Generalized Linear Models

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

April 24, 2007

The Big Picture

- In all of the linear models we have seen so far this semester, the *response variable* has been modeled as a *normal random variable*.

$$\text{(response)} = \text{(fixed parameters)} + \text{(normal random effects and error)}$$

- For many data sets, this model is inadequate.
- For example, if the response variable is *categorical* with two possible responses, it makes no sense to model the outcome as normal.
- Also, if the response is always a small positive integer, its distribution is also not well described by a normal distribution.
- *Generalized linear models* (GLMs) are an extension of linear models to model non-normal response variables.
- We will study *logistic regression* for *binary response variables* and *Poisson regression* for *small integer response variables*.

Generalized Linear Models

- A standard linear model has the following form:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e, \qquad e_i \sim \mathrm{N}(0, \sigma^2)$$

- The mean of *expected value* of the response is written this way.

$$\mathrm{E}[y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- We will use the notation $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ represent the *linear combination* of explanatory variables.
- In a standard linear model, $\mathrm{E}[y] = \eta$.
- In a GLM, there is a *link function* f between η and the mean of the response variable.

$$f(\mathrm{E}[y]) = \eta$$

- For standard linear models, the link function is the identity function $f(y) = y$.

Link Functions

- It is usually more clear to consider the *inverse of the link function*.

$$\mathrm{E}[y] = f^{-1}(\eta)$$

- This inverse is often the *function of a parameter* of a distribution of interest.
- When the response variable is binary (with values coded as 0 or 1), the mean is simply $\mathrm{E}[y] = \mathrm{P}\{y = 1\}$.
- A useful function for this case is

$$\mathrm{E}[y] = \mathrm{P}\{y = 1\} = \frac{\mathrm{e}^{\eta}}{1 + \mathrm{e}^{\eta}}$$

- Notice that the parameter is always between 0 and 1.
- The corresponding link function is called the *logit function*, $f(x) = \log(x/(1-x))$ and regression under this model is called *logistic regression*.

Deviance

- In standard linear models, we estimate the parameters by *minimizing the sum of the squared residuals*.
- This is equivalent to finding parameters that *maximize the likelihood*.
- In a GLM we also fit parameters by maximizing the likelihood.
- The *deviance* is equal to twice the difference between log likelihoods for reduced and full models.
- Estimation is equivalent to finding parameter values that *minimize the deviance*.
- For hypothesis testing, a test using the deviance is equivalent to a *likelihood ratio test*.

Logistic Regression

- Logistic regression is a natural choice when the response variable is categorical with *two possible outcomes*.
- Pick one outcome to be a "success", where $y = 1$.
- We desire a model to estimate the probability of "success" as a function of the explanatory variables.
- Using the *logit* function, the probability of success has the form

$$\mathrm{P}\{y = 1\} = \frac{\mathrm{e}^{\eta}}{1 + \mathrm{e}^{\eta}}$$

- We estimate the parameters so that this probability is high for cases where $y = 1$ and low for cases where $y = 0$.

Example

- In surgery, it is desirable to give enough anesthetic so that patients do not move when an incision is made.
- It is also desirable not to use much more anesthetic than necessary.
- In an experiment, patients are given different concentrations of anesthetic.
- The response variable is whether or not they move at the time of incision 15 minutes after receiving the drug.

Data

	Concentration					
	0.8	1.0	1.2	1.4	1.6	2.5
Move	6	4	2	2	0	0
No move	1	1	4	4	4	2
Total	7	5	6	6	4	2
Proportion	0.17	0.20	0.67	0.67	1.00	1.00

- Analyze in R with `glm` twice, once using raw data and once using summarized counts.

Binomial Distribution

- Logistic regression is related to the *binomial distribution*.
- If there are several observations with the same explanatory variable values, then the individual responses can be added up and the sum has a binomial distribution.
- Recall for the binomial distribution that the parameters are n and p and the moments are $\mu = p$ and $\sigma^2 = np(1-p)$.
- The probability distribution is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- Logistic regression is in the "binomial family" of GLMs.

R with Raw Data

```
> library(DAAG)
> str(anesthetic)

'data.frame': 30 obs. of  4 variables:
 $ move   : num  0 1 0 1 1 0 0 1 0 1 ...
 $ conc   : num  1 1.2 1.4 1.4 1.2 2.5 1.6 0.8 1.6 1.4 ...
 $ logconc: num  0.000 0.182 0.336 0.336 0.182 ...
 $ nomove : num  1 0 1 0 0 1 1 0 1 0 ...

> attach(anesthetic)
> aneRaw.glm = glm(nomove ~ conc, family = binomial(link = "logit"))
```


R with Raw Data

```
> summary(aneRaw.glm)

Call:
glm(formula = nomove ~ conc, family = binomial(link = "logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.76666  -0.74407   0.03413   0.68666   2.06900

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.469      2.418  -2.675  0.00748 **
conc           5.567      2.044   2.724  0.00645 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.455  on 29  degrees of freedom
Residual deviance: 27.754  on 28  degrees of freedom
AIC: 31.754

Number of Fisher Scoring iterations: 5
```


Fitted Model

- The fitted model is the following.

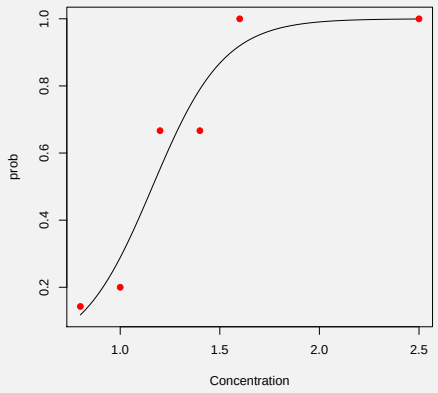
$$\eta = -6.47 + 5.57 \times (\text{concentration})$$

and

$$\mathrm{P}\{\text{No move}\} = \frac{\mathrm{e}^{\eta}}{1 + \mathrm{e}^{\eta}}$$

Plot of Relationship

```
> u = seq(0.8, 2.5, length = 201)
> eta = predict(aneRaw.glm, data.frame(conc = u))
> prob = exp(eta)/(1 + exp(eta))
> plot(u, prob, type = "l", xlab = "Concentration")
> conc = c(0.8, 1, 1.2, 1.4, 1.6, 2.5)
> prop = c(1/7, 1/5, 4/6, 4/6, 4/4, 2/2)
> points(conc, prop, pch = 16, col = "red")
```



Second Analysis

```
> noCounts = c(1, 1, 4, 4, 4, 2)
> total = c(7, 5, 6, 6, 4, 2)
> prop = noCounts/total
> concLevels = c(0.8, 1, 1.2, 1.4, 1.6, 2.5)
> aneTot.glm = glm(prop ~ concLevels, family = binomial, weights = total)
```


Second Analysis

```
> summary(aneTot.glm)

Call:
glm(formula = prop ~ concLevels, family = binomial, weights = total)

Deviance Residuals:
       1         2         3         4         5         6
 0.20147  -0.45367   0.56890  -0.70000   0.81838   0.04826

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.469      2.419  -2.675  0.00748 **
concLevels     5.567      2.044   2.724  0.00645 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 15.4334  on 5  degrees of freedom
Residual deviance:  1.7321  on 4  degrees of freedom
AIC: 13.811

Number of Fisher Scoring iterations: 5
```