Logistic Regression

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

February 14, 2008

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 additional models in Chapter 6.

Generalized Linear Models

- A standard linear model has the following form:

$$y = \beta_1 1 + \beta_2 x_2 + \cdots + \beta_k x_k + e, \qquad e_i \sim \mathrm{N}(0, \sigma^2)$$

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

$$\mathrm{E}[y] = \beta_1 1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- We will use the notation $\eta = \beta_1 1 + \beta_1 x_1 + \cdots + \beta_k x_k$ to represent the *linear combination* of explanatory variables.
- In a standard linear model, $\mathrm{E}[y] = \eta$.
- In a GLM, there is a *link function* g between η and the mean of the response variable.

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

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

Link Functions

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

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

- The mean of a distribution is usually either a parameter of a distribution or is a function of parameters of a distribution, which is what the this inverse function shows.
- 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*, $g(x) = \log(x/(1-x))$ 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 log likelihood* up to an additive constant.
- Estimation is equivalent to finding parameter values that *minimize the deviance*.

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 inverse *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 $\mu = np$ 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

```
> ane = read.table("anesthetic.txt", header = T)
> str(ane)

'data.frame': 30 obs. of  3 variables:
 $ movement: Factor w/ 2 levels "move","noMove": 2 1 2 1 1 2 2 1 2 1 .
 $ conc    : num  1 1.2 1.4 1.4 1.2 2.5 1.6 0.8 1.6 1.4 ...
 $ nomove  : int  1 0 1 0 0 1 1 0 1 0 ...

> aneRaw.glm = glm(nomove ~ conc, data = ane,
+     family = binomial(link = "logit"))
```


R with Raw Data

```
> library(arm)

arm (Version 1.1-1, built: 2008-1-13)
Working directory is /Users/bret/Desktop/s572/Spring2008/Notes
options( digits =  2 )

> display(aneRaw.glm, digits = 3)

glm(formula = nomove ~ conc, family = binomial(link = "logit"),
    data = ane)
            coef.est coef.se
(Intercept) -6.469    2.418
conc         5.567    2.044
---
  n = 30, k = 2
  residual deviance = 27.8, null deviance = 41.5 (difference = 13.7)
```


Fitted Model

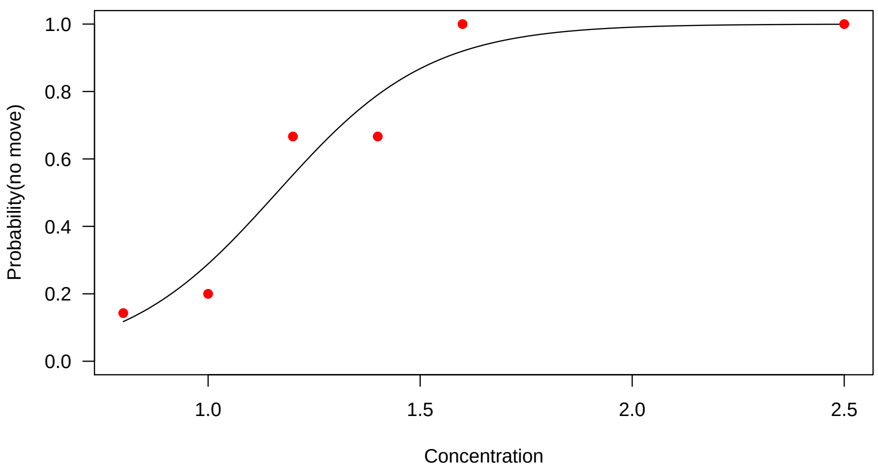
- The fitted model is the following.

$$\eta = -6.469 + 5.567 \times (\text{concentration})$$

and

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

Plot of Relationship



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)
> ane2 = data.frame(noCounts, total, prop,
+     concLevels)
> aneTot.glm = glm(prop ~ concLevels, data = ane2,
+     family = binomial, weights = total)
```


Second Analysis

```
> display(aneTot.glm)

glm(formula = prop ~ concLevels, family = binomial, data = ane2,
    weights = total)
            coef.est coef.se
(Intercept) -6.47     2.42
concLevels   5.57     2.04
---
  n = 6, k = 2
  residual deviance = 1.7, null deviance = 15.4 (difference = 13.7)
```