

ANALYZING HEALTH DATA WITH MISSING VALUES – A CASE STUDY

David Haziza, Karla Nobrega, Julie Bernier and Patricia Whitridge¹

This case study on missing data uses a sub-sample of the 1994 National Population Health Survey. The context of the exercise is the relationship between health status and health predictors. Health status is measured with either a general health question or the Health Utilities Index (HUI). The data represent persons, aged 20 to 65, living in a private household in the Prairie provinces.

The National Population Health Survey (NPHS) uses the Labour Force Survey sampling frame to draw the initial sample of approximately 20,000 households. The survey is designed to collect information on the health of the Canadian population and related socio-demographic information. The first cycle of data collection began in 1994 and continues every second year thereafter. The sample collection is distributed over four quarterly periods followed by a follow-up period and the whole process takes a year. The survey is designed to produce both cross-sectional and longitudinal estimates. In each household, some limited health information is collected from all household members and one person in each household is randomly selected for a more in-depth interview.

The questionnaires include content related to health status, use of health services, determinants of health, a health index, chronic conditions and activity restrictions. The use of health services is probed through visits to health care providers, both traditional and non-traditional, and the use of drugs and other medications is also collected. Health determinants include smoking, alcohol use and physical activity. As well, a section on self-care has also been included this cycle. The socio-demographic information includes age, sex, education, ethnicity, household income and labour force status.

Four exercises were presented for the students to consider in their analysis.

Exercise One: Assessing the response mechanism

Examine the data to assess whether the probability of response in HUI or General Health depends on any of the

auxiliary variables. Survey statisticians consider three types of response mechanisms: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). A MCAR response mechanism is one for which the probability of response does not depend on the variable of interest nor on any of the auxiliary variables. A MAR response mechanism is one where the probability of response depends on some auxiliary variables but not on the variable of interest. A NMAR response mechanism is a mechanism where the probability of response depends on the variable of interest and potentially on other variables that are not observed.

Exercise Two: Deciding on the method to deal with the missing data

In the presence of non-response, the analyst has several options:

- i) Do nothing
- ii) Use only complete respondents
- iii) Adjust the weights of the respondents
- iv) Use an imputation method.

Do nothing:

This option compels the analyst to work with an incomplete data file. Different analyses performed on the same data file could yield inconsistencies between the results. Some software deletes records that have missing values for any variable.

Use only complete respondents:

This option eliminates all incomplete records. This could lead to bias in the results, unless the response mechanism is MCAR. Valuable respondent information is lost. Original sampling weights cannot be used for inferences. The analyst needs to adjust for non-response, either total or partial.

Adjust the weights:

This option compensates for partial non-response. The analyst can use auxiliary information to form adjustment

¹ David Haziza, Karla Nobrega and Julie Bernier, Statistics Canada, Ottawa, Ontario, K1A 0T6 and Patricia Whitridge, Royal Canadian Mounted Police, Strategic Policy and Planning Branch, 1200 Vanier Parkway, Ottawa, Ontario, K1A 0R2.

classes. It is necessary to create new adjusted weights for each variable of interest. This could lead to inconsistent results of different analyses.

Use an imputation method:

This option completes partial responses. An ‘artificial’ value is produced for each missing value. A single weight is then retained for each respondent. The analyst can use partial information to improve the quality of the imputation. There are many, many methods possible!

Dangers of imputing:

Inference, especially point estimation, is valid only if additional underlying assumptions are satisfied. The relationships between variables can be modified if care is not taken. The variance of the estimator is underestimated if imputed data is treated as observed values.

Exercise Three: Analysis

Students have two choices of dependent variables: the HUI or the general health question. A multiple linear regression can be used to assess the relationship between HUI and the covariates. The general health question is an ordinal variable, so a log-linear analysis can be used to assess the

relationship between self-assessed health and the covariates.

Quality of Life:

Traditionally, quantity of life has been measured using life expectancies, mortality rates, preventable deaths and potential years of life lost. The simple increase in the quantity of life, however, does not give any indication of whether there is an accompanying increase, decrease, or maintenance of the health-related quality of life. It provides an incomplete “snapshot” of the health of a population. New models consider the expected number of years to be lived in specific health states. The Health Utilities Index (HUI) measures health status, health-related quality of life and utility scores.

Exercise Four: Estimation of Bias from Imputation

This exercise examines the impact of the response mechanism, the imputation method and the response rate on the point estimator. The Generalized System for Imputation Simulations (GENESIS v.1.0) was provided to students to aid them in their analysis. The assumptions on response mechanism must be evaluated. Assumptions on the variable of interest, especially surrounding the choice of model, must also be verified.