***Case Study-Classification Problem***

***Submitted by: T. Vijaya***

***INTRODUCTION***

We know that an outcome can be predicted from a number of predictor variables using multiple regression. Logistic regression is multiple regression but with an outcome variable that is a categorical dichotomy and predictor variables that are continuous or categorical. This simply means that we can predict which of two categories a person is likely to belong to given certain other information. The regression problem which we already know is called as a *classification* problem if the response is a *discrete* variable. In simpler words we want to classify an observation (univariate or multivariate) into one of several possible classes, or simply we want to estimate the probability given an observation that it belongs to one of the several possible classes.

Despite the similarities between linear regression and logistic regression, wecan’t apply linear regression directly to a situation in which the outcome variable is dichotomous. The reason is that one of the assumptions of linear regression is that the relationship between variables is linear. For linear regression to be a valid model, theobserved data should contain a linear relationship. When the outcome variable is dichotomous this assumption is usually violated. One way around this problem is to transform the data using logarithmic transformation. This has the effect of making the form of the relationship linear whilst leaving the relationship itself as non-linear (It’s a way of expressing a non-linear relationship in a linear way).The logistic regression equation is based on this principle, it expresses the multiple linear regression equation in logarithmic terms and thus overcomes the problem of violating the assumption of linearity. Due the fact that a discrete variable can’t be normally distributed the application of linear regression becomes invalid as the assumption of normality of the observations is no longer satisfied. Indeed the response follows a multinomial distribution.

Logistic regression is another approach to category prediction. This method carries fewer assumptions than does discriminant analysis: neither multivariate normality nor homogeneity of variance-covariance matrices isrequired. Discriminant analysis is also sensitive to the inclusion of qualitative independent variables such as blood group, gender or nationality, logistic regression on the other hand can cope with any number of qualitative regressors: in fact all predictors can be categorical. For these reasons logistic regression is fast overtaking discriminant analysis as the preferred technique for prediction of dichotomous category membership.

Moreover, the expectation of response becomes uninterruptable in terms of a linear function of the features and also the variances do not remain constant across observations and hence causing heteroscedasticity. Hence such problems are outside the ambit of linear regression. There are several tools namely *naïve Bayes classifier*, *logistic classifier*, *discriminant analysis*, *nearest neighbor approach*, *neural network* etc. are available to be deployed in such situations. We will try some of them. The classification problems are quite naturally divided into two types:

1. Binary, where response has two possible classes meaning by an observation either belongs to a class or it doesn’t, and
2. Multiclass, where the response has more than two possible classes.

Multinomial Logistic Regression is useful for situations in which you want to be able to classify subjects based on values of a set of predictor variables. This type of regression is similar to logistic regression, but it is more general because the dependent variable is not restricted to two categories.

1. **Binary Classification**

**1.1-Email Spam Filtering using NaïveBayes Classifier**

The problem of classifying an email as either *spam*or *non-spam* (*ham*) is an interesting and relevant problem in the present world. Now the question is how we go about it? How do we deal with the other problems? An email will have some information in itself about the fact that its spam or not. That information we have to find out and based on that we will estimate the probability of an email being spam. An email is nothing but a *text*, which has words, symbols and numbers but all is text. This entire textincluding the no. of persons to whom it’s been sent, the name of the persons to whom it’s been sent, the cc list, the bcc list (which we can’t see though), the subject line, the body message (header, main message, the signature and the postscripts) will have information about the message being spam or non-spam. So I hope you are getting motivated towards a basic philosophy. Now we will introduce and make use of a technique called as the *Naïve Bayes Classifier* for the purpose of email spam filtering.

* **Naïve Bayes Classifier**

Naïve Bayes classifier is a general technique for classification but we will study it in the context of spam filtering. So far, one thing is clear that the words of the message will tell us about it being spam or not. We can easily make a list of some common words which are generally seen in spam messages in a large frequency like congratulations, currency symbols, big numeric values, replica, derivative, claim, property, wealth etc. We have an idea of some common words. Or we can simply collect some 10-15 high frequency words from the spam emails in the training data.

One thing should be clear is that in order to make a *classifier/rule* for spam classification we obviously need some data, i.e. some emails (some of which are spam and some of which are non-spam), this available data in the context of machine learning (which I don’t expect you to know as of now) is called as the training data. This training is generally very large and maintains approximately a 50-50%ratio of the spam and non-spam messages. Now let’s see how does the naïve Bayes classifier work? It’s 3-step process, viz.

1) Computing the probability that the message is spam, knowing that a given word appears in this message;

2) Computing the probability that the message is spam, taking into consideration all of its words (or a relevant subset of them);

3) Dealing with rare words.

As a natural rule we do not consider the *stop words* like helping verbs, propositions etc. into consideration. For the purpose of above three steps we make use of the Bayes theorem in an absolute manner and hence the name Naïve Bayes Classifier.

1. **Computing the probability that the message is spam, knowing that a given word appears in this message:**

Let's suppose the suspected message contains the word ‘replica’. Most people who are used to receiving e-mail know that this message is likely to be spam, more precisely a proposal to sell counterfeit copies of well-known brands of watches (say). The spam detection software, however, does not "know" such facts; all it can do is compute probabilities. The formula used by the software to determine that is derived from Bayes' theorem:

, is the probability that a message is a spam, knowing that the word ‘replica’ is in it;, is the overall probability that any given message is spam;, is the probability that the word ‘replica’ appears in spam messages;, is the overall probability that any given message is not spam (is "ham");, is the probability that the word ‘replica’ appears in ham messages.  
**The ‘spamicity’ of a word:**

Recent statistics show that the current probability of any message being spam is 80% at the very least, which implies the following:

However, most Bayesian spam detection software makes the assumption that there is no *a priori* reason for any incoming message to be spam rather than ham, and considers both cases to be equally likely.

The filters that use this hypothesis are said to be "*not biased*", meaning that they have no prejudice regarding the incoming email. This assumption permits simplifying the general formula to the following:

This is functionally equivalent to asking: "*What percentage of occurrences of the word ‘replica’ appears in spam messages?* “This quantity is called "*spamicity*" (or "*spaminess*") of the word ‘replica’, and can be computed.

The number  used in this formula is approximated to the frequency of messages containing ‘replica’ in the messages identified as spam during the learning phase. Similarly,  is approximated to the frequency of messages containing ‘replica’ in the messages identified as ham during the learning phase. For these approximations to make sense, the set of learned messages needs to be big and representative enough. It is also advisable that the learned set of messages conforms to the 50% hypothesis about repartition between spam and ham, i.e. that the datasets of spam and ham are of same size. How this process does actually takes place? Suppose we have 500 spam emails and 500 ham emails (approximately half) we need to find then we can write it as follows:

Where is calculated as the relative frequency of the word in the spam messages. is taken a priori. Similarly we can calculate.

Of course, determining whether a message is spam or ham based only on the presence of the word ‘replica’ is error-prone, which is why Bayesian spam software tries to consider several words and combine their spamicities to determine a message's overall probability of being spam.

**b) Computing the probability that the message is spam, taking into consideration all of its words (or a relevant subset of them):**

Most Bayesian spam filtering algorithms are based on formulas that are strictly valid (from a probabilistic standpoint) only if the words present in the message areindependent events. This condition is not generally satisfied (for example, in natural languages like English the probability of finding an adjective is affected by the probability of having a noun), but it is a useful idealization, especially since the statistical correlations between individual words are usually not known. On this basis, one can derive the following formula from Bayes' theorem:

 is the probability that the suspect message is spam;is the probability  that it is a spam knowing it contains a first word (for example ‘replica’); is the probability  that it is a spam knowing it contains a second word (for example ‘watches’);is the probability  that it is a spam knowing it contains an *n*th word (for example ‘home’).

This is the formula referenced by *Paul Graham* in his 2002 article. Spam filtering software based on this formula is sometimes referred to as a*Naïve Bayes classifier.* The result  is typically compared to a given threshold to decide whether the message is spam or not. If islower than the threshold, the message is considered as likely ham, otherwise it is considered as likely spam. The probabilities which will be used for the purpose of classification are given in the 2nd and 3 rd columns of the following table:   
**Case 1:** Calculate the overall spamicity of the following emails and classify them as spam or non-spam. Assume that spam and non-spam emails are equally probable in nature

**Email 1:** *Congratulations on winning the $ 100,000,000 in the lottery. To claim the prize, send your contact details to* [*lucky@xyz.com*](mailto:lucky@xyz.com)*.*

*On computing the overall spamicity of email 1 using excel(calculations are shown below) we get 0.999784.Hence the probability of email 1 being spam is 0.999784.We have classified email 1 as spam.*

**Email 2:** *Everything is going fine. I will not be coming for summer holidays. Take care of yourself.*

*On computing the overall spamicity of email 2 using excel(calculations are shown below) we get 0.233577.Hence the probability of email 2 being spam is 0.233577.We have classified email 2 as non-spam or ham.*

***Table 1.1-Calculations***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Word** | **P(Word | Spam)** | **P(Word | Ham)** | **pi** | **1-pi** |
| Congratulations | 0.8 | 0.2 | 0.8 | 0.2 |
| winning | 0.7 | 0.4 | 0.636364 | 0.363636 |
| $ | 0.9 | 0.2 | 0.818182 | 0.181818 |
| 100000000 | 0.7 | 0.1 | 0.875 | 0.125 |
| lottery | 0.6 | 0.2 | 0.75 | 0.25 |
| claim | 0.6 | 0.3 | 0.666667 | 0.333333 |
| prize | 0.6 | 0.4 | 0.6 | 0.4 |
| send | 0.5 | 0.5 | 0.5 | 0.5 |
| you | 0.7 | 0.3 | 0.7 | 0.3 |
| contact | 0.5 | 0.5 | 0.5 | 0.5 |
| details | 0.6 | 0.6 | 0.5 | 0.5 |
| Everything | 0.2 | 0.7 | 0.222222 | 0.777778 |
| going | 0.2 | 0.7 | 0.222222 | 0.777778 |
| fine | 0.7 | 0.5 | 0.583333 | 0.416667 |
| I | 0.5 | 0.5 | 0.5 | 0.5 |
| coming | 0.5 | 0.5 | 0.5 | 0.5 |
| summer | 0.6 | 0.6 | 0.5 | 0.5 |
| holidays | 0.8 | 0.4 | 0.666667 | 0.333333 |
| Take | 0.7 | 0.6 | 0.538462 | 0.461538 |
| care | 0.2 | 0.2 | 0.5 | 0.5 |
| yourself | 0.8 | 0.7 | 0.533333 | 0.466667 |
|  | For 1 st mail | P | 0.999784 |  |
|  | For 2nd mail | P | 0.233577 |  |

**Logistic Regression**

Consider the multiple linear regression model (MLRM) for sample observations:

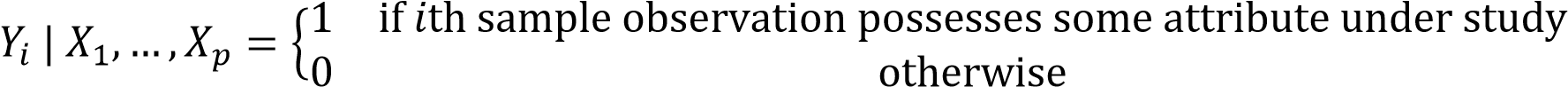


Equivalently,



Where and  .

Suppose  given  is a categorical random variable. For simplicity let  is an indicator variable defined as follows.



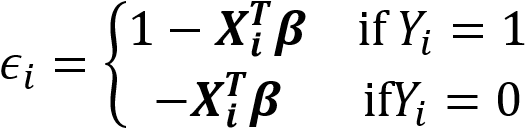
Clearly  where and .

In multiple linear regression  represents the conditional expectation of  given which is free to take any real value. But if the dependent variable is an indicator variable then it is inappropriate to represent  as **,** as expectation of an indicator variable is nothing but the probability of it taking value 1 which is bound to lie between 0 and 1.

Also in multiple linear regression is estimated using its conditional expectation given the independent variables, i.e.



Is a discrete random variable taking only two values and  is a continuous random variable which is free to take any real value. Hence it is inappropriate to use  as an estimator of  . Also when dependent variable is binary, error component is defined as follows in MLRM.



Hence  does not have a normal distribution. Moreover it can be seen that  is not constant. Hence CMLRM is not a good choice of model when response is categorical or specifically binary.

The idea to solve this problem is to replace by some other function of it in the model which can represent the conditional expectation or equivalently conditional probability of success of, i.e.

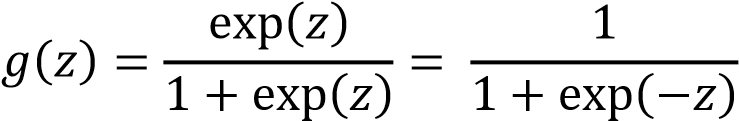


Some function of which is bound to lie between 0 and 1 will serve the purpose. Generally when response variable is binary, there is considerable empirical evidence indicating that the relationship between  and  is non-linear.

Hence we model the conditional probabilities using a non-linear function of the independent variables of the following form.



The function  is called as a link function. The most common link function is the **logit link function or logistic function or sigmoid function** which is defined as follows.



0

0.2

0.4

0.6

0.8

1

1.2

-15

-10

-5

0

5

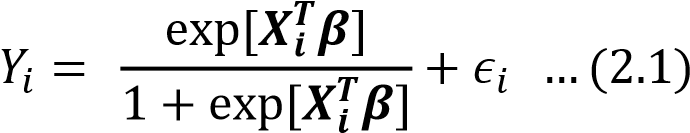
10

15

z

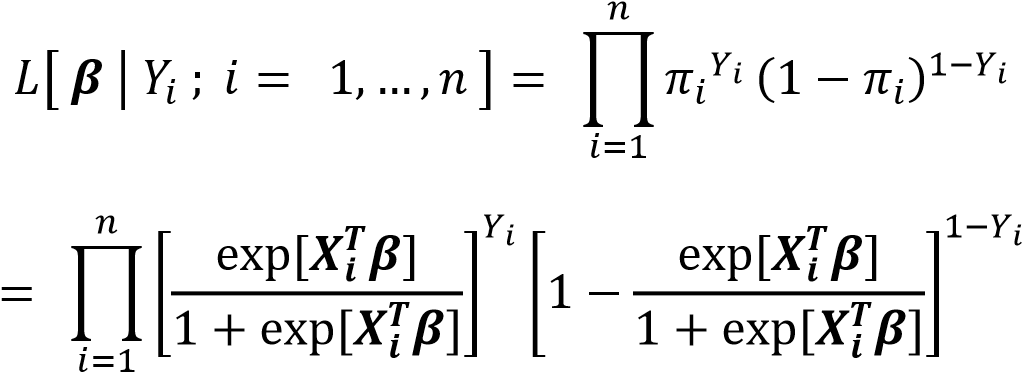
**Sigmoid Function**

In terms of the logit link function the model under study is actually given by

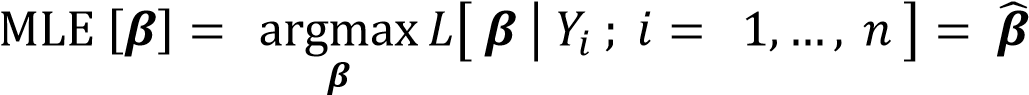


Observe that the above expression specifies a **non-linear regression model** which occurs when the dependent variable depends on independent variables through a non-linear function of unknown parameters. The stochastic model defined in (2.1) is called as a **logistic regression model**.

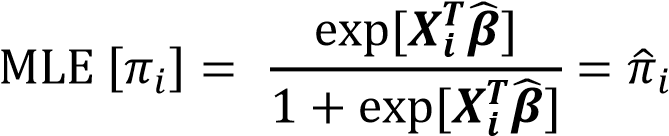
After specifying an appropriate model for modeling a binary response variable, the next task is to estimate the parameters. For estimation of parameters we use the famous maximum likelihood estimation (MLE). Consider the available observations  as independent observations from the conditional distribution of  given. The likelihood function of the unknown parameters  based on the observations  is given by the joint conditional density function of . As , the likelihood function is given by,



The maximum likelihood estimator of is defined as maxima of  with respect to, i.e.



Explicit maximization of likelihood function is not possible, although iterative methods like Iterative Reweighted Least Squares (IRLS) can be used. Hence using invariance property of MLE,



The estimated conditional probability of  is given by. Hence a classification rule can be given as follows:



**Note:** Logistic regression is actually a classification technique. Following are some common applications.

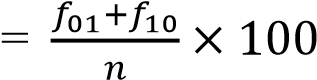
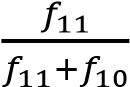
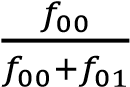
1. Spam Detection: Spam/Non-Spam
2. Tumor Examination: Malignant/Benign (iii) Online Transaction: Fraudulent/Non-Fraudulent (iv) Email Labeling: Work, Friends, Family etc.

# 1.1 **Model Adequacy Checking**

A simple  classification table termed as **confusion matrix** can be constructed as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Confusion Matrix** |  |  | **Estimated Response** | |
| **0** |  | **1** |
| **Observed Response** | **0** |  |  |  |
| **1** |  |  |  |

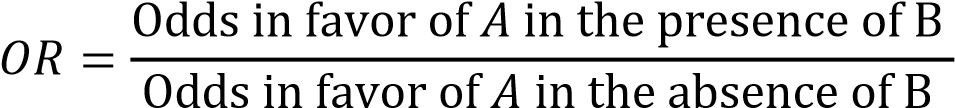
Whereand, is the frequency of the class where observed response is  and estimated response .Following simple measures of goodness can be defined.

1. Percentage of Misclassification: Less is better.
2. Sensitivity =: More is better.
3. Specificity =: More is better.

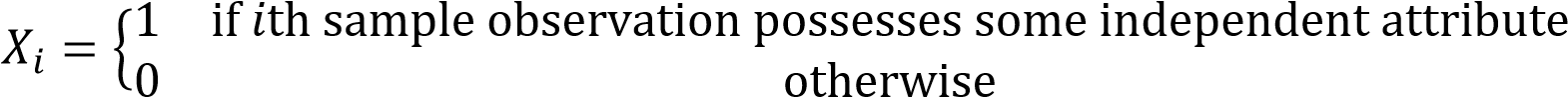
Hosmer and Lemeshow chi-square goodness of fit test used to test the goodness of logistic model.

# **Odds Ratio**

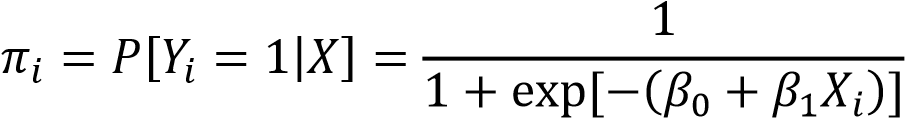
Odds Ratio (OR) is a measure of association between the two attributes and, say defined as follows:



Consider the case when there is single indicator independent variable, i.e.



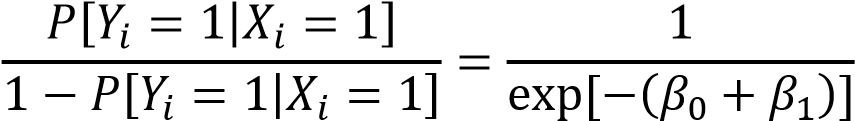
As defined earlier we have,



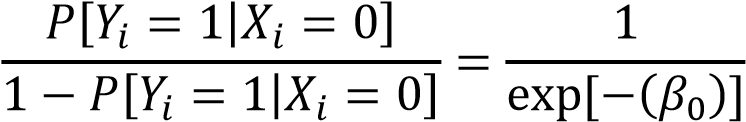
Consider the following table.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 𝑃  [  𝑌  𝑖  =  1  |  𝑋  𝑖  =  1  ]  =  1  1  +  exp  [  −  (  𝛽  0  +  𝛽  1  𝑋  𝑖  )  ] | 𝑃  [  𝑌  𝑖  =  1  |  𝑋  𝑖  =  0  ]  =  1  1  +  exp  [  −  (  𝛽  0  )  ] |
|  | 𝑃  [  𝑌  𝑖  =  0  |  𝑋  𝑖  =  1  ]  =  1  −  1  1  +  exp  [  −  (  𝛽  0  +  𝛽  1  𝑋  𝑖  )  ]  =  exp  [  −  (  𝛽  0  +  𝛽  1  )  ]  1  +  exp  [  −  (  𝛽  0  +  𝛽  1  )  ] | 𝑃  [  𝑌  𝑖  =  0  |  𝑋  𝑖  =  0  ]  =  1  −  1  1  +  exp  [  −  (  𝛽  0  )  ]  =  exp  [  −  (  𝛽  0  )  ]  1  +  exp  [  −  (  𝛽  0  )  ] |

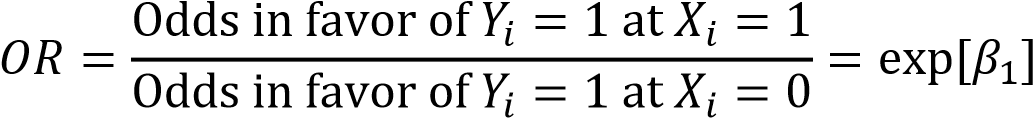
Odds in favor of 𝑌𝑖= 1 at 𝑋𝑖= 1 are given by



Odds in favor of 𝑌 = 1 at 𝑋 = 0 are given by



Hence Odds ratio is given by



Hence odds ratio in case of logistic regression has a nice interpretation in terms of the parameter of the independent variable. Suppose the response is presence/absence of lungs cancer and the independent variable is presence/absence of smoking and odds ratio exp [𝛽1] = 2 is interpreted as smoker is twice more likely to have lungs cancer as compared to a non-smoker.

It is basically an indicator of the change in odds resulting from a unit change in the predictor.

Our model will be constructed by an iterative maximum likelihood procedure. The program will start with arbitrary values of the regression coefficients and will construct an initial model for predicting the observed data. It will then evaluate errors in such prediction and change the regression coefficients so as to make the likelihood of the observed data greater under the new model. This model is repeated until the model converges-that is until the differences between the two are trivial.

The default method of conducting the regression is enter. In this all the covariates are placed into the regression model in one block, and parameter estimates are calculated for each block. The crucial statistic is the Wald statistic, which has a chi-square distribution and tells us whether the ß coefficient for that predictor is significantly different from 0.If the coefficient is significantly different from zero then we can assume that the predictor is making a significant contribution to the prediction of the outcome (Y).In the sense it is analogous to the t tests found in multiple regression.

***Wald statistic tests a regression coefficient (or the constant) for significance with null hypothesis: the value of the parameter is zero. That is the predictor does not make a significant contribution to regression.***

***Decision Rule: Reject H0 if p value<0.05***

The Wald statistic is defined as follows

Wald2 = (ß/S.E.)2 (SPSS actually quotes the Wald statistic squared)

And is distributed approximately as chi square with one degree of freedom.

The Wald statistic is used to ascertain whether a variable is a significant predictor of the outcome: however, it is probably more accurate to examine the likelihood ratio statistics. The reason why the Wald statistic should be used cautiously is because when the regression coefficient (ß) is large, the standard error tends to become inflated, resulting in the Wald statistic being underestimated. The inflation of the standard error increases the probability of rejecting a predictor as being insignificant when in reality it is making a significant contribution to the model (that is we are likely to make a type II error).

We saw in linear regression that the value of ß represents the change in the outcome resulting from a unit change in the predictor variable. The interpretation of this coefficient in logistic regression is very similar in that it represents the change in the logit of the outcome variable associated with one unit change in the predictor variable. The logit of the outcome is simply the natural logarithm of the odds of Y occurring.

The Hosmer Lemeshow goodness of fit statistic can be used to assess how well the chosen model fits the data. The Hosmer Lemeshow tests the null hypothesis that predictions made by the model fit perfectly with observed group memberships. Cases are arranged in order by their predicted probability on the criterion variable. These ordered cases are then divided into 10 groups of equal or near equal size ordered with respect to the predicted probability of the target event. For each of these groups we then obtain the predicted group memberships and the actual group memberships. This results in a 2x10 contingency table. A chi square statistic is computed comparing the observed frequencies with those expected under the linear model. A non -significant chi square indicates that the data fit the model well.

**Drawbacks**: It relies on a test of significance. With large sample sizes the test may be significant even when the fit is good. With small sample sizes it may not be significant, even with poor fit.

It is highly dependent on how the observations are grouped. If too few groups are used (5 or less) it almost always indicates that the model fits the data, this means that it’s usually not a good measure if you only have one or two categorical predictors. It is best used for continuous predictors.

***Assumptions*:**

* The data *Y*1, *Y*2... *Yn* are independently distributed, i.e., cases are independent.
* Distribution of *Yi* is *Bin* (*ni*, π*i*), i.e., binary logistic regression model assumes binomial distribution of the response. The dependent variable does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal...)
* Does NOT assume a linear relationship between the dependent variable and the independent variables, but it does assume linear relationship between the logit of the response and the explanatory variables; *logit* (π) = β0 + β*X*.
* Independent (explanatory) variables can be even the power terms or some other nonlinear transformations of the original independent variables.
* The homogeneity of variance does NOT need to be satisfied. In fact, it is not even possible in many cases given the model structure.
* Errors need to be independent but NOT normally distributed.
* It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.
* Goodness-of-fit measures rely on sufficiently large samples, where a heuristic rule is that not more than 20% of the expected cells counts are less than 5.

**For the Hosmer Lemeshow test -**We set the following hypothesis:

H0: Model adequately fits the data

H1: Model inadequately fits the data

**Prediction of Cancer due to Smoking using Logistic Regression**

Given the data on a binary response variable telling us whether the cancer is present or not and a single binary independent variable telling whether the person smokes or not we want to predict the possibility of cancer due to smoking. In nutshell we want to know “how more likely is a person to have cancer if he/she smokes rather he/she doesn’t”. We are supposed to do the following:

A study was performed on lung cancer possibility due to smoking habits. Data on presence/absence of two attributes viz. lung cancer and smoking was collected for 25 individuals.

**Case 3:**Consider the dataset **Smoking and Cancer.xlsx** and perform the following objectives.

1. Build a logistic regression model for cancer possibility using smoking as an independent variable.
2. Test for the Significance of independent variable.
3. Construct the Confusion (Classification) Table and report the percentage of correct classification in the given emails. Also calculate specificity and sensitivity of the model.
4. For each person obtain the probability of him/her having cancer and hence the predictionof cancer using the Logistic Classifier you have built.
5. Estimate the odds ratio and interpret it.

Let us first consider a simple (bivariate) logistic regression using lung cancer possibility as the dichotomous criterion variable and their smoking status as a dichotomous predictor variable. We have coded lung cancer possibility as: 0= Possibility of not having lung cancer and 1=Possibility of having lung cancerand smoking status(X –the predictor variable) as: 0=Does not smoke and 1=Smokes

Our regression model will be predicting the logit, that is the natural log of the odds of having cancer or not. Thatis ln(ODDS) =ln (Ŷ/1-Ŷ) =a+bX where Ŷ is the predicted probability of the event which is coded with 1 rather than with 0(possibility of not developing cancer), 1-Ŷ is the predicted probability of the other decision, and X is our predictor variable, smokingAfter doing binary logistic regression, we get the following output.

| ***Table 1.2-Variables in the Equation*** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 1a | X(Smoking) | .770 | .823 | .875 | 1 | .350 | 2.160 |
| Constant | -.182 | .606 | .091 | 1 | .763 | .833 |
| a. Variable(s) entered on step 1: X. | | | | | | | |

The variables in the equation output show us that the regression equation is: ln (ODDS) =

-0.182+.770(Smoking)

The Wald statistic gives the significance of each component of the logistic regression. Smoking is insignificant at p (>0.05)(Wald 0.875, DF=1, sig=0.350)Hence we accept our null hypothesis and conclude that smoking does not have a significant effect on the prediction. For the constant term (Wald 0.091, DF=1, sig=0.763).As p (>0.05) it is insignificant.

The value of B for smoking is 0.770 .This means that an increase in smoking level of one unit produces, on average an increase of 0.770 units in the logit (i.e. the natural log of the odds) in favor of having cancer. But an increase of 0.770 units in the logarithm corresponds to multiplication of the raw odds by Exp(.770)=2.160.In words, an increase of one unit in smoking, multiplies the likelihood of having cancer by 2.160.

We can now use this model to predict the odds that a subject with a given smoking status

Will get cancer or not.The Odds prediction equation is ODDS=e a+bx .If our subject does not smoke (smoking=0),then the ODDS =e -0.182+0.770(0) =0.8336.That is, a subject who does not smoke is only 0.8336 times as likely to get cancer as subject is to not getting cancer. If our subject smokes (smoking =1),then the ODDS =e -0.182+0.770(1)=1.8004.That is, a person who smokes is 1.8004 times more likely to get cancer as compared to the possibility of not getting cancer.

We can easily convert odds to probabilities. Fornon-smokers=(ODDS/1+ODDS)=0.4546

.That is, our model predicts that ***45%*** of non-smokers will get cancer. For smokers=0.6429.That is,our model predicts that ***64%*** of smokers will get cancer.

If odds ratio=1 then same probability of event occurring between two situations, ***If odds ratio>1 then probability of event occurring with unit increase in independent variables is higher than at original value of independent variable***. If odds ratio<1, probability of event occurring with unit increase in independent variable is lower than at original independent variable.

The last column gives us Exp (B).This is better known as the odds ratio predicted by the model. For our model, e .770 =2.160.That tells us that the model predicts that the odds of getting cancer are 2.160 times higher for smokers than they are fornon-smokers. For smokers the odds are 1.8004 and for non-smokers is 0.8336.The odds ratio is (1.8004/0.8336=2.160). We infer that suppose the response is presence/absence of lungs cancer and the independent variable is presence/absence of smoking and odds ratio exp [𝛽1] = 2.16 is interpreted as smoker is twice more likely to have lungs cancer as compared to a non-smoker.

We have not used Hosmer Lemeshow goodness of fit test because it needs more than one independent variable (the data contains only one predictor variable which is categorical). The results of our logistic regression can be used to classify subjects with respect to the possibility of cancer.Before we use this information to classify subjects, we need to have a decision rule. Our decision rule will take the following form: If the probability of the event is greater than or equal to some threshold, we shall predict that the event will take place. By default SPSS sets this threshold to 0.5.Using the default threshold,SPSS will classify a subject into will have cancer category if the estimated probability is 0.5 or more .SPSS will classify a subject into the wont have cancer category if the estimated probability is less than 0.5.

The classification table displays the number of observed cases that are correctly predicted by the model and the overall percentage of the cases that are predicted by the model.

From the classification table we calculated the percentage of misclassification as 40%.

The classification table shows us that this rule allows us to correctly classify 9/14=***64.3 %*** of the subjects where the predicted event (will have cancer) wasobserved. This is known as the ***sensitivity of prediction,*** theP (correct |event did occur), thatis, the percentage of occurrences correctly predicted. We also see that this rule allows us to correctly classify 6/11=***54.5%*** of the subjects where the predicted event was not observed. This is known as the ***specificity of prediction,*** the P (correct |event did not occur), that is, the percentage of non-occurrences correctly predicted. Overall our predictions were correct 15 out of 25 times, for an overall success rate of 60%.

| ***Table 1.3-Classification Table(a)*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Observed | | Predicted | | |
|  | Lung Cancer | | Percentage Correct |
|  | 0 | 1 |
| Step 1 | Lung Cancer | 0 | 6 | 5 | 54.5 |
| 1 | 5 | 9 | 64.3 |
| Overall Percentage | |  |  | 60.0 |
| a. The cut value is .500 | | | | | |

We could focus on error rates in classification. A false positive would be predicting that the event would occur when, infact, it did not. Our decision rule predicted possibility of having cancer 14 times. That prediction was wrong 5 times, for a false positive rate of 5/14=35.7%.A false negative would be predicting that the event would not occur when, infact, it did occur. Our decision rule predicted possibility of not having cancer 11 times. That prediction was wrong 5 times, for a false negative rate of 5/11=45.4%.

NOTE-To summarize the ***relationship between two categorical variables, we use a cross-tabulation (also called a contingency table).***A cross-tabulation (or crosstab for short) is a table that depicts the number of times each of the possible category combinations occurred in the sample data

We could have used a simple Pearson chi-square Contingency table analysis to answer the question whether or not there is a significant relationship between smoking and possibility of cancer.

| ***Table 1.4-Smoking \* Lung Cancer Cross-tabulation*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  |  | Lung Cancer | | Total |
|  |  |  | 0 | 1 |
| Smoking | 0 | Count | 6 | 5 | 11 |
| % within Smoking | 54.5% | ***45.5%*** | 100.0% |
| 1 | Count | 5 | 9 | 14 |
| % within Smoking | 35.7% | ***64.3%*** | 100.0% |
| Total | | Count | 11 | 14 | 25 |
| % within Smoking | 44.0% | 56.0% | 100.0% |

In the cross tabulation output we see that 64.3% of smokers and 45.5% of non-smokers developed cancer, just as predicted by our logistic regression.

| ***Table 1.5-Chi-Square Tests*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Value | Df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | .887a | 1 | .346 |  |  |
| Continuity Correctionb | .287 | 1 | .592 |  |  |
| Likelihood Ratio | .889 | 1 | .346 |  |  |
| Fisher's Exact Test |  |  |  | .435 | .296 |
| Linear-by-Linear Association | .851 | 1 | .356 |  |  |
| N of Valid Cases | 25 |  |  |  |  |
| a. 1cells (25.0%) have expected count less than 5. The minimum expected count is 4.84. | | | | | |
| b. Computed only for a 2x2 table | | | | | |

We notice that the Likelihood Ratio Chi Square is 0.889 on 1 df and the Pearson chi square is almost the same (0.887).We infer that this logistic regression is nearly equivalent to a simple Pearson chi square .We can add additional predictor variables, and those additional predictor variables can be either categorical or continuous (logistic regression).We can’t do that with a simple Pearson Chi square.

For each person we have obtained the probability of him/her having cancer and hence the prediction of cancer using the Logistic Classifier that we have built.

| ***Table 1.6-Case Summaries*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | Lung Cancer | Smoking | Predicted probability | Predicted group |
| 1 | | 1 | 0 | .45455 | 0 |
| 2 | | 0 | 0 | .45455 | 0 |
| 3 | | 1 | 1 | .64286 | 1 |
| 4 | | 0 | 1 | .64286 | 1 |
| 5 | | 0 | 0 | .45455 | 0 |
| 6 | | 1 | 0 | .45455 | 0 |
| 7 | | 1 | 0 | .45455 | 0 |
| 8 | | 0 | 0 | .45455 | 0 |
| 9 | | 0 | 1 | .64286 | 1 |
| 10 | | 1 | 1 | .64286 | 1 |
| 11 | | 0 | 0 | .45455 | 0 |
| 12 | | 0 | 0 | .45455 | 0 |
| 13 | | 1 | 1 | .64286 | 1 |
| 14 | | 0 | 1 | .64286 | 1 |
| 15 | | 1 | 1 | .64286 | 1 |
| 16 | | 1 | 0 | .45455 | 0 |
| 17 | | 1 | 1 | .64286 | 1 |
| 18 | | 0 | 1 | .64286 | 1 |
| 19 | | 1 | 1 | .64286 | 1 |
| 20 | | 1 | 0 | .45455 | 0 |
| 21 | | 1 | 1 | .64286 | 1 |
| 22 | | 1 | 1 | .64286 | 1 |
| 23 | | 0 | 0 | .45455 | 0 |
| 24 | | 0 | 1 | .64286 | 1 |
| 25 | | 1 | 1 | .64286 | 1 |

**The values in column 5 are the predicted groups with corresponding calculated probabilities in column 4.These values tell us that when a subject does not smoke(X=0),there is a probability of 0.45455 of developing lung cancer –approximately a 45% chance. However if the subject smokes(X=1),there is a probability of 0.64286 of developing lung cancer-approximately 64%.Consider that a probability of 0 indicates no chance of the subject developing lung cancer ,and a probability of 1 indicates that the subject will definitely develop cancer. Therefore the values obtained provide strong evidence for the role of smoking as a potential cause of cancer. Hence we have predicted the possibility of getting cancer for all the subjects depending on their smoking habits where (X=0 -does not smoke and X=1 -smokes).**

* 1. **Skull Type Prediction using Logistic Regression**

We are interested in predicting the type of skull of humans as one of two possible types I and II based on some five physical measures available related to the skulls.

**Case 4:** Consider the dataset **Skull Type Prediction.xlsx** and perform the following objectives.

1. Build a logistic regression model for classifying a human skull as Type I/Type II using the given independent variables.
2. Test for the Significance individual independent variables.
3. Test for the overall Logistic Regression using Hosmer and Lemeshow Test (It’s a Chi-Square Test).
4. Construct the Confusion (Classification) Table and report the percentage of correct classification in the given skulls. Also calculate specificity and sensitivity of the model.
5. For each skull obtain the probability of it being Type I or Type II, and hence predict the skull Type using the Logistic Classifier you have built.
6. For a set of five physical measures given for a new skull in the dataset**Skull Type Prediction – Validation Data.xlsx**predict the skull type using the Logistic Classifier you have built.

After conducting binary logistic regression we get the following output-

| ***Table 1.7-Hosmer and Lemeshow Test*** | | | |
| --- | --- | --- | --- |
| Step | Chi-square | Df | Sig. |
| 1 | 9.601 | 8 | .294 |

The important part of this test is the test statistic itself (9.601) and the significance value (0.294).A non-significant chi square (as p value is 0.294>0.05) indicates that the data fits the model well. Anon-significant value is indicative of a model that is predicting the real world data fairly well.As the p value is high; it indicates that all the systematic variance has been accounted for by the model: the rest is error.

| ***Table 1.8-Contingency Table for Hosmer and Lemeshow Test*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Skull type = .00 | | Skull type = 1.00 | | Total |
|  |  | Observed | Expected | Observed | Expected |
| Step 1 | 1 | 2 | 1.841 | 0 | .159 | 2 |
| 2 | 2 | 1.774 | 0 | .226 | 2 |
| 3 | 2 | 1.446 | 0 | .554 | 2 |
| 4 | 0 | 1.374 | 2 | .626 | 2 |
| 5 | 2 | 1.226 | 0 | .774 | 2 |
| 6 | 1 | 1.073 | 1 | .927 | 2 |
| 7 | 0 | .891 | 2 | 1.109 | 2 |
| 8 | 1 | .778 | 1 | 1.222 | 2 |
| 9 | 1 | .451 | 1 | 1.549 | 2 |
| 10 | 0 | .146 | 1 | .854 | 1 |

The 1 st column categorizes, in order of increasing magnitude, the probabilities assigned by the regression model into divisions known as deciles. The table shows the association between assigned probability and type of skull. In general there is close agreement between the expected frequencies (the assignments by the regression model and category assignment on the basis of the cutoff point of 0.05 for probability) and the observed or actual frequencies of subjects in those categories.

. The results of our logistic regression can be used to classify subjects with respect to the possibility of having type I or type II skull. Before we use this information to classify subjects, we need to have a decision rule. **Our decision rule will take the following form: If the probability of the event is greater than or equal to some threshold, we shall predict that the event will take place. By default SPSS sets this threshold to 0.5** .The classification table shows the proportion of correct assignments when the regression model has been applied to the data Overall our predictions were correct 14 out of 19 times for an overall success rate of 73.7%.From the classification table we calculated the percentage of misclassification as 26.3157%.

Sensitivity-P(correct |event did occur), thatis, the percentage of occurrences correctly predicted is 5/8=***62.5%.***The decision rule allows us to correctly classify 9/11=***81.8%*** of the subjects where the predicted event was not observed. This is known as the specificity of prediction, the P (correct |event did not occur), that is, the percentage of non-occurrences correctly predicted.

| ***Table 1.9-Classification Tablea*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Observed | | Predicted | | |
|  | Skull type | | Percentage Correct |
|  | .00 | 1.00 |
| Step 1 | Skull type | .00 | 9 | 2 | 81.8 |
| 1.00 | 3 | 5 | 62.5 |
| Overall Percentage | |  |  | 73.7 |
| a. The cut value is .500 | | | | | |

We could focus on error rates in classification. A false positive would be predicting that the event would occur when, in fact, it did not. Our decision rule predicted possibility of having skull type II 7 times. That prediction was wrong 2 times, for a false positive rate of 2/7=28.5%.A false negative would be predicting that the event would not occur when, in fact, it did occur. Our decision rule predicted possibility of having type I skull 12 times. That prediction was wrong 3times, for a false negative rate of 3/12=25%.

***Table 1.10-Variables in equation***

|  |  | B | S.E. | Wald | Df | Sig. | Exp(B) |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |
| Step 1a | X1(measure 1) | -.008 | .018 | .185 | 1 | .668 | .992 |
| X2(measure 2) | -.047 | .033 | 2.039 | 1 | .153 | .954 |
| X3(measure 3) | -.007 | .020 | .113 | 1 | .737 | .993 |
| X4(measure 4) | -.006 | .024 | .054 | 1 | .816 | .994 |
| X5(measure 5) | .022 | .020 | 1.245 | 1 | .264 | 1.022 |
| Constant | 1.149 | 2.826 | .165 | 1 | .684 | 3.155 |

**Logistic regression model is:ln (ODDS) =1.149-0.008X1-0.047X2-0.007X3-0.006X4+0.022X5**

From the above table we observe thatas p value is greater than 0.05, we accept our null hypothesis and conclude that none of the predictors makes a significant contribution to the regression model. The values of Exp (B) for X1(measure 1) indicates that if the percentage of measure 1 goes up by one, then the odds of having type I skull decrease (because exp b is less than one).Similarly for measure 2, 3 and 4.But if the percentage of measure 5 goes up by one, then the odds of having type I skull increase.

| ***Table 1.11-Case Summaries*** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Skull type | X1 | X2 | X3 | X4 | X5 | Predicted probability | Predicted group |
| 1 | | 1.00 | 13.00 | 29.00 | 82.00 | 24.00 | 60.00 | .58476 | 1.00 |
| 2 | | .00 | 79.00 | 1.00 | 1.00 | 56.00 | 63.00 | .82867 | 1.00 |
| 3 | | .00 | 5.00 | 77.00 | 47.00 | 45.00 | 26.00 | .07618 | .00 |
| 4 | | 1.00 | 100.00 | 16.00 | 80.00 | 60.00 | 98.00 | .72021 | 1.00 |
| 5 | | .00 | 55.00 | 65.00 | 3.00 | 20.00 | 2.00 | .08238 | .00 |
| 6 | | .00 | 91.00 | 55.00 | 20.00 | 59.00 | 68.00 | .25299 | .00 |
| 7 | | 1.00 | 47.00 | 31.00 | 31.00 | 52.00 | 19.00 | .32244 | .00 |
| 8 | | 1.00 | 17.00 | 43.00 | 61.00 | 45.00 | 79.00 | .52374 | 1.00 |
| 9 | | 1.00 | 30.00 | 54.00 | 11.00 | 83.00 | 60.00 | .30390 | .00 |
| 10 | | .00 | 45.00 | 17.00 | 63.00 | 79.00 | 10.00 | .34643 | .00 |
| 11 | | .00 | 1.00 | 40.00 | 97.00 | 51.00 | 94.00 | .60310 | 1.00 |
| 12 | | 1.00 | 69.00 | 44.00 | 40.00 | 47.00 | 84.00 | .47311 | .00 |
| 13 | | 1.00 | 95.00 | 19.00 | 33.00 | 2.00 | 53.00 | .61877 | 1.00 |
| 14 | | .00 | 83.00 | 47.00 | 75.00 | 68.00 | 9.00 | .08545 | .00 |
| 15 | | .00 | 78.00 | 57.00 | 86.00 | 19.00 | 88.00 | .30121 | .00 |
| 16 | | 1.00 | 94.00 | 1.00 | 50.00 | 44.00 | 88.00 | .85355 | 1.00 |
| 17 | | .00 | 7.00 | 54.00 | 15.00 | 23.00 | 67.00 | .45430 | .00 |
| 18 | | .00 | 77.00 | 45.00 | 34.00 | 32.00 | 75.00 | .42803 | .00 |
| 19 | | .00 | 30.00 | 37.00 | 81.00 | 92.00 | 3.00 | .14077 | .00 |

**The values in last column are the predicted groups with corresponding calculated probabilities in the second last column. Note that if the predicted probability is less than 0.5(the default threshold value in SPSS), the predicted group is type I skull (coding 0 in last column). We have classified 12 skulls as type I and 7 as type II.**

**For each skull we have obtained the probability of it being Type I or Type II, and also predicted the skull Type using the Logistic Classifier we have built.**

**For prediction-f (z) =ez/1+ez=π̂IWhere z=1.149-0.008X1-0.047X2-0.007X3-0.006X4+0.022X5**

**We have (171, 134, 130, 69,and 130)**

**On calculating π̂I we get (0.0068 which is less than 0.5) so the class is 0, that is type I skull.**

* 1. **Sentiment Analysis using Logistic Regression – What makes a US Presidential Candidate Win?**

What we are interested here in knowing that depending upon what and how a politician give speeches, his/her chances of winning the elections are affected. The idea here is similar to the email spam detection. The speech and more explicitly the content of the speech and it is delivery will have the information about the fact that the audience is convinced enough to vote for or against him/her. Sentiment Analysis is a discipline in itself; we are trying to understand the basics of to solve a particular problem. Commonly if politician is polite but passionate enough to serve the people, talks about development, remain optimist in his speech, talks about facts and figures related to government policies to explain his point to the audience is expected to win and vice-versa. But we want to examine the data.

The first aspect of the problem is to understand the data itself. I hope there is no confusion that we are going to use past data meaning by past win/loss statistics and the corresponding speeches. Clearly the response variable will indicate the win/loss information.

But what will be my independent variables? The independent variables will be the characteristics of the speech which may affect the win/loss which are commonly the following:

1. Proportion of words in the speech showing *Optimism*
2. Proportion of words in the speech showing *Pessimism*
3. Proportion of words in the speech showing the use of *Past*
4. Proportion of words in the speech showing the use of *Present*
5. Proportion of words in the speech showing the use of *Future*
6. Number of time he/she mentions his/her own party
7. Number of time he/she mentions his/her opposite parties

There are some more independent variables possible for which we need to understand the concept of big five personality traits which represent the personality traits of human which are the following:

1. Openness: *Curious, original, intellectual, creative and open to new ideas*.
2. Conscientiousness: *Organized, systematic, punctual, achievement oriented and dependable*.
3. Extraversion: *Outgoing, talkative, social and enjoys being in social situations*.
4. Agreeableness: *Affable, tolerant, sensitive, trusting, kind and warm*.
5. Neuroticism: *Anxious, irritable, temperamental and moody*.

Other than these big five personality traits the emotional content of the speech may also affect the win/loss. Thus we consider the following more independent variables.

1. Some measure indicating the content of speech showing *Openness*
2. Some measure indicating the content of speech showing *Conscientiousness*
3. Some measure indicating the content ofspeech showing *Extraversion*
4. Some measure indicating the content of speech showing *Agreeableness*
5. Some measure indicating the content of speech showing *Neuroticism*
6. Some measure indicating the content of speech showing *emotionality*

Once we get this data, task is all with the statistical analyst to make an efficient model with good predictive power.

**Case 5:** Consider the **US Presidential Data.xlsx**and perform the following objectives:

Build a logistic regression model for classifying win/loss using the given independent variables.

Test for the Significance individual independent variables.

1. Build a logistic regression model for classifying a human skull as Type I/Type II using the given independent variables.
2. Test for the Significance individual independent variables.
3. Test for the overall Logistic Regression using Hosmer and Lemeshow Test (It’s a Chi-Square Test).
4. Construct the Confusion (Classification) Table and report the percentage of correct classification in the given speeches. Also calculate specificity and sensitivity of the model.
5. For each speech obtain the probability of winning, and hence predict the win/loss status using the Logistic Classifier you have built.

On running binary logistic regression for the given data in SPSS we get:

| ***Table1.12-Hosmer and LemeshowTest*** | | | |
| --- | --- | --- | --- |
| Step | Chi-square | df | Sig. |
| 1 | 43.905 | 8 | .000 |

The important part of this test is the test statistic itself (43.905) and the significance value (0.000).A significant chi square (as p value is 0.000<0.05) indicates that the data does not fit the model well. Asignificant value is indicative of a model that is poorly predicting the real world data.

| ***Table 1.13-Contingency Table for Hosmer and Lemeshow Test*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Win Loss = 0 | | Win Loss = 1 | | Total |
|  |  | Observed | Expected | Observed | Expected |
| Step 1 | 1 | 127 | 122.009 | 25 | 29.991 | 152 |
| 2 | 113 | 102.257 | 39 | 49.743 | 152 |
| 3 | 91 | 87.973 | 61 | 64.027 | 152 |
| 4 | 57 | 74.678 | 95 | 77.322 | 152 |
| 5 | 62 | 62.229 | 90 | 89.771 | 152 |
| 6 | 55 | 49.828 | 97 | 102.172 | 152 |
| 7 | 25 | 39.285 | 127 | 112.715 | 152 |
| 8 | 28 | 29.436 | 124 | 122.564 | 152 |
| 9 | 16 | 19.326 | 136 | 132.674 | 152 |
| 10 | 21 | 7.978 | 135 | 148.022 | 156 |

The 1 st column categorizes, in order of increasing magnitude, the probabilities assigned by the regression model into divisions known as deciles. The table shows the association between assigned probability and win/loss status. In general there is close agreement between the expected frequencies (the assignments by the regression model and category assignment on the basis of the cutoff point of 0.05 for probability) and the observed or actual frequencies of subjects in those categories.

| ***Table 1.14-Classification Tablea*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Observed | | Predicted | | |
|  | Win Loss | | Percentage Correct |
|  | 0 | 1 |
| Step 1 | Win Loss | 0 | 354 | 241 | 59.5 |
| 1 | 157 | 772 | 83.1 |
| Overall Percentage | |  |  | 73.9 |
| a. The cut value is .500 | | | | | |

The classification table indicates how well the model predicts group membership. Before we use this information to classify subjects, we need to have a decision rule. **Our decision rule will take the following form: If the probability of the event is greater than or equal to some threshold, we shall predict that the event will take place. By default SPSS sets this threshold to 0.5** .The classification table shows the proportion of correct assignments when the regression model has been applied to the data. Overall our predictions were correct 1126 out of 1524 times for overall successes rate of 73.9%.From the classification table we calculated the percentage of misclassification as 26.1154%.

Sensitivity-P (correct |event did occur), that is, the percentage of occurrences correctly predicted is 772/929=***83.1%.***The decision rule allows us to correctly classify 354/595=***59.5%*** of the subjects where the predicted event was not observed. This is known as the specificity of prediction, the P (correct |event did not occur), that is, the percentage of non-occurrences correctly predicted.

| ***Table 1.15-Variables in the Equation*** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 1a | X1 | -3.567 | 2.090 | 2.912 | 1 | .088 | .028 |
| X2 | -28.451 | 2.951 | 92.952 | 1 | .000 | .000 |
| X3 | 2.080 | .763 | 7.439 | 1 | .006 | 8.002 |
| X4 | 4.138 | .725 | 32.559 | 1 | .000 | 62.668 |
| X6 | .008 | .006 | 1.756 | 1 | .185 | 1.008 |
| X7 | .016 | .013 | 1.466 | 1 | .226 | 1.016 |
| X8 | 311.989 | 53.759 | 33.680 | 1 | .000 | 3.128E135 |
| X9 | -.377 | .082 | 20.967 | 1 | .000 | .686 |
| X10 | .212 | .130 | 2.669 | 1 | .102 | 1.237 |
| X11 | -.583 | .188 | 9.649 | 1 | .002 | .558 |
| X12 | -.481 | .118 | 16.770 | 1 | .000 | .618 |
| X13 | .828 | .107 | 60.040 | 1 | .000 | 2.288 |
| Constant | .936 | .768 | 1.483 | 1 | .223 | 2.549 |
| a. Variable(s) entered on step 1: X1, X2, X3, X4, X6, X7, X8, X9, X10, X11, X12, and X13. | | | | | | | |

**Logistic regression model is:**

**ln(ODDS) =0.936-3.567X1-28.451X2+2.080X3+4.138X4+0.008X6+0.016X7+311.989X8-0.377X9+0.212X10-.583X11-0.481X12+0.828X13**

From the above table we observe that as p value is greater than 0.05,for X1,X6,X7,X10 and for the constant we accept our null hypothesis and conclude that these predictors do not make a significant contribution to the regression model. The values of Exp (B) for X1 indicates that if the percentage of measure1 (i.e. Optimism) goes up by one, then the odds of losingdecrease (because exp b is less than one).Similarly for all the measures for which Exp (B) is less than one. But if the percentage of measure 10(ie.emotionality) goes up by one, then the odds of losingincrease.

| ***Table 1.16-Case Summariesa*** | | | | |
| --- | --- | --- | --- | --- |
|  |  | Win Loss | Predicted probability | Predicted group |
| 1 | 1 | | .72361 | 1 |
| 2 | 1 | | .75652 | 1 |
| 3 | 1 | | .81174 | 1 |
| 4 | 1 | | .73705 | 1 |
| 5 | 1 | | .79277 | 1 |
| 6 | 1 | | .84662 | 1 |
| 7 | 1 | | .74538 | 1 |
| 8 | 1 | | .66512 | 1 |
| 9 | 1 | | .77181 | 1 |
| 10 | 1 | | .67257 | 1 |
| 11 | 1 | | .72340 | 1 |
| 12 | 1 | | .69840 | 1 |
| 13 | 1 | | .67125 | 1 |
| 14 | 1 | | .55977 | 1 |
| 15 | 1 | | .76468 | 1 |
| 16 | 1 | | .60683 | 1 |
| 17 | 1 | | .66167 | 1 |
| 18 | 1 | | .55822 | 1 |
| 19 | 1 | | .61113 | 1 |
| 20 | 1 | | .88805 | 1 |
| 21 | 1 | | .89074 | 1 |
| 22 | 1 | | .87926 | 1 |
| 23 | 1 | | .90907 | 1 |
| 24 | 1 | | .84911 | 1 |
| 25 | 1 | | .67924 | 1 |
| 26 | 1 | | .82010 | 1 |
| 27 | 1 | | .80423 | 1 |
| 28 | 1 | | .90430 | 1 |
| 29 | 1 | | .84226 | 1 |
| 30 | 1 | | .86952 | 1 |
| 31 | 1 | | .87081 | 1 |
| 32 | 1 | | .90870 | 1 |
| 33 | 1 | | .89486 | 1 |
| 34 | 1 | | .90952 | 1 |
| 35 | 1 | | .87786 | 1 |
| 36 | 1 | | .83934 | 1 |
| 37 | 1 | | .86685 | 1 |
| 38 | 1 | | .86518 | 1 |
| 39 | 1 | | .88238 | 1 |
| 40 | 1 | | .92768 | 1 |
| 41 | 1 | | .80554 | 1 |
| 42 | 1 | | .77595 | 1 |
| 43 | 1 | | .71316 | 1 |
| 44 | 1 | | .89093 | 1 |
| 45 | 1 | | .83370 | 1 |
| 46 | 1 | | .73309 | 1 |
| 47 | 1 | | .79226 | 1 |
| 48 | 1 | | .86210 | 1 |
| 49 | 1 | | .70734 | 1 |
| 50 | 1 | | .69685 | 1 |
| 51 | 1 | | .79396 | 1 |
| 52 | 1 | | .72540 | 1 |
| 53 | 1 | | .68232 | 1 |
| 54 | 1 | | .72140 | 1 |
| 55 | 1 | | .59070 | 1 |
| 56 | 1 | | .71287 | 1 |
| 57 | 1 | | .74980 | 1 |
| 58 | 1 | | .84728 | 1 |
| 59 | 1 | | .64936 | 1 |
| 60 | 1 | | .76343 | 1 |
| 61 | 1 | | .72045 | 1 |
| 62 | 1 | | .61689 | 1 |
| 63 | 1 | | .74863 | 1 |
| 64 | 1 | | .78652 | 1 |
| 65 | 1 | | .80460 | 1 |
| 66 | 1 | | .82958 | 1 |
| 67 | 1 | | .76584 | 1 |
| 68 | 1 | | .69951 | 1 |
| 69 | 1 | | .80702 | 1 |
| 70 | 1 | | .85557 | 1 |
| 71 | 1 | | .82967 | 1 |
| 72 | 1 | | .87596 | 1 |
| 73 | 1 | | .90388 | 1 |
| 74 | 1 | | .85383 | 1 |
| 75 | 1 | | .88329 | 1 |
| 76 | 1 | | .90451 | 1 |
| 77 | 1 | | .82226 | 1 |
| 78 | 1 | | .87158 | 1 |
| 79 | 1 | | .89157 | 1 |
| 80 | 1 | | .92757 | 1 |
| 81 | 1 | | .84141 | 1 |
| 82 | 1 | | .84559 | 1 |
| 83 | 1 | | .76183 | 1 |
| 84 | 1 | | .80375 | 1 |
| 85 | 1 | | .81535 | 1 |
| 86 | 1 | | .81694 | 1 |
| 87 | 1 | | .60807 | 1 |
| 88 | 1 | | .70535 | 1 |
| 89 | 1 | | .73985 | 1 |
| 90 | 1 | | .74851 | 1 |
| 91 | 1 | | .76387 | 1 |
| 92 | 1 | | .76992 | 1 |
| 93 | 1 | | .96263 | 1 |
| 94 | 1 | | .78835 | 1 |
| 95 | 1 | | .77144 | 1 |
| 96 | 1 | | .81636 | 1 |
| 97 | 1 | | .74372 | 1 |
| 98 | 1 | | .80556 | 1 |
| 99 | 1 | | .85364 | 1 |
| 100 | 1 | | .77558 | 1 |
| 101 | 1 | | .73536 | 1 |
| 102 | 1 | | .79196 | 1 |
| 103 | 1 | | .72696 | 1 |
| 104 | 1 | | .76558 | 1 |
| 105 | 0 | | .38450 | 0 |
| 106 | 0 | | .76703 | 1 |
| 107 | 0 | | .66615 | 1 |
| 108 | 0 | | .26678 | 0 |
| 109 | 0 | | .34893 | 0 |
| 110 | 0 | | .24597 | 0 |
| 111 | 0 | | .55249 | 1 |
| 112 | 0 | | .38595 | 0 |
| 113 | 0 | | .50878 | 1 |
| 114 | 0 | | .00437 | 0 |
| 115 | 0 | | .49874 | 0 |
| 116 | 0 | | .80182 | 1 |
| 117 | 0 | | .58909 | 1 |
| 118 | 0 | | .53095 | 1 |
| 119 | 0 | | .23832 | 0 |
| 120 | 0 | | .47547 | 0 |
| 121 | 0 | | .41860 | 0 |
| 122 | 0 | | .69545 | 1 |
| 123 | 0 | | .29519 | 0 |
| 124 | 0 | | .27535 | 0 |
| 125 | 0 | | .77341 | 1 |
| 126 | 0 | | .29322 | 0 |
| 127 | 0 | | .52065 | 1 |
| 128 | 0 | | .95873 | 1 |
| 129 | 0 | | .62926 | 1 |
| 130 | 0 | | .86413 | 1 |
| 131 | 0 | | .83664 | 1 |
| 132 | 0 | | .68357 | 1 |
| 133 | 0 | | .40935 | 0 |
| 134 | 0 | | .36645 | 0 |
| 135 | 0 | | .61266 | 1 |
| 136 | 0 | | .18117 | 0 |
| 137 | 0 | | .63389 | 1 |
| 138 | 0 | | .35139 | 0 |
| 139 | 0 | | .31305 | 0 |
| 140 | 0 | | .59432 | 1 |
| 141 | 0 | | .85404 | 1 |
| 142 | 0 | | .36742 | 0 |
| 143 | 0 | | .35421 | 0 |
| 144 | 0 | | .37921 | 0 |
| 145 | 0 | | .30347 | 0 |
| 146 | 0 | | .38333 | 0 |
| 147 | 0 | | .67858 | 1 |
| 148 | 0 | | .26365 | 0 |
| 149 | 0 | | .44005 | 0 |
| 150 | 0 | | .54865 | 1 |

We have taken only an extract of the data as we have total around 1500 observations. **The values in last column are the predicted groups with corresponding calculated probabilities in the second last column. Note that if the predicted probability is less than 0.5(the default threshold value in SPSS), the predicted group is 0 (Status –Loss) and if the predicted group is 1(Status-Win).** For each speech we have obtained the probability of winning, and also predicted the win/loss status using the Logistic Classifier we have built.

**Logistic regression** is used to predict a categorical (usually dichotomous) variable from a set of predictor variables. With a categorical dependent variable, discriminant function analysis is usually employed if all of the predictors are continuous and nicely distributed; logit analysis is usually employed if all of the predictors are categorical; and logistic regression is often chosen if the predictor variables are a mix of continuous and categorical variables and/or if they are not nicely distributed (logistic regression makes no assumptions about the distributions of the predictor variables)

**Discriminant Analysis**

Discriminant Analysis is a classification technique to classify an observation (univariate or multivariate) into one of several possible classes by means of “some” optimal way to separate different populations. Some real life examples of classification problem are loan classification – high risk, medium risk and low risk, warning systems for financial crisis, medical diagnostics – critical and non-critical patients.

Suppose we have as multivariate observations and let be the measurement space of all the multivariate observations. Further, suppose that each of the observations fall into one (exactly one) of the classes denoted as : Set of classes.

In discriminant analysis the basic aim is develop a systematic way of predicting the class membership of a multivariate observation by means of some optimal classification rule.

**Definition:** A classifier or a classification rule is a function defined on such that for every, is equal to one of the numbers.

Alternate way to look at the classifier is that it induces a partition of the entire measurement space such that

Let us concentrate on the binary classification problem. Suppose there are two populations and and an arbitrary multivariate observation comes from either of the two.

**Aim:** Let and be observations from and, the aim is to find some function say such that and look as different as possible then is the desired discriminant function to discriminate between and. The aim is to find some “optimal” discriminant function. One such optimal rule is given by fisher linear discriminant function.

* + 1. **Fisher Linear Discriminant Analysis**

Under the same setup assume that, where is the mean vector of the 1st population and is the covariance matrix for the 1st population, and where is the mean vector of the 2nd population and is the common covariance matrix for both the populations.

Further change and into two univariate populations by changing to some by means of “some”.

**Discrimination:**

We are interested in finding or choosing such that the separation between the two univariate populations is maximum, i.e. maximization of statistical distance between and with respect to.

A measure of statistical distance between the two populations is given by,

We want to obtain,

Assuming that is positive definite and defining. We have,

Using Cauchy Schwartz Inequality,

Hence we have the distance between two populations is always less than or equal to the Mahalanobis Distance. Is maximum when,

Thus we have the optimal which provides the maximum separation (discrimination) between the two populations.

The quantity is called the **Fisher Linear Discriminant Function(LDF),** which is an optimal separation between the two populations.

Now then next question of interest is that based on Fisher LDF how do we classify an arbitrary observation into one of the two possible classes?

**Classification:**

Realize that,

Note that

.

As a rule of classification for any new observation calculate, and assign to if is closer to than otherwise to which is an intuitive logical rule as we are trying to see if the FLDF is close to expectation of FLDF under or . Thus we have,

We can finally write that assign to if

And assign to otherwise.

Usually for practical problems the values of and are unknown, which we replace by their estimates. For estimation we need samples from both the populations, i.e. for some data points we need to have the classes being already assigned, i.e. we have data of the following form:

Where 1 or 2 representing the class or the population.

And is calculated as the pooled sample variance of the observations from both the samples. So finally the classifier in its executable form can be written as:

If then class is otherwise.

The FLDF of, i.e. with population characteristics being replaced by their sample counterparts is called as the sample FLDA.

**Assumptions of FLDA:**

The assumptions of discriminant analysis are the same as those for MANOVA. The analysis is quite sensitive to outliers and the size of the smallest group must be larger than the number of predictor variables.

[**Multivariate normality**](http://en.wikipedia.org/wiki/Multivariate_normal_distribution)**:** Independent variables are normal for each level of the grouping variable.

**Homogeneity of variance/covariance (**[**homoscedasticity**](http://en.wikipedia.org/wiki/Homoscedasticity)**):** Variances among group variables are the same across levels of predictors. This can be tested with Box's M statistic.

[**Multicollinearity**](http://en.wikipedia.org/wiki/Multicollinearity)**:** Predictive power can decrease with an increased correlation between predictor variables.

[**Independence**](http://en.wikipedia.org/wiki/Statistical_independence)**:** Participants are assumed to be randomly sampled, and a participant’s score on one variable is assumed to be independent of scores on that variable for all other participants.

It has been suggested that discriminant analysis is relatively robust to slight violations of these assumptions, and it has also been shown that discriminant analysis may still be reliable when using dichotomous variables (where multivariate normality is often violated)

**Some Discussion on the Tests and Routines to be used:**

1. **Wilk’s Lambda**

Wilk’s Lambda is a test for equality of group means and is used to test which independent variable contributes significantly to the discriminant function, i.e. which variable contributes significantly while discriminating between the 2 groups 0 & 1.

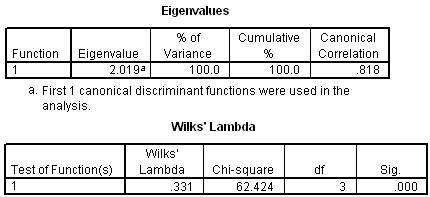
*Ho:* All the group means are statistically not significantly different between the 2 groups for a particular measure.

*H1:* All the group means are statistically significantly different between the 2 groups for a particular measure.

The value of Wilk’s Lambda varies between 0 and 1. The smaller its value, the more the corresponding variable contributes to the discriminant function.

**Overall significance of the discriminant function:**

Similar to multiple regression analysis, our first task is to determine whether or not there is a statistically significant relationship between the independent variables and the dependent variable. We navigate to the section of output titled "Summary of Canonical Discriminant Functions" to locate the following outputs:



The key statistic indicating whether or not there is a relationship between the independent and dependent variables is the significance test for Wilks' lambda. Wilks' lambda is the proportion of the total variance in the discriminant scores NOT explained by differences among the groups. In this example about 33% of the variance is not explained by group differences. Unlike R2, smaller values of Wilks' lambda are desirable. You can also note that Wilks’ Lambda (overall) is nothing but 1 – square of canonical correlation.

Wilks’ lambda is used to test the null hypothesis that the means of all of the independent variables are equal across groups of the dependent variable. If the means of the independent variables are equal for all groups, the means will not be a useful basis for predicting the group to which a case belongs, and thus there is no relationship between the independent variables and the dependent variable. If the chi-square statistic corresponding to Wilks' lambda is statistically significant we conclude that there is a relationship between the dependent groups and the independent variables. We should note that there is no correspondence between the size of Wilks’ lambda and the accuracy of the classifications based on the discriminant functions.

The information from the table of Eigenvalues is often cited in analyses using discriminant analysis, but it is not as important to us as the statistical test of Wilks’ lambda. The table of eigenvalues gives us information about the effectiveness of the discriminant functions. The eigenvalue is a ratio of the between-groups sum of squares to the within-groups or error sum of squares. The size of the eigenvalue is helpful for measuring the spread of the group centroids in the corresponding dimension of the multivariate discriminant space. Larger eigenvalues indicate that the discriminant function is more useful in distinguishing between the groups. The eigenvalues will always be listed in descending order since the solution in a discriminant analysis requires that the first discriminant function is the most capable in differentiating the groups; the second discriminant function is the second most useful function, etc.

The canonical correlation coefficient (.818) measures the association between the discriminant score and the set of independent variables. Like Wilks' lambda, it is an indicator of the strength of relationship between entities in the solution, but it does not have any necessary relationship to the classification accuracy which is our ultimate measure of the value of the model.

**Box’s M test:**

Box’s M test tests for the homogeneity of variance-covariance matrices of the 2 groups.

Covariance matrices of the 2 groups do not differ significantly.

Covariance matrices of the 2 groups differ significantly.

This is a very powerful test, so when the sample is large then even small differences are considered significant. Thus, in order to be lenient, we check for values of Log Determinants. If these values for the 2 groups are fairly close, we accept Ho and conclude that the 2 covariance matrices are not significantly different.

1. **Summary of Canonical Discriminant Functions**

**Eigenvalue and Canonical correlation:**

Eigenvalue represents the ratio of the between-group sum of squares to the within-group sum of squares of the discriminant score. It indicates the *relative discriminating power of the discriminant function*, i.e. how well the discriminating function discriminates between the 2 groups.

Canonical Correlation of discriminant function is the correlation of that function with discriminant scores. Since there are only 2 groups so only one discriminant function is generated which accounts for 100% of the explained variance. If Canonical Correlation is close to 1, it implies nearly all the variation in the discriminant scores can be attributed to the group differences. *Squared canonical correlation is the percentage of variation in the dependent, discriminated by the set of independents in discriminant analysis*.

**3 a Standardized Canonical Discriminant Function Coefficients:**

**Standardized Canonical Discriminant Function Coefficients** allows us to see the extent to which each of the predictors contribute to the ability of the discriminant function. It rescales the variables to unit standard deviation. If a coefficient lies in the neighborhood of 1 or -1, then it is a good explanator & if it lies in the neighborhood of 0 or 0.5, then it gives a moderate explanation. SPSS generates Unstandardized Canonical Discriminant Function Coefficients as well which gives the same information but as they are not standardized the same rule of interpretation is not applicable.

**Functions at Group Centroids:**

These are the estimated expected values of the FLDF for different groups. In machine learning terminology centroid is nothing but the mean. So, these are our and only.

**Classification Statistics:**

**Classification Function Coefficients:**

We need to calculate the optimum. Hence the difference between the values of classification function coefficients for two different groups will give us the entries of which we can use to multiply with the values of a new observation to calculate sample FLDF and then compare it with to classify it into one of the two groups.

**Classification Results:**

SPSS generates two classification tables which are as follows:

1. Original: which tells the fitting strength of the LDF or how well it performs for the given data.
2. Cross-Validated: which tells the predictive strength of the LDF or how well it performs for the new data.

**Skull Type Prediction using Discriminant Analysis**

We are interested in predicting the type of skull of humans as one of two possible types I and II based on some five physical measures available related to the skulls.

**Case 6:** Consider the dataset **Skull Type Prediction.xlsx** and perform the following objectives.

* 1. Test for normality of all five physical measures.
  2. Test for equality of the covariance matrices of different groups using Box’s M Test.
  3. Test if the group means are statistically significantly different between the 2 groups for all the physical measures and point out which measures contribute to the discriminant function significantly.
  4. Test for overall significance of the discriminant function.
  5. Obtain the standardized and unstandardized canonical discriminant function coefficients.
  6. Obtain the *values of discriminant function at the group centroids* and *classification function coefficients.*
  7. Obtain the entries of the vector, calculate the discriminant functions for the given skulls and hence obtain the predicted skull type for the given skulls.
  8. Construct the classification tables for fitting strength and predictive strength of the model. Also calculate sensitivity and specificity for both fitting strength and predictive strengths.
  9. For a set of five physical measures given for a new skull in the dataset **Skull Type Prediction – Validation Data.xlsx** predict the skull type using the Logistic Classifier you have built.

First of all we check for normality of all 5 measures. **We set the following hypothesis:**

**Ho: The observed data is normally distributed.**

**H1.: The observeddata is not normally distributed.**

**We are considering p value of KS test.**

| ***Table 6.1 Tests of Normality*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
|  | Statistic | df | Sig. | Statistic | df | Sig. |
| X1 | .174 | 19 | .133 | .909 | 19 | .071 |
| a. Lilliefors Significance Correction | | | | | | |

As p value>0.05 we may accept null hypothesis and conclude that observed data is normally distributed.

**Figure 6.1**

****

The data appears to be normal.

| ***Table 6.2 Tests of Normality*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
|  | Statistic | Df | Sig. | Statistic | df | Sig. |
| X2 | .112 | 19 | .200\* | .966 | 19 | .702 |
| a. Lilliefors Significance Correction | | | | | | |
| \*. This is a lower bound of the true significance. | | | | | | |

As p value>0.05 we may accept null hypothesis and conclude that observed data is normally distributed.

**Fig 6.2**



The data appears to be normal.

***Table 6.3 Test of normality***

|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Statistic | Df | Sig. | Statistic | df | Sig. |
| X3 | .131 | 19 | .200\* | .947 | 19 | .344 |
| a. Lilliefors Significance Correction | | | | | | | |
| \*. This is a lower bound of the true significance. | | | | | | | |

As p value>0.05 we may accept null hypothesis and conclude thatobserved data is normally distributed.

**Fig 6.3**



The data appears to be normal.

| ***Table 6.4 Tests of Normality*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
|  | Statistic | df | Sig. | Statistic | df | Sig. |
| X4 | .127 | 19 | .200\* | .978 | 19 | .913 |
| a. Lilliefors Significance Correction | | | | | | |
| \*. This is a lower bound of the true significance. | | | | | | |

As p value>0.05 we may accept null hypothesis and conclude that observed data is normally distributed.

**Fig 6.4**



The data appears to be normal.

***Table 6.5 Tests of normality***

|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
| --- | --- | --- | --- | --- | --- | --- |
|  | Statistic | df | Sig. | Statistic | df | Sig. |
| X5 | .191 | 19 | .066 | .890 | 19 | .032 |
| a. Lilliefors Significance Correction | | | | | | |

As p value>0.05 we may accept null hypothesis and conclude thatobserved data is normally distributed.

**Fig 6.5**



The data appears to be normal.

| ***Table 6.6 Group Statistics*** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Skull type | | Mean | Std. Deviation | Valid N (listwise) | |
| Unweighted | Weighted |
| .00 | X1 | 50.0909 | 34.42224 | 11 | 11.000 |
| X2 | 45.0000 | 21.35884 | 11 | 11.000 |
| X3 | 47.4545 | 34.91236 | 11 | 11.000 |
| X4 | 49.4545 | 24.45961 | 11 | 11.000 |
| X5 | 45.9091 | 36.01792 | 11 | 11.000 |
| 1.00 | X1 | 58.1250 | 36.17591 | 8 | 8.000 |
| X2 | 29.6250 | 17.31999 | 8 | 8.000 |
| X3 | 48.5000 | 24.77902 | 8 | 8.000 |
| X4 | 44.6250 | 23.94003 | 8 | 8.000 |
| X5 | 67.6250 | 25.15630 | 8 | 8.000 |
| Total | X1 | 53.4737 | 34.40667 | 19 | 19.000 |
| X2 | 38.5263 | 20.75885 | 19 | 19.000 |
| X3 | 47.8947 | 30.26897 | 19 | 19.000 |
| X4 | 47.4211 | 23.69087 | 19 | 19.000 |
| X5 | 55.0526 | 32.98733 | 19 | 19.000 |

This table gives means and standard deviations of each of the 5 measures for both the groups.

| ***Table 6.7 Tests of Equality of Group Means*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Wilks' Lambda | F | df1 | df2 | Sig. |
| X1 | .986 | .242 | 1 | 17 | **.629** |
| X2 | .859 | 2.794 | 1 | 17 | **.113** |
| X3 | 1.000 | .005 | 1 | 17 | **.943** |
| X4 | .989 | .184 | 1 | 17 | **.674** |
| X5 | .888 | 2.134 | 1 | 17 | **.162** |

We test the following hypothesis:

Ho: measure Xi does not contribute significantly to discriminant function

H1: measure Xi does contribute significantly to discriminant function.

As significance levels for all measures >0.05 we may accept null hypothesis that is measure Xi does not contribute significantly to discriminant function.

In the ANOVA table, the smaller the Wilks’s lambda, the more important the independent variable to the discriminant function.

|  | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Table 6.8 Covariance Matricesa*** | | | | | | | |
| Skull type | | X1 | X2 | X3 | X4 | X5 | |
| .00 | X1 | 1184.891 | -164.600 | -283.045 | -1.445 | 59.809 | |
| X2 | -164.600 | 456.200 | 14.000 | -266.600 | -31.700 | |
| X3 | -283.045 | 14.000 | 1218.873 | 313.873 | 47.545 | |
| X4 | -1.445 | -266.600 | 313.873 | 598.273 | -413.955 | |
| X5 | 59.809 | -31.700 | 47.545 | -413.955 | 1297.291 | |
| 1.00 | X1 | 1308.696 | -436.518 | -9.643 | -167.232 | 324.768 | |
| X2 | -436.518 | 299.982 | -174.500 | 179.839 | -100.304 | |
| X3 | -9.643 | -174.500 | 614.000 | -161.500 | 299.786 | |
| X4 | -167.232 | 179.839 | -161.500 | 573.125 | 89.554 | |
| X5 | 324.768 | -100.304 | 299.786 | 89.554 | 632.839 | |
| Total | X1 | 1183.819 | -292.985 | -158.836 | -75.822 | 204.418 | |
| X2 | -292.985 | 430.930 | -64.219 | -59.067 | -142.529 | |
| X3 | -158.836 | -64.219 | 916.211 | 110.269 | 148.839 | |
| X4 | -75.822 | -59.067 | 110.269 | 561.257 | -222.135 | |
| X5 | 204.418 | -142.529 | 148.839 | -222.135 | 1088.164 | |
| a. The total covariance matrix has 18 degrees of freedom. | | | | | | | |

This table displays variances on main diagonal and other elements are the covariance.

**Analysis 1**

**Box's Test of Equality of Covariance Matrices**

| ***Table 6.9Log Determinants*** | | |
| --- | --- | --- |
|  | Rank | Log Determinant |
| .00 | 5 | 32.677 |
| 1.00 | 5 | 29.947 |
| Pooled within-groups | 5 | 32.825 |
| The ranks and natural logarithms of determinants printed are those of the group covariance matrices. | | |

Before performing Fishers Discriminant analysis we use Box s M test to test for equality of population covariance matrices.

The log determinants values provide an indication of which groups covariance matrices differ most.

The larger the log determinant in the table, the more that group's covariance matrix differs. The "Rank" column indicates the number of independent variables in this case. Since discriminant analysis assumes homogeneity of covariance matrices between groups, we would like to have the determinants relatively equal.

| ***Table 6.10 Test Results*** | | |
| --- | --- | --- |
| Box's M | | 21.637 |
| F | Approx. | .947 |
| df1 | 15 |
| df2 | 908.297 |
| Sig. | **.511** |
| Tests null hypothesis of equal population covariance matrices. | | |
|  | | |

The hypothesis is set as follows:

Ho: ∑=∑12=∑21 the population covariance matrices are equal.

H1: ∑≠∑12≠∑21 the population covariance matrices are not equal.

This is a very powerful test, so when the sample is large then even small differences are considered significant. Thus, in order to be lenient, we check for values of Log Determinants. If these values for the 2 groups are fairly close, we accept Ho and conclude that the 2 covariance matrices are not significantly different and we can proceed with our analysis.

**Summary of Canonical Discriminant Functions**

| **Table 6.11Eigenvalues** | | | | |
| --- | --- | --- | --- | --- |
| Function | Eigenvalue | % of Variance | Cumulative % | Canonical Correlation |
| 1 | **.289a** | 100.0 | 100.0 | .473 |
| a. First 1 canonical discriminant functions were used in the analysis. | | | | |

An eigenvalue indicates the proportion of variance explained. (Between-groups sums of squares divided by within-groups sums of squares). A large eigenvalue is associated with a strong function. The canonical relation is a correlation between the discriminant scores and the levels of the dependent variable. A high correlation indicates a function that discriminates well. The present correlation of 0.473 is not extremely high (1.00 is perfect). For two group discriminant analysis there is one Eigen value. The eigenvalues assess relative importance because they reflect the percentage of variance explained in the dependent variable by the independent variable.

As our eigenvalue is 0.289 which gives the relative discriminating power of discriminant function, it is only 28.9% useful in distinguishing between the groups and here there is only one discriminating function which accounts for 100% of the explained variance.

| ***Table 6.12 Wilks' Lambda*** | | | | |
| --- | --- | --- | --- | --- |
| Test of Function(s) | Wilks' Lambda | Chi-square | df | Sig. |
| 1 | .776 | 3.678 | 5 | .597 |

We set the hypothesis as:

Ho: The 2 groups have same mean discriminant function scores or there is no relationship between independent and dependent variable.

H1: The 2 groups do not have same mean discriminant function scores or there is significant relationship between independent and dependent variable.

The key statistic indicating whether or not there is a relationship between the independent and dependent variables is the significance test for Wilks' lambda. Wilks' lambda is a measure of how well each function separates cases into groups. Smaller values of Wilks' lambda indicate greater discriminatory ability of the function. The associated chi-square statistic tests the hypothesis that the means of the functions listed are equal across groups. The small significance value indicates that the discriminant function does better than chance at separating the groups.

Wilks’ Lambda is the ratio of within-groups sums of squares to the total sums of squares. This is the proportion of the total variance in the discriminant scores not explained by differences among groups. A lambda of 1.00 occurs when observed group means are equal (all the variance is explained by factors other than difference between those means), while a small lambda occurs when within-groups variability is small compared to the total variability. A small lambda indicates that group means appear to differ. The associated significance value indicate whether the difference is significant. Here, the Lambda of 0.776 has a significant value (Sig. = 0.597); thus, the group means do not differ.

| ***Table 6.13 Standardized Canonical Discriminant Function Coefficients*** | |
| --- | --- |
|  | Function |
|  | 1 |
| X1 | .239 |
| X2 | .840 |
| X3 | .154 |
| X4 | .176 |
| X5 | -.592 |

The standardized discriminant function coefficients in the table serve the same purpose as beta weights in multiple regression (partial coefficient), they indicate the relative importance of the independent variables in predicting the dependent. They allow you to compare variables measured on different scales. Coefficients with large absolute values correspond to variables with greater discriminating ability. Standardizing the coefficients allows us to measure the relative standing of the measurements. It also indicates how heavily each variable is weighted in order to maximize discrimination of groups. X2 and X5 are weighted more than the rest.

This table allows us to see the extent to which each of the predictor variables is contributing to the ability of the discriminant function. We see that X2 and X5 have a very good contribution to the ability of discrimination between the 2 categories since their values are closer to 1 or -1. X3 has the least contribution.

| ***Table 6.14 Structure Matrix*** | |
| --- | --- |
|  | Function |
|  | 1 |
| X2 | .755 |
| X5 | -.659 |
| X1 | -.222 |
| X4 | .193 |
| X3 | -.033 |
| Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions  Variables ordered by absolute size of correlation within function. | |

1. The structure matrix table shows the correlations of each variable with each discriminant function.

2. Only one discriminant function is in this study.

3. By identifying the largest absolute correlations associated with each discriminant function the researcher gains insight into how to name each function.

The structure matrix shows how all the variables relate to each function at the same time. The output of discriminant function analysis illustrates that all the variables in the model predict group membership up to some extent even though small. Also each variable contributes some amount to each group at the same time. The structure matrix provides a way to study the usefulness of each variable in the discriminant function. These are the correlations of each independent variable with the standardized discriminant function.X2 and X5 are more correlated with the discriminant function than all the other measures.

| ***Table 6.15 Canonical Discriminant Function Coefficients*** | |
| --- | --- |
|  | Function |
|  | 1 |
| X1 | .007 |
| X2 | .042 |
| X3 | .005 |
| X4 | .007 |
| X5 | -.019 |
| (Constant) | -1.559 |
| Unstandardized coefficients | |

The ‘Canonical Discriminant Function Coefficients’ indicate the unstandardized scores concerning the independent variables. It is the list of coefficients of the unstandardized discriminant equation. Each measure’s discriminant score would be computed by entering variable values (raw data) for each of the variables in the equation.

The discriminant function coefficients are partial coefficients, reflecting the unique contribution of each variable to the classification of the criterion variable. These coefficients are used to assess the relative classifying importance of the independent variables. The coefficients are the coefficients of canonical variable. The coefficients are used to compute canonical variable scores for each case. These unstandardized weights show the relative significance of each variable based on its own scale of measurement.

| **Table 6.16 Functions at Group Centroids** | |
| --- | --- |
| Skull type | Function |
| 1 |
| .00 | **.433** |
| 1.00 | **-.596** |
| Unstandardized canonical discriminant functions evaluated at group means | |
|  | |

‘Functions at Group Centroids’ indicates the average discriminant score for subjects in the two groups. More specifically, the discriminant score for each group when the variable means (rather than individual values for each subject) are entered into the discriminant equation.

Functions of Group Centroids are mean discriminant scores for each of the dependent variable categories for each of the discriminant functions in the discriminant analysis. We want the means to be well apart to show that the discriminant function is clearly discriminating. The closer the means the more the errors of classification. The table above gives the means of both the groups, which are well apart so the discriminant function is clearly discriminating.

**Classification Statistics**

| ***Table 6.17 Classification Function Coefficients*** | | |
| --- | --- | --- |
|  | Skull type | |
|  | .00 | 1.00 |
| X1 | .100 | .093 |
| X2 | .233 | .189 |
| X3 | .055 | .050 |
| X4 | .139 | .132 |
| X5 | .062 | .081 |
| (Constant) | -14.614 | -13.093 |
| Fisher's linear discriminant functions | | |
|  | | |

Each column contains estimates of the coefficients for a classification function for 1 group. The functions are used to assign or classify cases into groups. To obtain a classification score for each case for each group, multiply each coefficient by the value of the corresponding variable, sum the products and add to get the score. A case is predicted as being a member of the group in which the value of its classification function is largest.

**Fig 6.6 Fig 6.7**



The canonical score plot shows how the first two canonical function classify observation between groups by plotting the observation score, computed via discriminant analysis**.** The plot provides a succinct summary of the separation of the observations. The clearer the observations are grouping to, the better the discriminant model is.

| ***Table 6.18 Classification Resultsb*,c** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | Skull type | Predicted Group Membership | | Total |
|  |  | .00 | 1.00 |
| Original | Count | .00 | 7 | 4 | 11 |
| 1.00 | 2 | 6 | 8 |
| % | .00 | 63.6 | 36.4 | 100.0 |
| 1.00 | 25.0 | 75.0 | 100.0 |
| Cross-validated | Count | .00 | 5 | 6 | 11 |
| 1.00 | 5 | 3 | 8 |
| % | .00 | 45.5 | 54.5 | 100.0 |
| 1.00 | 62.5 | 37.5 | 100.0 |
| a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case. | | | | | |
| b. **68.4%** of original grouped cases correctly classified. | | | | | |
| c**. 42.1%** of cross-validated grouped cases correctly classified. | | | | | |

‘Classification Results’ is a simple summary of number and percent of subjects classified correctly and incorrectly. The ‘leave-one out classification’ is a cross-validation method, of which the results are also presented. This table measures the degree of success of the classification for this sample. The no. of cases correctly classified or misclassified are displayed. It shows how well the combination of 4 independent variables classifies or predicts the discriminant function. It also shows that 68.4% of original grouped cases are correctly classified and 42.1% of cross validated grouped cases are correctly classified

.

**SENSITIVITY**

The sensitivity of a clinical test refers to the ability of the test to correctly identify those patients with the disease.

Formula

A test with 100% sensitivity correctly identifies all patients with the disease. A test with 80% sensitivity detects 80% of patients with the disease (true positives) but 20% with the disease go undetected (false negatives). A high sensitivity is clearly important where the test is used to identify a serious but treatable disease (e.g. cervical cancer).

**SPECIFICITY**

The specificity of a clinical test refers to the ability of the test to correctly identify those patients without the disease.

Formula

Therefore, a test with 100% specificity correctly identifies all patients without the disease. A test with 80% specificity correctly reports 80% of patients without the disease as test negative (true negatives) but 20% patients without the disease are incorrectly identified as test positive (false positives).

As discussed above, a test with a high sensitivity but low specificity results in many patients who are disease free being told of the possibility that they have the disease and are then subject to further investigation.

Sensitivity= (6/6+4) =0.6

I.e. a test with 60% sensitivity detects 60% of patients with the disease (true positives) but 40% with the disease go undetected (false negatives).

Specificity=7/ (7+2) =0.77

i.e. a test with 77% specificity correctly reports 77% of patients without the disease as test negative (true negatives) but 23% patients without the disease are incorrectly identified as test positive (false positives).

| ***Table 6.19 Case Summariesa*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | Predicted Group for Analysis 1 | Discriminant Scores from Function 1 for Analysis 1 | Probabilities of Membership in Group 0 for Analysis 1 | Probabilities of Membership in Group 1 for Analysis 1 |
| 1 | | 1.00 | -.77244 | .32927 | .67073 |
| 2 | | 1.00 | -1.73441 | .15424 | .84576 |
| 3 | | .00 | 1.81815 | .87602 | .12398 |
| 4 | | 1.00 | -1.18440 | .24314 | .75686 |
| 5 | | .00 | 1.69458 | .86153 | .13847 |
| 6 | | .00 | .66052 | .68213 | .31787 |
| 7 | | .00 | .25397 | .58543 | .41457 |
| 8 | | 1.00 | -.45431 | .40516 | .59484 |
| 9 | | .00 | .48143 | .64089 | .35911 |
| 10 | | .00 | .16713 | .56358 | .43642 |
| 11 | | 1.00 | -.74671 | .33515 | .66485 |
| 12 | | 1.00 | -.24014 | .45920 | .54080 |
| 13 | | 1.00 | -.91126 | .29852 | .70148 |
| 14 | | .00 | 1.69569 | .86166 | .13834 |
| 15 | | .00 | .32209 | .60234 | .39766 |
| 16 | | 1.00 | -1.94052 | .12855 | .87145 |
| 17 | | 1.00 | -.22034 | .46427 | .53573 |
| 18 | | 1.00 | -.11529 | .49125 | .50875 |
| 19 | | .00 | 1.22625 | .79346 | .20654 |

In the above table, if the values in the 5th column (Probabilities of Membership in Group 0) are less than the threshold value (default threshold value is 0.5 in SPSS) and the values in 6th column are greater than 0.5 then the predicted group is 1(ie.type II skull).For type I skull it is vice versa. The 2 ndcolumn gives observed skull types and 3 rd column gives the predicted group after doing discriminant analysis.

**We know that the discriminant function is:**

**L=-1.559+0.007\*X1+0.042\*X2+0.005\*X3+0.007\*X4-0.019\*X5**

**On using (171, 134, 130, 69, 130) if the value of L for a particular skull is greater than -0.0815(the average of centroids equals -0.0815) the skull is assigned to type I skull (coded 0) otherwise to type II skull (coded 0).For given measurements of skull, we get value of L to be 3.929>-0.0815 hence we classify it as type I skull.**

**Case 7: Comparative Study of Binary Logistic Regression and Binary Discriminant Analysis:**

The Skull Type Prediction Problem has already been solved using Logistic Regression and now you will be solving the same problem using a different technique Discriminant Analysis. Compare the two methodologies for Skull Type Prediction Problem on following grounds:

1. Classification of individual skull.
2. Confusion Matrix, Percentage of Correct Classification.
3. Sensitivity and Specificity.
4. Performance on the Validation Data.

***Table 7.1-Case Summaries of Discriminant analysis and logistic regression***

| S No. | Skull Type |  | Predicted Group for Analysis 1 |
| --- | --- | --- | --- |
| 1 | 1 | 1.00 | |
| 2 | 0 | 1.00 | |
| 3 | 0 | .00 | |
| 4 | 1 | 1.00 | |
| 5 | 0 | .00 | |
| 6 | 0 | .00 | |
| 7 | 1 | .00 | |
| 8 | 1 | 1.00 | |
| 9 | 1 | .00 | |
| 10 | 0 | .00 | |
| 11 | 0 | 1.00 | |
| **12** | **1** | **1.00** | |
| 13 | 1 | 1.00 | |
| 14 | 0 | .00 | |
| 15 | 0 | .00 | |
| 16 | 1 | 1.00 | |
| **17** | **0** | **1.00** | |
| **18** | **0** | **1.00** | |
| 19 | 0 | .00 | |

| Predicted group |
| --- |
| 1.00 |
| 1.00 |
| .00 |
| 1.00 |
| .00 |
| .00 |
| .00 |
| 1.00 |
| .00 |
| .00 |
| 1.00 |
| **.00** |
| 1.00 |
| .00 |
| .00 |
| 1.00 |
| **.00** |
| **.00** |
| .00 |

From the above table we observe that the predicted groups are identical except for 12, 17, 18th observation.

b) The classification table indicates how well the model predicts group membership the classification table shows the proportion of correct assignments when the regression model has been applied to the data

**By logistic regression**

| ***Table 7.2-Classification Tablea*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Observed | | Predicted | | |
|  | Skull type | | Percentage Correct |
|  | .00 | 1.00 |
| Step 1 | Skull type | .00 | 9 | 2 | **81.8** |
| 1.00 | 3 | 5 | **62.5** |
| Overall Percentage | |  |  | 73.7 |
| a. The cut value is .500 | | | | | |

**Type I skull cases were the most accurately classified, with 81.8% of the cases correct. The type 2 skulls were classified with a success rate of 62.5%.The** overall success rate of **73.7%.**

| ***Table 7.3 Classification Resultsb,c*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | Skull type | Predicted Group Membership | | Total |
|  |  | .00 | 1.00 |
| Original | Count | .00 | 7 | 4 | 11 |
| 1.00 | 2 | 6 | 8 |
| % | .00 | **63.6** | 36.4 | 100.0 |
| 1.00 | 25.0 | **75.0** | 100.0 |
| . **68.4%** of original grouped cases correctly classified | | | | | |

**By discriminant analysis we get:**

**Type I skull cases were the most accurately classified, with 75% of the cases correct. The type 2 skulls were classified with a success rate of 63.6%.It also shows that 68.4% of original grouped cases are correctly classified**

**That is overall success rate is 68.4%.**

c) **By logistic regression** -**Sensitivity**=***62.5%***

I.e. a test with 62.5% sensitivity detects **62.5%** of patients with the disease (true positives) but **37.5%** with the disease go undetected (false negatives

***Specificity*** =***81.8%***

i.e. a test with 81.8% specificity correctly reports **81.8%** of patients without the disease as test negative (true negatives) but **18.2%** patients without the disease are incorrectly identified as test positive (false positive).

**By Discriminant Analysis**:-**Sensitivity= (6/6+4) =0.6**

I.e. a test with 60% sensitivity detects **60%** of patients with the disease (true positives) but **40%** with the disease go undetected (false negatives).

**Specificity=7/ (7+2) =0.77**

i.e. a test with 77% specificity correctly reports **77%** of patients without the disease as test negative (true negatives) but **23%** patients without the disease are incorrectly identified as test positive (false positive).

**d)We know that the discriminant function is:**

**L=-1.559+0.007\*X1+0.042\*X2+0.005\*X3+0.007\*X4-0.019\*X5**

**On using (171, 134, 130, 69, 130) if the value of L for a particular skull is greater than -0.0815(the average of centroids equals -0.0815) the skull is assigned to type I skull (coded 0) otherwise to type II skull (coded 0).For given measurements of skull, we get value of L to be 3.929>-0.0815 hence we classify it as type I skull.**

**By logistic regression**

**For prediction**-f (z) =ez/1+ez=π̂I

**Where z=1.149-0.008X1-0.047X2-0.007X3-0.006X4+0.022X5**

**We have (171, 134, 130, 69 and 130)**

**On calculating π̂I we get (0.0068 which is less than 0.5) so the class is 0, that is type I skull.**

Results from discriminant analysis and logistic regression were compared using two data sets from a study on predictors of skull type data. Both techniques selected the same set of variables as important predictors and were of nearly equal value in classifying skull data into type I and type II. The logistic regression model made fewer classification errors. The magnitudes of the effects were considerably different for some variables Logistic regression slightly exceeds discriminant function in the correct classification rate and in terms of specificity and sensitivity also.

The overall classification rate for both was good, and either can be helpful in predicting the type of skulls.In conclusion, logistic regression and discriminant analyses were similar in the model analysis. In order to decide which method should be used, we must consider the assumptions for the application of each one.

**Multiclass Classification**

A classification problem is said to be multiclass classification problem if the response has more than two possible classes. We will be discussing only Multiclass Logistic Regression as a multiclass classification technique though the other two techniques viz. Naïve Bayes’ Classifier and Discriminant Analysis as well have their multivariate extensions

Logistic Regression can be used to solve a multiclass classification problem in following two ways:

a) By means of decomposing the multiclass classification problem into several binary classification problems.

b) By means of using multinomial probability distribution.

**Decomposing the multiclass classification problem into several binary classification problems:**

Suppose the response variable has three class viz. 1, 2 and 3 then the response is defined as follows:

We can define three binary variables using as follows:

Clearly there are three binary classification problems which model and respectively.

Using Binary Logistic Regression we can obtain and and can classify and but our goal is to classify the 3-class variable. We define the rule for multiclass classification as follows:

Where (# of classes) and (# of observations).

**2.2 Multinomial Distribution for Multiclass Logistic Regression Problem:**

Suppose the response variable has three class viz. 1, 2 and 3 then the response is defined as follows:

We can use the multinomial probability distribution to obtain the probability mass functions of s and hence the likelihood function of the sample observations. We can define appropriate links for different probabilities with the predictor variables. Then the likelihood function can be maximized using the IRLS technique and we can proceed further in a similar manner. SPSS does all of it in a built-in routine.

**Case 8:** Consider the dataset **Flower Species.xlsx** dataset which has data on Sepal Length, Sepal Width, Petal Length and Petal Width and Species Type for 150 different flowers and perform the following objectives.

1) Decompose the multiclass (3-class) classification problem into three Binary Classification Problems and perform the following for each problem:

a) Test for the Significance individual independent variables.

b) Test for the overall Logistic Regression using Hosmer and Lemeshow Test (It’s a Chi-Square Test).

c) Construct the Confusion (Classification) Table and report the percentage of correct classification for the given emails. Also calculate specificity and sensitivity of the model.

d) For each flower obtain the predicted probability and hence the predicted class using the Logistic Classifier you have built.

2) Obtain the multiclass predicted flower species for the original problem using the three sub problems.

3) Construct the classification matrix and report the percentage of correct classification.

**First of all we build a binary logistic model by coding setosa as 1 and the other two as 0.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | ***Table 8.1 Hosmer and Lemeshow Test*** | | | | | --- | --- | --- | --- | | Step | Chi-square | df | Sig. | | 1 | .000 | 8 | 1.000 | |

We set the following hypothesis:

Ho: The model fits the data well.

H1: The model does not fit the data well.

The Hosmer-Lemeshow tests the null hypothesis that predictions made by the model fit perfectly with observed group memberships. Cases are arranged in order by their predicted probability on the criterion variable. These ordered cases are then divided into ten (usually) groups of equal or near equal size ordered with respect to the predicted probability of the target event. For each of these groups we then obtain the predicted group memberships and the actual group memberships. This results in a 2 x 10 contingency table, as shown below. A chi-square statistic is computed comparing the observed frequencies with those expected under the linear model. A non-significant chi-square indicates that the data fit the model well. This procedure suffers from several problems, one of which is that it relies on a test of significance. With large sample sizes, the test may be significant, even when the fit is good. With small sample sizes it may not be significant, even with poor fit. Here, we see model fit is acceptable χ² (8) = 0.0, *p*=1.0, which indicates our model predicts values not significantly different from what we observed. (We want the *p-*value to be *greater than* the established cutoff (generally 0.05) to indicate good fit.)

| ***Table 8.2 Hosmer and Lemeshow contingency table*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Dependent 1 = .00 | | Dependent 1 = 1.00 | | Total |
|  |  | Observed | Expected | Observed | Expected |
| Step 1 | 1 | 15 | 15.000 | 0 | .000 | 15 |
| 2 | 15 | 15.000 | 0 | .000 | 15 |
| 3 | 15 | 15.000 | 0 | .000 | 15 |
| 4 | 15 | 15.000 | 0 | .000 | 15 |
| 5 | 15 | 15.000 | 0 | .000 | 15 |
| 6 | 15 | 15.000 | 0 | .000 | 15 |
| 7 | 10 | 10.000 | 5 | 5.000 | 15 |
| 8 | 0 | .000 | 15 | 15.000 | 15 |
| 9 | 0 | .000 | 15 | 15.000 | 15 |
| 10 | 0 | .000 | 15 | 15.000 | 15 |

The Contingency Table for Hosmer and Lemeshow Test simply shows the observed and expected values for each category of the outcome variable as used to calculate the Hosmer and Lemeshow chi-square.

| ***Table 8.3 Classification Tablea*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Observed | | Predicted | | |
|  | Dependent 1 | | Percentage Correct |
|  | .00 | 1.00 |
| Step 1 | Dependent 1 | .00 | **100** | 0 | 100.0 |
| 1.00 | 0 | **50** | 100.0 |
| Overall Percentage | |  |  | **100.0** |
| a. The cut value is .500 | | | | | |

The Classification Table above shows how well our full model correctly classifies cases. A perfect model would show only values in the diagonal--correctly classifying all cases. Adding across the rows represents the number of cases in each category in the actual data and adding down the columns represents the number of cases in each category as classified by the full model. The key piece of information is the overall percentage in the lower right corner which shows our model (with all predictors & the constant) is **100%** accurate; which is excellent.

| ***Table 8.4 Variables in the Equation*** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 1a | X1 | 8.666 | 14840.420 | .000 | 1 | 1.000 | 5800.142 |
| X2 | 6.637 | 6922.781 | .000 | 1 | .999 | 762.489 |
| X3 | -15.119 | 12366.721 | .000 | 1 | .999 | .000 |
| X4 | -16.272 | 17915.189 | .000 | 1 | .999 | .000 |
| Constant | -13.616 | 51859.146 | .000 | 1 | 1.000 | .000 |
| a. Variable(s) entered on step 1: X1, X2, X3, and X4. | | | | | | | |

The Variables in the Equation table (above), shows the logistic coefficient (B) for each predictor variable. The logistic coefficient is the expected amount of change in the logit for each, *df*, Sig. (*p*-value); as well as the Exp (B) .The one unit change in the predictor. The logit is what is being predicted; it is the odds of membership in the category of the outcome variable with the numerically higher value (here a 1, rather than 0). The closer a logistic coefficient is to zero, the less influence it has in predicting the logit. The table also displays the standard error, Wald statistic (and associated *p*-value) which is used to evaluate whether or not the logistic coefficient is different than zero.

We set the following hypothesis to test the individual significance of beta coefficients.

Ho: βi=0 the individual regressor is not significant.

H1:βi≠0 the individual regressor is significant.

The Exp (B) is the odds ratio associated with each predictor. We expect predictors which increase the logit to display Exp (B) greater than 1.0, those predictors which do not have an effect on the logit will display an Exp (B) of 1.0 and predictors which decrease the logit will have Exp (B) values less than 1.0. Note that the Exp (B) is wildly large for the x1 and x2 predictor. This is due to a combination of the strong relationship between that variable and the outcome variable. Generally, when using continuous variables as predictors, we do not see such large Exp (B).

**Now we build a second binary logistic model by coding versicolor as 1 and other 2 as 0.**

| ***Table 8.5 Hosmer and Lemeshow Test*** | | | |
| --- | --- | --- | --- |
| Step | Chi-square | df | Sig. |
| 1 | 8.524 | 8 | .384 |

We set the following hypothesis:

Ho: Model is a good fit to the data.

H1: Model is not a good fit to the data.

The [Hosmer & Lemeshow test](http://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/glossary/indexa039.html?selectedLetter=H#hosmer-and-lemeshow-test) of the goodness of fit suggests the model is a good fit to the data as *p=0.384*(*>.05*)*.*However the chi-squared statistic on which it is based is very dependent on sample size so the value cannot be interpreted in isolation from the size of the sample.

This test divides the data into several groups based on p cap values, then computes a chi-square from observed and expected frequencies of subjects falling in the two categories of the binary response variable within these groups.  Large chi-square values (and correspondingly small p-values) indicate a lack of fit for the model.

| ***Table 8.6 Contingency Table for Hosmer and Lemeshow Test*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Dependent 2 = .00 | | Dependent 2 = 1.00 | | Total |
|  |  | Observed | Expected | Observed | Expected |
| Step 1 | 1 | 15 | 14.546 | 0 | .454 | 15 |
| 2 | 15 | 13.998 | 0 | 1.002 | 15 |
| 3 | 14 | 13.396 | 1 | 1.604 | 15 |
| 4 | 12 | 12.445 | 3 | 2.555 | 15 |
| 5 | 11 | 11.158 | 4 | 3.842 | 15 |
| 6 | 10 | 10.227 | 5 | 4.773 | 15 |
| 7 | 5 | 8.882 | 10 | 6.118 | 15 |
| 8 | 7 | 7.372 | 8 | 7.628 | 15 |
| 9 | 8 | 5.152 | 7 | 9.848 | 15 |
| 10 | 3 | 2.823 | 12 | 12.177 | 15 |

The Contingency Table for Hosmer and Lemeshow Test simply shows the observed and expected values for each category of the outcome variable as used to calculate the Hosmer and Lemeshow chi-square.

The Hosmer-Lemeshow tests the null hypothesis that predictions made by the model fit perfectly with observed group memberships. Cases are arranged in order by their predicted probability on the criterion variable. These ordered cases are then divided into ten (usually) groups of equal or near equal size ordered with respect to the predicted probability of the target event.

For each of these groups we then obtain the predicted group memberships and the actual group memberships this results in a 2 x 10 contingency table, as shown below. A chi-square statistic is computed comparing the observed frequencies with those expected under the linear model. A nonsignificant chi-square indicates that the data fit the model well. This procedure suffers from several problems, one of which is that it relies on a test of significance. With large sample sizes, the test may be significant, even when the fit is good. With small sample sizes it may not be significant, even with poor fit.

**Classification Table for Block 1**

| ***Table 8.7 Classification Tablea*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Observed | | Predicted | | |
|  | Dependent 2 | | Percentage Correct |
|  | .00 | 1.00 |
| Step 1 | Dependent 2 | .00 | **86** | 14 | 86.0 |
| 1.00 | 25 | **25** | 50.0 |
| Overall Percentage | |  |  | **74.0** |
| a. The cut value is .500 | | | | | |

Classification Table is based on the model that includes the [explanatory variables](http://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/glossary/index34aa.html?selectedLetter=E#explanatory-variable). As we can see our model is now correctly classifying the outcome for **74%** of the cases. It classifies dependent variable of the 0 category (Flower Type coded as 0) correctly in 86% cases and dependent variable of the 1 category correctly in 50% of the cases. It correctly specifies flower type coded 0 better than skull type coded 1.

| ***Table 8.8 Variables in the Equation*** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 1a | X1 | -.245 | .650 | .143 | 1 | .706 | **.782** |
| X2 | -2.797 | .784 | 12.739 | 1 | .000 | **.061** |
| X3 | 1.314 | .684 | 3.691 | 1 | .055 | **3.720** |
| X4 | -2.778 | 1.173 | 5.609 | 1 | .018 | **.062** |
| Constant | 7.378 | 2.499 | 8.716 | 1 | .003 | **1601.165** |
| a. Variable(s) entered on step 1: X1, X2, X3, and X4. | | | | | | | |

The “Variables in the Equation” table shows the output resulting from including all of the predictor variables in the equation. In this table we see the coefficients, their standard errors, the Wald test statistic with associated degrees of freedom and p-values, and the exponentiated coefficient (also known as an odds ratio).X1 and X3 are statistically significant. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

**Now we build a third binary logistic model by coding flowers from virginica as 1 and other two as 0.**

| ***Table 8.9 Hosmer and Lemeshow Test*** | | | |
| --- | --- | --- | --- |
| Step | Chi-square | df | Sig. |
| 1 | .259 | 8 | 1.000 |

The test is a version of a goodness-of- fit chi-square test with a null hypothesis that the data fit the model adequately. Therefore, a p-value larger than 0.05 suggests an adequate model fit, while a small p-value indicates some problem with the model such as non-monotonicity, variance inappropriate for the binomial model at each combination of explanatory variables, or the need to transform one of the explanatory variables.

In our case, a p-value of 1.00 suggests no problem with model fit (but the test is not very powerful). In the event of an indication of lack of fit, examining the Contingency Table for Hosmer and Lemeshow Test may help to point to the source of the problem. This test is a substitute for residual analysis, which in raw form is uninformative in logistic regression because there are only two possible values for the residual at each fixed combination of explanatory variables.

| ***Table 8.10 Contingency Table for Hosmer and Lemeshow Test*** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Dependent 3 = .00 | | Dependent 3 = 1.00 | | Total |
|  |  | Observed | Expected | Observed | Expected |
| Step 1 | 1 | 15 | 15.000 | 0 | .000 | 15 |
| 2 | 15 | 15.000 | 0 | .000 | 15 |
| 3 | 15 | 15.000 | 0 | .000 | 15 |
| 4 | 15 | 15.000 | 0 | .000 | 15 |
| 5 | 15 | 15.000 | 0 | .000 | 15 |
| 6 | 15 | 14.996 | 0 | .004 | 15 |
| 7 | 10 | 9.768 | 5 | 5.232 | 15 |
| 8 | 0 | .235 | 15 | 14.765 | 15 |
| 9 | 0 | .001 | 15 | 14.999 | 15 |
| 10 | 0 | .000 | 15 | 15.000 | 15 |

The output has divided the data into 10 groups based on the outcome variable.These groups are defined by flower type.The observed frequencies are that 15 cases belong to other two types and 0 cases belong to versicolor.The observed and expected frequencies (based on the prediction model) match reasonably well for all the steps and is a desirable result.

| ***Table 8.11 Classification Tablea*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Observed | | Predicted | | |
|  | Dependent 3 | | Percentage Correct |
|  | .00 | 1.00 |
| Step 1 | Dependent 3 | .00 | 99 | 1 | 99.0 |
| 1.00 | 1 | 49 | 98.0 |
| Overall Percentage | |  |  | **98.7** |
| a. The cut value is .500 | | | | | |

Classification Table is based on the model that includes the explanatory variables. As we can see our model is now correctly classifying the outcome for 98.7% of the cases. It classifies dependent variable of the 0 category (Flower Type coded as 0) correctly in 99% cases and dependent variable of the 1 category correctly in 98% of the cases. It correctly specifies flower type coded 0 better than skull type coded 1.

| ***Table 8.12 Variables in the Equation*** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | B | S.E. | Wald | df | Sig. | Exp(B) |
| Step 1a | X1 | -2.465 | 2.394 | 1.060 | 1 | .303 | **.085** |
| X2 | -6.681 | 4.480 | 2.224 | 1 | .136 | **.001** |
| X3 | 9.429 | 4.737 | 3.962 | 1 | .047 | **12448.870** |
| X4 | 18.286 | 9.743 | 3.523 | 1 | .061 | **8.741E7** |
| Constant | -42.638 | 25.708 | 2.751 | 1 | .097 | **.000** |
| a. Variable(s) entered on step 1: X1, X2, X3, and X4. | | | | | | | |

The “Variables in the Equation” table shows the output resulting from including all of the predictor variables in the equation. In this table we see the coefficients, their standard errors, the Wald test statistic with associated degrees of freedom and p-values, and the exponentiated coefficient (also known as an odds ratio).X1 and X3 are statistically significant. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

**Case 9:** Consider the dataset **Flower Species.xlsx** dataset which has data on Sepal Length, Sepal Width, Petal Length and Petal Width and Species Type for 150 different flowers and perform the following objectives.

* 1. Built a multinomial logistic regression model for the given problem.
  2. Obtain the predicted class for each flower using the multinomial logistic regression.
  3. Construct the classification Matrix and report the percentage of correct classification.

**Multinomial Logistic Regression** is the linear regression analysis to conduct when the dependent variable is nominal with more than two levels.  Thus it is an extension of logistic regression, which analyses dichotomous (binary) dependents.  Multinomial logistic regression is used to model nominal outcome variables, in which the log odds of the outcomes are modelled as a linear combination of the predictor variables.

Like all linear regressions, the multinomial regression is a predictive analysis.  Multinomial regression is used to describe data and to explain the relationship between one dependent nominal variable and one or more continuous-level (interval or ratio scale) independent variables.

**Nominal Regression**

| ***Table 9.1 Model Fitting Information*** | | | | |
| --- | --- | --- | --- | --- |
| Model | Model Fitting Criteria | Likelihood Ratio Tests | | |
| -2 Log Likelihood | Chi-Square | df | Sig. |
| Intercept Only | 329.584 |  |  |  |
| Final | 11.899 | 317.685 | 8 | .000 |

We set the following hypothesis:

Ho: Full model has better fit than the null model.

H1: Full model does not have a better fit than a null model.

**Model** - This indicates the parameters of the model for which the model fit is calculated.  "Intercept Only" describes a model that does not control for any predictor variables and simply fits an intercept to predict the outcome variable. "Final" describes a model that includes the specified predictor variables and has been arrived at through an iterative process that maximizes the log likelihood of the outcomes seen in the outcome variable. By including the predictor variables and maximizing the log likelihood of the outcomes seen in the data, the "Final" model should improve upon the "Intercept Only" model.  This can be seen in the differences in the -2(Log Likelihood) values associated with the models.

1)**-2(Log Likelihood)** - This is the product of -2 and the log likelihoods of the null model and fitted "final" model. The likelihood of the model is used to test of whether all predictors' regression coefficients in the model are simultaneously zero.

2)**Chi-Square** - This is the Likelihood Ratio (LR) Chi-Square test that at least one of the predictors' regression coefficient is not equal to zero in the model. The LR Chi-Square statistic can be calculated by  -2\*L(null model) - (-2\*L(fitted model)) = 329.584 – 11.899 = 317.685, where *L(null model)* is from the log likelihood with just the response variable in the model (Intercept Only) and *L(fitted model)* is the log likelihood from the final iteration (assuming the model converged) with all the parameters.

3)**df** - This indicates the degrees of freedom of the chi-square distribution used to test the LR Chi-Square statistic and is defined by the number of predictors in the model.

Here, we see model fit is significant χ² (8) = 317.685, *p*< .05, which indicates our full model predicts significantly better, or more accurately, than the null model. We want the *p-*value to be *less than* your established cutoff (generally 0.05) to indicate good fit.

| ***Table 9.2 Likelihood Ratio Tests*** | | | | |
| --- | --- | --- | --- | --- |
| **Effect** | **Model Fitting Criteria** | **Likelihood Ratio Tests** | | |
| **-2 Log Likelihood of Reduced Model** | **Chi-Square** | **df** | **Sig.** |
| Intercept | 21.680 | 9.781 | 2 | .008 |
| X1 | 13.266 | 1.367 | 2 | .505 |
| X2 | 15.492 | 3.594 | 2 | .166 |
| X3 | 25.902 | 14.003 | 2 | .001 |
| X4 | 23.772 | 11.873 | 2 | .003 |
| The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0. | | | | |

The statistics in the Likelihood Ratio Tests table are the same types as those reported for the null and full models above in the Model Fitting Information table. Here however, each element of the model is being compared to the full model in such a way as to allow the research to determine if it (each element) should be included in the full model.

In other words, does each element (predictor) contribute meaningfully to the full effect? For instance, we see that the X1, X3 predictor displays a non-significant (*p*= .505, p=0.166) chi-square which indicates X1, X3 could be dropped from the model and the overall fit would NOT be significantly reduced. To be clear, if the *p-*value is *less than* your established cutoff (generally 0.05) for a predictor then that predictor contributes significantly to the full (final) model.

| ***Table 9.3 Parameter Estimates*** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dependenta** | | **B** | **Std. Error** | **Wald** | **df** | **Sig.** | **Exp(B)** | **95% Confidence Interval for Exp(B)** | |
| **Lower Bound** | **Upper Bound** |
| 1.00 | Intercept | 33.164 | 185175.457 | .000 | 1 | 1.000 |  |  |  |
| X1 | 11.864 | 54562.262 | .000 | 1 | 1.000 | 142063.517 | .000 | .b |
| X2 | 13.276 | 25968.390 | .000 | 1 | 1.000 | 583190.925 | .000 | .b |
| X3 | -26.896 | 25481.604 | .000 | 1 | .999 | 2.086E-12 | .000 | .b |
| X4 | -38.067 | .000 | . | 1 | . | 2.935E-17 | 2.935E-17 | 2.935E-17 |
| 2.00 | Intercept | 42.638 | 25.708 | 2.751 | 1 | .097 |  |  |  |
| X1 | 2.465 | 2.394 | 1.060 | 1 | .303 | 11.766 | .108 | 1284.293 |
| X2 | 6.681 | 4.480 | 2.224 | 1 | .136 | 797.026 | .123 | 5181847.251 |
| X3 | -9.429 | 4.737 | 3.962 | 1 | .047 | 8.033E-5 | 7.457E-9 | .865 |
| X4 | -18.286 | 9.743 | 3.523 | 1 | .061 | 1.144E-8 | 5.828E-17 | 2.246 |
| a. The reference category is: 3.00. | | | | | | | | | |
| b. Floating point overflow occurred while computing this statistic. Its value is therefore set to system missing. | | | | | | | | | |

The reference category plays the same role in multinomial logistic regression that it plays in the dummy-coding of a nominal variable: it is the category that would be coded with zeroes for all of the dummy-coded variables that all other categories are interpreted against.

The Parameter Estimates table (above), shows the logistic coefficient (B) for each predictor variable for each alternative category of the outcome variable. Alternative category meaning, not the reference category. The logistic coefficient is the expected amount of change in the logit for each one unit change in the predictor. The logit is what is being predicted; it is the odds of membership in the category of the outcome variable which has been specified (here the first value: 1 was specified, rather than the alternative values 2 or 3). The closer a logistic coefficient is to zero, the less influence the predictor has in predicting the logit. The table also displays the standard error, Wald statistic, *DF*, Sig. (*p*-value); as well as the Exp (B) and confidence interval for the Exp (B). The Wald test (and associated *p*-value) is used to evaluate whether or not the logistic coefficient is different than zero. The Exp (B) is the odds ratio associated with each predictor. We expect predictors which increase the logit to display Exp (B) greater than 1.0, those predictors which do not have an effect on the logit will display an Exp (B) of 1.0 and predictors which decease the logit will have Exp (B) values less than 1.0.

| ***Table 9.4 Classification*** | | | | |
| --- | --- | --- | --- | --- |
| Observed | Predicted | | | |
| 1.00 | 2.00 | 3.00 | Percent Correct |
| 1.00 | **50** | 0 | 0 | 100.0% |
| 2.00 | 0 | **49** | 1 | 98.0% |
| 3.00 | 0 | 1 | **49** | 98.0% |
| Overall Percentage | 33.3% | 33.3% | 33.3% | **98.7%** |

The Classification Table (above) shows how well our full model correctly classifies cases. A perfect model would show only values on the diagonal--correctly classifying all cases. Adding across the rows represents the number of cases in each category in the actual data and adding down the columns represents the number of cases in each category as classified by the full model. We see that setosa flower is classified with 100 % accuracy, versicolor flower is classified with 98% accuracy and virginica flower is classified with 98 % accuracy .The key piece of information is the overall percentage in the lower right corner which shows our model (with all predictors & the constant) is **98.7%** accurate; which is excellent.

**Case 10:** Comparative Study of the two approaches of Multiclass Classification for the Flower Species Prediction Problem – Perform the following objectives:

* 1. Compare the predicted class for each flower based on both the approaches.
  2. Compare the diagonal entries of the classification matrix based on both the approaches.
  3. Compare the percentage of correct classification for both the approaches.

a) Consider the table given below-

***Table 10.1 Case summaries***

|  | Dependent | Predicted Response Category | Estimated Classification Probability for the Predicted Category | Dependent 1 | Predicted probability | Predicted group | Dependent 2 | Predicted probability | Predicted group | Dependent 3 | Predicted probability | Predicted group |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .08 | .00 | 0 | .00 | .00 |
| 2 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .28 | .00 | 0 | .00 | .00 |
| 3 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .17 | .00 | 0 | .00 | .00 |
| 4 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .27 | .00 | 0 | .00 | .00 |
| 5 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .07 | .00 | 0 | .00 | .00 |
| 6 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .02 | .00 | 0 | .00 | .00 |
| 7 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .10 | .00 | 0 | .00 | .00 |
| 8 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .13 | .00 | 0 | .00 | .00 |
| 9 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .37 | .00 | 0 | .00 | .00 |
| 10 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .31 | .00 | 0 | .00 | .00 |
| 11 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .05 | .00 | 0 | .00 | .00 |
| 12 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .15 | .00 | 0 | .00 | .00 |
| 13 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .35 | .00 | 0 | .00 | .00 |
| 14 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .29 | .00 | 0 | .00 | .00 |
| 15 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .01 | .00 | 0 | .00 | .00 |
| 16 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .00 | .00 | 0 | .00 | .00 |
| 17 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .01 | .00 | 0 | .00 | .00 |
| 18 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .07 | .00 | 0 | .00 | .00 |
| 19 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .04 | .00 | 0 | .00 | .00 |
| 20 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .03 | .00 | 0 | .00 | .00 |
| 21 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .14 | .00 | 0 | .00 | .00 |
| 22 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .03 | .00 | 0 | .00 | .00 |
| 23 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .04 | .00 | 0 | .00 | .00 |
| 24 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .09 | .00 | 0 | .00 | .00 |
| 25 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .20 | .00 | 0 | .00 | .00 |
| 26 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .33 | .00 | 0 | .00 | .00 |
| 27 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .09 | .00 | 0 | .00 | .00 |
| 28 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .09 | .00 | 0 | .00 | .00 |
| 29 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .11 | .00 | 0 | .00 | .00 |
| 30 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .24 | .00 | 0 | .00 | .00 |
| 31 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .28 | .00 | 0 | .00 | .00 |
| 32 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .07 | .00 | 0 | .00 | .00 |
| 33 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .02 | .00 | 0 | .00 | .00 |
| 34 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .01 | .00 | 0 | .00 | .00 |
| 35 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .25 | .00 | 0 | .00 | .00 |
| 36 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .14 | .00 | 0 | .00 | .00 |
| 37 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .07 | .00 | 0 | .00 | .00 |
| 38 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .09 | .00 | 0 | .00 | .00 |
| 39 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .28 | .00 | 0 | .00 | .00 |
| 40 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .12 | .00 | 0 | .00 | .00 |
| 41 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .06 | .00 | 0 | .00 | .00 |
| 42 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .67 | 1.00 | 0 | .00 | .00 |
| 43 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .18 | .00 | 0 | .00 | .00 |
| 44 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .04 | .00 | 0 | .00 | .00 |
| 45 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .04 | .00 | 0 | .00 | .00 |
| 46 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .23 | .00 | 0 | .00 | .00 |
| 47 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .05 | .00 | 0 | .00 | .00 |
| 48 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .20 | .00 | 0 | .00 | .00 |
| 49 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .05 | .00 | 0 | .00 | .00 |
| 50 | 1 | 1.00 | 1.00 | 1 | 1.00 | 1.00 | 0 | .14 | .00 | 0 | .00 | .00 |
| 51 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .27 | .00 | 0 | .00 | .00 |
| 52 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .20 | .00 | 0 | .00 | .00 |
| 53 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .33 | .00 | 0 | .00 | .00 |
| 54 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .78 | 1.00 | 0 | .00 | .00 |
| 55 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .46 | .00 | 0 | .00 | .00 |
| 56 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .61 | 1.00 | 0 | .00 | .00 |
| 57 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .16 | .00 | 0 | .00 | .00 |
| 58 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .74 | 1.00 | 0 | .00 | .00 |
| 59 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .52 | 1.00 | 0 | .00 | .00 |
| 60 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .45 | .00 | 0 | .00 | .00 |
| 61 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .92 | 1.00 | 0 | .00 | .00 |
| 62 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .25 | .00 | 0 | .00 | .00 |
| 63 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .90 | 1.00 | 0 | .00 | .00 |
| 64 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .51 | 1.00 | 0 | .00 | .00 |
| 65 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .27 | .00 | 0 | .00 | .00 |
| 66 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .26 | .00 | 0 | .00 | .00 |
| 67 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .34 | .00 | 0 | .00 | .00 |
| 68 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .73 | 1.00 | 0 | .00 | .00 |
| 69 | 2 | 2.00 | .94 | 0 | .00 | .00 | 1 | .81 | 1.00 | 0 | .06 | .00 |
| 70 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .75 | 1.00 | 0 | .00 | .00 |
| 71 | 2 | 2.00 | .60 | 0 | .00 | .00 | 1 | .15 | .00 | 0 | .40 | .00 |
| 72 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .42 | .00 | 0 | .00 | .00 |
| 73 | 2 | 2.00 | .78 | 0 | .00 | .00 | 1 | .75 | 1.00 | 0 | .22 | .00 |
| 74 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .71 | 1.00 | 0 | .00 | .00 |
| 75 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .43 | .00 | 0 | .00 | .00 |
| 76 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .32 | .00 | 0 | .00 | .00 |
| 77 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .57 | 1.00 | 0 | .00 | .00 |
| 78 | 2 | 2.00 | .72 | 0 | .00 | .00 | 1 | .31 | .00 | 0 | .28 | .00 |
| 79 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .39 | .00 | 0 | .00 | .00 |
| 80 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .63 | 1.00 | 0 | .00 | .00 |
| 81 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .78 | 1.00 | 0 | .00 | .00 |
| 82 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .80 | 1.00 | 0 | .00 | .00 |
| 83 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .55 | 1.00 | 0 | .00 | .00 |
| 84 | 2 | 3.00 | .87 | 0 | .00 | .00 | 1 | .65 | 1.00 | 0 | .87 | 1.00 |
| 85 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .36 | .00 | 0 | .00 | .00 |
| 86 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .11 | .00 | 0 | .00 | .00 |
| 87 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .28 | .00 | 0 | .00 | .00 |
| 88 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .83 | 1.00 | 0 | .00 | .00 |
| 89 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .35 | .00 | 0 | .00 | .00 |
| 90 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .66 | 1.00 | 0 | .00 | .00 |
| 91 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .77 | 1.00 | 0 | .00 | .00 |
| 92 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .41 | .00 | 0 | .00 | .00 |
| 93 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .65 | 1.00 | 0 | .00 | .00 |
| 94 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .78 | 1.00 | 0 | .00 | .00 |
| 95 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .59 | 1.00 | 0 | .00 | .00 |
| 96 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .44 | .00 | 0 | .00 | .00 |
| 97 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .44 | .00 | 0 | .00 | .00 |
| 98 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .45 | .00 | 0 | .00 | .00 |
| 99 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .51 | 1.00 | 0 | .00 | .00 |
| 100 | 2 | 2.00 | 1.00 | 0 | .00 | .00 | 1 | .48 | .00 | 0 | .00 | .00 |
| 101 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .08 | .00 | 1 | 1.00 | 1.00 |
| 102 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .46 | .00 | 1 | 1.00 | 1.00 |
| 103 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .30 | .00 | 1 | 1.00 | 1.00 |
| 104 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .52 | 1.00 | 1 | 1.00 | 1.00 |
| 105 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .25 | .00 | 1 | 1.00 | 1.00 |
| 106 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .49 | .00 | 1 | 1.00 | 1.00 |
| 107 | 3 | 3.00 | .89 | 0 | .00 | .00 | 0 | .59 | 1.00 | 1 | .89 | 1.00 |
| 108 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .68 | 1.00 | 1 | 1.00 | 1.00 |
| 109 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .80 | 1.00 | 1 | 1.00 | 1.00 |
| 110 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .03 | .00 | 1 | 1.00 | 1.00 |
| 111 | 3 | 3.00 | .99 | 0 | .00 | .00 | 0 | .12 | .00 | 1 | .99 | 1.00 |
| 112 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .49 | .00 | 1 | 1.00 | 1.00 |
| 113 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .22 | .00 | 1 | 1.00 | 1.00 |
| 114 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .50 | 1.00 | 1 | 1.00 | 1.00 |
| 115 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .14 | .00 | 1 | 1.00 | 1.00 |
| 116 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .07 | .00 | 1 | 1.00 | 1.00 |
| 117 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .41 | .00 | 1 | 1.00 | 1.00 |
| 118 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .08 | .00 | 1 | 1.00 | 1.00 |
| 119 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .71 | 1.00 | 1 | 1.00 | 1.00 |
| 120 | 3 | 3.00 | .92 | 0 | .00 | .00 | 0 | .90 | 1.00 | 1 | .92 | 1.00 |
| 121 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .10 | .00 | 1 | 1.00 | 1.00 |
| 122 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .28 | .00 | 1 | 1.00 | 1.00 |
| 123 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .71 | 1.00 | 1 | 1.00 | 1.00 |
| 124 | 3 | 3.00 | .95 | 0 | .00 | .00 | 0 | .43 | .00 | 1 | .95 | 1.00 |
| 125 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .14 | .00 | 1 | 1.00 | 1.00 |
| 126 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .39 | .00 | 1 | 1.00 | 1.00 |
| 127 | 3 | 3.00 | .82 | 0 | .00 | .00 | 0 | .34 | .00 | 1 | .82 | 1.00 |
| 128 | 3 | 3.00 | .80 | 0 | .00 | .00 | 0 | .26 | .00 | 1 | .80 | 1.00 |
| 129 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .38 | .00 | 1 | 1.00 | 1.00 |
| 130 | 3 | 3.00 | .97 | 0 | .00 | .00 | 0 | .60 | 1.00 | 1 | .97 | 1.00 |
| 131 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .61 | 1.00 | 1 | 1.00 | 1.00 |
| 132 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .09 | .00 | 1 | 1.00 | 1.00 |
| 133 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .31 | .00 | 1 | 1.00 | 1.00 |
| 134 | 3 | 2.00 | .80 | 0 | .00 | .00 | 0 | .63 | 1.00 | 1 | .20 | .00 |
| 135 | 3 | 3.00 | .97 | 0 | .00 | .00 | 0 | .89 | 1.00 | 1 | .97 | 1.00 |
| 136 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .22 | .00 | 1 | 1.00 | 1.00 |
| 137 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .05 | .00 | 1 | 1.00 | 1.00 |
| 138 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .35 | .00 | 1 | 1.00 | 1.00 |
| 139 | 3 | 3.00 | .67 | 0 | .00 | .00 | 0 | .24 | .00 | 1 | .67 | 1.00 |
| 140 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .15 | .00 | 1 | 1.00 | 1.00 |
| 141 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .10 | .00 | 1 | 1.00 | 1.00 |
| 142 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .06 | .00 | 1 | 1.00 | 1.00 |
| 143 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .46 | .00 | 1 | 1.00 | 1.00 |
| 144 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .13 | .00 | 1 | 1.00 | 1.00 |
| 145 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .05 | .00 | 1 | 1.00 | 1.00 |
| 146 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .10 | .00 | 1 | 1.00 | 1.00 |
| 147 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .53 | 1.00 | 1 | 1.00 | 1.00 |
| 148 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .21 | .00 | 1 | 1.00 | 1.00 |
| 149 | 3 | 3.00 | 1.00 | 0 | .00 | .00 | 0 | .05 | .00 | 1 | 1.00 | 1.00 |
| 150 | 3 | 3.00 | .98 | 0 | .00 | .00 | 0 | .32 | .00 | 1 | .98 | 1.00 |

**Note that if the predicted probability is less than 0.5(the default threshold value in SPSS), the predicted group is the group with (coding 0). By comparing the multinomial logistic regression approach and the logistic regression approach (where y1 is dependent variable=setosa), we observe that the predicted class is same for all the flowers. It is not evident due to the difference in coding (for multinomial we have 1-setosa, 2-versicolor and 3-virginica whereas in logistic we have 1-setosa and 0—versicolor and virginica).**

**By comparing the multinomial logistic regression approach and the logistic regression approach (where y2 is dependent variable=versicolor) we observe that there are differences in the predicted groups. When we have probability>0.5, it belongs to versicolor (coded1 in 10 th column) .Similarly for virginica (dependent variable y3)**

| ***Table 10.2 Classification Tablea*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Observed | | Predicted | | |
|  | Dependent 1 | | Percentage Correct |
|  | .00 | 1.00 |
| Step 1 | Dependent 1 | .00 | **100** | 0 | 100.0 |
| 1.00 | 0 | **50** | 100.0 |
| Overall Percentage | |  |  | **100.0** |
| a. The cut value is .500 | | | | | |

b) A more useful measure to assess the utility of a logistic regression model is classification accuracy, which compares predicted group membership based on the logistic model to the actual, known group membership, which is the value for the dependent variable.

When we code setosa as 1 and others as 0. It classifies dependent variable of the 0 category (flower Type coded as 0) correctly in 100% cases and dependent variable of the 1 category correctly in 100% of the cases. It correctly specifies both the flower types.

| ***Table 10.3 Classification Tablea*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Observed | | Predicted | | |
|  | Dependent 2 | | Percentage Correct |
|  | .00 | 1.00 |
| Step 1 | Dependent 2 | .00 | **86** | 14 | 86.0 |
| 1.00 | 25 | **25** | 50.0 |
| Overall Percentage | |  |  | **74.0** |
| a. The cut value is .500 | | | | | |

When we code versicolor as 1 and others as 0. It classifies dependent variable of the 0 category (Flower Type coded as 0) correctly in 86% cases and dependent variable of the 1 category correctly in 50% of the cases. It correctly specifies flower type coded 0 better than flower type coded 1.

| ***Table 10.4 Classification Tablea*** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Observed | | Predicted | | |
|  | Dependent 3 | | Percentage Correct |
|  | .00 | 1.00 |
| Step 1 | Dependent 3 | .00 | 99 | 1 | 99.0 |
| 1.00 | 1 | 49 | 98.0 |
| Overall Percentage | |  |  | **98.7** |
| a. The cut value is .500 | | | | | |

When we code virginica as 1 and others as 0. It classifies dependent variable of the 0 category (flower Type coded as 0) correctly in 99% cases and dependent variable of the 1 category correctly in 98% of the cases. It correctly specifies flower type coded 0 better than flower type coded 1.

So the binary model with setosa is the best binary logistic model as compared to the other two. This is because it correctly classifies both the flower types in all the cases.

| ***Table 10.5 Classification*** | | | | |
| --- | --- | --- | --- | --- |
| Observed | Predicted | | | |
| 1.00 | 2.00 | 3.00 | Percent Correct |
| 1.00 | **50** | 0 | 0 | 100.0% |
| 2.00 | 0 | **49** | 1 | 98.0% |
| 3.00 | 0 | 1 | **49** | 98.0% |
| Overall Percentage | 33.3% | 33.3% | 33.3% | **98.7%** |

But in a multinomial logistic model, we have one model which classifies all cases well. It classifies flower type setosa, versicolor and virginica 100%, 98%, 98% correctly. It is better as compared to binary logistic as it specifies almost all cases (all flower types) correctly most of the time.

c) We build a binary logistic model by coding setosa as 1 and the other two as 0.From the table 10.2 we can say that full model correctly classifies 100% of the cases correctly. From table 10.3 we can say that on coding versicolor as 1 and rest as 0, the model can correctly identify only 74% of the cases. From table 10.4 we build a third logistic regression model by coding virginica as 1 and rest as 0 and this model can correctly identify 98.7% of the cases. Finally from table 10.5 we combine the three to form a multiclass model, which could correctly classify 98.7% of all the cases. The multinomial logistic model is better because it classifies all cases correctly most of the time (98.7%).

**Bibliography**

1) Abhishek Sir‘s notes.

2) PASW Statistics 17 Made Simple by Paul R. Kinnear and Colin.D.Gray

3) [www.ats.ucla.edu](http://www.ats.ucla.edu)

4) Google has been used generously to verify facts.

5) Discovering Statistics using SPSS by Andy Field