UK

Applied Statistics Lab

qipsr quantitative initiative for policy and social research

Applied Survey Data Analysis

Module 3: Using svyset commands in Stata

March 30, 2013

Outline

- Preparing for complex survey data analysis.
- Declaring the complex sample design features of your survey to Stata using the **svyset** command.
- Creating summary statistics such as frequencies, means, and cross-tabulations incorporating complex survey design (**svy**: commands).
- Conducting sub-population analysis correctly.
- Fitting OLS or logistic models to complex sample survey data.

Opening the Data Sets

- Two data sets are provided for the lab:

  -ANES Times Series 2008.dta

  -GSS 2010 Merged_r2b.dta

- The data sets are located in the virtual drive: s:\workshop

- Copy and paste the data sets in the desktop or your USB drive.

```
cd “C:\Users\pdze222\Desktop”
```

Do File

- Copy and paste the Do file: “Do file Module 3 ANES survey” in the desktop or your USB drive
- Open Stata: Click on Start- All programs-Stata
- Open the do-file
- Change the current working directory to the specified drive and directory.

Preparing for Complex Sample Survey Data Analysis

- Survey data are characterized by the following:
  - Sampling weights- probability weights—pweights
  - Clustering
  - Stratification
- Read the technical information of the survey to:
  - Identify the correct weight variables for the analysis
  - Identify the stratum and cluster codes in complex survey data sets.

ANES Sample Design

- *Oversampling*: Members of minority groups, such as blacks and Hispanics in the 2008 Time Series, have been sampled at a higher rate than their proportion in the population. [weight=V080102]

- *Stratified cluster sampling*
  - **9 Strata**: Eight Largest Metropolitan Satistical Areas (MSAs) (*New York, Los Angeles, Chicago, Philadelphia, Dallas, Miami, Houston, and Washington)* and remaining counties [strata=V081206]
  - **Clusters**
    - Counties (PSUs) [psu=V081205]
    - Census Tracts within selected counties
    - Census Block groups (CBG) within selected Census Tracts
    - Residential mailing address within select CBG
    - Respondents within screened and eligible household

GSS Sample Design

- The General Social Survey (GSS) is an area-probability sample that uses the NORC National Sampling Frame for an equal-probability multi-stage cluster sample of housing units for the entire United States
  - Clusters:
    - National Frame Areas, (NFAs), each of which is composed of one or more counties
    - Segments, each of which is either a block, a group of blocks, or an entire census tract
    - Housing units

- Poststratification: Two PSUs per strata
  - In certainty NFAs, segments are paired into strata with one segment assigned to VPSU = 1 while the other segment is assigned to VPSU = 2. Often, small segments are combined into one VPSU.
  - Non-certainty NFAs are paired into strata with one NFA assigned to VPSU = 1 while the other NFA is assigned to VPSU = 2.

PSU= vpsu strata=vstrat weight=wtcombnr

Applied Statistics Lab

quantitative initiative for policy and social research

Declaring your Sample Design using svyset

- The command svyset (declare data as survey data) is used to identify the sample design features of your data to Stata.

Single-stage design syntax:

```
svyset [psu] [weight] [, design_options options]
```

Survey setting for ANES

```
svyset V081205 [pweight=V080102], strata(V081206) vce(linearized)
```

Survey setting for GSS

```
svyset vpsu [weight=wtcombnr], strata (vstrat) vce(linearized)
```

Survey Describe

- Use svydescribe to describe the first stage of the survey dataset

```
svydescribe
```

- Specify the variable to get svydescribe to report on where it contains missing values and how this affects the estimation sample.

```
svydescribe voteobama
```

Singleton Stratum

- Once you have identified cases in singleton stratum, there are a number of ways of dealing with singleton Strata:
  - Singleton strata can be treated as missing and error, deleting them from your sample.
  - Singleton strata can be grouped with other singleton strata to treat them like they belong to the other strata.
  - Singleton strata can be specified as ‘certainty units’ that are centered and/or scaled using the **singleunit (method)** option in the svyset command. Type help svyset.

svyset [psu] [weight] , strata (stratacode) singleunit(certainty)

Applied Statistics Lab

quantitative initiative for policy and social research

Summary Statistics

- First, estimate the mean of feeling thermometer variable for Obama (obamaft) assuming simple random sampling

mean *obamaft*

- Now see what happens when we take into account the design of the survey.

svy: mean *obamaft*

- Use *estat* effects to report the design effects DEFF and DEFT for the mean estimates

estat effects

Comparing Means

- To estimate the mean of *obamaft* for each subpopulation identified by the categories of the female variable (0=male and 1=female).

```
svy: mean obamaft, over(female)
```

- Perform differences of means analysis (similar to ANOVA)

```
lincom [obamaft]0-[obamaft]1
```

Frequencies and Cross-Tabulation

These are produced using the ‘svy: tab’ command (tabulate). Type:

```
svy: tab voteobama
```

For two-way tables (cross-tabs):

```
svy: tab voteobama female
```

Standard table options such as row, standard errors and confidence intervals can be specified:

```
svy: tab voteobama female , row  se ci
```

Sub-Population Analysis

- In many analyses, you may wish to focus on a sub-population, such as men or women, or a specific age group. A standard approach to this would be to use the ‘if’ command (e.g. tab voteobama if sex ==1), or to drop unwanted cases.

- However, svyset commands require information on the entire population size to calculate standard errors. Such approaches should therefore be avoided, and instead, the ‘subpop’ command should be used (although in practice it often does not make much difference).

Examples of Sub-Population Analysis

We first need a binary variable coded as 1=subpopulation of interest, 0 = otherwise (missing if we don’t know).

In the data, we will generate a recode of our variable ‘sex’ (currently 1=men 2==women, recoding women to 0). Type:

```
gen male = sex == 1 if !missing(sex)
```

Next, apply the subpop command:

```
svy, subpop (male): tab voteobama
svy, subpop (male): regress obamaft  black
```

Fitting a Logistic Model to Complex Sample Survey Data

- **Dependent Variable: Voting for Obama in the 2008 election**

- Coded 1 if the respondent voted for Obama and 0 if the respondent voted for another candidate.
- Nonvoters excluded

Independent Variables

- Party ID *(1 strong Democrat to 7 strong Republican)*
- Feeling thermometers for Obama and McCain *(Scale=1-100)*
- Sex and educational attainment
- Respondent is black
- Belief about the Bible being the word of God
- Belief whether the Iraq war was worth the cost *(worth/not worth it)*
- Belief whether homosexuals should be allowed to serve in the armed forces *(5 strongly should not be allowed to 1 strongly should be allowed)*
- Interviewer's assessment of whether the respondent seemed well informed. *(1 very low to 5 very high)*

Logistic Regression Analyses with the 2008 ANES Time Series Data

	Model 1			Model 2		
	SRS SEs			Design-based SEs		
	Coef.	s.e.	p	Coef.	s.e.	p
Party ID	-3.4***	0.47	0.000	-3.7***	0.39	0.000
Obama FT	7.8***	0.72	0.000	8.2***	1.31	0.000
MC Cain FT	-5.4***	0.73	0.000	-5.4***	0.77	0.000
Bible word of God	-1.1**	0.40	0.006	-1.2*	0.55	0.032
Gays in Military	0.6	0.40	0.127	0.9*	0.37	0.021
Iraq war worth costs	-0.8*	0.31	0.012	-0.7	0.36	0.056
Appeared informed	1.1*	0.52	0.036	1.0	0.66	0.119
Education	-3.7***	1.05	0.000	-4.2***	1.19	0.001
Female	0.2	0.27	0.382	0.0	0.25	0.955
Black	2.5***	0.63	0.000	2.6***	0.68	0.000
Constant	2.7**	1.05	0.010	2.9**	1.01	0.005
* p<0.05, ** p<0.01, *** p<0.001						

UK
Applied Statistics Lab

qipsr quantitative initiative for policy and social research

Predicted Probabilities using margins and marginsplot

- “Margins,” allows calculating confidence intervals associated with predicted probabilities after the command svy: logit.

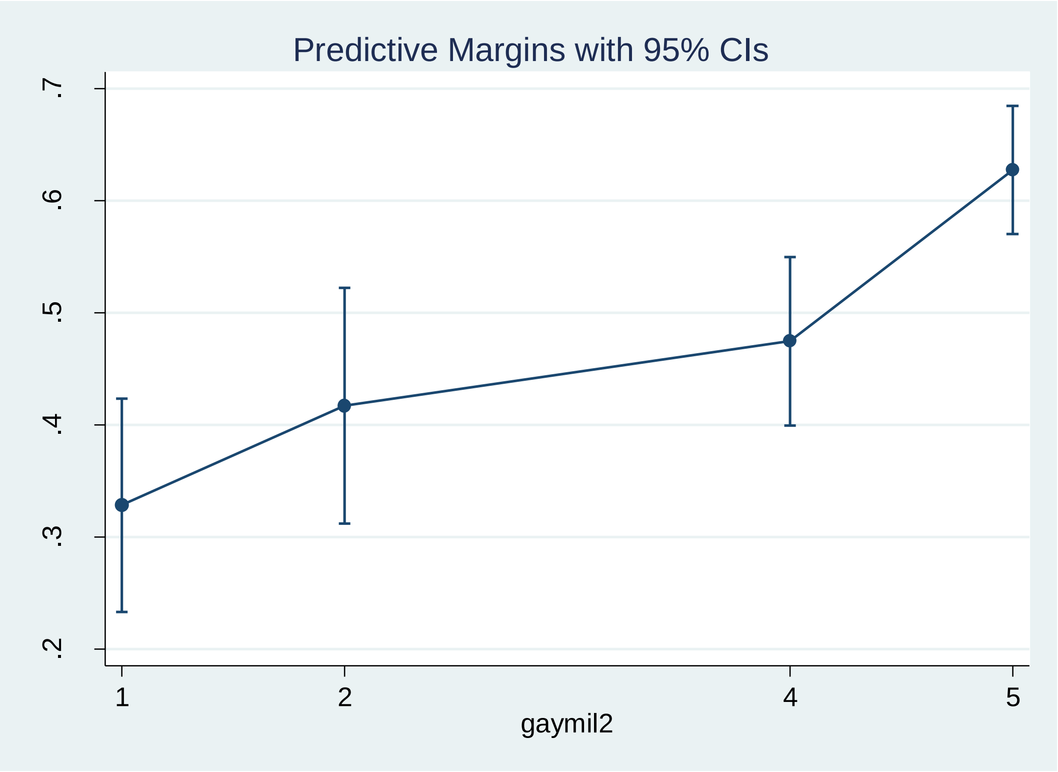
- Specify the linearization method vce (unconditional) instead of the default delta method vce (delta) to account for complex surveys

- **Examples :**

```
margins, over(gaymil2) subpop(subp) vce (unconditional)
marginsplot
marginsplot, recast(line) recastci(rarea)
```

Create Chart using marginsplot

Stata syntax: marginsplot



UK

Applied Statistics Lab

qipsr

quantitative initiative for policy and social research