**Assignment #6**

Daniel S Prusinski

**Introduction:**

A weakness in regression analysis is the tendency to build models that over-fit the data. Cross validation is a technique that splits the data and allows one to test the regression model on data that has not been associated with building the model. In this assignment, cross-validation will be utilized to assess the best multiple variable logistic regression model. Techniques such as backward selection, assessing goodness-of-fit, lift charts, and the KS test statistic all aid in selecting the best model.

**In-Sample Results:**

Throughout this assignment, two models will be compared. The first model is chosen based on management’s decision and will be called Model 1. The second model, Model 2, is based on a statistical technique that analyzes all the variables in the model and chooses the best variables based on a p-value set at the user’s desire, which is .05 for this exercise. The data being used to formulate the models are comprised of 70% of the total data. The output from running this procedure can be seen below.

| **Summary of Backward Elimination** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Step** | **Effect Removed** | **DF** | **Number In** | **Wald Chi-Square** | **Pr > ChiSq** |
| **1** | **A8** | 1 | 16 | 0.0011 | 0.9733 |
| **2** | **A6\_q** | 1 | 15 | 0.0427 | 0.8364 |
| **3** | **A7\_h** | 1 | 14 | 0.1323 | 0.7161 |
| **4** | **A3** | 1 | 13 | 0.3335 | 0.5636 |
| **5** | **A10\_t** | 1 | 12 | 0.3828 | 0.5361 |
| **6** | **A7\_v** | 1 | 11 | 0.4865 | 0.4855 |
| **7** | **A6\_k** | 1 | 10 | 0.6272 | 0.4284 |
| **8** | **A1\_a** | 1 | 9 | 0.8127 | 0.3673 |
| **9** | **A7\_bb** | 1 | 8 | 1.0080 | 0.3154 |
| **10** | **A6\_w** | 1 | 7 | 1.9597 | 0.1616 |
| **11** | **A12\_t** | 1 | 6 | 1.6777 | 0.1952 |
| **12** | **A2** | 1 | 5 | 1.9496 | 0.1626 |

| **Analysis of Maximum Likelihood Estimates**  **Variables that are staying the model.** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -4.0846 | 0.5365 | 57.9692 | <.0001 |
| **A11** | 1 | 0.2291 | 0.0680 | 11.3368 | 0.0008 |
| **A15** | 1 | 0.000597 | 0.000224 | 7.1232 | 0.0076 |
| **A4\_u** | 1 | 0.8924 | 0.4054 | 4.8450 | 0.0277 |
| **A7\_ff** | 1 | -2.1065 | 0.9038 | 5.4325 | 0.0198 |
| **A9\_t** | 1 | 4.0210 | 0.4378 | 84.3549 | <.0001 |

From the results above, it can be seen that five variables had a p-value less than .05, and as a result will be the variables in the second model. From the past EDA in assignment five, I chose A11 as my optimal model and seeing that A11 and A9\_t are both in this model my conclusion is that this model will be very strong.

The next step in assessing the two models is to compare the inferential statistics between the two models. Below is the “Model Fit Statistic”.

| **Model Fit Statistics Model 1** | | |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 620.703 | 340.739 |
| **SC** | 624.812 | 357.176 |
| **-2 Log L** | 618.703 | 332.739 |

| **Model Fit Statistics Model 2** | | |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 620.703 | 289.616 |
| **SC** | 624.812 | 314.271 |
| **-2 Log L** | 618.703 | 277.616 |

This output compares how well the models fit the data. A high -2 Log L value equates to a worse fit. It is assumed that Model 2 will fit better given the fact it has higher covariates, but the AIC and SC penalize a model for having more covariates. Interestingly, Model 2 has a lower value for each criterion than Model 1. The Global Null hypothesis tests that all the explanatory variables have coefficients equal to zero. It can be seen that both variables have at least one coefficient that does not equal zero. Both models also have a significant p-value. Model 2 has much higher scores, and its coefficients are likely to be more form fitting on the data. This could lead to over-fitting which will be analyzed later.

| **Model 1 Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 285.9640 | 3 | <.0001 |
| **Score** | 246.5494 | 3 | <.0001 |
| **Wald** | 151.7473 | 3 | <.0001 |

| **Model 2 Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 341.0870 | 5 | <.0001 |
| **Score** | 267.3283 | 5 | <.0001 |
| **Wald** | 131.5481 | 5 | <.0001 |

The maximum likelihood analysis pared with the odds ratio estimates reveal statistically significant individual coefficients and their prospective magnitudes. In Model 1, A2 and A3 lack the statistical significance at the .05 threshold, and I would alert this point to management. Model 2 only has one variable that is not statistically significant, but it is rather close to the .05 threshold. In order to better understand the magnitude of the coefficients, interpreting the odds ratio is helpful. The odds ratio of a coefficient communicates that the predicted odds for that coefficient are the Point Estimate times the odds compared to that specific non-coefficient. For example, A9\_t has 53 times the odds of non A9\_t values of being 1. The magnitude for A9\_t in model 1 is huge compared to Model 2. Also, A15 almost has a non-existent coefficient.

| **Model 1 Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -3.6287 | 0.5051 | 51.6051 | <.0001 |
| **A9\_t** | 1 | 3.9836 | 0.3302 | 145.5842 | <.0001 |
| **A2** | 1 | 0.0227 | 0.0127 | 3.1641 | 0.0753 |
| **A3** | 1 | 0.0527 | 0.0314 | 2.8241 | 0.0929 |

| **Model 2 Analysis of Maximum Likelihood Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** | 1 | -3.5542 | 0.4514 | 61.9895 | <.0001 |
| **A11** | 1 | 0.2229 | 0.0607 | 13.5025 | 0.0002 |
| **A15** | 1 | 0.000555 | 0.000207 | 7.1670 | 0.0074 |
| **A4\_u** | 1 | 0.6854 | 0.3706 | 3.4200 | 0.0644 |
| **A7\_ff** | 1 | -2.1243 | 0.8686 | 5.9816 | 0.0145 |
| **A9\_t** | 1 | 3.6107 | 0.3630 | 98.9523 | <.0001 |

| **Model 1 Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **A9\_t** | 53.712 | 28.122 | 102.590 |
| **A2** | 1.023 | 0.998 | 1.049 |
| **A3** | 1.054 | 0.991 | 1.121 |

| **Odds Ratio Estimates** | | | |
| --- | --- | --- | --- |
| **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | |
| **A11** | 1.250 | 1.110 | 1.407 |
| **A15** | 1.001 | 1.000 | 1.001 |
| **A4\_u** | 1.985 | 0.960 | 4.103 |
| **A7\_ff** | 0.120 | 0.022 | 0.656 |
| **A9\_t** | 36.992 | 18.161 | 75.349 |

The goodness-of-fit statistics include the percent concordant, percent discordant, Somer’s D, Gamma, and Tau-a. The output for these statistics are listed below.

| **Model 1 Association of Predicted Probabilities and Observed Responses** | | | |
| --- | --- | --- | --- |
| **Percent Concordant** | 89.1 | **Somers' D** | 0.787 |
| **Percent Discordant** | 10.5 | **Gamma** | 0.790 |
| **Percent Tied** | 0.4 | **Tau-a** | 0.390 |
| **Pairs** | 50049 | **c** | 0.893 |

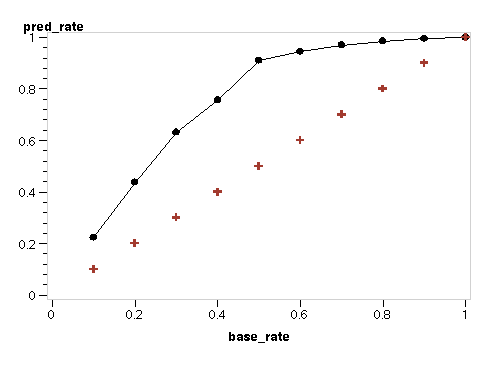
| **Model 2 Association of Predicted Probabilities and Observed Responses** | | | |
| --- | --- | --- | --- |
| **Percent Concordant** | 92.6 | **Somers' D** | 0.871 |
| **Percent Discordant** | 5.5 | **Gamma** | 0.888 |
| **Percent Tied** | 1.9 | **Tau-a** | 0.432 |
| **Pairs** | 50049 | **c** | 0.936 |

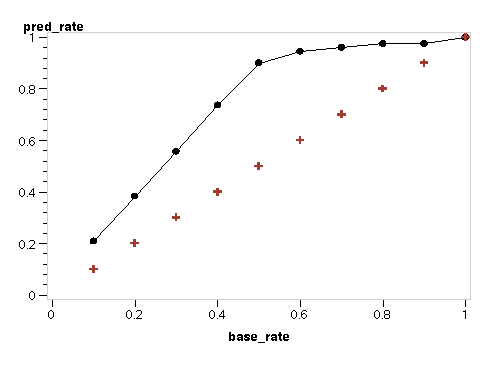
Both models have high percent concordant values. I look forward to analyzing this information on the test data. Model 2 is slightly better, and this comes as no surprise based on the prior analysis.

For Model 1, the lift model reflects what is seen in the lift chart. When targeting 50% of the population, the lift is around 40%. For Model 2, the results are very similar, except the lift is 1 percent greater.

| **Model 2** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs** | **score\_decile** | **Y\_Sum** | **Nobs** | **cum\_obs** | **model\_pred** | **pred\_rate** | **base\_rate** | **life** |
| **1** | 1 | 45 | 45 | 45 | 45 | 0.22388 | 0.1 | 0.12388 |
| **2** | 2 | 43 | 45 | 90 | 88 | 0.43781 | 0.2 | 0.23781 |
| **3** | 3 | 39 | 45 | 135 | 127 | 0.63184 | 0.3 | 0.33184 |
| **4** | 4 | 25 | 31 | 166 | 152 | 0.75622 | 0.4 | 0.35622 |
| **5** | 5 | 31 | 56 | 222 | 183 | 0.91045 | 0.5 | 0.41045 |
| **6** | 6 | 7 | 48 | 270 | 190 | 0.94527 | 0.6 | 0.34527 |
| **7** | 7 | 5 | 44 | 314 | 195 | 0.97015 | 0.7 | 0.27015 |
| **8** | 8 | 3 | 60 | 374 | 198 | 0.98507 | 0.8 | 0.18507 |
| **9** | 9 | 2 | 25 | 399 | 200 | 0.99502 | 0.9 | 0.09502 |
| **10** | 10 | 1 | 51 | 450 | 201 | 1.00000 | 1.0 | 0.00000 |

| **Model 1** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs** | **score\_decile** | **Y\_Sum** | **Nobs** | **cum\_obs** | **model\_pred** | **pred\_rate** | **base\_rate** | **life** |
| **1** | 1 | 42 | 45 | 45 | 42 | 0.20896 | 0.1 | 0.10896 |
| **2** | 2 | 35 | 45 | 90 | 77 | 0.38308 | 0.2 | 0.18308 |
| **3** | 3 | 35 | 45 | 135 | 112 | 0.55721 | 0.3 | 0.25721 |
| **4** | 4 | 36 | 45 | 180 | 148 | 0.73632 | 0.4 | 0.33632 |
| **5** | 5 | 33 | 45 | 225 | 181 | 0.90050 | 0.5 | 0.40050 |
| **6** | 6 | 9 | 45 | 270 | 190 | 0.94527 | 0.6 | 0.34527 |
| **7** | 7 | 3 | 45 | 315 | 193 | 0.96020 | 0.7 | 0.26020 |
| **8** | 8 | 3 | 45 | 360 | 196 | 0.97512 | 0.8 | 0.17512 |
| **9** | 9 | 0 | 45 | 405 | 196 | 0.97512 | 0.9 | 0.07512 |
| **10** | 10 | 5 | 45 | 450 | 201 | 1.00000 | 1.0 | 0.00000 |





The ‘Out-of-Sample Results’ section should contain a lift chart table and graph for each model for the testing data set with a discussion of their predictive accuracy and a recommendation for one model over the other model. The document should be submitted in pdf format.

| **Model 2** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs** | **score\_decile** | **Y\_Sum** | **Nobs** | **cum\_obs** | **model\_pred** | **pred\_rate** | **base\_rate** | **life** |
| **1** | 1 | 19 | 20 | 20 | 19 | 0.20000 | 0.1 | 0.10000 |
| **2** | 2 | 19 | 20 | 40 | 38 | 0.40000 | 0.2 | 0.20000 |
| **3** | 3 | 18 | 21 | 61 | 56 | 0.58947 | 0.3 | 0.28947 |
| **4** | 4 | 14 | 17 | 78 | 70 | 0.73684 | 0.4 | 0.33684 |
| **5** | 5 | 15 | 23 | 101 | 85 | 0.89474 | 0.5 | 0.39474 |
| **6** | 6 | 5 | 21 | 122 | 90 | 0.94737 | 0.6 | 0.34737 |
| **7** | 7 | 0 | 16 | 138 | 90 | 0.94737 | 0.7 | 0.24737 |
| **8** | 8 | 3 | 32 | 170 | 93 | 0.97895 | 0.8 | 0.17895 |
| **9** | 9 | 0 | 12 | 182 | 93 | 0.97895 | 0.9 | 0.07895 |
| **10** | 10 | 2 | 21 | 203 | 95 | 1.00000 | 1.0 | 0.00000 |

| **Model 1** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs** | **score\_decile** | **Y\_Sum** | **Nobs** | **cum\_obs** | **model\_pred** | **pred\_rate** | **base\_rate** | **life** |
| **1** | 1 | 14 | 20 | 20 | 14 | 0.14737 | 0.1 | 0.04737 |
| **2** | 2 | 17 | 20 | 40 | 31 | 0.32632 | 0.2 | 0.12632 |
| **3** | 3 | 15 | 21 | 61 | 46 | 0.48421 | 0.3 | 0.18421 |
| **4** | 4 | 18 | 20 | 81 | 64 | 0.67368 | 0.4 | 0.27368 |
| **5** | 5 | 16 | 20 | 101 | 80 | 0.84211 | 0.5 | 0.34211 |
| **6** | 6 | 10 | 21 | 122 | 90 | 0.94737 | 0.6 | 0.34737 |
| **7** | 7 | 2 | 20 | 142 | 92 | 0.96842 | 0.7 | 0.26842 |
| **8** | 8 | 0 | 21 | 163 | 92 | 0.96842 | 0.8 | 0.16842 |
| **9** | 9 | 1 | 20 | 183 | 93 | 0.97895 | 0.9 | 0.07895 |
| **10** | 10 | 2 | 20 | 203 | 95 | 1.00000 | 1.0 | 0.00000 |

