DISCUSSION OF "AGGREGATING DISTRIBUTIONAL TREATMENT EFFECTS"

Michal Kolesár (Princeton University)

Jul 2020

- 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

GENERAL CONSIDERATIONS WHEN "AGGREGATING EVIDENCE"

- To focus ideas, suppose we're interested in scalar θ_k (e.g. QTE at particular quantile), for sites $k = 1, \dots, K$. Model for data at site k, $Y_{ik} \sim f_k(\cdot \mid \theta_k)$.
- For simplicity, suppose model delivers site-specific estimates $\hat{\theta}_k \mid \theta_k \sim \mathcal{N}(\theta_k, \sigma_k^2)$
- 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 θ_k
- Possible goals of aggregating evidence $\{\hat{\theta}_1, \dots, \hat{\theta}_k\}$:
  1. Estimate $E[\theta_k]$ (overall average QTE)
  2. Predict θ_{K+1} at new site
  3. "Borrow strength" from other sites to improve estimates $\hat{\theta}_1, \dots, \hat{\theta}_k$
  4. Estimate g, or features of it, say $\text{var}(\theta_k)$ (learn about TE heterogeneity)

- "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:

Partial Pooling										
Average	-1.3	-1.3	-1	-0.6	0	1	2.3	4.3	7.7	16.9
	(-12.9,10.7)	(-12.3,8.4)	(-11.8,8.5)	(-10.9,9.2)	(-10.3,10.5)	(-10.5,13.6)	(-11.9,20.8)	(-15.5,35.8)	(-23.6,63.8)	(-48.9,163.9)
Full Pooling										
Average	-3.9	0.2	-0.9	-1.8	-1.3	2.5	3.6	6.1	6.4	13.9
	(-6.8,-0.9)	(-2.4,2.9)	(-3.7,1.9)	(-5,1.4)	(-4.8,2.2)	(-1.4,6.3)	(-0.8,7.9)	(0.2,11.9)	(-1.8,14.6)	(-6.1,33.9)

Goal 2: Predict θ_{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 $1, \ldots, K$
  - 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 τ and all outcomes…

Goal 3: "Borrow strength" from other sites to improve estimates

- Shrinkage/Hierarchical models not appropriate if want good (frequentist) MSE individually for all estimates $\hat{\theta}_k$, $E[(\hat{\theta}_k - \theta_k)^2 \mid \theta_k]$
    - 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 $K^{-1} \sum_{i=1}^{K} E[(\hat{\theta}_k - \theta_k)^2 \mid \theta_k]$.
    - 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)

Quantile:	65th	75th	85th	95th
No Pooling				
Bosnia	-16.3	-34.4	-64.5	104
	(-46.2,13.6)	(-74.9,6.1)	(-131.1,2.2)	(-77.4,285.5)
India	2.2	4.6	8.2	40.1
	(-6.3,10.7)	(-6.3,15.6)	(-7.5,24)	(-4.5,84.7)
Mexico	5.5	11	13.2	16.6
	(0,11)	(2.7,19.2)	(1.8,24.7)	(-6.7,39.9)
Mongolia	-0.9	-2.5	-12.8	87.4
	(-32.7,30.9)	(-42.1,37.1)	(-70.3,44.8)	(-40.6,215.4)
Morocco	3.7	0.4	-6.4	-54
	(-7.3,14.7)	(-12.4,13.2)	(-23.8,11)	(-104,-4)
Partial Pooling				
Bosnia	-1.1	2.6	11.8	52.4
	(-14.7,8.6)	(-19.4,20.9)	(-30.4,52.1)	(-75.8,188.3)
India	2.4	4.3	7.5	16
	(-4,9)	(-4,12.8)	(-4.2,19.6)	(-5.6,37.9)
Mexico	3.9	8	15.1	34.1
	(-1.7,9.3)	(0.7,15)	(4.8,25.1)	(15.5,52.7)
Mongolia	5.8	10.3	18	38.4
	(-5.7,26.5)	(-7.3,36.2)	(-10.6,55)	(-22.4,108)
Morocco	-2.2	-4.6	-8.7	-18.8
	(-9.6,5)	(-14,4.4)	(-21.7,4)	(-41.5,3.3)
Average	2.3	4.3	7.7	16.9
	(-11.9,20.8)	(-15.5,35.8)	(-23.6,63.8)	(-48.9,163.9)

- 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 $K = 7 \ldots$
- 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 τ 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…

REFERENCES

Armstrong, T., Kolesár, M., & Plagborg-Møller, M. (2020). *Robust empirical Bayes confidence intervals.* (Tech. rep. No. 2004.03448). arXiv. Retrieved from https://arxiv.org/abs/2004.03448

Chernozhukov, V., Fernández-Val, I., Melly, B., & Wüthrich, K. (2020). Generic inference on quantile and quantile effect functions for discrete outcomes. *Journal of the American Statistical Association*, *115*(529), 123–137. doi:10.1080/01621459.2019.1611581

Chetty, R., & Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility II: County-level estimates. *The Quarterly Journal of Economics*, *133*(3), 1163–1228. doi:10.1093/qje/qjy006

Efron, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika*, *103*(1), 1–20. doi:10.1093/biomet/asv068

Efron, B. (2019). Bayes, Oracle Bayes and empirical Bayes. *Statistical Science*, *34*(2), 177–201. doi:10.1214/18-STS674

James, W., & Stein, C. M. (1961). Estimation with quadratic loss. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 361–379). Berkeley, CA: University of California Press. Retrieved from https://projecteuclid.org/euclid.bsmsp/1200512173

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics*, *32*(1), 143–155. doi:10.1016/0304-4076(86)90016-3

Vivalt, E. (2020). How much can we generalize from impact evaluations? *Journal of the European Economic Association*, *forthcoming*.