Soc 73994, Homework #4
Complex Survey Designs; Multiple Imputation

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised January 25, 2021

All answers should be typed and mailed to the TA. Be sure your response includes your name, the date, and a clear title, e.g. Homework # 4. If there is a huge amount of output for any analyses you run yourself, you may want to be selective in what you copy and paste into your assignment (but make sure you include enough so it is clear what commands you executed, e.g. you might show all the commands but only parts of the output).

This assignment has two parts. First, you will do some very basic analyses with svyset data. Second, you will be asked to use multiple imputation. With both I will ask you to work through an example and then do similar analyses with a data set of your choosing. Ideally you will use your own data but if your own data aren't appropriate you can use something else. If you aren't sure what a command does look it up in the manuals or handouts. (Thanks to Chris Quiroz for helping to prepare this assignment.)

1. Survey data analysis

a. Run the following commands that use the American National Election Study, 2008 (ANES 2008). You can add other commands besides the ones shown here if you think they would be helpful, e.g. you could do a twoway tabulation or use spost13 commands like `mtable`.
   - Briefly explain the results from the descriptive statistics (the `mean` and `proportion` commands). Indicate whether it appears that some groups were oversampled (you might be able to do this by comparing weighted and unweighted proportions). Also indicate how results might be affected/ distorted if you did not weights.
   - Discuss the results from the two logistic regressions. Explain how the results differ (if they do) and why.
   - Discuss and explain the results produced by the `margins` commands. Indicate whether or not weighting affected the results.

```
version 13.1
use https://www3.nd.edu/~rwilliam/statafiles/anes_codeddata, clear
* Describe the vars that will be used
des pres2008 female race age
*Svy set the data
svyset CASEID [pweight = V080102A]
*Descriptive stats- First do the continous var
mean age
estat sd
svy: mean age
estat sd
* Now do the categorical variables.
```

```
proportion race female
svy: proportion pres2008 race female
*Logistic Regression & Margins With and Without Svy;
logit pres2008 i.female i.race age, nolog
margins female
svy: logit pres2008 i.female i.race age
margins female
```

b. Now choose a data set with a complex survey design that is or can be svyset. If you don't have anything yourself, you can use the nhanes2f data, e.g. `webuse nhanes2f, clear`. Or, you could use the ANES data but with different variables. Briefly describe the variables; it is ok to copy and paste from previous homeworks. Run analyses similar to those above, e.g.
   - Run a few descriptive statistics. You can use other commands besides the ones shown here if you think they would be helpful, e.g. `svy: tabulate`. Briefly describe the results.
   - Run an estimation command (e.g. logit) both with and without the svy: prefix. Explain how the results differ and why.
   - Compute adjusted predictions (or whatever else you want) with the `margins` and/or spost13 commands, e.g. `mtable`. Interpret the results.

2. Multiple imputation

a. This part of the problem is adapted from Paul Allison's 2009 book *Fixed Effects Regression Models*. Data are from the National Longitudinal Study of Youth (NLSY). This subset of the data set has 1151 teenage girls who were interviewed annually for 5 years beginning in 1979. Only the fifth and final wave is used here. I have modified the data set so that some values are missing.

- pov is coded 1 if the subject was in poverty during that time period, 0 otherwise.
- age is the age at last interview.
- mother is coded 1 if the respondent currently has at least 1 child, 0 otherwise.
- spouse is coded 1 if the respondent is currently living with a spouse, 0 otherwise.
- hours is the hours worked during the week of the survey.

Run the following code. Briefly explain the purpose of each command and what, if anything, the results from it showed you. Among other things, you should explain the imputation method that was used, and why. Also note what effect using imputation had on the results. In particular, how many more cases got utilized? How much did the MI results differ from the results you would have gotten without imputation? Do you think imputation was worth the trouble?

```
* Basic MI problem
version 13.1
use "https://www3.nd.edu/~rwilliam/statafiles/mdpov2.dta", clear
sum pov age mother spouse hours
mi set mlong
mi misstable summarize
mi misstable patterns
mi register imputed age mother
mi impute chained (regress) age (logit) mother = pov spouse hours, add(20) rseed(2232)
```

```
mi xeq 0: logit pov age i.mother i.spouse hours, nolog
mi estimate, dots: logit pov age i.mother i.spouse hours
mimrgns spouse, predict(pr)
```

b. Now do something similar with a data set of your own choice. Briefly describe the variables; it is ok to copy and paste from previous homeworks. If by some miracle the data set for your paper has little or no missing data then use something else, e.g. the ANES data (you can try to use both svy: and mi: but that can get very complicated so for now you may wish to ignore any svysetting in your data). You of course need to impute values for at least one independent variable but you don't have to make it much more complicated than that unless you want to.

As with part A, briefly explain the purpose of each command and what, if anything, the results from it showed you. Among other things, you should explain the imputation method that was used, and why. Also note what effect using imputation had on the results. In particular, how many more cases got utilized? How much did the MI results differ from the results you would have gotten without imputation? Do you think imputation was worth the trouble?