Recap: overdispersion etc.

$(y_{ij} - \bar{y}_i)$

- saturated model: $y_i \sim Bin(n_i, p_i), \quad \tilde{p}_i = y_i/n_i,$

  $\hat{\mu}_i = \bar{y}_i$

  $$\ell(\tilde{p}) = \Sigma\{y_i \log(y_i/n_i) + (n_i - y_i)\log(1 - y_i/n_i)\}$$

  $\bar{y}_i = \mu_i + \bar{\varepsilon}_i$

- what's the saturated model for linear regression? what is the maximized log-likelihood for this model? $i = 1, \dots, n$
- with binomial data, large-ish n_i, residual deviance compares regression model to saturated model
- if it's too large, we have the wrong model
- lack of independence among individual Bernoullis; a few outliers; wrong predictors ELM p. 43,4
- estimate $\tilde{\phi} = X^2/(n-p)$ ELM p. 45
- inflate variance $\hat{\beta} \,\dot{\sim}\, N(\beta, \tilde{\phi}(X^{\mathrm{T}}WX))$ instead of $N(\beta, X^{\mathrm{T}}WX)$

... overdispersion

```
> summary(bmod)

Call:
glm(formula = cbind(survive, total - survive) ~ location + period,
    family = binomial, data = troutegg)
...
period8      -2.3256     0.2429  -9.573  < 2e-16 ***
period11     -2.4500     0.2341 -10.466  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1021.469  on 19  degrees of freedom
Residual deviance:   64.495  on 12  degrees of freedom
AIC: 157.03

> summary(bmod2)

Call:
glm(formula = cbind(survive, total - survive) ~ location + period,
    family = quasibinomial, data = troutegg)

period8      -2.3256     0.5609  -4.146 0.001356 **
period11     -2.4500     0.5405  -4.533 0.000686 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 5.330358)
```

... overdispersion

SM §10.6, p.512

- $Y \mid \epsilon \sim \text{Bin}(m, \epsilon p)$
- $\text{E}(\epsilon) = 1, \quad \text{var}(\epsilon) = \xi$
- $\text{E}(Y) = \text{E}\{\text{E}(Y \mid \epsilon)\} = \text{E}(mp\epsilon) = mp$
- $\text{var}(Y) = \text{var}\{\text{E}(Y \mid \epsilon)\} + \text{E}\{\text{var}(Y \mid \epsilon)\}$

$$\text{var}(mp\epsilon) + E[mp\epsilon(1-p\epsilon)]$$

$$= m^2p^2\xi + mp - mp^2E\epsilon^2$$

- $\text{var}(Y) = m\{p(1-p) + \xi p^2(m-1)\}$ ntbc
- variance is larger than $mp(1-p)$ see also ELM p.44
- can't be detected if $m = 1$ m plays the role of n_i

... overdispersion

SM §10.6, p.512

- $Y \mid \epsilon \sim \text{Bin}(m, \epsilon p)$
- $E(\epsilon) = 1, \quad \text{var}(\epsilon) = \xi$
- $E(Y) = E\{E(Y \mid \epsilon)\} = E(mp\epsilon) = mp$
- $\text{var}(Y) = \text{var}\{E(Y \mid \epsilon)\} + E\{\text{var}(Y \mid \epsilon)\}$

$= mp(1-p) \cdot \phi \qquad \phi = ?$

- $\text{var}(Y) = m\{p(1-p) + \xi p^2(m-1)\}$ ntbc
- variance is larger than $mp(1-p)$ see also ELM p.44
- can't be detected if $m = 1$ m plays the role of n_i

... matched case-control studies

- suppose we have $1 : M$ matching one case, M matched controls
- for person i in matched set j, we have
  $y_{ij}, x_{ij}, \quad i = 0, 1, \ldots, M$

  $j = 1, \ldots, n$ y_{ij} $i = 0, .., M$
- model:

$$\log \frac{p_j(x_{ij})}{1 - p_j(x_{ij})} = \alpha_j + x_{ij}^{\mathrm{T}} \beta$$

$n + p$ pars

Mn people

- different intercept for each matched set confounding variables
- same effect of covariates across patients and sets β
- data: in matched set j, we have 1 case (person 0) and M controls (persons $1, \ldots, M$)
-

$$\begin{aligned} \Pr(y_{0j} = 1 \mid \Sigma_{i=1}^{M} y_{ij} = 1) &= \frac{\Pr(y_{0j} = 1, y_{1j} = 0, \ldots, y_{Mj} = 0)}{\Pr(y_{1j} = 0, \ldots, y_{Mj} = 0)} \\ &= \frac{\exp(x_{0j}^{\mathrm{T}} \beta)}{\Sigma_{i=0}^{M} \exp(x_{ij}^{\mathrm{T}} \beta)} \end{aligned}$$

... matched case-control studies

- suppose we have $1 : M$ matching one case, M matched controls
- for person i in matched set j, we have
  $y_{ij}, x_{ij}, \quad i = 0, 1, \ldots, M$
- model:

$$\log \frac{p_j(x_{ij})}{1 - p_i(x_{ij})} = \alpha_j + x_{ij}^{\mathrm{T}}\beta$$

- different intercept for each matched set confounding variables
- same effect of covariates across patients and sets β
- data: in matched set j, we have 1 case (person 0) and M controls (persons $1, \ldots, M$)
-

Fix j

$$\Pr(y_{0j} = 1 \mid \Sigma_{i=0}^{M} y_{ij} = 1) = \frac{\Pr(y_{0j} = 1, y_{1j} = 0, \ldots, y_{Mj} = 0)}{\Pr(y_{1j} = 0, \ldots, y_{Mj} = 0)}$$

Pr(case is case | ex. 1 case) $= \dfrac{\exp(x_{0j}^{\mathrm{T}}\beta)}{\Sigma_{i=0}^{M} \exp(x_{ij}^{\mathrm{T}}\beta)}$

... matched case-control studies

$$\Pr(Y_0 = 1 \mid Y_0 + Y_1 = 1)$$

$1 \quad 0 \mid 1$

$(1 \quad 0) \leftarrow \Sigma = 1$

$(0 \quad 1) \leftarrow$

$$= \frac{P(Y_0 = 1, Y_1 = 0)}{\Pr(Y_0 = 1, Y_1 = 0) + \Pr(Y_0 = 0, Y_1 = 1)}$$

$$= \frac{e^{\alpha + \beta x_0}}{e^{\alpha + \beta x_0} + e^{\alpha + \beta x_1}}$$

$$= \frac{\dfrac{e^{\alpha + \beta x_0}}{1 + e^{\alpha + \beta x_0}} \cdot \dfrac{1}{1 + e^{\alpha + \beta x_1}}}{e^{\alpha + \beta x_0} / 1 + e^{(\ )} \cdot 1 / 1 + e^{(\ )} \quad + \quad \cdots}$$