It's a Little Different with Survey Data

Christine Wells
Statistical Consulting Group
Academic Technology Services
University of California, Los Angeles

Thursday, November 13, 2008

Table of contents

Introduction

- How much of what you learned about analyzing experimental and/or observational data can be applied to analyzing survey data?
- Example data: NHANES III Adult data set
- NHANES III (National Health And Nutrition Examination Survey) collected from 1988-1994
- Stratified multi-stage cluster sample of approximately 34,000 people
- The data set includes sampling weights, as well as both (pseudo) PSU/strata variables and replicate weights
- Sampling weight affects point estimates, PSU/strata or replicate weights affect standard errors

The **svyset** command with PSU and strata variables

```
. svyset sdppsu6 [pweight = wtpfqx6], strata(sdpstra6) singleunit(centered)

      pweight: wtpfqx6
          VCE: linearized
  Single unit: centered
     Strata 1: sdpstra6
         SU 1: sdppsu6
        FPC 1: <zero>
```

Definition of stratified

- With non-survey data: Part of the analysis
- With survey data: Part of the sampling

The **svyset** command with replicate weights

```
. display 1-(1/sqrt(1.7))
.23303501
. svyset [pweight = wtpfqx6], brr(wtpqrp1 - wtpqrp52) fay(.23303501) ///
  vce(brr) mse singleunit(centered)
      pweight: wtpfqx6
          VCE: brr
          MSE: on
    brrweight: wtpqrp1 wtpqrp2 wtpqrp3 wtpqrp4 wtpqrp5 wtpqrp6 wtpqrp7
               wtpqrp8 wtpqrp9 wtpqrp10 wtpqrp11 wtpqrp12 wtpqrp13 wtpqrp14
               wtpqrp15 wtpqrp16 wtpqrp17 wtpqrp18 wtpqrp19 wtpqrp20 wtpqrp21
               wtpqrp22 wtpqrp23 wtpqrp24 wtpqrp25 wtpqrp26 wtpqrp27 wtpqrp28
               wtpqrp29 wtpqrp30 wtpqrp31 wtpqrp32 wtpqrp33 wtpqrp34 wtpqrp35
               wtpqrp36 wtpqrp37 wtpqrp38 wtpqrp39 wtpqrp40 wtpqrp41 wtpqrp42
               wtpqrp43 wtpqrp44 wtpqrp45 wtpqrp46 wtpqrp47 wtpqrp48 wtpqrp49
               wtpqrp50 wtpqrp51 wtpqrp52
          fay: .23303501
  Single unit: centered
     Strata 1: <one>
         SU 1: <observations>
```

Subsetting, AKA subpop'ing

- Under certain circumstances, deleting cases can mess up the calculation of the standard errors
- `subpop` option (dummy variable: 1 = in subpop; 0 = not in subpop)
- `over` option (can use a polychotomous variable)
- `over` not avialable for all commands (bummer!)
- Can combine them (way cool!)

Examples with subpopulations

```
. svy, subpop(female):  mean edu
. svy: mean edu, over(race)
. svy, subpop(female if mar_status==1):  mean edu, over(race)
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      49       Number of obs     =      20007
Number of PSUs   =      98       Population size   =  187354934
                                 Subpop. no. obs   =       4740
                                 Subpop. size      = 51858956.6
                                 Design df         =         49

        white: race = white
        black: race = black
        other: race = other

--------------------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
edu          |
       white |   12.68359   .1128168      12.45687     12.9103
       black |   11.68496   .1429806      11.39763    11.97229
       other |   12.09518   .7607426      10.56641    13.62395
--------------------------------------------------------------
```

Descriptives with categorical variables

- Categorical variables: frequencies, crosstabulations, chi-square
- Use sampling weights for all descriptives and analyses, but look at non-weighted counts
- Frequencies and two-way crosstabulations are easy to do
- Chi-square easy to do but may not mean much

Example with categorical variables

```
. svy: tab pet, count cellwidth(15) format(%15.2f) obs
(running tabulate on estimation sample)

Number of strata    =        49                Number of obs     =      19975
Number of PSUs      =        98                Population size   =  187026362
                                               Design df         =         49

-------------------------------------------
      pet |           count             obs
----------+--------------------------------
       no |    108455905.98        13480.00
      yes |     78570456.49         6495.00
          |
    Total |    187026362.47        19975.00
-------------------------------------------
  Key:  count           =  weighted counts
        obs             =  number of observations
```

Example with categorical variables

```
. svy: tab military pet
(running tabulate on estimation sample)

Number of strata   =        49                  Number of obs     =     19878
Number of PSUs     =        98                  Population size   = 186223378
                                                Design df         =        49

-------------------------------
          |        pet
 military |    no    yes  Total
----------+--------------------
       no | .4946  .3492  .8438
      yes | .0844  .0719  .1562
          |
    Total | .5789  .4211      1
-------------------------------
  Key:  cell proportions

  Pearson:
    Uncorrected   chi2(1)         =   22.9203
    Design-based  F(1, 49)        =    8.6228     P = 0.0050
```

Descriptives with continuous variables

- Continuous variables: means, standard deviations and correlations
- Means are easy to do
- Standard deviations - can do in Stata 10, but not commonly available
- Correlations - user-written .ado (`corr_svy`) and available in some packages
- Medians and percentiles - available in some packages
- Graphing is difficult because of the sampling weights and the large sample size

Descriptives with continuous variables

```
. svy: mean edu
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      49        Number of obs     =      19772
Number of PSUs   =      98        Population size   =  185855207
                                  Design df         =         49

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
         edu |   12.32709   .0903272      12.14557    12.50861
--------------------------------------------------------------
```

Descriptives with continuous variables

```
. estat sd

-------------------------------------
             |       Mean   Std. Dev.
-------------+-----------------------
         edu |   12.32709    3.372697
-------------------------------------

. estat sd, var

-------------------------------------
             |       Mean    Variance
-------------+-----------------------
         edu |   12.32709    11.37509
-------------------------------------
```

Linear regression - Similarities

- Linear regression is easy to do
- Categorical predictor variables (`xi` prefix and `test` command)
- Interpretation of regression coefficients
- R-square
- Can compare nested models with Wald test (`test` command)
- Can output residuals and predicted values

Linear regression - Example

```
. xi: svy: reg times_friends times_neighbors military i.race
i.race            _Irace_1-3          (naturally coded; _Irace_1 omitted)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        49                 Number of obs      =     19843
Number of PSUs     =        98                 Population size    = 186021259
                                               Design df          =        49
                                               F(   4,     46)    =     38.08
                                               Prob > F           =    0.0000
                                               R-squared          =    0.1457

------------------------------------------------------------------------------
             |             Linearized
times_frie~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
times_neig~s |   .3869599   .0769275     5.03   0.000     .2323682    .5415516
    military |  -32.71182   3.736655    -8.75   0.000    -40.22091   -25.20273
    _Irace_2 |   10.68002   4.046985     2.64   0.011     2.547298    18.81274
    _Irace_3 |  -13.13257   11.39838    -1.15   0.255    -36.03848    9.773329
       _cons |   105.9647   5.820172    18.21   0.000     94.26861    117.6608
------------------------------------------------------------------------------
```

Linear regression - Example

```
. test _Irace_2 _Irace_3

Adjusted Wald test

 ( 1)  _Irace_2 = 0
 ( 2)  _Irace_3 = 0

       F(  2,    48) =    3.60
            Prob > F =    0.0351
```

Linear regression - Differences

- Not OLS, but weighted least squares
- No standardized regression coefficients
- No adjusted R-square
- No easy way to compare non-nested models
- Most diagnostics not yet available

Linear regression - Diagnostics

- Need to account for the sampling plan
- Most regression diagnostic commands do not work with survey data
- Very cool research currenly being done
  - Outliers, leverage and influence

Logistic regression - Similarities

- Logistic regression is easy to do
- Categorical predictor variables (`xi` prefix and `test` command)
- Interpretation of logistic regression coefficients
- Can get odds ratios
- Can compare nested models with Wald test (`test` command)
- Can output predicted values

Logistic regression - Example

```
. xi: svy: logit clubs female military i.race
i.race            _Irace_1-3          (naturally coded; _Irace_1 omitted)
(running logit on estimation sample)

Survey: Logistic regression

Number of strata    =        49                Number of obs      =      19861
Number of PSUs      =        98                Population size    =  186138472
                                               Design df          =         49
                                               F(   4,      46)   =      32.43
                                               Prob > F           =     0.0000

------------------------------------------------------------------------------
             |             Linearized
       clubs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .0197592   .0413089     0.48   0.635     -.063254    .1027725
    military |   .5781711   .0716061     8.07   0.000     .4342732    .7220689
    _Irace_2 |  -.5605115   .0619599    -9.05   0.000    -.6850245   -.4359985
    _Irace_3 |  -.7896878   .1673143    -4.72   0.000    -1.125918   -.4534571
       _cons |  -.5533257   .0426134   -12.98   0.000    -.6389606   -.4676908
------------------------------------------------------------------------------
```

Logistic regression - Example

```
. test  _Irace_2 _Irace_3

Adjusted Wald test

 ( 1)  _Irace_2 = 0
 ( 2)  _Irace_3 = 0

       F(  2,    48) =   43.67
            Prob > F =    0.0000
```

Logistic regression - Differences

- Assessing model fit even more difficult
  - Not even a pseudo-R-square
- Testing non-nested models is difficult
  - No log-likelihood, AIC or BIC
- Diagnostics generally not available

Conclusions

- Much of what you know about data analysis applies to the analysis of survey data
- Don't expect to be able to do everything that you can do with non-survey data
  - Diagnostics for various types of regression models are still being developed
  - Model fit and model comparison are often difficult
  - Caution necessary when considering multiple imputation
  - Some techniques frequently requested by our clients
    - Multivariate (e.g., MANOVA, EFA)
    - Effect sizes
    - Graphical techniques

References

Berk, Richard A. (2004). Regression Analysis: A Constructive Critique. Thousand Oaks, CA: Sage Publications.

Berry, William D. Understanding Regression Assumptions. (1993). Thousand Oaks, CA: Sage Publications.

Brewer, Ken. (2002). Combined Survey Sampling Inference: Weighing Basus Elephants. London: Arnold Publishers.

Chambers, R. L. and Skinner, C. J. (Editors). (2003). Analysis of Survey Data. New York: John Wiley and Sons.

Cochran, William G. (1977). Sampling Techniques, Third Edition. New York: John Wiley and Sons.

References continued

Kalton, Graham, and Heeringa, Steven (Editors). (2003). Leslie Kish: Selected Papers. New York: John Wiley and Sons.

Kish, Leslie. (1995). Survey Sampling. New York: Wiley Classics Library.

Korn, Edward L. and Graubard, Barry I. (1999). Analysis of Health Surveys. New York: John Wiley and Sons.

Korn, Edward L. and Graubard, Barry I. (1998). Statistical Computing and Graphics: Scatterplots with Survey Data, The American Statistician. 52(1): 58-69.

References continued

Li, Jianzhu (2006). Influence Analysis in Linear Regression with Sampling Weights. Paper presented at The Joint Statistical Meetings, Section on Survey Research Methods, Seattle, WA. http://www.amstat.org/sections/srms/proceedings/y2006/Files/JSM2006-000435.pdf

Li, Jianzhu (2008). Linear Regression Diagnostics in Cluster Samples. Paper presented at The Joint Statistical Meetings, Section on Survey Methods, Denver, CO. http://www.amstat.org/sections/srms/Proceedings/y2007f.html

Lohr, Sharon L. (1999). Sampling: Design and Analysis. New York: Duxbury Press.