Introduction

Similar to traditional linear regression, logistic regression is a method for relating a set of independent predictor variables to a dependent outcome variable. In logistic regression, however, the dependent variable is binary \{0,1\} rather than continuous. The beta coefficients take on special meaning in terms of the logarithm of the odds (logit) and odds ratios for whether or not an event occurs.

Logistic Coefficients

Let \( Y_j \) be an indicator variable equal to one when an event occurs and zero otherwise. Then \( P[E | X_{1j}, X_{2j}, ... , X_{k-1,j}; \beta_0, \beta_1, ... , \beta_{k-1,j}] \) represents the probability of the event \( E \) given the vector set \( X \) and \( \beta \). Here, \( X_j \) represents the independent variable for the \( j \)th observation, \( \hat{\beta}_0 \) the intercept parameter, and \( \hat{\beta}_1, ... , \hat{\beta}_{k-1} \) the set of \( k-1 \) regression coefficients that maximize the function,

\[
\sum Y_j (\beta_0 + \beta_1 X_{1j} + ... + \beta_{k-1,j} X_{k-1,j})
\]

\[-\sum \ln (1 + \exp(\beta_0 + \beta_1 X_{1j} + ... + \beta_{k-1,j} X_{k-1,j})) \]

In terms of the logistic model \((k=2), \{x=0,1\}\), if

\[
P[E | X; \hat{\beta}] = \left[1 + \exp(-x\hat{\beta})\right]^{-1}
\]
y = \begin{cases} \text{Event (E+1)} \\ \text{Non-Event} \end{cases}

Then,

$$\psi_\xi = \frac{\text{Odds} \{ x=1 \}}{\text{Odds} \{ x=0 \}} = \frac{\hat{P}_\xi [E \mid \{ x=1 \}] / \hat{P}_\xi [\bar{E} \mid \{ x=1 \}]}{\hat{P}_\xi [E \mid \{ x=0 \}] / \hat{P}_\xi [\bar{E} \mid \{ x=0 \}]}$$

$$= \frac{[1+\exp(-\beta_0-\beta_\xi)]^{-1}}{[1+\exp(-\beta_0)]^{-1}} \frac{\hat{\beta}_\xi^{-1}}{1-[1+\exp(-\beta_0-\beta_\xi)]}$$

$$= \exp(\hat{\beta}_\xi)$$

which is the logarithm of the odds ratio. When other covariates are present in the model, the odds ratio holds controlling for these additional factors. Further,

$$\log \frac{[1+\exp(-X \hat{\beta})]^{-1}}{1-[1+\exp(-X \hat{\beta})]} = X \hat{\beta}$$

implying that \( \hat{\beta}_0 \) estimates the log odds of an event given the set of independent variables \( X = 0 \), while \( \hat{\beta}_{k-1} \) measures the increase in risk per unit change in \( X_{k-1} \). An estimate for relative risk is given as,

$$\delta_\xi = \frac{\hat{P}_\xi [E \mid \{ x=1 \}]}{\hat{P}_\xi [E \mid \{ x=0 \}]}$$

$$= \frac{1+\exp(-\hat{\beta}_0)}{1+\exp(-\hat{\beta}_0-\hat{\beta}_\xi)}$$
Assessing Model Fit
Due to the interpretability of the coefficients and the fact that the
covariates, X, do not need to have a multivariate normal distribution,
logistic regression is generally preferred over other multivariate
techniques, such as discriminant analysis, when dealing with binary
outcome data. In practice, however, logistic regression is frequently used
inappropriately to analyze data that simply do not fit the logistic model.
Several methods are available to assess model fit. These include diagnostic
plots of the studentized residuals, leverage points and Cook’s distance, in
addition to goodness-of-fit techniques, such as the classical Chi-Square test
and Hosmer-Lemeshow’s adaptation for small samples. When the model
contains at least one continuous covariate (e.g., age, weight, height), the
Efird-David Split-Range test provides an effective non-parametric means
for checking if the model has changed over the duration of the study.
Observing study entrance times, the sample is divided into equally sized
subsets $n_1 = n_2$, with each being fitted to the logistic model. The resulting
$P[E|X; \hat{\beta}]$ values are labelled and ranked. One then notes the position of
the lowest and highest values from the first subset to obtain the range
score. Letting $N=2n$, the rejection regions for the Split-Sample Range Test
are given in Table 1. Given a sufficiently large sample size, such that the
true ranking order holds, a small p-value suggests that the model may be
unstable over time.

<table>
<thead>
<tr>
<th>Sup $R &gt; P[R \leq r] \leq .05$</th>
<th>Sup $R &gt; P[R \leq r] \leq .01$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n \leq 4$, $r = {1}$</td>
<td>$n \leq 5$, $r = {1}$</td>
</tr>
<tr>
<td>$5 \leq n \leq 6$, $r = N - 6$</td>
<td>$n = 6$, $r = N - 7$</td>
</tr>
<tr>
<td>$7 \leq n \leq 26$, $r = N - 7$</td>
<td>$7 \leq n \leq 10$, $r = N - 8$</td>
</tr>
<tr>
<td>$n \geq 27$, $r = N - 8$</td>
<td>$11 \leq n \leq 20$, $r = N - 9$</td>
</tr>
<tr>
<td></td>
<td>$21 \leq n \leq 204$, $r = N - 10$</td>
</tr>
<tr>
<td></td>
<td>$n \geq 205$, $r = N - 11$</td>
</tr>
</tbody>
</table>

* When ties are present the distribution can be determined by
enumeration or simulation.
Determining the Importance of Covariates

One way to determine the importance of a covariate is to examine its effect on model discrimination. This is accomplished by the user choosing a cutoff value \( p \). When \( P[E \mid X; \hat{\beta}] > p \) the event is predicted to occur. Discrimination is expressed by the relationship between sensitivity (proportion of true positives predicted to be positive) and specificity (proportion of true negatives predicted to be negative). When sensitivity is plotted against 1-specificity (ROC curve), models that manifest good discrimination typically are positioned toward the upper left corner and have greater area under the curve. Area under the ROC curve increases when informative prognostic covariates are added to the model and decreases if the additional covariates fail to contribute to overall discrimination. Related measures include the False Positive Rate (proportion of predicted positives that are truly negative), False Negative Rate (proportion of predicted negatives that are truly positive), and the percentage of predicted values classified correctly.

Significance of a covariate can be assessed quantitatively (for large sample sizes) using the Likelihood Ratio test,

\[
G^2_{X_{k-1}} = 2 \left( \ln L_{X_{k-1}} - \ln L_{X_k} \right)
\]

which is twice the logarithm of the likelihood function calculated with all the independent variables, divided by the likelihood calculated with the covariate \( X_{k-1} \) removed.

Population-Adjusted Logistic Regression

When stable population variables (e.g., census track information like age and sex) are available, this information can be incorporated into the logistic model for a sample subset to improve model fit. The method uses Bayesian probability expansions to adjust expected outcome values with respect to one or more population covariates. For example, let \( P[i, kl j] \) be the population proportion in exposure group \( j \) (e.g. smoker vs non smoker) who are in stratum \( i \) (e.g. male vs female) and stratum \( k \) (e.g. young vs. old). Such information is easily obtained via a telephone interview. On the
other hand, a physical examination is needed to determine if an individual has a particular condition such as lung cancer. To minimize study cost, a random sample is taken to determine this information. Let \( \hat{P}(c \mid i, j, k) \) be the logistic model estimated from the sample and \( P(i, k \mid j) \) be the known population probabilities. Then the proportion of smokers with cancer \( P(c \mid j) \) can be estimated by \( \sum \sum P(c, i, k \mid j) \). This is an unbiased estimate as shown by the following argument. Assuming

\[
E[\hat{P}(c \mid i, j, k)] = P(c \mid i, j, k)
\]

then,

\[
E\left[ \sum_i \sum_k P(i, k \mid j) \hat{P}(c \mid i, j, k) \right] = \sum_i \sum_k P(c, i, k \mid j) = P(c \mid j).
\]

The variance of the estimates can be found accordingly. Let \( X \) be an \( 8 \times 4 \) design matrix where the entries in column 1 are ones (intercept), column 2 are the explicit mean ages for the young and old groups, column 3 are 0 if female and 1 if male, and column 4 are 0 if non-smoker and 1 if smoker. Given that \( \hat{\beta} \) is a \( 4 \times 1 \) vector and that \( \hat{\Sigma} \) (the covariance matrix) is a \( 4 \times 4 \) matrix, it follows that the covariance of the linear combination \( X\hat{\beta} = X\hat{\Sigma}X' \). Now take the logistic function element-wise in the above combination \( L_k \),

\[
\hat{P}_{(8 \times 1)}(c) = \left[ \frac{1}{1 + \exp(-L_k)} \right]
\]

and create a diagonal matrix \( D_{(8 \times 8)} \) with the elements

\[
(\hat{P}_k(1-\hat{P}_k)).
\]

Then the covariance of the logistic elements \( \hat{P} \) is
Rewriting the estimates for each stratum combination in matrix form as

\[ \mathbf{W}(4 \times 8) \mathbf{\hat{P}}(8 \times 1) \]

the covariance estimates for the above is given as

\[ \mathbf{WDX\hat{X}DW}. \]

Further study is needed to determine the efficiency of population adjusted estimates.

**Example**

MRI-CBV, a non-invasive imaging technique useful in assessing the progression of brain tumors, is being studied at Harvard Medical School. Preliminary results indicate that this procedure can play an important role in the prognosis of event-free survival, supplementing information obtained from surgical biopsy (Kaplan-Meier, \( p=0.0006 \); see Figures 1A & 1B). In certain cases, however, a biopsy may be infeasible, due to the critical location of the tumor or the general health of the patient. Under these conditions, the physician is interested in determining if the MRI-CBV can provide prognostic information consistent with tumor grade (determined via biopsy). The data under examination consists of 29 patients having scores for both MRI-CBV and tumor grade. The MRI-CBV is classified as low, if the value is either 0, 1, or 2, and as high, if the result is 3 or 4. Low tumor grade corresponds to the values 1 and 2, while high tumor grade includes the values 3 and 4. Fisher's Exact 2-Tail test (see Figure 3 for computer output) shows that the two factors are highly related (\( p=0.00251 \)). Using logistic regression, the estimated odds ratio of Low MRI-CBV : High MRI-CBV with respect to a low tumor grade outcome is 16:1 (CI\(_{95}\) : 2.818, 93.702; \( p=0.0038 \)). Incorporating the factor age (\( \leq 50 \) vs > 50) improves model fit (Likelihood Ratio Test, \( p=0.0017 \)) as illustrated by the increased area under the ROC curve in Figure 2. Controlling for age, the odds ratio for MRI-CBV increases to 22:1 (CI\(_{95}\) : 2.151, 229.102, \( p=0.0092 \)).
The author gratefully acknowledges Laurel Beckett for introducing him to Logistic Regression and Population Projection Techniques.

References


Figure 1B

Overall Survival (Event-Free)
By Grade

Log-Rank p = 0.0008
- - - 1 & 2
----- 3 & 4

Figure 2

ROC Curves W.R.T. Grade
MRI - CBV Univariate Model VS Age Adjusted Model

Likelihood Ratio Test p = 0.0017
Area Increase = 10.7155
Figure 3: Computer Output

### Table of MRI34F by GRAD34F

<table>
<thead>
<tr>
<th>MRI34F</th>
<th>GRAD34F</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-mri012</td>
<td>13</td>
<td>44.83%</td>
<td>6.90</td>
<td>51.72%</td>
<td>86.67%</td>
<td>133.33%</td>
</tr>
<tr>
<td>1-mri34</td>
<td>4</td>
<td>13.79%</td>
<td>34.48%</td>
<td>48.28%</td>
<td>28.57%</td>
<td>71.43%</td>
</tr>
<tr>
<td>2-mri54</td>
<td>10</td>
<td>33.33%</td>
<td>83.33%</td>
<td>10.00%</td>
<td>23.53%</td>
<td>16.67%</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>58.62</td>
<td>41.38</td>
<td>100.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

xtabs of mri34f by grad34f

### Statistics for Table of MRI34F by GRAD34F

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-Square</td>
<td>1</td>
<td>10.076</td>
<td>0.002</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>10.804</td>
<td>0.001</td>
</tr>
<tr>
<td>Continuity Adj. Chi-Square</td>
<td>1</td>
<td>7.823</td>
<td>0.005</td>
</tr>
<tr>
<td>Mantel-Haenszel Chi-Square</td>
<td>1</td>
<td>9.772</td>
<td>0.002</td>
</tr>
<tr>
<td>Fisher's Exact Test (Left)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Right)</td>
<td></td>
<td></td>
<td>2.13E-03</td>
</tr>
<tr>
<td>(2-Tail)</td>
<td></td>
<td></td>
<td>2.51E-03</td>
</tr>
<tr>
<td>Phi Coefficient</td>
<td></td>
<td></td>
<td>0.589</td>
</tr>
<tr>
<td>Contingency Coefficient</td>
<td></td>
<td></td>
<td>0.508</td>
</tr>
<tr>
<td>Cramer's V</td>
<td></td>
<td></td>
<td>0.508</td>
</tr>
</tbody>
</table>

### Estimates of the Common Relative Risk (Row1/Row2)

<table>
<thead>
<tr>
<th>Type of Study</th>
<th>Method</th>
<th>Value</th>
<th>Confidence Bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-Control</td>
<td>Mantel-Haenszel</td>
<td>16.250</td>
<td>2.818-93.702</td>
</tr>
<tr>
<td>(Odds Ratio)</td>
<td>Logit</td>
<td>16.250</td>
<td>2.462-107.242</td>
</tr>
</tbody>
</table>

### Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; Chi-Square</th>
<th>Standardized Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCPT</td>
<td>-0.9163</td>
<td>0.5916</td>
<td>2.3988</td>
<td>0.1214</td>
<td>0.1214</td>
</tr>
<tr>
<td>MRI34L</td>
<td>2.7881</td>
<td>0.9628</td>
<td>8.3863</td>
<td>0.0038</td>
<td>0.781717</td>
</tr>
</tbody>
</table>

### Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; Chi-Square</th>
<th>Standardized Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCPT</td>
<td>-3.3682</td>
<td>1.5821</td>
<td>4.5322</td>
<td>0.0333</td>
<td>0.869226</td>
</tr>
<tr>
<td>MRI34L</td>
<td>3.1302</td>
<td>1.1304</td>
<td>6.7783</td>
<td>0.0092</td>
<td>0.660704</td>
</tr>
<tr>
<td>AGE50L</td>
<td>2.9069</td>
<td>1.5290</td>
<td>3.6146</td>
<td>0.0573</td>
<td>0.660704</td>
</tr>
</tbody>
</table>

NESUG '92 Proceedings