NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

10-18

How to use Replicate Weights in Health Survey Analysis using the National Nutrition and Physical Activity Survey as an Example

Carole L. Birrell, David G. Steel, Marijka J. Batterham and Ankur Arya

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia Phone +61 2 4221 5435, Fax +61 2 4221 4845.
Email: karink@uow.edu.au

Abstract

Objective

To conduct nutrition related analyses on large-scale health surveys, two aspects of the survey must be incorporated into the analysis: the sampling weights and the sampling design; a practice which is not always observed. The aim of this paper is to compare three analyses: 1. unweighted; 2. weighted but not accounting for the complex design; and 3. weighted and accounting for the complex design.

Design

Analyses with and without the proper use of sampling weights and replicate weights are conducted using Stata. Descriptive statistics are computed and a logistic regression investigating whether chosen explanatory variables are associated with being overweight/obese.

Setting

Cross-sectional health survey with complex sampling design when replicate weights are supplied.

Subjects

Responding adults from the National Nutrition and Physical Activity Survey (NNPAS) part of the Australian Health Survey (2011-13).

Results

An unweighted analysis gives both a biased estimate and incorrect standard errors. Adjusting for the sampling weights give unbiased estimates but incorrect standard errors. Incorporating both the sampling weight and design results in unbiased estimates and the correct standard error. This can have effects on interpretation, for example, the odds ratio for being a current smoker in the unweighted analysis was 1.20 (95%CI 1.06,1.37) t=2.89, P=0.004, suggesting a statistically significant relationship with being overweight/obese. When the sampling weights and complex design are incorporated the results are no longer significant, odds ratio =1.06 (95% CI 0.89,1.27) t=0.71, P=0.480.

Conclusions

Correct incorporation of the sampling weights and design are crucial for valid inference from survey data.

Keywords: Complex survey design, replicate weights, survey sampling, sampling weights, health surveys, BMI.

1 Introduction

Many nutrition and public health researchers make use of data obtained from large-scale surveys to estimate relationships between nutrition and the health status of the population and particular sub-groups, and to inform health policies. The Australian Health Survey (AHS) and the United States National Health and Nutrition Examination Survey (NHANES) are two health surveys collected using a complex survey design. Complex sampling procedures may involve use of many design features such as geographic stratification, multistage sampling involving clustering, and the disproportionate sampling of certain ethnic or age groups. In order to validly generalise the results to the relevant population, the study design features must be incorporated into the estimation and analysis.

Analysis of data resulting from a complex sample survey to produce unbiased estimates and estimated standard errors (SEs) which account for the sampling design can be complicated[1]. It requires the use of the individual's sampling weight and the sampling design variables; the resulting estimates are called *design-based* estimates[2]. The sampling weight is based on the inverse of the probability of selection and will often vary considerably between individuals, due to the sampling design and post-survey adjustments. It can be considered as the number of units (such as individuals) in the population that the unit represents. If the sampling weight is ignored in the analysis, this is equivalent to setting all the weights to be equal to one, producing biased estimates of population quantities such as means, totals and proportions. For more discussion on survey weights see Levy & Lemeshow[2], ch 16; Valliant *et al.*[1] and Valliant & Dever[3]. Using the sampling weights but ignoring the sampling design will result in biased estimates of the SEs associated with the estimated population quantities, resulting in invalid inferences[4–6].

A statistical agency may release unit-level data for public use with different levels of confidentiality protection. There are essentially two ways the unit-level data, often called the Confidentialised Unit Record File (CURF), are released: with or without the sampling design variables. The purpose of the latter approach is to protect the identity of the respondents. Instead of the sampling design variables, a set of replicate weights are supplied; the number of which may vary from survey to survey. To obtain unbiased estimates and valid estimates of SEs, the researcher needs to use either: (1) the individual's sampling weight as well as the sampling design variables; or (2) the individual's sampling weight in conjunction with the set of replicate weight variables (Section 2.1). In the estimation of valid SEs, a Taylor series linearisation method is applied in the first approach; and for the second approach, a replication method such as the jackknife method is required. An example of the two procedures using the NHANES data can be found in StataCorp[7], p116–117. The importance of using the sampling weights and the sample design variables as in (1) is demonstrated in Saylor *et al.*[8] and Kim *et al.*[9] with particular reference to NHANES and the Korean NHANES respectively.

Researchers new to survey analysis often struggle to understand the weighting procedure and how this should be incorporated into the analysis. The focus of this paper is to answer the following questions when the data supplied includes the replicate weights rather than the sampling design variables, as in (2):

1. *What happens if I don't use the sampling weights or the design information in my analysis?*;
2. *How do I carry out analyses such as estimation of means, proportions and their SEs; and estimates of coefficients for a logistic regression model.*
3. *How do I obtain estimates for subgroups when data is sampled using a complex survey design?*
4. *How do I set up the code to include to incorporate the replicate weights in Stata?*

The data from the Australian Health Survey (2011-13) (AHS) will be used to answer these questions showing results for three analyses: (1) unweighted; (2) weighted but

not accounting for design; and (3) weighted and accounting for complex sample design. This paper is structured as follows: in Section 2, a description of replicate weights, the AHS sampling design and variables to be used are described, along with the details of the statistical analyses performed. In Section 3, results for the three methods are provided with discussion of results in Section 4.

2 Experimental Methods

2.1 Replicate Weights

In the replication approach, sub-samples are selected from the original sample in a way that reflects the sample design and estimation methods used, analysis is carried out on each sub-sample and the variance between these estimates is used to estimate the variance and SE of the required parameter estimate from the full sample[10]. The replicate weights may be constructed by the statistical agency using a jackknife procedure which depends upon the sample design. Often in a multistage design, each replicate includes all but one primary sampling unit (PSU) and the total number of replicates is the number of PSUs in the design[6]. If the sample design involves a large number of PSUs, there will be a large number of replicates. An alternative is the delete-a -group jackknife method[10,11] where each replicate is formed by deleting one in R of the PSUs, where R is the number of replicates required. For more detail in how to generate replicate weights (in Stata) given the sample design refer to Valliant & Dever[3], Section 5.4.

When the survey data set does not provide the survey design variables, the number of PSUs (the top level cluster variable) is often not provided. Instead, the number of replicate weights and the variable names will be specified in the user documentation. When the statistical agency constructs and supplies the set of replicate weights in a CURF, it simplifies the task for the analyst in as much as the variables pertaining to the sampling design as in approach (1) and the syntax required in statistical software to use them are not required. However, for approach (2), the data analyst must know how to use the replicate weights, a demonstration of which is given in this paper.

For the AHS, the set of replicate weight consists of $R = 60$ variables, in addition to the individual's sampling weight (referred to as the person weight). Each of these 60 replicate weight variables will have a collection of rows or individuals where the weight is set to zero, such that no two variables will have the same rows set to zero but across the 60 variables, each case will appear as zero in one variable only. The collection of rows which are set to zero for a replicate weight variable indicate those individuals that are deleted to form the replicate. Because replicates are formed by deleting PSUs, the number of rows set to zero in each variable may vary. In each replicate weight variable, the remaining non-zero weights are adjusted to sum to the number of units in the population; so the sum of the weights for each of these variables is identical. It would be incorrect to only use a subset of the full set of replicate weight variables. This set of replicate weights is then used in the jackknife variance estimation for the parameters of interest. For more details about standard errors and the replicate weights technique for the AHS see see the AHS User's Guide[10] and for an introduction to jackknife estimation see Abdi & Williams[12].

2.2 Data Description

The Australian Health Survey (2011-13) (AHS) combines three national health surveys collected by the Australian Bureau of Statistics (ABS), namely: National Health Survey (NHS); National Nutrition and Physical Activity Survey (NNPAS); and National Health Measures Survey (NHMS) which is a biomedical information component. Information collected includes health status, risk factors, actions and socioeconomic circumstances. More detailed information about the structure of the AHS may be found in the AHS First Results Report[13].

For the purpose of this paper, variables used will be measures taken from NNPAS, as this is the survey generally of most interest for nutrition related questions. The sampling design used a stratified multistage area sample of private dwellings, collecting information by face-to-face interview. The strata are Statistical Divisions within each state and territory; each stratum comprises a number of Census Collection Districts (CDs) which were used as PSUs and consists of an average 250 dwellings. The CDs

were sampled within each stratum and then dwellings within a sample of a selected block in each selected CD were selected. A total of 3047 PSUs were selected ; persons were then randomly selected from each dwelling such that one adult and one child aged 2-17 years were selected where possible. Oversampling, (i.e. higher sampling rate) of older adults (65+ years) was also carried out. More detail of the sampling design may be found in the AHS Users' Guide[10]. This complex survey design is typical of many national surveys. The total responding sample ($n = 12153$) comprised both adults and children aged 2+ years, and our analysis has been limited to adults (aged 18+ years; $n = 9435$).

The survey included the collection of measured height (in cm) and weight (in kg) and then Body Mass Index (BMI) was calculated as the weight in kilograms divided by the square of the height in metres. BMI values are categorised according to the World Health Organisation (WHO) and the National Health and Medical research Council (NHMRC) guidelines. These categories are: *underweight* (< 18.5), *normal* (18.50-24.99), *overweight* (25.00 to 29.99) and *obese* (30+)[10]. The relevant original variable names in the NNPAS CURF are: weight (*PHDKGWBC*), height (*PHDCMHBC*), measured BMI (*BMISC*) and BMI categories (*BMICATHY*).

There are three types of sampling weights supplied in the NNPAS dataset: household weight; two person weights (for all responding persons and biomedical sample only). For estimating mean BMI and proportions of persons categorised as overweight or obese, the person weight (*NPAFINWT*) applied to all responding persons is appropriate. The 60 replicate weights are named *WPM0101 - WPM0160*.

2.3 Statistical Analysis

Estimating descriptive statistics and their SEs for a mixture of variable types was conducted: continuous variables *Height (in cm), Weight (in kg)* and *BMI*; categorical variables *Overweight or Obese* and *Current smoker*. Coefficients for a logistic regression model for the binary variable of *Overweight or Obese* were also estimated. Three methods of statistical analyses were conducted:

(1) Unweighted: without sampling weights or replicate weights;

(2) Weighted: with sampling weights but without accounting for the complex design; equivalent to weighted analysis assuming simple random sampling (SRS); and

(3) Complex design: with sampling weights and accounting for the complex design using a jackknife procedure with the replicate weights.

For the three continuous variables *Height (in cm), Weight (in kg)* and *BMI*, the estimated mean and SE were determined. A binary variable identifying adults ($\geq$ 18 years) was first created from the continuous age variable (*AGEC*); then a binary variable identifying overweight or obese adults was created. For the categorical variable for smoking *SMOKEQ1*, the percentage of current smokers is estimated for the adult population. A logistic regression model for the status of overweight or obese adults is applied using the covariates: *sex*, *age (in years)*, highest year of school completed (*SchEd*), total minutes undertaken physical activity in last week (*PhysActMin*), remoteness of area category (*ARIABC*) and current smoker (*SMOKEQ1*). Reference categories: for *sex* is male; for *SchEd* is year 12 or equivalent; for *ARIABC* is major city; for *Smoker* is yes. Stata 15 was used for all analyses, the commands for `mean`, `proportion`, and `logistic` were used with the appropriate `svy` command settings for the three methods given in the appendix. The 60 jackknife replicate weight variables are defined in Stata with the `jkrweight` option in the `svyset` command.
The formula used for the three methods are shown here for the estimate of the population mean and its variance.

(1) Unweighted: the familiar sample mean of a single variable y, denoted by $\bar{y}$, and its estimated variance assuming a simple random sample without replacement of size n from a target population of size N, and sample variance s^2, is calculated without the sampling weights or the replicate weights. If y_i is the ith observation $(i = 1 \ldots n)$ from the sample, then the sample mean and estimated variance is given by:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad v(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

(2) Weighted: If the sampling weight for an individual in the sample is denoted by w_i $(i = 1 \dots n)$ where the sum of the weights is equal to the population size N, $\sum_{i=1}^{n} w_i = N$, the estimator of the population mean is the mean of the weighted observations; and the variance is the equivalent to weighted analysis assuming simple random sampling (SRS) such that:

$$\hat{\theta} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \qquad v(\hat{\theta}) = \left(1 - \frac{n}{N}\right)\left(\frac{n}{n-1}\right)\frac{1}{N^2}\sum_{i=1}^{n} w_i^2 (y_i - \hat{\theta})^2$$

(3) Complex design: the sample weights are used to calculate the weighted mean as given in (2). The replicate weight variable for each replicate group is used to obtain each of the R replicate estimates of the mean resulting in $\hat{\theta}_1 \dots \hat{\theta}_R$. The variance estimate of $\hat{\theta}$ is then given by $v^*(\hat{\theta})$:

$$v^*(\hat{\theta}) = m \sum_{r=1}^{R} (\hat{\theta}_r - \hat{\theta})^2 \tag{1}$$

where the jackknife multiplier (m) is given by $m = \frac{R-1}{R}$. For the AHS data, a delete-a-group jackknife method of replicate weighting is used producing $R = 60$ replicate weights, so $m = 59/60$ is the required multiplier[10].

For the coefficients in a logistic regression, the variance of the unweighted estimates was estimated using standard methods. For weighted analysis, the variance ignoring the sample design was estimated using a linearization approach[14]. The jackknife approach uses $v^*(\hat{\theta})$ defined in (1) where $\hat{\theta}_r$ is the estimate obtained using the weights for replicate r.

2.4 Estimates for subgroups

Often a researcher is interested in estimating a quantity such as a mean or proportion for a subgroup of the population, for example, the mean *BMI* by *Sex* may be of interest. In this section, we focus on how to carry out such analysis when the jackknife replication method is to be used. The subgroup analyses for the mean *BMI* and the proportion of *Overweight or Obese* and *Smoker* by *Sex* are conducted.

When analysing survey data for particular subgroups, it is strongly recommended that the full data set be used rather than restricting the data to the particular cases belonging to the subgroup. The Stata manual for survey data[7] describes the use of command options `subpop` and `over` when estimating parameters for subgroups of the population rather than restricting the number of cases using conditional `if` or `in` qualifiers. The `subpop` option can be used to break down estimates into two groups using a binary variable with zero/nonzero values such as 0/1. The `over` option allows a breakdown by a categorical variable with two or more categories. West *et al.*[15] explain the conceptual differences between the 'conditional' approach and the design (or 'unconditional') approach. The latter allows for the variance arising from the randomness of subgroup representation in the sample.

The Stata manual[7] p72 states that 'every (replicate) weight variable produces a replicate, even if it does not contain an observation within the subpopulation specified in the `subpop` option'. However, if the `if` or `in` qualifiers are used instead, only the PSUs that have at least one observation within the subgroup will produce a replicate. However, West *et al.*[15], p527 state that a conditional approach for complex design estimation of variances will produce the correct standard errors assuming the replicate weights have been correctly produced according to the complex sampling design. Valliant *et al.*[1], p421 note that the jackknife correctly handles subgroup estimation without the need to explicitly give people not in the subgroup a zero response variable. They also note that for linearization approach to variance estimation, restricting the data set to the subgroup is generally a mistake.

3 Results

The results for the descriptive statistics for the 5 chosen variables are listed in Table 1. The unweighted (1) estimates of the means and percentages differ from the weighted and complex design, as expected. For *Height* the unweighted mean is lower but for the other listed variables, it is higher. The standard errors across the three methods are different: the difference being larger between Methods (1) and (2) than between

Methods (2) and (3). What is interesting is the difference between the SEs for methods (2) and (3). For *Height*, *Weight* and *BMI*, the SEs have decreased, however, increases in SE are evident for *Overweight or Obese* (from 0.74 to 0.78) and for *Current Smoker* (0.52 to 0.53). Method (1) gives a biased estimate of both the mean (or percentage) and the SE; Method (2) gives an unbiased estimate of the mean (or percentage) but a biased estimate of the associated SE; and Method (3) provides an unbiased estimate of both the mean (or percentage) and its SE.

The results for the logistic regression model of whether or not an adult is *Overweight or Obese* are provided in Table 2. The odds ratio (OR), standard error (SE), the t-statistic, related p-value and the 95% confidence interval (CI) are shown for each of the six covariates for each of the three methods. Similarly to Table 1, the results for Method (1) unweighted analysis provides a different parameter estimate for the odds ratio; but for methods (2) and (3), the estimated OR are equal for the corresponding covariate.

The estimated SEs for the corresponding covariates differ across the three methods. Comparing the SEs between Method (1) and Method(2), it is clear that for Method (2), the SEs are all higher than for Method (1). All but one of the SEs are higher for the complex design (3) results than for the weighted (2) results; with the SE for *Smoker* being the exception. The most notable difference is for the covariate *Smoker*. For *Smoker*, the unweighted method gives an *OR=1.20* which is statistically significantly higher than 1.0 (assuming a 5% level) with *t=2.89, P=0.004* and *95% CI: (1.061, 1.365)*. However, for the complex design method, the result is not significant *t=0.71, P=0.480* and *95% CI: (0.894, 1.267)*, thus underlining that invalid inferences can be made if analysis does not take the complex design into account. Also noteworthy are the results for the variable for Remoteness of area category: (*ARIABC*). Method (1) reports, for the *Other* category, *SE=0.0805, t=2.03 and P=0.043* whereas the corresponding results for Method (3) gives *SE=0.1460, t=2.47 and P=0.016*. The results for Method (2) are similar to those for Method (3) with SE slightly higher for Method (3).

To summarise the different SEs between methods for the same covariate, the ratio of

the SE for Method (2) to the SE for Method (1) found a minimum ratio of 1.12 (for *Smoker-No*), a maximum of 1.78 (for *ARIABC-Other*) and a median of 1.36 across the covariates. The ratio of the SE for Method (3) to SE for Method (2) found a minimum ratio of 0.99 (for *Smoker-No*), a maximum of 1.19 (for *SchEd-Year 10*) and a median of 1.07.

The subgroups analysis for the mean *BMI* and proportion of *Overweight or Obese* and *Smoker* by *Sex* were carried out for the jackknife variance estimates. However, since the same SE is produced for complex design estimates whether or not the 'conditional' or 'unconditional' approach is adopted, the results are not shown here. It is suggested that the analyst becomes familiar with, and uses the 'unconditional' approach for good practice, whether or not replicate weights are supplied in the data[5]. If the survey design variables are supplied rather than the replicate weights, then the conditional approach will produce incorrect SEs, hence the 'unconditional' approach is required. Stata code for this analysis is shown in the appendix.

4 Discussion

When reading the literature on secondary analyses of national health surveys, it is sometimes not clear whether or not the reported estimates are the weighted estimates and whether or not the analysis accounted for the complex survey design. Bell *et al.*[4] carried out a review of 1003 published papers reporting empirical research from 1995 to 2010 in three health surveys. They found that '60% of articles reported accounting for design effects and 61% reported using sample weights'. For an Australian example, Allman-Farinelli *et al.*[16] examine BMI and the prevalence of overweight and obesity by occupation using National Health Survey (NHS) 2004-05 data collected by the ABS. It is reported that the person sampling weight is used in the analysis but there is no mention of the use of the method to obtain the reported standard errors that account for the complex sample design and how the restriction to adults aged 20-64 years was handled. The AHS data from 2011-12 was used in a study on cardiovascular health by Peng *et al.*[17]. Poisson and logistic regression analyses were conducted on a

restricted subgroup of the core sample with analysis applying the biomedical sample weights and jackknife method as recommended by ABS[18].

Saylor *et al.*[8] demonstrate the importance of using the sampling weights and the survey's complex sample design in any statistical analysis with particular reference to NHANES. The sampling design variables for NHANES, including the stratification and cluster variables, are supplied in the data files in addition to the sampling weight. The authors undertake analyses in SPSS including descriptive statistics, linear and logistic regression using three methods: unweighted, weighted and complex samples. They conclude that accurate estimates of means and frequencies are produced if using weights without the complex sample design information but that 'weighting alone leads to inappropriate population estimates of variability'[8] p236. Similarly, Kim *et al.*[9] report that only 19.8% of the 247 research articles the Korean National Health and Nutrition Examination Survey (KNHANES) cited in PubMed from 2007 -2012 correctly used survey design analysis. Using SAS and SUDAAN, Kim *et al.*[9] compare the estimates of levels of lead, cadmium and mercury in the blood and the associated SEs as well as odds ratios (and 95% confidence intervals) for hypertension and osteoporosis for particular subgroups using both unweighted and a weighted analysis accounting for the complex design features. The results highlight the differences in the parameter estimates if weighting is not applied and the tendency for SEs to be underestimated and the CIs to be invalid.

The weighted SRS SE estimator [Method (2)] treats the data as a simple random sample of weighted values. This estimator at least partially accounts for the use of weights but does not reflect the effect of stratification and clustering in the sample design or the use of post-stratification in the estimation. Ignoring the effect of stratification will mean that the estimator will tend to overestimate the true SE, while ignoring the clustering and post-stratification will tend to underestimate the SE. The net effect of these factors will depend on the particular design used and the variable being considered and means that it is possible for the weighted SRS SE estimator [Method (2)] to be larger or smaller than the SE estimates obtained using the replicate weights [(Method (3)], which properly account for these effects. The clustering in the

NHS is not high, with an average of less than 7 dwellings selected per PSU and so we would not expect a large increase in SE due to the clustering in the sample, but some increase is evident in the complex variance analysis. A larger SE will result in a wider confidence interval and in turn will reduce the power of an analysis[19].

4.1 Conclusion

This paper discusses the results of three approaches to secondary analysis of complex survey data which have replicate weight variables supplied in the data rather than the survey design variables, such as the variables indicating the strata and cluster to which people belong. These are important considerations for nutrition related analyses in surveys employing replicate weights.

The first question: *What happens if I don't use the sampling weights or the design information in my analysis?* is answered in two parts. If the sampling weights are not used in the analysis, biased estimates are produced which will lead to incorrect conclusions. This is demonstrated by the differences in the estimates produced with the unweighted and weighted methods. In addition, if the complex design is not included, which translated to not using the replicate weight variables, the standard errors will also be incorrect. The use of these incorrect standard errors may result in incorrect inferences and conclusions.

The second question posed was: *How do I carry out analyses such as estimation of means, proportions and their SEs; and estimates of coefficients for a logistic regression model.* This paper demonstrates the use of replicate weights for analysing complex survey data by way of several examples using data from the Australian Health Survey (AHS). The AHS data has 60 replicate weight variables supplied with the data file. Other researchers of AHS data or other surveys with replicate weights may use this analysis as an example.

For the third question: *How do I obtain estimates for subgroups when data is sampled using a complex survey design?* We found no difference when the conditional or 'unconditional' approach was adopted using the replicate weights; thereby showing that the approach using the replicate weights is robust and simplifies procedures for

the analyst. However, for good practice, we suggest that the analyst becomes familiar with the unconditional approach to the analysis of subgroups when using other approaches to variance estimation.

The Stata code for all analyses is provided, which answers the last question posed: *How do I set up the code to include to incorporate the replicate weights in Stata?*. This code may be used as an example for researchers performing similar analysis. It is recommended that the analyst refers to the user's guide for the particular survey and the relevant Stata documentation to determine the type of replication method carried out and the number of replicate weights to apply. Some further examples may be found in chapter 5 of Valliant & Dever[3].

References

1. Valliant R, Dever JA, Kreuter F (2013) *Practical Tools for Designing and Weighting Survey Samples*. Statistics for Social and Behavioral Sciences, New York: Springer.

2. Levy PS, Lemeshow S (2008) *Sampling of Populations: Methods and Applications*. Wiley Series in Survey Methodology, Hoboken, New Jersey: John Wiley & Sons, 4th edition.

3. Valliant R, Dever JA (2018) *Survey Weights - A Step-By-Step Guide to Calculation*. College Station, Texas: Stata Press.

4. Bell BA, Onwuegbuzie AJ, Ferron JM, *et al.* (2012) Use of design effects and sample weights in complex health survey data: A review of published articles using data from 3 commonly used adolescent health surveys. *American Journal of Public Health* **102**, 1399–1405.

5. Heeringa SG, West BT, Berglund PA (2010) *Applied Survey Data Analysis*. Statistics in the Social and Behavioral Sciences Series, Boca Raton, FL, U.S.A.: Chapman and Hall / CRC, Taylor and Francis Group.

6. Campbell RT, Berbaum ML (2010) *Handbook of Survey Research*, chapter Analysis of Data from Complex Surveys. Bingley UK: Emerald Group Publishing Limited, 2nd edition, pp. 221–259.

7. StataCorp (2017) *Stata Survey Data Reference Manual, Release 15*. StataCorp LLC, College Station, Texas.

8. Saylor J, Friedmann E, Lee HJ (2012) Navigating complex sample analysis using national survey data. *Nursing Research* **61**, 231–237.

9. Kim Y, Park S, Kim NS, *et al.* (2013) Inappropriate survey design analysis of the Korean National Health and Nutrition Examination Survey may produce biased results. *Journal of Preventive Medicine & Public Health* **46**, 96–104.

10. ABS (2012) Australian Health Survey: Users' Guide, 2011-13, Cat. no. 4363.0.55.001. Technical report, Australian Bureau of Statistics (ABS), Canberra, accessed 4/12/2017, Available at http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4363.0.55.001Main+Features12011-13?OpenDocument.

11. Kott PS (2001) The delete-a-group jackknife. *Journal of Official Statistics* **17**, 521–526.

12. Abdi H, Williams LJ (2010) *Encyclopedia of Research Design*, chapter Jackknife. Thousand Oaks, CA: Sage.

13. ABS (2012) Australian Health Survey: First Results, 2011-12, Cat. no. 4364.0.55.001. Technical report, Australian Bureau of Statistics (ABS), Canberra, accessed 4/12/2017. Available at http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/4364.0.55.001main+features12011-12.

14. Binder DA (1983) On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review* **51**, 279–292.

15. West BT, Berglund P, Heeringa SG (2008) A closer examination of subpopulation analysis of complex-sample survey data. *The Stata Journal* **8**, 520–531.

16. Allman-Farinelli MA, Chey T, Merom D, *et al.* (2010) Occupational risk of overweight and obesity: an analysis of the Australian Health Survey. *Journal Of Occupational Medicine and Toxicology* **5**. Published Online: DOI: 10.1186/1745-6673-5-14.

17. Peng Y, Wang Z, Dong B, *et al.* (2017) Life's simple 7 and ischemic heart disease in the general Australian population. *PLoS ONE* **12**, 1–9. Available at https://doi.org/10.1371/journal.pone.0187020.

18. ABS (2005) Technical Manual: National Aboriginal and Torres Strait Islander Health Survey, Expanded CURF, 2004-05, Cat. no. 4715.0.55.002. Technical

report, Australian Bureau of Statistics (ABS), Canberra, accessed 1/2/2018. Available at http://www.abs.gov.au/AUSSTATS/abs@.nsf/Latestproducts/4715.0.55.002MainFeatures3002004-05.

19. Burden S, Probst Y, Steel D, *et al.* (2012) The impact of complex survey design on prevalence estimates of intakes of food groups in the Australian National Children's Nutrition and Physical Activity Survey. *Public Health Nutrition* **15**, 1362–1372.

A Stata code: Estimates of Means and Proportions

The code in this section relates to the results in Table 1. In AHS data, some variables have been given missing codes of 98, 99, 997, 998 and 999. These values were replaced with appropriate codes for missing observations in Stata such as .a, .b, .c (Stata code not given). For convenience, the variable for weight (*PHDKGWBC*) was renamed to *Weight_kg*. Similarly, the variable for height (*PHDCMHBC*) was renamed to *Height_cm*.

A dummy variable to indicate adults was created:

```
gen Adults=0 if AGEC<18
 replace Adults=1 if AGEC>=18 & AGEC<.
 label define Adultslabel 0 "Under 18" 1 "18 or over"
 label values Adults Adultslabel
```

A dummy variable to indicate the BMI category of overweight or obese was also created:

```
gen OverObese = 0
replace OverObese = 1 if BMISC >= 25 & BMISC<.
replace OverObese = .a if BMISC==.a
replace OverObese = .b if BMISC==.b
label define OverOblabel 1 "Overweight or Obese" 0 "Not Overweight or
Obese" .a "Measurement not taken - refusal" .b "Measurement not taken -
other reason"
label values OverObese OverOblabel
codebook OverObese, tabulate(20)
```

A.1 Method (1): Unweighted

Unweighted results are obtained using standard procedures without sampling weights or accounting for design features.

```
mean Height_cm, over(Adults)
```

```
mean Weight_kg, over(Adults)
mean BMISC, over(Adults)
proportion OverObese, over(Adults)
proportion SMOKEQ1, over(Adults)
```

Alternatively, the same results may be obtained using the following code assuming simple random sampling:

```
svyset, clear
svyset _n

svy, subpop(Adults):  mean Height_cm
svy, subpop(Adults):  mean Weight_kg
svy, subpop(Adults):  mean BMISC
svy, subpop(Adults):  proportion OverObese
svy, subpop(Adults):  proportion SMOKEQ1
```

A.2 Method (2): Weighted

Weighted results include the sampling weights but does not account for the complex sample design. *NPAFINWT* are the sampling weights supplied with the data.

```
svyset, clear
svyset _n [pweight=NPAFINWT]

svy, subpop(Adults):  mean Height_cm
svy, subpop(Adults):  mean Weight_kg
svy, subpop(Adults):  mean BMISC
svy, subpop(Adults):  proportion OverObese
svy, subpop(Adults):  proportion SMOKEQ1
```

A.3 Method (3): Complex Design

Results for Method (3) are weighted and account for the complex design: utilises the sampling weights, *NPAFINWT*, and the 60 replicate weights *WPM0101 -WPM0160* supplied with the data.

```
svyset, clear
local mult =59/60
svyset [pweight=NPAFINWT], jkrweight(WPM01*, multiplier('mult'))
vce(jackknife)

svy, subpop(Adults):  mean Height_cm
svy, subpop(Adults):  mean Weight_kg
svy, subpop(Adults):  mean BMISC

svy, subpop(Adults):  proportion OverObese
svy, subpop(Adults):  proportion SMOKEQ1
```

B Stata code: Logistic Regression

The code in this section relates to the results in Table 2. For convenience, the following variables were renamed: *LVHNSQBC* was renamed to *NonSchEd*; *HYSCHCBC* was renamed to *SchEd*; and *EXLEVELN* was renamed to *PhysActMin*.

B.1 Method (1): Unweighted

```
svyset, clear
svyset _n

svy, subpop(Adults):  logistic OverObese SEX AGEC PhysActMin i.SchEd
i.ARIABC i.SMOKEQ1
```

B.2 Method (2): Weighted

```
svyset, clear
```

```
svyset _n [pweight=NPAFINWT]
svy, subpop(Adults): logistic OverObese SEX AGEC PhysActMin i.SchEd
i.ARIABC i.SMOKEQ1
```

B.3 Method (3): Complex Design

```
svyset, clear
local mult =59/60
svyset [pweight=NPAFINWT], jkrweight(WPMO1*, multiplier(‘mult’))
vce(jackknife)
svy, subpop(Adults): logistic OverObese SEX AGEC PhysActMin i.SchEd
i.ARIABC i.SMOKEQ1
```

Table 1: Results for estimates of mean Height (in cm), Weight (in kg), BMI, percentage of Overweight or Obese adults (BMI $\geq$ 25) and percentage of Current Smokers for all adults with associated standard errors (SE) are shown for three methods: (1) Unweighted; (2) Weighted; and (3)Complex design.

	n	Parameter Estimate	(1) Unweighted	(2) Weighted	(3) Complex design
Height (in cm)	8057	Mean	168.5297	169.1566	169.1566
		SE	0.1117	0.1518	0.1187
Weight (in kg)	8009	Mean	78.4144	78.0985	78.0985
		SE	0.1993	0.2686	0.2503
BMI	7958	Mean	27.5417	27.2245	27.2245
		SE	0.0624	0.0816	0.0801
Overweight/Obese	7958	Percentage	64.0990	61.9995	61.9995
		SE	0.5378	0.7438	0.7752
Smoker	9435	Percentage	18.9189	17.6150	17.6150
		SE	0.4032	0.5244	0.5287

Table 2: Results for logistic regression for all adults (n=7874): whether or not an adult is Overweight or Obese given 6 explanatory variables are shown for three methods: (1) Unweighted; (2) Weighted; and (3) Complex design. Odds ratio (OR) standard error (SE) the t-statistic, related p-value and 95% confidence interval (CI) are shown.

		OR	SE	t	P	95% CI	
				(1) Unweighted			
Sex	F	0.5188	0.0257	-13.26	0.000	0.4708	0.5716
Age		1.0201	1.743E-3	11.67	0.000	1.0167	1.0236
PhysActMin		0.9996	0.833E-4	-5.04	0.000	0.9994	0.9997
SchEd	Year 11	1.4583	0.1247	4.41	0.000	1.2332	1.7244
	Year 10	1.3421	0.0877	4.50	0.000	1.1808	1.5254
	Year 9	1.4503	0.1624	3.32	0.001	1.1643	1.8064
	Year 8 or below	1.1046	0.1215	0.90	0.366	0.8904	1.3703
ARIABC	Inner regional	1.1491	0.0739	2.16	0.031	1.0130	1.3034
	Other	1.1523	0.0805	2.03	0.043	1.0047	1.3215
Smoker	No	1.2037	0.0773	2.89	0.004	1.0614	1.3650
Constant		1.5197	0.1775	3.58	0.000	1.2087	1.9108
				(2) Weighted			
Sex	F	0.5091	0.0337	-10.19	0.000	0.4470	0.5797
Age		1.0262	2.378E-3	11.15	0.000	1.0215	1.0308
PhysActMin		0.9996	1.145E-4	-3.90	0.000	0.9993	0.9998
SchEd	Year 11	1.4171	0.1624	3.04	0.002	1.1320	1.7741
	Year 10	1.3097	0.1154	3.06	0.002	1.1020	1.5565
	Year 9	1.4918	0.2242	2.66	0.008	1.1111	2.0030
	Year 8 or below	1.2022	0.1760	1.26	0.208	0.9023	1.6017
ARIABC	Inner regional	1.3775	0.1203	3.67	0.000	1.1608	1.6346
	Other	1.3156	0.1429	2.52	0.012	1.0632	1.6278
Smoker	No	1.0640	0.0939	0.70	0.482	0.8951	1.2649
Constant		1.2413	0.1980	1.35	0.175	0.9080	1.6969
				(3) Complex design			
Sex	F	0.5091	0.0371	-9.26	0.000	0.4399	0.5890
Age		1.0262	2.705E-3	9.80	0.000	1.0208	1.0316
PhysActMin		0.9996	1.174E-4	-3.80	0.000	0.99938	0.9998
SchEd	Year 11	1.4171	0.1834	2.69	0.009	1.0939	1.8360
	Year 10	1.3097	0.1372	2.58	0.013	1.0621	1.6150
	Year 9	1.4918	0.2377	2.51	0.015	1.0846	2.0520
	Year 8 or below	1.2022	0.1811	1.22	0.227	0.8893	1.6251
ARIABC	Inner regional	1.3775	0.1287	3.43	0.001	1.1426	1.6606
	Other	1.3156	0.1460	2.47	0.016	1.0536	1.6427
Smoker	No	1.0640	0.0929	0.71	0.480	0.8935	1.2671
Constant		1.2413	0.2217	1.21	0.231	0.8683	1.7745

Notes: Reference categories: for *Sex* is Male; for *SchEd* is Year 12 or equivalent; for *ARIABC* is Major city; for *Smoker* is Yes.