Logistic Regression Model Report:

Hepatitis Data Set

YoonSeok Kim (212752473)

MATH 4330: Applied Categorical Data Analysis

Dr. Steven Wang

December 19, 2016

Introduction :

For this research paper, we will be studying the effects of various symptoms and tests which have the possibility to affect a patient's mortality due to hepatitis. We will be testing whether these effects can be used in a model to determine the probability of a given patient dying. There were 20 attributes in total in the data set; 19 independent variables and 1 dependent variable. Independent variables include both categorical as well as continuous variables. The response variable is categorical and represents whether the patient lives or dies. Statistical tests such as a correlation matrix, logistic regression modeling, and forward selection were used to test the hypothesis: determining which variables are significant and would be best to include in the model. Analysis between the independent variables (categorical and continuous) and their relation to the response variable was done through statistical tests in order to determine the best model to fit the data. In conclusion, our final model was defined as:

$$\text{Log}\left(\frac{\text{Prob (Patient Dies: LD=0)}}{\text{Prob (Patient Lives: LD=1)}}\right) = 0.9163\text{-}1.6502(\text{asc})\text{-}1.9108(\text{spid})(\text{asc})\text{-}0.2231(\text{spid})$$

Dataset Description :

Hepatitis is defined to be the inflammation of the liver. Hepatitis can be caused by viruses and can lead to serious health consequences. There are several different forms of the virus, including types A, B, C, D, E, and G.[1] Data retrieved from the UCI Machine Learning Repository: Hepatitis Data Set, donated by G. Gong from Carnegie-Mellon University 1988, studies how a variety independent symptoms and tests can affect a patient's mortality. In the data set, there is 155 observations and 20 variables in total. There are two types of independent

[1] *Public Health Agency of Canada*. Hepatitis, 2015, http://www.phac-aspc.gc.ca/hep/index-eng.php

variables; categorical and continuous. Categorical variables take the value zero or one, whereas continuous variables take the value of a rational number. There are 14 categorical variables as follows: mortality (which is the response variable), sex, steroid, antivirals, fatigue, malaise, anorexia, big liver, firm liver, spleen palpable, spiders, ascites, varices, and histology. There are six continuous variables which are age, bilirubin, alkaline phosphate, serum glutamic oxaloacetic transaminase, albumin, and prothrombin time. See table 1 below for important variable definitions:

Variable	Value in SAS	Definition
Live or Die (LD): Response Variable	LD=1, Patient Lives LD=0, Patient Dies	Whether the patient lives or dies.
Age	Numerical	The patient's age.
Antivirals (antiv)	Antiv=0 if No, Antiv=1 if Yes	Class of medication for viral infections which could impact the severity of Hepatitis.
Liver Big (livb)	Livb=0 if No, Livb=1 if Yes	Whether or not the liver is enlarged/ inflamed.
Liver Firm (livf)	Livf=0 if No, Livf=1 if Yes	Whether or not the liver is firm to touch.
Spleen Palpable (spleen)	Spleen=0 if No, Spleen=1 if Yes	The Spleen is located near the liver. If the liver is enlarged/ inflamed, it could affect whether the spleen is able to be felt.
Spiders (spid)	Spid=0 if No, Spid=1 if Yes	Collection of small blood vessels clustered very close to the surface of the skin.
Ascites (asc)	Asc=0 if No, Asc=1 if Yes	Accumulation of fluid in the abdominal region, leading to abdominal swelling

Table 1

METHOD :

Among the six continuous variables, protime was excluded from our data set as 43% of the observations were missing. We believed this would make it difficult to adequately test the true effects of protime on the response variable, therefore it was removed.

The correlation matrix was used to decide whether or not an interaction term should be added into the model at the beginning. Following this, forward selection was run with two different α_{entry} levels (0.1 and 0.05). This allowed us to determine which independent variables should be included in the final model as well as analyzing the SAS outputs to determine which α_{entry} level is best to use. Furthermore, the Likelihood Ratio Test was used to decide which model from forward selection best fit the data. Interaction between the significant variables found using forward selection was tested to determine if the interaction term resulted in better predictions. Analysis of the classification table, which includes percentages false positives and false negatives from the models, is then used to determine which model best fits the data set. Due to nature of the model, data, underlying probabilities defined by false negatives or false positives, it is in our best interest as well as the best interest of the patients to minimize the percentage of false negatives.

MODEL CREATION AND ANALYSIS :

A Correlation Matrix was used to determine if there should be any interaction variables at the beginning of the selection process. The SAS output generated results having no correlation coefficient greater than the absolute value of 0.59; therefore multicollinearity was not an issue. Because of this fact, there were no interaction term added at the beginning of model selection.

Forward selection was used to decide which independent variables should be included in the final model. Two α_{entry} levels were tested, 0.10 and 0.05, for the selection process. In order to determine which α_{entry} level would give us the most accurate/ significant variables in the model, we used the Likelihood Ratio Test. Table 2 below outlines the SAS output.

For α_{entry}=0.1:

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	103.992	77.734
SC	106.710	88.608
-2 Log L	101.992	69.734

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.2493	0.6442	12.1928	0.0005
spid	0	1	1.5261	0.6563	5.4075	0.0201
asc	0	1	2.0139	0.7447	7.3138	0.0068
hist	0	1	-1.3112	0.7512	3.0466	0.0809

For α_{entry}=0.05:

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	103.992	79.041
SC	106.710	87.196
-2 Log L	101.992	73.041

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.0428	0.5287	33.1282	<.0001
spid	0	1	1.8708	0.6252	8.9546	0.0028
asc	0	1	2.5179	0.7087	12.6240	0.0004

Table 2

For α_{entry}=0.1, -2logL= 69.734 and the selection process identified three significant independent variables; *spid*, *asc*, and *hist*. For α_{entry}=0.05, -2logL= 73.041 and the selection process identified two significant independent variables; *spid* and *asc*.

The 'Complex Model' from α_{entry}=0.1 is given by:

$$\mathbf{L1} = -2.2493+1.5261(spid)+2.0139(asc)-1.3112(hist)$$

The 'Simple Model' from α_{entry}=0.05 is given by:

$$\mathbf{L0} = -3.0428+1.8708(spid)+2.5179(asc)$$

The Likelihood Ratio Test found below allows us to determine whether the coefficient of *hist*=0.

$$-2 \text{ Log LRT} = -2\text{Log L0} - (-2\text{Log L1}) = 73.041-69.734 = 3.307.$$

Since 3.307 is not significant for a Chi-Squared distribution with *df=1*, we cannot reject that the coefficient of *hist*=0. As a result, we found our 'Simple Model: L0' from α_{entry}=0.05 to be the best model to use.

Next, we analyzed the results of a model which included the interaction term of the variables *spid* and *asc*. The quadratic terms of these variables are not needed as they are both categorical variables which take on the values 0 or 1 so, the quadratic terms would give the same variable values. When running a forward selection process with α_{entry}=0.05, the results identified the interaction variable *spid* x *asc*= *spidasc* as significant.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.5115	0.7176	23.9471	<.0001
asc	0	1	1.5529	0.6955	4.9850	0.0256
spidasc	0	1	2.5465	0.8292	9.4314	0.0021

Figure 1

We can therefore hypothesis that our final model should include the term *spidasc.*

Classification tables were used to compare results of false negatives and false positives for our models. For this data set False Positive= Probability [model says patient dies | Patient lives] and False Negative= Probability [model says patient lives | Patient dies]. As previously stated, it is in the best interest of the Patients to minimize false negatives in order to prevent giving the Patient and their family a false hope of survival. The findings for each model are outlined below in table 3.

Model Variables	Cutoff Point	% Correct	% of False Pos.	% of False Neg.
Spid, Asc	0.500	83.3	28.6	15.4
Asc, SpidxAsc	0.500	84.7	30.0	13.1
Spid, Asc, SpidxAsc	0.500	84.7	30.0	13.1

Table 3

As it can be seen, the models with the interaction term have a higher percentage correct, lower percentage of false negatives (which is desired), and only slightly higher percentage of false positives compared to the model without the interaction term.

CONCLUSION :

As a result, our final model will include the variables *spid*, *asc*, and *spidasc* where all 3 of the remaining variables are categorical. Through SAS, the final model was determined to be:

$$\text{Log}\left(\frac{\text{Prob (Patient Dies: LD=0)}}{\text{Prob (Patient Lives: LD=1)}}\right) = 0.9163-1.6502(asc)-1.9108(spid)(asc)-0.2231(spid)$$

It can be seen that the coefficients of this model are significantly different from the model SAS found through forward selection when the interaction term was added (Figure 1). We believe this is because SAS found only *asc* and *spidasc* to be significant. When we had SAS keep the *spid* term, the coefficients drastically changed to accommodate for the variable that forward selection did not keep and did not find significant in the model. The model still however gave a higher percentage correct and a lower percentage of false negatives (Table 3) compared to the model with only *spid* and *asc* which makes this model the best fit for the data.

Dataset Retrieved from:

http://mlr.cs.umass.edu/ml/datasets/Hepatitis

Code Appendix:

```
'Import Excel file as: 'hepatitis' into SAS';

proc print data=hepatitis;
run;
quit;

*take out protime variable since 67/155 are missing values;
data hep2;
set hepatitis;
drop prot;
run;
quit;
proc print data=hep2;
run;
quit;

* Correlation Matrix ;
proc corr data=hep2
rank;
run;
quit;

title 'Frequncy tables';
proc freq data=hep2;
tables ld sex steoid antiv fatigue malai anor livb livf spleen spid asc vari
hist;
run;

title 'Correlation of continuous variables';
proc corr data=hep2
plots=matrix(histogram)
rank;
var age bili alk sgot albu ;
run;

Title 'Forward selection for logistic regression model testing alpha=0.05 or
0.1';
proc logistic data=hep2;
class sex steoid antiv fatigue malai anor livb livf spleen spid asc vari hist
/ param=ref;
model LD = sex steoid antiv fatigue malai anor livb livf spleen spid
asc vari hist age bili alk sgot albu
/ selection=forward
slentry=0.1
details;
run;

proc logistic data=hep2;
class sex steoid antiv fatigue malai anor livb livf spleen spid asc vari hist
```

```
/ param=ref;
model LD = sex steoid antiv fatigue malai anor livb livf spleen spid
asc vari hist age bili alk sgot albu
/ selection=forward
slentry=0.05
details ctable;
run;

*Creating new data set to test spid*asc;
title 'Data set with spid*asc';
data hep3;
set hep2;
spidasc=spid*asc;
run;
proc print data=hep3;
run;
quit;

Title 'Forward selection for logistic regression model-with spid*asc';
proc logistic data=hep3;
class sex steoid antiv fatigue malai anor livb livf spleen spid asc vari hist
spidasc
/ param=ref;
model LD = sex steoid antiv fatigue malai anor livb livf spleen spid
asc vari hist age bili alk sgot albu spidasc
/ selection=forward
slentry=0.05
details ctable;
run;

Title 'table with cutoff probability=0.5 for our final model: spid, asc,
spid*asc';
proc logistic data=hep3;
model LD= asc spidasc spid/
ctable pprob=0.5 ;
run;
```