## TITLE: Measures of Selection Bias in Regression Coefficients Estimated from Non-Probability Samples

AUTHORS: Brady T. West, Roderick J.A. Little, Rebecca R. Andridge, Philip S. Boonstra, Fernanda Alvarado-Leiton, Anita Pandit, and Erin B. Ware

## ABSTRACT

Selection bias in survey estimates is a major concern, particularly for non-probability samples. Recent developments have provided survey researchers with model-based indices of the potential selection bias in estimates of means and proportions computed from non-probability samples that may be subject to non-ignorable selection mechanisms. To our knowledge, there are currently no systematic approaches for measuring selection bias for regression coefficients, a problem of great practical importance. Generalizing recent developments, we derive novel measures of selection bias for estimated coefficients in linear and probit regression models. The measures arise from normal pattern-mixture models that allow analysts to examine the sensitivity of their inferences to assumptions about the extent of the non-ignorable selection, and they leverage auxiliary variables available for the population of interest that provide information about the variable being modeled when conditioning on the predictors of interest.

After reviewing conceptual and technical details, we describe a simulation study that demonstrates the effectiveness of the measures across a variety of different scenarios. We then apply the measures to data from two real studies. In the first application, we analyze data from the Genes for Good project, which aims to predict health indicators with polygenic scores computed from a large volunteer sample that is recruited via Facebook and provides genetic information by mail. We demonstrate that our measures can effectively detect bias in estimates of regression model coefficients based on these data, using the U.S. Health and Retirement Study as a population benchmark. In the second application, we analyze data from the U.S. National Survey of Family Growth, examining predictors of months worked in the past year among a hypothetical non-probability sample of smartphone users with lower education. Using the full NSFG sample as a population benchmark, we again demonstrate the ability of these measures to indicate potential selection bias.

**Key Words:** Selection Bias, Non-Probability Sampling, Non-Ignorable Selection, Linear Regression, Probit Regression