Further notes on GEE

- Intuitively: to restore the unbiasedness of the estimating equation for the complete population we need to weight the contribution of Y_{ij} by the inverse of π_{ij}.
- For unbiasedness of the estimating equation we require consistent estimation of the dropout probabilities, given the history of responses and covariates.
- The method can be extended to the case of informative dropout, in the form of a sensitivity analysis.

2009 Jon Wakefield, Stat/Biostat 571

Bayesian Inference via Data Augmentation

Data augmentation is a auxiliary variable method that treats the missing observations as unknown parameters – this can lead to simple MCMC schemes.

General formulation: we have posterior

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{Y}^M \mid \boldsymbol{Y}^O) &= p(\boldsymbol{\theta} \mid \boldsymbol{Y}^M, \boldsymbol{Y}^O) p(\boldsymbol{Y}^M \mid \boldsymbol{Y}^O) \\ &= p(\boldsymbol{Y}^M \mid \boldsymbol{\theta}, \boldsymbol{Y}^O) p(\boldsymbol{\theta} \mid \boldsymbol{Y}^O) \end{aligned}$$

MCMC scheme:

1. Auxiliary variables:
$$\boldsymbol{Y}^M \sim p(\boldsymbol{Y}^M \mid \boldsymbol{Y}^O, \boldsymbol{\theta}).$$
2. Model parameters:
$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{Y}^O, \boldsymbol{Y}^M).$$

The auxiliary variable scheme may be modified to $p(\boldsymbol{Y}^M \mid \boldsymbol{Y}^O, \boldsymbol{\theta}) \sim p(\boldsymbol{Y}^M \mid \boldsymbol{\theta})$, depending on the missing data model, as we now illustrate.

Example: Censoring Model

Suppose we have data Y_i measured at times t_i, $j = 1, ..., n$, but measurements *below the lower limit of detection*, D (assumed known) are not recorded. Also suppose that the data generating model (likelihood) is:

$$Y \mid \boldsymbol{\beta}, \sigma \sim_{ind} \mathrm{N}(\eta(\boldsymbol{\beta}, t), \sigma^2).$$

Clearly setting such measurements to zero or ignoring the measurements will lead to bias in estimation.

Figure 42 illustrates for a set of simulated data in which the true slope was -0.01; the slope estimates are -0.0099, -0.0095 and -0.0087 for the full data, set equal to D and ignored schemes, respectively.

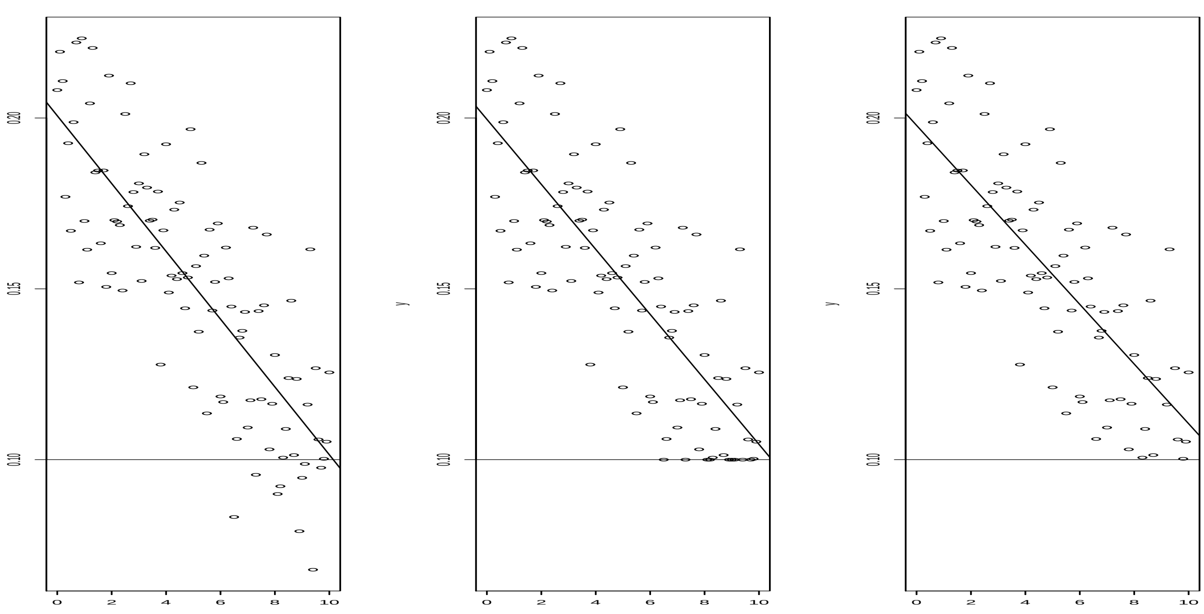

Figure 42: All data (left), assigned to lower limit (middle), ignored (right). Horizontal line is the lower limit of detection.

Suppose that the last c measurements are censored, the remaining $n-c$ being uncensored. Then

$$\begin{aligned} p(\boldsymbol{y} \mid \theta) &= \prod_{i=1}^{n-c} p(y_i \mid \boldsymbol{\beta}, \sigma^2) \prod_{i=c+1}^{n} \Pr(Y_i < D \mid \boldsymbol{\beta}, \sigma^2) \\ &= \prod_{i=1}^{n-c} \phi\left(\frac{y_i - \eta(\boldsymbol{\beta}, t_i)}{\sigma}\right) \prod_{i=c+1}^{n} \Phi\left(\frac{D - \eta(\boldsymbol{\beta}, t_i)}{\sigma}\right) \end{aligned}$$

where

$$\phi(z) = (2\pi)^{-1/2} \exp(-z^2/2)$$

and

$$\Phi(z_0) = \Pr(Z < z_0) = \int_{-\infty}^{z_0} \phi(z)\, dz$$

where Z is an N$(0,1)$ random variable.

To perform likelihood or Bayesian inference we need to numerically evaluate the distribution function of a normal distribution for each likelihood calculation.

Data Augmentation Scheme

Letting $\boldsymbol{Y}^O = \{Y_i, i = 1, ..., n-c\}$ and $\boldsymbol{Y}^M = \{Y_i, i = n-c+1, ..., n\}$, we iterate between

1. $y_i \mid \boldsymbol{\beta}, \sigma \sim \text{TruncNorm}(\eta(\boldsymbol{\beta}, t_i), \sigma^2)$, on $(-\infty, D)$, $i = n-c+1, ..., n$.
2. $\boldsymbol{\beta} \mid y_1, ..., y_n, \sigma^2 \propto \prod_{i=1}^n p(y_i \mid \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta})$. Usual (uncensored) posterior.
3. $\sigma^2 \mid y_1, ..., y_n, \boldsymbol{\beta} \propto \prod_{i=1}^n p(y_i \mid \boldsymbol{\beta}, \sigma^2)\pi(\sigma^2)$. Usual (uncensored) posterior.

CONCLUSIONS

We have looked at regression modeling for dependent data and have examined three approaches to inference:

1. Likelihood-based:
   - Likelihood inference.
   - Bayesian inference.
2. Generalized Estimating Equations.

Issues:

- Assumptions for valid inference.
- Efficiency.
- Computation.
- Parameter interpretation.
- Flexibility in dealing with different types of or missing data.

Likelihood Approach

We have examined Mixed Effects Models in which random effects are introduced to induce dependencies.

Non-Linear Mixed-Effects Models:

$$\boldsymbol{y}_i = \boldsymbol{f}_i(\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{x}_{ij}) + \boldsymbol{\epsilon}_i,$$

for mean function $\boldsymbol{f}_i(\cdot)$.

Generalized Linear Mixed Effects Models: $Y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \alpha \sim p(\cdot)$ where $p(\cdot)$ is a member of the exponential family and, if $\mu_{ij} = \text{E}[Y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \alpha]$, then we have a link function $g(\cdot)$, with

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i,$$

with $\boldsymbol{b}_i \sim_{iid} \text{N}(\boldsymbol{0}, \boldsymbol{D})$.

For linear models we need an appropriate marginal mean-variance model, for GLMM and NLMEMs it is more tricky...

In GLMMs and NLMEMs we require integration over the random effects — not always trivial, and can be an issue — for binary data models this is still a big problem. Asymptotics needed for inference.

Regular likelihood ratio tests are available for regression fixed effects — for variance components the null distribution is of non-standard form. Mixtures of χ^2's theoretical distributins are available for some null/alternatives, otherwise simulate data under the null to determine significance.

For variance components, asymptotic interval estimates may not be accurate.

The choice of random effects is guided in part by data availability — if we have small clusters then fewer random effects are supported.

Bayesian Approach

Takes the likelihood and adds priors to $\boldsymbol{\alpha}$.

MCMC/INLA needed for inference — no dependence on asymptotics.

Generalized Estimating Equations

Take as estimator $\widehat{\boldsymbol{\beta}}$ that which satisfies

$$\boldsymbol{G}(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}) = \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0},$$

where $\boldsymbol{D}_i = \frac{\partial \mu_i}{\partial \beta}$, $\boldsymbol{W}_i = \boldsymbol{W}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is the working covariance model, $\mu_i = \mu_i(\boldsymbol{\beta})$ and $\widehat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$

We obtain an appropriate standard error so long as we have independence between "units" — m is the number of independent units.

Parameter Interpretation

In the mixed effects models the regression coefficients have a conditional interpretation, i.e. conditional on the random effects.

In GEE the regression coefficients have a marginal interpretation, i.e. averaged across individuals within populations with specific values of covariates.

Parameter Interpretation for a Linear Model

Consider the random intercepts LMEM:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_j + b_i + \epsilon_{ij} \quad (49) \\ \mathrm{E}[Y_{ij}|b_i] &= \beta_0 + \beta_1 t_j + b_i \quad (50) \\ \mathrm{E}[Y_{ij}] &= \beta_0 + \beta_1 t_j \quad (51) \end{aligned}$$

with $b_i \sim \mathrm{N}(0, \sigma_0^2)$, independent of $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_\epsilon^2)$.

The expectation in (50) is with respect to "measurement error", and is within-unit averaging.

The expectation in (50) is with respect to the between-unit distribution of the random effects.

Note that in each expectation we have explicitly conditioned on t_j.

In this model, from (51), we can say that β_1 is the average change, across the population from which the units were sampled, in (average) response given a unit increase in t.

Because we have random intercepts only, we can also say that, from (50), β_1 is the change in average response given a unit increase in t, for each individual.

Now consider the the random intercepts and slopes LMEM:

$$Y_{ij} = (\beta_0^\star + b_{i0}) + (\beta_1^\star + b_{i1})t_j + \epsilon_{ij} \quad (52)$$

$$\mathrm{E}[Y_{ij}|b_i] = (\beta_0^\star + b_{i0}) + (\beta_1^\star + b_{i1})t_j \quad (53)$$

$$\mathrm{E}[Y_{ij}] = \beta_0^\star + \beta_1^\star t_j \quad (54)$$

with $\boldsymbol{b}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

In this model $\beta_1^\star$ is the average change, across the study population, in average response given a unit increase in t. Because we have random slopes, $\beta_1^\star + b_{i1}$ is the change in average response given a unit increase in t, for individual i.

Because of linearity we can say that $\beta_1^\star$ is the average of the individual changes in (average) response for a unit increase in t.

Consider the GEE model:

$$\mathrm{E}[Y_{ij}] = \gamma_0 + \gamma_1 t_j$$

where the expectation is over within- and between-unit distributions.

Note that we have not made any assumptions about these distributions.

In this model γ_1 is the average change, across the study population, in (average) response given a unit increase in t.

Parameter Interpretation for a Non-Linear Model

For simplicity we consider a binomial GLMM with a single measurement in each unit:

$$Y_i | p_i \sim \text{Binomial}(N_i, p_i) \quad (55)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + b_i, \quad b_i \sim \text{N}(0, \sigma^2) \quad (56)$$

We have

$$\text{E}[Y_i | p_i] = \text{E}[Y_i | b_i] = N_i p_i(b_i)$$

where this expectation is with respect to the binomial distribution.

We have

$$\text{E}_{y,b}[Y_i] = \text{E}_b \text{E}_{y|b}[Y_i | p_i]] = N_i \text{E}_b[p_i(b_i)] \approx N_i \frac{\exp(\beta_0 / \sqrt{c^2\sigma^2 + 1})}{1 + \exp(\beta_0 / \sqrt{c^2\sigma^2 + 1})}$$

where $c = 16\sqrt{3}/(15\pi)$. We also have

$$\text{E}\left[\frac{p_i}{1-p_i}\right] = \exp(\beta_0 + \sigma^2/2)$$

Interpretation of random effects distribution: Exchangeability for a Bayesian. Superpopulation for a frequentist, i.e. hypothetical.

Generalized Estimating Equations

The generic estimating equation for a $p \times 1$ parameter is:

$$\sum_{i=1}^{m} \boldsymbol{D}_i^{\text{T}} \boldsymbol{W}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$$

where

- $\boldsymbol{D}_i$ is the $p \times n_i$ matrix of derivaties $\frac{\partial \mu_i}{\partial \gamma_j}$,
- $\boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\boldsymbol{\gamma})$ and
- $\boldsymbol{W}_i = \boldsymbol{\lambda}_i(\boldsymbol{\alpha})^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{\lambda}_i(\boldsymbol{\alpha})^{1/2}$ is the $n_i \times n_i$ working covariance matrix for unit i, with
- $\boldsymbol{\lambda}_i(\boldsymbol{\alpha}) = \text{diag}[\text{var}(Y_{i1}), ..., \text{var}(Y_{in_i})]^{\text{T}}$, where the variances are the "nominal" forms suggested by the family, e.g. binomial.
- $\boldsymbol{R}_i$ is a working correlation structure, for example, independence or exchangeable.

Simple Example

In the "binomial" example:

$$\mathrm{E}[Y_i] = \mathrm{E}[Y_i|N_i] = N_i p = N_i \frac{\exp(\gamma_0)}{1+\exp(\gamma_0)}$$

where the expectation is with respect to within-unit and between-unit sampling.

Note that we could also condition on covariates: $\mathrm{E}[Y_i|\boldsymbol{x}_i]$.

Parameter interpretation: $\exp(\gamma_0)$ is the odds of the event across the study population.

It is not directly comparable with $\exp(\beta_0 + \sigma^2/2)$ since the latter is the average of the odds.

We give an example to demonstrate.

Numerical Example: The illiteracy data for native-born whites

As an example we consider the native-born white data across states:

$$\sum_{i=1}^{m} D_i W_i^{-1}(Y_i - \mu_i)$$

We have

$$\begin{aligned}
\mu_i &= N_i p = N_i \frac{\exp(\gamma_0)}{1+\exp(\gamma_0)} \\
D_i &= N_i p(1-p) = N_i \frac{\exp(\gamma_0)}{[1+\exp(\gamma_0)]^2} \\
W_i &= \alpha N_i \frac{\exp(\gamma_0)}{[1+\exp(\gamma_0)]^2}
\end{aligned}$$

so that α is the scale parameter that multiplies the nominal binomial variance.

The estimating equation is:

$$G(\gamma_0) = \frac{1}{\alpha} \sum_{i=1}^{m} \left(Y_i - N_i \frac{\exp(\gamma_0)}{1+\exp(\gamma_0)} \right) = 0$$

so that the solution is

$$\widehat{\gamma}_0 = \log \left[\frac{\sum_i Y_i}{\sum_i (N_i - Y_i)} \right].$$

This confirms that we are estimating the odds of illiteracy in the US.

The variance is given by

$$\text{var}(\widehat{\gamma}_0) = A^{-1} B A^{-1}$$

with

$$\begin{aligned} A &= \text{E}\left[\frac{\partial G}{\partial \gamma_0} \right] = -\frac{1}{\alpha} \sum_{i=1}^{m} N_i \frac{\exp(\gamma_0)}{[1+\exp(\gamma_0)]^2} = -\frac{N_+}{\alpha} p(1-p) \\ B &= \text{cov}(G) = \frac{1}{\alpha^2} \sum_{i=1}^{m} \text{var}(Y_i) = \frac{1}{\alpha^2} \sum_{i=1}^{m} (Y_i - N_i p)^2 = \frac{1}{\alpha^2} \sum_{i=1}^{m} N_i^2 \left(\frac{Y_i}{N_i} - p \right)^2 \end{aligned}$$

Estimation of α

Various options are available for estimation of α, including the quasi-likelihood estimator, method of moments type estimator. Since

$$\alpha = \frac{1}{m} \text{E}\left[(\boldsymbol{Y} - \boldsymbol{\mu})^{\text{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) \right] = \frac{1}{m} \text{E}\left[\sum_{i=1}^{m} \frac{(Y_i - \mu_i)^2}{V_i} \right]$$

(where $\boldsymbol{V}$ is the nominal variance and is diagonal), an unbiased estimator would be

$$\widehat{\alpha} = \frac{1}{m} \sum_{i=1}^{m} \frac{(Y_i - \mu_i)^2}{V_i}$$

A "degrees of freedom" (but in general biased) estimator is:

$$\begin{aligned} \widehat{\alpha} &= \frac{1}{m-1} \sum_{i=1}^{m} \frac{(Y_i - N_i \widehat{p})^2}{N_i \widehat{p}(1-\widehat{p})} \\ &= \frac{1}{m-1} \sum_{i=1}^{m} \frac{N_i \left(\frac{Y_i}{N_i} - \widehat{p} \right)^2}{\widehat{p}(1-\widehat{p})} \end{aligned}$$

Forms of Variance

Under the model, $A = -B$, and

$$\text{var}(\widehat{\gamma}_0) = -A^{-1} = \frac{\alpha}{N_+ p(1-p)}$$

which is estimated by:

$$\widehat{\text{var}}(\widehat{\gamma}_0) = \frac{\widehat{\alpha}}{N_+ \widehat{p}(1-\widehat{p})}$$

The robust sandwich estimator is given by:

$$\text{var}(\widehat{\gamma}_0) = \frac{B}{A^2} = \frac{\sum_{i=1}^m N_i^2 \left(\frac{Y_i}{N_i} - p \right)^2}{N_+^2 p^2 (1-p)^2}$$

which is estimated by

$$\widehat{\text{var}}(\widehat{\gamma}_0) = \frac{\sum_{i=1}^m N_i^2 \left(\frac{Y_i}{N_i} - \widehat{p} \right)^2}{N_+^2 \widehat{p}^2 (1-\widehat{p})^2}$$

Numerical Calculations

```
> illitw <- illit[race==1]
> totalw <- total[race==1]
> statew <- state[race==1]
> geemod0w <- gee(cbind(illitw,totalw-illitw)~1,id=statew,family=binomial,
  corstr="exchangeable")
> summary(geemod0w)
Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Exchangeable
gee(formula = cbind(illitw, totalw - illitw) ~ 1, id = statew,
    family = binomial, corstr = "exchangeable")
Coefficients:
             Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept) -4.202025  0.1738016 -24.17714   0.1847846 -22.74012
Estimated Scale Parameter:  32831.03
# Noe let's construct by hand:
> Nsum <- sum(totalw)
> phat <- sum(illitw)/Nsum
> gamma0hat <- log(phat/(1-phat))
> gamma0hat
[1] -4.202025
> resid <- illitw/totalw-phat
```

```
> alphahat <- (1/48)*sum(totalw*resid^2)/(phat*(1-phat))
> alphahat
[1] 32831.03
> Aterm <- Nsum*phat*(1-phat)/alphahat
> Bterm <- sum(totalw^2*resid^2)/alphahat^2
> robustvar <- Bterm/Aterm^2
> sqrt(robustvar)
[1] 0.1847846
> modelvar <- 1/Aterm
> sqrt(modelvar)
[1] 0.1738016
```


Model Checking

For LMEMs and NLMEMs model checking can be carried out reasonably well so long as their are some individuals with larger n_i — individual fits may then be carried out, with $\widehat{\boldsymbol{\beta}}_i$'s being examined.

Fo GLMMs with binary data it is very difficult to diagnose problems with the model — multiple observations within clusters are more conducive to diagnosis of problems.

An important assumption is of a constant random effects distribution across covariate groups.

Final Comment

Dependent data are complex and difficult to analyze, but don't be afraid to apply different techniques.

Each of likelihood, Bayes and GEE have strengths and weaknesses, but can often be used in a complementary fashion.

Care is required in interpretation of parameters, however.

The End!