Appendix Y Statistical methods for the comparison of dietary intake

Ivonne Solis-Trapala

Y.1 Introduction

This appendix provides an outline description of the statistical methods used for the comparisons of dietary intake from Years 1 to 4 of the NDNS Rolling Programme (RP). The statistical analyses require estimating the difference of mean intake of non-overlapping subpopulations, defined by income and fieldwork years.

The NDNS RP sample requires weights to adjust for differences in sample selection and response. The statistical analysis of data generated from this complex survey design requires taking the sample design (ie sample stratification, clustering and weighting) into account to yield valid estimates of the population parameters. A detailed description of the weighting and sampling procedures is provided in Appendix B.

Y.2 Comparison of dietary intake between non-overlapping subpopulations

This section outlines the statistical methods used to estimate the differences between mean intakes of key foods and nutrients from non-overlapping subpopulations. The relevant analyses included differences between means for equivalised household income quintiles split by age and sex (see Chapter 9). Equivalised household income was derived from the variable “income” so that the differences in the household’s size and composition are taken into account to yield a representative income. The comparisons among equivalised household income quintiles used the highest income group as the reference group.

In addition, NDNS RP data for Years 1 to 4 was split to form two groups (survey period 1: Years 1&2 and survey period 2: Years 3&4). The same weights and design variables created for the Years 1 to 4 dataset were applied to the appropriate subsets of the data.[1] Analysis of mean daily intake of key nutrients and foods compared Years 1&2 with Years 3&4 across five age groups, overall and by sex. The age groups were 1.5 to 3 years (sex combined only), 4 to 10 years, 11 to 18 years, 19 to 64 years and 65 years and over (see Chapter 10, section 1).

The comparisons described above involve comparing either means of continuous variables (mean differences in energy and nutrient intakes) or differences of proportions (such as the percentage of the sample with an intake below the LRNI) among groups, defined by survey periods (Years 1&2 compared with Years 3&4) or equivalised household income (quintiles), overall and by subgroups (sex and consumers/non-consumers of alcohol). The mean differences for the continuous variables were estimated through multivariate linear regression models and

differences of proportions through logistic regression models. The statistical analyses were undertaken following three stages: exploratory analyses, estimation of mean differences and diagnostic procedures (ie assessment of model assumptions and goodness of fit). All the analyses including the graphical tools and diagnostic procedures took into account the complex survey design.

Y.2.1 Exploratory analyses

The observed distribution of the continuous variables was screened through histograms, Q-Q plots and boxplots. These graphical tools showed the shape of the distribution and highlighted the presence of outliers. These were investigated as well as their impact on the regression analyses. In cases where the variable had small variability and hence took a reduced range of values (eg fish or alcohol consumption), the variable was dichotomised using the population median as the cut-off value and analysed through logistic regression.

Y.2.2 Estimation of differences of means

Multivariate linear regression models were used for continuous measurements of nutrient or food intake. The purpose of the analyses was to perform simple study-domain comparisons rather than investigating the relationship between nutrient or food intake and age or gender. Therefore, only categorical variables needed to be defined to represent the comparison groups (Years 1&2 compared with Years 3&4) or equivalised household income (quintiles), the study domains (age, sex and consumers/non-consumers of alcohol) and their interactions. The regression coefficients estimate the subgroup differences that exist in the population. This approach is equivalent to estimating each difference of means by study domain, provided that the full sample is used for the estimation of standard errors. The use of regression models allows the analyst to estimate the mean differences simultaneously. For illustration, consider the comparison of mean intakes of fruit in grams between survey period 1 (Years 1&2) and 2 (Years 3&4) across age groups. The response variable is total fruit intake and the independent variables are: age (categorical variable for 1.5 to 3 years, 4 to 10 years, 11 to 18 years, 19 to 64 years and 65 years and over), survey period (categorical variable for survey periods 1 and 2) and the interaction between age and survey period. The variable “age” has four associated regression coefficients (B11, B12, B13 and B14), the indicator variable “survey period” has one regression coefficient (B2) and the interaction term generates four regression coefficients (B31, B32, B33 and B34), the intercept is denoted by B0. The target differences of means are functions of these parameters as described in Table Y.1. Tests of hypothesis for these differences can be undertaken by use of the estimated regression parameters and their covariance matrix.

Table Y.1 Comparison of mean intakes of fruit in grams between survey periods 1 and 2 across age groups in terms of linear regression parameters

Age group (years)	Mean intake (survey period 1)	Mean intake (survey period 2)	Difference of means (survey period 2 minus period 1)
1.5-3	B0	B0+B2	B2
4-10	B0+B11	B0+B11+B2+B31	B2+B31
11-18	B0+B12	B0+B12+B2+B32	B2+B32
19-64	B0+B13	B0+B13+B2+B33	B2+B33
65 years and over	B0+B14	B0+B14+B2+B34	B2+B34

In this example the linear regression model can be expressed as:

$$y_{hij} = B0 + \sum_{r=1}^{4} B1rx1r_{hij} + B2x2_{hij} + \sum_{r=1}^{4} B3rx1r_{hij} \cdot x2_{hij} + \varepsilon_{hij},$$

where y_{hij} represents the observed total fruit intake for the *j*-th individual in the *i*-th primary sampling unit of the *h*-th stratum; x1r (r=1,2,3,4) are indicators for age groups, with the first group used as reference category; x2 is an indicator for survey period 2 and ε_{hij} is the error term.

The regression coefficients in this model were estimated using probability weighted least squares[2] and their covariance matrix was estimated using a Taylor linearization method.

Y.2.3 Estimation of differences of proportions

Logistic regression models the probability describing the possible outcome of a binary variable as a function of explanatory variables, using a logistic transformation. In this model, the logarithm of the odds of occurrence (eg odds of meeting the “5-a-day” guideline for fruit and vegetable intake[3]) is expressed as a linear function of explanatory variables. Differences in proportions were estimated using logistic regression analyses for the observed proportions. The terms in the linear predictor of the logistic regression models were defined as described in the previous section; however, the regression coefficients have different interpretations. Here, they represent group differences expressed in terms of log odds ratios. For example, to analyse the changes in proportions of people meeting the “5-a-day” guideline between survey periods 1 and 2, for a given age group (eg 19 to 64 years), we obtain an estimate of the ratio of the odds of meeting the “5-a-day” guideline at survey period 2 and the odds of meeting the “5-a-day” guideline at survey period 1 (analogous to B2+B33 in Table Y.1), in the logarithmic scale. An estimated log odds ratio of zero indicates no changes in the proportion of people meeting the “5-a-day” guideline, while negative/positive values correspond to decreases/increases in the

proportion. The regression parameters in these models were estimated using a pseudo-likelihood approach[4] and their covariance matrix was estimated using a Taylor linearization method.

Y.2.4 Diagnostic procedures

The linearity assumption between the dependent variable and the explanatory variables is crucial in multiple regression analyses; however, the use of categorical variables as independent explanatory variables does not require the assumption of a linear relationship with the dependent variable. Similarly, the logistic regressions specified above do not require a linear relationship between the log odds and the explanatory variables. Therefore, checks for departures from linearity were not undertaken. The goodness of fit of the multivariate linear models was examined using the concept of explained variation (R-squared).

The statistical analyses described above were performed using the survey package in the statistical program R.[5,6]

The statistical analyses described in this appendix are for descriptive purposes rather than analytical, ie they are not intended to estimate the associations among many variables. Therefore, corrections for multiple comparisons were not necessary. Bonferroni procedures may be applicable in other situations involving simultaneous testing of regression coefficients when the number of independent variables in the regression analysis is large compared to the number of sampled PSUs.[7]

---

[1] Although the weights were not specifically designed for this type of sub-group analysis, it was possible to use the Years 1 to 4 weights and design variables for just 2 years' data (Years 1&2 or Years 3&4), as:

- The selection weights correct for any differences in sampling strategy across survey years,
- We did not find evidence that response behaviour had changed significantly between the two survey periods.

However, to use subsets of any other combination of years of the dataset, the weights and design variables would have to be reviewed to ensure that the subset of data is still representative of the UK population when the Years 1 to 4 weights and design variables have been applied.

[2] Holt, D., Smith, T.M.F. and Winter, P.D. (1980) Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society A,* **143,** 474 –487.

[3] Appendix A provides further details regarding the "5-a-day" guidelines for those aged 11 years and over. "5-a-day" portions of fruit and vegetables were not calculated for children aged 10 years and younger.

[4] Skinner, C.J. (1989) Domain means, regression and multivariate analysis. In *Analysis of complex surveys* (eds C.J. Skinner, D. Holt and T.M.F. Smith). Chichester: Wiley.

[5] Lumley, T. (2012) "survey: analysis of complex survey samples". R package version 3.28-2.
Lumley, T. (2004) Analysis of complex survey samples. *Journal of Statistical Software,* **9**(1): 1-19

[6]R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[7] Korn, E.L., Graubard, B.I.(1990) Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni *t* statistics. *The American Statistician,* **44,** 270 –276.