Urvalsmetoder och Estimation 6

Sampling and Estimation 6

2011-02-21

7. Analysis of data with non-response

7.1 Missing values in general

There are many good methods to deal with data having missing data - but no perfect. It is quite natural since the missing data may be anything and you will never know.

Non response is a serious problem

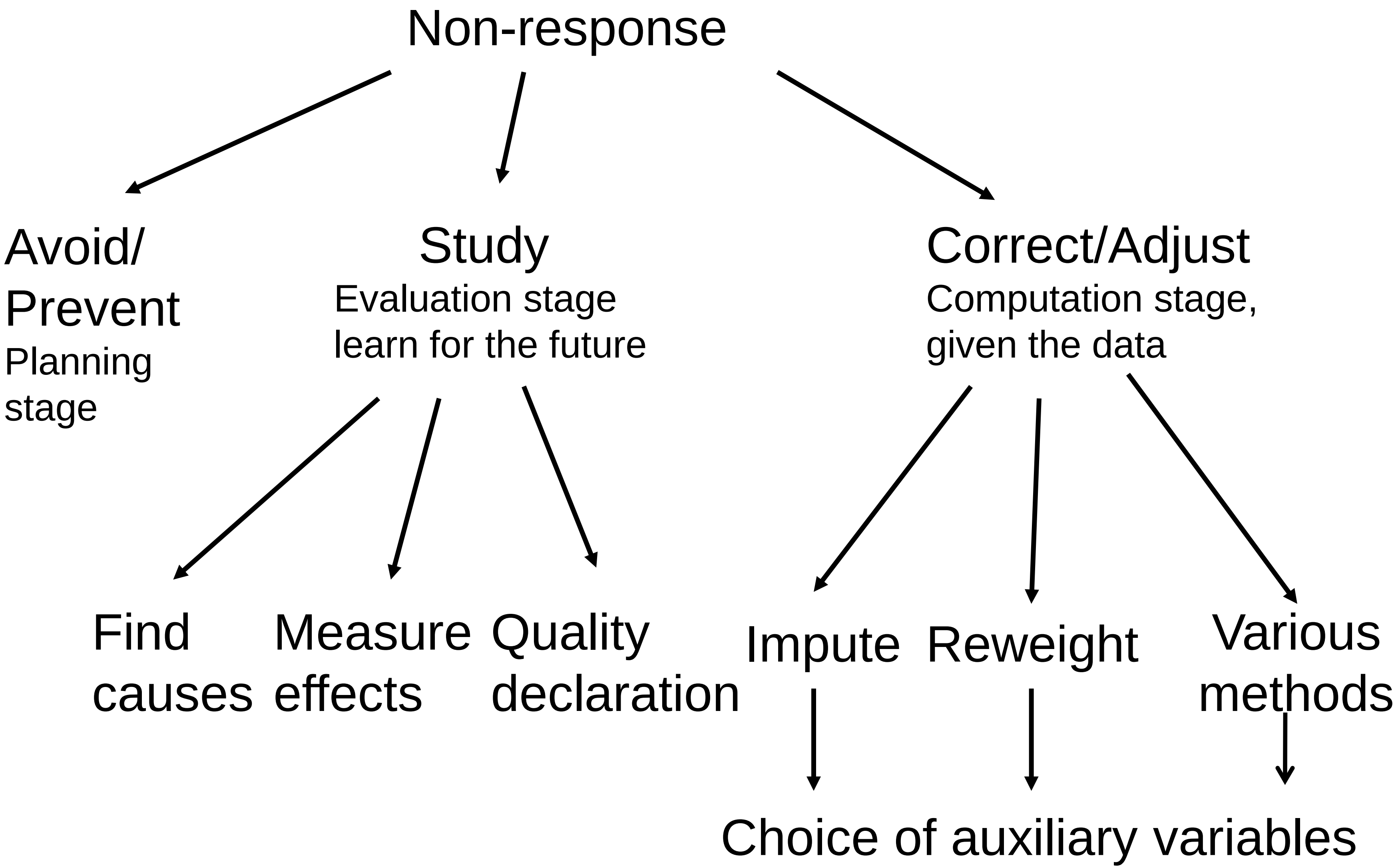
Nonresponse for some Swedish telephone interview studies:
(CATI, performed by Statistics, Sweden)

- 23.7 %   Labour Force Survey   Aug 2010
  (refusals 10,8 %, non-contact 12.0 %, other 0.9 %)
- 32.1 %   Party sympathy survey May 2010
  (refusals 13.6 %, non contact 15.6 %, other 3.5 %)
- 35.3%   The Households' Economy (2008)
  (refusals 16.9 %, non contact 14,9 %, other 3.3 %)
- 43.3 %   Citizen study
  (average of 430 municipal studies, 2005-2009)
- 49 %   Households' expenditure study, 2009

The Survey Climate

- Non-response has been steadily increasing in Sweden and internationally during the last 35 years.
  - For LFS from about 5% 1975 to 23.7 % today.
  - Higher in some important groups: Young, Socially weak (e.g. homeless), Immigrants
- Tore Dalenius: never accept over 10 % and never above interesting variables.
- People (compared to 20 years ago) tend to be
  - More positive to statistics, not afraid of large data-bases and computers
  - More self-centered and to show less solidarity with the society

Outline of non-response literature



Non-response

Avoid/ Prevent
Planning stage

Study
Evaluation stage
learn for the future

Correct/Adjust
Computation stage,
given the data

Find causes

Measure effects

Quality declaration

Impute

Reweight

Various methods

Choice of auxiliary variables

Use a combination of measures!

- I will mainly deal with methods to correct data with missing values.
- In practice, this is only one aspect. It must be weighted against other measuures to decrease the non-response effects and increase the quality

7.2.1 Preventing non-response

- Designing surveys to reduce non-response
  - Support from influential people/organisations
  - Compulsary response
  - Proxies – On-line editing
  - More efforts in difficult strata (e.g. earlier start)
  - Call-backs – reminders – small gift – payment
  - Sub-sampling in the non-response
  - Panel care – decrease burden – Adapt to accounting practices
  - Interviewer education …
- Chose other frames or modes e.g.
  - web panels, RDD
  - Multimode surveys. Different modes for different groups, during different phases …
  - register studies …

- Chose other methods, frames or modes e.g.
  - Web access panels
  - RDD
  - Multimode surveys.
    - Different modes for different groups
      - Personalise the cover letter
      - Use the web to approach the young
      - …
    - Different modes during different phases
  - Register studies …

7.2.2 Reasons for missing values

- In personal interviews mainly due to refusals or not found. (To a smaller extent also sick, language problems or unable to answer)
- These groups usually differ. Not found is usually a much worse problem than refusals.
- For mail surveys the main reasons are that people forget/postpone or that the questionnaire is too complicated or the subject too unimportant or uninteresting. People outside the frame
  - Overcoverage usually respond less.
  - In a study of political interest, those that are politically active respond often fast and the uninterested late and after several reminders or not at all.
  - Recent study of pets in Sweden by Statistics Sweden gave more than double the expected number of pets.
- Technical reasons e.g. overcoverage and doubles in the frames.

Reasons for missing values

- Knowing the reason is important when interpreting the results
- Thus in Quality declaration:
  - If possible present non-response after reason and for important subgroups (age, sex, region …)
  - Always present uncorrected estimates. Only presenting the non-response corrected estimate means that the reader cannot use such knowledge.

- Item non-response (Partially missing)

  There are data on some variables - but not all - for some units

  - The respondent does not answer some questions (or gives illegible answers).
    - Badly constructed interview/questionnaire (respondent forgets to answer or misses the question),
    - Difficult questions (respondent does not know what to answer) due to bad formulation or difficult issue
    - Sensitive question (respondent does not want to answer)
    - There is an answer ”don´t know”
    - Does not answer, when there is nothing to report
    - etc
  - Data from a register, covering only part of the population
  - The question was never asked to that unit

- Structurally missing

  Variables, which logically do not exist.

  - Age of the children if the unit has no children
  - Area of apartment/dwelling for homeless
  - Value of real estate or stock for companies having no real

  Missing by design

  - split questionnaires

- Dark numbers (Mörkertal)

  You do not even know that the unit exists.

  - Called frame error if the register is used as a frame in sampling.
  - But dark numbers if the object of statistics is to find the total number of something
    - E.g. establishments in Sweden, crimes or traffic accidents, persons suffering from KOL

7.2.3 Classification of assumptions

Some assumptions on the non-response must always be made

A typical Tore Dalenius illustration. What is possible without assumptions?:

- Suppose we intended to study 1000 persons and ask about unemployment
- 800 answers (response rate 80%)
- 6% of them unemployed (48 persons)
- Highest possible number of unemployed in sample: 248
- True value in the sample lies somewhere between 4.8% and 24.8%
- A 95% interval is (4.8-2*0.7, 24.8+2*1.35) = (3.4, 27.5)

Main notations

Not always used. Sometimes other notations may be used if suitable

- U population
- S intended sample
- R the obtained sample (response set)
- Y Studied variable
- X Auxilary information / variables
- Z Response indicator
- π_i, π_{ij} first and second order inclusion probabilities
- q_i response probabilities
- ω weights (designweights, $= 1/\pi_i$)
- f prediction function
- g used transformation of the auxiliary information

Rubin’s classification of non-response

Often classified into three categories

- MCAR, Missing Completely At Random
  - The non-response mechanism and the study-variables are independent (The nicest situation)
  - (Y,X) $\perp$ Z where Z is the random response indicator
- MAR, Missing At Random
  - The non-response and study-variables are independent conditio-nally on what is known (the auxiliary variables X and what is observed. The last part needs some care to avoid a vicious circle)
  - $Y \perp Z \mid X$ or sometimes $Y \perp Z \mid g(X)$ for some function g
- NMAR, Not Missing At Random
  - The non-response depends on the study-variable and this cannot be handled by using only what is known.

MCAR - Ignorable non-response

- Forget about the nonresponse and assume that this was the intended sample
  - Usually: If the sample is drawn by SRS and the observed size is r, instead of the intended size n, analyse as if intended size was r.
  - Similarly for stratified sampling. Analyse as if observed stratum sizes r_h were the intended stratum sizes n_h.

- If a general sampling scheme was used second order inclusion probabilities are needed
  - But one may always assume independence (Allowed since MAR assumes no information in missingness).
  - Set $\pi_i^*=\pi_i r/n$ and $\pi_{ij}^*=\pi_{ij}\, r(r-1)/n(n-1)$ (the second factor corresponds to the assumption that response being SRSwor from S. Other suggestions exist)
  - Use ordinary formulas for πps-sampling like HT-estimators and SYG-variance estimator or HT-ratio-estimators.

MAR

Sometimes also called ignorable non-response.

- If the correct method and auxiliary variables are chosen we may ignore non-response.
- But one cannot ignore the non-response from start

Almost all available correction methods remove non-response bias completely if and only if some type of MAR is assumed.

MAR never holds exactly but may be close to the truth and then the methods work quite well.

In the following we assume that there are auxiliary variables and discuss metods that work under MAR

Take into account what you know
-
The problem is the factors that you do not know

A rule for compensation methods

If MAR holds it is enough to do the first (but in an appropriate way)

- One formal definition of MAR:

Let Q be a subset of elements of U.

For all Q such that P(R=Q) > 1 (i.e. Q might be the response set) it holds that

Z and Y are independent given $\{Y_i; i€Q\} \cup X$

Remember that Z is the missingness indicator telling which units (outside Q) that would respond if asked and Y is the value of all these units.

NMAR - Non-ignorable non-response

- The only way to handle NMAR is to model the nonresponse and/or to make follow-up studies of the non-response.
- We leave NMAR for the moment

7.3 Non-response, adjustment

7.3.1 Many methods to handle MAR – to correct the estimates.

- Reweighting
  - – Original estimate $t_{yS} = \Sigma_{i€S}\ \omega_i\ y_i = \Sigma_{i€S}\ y_i/\pi_i$
  - – Final estimate $t_{yR} = \Sigma_{i€R}\ \omega*_i\ y_i$

- Imputation
  - – Predict the values of all missing data in some sensible way using what is known $y*_i = f(S,X_i)$
  - – Final estimate $t_{yI} = \Sigma_{i€R}\ \omega_i\ y_i + \Sigma_{i€S\text{-}R}\ \omega_i y*_i$

- Both reweighting and imputation
  - Post stratification
  - Generalised regression estimators

- Reweighting
  - RHG-groups (Response homogeneity classes)
  - Weighting classes
  - Calibration
  - Raking or iterative proportional fitting
  - Propensity (scores)
  - Reweighting with estimated response probabilities
  - Estimate response probabilities by asking for them

- Imputation

(Used also for item non-response)

  - Mean value imputation
  - Regression imputation
  - Plus random error
  - Nearest neighbour
  - Two classifications
    - Model or real donor
    - Hot or cold deck
  - Multiple imputation
  - Mass imputation (Imputation to the population level)

7.3.2 Poststratification

- Suppose we have an auxiliary variable X which is discrete or which can be classfied. In that case one may poststratify after X (or the grouped/classified variable).
  - For instance last years's value, sex, age, region sometimes also income group or anything where the total for the full frame is known
- For continuous auxiliary variables a classification in at least 5 - 7 groups is often a good procedure
  - e.g. age in many groups because both ends behave differently
- If the situation is MAR, this removes all non-response bias. If NMAR it usually decreases the bias but it is not removed completely

7.3.3 Generalised regression estimators (GREG)

- Do as we did for regression estimators earlier.
  - Use the response set R to find a relation between X and Y e.g. Y* = a + bX
  - Replace all missing data in S – R by Y*.
  - Analyse
- In practice, a classification of X followed by poststratification is usually better than simple linear regression since the relation is seldom linear and the model is estimated from R and used in S-R, where the relation may be different

These methods can be seen as both reweighting and imputation!

- Poststratification may be seen as
  - changing the weights of all units in R from $1/\pi_i$ to N_h/r_h
  - replacing all non-respondents in the strata by the stratum mean

- Regression estimates may be seen as
  - changing the weight of all units in R from $1/\pi_i$ to $1/\pi_i + \Sigma_U X_j \Sigma_R (X_i - Xbar) / \Sigma_R (X_i - Xbar)^2$
  - replacing all non-respondents by $Y_i^* = a^* + b^* X_i$

- Note that the assumption of MAR in connection with different adjustment methods may mean different things.
  (That model residuals are independent of non-response may mean different things for different models, since e.g. fixed levels within poststrata and a linear regression model are different)

- In practice a classification of X followed by poststratification is often better than simple linear regression
  (since relations are seldom nonlinear and the model is estimated from R and used in S-R, where the relation may be different)

- The models that we will describe later will also require other things
  (but we use the name MAR or ignorable non-response)

7.4 Size of non-response bias

Uncorrected estimates

(Assume SRS for simplicity)

Let p_i be the response probability of respondent i
(In some literature called propensity)

The bias is approximately

$$\Sigma_U p_i(y_i – \text{y-bar}) / \Sigma_U p_i = \text{cov}(p,y) / \text{p-bar}$$

Corrected estimates

Regression estimates with one auxiliary variable

The bias is approximately
$\Sigma_U\ p_i(y_i – \text{y-bar} - b^*(x_i – \text{x-bar})) / \Sigma_U\ p_i$ =
cov(p,y) - cov(p,x)*cov(x,y)/var(x) / p-bar

Thus the relative bias decreases as from $\rho_{p,y}$ to $\rho_{p,y} - \rho_{p,x}\ \rho_{y,x}$

Holds for most correction methods even though we derived it for SRS and regression estimates.

Correspondingly for multiple regression $\text{cov}(p_i,y_i)$/p-bar to
$(\text{cov}(p,y) – \Sigma_{y,X}\ \Sigma_{X,X}^{-1}\ \Sigma_{y,p})$/p-bar
Where the Σ's are variance/covariance matrices/vectors

Thus the auxiliary variable should be related to both response probability and the study variable. Both relations are, as we saw, needed.

- If X is only related to the study variable most methods decrease the variance but leaves the non-response bias unchanged

- If X is only related to the response probability most methods have usually only a marginal effect (usually a slight increase) on variance.

- If some but not all study variables are related to X it may be a good procedure to poststratify in the same way for all study variables. (You get consistency and looses very little efficiency)

It is important to find the best auxiliary variables.

- The compensation methods differ, as we saw and shall see, technically quite a lot. In practice the resulting estimates do not differ so much. Since all methods are developed by sensible persons, they behave sensibly (i.e. similarly).
- It is usually more important to decide which auxiliary variables to use and how (transformations, classifications, interaction) than which method
- However, only a few methods give exactly the correct variance estimators under MAR.
- Most methods underestimate the variance.

(In particular they seldom takes the randomness in the non-response mechanism into account and the model estimation uncertainty)

Chosing auxiliary variables in practice

- Try adjustments after some explaining variables
  - If no method changes the estimate much – choose the best one (the one that changes the estimate most) and feel confident in the result
  - If the results are changed considerably by one or more of the methods. Choose the one, which changed the estimate most. But do not feel confident. The non-response obviously affects the result and there may remain factors influencing the estimates to some extent
  - You can seldom remove more than 75% of the bias using the explaining variables at hand

7.5 Methods based on weighing

7.5.1 Calibration

- Suppose that the estimator should have been
  $t_{yS} = \Sigma_{i€S}\ \omega_i\ y_i$
  if there had been no response (often $\omega_i = 1/\pi_i$)
- Now find new weights ω^*_i which are as close to ω_i as possible and such that $\Sigma_{i€R}\ \omega^*_i = N$ and
  $\Sigma_{i€R}\ \omega^*_i\ x_{j,i} = T_{Xj}$
  for some selected auxiliary variables X_j; j=1,…k
- Use the estimate
  $t_{yR} = \Sigma_{i€R}\ \omega_i^*\ y_i$

- ”As close as possible” is often taken to mean the Euclidean distance $\Sigma_{i€R}\ (\omega^*_i - \omega_i)^2$
- Minimisation is often simple to do by the technique of Lagrange multiplicators

$\Sigma_{i€R}\ (\omega^*_i - \omega_i)^2 + \lambda_0(\ \Sigma_{i€R}\ \omega^*_i - N) + \Sigma_j \lambda_j (\Sigma_{i€R}\ \omega^*_i\ x_{j,i} - T_{Xj})$

We can take the derivative with the respect to each of the unknown ω^*_i but we solve it in compact matrix form

$$= (W-W^*)'(W-W^*)+\Lambda'(X'W^*-T)$$

Where W and w* are column vectors with N weights each, Λ is a column vector with the k+1 Lagrange multiplicators, X is a Nxk+1 matrix with auxiliary variables (elements in the first row 1) and T a column vector with the known k+1 totals

Differentiate with respect to the vector W* gives

$$W-W^*+X\Lambda=0$$

A multiplication with X' gives $X'(W-W^*)+X'X\Lambda=0.$

Since $X'W^*=T,$ this gives $\Lambda=(X'X)^{-1}(T-X'W^*)$.

Inserting this, we get the usual general regression estimator (GREG); in matrix form $W^*=W+X(X'X)^{-1}(X'W-T)$

- Sometimes one starts with other weights like $\omega_i = n/(r\pi_i)$

- Sometimes weighted distance functions like $\Sigma_{i€R} (\omega*_i - \omega_i)^2 / \sigma_i^2$

- Often calibrated estimators turn out to be poststratified estimators or some type of regression (GREG-) estimators. (the last is true for the Euclidean distance as we saw above)

- Calibration was first presented (by deVille-Särndal) as a means to reduce sampling error but is mainly used today to reduce non-response bias and to get consistent estimates

On auxiliary variables for calibration

- The auxiliary variables may be
  - Indicators of different categories like males or Stockholm
  - size variables like turnover, number of employees last year or taxed income
  - For most continuous variables, like age and income, a classification scheme is usually preferable, i.e. use a dummy variable for age below 20, age between 21 and 35, a.s.o.
  - It is dangerous to calibrate directly with very skew variables.

7.5.2 Propensity

- Estimate the response probability to respond as a function of the auxiliary variables X using the sample S (the missingness indicator is known for all S)

- E.g. Use logistic regression with response indicator Z_i = 1 if i€R and 0 otherwise finding

  p_i*(X) =(if X two-dimensional) = $p_i^*(x_{i1},x_{i2})$ = $\exp(c^*+d_1^*x_{i1}+d_2^*x_{i2})/(1+\exp(c^*+d_1^*x_{i1}+d_2^*x_{i2}))$.

- Or another model like probit regression or splines.

- This function is called propensity. Propensity is often interpreted as a synonyme to probability, but the name is chosen because it is not a probability. In medical trials it may be the propensity to go to a doctor if you get some decease or in Web-surveys the penetration of computer use.

- Replace the inclusion probabilities π_i by $\pi_i*p_i^*(X)$

- Use HT-estimator (or similar like HT-ratio-estimator) setting the weights to $\omega_i = 1/(\pi_i*p_i^*(X))$

7.5.3 Propensity scores

- Derive the propensity as above
- Order the sample after increasing propensity
- Divide into 5 (or another number of) groups
- Estimate, using these groups as groups with constant response probability. (I e. like using poststratification after these groups)

- The group indicator is called propensity score

- The propensity idea was first developed for medical studies. You want to compare two treatments which are allocated by a doctor based on some data. To compensate for the fact that some treatments are more common for severe cases propensity is used to estimate the tendency for doctors to give patients with different background data different treatments
- It is also often used in Web surveys. (e.g. Web access panels)

Propensity score -Web access panels

- Background data (X)
  - Make a traditional survey. Ask everyone about some background (auxiliary) variables i.e. life style. (Use in many studies)
- The study (Y)
  - Ask a sample (not necessarily a probability sample) about the study variables and the background questions.
  - In the pooled data set estimate the propensity to belong to the study as a function of the background questions. (logistic regression)
  - Divide the study data in five (or more) equal groups after the propensity. (These groups constitute the propensity score)
  - Estimate as if poststratified in those groups
- Under MAR, this will give an unbiased estimate (disregarding estimation problems and the discretisation effects)

Propensity score - properties

- Advantages
  - You may select specially select the best auxiliary information which is relevant both for nonresponse and the studyvariables in general.
  - May get a better correction than what you can get using already available data
  - The method is fast in the second stage (Web surveys with this method are usually presented within less than one week after survey date)
- Disadvantages:
  - Expensive in the first stage.
  - If the reference population (first stage) is not updated you will poststratify to an old population. Dangerous especially if it contains really time changing data (e.g. seen the last Harry Potter movie)
  - If there is non-response in first survey you do not correct for that. (Could be done but not straight forward)
  - If the propensity depends strongly on the selected variable you will get too small post-strata giving large uncertainties.
  - Seldom reduces the variance as ordinary methods like poststratification or calibration does.

7.5.4 RHG-groups

- Suppose that the sample can be divided into H groups with MCAR within groups. They are called Response Homogeneity Groups (RHG). They may depend on both the auxiliary variable, X, and the sample S, e.g. through the interviewer allocation

- Let there be r_h respondents out of n_h in RHG-group h.

- Set $\pi_i^* = \pi_i r_h / n_h$
  $\pi_{ij}^* = \pi_{ij} r_h (r_h - 1) / n_h (n_h - 1)$ for i and j in RHG-group h
  $\pi_{ij}^* = \pi_{ij} r_h r_g / n_h n_g$ for i and j in different groups, h and g.

- Use the weights $1/\pi_i^*$ e.g. use the HT-estimator and SYG-estimator

7.5.5 Weighting class

- Similar to RHG-groups

- Instead of estimating probability of reponse in each group by using r_h/n_h, use the sample weights, when estimating the response probabilities $(\Sigma_{Rh} 1/\pi_i) / (\Sigma_{Sh} 1/\pi_i)$

- If the assumption of constant response probability is true the RHG-estimate is better, but not generally. The weighting class technique is probably more robust.

7.6 Methods based on imputation

7.6.1 Real or Model donor imputation – Real donor

- Look for units in R with the same (or similar) X-values as the non-respondent (or similar observed values for item non-response).
- Draw one of these units (often with som random mechanism)
- Replace the non-respondent's values by those of this unit. Use standard estimation techniques on the completed data set.

- If always closest called Nearest neighbour imputation

- Advantages: All imputed values are realistic (not 0.4 children). The data can be handled by standard statistical packages. Variance estimation will be more correct compared to mean or regression imputation (see below)
- Disadvantage: You impute a random, maybe false, value, which means that you introduce an estimation error that was not there before

Model donor imputation

- From the data estimate a model for f*(x) = E(Y|X=x); (Y = f*(x) + e), e.g.
  - $f^*(x_i)$ = y-bar
  - $f^*(x_i) = a^* + b_1{*}x_{1i} + b_2{*}x_{2i}$
- Replace all non-responding units (in S-R) by $Y_i^* = f^*(x_i)$
- Estimate with the usual estimate. (But the formulas for the variance must be changed).
- Since the value is taken from a model, this situation is called model donor

- There is no random error in the imputed value. It lies on the ”regression” line)
- In order to make variance estimation better one sometimes imputes $f^*(x_i) + \varepsilon_i$ where ε_i is a random number with a suitable variance
- But this means that the location estimator will contain a home-made error, which is a problem

Real or model donor – an example

Data

Nr	var	Gender	Income
1	17	2	14180
2	16	1	-
3	14	-	27690
4	10	1	-
5	18	2	16189
6	10	2	23457

…

Real donor

Nr	var	Gender	Income
1	17	2	14180
2	16	1	14180
3	14	1	27690
4	10	1	23457
5	18	2	16189
6	10	2	23457

”Nearest Neighbour”

Model donor

Nr	var	Gender	Income
1	17	2	14180
2	16	1	20379
3	14	1,6	27690
4	10	1	20379
5	18	2	16189
6	10	2	23457

Mean value imputation

Standard confidence interval formula

- $t_{a/2}(n\text{-}1) * (\Sigma_{i€S} (y_i – y\text{-bar})^2/(n\text{-}1))^{1/2} /n^{1/2}$

- There are three reasons why the mean value imputation (simplest model donor case) gives too short intervals
  - Degrees of freedom to large (n-1) instead of r-1)
  - Variance estimator contains 0-terms (n-r terms: (y-bar – y-bar)^2)
  - Divided by $n^{1/2}$ ($n^{1/2}$ instead of $r^{1/2}$)

- One or more of the reasons hold for all model imputation methods to some extent (except multiple imputation)

Model donor - examples

X	17	16	14	10	18	10	15	22	37	48
Y	235	-	173	-	163	142	315	277	-	423

Mean 247 standard deviation 100 confidence interval 287 +/- 92

According to standard procedures based on seven values

Mean value imputation

Y	142	247	277	247	163	235	315	173	247	423

Mean 247 standard deviation 83 confidence interval 247 +/- 52

Gives good estimates of means or totals but too small variances

Regression imputation

Y	142	240	277	208	163	235	315	173	350	423

Mean 253 standard deviation 89 confidence interval 253 +/- 56

Too small variation in the data set and standard statistical packages will assume that there are more observations than in reality. Also correlations and other relations will be stronger and more often significant with this technique.

Model donor examples - continued

X	17	16	14	10	18	10	15	22	37	48
Y	235	-	173	-	163	142	315	277	-	423

Mean 247 standard deviation 100 confidence interval 287 +/- 92

Regression imputation

Y	142	240	277	208	163	235	315	173	350	423

Mean 253 standard deviation 89 confidence interval 253 +/- 56

If it done correctly the estimates of mean and variance will be unbiased but the results will depend on the randomness and cannot be replicated. The intervals will be too long since the imputation error is not included in the intervals

Regression plus randomness

Y	142	337	277	221	163	235	315	173	327	423

Mean 261 standard deviation 92 confidence interval 260 +/- 58

This randomness leads to a situation where all estimates are unbiased, but with a lot of extra randomness. To solve this you may repeat this imputation B times and take the avearge of them. ”Multiple imputation”

Another way to get more randomness is to use real donors with a suitable distance measure . But the variances will still be too optimistic.

7.6.2 Multiple imputation

- Step 1.
- Impute, not expected/predicted values as in the model donor case, but random values drawn from the full conditional distribution of Y given X.
- You must first draw regression parameters to take the uncertainty about them into account. Assuming normality this means draw σ^2 from inverse chi-square(r-1. s^2), , b from N(S_{xy}/S_{xx}. σ^2/S_{xx}) and a+b$\bar{x}$ from N($\bar{y}$, σ^2/r).

- This introduces a random error, but all imputed values are realistic values if the model is correct. They are as likely to be correct as other variables
- Estimate as if there were no non-response

- Step 2. Repeat this imputation and estimate B times (say 10 or 100 times)

- Step 3. Find the mean t_{MI} and variance V_1 of all these B estimates and the mean V_2 of all the B variance estimates.
- Then give this mean t_{MI} as the estimate and the sum of the two variances $Var^*(t_{MI}) = V_1(1+1/B) + V_2$ as the variance estimator of the estimate.

- The law of large number guarantees that your estimate does not depend on the random drawings.
- B should be chosen so large that V_1 /B is small compared to the total variance. (Thus it is often omitted)
- V_1 is the extra error, due to the non-response. V_1/B is due to taking too few imputation rounds. V_2 is the error thet we should have had if there was no non-response.

Multiple imputation is Bayesian theory

- ”the full conditional distribution of Y given X”. This is a Bayesian notion
- Classical people do multiple imputation but seldom completely correct. Distribution of Y contains unknown parameters and the uncertainty about them must also be included.
- But since the Bayesian estimator under normality and vague prior is the ML-estimates is the LS-estimate this can be done also from a classical point of view
- Sometimes the posterior is complicated and then it is possible to use MCMC-techniques, imputing after each step. B must then be chosen larger since the imputed values are no longer independent.

7.6.3 More on imputation

- Sometimes the imputed values are last year’s value. This is often called real donor even though it is a function of the auxiliary variables

- Another common classification is
  - Hot deck imputation - The value comes from the same sample S (e.g. Mean, nearest neighbour or a derived value from other answers at item non-response)
  - Cold deck imputation - The imputed value comes from an old or at least another data-set (e.g last year’s value or a value from a register like education or taxed income)

Legal aspects in Sweden

- You are not allowed to impute data values in Swedish registers for individuals.
- You are not allowed to deliberately include any fals values in personal databeses but
  - You may impute during the analysis phase
  - You may include new variables called derived values. And during the analysis tell the program to fetch that value if the correct one is missing and use it for imputation.
  - You may impute de-identified registers
  - You may impute in other registers like establishment registers

Imputation to the population level

- Impute not only all non-respondents in the sample – but all units in the population U (except of course those observed in R).
- May require too much computations.

7.7 NMAR – Modeling non-response

Apart from weighting and imputation there exist various methods.

- All of them depend heavily on modelling and the choosen model.
- Many of them work also in the NMAR case. Many different methods. Two examples:
- Some years ago MacFadden and Heckman received the Nobel price for such models. (e.g. they treated questions like what is the value of an academic education. Those who get an academic education have other positive properties which probably should have lead to income differences even without an academic education)

7.7.1 Another example, unemployed

- Two definitions:
  1. According to Statistics, Sweden (LFS-definition). Sample Survey ~20 000 persons per month (60 000 per quarter). (Today almost ~30 000 per month)
  2. Swedish Public Employment Service (AMS) has a register with all persons getting unemployment benefits or are looking for a job through them (compulsary for getting unemployment benefits)

MAR or NMAR?

- MAR
  - Non-response does not depend on LFS-status given AMS-status (poststratification, calibration, …)

    P(NR|LFS,AMS) = P(NR|AMS)

- NMAR, two out of many reasonable versions
  1. Treat the two unemployment definitions symmetrically. Assume that the sample is equally biased for both definitions (relative bias). (The bias for AMS-unemployment is known, since we have access to a total register)
  2. Nonresponse does not depend on AMS-status given the LFS-status (LFS-status is believed to be more central than the AMS-status. Inverse MAR)

     P(NR|LFS, AMS) = P(NR|LFS)

Data

- Macrodata from the Swedish Labour Force Survey, all four quarters 2008. (Non-response ~21 %, about ½ of it non-contact, refusals a little less)

- Auxiliary information AMS-status and one of the following
  - Age - gender (5 x 2 = 10 groups)
  - Country of birth (4 groups)
  - Industry (NACE) (7 groups + unknown)
  - Region (26 regions in Sweden)

Results, first quarter, 2008

using only AMS-status as auxiliary

Raw estimate	4.50
Poststratified	4.61
Same relative bias-corr.	4.80
Inverse MAR	4.94

- Unemployed out of the total studied population frame
- Sampling standard deviation 0.09. (But the random error of the differences is smaller)

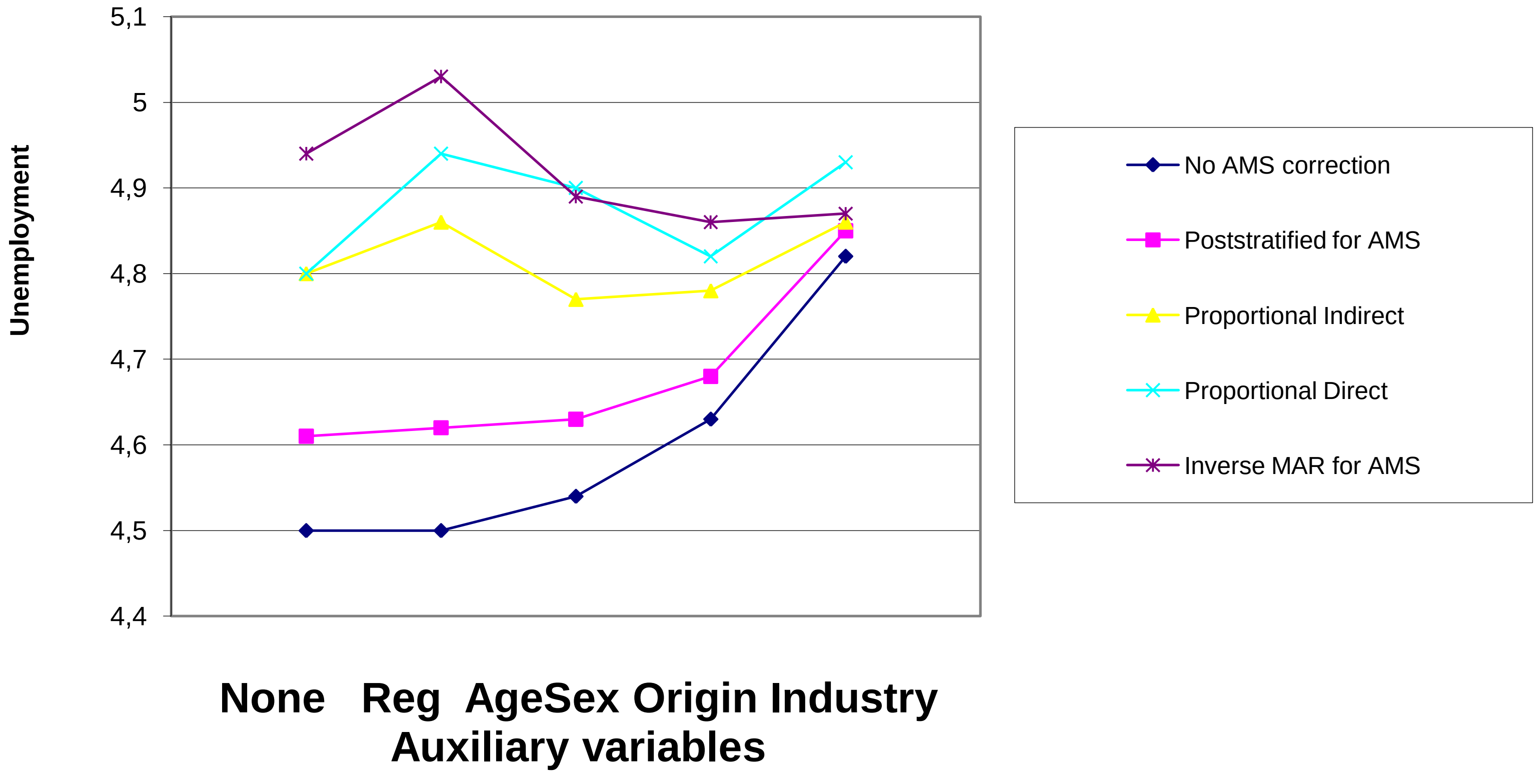
Results with more auxiliary information

First quarter 2008

Auxiliary data	Raw est	Post strat (Calibr)	Corrected estimate		Inverse MAR
			Indirect	Direct	
None	4.50	4.61	4.80		4.94
Age, Sex	4.54	4.63	4.77	4.90	4.89
Origin	4.63	4.68	4.78	4.82	4.86
Industry	4.82	4.85	4.86	4.93	4.87
Region	4.50	4.60	4.86	4.94	5.03

There is no known true value but the corrected estimate (assuming the same bias in both unemployment definitions) seems to be more stable. And the bias correcting effect of using more auxiliary information is not obvious

Figure A1a. Unemployment, different estimators, 1st quarter 2008



Results with auxiliary information

Differences between first quarter and year average, 2008

Auxiliary data	Raw est	Post strat (Calibr)	Corrected estimate		Inverse MAR
			Indirect	Direct	
None	0.06	0.04	-0.03		-0.09
Age, Sex	0.06	0.04	-0.02	-0.05	-0.09
Origin	0.06	0.05	-0.01	0.06	-0.06
Industry	0.05	0.02	-0.04	-0.08	-0.09
Region	0.06	0.05	-0.01	-0.04	-0.07

The assumptions on the non-response behaviour seems to be much more important for the estimate of changes than to use the auxiliary information. Changes in opposite direction

7.7.3 Conclusions

- NMAR is often a quite reasonable assumption. Different realistic assumptions on the non-response lead to quite different corrections.
- The choice of response models is often more important than to use the available auxiliary data in a MAR-setting like post stratification
- In NMAR, it is often difficult to get a full correction by increasing the amount of auxiliary variables.
- In quality declarations, the effects of some different but reasonable, correction methods should be given
- Methods developed for stock (level-) estimates are usually not suitable for estimates of change
- An old rule of thumbs: When you have applied your best (MAR-) correction method, about one quarter of the bias remains (In a variance sense)

Other methods

- Another form mail questionnaires:

- Sometimes you have the answers sorted after arrival date or number of reminders. A reasonable model may build a model
  - Where the response probability depends on the numbers of reminders
  - Where you assume that the non response is more similar to the late answers.
  - Where you estimate the response error by assuming that the non-response differs from the response by more than early and late responders.

- There are many studies indicating that uncertain or uninterested responders are overrepresented among late answers and in the non-response.

7.8. When is a sample representative?

- How to decide, from the data. People at the National Statistical Institutes have been asking that question
- My answer:
  - Take all interesting auxiliary variables that you have.
  - Estimate their totals directly or
  - Estimate their totals using the others as auxiliaries.

  Compute their squared relative biases (bias^2/Var). Do not assume that your error is smaller than the largest of these or
- Or be a Bayesian and assume that they are an iid sample from biases of all interesting variables. Use a χ^2-distribution.
- Those quantities answer questions on how skew an estimate may be (e.g. you may use 95% quantiles). With or without using the auxiliary information and assuming MAR

When is a sample representative?

- How to decide from data?
- R-indicators (Schouten) have also been suggested
- One version is based on
  $\mathrm{Var}(p_x{}^*(X)) = (1/n)\Sigma_{x€S}\ (p_x{}^*(X) – p)^2$
  (for simplicity formulated for SRS)
- e.g. R = 1 – Var/(p(1-p))
  (where p is the overall response rate)
  or 1 – 4Var or 1 – 2Dev

Missing values in general

There are many good methods to deal with data having missing data - but no perfect. It is quite natural since the missing data may be anything and you will never know.