Estimating Causal Effects with Zero-Inflated Outcomes

Luis F. Campos, Lo-Hua Yuan, Luke W. Miratrix, Avi Feller

October 1, 2017

Background

Outcomes with excessive zeros commonly occur in randomized experiments across many fields in social science. For example, we might estimate the consequence of periodic surveys on students' participation activity during a massive online open course (MOOC). If we use the number of posts in online discussions as a measure of engagement, the majority of students who initially sign up for the MOOC will not partake in any online discussions and hence have zero engagement. But the distribution of engagement levels for those who do participate is right-skewed with a very long tail[1]. This is further complicated by introducing randomized experiments targeted at increasing engagement. A treatment effect might show itself in several ways: a smaller proportion of disengaged students, an overall increase in mean engagement, or an increase in mean engagement amongst those who showed any sign of engagement. Various statistical models have been developed to handle count data with excess zeros, the two basic ones being the zero-inflated Poisson model and the hurdle model.

Although extensive progress has been made on how to fit these models, the statistics literature provides scarce guidance on how to conduct causal inference for data with excess zeros. Typically, papers presenting results from a zero-inflated or hurdle regression analysis show separate treatment effect estimates for the susceptible probability (corresponding to the structural zeros model) and the susceptible population mean (corresponding to the count model). While the causal estimands are relatively straightforward for some models, e.g. the hurdle model, other models emit more complex estimands and thus complicate the analysis as well. For example, the zero-inflated case lends itself to principal stratification because some zeros will be structural while others will come from the count model. When comparing the susceptible probabilities the effect of interest is called the extensive marginal effect (the difference in proportion of structural zeros). Alternatively, when comparing the susceptible population means the effect of interest is called the intensive marginal effect (the difference in means for non-structural zeros). Furthermore, in randomized experiments and observational studies of an exposure, the primary interest is typically in the comparison between treatment groups based on the overall mean, possibly adjusted for baseline covariates.

Previous Work

Despite a plethora of zero-inflated outcome data in real-world applications, there is a paucity of publications in the causal inference literature that discuss the special considerations we need to address when estimating treatment effects for data with excessive zeros. The majority of papers dealing with zero-inflated outcomes take a heavily model-dependent approach to estimate treatment effects (Böhning *et al.* (1999)[2], Yau and Lee (2001)[10], Heroux *et al.* (2014)[11] and DeSantis *et al.* (2014)[12]). Recently, Lee (2017)[4] consider extensive and intensive causal effects via principal stratification and modeling, while Keele and Miratrix (2017)[5] consider randomization-based inference for an assortment of hypothesis tests generally considered for zero-inflated outcome data. Most previous work is fundamentally rooted in the zero-inflated Poisson model, and with the exception of a few works (DeSantis (2014), Lee (2017) and Keele and Miratrix (2017)), they generally do not define causal effects in terms of the Neyman-Rubin potential outcomes framework.

Contributions

We link the causal inference literature with literature on excess-zero models. With explicit focus on a potential outcomes framework, we deliberately define the causal estimand of interest as a step distinct from, and *a priori* of, the specification of a regression model or estimator. The potential outcomes framework is a powerful tool used to draw conclusions from experimental data. Working with complex data and models, e.g. those with excess-zeros, can make it difficult to assess these causal quantities. So we first investigate both model-free and model-based methods for estimating causal intention-to-treat (ITT) effects and conduct a small simulation study to compare the finite sample performance of various types of confidence intervals. We then use the principal stratification framework to define the impact of the intervention on those individuals who would have a non-zero outcome regardless of treatment assignment, the intensive marginal causal effect. While much of the previous literature has focused on the zero-inflated Poisson model, we work in the most general case of zero-inflated distributions and hurdle models. For the MOOC example we define the ITT, extensive and intensive causal effects. We use several models to illustrate how different estimates target particular estimands and thus have a particular interpretation. Knowing the extensive and intensive causal effects can help shape new interventions by separating the effect on overall participation from the volume of engagement. For example, if there are only modest gains from the intervention to the intensive marginal effect, we may abandon attempting to increase the volume of participation in favor of treatments that only target overall participation. The results will allow for targeted causal inference in massive open online courses and help not only draw strong conclusions from the experiment but also help shape future development.

References

[1] Anne Lamb, Jascha Smilack, Andrew Ho, Justin Reich (2014). "Addressing Common Analytic Challenges to Randomized Experiments in MOOCs: Attrition and Zero-Inflation." *L@S 2015*. March 14–18, 2015, Vancouver, BC, Canada.

[2] Böhning, D., Dietz, E., Schlattmann, P., Mendonça, L. and Kirchner, U. (1999). "The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **162**: 195–209.

[3] Holland, Paul W. (1986). "Statistics and Causal Inference." *Journal of the American Statistical Association.* **81**(396): 945-960.

[4] Lee, Myoung-jae (2017) "Extensive and intensive margin effects in sample selection models: racial effects on wages". *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, Part 3, pp. 817–839.

[5] Luke Keele, Luke W. Miratrix (2017). "Randomization Inference for Outcomes with Clumping at Zero." Paper under revision at *Annals of Applied Statistics.*

[6] Imbens, G., Rubin, D. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* 2015.

[7] Rubin, Donald. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology* **66**(5), 688.

[8] Rubin, D. B. (1980). "Discussion of 'Randomization analysis of experimental data in the Fisher randomization test' by Basu." *J. Amer. Statist. Assoc.* **75**, 591–593.

[9] Lambert, D. (1992). "Zero-inflated poisson regression, with an application to defects in manufacturing." *Technometrics*, **34**(1):1–14.

[10] Yau KKW, Lee AH. (2001). "Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention program." *Statistics in Medicine.* **20**:2907–2920.

[11] Héroux, J., Moodie, E. E.M., Strumpf, E., Coyle, N., Tousignant, P. and Diop, M. (2014). "Marginal structural models for skewed outcomes: identifying causal relationships in health care utilization." *Statist. Med.*, **33**: 1205–1221.

[12] DeSantis SM, Lazaridis C, Ji S, Spinale FG (2014). "Analyzing Propensity Matched Zero-Inflated Count Outcomes in Observational Studies." *Journal of Applied Statistics*.

[13] Statistical Consulting Group, UCLA Academic Technology Services. "Regression models with Count Data." April 2007. `http://www.ats.ucla.edu/stat/stata/seminars/count_presentation/count.htm`.