SURVEY REGRESSION

Stratification

- Dividing the population into relatively homogenous groups (strata) and sampling a predetermined number from each stratum will increase precision for a given sample size

Clustering

- Dividing the population into groups and sampling from a random subset of these groups (e.g. geographical locations) will decrease precision for a given sample size but often increase precision for a given cost.

Unequal Sampling

- Sampling small subpopulations more heavily will tend to increase precision relative to a simple random sample of the same size.

Finite Population

- Sampling all of a population or stratum results in an estimate with no variability, and sampling a substantial fraction of a stratum results in decreased variability in comparison to a sample from an infinite population.

Linear regression of Survey Data

- $Y = X\beta + \epsilon$
  - $Y \to M \times 1$
  - $X \to M \times K$
  - $\beta \to K \times 1$
- In the standard linear regression model ϵ is generally $N(0, \Sigma)$ where $\Sigma = \sigma^2 I_M$
- This is not the case in survey regression

Data Structure

- Suppose a sample of size M is drawn from a population of size T.
- The population is divided into H strata.
- In each strata h, n_h clusters are randomly sampled.
- From each cluster(h_j) a random sample of m_{h_j} individuals is selected.

Sampling weights

- Each sample hji can be attributed to a strata h, a cluster j and the individual i.

- Based on the design let p_{hji} be the probability of the element appearing the sample.

Missing regressors

- What if the model is incomplete?
- This is not a problem in standard regression as its contribution is treated as random.
- But in survey regression these missing regressors may be related to the sampling weights and may not be truly random.
- $Y = X\beta + Z + E$
  - Where Z contains the contribution from the missing regressors.

Estimation of β

- The estimation of β is straight forward as it is the weighted least squares estimate.
- $\hat{\beta} = (X^T W X)^{-1} \mathrm{X}^T \mathrm{WY}$
  - where W is the $\mathrm{M} \times M$ diagonal matrix of sampling weights

Better than OLS estimate

- $E(\hat{\beta}) = \beta + (X^T W X)^{-1} \mathrm{X}^T \mathrm{WE}(\mathrm{Z})$
- If E(Z)=0 i.e. no missing regressors both OLS and WLS would be unbiased estimators.
- However when $E(Z) \neq 0$ both the estimators are biased.
- But the bias of the WLS estimator falls sharply with sample size and is negligible for large sample sizes.

Variance Computation

- Similarly the variance of WLS can be computed.
- Several estimates for the variance were proposed which had better asymptotic and robust properties than the WLS variance estimator.
- Variance estimates of WLS are typically higher than OLS.
- The bias-variance trade off.

National Health Interview Survey

- Nationally representative sample of the civilian, noninstitutionalized population of the United States
- Face-to-face personal interviews
- Continuously conducted since 1957
- Topics cover a broad range of health issues and provide information on both acute and chronic conditions

NHIS sample design

- Has a “complex” design (multistage design includes clustering, stratification)
- 428 primary sampling units (PSUs) drawn from approximately 1800 geographic areas covering the 50 states and the District of Columbia
- PSUs are individual counties or contiguous groups of counties and vary in size from a few hundred to several million.
- Sampling geographic areas helps to control survey costs

NHIS sample design

- Approximately 40,000 households containing almost 100,000 persons are selected from Census-defined tracts and block groups
- Currently oversampling Blacks, Hispanics, Asians, and elderly minorities in these groups
- Detailed health information collected from one sample adult and one sample child per household

NHIS sample weights

- Weights composed of three components:
  - The reciprocal of the probability of selection
  - A household nonresponse adjustment
  - Post-stratification adjustment to the U.S. population by age, sex, and race ethnicity

Inverse Probability Example

- Suppose that there is a population of 100,000 people, and there is enough money in the grant to collect data from 1,000 people.
- 20% indigenous population.
- The population is divided into two regions, (A and B).
- Region A has 25,000 people, 50% of whom are indigenous.
- Region B has 75,000 people, and 10% are indigenous.
- They will choose a 2% sample of people (n=500) from Region A and .67% (n = 500).from Region B from (n = 500).

- The likelihood of a person in Region A being selected is 500/25,000. Each person in Region A represents 50 people (25,000/500 = 50).
- The chance of a person in Region B being selected in 500/75,000. Each person in Region B represents 150 people (75,000/500 = 150).
- Note that the weight for people in Region A are lower than those in Region B. People in Region A are overrepresented in the sample, and people in Region B are under-represented in the sample.

Non-response Weighting Example

- Question: Have you ever visited Houston? (Y/N)
- Population information available about age (18-30) (31-64) (65-older)
- Comparison of respondents and population
- Weighting

Age	Resp	Popul	Weight
18-30	20%	30%	30/20=1.5
31-64	70%	50%	50/70=0.7
65+	10%	20%	20/10=2.0

Houston	18-30	31-64	65+	Unw	W*
Yes	20% (4)	50% (35)	10% (1)	40% (40)	33% (33)
No	80% (16)	50% (35)	90% (9)	60% (60)	67% (67)
N	20	70	10	100	100

Yes: 4*1.5 + 35*.7 + 1*2.0=6+24.5+2=32.5

No: 16*1.5 + 35*.7 + 9*2.0=24+24.5+18=66.5

Post Stratification weights

- Typically used to adjust for minor differences in nonresponse by demographic subgroup.
- Bring the sample proportions in demographic subgroups into agreement with the population proportion in the subgroups.
- Requires auxiliary dataset to use as a comparison.

	Sample Percent	Population Percent	Weight
Male	42%	49%	1.16
Female	58%	51%	.879

Survey Analysis for NHIS data

- Create a survey object
- svy<-svydesign(id=~PSU_P, strata=~STRAT_P, nest=TRUE, weights=~WTFA_SA, data=Y)

- Logistic Regression
- svyglm(r1~AGE_P+SEX+…, svy, family="binomial")