Evaluating Causal Inference Techniques Using the g-formula in Two Different Implementations

Morgan F. Breitmeyer

February 20, 2017

1 Research Questions of Interest

In the realm of observational data, causal inference is frequently used to identify causal relationships. The goal is to effectively simulate a randomized trial experiment retroactively. The most widely used method for doing this is through using the g-formula, which parametrically estimates an average casual effect. This method depends on finding parametric models, which are marginal distributions of covariates in the data which can be empirically estimated. Subsequently, a Monte Carlo simulation is used on the estimated parameters so estimates of the causal effects can be measured.

The goal of the thesis is to implement a sequential version of the g-formula as a proof of concept. This has yet to be done in a sophisticated manner and will hopefully be scalable for wider research. In order to prove its efficacy, it will be applied to a dataset of interest, likely something of the epidemiological sort. The two implementations of the g-formula will be compared on this dataset and hopefully conclusions will be drawn. The biggest questions are whether implementation of this version of the g-formula is possible, whether it will be successful, and if it's more efficient than the existing implementation such that it could be scalable to large data sets.

I will be advised by Dave Harrington and Miguel Hernan (and perhaps Jamie Robbins in a less official capacity).

2 Outline

- November 10: Literature review and developing a deep understanding of causal inference and g-formulas. This will include building out why causal inference is of interest, the intuition behind both implementations of the g-formula, and outlining the limitations and assumptions of these methods. This part of the process may take longer than just November, but will at least be a significant start.
- End of 2016/early 2017: Hopefully, I will have the implementation of the formula almost complete. I will be using python to do this.
- January 2017: Do analysis of a dataset using the two methods. Building out the two models and their likelihood models will likely take quite a bit of time so I want to leave room in here for this.

- February/March 2017: Wrapping everything up and explaining all of the analyses and how the implementations worked/differed/etc.

References

[1] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[2] Miguel A Hernan and James M Robins. *Causal Inference*. Chapman & Hall/CRC, 2016.

[3] Sara Lodi, Andrew Phillips, Roger Logan, Ashley Olson, Dominique Costagliola, Sophie Abgrall, Ard van Sighem, Peter Reiss, José M Miró, Elena Ferrer, et al. Comparative effectiveness of immediate antiretroviral therapy versus cd4-based initiation in hiv-positive individuals in high-income countries: observational cohort study. *The Lancet HIV*, 2(8):e335–e343, 2015.

[4] James D Wright. International encyclopedia of the social and behavioral sciences. 2015.

[5] Jessica G Young, Lauren E Cain, James M Robins, Eilis J O?Reilly, and Miguel A Hernán. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Statistics in biosciences*, 3(1):119–143, 2011.