AGRODEP

Household survey data course

Dakar, 8-10 October 2012

Cluster Effects

Cluster sampling

- Population divided into groups of the units of analysis called clusters
- Generally used in multi-stage sampling
- Examples:
  - Sample of enumeration areas – cluster of households – selected at first sampling stage for household surveys
  - Education survey – schools or classes can be defined as clusters of students
- Three or more sampling stages – different levels of clustering

Advantages of clustering

- Reduces costs
  - concentrating survey efforts in sample clusters
  - updating of frame (listing) only needed in sample clusters

- Facilitates logistics, operational considerations, supervision and quality control

- Generally not feasible to have a simple random sample of units

Some practical considerations

- Take (hhs/cluster=*m*) vs. no. of clusters
  - Optimal take, equal workload, …
- Intracluster correlation unknown. Rule of thumb: 0.1-0.3 unless …
- Other rule of thumb: Deff = 2
- Cost function often unknown.
- Lost efficiency and sample size increase

Cluster effect

Standard error grows when the sample of size n is drawn from k PSUs, with m households in each PSU ($n = k \bullet m$)

$$\underbrace{e_{TSS}^{2}}_{\textit{Two Stage Sample}} = \underbrace{e_{SRS}^{2}}_{\textit{Simple Random Sample}} * \underbrace{\left[1+\overset{\textbf{Intra-cluster correlation coefficient}}{\rho}(m-1)\right]}_{\textbf{Cluster effect}}$$

Cluster effect

$$\hat{\rho} = \frac{\sum_{c=1}^{C} \sum_{j=1}^{m} \sum_{k \neq j}^{m} \left(x_{jc} - \bar{x}\right)\left(x_{kc} - \bar{x}\right)}{C\, m(m-1)\hat{s}^2}$$

C = number of clusters

- Typical number of households per cluster:
  - 10 to 15 sample households for socioeconomic and LSMS surveys
  - 20 to 25 households for Demographic Surveys

Cluster effect

- The cluster effect increases with the intraclass correlation coefficient (ϱ) and the number of sampling units per cluster
- The intraclass correlation coefficient is
  - Very high (> 0.2) for variables of infrastructure
  - High (~ 0.05) for socioeconomic variables
  - Low (< 0.02) for demographic variables

Cluster effect

For a total sample size of 12,000 households

Number of PSUs	Number of households per PSU	Intra-cluster correlation coefficient 0.01	0.02	0.05	0.10
3000	4				
2000	6				
1500	8				
1000	12				
800	15				
600	20			1.95	
400	30				
300	40				
200	60				
150	80				
100	120				

Cluster effect

For a total sample size of 12,000 households

Number of PSUs	Number of households per PSU	Intra-cluster correlation coefficient			
		0.01	0.02	0.05	0.10
3000	4				
2000	6				
1500	8				
1000	12				
800	15				
600	20	1.19	1.38	1.95	2.90
400	30				
300	40				
200	60				
150	80				
100	120				

Cluster effects

For a total sample size of 12,000 households

Number of PSUs	Number of households per PSU	Intra-cluster correlation coefficient			
		0.01	0.02	0.05	0.10
3000	4			1.15	
2000	6			1.25	
1500	8			1.35	
1000	12			1.55	
800	15			1.70	
600	20	1.19	1.38	1.95	2.90
400	30			2.45	
300	40			2.95	
200	60			3.95	
150	80			4.95	
100	120			6.95	

Cluster effect

For a total sample size of 12,000 households

Number of PSUs	Number of households per PSU	Intra-cluster correlation coefficient			
		0.01	0.02	0.05	0.10
3000	4	1.03	1.06	1.15	1.30
2000	6	1.05	1.10	1.25	1.50
1500	8	1.07	1.14	1.35	1.70
1000	12	1.11	1.22	1.55	2.10
800	15	1.14	1.28	1.70	2.40
600	20	1.19	1.38	1.95	2.90
400	30	1.29	1.58	2.45	3.90
300	40	1.39	1.78	2.95	4.90
200	60	1.59	2.18	3.95	6.90
150	80	1.79	2.58	4.95	8.90
100	120	2.19	3.38	6.95	12.90