Week 5. Weighting

Why applies weighting?

1. Weighting factors are used in sampling to make samples match the population. Weighting is necessary when each observation has a different probability of selection in a sample.
2. If all observations have the same probability of selection as assumed by the pure random sampling, no weight is necessary (except for population projections). For population projections, a simple multiplication is enough (e.g., if pop size = 10,000 and sample size = 200, multiply the sample by 500(=10,000/200) will yield a population projection).

Four weighting methods in Stata

1. `pweight`: Sampling weight.
   (a) This should be applied for all multi-variable analyses.
   (b) Effect: Each observation is treated as a randomly selected sample from the group which has the size of weight.
2. `aweight`: Analytic weight.
   (a) This is for descriptive statistics.
   (b) If `pweight` option is not available, use `aweight` in multi-variable analyses.
   (c) Effect: Each observation is treated as the mean of a group which has the size of weight.
3. `fweight`: Frequency weight (= weight in SPSS).
   (a) Use this weight when population projection is needed. That is, if you need to compute actual population sizes, use `fweight`.
   (b) Effect: By weighting, each observation is duplicated by the size of weight. It has an effect of raising the sample size.
4. `iweight`: Importance weight.
   (a) Don't use it in demographic and sociological studies unless there are strong theoretical or methodological reasons to believe certain cases are more important than the others.

Which weighting option should be used?

1. `pweight` vs. `aweight`
   (a) As noted above, only standard errors change when you apply pweight or aweight. The estimated coefficients are the same.
   (b) The following two regression will yield the exactly same results:
   - reg y x1 x2 [pw=wt]

- reg y x1 x2 [aw=wt], vce(robust)

That is, the estimated coefficients and standard errors between two models are identical.

(c) That is, pweight option is the same with the robust standard error option of aweight.

(d) In random sampling, each observation has the same sampling probability. In this case, each observation contributes to σ^2 by the amount of σ^2/N. When each individual has a varying probability of sample selection, each observation represents different number of population, and thus the amount of contribution to σ^2 varies across observations. Pweight in Stata adjusts the differentiated contribution by each observation by applying so-called Sandwitch White Estimator (don't ask), which is the same as robust standard errors.

(e) Thus, pweight option is not necessary for descriptive statistics because we don't compute st. error for descriptive statistics.

(f) Which weight should we use: `pweight` or `aweight`?

i. When unit of obs is individual and each variable indicates the characteristics of individuals (this is almost always the case), use `pweight`.

ii. But, when the main treatment(=explanatory) variable represents the characteristic of a group, you can use `aweight` even if unit of obs is individual. This is because there is no variation across individuals within that group. In this case, what the dependent variable actually indicates is the group mean rather than an individual characteristic.

iii. Since there is no variance across individuals within a group, the standard errors with aweight are smaller than the standard errors with pweight. As a result, the p-values are smaller with aweight than those with pweight. Your estimated coefficients are more likely to be statistically significant with aweight than with pweight. Because of this property, it is tempting to use aweight when the estimated coefficients are insignificantly by a slight margin when pweight is applied. But this is a statistical cheating.

iv. Simply put, use `pweight` for all multi-variable analyses unless you have strong reasons to believe that aweight is better.

2. `fweight`

(a) Use `fweight` only for population projections.

(b) For example, if you'd like to estimate the actual population size of Korean immigrants rather than the % of Korean immigrants in American population using ACS (American Community Survey), apply `fweight`. Stata will report actual population size.

(c) The effect of fweight is the same as increasing sample size by the factor of fweight. Thus, N becomes "original N * mean of weight."

(d) In Stata, all fweight factors should be integer numbers. No fraction is allowed.

(e) Because N is artificially increased with fweight, SE ($= \sigma/\sqrt{N}$) will decrease artificially.

(f) If you multiply the standard error of fweight by $\sqrt{\frac{n-k}{newN-k}}$, then the standard error becomes the same as with the standard error of aweight.

(g) SPSS weight was equal to the fweight in Stata (not sure whether this is still ture).

Bottom Line

1. Weighting must be applied. Otherwise, your results are biased in most cases. An exception is if all variables that are associated with weighting factors are controlled for, then the estimated coefficients are identical regardless of the application of weighting. Even in this case, however, the standard errors are not correct if weighting is not properly applied.
2. In Stata, the estimated coefficients are identical regardless of which weighting options are applied. All weighting methods will yield the same coefficients estimated.
3. But standard errors vary by weighting options.
4. Usually, SE errors with `pweight` > SE with `aweight` > SE with `fweight`
5. Use `pweight` for all multi-variable analyses as long as `pweight` is available.
6. Use `aweight` for descriptive statistics.
7. Use `aweight` for population projection.
8. Don't use `iweight`.