Hypothesis Testing for Personalizing Treatment

Huitian Lei and Susan Murphy

Abstract

In personalized treatment the recommended treatment is based on patient characteristics. Given pre-specified subgroups, we define the subgroup indicator as useful for personalized decision making if for particular subgroups there is sufficient evidence to recommend one treatment, while for other subgroups, either there is sufficient evidence to recommend a different treatment, or there is insufficient evidence to recommend a particular treatment. We propose a two-stage hypothesis testing procedure to evaluate if a subgroup indicator is useful in personalized decision making. In the first stage of the procedure, we utilize the test statistic for testing treatment-subgroup interaction. If the first stage test statistic exceeds the critical value, we proceed to the second stage of the procedure and utilize test statistics for testing subgroup treatment effects. We control a generalized Type I error rate.

Simple Setting

- Two pre-defined subgroups: subgroup 1 and subgroup 2
- Two active treatments: treatment A and treatment B
- The mean treatment response in subgroup i under treatment TR: μi,T, where i \in \{1, 2\}, TR \in \{A, B\}
- Subgroup treatment effects: θ = (θ1, θ2), where θ1 = μL,A - μL,B and θ2 = μH,A - μH,B.

Motivation and Objectives

Current Subgroup Analysis: Testing qualitative interaction
- Gail and Simon (1985): likelihood ratio test of qualitative interaction: θ1θ2 < 0
- The hypothesis: H0: θ1θ2 ≥ 0
- Qualitative interaction is not the only treatment-subgroup interaction informative in personalizing treatment (e.g., Simon 2002).

Current Subgroup Analysis: Testing subgroup hypotheses
- H1: θ1 = 0 and θ2 = 0
- H0: θ1 = θ2, H1: θ1 = 0 and H2: θ1 = 0
- Usually control the familywise size error rate. Mismatch between the controlled errors and the scientific goal.

Objective of the decision procedure
- Match the decision procedure and controlled errors with the scientific goal
- Distinguish different types of interaction informative in personalizing treatment
- Same power to detect qualitative interactions as Gail and Simon test

The Procedure

The hypothesis that the subgroup indicator is not useful for personalizing treatment:

H : θ1 = θ2 = 0, or θ1θ2 > 0

The complement of H consists of:
- θ1 ≠ 0, θ2 = 0
- θ1 = 0, θ2 ≠ 0
- θ1θ2 < 0

The Decision Space

<table>
<thead>
<tr>
<th>Decision</th>
<th>Decision 1</th>
<th>Decision 2</th>
<th>Decision 3</th>
<th>Decision 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>θ1 = θ2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>θ1 ≠ 0, θ2 = 0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>θ1 ≠ 0, θ2 ≠ 0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>θ1 = 0, θ2 ≠ 0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>θ1 ≠ 0, θ2 ≠ 0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The Controlled Errors

- Data from two-arm randomized clinical trials
- Assume equal unknown variance

\[ p_1, p_2 \text{ are sample subgroup proportions. } n_1 \text{ and } n_2 \text{ are subgroup sample sizes. } \bar{X}_{1A}, \bar{X}_{2A}, \bar{X}_{1B}, \bar{X}_{2B} \text{ are cell means. } \sigma^2 \text{ is the estimated pooled variance.} \]

\[ c_0 \text{ and } c_2 \text{ are the critical values in stage I and stage II.} \]

- In stage I, utilize test statistic \( T_1 = \frac{\bar{X}_{1A} - \bar{X}_{1B}}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \) and compare it with critical value \( \pm c_0. \)

- If \( T_1 > c_0 \) or \( T_1 < -c_0 \), proceed to stage II. Otherwise if \( |T_1| \leq c_0 \), stop the testing procedure and make decision 1.

- In stage II, utilize test statistics \( \bar{T}_1 = \frac{\bar{X}_{1A} - \bar{X}_{1B}}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \) and \( \bar{T}_2 = \frac{\bar{X}_{2A} - \bar{X}_{2B}}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \) and compare them with critical value \( \pm c_1. \)

The Critical Values at α = 0.05

Evaluating the Power

- Same power to detect qualitative interactions as Gail and Simon test when 0.10 ≤ p1 ≤ 0.50.

\[ \{1, T_2 : \sqrt{\frac{\bar{T}_1}{\hat{\theta}_1}} - \sqrt{\frac{\bar{T}_2}{\hat{\theta}_2}} = \bar{T}_1 > c_0, T_2 > c_0, T_2 < -c_1\} \]

\[ \{1, T_2 : \sqrt{\frac{\bar{T}_1}{\hat{\theta}_1}} - \sqrt{\frac{\bar{T}_2}{\hat{\theta}_2}} = \bar{T}_1 < -c_0, T_2 > c_0, T_2 < c_1\} \]

\[ \{1, T_2 : \sqrt{\frac{\bar{T}_1}{\hat{\theta}_1}} - \sqrt{\frac{\bar{T}_2}{\hat{\theta}_2}} = \bar{T}_1 > c_0, T_2 < -c_1\} \]

\[ \{1, T_2 : \sqrt{\frac{\bar{T}_1}{\hat{\theta}_1}} - \sqrt{\frac{\bar{T}_2}{\hat{\theta}_2}} = \bar{T}_1 < -c_0, T_2 < c_1\} \]

- Better power to detect \( \theta_1 \neq 0, \theta_2 = 0 \) or \( \theta_1 = 0, \theta_2 \neq 0 \), compared to a likelihood ratio test of \( H : \theta_1 = \theta_2 = 0 \), or \( \theta_1\theta_2 > 0 \).

- Less powerful to detect qualitative interactions when one subpopulation is rare (e.g., \( p_1 = 0.05 \)).

Future work and Acknowledgement

- Generalize the proposed procedure to multi-subgroup settings.
- The treatment effects are estimated by cell means. A regression model which includes useful baseline variables may significantly reduce the variance of the estimated treatment effects.

We acknowledge NIDRA grant 550-DA-010075 for support.