PENNSTATE

? ? ?

To Weight or Not to Weight, That is the Question:
Survey Weights and Multivariate Analysis

Rebekah Young

Department of Biostatistics

University of Washington

David R. Johnson

Department of Sociology

The Pennsylvania State University

POPULATION RESEARCH INSTITUTE

Survey Sample Weights

- Weights are common in sample surveys
- Used to adjust for
  - Sample design (oversampling of some groups or disproportionate stratification)
  - Nonresponse
  - Post-stratification weights bring sample back to being representative of the population on a select set of demographic characteristics

The Good Side of Weights

- Allow us to claim that our results are representative of the population (at least that they have the same distribution on major demographic characteristics as the population)
- Used to adjust descriptive information on the sample

The Bad Side of Weights

- They introduce a substantial design effect into our data
- Weights will increase the standard errors of our statistics making our findings less precise and more variable
- The larger the variability in the weights the larger the design effects
- This will affect most kinds of statistical analysis
  - Descriptive statistics (means, percentages)
  - Regression coefficients
  - Other multivariate coefficients

PENNSTATE

Using Weights in Regression Models

- General consensus in the literature that weights should be use for descriptive statistics (e.g., Kish & Frankel, 1974)
- Less consensus on whether weights should be routinely used in multivariate models, such as regression (e.g., Gelman, 2007 and comments; Kott, 2007; Winship & Radbill, 1994)

Study Objectives

- Review methodological issues with use of weights in multivariate analysis
- Simulate difference between weighted and unweighted regression models under different conditions
- Develop guidelines for researchers

Model-based methods vs. Use of Weights

- A model-based strategy does not use the weights but includes the variables used to construct the weights as variables in the analysis
  - A correctly specified model-based procedure would use both additive and interaction effects between the substantive variables and the weight-construction variables
  - This will provide unbiased and consistent parameter estimates with smaller standard errors then weighting the data
  - If correctly specified, the model-based approach has the advantage of providing more efficient estimates

Problems with Model-based Approach

- When using archived datasets, all factors that were used to compute the weights are often not available
  - Post-stratification weights often use PSU or sampling error estimates that are not available in the dataset
  - Weights are often altered by trimming and other transformations that distort the relationship to the weighting variables
  - Complete description of the construction of the weights is often not available. Weights might use a interactive rather than additive model (e.g. age x gender)
  - If some variables of substantive importance are also used in the weights, may create problems in testing models (analysis models become more complex)

When should Weights be used?

- Some general recommendations in the literature
  - When the weight is a function of the dependent variable
  - Not enough is known about how the weights were constructed
  - Large samples so loss of statistical power not as much of a concern. (weighted data are generally unbiased but inefficient)

Ways to Test if Weights are Needed

- Sensitivity analysis—compare the coefficients of interest from analysis with and without weights. If not significantly different then weights unnecessary.
- Add the weight and the interaction of the weight with each independent variable to the model. If these do not add significant amount of explained variance to the model, then weights not necessary (A Stata ado – *wgttest* - is available that does this for you).
- When the weights have no effect on the parameter estimates they are called ignorable.

Simulation

- Conducted a simulation of different strategies of using a weight in a regression model
- Used General Social Survey (GSS) data to produce a large dataset (N=75,834)with educational attainment and Race/ethnicity of the respondent as observed in the GSS
- This dataset was used as the target to 1,000 random probability samples of 500 cases each using different realistic sampling strategies

Sampling Strategy for Simulation

- Randomly oversample so 50% of sample was white and 50% non-white
- Educational categories were differentially sampled so that the greater the education the higher the response rate
- Generated a dependent variable under three conditions:
  - Unrelated to education and race
  - Moderately related to education and race
  - Strong relationship to education and race
- Generated an independent variable in three conditions
  - Likert scale, ordered categorical, and Binary
- Generated two models in in which interactions were included between the effects of education and race/ethnicity with the dependent variable

Summary of Findings

- Our simulation showed that a model-based approach performs well when the individual weight variables, but not their interactions, are significant predictors of the dependent variable in the population.
- When the dependent variable is more strongly related to the interactions among the weight variables, however, an unweighted model-based approach leads to biased coefficients in our simulation.
- Our findings are consist with the literature that a model-based approach will always yield smaller standard errors than a weighted approach.
- Including interactions among the weighting variables in a model based approach has no effect on the estimates and is probably unnecesary.

Cautions

- We could not simulate all conditions or situations so these conclusions should be tempered as alternative models may strongly favor one approach over the other
- Because there is still substantial discussion and controversy in the mathematical statistics literature about the use of weights, additional work is needed

Recommendations

- Compare basic substantive model with and without weights and including model-based variables. If same coefficients, unweighted data would be best because of smaller standard errors.
- Test if a model with the weight and the interactions of the weight with all independent variables included adds significant explained variance. If yes, then use the weighted data or use a model-based approach.

References

- Gelman, Andrew. 2007. “Struggles with Survey Weights and Regression Modeling.” *Statistical Science* 22:153-164. (plus comments)
- Kish, Leslie & Martin R. Frankel. 1974. “Inference from Complex Samples.” *Journal of the Royal Statistical Society* 36:1-37.
- Kott, Phillip S. 2007. “Clarifying some Issues in the Regression Analysis of Survey Data.” *Survey Research Methods* 1:11-18.
- Winship, Christopher & Larry Radbill. 1994. “Sampling Weights and Regression Analysis.” *Sociological Methods & Research* 23:230-257.

Thank You!

- If you would like a copy of the paper with more details of the simulation and references to the literature on weighting options, please email either of us.

- David R. Johnson: drj10@psu.edu
- Rebekah Young: rlyoung@uw.edu