Multiple Logistic models with Categorical predictors

1. Logit models with Multiway Contingency Tables
2. AIDS and AZT Example
3. Problem 5.35

Homework: Ch.5, Sec. 4
Problem: 5.37
Multiple Logistic regression

Let us consider the model with multiple explanatory variables

$$x = (x_1, x_2, ..., x_p) \quad \text{if} \quad p \geq 2$$

Assume the response $Y$ is binary and

$$\text{logit}[\pi(x)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p.$$  

$\beta_i$ reflects the effect of $x_i$ on the log odds that $Y=1$ controlling the others $x_j$, $j \neq i$.

Namely, $e^{\beta_i}$ is the multiplicative effect on the odds of a 1-unit increase in $x_i$ at fixed levels of others $x_j$.

Remark #1. An explanatory variables can be qualitative, using dummy variables for categories.
Logit Models for Multiway Contingency Tables

Assume that there are two categories for each explanatory variable $X$ and $Z$ ($p=2$)

$$x_1 = 1$$
$$x_2 = 0 \quad \text{for one factor } X$$

and

$$z_1 = 1$$
$$z_2 = 0 \quad \text{for another factor } Z.$$  

Consider the model

(1) \[ \text{logit } [P(Y=1)] = \alpha + \beta_1 x_i + \beta_2 z_k \]

$x_i \in \{0, 1\}$, $z_k \in \{0, 1\}$.

Note that at a fixed level $z_k$ of $Z$, the effect of changing categories of $X$ is

$$[\alpha + \beta_1 x_i(1) + \beta_2 z_k] - [\alpha + \beta_1 x_i(0) + \beta_2 z_k] = \beta_1$$
i.e. $e^\beta_1$ is the conditional Odds Ratio between $X$ and $Y$ given $Z$:

$$
\frac{\Pr(Y=1 \mid x_i=1, z_k)}{1 - \Pr(Y=1 \mid x_i=1, z_k)} / \frac{\Pr(Y=1 \mid x_i=0, z_k)}{1 - \Pr(Y=1 \mid x_i=0, z_k)} = e^\beta_1
$$

or

$$
\frac{\Pr(Y=1 \mid x_i=1, z_k)}{\Pr(Y=1 \mid x_i=0, z_k)} = e^\beta_1
$$

So that controlling for $Z=z_k$, the Odds of success when $X=1$ equal $e^\beta_1$ times the odds of success when $X=0$.

\[\begin{array}{c|c|c}
  & Y=1 & Y=0 \\
\hline
  X=1 & \Pr(Y=1 \mid x_i=1, z_k) & 1 - \Pr(Y=1 \mid x_i=1, z_k) \\
  X=0 & \Pr(Y=1 \mid x_i=0, z_k) & 1 - \Pr(Y=1 \mid x_i=0, z_k) \\
\end{array}\]

\[\begin{array}{c|c|c}
  & \Pr(Y=1 \mid x_i=1, z_k) & \Pr(Y=1 \mid x_i=0, z_k) \\
\hline
  L_{1|z_k} & \frac{\Pr(Y=1 \mid x_i=1, z_k)}{1 - \Pr(Y=1 \mid x_i=1, z_k)} & \frac{\Pr(Y=1 \mid x_i=0, z_k)}{1 - \Pr(Y=1 \mid x_i=0, z_k)} \\
\end{array}\]
From (a) we conclude that the conditional Odds ratio is the same at each level of \( Z = 1 \) and \( Z = 0 \).

\[ \Theta_{XY}(1) = \Theta_{XY}(0) \]

i.e. \( X \) and \( Y \) have homogeneous \( XY \) association.

\textbf{Remark #2}: This is a consequence of the lack of interaction term in

\[ \log(\frac{P(Y=1 | X, Z)}{1-P(Y=1 | X, Z)}) = \beta_0 + \beta_1 X_i + \beta_2 Z_k \]

\[ x_i = 1, 0, \quad z_k = 0, 1 \]

\textbf{Remark #3}: Under Hypothesis

\[ H_0: \beta_1 = 0 \]

\[ \frac{\sigma_{\text{err}}^2}{\sigma_{\text{err}}^2} = 1 \quad \text{for any } z_k = 1, 0 \]

so that under \( H_0 \), \( \beta_1 = 0 \) \( X \) \& \( Y \mid Z = z_k \) are independent.
Now let us assume that the factor $X$ has $I-1$ dummy variables but factor $Z$ - $K$ categories with $K-1$ dummy variables.

(2) Model: \[
\text{logit} \left[ P(Y=1) \right] = \alpha + \beta_i^X + \beta_k^Z \]

$\beta_i^X, \ldots, \beta_{I-1}^X$ - the effects parameters of $X$

$\beta_1^Z, \ldots, \beta_{K-1}^Z$ - the effects parameters of $Z$.

Testing conditional independence of $X \mid Y$ given $Z$ corresponds to testing

$H_0: \beta_i^X = \beta_1^Z = \ldots = \beta_{I-1}^X = \beta_1^Z$

Note that for each factor one parameter is redundant, we assume there

$\beta_1^X = \beta_k^Z = 0$

Remark #4. When each factor $X$ and $Z$ have two categories $I=2$, $K=2$, then (1) and (2) models are equivalent to each other with

$\beta_2^X = \beta_2^Z = 0$ and $\beta_1^X = \beta_1^Z$, but $\beta_1^X = \beta_1^Z$. 

2. Example 1: AIDS and AZT

Consider the Table 5.5.

338 veterans

whose immune systems were beginning to falter after infection with the AIDS virus were randomly assigned either to receive AZT immediately or to wait until their T cells showed severe immune weakness.

<table>
<thead>
<tr>
<th>Table 5.5:</th>
<th>AIDS Symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>AZT use</td>
<td>Yes</td>
</tr>
<tr>
<td>White</td>
<td></td>
</tr>
<tr>
<td>(x=1) Yes</td>
<td>14</td>
</tr>
<tr>
<td>(x=0) No</td>
<td>32</td>
</tr>
<tr>
<td>Black</td>
<td></td>
</tr>
<tr>
<td>(x=1) Yes</td>
<td>11</td>
</tr>
<tr>
<td>(x=0) No</td>
<td>12</td>
</tr>
</tbody>
</table>

Model: \( \text{logit}[P(Y=1)] = \alpha + \beta_1 x_1 + \beta_2 z_2 \)

we have 2x2x2 Table
Interpretation of parameters: $\alpha$, $\beta_1$, and $\beta_2$

\[ \log \frac{\pi_{x_i z_k}}{1-\pi_{x_i z_k}} = \alpha + \beta_1 x_i + \beta_2 z_k \]

let $x_2 = 0$, i.e., no immediate use of AZT

and $z_2 = 0$, i.e., the subject is Black

So that

\[ \alpha = \log \frac{\pi_{00}}{1-\pi_{00}} \quad , \quad \pi_{00} = P(Y=1 \mid x=0 , z=0) \]

$\alpha$ is the log odds of developing AIDS symptoms for Black subjects without immediate use of AZT.

\[ \beta_1 = \left[ \alpha + \beta_1 (1) + \beta_2 (1) \right] - \left[ \alpha + \beta_1 (0) + \beta_2 (1) \right] \]

with $x_1 = 1$ and $x_2 = 0$
given $z_1 = 1$: White subject

i.e.,

\[ \beta_1 = \log \frac{\pi_{11}}{1-\pi_{11}} - \log \frac{\pi_{01}}{1-\pi_{01}} \]
\( \beta_1 \) is the increment of log odds for those (White subjects) with immediate AZT use.

or, since

\[
\beta_1 = \left[ \beta_1(1) + \beta_2(0) \right] - \left[ \beta_1(0) + \beta_2(0) \right]
\]

with \( x_1 = 1 \)

\[
\beta_1 = \log \frac{\pi_10}{1 - \pi_10} - \log \frac{\pi_00}{1 - \pi_00}
\]

or

\[
\beta_1 = \log \frac{\pi_{10}}{1 - \pi_{10}} - \log \frac{\pi_{00}}{1 - \pi_{00}}
\]

So that

\( \beta_1 \) is the increment of log odds for those (Black subjects) with immediate AZT use.

Finally: \( \beta_1 \) is the increment of log odds of developing AIDS symptoms for those with immediate AZT use.
By the same way one can show that \( \beta_2 \) is the increment to the log odds developing AIDS symptoms for white subjects (by considering the difference with \( z_i = 1 \) and \( z_k = 0 \)).

**Fitted odds:** Table 5.6

\[
\log\left( \frac{P_{x_i z_k}}{1 - P_{x_i z_k}} \right) = \alpha + \beta_1 x_i + \beta_2 z_k 
\]

\[
= -1.0736 + (-.7195)x_i + .0555z_k
\]

So that

\[
\frac{P_{x_i z_k}}{1 - P_{x_i z_k}} = \frac{e}{1 + e^{(-1.0736 + (-.7195)x_i + .0555z_k)}}
\]

<table>
<thead>
<tr>
<th>Race</th>
<th>AZT</th>
<th>Yes</th>
<th>No</th>
<th>Trials</th>
<th>( \frac{\Lambda}{\Pi_{x_i z_k}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>14</td>
<td>93</td>
<td>107</td>
<td>( \frac{\Lambda}{\Pi_{11}} = .15 )</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>32</td>
<td>81</td>
<td>113</td>
<td>( \frac{\Lambda}{\Pi_{01}} = .26540 )</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>11</td>
<td>52</td>
<td>63</td>
<td>( \frac{\Lambda}{\Pi_{10}} = .1427 )</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>12</td>
<td>43</td>
<td>55</td>
<td>( \frac{\Lambda}{\Pi_{00}} = .25472 )</td>
</tr>
</tbody>
</table>
The estimated OR between immediate AZT use and development of AIDS symptoms is
\[ e^{\beta_1} = e^{-0.7195} = 0.487 \]

95% CI for \( \beta_1 \): \[ -0.7195 \pm 1.96 \times 0.279 \]

Wald: \[ (-1.2663, -0.17266) \]

CI for the effect \( e^{\beta_1} \) (Odds Ratio): \[ -0.7195 \pm 1.96 \times 0.279 \]

or \[ (0.28, 0.84) \]

Likelihood Ratio CI for \( e^{\beta_1} \): \[ (0.279, 0.835) \]

CI for \( e^{\beta_2} \): \[ (0.605, 1.884) \]
Testing Hypothesis

$H_0: \beta_1 = 0$

i.e. conditional independence of AZT treatment & development of AIDS symptoms controlling for race.

Wald Test:

$$W = \left( \frac{\hat{\beta}_1 - \hat{\beta}_0}{SE(\hat{\beta}_1)} \right)^2 = \left( \frac{-0.7195}{0.279} \right)^2 = 6.65$$

$W$ under $H_0$ has $\chi^2_{1 \text{ d.f.}} = 1$

p-value $= P(\chi^2_1 \geq 6.65) \approx 0.01$

LR Test: $G^2 = 6.9$

for testing $H_0: \beta_1 = 0$

$H_0$ is rejected, showing the evidence of association
Problem 5.35: Consider $1 \times 2$ contingency Table and model:

$$\log \frac{\pi_i}{1-\pi_i} = \alpha + \beta_i$$

(a) Given $\{\pi_i > 0\}$, show how to find $\{\beta_i\}$ satisfying $\beta_1 = 0$.

Since $\beta_1 = 0$, then

$$\log \frac{\pi_i}{1-\pi_i} = \alpha$$

and

$$\beta_i = \log \frac{\pi_i}{1-\pi_i} - \alpha = \log \frac{\pi_i}{1-\pi_i} - \log \frac{\pi_i}{1-\pi_i}$$

i.e.

$\beta_i$ - is the log OR for rows $i \neq 1$.

(b) When $\beta_1 = \ldots = \beta_{10}$, then

$\beta_i$ - is the same for rows $i = 1, 2, \ldots, 10$

$$P(Y=1 | X=i) = \text{constant}$$

i.e. $X \& Y$ are independent
The general form of the likelihood equations with covariates \( x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \) are

\[
\sum_i Y_i x_{ij} - \sum_i n_i \pi_i x_{ij} = 0, \quad j = 1, 2, \ldots, p
\]

where \( Y_i \sim \text{Bin}(\pi_i, n_i) \), \( i = 1, 2, \ldots, I \).

In our case under \( H_0: \pi_1 = \pi_2 = \ldots = \pi_I = \pi \)

Assume now that for \( j = 1 \)

\[ x_{i1} = 1, \quad i = 1, 2, \ldots, I \]

then we have

\[
\sum_{i=1}^I Y_i - 1 - \sum_{i=1}^I n_i \pi \cdot 1 = 0
\]

or

\[
\lambda = \frac{\sum_{i=1}^I Y_i}{\sum_{i=1}^I n_i} = \frac{\sum_{i=1}^I Y_i}{n}
\]

Remark: See also equation (4.25) with

\[ n_i y_i = Y_i, \quad \text{and} \quad y_i = \text{the sample proportion of } y = 1 \text{ on the row } i. \]
STAT 555  Spring 2005

Ch. 5: Solutions

Problem 5.1:

(c) at $L_1 = 8$, $\hat{P} = 0.068$, so rate of change in $\hat{P}$ is

$\hat{P} \hat{P} (1-\hat{P}) = (0.1449) \cdot (0.068) / (1 - 0.068) = 0.009$

e $e^{0.009} = 1.01$

(e) $e^{0.1449} = 1.16$

(g) Wald statistic: $W = 5.96$, d.f. = 1

For $H_0: \beta \neq 0$ the p-value = 0.0196

(h) LR statistic: $LR = 8.2988$, d.f. = 1

p-value = 0.004

Problem 5.3:

Fitted Model: $\text{logit}(\hat{P}) = -3.866 + 0.397(\text{snoring})$

Fitted probabilities: 0.021; 0.044; 0.083; 0.132

Effect: 1.49 for one-unit change in snoring

Effect: 2.21 for two-unit change

Goodness-of-fit Test: $G^2 = 2.8$, d.f. = 2
Problem 5.5: The Cochran-Armitage test is better to use when there truly is a linear trend.

Problem 5.9:

(a) Black defendants with white victims had estimated probability .23.

(b) Given defendant's race, the CI for odds is the penalty:

$(3.7, 41.2)$

(c) Wald Test: $T_w^2 = 5.6, d.f.=1$

$LR = 5.0, d.f. = 1$

p-value for LR test = .025

(d) $G^2 = .38, \chi^2 = .20, d.f. = 1$