 * D_1 + \beta_3 * D2$.

Thus the two equations become different once we involve design variables. Note that the coefficient β_2 of X_2 and the coefficient β_2 of D1 are different.

7.1 Fitting of the logistic regression model

As performed earlier in case of simple logistic regression we need to find the values of **β**= { $\beta_0, \beta_1, \beta_{2,\ldots,}\beta_p$}

Again we define the likelihood function and on differentiating the likelihood function by each of the beta terms and equating it to zero we get p+1 equations and on solving those equations we get values for **β**= { $\beta_0,\beta_1, \beta_{2,\ldots,}\beta_p$}.

The likelihood function in case of multiple regression model is

$$l(\beta)=\prod_{i=1}^{n}\pi(X_i)^{y_i} *(1-\pi(X_i))^{1-y_i}$$

where $\pi(X_i)$ represents the probability that Y=1 given X_i.

Mathematically $\pi(X_i)=\dfrac{e^{\beta_0+\beta_1*X_1+\beta_2*X_2+\ldots+\beta_p*X_p}}{1+e^{\beta_0+\beta_1*X_1+\beta_2*X_2+\ldots+\beta_p*X_p}}$

Thus, in a similar way we can find the fit a multiple logistic regression model. **We can obtain the values of the coefficients using Matlab.**

7.2 Testing for the significance of the model

The first null hypothesis that we assume is each $\beta_i=0$ except $i\neq0$. In other words all the independent variables are insignificant. In order to check the null hypothesis we obtain the deviance of the model containing all the variables. This value can again be obtained from Matlab.

The log-likelihood is equal to –deviance/2. Again applying the formula,

G= 2*(log-likelihood of model with X – [n_1 *ln(n_1)+n_0*ln(n_0) – n*ln(n)])

The distribution of G is chi-square with ‘p’ degrees of freedom since according to the null hypothesis we consider all the ‘p’ co-variates as zero. If the p-value associated with this hypothesis test is less than equal to 0.05 we can say that at least one of the β_i is not zero.

After fitting the model we check the p-value for each coefficient’s significance. We can get the p-values as output from the **glmfit()** function in Matlab.

If the p-value associated with each of the coefficient is below a certain threshold level, generally 0.05, we say that the null hypothesis that the corresponding coefficient is zero is false and thus rejected. So that variable is included for further analysis and on the other hand

if null hypothesis is correct we perform the following test in order to confirm their exclusion from the model.

Let us suppose that after checking the p-value associated with each variable, we confirm that β_i and β_j is of little significance in the model and thus is a candidate for exclusion from the model. Now we need to ascertain whether the model containing β_i and β_j is no better than the model not containing β_i and β_j. **If it is found** that the model with the two variables included is no better than in which those two variables are excluded then we reject those two variables otherwise we keep them in the model. Note that **we try to exclude all those variables which are not significant**.

To ascertain whether β_i and β_j is to be excluded or not we need to perform the likelihood ratio test. The value of G will be calculated as

G=2*(log-likelihood of the model containing β_i and β_j – log-likelihood of the model without β_i and β_j)

In order to obtain the log likelihood without β_i and β_j just find the deviance of the model not containing β_i and β_j but keeping other variables in the model. The deviance can be found out using glmfit() function in Matlab. From the deviance we can calculate the value of G. Note that now since we are considering the hypothesis that both β_i and β_j are equal to zero the degree of freedom for the chi-square distribution of G has **degree of freedom equal to 2**.

So in this way we remove all those variables which are not significant and keep all those which are either clinically significant or statistically significant.

7.3 Confidence Interval Estimation

The confidence interval for each of the p+1 parameters can be obtained as explained in the simple logistic regression model.

8.0 Logistic Regression Model: Interpretation

Once we are done with fitting the logistic Regression model we must concentrate on the **interpretation of the beta parameters** that are obtained after applying logistic regression. The interpretation of the beta parameters is the main area of focus and it is the reason why we are applying logistic regression. "What do the estimated coefficients in the model tell us about the research questions that motivated the study"? is our primary query.

In the logistic regression model, the slope coefficient is the change in the logit corresponding to a one unit change in the independent variable i.e.
β_1=logit(Y=1|x+1) - logit(Y=1|x) in case of simple logistic regression.

The interpretation of the independent variable's slope i.e. its associated beta value depends upon the nature of the independent variable. The interpretation changes with nature. The following section involves interpretation as per the variable's nature.

8.1 Dichotomous variable: In this the variable is coded as 0 or 1. As an example let us consider a simple logistic regression model logit(Y=1) = $\beta_0+\beta_1*X1$ where X1 is a dichotomous variable. Now logit(Y=1|X1=1) = $\beta_0+\beta_1$ and logit(Y=1|X1=0) = β_0.
Thus

$$\textbf{logit(Y=1|X1=1) – logit(Y=1|X1=0) = } \boldsymbol{\beta_1}$$

The above equation can now be expressed as **ln(odds of Y=1|X1=1) – ln(odds of Y=1|X1=0) = β_1** and can be further modified as

$$\ln\left(\frac{odds\_of\_Y=1\,|\,X1=1}{odds\_of\_Y=1\,|\,X1=0}\right)=\beta_1$$

Now odds of Y=1|X1=1/ odds of Y=1|X1=0 can be defined as odds ratio of Y=1 given X1.

Therefore,

$$\textbf{Odds Ratio} = e^{\beta 1}$$

The odds ratio is the most widely used measure of association as it can help in meaningful prediction. The above formula can be interpreted that the odds of Y=1 given X1=1 is $e^{\beta 1}$ times the odds of Y=1 given X1=0.

For example let us consider an example which relates the presence of myopia of a child with the independent variable "Mother or Father also has myopia". Now the independent variable can be coded as "Mother or Father has myopia"=1 if either the father or mother also had myopia or on the other hand "Mother or Father has myopia"=0 if both father and mother did not have myopia. Now for example studies show that the Odds Ratio comes out to be 3. It can be interpreted as **"The odds of the child suffering of myopia if either one of their parents also suffers from it is 3 times the odds of the child suffering of myopia if both of their parent did not have myopia"**.

Now we can also find out the confidence interval for the odds ratio. Since confidence interval of 100(1-α) % for the log odds ratio β_1 is CI= $\beta_1 \pm z_{1-\alpha/2}*SE(\beta_1)$, the Confidence Interval for the odds ratio can be obtained just by exponentiation of the end points of the CI of β_1.

For a dichotomous variable coding it with 0 and 1 is not mandatory but should be followed because if some other coding let us say 'a' and 'b' are used then it starts playing a major role in all the estimations. Thus to avoid any unnecessary calculation we use 0 and 1.

8.2 Continuous variable

Suppose we have a logit model, logit(Y=1) = $\beta_0+\beta_1*X_1$ where X_1 is a continuous variable. Now logit(Y=1|X_1=a+10) = $\beta_0+\beta_1*(a+10)$ and logit(Y=1|X_1=a) = $\beta_0+\beta_1*a$. Thus we have

$$\textbf{logit(Y=1|X}_1\textbf{=a+10)-logit(Y=1|X}_1\textbf{=a) = 10*}\ \boldsymbol{\beta_1}$$

As calculated in case of dichotomous variable, in a similar way we can say that **Odds Ratio= $e^{10*\beta 1}$**. In general we can say that for 'c' units of change the Odds Ratio= **$e^{c*\beta 1}$**. As we can see the odds ratio is **independent of the value of a**. Thus, we can say that odds of Y=1 when X1 is increased to a+10 is **$e^{10*\beta 1}$** times the odds of Y=1 when X_1 is equal to a.

Confidence interval (CI) for odds ratio = $\exp(c*\beta_1+ z_{1-\alpha/2}*|c|*SE(\beta_1))$

8.3 Categorical variable

This is the most important type of variable when it comes to interpretation of the associated beta parameters. In case of a categorical variable we have 'k' levels. In order to interpret these types of variables, we consider any one of the levels as the reference level and find the odds ratio of another level of the same variable with respect to the reference level.

Suppose we have the following logit model logit(Y=1) = $\beta_0+\beta_1*X_1$ where X_1 is a categorical variable with 3 levels. Those 3 levels can be coded by two design variables D_1 and D_2. We code D1 and D2 in the following way,

Level 0	D_1=0	D_2=0
Level 1	D_1=1	D_2=0
Level 2	D_1=0	D_2=1

Thus as explained earlier we replace the X_1 term with D_1 and D_2 in the above equation. The equation now transforms into logit(Y=1) = $\beta_0+\beta_1*D_1+\beta_2*D_2$.

Now for first level, D_1=0 and D_2=0, so logit(Y=1) = β_0
for second level, D_1=1 and D_2=0, so logit(Y=1) = $\beta_0+\beta_1$
for third level, D_1=0 and D_2=1, so logit(Y=1) = $\beta_0+\beta_2$

Since our reference level is the first one, we subtract the logit of first level from that of both second and third level considered individually.

Thus in this way we arrive at the following conclusion, Odds ratio involving second and reference level= $e^{\beta 1}$. Similarly Odds ratio involving third and reference level = $e^{\beta 2}$.

We can thus say that the odds of Y=1 when X1=Level 1 is $e^{\beta 1}$ times the odds of Y=1 when X1=Level 0 or the reference level. Similarly the odds of Y=1 when X1=Level 2 is $e^{\beta 2}$ times the odds of Y=1 when X1=Level 0 or the reference group.

In a similar way we can deal with categorical variables with k number of levels.

9.0 Statistical Interaction&Confounding

Statistical Interaction

Suppose we have a logit model equation logit(Y=1) = $\beta_0+\beta_1*X1+\beta_2*X2$. To make the model fit better we must check for statistical interaction between the co-variates. The interaction term involving X1 and X2 is X1*X2. Thus we need to include the interaction term to the model for checking its significance. Thus the model becomes **logit(Y=1) = $\beta_0+\beta_1*X1+\beta_2*X2+\beta_3*X1*X2$.** Once we fit this model we will get the beta values and also the p-value for its Wald's statistic. If the p-value for its Wald's statistic is less than or equal to 0.05 we reject the **null hypothesis that $\beta_3=0$** and thus the interaction term is significant and needs to be included in the model.

We also check whether the model containing {X1, X2, X1*X2} is a better model than {X1,X2}. To do this, we use the likelihood ratio test. The value of G associated with it is G=2*(log-likelihood with variable X1*X2 – log-likelihood without variable X1*X2}. G will follow a chi-square statistic with degree of freedom equal to 1. If the p-value associated with G is less than 0.05 then we are sure to include X1*X2 in the model.

If there are more than two variables originally in the model then each and every **pairs of statistical interaction** must be considered and checked for its significance but **term by term**. For example if the original model had three independent variables namely X1,X2,X3 then the following three interaction terms must be checked for its significance term by term.

i) {X1*X2}
ii) {X1*X3}
iii) {X2*X3}

Statistical Confounding

Confounding variables are those variables which have not been included in the model but it has an effect on the dependent variable.
For example, we want to test whether men or women are taller. The independent variable is gender and the dependent variable is height. Anything else that effects height is a confounding variable in this study. An obvious example would be age. The more different age groups you measure, the more variation you will see in heights. Nationality would be

another. If you are lucky, confounding variables will only increase variance but if you are unlucky, they could introduce bias.
So, we need to check for the confounding variables in a model and controlling confounding variables reduces its effect drastically. In the above example age, which is a confounding variable, can be controlled by measuring heights of people whose age is the same.

10.0 Model Building Strategy

This section is a summary of Chapter 4 of the book *Applied Logistic Regression* by Hosmer, Lemeshow.
In this part, we shall discuss the sequential steps that need to be taken in order to build a model which fits the data in the best way possible.

a) **Uni-variable analysis of each independent variable**: This step is performed in order to identify important co-variates. In this step each and every independent variable is used to model the outcome data one by one. The variables must be selected one at a time. The p-value obtained for each of the variable is observed. The rule of thumb in this step is that reject all those variables whose p-value is greater than 0.25 and not the traditional value of 0.05. We select the value as high as 0.25 because we may be excluding those variables which are not significant when considered alone but may become significant with some other variable(Statistical Confounding). In order to remain safe p value as high as 0.25 is selected.

An important concept which should be discussed is that whenever the independent variable under analysis is a categorical variable with k levels then there would be k beta parameters involved in the uni-variable analysis. Of the k beta parameters, some of the variables may have p-value greater than 0.25 and some may have less. So, in this situation we go through all the p-values and even if one of the p-value is less than 0.25 we decide to keep the categorical variable in the model. Additionally we should also check for log likelihood ratio test in order to check if excluding the categorical variable does not lead to degradation of the model. If the model including the categorical variable is no better than the model excluding it we reject the categorical variable.

b) **Fit a multiple logistic regression model using the variables selected in step a)**: Now after step a) we have all those variables which are expected to play a significant role in the model building process. We now collect all those variables and fit a multiple logistic regression model with all those variables. After fitting the model, we check for the significance of each of the independent variables at the traditional level of p-value less than equal to 0.05. Suppose we observe a variable X_j having a p value>0.05. Thus we decide to remove X_j from the model. Now we check whether the new model is better than the previous model. To check we need to apply log-likelihood ratio test. The value of G shall be defined as

G= 2*(log-likelihood of model containing X_j - log-likelihood of model not containing X_j)

Note that all other variables except X_j will remain as it is in the model. If the p-value associated with G is greater than 0.05 then it means that the model containing X_j is no better than the model without X_j and thus we exclude X_j from the model. Repeat this process of

deleting, refitting and verifying until it appears all the important variables are included while others are excluded.

Once we are done with it, we fit all those variables which were excluded in the first step. assess the joint significance of the of the variables that were not selected in the first step. This step is necessary since it helps to identify the confounding variables. The model obtained after this step is called the preliminary main effects model.

c) **Check the assumption of linearity for each continuous co-variate:**

We need to check whether the logit model is linear in the continuous variables. To do this, we need draw the scatter plot diagram. The scatter plot diagram can be obtained from softwares like SYSTAT or STATA. If the smoothed scatter plot looks linear then the relationship is linear and if the plot looks non-linear then we have to find a function such that the logit of the outcome is linear in the function of the independent variable. At the end of this step, we obtain the main effects model.

d) **Check for interaction terms:**

Now that we are ready with our main effects model, we now proceed towards checking interaction terms that may be significant. Create a list of those pairs of independent variables that have some scientific basis to interact with each other. This list may or may not include all the possible pairs.

Add the interaction term one by one to the main effects model and check for the significance of the term by using the p-value obtained from the Wald's test and also by the log-likelihood ratio test. Those interaction terms which have a p-value less than 0.05 and whose inclusion in the main effects model leads to a better model (log-likelihood test), should be included in the model. After having done this, we get our preliminary final model

e) **Assessing goodness of fit of the preliminary final model:**

Now that we have selected our model we need to assess how much correct the model is. In other words, we need to check whether the probabilities predicted by the obtained model accurately or nearly accurately reflect the true outcome of the data.

Hosmer and Lemeshow test(Hosmer,Lemeshow,Sturdivant,3rd ed, p. 157) can be applied on the model in order to assess the fit of the model. Hosmer and Lemeshow test is based on grouping the estimated probabilities. The grouping is done based on percentiles of the estimated probabilities. Another way of grouping is based on fixed values of the estimated probabilities. Since the first grouping strategy is better than the second one explanation of the second method is not provided.

In the first step, we group the estimated probabilities such that the first group will have $n' = n/10$ smallest estimated probabilities. The second group will have the next greatest n' estimated probabilities and the last group will contain the largest n' estimated probabilities.

Then we build the following table:

Decile	Cut point	Obs(Y=1)	Exp(Y=1)	Obs(Y=0)	Exp(Y=0)	Total
1	Cut-point$_1$	No. of obs with Y=1 in 1st decile	Sum of probabilities for 1st decile	No. of obs with Y=0 in 1st decile	Sum of(1-probability) in 1st decile	Total number of obs in 1st decile
2	Cut-point$_2$	No. of obs with Y=1 in 2nd decile	Sum of probabilities for 2nd decile	No. of obs with Y=0 in 2nd decile	Sum of(1-probability) in 2nd decile	Total number of obs in 2nd decile
3	Cut-point$_3$	No. of obs with Y=1 in 3rd decile	Sum of probabilities for 3rd decile	No. of obs with Y=0 in 3rd decile	Sum of(1-probability) in 3rd decile	Total number of obs in 3rd decile
4	Cut-point$_4$	No. of obs with Y=1 in 4th decile	Sum of probabilities for 4th decile	No. of obs with Y=0 in 4th decile	Sum of(1-probability) in 4th decile	Total number of obs in 4th decile
5	Cut-point$_5$	No. of obs with Y=1 in 5th decile	Sum of probabilities for 5th decile	No. of obs with Y=0 in 5th decile	Sum of(1-probability) in 5th decile	Total number of obs in 5th decile
6	Cut-point$_6$	No. of obs with Y=1 in 6th decile	Sum of probabilities for 6th decile	No. of obs with Y=0 in 6th decile	Sum of(1-probability) in 6th decile	Total number of obs in 6th decile
7	Cut-point$_7$	No. of obs with Y=1 in 7th decile	Sum of probabilities for 7th decile	No. of obs with Y=0 in 7th decile	Sum of(1-probability) in 7th decile	Total number of obs in 7th decile
8	Cut-point$_8$	No. of obs with Y=1 in 8th decile	Sum of probabilities for 8th decile	No. of obs with Y=0 in 8th decile	Sum of(1-probability) in 8th decile	Total number of obs in 8th decile
9	Cut-point$_9$	No. of obs with Y=1 in 9th decile	Sum of probabilities for 9th decile	No. of obs with Y=0 in 9th decile	Sum of(1-probability) in 9th decile	Total number of obs in 9th decile
10	Cut-point$_{10}$	No. of obs with Y=1 in 10th decile	Sum of probabilities for 10th decile	No. of obs with Y=0 in 10th decile	Sum of(1-probability) in 10th decile	Total number of obs in 10th decile

The Hosmer-Lemeshow statistic of goodness of fit is defined as

$$\hat{C} = \sum_{k=1}^{g} \left[\frac{(o_{1k} - \hat{e}_{1k})^2}{\hat{e}_{1k}} + \frac{(o_{0k} - \hat{e}_{0k})^2}{\hat{e}_{0k}} \right]$$

Here g represents the number of deciles group. Here o_{1k} represents the number of outcomes with Y=1 in the k^{th} decile. $\hat{e}_{1k}$ represents the sum of probabilities in the k^{th} decile. o_{ok} represents the number of outcomes with Y=0 in the k^{th} decile. $\hat{e}_{0k}$ represents the sum of (1-probabilities) in the k^{th} decile.

The third column in the above mentioned table represents the values of o_{1k} . The fourth column represents $\hat{e}_{1k}$. The fifth column represents o_{ok} while the sixth column represents $\hat{e}_{0k}$.

So the value of $\hat{C}$ can be obtained by adding the value (((third column-fourth column)^2)/ third column)+(((fifth column-sixth column)^2)/fifth column) for each of the values of k varying from 1 to number of groups g.

Now $\hat{C}$ will follow a chi-square distribution with degree of freedom=g-2.

If the p-value associated with $\hat{C}$ is high then the model provides a good fit.

11. Standard Example: GLOW Study

11.1 Introduction

One of the most standard data sets on which we can apply logistic regression to understand its methodology is GLOW500 data. GLOW stands for Global Longitudinal Study of Osteoporosis in Women over 55 years of age being coordinated at the Centre for Outcomes Research(COR) at the University of Massachusetts/Worcester. The major goals of the study are to use the data to provide insights into the management of fracture risk, patient experience with prevention and treatment of fractures and distribution of risk factors among older women on an international scale over the follow up period. The data set can be obtained from the following link:- *http://www.umass.edu/statdata/statdata/data/glow/index.html*

11.2 Description of variables

The code sheet for the variables included in the GLOW study is provided below

Variable	Description	Codes/Values	Name
1	Identification Code	1-n	SUB_ID
2	Study site	1-6	SITE_ID
3	Physician ID code	128 codes	PHY_ID
4	History of prior fracture	1=yes 0=no	PRIORFRAC

5	Age at enrolment	Years	AGE
6	Weight at enrolment	Kilograms	WEIGHT
7	Height at enrolment	Centimetres	HEIGHT
8	Body mass Index	kg/m^2	BMI
9	Menopause before age 45	1=yes 0=no	PREMENO
10	Mother had hip fracture	1=yes 0=no	MOMFRAC
11	Arms needed to stand from chair	1=yes 0=no	ARMASSIST
12	Former or current smoker	1=yes 0=no	SMOKE
13	Self-reported risk of fracture	1=less than others of same age 2=same as others of same age 3=greater than others of same age	RATERISK
14	Any fracture in first year	1=yes 0=no	FRACTURE

The variables 1, 2, 3 are of no use in our analysis. These variables may be used to uniquely identify a patient and the physician who treated the patient. Study site is also irrelevant in our case.

So the list of independent variables is:

a) History of prior fracture: Whether the patient has suffered fracture earlier. This variable is coded as PRIORFRAC.
b) Age at enrolment: Age of the patient during admission. This variable is coded as AGE.
c) Height of enrolment: Height of the patient during admission. This variable is coded as HEIGHT.
d) Body mass index: This variable is coded as BMI.
e) Menopause before age 45: This variable is coded as PREMENO.
f) Mother had hip fracture: This variable is coded as MOMFRAC.
g) Arms needed to stand from a chair: This variable is coded as ARMASSIST.
h) Former or current smoker- This variable is coded as SMOKE.
i) Self-Reported risk of fracture- This variable is coded as RATERISK

The dependent variable is any Fracture in first year of follow up.

We must note that the variables PRIORFRAC, PREMENO, MOMFRAC, ARMASSIST and SMOKE are all dichotomous variables.

The variables AGE, WEIGHT, HEIGHT and BMI are continuous variables while the only categorical variable is RATERISK.

RATERISK is a categorical variable with 3 levels. Level 1 represents that the patient feels that the risk involved is less than others of the same age. Level 2 represents that the patient feels that risk is same as others of the same age and the last level i.e. level 3 represents that patient feels that risk is more than other patients of the same age.
To represent RATERISK, we need 2 design variables $RATE\_RISK_1$ and $RATE\_RISK_2$ which shall be coded in the following manner:

Level	$RATE\_RISK_1$	$RATE\_RISK_2$
1	0	0
2	1	0
3	0	1

11.3 Selection of significant covariates and model building

a) The first step as mentioned in the model building strategy is to perform uni-variable analysis of all the co-variates one by one. The result of the analysis is shown in the following table. Note that the coefficients obtained are from the model containing only that independent variable.

Variable	Coeff.	Standard_Error	p-value
AGE	0.053	0.0116	<0.001
WEIGHT	−0.0052	0.0064	0.415
HEIGHT	−0.052	0.0171	0.002
BMI	0.006	0.0172	0.738
PRIORFRAC	1.064	0.2231	<0.001
MOMFRAC	0.661	0.2592	0.022
PREMENO	0.051	0.2810	0.0845
ARMASSIST	0.709	0.2098	0.001
SMOKE	−0.308	0.4358	0.469
RATE_RISK			
$RATE\_RISK_1$	0.546	0.2664	0.003
$RATE\_RISK_2$	0.909	0.2711	0.001

We reject all those variables whose p-value is greater than 0.25. We observe that there are four such variables WEIGHT, BMI, PREMENO and SMOKE.

b) We now fit a multi-variable model which contains all those variables which were not rejected in the first step.

The result of the multi-variable model fit is shown in the next table.

Variable	Coeff.	Standard Error	p-value
AGE	0.034	0.0130	0.008
HEIGHT	-0.044	0.0183	0.016
PRIORFRAC	0.645	02461	0.009
MOMFRAC	0.621	0.3070	0.043
ARMASSIST	0.446	0.2328	0.056
$RATE\_RISK_1$	0.422	0.2792	0.131
$RATE\_RISK_2$	0.707	0.2934	0.016
Constant	2.709	3.2299	0.402

We note that the p-value of RATE_RISK1 is 0.131 indicating that the variable is not significant while on the other hand RATE_RISK2 is shown to be significant. Note that since RATE_RISK1 and RATE_RISK2 are design variables of a single categorical variable, RATE_RISK, we have to either include both the design variables or exclude both. To check whether RATE_RISK1 and RATE_RISK2 provide a better model we perform the log-likelihood ratio test. On performing the log likelihood test we find that the two variables play a significant role and thus we keep both RATE_RISK1 and RATE_RISK2 in the model for further analysis.

We note that while RATE_RISK is not a confounder the variable ARMASSIST is an important co-variate. No other variables are candidates for exclusion and thus we continue with our current model.

c) We now add all those variables which were excluded in the first step one by one and check whether the model obtained by including those variables are better than the model without it. In this example, the variables excluded in step 1 were WEIGHT, BMI, PREMENO and SMOKE. We see that on addition of the each of the variables, its coefficient did not become significant. Thus we permanently exclude those variables from the model.

Since the p-value of $RATE\_RISK_1$ is not significant, one idea is to combine level 1 and level 2 so that now a new design variable $RATE\_RISK_3$ is now a dichotomous variable. Now $RATE\_RISK_3$=0 denotes that the rate of risk is either less than or equal to other women and $RATE\_RISK_3$=1 denotes that rate of risk is greater than other women. So, the results of fitting the multi variable model with $RATE\_RISK_1$ and

$RATE\_RISK_2$ $RATE\_RISK_3$ are replaced by given below,

Variable	Coeff.	Standard Error	p-value
AGE	0.033	0.0129	0.01
HEIGHT	-0.046	0.0181	0.011
PRIORFRAC	0.664	0.2452	0.007
MOMFRAC	0.664	0.3056	0.030
ARMASSIST	0.473	0.2313	0.041
$RATE\_RISK_3$	0.458	0.2381	0.054
Constant	3.407	3.1770	0.284

Now we check for linearity between the logit and the continuous variables. Here we have only two continuous variables AGE and HEIGHT. In this particular example both the continuous variables have a linear relationship with the logits of the outcome.

d) The next step is to search for statistical interactions between the variables in the main effects model. We may consider all such pairs of variables which may be statistically significant or clinically significant. In this example, we consider each and every possible pairs and find out its significance in the main effects model.

Interaction	Log-likelihood	G	p
Main effects model	-254.9089		
AGE*HEIGHT	-254.8422	0.13	0.715
AGE*PRIORFRAC	-252.3921	5.03	0.025
AGE*MOMFRAC	-254.8395	0.14	0.710
AGE*ARMASSIST	-254.8358	0.15	0.702
$AGE*RATE\_RISK_3$	-254.3857	1.05	0.306
HEIGHT*PRIORFRAC	-254.8024	0.21	0.645
HEIGHT*MOMFRAC	-253.7043	2.41	0.121
HEIGHT*ARMASSIST	-254.1112	1.60	0.207
$HEIGHT*RATE\_RISK_3$	-254.4218	0.97	0.324
PRIORFRAC*MOMFRAC	-253.5093	2.80	0.094
PRIORFRAC*ARMASSIST	-254.7962	0.23	0.635
$PRIORFRAC*RATE\_RISK_3$	-254.8476	0.12	0.726
MOMFRAC*ARMASSIST	-252.5179	4.78	0.029
$MOMFRAC*RATE\_RISK_3$	-254.6423	0.53	0.465
$ARMASSIST*RATE\_RISK_3$	-253.7923	2.23	0.135

Here we see that only AGE*PRIORFRAC and MOMFRAC*ARMASSIST and PRIORFRAC*MOMFRAC interactions have significance at the 10% level.
We may consider traditional level of 0.05 but since we PRIORFRAC*MOMFRAC is significant at 10% level; we choose that as our limit. Thus we will include these three interactions terms in the model.
Now we fit the main model along with the interactions. The results obtained on fitting the model is

Variables	Coeff.	Std Error	p-value
AGE	0.058	0.0166	0.000
HEIGHT	-0.049	0.0184	0.008
PRIORFRAC	4.598	1.8780	0.014
MOMFRAC	1.472	0.4229	0.000
ARMASSIST	0.626	0.2538	0.014
$RATE\_RISK_3$	0.474	0.2410	0.049
AGE*PRIORFRAC	-0.053	0.0259	0.040
PRIORFRAC*MOMFRAC	-0.847	0.6475	0.191
MOMFRAC*ARMASSIST	-1.167	0.6168	0.058
Constant	1.959	3.3272	0.556

As we can observe that the interaction term PRIORFRAC*MOMFRAC has become insignificant. Thus we reject this interaction term and hence there will be two interaction terms in addition to the main model. The results of fitting the new model excluding the term PRIORFRAC*MOMFRAC is

Variable	Coeff.	Std Error	p-value
AGE	0.057	0.0165	0.001
HEIGHT	-0.047	0.0183	0.011
PRIORFRAC	4.612	1.8802	0.014
MOMFRAC	1.247	0.3930	0.002
ARMASSIST	0.644	0.2519	0.011
$RATE\_RISK_3$	0.469	0.2408	0.051
AGE*PRIORFRAC	-0.055	0.0259	0.033
MOMFRAC*ARMASSIST	-1.281	0.6230	0.040
Constant	1.717	3.3218	0.605

Now we have gone through all the procedures of model building and now we need to check how good the model fits. That means, how close is it to the observed data. In order to find this we need to find the value of Hosmer Lemeshow goodness of fit. First we need to construct the table in order to find its value.

Decile	Cut-point	Obs(Y=1)	Exp(Y=1)	Obs(Y=0)	Exp(Y=0)	Total
1	0.085	3	3.3	47	46.7	50
2	0.111	4	4.9	46	45.1	50
3	0.141	7	6.3	43	43.7	50
4	0.176	11	8.1	40	42.9	51
5	0.208	7	9.4	42	39.6	49
6	0.249	13	11.4	37	38.6	50
7	0.323	9	14.3	41	35.7	50
8	0.389	19	17.6	31	32.4	50
9	0.483	25	21.8	25	28.2	50
10	0.747	27	28.0	23	22.0	50

From the above table we can calculate the value of $\hat{C}$ which is this case is found out to be $\hat{C}$= 6.39. Now $\hat{C}$ will follow a chi-square distribution with degree of freedom=8. So the p-value associated with the $\hat{C}$= 6.39 is 0.603. So it indicates that the **model fits very well**.

Now having confirmed the model, we can write the corresponding logit equation as

logit(Y=1)=1.717+0.057*AGE-0.047*HEIGHT+4.612*PRIORFRAC+1.247*MOMFRAC +0.644*ARMASSIST+0.469*RATE_RISK$_3$-0.055*AGE*PRIORFRAC-1.281* MOMFRAC*ARMASSIST.

Converting into odds we get

(Odds of Y=1)= exp(1.717+0.057*AGE-0.047*HEIGHT+4.612*PRIORFRAC+1.247*MOMFRAC +0.644*ARMASSIST+0.469*RATE_RISK$_3$-0.055*AGE*PRIORFRAC-1.281* MOMFRAC*ARMASSIST)

Converting into probability we get

(Probability of Y=1) =(Odds of Y=1)/ (1+(Odds of Y=1))

11.4 Interpretation of Model Parameters

After framing the logit equation, interpretation of the model parameters is the next vital step.

a) AGE

In case of AGE, we observe that there is an interaction term with PRIOFRAC. So we need to provide a value of PRIORFRAC before we could estimate the effect of age.

Let us suppose age of a patient A is X and value of PRIORFRAC be P where P is either 0 or 1.

The logit equation corresponding to patient A will be

logit(Y=1) = 1.717+ 0.057*X-0.047*HEIGHT+4.612*PRIORFRAC+1.247*MOMFRAC
+0.644*ARMASSIST+0.469*RATE_RISK3-0.055*X*P-1.281*MOMFRAC*ARMASSIST

Now for another patient B whose age is X+10 and value of PRIORFRAC is Q where Q is either 0 or 1.

The logit equation corresponding to patient B will be
logit(Y=1|AGE=X+10,PRIORFRAC=Q) =1.717+ 0.057*(X+10)-0.047*HEIGHT
+4.612*PRIORFRAC+1.247*MOMFRAC+0.644*ARMASSIST+0.469*RATE_RISK3-
0.055*(X+10)*Q-1.281*MOMFRAC*ARMASSIST

Subtracting equation 1 from 2 we get

logit(Y=1|AGE=X+10,PRIOFRAC=Q)-logit(Y=1|AGE=X,PRIORFRAC=P) =
0.57+0.055*X*(P-Q)-0.55*Q

Thus the odds ratio OR becomes

$$OR= \exp(0.57+0.055*X*(P-Q)-0.55*Q)$$

Now when P=Q, Odds Ratio does not depend upon AGE otherwise the AGE of the patient plays a vital role in the calculation of Odds Ratio.

Mathematically the above Odds Ratio predicts that the Odds of the patient B having fracture in first year of follow up is exp(0.57+0.055*X*(P-Q)-0.55*Q) times the odds of patient A.

b)HEIGHT

The odds of a patient A having fracture within first year of follow up is 0.625 times the odds of another patient B where the age of patient A is 10 cm more than patient B. It also means that the odds of a fracture are 37.5% lower in patient A as compared to patient B. Note that since HEIGHT does not have any interaction term we can determine its effect without having to consider other variables.

c) PRIORFRAC:

Note we have an interaction term between PRIORFRAC and AGE.

Suppose a patient A whose PRIORFRAC value is P and AGE is X.
Another patient B whose PRIOFRAC value is Q and AGE is Y.

The logit equation of patient A is

logit(Y=1|PRIORFRAC=P,AGE=X) =1.717+0.057*X-0.047*HEIGHT+4.612*P
+1.247*MOMFRAC+0.644*ARMASSIST+0.469*RATE_RISK$_3$-0.055*X*P-1.281*
MOMFRAC*ARMASSIST

The logit equation of patient B is

logit(Y=1|PRIORFRAC=Q,AGE=Y) = 1.717+0.057*Y-0.047*HEIGHT +4.612*Q+1.247*MOMFRAC+0.644*ARMASSIST+0.469*RATE_RISK$_3$-0.055*Y*Q-1.281* MOMFRAC*ARMASSIST

Subtracting the first from the second we get

logit(Y=1|PRIORFRAC=Q,AGE=Y)-logit(Y=1|PRIORFRAC=P,AGE=X) =0.057*(Y-X)+0.055(Y*Q-X*P)+4.612*(Q-P)

Thus the OR=exp(0.057*(Y-X)+0.055(X*P-Y*Q)+ 4.612*(Q-P))

Suppose that the age of the two patients are same then X=Y

OR becomes exp(0.055*X*(Q-P)+4.612(Q-P))

Thus the odds of a patient B having fracture within first year of follow up is exp(0.055*X*(Q-P) +4.612(Q-P)) times that of patient A assuming the age of both of them are the same.

d) MOMFRAC

In the model MOMFRAC has an interaction term with ARMASSIST. Thus both the variables need to be considered for interpretation of MOMFRAC

Suppose a patient A has MOMFRAC=M and ARMASSIST= A and another patient B has MOMFRAC=N and ARMASSIST=B.

The logit equation for patient A is

logit(Y=1|MOMFRAC=M,ARMASSIST=A) = 1.717+0.057*AGE-0.047*HEIGHT +4.612*PRIORFRAC+1.247*M+0.644*A+0.469*RATE_RISK$_3$-0.055*AGE*PRIORFRAC - 1.281* M*A.

The logit equation for patient B is

logit(Y=1|MOMFRAC=N,ARMASSIST=B) = 1.717+0.057*AGE-0.047*HEIGHT +4.612*PRIORFRAC+1.247*N+0.644*B+0.469*RATE_RISK3-0.055*AGE*PRIORFRAC -1.281* N*B

Subtracting equation 1 from 2

logit(Y=1|MOMFRAC=N,ARMASSIST=B)-logit(Y=1|MOMFRAC=M,ARMASSIST=A)= 1.247*(N-M)-1.281*(N*B-M*A)+0.644*(B-A)

Thus the OR=exp(1.247*(N-M)-1.281*(N*B-M*A)+0.644(B-A))

If we consider A=B, then

OR= exp(1.247*(N-M)-1.281*A*(N-M))

Thus the odds of a patient B having fracture within first year of follow up is exp(1.247*(N-M)-1.281*A*(N-M)) times that of patient A assuming the value of ARMASSIST of both of them are the same.

e) ARMASSIST

We can deduce the OR for ARMASSIST from the logit equation obtained from MOMFRAC's interpretation. We note that

logit(Y=1|MOMFRAC=N,ARMASSIST=B)-logit(Y=1|MOMFRAC=M,ARMASSIST=A) = 1.247*(N-M)-1.281*(N*B-M*A)+0.644*(B-A)

Thus the OR=exp(1.247*(N-M)-1.281*(N*B-M*A)+0.644(B-A))

If we consider both the patient have the same value for MOMFRAC

Thus the OR for ARMASSIST is exp(-1.281*N*(B-A)+0.644(B-A))

Thus the odds of a patient B having fracture within first year of follow up is exp((-1.281*N*(B-A)+0.644(B-A)) times that of patient A assuming the value of MOMFRAC of both of them are the same.

f) $RATE\_RISK_3$

Since RATE_RISK3 is not involved in any interaction term we can directly get the odds ratio.

The odds ratio in this case is OR=exp(0.469)= 1.5984 the beta coefficient associated with RATE_RISK3 is 0.469.

It means the odds of a patient A having fracture within one year of follow up is 1.59 times of another patient B given that the patient A has reported a fracture risk greater than other women whereas patient B has reported a risk of fracture either less than or equal to another women. In other words the odds of patient A getting fractured is 59% higher as compared to patient B.

12.0 Model building of given dataset

12.1 Dataset 1

Details of variables in Dataset 1

a) Outcome variable: Y which is dichotomous. Y is vector consisting of 336 observations.
b) Independent variables: X_1, X_2, X_3, X_4 all of which are continuous and each of them containing 336 observations.

The following Matlab code finds the correlation matrix and plots the correlation among the predictor variables.

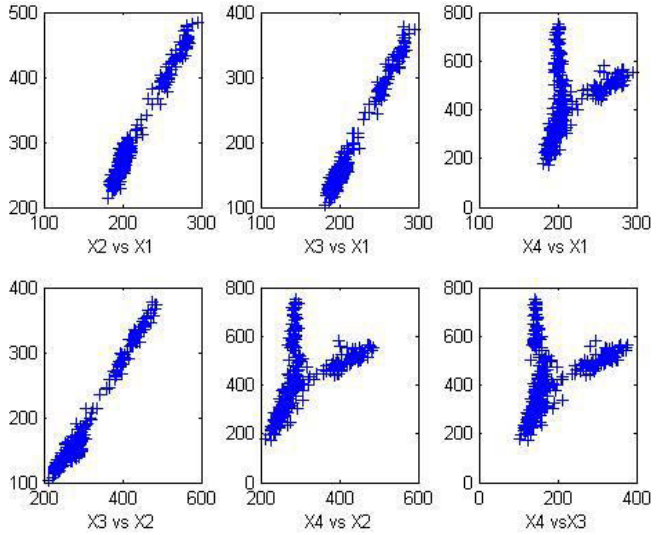
Matlab Code:

```
load datas.mat % the datas.mat file contains the dataset
X=[A B C D];% A B C D are the independent variables
M=corr(X) % M contains the correlation matrix of {A,B,C,D}
subplot(2,3,1);
plot(A,B,'+'); %plotting B vs A
xlabel('X2 vs X1');
subplot(2,3,2);
plot(A,C,'+'); %plotting C vs A
xlabel('X3 vs X1');
subplot(2,3,3);
plot(A,D,'+');  %plotting D vs A
xlabel('X4 vs X1');
subplot(2,3,4);
plot(B,C,'+');  %plotting C vs B
xlabel('X3 vs X2');
subplot(2,3,5);
plot(B,D,'+');  %plotting D vs B
xlabel('X4 vs X2');
subplot(2,3,6);
plot(C,D,'+');  %plotting D vs C
xlabel('X4 vsX3');
```

Output:

```
M =   1.0000   0.9853   0.9897   0.4052
      0.9853   1.0000   0.9853   0.5038
      0.9897   0.9853   1.0000   0.3718
```

0.4052 0.5038 0.3718 1.0000



From correlation matrix M the following deductions can be made

a) The correlation coefficient between X_2 and X_1 is 0.9853. The graph X_2 vs X_1 is almost linear indicating that these two variables are strongly correlated.

b) The correlation coefficient between X_3 and X_1 is 0.9897. The graph X_3 vs X_1 is almost linear indicating that these two variables are strongly correlated.

c) The correlation coefficient between X_3 and X_2 is 0.9853. The graph X_3 vs X_2 is almost linear indicating that these two variables are strongly correlated.

d) The variable X_4 is weakly or moderately correlated with the other variables.

Dataset 1 violates the third assumption of logistic regression which states that the independent variables must not be correlated.

Thus in order to apply logistic regression on this dataset, only one out of the three correlated variables must be used for modelling along with X_4 since introduction of other correlated variables makes the model redundant.

Thus the possible sets of independent variables are as follows:

a) X_1 ,X_4
b) X_2, X_4
c) X_3, X_4

Now the model building strategy is followed to predict correct model for each set of independent variables.

a) $\{X_1, X_4\}$

The results of uni-variable analysis are tabulated below

Uni-variable analysis of variable	Logit equation	β_0	β_1	p-value for β_0	p-value for β_1	Deviance	Log likelihood
X_1	logit(Y=1)= $\beta 0+\beta 1*X_1$	10.0818	-0.0501	1.9846e-09	1.4678e-09	376.6153	-188.3077
X4	logit(Y=1)= $\beta 0+\beta 1*X_4$	3.8266	-0.0102	1.0994e-14	2.5827e-17	352.4087	-176.2044

Since p-value of both β_0 and β_1 is less than 0.25 we keep it in the model for the next step which is bi-variable analysis of $\{X_1, X_4\}$.

The corresponding logit equation is $\text{logit}(Y=1) = \beta_0+\beta_1*X1+\beta_2*X4$. The results of fitting are tabulated below

Variable	β-value	p-value
X_1	-0.0315	5.3101e-05
X_4	-0.0070	3.0531e-09
Constant	9.1885	9.9902e-10

Deviance = 330.1877

Log-likelihood= -165.0938

After this step, we confirm whether the model $\{X_1, X_4\}$ is better than only $\{X_1\}$ or only $\{X_4\}$.

In case of {X1} and {X1,X4} we see that we are concerned about the hypothesis that β2=0. So,

log likelihood with variable X4= -165.0938
log likelihood without variable X4= -188.3077

So g=2*(-165.0938+188.3077)
= 46.4278

Under the hypothesis that β2 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is $P\{\chi 2(1)>46.4278) = <.001$. Thus the hypothesis that β2 is zero is rejected and we confirm that {X1,X4} is a better model that {X1} only.

In case of {X4} and {X1,X4} we see that we are concerned about the hypothesis that β1=0. So,

log likelihood with variable X1= -165.0938
log likelihood without variable X1= -176.2044

So g=2*(-165.0938+176.2044)
= 22.2212

Under the hypothesis that β1 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is P{χ2(1)>22.2212) = <0.05.

Thus the hypothesis that β1 is rejected and we confirm that {X1, X4} is a better model that {X4} only.

Next we consider an interaction term X_1*X_4.

The corresponding logit equation will be logit(Y=1) = β0+β1*X1+β2*X4+β3*X1*X4

The result of fitting the above model is tabulated below

Variable	β-value	p-value
X_1	0.7730	6.6251e-13
X_4	0.3634	7.9442e-12
X_1*X_4	-0.0019	3.0545e-12
Constant	-147.4110	1.5865e-12

Deviance = 234.5776

Log likelihood= -117.2888

Now under the hypothesis that β3=0 we obtain the corresponding value of g

g=2*(-117.2888+165.0938)
= 95.6100

Under the hypothesis that β3 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is P{χ2(1)>95.6100) = <0.001.

Thus the hypothesis that β3=0 is rejected and we confirm that {X1, X4, X1*X4} is a better model that {X1, X4} only.

b) $\{X_2, X_4\}$

The results of uni-variable analysis are tabulated below

Uni-variable analysis of variable	Logit equation	β_0	β_1	p-value for β_0	p-value for β_1	Deviance	Log likelihood
X_2	logit(Y=1)= $\beta 0+\beta 1*X_2$	6.7497	-0.0245	1.1961e-10	7.3027e-11	360.2780	- 180.1390
X4	logit(Y=1)= $\beta 0+\beta 1*X_4$	3.8266	-0.0102	1.0994e-14	2.5827e-17	352.4087	-176.2044

Since p-value of both β_0 and β_1 is less than 0.25 we keep it in the model for the next step which is bi-variable analysis of $\{X_1, X_4\}$.

The corresponding logit equation is logit(Y=1) = $\beta_0+\beta_1*X_2+\beta_2*X_4$. The results of fitting are tabulated below

Variable	β-value	p-value
X_1	-0.0138	5.8161e-05
X_4	-0.0062	1.2074e-05
Constant	6.3038	1.8022e-13

Deviance= 331.0544

Log-likelihood= - 165.5272

After this step, we confirm whether the model $\{X_2, X_4\}$ is better than only $\{X_2\}$ or only $\{X_4\}$.

In case of {X2} and {X2, X4} we see that we are concerned about the hypothesis that β2=0. So,

log likelihood with variable X4= -165.5272
log likelihood without variable X4= -180.1390

So g=2*(-165.5272+180.1390)
= 29.2236

Under the hypothesis that β2 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is P {χ2 (1)>29.2236) = <.001.
Thus the hypothesis that β2 is zero is rejected and we confirm that {X2, X4} is a better model that {X2} only.

In case of {X4} and {X2, X4} we see that we are concerned about the hypothesis that β1=0. So,

log likelihood with variable X2= -165.5272
log likelihood without variable X2= -176.2044

So g=2*(-165.5272+176.2044)
= 21.3544

Under the hypothesis that β1 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is P{χ2(1)>21.3544) = <0.05.
Thus the hypothesis that β1 is rejected and we confirm that {X2, X4} is a better model that {X4} only.

Next we consider an interaction term X_2*X_4.

The corresponding logit equation will be logit(Y=1) = β0+β1*X2+β2*X4+β3*X2*X4

The result of fitting the above model is tabulated below

Variable	β-value	p-value
X_2	0.2938	1.6159e-14
X_4	0.1973	1.5547e-12
X_2*X_4	-0.000771	1.3410e-13
Constant	-73.2749	9.8256e-14

Deviance = 209.3288

Log likelihood = -104.6644

Now under the hypothesis that β3=0 we obtain the corresponding value of g

g=2*(-104.6644+165.5272)
= 121.7256

Under the hypothesis that β3 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is P{χ2(1)>121.7256) = <0.001.

Thus the hypothesis that β3 is rejected and we confirm that {X2, X4,X2*X4} is a better model that {X2,X4}only.

c) {X3, X4}

The results of uni-variable analysis are tabulated below

Uni-variable analysis of variable	Logit equation	β_0	β_1	p-value for β_0	p-value for β_1	Deviance	Log likelihood
X_2	logit(Y=1)= $\beta 0+\beta 1*X_3$	2.6811	-0.0175	1.7341e-08	1.3921e-09	387.4950	-193.7475
X4	logit(Y=1)= $\beta 0+\beta 1*X_4$	3.8266	-0.0102	1.0994e-14	2.5827e-17	352.4087	-176.2044

Since p-value of both β_0 and β_1 is less than 0.25 we keep it in the model for the next step which is bi-variable analysis of $\{X_3, X_4\}$.

The corresponding logit equation is $\text{logit}(Y=1) = \beta_0+\beta_1*X_3+\beta_2*X_4$. The results of fitting are tabulated below

Variable	β-value	p-value
X_3	-0.0115	1.0106e-04
X_4	-0.0075	2.2084e-10
Constant	4.7697	1.4260e-16

Deviance= 332.9560

Log likelihood = -166.4780

After this step, we confirm whether the model $\{X_3, X_4\}$ is better than only $\{X_3\}$ or only $\{X_4\}$.

Now checking whether the model {X3,X4} is better than {X3} and {X4} both.
To compare with {X3} we note

Log likelihood with variable X4=-166.4780
Log likelihood without variable X4 = -193.7475

g=2*(-166.4780+193.7475)
= 54.5390

Under the hypothesis that β2 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is $P\{\chi 2(1)>54.5390) = <.001$.
Thus the hypothesis that β2 is zero is rejected and we confirm that {X3, X4} is a better model that {X3} only.

To compare with {X4} we note

Log likelihood with variable X3= -166.4780
Log likelihood without variable X3= -176.2044

g=2*(-166.4780+176.2044)
= 19.4528

Under the hypothesis that β1 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is P{χ2(1)>19.4528) = <.05.
Thus the hypothesis that β2 is zero is rejected and we confirm that {X3, X4} is a better model that {X3} only.

Now we consider an interaction term {X3*X4}.

The corresponding logit equation will be logit(Y=1) = $\beta_0+\beta_1$*X3+β_2*X4+β_3*X3*X4.

The result of fitting the above model is tabulated below

Variable	β-value	p-value
X_3	0.3046	9.1531e-12
X_4	0.0913	3.3362e-10
X_3*X_4	-0.000716	1.2236e-11
Constant	-38.4622	1.8748e-10

Deviance = 245.9717

Log likelihood = -122.9858

Now under the hypothesis that β3=0 we obtain the corresponding value of g

g=2*(-122.9858+166.4780)
= 86.9844

Under the hypothesis that β3 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is P{χ2(1)>86.9844) = <0.001.
Thus the hypothesis that β3 is rejected and we confirm that {X3, X4, X3*X4} is a better model that {X3, X4} only.

Thus the new list of models is:

a) logit(Y=1) = -147.4110+ 0.7730*X1+ 0.3634*X4+ -0.0019*X1*X4
b) logit(Y=1) = -73.2749+ 0.2938*X2+ 0.1973*X4 -0.000771*X2*X4

c) logit(Y=1) = -38.4622+ 0.3046*X3+0.0913*X4-0.000176*X3*X4

The last step is to assess the fit of the model by using the Hosmer Lemeshow goodness of fit statistic $\hat{C}$.

The values of $\hat{C}$ for the three models are

$\hat{C}$ for the first model= 9.8199
$\hat{C}$ for the second model= 11.5877
$\hat{C}$ for the third model= 23.6997

Since we calculate $\hat{C}$ using the concept of 'deciles of risk' grouping, $\hat{C}$ follows a chi-square distribution statistic with degree of freedom equal to 8(10-2),

P ($\chi 2$ (8)>9.8199) = 0.27789488
P ($\chi 2$ (8)>11.5877) = 0.17056927
P ($\chi 2$ (8)>23.6997) = 0.0025730

Larger the p-value better is the fit of the model. Thus from the above calculated p-values, first model fits the data best.

Thus we reject the other two models and finally select the first model as our final model.

Thus the final logit equation becomes,

logit(Y=1) = -147.4110+ 0.7730*X_1+ 0.3634*X_4-0.0019*X_1*X_4

Since we have rejected the correlated variables X_2 and X_3, there may be certain combinations of {X1,X2,X3} which along with X_4 may provide us a better model.

Various combinations were checked and the value of $\hat{C}$ for those models are shown below

Model Number	Model	$\hat{C}$	p-value
1	{X_1+X_2-X_3,X_4, (X_1+X_2-X_3)*X_4}	7.2168	0.1061
2	{X_1+X_2+X_3,X_4, (X_1+X_2+X_3)*X_4}	15.6379	0.016
3	{X_1*X_2/X_3, X_4, (X_1*X_2/X_3)*X_4}	55.4644	0
4	{X_1-X_2+X_3, X_4,(X_1+X_2+X_3)*X_4}	113.589	0
5	{X_1*X2*X_3,X_4,(X_1*X_2*X_3)*X_4}	16.2231	0.0133

It is observed that the model {X_1+X_2-X_3,X_4, (X_1+X_2-X_3)*X_4} gives a better fit as compared to our assumed best fitting model of {X_1,X_4,X_1*X_4}.

Thus we choose the model containing the terms {X_1+X_2-X_3, X_4, (X_1+X_2-X_3)*X_4}.

The logit equation corresponding to the above model is

logit(Y=1) = -145.1799+0.4602*(X_1+X_2-X_3)+0.4957*X_4-0.0014*(X_1+X_2-X_3)*X_4

Applying inverse logit, we get

(Odds of Y=1) =

$$e\text{^}(-145.1799 + 0.4602*(X_1 + X_2 - X_3) + 0.4597*X_4 - 0.0014*(X_1 + X_2 - X_3)*X_4$$

Thus,

(Probability of Y=1) =

$$\frac{e\text{^}(-145.1799 + 0.4602*(X_1 + X_2 - X_3) + 0.4597*X_4 - 0.0014*(X_1 + X_2 - X_3)*X_4}{1 + e\text{^}(-145.1799 + 0.4602*(X_1 + X_2 - X_3) + 0.4597*X_4 - 0.0014*(X_1 + X_2 - X_3)*X_4}$$

Interpretation:

a) X_1:

The logit equation is

$$\text{logit}(Y=1) = -145.1799+0.4602*(X_1+X_2-X_3)+0.4957*X_4-0.0014*(X_1+X_2-X_3)*X_4$$

Now let us interpret the effect of X_1 on the model. To do that, we increase X_1 by 10 keeping other variables constant.

Therefore,

$$\text{logit}(Y=1|X_1=X) = -145.1799+0.4602*(X+X_2-X_3)+0.4957*X_4-0.0014*(X+X_2-X_3)*X_4$$

$$\text{logit}(Y=1|X_1=X+10) = -145.1799+0.4602*(X+10+X_2-X_3)+0.4957*X_4-0.0014*(X+10+X_2-X_3)*X_4$$

Subtracting first equation from second we get

$$\text{logit}(Y=1|X_1=X+10)-\text{logit}(Y=1|X_1=X) = 4.602-0.014*X_4$$

Thus the odds ratio is $OR=e^{4.602-0.014*X_4}$

Thus the effect of X1 on the odds ratio is dependent on the value of X_4. The above equation means that the odds that Y=1 given $X_1=X+10$ is $e^{4.602-0.014*X_4}$ times the odds that Y=1 given $X_1=X$.

b) X_2:

Since we found that X_1 and X_2 are correlated variables their effect on the odds ratio must be similar. The logit equation confirms the above statement and thus $OR= e^{4.602-0.014*X_4}$

c) X_3:

Again we obtain that the odds ratio for 10-unit increase in X_3 is $OR=e^{-4.602+0.014*X4}$. The OR obtained for X_3 is inverse of those obtained for X_1 and X_2.

d) X_4:

We start from the original equation

$$\text{logit}(Y=1) = -145.1799+0.4602*(X_1+X_2-X_3)+0.4957*X_4-0.0014*(X_1+X_2-X_3)*X_4$$

Let us consider the effect of 10-unit increase in X_4.

$$\text{logit}(Y=1|X_4=X)= -145.1799+0.4602*(X_1+X_2-X_3)+0.4957*X-0.0014*(X_1+X_2-X_3)*X$$

$$\text{logit}(Y=1|X_4=X+10)= -145.1799+0.4602*(X_1+X_2-X_3)+0.4957*(X+10)-0.0014*(X_1+X_2-X_3)*(X+10)$$

Therefore subtracting 1 from 2 we get

$$\text{logit}(Y=1|X_4=X+10)- \text{logit}(Y=1|X_4=X)= 4.957-0.014*(X_1+X_2-X_3)$$

Thus the odds ratio OR becomes $OR= e^{4.957-0.014*(X1+X2-X3)}$

Thus the effect of X_4 is dependent on the values of the other three variables. The above equation means that the odds of Y=1 given $X_4= X+10$ is $e^{4.957-0.014*(X1+X2-X3)}$ times the odds of Y=1 given $X_4=X$.

12.2 Dataset 2

Details of variables in Dataset 2

a) Outcome variable: Y which is dichotomous. Y is vector consisting of 445 observations.
b) Independent variables: X_1, X_2, X_3, X_4 all of which are continuous and each of them containing 445 observations.

The following Matlab code finds the correlation matrix and plots the correlation among the predictor variables.

Matlab Code:

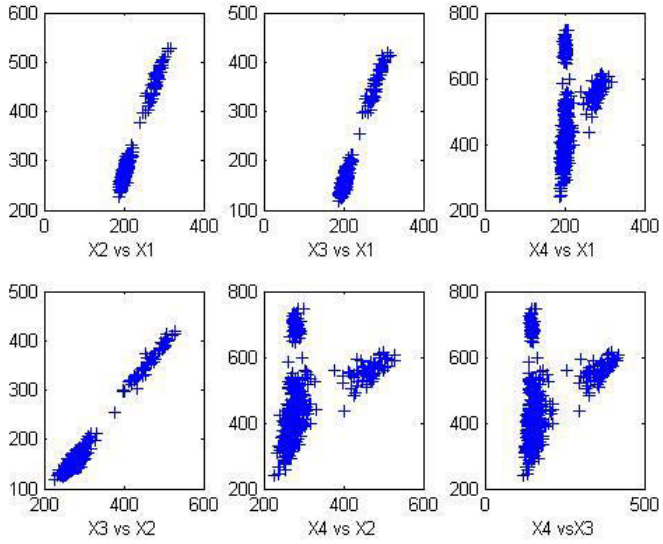
```
load data_new.mat
A=[X1 X2 X3 X4];
M=corr(A)
subplot(2,3,1);
plot(X1,X2,'+');
xlabel('X2 vs X1');
```

```
subplot(2,3,2);
plot(X1,X3,'+');
xlabel('X3 vs X1');
subplot(2,3,3);
plot(X1,X4,'+');
xlabel('X4 vs X1');
subplot(2,3,4);
plot(X2,X3,'+');
xlabel('X3 vs X2');
subplot(2,3,5);
plot(X2,X4,'+');
xlabel('X4 vs X2');
subplot(2,3,6);
plot(X3,X4,'+');
xlabel('X4 vsX3');
```

Output:

Correlation Matrix =

1.0000	0.9911	0.9902	0.3313
0.9911	1.0000	0.9912	0.3704
0.9902	0.9912	1.0000	0.2808
0.3313	0.3704	0.2808	1.0000



Again we can deduce the fact that X_1, X_2, X_3 are strong correlated and we follow the methodology employed in the first dataset to find out the best fitting model.

The three possible models we will consider again are

a) $\{X_1, X_4\}$
b) $\{X_2, X_4\}$
c) $\{X_3, X_4\}$

Thus we shall perform logistic regression on each of these models as in the first dataset.

a)$\{X_1, X_4\}$

The results of uni-variable analysis are tabulated below

Uni-variable analysis of variable	Logit equation	β_0	β_1	p-value for β_0	p-value for β_1	Deviance	Log likelihood
X_1	logit(Y=1)= $\beta 0+\beta 1*X_1$	14.2294	-0.0641	6.9383e-14	4.9020e-12	398.5817	-199.2908
X4	logit(Y=1)= $\beta 0+\beta 1*X_4$	13.7945	-0.0260	1.1851e-24	1.4029e-23	231.1564	-115.5782

Since p-value of both β_0 and β_1 is less than 0.25 we keep it in the model for the next step which is bi-variable analysis of $\{X_1, X_4\}$.

The corresponding logit equation is logit(Y=1) = $\beta_0+\beta_1*X1+\beta_2*X4$. The results of fitting are tabulated below

Variable	β-value	p-value
X_1	-0.0653	7.3496e-08
X_4	-0.0222	4.9944e-23
Constant	26.2387	2.5466e-18

Deviance = 151.4291

Log-likelihood= -75.7145

As done in the first data set, we can employ similar technique to check whether the model containing $\{X_1, X_4\}$ is better than $\{X_1\}$ and $\{X_4\}$. After carrying out the pre-defined steps we conclude that model $\{X_1, X_4\}$ is better than the other two.

Next we consider an interaction term X_1*X_4.

The corresponding logit equation will be logit(Y=1) = $\beta 0+\beta 1*X1+\beta 2*X4+\beta 3*X1*X4$

The result of fitting the above model is tabulated below

Variable	β-value	p-value
X_1	0.9153	5.0946e-05
X_4	0.3712	5.2908e-05
X_1*X_4	-0.0020	2.2933e-05
Constant	-170.2149	1.3728e-04

Deviance = 130.4453

Log likelihood = -65.2227

Now under the hypothesis that β3=0 we obtain the corresponding value of g

g=2*(-65.2227+75.7145)
= 20.9838

Under the hypothesis that β3 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is P{χ2(1)>20.9838) = <0.001.

Thus the hypothesis that β3=0 is rejected and we confirm that {X1, X4, X1*X4} is a better model that {X1, X4} only.

b){X_2, X_4}

The results of uni-variable analysis are tabulated below

Uni-variable analysis of variable	Logit equation	β_0	β_1	p-value for β_0	p-value for β_1	Deviance	Log likelihood
X_2	logit(Y=1)= β0+β1*X_2	9.7958	-0.0307	1.2849e-13	5.4155e-11	384.5768	-192.2884
X4	logit(Y=1)= β0+β1*X_4	13.7945	-0.0260	1.1851e-24	1.4029e-23	231.1564	-115.5782

Since p-value of both β_0 and β_1 is less than 0.25 we keep it in the model for the next step which is bi-variable analysis of {X_2, X_4}.

The corresponding logit equation is logit(Y=1) = β_0+β_1*X2+β_2*X4. The results of fitting are tabulated below

Variable	β-value	p-value
X_2	-0.0258	2.2837e-08
X_4	-0.0215	2.5109e-22
Constant	19.8731	1.6668e-24

Deviance = 155.5273

Log-likelihood= -77.7637

As done in the first data set, we can employ similar technique to check whether the model containing {X_2, X_4} is better than {X_2} and {X_4}. After carrying out the pre-defined steps we conclude that model {X_2, X_4} is better than the other two.

Next we consider an interaction term X_2*X_4.

The corresponding logit equation will be logit(Y=1) = β0+β1*X2+β2*X4+β3*X2*X4

The result of fitting the above model is tabulated below

Variable	β-value	p-value
X_2	0.4420	6.4545e-10
X_4	0.2354	2.5931e-09
X_2*X_4	-0.0009443	2.8692e-09
Constant	-106.6762	9.5543e-09

Deviance = 107.6268

Log likelihood = -53.8134

Now under the hypothesis that β3=0 we obtain the corresponding value of g

g=2*(-53.8134+ 77.7637)
= 47.9005

Under the hypothesis that β3 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is P{χ2(1)>47.9005) = <0.001.

Thus the hypothesis that β3=0 is rejected and we confirm that {X2, X4, X2*X4} is a better model that {X2, X4} only.

c) $\{X_3, X_4\}$

The results of uni-variable analysis are tabulated below

Uni-variable analysis of variable	Logit equation	β_0	β_1	p-value for β_0	p-value for β_1	Deviance	Log likelihood
X_3	logit(Y=1)= $\beta 0+\beta 1*X_3$	4.2899	-0.0198	2.8783e-23	2.4575e-15	422.5374	-211.2687
X4	logit(Y=1)= $\beta 0+\beta 1*X_4$	13.7945	-0.0260	1.1851e-24	1.4029e-23	231.1564	-115.5782

Since p-value of both β_0 and β_1 is less than 0.25 we keep it in the model for the next step which is bi-variable analysis of $\{X_3, X_4\}$.

The corresponding logit equation is $\text{logit}(Y=1) = \beta_0+\beta_1*X3+\beta_2*X4$. The results of fitting are tabulated below

Variable	β-value	p-value
X_3	-0.0237	3.3849e-08
X_4	-0.0228	3.4742e-24
Constant	16.9460	4.4105e-27

Deviance = 153.1368

Log-likelihood= -76.5684

As done in the first data set, we can employ similar technique to check whether the model containing $\{X_3, X_4\}$ is better than $\{X_3\}$ and $\{X_4\}$. After carrying out the pre-defined steps we conclude that model $\{X_3, X_4\}$ is better than the other two.

Next we consider an interaction term X_3*X_4.

The corresponding logit equation will be logit(Y=1) = β0+β1*X3+β2*X4+β3*X3*X4

The result of fitting the above model is tabulated below

Variable	β-value	p-value
X_3	0.4464	2.5771e-04
X_4	0.1147	0.0012
X_3*X_4	-0.00093903	1.6611e-04
Constant	-51.8918	0.0026

Deviance = 132.8193

Log likelihood = -66.4097

Now under the hypothesis that β3=0 we obtain the corresponding value of g

g=2*(-66.4097+76.5684)
= 20.3175

Under the hypothesis that β3 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is P{χ2(1)>20.3175) = <0.001.

Thus the hypothesis that β3=0 is rejected and we confirm that {X3, X4, X3*X4} is a better model that {X3, X4} only.

Thus the new list of models is:

a) logit(Y=1) = -170.2149+ 0.9153*X1+ 0.3712*X4-0.0020*X1*X4
b) logit(Y=1) = -106.6762+ 0.4420*X2+ 0.2354*X4 -0.0009443 *X2*X4
c) logit(Y=1) = -51.8918+0.4464*X3+0.1147*X4-0.00093903*X3*X4

The last step is to assess the fit of the model by using the Hosmer Lemeshow goodness of fit statistic $\hat{C}$.

The values of $\hat{C}$ for the three models are

$\hat{C}$ for the first model= 21.3858
$\hat{C}$ for the second model= 10.0821
$\hat{C}$ for the third model= 40.5402

Since we calculate $\hat{C}$ using the concept of 'deciles of risk' grouping, $\hat{C}$ follows a chi-square distribution statistic with degree of freedom equal to 8(10-2),

P ($\chi 2$ (8)>21.3858) = 0.0023
P ($\chi 2$ (8)>10.0821) = 0.069
P ($\chi 2$ (8)>40.5402) = 0

Larger the p-value better is the fit of the model. Thus from the above calculated p-values, second model fits the data best.

Thus the final logit model is

logit(Y=1) = -106.6762+ 0.4420*X2+ 0.2354*X4 -0.0009443 *X2*X4

Again checking the various combinations which may provide a fit better than the above logit model is shown below

Model Number	Model	$\hat{C}$	p-value
1	$\{X_1+X_2-X_3,X_4, (X_1+X_2-X_3)*X_4\}$	29.3126	0.0001
2	$\{X_1+X_2+X_3,X_4, (X_1+X_2+X_3)*X_4\}$	5.1530	0.1084
3	$\{X_1*X_2/X_3, X_4, (X_1*X_2/X_3)*X_4\}$	155.3948	0
4	$\{X_1-X_2+X_3, X_4,(X_1-X_2+X_3)*X_4\}$	328.8601	0
5	$\{X_1*X2*X_3,X_4,(X_1*X_2*X_3)*X_4\}$	19.7892	0.0041

It is observed that the model $\{X_1+X_2+X_3,X_4, (X_1+X_2+X_3)*X_4\}$ gives a better fit as compared to our assumed best fitting model of $\{X_2,X_4,X_2*X_4\}$.

Thus we choose the model containing the terms $\{X_1+X_2+X_3, X_4, (X_1+X_2+X_3)*X_4\}$.

The logit equation corresponding to the above model is

logit(Y=1) = $-117.3236+0.2116*(X_1+X_2+X_3)+0.2547*X_4-0.00044763*(X_1+X_2+X_3)*X_4$

Applying inverse logit we get

(Odds of Y=1)=

$e\text{^}(-117.3236+0.2116*(X_1+X_2+X_3)+0.2547*X_4-0.00044763*(X_1+X_2+X_3)*X_4$
Thus,

(Probability of Y=1) =

$$\frac{e\text{^}(-117.3236+0.2116*(X_1+X_2+X_3)+0.2547*X_4-0.00044763*(X_1+X_2+X_3)*X_4}{1+e\text{^}(-117.3236+0.2116*(X_1+X_2+X_3)+0.2547*X_4-0.00044763*(X_1+X_2+X_3)*X_4}$$

Interpretation:

a) X_1:

The logit equation is

logit(Y=1) = -117.3236+0.2116*(X_1+X_2+X_3)+0.2547*X_4-0.00044763*(X_1+X_2+X_3)*X_4
Now let us interpret the effect of X_1 on the model. To do that, we increase X_1 by 10 keeping other variables constant.

Therefore,

logit(Y=1|X_1=X) = -117.3236+0.2116*(X+X_2+X_3)+0.2547*X_4-0.00044763*(X+X_2+X_3)*X_4

logit(Y=1|X_1=X+10) =-117.3236+0.2116*(X+10+X_2+X_3)+0.2547*X_4-0.00044763*(X+10+X_2+X_3)*X_4

Subtracting first equation from second we get

logit(Y=1|X_1=X+10)-logit(Y=1|X_1=X)= 2.116-0.0044763*X_4

Thus the odds ratio is OR=$e^{2.116-0.0044763*X_4}$

Thus the effect of X1 on the odds ratio is dependent on the value of X_4. The above equation means that the odds that Y=1 given X_1=X+10 is $e^{2.116-0.0044763*X_4}$ times the odds that Y=1 given X_1=X.

b) X_2:

Since we found that X_1 and X_2 are correlated variables their effect on the odds ratio must be similar. The logit equation confirms the above statement and thus OR= $e^{2.116-0.004763*X_4}$

c) X_3:

Again we obtain that the odds ratio for 10-unit increase in X_3 is OR=$e^{2.116-0.0044763*X4}$

d) X_4:

We start from the original equation

logit(Y=1|X_4=X) = -117.3236+0.2116*(X_1+X_2+X_3)+0.2547*X-0.00044763*(X_1+X_2+X_3)*X

Let us consider the effect of 10-unit increase in X_4.

$\text{logit}(Y=1|X_4=X+10)= -117.3236+0.2116*(X_1+X_2+X_3)+0.2547*(X+10)-0.00044763*(X_1+X_2+X_3)*(X+10)$

Therefore subtracting 1 from 2 we get

$\text{logit}(Y=1|X_4=X+10)- \text{logit}(Y=1|X_4=X)= 2.547-0.0044763*(X_1+X_2+X_3)$

Thus the odds ratio OR becomes $OR= e^{2.547-0.0044763*(X1+X2+X3)}$

Thus the effect of X_4 is dependent on the values of the other three variables. The above equation means that the odds of Y=1 given $X_4= X+10$ is $e^{2.547-0.0044763*(X1+X2+X3)}$ times the odds of Y=1 given $X_4=X$.

12.3 Combined dataset:

This dataset is the combination of the Dataset1 and Dataset2. The details of this dataset are as follows:

a) Outcome variable: Y which is dichotomous. Y is vector consisting of 781 observations.

b) Independent variables: X1, X2, X3, X4 all of which are continuous and each of them containing 781 observations.

The following Matlab code finds the correlation matrix and plots the correlation among the predictor variables.

Matlab Code:

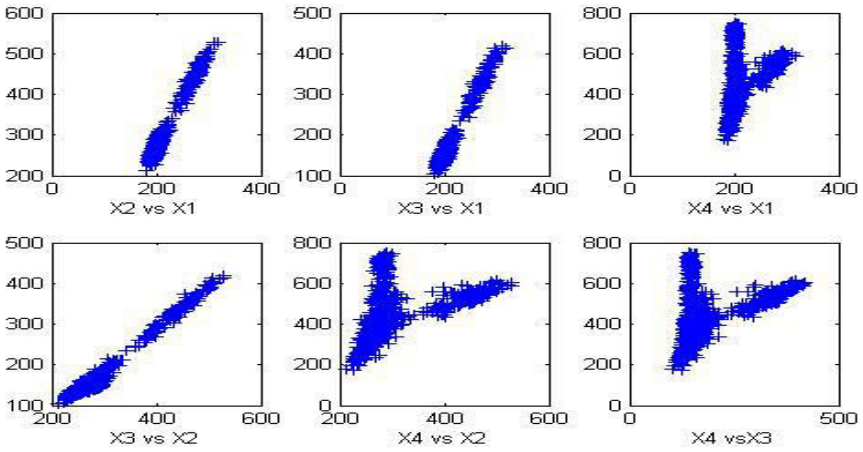
```
load data_combine.mat
A=[A1 A2 A3 A4];
M=corr(A)
subplot(2,3,1);
plot(A1,A2,'+');
xlabel('X2 vs X1');
subplot(2,3,2);
plot(A1,A3,'+');
xlabel('X3 vs X1');
subplot(2,3,3);
plot(A1,A4,'+');
xlabel('X4 vs X1');
subplot(2,3,4);
plot(A2,A3,'+');
xlabel('X3 vs X2');
subplot(2,3,5);
```

```
plot(A2,A4,'+');
xlabel('X4 vs X2');
subplot(2,3,6);
plot(A3,A4,'+');
xlabel('X4 vsX3');
```

Output:

Correlation matrix M=

1.0000	0.9886	0.9898	0.3539
0.9886	1.0000	0.9887	0.4193
0.9898	0.9887	1.0000	0.3108
0.3539	0.4193	0.3108	1.0000



As predicted the combined dataset's X_1, X_2, X_3 will also be correlated and thus we follow the steps as in the previous example. Since X_4 is weakly correlated with every other variable we include it in each of the models described below.

The three models to be taken into consideration are

a) $\{X_1, X_4\}$
b) $\{X_2, X_4\}$
c) $\{X_3, X_4\}$

Thus we shall perform logistic regression on each of these models as in the first dataset.

a)$\{X_1,X_4\}$

The results of uni-variable analysis are tabulated below

Uni-variable analysis of variable	Logit equation	β_0	β_1	p-value for β_0	p-value for β_1	Deviance	Log likelihood

X_1	logit(Y=1)= β0+β1*X_1	11.9499	-0.0558	5.5479e-24	5.1519e-22	832.8016	-416.4008
X4	logit(Y=1)= β0+β1*X_4	5.8085	-0.0121	3.5142e-43	1.5032e-41	772.9817	-386.4908

Since p-value of both β_0 and β_1 is less than 0.25 we keep it in the model for the next step which is bi-variable analysis of $\{X_1, X_4\}$.

The corresponding logit equation is logit(Y=1) = $\beta_0+\beta_1*X1+\beta_2*X4$. The results of fitting are tabulated below

Variable	β-value	p-value
X_1	-0.0439	6.9352e-15
X_4	-0.0093	4.5848e-29
Constant	13.8655	4.3677e-31

Deviance = 662.6630

Log-likelihood= -331.3315

As done in the first data set, we can employ similar technique to check whether the model containing $\{X_1, X_4\}$ is better than $\{X_1\}$ and $\{X_4\}$. After carrying out the pre-defined steps we conclude that model $\{X_1, X_4\}$ is better than the other two.

Next we consider an interaction term X_1*X_4.

The corresponding logit equation will be logit(Y=1) = β0+β1*X1+β2*X4+β3*X1*X4

The result of fitting the above model is tabulated below

Variable	β-value	p-value
X_1	0.7790	1.0262e-24
X_4	0.3530	1.9296e-24
X_1*X_4	-0.0018	2.5796e-25
Constant	-148.1900	1.2642e-23

Deviance = 499.7194

Log likelihood = -249.8597

Now under the hypothesis that β3=0 we obtain the corresponding value of g

g=2*(-249.8957+331.3315)
= 162.8716

Under the hypothesis that β3 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is P{χ2(1)>162.8716) = <0.001.

Thus the hypothesis that β3=0 is rejected and we confirm that {X1, X4, X1*X4} is a better model that {X1, X4} only.

b) {X_2, X_4}

The results of uni-variable analysis are tabulated below

Uni-variable analysis of variable	Logit equation	β_0	β_1	p-value for β_0	p-value for β_1	Deviance	Log likelihood
X_2	logit(Y=1)= β0+β1*X_2	8.1288	-0.0269	1.8341e-25	2.1543e-22	802.3543	-401.1771
X4	logit(Y=1)= β0+β1*X_4	5.8085	-0.0121	3.5142e-43	1.5032e-41	772.9817	-386.4908

Since p-value of both β_0 and β_1 is less than 0.25 we keep it in the model for the next step which is bi-variable analysis of {X_2, X_4}.

The corresponding logit equation is logit(Y=1) = β_0+β_1*X2+β_2*X4. The results of fitting are tabulated below

Variable	β-value	p-value
X_2	-0.0186	4.2871e-15
X_4	-0.0087	1.4965e-24
Constant	9.8248	8.9600e-42

Deviance = 666.2849

Log-likelihood= -333.1425

As done in the first data set, we can employ similar technique to check whether the model containing {X_2, X_4} is better than {X_2} and {X_4}. After carrying out the pre-defined steps we conclude that model {X_2, X_4} is better than the other two.

Next we consider an interaction term X_2*X_4.

The corresponding logit equation will be logit(Y=1) = β0+β1*X2+β2*X4+β3*X2*X4

The result of fitting the above model is tabulated below

Variable	β-value	p-value
X_2	0.3156	1.0822e-30
X_4	0.1967	1.6814e-29
X_2*X_4	-0.00076881	2.1057e-31
Constant	-78.1603	2.1810e-28

Deviance = 425.1497

Log likelihood = -212.5749

Now under the hypothesis that β3=0 we obtain the corresponding value of g

g=2*(-212.5749+333.1425)
= 241.1352

Under the hypothesis that β3 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is P{χ2(1)>241.1352) = <0.001.

Thus the hypothesis that β3=0 is rejected and we confirm that {X2, X4, X2*X4} is a better model that {X2, X4} only.

c) {X_3, X_4}:

The results of uni-variable analysis are tabulated below

Uni-variable analysis of variable	Logit equation	β_0	β_1	p-value for β_0	p-value for β_1	Deviance	Log likelihood
X_3	logit(Y=1)= β0+β1*X_3	3.5403	-0.0187	1.4053e-29	8.5133e-24	864.9442	-432.4721
X4	logit(Y=1)= β0+β1*X_4	5.8085	-0.0121	3.5142e-43	1.5032e-41	772.9817	-386.4908

Since p-value of both β_0 and β_1 is less than 0.25 we keep it in the model for the next step which is bi-variable analysis of {X_3, X_4}.

The corresponding logit equation is logit(Y=1) = β_0+β_1*X3+β_2*X4. The results of fitting are tabulated below

Variable	β-value	p-value
X_3	-0.0163	6.1648e-15
X_4	-0.0098	1.9243e-32
Constant	7.7229	2.6221e-49

Deviance = 668.7901

Log-likelihood= -334.3951

As done in the first data set, we can employ similar technique to check whether the model containing $\{X_3, X_4\}$ is better than $\{X_3\}$ and $\{X_4\}$. After carrying out the pre-defined steps we conclude that model $\{X_3, X_4\}$ is better than the other two.

Next we consider an interaction term X_3*X_4.

The corresponding logit equation will be logit(Y=1) = β0+β1*X3+β2*X4+β3*X3*X4

The result of fitting the above model is tabulated below

Variable	β-value	p-value
X_3	0.3288	2.0634e-20
X_4	0.0972	4.6381e-18
X_3*X_4	-0.00075447	8.8626e-21
Constant	-40.9851	4.0895e-17

Deviance = 527.5037

Log likelihood = -263.7518

Now under the hypothesis that β3=0 we obtain the corresponding value of g

g=2*(-263.7518+334.3951)
= 141.2866

Under the hypothesis that β3 is equal to zero, g will follow a chi square distribution with degree of freedom equal to 1. Thus from the chi square table, the p value associated with the test is P{χ2(1)>141.2866) = <0.001.

Thus the hypothesis that β3=0 is rejected and we confirm that {X3, X4, X3*X4} is a better model that {X3, X4} only.

Thus the final three models obtained are

a) logit(Y=1) = -148.1900+0.7790*X1+0.3530*X4-0.0018*X1*X4
b) logit(Y=1) = -78.1603+0.3156*X2+0.1967*X4 -0.00076881*X2*X4
c) logit(Y=1) = -40.9851+0.3288*X3+0.0972*X4-0.00075447*X3*X4

The last step is to assess the fit of the model by using the Hosmer Lemeshow goodness of fit statistic $\hat{C}$.

The values of Ĉ for the three models are

Ĉ for the first model= 30.3395
Ĉ for the second model= 14.0149
Ĉ for the third model= 21.7568

Since we calculate Ĉ using the concept of 'deciles of risk' grouping, Ĉ follows a chi-square distribution statistic with degree of freedom equal to 8(10-2),

P (χ2 (8)>30.3395) = 0.0001
P (χ2 (8)>14.0149) = 0.026
P (χ2 (8)>21.7568) = 0.002

Larger the p-value better is the fit of the model. Thus from the above calculated p-values, second model fits the data best.

Thus the final logit model is

logit(Y=1) = -106.6762+ 0.4420*X2+ 0.2354*X4 -0.0009443 *X2*X4

Again checking the various combinations which may provide a fit better than the above logit model is shown below

Model Number	Model	Ĉ	p-value
1	$\{X_1+X_2-X_3,X_4, (X_1+X_2-X_3)*X_4\}$	6.5273	0.1108
2	$\{X_1+X_2+X_3,X_4, (X_1+X_2+X_3)*X_4\}$	14.9473	0.0198
3	$\{X_1*X_2/X_3, X_4, (X_1*X_2/X_3)*X_4\}$	187.1252	0
4	$\{X_1-X_2+X_3, X_4,(X_1-X_2+X_3)*X_4\}$	445.8106	0
5	$\{X_1*X2*X_3,X_4,(X_1*X_2*X_3)*X_4\}$	41.5027	0

It is observed that the model $\{X_1+X_2-X_3,X_4, (X_1+X_2-X_3)*X_4\}$ gives a better fit as compared to our assumed best fitting model of $\{X_2,X_4,X_2*X_4\}$.

Thus we choose the model containing the terms $\{X_1+X_2-X_3, X_4, (X_1+X_2-X_3)*X_4\}$.

The logit equation corresponding to the above model is

logit(Y=1) = $-128.3361+0.4113*(X_1+X_2-X_3)+0.3644*X_4-0.0011*(X_1+X_2-X_3)*X_4$

Applying inverse logit we get

(Odds of Y=1)=

$$e\text{^}(-128.3361+0.4113*(X_1+X_2-X_3)+0.3644*X_4-0.0011*(X_1+X_2-X_3)*X_4$$

Thus,

(Probability of Y=1) =

$$\frac{e\text{^}(-128.3361+0.4113*(X_1+X_2-X_3)+0.3644*X_4-0.0011*(X_1+X_2-X_3)*X_4}{1+e\text{^}(-128.3361+0.4113*(X_1+X_2-X_3)+0.3644*X_4-0.0011*(X_1+X_2-X_3)*X_4}$$

Interpretation:

a) X_1:

We start with the obtained logit model.

logit(Y=1) = -128.3361+0.4113*(X_1+X_2-X_3)+0.3644*X_4-0.0011*(X_1+X_2-X_3)*X_4

Considering the effect of 10-unit increase in X_1 we get the following logit equations

logit(Y=1|X_1=X) = -128.3361+0.4113*(X+X_2-X_3)+0.3644*X_4-0.0011*(X+X_2-X_3)*X_4

logit(Y=1|X_1=X+10)= -128.3361+0.4113*(X+10+X_2-X_3)+0.3644*X_4-0.0011*(X+10+X_2-X_3)*X_4

Subtracting first equation from second we get,

logit(Y=1|X_1=X+10)-logit(Y=1|X_1=X) = 4.113-0.011*X_4

Thus the odds ratio is OR=$e^{4.113-0.011*X_4}$

The above equation means that the odds of Y=1 given X_1=X+10 is $e^{4.113-0.011*X_4}$ times the odds of Y=1 given X_1=X.

b) X_2:

We again find that the OR for X_2 remains same as OR for X_1. Therefore OR for X_2= $e^{4.113-0.011*X_4}$.

c) X_3:

Similarly we find out that the odds ratio for X_3 is OR= $e^{-4.113+0.011*X_4}$.

d) X_4:

We start with the obtained logit model.

logit(Y=1) = -128.3361+0.4113*(X_1+X_2-X_3)+0.3644*X_4-0.0011*(X_1+X_2-X_3)*X_4

Considering the effect of 10-unit increase in X_4 we get the following logit equations

logit(Y=1|X_4=X) = -128.3361+0.4113*(X_1+X_2-X_3)+0.3644*X-0.0011*(X_1+X_2-X_3)*X

logit(Y=1|X_4=X+10)= -128.3361+0.4113*(X_1+X2-X_3)+0.3644*(X+10)-0.0011*(X_1+X_2-X_3)*(X+10)

Subtracting first equation from second we get,

logit(Y=1|X_1=X+10)-logit(Y=1|X_1=X) = 3.644-0.011*(X_1+X_2-X_3)

Thus the odds ratio is OR=$e^{3.644-0.011*(X1+X2-X3)}$

The above equation means that the odds of Y=1 given X_4=X+10 is $e^{3.644-0.011*(X1+X2-X3)}$ times the odds of Y=1 given X_4=X.

13.0 Testing and Validation of the Models

As per previous section, we obtained the best fitting models for the two datasets and the combined dataset.

13.1 Testing and validation-First model

Model for first data set is

(Probability of Y=1) =

$$\frac{e\text{^}(-145.1799+0.4602*(X_1+X_2-X_3)+0.4597*X_4-0.0014*(X_1+X_2-X_3)*X_4}{1+e\text{^}(-145.1799+0.4602*(X_1+X_2-X_3)+0.4597*X_4-0.0014*(X_1+X_2-X_3)*X_4}$$

Matlab Code:

```
load datas.mat%load the data
%the first 10 observations and last 10 observations shall be used
%for model testing and the other observations shall be used for
%model building
V1=A(11:326);
V2=B(11:326);
V3=C(11:326);
V4=D(11:326);
%our model
X3=[V1+V2-V3 V4 (V1+V2-V3).*V4];

V5=[A(1:10);A(327:336)]
V6=[B(1:10);B(327:336)];
V7=[C(1:10);C(327:336)];
V8=[D(1:10);D(327:336)];
Y_new=[Y(1:10);Y(327:336)];
%building the model
```

```
[Bs,dev,stats]=glmfit(X3,Y(11:326),'binomial','link','logit');
X4=[V5+V6-V7 V8 (V5+V6-V7).*V8];
C=[Y_new  glmval(Bs,X4,'logit')]% c vector will show original value of y%and the
probability of Y=1 as returned by our model
```

Output:

Index	Expected Y	Model prediction of Probability of Y=1
1	1.0000	0.8128
2	1.0000	0.9417
3	1.0000	0.8027
4	1.0000	0.9307
5	1.0000	0.7340
6	1.0000	0.3066
7	1.0000	0.8857
8	1.0000	0.7493
9	1.0000	0.8851
10	1.0000	0.7285
11	0	0.0006
12	0	0.0040
13	0	0.0000
14	0	0.0000
15	0	0.0000
16	0	0.0000
17	0	0.0000
18	0	0.0001
19	0	0.0000
20	0	0.0000

Assume the threshold value for probability is 0.6. It means if the predicted probability is greater than equal to 0.6 then Y=1 otherwise Y=0. Going by this assumption our model correctly predicts the value of Y=1 for all the values except for the 6th index where the probability of Y=1 is 0.3066.

Thus our model is 95% accurate for the given set of test observations.

13.2 Testing and Validation: Second Model

Model for second data set is

(Probability of Y=1) =

$$\frac{e\text{^}(-117.3236+0.2116*(X_1+X_2+X_3)+0.2547*X_4-0.00044763*(X_1+X_2+X_3)*X_4}{1+e\text{^}(-117.3236+0.2116*(X_1+X_2+X_3)+0.2547*X_4-0.00044763*(X_1+X_2+X_3)*X_4}$$

Matlab Code:

```
load data_new.mat%load the data
%the first 10 observations and last 10 observations shall be used
%for model testing and the other observations shall be used for
%model building
V1=X1(11:435);
V2=X2(11:435);
V3=X3(11:435);
V4=X4(11:435);
%our model
L3=[V1+V2+V3 V4 (V1+V2+V3).*V4];
V5=[X1(1:10);X1(436:445)]
V6=[X2(1:10);X2(436:445)];
V7=[X3(1:10);X3(436:445)];
V8=[X4(1:10);X4(436:445)];
Y_new=[Y(1:10);Y(436:445)];
%building the model
[Bs,dev,stats]=glmfit(L3,Y(11:435),'binomial','link','logit');
L4=[V5+V6+V7 V8 (V5+V6+V7).*V8];
C=[Y_new glmval(Bs,L4,'logit')]% C vector will show original value of y
%and the probability of Y=1 as returned by our model
```

Output:

Index	Expected Y	Model prediction of Probability of Y=1
1	1.0000	0.9809
2	1.0000	0.9169
3	1.0000	0.9400
4	1.0000	0.9911
5	1.0000	0.9834
6	1.0000	0.9606
7	1.0000	0.9768
8	1.0000	0.9876
9	1.0000	0.9924
10	1.0000	0.9939
11	0	0.0000
12	0	0.0000
13	0	0.0000
14	0	0.0001
15	0	0.0000
16	0	0.4805
17	0	0.0000
18	0	0.0000

19	0	0.0000
20	0	0.0009

Assume the threshold value for probability is 0.6. It means if the predicted probability is greater than equal to 0.6 then Y=1 otherwise Y=0. Going by this assumption our model correctly predicts the value of Y=1 for all the values.

Thus our model is 100% accurate for the given set of test observations.

13.3 Testing and validation-Third model

Model for combined (third) data set is

(Probability of Y=1) =

$$\frac{e\text{^}(-128.3361+0.4113*(X_1+X_2-X_3)+0.3644*X_4-0.0011*(X_1+X_2-X_3)*X_4}{1+e\text{^}(-128.3361+0.4113*(X_1+X_2-X_3)+0.3644*X_4-0.0011*(X_1+X_2-X_3)*X_4}$$

Matlab Code:

```
load data_combine.mat%load the data
%the first 10 observations and last 10 observations shall be used
%for model testing and the other observations shall be used for
%model building
V1=A1(11:771);
V2=A2(11:771);
V3=A3(11:771);
V4=A4(11:771);
%our model
X3=[V1+V2-V3 V4 (V1+V2-V3).*V4];
V5=[A1(1:10);A1(772:781)];
V6=[A2(1:10);A2(772:781)];
V7=[A3(1:10);A3(772:781)];
V8=[A4(1:10);A4(772:781)];
Y_new=[Y2(1:10);Y2(772:781)];
%building the model
[Bs,dev,stats]=glmfit(X3,Y2(11:771),'binomial','link','logit');
X4=[V5+V6-V7 V8 (V5+V6-V7).*V8];
C=[Y_new glmval(Bs,X4,'logit')]% C vector will show original value of y
%and the probability of Y=1 as returned by our model
```

Output:

Index	Expected Y	Model prediction of Probability of Y=1
1	1.0000	0.8503
2	1.0000	0.9536
3	1.0000	0.8653

4	1.0000	0.9381
5	1.0000	0.9114
6	1.0000	0.8420
7	1.0000	0.9423
8	1.0000	0.9118
9	1.0000	0.9480
10	1.0000	0.6556
11	0	0.0000
12	0	0.0000
13	0	0.0000
14	0	0.0001
15	0	0.0000
16	0	0.0364
17	0	0.0000
18	0	0.0000
19	0	0.0001
20	0	0.0036

Assume the threshold value for probability is 0.6. It means if the predicted probability is greater than equal to 0.6 then Y=1 otherwise Y=0. Going by this assumption our model correctly predicts the value of Y=1 for all the values.

Thus our model is 100% accurate for the given set of test observations.

14.0 Comparison with regression tree model

Matlab provides a built-in class ‘classregtree’ whose constructor named **classregtree(X,Y)** creates a binary decision tree corresponding to the values of predictor variables X and response or outcome variable Y.

14.1 First dataset:

The following code shows the graphical description of the binary regression tree constructed out of the first dataset.

Matlab Code:

```
load datas.mat%load the dataset
X=[A B C D];%A,B,C,D are the predictor variables
tree=classregtree(X,Y);%construct a regression tree
view(tree);%view the regression tree
```

Output:

Since the tree is very large, the tree cannot be shown here.

We now perform a comparison between the model obtained using logistic regression and the binary regression tree.

Matlab Code:

```
load datas.mat%load the data set of x and y
V1=A(11:326);%X1(11:326) will help in model building and X1(1:10) and X1(327:336) will
be used for model testing
V2=B(11:326);%same for X2
V3=C(11:326);%same for X3
V4=D(11:326);%same for X4
X3=[V1+V2-V3 V4 (V1+V2-V3).*V4];%logistic model
V5=[A(1:10);A(327:336)];%data of X1 which will help in testing
V6=[B(1:10);B(327:336)];%data of X2 which will help in testing
V7=[C(1:10);C(327:336)];%data of X3 which will help in testing
V8=[D(1:10);D(327:336)];%data of X4 which will help in testing
X4=[V5+V6-V7 V8 (V5+V6-V7).*V8];
[Bs,dev,stats]=glmfit(X3,Y(11:326),'binomial','link','logit');%applying logistic regression
C_logit=glmval(Bs,X4,'logit');%getting the values of probability of Y=1 for the test data
tree=classregtree([V1 V2 V3 V4],Y(11:326));%construct a regression tree
C_tree=eval(tree,[V5 V6 V7 V8]);%evaluate probability of Y=1 for the test data
Y_new=[Y(1:10);Y(327:336)];%expected Y-values
[Y_new C_logit C_tree]%displaying expected Y-values,probability of Y=1
%for logistic regression and probability of Y=1 as predicted by regression
%tree
```

Output:

Expected Y	Probability of Y=1(logistic model)	Predicted Y(logistic model) (threshold=0.6)	Probability of Y=1(regression tree model)	Predicted Y(regression tree) (threshold=0.6)
1	0.8128	1	1.00	1
1	0.9417	1	1.00	1
1	0.8027	1	1.00	1
1	0.9307	1	1.00	1
1	0.7340	1	0.40	0
1	0.3066	0	0.50	0
1	0.8857	1	1.00	1
1	0.7493	1	1.00	1
1	0.8851	1	1.00	1
1	0.7285	1	1.00	1
0	0.0006	0	0.00	0
0	0.0040	0	0.00	0
0	0.0000	0	0.00	0
0	0.0000	0	0.00	0
0	0.0000	0	0.00	0

0	0.0000	0	0.00	0
0	0.0000	0	0.00	0
0	0.0001	0	0.00	0
0	0.0000	0	0.00	0
0	0.0000	0	0.00	0

We can observe that our model correctly predicts 19 out of the 20 test values while the regression tree predicts 18 out of the 20 correctly. Thus our statistical logistic model performs better than the algorithmic regression tree model.

14.2 Second dataset:

The comparison between the obtained logistic model for the second dataset and the regression tree for the corresponding dataset is done as follows:

Matlab Code:

```
load data_new.mat%load the data set of x and y
V1=X1(11:435);%X1(11:436) will help in model building and X1(1:10) and X1(327:336)
will be used for model testing
V2=X2(11:435);%same for X2
V3=X3(11:435);%same for X3
V4=X4(11:435);%same for X4
L3=[V1+V2+V3 V4 (V1+V2+V3).*V4];%logistic model
V5=[X1(1:10);X1(436:445)];%data of X1 which will help in testing
V6=[X2(1:10);X2(436:445)];%data of X2 which will help in testing
V7=[X3(1:10);X3(436:445)];%data of X3 which will help in testing
V8=[X4(1:10);X4(436:445)];%data of X4 which will help in testing
L4=[V5+V6+V7 V8 (V5+V6+V7).*V8];
%Y_new=[Y(1:10);Y(327:336)];
[Bs,dev,stats]=glmfit(L3,Y(11:435),'binomial','link','logit');%applying logistic regression
C_logit=glmval(Bs,L4,'logit');%getting the values of probability of Y=1 for the test data
tree=classregtree([V1 V2 V3 V4],Y(11:435));%construct a regression tree
C_tree=eval(tree,[V5 V6 V7 V8]);%evaluate probability of Y=1 for the test data

Y_new=[Y(1:10);Y(436:445)];%expected Y-values
[Y_new C_logit C_tree]%displaying expected Y-values,probability of Y=1
%for logistic regression and probability of Y=1 as predicted by regression
%tree
```

Output:

Expected Y	Probability of Y=1(logistic model)	Predicted Y(logistic model) (threshold=0.6)	Probability of Y=1(regression tree model)	Predicted Y(regression tree) (threshold=0.6)
1	0.9809	1	1.00	1
1	0.9169	1	1.00	1

1	0.9400	1	1.00	1
1	0.9911	1	1.00	1
1	0.9834	1	1.00	1
1	0.9606	1	1.00	1
1	0.9768	1	1.00	1
1	0.9876	1	1.00	1
1	0.9924	1	1.00	1
1	0.9939	1	1.00	1
0	0.0000	0	0.00	0
0	0.0000	0	0.00	0
0	0.0000	0	0.00	0
0	0.0001	0	0.00	0
0	0.0000	0	0.00	0
0	0.4805	0	0.00	0
0	0.0000	0	0.00	0
0	0.0001	0	0.00	0
0	0.0000	0	0.00	0
0	0.0009	0	0.00	0

We can observe that our model correctly predicts 20 out of the 20 test values while the regression tree also predicts 20 out of the 20 correctly. Thus our statistical logistic model performance is as good as algorithmic regression tree model.

14.2 Third dataset:

The comparison between the obtained logistic model for the third dataset and the regression tree for the corresponding dataset is done as follows:

Matlab Code:

```
load data_combine.mat%load the data set of x and y
V1=A1(11:771);%X1(11:326) will help in model building and X1(1:10) and
X1(327:336) will be used for model testing
V2=A2(11:771);%same for X2
V3=A3(11:771);%same for X3
V4=A4(11:771);%same for X4
L3=[V1+V2-V3 V4 (V1+V2-V3).*V4];%logistic model
V5=[A1(1:10);A1(772:781)];%data of X1 which will help in testing
V6=[A2(1:10);A2(772:781)];%data of X2 which will help in testing
V7=[A3(1:10);A3(772:781)];%data of X3 which will help in testing
V8=[A4(1:10);A4(772:781)];%data of X4 which will help in testing
L4=[V5+V6-V7 V8 (V5+V6-V7).*V8];
%Y_new=[Y(1:10);Y(327:336)];
[Bs,dev,stats]=glmfit(L3,Y2(11:771),'binomial','link','logit');%applying
logistic regression
C_logit=glmval(Bs,L4,'logit');%getting the values of probability of Y=1 for
the test data
tree=classregtree([V1 V2 V3 V4],Y2(11:771));%construct a regression tree
C_tree=eval(tree,[V5 V6 V7 V8]);%evaluate probability of Y=1 for the test
data
```

```
Y_new=[Y2(1:10);Y2(772:781)];%expected Y-values
[Y_new C_logit C_tree]%displaying expected Y-values,probability of Y=1
%for logistic regression and probability of Y=1 as predicted by regression
%tree
```

Output:

Expected Y	Probability of Y=1(logistic model)	Predicted Y(logistic model) (threshold=0.6)	Probability of Y=1(regression tree model)	Predicted Y(regression tree) (threshold=0.6)
1	0.8503	1	0.75	1
1	0.9536	1	1.00	1
1	0.8653	1	0	0
1	0.9381	1	1.00	1
1	0.9114	1	1.00	1
1	0.8420	1	1.00	1
1	0.9423	1	1.00	1
1	0.9118	1	1.00	1
1	0.9480	1	0.75	1
1	0.6556	1	0.6667	1
0	0.0000	0	0.00	0
0	0.0000	0	0.00	0
0	0.0000	0	0.00	0
0	0.0000	0	0.00	0
0	0.0000	0	0.00	0
0	0.0364	0	0.00	0
0	0.0000	0	0.00	0
0	0.0000	0	0.00	0
0	0.0001	0	0.00	0
0	0.0036	0	0.00	0

We can observe that our model correctly predicts 20 out of the 20 test values while the regression tree also predicts 19 out of the 20 correctly. Thus our statistical logistic model performs better than the algorithmic regression tree model.

15.0 Conclusion

Even though algorithmic regression tree model is fairly accurate in its predictions, we can achieve better results with the help of logistic regression model. The model building strategy needs to be strictly followed after all the correlated (redundant) variables are excluded. If no variables are correlated then include all the variables in the model building strategy. Some other possible models should be also verified. The goodness of fit gives us an absolute idea about the fitting of the model to the data.

Even though regression trees are known for its efficiency when a large number of predictor variables are present, it is also subject to instability. A small change in the training set may lead to a different choice when building a node, which in turn may represent a dramatic change in the tree, particularly if the change occurs in top level nodes. Thus it is advisable that while dealing with continuous variables we apply logistic regression. Since our data set included four continuous variables use of Logistic regression model is completely justified.

So a good model building strategy will always provide us a model with good predictive power.

References:

a) Hosmer, D., Lemeshow S. and Sturdivant R.,(2013), *Applied Logistic Regression*,(3 ed.), New Jersey, John-Wiley &Sons, Inc.

b) Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1), 3-14.

c) Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1), 35-39.

d) Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model, *Statistics in medicine*, 16(9), 965-980.

e) Pyke, S. W., & Sheridan, P. M. (1993). Logistic regression analysis of graduate student retention. *Canadian Journal of Higher Education*, 23(2), 44-64.

f) Saha, G. (2011). Applying logistic regression model to the examination results data. *Journal of Reliability and Statistical Studies*, 4(2), 1-13.

g) Bender, R., & Grouven, U. (1997). Ordinal logistic regression in medical research. *Journal of the Royal College of Physicians of London*, 31(5), 546-551.

h) Kumar R., Nandy S., Agarwal R. and Kushwaha S.P.S, (2014), Forest cover dynamics analysis and prediction modeling using logistic regression model, *Ecological Indicators*,v. 45, p. 444-445

i) Pradhan B.(2010), Remote sensing and GIS-based landslide hazard analysis and cross-validation using multivariate logistic regression model on three test areas in Malaysia, *Advances in Space Research*, v. 45 p. 1244-1256

j) Brenning A. and Trombotto D.,(2006), Logistic regression modeling of rock glacier and glacier distribution: Topographic and climatic controls in the semi-arid Andes, *Geomorphology*, v. 81, p. 141-154

k) Yang X. , Skidmore A.K. , Melick D.R. , Zhou Z. and Xu J., (2006), Mapping non-wood forest product (matsutake mushrooms) using logistic regression and a GIS expert system, *ecological modelling*,v. 198 p. 208-218

l) Singh A. and Kushwaha S.P.S,(2011), Refining logistic regression models for wildlife habitat suitability modeling—A case study with muntjak and goral in the Central Himalayas, India, *Ecological Modelling*, v. 222 p. 1354-1366

m) Umar Z., Pradhan B., Ahmad A., Jebur M.N., Tehrany M.S.,(2014), Earthquake induced landslide susceptibility mapping using an integrated ensemble frequency ratio and logistic regression models in West Sumatera Province, Indonesia, *Catena*, v. 118 p. 124-135

n) LaVange, L. M., Iannacchione, V. G., & Garfinkel, S. A. (1986). An application of logistic regression methods to survey data: Predicting high cost users of medical care. In *Proc. Survey Research Methods Section*

o) Lemeshow S. and Archer K.J.,(2006), Goodness-of-fit test for a logistic regression model fitted using survey sample data, *The Stata Journal*, v. 6, Number 1, p. 97-105

p) Bera D and Nayak M.M,(2012), Mortality Risk Assessment for ICU patients using Logistic Regression,*Computing in Cardiology*, v. 39 p.493-496.

q) Introduction to logistic regression: *http://logisticregressionanalysis.com/*

Appendix:

Proof that the Mean Minimizes the Sum of Squared Errors

Let an arbitrary estimate or model value be $Y\hat{}$ and let the mean or arithmetic average be $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$. For an arbitrary estimate, the sum of squared errors is SSE= $\sum_{i=1}^{n}(Y_i - \hat{Y})^2$.

Obviously, Y -Y = 0 , so we can add this representation of zerowithin the parentheses without changing the sum of squared errors. That is, SSE= $\sum_{i=1}^{n}(Y_i + \overline{Y} - \overline{Y} - \hat{Y})^2$.

Rearranging the terms slightly and regrouping, we get SSE= $\sum_{i=1}^{n}((Y_i - \overline{Y}) + (\overline{Y} - \hat{Y}))^2$

Squaring the term inside the summation gives SSE=

$$\sum_{i=1}^{n}((Y_i - \overline{Y})^2 + (\overline{Y} - \hat{Y})^2 + 2((Y_i - \overline{Y})(\overline{Y} - \hat{Y}))$$

Breaking the sums apart yields SSE= $\sum_{i=1}^{n}(Y_i - \overline{Y})^2 + \sum_{i=1}^{n}(\overline{Y} - \hat{Y})^2 + \sum_{i=1}^{n}2(\ (Y_i - \overline{Y})(\overline{Y} - \hat{Y})\ )$

The second term contains no subscripts so the sum can be replaced with $n(\overline{Y} - \hat{Y})^2$ and the factor with no subscripts in the third term can be moved outside the summation giving

SSE= $\sum_{i=1}^{n}(Y_i - \overline{Y})^2 + n(\overline{Y} - \hat{Y})^2 + 2\ (\overline{Y} - \hat{Y})\left[\sum_{i=1}^{n}(Y_i - \overline{Y})\right]$

The bracketed term is the sum of the deviations about the mean and must therefore equal zero. So, the above equation reduces to

SSE= $\sum_{i=1}^{n}(Y_i - \overline{Y})^2 + n(\overline{Y} - \hat{Y})^2$

Clearly, if our model estimate equals the mean, then Y - Yˆ = 0 and
the last term in the above equation is zero. Thus, when the model estimate is the mean, SSE= $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ and if the model estimate equals anything other than the mean, the sum of squared errors would be increased by n(Y - Yˆ)2 . Hence, the mean minimizes the sum of squared errors; any other estimate necessarily gives a larger sum of squared errors.