Normalized weights: is using them enough?

Session Outline

- Main motivation behind using normalized (standardized) weights?
- Some problems associated with the use of weights
- How to compute normalized weights?
- Is it enough to use normalized weights?

Normalized weights: Is using them enough?

- Not so long ago, most statistical software programs that used a model-based approach did not offer the possibility of doing an analysis using a design-based approach.
- As a result, we were faced with the following choices:
  - Learn to use a new software
  - Program your own macros
  - Try to get the maximum out of the usual software (and accept the possibility of errors)

Normalized weights: Is using them enough?

- The use of normalized (standardized) weights is an attempt to make adjustments in order to continue using one's usual software.
- Normalized weights consider the survey weights, but not the other aspects of the design (stratification, cluster sampling, calibration, etc.). This is a modification of the model-based approach (to include weights) or an **incomplete** application of the design-based approach.

Normalized weights: Is using them enough?

- To ensure that the estimates of the
  - the
  - the
  - nbiased
  - ution,
  - a surve

Normalized weights: Is using them enough?

- The use of t eight with certain procedures i not designed for sample surve le, SAS or SPSS) can produce s lts.
- This is due to the fact that the software associates the sum of the weights with the number of observations at its disposal.

$\Rightarrow$ An overestimated statistical power!

Normalized weights: Is using them enough?

- Classic cases:
  - Independence test with SAS's PROC FREQ
  - Logistical regression with PROC LOGISTIC

Example from the NLSCY Cycle 6:

> Concepts will be covered in depth tomorrow, but suppose for now that we are interested in verifying whether the extent to which computers are used by teenagers is linked to the work/study situation of the parent(s)…

Normalized weights: Is using them enough?

- Results:
  - The SAS FREQ procedure with the *chisq* option give us a X^2 value of 8,929.7088 with an associated p-value of less than 0.0001.
  - From this, we should definitely conclude that the work/study situation of the parent(s) and the extent to which computers get used by teenagers are strongly linked.
  - Fortunately for us, before making this the headline of a report, we notice the following note in the output:

    Effective Sample Size = 1,816,357.2108

  - How do we adjust this? By using normalized weights!

Normalized weights: Is using them enough?

- What is a normalized weight?
  - It is a rescaled version of the survey weight.

  - The variable that contains the normalized weights has the following property: its sum corresponds to the exact number of units involved in the analysis. Therefore, the actual number of observations is closer to what it should be.

Normalized weights: Is using them enough?

- An example of normalization:

Identifier	Survey weight	Normalized weight
1	1.00	0.25
2	3.00	0.75
3	4.00	1.00
4	4.00	1.00
5	6.00	1.50
6	6.00	1.50
Total	24.00	6

Normalized weights: Is using them enough?

- How to normalize weights
  - Mathematically:
    - Simply divide the survey weight of each unit used in the analysis by the (unweighted) average of the survey weights of all the analyzed units.

$$w_k^{std} = \frac{w_k^{final}}{\overline{w}^{final}}$$

    - In the previous example, there are 6 observations and the sum of the survey weights is 24, making the average 4. Therefore, we divide each weight by 4.

Normalized weights: Is using them enough?

- How to normalize weights
  - Using a code similar to the following one will do the job in SAS:

```
proc sql;
   create table data2 as
   select *, finalweight/mean(finalweight) as stdweight
   from data
   where in_analysis=1;
/*Here, we suppose that the units that are part of the
   analysis have been flagged with in_analysis=1.*/
quit;
```

Normalized weights: Is using them enough?

- Is it enough to normalize?

  For complex surveys, the effective number of units is usually less than the number of observations in the sample. This is generally linked to the cluster effects (correlation between the observations of the same cluster) and sometimes also to stratification (ineffective stratification to ensure representativeness).

Normalized weights: Is using them enough?

- Is it enough to normalize?

In these cases, standardization results in:

- overestimation of the effective number of observations
- underestimation of variability
- too many significant results

Normalized weights: Is using them enough?

- Is it enough to normalize?

  To make corrections once again, some normalized weight users adopt a rule of thumb and use a more conservative significance level (1% instead of 5%) before declaring a significant result.

  However, this remains a rule of thumb. It is sometimes too strict and other times not strict enough…

Normalized weights: Is using them enough?

- Back to the example of the computer use and the work/study situation:
  - Result after standardization:
    - SAS: a X^2 value of 25.9481 ($p<0.0001$).
    - We would again conclude that they are strongly related, even if we were to adopt the 1% rule of thumb.

Normalized weights: Is using them enough?

- Example of the computer use and the work/study situation:
  - Result with a software program using a design-based approach:
    - SUDAAN: a X^2 value of 1.75 (p=0.1212).
    - Conclusion: The link is not significant.

Normalized weights: Is using them enough?

- Conclusion:
  - With **software programs that use a model-based approach, normalization is an attempt to salvage use of a certain number of procedures.**
  - It is an **incomplete application** of the design-based approach because it considers weights, but not the other aspects of the design.

Normalized weights: Is using them enough?

- Conclusion:
  - It generally results in **underestimation of the variance** of estimations and in too many results declared as significant.
  - Some users adopt a **rule of thumb** to try and compensate for this (or compensate based on the design effects). But this approach can **sometimes be too conservative, and other times not conservative enough**.

Normalized weights: Is using them enough?

- Conclusion:
  - If possible (a design-based approach has been developed for the analysis you are doing) and available (working within a RDC, not with a PUMF), bootstrap weights should be used.
  - Many softwares or sets of macros can help you deal with bootstrap weights: SUDAAN, STATA, WesVar, Bootvar…

Summary table of analysis tools based on the design, available in certain selected software programs

Software	SUDAAN 9.0	WesVar 4.2	Stata 9.0	Bootvar	SAS 9..1
Variance estimation method	**BRR (Bootstrap) Jackknife Taylor method**	**BRR (Bootstrap) Jackknife**	**BRR (Bootstrap) Jackknife Taylor method**	**Bootstrap (BRR)**	**Taylor method**
Modelling					
Linear regression	*proc regress*	Yes	*Svyreg*	*%regress*	*proc surveyreg*
Logistical regression	*proc logistic* (*rlogist*)	Yes	*Svylogit*	*%logreg*	*proc surveylogistic*
Generalized logistical models	*Proc multilog*	Yes	*Svymlog*	No	*proc surveylogistic*
Models of proportional odds	*Proc multilog*	No	*Svyolog*	No	*proc surveylogistic*
Poisson and log-linear models	*Proc loglink*	No	*Svypois*	No	No
Probit regression	No	No	*Svyprobt*	No	*proc surveylogistic*
Ordered probit regression	No	No	*Svyoprob*	No	*proc surveylogistic*
Proportional risk models	*proc survival*	No	No	No	No
Regression by instrumental variables	No	No	*Svyireg*	No	No
Regression by intervals	No	No	*Svyintrg*	No	No
Heckman models	No	No	*Svyheck*	No	No
Descriptive statistics					
Average	*Proc descript*	Yes	*Svymean*	*%ratio*	*proc surveymeans*
Totals	*proc descript*	Yes	*Svytotal*	*%total*	*proc surveymeans*
Proportions	*proc descript*	Yes	*Svyprop*	*%ratio*	*proc surveymeans*
Ratios	*proc ratio*	Yes	*Svyratio*	*%ratio*	*proc surveymeans*
Independence tests	*proc crosstab*	Yes	*Svytab*	*%chi2*	*proc surveyfreq*
Quantiles	*proc descript*	Yes	No	*%prcntle*	No
Plausible values/Multiple imputation	Some	Some	No	No	No

Normalized weights: Is using them enough?

- Conclusion:
  - NOTE: **With software programs that use a design-based approach, normalization is not necessary and could actually lead to errors (estimates of totals for example).**

Non-response Session

Dealing with non-response in NLSCY

Session Outline

- What is non-response?
- Types of non-response
- Why is non-response an issue for data analysts?
- Non-response and NLSCY
- Techniques for dealing with partial non-response (with examples)
- A quick wrap-up

What is non-response?

- Non-response is a situation that occurs when information from sampled units is unavailable.

- Data can be missing for some or for all questions.

Types of non-response

- Total non-response
- Wave non-response
- Partial non-response
  - Item
  - Component

Types of non-response

- Total non-response
  - No information is collected
  - Insufficient information is collected, so that the information collected is considered useless
- Wave non-response
  - Information about a respondent is available but not for every cycle, due to total non-response in a given cycle
    - NLSCY: mainly with the original cohort

Types of non-response

- Partial non-response (item)
  - Some individual questions are not answered
  - Some individual questions are not asked but should have been
- Partial non-response (component)
  - The NLSCY is divided into different groups of questions related to various topics; an entire section may be missing (e.g.: self-complete component)

Some reasons for total and wave non-response

- Some reasons related to the survey process
  - Timing
  - Poor frame information
  - Interviewer or field errors
  - Coverage and collection rules
- Some reasons related to circumstances
  - Weather
  - Language issues
  - Difficulty in tracing individuals
- Others reasons related to respondents
  - Unable to participate
  - Unwillingness to participate

Some reasons for partial non-response

- Refusal
- Don’t know
- Sensitive questions (skipped if respondent is uncomfortable)
- Respondent fatigue or time available
- Questions not asked by mistake

Why is non-response an issue for data analysts?

- A factor not taken into account in the derivation of a statistical result may cause a systematic distortion we call bias.
- It refers to the difference between the true value of the parameter of interest and the expected value of an estimator.

Why is non-response an issue for data analysts?

- Could non-response result in a possible bias?
  - Non-respondents often have characteristics that are different than those of the respondents. This can result in biased estimates if not corrected for (if ignored).
- Conclusion: possibly reporting erroneous results.

Non-response and NLSCY

- Total and wave non-response
  - Statistics Canada re-weights respondents to account for the non-respondents (see User Guide).
  - ***You don't have to do anything, just use the weights***
- Partial non-response
  - Statistics Canada imputes the income variables and a few other variables depending on the cycle (see User Guide).
  - ***Handling of other variables is up to you…***

Why is it up to you and not Statistics Canada?

- Treating all partial non-responses in NLSCY is not the best option for you nor for Statistics Canada.
  - Too many variables to treat (approx. 1,500 variables).
    - Would delay the release of the data.
    - NLSCY data users typically work with different subsets of variables.

Why is it up to you and not Statistics Canada?

- Non-response treatment often depends on the context
  - Different strategies may be valid.
  - Choice of strategy likely depends on the type of analysis conducted, the variables involved in the analysis and the domain of interest.
  - Knowing the specific context of their analysis, the NLSCY data users are better placed to determine the appropriate non-response treatment strategy to adopt.

Why is it up to you and not Statistics Canada?

- Example #1:
  - A colleague of yours has performed imputation of missing math scores using the mean score of all respondents.
  - You are interested in math scores, yes, but even more in scores within the different types of schools.
  - Results could potentially be very different than if the imputation had been performed within the classification considered.

Why is it up to you and not Statistics Canada?

- Example #2:
  - An analyst is interested in measuring the correlation between two variables (say PMK depression score and kid anxiety score), each of them including missing values.
  - Imputing variables separately could potentially distort the correlation between these two variables.

Why is it up to you and not Statistics Canada?

- Answers to the Why:
  - It is not possible to envision all relevant classifications of interest.
  - It is not really feasible to foresee all possible pairings (groupings) of variables.
  - And even if it was, that would mean releasing multiple times the same variable, but with different imputed values each time. Imagine the size of the files…

Partial non-response in NLSCY

- Missing data for variables are identified as:

	Values
Don’t know	7, 97, 997…
Refusal	8, 98, 998…
Not stated	9, 99, 999…

- Note: These are different from:

	Values
Not applicable or valid skip	6, 96, 996…

Not Applicable / Valid Skip

- Related to the coverage of the question
  - Identify people to whom the question doesn't apply.
    - Codebook
    - Questionnaire
    - Data file
- Usually excluded from the analysis right from the start.

Not Applicable / Valid Skip

- Example: Cycle 6 Self-Complete
  - Question: During the past 12 months, how often have you volunteered or helped without pay?
  - Coverage: Respondents 12 to 15 years old who have volunteered during the past 12 months.
  - File contains all 10-17 year olds.
  - Not applicable / valid skip:
    - 10-11, 16-17 year olds
    - 12-15 year olds who have not volunteered in the past 12 months

Techniques for dealing with partial non-response

What are your options?

- The question is…
  - How to make inferences when there are missing data for the sampled units?
- Things to consider:
  - Extent of non-response
  - Type of data (continuous vs categorical)
  - Number of variables of interest
  - Type of analysis (descriptive vs analytical)
  - Analysis context

What are your options?

- You basically have 5 options:
  a) Ignore partial non-response (only use records with a response)
  b) Report partial non-response as a category
  c) Profile the partial non-respondents
  d) Re-weight the records that have a response to account for the partial non-respondents
  e) Impute partial non-response (replace missing values with plausible values)

Option a) Ignore partial non-response

- Simply eliminate partial non-respondent records.
  - Throws away information (answered questions), especially in modeling procedures.
- Inferences to the whole population assume that response is missing completely at random (MCAR) (i.e. non-response does not depend on any covariate nor on the variable of interest, or said differently, that non-respondents are similar in every way to respondents).
  - Results in biased inferences if not true (rarely true…).

Option a) Ignore partial non-response

- Limits types of inference (models, means, proportions, but not totals (at least directly)).

- More viable if the extent of non-response is low.
  - Sometimes the only option when the non-response is very low (because it is impossible to know enough about the non-respondents to profile them).

Option a) Ignore partial non-response

- Inferences only to the subpopulation represented by the respondents can be made.
  - Interpretation not clear (abstract concept)
    - This “subpopulation of respondents” likely does not exist in reality…
    - Are such inferences meaningful???

Option b) Report partial non-response as a category

- Report missing values (partial non-response) as a valid category in tables or in models.
  - It complicates the interpretation of the numbers showing up in the table but gives an indication of the quality of the data and of the possible shifts in the observed values.

	Low	Medium	High	Non-Response
Example 1	10%	30%	50%	10%
Example 2	18%	20%	22%	40%

Option b) Report partial non-response as a category

- Applies only to categorical data.
- Allows inferences to the whole population.
- May reduce the power to detect certain effects, since it splits cases that in fact may be associated with the same level of a given variable.

Option c) Profile the partial non-respondents

- This is actually the first step in applying option d).
- Consider the partial non-respondents of a set of given questions as a sub-population of interest.
- Consider the full respondents of the same set of questions as the other sub-population of interest.

Option c) Profile the partial non-respondents

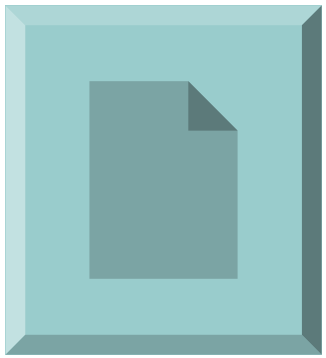
- Determine how these sub-populations differ with respect to other key/related variables to which both groups responded.
  1) Variables related to the **response status**
  2) Variables related to the **values themselves** of the variable(s) with non-response that we want to treat
     - Not practical when a lot of variables to analyze

Option c) Profile the partial non-respondents

- Report on your findings, so that it gives an indication to the reader about the possible sources of bias in the reported results.
- Inference to the whole population? Not really, but could be used as a conditional one:
  - “If there was no non-response bias, here is what would be reported. And here are the most likely sources of non-response bias.”
- Compared to a), the reader is more informed of how the numbers could change.

Option c) Profile the partial non-respondents

- There is an example, applying this option, that is provided in the Cycle 6 user guide:



Option d) Re-weighting

- Ignore partial non-respondent records.

- BUT the weights of the respondents are increased to account for non-respondents.
- Groups based on the variables identified in option c) are formed.
- There are several techniques (Cross-classification, Scoring method, CHAID algorithm…) to form these groups, and some constraints are usually imposed (number of groups, response level in each groups…)

Option d) Re-weighting

- Weights of the respondents in each ‘cell’ are increased by a factor corresponding to the inverse of the weighted response rate within the cell.
- Example:

Group (age by gender by LICO situation)	Weighted Response Rate	Adjustment Factor
10 year-old girl below LICO	0.80	1.25

Option d) Re-weighting

- Used for total and wave non-response (case for NLSCY).
- Interesting for component non-response (missing a whole section of a questionnaire: similar to total NR).
- Less appealing for item non-response.
  - Throws away answered questions (only “total respondents” are kept in the analysis)
- But still easier to apply…

Option d) Re-weighting

- Allows inferences to the whole population.
- Necessitates that a similar redistribution of weights is done with each of the bootstrap replicate.
- Does not require any other adjustment for the computation of a proper design-based variance.

Option d) Re-weighting

- Note regarding control totals…
  - The final NLSCY weights are adjusted so that NLSCY population totals agree with official totals at Statistics Canada (see post-stratification in User Guide).
    - Weight adjustments are made for each age-gender-province combination.
  - When re-weighting to adjust for partial NR, control totals are not respected anymore.

Option d) Re-weighting

- Concerned about control totals not being respected anymore?
  - Re-compute the age-gender-province weight adjustments (post-stratification).
    - Extra work…
  - Or consider treating partial NR with imputation (next option) instead of re-weighting.

Option e) Imputation

- Replace missing values with plausible ones.
  - May be seen as micro level estimation
  - Several approaches & methods available
- Allows inferences to the whole population.
- Usually for item non-response (case for income variables in NLSCY).
  - More appealing than re-weighting since we keep questions with responses (fill in the gaps).

Option e) Imputation

- Imputation can artificially reduce the estimated variance.
- Typical variance approaches should require some adjustments in order to compute a proper design-based variance.
- Depending on the approach used (multiple imputation or single imputation), commonly-used software may or may not allow the user to make these required adjustments.
- Important to report imputation with the results (methods used, imputation rate, etc.).

A general recommendation

- We can never be certain that non-response bias is totally eliminated with a non-response adjustment.
  - We may not have considered all the variables related to non-response in our adjustment.
    - Some of these variables may not even be on the file.

A general recommendation

- We can never be certain that non-response bias is totally eliminated with a non-response adjustment.
  - Non-response may be related to the analysis variable: data is not missing at random (NMAR)
    - Cannot completely adjust for non-response.
    - Can partly adjust for non-response if it is also related to known variables (e.g. socio-demographic variables).

A general recommendation

- Our goal should be to reduce the non-response bias rather than to eliminate it.
  - Any remaining bias that we know of or suspect should be reported with the analysis.

Example 1: The population and sample

- A school of 50 children, 80% of whom are 15 years old.

- For a survey, a simple random sample (SRS) of 5 kids was drawn from the school.

Example 1: The data set and the analysis question

- One question in the survey asks 15-year-olds if they smoke cigarettes.
- Want to estimate the number of 15 years old who smoke.

Child	Age	Gender	Smoker	Weight
1	15	F	1	10
2	15	M	0	10
3	15	F	1	10
4	16	M	6	10
5	15	F	9	10

1: means the child smokes
0: means the child does not smoke
6: Not applicable (e.g., child is not 15 years old)
9: Not-stated

Example 1: Ignoring non-respondents

- Want to estimate the number of 15-year-olds who smoke.
  - Remove Not-applicable (6) cases
  - Deal with the non-response (9)
    - Option a) Ignore the non-respondent

Child	Age	Gender	Smoker	Weight
1	15	F	1	10
2	15	M	0	10
3	15	F	1	10
4	16	M	6	10
5	15	F	9	10

For the subpopulation of 30 kids represented by those who responded, the estimated number of 15-year-olds who smoke is

20

Example 1: Ignoring non-respondents

- For the subpopulation of 30 kids represented by those who responded, the estimated number of 15-year-olds who smoke is **20.**

- Notice that because we are estimating a total, if we ignore non-respondents then we can only make inferences about the subpopulation

- If we were estimating a proportion, we could justify making inferences about the entire population if at least one of the following holds:
  - the non-response rate was very low

  OR
  - the response was missing completely at random (MCAR)

Example 1: Reporting non-response as a valid category

- Want to estimate the number of 15-year-olds who smoke.
  - Remove Not-applicable (6) cases
  - Deal with the non-response (9)
    - Option b) Report NR as a valid category

The distribution of the smoking status for 15-year-olds in the school is …

Child	Age	Gender	Smoker	Weight
1	15	F	1	10
2	15	M	0	10
3	15	F	1	10
4	16	M	6	10
5	15	F	9	10

Smoke	Don't smoke	Unknown
20	10	10
50%	25%	25%

Example 1: The population and sample

- Option c) Profiling the non-response

  Here, there is a single non-respondent, and all we know about it, is that it is a girl. But, if this was done on a larger scale, the analyst could report something like: “Non-response was evaluated, and it appeared to be mostly related to the following variables: gender… Results were not controlled for these differences and could therefore include some non-response bias.

Child	Age	Gender	Smoker	Weight
1	15	F	1	10
2	15	M	0	10
3	15	F	1	10
4	16	M	6	10
5	15	F	9	10

Example 1: Re-weighting

- Want to estimate the number of 15-year-olds who smoke.
  - Remove Not-applicable (6) cases
  - Deal with the non-response (9)
    - Approach d): Re-weight to adjust for non-response using gender groupings

Child	Age	Sex	Smoker	Weight	
1	15	F	1	10	*30/20
2	15	M	0	10	*10/10
3	15	F	1	10	*30/20
4	16	M	6	10	
5	15	F	9	10	

After re-weighting to compensate for non-response, the estimated number of 15-year-olds in the school who smoke is

30

Example 1: Imputation

- Want to estimate the number of 15-year-olds who smoke.
  - Remove Not-applicable (6) cases
  - Deal with the non-response (9)
    - Option e) Impute the missing value
      - We will impute based on gender (it means we impute a 1 for F and 0 for M)

Child	Age	Gender	Smoker	Weight
1	15	F	1	10
2	15	M	0	10
3	15	F	1	10
4	16	M	6	10
5	15	F	~~9~~1	10

After imputing for non-respondents, the estimated number of 15-year-olds in the school who smoke is

30.

A quick wrap-up

- Assess the impact of partial non-response in your analysis.
  - Extent of partial non-response
  - Variables explaining the response status
- Then, take appropriate actions.
  - Little non-response, no variables explaining non-response, or making inferences to the subpopulation represented by the respondents
    - You may consider ignoring partial non-response

A quick wrap-up

- Otherwise, your partial non-response treatment options are…
  - b) Reporting partial non-response as a category
    - ☺ Quick and simple
    - ☹ Maybe difficult to interpret
    - ☹ Does not apply to all analysis
  - c) Profiling the non-response
    - ☺ More work than a), but more informative
    - ☹ Less work than d) or e) but doesn't really allow inference to the whole population

A quick wrap-up

- Your partial non-response treatment options are…
  - d) Re-weighting
    - 😐 Relatively simple but more work
    - ☹ Throws away responded items
    - ☹ Control totals not respected anymore
  - e) Imputation
    - ☺ Takes advantage of all responded items
    - ☹ More complex and more work
    - ☹ Risk of generating incoherence at the record level

A quick wrap-up

- Report your partial non-response treatment strategy:
  - Chosen strategy with justifications (ignoring partial non-response is in fact a strategy)
  - Part of the scientific process (other researchers can reproduce your results)

A quick wrap-up

- Report your partial non-response treatment strategy:
  - Be prepared to provide the reader with any relevant information
    - Response and imputation rates
    - Variables considered for potential non-response bias
    - Construction and size of re-weighting or imputation groups
    - etc.

A quick wrap-up

- Is it more work to adjust for partial non-response in the analysis than to ignore it?
  - Yes… but your analysis is more reliable (non-response bias is addressed) and practical (conclusions apply to the whole population)
- Remember, conclusions obtained after ignoring partial non-response apply to the *subpopulation represented by the respondents* (abstract concept)
  - Unless you are able to show that the response is missing completely at random…

Combining multiple cohorts within NLSCY

Session Outline

- Why do people consider combining?
- When is it feasible to combine?
- The most typical approach
  - ‘Pooling approach’
- Some examples

Why do people consider combining?

- Sample size provided by a single cohort is too small.
  - Estimates are not publishable, based on the CV criteria required for the survey
  - Wish to add to the statistical power in order to possibly better detect some effects
- Feeling that the quantities of interest are stable over time.

Why do people consider combining?

- Example:
  - If the objective is to describe some aspects of the life of 1-year-old kids living in households below the LICO (say the proportion that have a low motor social development score), then it is likely that the sample available from individual cohorts would be too small.

When is it feasible to combine?

- If there is a link between the researcher's target population and the survey populations being combined:
  - If the population of interest is all of the survey populations being combined.
  - If each of the survey populations can be thought to represent the population of interest.
  - If none of the survey populations are representative, but a summary measure from these populations is meaningful.

When is it feasible to combine?

- If you believe that the model you are suggesting could have generated the data for each of the survey populations.
  - Typically, the model will include finite-population specific parameters (at least initially).

The most typical approach

- The ‘pooling approach’ steps:
  1. Pool data from the different cohorts into a single file.
  2. Create a weight variable appropriate for the pooled data, your target population and the quantities of interest.
  3. Create new bootstrap weights according to the strategy opted for in 2.

The most typical approach

- The ‘pooling approach’ steps:
  4. Validate the assumptions that quantities of interest are the same in the different cohorts being combined.
  5. Carry out estimation and inference on the pooled data using techniques that would be appropriate for data from a single sample.

The most typical approach

- The ‘pooling approach’:

  Notes:

  - Pooling of samples that are not independent has additional complications.
    - If the same child (PERSRUK) happens to be part of the two samples being combined, then the samples are definitely not independent.

The most typical approach

- The ‘pooling approach’:

  Notes:

  - There are a variety of ways to create a weight variable and bootstrap weights in the ‘pooling approach’. In many cases, pooling the weight variables from the different cohorts, without adjusting, is a good choice.
    - Among things to consider:
      - Respective sample size available for each cohort and respective target population sizes
      - Are all cases from each cohort used or are some cases put aside?

Some examples

- First example:
  - The focus of the analysis will be babies' sleep. We will restrict the analysis to 0-year-olds (not 1-year-olds) to minimize the impact of recall errors.
  - The 'sleep' questions are quite new. Most of them have only been asked since Cycle 4, and some of them only since Cycle 5.
  - We should check the literature for any evidence of a change in recommended babies sleeping positions between 2002 and 2004.

Some examples

- First example:

  We will define the target population as the population of 0-year-olds who are born between 2002 and 2004.

  We could build ourselves a bigger sample by combining the 0-year-olds from Cycle 5 and the ones from Cycle 6.

  The sample sizes from each cycle are very similar, and so are the survey population sizes. All 0-year-olds from both cohorts will be part of the analysis.

Some examples

- First example:

  So here, we could simply use the cross-sectional weights and the cross-sectional bootstrap weights associated with each unit at the time they are 0 year old.

  Even if no evidence of changes in recommendations to parents was found, we should still perform some work to check that quantities of interest are measuring the same thing in the two cohorts combined.

Some examples

- Second example:

  The focus of the analysis is babies that were breast fed for more than 6 months and when they reached certain developmental milestones. We want to compare them to other breast fed babies and to other babies in general.

  The target population will be kids that had their first birthday between 2000 and 2002. But we will take a look at these kids once they are 3 years old, since this gives them the time to reach several developmental milestones. It is felt that the recommended practices for feeding babies were relatively constant over the 1999-2001 period.

Some examples

- Second example:

  We will therefore combine the 3-year-olds from Cycle 5 and Cycle 6.

  Sample sizes are relatively homogeneous, target population sizes are also homogeneous. All longitudinal 3-year-olds from both cohorts will be part of the analysis.

  We could simply use the longitudinal weights and the longitudinal bootstrap weights associated with each unit at the time they are 3 years old.

Some examples

- Second example:

  Let's assume that the sample size is still not large enough. How could we have obtained more cases?

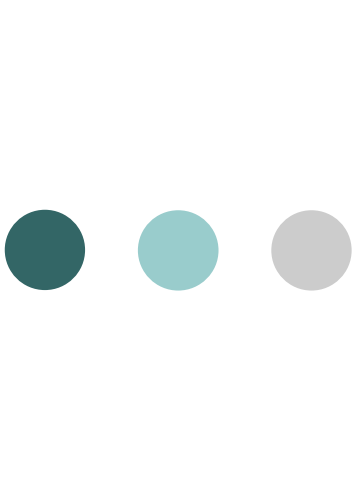
  - Taking younger kids as well (2-year-olds) may have led to censoring issues (some kids wouldn't have reached the milestones yet).

Some examples

- Second example:

  Let's assume that the sample size is still not large enough. How could we have obtained more cases?

  - Taking 3-year-olds from Cycle 4, or even 4-year-olds from Cycles 5 and 6, could have been done, but the differences in the respective sample sizes would likely have necessitated creating new weights. We would also need to check the stability of the 'environment' prior to 1999.



A full example

A full example

- We have selected the topic of our next paper: “Use of computers by teenagers: who are the ‘heavy’ users?”
- We have done a literature review and have identified through it a list of characteristics / factors that would apparently be linked to the extent to which computers are used by teenagers.
- We would like to use Canadian survey data to put to the test these theories / beliefs.

A full example

- We have identified NLSCY as a possible source of Canadian data to answer this question.
- We have read the NLSCY documentation. We are now all set to perform the analysis.
- We will now go through some of the steps we would have to deal with.

A full example

- Here are some of the objectives we had in mind for the paper:
  - To provide the reader with a snapshot of the extent to which teenagers use computers nowadays.
  - To compare current use to past use of computers by teenagers.
  - To identify some factors linked to the extensive use of computer by teenagers.

A full example

- As part of the Cycle 6 self-complete component, a component for children 10 to 17 years old, there is a section on activities.
- One of the questions asked is the following:

  FATCeQ21

  - On average, about how many hours a day do you spend on a computer (doing work, playing games, e-mailing, chatting, surfing the Internet, etc.)?

A full example

- Notice that the name of the variable (**FATCeQ21**) includes a lower case ‘e’ as the fifth character.
  - This lower case letter refers to the NLSCY cycle (‘e’=5, ‘f’=6, ‘d’=4, etc.) in which the variable first appeared on the file or the cycle in which changes to a previously asked question were made.

A full example

- In this case, consulting the Cycle 4 and Cycle 5 codebooks would allow us to confirm that this is a new question, used only since Cycle 5.

- So, the first constraint imposed by the data available on our first objective is that it really won't be possible to go back in the past, at least not more than 2 years.

A full example

- Here are the responses to the question as reported in the Cycle 6 codebook for the 10-17:

Value	Labels	Freq	Wtd
01	I don't use a computer	54	18,782
02	Less than 1 hour a day	2,345	722,791
03	1 or 2 hours a day	2,177	784,766
04	3 or 4 hours a day	764	291,690
05	5 or 6 hours a day	180	72,835
06	7 or more hours a day	78	25,978
96	Valid skip / Not Applicable	1,331	611,323
99	Not stated	1,267	512,195
		======	=========
		8,196	3,040,360

Coverage: Respondents 10 to 15 years old

A full example

- Notice the ‘Coverage’ statement in the codebook, below any table reporting the results associated with a given variable. In the case of **FATCeQ21**, it states:

  Coverage: Respondents 10 to 15 years old

- So, a second constraint imposed by the data available on our first objective is that our definition of a teenager will have to be limited to, at most, children between 10 and 15 years old.

A full example

- At the very end of the Cycle 6 codebook for the 10-17, just before the index, there is a section that describes the weights that are available to you. More information about these weights can also be found in the User Guide.
- Notice that only two types of weights are available:
  - FWTCW01L: Child longitudinal weight
  - FWTCWd1L: Child longitudinal all cycles (funnel) weights

A full example

- Since a cross-sectional weight is not available, but rather only longitudinal weights, it means that it will not be possible to provide a snapshot of the extent to which teenagers of 10-15 years old in 2004 used computers.
- A cross-sectional weight is not provided with the original cohort since it has been established that the population of children has changed too much, since the sample was drawn in 1994, to be considered representative of the 2004 population of children.

A full example

- Therefore, all inferences will need to be made with respect to the longitudinal population.

- In our case, it means that conclusions will hold only for the population of children aged 0 to 5 in 1994, as they are when they reach 10-15 years old in 2004.

A full example

- Let’s now go back to the responses for the question as reported by the Cycle 6 codebook for the 10-17:

Value	Labels	Freq	Wtd
01	I don't use a computer	54	18,782
02	Less than 1 hour a day	2,345	722,791
03	1 or 2 hours a day	2,177	784,766
04	3 or 4 hours a day	764	291,690
05	5 or 6 hours a day	180	72,835
06	7 or more hours a day	78	25,978
96	Valid skip / Not Applicable	1,331	611,323
99	Not stated	1,267	512,195
		======	=========
		8,196	3,040,360

Coverage: Respondents 10 to 15 years old

A full example

- We can note a few things:

  - Since the file contains information from the 10-17 year-olds, and the coverage for the question is 10-15 year-olds, there are a certain number (1,331) of Not Applicable (96) responses to this question. This should correspond to the number of children that are 16-17 years old and these records should be removed from the analysis file.

A full example

- We can note a few things:
  - If we had more time to look at the data, we might notice that the number (1,585) of children 16-17 years old do not match the count (1,331) of Not Applicable for this question. Actually, all Not Applicable counts are erroneous on the file. This is due to a (Cycle 6 only) processing error. Cases for which there is no self-complete data at all were assigned a Not Stated code, regardless of the coverage of the question.

A full example

- We can note a few things:
  - Therefore, when using the Cycle 6 file for the 10-17, the removal of the ‘Not Applicable’ should always be based on the coverage statement, and not on the values taken by the variable of interest.

A full example

- An additional note:
  - We would also likely want to remove children that are longitudinally in scope, but for whom data could not be collected. This includes children who have died or who no longer reside in Canada.
  - These people can be identified by the variable FLWTCD on the LONG file.
  - In the case of an analysis using the file for the 10-17, these records were already removed from the file, so this additional step may be disregarded.

A full example

- Once all Not applicable cases have been removed, the responses look like this:

Value	Labels	Freq	Wtd
01	I don't use a computer	54	18,782
02	Less than 1 hour a day	2,345	722,791
03	1 or 2 hours a day	2,177	784,766
04	3 or 4 hours a day	764	291,690
05	5 or 6 hours a day	180	72,835
06	7 or more hours a day	78	25,978
99	Not stated	1,013	373,322
		======	=========
		6,611	2,290,163

A full example

- We can note a few things:
  - There is a certain amount of non-response (Not stated ’99’). It represents about 15% of the responses.
  - We will go over how to deal with those soon.
  - But first, let’s have a look at the unweighted and the weighted distribution of responses.

A full example

	Unweighted		Weighted	
Values to **FATCeQ21**	Freq.	Perc. (%)	Freq.	Perc. (%)
I don't use a computer	54	0.82	18,782.5	0.82
Less than 1 hour a day	2,345	35.47	722,790.7	31.56
1 to 2 hours a day	2,177	32.93	784,765.5	34.27
3 to 4 hours a day	764	11.56	291,689.6	12.74
5 to 6 hours a day	180	2.72	72,834.7	3.18
7 or more hours a day	78	1.18	25,978.3	1.13
Not stated	1,013	15.32	373,321.8	16.30

A full example

- Comparing the two tables:
  - The weighting modifies the frequency distribution. This is to correct for the distortion inherent to the sample that Claude talked about yesterday.
  - Note that the most frequent category is now 1-2 hours a day, and not less than 1 hour a day.

A full example

- Comparing the two tables:
  - Also note that the weighting has an impact on the frequency of the ‘Not stated’. This is a sign that the non-response is at least partially related to some aspects of the design.
  - Among the variables considered for your non-response treatment strategy, you should include the design variables that appear to be related to the non-response (and or to the values of the variables).

A full example

- Comparing the two tables:
  - What could partly explain the differences in the sample distribution (unweighted distribution) and the population distribution (weighted distribution)?
  - In order to answer this question, you should explore the different distortions of the sample and try to find one (or a few) that is (are) related to the variable of interest.

A full example

- Comparing the two tables:
  - The most common/important distortions are:
    - Typically, Atlantic provinces are over-represented, as are Manitoba and Saskatchewan. Quebec and Ontario are under-represented.
    - Rural and smaller urban areas are over-represented. CMAs/CAs are under-represented.
    - With respect to the original cohort, older kids are under-represented and younger kids are over-represented.

A full example

- Comparing the two tables:
  - For this analysis, age being strongly linked to the use of computers by children, it is likely that this is the main source of distortion.
  - Rural / Urban residence is definitely another source of distortion.

A full example

- We will set aside for a while our first objective and focus more on our third objective, identifying some factors linked to the ‘heavy use’ of computers by teenagers.
  - We have seen before that our definition of teenagers will have to be limited to children 10-15 years old.
  - Therefore, we should be looking for possible factors that are available at the same time for all children from that age group.

A full example

- It is not impossible to consider factors that are not available at a given time for all children.
- The longitudinal survey makes it possible to go back in the past to get some information at a different time for part of the sample, but this type of information will likely have some drawbacks.
  - It would possibly require that such a factor is stable in time and that its effect over time remains unchanged.
  - These two assumptions are quite restrictive and likely rarely met in practice.

A full example

- For modeling purposes, one of the first things we should do is define what ‘heavy use’ is.
- Of course, this definition will need to be based on the data available and on existing measures present in the literature.
- The results will likely be strongly linked to how the different concepts were defined, so it is very important to state the definitions clearly.

A full example

- In our case, we will be using the following concepts:
  - We will define ‘heavy’ users as children using a computer 3 hours or more a day.
  - We will define the ‘light’ users as children using a computer at most 2 hours a day.
  - We will put aside the ‘none’ users, since they are likely to have a different profile. Once we are done with our treatment of non-response, we will treat them as ‘Not Applicable’ and remove them from our analysis.

A full example

- Similar work will need to be done with each of the potential factors identified by the literature review. Here, considerations could be given to the model fit or the data available in assessing how certain factors should be used.

- In addition, some of the potential factors identified may not be directly available in the survey. A possible option is then to try to identify available variables that could serve as ‘proxies’ for these unavailable factors.

A full example

- Suppose that after an extensive literature review (including Lafortune (2008), Lafortune (2008) and Lafortune (2008)), the following factors have been identified as being associated with the extent to which children are using computers:
  - Age
  - Gender
  - Computer available in the house and at school
  - Amount of spare time parents have to be with their kids
  - Parental monitoring

A full example

- Within NLSCY Cycle 6 data, we have variables for most of these potential factors:
  - **FMMCQ01** (Age of child)
  - **FMMCQ02** (Gender of child)
  - **FATCeQ22** (Is there a computer in your home?)
  - **FPMCcS3** (Parental monitoring score)

- We would like to find proxies for **Computers available in school** and **Amount of spare time parents have to be with their kids**.

A full example

- Proxies will never be as good as the desired variables, and this should always be kept in mind when reporting results.
- As proxies for **Computers available in school** and **Amount of spare time parents have to be with their kids**, we will use:
  - **FEDCbQ0** (What type of school is this child currently in?)
  - **FLFHD49B** (Current work/study situation of PMK and spouse)

A full example

Variable name	Variable type
Age of child	continuous(?), 10-15
Gender of child	2-level categorical
Computer in the home	2-level categorical
Parental monitoring	continuous, 0-20
Type of school	7-level categorical
Work/study situation of parents	6-level categorical

Plus missing values

A full example

- A decision regarding how to use those variables is then needed.
  - Should a continuous variable be used as is, or should it be categorized?
  - Should some of the categories be collapsed?
  - How to determine a cut-off point?

- These won't be discussed at length here, but this should be part of every analysis.

A full example

- In our case, we will group together categories with very low frequencies for **FEDCbQ0**.

Type of school currently in?

FEDCbQ0	Frequency	Percent	Cumulative Frequency	Cumulative Percent

This information had to be suppressed…

A full example

- Monitoring has been shown to be related to child outcomes – there is a correlation between a lack of supervision and negative outcomes.
- We will therefore group the most at-risk children together and dichotomize the scores of the parental monitoring scale (**FPMCcS3**), using the first decile (that is the lowest 10%, the lowest monitored kids) as the cut-off point.

A full example

- Once the core set of variables has been identified, the extent of non-reponse can be assessed and a strategy for non-response treatment can be adopted.
- There are 5,328 records with a full response. Some of these records will be put aside (the ‘none’ users) once the non-respondents have been dealt with.
- There are about 1,300 ‘partial non-response’ cases.

A full example

- Non-response is spread out in the following way:
  - 1,013 records with no information on computer use (and for most of them, little other information: in fact for 618 records, there is no information from the child).
  - 211 additional cases with parental monitoring missing (but with information on computer use)

A full example

- Non-response is spread out in the following way:
  - 53 cases have missing information on the work/study situation, with information on the other variables.
  - Some information had to be suppressed…

A full example

- Non-response affects all of the variables of interest for our research question

Variable	Records with a missing value
Computer use	1,013
Parental monitoring	+ 211
Work/study situation	+ 53
Some information had to be suppressed…	
Total	

A full example

- Option a) Ignore partial non-response
  - Here, we consider only the respondents to the full set of questions that are used, without any adjustments (5,278 children).

A full example

- Option a) Ignore partial non-response

hours a day do you spend on a computer

FATCeQ21	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Less than 1 hour a day	684216.4	37.67	684216.4	37.67
1 or 2 hours a day	753541.7	41.49	1437758	79.16
3 or 4 hours a day	283800.3	15.62	1721558	94.78
5 or 6 hours a day	69954.27	3.85	1791513	98.63
7 or more hours a day	24844.51	1.37	1816357	100.00

Can't make inference about totals directly

Valid only if response is MCAR

A full example

- Option a) Ignore partial non-response

Variance as measured by using normalized weights in SAS

Selected Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
FMMCQ01	10 vs 15	0.209	0.157	0.278
FMMCQ01	11 vs 15	0.415	0.328	0.524
FMMCQ01	12 vs 15	0.545	0.437	0.679
FMMCQ01	13 vs 15	0.554	0.447	0.687
FMMCQ01	14 vs 15	0.880	0.718	1.077
computer_inhouse	0 vs 1	0.139	0.062	0.312
FLFHD49B	1 vs 6	1.033	0.694	1.539
FLFHD49B	2 vs 6	0.982	0.557	1.730
FLFHD49B	3 vs 6	0.984	0.631	1.532
FLFHD49B	4 vs 6	0.756	0.335	1.707
FLFHD49B	5 vs 6	1.500	0.988	2.277
parentalmonitoring	0 vs 1	1.395	1.118	1.742

Valid only if response is MCAR

A full example

- Option a) Ignore partial non-response

Variance as measured by SUDAAN using the bootstrap weights

Selected Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
FMMCQ01	10 vs 15	0.21	0.14	0.31
FMMCQ01	11 vs 15	0.41	0.29	0.60
FMMCQ01	12 vs 15	0.54	0.37	0.81
FMMCQ01	13 vs 15	0.55	0.38	0.82
FMMCQ01	14 vs 15	0.88	0.61	1.27
computer_inhouse	0 vs 1	0.14	0.05	0.37
FLFHD49B	1 vs 6	1.03	0.50	2.13
FLFHD49B	2 vs 6	0.98	0.38	2.53
FLFHD49B	3 vs 6	0.98	0.43	2.24
FLFHD49B	4 vs 6	0.76	0.19	2.95
FLFHD49B	5 vs 6	1.50	0.71	3.17
parentalmonitoring	0 vs 1	1.40	0.98	1.99

Valid only if response is MCAR

A full example

- Option a) Results with survey weights, normalized weights and bootstrap weights

Main Effects	Survey weights p-values	Normalized weights p-values	Bootstrap weights p-values
Age	<0.0001	<0.0001	0.0000
Gender	<0.0001	0.1801	0.4291
Type of school	<0.0001	0.0699	0.5495
Computer in the house	<0.0001	<0.0001	0.0000
Work/study	<0.0001	0.0014	0.1680
Parental monitoring	<0.0001	<0.0001	0.0127

A full example

- Option b) Report partial non-response as a category

  - Here, we use all children that are applicable (in our case, 10-15 year-olds for a total of 6,611 children).

A full example

- Option b) Report partial non-response as a category

Need to include these cases, since the Not stated could go there

hours a day do you spend on a computer

FATCeQ21	Frequency	Percent	Cumulative Frequency	Cumulative Percent
I don't use a computer	18782.46	0.82	18782.46	0.82
Less than 1 hour a day	722790.7	31.56	741573.2	32.38
1 or 2 hours a day	784765.5	34.27	1526339	66.65
3 or 4 hours a day	291689.6	12.74	1818028	79.38
5 or 6 hours a day	72834.72	3.18	1890863	82.56
7 or more hours a day	25978.27	1.13	1916841	83.70
Not stated	373321.8	16.30	2290163	100.00

Allows the reader to evaluate worst case scenarios

Difficult to interpret the percentages

A full example

- Option b) Report partial non-response as a category

  - We can't really perform the logistic regression model as we did with option a).
    - For non-response to explanatory variables, we could model by simply adding a 'missing' category (when there are enough cases).
    - But for non-response to the dependent variable, we would need to use a multi-logistic regression model.

A full example

- Option c) Profile the partial non-respondents

  - Here, we want to further explore who those 1,300 partial non-respondents are and try to identify some of their characteristics.
  - We should also look at patterns among the different classes of partial non-respondents (no children data, no computer use answer, other types of partial non-response) to see if the same strategy can apply to all.

A full example

- Option c) Profile the partial non-respondents

Type of partial non-response BY Age of child

Row Pct	10 Years	11 Years	12 Years	13 Years	14 Years	15 Years
No children data	16.45	16.49	17.04	14.79	18.01	17.22
No computer use data	21.41	21.13	14.94	16.42	15.85	10.25
Other	26.09	18.28	22.19	13.00	8.41	12.02
Respondents	14.86	15.81	16.64	17.51	17.60	17.58
Total	15.78	16.25	16.79	16.99	17.23	16.96

A full example

- Option c) Profile the partial non-respondents

Type of partial non-response by Income categories

Row Pct	Below Lico	Below 1.5*LICO	Above 1.5*LICO
No children data	15.75	18.86	65.39
No computer use data	32.65	19.64	47.71
Other	16.11	12.33	71.56
Respondents	12.55	16.01	71.45
Total	14.08	16.38	69.54

A full example

- Option c) Profile the partial non-respondents

Type of partial non-response by Rural/Urban indicator

Row Pct	Less Urban	More Urban
No children data	48.75	51.25
No computer use data	60.64	39.36
Other	55.76	44.24
Respondents	47.36	52.64
Total	48.52	51.48

A full example

- Option c) Profile the partial non-respondents
  - The factors that appear to be linked the most to the partial non-response are: age, income and size of area of residence (rural/urban).
  - Results reported under option a) could be affected by the difference in the respective profile (especially if the variables identified above are also linked to the variable of interest).

A full example

- Option d) Re-weight the records with a response to account for the partial non-respondents

  - Here, we will use the variables identified in option c) to inflate the weights of the respondents, so that they also represent the partial non-respondents.

  - We first need to create an indicator variable to easily identify respondents.

A full example

- Option d) Re-weight the records with a response to account for the partial non-respondents
  - SAS code useful for re-weighting:

```
%macro write_adj;
    %do i=1 %to 1000;
        ,sum(bsw&i)/sum(bsw&i*respondent)
        as adj&i
    %end;
%mend;
```

A full example

- Option d) Re-weight the records with a response to account for the partial non-respondents
  - SAS code useful for re-weighting:

```
%macro write_newboot;
    %do i=1 %to 1000;
        ,bsw&i*calculated adj&i as
        new_bsw&i
    %end;
%mend;
```

A full example

- Option d) Re-weight the records with a response to account for the partial non-respondents
  - SAS code useful for re-weighting:

```
proc sql;
   create table filenamev2 as
   select *, sum(fwtcw01l)/sum(fwtcw01l*respondent) as
   adjustment %write_adj, fwtcw01l*calculated
   adjustment as new_fwgt %write_newboot
   from filename
   group by fmmcq01, catrev, rural_urban;
quit;
```

A full example

- Option d) Re-weight the records with a response to account for the partial non-respondents
  - The 36 adjustments (6 X 3 X 2 = 36) created ranged from 1.1 to 1.6 and the smallest group contained 19 respondents.

A full example

- Option d) Re-weight the records with a response to account for the partial non-respondents
  - A similar code could have been written to make sure that the control totals would match the previous counts.
  - The groupings for the control totals would be defined by province of residence, gender and age.

A full example

- Option d) Re-weight the records with a response to account for the partial non-respondents

hours a day do you spend on a computer

FATCeQ21	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Less than 1 hour a day	867235.9	38.25	867235.9	38.25
1 or 2 hours a day	935328.1	41.25	1802564	79.50
3 or 4 hours a day	347297.4	15.32	2149861	94.82
5 or 6 hours a day	86306.34	3.81	2236168	98.63
7 or more hours a day	31166.5	1.37	2267334	100.00

A full example

- Option d) Re-weight the records with a response to account for the partial non-respondents

Main Effects	Survey weights p-values	Normalized weights p-values	Bootstrap weights p-values
Age	<0.0001	<0.0001	0.0000
Gender	<0.0001	0.1952	0.4378
Type of school	<0.0001	0.0859	0.5630
Computer in the house	<0.0001	<0.0001	0.0002
Work/study	<0.0001	0.0020	0.1945
Parental monitoring	<0.0001	<0.0001	0.0106

A full example

- Option d) Re-weight the records with a response to account for the partial non-respondents
  - Results haven't changed much…Why???
    - About 19% of partial non-response in total.
    - Half of them (no children data) very similar to respondents  about 10% to play with…
    - Variables used to create the re-weighting groupings are relatively poorly related to the variable of interest: no huge discrepancy in the extent of computer use according to the different level of the re-weighting variables.
    - Age was already partly 'accounted' for by the model.

A full example

- Note that we could have opted for other strategies:
  - Re-weight only cases with no children data, and impute all other partial non-response.
  - Re-weight only cases with no children data, impute cases with no computer use answer and use other partial non-response as a valid category.
  - Impute everything.

A full example

- Now, we could be interested in two things that would require using data from Cycle 5 and Cycle 6:
  - Reporting about the change in the way computers are used by teenagers.
  - Propose a model that would fit both Cycle 5 and Cycle 6 computer use behaviors.

A full example

- Let's concentrate on the first task.
  - First, we will need to check on the availability of the variables and make sure that these variables were not subject to changes between Cycle 5 and Cycle 6.
  - Then, if possible, we will need to define the variables in the same way.
  - Then, we will need to deal with the non-response. Note here that the reweighting groups could be entirely different or exactly the same.

A full example

- In the case of the Cycle 5 data, we will use the same re-weighting groups, as they appear to be related to the fact of responding or not.

Table of C5 respondent status by Income Categories

Row Pct	Below LICO	Below 1.5*LICO	Above 1.5*LICO
0	21.22	15.74	63.04
1	13.92	14.42	71.66
Total	15.94	14.79	69.27

A full example

- Before putting the data together and going ahead with a combined model, it is important to perform some testing about the assumption that the quantities of interest are the same. By completing the first task, we are exactly doing so.
- Computer use has evolved over the years, so it is likely that the responses could be affected by time.

A full example

- Profile of the use of computers by teenagers:

	Cycle 5 (cohort of 2-7 in 1994)	Cycle 6 (Cohort of 0-5 in 1994)
'Light' users	82.44% (1.02%)	79.50% (0.95%)
'Heavy' users	17.56% (1.02%)	20.50% (0.95%)

A full example

- Be careful when performing the testing!
  - Here, the two samples are highly dependent. The testing will need to take this into consideration.
    - Choose a test and a method that incorporates a covariance component.
    - In this case, we would conclude that the difference is significant (p-value=0.0293)

A full example

- Let’s summarize our findings:
  - When the 0-5 year-olds of the 1994 cohort became teenagers (10-15 years old) in 2004, about 20% of them were ‘heavy’ computer users (3 hours or more a day).
  - Being older and having a computer at home increased the odds of being a ‘heavy’ user.

A full example

- Let’s summarize our findings:
  - The children that are the least monitored were also more likely of being ‘heavy’ users.
  - Boys and girls were not different in their odds of being a ‘heavy’ user.

A full example

- Let’s summarize our findings:
  - The two variables (type of school and work/study situation) used as proxies for ‘having access to computers at school’ and ‘Amount of spare time of parents for kids’ were not linked to computer use.

A full example

- Let’s summarize our findings:
  - The extent to which computers get used by teenagers seems to be quickly expanding: there was an increase of about 3% in the proportion of ‘heavy’ computer users in the 1994 cohort of 0-5 year-olds compared to the 1994 cohort of 2-7 year-olds (when both groups respectively reached 10-15 years old).