Center for Assessment Summer Internship 2021

Tackling Technical Challenges to Analyzing 2021 Assessment Data

Allie Cooperman

Supporting Research by Damian Betebenner and Adam Van Iwaarden

June 18, 2021

Internship "Theme"

If we are going to measure student learning and achievement using 2021 assessment data...

how can we best ensure an "apples-to-apples" comparison across samples?

Potential Roadblocks to Making Appropriate Comparisons

- Due to differential participation rates, "opt-out" testing, and other factors, there may be large amounts of missingness in the 2021 assessment data.
- Changing enrollment patterns may result in two samples (e.g., 2019 and 2021 students) with substantially different demographic compositions.
- Can we "adjust" our scale score and student growth percentile (SGP) analyses to foster more comparable samples?
- In what data contexts are these adjustments plausible?

Potential Solutions for Making Appropriate Comparisons

Missing Data $\rightarrow$ Multiple Imputation

Covariate Imbalance $\rightarrow$ Propensity Score Weighting

Update Overview

Multiple Imputation (MI)

- Fit a series of regression models to identify factors associated with MI efficacy.
- Summarized results from a new simulation evaluating MI when a COVID-19 impact is present.

Propensity Score Weighting (PSW)

- Learning the basics of PSW for cross-sectional studies.
- Demonstrated how to apply PSW to non-hierarchical and two-level educational assessment data using R.

Reproducibility

- Continually getting to know the basics of GitHub.
- Created a basic personal webpage on GitHub.

Multiple Imputation: Simulating a COVID-19 Impact

- Data were systematically removed according to varying missingness percentages and types.

- Six MI methods were compared:
  - **Previously Examined**: Cross-sectional L2PAN, longitudinal L2PAN, quantile regression, predictive mean matching
  - **New**: Random forest and multilevel predictive mean matching

- MI efficacy was evaluated in terms of (a) percent bias, (b) simplified confidence interval coverage rates (Vink & van Buuren, 2014), and the simplified F_1 statistic (van Buuren, 2018).

Multiple Imputation: Simulating a COVID-19 Impact

- Many trends replicated from the "no impact" simulation, with cross-sectional L2PAN more often outperforming the other methods.
- MI methods tended to function more similarly (as poor-performing methods were removed and more viable candidates were introduced).
- There were noticeable differences by grade, with higher bias and lower coverage rates for imputed scale scores among grades 3 and 4 when data were missing at random based on status and growth.

Multiple Imputation Simulation Take-Aways

MI (with cross-sectional L2PAN) appears to be a viable method for dealing with missing educational assessment when

- Less than 50% of data are missing
- Data are missing completely at random or missing at random based on more factors than just status and growth
- School or grade/content area sizes are relatively large

MI's accuracy will likely differ by school and grade.

"Asterisks" for Applying Multiple Imputation

- Missingness patterns should be examined prior to addressing the missing data, and diagnostic checks included to evaluate MI's performance in a given data set.

- Analyses can be run with and without including MI, highlighting whether inferences generalize across the methods.

- It is difficult (if not impossible?) to identify "one-size-fits-all" guidelines for when MI should be used; rather, analyses will likely be individualized based on the idiosyncrasies of a data set.

What Is Propensity Score Weighting?

A **propensity score** (Rosenbaum & Rubin, 1983, 1984; as cited in Li et al., 2013) is defined as

$$P(T_i = 1|\mathbf{x}_i)$$

"the probability of receiving a treatment conditional on a set of observed covariates" (Lee et al., 2010, p. 337; Rosenbaum & Rubin, 1983).

Propensity score weighting uses propensity scores to make the covariate distributions between two samples more similar (Desai & Franklin, 2019; Li et al., 2013).

Propensity Score Weighting for Educational Assessments

- Researchers and policymakers often want to compare mean scale scores and SGPs across years (e.g., fifth graders in 2019 compared to fifth graders in 2021).
- However, the student compositions in the two samples may look dramatically different due to factors like enrollment changes.
- Propensity score weighting has been used in previous cross-sectional education studies (e.g., Liu et al., 2016), and may be an alternative to proposed methods like Ho's (2021) Fair Trend metric.

Basic Steps of Propensity Score Weighting

1. Select a set of important covariates to balance.
2. Estimate propensity scores by regressing the grouping variable on the covariates.
3. Compute weights and evaluate covariate balance.
4. Apply weights to the analysis (e.g., estimating mean scale score differences).
5. Perform sensitivity analyses.

e.g., Desai & Franklin, 2019; Leite et al., 2015; Liu et al., 2016; Ridgeway et al., 2021

Basic Steps of Propensity Score Weighting

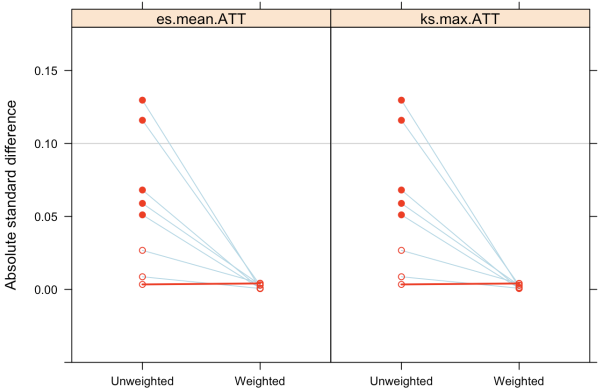

Figure 1: Example figure from the ‘twang‘ R package (Cefalu et al., 2021) showing standardized differences on a set of covariates with and without propensity score weighting

Propensity Score Weighting Methods

Numerous methods for propensity score weighting have been proposed, including

- Different weighting and estimation approaches (e.g., logistic regression, gradient boosted decision trees, etc.);
- Applications to different estimands;
- Approaches for multilevel data (e.g., estimating propensity scores using a random intercept and slope model); and
- Approaches for longitudinal studies where selective attrition is a concern

Desai & Franklin, 2019; Burgette et al., 2016; Lee et al., 2010; Leite et al., 2015; Li et al., 2013; Weuve et al., 2012

Pondering Questions and Future Directions

- In what contexts might propensity score weighting work well for our research questions of interest?
- How does propensity score weighting compare to Ho's (2021) Fair Trend metric?
- What other factors influence the efficacy of these methods?
- What other technical challenges may arise when analyzing 2021 assessment data (e.g., comparing assessment modalities)?

References

- Burgette, J. M., Preisser, J. S., & Rozier, R. G. (2016). Propensity score weighting: An application to an Early Head Start dental study. *Journal of Public Health Dentistry, 76*(1), 17-29. https://doi.org/10.1111/jphd.12106
- Cefalu, M., Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., & Burgette, L. (2021). twang: Toolkit for weighting and analysis of nonequivalent groups. R package version 2.1. https://CRAN.R-project.org/package=twang
- Desai, R. J., & Franklin, J. M. (2019). Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: A primer for practitioners. *The British Medical Journal, 367*, 15657. https://doi.org/10.1136/bmj.l5657
- Ho, A. (2021, Feburary 26). *Three test-score metrics that all states should report in the COVID-19-affected spring of 2021.* Harvard Graduate School of Education. https://scholar.harvard.edu/files/andrewho/files/threemetrics.pdf
- Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine, 29*(3), 337-346. https://doi.org/10.1002/sim.3782

References

- Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research, 50*(3), 265-284. https://doi.org/10.1080/00273171.2014.991018
- Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine, 32*(19), 3373-3387. https://doi.org/10.1002/sim.5786
- Liu, O. L., Liu, H., Roohr, K. C., & McCaffrey, D. F. (2016). Investigating college learning gain: Exploring a propensity score weighting approach. *Journal of Educational Measurement, 53*(3), 352-367. https://doi.org/10.1111/jedm.12112
- Ridgeway, G., McCaffrey, D., Morral, A., Cefalu, M., Burgette, L., Pane, J., & Griffin, B. A. (2021, June 5). *Toolkit for weighting and analysis of nonequivalent groups: A guide to the twang package*. Retrieved from https://cran.r-project.org/web/packages/twang/vignettes/twang.pdf
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

References

- Rosenbaum, P., & Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524.

- van Buuren, S. (2018). *Flexible imputation of missing data*. CRC Press. https://stefvanbuuren.name/fimd/

- Vink, G., & van Buuren, S. (2014). Pooling multiple imputations when the sample happens to be the population. arXiv Pre-Print 1409.8542.

- Weuve, J., Tchetgen, E. J. T., Glymour, M. M., Beck, T. L., Aggarwal, N. T., Wilson, R. S., Evans, D. A., & Mendes de Leon, C. F. (2012). Accounting for bias due to selective attrition: The example of smoking and cognitive decline. *Epidemiology, 23*(1), 119-128. https://doi.org/10.1097/EDE.0b013e318230e861