HBAT Logistic Regression (August 2019)

Jason Huggy, Student, Lewis University

I. INTRODUCTION

The purpose of this study was to utilize multiple logistic regression to analyze the HBAT dataset found in *Multivariate Data Analysis* [1]. The data set includes 99 lines of unlabeled data, and 23 variables. Logistic regression analyzes the relationship between multiple independent variables and a non-metric dependent variable. Logistic regression can only be used for binary classification; therefore, in this study there are only two possible outcomes for a predicted value. In logistic regression, the primary purpose is to build a regression model that can be used to assess the probability of an observation to belong to one of two categories. Logistic regression uses the logistic coefficient in order to build the model, which is very similar to multiple regression. In this study, SAS Enterprise Guide (EG) is used to predict all outcomes and to analyze the relationship between variables.

II. METHODOLOGY

To conduct this study, SAS Enterprise Guide is used to produce all data tables. To begin, the HBAT data is imported into EG and verified to be free of missing values or errors. A quick examination of the data shows that the first five columns of data, as well as the last, are nonmetric features. The rest of the features are all metric. For this study variable 6 through 18 are used for the base model, and then a select few independent variables are used in the stepwise estimation model. The dependent variable is a classification variable labeled as either 0 or 1 and comes from variable 4 of the HBAT data.

A. Splitting the Data

In order to conduct this study, the HBAT data was split into a training set and a testing set. The training set contains 60 percent of the data, while the testing set contains the other 40 percent. This allows the initial model to be built with the training set, which will determine the best fit for a logistic regression model. The testing set is then applied to validate the effectiveness of the model.

B. Buidling the Logistic Regression Models

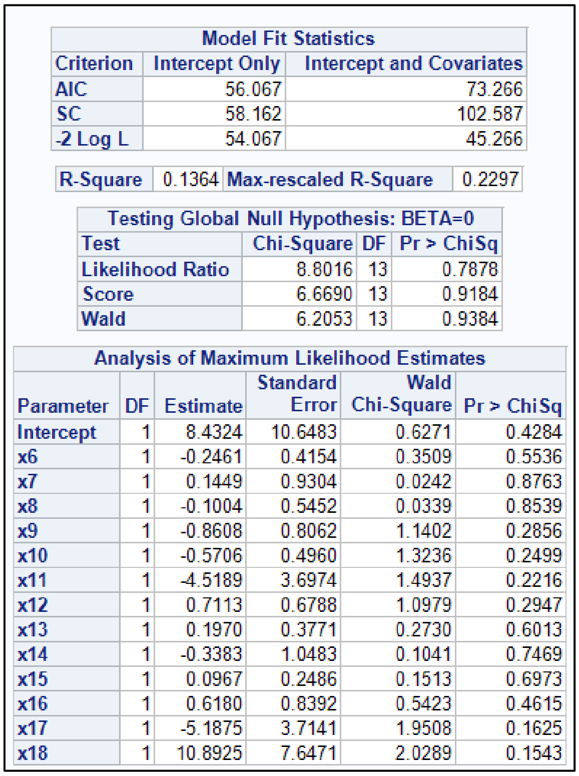
To build the logistic regression model in EG, the *Logistic Regression* function is used. Variables 6 through 18 are selected as the independent variables, and variable 4 is selected as the dependent variable. To begin, the base method is used to build a logistic regression model. This model uses all the independent variables listed above. The training data, which contains 60 percent of the observations, is used to train the model and determine the logistic coefficients to be used to predict the test values. The testing set is selected as the data for the program to predict once the model is built.

The process is then repeated but using the stepwise estimation process. This process selects the independent variable with the greatest level of significance and adds it to the model. Typically, this is repeated until the goodness of fit of the model is no longer improved. The goodness of fit of the model can be measured using the -2LL value (the log likelihood value). However, in SAS Enterprise Guide, each variable is added based on its level of significance. To get the model to go two steps, for the purpose of analyzing the results, the p-value had to be changed to 0.25 otherwise EG couldn’t even run stepwise estimation.

III. RESULTS AND DISCUSSION

After running the logistic regression function in EG, the results are analyzed. The first thing to notice is that the training data contained data with mostly zeros as their classification. 50 observations were zeros while only 10 were ones. This is a problem because it doesn’t allow the model to get enough exposure to cases marked as ones. Problems like this make it harder for the model to distinguish a difference between the two categories because there is mostly data to support the zeros but not the ones. This ends up affecting the entire study as far as accuracy.

For the initial logistic regression model, where all variables are used, the results were not very good. As can be seen in Table 1, there was not a single variable in the training data with a level of significance less than 0.05. The -2LL is not horrible though, being 45.266. The lower the -2LL is, the greater the goodness of fit. The pseudo R-squared is only 0.22, which means the model only explains about 22 percent of the variation in the data. While this is a good starting point, stepwise estimation is used to generate the logistic regression model in hopes of improving the model.



Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	56.067	73.266
SC	58.162	102.587
-2 Log L	54.067	45.266

R-Square	0.1364	Max-rescaled R-Square	0.2297

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.8016	13	0.7878
Score	6.6690	13	0.9184
Wald	6.2053	13	0.9384

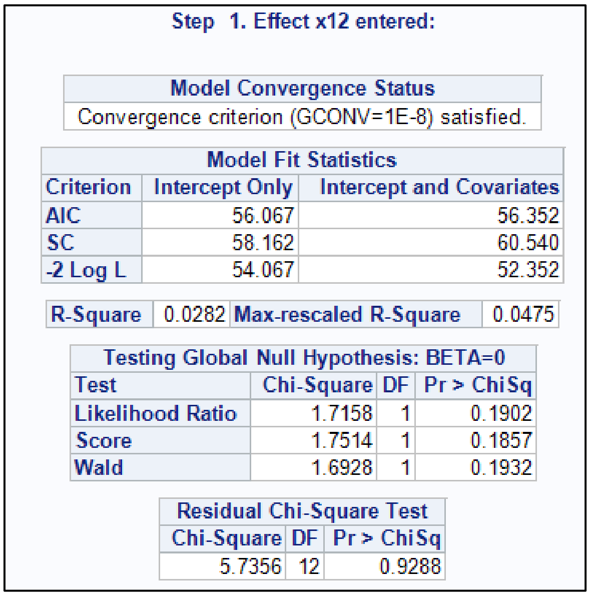
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	8.4324	10.6483	0.6271	0.4284
x6	1	-0.2461	0.4154	0.3509	0.5536
x7	1	0.1449	0.9304	0.0242	0.8763
x8	1	-0.1004	0.5452	0.0339	0.8539
x9	1	-0.8608	0.8062	1.1402	0.2856
x10	1	-0.5706	0.4960	1.3236	0.2499
x11	1	-4.5189	3.6974	1.4937	0.2216
x12	1	0.7113	0.6788	1.0979	0.2947
x13	1	0.1970	0.3771	0.2730	0.6013
x14	1	-0.3383	1.0483	0.1041	0.7469
x15	1	0.0967	0.2486	0.1513	0.6973
x16	1	0.6180	0.8392	0.5423	0.4615
x17	1	-5.1875	3.7141	1.9508	0.1625
x18	1	10.8925	7.6471	2.0289	0.1543

Table 1. Results from base Logistic Regression Model.

As stated, none of the variables used in the training data carry much significance; therefore, the significance to be allowed into the model during stepwise estimation was raised to 0.25 in order to still facilitate the process. Ideally, I would only choose variables with a level of significance of 0.05 or less, or variables that can lower the -2LL. However, this is not possible with this set of training data. Therefore, the study is continued knowing the results will not be as good.

Table 2 shows the results after one step of the stepwise estimation method. Variable 12 is the first variable to be added because its level of significance was less than 0.25. As can be seen, the results compared to the base model, actually get worse. The -2LL goes up to 52.352 and the pseudo R-squared goes down to 4 percent. This step alone is not going to be good enough, so the process continues.



Step 1. Effect x12 entered:

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	56.067	56.352
SC	58.162	60.540
-2 Log L	54.067	52.352

R-Square	0.0282	Max-rescaled R-Square	0.0475

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1.7158	1	0.1902
Score	1.7514	1	0.1857
Wald	1.6928	1	0.1932

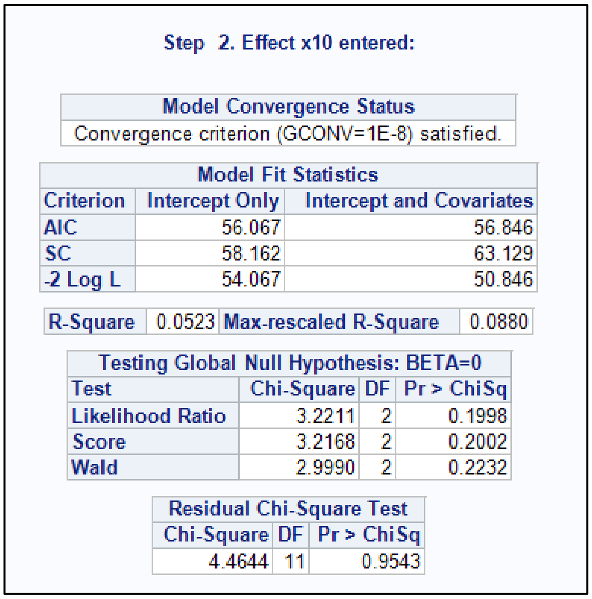
Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
5.7356	12	0.9288

Table 2. Results from the first step of the stepwise estimation process. Variable 12 is added first.

Table 3 shows the results after the second, and final, step of the stepwise estimation method. Variable 10 is added because its level of significance was less than 0.25, but it is higher than variable 12. In this step both the -2LL and pseudo R-squared improve, but not by much. The -2LL lowers to 50.846. Though these variables were the only two variables with a significance under 0.25, they do not do a great job of creating a good logistic regression model, because the model only makes one correct prediction on the test data.

To see what would happen, I continued the stepwise method manually. In the end, the best pseudo R-squared and -2LL results from using all 13 variables. The only thing is that most of the variables carry little weight in the final model. Therefore, while the model is created, there is little meaning to gain from this study.



Step 2. Effect x10 entered:

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	56.067	56.846
SC	58.162	63.129
-2 Log L	54.067	50.846

R-Square	0.0523	Max-rescaled R-Square	0.0880

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3.2211	2	0.1998
Score	3.2168	2	0.2002
Wald	2.9990	2	0.2232

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
4.4644	11	0.9543

Table 3. Results from the second step of the stepwise estimation process. Variable 10 is added.

IV. CONCLUSION

While using the base model results in a -2LL of 45.26, the model still suffers when trying to predict new values because of the initial sample size. In the end, the training data didn’t allow for the model to include the variation it needed. This also made it so that none of the independent variables displayed a very high significance; therefore, it is difficult to assess the impact each variable has on the dependent. In the future, a larger sample should be used that shows more variability. Each category needs to have more observations in order for the model to be trained well. Overall, logistic regression is a great method to utilize for binary classification; however, it did not prove successful in this study.

REFERENCES

[1] J. Hair, W. Black, B. Babin, R. Anderson, *Multivariate Data Analysis: Pearson New International Edition*. 2013.