Utility of synthetic microdata generated using tree-based methods

Beata Nowok

Administrative Data Research Centre – Scotland
The Scottish Health Survey
2008 & 2010
Physical Activity and Mental Health in Scotland
Tree-based methods

Generating synthetic data

\[ Y_j \sim (Y_0, Y_1, \ldots, Y_{j-1}) \]

Utility evaluation
Tree-based methods

- Classification and regression trees (CART)
- Bagging
- Random forest
Build a tree (binary splits)
\( Y_j \sim (Y_0, \ldots, Y_{j-1}) \)

Generate \( Y_j \) by:
- Running \( Y_0, \ldots, Y_{j-1} \) down the tree
- Sampling \( Y_j \) from the leaves

Life satisfaction

- health status = fair / poor
- age < 50
- sex = male
Build a tree (binary splits)
\( Y_j \sim (Y_0, \ldots, Y_{j-1}) \)

Generate \( Y_j \) by:
- Running \( Y_0, \ldots, Y_{j-1} \) down the tree
- Sampling \( Y_j \) from the leaves
CART splitting rules

- Association with the response variable
- Gini index and deviance
Random forest and bagging

Collection of trees:
- Bootstrap samples from the original data set
- Sample of predictors as split candidates
Synthetic data utility

- **General** measures:
  - consider overall microdata similarity
  - predict effectiveness without fitting all models

- **Specific** measures:
  - consider analysis-specific similarity
  - high specific utility is desirable for researchers in preliminary analysis
General utility measures

Group \sim (Y_0, Y_1, \ldots, Y_j)

Probability of being in the synthetic data
Analysis-specific utility measures

- Comparison of models fit with both the observed data and the synthetic data
  - Overlap of coefficient confidence intervals
  - Standardized difference of coefficient estimates
Synthetic data

- SHeS 2008 & 2010, 46 variables
- Replacing all values for all variables
- 10 synthetic data sets for each method:
  - \( \text{CART}_1 \) (correlation)
  - \( \text{CART}_2 \) (Gini index and deviance)
  - [RF] Random forest (100 trees)
  - [BAG] Bagging (100 trees)
  - [PARA] Parametric
  - [SAMP] Sampling
Estimated regression models

- **Linear: mental well-being**
  
  \[ Y_{wemwbs} \sim \text{Sex} + \text{age} + \text{age}^2 + \text{sptinPAg} + \text{hrstot10} + \text{AllNature} + \text{Gym} + \text{Other} + \text{Longill08} + \text{ simd15 09} + \text{maritalg} \]

- **Logistic: life satisfaction**
  
  \[ Y_{lsg} \sim \text{Sex} + \text{ag16g10} + \text{ParentInf} + \text{hrsptg10} + \text{eqvinc} + \text{maritalg} + \text{XOwnRnt08} + \text{econac08} + \text{drkcat3} + \text{cigst3} + \text{porftvg3} + \text{AllNature} \]

- **Poisson: time of sport activities**
  
  \[ hrsspt10 \sim \text{Sex} + \text{ag16g10} + \text{hrstot10} + \text{URINDSC} + \text{XCare} + \text{Longill08} + \text{eqvinc} + \text{maritalg} + \text{Active} + \text{ParentInf} + \text{econac08} \]
## General and analysis-specific utility

<table>
<thead>
<tr>
<th></th>
<th>$CART_1$</th>
<th>$CART_2$</th>
<th>$BAG$</th>
<th>$RF$</th>
<th>$PARA$</th>
<th>$SAMP$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General utility</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Propensity score measure</td>
<td>0.14</td>
<td>0.11</td>
<td>0.12</td>
<td>0.13</td>
<td>0.15</td>
<td>0.21</td>
</tr>
<tr>
<td><strong>Analysis-specific utility</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Model 1: linear</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean $\hat{\beta}$ stand. difference</td>
<td>8.40</td>
<td>2.76</td>
<td>4.93</td>
<td>6.51</td>
<td><strong>1.77</strong></td>
<td>11.63</td>
</tr>
<tr>
<td>Mean 95% CI overlap</td>
<td>0.42</td>
<td>0.77</td>
<td>0.60</td>
<td>0.52</td>
<td><strong>0.86</strong></td>
<td>0.31</td>
</tr>
<tr>
<td><strong>Model 2: logistic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean $\hat{\beta}$ stand. difference</td>
<td>6.95</td>
<td>3.72</td>
<td>3.73</td>
<td>5.30</td>
<td><strong>1.72</strong></td>
<td>10.98</td>
</tr>
<tr>
<td>Mean 95% CI overlap</td>
<td>0.50</td>
<td>0.71</td>
<td>0.70</td>
<td>0.60</td>
<td><strong>0.86</strong></td>
<td>0.33</td>
</tr>
<tr>
<td><strong>Model 3: Poisson</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean $\hat{\beta}$ stand. difference</td>
<td>7.32</td>
<td>4.29</td>
<td>4.04</td>
<td>4.01</td>
<td><strong>3.89</strong></td>
<td>9.48</td>
</tr>
<tr>
<td>Mean 95% CI overlap</td>
<td>0.47</td>
<td>0.66</td>
<td>0.62</td>
<td>0.64</td>
<td><strong>0.69</strong></td>
<td>0.39</td>
</tr>
</tbody>
</table>

*NOTE: BAG - bagging, RF - random forest, PARA - parametric, SAMP - random sampling with replacement; bold indicates best model*
Analysis-specific utility
Analysis-specific utility

[Box plots showing 95% CI overlap for different methods and models (linear, logistic, Poisson).]
Conclusions

- Useful completely synthetic data can be generated using automated methods
- More research needed to validate synthesising methods in complex settings
- Tree-building procedure matters