Analyzing Data Collected with Complex Survey Design

Hsueh-Sheng Wu

CFDR Workshop Series

Summer 2011

BGSU
Center for Family and Demographic Research

Outline

- Introduction to complex survey design
- Relation between complex survey design and data analysis
- Examples of analysis
  - Analyzing Add Health data
  - Analyzing American Community Survey
- Conclusions

Introduction to Complex Survey Design

- Confusion over complex survey design
  - Survey design discusses both questionnaire design and sample design
  - Complex survey design focuses on complex sample design only

- Different Sampling designs
  - Non-probability sampling: Convenience Sampling, Purposive Sampling, or Snowball sampling
  - Probability sampling: Simple random sampling, Systematic Random Sampling, Stratified sampling, Cluster sampling

Introduction to Complex Survey Design (Continued)

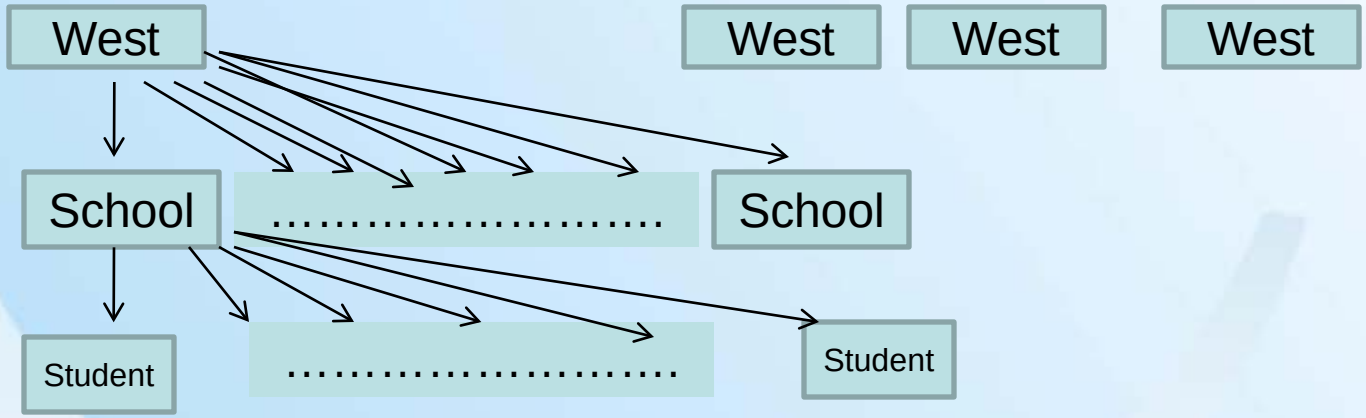
- Complex sample design
  - Multistage sampling: Sample is collected with more than one probability sampling methods
  - Multiphase sampling: Sample is collected by more than on multistage-sampling methods

Example of Multistage Sampling

Add Health data were collected by a three-stage sampling

- Stratified by region
- Select schools within each region
- Select individuals within schools

Diagram



- Sample probability for a student is p1*p2*p3, where p1, p2, and p3 indicates the probability of being selected for a region, a school, and a student
- Respective weights are provided for region, schools, and individuals

Example of Multiphase Sampling

- American Community Survey (ACS) uses two-phase three-stage sampling
  - Main processing:
    - In each August, select housing units from National Master Addressing file (MAF) created by Census Bureau
    - Assign housing units to five existing sub-frames
    - Decide which sub-frame to use
    - Select housing units
    - Select individuals
  - Supplement phase:
    - In each January, select housing units that are not listed in the National Master Addressing file (MAF) in the previous August.
    - Assign housing units to five existing sub-frames
    - Decide which sub-frame to use
    - Select housing units
    - Select individuals

Relation between Complex Survey Design and Data Analysis

- Statistical methods often assume that data were collected from a sample of respondents that are selected by simple random sampling. If data collected by complex survey design are analyzed as if they were collected by simple random sampling, the estimate of parameters and their standard errors are likely to be wrong.

- When software has information on the sample design, some of them can adjust the possible bias from the complex survey design. But different data sets may provide different information about the sample design.
  - Data with multiple stage sampling (e.g., Add Health): The weights on PSU, Strata, and individual are provided
  - Data with multiple stage sampling but do not provide information on PSU and Strata (e.g., ACS and CPS), but provide replicate weights and the weights for the housing units and individual respondents.

Analyzing Add Health Data

- use “R:\add_health.dta”, clear

  This command line asks Stata to read in Add Health data.

- svyset psuscid [pweight = gswgt2], strata(region)

  This -svyset- command specifies the design of the data set. The psucid is the PSU indicator. The -pweight = gswgt2- statement specifies that the personal weight is determined by the variable, gswgt2. The -strata(region)- statement specifies that the strata weight is defined by the variable, region

- svy: mean X1

  The -svy: mean- command calculate the mean of a continuous variable X1, while controlling for the complex design effect

- svy: tab X2 X3

  The -svy: tab- command creates the cross-tabulation between two categorical variables (i.e., X2 and X3), while controlling for the complex design effect

Analyzing Add Health Data

- svy: reg Y X1 X2 X3

  The –svy: reg- command specifies that a regression analysis is to be conducted, and data are collected with complex survey design defined by the previous svyset command. The -Y X1 X2 X3- command specifies that Y is a continuous dependent variable, while X1, X2, and X3 are independent variables.

- svy: logit Y X1 X2 X3

  The –svy:logit- command specifies that a logistic regression analysis is to be conducted, and data are collected with complex survey design defined by the previous svyset command. The -Y X1 X2 X3- command specifies that Y is the dichotomous dependent variable, while X1, X2, and X3 are independent variables.

Analyzing Add Health Data

- svy: ologit Y X1 X2 X3

  The -svy:ologit- command specifies that an ordered logistic regression analysis is to be conducted, and data are collected with complex survey design defined by the previous svyset command. The -Y X1 X2 X3- command specifies that Y is an ordinal dependent variable, while X1, X2, and X3 are independent variables.

- svy: mlogit Y X1 X2 X3

  The -svy:mlogit- command specifies that a multinomial logistic regression analysis is to be conducted, and data are collected with complex survey design defined by the previous svyset command. The -Y X1 X2 X3- command specifies that Y is the nominal dependent variable, while X1, X2, and X3 are independent variables.

Analyzing Add Health Data

Table 1. Additaional Analyses for data collected with complex survey design

Comands	Analysis	Comands	Analysis
svy: biprobit	Bivariate probit regression for survey data	svy: nl	Nonlinear least-squares estimation for survey data
svy: clogit	Conditional (fixed-effects) logistic regression for survey data	svy: oprobit	Ordered probit regression for survey data
svy: cloglog	Complementary log-log regression for survey data	svy: poisson	Poisson regression for survey data
svy: cnreg	Censored-normal regression for survey data	svy: probit	Probit regression for survey data
svy: cnsreg	Constrained linear regression for survey data	svy: proportion	Estimate proportions for survey data
svy: glm	Generalized linear models for survey data	svy: ratio	Estimate ratios for survey data
svy: gnbreg	Generalized negative binomial regression for survey data	svy: scobit	Skewed logistic regression for survey data
svy: heckman	Heckman selection model for survey data	svy: slogit	Stereotype logistic regression for survey data
svy: heckprob	Probit model with sample selection for survey data	svy: stcox	Cox proportional hazards model for survey data

Analyzing Add Health Data

Table 2. Additaional Analyses for data collected with complex survey design

Comands	Analysis	Comands	Analysis
svy: hetprob	Heteroskedastic probit regression for survey data	svy: streg	Parametric survival models for survey data
svy: intreg	Interval regression for survey data	svy: tobit	Tobit regression for survey data
svy: ivprobit	Probit model with endogenous regressors for survey data	svy: total	Estimate totals for survey data
svy: ivregress	Single-equation instrumental-variables regression for survey data	svy: treatreg	Treatment-effects regression for survey data
svy: ivtobit	Tobit model with endogenous regressors for survey data	svy: truncreg	Truncated regression for survey data
svy: logistic	Logistic regression for survey data, reporting odds ratios	svy: zinb	Zero-inflated negative binomial regression for survey data
svy: logit	Logistic regression for survey data, reporting coefficients	svy: zip	Zero-inflated Poisson regression for survey data
svy: mean	Estimate means for survey data	svy: ztnb	Zero-truncated negative binomial regression for survey data
svy: mprobit	Multinomial probit regression for survey data	svy: ztp	Zero-truncated Poisson regression for survey data
svy: nbreg	Negative binomial regression for survey data		

Analyzing ACS Data

- svyset [iw=perwt], jkrweight(repwtp1-repwtp80, multiplier(.05))
  vce(jackknife) mse

  The -svyset- command describes the survey design of the ACS.

  The -[iw=perwt]- command specifies that the sampling weight variable is "perwt".

  The -jkrweight(repwtp1-repwtp80, multiplier(.05))- command instructs Stata that there are 80 replicate weight variables, including repwtp1 through repwtp80 and these variables are used in the Jackknife method to estimate the variance of parameters.

  The -multiplier(.05)- command is decided by the formula provided by Census Bureau.

  The -vce(jackknife) mse - command specifies that a Jackknife method is used to calculate variance and mean square error.

Analyzing ACS Data

- svy jackknife: mean X1

  The -svy:mean- command calculates the mean of a continuous variable X1, while using the jackknife method and replicate weights to calculate the variance and mean square errors.

- svy jackknife:tab X1 X2

  The -svy:tab- command creates the cross-tabulation between X1 and X2 while using the jackknife method and replicate weights to calculate the variance and mean square errors.

Analyzing ACS Data

- svy jackknife: reg Y X1 X2 X3

  The –svy:reg- command specifies that a regression analysis is to be conducted, and data are collected with complex survey design defined by the previous svyset command. The -Y X1 X2 X3- command specifies that Y is a continuous dependent variable, while X1, X2, and X3 are independent variables.

- svy jackknife: logit Y X1 X2 X3

  The –svy:logit- command specifies that a logistic regression analysis is to be conducted, and data are collected with complex survey design defined by the previous svyset command. The -Y X1 X2 X3- command specifies that Y is the dichotomous dependent variable, while X1, X2, and X3 are independent variables.

Analyzing ACS Data

- svy jackknife: ologit Y X1 X2 X3

  The –svy:ologit- command specifies that an ordered logistic regression analysis is to be conducted, and data are collected with complex survey design defined by the previous svyset command. The -Y X1 X2 X3- command specifies that Y is a ordinal dependent variable, while X1, X2, and X3 are the independent variables.

- svy jackknife: mlogit Y X1 X2 X3

  The –svy:mlogit- command specifies that a multinomial logistic regression analysis is to be conducted, and data are collected with complex survey design defined by the previous svyset command. The -Y X1 X2 X3- command specifies that Y is the nominal dependent variable, while X1, X2, and X3 are the independent variables.

- Additional analyses are available, if you follow those listed on pages 11 and 12.

Conclusions

- Many survey data are not collected with simple random sampling and thus each individual does not have equal weight, which influences the estimation of population parameters and their standard errors.

- Data using complex survey design can be divided into two types: one provides the information on PSU, Strata, and sampling weights; while the other provides replicate and sampling weights. These two types of data requires different setups in the –svyset- commands in Stata.

- After the -svyset- command is setup, you can analyze data using the –svy- command with the regular analysis command

- For further question, feel free to contact me at wuh@bgsu.edu or stop by my office (5D Williams Hall).