Appendix Y Statistical methods for the comparison of dietary intake

Jianhua Wu, Petros Gousias, Nida Ziauddeen, Sonja Nicholson and Ivonne Solis-Trapala

Y.1 Introduction

This appendix provides an outline description of the statistical methods used for the comparisons of dietary intake in the Scotland report for 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 SIMD) as well as overlapping subpopulations, (defined by country).

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 (i.e. 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 subpopulations

This section outlines the statistical methods used to estimate the differences between mean intakes of key foods and nutrients from non-overlapping or overlapping subpopulations. The relevant analyses included differences between means for equivalised household income quintiles and for Scottish Index of Multiple Deprivation (SIMD)[1] quintiles split by age for those aged 4 to 10 years, 11 to 18 years and 19 to 64 years (see Chapter 9). Equivalised household income was derived to account for the differences in the household's size and composition and thus yield a representative income. The comparisons for equivalised household income and SIMD quintiles used the highest income group as the reference group. In addition, NDNS RP Scotland data for Years 1 to 4 has been compared to NDNS RP UK data for Years 1 to 4, of which NDNS RP Scotland data for Years 1 to 4 is a subset. A set of weights and design variables were generated for the NDNS RP Scotland Years 1 to 4 dataset. These are described in more detail in Appendix B.

Analysis of mean daily intake of key nutrients and foods compared NDNS RP Scotland data for Years 1 to 4 to NDNS RP UK data for Years 1 to 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).

The comparisons described above involved 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 non-overlapping groups (see Chapter 9), defined by equivalised household income or SIMD (quintiles), or between overlapping groups (see Chapter 10), defined by countries (NDNS RP Scotland data for Years 1 to 4 compared with NDNS RP UK data for Years 1 to 4). 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 (i.e. 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 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 (e.g. 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 between non-overlapping subpopulations

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 sex. Therefore, only categorical variables needed to be defined to represent the comparison groups, such as equivalised household income (quintiles) or SIMD (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 among equivalised household income (quintiles) across age groups. The response variable is total fruit intake and the independent variables are: age (categorical variable for 4 to 10 years, 11 to 18 years and 19 to 64 years), equivalised household income (categorical variable for quintiles 1, 2, 3, 4 and 5, with quintile 5 the highest income group) and the interaction between age and equivalised household income. The variable "age" has two associated regression coefficients (B11 and B12), the indicator variable "quintile" has four regression coefficients (B2, B3, B4 and B5) and the interaction term generates eight regression coefficients (B21, B22, B31, B32, B41, B42, B51 and B52), the intercept is denoted by B0. The target differences of means are functions of these parameters as described in Table Y.1 (only differences between quintiles 1 and 5 are shown for illustration). 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 equivalised household income (quintiles) across age groups in terms of linear regression parameters

Age group (years)	Mean intake (quintile 5)	Mean intake (quintile 1)	Difference of means (quintile 1 minus quintile 5)
4-10	B0	B0+B2	B2
11-18	B0+B11	B0+B11+B2+B21	B2+B21
19-64	B0+B12	B0+B12+B2+B22	B2+B22

Note: this table only shows the model mean intake and mean differences for quintiles 1 and 5.

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

$$y_{hij} = B0 + \sum_{r=1}^{2} B1r\ x1r_{hij} + \sum_{t=2}^{5} Bt\ x2t_{hij} + \sum_{t=2}^{5}\sum_{r=1}^{2} Btr\ x1r_{hij}\ x2t_{hij} + \varepsilon_{hij}$$

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) are indicators for age groups, with the first group used as reference category; x2t (t = 2, 3, 4, 5) is an indicator for equivalised household income (quintiles), with quintile 5 used as reference category; 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 (e.g. 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 equivalised household income quintile 1 and quintile 5, for a given age group (e.g. 19 to 64 years), we obtain an estimate of the ratio of the odds of meeting the "5-a-day" guideline at quintile 1 and the odds of meeting the "5-a-day" guideline at quintile 5 (analogous to B2+B22 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, i.e. 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]

Y.2.5 Comparison of dietary intake between overlapping subpopulations

The comparisons between NDNS RP Scotland data for Years 1 to 4 and NDNS RP Years 1 to 4 data for the UK as a whole involve comparing either means or proportions between overlapping subpopulations. The mean estimates for Scotland and the UK as a whole are analogous to the analyses described in section Y.2.2. The mean difference is the subtraction of the mean intake for Scotland and the mean intake for the UK. However, estimation of the standard error of the mean difference requires consideration of the overlapping of the sample.

For illustration, consider the comparison of mean intakes of fruit in grams between Scotland and the UK across age groups, where Scotland is a subset of the UK as a whole. Suppose the mean intakes of fruit in grams for the 4 to 10 years age group for Scotland and the UK is $\bar{y}_1$ and $\bar{y}_2$, respectively. The standard error of the mean difference $d = \bar{y}_1 - \bar{y}_2$ can be calculated using the formula below:

$$s.e.(d) = \frac{r}{t}\sqrt{var(\bar{y}_{1\notin 2}) + var(\bar{y}_1)}$$

Where r refers to the weighted sample size of the UK after excluding Scotland and t refers to the weighted sample size of the UK as a whole; $var(\bar{y}_{1\notin 2})$ represents the variance of mean intakes of fruit for the UK excluding Scotland and $var(\bar{y}_1)$ represents the variance of mean intakes of fruit for Scotland.

The Z-score for testing whether the mean difference is significantly different from zero can be obtained by

$$Z = \frac{d}{s.e.(d)}$$

[1] http://www.scotland.gov.uk/Topics/Statistics/SIMD/ (accessed 31/07/14)

[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.