Multiple Logistic Regression

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

April 26, 2007

Multiple Logistic Regression

- *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*.

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 inflamation 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.

```
> cows = read.table("cows", header = T)
> str(cows)

'data.frame': 74 obs. of  5 variables:
 $ date          : Factor w/ 74 levels "1/1/07","1/10/07",..: 1 22 25 
 $ numCows       : num  1177 1174 1178 1182 1190 ...
 $ numMastitis   : num  38 35 34 35 33 36 32 31 32 32 ...
 $ numNewMastitis: num  2 5 4 5 4 0 4 4 3 8 ...
 $ milk          : num  87.8 86.8 86.1 85.4 85.2 84.8 85.6 85.6 86.2 8

> cows = data.frame(time = 1:74, cows)
> str(cows)

'data.frame': 74 obs. of  6 variables:
 $ time          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ date          : Factor w/ 74 levels "1/1/07","1/10/07",..: 1 22 25 
 $ numCows       : num  1177 1174 1178 1182 1190 ...
 $ numMastitis   : num  38 35 34 35 33 36 32 31 32 32 ...
 $ numNewMastitis: num  2 5 4 5 4 0 4 4 3 8 ...
 $ milk          : num  87.8 86.8 86.1 85.4 85.2 84.8 85.6 85.6 86.2 8

> attach(cows)
```

First GLM Analysis

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

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
```


Second GLM Analysis

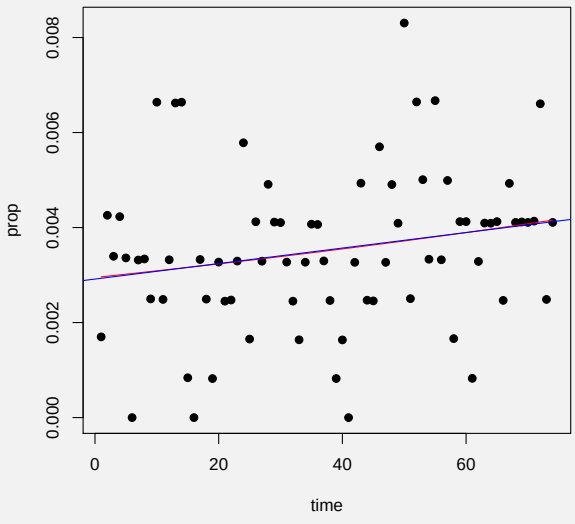
```
> fit2 = glm(prop ~ time + I(time^2), family = binomial, weights = numCows)
> summary(fit2)

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
```

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.

Plots

```
> 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.

Data

	Block			
Treatment	1	2	3	4
Control	86	90	88	87
Arasan	98	94	93	89
Spergon	96	90	91	92
Semesan	97	95	91	92
Fermate	91	93	95	95

Lines show treatment

	Block			
Treatment	1	2	3	4
Control	86	90	88	87
Arasan	98	94	93	89
Spergon	96	90	91	92
Semesan	97	95	91	92
Fermate	91	93	95	95

Lines show blocking

Model

- A model is

$$\eta_{ij} = \mu + \alpha_i + \beta_j, \qquad \mathrm{P}\{\text{seed } ij \text{ germinates}\} = \frac{\mathrm{e}^{\eta_{ij}}}{1 + \mathrm{e}^{\eta_{ij}}}$$

  where:
  - μ is an intercept,
  - α_i is the effect of treatment i where $\sum_i \alpha_i = 0$.
  - and β_j is the effect in block j where $\sum_j \beta_j = 0$.

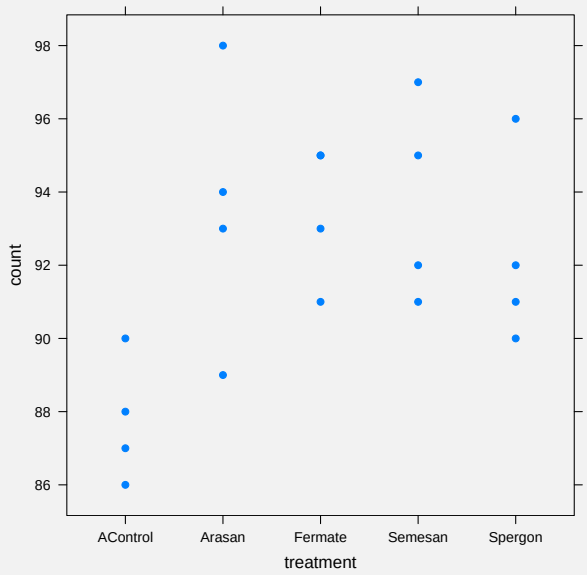
Data

```
> 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 ...
> seed
   count treatment block
1     86  AControl    b1
2     90  AControl    b2
3     88  AControl    b3
4     87  AControl    b4
5     98    Arasan    b1
6     94    Arasan    b2
7     93    Arasan    b3
8     89    Arasan    b4
9     96   Spergon    b1
10    90   Spergon    b2
11    91   Spergon    b3
12    92   Spergon    b4
13    97   Semesan    b1
14    95   Semesan    b2
15    91   Semesan    b3
16    92   Semesan    b4
17    91   Fermate    b1
18    93   Fermate    b2
19    95   Fermate    b3
20    95   Fermate    b4
```

Plot of Data

```
> library(lattice)
> fig1 = xyplot(count ~ treatment, pch = 16)
> print(fig1)
```

- It looks like each treatment improves germination rate over the control.



Logistic Regression Analysis

```
> options(contrasts = c("contr.sum", "contr.poly"))
> prop = count/100
> seed.glm = glm(prop ~ treatment + block, family = binomial, weights = rep(100,
+     length(prop)))
> summary(seed.glm)

Call:
glm(formula = prop ~ treatment + block, family = binomial, weights = rep(100,
    length(prop)))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5209  -0.4512   0.2006   0.6496   1.6685

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.50532    0.08622  29.057  < 2e-16 ***
treatment1  -0.52907    0.14601  -3.624 0.000291 ***
treatment2   0.16918    0.17901   0.945 0.344616
treatment3   0.16918    0.17901   0.945 0.344616
treatment4   0.21112    0.18156   1.163 0.244910
block1       0.21328    0.15447   1.381 0.167356
block2       0.02734    0.14631   0.187 0.851783
block3      -0.08223    0.14199  -0.579 0.562502
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.237  on 19  degrees of freedom
Residual deviance: 14.634  on 12  degrees of freedom
AIC: 105.42

Number of Fisher Scoring iterations: 4
```

Germination Probabilities

```
> cf = coef(seed.glm)
> cf

(Intercept)  treatment1  treatment2  treatment3  treatment4      block1
 2.50532055 -0.52907338  0.16917897  0.16917897  0.21111575  0.21328163
     block2      block3
 0.02733672 -0.08223199

> eta = cf[1] + c(cf[2:5], -sum(cf[2:5]))
> prob = round(exp(eta)/(1 + exp(eta)), 3)
> prob

treatment1 treatment2 treatment3 treatment4
     0.878      0.936      0.936      0.938      0.923
```

Treatment	Control	Arasan	Spergon	Semesan	Fermate
Probability	0.878	0.936	0.936	0.938	0.923