***Case Study-Classification Problem***

***Submitted by: T. Sujaya***

***INTRODUCTION***

We know that in multiple regression, the values of one variable (the target, criterion or dependent variable Y) are estimated from those of two or more independent variables (or regressors X1,…Xp).Logistic regression or logit regression ,or logit model is a regression model where the dependent variable is categorical and predictor variables are continuous or categorical. This simply means that the dependent variable is binary, i.e. can take only two values, such as pass/fail or win/lose. Cases with more than two categories are referred to as multinomial logistic regression. 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.

Logistic regression can be seen as a special case of generalized linear model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions from those of linear regression. In particular the key differences of these two models can be seen in the following two features of logistic regression. First the conditional distribution y|x is a Bernoulli distribution than a Gaussian distribution, because the dependent variable is binary. Second the predicted values are probabilities and are therefore restricted to (0, 1) through the logistic distribution function because logistic regression predicts the probability of particular outcomes. Logistic regression is an alternative to Fisher’s 1936 classification method, linear discriminant analysis. If the assumptions of linear discriminant analysis hold, application of Bayes rule to reverse the conditioning results in the logistic model, so if linear discriminant assumptions are true, logistic regression assumptions must hold. The converse is not true, so the logistic model has fewer assumptions than discriminant analysis and makes no assumption on the distribution of the independent variables. 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. 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.

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.
3. **Binary Classification**
   1. **Email Spam Filtering using Naïve Bayes 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 text including 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 are independent 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:

***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 email 1 | P | 0.999784=0.9998 | |
|  | For email 2 | P | 0.233577=0.2336 | |

**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 calculating the overall spamicity of email 1 using excel (calculations are shown above in columns 4 and 5) we get 0.9998.Hence the probability of email 1 being spam is 0.9998.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 above in columns 4 and 5) we get 0.2336.Hence the probability of email 2 being spam is 0.2336.We have classified email 2 as non spam or ham.*

**Logistic Regression**

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

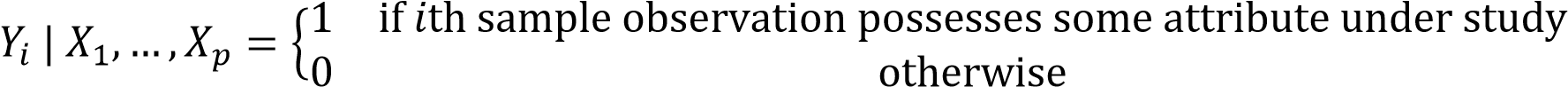


Equivalently,



Where  and  .

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



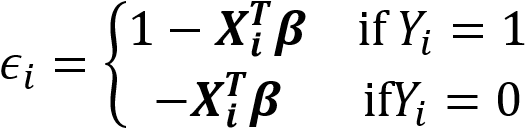
Clearly  where  and .

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.

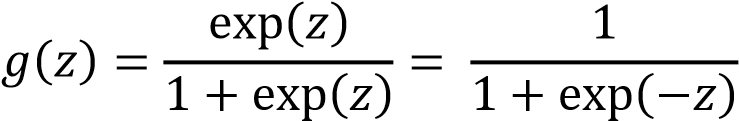


Some function of which is bound to lie between 0 and 1 will serve the purpose. Generally when response variable is binary, there is considerable empirical evidence indicating that the relationship between  and  is non-linear.

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

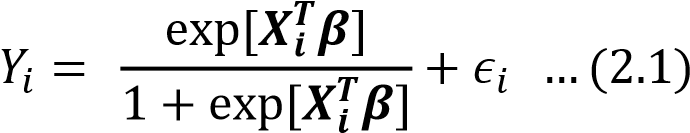
10

15

Z

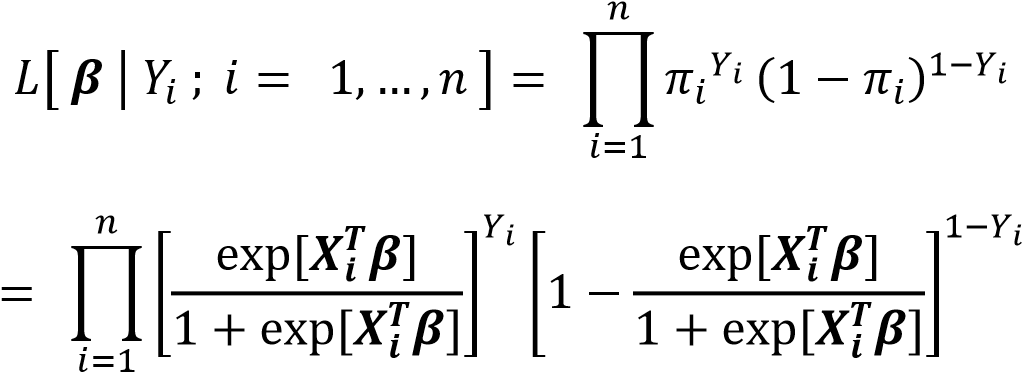
**Sigmoid Function**

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

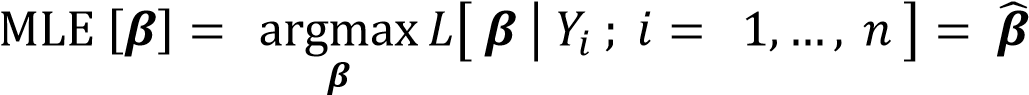


Observe that the above expression specifies a **non-linear regression model** which occurs when the dependent variable depends on independent variables through a non-linear function of unknown parameters. The stochastic model defined in (2.1) is called as a **logistic regression model**.

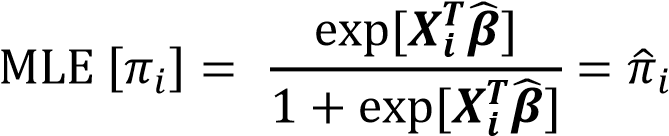
After specifying an appropriate model for modeling a binary response variable, the next task is to estimate the parameters. For estimation of parameters we use the famous maximum likelihood estimation (MLE). Consider the available observations  as independent observations from the conditional distribution 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.

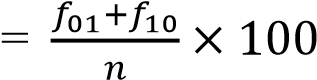
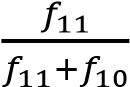
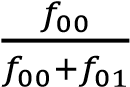
1. Spam Detection: Spam/Non-Spam
2. Tumor Examination: Malignant/Benign (iii) Online Transaction: Fraudulent/Non-Fraudulent (iv) Email Labeling: Work, Friends, Family etc.

# 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** |  |  |  |

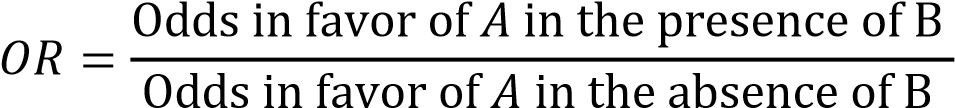
Where and, is the frequency of the class where observed response is  and estimated response .Following simple measures of goodness can be defined.

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

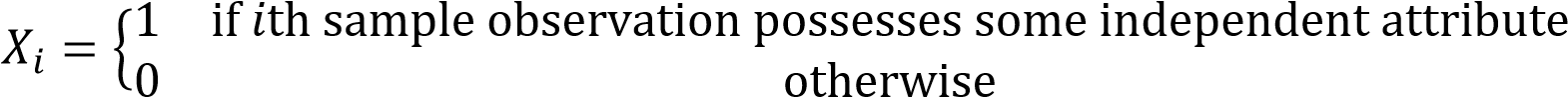
Hosmer and Lemeshow chi-square goodness of fit test used to test the goodness of logistic model.

# **Odds Ratio**

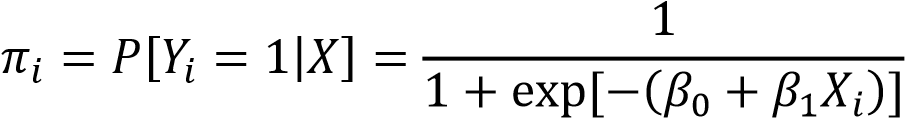
Odds Ratio (OR) is a measure of association between the two attributes  and, say defined as follows:



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



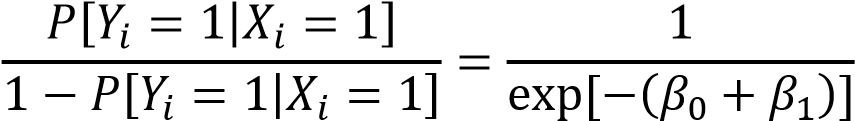
As defined earlier we have,



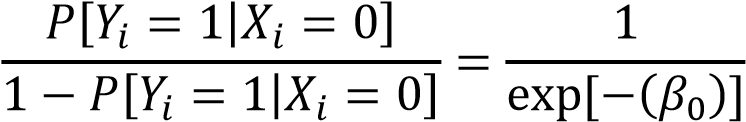
Consider the following table.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 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  )  ] |

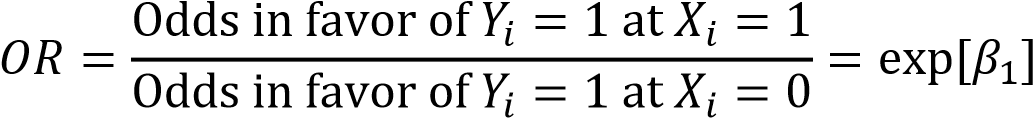
Odds in favor of 𝑌𝑖= 1 at 𝑋𝑖= 1 are given by



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



Hence Odds ratio is given by



Odds ratio is a ratio of the odds at two values of the independent variable that are one unit apart. It indicates how many times higher the odds of occurrence are for each one unit increase in the independent variable

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.

Logistic regressions predict likelihoods, measured by probabilities, odds or log-odds. But there is a distinction between probability and odds. A probability is the ratio of the number of occurrences to the total number of possibilities whereas odds is the ratio of number of occurrences to the number of non occurrences. Probabilities range from 0 to 1, whereas odds range from 0 to infinity. It is basically an indicator of the change in odds resulting from a unit change in the predictor. As such it is similar to the b coefficient in logistic regression but easier to understand (because it doesn’t require a logarithmic transformation).When the predictor variable is categorical exp B is easier to explain. The proportionate change in odds is exp b, so we can interpret exp b in terms of the change in odds: if the value is equal to one then it indicates equal probability of occurrence and non occurrence and if the value is greater than 1 then it indicates that as the predictor increases, the odds of the outcome occurring increase. Conversely, a value less than 1 indicates that as the predictor increases, the odds of the outcome occurring decrease.

The default method of conducting regression is ‘enter’. This is same as forced entry in multiple regression in that all of the covariates are placed into the regression model in one block, and parameter estimates are calculated for each block. When constructing the binary dependent variable, it is important that the categories be mutually exclusive, so that a case cannot be in both categories at the same time. 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 (Model in step 0 as given by SPSS which usually doesn’t contain the predictors). 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 (model in step 0 and model in step 1) are trivial.

There is an important statistic called the Wald Chi-Square statistic, which tests the unique contribution of each predictor and tells us whether the ß coefficient for that predictor is significantly different from 0.If the test fails to reject the null hypothesis, this suggests that removing the variables from the model will not substantially harm the fit of that model, since a predictor with a coefficient that is very small relative to its standard error is generally not doing much to help predict the dependent variable. It tests how far the estimated parameters are from zero(or any other value under the null hypothesis) in standard errors, similar to the hypothesis tests typically printed in regression output. The difference is that the Wald test can be used to test multiple parameters simultaneously, while the tests typically printed in regression output only test one parameter at a time.

***Wald statistic tests a regression coefficient (or the constant) for significance with null hypothesis=0, the value of the (parameter) coefficient 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.

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.

 Hosmer-Lemeshow goodness-of-fit statistic- This goodness-of-fit statistic is more robust than the traditional goodness-of-fit statistic used in logistic regression, particularly for models with continuous covariates and studies with small sample sizes. It is based on grouping cases into deciles of risk and comparing the observed probability with the expected probability within each decile. It is used frequently in risk prediction models. The Hosmer Lemeshow test assesses whether or not the observed event rates match expected event rates in subgroups of the model population. 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.

The important part of this test is the test statistic itself and the significance value. This statistic tests the hypothesis that the observed data are significantly different from the predicted values of the model. So in effect we want a non significant value for this test (because this would indicate that the model does not differ significantly from the observed data).

Drawbacks of Hosmer Lemeshow test-

1) It is a conservative statistic, i.e.its value is smaller than what it ought to be and therefore rejection probability of the null hypothesis is smaller.

2) It has low power in predicting a certain types of lack of fit such as non linearity in explanatory variable.

3) 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.

4)It is more useful when there is more than one predictor and/or continuous predictors in the model.

**For the Hosmer Lemeshow test** we set up the following hypothesis:

H0: The current model fits well

H1: The current model does not fit well

**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 prediction of 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 dependent 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 cancer and smoking status(X –the predictor variable) as: 0=Does not smoke and 1=Smokes

Our regression model will be predicting the logit. That is

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 ,smoking

The output of the logistic regression will be arranged in terms of the blocks that are specified below.

| ***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 Output labeled as variables in the equation tells us the parameters of the model when smoking is the predictor variable. The regression equation is:

ln (ODDS)= -0.182+.770X where (X=Smoking)

The Wald statistic gives the unique contribution of each component of the logistic regression.

The significance value of the Wald statistic for the predictor X indicates acceptance of our null hypothesis that smoking does not significantly predict the possibility of lung cancer. [p value (>0.05) (Wald 0.875, df =1, sig=0.350)].For the constant term (Wald 0.091, df =1, sig=0.763).As p (>0.05) it is also not significant.

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. For non 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.

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 favour 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, if the percentage of smoking goes up by one, then the odds of having cancer increase by 2.160.

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 for non 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 lung cancer as compared to a non-smoker.

SPSS did not produce Hosmer Lemeshow’s goodness of fit test .The reason is that this test can’t be calculated when there is only one predictor and that predictor is a categorical dichotomy. 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 shows us the number of observed cases that are correctly classified by using the single predictor smoking in the model and the overall percentage of the cases that are predicted correctly 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) was observed.This is known as the ***sensitivity of prediction,*** the P (correct |event did occur), that is, 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, in fact, 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, in fact, 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%.

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’t add predictor variables with a simple Pearson Chi square

whereas in logistic regression we can add additional predictor variables, and those additional predictor variables can be either categorical or continuous.

NOTE-To summarize the ***relationship between two categorical variables, we use a cross-tabulation (also called a contingency table).*** Crosstabs' statistics and measures of association are computed for two-way tables only.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

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 |

*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 The values in column 5 are the predicted groups with corresponding predicted 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%. Therefore the values obtained provide strong evidence for the role of smoking as a potential cause of lung cancer. Hence we have predicted the possibility of getting lung 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.

The output of the logistic regression will be arranged in terms of the blocks that are specified below.

| ***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 current model does not differ significantly from the observed data.

| ***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 categorises, 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), that is, 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**

The Wald statistic gives the unique contribution of each component of the logistic regression.

The significance value of the Wald statistic (which is greater than 0.05) for all the predictors indicates acceptance of our null hypothesis that none of the predictors significantly predict the possibility of lung cancer.

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.As exp b is greater than one the odds of having type I skull increase as the percentage of measure 5 goes up by one.

| ***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=π̂I

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

We have (171, 134, 130, 69, 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 of speech 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.

The output of the logistic regression will be arranged in terms of the blocks that are specified below.

| **Table1.12-Hosmer and Lemeshow Test** | | | |
| --- | --- | --- | --- |
| 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 current model differs significantly from the observed data. A significant 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 | | | | | |

*. 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 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. 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, 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 the contribution of these predictors is not significant to the regression model. The values of Exp (B) for X1 indicates that if the percentage of X1 goes up by one, then the odds of loosing decrease (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 X6 goes up by one, then the odds of losing the election increase.

| **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.*