MATH 523: QUASI-BINOMIAL MODELS

The quasi-binomial model assumes a mean-variance structure of the form

$$\begin{aligned} \mathbb{E}[Y_i] &= m_i\pi_i \\ \text{Var}[Y_i] &= \phi m_i\pi_i(1-\pi_i). \end{aligned}$$

for $i = 1, \ldots, n$. Setting $Y_i^\star = Y_i/m_i$, we have

$$\begin{aligned} \mathbb{E}[Y_i^\star] &= \pi_i \\ \text{Var}[Y_i^\star] &= \phi w_i\pi_i(1-\pi_i) = \phi w_i v(\pi_i) \end{aligned}$$

say, where $w_i = 1/m_i$. Under the logistic link, where

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i\beta = \eta_i$$

we have $g(t) = \log(t/(1-t))$, $\dot{g}(t) = 1/(t(1-t))$, so that the quasi-likelihood estimating equation is, as for the binomial case (see previous notes, and the handout on Iteratively Reweighted Least Squares)

$$\sum_{i=1}^n \frac{1}{w_i}\frac{(y_i^\star - \pi_i)}{\dot{g}(\pi_i)v(\pi_i)}x_i^\mathsf{T} = \sum_{i=1}^n m_i(y_i^\star - \pi_i)x_i^\mathsf{T} = \sum_{i=1}^n (y_i - m_i\pi_i)x_i^\mathsf{T} = 0_p$$

If β contains p parameters, the estimate of the dispersion parameter is

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^n \frac{(y_i^\star - \widehat{\pi}_i)^2}{w_i V(\widehat{\pi}_i)} = \frac{1}{n-p}\sum_{i=1}^n \frac{(y_i^\star - \widehat{\pi}_i)^2}{[\widehat{\pi}_i(1-\widehat{\pi}_i)/m_i]} = \frac{1}{n-p}\sum_{i=1}^n \frac{(y_i - m_i\widehat{\pi}_i)^2}{m_i\widehat{\pi}_i(1-\widehat{\pi}_i)}$$

In the standard Binomial model $\phi = 1$, so if $\widehat{\phi} > 1$, this is evidence of overdispersion.

If $\widehat{\phi} > 1$, then the standard errors of the estimates of the coefficients will be inflated, as the following example indicates. To generate overdispersion, set $n = 100, m_i = 20$, then

- $\pi_i \sim Beta(2,2)$
- $Y_i|\pi_i \sim Binomial(m_i, \pi_i)$

and then fit the constant response models.

```
n<-100
pi<-rbeta(n,2,2)
m<-20
Y<-rbinom(n,m,pi)

summary(glm(cbind(Y,m-Y)~1,family=binomial))
summary(glm(cbind(Y,m-Y)~1,family=quasibinomial))
```

The results are as follows:

```
> summary(glm(cbind(Y,m-Y)~1,family=binomial))

Call: glm(formula = cbind(Y, m - Y) ~ 1, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2276  -1.3077   0.0447   1.3964   4.4886

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.02000    0.04472  -0.447    0.655

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 445.06  on 99  degrees of freedom
Residual deviance: 445.06  on 99  degrees of freedom
AIC: 765.77

Number of Fisher Scoring iterations: 3


> summary(glm(cbind(Y,m-Y)~1,family=quasibinomial))

Call: glm(formula = cbind(Y, m - Y) ~ 1, family = quasibinomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2276  -1.3077   0.0447   1.3964   4.4886

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02000    0.08988  -0.223    0.824

(Dispersion parameter for quasibinomial family taken to be 4.038788)

    Null deviance: 445.06  on 99  degrees of freedom
Residual deviance: 445.06  on 99  degrees of freedom
AIC: NA
```

Note that the standard error doubles from 0.04472 (line 11) to 0.08988 (line 32), and that the estimate of the dispersion is $\widehat{\phi} = 4.0388$); indeed, the standard error increases by $\sqrt{\widehat{\phi}}$.