Overdisperse Binomials

Sanford Weisberg

Department of Applied Statistics, University of Minnesota, St. Paul, MN 55108-6042.

Supported by National Science Foundation Grant DUE 96-52887

May 21, 1999

Abstract

This note describes the fitting of overdisperse binomials using *Arc* using methodology outlined by Williams (1982). *Arc* is the computer program that is described in Cook and Weisberg (1999), and is available from the web site www.stat.umn.edu/arc.

1 Introduction

Suppose we observe the number of successes y_i in m_i trials, $i = 1, \ldots, r$, such that

$$\begin{aligned} y_i|p_i &\sim \text{Bin}(m_i, p_i) \\ p_i &\sim \text{Beta}(\gamma, \delta) \end{aligned}$$

Under this model, each of the r binomials has a different probability of success p_i, where p_i is a random draw from a beta distribution. The mean and variance of the p_i are given by

$$\begin{aligned} \text{E}(p_i) &= \frac{\gamma}{\gamma+\delta} = \theta \\ \text{Var}(p_i) &= \frac{\theta(1-\theta)}{\gamma+\delta-1} = \phi\theta(1-\theta) \end{aligned}$$

We will assume that $\gamma > 1$ and $\delta > 1$, so that the beta density is equal to zero at both zero and one, and thus $0 < \phi \leq 1/3$. From this, we can calculate the unconditional mean and variance of the y_i to be

$$\begin{aligned} \text{E}(y_i) &= m_i\theta \\ \text{Var}(y_i) &= m_i\theta(1-\theta)(1+(m_i-1)\phi) \end{aligned} \tag{1}$$

so unless $m_i = 1$ or $\phi = 0$, the unconditional variance of y_i is larger than binomial variance.

Identical expressions for the mean and variance of y_i can be obtained if we assume that the m_i counts on the i-unit are dependent, with the same correlation ϕ. In this case $-1/(m_i - 1) < \phi \leq 1$, with negative ϕ corresponding to underdispersion.

Table 1: The *Orobanche* Data, from Crowder (1978).

Variety:	O.a 75				O.a 73			
Host:	Bean		Cucumber		Bean		Cucumber	
	y	n	y	n	y	n	y	n
Slide 1	10	39	5	6	8	16	3	12
Slide 2	23	62	53	74	10	30	22	41
Slide 3	23	81	55	72	8	28	15	30
Slide 4	26	51	32	51	23	45	32	51
Slide 5	17	39	46	79	0	4	3	7
Slide 6			10	13				

Table 2: Fit of the Saturated Model.

```
Data set = Orobanche, Name of Fit = B1
Binomial Regression
Mean function = Logistic
Response      = y
Predictors    = (v h v*h)
Trials        = n
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE
Constant  -0.558172        0.126020       -4.429
v          0.145927        0.223164        0.654
h          1.31818         0.177466        7.428
v.h       -0.778104        0.306431       -2.539

Sigma hat:                    1.
Number of cases:             21
Degrees of freedom:          17
Pearson X2:               31.651
Deviance:                 33.278
```

2 Logistic models

Now suppose that we have observed counts $y_i(x_j)$ in m_{ij} trials, where x_j is the j set of predictors, and i indexes the replicate of the predictors. For example, we consider two varieties of the parasitic plant *Orobanche* each grown on one of two media, either beans or cucumbers. Several slides for each of the $2 \times 2 = 4$ treatment combinations were prepared, and on each slide the number of seeds that germinate were recorded. We could observe overdispersion if (1) the probability of germination varied from slide to slide (a random effect); or (2) the seeds on a slide were correlated. In either case, we need to take account of the possibility of overdispersion. The data in Table 1 are in the file `orabanche.lsp` on the *Arc* website.

The saturated model is shown in Table 2. The outstanding feature of the fit is that the *saturated model fails to fit*! Examination of the counts in Table 1 shows the reason:

The variation between slides is too large to be explained by binomial variation alone. We would like to fit a logistic-like model, but with variance function given by (1). Methodology for this is provided by Williams (1982). We briefly summarize his ideas.

If we knew ϕ, the mean and variance equations (1), now viewed as a function of the predictors would be appropriate for a "weighted" logistic regression, with weights given by $w_i = (1 + \phi(m_i - 1))^{-1}$. We could then use the logistic regression paradigm with weights to get estimates of parameters, and rely of the results concerning quasi-likelihood (see, for example, McCullagh and Nelder, 1989, Chapter 9) for optimality properties of the resulting estimates.

If ϕ is unknown but the $m_i = m$ are all equal, then the usual unweighted coefficient estimates are correct, but standard errors are too small because they fail to account for the multiplier $1 + (m - 1)\phi$. We can estimate this constant by $X^2/(n - k)$, where X^2 is Pearson's Chi-square statistic, n is the number of binomials observed, and k is the number of parameters fit, and then multiplying each of the standard errors by the square root of this quantity.

For the general case of ϕ unknown and the m_i unequal, Williams (1982) presents an iterative algorithm (summaried in Section 4) that will estimate ϕ, and hence the necessary binomial weights $(1 + \phi(m_i - 1))^{-1}$, based on the expectation of X^2; the details are given by Williams.

3 Fitting Using Arc

To fit extra-binomial variation in *Arc*, first select "Settings" from the Arc menu, then select the item for binomial-extra-variance, and update the setting (click on the item, and then select "Update" from the Settings menu). This will make the menu item available to all binomial regressions.

Next, fit the model of interest (such as the saturated model shown in Table 1. From this model's menu, select the item "Fit extra-binomial variance;" the result is shown in Table 3. The new information is the estimate $\hat{\phi} = 0.0249$ as the over-dispersion parameter. This is estimated by equating X^2 to its degrees of freedom, so the goodness of fit statistics are no longer meaningful

We see that the Wald statistic for the interaction in Table 3 is about -1.88, giving a significance level (compared to the normal distribution) somewhat larger than 0.05. As a consequence, we may feel comfortable fitting a model without the interaction, but using $\hat{\phi}$ to estimate weights. Select "Fit binomial" from the Graph&Fit menu. The program will assume you want to use the extra binomial weights in the fitting; remove the interaction from the model, to get the output shown in Table 4 as the final summary of these data.

4 Algorithm

Let h_i be the estimate of the i-leverage; in logistic models, the leverage depends on the variance $v_i = m_i\theta_i(1 - \theta_i)$ and on the current estimate weight w_i.

Table 3: Fitting Extra Binomial Variance to the Saturated Model.

```
Data set = Orobanche, Name of Fit = B1
Binomial Regression
Mean function = Logistic
Response      = y
Predictors    = (v h v*h)
Weights       = Extra-bin-var
Trials        = n
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE
Constant  -0.535411        0.193740       -2.764
v          0.0700884       0.311455        0.225
h          1.32979         0.278161        4.781
v.h       -0.819557        0.435205       -1.883

Sigma hat:                     1.
Number of cases:              21
Degrees of freedom:           17
Pearson X2:               17.000
Deviance:                 18.442
Extra variation           0.0249
```

Table 4: Fit Using Estimated Weights.

```
Data set = Orobanche, Name of Fit = B3
Binomial Regression
Mean function = Logistic
Response      = y
Predictors    = (v h)
Weights       = Extra-bin-var
Trials        = n
Coefficient Estimates
Label      Estimate        Std. Error     Est/SE
Constant  -0.377970        0.169103       -2.235
v         -0.351546        0.216039       -1.627
h          1.00542         0.210368        4.779

Sigma hat:                     1.
Number of cases:              21
Degrees of freedom:           18
Pearson X2:               21.135
Deviance:                 22.401
```

1. Initially estimate $\phi = [X^2 - (n-k)]/[\sum[(m_i - 1)(1 - h_i)]]$ assuming that $\phi = 0$, and so the $w_i = 1$.

2. Refit with weights $w_i = (1 + \phi(m_i - 1))^{-1}$. This will update the leverages h_i.

3. Re-estimate ϕ from

$$\phi = [X^2 - \sum\{w_i(1 - h_i)\}]/[\sum\{w_i(m_i - 1)(1 - h_i)\}]$$

4. Repeat 2 and 3 as needed, or until $X^2 \approx$ df.

5 References

Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. New York: Wiley.

Crowder, M. J. (1978). Beta-binomial anova for proportions. *Applied Statistics*, 27, 34–37.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, 31, 144–148.