Reference: Agresti, Chapter 8

Slide 1

Quasi-likelihood methods

- Quasi-likelihood methods allow statistical modeling by making assumptions about the link function and the relationship between the first two moments, but without fully specifying the complete distribution of the response.
- Quasi-likelihood estimation specifies a link function and linear predictor like a GLM

$$g(\mu_i) = \sum_j \beta_j x_{ij}$$

but it does not assume a particular probability distribution for y_i.

Slide 2

- For a GLM, the score equations are

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{v(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0 \qquad j = 1, \ldots, p.$$

- The quasi-likelihood parameter estimates $\hat{\boldsymbol{\beta}}$ are the solutions of **quasi-score** equations that resemble the score equations of GLMs with $v(\mu_i)$ replaced by the appropriate variance function.
  - For example, for count data, we may set

$$v(\mu_i) = \phi \mu_i$$

  for some unknown constant ϕ.
- The quasi-score equations are not score equations without the extra assumption that y_i has a distribution in the exponential family.

Slide 3

- QL estimators have similar properties with ML estimators.
- The QL estimator $\hat{\boldsymbol{\beta}}$ is asymptotically normal with covariance matrix approximated by

$$var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'W\boldsymbol{X})^{-1}$$

which is ϕ times the variance from the ordinary GLM, since

$$w_i = \frac{1}{var(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \frac{1}{\phi v(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

Slide 4

Quasi-likelihood approach of variance inflation

A simple quasi-likelihood approach consists of

1. Fit the ordinary GLM
2. Multiply the SE of the $\hat{\beta}_j$ by $\sqrt{\chi^2/(n-p}$

The motivation is the following:

- when a variance function has the form $\phi v(\mu_i)$, the corresponding Pearson χ^2 statistic is

$$\frac{1}{\phi}\sum_{i=1}^{N}\frac{(y_i-\hat{\mu}_i)^2}{\phi v(\hat{\mu}_i)} = \frac{1}{\phi}\chi^2 \sim \ \chi^2_{N-p}$$

when there are p parameters in the linear predictor. Thus,

$$E[\chi^2/\phi] \approx N-p \qquad \Rightarrow \qquad E[\chi^2/(N-p)] \approx \phi$$

Slide 5

leading to

$$\hat{\phi} = \frac{\chi^2}{N-p}$$

Note that this method is appropriate only if the model describes well the relationship between $E[y_i]$ and the explanatory variables.

Slide 6

Quasi-likelihood variance-inflation for count data

For overdispersed count data with a mean-variance relation of the form

$$v(\mu_i) = \phi\mu_i$$

the adjusted covariance matrix for $cov(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'W\boldsymbol{X})^{-1}$ with

$$w_i = \frac{1}{var(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 = \frac{\mu_i^2}{\phi\mu_i} = \frac{\mu_i}{\phi}.$$

The Pearson statistic is

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

and the variance-inflation estimate is

$$\hat{\phi} = \frac{\chi^2}{N-p}.$$

Slide 7

Example: Let's consider the horseshoe crab satellite data and use the female crab's weight to predict the number of male satellites.

```
> attach(Crabs)
fit.pois = glm(satellite ~ weight, family=poisson)

> summary(fit.pois)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.42841    0.17893  -2.394   0.0167 *
weight       0.58930    0.06502   9.064   <2e-16 ***
---
    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 560.87  on 171  degrees of freedom
AIC: 920.16

# variance inflation estimate
```

Slide 8

```
X2.pois = sum(resid(fit.pois, type="pearson")^2)
phi.hat = X2.pois/df.residual(fit.pois)
> phi.hat
[1] 3.133893
```

Slide 9

- The Poisson GLM equation is

$$\log \hat{\mu}_i = -0.428 + 0.589 \; weight_i$$

  with $SE(\hat{\beta}_1) = 0.065$.

- The variance inflation estimate is

$$\hat{\phi} = \frac{\chi^2}{N-p} = \frac{535.9}{171} = 3.13$$

- A more plausible standard error for $\hat{\beta}_1$ account for overdispersion is

$$SE(\hat{\beta}_1) = \sqrt{\hat{\phi}} \times 0.065 = 0.115$$

Slide 10

R can do this calculation for us if we use the `quasi` family:

```
fit.quasi = glm(satellite ~ weight, family=quasi(link="log", variance="mu"))

> summary(fit.quasi)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4284     0.3168  -1.352    0.178
weight        0.5893     0.1151   5.120 8.17e-07 ***
---
(Dispersion parameter for quasi family taken to be 3.134159)
```

Slide 11

Quasi-likelihood variance-inflation for binary data

- The inflated-variance quasi likelihood approach is appropriate only for grouped binary data.
- It uses variance function for the proportions y_i

$$v(\pi_i) = \phi \frac{\pi_i(1-\pi_i)}{n_i}$$

- The quasi likelihood estimates are the same as the ML estimates for the binomial GLM and the asymptotic covariance is multiplied by ϕ.
- The Pearson statistic and the variance-inflation estimate are

$$\chi^2 = \sum_i \frac{(y_i - \hat{\pi}_i)^2}{[\hat{\pi}_i(1-\hat{\pi}_i)]/n_i} \qquad \hat{\phi} = \frac{\chi^2}{N-p}$$

Slide 12

Quasi-likelihood for correlated Bernoulli trials

Let $y_{i1}, y_{i2}, \ldots, y_{in_i}$ denote the n_i Bernoulli trials for observation i, $y_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and let $\rho = cor(y_{i,j}, y_{i,k})$ for $j \neq k$

$$\begin{aligned}
Var(y_i) &= Var\left(\frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}\right) \\
&= \frac{1}{n_i^2}\left[\sum_{j=1}^{n_i} Var(y_{ij}) + 2\sum_{j=1}^{n_i}\sum_{k=1}^{j-1} Cov(y_{ij}, y_{ik})\right] \\
&= \frac{1}{n_i^2}\left[n_i\pi_i(1-\pi_i) + 2\binom{n_i}{2}\rho\pi_i(1-\pi_i)\right] \\
&= \frac{\pi_i(1-\pi_i)}{n_i}\left[1 + \rho(n_i - 1)\right]
\end{aligned}$$

Note that this variance function is similar to the one obtained using the beta-binomial model.

Slide 13

- The quasi-likelihood approach can use this variance function

$$v(\mu_i) = \frac{\pi_i(1-\pi_i)}{n_i}\left[1+\rho(n_i-1)\right], \qquad |\rho| \le 1.$$

- The estimates using this approach differ from the ML estimates because the multiple of the binomial variance does not drop out of the quasi-likelihood scores.
- An iterative two-step process can be used:
  1. solve the quasi-likelihood score equations for $\boldsymbol{\beta}$ for a given $\hat{\rho}$
  2. solve for $\hat{\rho}$ using the updated $\hat{\boldsymbol{\beta}}$ in the following equation

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \hat{\pi}_i)^2}{\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i}\left[1+\hat{\rho}(n_i-1)\right]} = N - p.$$

Slide 14

Example: Let's consider the teratology example saved under `Rats.txt` that we used in the beta-binomial lecture. Recall that

- y_{ij} denote the proportion of dead among the n_{ij} fetuses in litter j for treatment group i.
- z_{ij} be an indicator for the placebo group, i.e., $z_{ij} = 1$ if litter j got placebo and 0 otherwise.
- h_{ij} is the hemoglobin level for litter j in treatment i.

Slide 15

The logistic regression output is

```
 Rats$placebo = ifelse(group==1, 1, 0)
 fit.logit = glm(s/n ~ placebo+h, weights=n, family=binomial, data=Rats)

> summary(fit.logit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.62391    0.78996  -0.790   0.4296
placebo      2.65088    0.48238   5.495  3.9e-08 ***
h           -0.18713    0.07428  -2.519   0.0118 *
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 509.43  on 57  degrees of freedom
Residual deviance: 170.57  on 55  degrees of freedom
AIC: 248.04
```

Slide 16

- The Pearson χ^2 statistic is given by

  ```
  > sum(resid(fit.logit, type="pearson")^2)
  [1] 159.815
  ```

  with $df = 58 - 3 = 55$.

- The variance inflation estimate is

$$\hat{\phi} = \frac{\chi^2}{N-p} = \frac{159.815}{55} = 2.906$$

  so the standard errors for $\hat{\boldsymbol{\beta}}$ accounting for overdispersion are

  $SE(\hat{\beta}_1) = \sqrt{2.906} \times 0.4824 = 0.822 \qquad SE(\hat{\beta}_2) = \sqrt{2.906} \times 0.0743 = 0.127$

Slide 17

Using the `quasi` family in R:

```
fit.quasi = glm(s/n ~ placebo+h, weights=n,
       family=quasi(link="logit", variance="mu(1-mu)"), data=Rats)

 > summary(fit.quasi)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6239     1.3466  -0.463  0.64495
placebo       2.6509     0.8223   3.224  0.00213 **
h            -0.1871     0.1266  -1.478  0.14514
---
(Dispersion parameter for quasi family taken to be 2.905728)

    Null deviance: 509.43  on 57  degrees of freedom
Residual deviance: 170.57  on 55  degrees of freedom
AIC: NA
```

Slide 18

Using the quasi-likelihood approach with a beta-binomial type variance gives $\hat{\rho} = 0.1985$:

```
library(aod)
fit.quasibb = quasibin(cbind(s, n-s) ~ placebo+h, data=Rats)

> fit.quasibb
Quasi-likelihood generalized linear model
-----------------------------------------
quasibin(formula = cbind(s, n - s) ~ placebo + h, data = Rats)

Fixed-effect coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.7237     1.3785 -0.5250   0.5996
placebo       2.7573     0.8522  3.2355   0.0012
h            -0.1758     0.1284 -1.3692   0.1709
```

Slide 19

```
Overdispersion parameter:
   phi
0.1985
```

Slide 20

Recall that the beta-binomial led to

```
library(VGAM)
fit.betabin = vglm(cbind(s, n-s) ~ placebo+h, betabinomial, data=Rats)

> summary(fit.betabin)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  -0.5009     1.1907  -0.421 0.673974
(Intercept):2  -1.1676     0.3251  -3.592 0.000328 ***
placebo         2.5601     0.7642   3.350 0.000807 ***
h              -0.1546     0.1085  -1.424 0.154354
```

Slide 21

The following table summarizes the results for the four analyses:

Parameter	logistic GLM	QL with $v(\pi_i) = \phi \frac{\pi_i(1-\pi_i)}{n_i}$	QL with $v(\pi_i) = \frac{\pi_i(1-\pi_i)}{n_i}[1+\rho(n_i-1)]$	beta-binomial
Intercept	-0.62 (0.79)	-0.62 (1.35)	-0.72 (1.38)	-0.50 (1.19)
Placebo	2.65 (0.48)	2.65 (0.82)	2.76 (0.85)	2.56 (0.76)
Hemoglobin	-0.19 (0.07)	-0.19 (0.13)	-0.18 (0.13)	-0.15 (0.11)
Overdispersion	None	$\hat{\phi} = 2.906$	$\hat{\rho} = 0.1985$	$\hat{\rho} = 0.237$

The QL approaches and the beta-binomial model have similar standard errors, much larger than the logistic regression.