Predictive Modeling

Logistic Regression

• Consider the Multiple Linear Regression Model:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \]

where the response variable \( y \) is dichotomous, taking on only one of two values:

\[ y = 1 \quad \text{if a success} \]
\[ = 0 \quad \text{if a failure} \]
Logistic Regression

• The response variable, \( y \), is really just a Bernoulli trial, with
  \[ E(y) = \pi \]
  where
  \( \pi \) = probability of a success on any given trial
• \( \pi \) can only take on values between 0 and 1

Logistic Regression

• Thus, the Multiple Regression Model
  \[ \pi = \mu_{y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_{ik} \]
  is not appropriate for a dichotomous response variable, since this model assumes \( \pi \) can take on any value, but in fact it can only take on values between 0 and 1.
Logistic Regression

• When the response variable is dichotomous, a more appropriate linear model is the Logistic Regression Model:

\[
\logit(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}
\]

Logistic Regression

• The ratio \( \frac{\pi}{1 - \pi} \) is called the odds, and is a function of the probability \( \pi \)
Logistic Regression

- Consider the model with one predictor (k=1):

\[
\text{Logit} \quad \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x
\]

\[
\text{Odds} \quad \frac{\pi}{1-\pi} = e^{\alpha + \beta x} = e^\alpha \left(e^\beta\right)^x
\]

- Every one unit increase in x increases the odds by a factor of \(e^\beta\)
Logistic Regression

\[ x = 1 \quad \frac{\pi}{1 - \pi} = e^{\alpha (e^{\beta})} \]

\[ x = 2 \quad \frac{\pi}{1 - \pi} = e^{\alpha (e^{\beta})^2} = e^{\alpha (e^{\beta})(e^{\beta})} \]

\[ x = 3 \quad \frac{\pi}{1 - \pi} = e^{\alpha (e^{\beta})^3} = e^{\alpha (e^{\beta})(e^{\beta})(e^{\beta})} \]

Logistic Regression

- Thus, \( e^\beta \) is the odds ratio, comparing the odds at \( x+1 \) with the odds at \( x \)
- An odds ratio equal to 1 (i.e., \( e^\beta = 1 \)) occurs when \( \beta = 0 \), which describes the situation where the predictor \( x \) has no association with the response \( y \).
Logistic Regression

• As with regular linear regression, we obtain a sample of n observations, with each observation measured on all k predictor variables and on the response variable.
• We use these sample data to fit our model and estimate the parameters

\[ p = \frac{e^{a + bx}}{1 + e^{a + bx}} \]

for a single predictor.
Logistic Regression

- The estimate of $\pi$ is
  \[ p = \frac{e^{a+bx}}{1+e^{a+bx}} \]

Example: Coronary Heart Disease

<table>
<thead>
<tr>
<th>Age</th>
<th>Age Group (x)</th>
<th>CHD Present</th>
<th>N</th>
<th>p</th>
<th>p(1-p)</th>
<th>log(p(1-p)) = a+bx</th>
<th>$\hat{p}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>0.10</td>
<td>0.09834</td>
<td>-2.31929</td>
<td>0.08664</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>2</td>
<td>15</td>
<td>0.13</td>
<td>0.17184</td>
<td>-1.76117</td>
<td>0.16864</td>
</tr>
<tr>
<td>35</td>
<td>3</td>
<td>3</td>
<td>12</td>
<td>0.25</td>
<td>0.30028</td>
<td>-1.20365</td>
<td>0.23093</td>
</tr>
<tr>
<td>40</td>
<td>4</td>
<td>4</td>
<td>15</td>
<td>0.33</td>
<td>0.52470</td>
<td>-0.64493</td>
<td>0.34413</td>
</tr>
<tr>
<td>45</td>
<td>5</td>
<td>5</td>
<td>13</td>
<td>0.46</td>
<td>0.91685</td>
<td>-0.08681</td>
<td>0.47831</td>
</tr>
<tr>
<td>50</td>
<td>6</td>
<td>6</td>
<td>13</td>
<td>0.63</td>
<td>1.60209</td>
<td>0.47131</td>
<td>0.61569</td>
</tr>
<tr>
<td>55</td>
<td>7</td>
<td>7</td>
<td>17</td>
<td>0.76</td>
<td>2.79947</td>
<td>1.02943</td>
<td>0.73681</td>
</tr>
<tr>
<td>60</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>0.80</td>
<td>4.89175</td>
<td>1.58755</td>
<td>0.83027</td>
</tr>
</tbody>
</table>

$\text{Exp}(b) = 1.74738 \quad b = 0.55812$

(Multiplicative) (Additive)
Example: Coronary Heart Disease

Parameter Estimation

- A 100(1-\(\alpha\))% Confidence Interval for \(\beta\) is:

\[
\hat{\beta} \pm z_{\alpha/2} \times \text{a.s.e.}(\hat{\beta})
\]

where \(z_{\alpha/2}\) is a critical value from the standard normal distribution, and a.s.e. stands for the asymptotic standard error.
Parameter Estimation

• A 100(1-\(\alpha\))% Confidence Interval for the odds ratio \(e^\beta\) is:

\[ e^{\hat{\beta} \pm z_{\alpha/2} \times \text{a.s.e.}(\hat{\beta})} \]

Hypothesis Testing

\(H_0: \beta = 0\)
\(H_1: \beta \neq 0\)

The test statistics for testing this hypothesis are \(\chi^2\) statistics. There are 3 that are commonly used:
Hypothesis Testing

Wald Test
\[ Q_W = \left( \frac{\hat{\beta}}{\text{a.s.e.}(\hat{\beta})} \right)^2 \sim \chi_1^2 \]

Likelihood Ratio Test
\[ Q_{LR} = -2(L_{\alpha} - L_{\alpha,\beta}) \sim \chi_1^2 \]

Score Test
\[ Q_S \sim \chi_1^2 \]

Goodness of Fit

- Let \( m = \# \) of levels of \( x \) (\( m=8 \) for CHD example)
- Let \( n_i = \) number of observations in the \( i^{th} \) level of \( x \)
- Let \( k = \) number of parameters (\( k=2 \) for CHD example)

\[ X_{Pearson}^2 = \sum_{i=1}^{m} n_i \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} \sim \chi_{m-k}^2 \]

\[ X_{Deviance}^2 = 2 \sum_{i=1}^{m} n_i p_i \log \left( \frac{p_i}{\hat{\pi}_i} \right) \sim \chi_{m-k}^2 \]
Example: Coronary Heart Disease

Logistic Regression Example
Coronary Heart Disease vs. Age
(first 25 observations, out of 100)
SAS Code:
Proc Logistic

Proc Logistic Data=CHD100 Descending;
  Model CHD=AgeGroup;
Run;

Proc Logistic Output: CHD Example

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>28.4851</td>
<td>1</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Score</td>
<td>26.0782</td>
<td>1</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Wald</td>
<td>21.4281</td>
<td>1</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-2.8770</td>
<td>0.6233</td>
<td>21.3024</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>AgeGroup</td>
<td>1</td>
<td>0.5580</td>
<td>0.1206</td>
<td>21.4281</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Odds Ratio Estimates

<table>
<thead>
<tr>
<th>Effect</th>
<th>Point Estimate</th>
<th>95% Wald Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgeGroup</td>
<td>1.747</td>
<td>1.380</td>
</tr>
</tbody>
</table>
### Proc Logistic Output: CHD Example

#### Deviance and Pearson Goodness-of-Fit Statistics

<table>
<thead>
<tr>
<th>Criterion</th>
<th>DF</th>
<th>Value</th>
<th>Value/DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>6</td>
<td>0.2164</td>
<td>0.0361</td>
<td>0.9998</td>
</tr>
<tr>
<td>Pearson</td>
<td>6</td>
<td>0.2178</td>
<td>0.0363</td>
<td>0.9998</td>
</tr>
</tbody>
</table>

Number of unique profiles: 8