Topic 30 - Logistic Regression

STAT 525 - Fall 2013

Outline

• Logistic Regression
  – Background
  – Model
  – Inference
  – Diagnostics

Background

• In many applications, the response variable $Y$ has only two possible outcomes, labeled numerically 0 and 1
  – Not diseased ($Y = 0$) vs Disease ($Y = 1$)
  – Unemployed ($Y = 0$) vs Employed ($Y = 1$)
• Response is binary or dichotomous
• Can model response using Bernoulli distribution
  \[
  \Pr(Y_i = 1) = \pi_i \\
  \Pr(Y_i = 0) = 1 - \pi_i \\
  \rightarrow \quad E(Y_i) = \pi_i
  \]
• Goal is to link $E(Y_i) = \pi_i$ with covariates $X_i$

Just use the linear link?

• Suppose we consider the linear model (with one $X_i$)
  \[
  Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i
  \]
• If $Y_i$ only takes values 0 and 1, then
  – We have non-normal error terms
    when $Y_i = 0$: $\varepsilon_i = -\beta_0 - \beta_1 X_i$
    when $Y_i = 1$: $\varepsilon_i = 1 - \beta_0 - \beta_1 X_i$
  – We have nonconstant variance
    \[
    \text{Var}(Y_i) = \pi_i (1 - \pi_i)
    \]
  – Need parameter bounds so $0 \leq \pi_i \leq 1$
Logistic Response Function

- Need alternate function to link $E(Y_i) = \pi_i$ and $X$
- Common to consider the sigmoidal function

$$E(Y_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

$$= (1 + \exp(-\beta_0 - \beta_1 X_i))^{-1}$$

- Example of a nonlinear model
- Other sigmoidal functions (e.g., normal CDF) possible

Properties of Logistic Function

- Monotonic increasing/decreasing function
- Restricts $0 \leq E(Y_i) \leq 1$
- Can be linearized through the logit transformation

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X_i$$

- Use logit link function to relate $\pi_i$ with $X_i$
- Other link functions possible (i.e., probit, complementary log-log)

Logistic model

- Assume the $Y_i$ are independent Bernoulli random variables with means $\pi_i$
- Could express model as $Y_i = E(Y_i) + \varepsilon_i$ but the error terms $\varepsilon_i$ depend on the Bernoulli distribution of $Y_i$
- Better to express

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 X_i$$

Estimation

- Given the distributions of $Y_i$, we can formulate the likelihood function and obtain MLE estimates

$$\log(L) = \log\left(\prod \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i}\right)$$

$$= \sum Y_i \log(\pi_i) + \sum (1 - Y_i) \log(1 - \pi_i)$$

$$= \sum Y_i \log(\frac{\pi_i}{1 - \pi_i}) + \sum \log(1 - \pi_i)$$

$$= \sum Y_i (\beta_0 + \beta_1 X_i) - \sum \log(1 + \exp(\beta_0 + \beta_1 X_i))$$

- MLEs do not have closed forms
- SAS performs iterative reweighted least squares
### Interpretation

- Given $b_0$ and $b_1$, can calculate $\hat{\pi}_i$
  
  $$\hat{\pi}_i = \frac{\exp(b_0 + b_1 X_i)}{1 + \exp(b_0 + b_1 X_i)}$$

- $\hat{\pi}_i$ is the estimated probability of individual $i$ having response $Y_i = 1$

- $b_1$ is now the “slope” of the logit relationship
  
  $$\logit(\hat{\pi}(X_i + 1)) - \logit(\hat{\pi}(X_i)) = b_1$$

- Logit transform is the log of the odds

- Thus, $\exp(b_1)$ becomes the odds ratio

- Popular summary in epidemiologic studies

### Repeat Observations

- Particularly in designed experiments will have repeat observations at certain levels of $X$

- Allows for assessment of model fit (recall Section 3.7)

- Label $X$ levels as $X_1, X_2, \ldots, X_c$ such that there are $n_j$ replicates at level $X_j$ with $Y_{ij}$ the $i^{th}$ replicate at $X_j$

- Can simplify log-likelihood function by considering the sums $Y_{.j}$, which are binomially distributed

  $$Y_{.j} \sim \text{Binomial}(n_j, \pi_j)$$

  $$\logit(\pi_j) = \beta_0 + \beta_1 X_j$$

### Example Page 625

- Board of directors interested in estimating the effect of a due increase on membership

- Randomly surveyed $n = 30$ members

- $X_i$ is the due increase posited to the member

- $X_i$ varied between $30$ and $50$

- Repeat observations for certain $X_i$

- $Y_i$ is whether the member would continue membership

### SAS Commands

```sas
data a1;
  infile 'u:\.www\datasets525\CH14PR07.txt';
  input norenew increase;
  renew=1-norenew;
proc print data=a1;
  var renew increase;
symbol1 v=circle i=sm70;
proc gplot;
  plot renew*increase;
```

```
The Data - Single Trials

<table>
<thead>
<tr>
<th>Obs</th>
<th>renew</th>
<th>increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>37</td>
</tr>
</tbody>
</table>

and so on...

SAS Commands

***Converting data set to events/trials format****

```sas
proc sort data=a1; by renew;
proc freq data=a1; tables increase / out=a1c; by renew;
```

data renew0;
set a1c; if renew=0; n0=count; drop count;
data renew1;
set a1c; if renew=1; n1=count; drop count;
data a1c;
merge renew0 renew1; by increase;if n0=. then n0=0; if n1=. then n1=0;
tot=n0 + n1;
proc print data=a1c;
var increase n1 tot;
run;
```

The Data - Grouped trials

<table>
<thead>
<tr>
<th>Obs</th>
<th>increase</th>
<th>n1</th>
<th>tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>33</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>34</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>35</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>36</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>37</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>38</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>39</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>40</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>45</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>46</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>47</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>48</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>49</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>50</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
```
**SAS Commands**

```sas
proc logistic data=a1 descending;
   model renew = increase;      /*single trial form*/
   output out=a2 p=pred;
proc print;
symbol1 v=circle i=none;
symbol2 v=star i=sm30;
proc gplot data=a2;
   plot renew*increase pred*increase /overlay;
proc logistic data=a1c;
   model n1 / tot = increase;  /*grouped trial form*/
run;
```

**Output**

The LOGISTIC Procedure  ***Single trials model***

Response Profile

<table>
<thead>
<tr>
<th>Ordered Value</th>
<th>Total Value</th>
<th>renew</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
</tbody>
</table>

Probability modeled is renew=1.  ******Be wary******

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Only</th>
<th>Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>43.455</td>
<td>41.465</td>
</tr>
<tr>
<td>SC</td>
<td>44.857</td>
<td>44.267</td>
</tr>
</tbody>
</table>
| -2 Log L       | 41.455 | 37.465    | 3.990

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>3.9906</td>
<td>1</td>
<td>0.0458</td>
</tr>
<tr>
<td>Score</td>
<td>3.8265</td>
<td>1</td>
<td>0.0504</td>
</tr>
<tr>
<td>Wald</td>
<td>3.5104</td>
<td>1</td>
<td>0.0610</td>
</tr>
</tbody>
</table>

Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF Estimate</th>
<th>Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>4.8075</td>
<td>2.6558</td>
<td>3.2769</td>
</tr>
<tr>
<td>increase</td>
<td>1</td>
<td>-0.1251</td>
<td>0.0668</td>
<td>3.5104</td>
</tr>
</tbody>
</table>

Odds Ratio Estimates

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>95% Wald</th>
<th><strong>Wald test like squared</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Point</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>increase</td>
<td>0.882</td>
<td>0.774</td>
<td>1.006</td>
</tr>
</tbody>
</table>

**Scatterplot/Fit**

![Scatterplot/Fit](image-url)
### Logistic Regression Output

**The LOGISTIC Procedure**  
***Grouped trials model***

#### Response Profile

<table>
<thead>
<tr>
<th>Ordered</th>
<th>Binary</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Event</td>
<td>142</td>
</tr>
<tr>
<td>2</td>
<td>Nonevent</td>
<td>16</td>
</tr>
</tbody>
</table>

#### Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

#### Model Fit Statistics

**Intercept and Covariates**

<table>
<thead>
<tr>
<th>Criterion</th>
<th>DF Estimate</th>
<th>Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>4.8075</td>
<td>2.6558</td>
<td>3.2769</td>
</tr>
<tr>
<td>increase</td>
<td>-0.1251</td>
<td>0.0668</td>
<td>3.5104</td>
<td>0.0610</td>
</tr>
</tbody>
</table>

#### Odds Ratio Estimates

- **increase**: 0.882 (0.774, 1.006)

### Alternate Approaches

Logistic regression is an example of a generalized model so we can consider generalized linear model procedures:

- **proc genmod data=a1 descending;**
- **proc genmod data=a1c;**

### GENMOD Output

#### Criteria For Assessing Goodness Of Fit

**Deviance**
- **DF 19**
- **Value 22.1885**
- **Value/DF 1.1678**

**Scaled Deviance**
- **DF 19**
- **Value 22.1885**
- **Value/DF 1.1678**

**Pearson Chi-Square**
- **DF 19**
- **Value 17.9966**
- **Value/DF 0.9472**

**Scaled Pearson X2**
- **DF 19**
- **Value 17.9966**
- **Value/DF 0.9472**

**Log Likelihood**
- **Value -18.7324**

**Full Log Likelihood**
- **Value -14.3379**

**AIC (smaller is better)**
- **Value 32.6759**

**AICC (smaller is better)**
- **Value 33.1203**

**BIC (smaller is better)**
- **Value 35.4783**

#### Analysis Of Maximum Likelihood Parameter Estimates

**Intercept**
- **DF 1**
- **Estimate 4.8075**
- **Error 2.6558**
- **Chi-Square 3.2769**
- **Pr > ChiSq 0.0703**

**increase**
- **DF 1**
- **Estimate -0.1251**
- **Error 0.0668**
- **Chi-Square 3.5104**
- **Pr > ChiSq 0.0610**

**Scale**
- **DF 0**
- **Estimate 1.0000**
- **Error 1.0000**
GLIMMIX Output

Fit Statistics

-2 Log Likelihood 37.46
AIC (smaller is better) 41.46 **Single trial model
AICC (smaller is better) 41.91
BIC (smaller is better) 44.27 **Likelihood and model
CAIC (smaller is better) 46.27 selection criteria
HQIC (smaller is better) 42.36 based on single trial
Pearson Chi-Square 30.10 likelihood
Pearson Chi-Square / DF 1.08

Type III Tests of Fixed Effects

Effect DF DF Chi-Square F Value Pr > ChiSq Pr > F
increase 1 28 3.51 3.51 0.0610 0.0715

**Will not discuss F test but Den df = 30-1-1**

Comparison of Global Tests

- **Likelihood ratio test** compares \(-2 \log(L)\) of two nested models. The df equals the difference in the number of model parameters.

- **Wald test** requires fit only to full model. Assesses how far estimated parameters are from hypothesized values in terms of standard errors. Uses asymptotic normality of MLEs. When considering only one parameter, test just the squared-z test.

- **Score test** requires fit only to \(H_0\) model. Test is based on the slope of the log-likelihood function at the values specified by the null hypothesis. Describes expected change in chi-squared statistic if variables were added.

Deviance

- Recall hypothesis testing can be done using a likelihood ratio test, which is similar to the general linear test.

- Consider the grouped trials form. The deviance of the fitted model is the difference in the log-likelihood of the fitted model and the most general model, which has a parameter \((\pi_j)\) for each \(X_j\).

- The MLEs for this full model are \(p_j = Y_{.j}/n_{.j}\) for \(j = 1, 2, \ldots, c\) so the deviance for the fitted model is:

\[
-2 \left( \log(L(R)) - \log(L(F)) \right) - 2 \sum \left( Y_{.j} \log \left( \frac{\hat{\pi}_j}{p_j} \right) + (n_{.j} - Y_{.j}) \log \left( \frac{1 - \hat{\pi}_j}{1 - p_j} \right) \right)
\]
Using the Deviance

- If the logistic model is correct, then the deviance will be approximately chi-square with $c - p$ df.
- Since the mean of the chi-square is its df, we’d expect the deviance/df to be approximately 1. This is commonly used to assess goodness of fit.
- Can also use deviance to compare two hierarchical models

\[
\text{DEV}(X_{\eta}, \ldots, X_{\eta-1} | X_0, \ldots, X_{\eta-1}) = \text{DEV}(X_0, \ldots, X_{\eta-1}) - \text{DEV}(X_0, \ldots, X_{p-1})
\]

- Partial deviance approx $\chi^2$ with $p - q$ df
- Must compute by hand or with GENMOD
- Will use this in multiple logistic regression

Pearson Goodness of Fit Test

- Alternative goodness of fit test using same principals as deviance goodness of fit test.
- For each observed $X_j$, can compute the expected number of events and nonevents.

\[
X^2 = \sum_c \left( \frac{(Y_j - n_j \hat{\pi}_j)^2}{n_j \hat{\pi}_j} + \frac{(n_j - Y_j - n_j(1 - \hat{\pi}_j))^2}{n_j(1 - \hat{\pi}_j)} \right)
\]

- This also will approximately follow a chi-square distribution with $c - p$ df.

Wald’s Test

- Can also use Wald’s Test for $H_0 : L' \beta = C$
- Described on page 578 for single parameter tests

\[
S = (L'\hat{\beta} - C)'(L'\hat{\Sigma}L)^{-1}(L'\hat{\beta} - C)
\]

where $\hat{\Sigma}$ is the estimate covariance matrix of $\hat{\beta}$

- Under $H_0$, $S \sim \chi^2_r$ where $r$ is rank of $L$
- Available in GENMOD and LOGISTIC

Alternative Goodness of Fit Test

- Have previously described goodness of fit tests when there is replication
- For unreplicated studies, can consider the Hosmer-Lemeshow goodness of fit
  - Group observations into classes (usually around 10) according to fitted logit values.
  - Assess overall fit to each class using a Pearson goodness of fit approach.
Hosmer-Lemeshow Goodness of Fit Test

- Divide obs up into \( \approx 10 \) groups of equal size based on percentiles of the estimated probabilities
- Expected # of 1’s is \( \sum \hat{\pi}_i \)
- Expected # of 0’s is \( n_i - \sum \hat{\pi}_i \)
- Compare expected with observed through

\[
\chi^2 = \sum \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
\]

- In the example

\[
E_{11} = (0.19056 + 0.19056 + 0.21060) = 0.59 \rightarrow E_{10} = 2.41
\]

\[
E_{21} = (0.23214 + 0.25518 + 0.27967) = 0.77 \rightarrow E_{20} = 2.23
\]

SAS Commands

```sas
data a1;
  infile 'u:\www\datasets525\CH14PR07.txt';
  input norenew increase;
  renew = 1 - norenew;
  
proc logistic data=a1 descending;
  model renew = increase / lackfit;
  output out=a2 p=pred;
  
run;
proc print;
run;
```

Additional Output

| Percent Concordant | 68.3 | Somers’ D | 0.402 |
| Percent Discordant | 28.1 | Gamma     | 0.417 |
| Percent Tied       | 3.6  | Tau-a     | 0.207 |
| Pairs              | 224  | c         | 0.701 |

Partition for the Hosmer and Lemeshow Test

<table>
<thead>
<tr>
<th>Group</th>
<th>Total Observed</th>
<th>Expected</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0.59</td>
<td>2</td>
<td>2.41</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.77</td>
<td>2</td>
<td>2.23</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.92</td>
<td>2</td>
<td>2.08</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1.08</td>
<td>2</td>
<td>1.92</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>1.77</td>
<td>2</td>
<td>2.23</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>1.54</td>
<td>1</td>
<td>1.46</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>2.39</td>
<td>2</td>
<td>1.61</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>1.99</td>
<td>1</td>
<td>1.01</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>2.94</td>
<td>1</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Hosmer and Lemeshow Goodness-of-Fit Test

<table>
<thead>
<tr>
<th>Chi-Square DF Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6526 7 0.9152</td>
</tr>
</tbody>
</table>

Output

<table>
<thead>
<tr>
<th>Obs</th>
<th>no renew</th>
<th>increase</th>
<th>renew</th>
<th><em>LEVEL</em></th>
<th>pred</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>50</td>
<td>0</td>
<td>1</td>
<td>0.19056</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>50</td>
<td>0</td>
<td>1</td>
<td>0.19056</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>49</td>
<td>1</td>
<td>1</td>
<td>0.21060</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>48</td>
<td>0</td>
<td>1</td>
<td>0.23214</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>47</td>
<td>0</td>
<td>1</td>
<td>0.25518</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>46</td>
<td>1</td>
<td>1</td>
<td>0.27967</td>
</tr>
</tbody>
</table>

Hosmer and Lemeshow Test

<table>
<thead>
<tr>
<th>Chi-Square DF Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6526 7 0.9152</td>
</tr>
</tbody>
</table>
Measures of Agreement

- Have $N$ observations
- Consider all pairs of distinct responses
  - In this example $t = 16 \times 14 = 224$
- Compare predicted probabilities
  - Concordant if $\hat{\pi}_{Y=1} > \hat{\pi}_{Y=0}$
  - Discordant if $\hat{\pi}_{Y=1} < \hat{\pi}_{Y=0}$
  - Tie if $\hat{\pi}_{Y=1} = \hat{\pi}_{Y=0}$
- Measures of agreement
  - Somers’ $D$: $(\#C - \#D)/t$
  - Goodman-Kruskal Gamma: $(\#C - \#D)/(\#C + \#D)$
  - Kendall’s Tau-a: $(\#C - \#D)/(0.5N(N-1))$
  - $c$: $(\#C + 0.5(t-\#C-\#D))/t$

Background Reading

- KNNL Chapter 14
- knnl555.sas
- KNNL Chapter 14