ABSTRACT

Logistic regression is a popular tool among epidemiologists and other data analysts and while the SAS system is widely used, few investigators make use of the readily available plots that assess model fit. The purpose of this presentation is to demonstrate the steps necessary to generate these plots. We provide a brief overview of logistic regression diagnostics and introduce a sample data set. Using the Output statement in PROC LOGISTIC we show how to obtain the estimated logistic probabilities, change in Pearson chi-square, and Cook distance parameters so simple plots can be generated for assessing the presence of ill-fitting and/or influential observations. The plots provide a graphical approach for assessing model adequacy. While interpretation of the plots is somewhat subjective, these simple plots are a valuable tool for screening out inadequately fitting logistic regression models which may lead to incorrect or misleading inferences.

INTRODUCTION

Examination of data for potential difficulties is a fundamental component of data analysis, yet it is one that is often overlooked. Unusual data points in a logistic regression model can be problematic because they may unduly influence the results of the analysis. They may also be a signal to alert the data analyst that important data items are missing from the regression model.

Hosmer et al. (1991) discuss the need for routine regression diagnostics in logistic regression modelling. The authors present examples of plots that are useful for identification of influential observations and poorly fit observations. Essentially, two plots capture the key elements identified by Hosmer et al. (1991) for assessing the fit of a logistic regression model. These plots: 1) change in Pearson chi-square by estimated logistic probability for identification of observations that fit poorly (outliers) and, 2) Cook distance by estimated logistic probability to identify observations that strongly influence the expressed regression coefficients.

The SAS technical report for the LOGISTIC Procedure
provides descriptions of the regression diagnostics available (SAS Institute, 1990). In order to generate the plots identified by Hosmer et al (1991) an output data set is created using an Output statement. In the Output statement each of the desired diagnostics are specified. PROC PLOT provides the plots of interest.

DEFINITIONS

Estimated logistic probability. This value represents the model estimate for the probability that a given subject has the value of 1 for the dependent variable.

Change in Pearson chi-square. This value is derived for the standardized Pearson residual and represents the amount by which the Pearson chi-square value for the model would change if this individual were deleted from the model.

Cook distance. This value is the counterpart of the value of the same name in linear regression and is defined different ways. The SAS technical report (SAS Institute, 1990) presents two forms one termed C and one termed CBAR. The version Hosmer et al. (1991) prefer is CBAR.

SAMPLE DATA

These data were taken from Appendix 1 of Hosmer and Lemeshow (1989). They were derived from a study of low infant birthweight at Baystate Medical Center, Springfield, MA, during 1986. The authors use this data set to illustrate a variety of points relating to model building strategy as well as regression diagnostics. The fitted logistic regression model we selected is the same one presented in Table 4.10 of Hosmer and Lemeshow (1989, page 101). We chose this model since it is the same model that the authors use in their discussion of regression diagnostics and this will facilitate comparisons between our results and the authors results.

PROGRAM

libname 'a:';
data lbw; set in.lbw;
low2 = 2 - low; *this changes a 1-0 variable to 1-2 variable;
*we create two design variables from an ordinal variable with 3 levels;
race2 = 0; if race = 2 then race2 = 1;
race3 = 0; if race = 3 then race3 = 1;
*we dichotomize as was done in chapter 4;
ptd = 0; if ptl >= 1 then ptd = 1;
lwd = 0; if lwt < 110 then lwd = 1;
*appropriate interaction terms constructed;
age_lwd = age*lwd;
smok_lwd = smoke*lwd;
proc logistic; model low2 = age race2 race3 smoke
ht ui lwd ptd age_lwd
RESULTS

Table 1 shows the output from the model specified above. These results are the same as those obtained by Hosmer and Lemeshow (1989) as presented in table 4.10 of their text.

Figure 1 presents a plot of poorly fitting observations (change in Pearson chi-square by estimated logistic probability). Observations that are located toward the upper aspects of the plot are not explained well by the model.

Figure 2 is a plot of influential observations (Cook distance by estimated logistic probability). Similar to figure 1 the most influential observations are located in the upper right and upper left portions of the plot.

After examination of the plots, we select a value for change in Pearson chi-square which seems "high" and print out the observations which exceed this value.

As can be seen in table 2, there were nine individuals with varying levels of poor fit.

Similar to the approach taken above for poorly fitting observations, a level for Cook distance which seems "high" is selected and these data are examined more closely.

Table 3 presents the data for the four most influential observations. Since the assessment of influence involves both leverage and poor fit, it is not surprising to see that two of the observations have extremely high values for change in Pearson chi-square (11.6 and 9.6). The other two observations involve mothers with relatively low weights (95 and 100 pounds) who gave birth to normal weight infants. The model predicted each of these women to have low-birthweight infants.

CONCLUSION

This brief presentation has not provided the technical detail that is required for a full discussion of regression diagnostics in the logistic model. The present focus has been on two simple plots that offer valuable information for data analysts involved with logistic regression. In conjunction with the references cited, these plots can supplement a discussion of logistic regression diagnostics.
REFERENCES


SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. * indicates USA registration.

Table 1. Maximum Likelihood Estimates and Standard Errors (SE).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>-0.51</td>
<td>1.09</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.08</td>
<td>0.05</td>
</tr>
<tr>
<td>RACE2</td>
<td>1.08</td>
<td>0.52</td>
</tr>
<tr>
<td>RACE3</td>
<td>0.76</td>
<td>0.46</td>
</tr>
<tr>
<td>SMOKE</td>
<td>1.15</td>
<td>0.46</td>
</tr>
<tr>
<td>HT</td>
<td>1.36</td>
<td>0.66</td>
</tr>
<tr>
<td>UI</td>
<td>0.73</td>
<td>0.48</td>
</tr>
<tr>
<td>LWD</td>
<td>-1.73</td>
<td>1.87</td>
</tr>
<tr>
<td>PTD</td>
<td>1.23</td>
<td>0.47</td>
</tr>
<tr>
<td>AGE_LWD</td>
<td>0.15</td>
<td>0.08</td>
</tr>
<tr>
<td>SMOK_LWD</td>
<td>-1.41</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Table 2. Output of observations with poor fit.

<table>
<thead>
<tr>
<th>ID</th>
<th>LOW</th>
<th>AGE</th>
<th>LWT</th>
<th>RACE</th>
<th>HT</th>
<th>DIFCHI</th>
</tr>
</thead>
<tbody>
<tr>
<td>98</td>
<td>0</td>
<td>22</td>
<td>95</td>
<td>3</td>
<td>1</td>
<td>4.0</td>
</tr>
<tr>
<td>99</td>
<td>0</td>
<td>30</td>
<td>107</td>
<td>3</td>
<td>0</td>
<td>11.6</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>29</td>
<td>130</td>
<td>1</td>
<td>0</td>
<td>9.6</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>27</td>
<td>150</td>
<td>3</td>
<td>0</td>
<td>7.7</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>25</td>
<td>115</td>
<td>3</td>
<td>0</td>
<td>6.5</td>
</tr>
<tr>
<td>36</td>
<td>1</td>
<td>24</td>
<td>138</td>
<td>1</td>
<td>0</td>
<td>12.7</td>
</tr>
<tr>
<td>57</td>
<td>1</td>
<td>15</td>
<td>110</td>
<td>1</td>
<td>0</td>
<td>6.1</td>
</tr>
<tr>
<td>63</td>
<td>1</td>
<td>23</td>
<td>120</td>
<td>3</td>
<td>0</td>
<td>5.5</td>
</tr>
<tr>
<td>77</td>
<td>1</td>
<td>26</td>
<td>190</td>
<td>1</td>
<td>0</td>
<td>4.8</td>
</tr>
</tbody>
</table>

Table 3. Output of influential observations.

<table>
<thead>
<tr>
<th>ID</th>
<th>LOW</th>
<th>AGE</th>
<th>LWT</th>
<th>RACE</th>
<th>HT</th>
<th>CBAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>98</td>
<td>0</td>
<td>22</td>
<td>95</td>
<td>3</td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>99</td>
<td>0</td>
<td>30</td>
<td>107</td>
<td>3</td>
<td>0</td>
<td>0.8</td>
</tr>
<tr>
<td>107</td>
<td>0</td>
<td>31</td>
<td>100</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>29</td>
<td>130</td>
<td>1</td>
<td>0</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Figure 1. Plot of DIFCHI*PREDE. Legend: A = 1 obs, B = 2 obs, etc.
Figure 2. Plot of CBAR*PRED. Legend: A = 1 obs, B = 2 obs, etc.