DISCUSSION OF “AGGREGATING DISTRIBUTIONAL TREATMENT EFFECTS”

Michal Kolesár (Princeton University)

Jul 2020
SUMMARY

• Paper proposes model for aggregating distributional treatment effects from multiple RCTs
  • Explicitly deals with point masses at zero for some outcomes (profit)
  • Bayesian implementation makes inference straightforward
  • Methodology requires access to original microdata; not a standard “metastudy”

• Nice illustration of:
  1. Value of moving past the ATE
  2. Value of estimating “precise null effects”

• My discussion will focus on:
  1. General considerations when “aggregating evidence”
  2. Quantile treatment effects with non-continuous outcomes
To focus ideas, suppose we’re interested in scalar $\theta_k$ (e.g. QTE at particular quantile), for sites $k = 1, \ldots, K$. Model for data at site $k$, $Y_{ik} \sim f_k(\cdot | \theta_k)$.

For simplicity, suppose model delivers site-specific estimates $\hat{\theta}_k | \theta_k \sim \mathcal{N}(\theta_k, \sigma_k^2)$

Hierarchical models complement this with assumption that across sites $\theta_k \sim g$.

- Not restrictive if left unrestricted, e.g. $g$ could be empirical distribution of $\theta_k$

Possible goals of aggregating evidence $\{\hat{\theta}_1, \ldots, \hat{\theta}_k\}$:

1. Estimate $E[\theta_k]$ (overall average QTE)
2. Predict $\theta_{K+1}$ at new site
3. “Borrow strength” from other sites to improve estimates $\hat{\theta}_1, \ldots, \hat{\theta}_k$
4. Estimate $g$, or features of it, say $\text{var}(\theta_k)$ (learn about TE heterogeneity)
GOAL 1: ESTIMATE $E[\theta_k]$

• “Aggregate results” in slides 15–17 of presentation
• Naive approach: report $K^{-1} \sum_{i=1}^{K} \hat{\theta}_k$, or do “full pooling”
• Hierarchical model estimates typically very similar. Consider partial vs full pooling estimates for QTE on consumption from paper:

<table>
<thead>
<tr>
<th></th>
<th>Partial Pooling</th>
<th>Full Pooling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average</strong></td>
<td>-1.3</td>
<td>-3.9</td>
</tr>
<tr>
<td></td>
<td>(-12.9,10.7)</td>
<td>(-6.8,-0.9)</td>
</tr>
<tr>
<td></td>
<td>-1.3</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>(-12.3,8.4)</td>
<td>(-2.4,2.9)</td>
</tr>
<tr>
<td></td>
<td>-1</td>
<td>-0.9</td>
</tr>
<tr>
<td></td>
<td>(-11.8,8.5)</td>
<td>(-3.7,1.9)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>-1.8</td>
</tr>
<tr>
<td></td>
<td>(-10.9,9.2)</td>
<td>(-5.0,1.4)</td>
</tr>
<tr>
<td></td>
<td>-0.6</td>
<td>-1.2</td>
</tr>
<tr>
<td></td>
<td>(-10.3,10.5)</td>
<td>(-4.8,2.2)</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>(-10.5,13.6)</td>
<td>(-1.4,1.6)</td>
</tr>
<tr>
<td></td>
<td>2.3</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>(-11.9,20.8)</td>
<td>(-0.8,7.9)</td>
</tr>
<tr>
<td></td>
<td>4.3</td>
<td>6.1</td>
</tr>
<tr>
<td></td>
<td>(-15.5,35.8)</td>
<td>(0.2,11.9)</td>
</tr>
<tr>
<td></td>
<td>7.7</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td>(-23.6,63.8)</td>
<td>(-1.8,14.6)</td>
</tr>
<tr>
<td></td>
<td>16.9</td>
<td>13.9</td>
</tr>
<tr>
<td></td>
<td>(-48.9,163.9)</td>
<td>(-6.1,33.9)</td>
</tr>
</tbody>
</table>
GOAL 2: PREDICT $\theta_{K+1}$ AT NEW SITE

• “Predicted quantile effects” in slides 18–21 of presentation
• Requires site $K + 1$ to be drawn from same distribution as sites 1, . . . , $K$
  • Reasonable in observational studies
  • Here across $k$: not just different location, but also different NGOs, loan contracts, interest rates, randomization units and encouragement designs
  • Requires new site not to learn from results in existing studies
• Can again use naive approach, predict $K^{-1} \sum_{i=1}^{K} \hat{\theta}_k$
  • Similarly to Goal 1, value of hierarchical model mostly in delivering uncertainty assessment for prediction (but not robust to misspecification in $g$)
  • Turns out posterior mean $\hat{\theta}_k(\tau) = 0$ for all quantiles $\tau$ and all outcomes…
• Shrinkage/Hierarchical models not appropriate if want good (frequentist) MSE individually for all estimates $\hat{\theta}_k, E[(\hat{\theta}_k - \theta_k)^2 | \theta_k]$
  • This is why we don’t do shrinkage in, say, linear regression: shrinkage introduces bias, can make MSE for individual estimates worse

• Shrinkage appropriate if prioritize favorable group performance over protecting individual performance, i.e. want good average MSE $K^{-1} \sum_{i=1}^{K} E[(\hat{\theta}_k - \theta_k)^2 | \theta_k]$.
  • Overall variance reduction can outweigh overall increase in bias $\implies$ lower average MSE: for James and Stein (1961) shrinkage (motivated by assuming $g$ Gaussian), this is true irrespective of true $g$
  • As with Goal 2, uncertainty assessment not robust to misspecification in $g$, though possible to “robustify” CIs (Armstrong, Kolesár, & Plagborg-Møller, 2020)
• Model in paper also shrinks more extreme quantiles (Bosnia even past overall mean—is this due to smoothing across quantiles?)

• What are the overall gains in precision of estimates? What are the gains from doing this aggregation exercise?
What do we learn about $g$ (i.e. TE heterogeneity) from the data? How variable are TE across sites, relative to prior? Paper only notes that it rejects degenerate $g$.

In principle, could estimate $g$ nonparametrically (large nonparametric empirical Bayes literature) or flexibly (Efron, 2016, 2019), but here $K = 7$ ...

Ideally, with larger $K$, could try to understand reasons for heterogeneity by letting $g$ depend on site-specific covariates (as, e.g., in Chetty & Hendren, 2018; Vivalt, 2020)
Paper takes non-continuity in outcome data seriously: point mass at zero for some variables (e.g. profit)

What goes wrong when we ignore it and use standard quantile regression?

- Quantile estimator $\hat{\theta}_k(\tau)$ for quantiles $\tau$ where CDF jumps no longer asymptotically normal
- But, in a sense, discreteness is good news since estimator converges at faster than $\sqrt{n}$-rate, and puts point mass on $F^{-1}(\tau)$ (intuition: it’s “obvious” from data that there is a jump)
- Could use the same estimator, but validity of inference may be affected

Paper overcomes this by using parametric model $f_k$ for $Y_{ik}$ that allows for point mass at 0.

- Natural given Bayesian setting
- But would we use $f_k$ for estimating QTE at single site? Lose attractive robustness properties of quantile regression (what if model for tails misspecified?)
- Hard to incorporate covariates
• (Frequentist) alternatives to parametric modeling:
  • Use usual estimator, but make sure inference remains valid in presence of mass points (use recent method by Chernozhukov, Fernández-Val, Melly, and Wüthrich (2020): construct confidence bands for CDF, then “flip” the picture; or use conservative normal approximation)
  • Can we directly model extensive margin decision, say using latent variables as in Powell (1986)?
• But I have not thought through the difficulties of nesting these suggestions within a hierarchical framework…


