Improving Small-Sample Inference in Group Randomized Trials with Binary Outcomes

Philip M. Westgate

Doctoral Candidate
Department of Biostatistics
University of Michigan

Job Talk Seminar
January 25, 2011
Department of Biostatistics
Vanderbilt University School of Medicine
Outline

- Introduction
- Statistical Background
- Test Size
- Simulations and Application
- Summary and References
- Dissertation Work Based on QIF
Group Randomized Trials (GRTs)

Common Properties of Group Randomized Trials:

- Randomize groups, or clusters, of subjects
- Number of clusters, $n$, typically small
- Outcome of interest is at subject-level

Scenario:

- Subject-level outcomes are binary indicators of a desired outcome ("success")
- Clusters are randomized to treatment or control
- No other covariates are considered
Example

Breast Screening Study [Atri et al., 1997]
Intervention: 2 hour training session for receptionists at medical care practices
Goal: Increase breast screening rates in women who failed to attend their initial appointment
Outcome of Interest: Number of women in practice $i$, $y_i$, who receive breast screening, out of the $n_i$ women failing to attend their initial appointment.
Practices were randomized to intervention (12) or control (14).
Purpose

Demonstrate methods to obtain test sizes at their nominal values when testing for a marginal treatment/intervention effect using a Wald test statistic.
Notation

- $n_i$ denotes the size of the $i^{th}$ cluster
- $Y_i$ - number of successes in cluster $i$
- $\pi_i = Prob(X_{ij} = 1) = E[Prob(X_{ij} = 1 | p_i)]$
- $p_i$ - actual probability of success for any given subject in cluster $i$
- $E(Y_i) = n_i \pi_i$, $Var(Y_i) = n_i \pi_i (1 - \pi_i) [1 + (n_i - 1) \rho]$ (Over-Dispersion)
- $\rho$ - Intra-Cluster Correlation (ICC) (currently assumed constant and known)
- $Z_i$ indicates treatment assignment for cluster $i$ (1-Treatment, 0-Control)
- $\text{logit}(\pi_i) = \beta_0 + \beta_1 Z_i$
- $H_0 : \beta_1 = 0$
- $\pi_i = \pi_T = \pi_T(\beta)$ if randomized to Treatment, else $\pi_i = \pi_C = \pi_C(\beta)$
Quasi-Likelihood

- Only requires the mean and variance structures for $Y_i$ to be correctly modeled
- Statistical results rely upon asymptotic theory (which may not hold for GRT scenarios)
- Potential Problems with Empirical/Sandwich [Liang and Zeger, 1986; Mancl and DeRouen, 2001] and bootstrap [Mancl and DeRouen, 2001] SEs, obtaining critical values, and bias in MQL estimates of $\hat{\beta}_{MQL} = [\hat{\beta}_{0MQL}, \hat{\beta}_{1MQL}]^T$ [Cordeiro and Demetrio, 2008]
- Model-Based SE may be best
Standard Error (SE) and Wald Statistic

Estimating $\hat{\beta}_{MQL}$: $\sum_{i=1}^{n} [1, Z_i]' \frac{Y_i - n_i \pi_i}{1 + (n_i - 1) \rho} = 0$

Model-Based SE for $\hat{\beta}_{1MQL}$:

$SE(\hat{\beta}_{1MQL}) = SE_{\hat{\beta}_{1MQL}}[\pi_C(\beta), \pi_T(\beta)] = \sqrt{\left[ \sum_{i=1}^{u} \frac{n_i \pi_C(\beta)[1 - \pi_C(\beta)]}{1 + (n_i - 1) \rho} \right]^{-1} + \left[ \sum_{i=u+1}^{n} \frac{n_i \pi_T(\beta)[1 - \pi_T(\beta)]}{1 + (n_i - 1) \rho} \right]^{-1}}$

$\hat{SE}_{MQL}(\hat{\beta}_{1MQL}) = SE_{\hat{\beta}_{1MQL}}[\pi_C(\hat{\beta}_{MQL}), \pi_T(\hat{\beta}_{MQL})]$

Wald Statistic: $W_{Reg} = \hat{\beta}_{1MQL}/\hat{SE}_{MQL}(\hat{\beta}_{1MQL})$
Empirical distribution of $W_{Reg}$, $n = 10/Arm$, $\pi_C = \pi_T = \rho = 0.05$
Causes of Incorrect Test Size

- How much smaller test size is than its nominal value depends on $\text{Var}(W_{\text{Reg}})$
- $\text{Var}(W_{\text{Reg}})$ depends on the variances and covariance of $\hat{\beta}_{1\text{MQL}}$ and $\text{SE}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})$
- The variance of $\text{SE}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})$ depends on $\text{Var}(\hat{\beta}_{\text{MQL}})$, which in turn depends on cluster sizes, $n$, $\rho$, $\pi_C$ and $\pi_T$.
- Decreases in $n$ and cluster sizes, increases in $\rho$ and the distances marginal probabilities are from 0.5: Associated with larger $\text{Var}(\hat{\beta}_{\text{MQL}})$, and therefore larger $\text{Var}[\text{SE}_{\text{MQL}}(\hat{\beta}_{1\text{MQL}})]$
As $\text{Var}(\hat{\beta}_{1MQL})$ increases, there are more extreme values for $\hat{\beta}_{1MQL}$, and these large values are associated with larger values for $\hat{\text{SE}}_{MQL}(\hat{\beta}_{1MQL})$

This causes $W_{Reg}$ to be smaller than desired, therefore:

- Reducing $\text{Var}(W_{Reg})$
- Causing the tails in the distribution of $W_{Reg}$ to become lighter
- Test size decreases
Test size decreases as $\text{Var}[\hat{SE}_{MQL}(\hat{\beta}_{1MQL})]$ increases, meaning test size decreases away from its nominal value as

- $n$ decreases,
- cluster sizes decrease,
- $\rho$ increases,
- $\pi_C, \pi_T$ move further from 0.5
Correcting Test Size

- Cordeiro and Demetrio (2008) gave formulas for bias
- Biases are given by:

\[
Bias(\hat{\beta}_{0\text{MQL}}) = \frac{2\pi C(\beta) - 1}{2\pi C(\beta)[1 - \pi C(\beta)] \sum_{i=1}^{u} q_i}
\]

\[
Bias(\hat{\beta}_{1\text{MQL}}) = \frac{2\pi T(\beta) - 1}{2\pi T(\beta)[1 - \pi T(\beta)] \sum_{i=u+1}^{n} q_i} - Bias(\hat{\beta}_{0\text{MQL}}),
\]

\[
q_i = n_i/[1 + (n_i - 1)\rho]
\]
Note: Amount of Bias Increases as

- $n$ decreases,
- cluster sizes decrease,
- $\rho$ increases,
- $\pi_C, \pi_T$ move further from 0.5

Bias-Corrected Estimates (BCEs) given by:

$$\hat{\beta}_{BC} = [\hat{\beta}_{0MQL} - \widehat{Bias}(\hat{\beta}_{0MQL}), \hat{\beta}_{1MQL} - \widehat{Bias}(\hat{\beta}_{1MQL})]' = [\hat{\beta}_{0BC}, \hat{\beta}_{1BC}]'$$
Var(\(W_{\text{Reg}}\)) can be less than one, resulting in test sizes smaller than \(\alpha\), the nominal value, when using \(N(0,1)\) critical values.

Can fix this by modifying \(\widehat{SE}(\hat{\beta}_1^\text{MQL})\).

This modification needs to depend on \(n\), cluster sizes, \(\rho\), \(\pi_C\) and \(\pi_T\), since test size depends on these quantities.

Resulting Wald statistic should always have a variance of approximately 1, giving nominal test size.
Solution

- $\text{Bias}(\hat{\beta}_{0\text{MQL}})$, and $\text{Bias}(\hat{\beta}_{1\text{MQL}})$ are functions of $n$, cluster size, $\rho$, $\pi_C$ and $\pi_T$

- Define $\tilde{\beta}^N_k = \left(\frac{\hat{\beta}_{kBC}}{\hat{\beta}_{k\text{MQL}}}\right)^N \hat{\beta}_{k\text{MQL}}$, $k = 0, 1$, for any non-negative real number $N$

- Modified SE estimate is
  $$\tilde{SE}^N(\hat{\beta}_{1\text{MQL}}) = \text{SE}_{\hat{\beta}_{1\text{MQL}}} \left[ \pi_C(\tilde{\beta}^N), \pi_T(\tilde{\beta}^N) \right]$$
  in which
  $$\tilde{\beta}^N = [\tilde{\beta}_0^N, \tilde{\beta}_1^N]$$

- Use the pseudo-Wald statistic $\tilde{W}_N = \hat{\beta}_{1\text{MQL}}/\tilde{SE}^N(\hat{\beta}_{1\text{MQL}})$

- Asymptotically, $\tilde{W}_N \xrightarrow{d} N(0, 1)$ under the null hypothesis
How do we choose N, assuming modifying the SE estimate in this manner is appropriate?

Exploratory Simulations:
- 10,000 simulations/setting
- Cluster sizes varied uniformly from 25 to 150 subjects
- Data generated from a beta-binomial distribution
- Compare test size for $W_{\text{Reg}}, \tilde{W}_N; N = 1, 1.5, 2$
- $N=1.5$ performed best
Table: Simulated test sizes using the given Wald statistic and $N(0, 1)$ critical values. The ICC is known.

<table>
<thead>
<tr>
<th>$n/2$</th>
<th>$\pi$</th>
<th>ICC</th>
<th>$W_{\text{Reg}}$</th>
<th>$W_1$</th>
<th>$W_{1.5}$</th>
<th>$W_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.05</td>
<td>0.05</td>
<td>0.0360</td>
<td>0.0459</td>
<td>0.0509</td>
<td>0.0557</td>
</tr>
<tr>
<td>20</td>
<td>0.05</td>
<td>0.05</td>
<td>0.0407</td>
<td>0.0453</td>
<td>0.0464</td>
<td>0.0485</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>0.05</td>
<td>0.0436</td>
<td>0.0461</td>
<td>0.0478</td>
<td>0.0491</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>0.05</td>
<td><strong>0.0470</strong></td>
<td>0.0490</td>
<td>0.0499</td>
<td>0.0509</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>0.10</td>
<td>0.0390</td>
<td>0.0462</td>
<td>0.0502</td>
<td>0.0534</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>0.10</td>
<td>0.0429</td>
<td>0.0458</td>
<td>0.0471</td>
<td>0.0490</td>
</tr>
<tr>
<td>10</td>
<td>0.20</td>
<td>0.05</td>
<td><strong>0.0481</strong></td>
<td>0.0492</td>
<td>0.0495</td>
<td>0.0500</td>
</tr>
<tr>
<td>20</td>
<td>0.20</td>
<td>0.10</td>
<td><strong>0.0496</strong></td>
<td>0.0502</td>
<td>0.0503</td>
<td>0.0505</td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>0.05</td>
<td><strong>0.0518</strong></td>
<td>0.0523</td>
<td>0.0525</td>
<td>0.0528</td>
</tr>
</tbody>
</table>
Empirical Distribution of $\tilde{W}_{1.5}$, $n = 10/\text{Arm}$, $\pi_C = \pi_T = \rho = 0.05$
The following two issues still need to be handled in practice:

- Find a consistent estimate, $\hat{\rho}$, for the ICC.
- Find appropriate critical values to handle the effect estimating $\rho$ has on the distribution of the Wald statistic.

We use:

- ANOVA to estimate the ICC (Donner and Donald, 1988)
- The $t$-distribution with $f(n) = n$ degrees of freedom ($df$), $t_n$, to obtain critical values

Note: Estimating the ICC increases test size
Simulations

Define:

- $\hat{SE}_E(\hat{\beta}_{1MQL})$ - empirical SE estimate
- $\hat{SE}_{EBC}(\hat{\beta}_{1MQL})$ - bias-corrected empirical SE estimate
- $W_E = \hat{\beta}_{1MQL}/\hat{SE}_E(\hat{\beta}_{1MQL})$
- $W_{EBC} = \hat{\beta}_{1MQL}/\hat{SE}_{EBC}(\hat{\beta}_{1MQL})$

Compare performances of:

- $W_{Reg}$, $\tilde{W}_{1.5}$, $W_E$, and $W_{EBC}$ as test statistics
- Critical values from $N(0, 1)$ and $t_n$
**Table: Simulated Test Sizes**

<table>
<thead>
<tr>
<th>$n/2$</th>
<th>$\pi$</th>
<th>ICC</th>
<th>$W_{\text{Reg}}$</th>
<th>$t_n$</th>
<th>$\tilde{W}_{1.5}$</th>
<th>$t_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.05</td>
<td>0.05</td>
<td>0.0506</td>
<td>0.0376</td>
<td>0.0662</td>
<td>0.0478</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>0.10</td>
<td>0.0509</td>
<td>0.0443</td>
<td>0.0557</td>
<td>0.0486</td>
</tr>
<tr>
<td>10</td>
<td>0.20</td>
<td>0.10</td>
<td>0.0614</td>
<td>0.0462</td>
<td>0.0645</td>
<td>0.0491</td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>0.05</td>
<td>0.0655</td>
<td>0.0514</td>
<td>0.0660</td>
<td>0.0520</td>
</tr>
<tr>
<td>20</td>
<td>0.50</td>
<td>0.10</td>
<td>0.0567</td>
<td>0.0499</td>
<td>0.0567</td>
<td>0.0499</td>
</tr>
</tbody>
</table>

Philip M. Westgate

Improving Small-Sample Inference in Group Randomized Trials
### Table: Simulated Test Sizes

<table>
<thead>
<tr>
<th>( n/2 )</th>
<th>( \pi )</th>
<th>ICC</th>
<th>( W_E )</th>
<th>( t_n )</th>
<th>( W_{EBC} )</th>
<th>( t_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.05</td>
<td>0.05</td>
<td>0.1029</td>
<td>0.0836</td>
<td>0.0715</td>
<td>0.0548</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>0.10</td>
<td>0.0713</td>
<td>0.0635</td>
<td>0.0586</td>
<td><strong>0.0509</strong></td>
</tr>
<tr>
<td>10</td>
<td>0.20</td>
<td>0.10</td>
<td>0.0836</td>
<td>0.0693</td>
<td>0.0589</td>
<td>0.0452</td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>0.05</td>
<td>0.0806</td>
<td>0.0662</td>
<td>0.0550</td>
<td>0.0418</td>
</tr>
<tr>
<td>20</td>
<td>0.50</td>
<td>0.10</td>
<td>0.0635</td>
<td>0.0556</td>
<td><strong>0.0501</strong></td>
<td>0.0437</td>
</tr>
</tbody>
</table>
Application

Breast Screening Study

Table: Number of women having been screened, $y_i$, out of $n_i$ possible women in practice $i$, $i = 1, 2, \ldots 26$.

<table>
<thead>
<tr>
<th>Practice, $i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>\ldots</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls $y_i$</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>\ldots</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>$n_i$</td>
<td>145</td>
<td>43</td>
<td>59</td>
<td>22</td>
<td>\ldots</td>
<td>26</td>
<td>35</td>
<td>25</td>
<td>179</td>
<td>38</td>
</tr>
<tr>
<td>Practice, $i$</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>\ldots</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>Interventions $y_i$</td>
<td>5</td>
<td>28</td>
<td>10</td>
<td>2</td>
<td>\ldots</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>$n_i$</td>
<td>44</td>
<td>156</td>
<td>19</td>
<td>80</td>
<td>\ldots</td>
<td>39</td>
<td>103</td>
<td>56</td>
<td>139</td>
<td>201</td>
</tr>
</tbody>
</table>

Data obtained from Turner et al. (2001)
### Table: Analysis Results

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>$N(0, 1)$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$p$-value</td>
<td>$95%$ CI</td>
<td>$p$-value</td>
</tr>
<tr>
<td>$\hat{\beta}_{1MQL}$</td>
<td>1.138</td>
<td>0.023</td>
<td>(0.160, 2.116)</td>
<td>0.031</td>
</tr>
<tr>
<td>$\widehat{SE}<em>{MQL}(\hat{\beta}</em>{1MQL})$</td>
<td>0.499</td>
<td>0.017</td>
<td>(0.200, 2.076)</td>
<td>0.025</td>
</tr>
<tr>
<td>$\widetilde{SE}<em>{E}(\hat{\beta}</em>{1MQL})$</td>
<td>0.479</td>
<td>0.006</td>
<td>(0.322, 1.955)</td>
<td>0.011</td>
</tr>
<tr>
<td>$\widehat{SE}<em>{EBC}(\hat{\beta}</em>{1MQL})$</td>
<td>0.417</td>
<td>0.011</td>
<td>(0.261, 2.015)</td>
<td>0.017</td>
</tr>
</tbody>
</table>
Alternative Model-Based Wald Statistic

Approximation: $\hat{\beta}_{BC} \propto \hat{\beta}_{ML}$ [King and Zeng, 2001]

- $\hat{\beta}_{1BC}$ in numerator
- $(\hat{\beta}_{1BC}/\hat{\beta}_{1MQL}) \tilde{SE}_{1.5}(\hat{\beta}_{1MQL})$ in denominator
- Test size remains unchanged
- Point estimate is approximately unbiased
- Narrower confidence intervals
Summary

- Estimating the QL Model-based SEs can cause test size to fall below its nominal value.
- The difference between the true test size and its nominal value depends on the number of clusters and their sizes, the ICC, and the true marginal probabilities.
- We use the fact that bias of the MQL estimates also depends on these factors to modify the SE used in the Wald statistic.
- The use of $\tilde{W}_{1.5}$ with $t_n$ critical values results in test sizes at their nominal value when the ICC is correctly modeled.
References


Quadratic Inference Functions (QIF)

- Proposed by Qu, Lindsay, and Li (2000, *Biometrika* 87, 823-836)
- Theoretically: Improves upon or maintains the efficiency of GEE
- Realistically: GEE may produce estimates with greater precision than QIF in small-sample settings (e.g. GRTs)
- Second Manuscript: Study of QIF’s theoretical and empirical nature that leads to potentially poor performance. The major influences are cluster size imbalance and covariates via the empirical weight matrix employed with QIF.
- Third Manuscript: Improving the small-sample estimation performance of QIF by modifying the empirical weight matrix.
QIF Basics

GEE [Liang, K Y. and Zeger S L., (1986); Biometrika 73, 13-22] given as

\[ \sum_{i=1}^{n} D_i^T A_i^{-1/2} R_i^{-1} A_i^{-1/2} (Y_i - \mu_i) = 0, \]

QIF typically uses \( R_i^{-1} = \sum_{r=1}^{2} \gamma_{ri} M_{ri} \)

Splits GEE into the sum of two unbiased estimating equations:

\[ \sum_{i=1}^{n} D_i^T A_i^{-1/2} R_i^{-1} A_i^{-1/2} (Y_i - \mu_i) = \sum_{i=1}^{n} \gamma_{1i} g_{1i} + \sum_{i=1}^{n} \gamma_{2i} g_{2i} \]
Places one on top of the other, but ignores estimating $\gamma_{ri}$, and thus the correlation parameter:

$$\bar{g}_n = \frac{1}{n} g_n = \frac{1}{n} \sum_{i=1}^{n} g_i = \left[ \frac{1}{n} \sum_{i=1}^{n} g_{1i} \right]$$

QIF defined as

$$Q_n = n\bar{g}_n^T C_n^{-1} \bar{g}_n, \quad C_n = \frac{1}{n} \sum_{i=1}^{n} g_i g_i^T$$

Based on Generalized Method of Moments [Hansen, L. P. (1982); *Econometrica* **50**, 1029-1054]
Corresponding estimating equations (EE) given by

\[ \sum_{i=1}^{n} \nabla \bar{g}_n^T C_n^{-1} g_i \]

Focus of second part of my dissertation:

Two Influences on Estimation Performance
- EE Class
- Empirical Nature of EE via $C_n$
EE Class

Importance:

• Gives QIF an efficiency advantage over GEE when both methods’ EE are in the same class

• Imbalance in cluster sizes causes GEE and EE from QIF to be in different classes (my theoretical focus)

• QIF may not have an efficiency advantage anymore

• Solution: create an alternate QIF version such that its EE are in the same class as GEE
Empirical Nature of EE

- $C_n$ is an empirical estimate used inside QIF’s EE
- Covariates and Imbalance in cluster sizes affect the amount of empirical information in $C_n$ used to weight outcomes from each cluster inside the EE (my applied focus)
- These EE can therefore be more variable than GEE
- Potentially results in a poor estimation performance by QIF
- Poor performance is most notable in small-sample settings (e.g. GRTs)
Improving QIF Estimation Performance

Focus of the last part of my dissertation

Solutions I am currently studying

• Modify $C_n$ to combat effect of imbalanced cluster sizes by
• Averaging out the effect of cluster size on $C_n$, or
• Using a model-based estimate for $C_n$, or
• A combination of $C_n$ and the previous two estimates