Baseline Equivalency Measures in High Dimensions

Mark M. Fredrickson   Ben Hansen

October 1, 2019

1 Context

Assessments of baseline equivalency of intervention and control groups, "balance," play a critical role in evaluating educational interventions. The What Works Clearinghouse (WWC) of the Institute of Educational Studies (IES) classifies research designs on their ability to justify causal claims. The highest standard, "Meets WWC Design Standards Without Reservations", is reserved for randomized controlled trials (RCTs) without substantial attrition (IES, 2017), though specific subject matter protocols may also require demonstration of baseline equivalency on the primary outcome measures (e.g., IES, 2018). Randomized controlled trials with high rates of attrition and observational quasi-experimental designs may receive the "Meets WWC Design Standards With Reservations" designation if they able to demonstrate balance on critical baseline variables.

Balance assessments are also useful in study design. Good balance increases statistical efficiency, and comparable treatment and control groups makes equivalency easier to show if attrition is substantial. Blocking, grouping units by baseline variables, ensures balance on those variables, and balance assessments reveal whether differences on other variables are controlled as well. Extending this principle, "restricted randomization" limits the treatment regimen to those with no more than a given amount of imbalance (Morgan and Rubin, 2012).

To quantify baseline equivalency, a common choice to combine the imbalance across several variables is the (pre-treatment) Mahalanobis distance between treated and control group means (Hansen and Bowers, 2008; Morgan and Rubin, 2012):

$$M = \frac{n_1 n_0}{(n_1 + n_0)} \left(\bar{X}_1 - \bar{X}_0\right)' \Sigma^{-1} \left(\bar{X}_1 - \bar{X}_0\right),$$

where n_j and $\bar{X}_j$ are the number and means of the treatment ($j = 1$) and control ($j = 0$) assignment units, often schools in the educational setting. Here Σ^{-1} is an inverted sample covariance matrix for the baseline variables.

In asymptotics with a fixed number of covariates and the number of assignment units tending to infinity, $\left(\bar{X}_1 - \bar{X}_0\right)' \Sigma^{-1/2}$ tends to a multivariate Normal distribution, which implies M has a well-behaved distributional limit (Li et al., 2018), ordinarily χ^2 on rank(Σ) degrees of freedom (Hansen and Bowers, 2008). Recent developments in high dimensional probability caution that these results may not hold when the number of variables also increases without bound. Insofar as high dimensional balance assessment has been considered, previous authors suggest the use of a generalized inverse or alternative covariance matrix instead of Σ^{-1} to solve the numerical problems, but fail to consider that the usual limiting distribution may no longer apply (Hansen and Bowers, 2008; Morgan and Rubin, 2012; Branson and Shao, 2018).

2 Research Questions

Our first research question is, "Does the usual asymptotic approximation of M using a χ^2 distribution hold when the number of baseline variables approaches or exceeds the number of assignment units?" The number of variables in educational interventions can be quite large relative to the number of assignment units, particularly for cluster randomized trials. The WWC protocol for basic reading interventions, for example, requires demonstrating balance on twelve background and demographic variables in addition to pre-test or proxy measures of all primary outcomes (IES, 2014). Including interactions with key subgroups multiplies the number of variables required.

If the usual approximation proves inadequate, our next research question is, "Can the true distribution of M be better approximated using higher moments of the distribution?" The asymptotic approximation based on a fixed number of variables uses only the first moment of the distribution. Theory suggests the first moment of the distribution would remain unchanged, but the second and higher moments may provide useful information on the shape of the true distribution, allowing for more accurate approximations.

Finally, our third research questions asks, "Are there alternative statistics that provide better balance assessments when the number of variables is large?" If the distribution of M sufficiently degenerates, it becomes uninformative about baseline equivalency or which randomizations minimize imbalance. Modifications of M or alternative measures of discrepancy between the treated and control groups may be preferable in the high-dimensional setting.

3 Setting

We investigate our first research question by analyzing a school based RCT reported by Gamoran et al. (2012) fielded in Phoenix, Arizona and Austin, Texas.

4 Subjects

We focus on the 1477 first grade children and their families recruited at 26 participating schools in Phoenix, AZ.

Variable	Number of Categories	Level
Date of Birth	7	Student
Gender	3	Student
Gender Impute from Records	3	Student
Race/Ethnicity	6	Student
Race/Ethnicity Impute from Records	3	Student
Free/Reduce Lunch	4	Student
English Language Learner	3	Student
Special Ed in 1st Grade	3	Student
Attended Kindergarten	3	Student
Retained in Kindergarten	3	Student
Percentage Meets Average Yearly Progress in Math	3	School
Percentage Meets Average Yearly Progress in Reading	3	School
Percentage Free/Reduce Lunch	3	School
Percentage Race/Ethnicity White	2	School
Percentage Race/Ethnicity Black	2	School
Percentage Race/Ethnicity Hispanic	2	School
Percentage Race/Ethnicity Other	2	School

Table 1: Variables available for reanalysis of study design for Gamoran et al. (2012).

5 Intervention

Within treatment schools, families were encouraged to attend events based on the FAST social capital building curriculum. Additional details are given Gamoran et al. (2012).

6 Research Design

Schools were located in three Phoenix area districts (with six, eight, and twelve schools). We consider the distribution of M for randomizations within these blocks. Table 1 gives the variables included in the data release. While several variables would conceptually be continuous, the data were released with categorical variables to ensure anonymity of subjects. Our analysis adds variables one at a time, so we performed principal components analysis to create 25 linearly independent variables ordered by variance.

7 Analysis

We randomly generated 10,000 treatment allocations and computed M. Figure 1 presents the empirical distribution of M as the number of variables is increased. As the number of variables approaches the number of assignment units, the distribution tends to cluster near its mean more closely than the theoretical χ^2 distribution. After 22 variables, the number of schools less the number of blocks, the distribution becomes a constant.

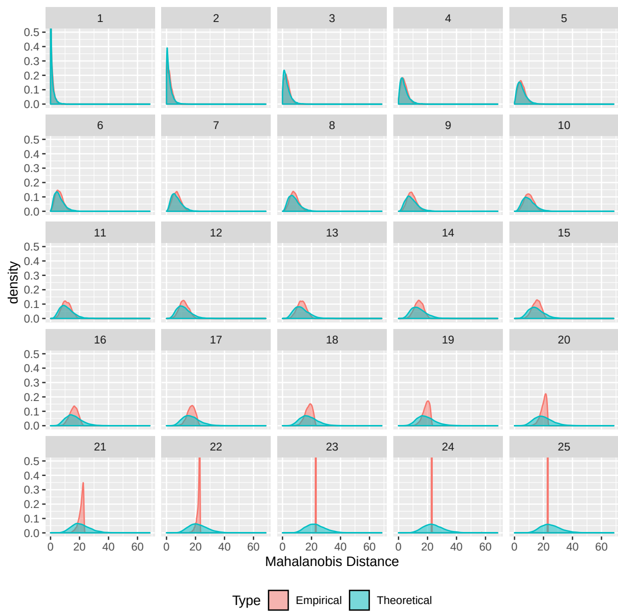

Figure 1: Empirical and theoretical (χ^2_k) Mahalanobis distance distributions for the 26 schools in Phoenix, AZ from the study by Gamoran et al. (2012) as the number of baseline variables is increased.

8 Conclusions

Our first research question asked if the usual limiting distribution of M holds in typical educational RCTs, such as those funded by the IES. The reanalysis of the Gamoran et al. data suggests the usual χ^2 approximation suffers when the number of baseline variables approaches or exceeds the number of sample units. We have begun the process of characterizing the distribution of M and have developed computationally tractable expressions for the second moment. We are currently investigating the usefulness of the second moment to improve the approximation. We are exploring statistics that remain robust in the presence of many baseline variables. Recent years have seen an increase in statistical methods for high dimensional settings. While the assumptions of these methods are inappropriate for RCTs, similar techniques may lead to developments for educational interventions.

References

Branson, Z. and Shao, S. (2018). Ridge rerandomization: An experimental design strategy in the presence of collinearity. *arXiv e-prints*, page arXiv:1808.04513.

Gamoran, A., Turley, R. N. L., Turner, A., and Fish, R. (2012). Differences between hispanic and non-hispanic families in social capital and child development: First-year findings from an experimental study. *Research in Social Stratification and Mobility*, 30(1):97 – 112. Inequality across the Globe.

Hansen, B. B. and Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23:219.

IES (2014). Review protocol for beginning reading interventions. Technical report, Institute of Education Sciences. Version 3.

IES (2017). What works clearinghouse™ standards handbook. Technical report, Institute of Education Sciences. Version 4.0.

IES (2018). Review protocol for primary mathematics. Technical report, Institute for Educational Studies. Version 4.0.

Li, X., Ding, P., and Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162.

Morgan, K. L. and Rubin, D. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282.