Multiple Logistic Regression

Bret Larget
Departments of Botany and of Statistics
University of Wisconsin—Madison
April 26, 2007

Multiple logistic regression is an extension of logistic regression to the case where there may be multiple explanatory variables.

- The basic idea is the same, where the probability of one outcome is modeled as a function of the linear combination of several explanatory variables.
- A special case of multiple logistic regression is when the probability varies as a polynomial function of a single quantitative explanatory variable.
- This is similar to polynomial regression.

Cow Example

Mastitis in Cows

- In an example from a student in class, we have daily data on the number of cows from a dairy herd that experience new cases of mastitis, or inflammation of the udder.
- Mastitis is a costly problem for dairy farmers.
- We wish to examine the trend in the rate of mastitis over time.
- We will consider possible nonlinear trends in time.

Data

- We will model the new cases of mastitis as the response variable.
- The size of the herd changes slightly each day.
- We account for the changes in herd size, but do not model individual cows.
- We create a new variable called time which is days since the beginning of the year.
First GLM Analysis

```r
> prop = numNewMastitis/numCows
> fit1 = glm(prop ~ time, family = binomial, weights = numCows)
> summary(fit1)
```

```r
Call: glm(formula = prop ~ time, family = binomial, weights = numCows)
Deviance Residuals:
Min 1Q Median 3Q Max
-2.95874 -0.58793 0.02157 0.42698 2.24759
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.823143 0.119411 -48.765 <2e-16 ***
time 0.004649 0.002653 1.752 0.0797 .
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 83.08 on 73 degrees of freedom
Residual deviance: 80.00 on 72 degrees of freedom
AIC: 316.22
Number of Fisher Scoring iterations: 4
```

Comments

- For this data set, there is slight evidence of an increasing trend of mastitis rate in time, but no need for a quadratic model.
- In fact, for this data a regular linear model would have sufficed.
- The next plot compares the logistic regression and simple linear regression models.

Second GLM Analysis

```r
> fit2 = glm(prop ~ time + I(time^2), family = binomial, weights = numCows)
> summary(fit2)
```

```r
Call: glm(formula = prop ~ time + I(time^2), family = binomial, weights = numCows)
Deviance Residuals:
Min 1Q Median 3Q Max
-2.919315 -0.547581 -0.005287 0.456351 2.294789
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.763e+00 1.827e-01 -31.535 <2e-16 ***
time 7.236e-05 1.089e-02 0.007 0.995
I(time^2) 5.959e-05 1.377e-04 0.433 0.665
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 83.080 on 73 degrees of freedom
Residual deviance: 79.814 on 71 degrees of freedom
AIC: 318.04
Number of Fisher Scoring iterations: 4
```

Plots

```r
> fit3 = lm(prop ~ time)
> plot(time, prop, pch = 16)
> eta = predict(fit1, data.frame(time = time))
> prob = exp(eta)/(1 + exp(eta))
> lines(time, prob, col = "red")
> abline(fit3, col = "blue")
```
Seed Germination Experiment

- We studied this seed germination data earlier in the semester.
- In an experiment, four sites were selected where the soil and climate conditions were expected to be very similar within the site.
- Here we will treat each site as a block.
- Within each block, five plots were identified.
- The treatment was applying a seed disinfectant to seeds. There were four different treatments (brands) plus a control.
- The researchers planted 100 seeds from a single treatment in each plot.
- The response is the number of seeds that germinated.

Model

- A model is
  \[ \eta_{ij} = \mu + \alpha_i + \beta_j, \quad P\{\text{seed } ij \text{ germinates}\} = \frac{e^{\eta_{ij}}}{1 + e^{\eta_{ij}}} \]
  where:
  - \( \mu \) is an intercept,
  - \( \alpha_i \) is the effect of treatment \( i \) where \( \sum_i \alpha_i = 0 \).
  - and \( \beta_j \) is the effect in block \( j \) where \( \sum_j \beta_j = 0 \).

Data

```r
> seed = read.table("seed.txt", header = T)
> attach(seed)
> str(seed)
'data.frame': 20 obs. of 3 variables:
  $ count: int 86 90 88 87 98 94 93 89 96 90 ...  
  $ treatment: Factor w/ 5 levels "AControl","Arasan",..: 1 1 1 1 2 2 2 2 5 5 ...  
  $ block : Factor w/ 4 levels "b1","b2","b3",..: 1 2 3 4 1 2 3 4 1 2 ...  
```

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Block 1</th>
<th>Block 2</th>
<th>Block 3</th>
<th>Block 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>86</td>
<td>90</td>
<td>88</td>
<td>87</td>
</tr>
<tr>
<td>Arasan</td>
<td>98</td>
<td>94</td>
<td>93</td>
<td>89</td>
</tr>
<tr>
<td>Spergon</td>
<td>96</td>
<td>90</td>
<td>91</td>
<td>92</td>
</tr>
<tr>
<td>Semesan</td>
<td>97</td>
<td>95</td>
<td>91</td>
<td>92</td>
</tr>
<tr>
<td>Fermate</td>
<td>91</td>
<td>93</td>
<td>95</td>
<td>95</td>
</tr>
</tbody>
</table>

Lines show treatment

Lines show blocking
It looks like each treatment improves germination rate over the control.

Germination Probabilities

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Control</th>
<th>Arasan</th>
<th>Spergon</th>
<th>Semesan</th>
<th>Fermate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.878</td>
<td>0.936</td>
<td>0.936</td>
<td>0.938</td>
<td>0.923</td>
</tr>
</tbody>
</table>