NIH Diversity Program Consortium (DPC) Methods Guide for Research Publications Using Consortium-Wide Data

OVERVIEW

This document serves as an introduction to common social science statistical methods used for analyzing large sample, cross-sectional and longitudinal survey data. It is not meant to be all-inclusive or methodologically prescriptive; rather, it is intended to provide basic information and resources on methods that can be used to address questions based on Consortium-Wide Evaluation Plan (CWEP) data and the DPC Hallmarks of Success. It may also assist in completing the required sections of the DPC Manuscript Proposal form to request CWEP data.

The following section of this document lists potential empirical research questions aimed at examining Student- and Faculty-level outcomes that correspond to the DPC Hallmarks of Success. These questions are framed broadly to capture the experiences of students and faculty at BUILD and non-BUILD institutions. They leave room for developing more specific research questions. For each question, the corresponding Hallmarks of Success and surveys are identified (see "Mapping Data Elements to Hallmarks of Success"). These surveys are:

- TFS= HERI The Freshman Survey
- CSS = HERI College Senior Survey
- FAC = HERI Faculty Survey
- FAC STEM = HERI Faculty Survey STEM Module
- FAC MENTOR = HERI Faculty Survey Mentoring Module
- D SAFS = DPC Student Annual Follow up Survey
- D FAFS = DPC Faculty Annual Follow up Survey

Based on the measured outcomes, possible models for analyzing the data are also identified. The final section of this document provides a description and resources for each statistical method. These methods include:

- Multiple Linear Regression
- Hierarchical Linear Modeling
- Binomial Logistic Regression
- Ordinal Logistic Regression
- Multinomial Logistic Regression
- Sequential Logistic Regression
- Structural Equation Modeling
- Difference in Difference Estimation

The information provided on these methods serves as a general overview; the statistical models used will depend on the nature of the question and the fit to the data, such as how the explanatory variables and outcome are measured or constructed (for guidance, see "Choosing the Correct Statistical Test" under the Other Resources section). A variety of data issues will also need to be considered, including but not limited to linearity and non-linearity, sample size and power, non-independence or clustering of observations, and missingness.

POTENTIAL EMPIRICAL RESEARCH QUESTIONS

STUDENTS

- What factors are associated with pursuing a biomedical science degree (versus a non-biomedical science degree) for BUILD students as compared to non-BUILD students?
    - Relevant Hallmarks of Success: STU-7, STU-8, STU-9

    - Survey Sources of Measures: TFS, CSS, D-SAFS
    - Possible Method(s) Based on Data: Multinomial Logistic Regression, Ordinal Logistic Regression, Binomial Logistic Regression, Sequential Logistic Regression
- What factors are associated with students' high levels of participation in research training?
    - Relevant Hallmarks of Success: STU-11, STU-14
    - Survey Sources of Measures: CSS, D-SAFS
    - Possible Method(s) Based on Data: Ordinal Logistic Regression, Binomial Logistic Regression, Hierarchical Linear Modeling
- What factors are associated with students' high levels of satisfaction with faculty mentorship?
    - Relevant Hallmarks of Success: STU-4
    - Survey Sources of Measures: CSS, D-SAFS
    - Possible Method(s) Based on Data: Ordinal Logistic Regression, Hierarchical Linear Modeling
- How do BUILD students perceive themselves as socially integrating or fitting in on campus, and how does it affect their academic performance or research engagement?
    - Relevant Hallmarks of Success: STU-1, STU-2, STU-3, STU-5, STU-6
    - Survey Sources of Measures: TFS, CSS, D-SAFS
    - Possible Method(s) Based on Data: Structural Equation Modeling, Ordinal Logistic Regression
- To what extent are interventions associated with BUILD students' intent to pursue a career in biomedical research as compared to non-BUILD students?
    - Relevant Hallmarks of Success: STU-7, STU-9, STU-10, STU-16, STU-17, STU-18
    - Survey Sources of Measures: CSS, D-SAFS
    - Possible Method(s) Based on Data: Multinomial Logistic Regression, Binomial Logistic Regression, Sequential Logistic Regression, Difference in Difference Estimation
- What factors are associated with students' biomedical education and career trajectories?
    - Relevant Hallmarks of Success: STU-7, STU-8, STU-9, STU-16, STU-17, STU-18
    - Survey Sources of Measures: CSS, D-SAFS
    - Possible Method(s) Based on Data: Multinomial Logistic Regression, Binomial Logistic Regression, Sequential Logistic Regression

FACULTY

- What factors are associated with BUILD faculty members' mentoring self-efficacy as compared to non-BUILD faculty members?
    - Relevant Hallmarks of Success: FAC-1, FAC-2, FAC-3, FAC-4, FAC-16
    - Survey Sources of Measures: FAC, FAC-STEM, FAC-MENTOR, D-FAFS
    - Possible Method(s) Based on Data: Ordinal Logistic Regression, Multiple Linear Regression, Binomial Logistic Regression, Structural Equation Modeling
- What factors are associated with faculty members' increased mentorship of diverse students and early career researchers?
    - Relevant Hallmarks of Success: FAC-3, FAC-4, FAC-5
    - Survey Sources of Measures: FAC, FAC-MENTOR, D-FAFS
    - Possible Method(s) Based on Data: Ordinal Logistic Regression, Multiple Linear Regression, Binomial Logistic Regression, Structural Equation Modeling
- What factors or interventions are associated with high-quality faculty mentorship?
    - Relevant Hallmarks of Success: FAC-3, FAC-4, FAC-5, FAC-17

    - Survey Sources of Measures: FAC, FAC-MENTOR, D-FAFS
    - Possible Method(s) Based on Data: Ordinal Logistic Regression, Multiple Linear Regression, Binomial Logistic Regression, Structural Equation Modeling, Difference in Difference Estimation
- What factors or interventions are associated with high research productivity for BUILD faculty as compared to non-BUILD faculty?
    - Relevant Hallmarks of Success: FAC-8, FAC-9, FAC-10, FAC-12
    - Survey Sources of Measures: FAC, D-FAFS
    - Possible Method(s) Based on Data: Binomial Logistic Regression, Ordinal Logistic Regression, Multiple Linear Regression, Difference in Difference Estimation
- To what extent are professional development activities associated with faculty members' increased research productivity?
    - Relevant Hallmarks of Success: FAC-8, FAC-9, FAC-11, FAC-12
    - Survey Sources of Measures: FAC, D-FAFS
    - Possible Method(s) Based on Data: Binomial Logistic Regression, Ordinal Logistic Regression, Difference in Difference Estimation

DESCRIPTIONS AND RESOURCES ABOUT STATISTICAL METHODS

Multiple Linear Regression: can be used to model the relationship between multiple explanatory variables and an outcome variable, when several assumptions and conditions are met, including: linearity, homoscedasticity, uncorrelated residuals, absence of multicollinearity, normality of residuals. For censored dependent variables, a related method is tobit regression.

- Gujarati, Damodar N. 2018. *Linear Regression: A Mathematical Introduction*. Vol. 177. Thousand Oaks, CA: Sage Publications. https://us.sagepub.com/en-us/nam/linear-regression/book262447
- Lewis-Beck, Colin and Michael Lewis-Beck. 2015. *Applied Regression: An Introduction*. Vol. 22. Thousand Oaks, CA: Sage Publications. https://us.sagepub.com/en-us/nam/applied-regression/book244616
- Berry, William D. and Stanley Feldman. 1985. *Multiple Regression in Practice*. No. 50. Thousand Oaks, CA: Sage Publications. https://us.sagepub.com/en-us/nam/multiple-regression-in-practice/book471

Hierarchical Linear Modeling (HLM)/Multi-Level Modeling (MLM): a form of ordinary least squares (OLS) regression for analyzing data when the explanatory variables vary at more than one level (e.g., hierarchical or nested data), such as students within classrooms within universities. Note: this method applies to analysis that merges CWEP and Institutional Record data.

- Raudenbush, Stephen W. and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd ed. Thousand Oaks, CA: Sage Publications.
- Snijders, Tom A. B. and Roel J. Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Thousand Oaks, CA: Sage Publications.
- Multi-Level Modeling: https://www.mailman.columbia.edu/research/population-health-methods/multi-level-modeling

Logistic Regression (Binomial, Ordinal, and Multinomial Logistic Regression): for analyzing outcomes that are measured as binary, ordered (i.e., Likert scale), or categorical variables. Assumptions of this method include: appropriate outcome structure, independent observations, absence of multicollinearity, linearity of independent variables and log odds. A related method is probit regression.

- Stoltzfus, Jill C. 2011. "Logistic Regression: A Brief Primer." *Academic Emergency Medicine* 18(10): 1099-1104. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1553-2712.2011.01185.x
- Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.
- Long, J. Scott and Jeremy Freese. 2006. *Regression Models for Categorical and Limited Dependent Variables Using Stata*. College Station, TX: Stata Press.

Sequential Logistic Regression: estimates a series of logistic regression models simultaneously, taking into account passing through a sequence of transitions. Unlike with multinomial logistic regression, sequential logistic regression does not assume that the chance of falling into any outcome category is equal.

- Buis, Maarten L. 2011. "The Consequences of Unobserved Heterogeneity in a Sequential Logit Model." *Research in Social Stratification and Mobility* 29(3): 247-262. https://www.sciencedirect.com/science/article/pii/S0276562410000703?via%3Dihub
- Buis, Maarten L. 2017. "Not All Transitions Are Equal: The Relationship Between Effects on Passing Steps in a Sequential Process and Effects on the Final Outcome." *Sociological Methods & Research* 46(3): 649-680. https://journals.sagepub.com/doi/abs/10.1177/0049124115591014
- seqlogit command in STATA: http://maartenbuis.nl/software/seqlogit.html

Structural Equation Modeling (SEM): for analysis of unobserved constructs (latent variables) derived from observed variables (e.g., psychological and psychosocial measures). Structural Equation Modeling can be a confirmatory technique for determining and validating a proposed causal process or model. Related methods involving latent variables include Factor Analysis and Path Analysis.

- Schumacker, Randall E. and Richard G. Lomax. 2016. *A Beginner's Guide to Structural Equation Modeling*. New York, NY: Routledge.
- Maruyama, Geoffrey. 1998. *Basics of Structural Equation Modeling*. Thousand Oaks, CA: Sage Publications.
- Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. New York, NY: John Wiley.
- Loehlin, John C. 1998. *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*. 3rd ed. Mahwah, N.J.: Lawrence Erlbaum Associates.
- Duncan, Otis Dudley. 1966. "Path Analysis: Sociological Examples." *American Journal of Sociology* 72(1): 1-16. https://www.journals.uchicago.edu/doi/abs/10.1086/224256
- Alwin, Duane F., and Robert M. Hauser. 1975. "The Decomposition of Effects in Path Analysis." *American Sociological Review* 40(1): 37-47. https://www.jstor.org/stable/2094445?seq=1#metadata_info_tab_contents
- Exploratory Factor Analysis: https://www.mailman.columbia.edu/research/population-health-methods/exploratory-factor-analysis
- Path Analysis: https://www.mailman.columbia.edu/research/population-health-methods/path-analysis

Difference in Difference (DID) Estimation: used to estimate the effect of an intervention when panel or repeated cross-sectional data are available for both treatment and control groups.

- Difference in Difference Estimation: https://www.mailman.columbia.edu/research/population-health-methods/difference-difference-estimation
- Wing, Coady, Kosali Simon, and Ricardo A. Bello-Gomez. 2018. "Designing Difference in Difference Studies: Best Practices for Public Health Policy Research." *Annual Review of Public*

Health 39(1): 453-469. https://www.annualreviews.org/doi/full/10.1146/annurev-publhealth-040617-013507

Other Resources

- Choosing the Correct Statistical Test: https://stats.idre.ucla.edu/other/mult-pkg/whatstat/
- General Modeling Techniques: https://www.mailman.columbia.edu/research/population-health-methods/techniques
- Propensity Score Methods: addresses bias in observational studies; enables comparisons between treatment and non-treatment groups.
    - Propensity Score: https://www.mailman.columbia.edu/research/population-health-methods/propensity-score
    - Austin, Peter C. 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research* 46(3): 399-424. https://www.tandfonline.com/doi/full/10.1080/00273171.2011.568786
    - Austin, Peter C. 2010. "Statistical Criteria for Selecting the Optimal Number of Untreated Subjects Matched to Each Treated Subject When Using Many-to-One Matching on the Propensity Score." *American Journal of Epidemiology* 172(9): 1092-1097. https://academic.oup.com/aje/article/172/9/1092/147493
- Agresti, Alan. 1990. *Categorical Data Analysis*. New York, NY: John Wiley and Sons.
- Powers, Daniel and Yu Xie. 1999. *Statistical Methods for Categorical Data Analysis*. San Diego, CA: Academic Press.
- Hancock, Gregory R., Laura M. Stapleton, and Ralph. O. Mueller. 2019. *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (2nd ed.) New York, NY: Routledge.