Urvalsmetoder och Estimation 8

Sampling and Estimation 8

2012-02-27

Today

14 -14.45 Lecture by DT

15 - ca 16 Guest lecture on LFS Frida Videll

Ca 16.15 - 17 Lecture by DT

Ca 17- 17.45 Exercise by DT (NP is still sick)

Examination

- 10 assignment problems give maximally 100 points
- The presentation gives maximally 25 points
- For the assignments to get the same weight this is multiplied by 100/125
- Written exam gives 100 points. Added together with the asignemnts give maximally 200.
- 100-119 E 120-139 D, 140-159 C, 160-179 D, 180-200 A
- Example Assignments 1-3 gave 70. presentation 20 and written test 77. The total is (70+20)*100/125 + 77 = 149 corresponding to a C

8. Coordinated Samples
(Samordnade urval)

- Why coordinate samples and surveys?
  - The Response Burden
    - Distribute it more equally
    - Decrease it
    - Negative coordination, when little overlap
  - Information on relations
    - Over time, longitudinal studies
    - Between surveys. If the same persons participate in two surveys, one may study relations between their study variables e.g. health and economic variables or sports participation
    - Positive coordination, with large overlap

8.1 Coordination within surveys

Two independent surveys

No or negative coordination

Panel Questionblock	1	2	3	4	5	6
1	X	X	X			
2	X	X	X			
3				X	X	X
4				X	X	X

Coordinated surveys

Three surveys with a common basic block

The common part can be used for specially important question where a larger sample is needed or as a background for calibration or finding relations implicitly.

SILC and the Swedish social statistical system are planned like this

Panel Questionblock	1	2	3	4	5	6
1	X	X	X	X	X	X
2	X	X				
3			X	X		
4					X	X

Coordinated surveys -split questionnaires

Split questionnaires are nowadays easy to administrate with web-surveys or CATI.

All bivariate distributions studied

Observe that all pairs of blocks appear for some panel

Panel Questionblock	1	2	3	4	5	6
1	X	X	X			
2	X			X	X	
3		X		X		X
4			X		X	X

8.2 Longitudinal studies

Longitudinal studies is a term for studies where one follows the same units over time (The opposite is cross-sectional studies), (not only sample surveys)

Typical examples: You want to follow up what happens to those firms that were reconstructed during the financial crises or were fined from environmental reasons or have female members of the board

To follow units over time means that you must be able to know what is meant by the same unit. What to do at takeovers, fusions, bankruptcies with a following reconstruction and spin-offs (easy for persons but not for enterprises or households).

- Rotating samples (e.g. Labour Force Survey, Party preference study)
- The same persons are asked the same questions several times (c.f. Survey of Living conditions asked every 8th year, EU-SILC all household members are followed for four years)
- The same persons are followed but different questions are asked (c.f. the IDA-project (Psychology dep SU). School-children in the third grade 1965 are followed through life – school –education – working life - migration - family formation – housing –children etc)

Attrition

- A special problem with longitudinal studies is attrition which is a special case of nonresponse. If a sample is studied for 10 periods and each time 10% disappears the total response will be $(1-0.10)^{10}$ = 34 %
- Not so much in Sweden with good registers but in countries with less information like developing countries or USA
- Even in a two period situation the nonresponse may be doubled if the nonresponses are independent at both time points. 25% at two time points means 44 % in total

Epidemiology

- Prospective - Forward in time.
  - A study has been made a long time ago (i.e. At ”mönstringen” all male Swedish 19 years old were tested and measured).
    - Make a study now and relate their health and success in life to their obesity at 19 years age
  - Exposure-control.
    1. Take a number of exposed persons (e.g. workers at a special factory e.g. Rönnskärsverken, 1990) (Exposure group).
    2. Find similar persons in the whole population (random or matched i.e. each person gets one or two twins, same age, education, type of work etc. Control group)
    3. Study their health today looking for differences between the exposure and control groups. i.e. for effects of the work environment

- Retrospective - looking back in time
  - Ex. Case-control
    1. You have a number of cases, (e.g. of lung cancer) (Case group)
    2. Select a number of similar individuals in the whole population so that. (at random or matched i.e. each case gets one or two twins after e.g. age, education, marital status, housing) (Control group)
    3. Ask all in the case and control groups. See if the groups differ in other relevant aspects e.g. (in previous smoking habits).
  - Much used but considered inferior to prospective studies since there is a risk for response bias.

Analysis - simple

- The usual approach to these forms of analysis are odds ratios (OR) and logodds ratios
- OR = $n_{11}n_{22}/(n_{12}n_{21})$
- Logodds ratio = ln(OR)
- St dev* (logodds ratio) = (appr) Root($1/n_{11} + 1/n_{12} + 1/n_{12} + 1/n_{22}$)

Number	Cases	Healthy	Sum
Exposed	n_{11}	n_{12}	$n_{1.}$
Not exposed	n_{21}	n_{22}	$n_{2.}$
Sum	$n_{.1}$	$n_{.2}$	$n_{..}$

- When the logodds ratio is positive there is a positive relation between exposure and incidence and 0 means independence
- The analysis is independent of whether the data are obtained by case-control, exposure control or as a SRS sample from the population

An example – a study on the effect of class size

10000 persons 30 years old were randomly selected and the size of their class in fifth grade and their wages today were found from registers.

Figures are faked but I recently heard about a study to this effect on TV.

Number	Income above median	Income below median	Sum
Class size > 25 in fifth grade	1852	1271	3123
Class size ≤ 25 in fifth grade	3148	3729	6877
Sum	5000	5000	10000

OR = 1.73 ln(OR) = 0.55 s.e. = rot(0.0019) = 0.044

Usually interpreted as saying that the effect on the income lies between 0.46 and 0.64. It is clearly significant but of moderate size.

8.3 Rotating panels

- Every unit is included in the survey a predetermined number of times (waves) e.g. four years and a fourth of all firms are replaced every year.
- A common sampling scheme for permanent or a intermittent surveys is rotating panels (cf Labour Force Survey)

Ex: Four active rotating panels

Time Panel	2000	2001	2002	2003	2004	2004	2005	2006	2007
A	X								
B	X	X							
C	X	X	X						
D	X	X	X	X					
E		X	X	X	X				
F			X	X	X	X			
G				X	X	X	X		
H					X	X	X	X	
I						X	X	X	X
J							X	X	X
K								X	X
L									X

Advantages with rotating panels

- The respondent burden is decreased in this way
  - Since it is always most work the first time you are in the study
  - Basic questions are asked only once. Or you ask only if there is a change since last time
  - But no unit will always be in the study. This would be considered unfair and there may be an effect on the behaviour of the firm if you are in such a study for a long time.
  - Smaller tracking and contacting costs after the first wave. (The interviewers may know whom to phone)
- You can study the development over the years in another way.
- E.g short period salary statistics, LFS (the labour force survey), EU-SILC

Estimation with rotating samples with k active and equally large panels

Composite estimators

- Let X_{ti} be the estimate of the mean at time t from the i:the panel
- Suppose that all these estimates have the same variance, σ^2, and that the correlation decreases exponentially between times within panels $\rho^{|t1-t2|}$ (Large firms are usually large also next year)
- A simple estimate of the mean at time t is then the mean of all panels $\Sigma_i X_{ti}/k$ with the variance σ^2/k (no correlation between panels)
- The variance for the difference between two time points, t och t+1, will then be $2(1 - ((k-1)/k)\,\rho)\,\sigma^2/k$ (Prove it!)
- The random error decreases with the number of panels i.e. the period of rotation, k. The variance without any overlap (two independent samples) would have been $2\,\sigma^2/k$
- E.g. with k = 4 and $\rho = 0.9$ the gain is a factor 0.325

- But it is possible to do something even better (but it is seldom done)
- The difference between the first and second time point can be estimated in two ways:
  - The difference between the common panels $D_1 = \Sigma_{i=2}^{k} (X_{2i} - X_{1i})/(k-1)$ with variance $2(1-\rho)\,\sigma^2/(k-1)$
  - The difference between the new and old panel $D_2 = (X_{2k+1} - X_{11})$ with variances $2\sigma^2$
  - If these are weighted together with optimal weights (inversely proportional to their variance) one gets $(D_1 + (1-\rho)/(k-1)\,D_2)/(1 + (1-\rho)/(k-1))$ with the variance $2\sigma^2/(1 + (k-1)/(1-\rho))$ (Prove it!)
- With k = 4 och = 0.9 the gain will be a factor 0.129
- Can you explain why this is seldom used?

- One does not want to change already published estimates.
- And it is natural (but not optimal) to estimate the level one year with the average of all the values observed that year
- One wants to have consistency, the estimate of the change should be the difference between the two level estimates. But as we saw one looses precision by requiring this.
- It is of course possible to get optimal and consistent linear estimates at any given time (Just a projection of the data), but it is not possible if one also requires that estimates should not be changed.

- It is possible improve other estimators too
  (One may for example estimate the level at time 2 better by using that we know if the three remaining panels were higher than the leaving one).
- We have two estimates of the level at time 2:
  - $E_1 = X_{2,k+1}$ and
  - $E_2 = \Sigma_{i=2}^{k} X_{2i}/(k-1)$.
  - We further have one estimate of 0: $E_3 = \Sigma_{i=2}^{k} X_{1i}/(k-1) - X_{11}$.
- Any expression of the form $a\, E_1 + (1-a)\, E_2 + b\, E_3$ is an unbiased estimator of the level at time 2.
- It is straight forward to compute its variance and minimise it.

Transition probabilities – discrete data

- A similar problem occurs when data ar on a ”nominal scale”.
- Let $p_{ij} = n_{ij}^{(2-k)}/n_{i-}^{(2-k)}$ be an estimate of the proportion of category i that next year will be in category j (2-k means the units present at both time points)
- But $\Sigma_i\, n_{i-}^{(1-k)}\, p_{ij}$ does not equal $n_{-j}^{(2-k+1)}$
  (1-k and 2-k+1 means all units present at time 1 resp 2)
- This inconsistency in the estimates are usually accepted, however.

8.4 Changing populations

- Are always problematic.
- Compare - We are interested in the chemical industry
  - Sample from all firms according to last year’s register – remove overcoverage.
  - Sample only firms which last year were in the chemical industry and remove overcoverage.
  - The second procedure gives a smaller industry, why?
  - What about newly started?
- A consistency criterion may be that the total estimate for one year should be the same from different surveys. But what if they are done at different times for administrative reasons?

8.5 Screening

- Screening. You want in particular to study enterprises with a special property
  - e.g. newly made investments for environmental reasons, have a female managing director, have disabled employees or have used special EU-money.
- In the main survey you can ask about if the the firm belongs to this group. Later you return with a more detailed survey on that issue.
  - Sometimes the last step can be a planned subsample in the selected group. E.g. For the second step choose equally many with male and female managing director – or even matched pairs. This will lead to more efficient comparisons

An example

- 500 enterprises are studied in a SRS-sample from 5000 firms in the first round
- In one important respect they can be classified into four classes with 250, 150, 75 and 25 firms after a variable observed in the first round
- 100 firms are selected for the second round, 25 from each group.
- Observed stratum means and variances in the second round are 5, 25, 30, 145 and 5, 4, 20, 200, resp.
- Now estimate the total
  - Mean (250*5+150*25+75*30+25*145)/500=21.75
  - Its variance is more complicated (see next page).
- This approach can also be used for comparing the means in different groups in an efficient way

Variance estimation

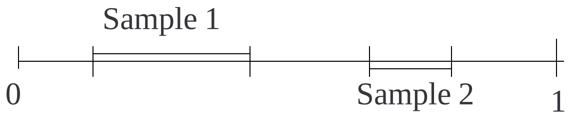
- First consider the variance if the value of all units in the same group had been the same (i.e. 250 units with value 5, 150 with 25, …). Standard SRS-formulas give 1.65
- Next compute the variances in the second step wihin each group (drawing 25 from 250 and … with SRS). 0.18, 0.133, 0.533, 0. Weighting them together gives 0.069.
- The sum of the two components gives 1.72.
- The variance if only one SRS-sample with 100 units had been drawn would have been 8.39. This two stage sampling procedure has improved precision considerably.

8.6 Permanent Random Numbers

Given a frame, which will be used many times (e.g. The Swedish Business Register).

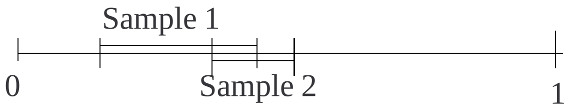
- Every unit, *i*, gets a uniformly distributed random number, U_i, (in (0,1)) .
- Idea of negative coordination illustrated for SRS.
  - Pick the n_1 units with the lowest random variables for the first survey.
  - Pick the next n_2 random variables for the second survey

Negative coordination – No elements in common



The elements' permanent random numbers

Positive coordination – 50 of % of sample 2 in common



The elements' permanent random numbers

Permanent Random Numbers

- It is easy to see that one can get any amount of overlap between different studies
- Since the numbers are permanent it is easy to make longitudinal studies.
- When new units enter the register they get random numbers and if it is in the designated interval the new unit should e.g. be included in next wave of a longitudinal study.