1 Research Methods 2:

1.1 Econometric Methods and Surveys

When analysing survey data using statistic/econometrics can distinguish two approaches, descriptive and modelling, and the distinction can be important.

Descriptive: where concerned to describe the data being analysed

Modelling: where concerned with developing and testing models and inference

This is particularly the case when dealing with the problem of surveys that have different probabilities of selection

Question is whether need to take account of weights in regression?

and the answer is:

It depends on what you are doing

1.1.1 Survey designs and regression

consider N_s a population of households and n_s sample of households

$$w_{is} = \left(\frac{N_s}{n_s}\right)$$

and so

$$\overline{x}_w = \frac{\sum_{s=1}^{S}\sum_{i=1}^{n_s} w_{is}x_{is}}{\sum_{s=1}^{S}\sum_{i=1}^{n_s} w_{is}} = \frac{\sum_{s=1}^{S} N_s\overline{x}_s}{\sum_{s=1}^{S} N_s} = \sum_{s=1}^{S}\frac{N_s}{N}\overline{x}_s = \overline{x}$$

Within each sector

$$y_s = \alpha_s + \beta_s x_s + u_s$$

Population weighted average is

$$\beta = \frac{1}{N}\sum_{s=1}^{S} N_s\beta_s$$

So if estimate the regression for each sector the population weighted estimate is

$$\widehat{\beta} = \sum_{s=1}^{S} \frac{N_s}{N} \widehat{\beta}_s$$

such regressions are regularly used when sectors are broad, but when there are few households in sector the parameters will be estimated imprecisely.

Even then worth considering as the variance is

$$var\left(\widehat{\beta}\right) = \sum_{s=1}^{S} \left(\frac{N_s}{N}\right)^2 var\left(\widehat{\beta}_s\right) = \sum_{s=1}^{S} \left(\frac{N_s}{N}\right)^2 \frac{\sigma_s^2}{\sum_{i=1}^{n_s} x_i^2}$$

where σ_s^2 is residual variance in stratum.

because the population fractions are squared (and is a fraction) β will be more precisely estimated than the individual β_ss

$$var\left(\widehat{\beta}\right) < var\left(\widehat{\beta}_s\right)$$

Common for researchers to estimte on all observations at once, either using the inflation factors to calciulate a weighted least squares or ignoring them.

In general OLS estimates will not yield any parameters of interest

- if all the β_s are the same then the OLS estimate will be consistent for the common β
- Even if the structure of the explanatory variables in each stratum is the same sample weighted average will be inconsistent unless the sample is a simple random sample with equal probabilities of acceptance
- It is a problem of population heterogeneity rather than sample design
- This mirrors inconsistency of unweighted mean for population mean

So:

- If population is homogeneous, meaning the coefficients are identical for each stratum both weighted and unweighted are consistent and unweighted is prefered as more efficient (Gauss Markov)
- If population not homogeneous both estimates are inconsistent
- So in neither case is there an argument for weighting

Weighted estimates justified:

- If have many strata and suspect heterogeneity but it is not systematically linked to other varaibles. Weighted estimates consistent if variation in parameters is random and independent of xs and if number of strata is large enough for weights mean to be zero.
- If consider regression as descriptive rather than structural. Are summarising characteristics of the population by the regression, heterogeneity and all

But if trying to estimate behavioral relation and if this is different in different parts of the population -weighting is at best useless.

1.1.2 Recommendations for practice

- If regressions primarily descriptive -exploring association by looking at mean of one variable conditional on others, then use weights and correct the standard errors for the design
- If modelling and concerned with heterogeneity and its interaction with sample design the more complex standard errors can get explicit formulas or use bootstrap, but program bootstrap to reflect sample design
- In practice it is clustering that has the largest effect on standard errors, conventional formulae overstate precision by ignoring dependence of observations within same cluster/psu. This is true for both descriptive and structural estimation

1.1.3 Dealing with heterogeneity and design

- One extreme is standard approach for modeller to assume homogenous behaviour across the subunits and pool the data ignoring weights.

  in fact wise to calculate both weighted and unweighted and compare or test differences using auxilliary regression with

  $$Y_i = \alpha + \beta X_i + \gamma W X_i + v_i$$

  and testing $\gamma = 0$
- Other extreme is consider behaviour to differ across subunits and estimate separate regressions for each and then combine results using population weights when distribution between groups are of interest can test for differences using covariance analysis.

- When there are many sectors can assume intersectoral heterogeneity with random variation in parameters: weighted and unweighted resids will be heteroscedastic and dependent and neither weighted nor unweighted will be consistent
- When explanatory variables differ within clusters or their are unequal numbers of observations in each cluster, although OLS is inefficient the efficiency loss is typically small
  - but may still be large enough to justify correction
- Could estimate OLS and use residuals to correct for clusters using GLS type estimator