BIOSTAT 651
Notes #14: Overdispersion (revised)

- Lecture Topics:
  - Causes of overdispersion
  - Estimating scale parameter
  - Random effect models
  - Generalized estimating equations

Overdispersion

- *Overdispersion*: variance exceeds that under the assumed model
  - e.g., Binomial data

    $Var(Y_i) > v(\mu) = n\mu(1 - \mu)$

  - e.g., Poisson response

    $Var(Y_i) > v(\mu) = \mu$

- Under-dispersion can also occur
  - less common

Overdispersion: Causes

- Overdispersion can result for several reasons
  - heterogeneous populations
  - unmeasured covariate
  - events *within cell* are correlated

Overdispersion: Causes

- Example: Population Heterogeneity
  - Suppose there exists a binary covariate, Z_i, and that

$$
\begin{aligned}
Y_i|Z_i = 0 &\sim Poisson(\lambda_0) \\
Y_i|Z_i = 1 &\sim Poisson(\lambda_1) \\
P(Z_i = 1) &= \pi
\end{aligned}
$$

$$
\begin{aligned}
E(Y_i) &= \pi\lambda_1 + (1-\pi)\lambda_0 = \mu \\
Var(Y_i) &= E(\lambda_1 Z_i + \lambda_0(1 - Z_i)) \\
&+ Var(\lambda_1 Z_i + \lambda_0(1 - Z_i)) \\
&= \mu + (\lambda_1 - \lambda_0)^2\pi(1-\pi)
\end{aligned}
$$

Overdispersion: Impact on Analysis

- As implied, overdispersion generally involves the assumed variance structure being inconsistent with that actually underlying the data
  - in GLMs, $\widehat{\boldsymbol{\beta}}$ is generally unbiased
  - however, $\widehat{SE}(\widehat{\beta}_j)$ may be substantially biased
- As a result:
  - hypothesis tests tend to be anti-conservative
  - CIs tend to be artificially narrow
- For under-dispersion, effect is in the opposite direction

Overdispersion: Case of Poisson Model

- Any GLM is subject to misspecification
- Poisson model is especially susceptible

Accommodating Overdispersion

- Several methods have been developed for handling overdispersion
  - estimating scale parameter (quasi-likelihood)
  - random effects models
  - generalized estimating equations

- Each method involves modifying the original
  - (i) model assumptions
  - (ii) estimation procedures

Quasi-likelihood

- By our previously derived Exponential family and GLM results:

$$V(Y_i) \quad = \quad a(\phi)\, v(\mu_i)$$

- Suppose we relax the assumption that $a(\phi) = 1$
  - e.g., common to assume that $a(\phi) = \phi$,
    $E[Y_i] = \mu_i$
    $V(Y_i) = \phi\, v(\mu_i)$
  - Note: Impact on score and information:

    $U(\boldsymbol{\beta}) = \frac{1}{\phi} X^T (Y - \mu)$

    $J(\boldsymbol{\beta}) = \frac{1}{\phi} X^T V X$

Estimating Scale Parameter

- As implied, $\widehat{\boldsymbol{\beta}}$ can still be computed by solving

$$\mathbf{X}^T(\mathbf{Y} - \boldsymbol{\mu}) = \mathbf{0}$$

- We now need a method for estimating ϕ
- Pearson Chi-square statistic:

$$X_P^2(\phi) = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{V}(Y_i)}$$

  - under some conditions

$$X_P^2(\phi) \sim \chi^2_{n-q}$$

Estimating Scale Parameter (continued)

- Then, using the fact that

$$V(Y_i) \quad = \quad a(\phi)\, v(\mu_i)$$

and the assumption that

$$a(\phi) \quad = \quad \phi$$

along with MoM concepts suggests setting

$$X_P^2(\phi) = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{\phi\, v(\widehat{\mu}_i)} \quad \approx \quad (n-q)$$

which implies the Pearson-based scale estimator:

$$\widehat{\phi}_P \quad = \quad \frac{X_P^2(1)}{n-q}$$

Estimating Scale Parameter: Examples

- e.g., $Y_i \sim \text{Poisson}(\mu_i)$:

  $\widehat{\phi}_P =$

- e.g., $Y_i \sim \text{Binomial}(n_i, \pi_i)$

  $\widehat{\phi}_P =$

- Note: can also use an estimator based on the Deviance:

$$\widehat{\phi}_D = \frac{D}{n-q}$$

  - often gives similar results

Impact of Scale Parameter on Inference

- Estimation of $\boldsymbol{\beta}$ is unaffected

  Q: Why?

- Standard errors are modified:

  uncorrected: $\widehat{SE}(\widehat{\beta}_j)$

  corrected:

Dispersion Parameter: SAS Code

- Easy to estimate ϕ in SAS

  e.g., Poisson regression:

```
PROC GENMOD DATA=dialysis;
 MODEL admits = diab age / DIST=Poisson LINK=log
                           OFFSET=log_yrs
                           PSCALE;
RUN;
```

- For $\widehat{\phi}_D$ estimator, use DSCALE option

Random effect model

- To use a hierarchical random effect model for over dispersion data
  - Binomial data: Beta-binomial regression
  - Count data: Negative-binomial regression

Negative binomial regression

- Introduce a random effect term to model extra variation.

$$
\begin{aligned}
Y_i|\theta_i &\sim Poisson(\theta_i) \\
\theta_i &= \exp(X_i\beta + \epsilon_i) \\
&= \exp(X_i\beta)\exp(\epsilon_i) = \mu_i z_i \qquad (1)
\end{aligned}
$$

- z_i follows gamma distribution $\Gamma(shape = \delta, rate = \delta)$, so $E(z_i) = 1$

Negative binomial regression

$$P(Y_i = y|\mu_i, \delta) \quad = \quad \frac{\Gamma(\delta + y)}{\Gamma(\delta)\Gamma(y+1)} \left(\frac{\delta}{\delta + \mu_i} \right)^{\delta} \left(\frac{\mu_i}{\delta + \mu_i} \right)^{y}$$

- Mean and variance:

$$\begin{aligned} E(Y_i) &= \mu_i = \exp(X_i \beta) \\ Var(Y_i) &= \mu_i + \frac{1}{\delta}\mu_i^2 \end{aligned}$$

- Additional parameter δ in the variance.

Negative binomial regression: SAS code

- Use dist=negbin

```
PROC GENMOD DATA=dialysis;
 MODEL admits = diab age / DIST=negbin
                           OFFSET=log_yrs;
RUN;
```

Generalized Estimating Equations

- Consider the score function for a canonical GLM,

$$U(\boldsymbol{\beta}) \quad = \quad \frac{1}{a(\phi)}\mathbf{X}^T(\mathbf{Y} - \boldsymbol{\mu})$$

  obtained through standard ML theory

- Suppose now that the model assumptions are in question
  - e.g., quite possible that $E[Y_i] = \mu_i$, but that other properties connected with the GLM may not hold
  - e.g., Poisson case: variance structure

GEE: Poisson Case

- Suppose that Y_i is an event count
  - seems natural to assume $Y_i \sim \text{Poisson}(\mu_i)$
  - and, under a (canonical) GLM:
    $E[Y_i] = T_i \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}$
- The effect of covariates on $E[Y_i] = \mu_i$ is of chief interest

- However, $V(Y_i)$ is likely of little inherent interest
  - assumptions regarding $V(Y_i)$ are best avoided
  - In GLM we assume $V(Y_i) = \mu_i$
  - need to do a valid inference when $V(Y_i) \neq \mu_i$

GEE: Intro

- Recall: Method-of-Moments estimation
  - equate sample and population moments
  - solve for parameters of interest
- GEE (Liang & Zeger, 1986): Solve

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} D_i^T V_i^{-1}(Y_i - \mu_i) = 0$$

  - $D_i = \partial\mu_i / \partial\beta_i^T$
  - V_i: assumed variance of Y_i
- GEE can be viewed as a regression analog of MoM
  - i.e., provided that the model for the mean structure is correct,

    $S(\beta)$ is a zero-mean estimating equation

- In univariate data, $S(\beta)$ is the same as the score function $U(\boldsymbol{\beta})$ when V_i is an assumed variance in GLM
  - $D_i = 1/g'(\mu_i)\mathbf{x}_i^T$
  - $V_i = a(\phi)v(\mu_i)$

$$
\begin{aligned}
S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{Y_i - \mu_i}{a(\phi)v(\mu_i)g'(\mu_i)}\mathbf{x}_i \\
&= U(\boldsymbol{\beta}) \qquad (2)
\end{aligned}
$$

- With canonical link function $(v(\mu_i) = 1/g'(\mu_i))$

$$
\begin{aligned}
S(\boldsymbol{\beta}) &= \frac{1}{a(\phi)} \sum_{i=1}^{n} (Y_i - \mu_i)\mathbf{x}_i \\
&= \frac{1}{a(\phi)} \mathbf{X}^T(\mathbf{Y} - \boldsymbol{\mu}) \qquad (3)
\end{aligned}
$$

GEE: Applicability

- GEE is usually thought of as a method for correlated responses
  - MLE requires a fully-specified model
  - not so easy in some cases . . .

- GEE only requires specification of the *mean* and *covariance* terms
  - consistency of $\widehat{\boldsymbol{\beta}}$ requires that the mean be modeled correctly
  - does *not* require that covariance be correctly specified

- More generally, GEE can be used with univariate data, in order to avoid pitfalls of model mispecification

GEE: Properties

- Key GEE result: a zero-mean estimating equation should yield a consistent estimator of $\boldsymbol{\beta}$
  - requires that the assumptions that lead to the zero-mean property hold
  - note: *not* required that the estimating equation be based on ML
- GEE $S(\beta)$ is a zero-mean estimating equation

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} D_i^T V_i^{-1} (Y_i - \mu_i) = 0$$

  - $E(S(\beta_0)) = 0$ has mean zero provided that $E[Y_i|\mathbf{x}_i] = \mu_i$
  - This mean zero property does not depend on V_i
  - If $V_i = a(\phi)v(\mu_i)$ in GLM, $S(\boldsymbol{\beta})$ is the same as the score function from the log-likelihood

- In GEE, $\boldsymbol{\beta}$ is estimated by the solution to $S(\boldsymbol{\beta}) = \mathbf{0}$ without framing S as the derivative of the log-likelihood function
  - note: standard ML-based inference procedures do not apply to GEE estimators

GEE: Inference

- Provided that $E[S(\boldsymbol{\beta}_0)] = \mathbf{0}$ and under some (mild) regularity conditions:
  - $\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$
  - $V(\widehat{\boldsymbol{\beta}}) \approx H(\boldsymbol{\beta}_0)$

$$
\begin{aligned}
H(\boldsymbol{\beta}_0) &= H_1(\boldsymbol{\beta}_0)^{-1} H_2(\boldsymbol{\beta}_0) H_1(\boldsymbol{\beta}_0)^{-1} \\
H_1(\boldsymbol{\beta}_0) &= \sum_{i=1}^{n} D_i^T V_i^{-1} D_i \\
H_2(\boldsymbol{\beta}_0) &= \sum_{i=1}^{n} D_i^T V_i^{-1} Var(Y_i) V_i^{-1} D_i
\end{aligned}
$$

- When V_i is correctly specified: $V_i = Var(Y_i)$
  - $H_1(\boldsymbol{\beta}_0)^{-1} H_2(\boldsymbol{\beta}_0) H_1(\boldsymbol{\beta}_0)^{-1} = H_1(\boldsymbol{\beta}_0)$
- Canonical link with $V_i = Var(Y_i)$:
  - $H_1(\boldsymbol{\beta}_0) = X^T V X / a(\phi) = J(\boldsymbol{\beta}_0)$
- If V_i is misspecified, the variance would be larger, so the efficiency decreases.

GEE: Inference (continued)

- When estimated through GEE, $V(\widehat{\boldsymbol{\beta}})$ can be estimated by the *robust* variance estimator,

$$\widehat{V}(\widehat{\boldsymbol{\beta}}) = H_1(\widehat{\boldsymbol{\beta}})^{-1}\widehat{H}_2(\widehat{\boldsymbol{\beta}})H_1(\widehat{\boldsymbol{\beta}})^{-1}$$

$$\widehat{H}_2(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} D_i^T V_i^{-1}(y_i - \widehat{\mu}_i)(y_i - \widehat{\mu}_i)^T V_i^{-1} D_i$$

  - also known as the *sandwich* estimator

Generalized Estimating Equations: SAS Code

- *Working independence* assumption:

  e.g., Poisson regression GEE:

```
PROC GENMOD DATA=dialysis;
 CLASS idnum;
 MODEL admits = diab age / DIST=Poisson LINK=log
                           OFFSET=log_yrs;
 REPEATED SUBJECT=idnum / TYPE=ind;
RUN;
```

GEE: Efficiency

- Increase efficiency by correctly specifying the variance
- For longitudinal data (correlated outcomes), typically

$$\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}$$

  - several options for specifying $\mathbf{R}_i$
  - correct $\mathbf{R}_i$ increase the efficiency
  - Although you misspecified $\mathbf{R}_i$, you still can do a valid inference.