Overdispersion

Motivation

Overdispersion: True variance in data exceeds that under assumed model - especially common with count data.

Examples:

- Binomial $V[Y_i] > np(1-p)$
- Poisson $V[Y_i] > \mu_i$

Causes

- Heterogenous populations/ unmeasured covariate
- events within cell are correlated

Effects

- $\hat{SE}(\hat{\boldsymbol{\beta}})$ may be substantially biased
- hypothesis tends to be anti-conservative
- CIs tend to be artificially narrow
- underdispersion causes the effect in the opposite direction

Detection of Dispersion

if adjusted pearson chi-square or adjusted deviance is much larger than 1, disperson problem is apparent

There are three solutions to overdisperson: (1) Estimate Scale Parameter, (2) Random Effect Model, (3) Generalized Estimating Equation

Estimate Scale Parameter

Recall the definition that $V[Y_i] = a(\phi)v(\mu_i)$, where $a(\phi) = 1$.
We now assume that $a(\phi) = \phi$

$$
\begin{aligned}
a(\phi) = \phi &\Rightarrow E[Y_i] = \mu_i; V[Y_i] = \phi v(\mu_i) \\
U(\boldsymbol{\beta}) &= \frac{1}{\phi}\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{\mu}) \\
J(\boldsymbol{\beta}) &= \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}
\end{aligned}
$$

Estimating the Scale Parameter

- Nothing changes in the estimation of the point estimate of $\hat{\boldsymbol{\beta}}$
- We now need a method for estimating ϕ
- Pearson Chi-Square statistic is used

$$X_P^2(\phi) = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\hat{V}(Y_i)} \sim \chi^2_{n-q}$$

Since $V(Y_i) = a(\phi)v(\mu_i) \Rightarrow$

$$\begin{aligned} X_P^2(\phi) &= \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\hat{V}(Y_i)} \\ &= \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\phi v(\hat{\mu}_i)} \approx (n-q) \\ \Longleftrightarrow \ \hat{\phi}_P &= \frac{X_P^2(\phi = 1)}{n-q} \end{aligned}$$

- Equivalently, the Deviance scale estimator is

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

Examples

- $Y_i \sim$ Poisson (μ_i)

  $\hat{\phi}_P = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \frac{1}{n-q}$
- $Y_i \sim$ Binomial (n_i, π_i)

  $\hat{\phi}_P = \sum_{i=1}^{n} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)} \frac{1}{n-q}$
- Corrected Standard Errors of $\hat{\boldsymbol{\beta}}$ becomes:

  corrected SE $= \sqrt{\hat{\phi}}\hat{SE}(\hat{\beta})$

Using the `PSCALE` or `DSCALE` options in proc genmod in SAS will adjust for these automatically.

Random Effect Model

Two examples possible within the context of our class, only Negative Binomial explored here:

1. Binomial data : beta - binomial regression
2. Count data : negative - binomial regression

Negative Binomial Hierarchical Model

- Model: $Y_i|\theta_i \sim$ Poisson (θ_i)

- $\theta_i = \exp\{\boldsymbol{X}\beta + \epsilon_i\} = e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\exp\{\epsilon_i\} = \mu_i z_i$
- Assume $z_i \sim$ Gamma(shape=δ, rate = δ) ; $E[z_i] = 1$

Note that this implies that

$$P(Y_i = y|\mu_i, \delta) = \frac{\Gamma(\delta + y)}{\Gamma(\delta)\Gamma(y+1)}\Big(\frac{\delta}{\delta + \mu_i}\Big)^{\delta}\Big(\frac{\mu_i}{\delta + \mu_i})^{y} \qquad (*)$$

with mean and variance

$$E[Y_i] = \mu_i = e^{X_i\boldsymbol{\beta}} \qquad\qquad V[Y_i] = \mu_i + \frac{1}{\delta}\mu_i^2$$

Example SAS code

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

Generalized Estimating Equation

The intuition behind GEE is to equate sample and population moments, and then solve for parameters of interest [1]

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

$$D_i = \frac{\partial u_i}{\partial \beta_i^T}$$

V_i : assumed variance of Y_i

Properties

- In univariate data, $S(\boldsymbol{\beta})$ is the same as the score function $U(\boldsymbol{\beta})$ when V_i is an assumed variance in GLM

  $D_i = \frac{1}{g'(\mu_i)}\boldsymbol{x}_i$

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

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

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

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

[1] Liang & Zeger, 1986

Applicability

- MLE requires a lot of structure; GEE less so
- GEE's advantage is that it only requires specification of the *mean* and *covariance*
- Consistency of $\hat{\boldsymbol{\beta}}$ requires mean modeled correctly, but not covariance
- GEE is also used for handling correlated responses
- GEE can be used with univariate data, to avoid pitfalls of model misspecification

Model Fitting

- $\hat{\boldsymbol{\beta}}$ is derived by finding the solution of $S(\boldsymbol{\beta}) = 0$
- Provided that $E[S(\hat{\beta}_0)] = \mathbf{0}$, and assuming mild regularity conditions

$$\hat{\boldsymbol{\beta}}_0 \to \boldsymbol{\beta}_0$$

$V(\hat{\boldsymbol{\beta}}) \approx H(\boldsymbol{\beta}_0)$ where

$$H(\boldsymbol{\beta}_0) = H_1(\boldsymbol{\beta}_0)^{-1} H_2(\boldsymbol{\beta}_0) H_1(\boldsymbol{\beta}_0)^{-1}$$

$$H_1(\boldsymbol{\beta}_0) = \sum_{i=1}^{n} D_i^T V_i^{-1} D_i$$

$$H_2(\boldsymbol{\beta}_0) = \sum_{i=1}^{n} D_i^T V_i^{-1} V[Y_i] V_i^{-1} D_i$$

$$V_i = V[Y_i] \Rightarrow H_1(\boldsymbol{\beta}_0)^{-1} H_2(\boldsymbol{\beta}_0) H_1(\boldsymbol{\beta}_0) = H_1(\boldsymbol{\beta}_0)$$

- Using the canonical link would imply that $V_i = V[Y_i]$

$$H_1(\boldsymbol{\beta}_0) = \frac{1}{a(\phi)} \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X} = J(\boldsymbol{\beta}_0)$$

- if V_i is misspecified the variance would be larger, the efficiency would decrease
- when estimated through GEE $V(\hat{\boldsymbol{\beta}})$ is estimated by the *robust* variance estimator (also called sandwhich estimator)

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

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

Sample Code:

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

Efficiency

- Increase efficiency by correctly specifying the variance
- for longitudinal data (correlated outcomes) change specification to

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

- There are several options for specifying $\boldsymbol{R}_i$
- correct $\boldsymbol{R}$ increases efficiency, misspecifying doesn't ruin validity of inference

Worked Example: Seizure Data

A clinical trial was conducted in order to evaluate the impact of Progabide on the frequency of epileptic seizures. Patients were randomized to either receive or not receive Progabide. The data set contains information on:

- Age at start of study; AGE
- baseline seizure count ; BASE
- treatment indicator ; Z
- seizure counts in each of 4 two-week periods: $Y_1, ...Y_4$
- outcome $Y_i = \sum_{j=1}^{4} Y_{ij}$

Model Fitting Questions

a Fit a Poisson Regression Model

```
title "Poisson regression";
proc genmod data=seizure1;
model Y_tot = age base Z / dist=Poisson
link=log;
run;
```

b Evidence of Dispersion?

Yes- via a high Adjusted Deviance and Adjusted Pearson $\chi^2 \approx 11 >> 1$

c Re-estimate and re-test the treatment effect by estimating the scale parameter

```
title "Quasi-likelihood";
proc genmod data=seizure1;
model Y_tot = age base Z / dist=Poisson
link=log Pscale; run;
```

d Negative Binomial Regression

```
title
"Negative binomial regression";
proc genmod data=seizure1;
model Y_tot = age base Z / dist=negbin
link=log;
run;
```

e Generalized Estimate Equation

```
title
"GEE";
proc genmod data=seizure1;
class idnum;
model Y_tot = age base Z / dist=Poisson
link=log;
repeated subject=idnum / type=ind ;
run;
```