The University Of Sheffield.

Methods to process the Health Survey for England data for the modelling of policy effects on tobacco and alcohol consumption

Version 1.1.0

May 2021

Duncan Gillespie, Colin Angus, Laura Webster & Alan Brennan

School of Health and Related Research (ScHARR), The University of Sheffield

Address for correspondence:
Dr Duncan Gillespie
Section of Health Economics and Decision Science,
School for Health and Related Research,
The University of Sheffield,
Regent Court, Regent Street, Sheffield, S1 4DA, UK
Email: duncan.gillespie@sheffield.ac.uk

WARNING

This is a working version that is subject to review and future updated and extended versions are likely.

Summary

This methodology report sets out a description of how the data related to tobacco and alcohol consumption in the Health Survey for England are processed for use in the Sheffield Tobacco and Alcohol Policy Modelling (STPM). The report covers: (1) the processing of alcohol consumption data; the processing of tobacco consumption data; the processing of socio-demographic covariates relevant to our tobacco and alcohol modelling; the imputation of missing data.

1 Background

The Sheffield Alcohol Policy Model (SAPM) has been used to examine the effects of pricing policies, advertising restrictions and advice on why and how to reduce drinking (**???**; **???**) (see the range of publications and projects on the Sheffield Alcohol Research Group website). Patterns of alcohol consumption are informed primarily by the Health Survey for England data.

2 Approach to coding

2.1 Data storage and management

The data is stored in `X:/ScHARR/PR_Consumption_TA/Data/`. The following code will read, clean, filter and combine the data.

3 Alcohol consumption

Alcohol consumption data in the Health Survey for England (HSE) is recorded in four main forms:

- How often someone usually drinks
- For adults considering the last 12 months, what they drink on average
- For adults considering the last week, when they drank the most
- For children considering the last week, what they drank

Both adults and children have data on whether they drink alcohol or not, and on the frequency of drinking. The main difference between the recording of data for adults and children is that adults have a lot of data on how much and what they drink, but children only have data on the amount drunk in the last week.

The recording of data varies among years of the HSE. We consider years from 2001 onwards. The main features of these changes in recording are:

- Adult drinking in the last 12 months is only recorded for years 2001, 2002, and 2011 onwards.
- Adult drinking in the last 12 months is recorded in different terms for 2001/2 and 2011+.
- In 2007, the way that wine was recorded changed from asking how many glasses (with a size of 125ml assumed) to asking how many glasses of either 125ml, 175ml or 250ml. Therefore the post HSE 2007 unit calculations are not directly comparable to previous years' data.

Due to the variability in recording, we only consider data on the amount drunk by adults and children from 2011 onwards.

We analyse beverage-specific alcohol consumption in terms of beer (combining normal beer, strong beer), wine (combining wine and sherry), spirits, and alcopops.

3.1 Whether someone drinks and frequency of drinking

Calculated for adults (aged 16 years or older) and children (aged 8 to 15 years) by the function `alc_drink_now_allages()`. We combine the information on drinking frequency from adults and children into a single variable.

We calculate the variable `drinks_now`, which classes someone as either a drinker or a non-drinker. Adults are classed as drinkers if they reported drinking at all in the last 12 months, even if reporting only having 1-2 drinks a year (according to the variable `dnoft`). Note that this definition of a non-drinker can vary among surveys, e.g. some surveys class only having 1-2 drinks a year as a non-drinker, and this could lead to variation in estimates of the number of non-drinkers.

We calculate the variable `drink_freq_7d`, which is a numerical variable that described drinking frequency. Adult drinking frequency is also inferred from the variable `dnoft`: the function `alc_drink_freq()` converts the categorical responses into the expected number of days in a week that someone drinks.

- "Almost every day" = 7 days a week
- "Five or six days a week" = 5.5 days a week
- "Three or four days a week" = 3.5 days a week
- "Once or twice a week" = 1.5 days a week
- "Once or twice a month" = 0.375 days a week
- "Once every couple of months" = 0.188 days a week
- "Once or twice a year" = 0.029 days a week

Missing data on whether or not someone currently drinks (`drinks_now`) is supplemented by responses to if currently drinks or if always non-drinker (the variables `dnnow`, `dnany` and `dnevr`).

For children (aged 8-15 years) we infer whether someone drinks or not (`drinks_now`) from the variable `adrinkof`. Someone is a non-drinker if they responded `never` to `adrinkof`. The categorical responses are converted into the expected number of days in a week that someone drinks as follows

- "Almost every day" = 7 days a week
- "Twice a week" = 2 days a week
- "Once a week" = 1 days a week
- "Once a fortnight" = 0.5 days a week
- "Once a month" = 0.25 days a week
- "Only a few times a year" = 0.058 days a week

Missing data on whether or not a child currently drinks (`drinks_now`) is supplemented by responses to when they last had an alcoholic drink (`adrlast`): if the last drink was less than six months ago, then we classify them as a drinker; if the last drink was six months or more ago, then we classify them as a non-drinker.

3.2 Average amount of alcohol consumed

3.2.1 Assumptions about serving size and alcohol content

Some standard assumptions are made about the volume and alcohol content of the beverages that are reported to be drunk. The values that we use for these assumptions are based on those used by Natcen to create the derived variables for units of alcohol consumed in the HSE. We have made our own adjustments to the values used based on further information from market research data and figures from academic publications.

Alcohol content assumptions are the expected percentages of alcohol that each beverage contains (alcohol by volume, ABV). We use separate values for normal beer (4.4%), strong beer (8.4%), spirits (38%), sherry (17%), wine (12.5%), and alcopops (also known as "ready to drink" or RTD) (4.5%).

Beverage volume assumptions are the expected volumes (ml) of different beverage containers / serving sizes. We use separate values for normal and strong beer (half pint 284ml, small can 330ml, large can 440ml, bottle 330ml), spirits (serving 25ml), sherry (serving 50ml), wine (small glass 125ml, standard glass 175ml, large glass 250ml, bottle 750ml), and alcopops (small can 250ml, small bottle 275ml, large bottle 700ml).

3.2.2 Adult average weekly consumption in the last 12 months

We estimate the average amount drunk in a week (`weekmean`) in terms of UK standard units of alcohol (1 unit = 10ml or 8g pure ethanol). The average amount drunk is then categorised as follows:

- `abstainer` = 0 units/week
- `lower_risk` drinker = less than 14 units/week
- `increasing_risk` drinker = 14 or more units/week but less than 35 units/week for females or less than 50 units/week for males
- `higher_risk` drinker = 35 or more units/week for females or 50 or more units/week for males

Separate variables are produced describing the average weekly units in four beverage categories: `beer_units` (including cider), `wine_units` (including sherry), `spirit_units`, `rtd_units` (this is alcopops). Further variables on beverage preference are produced that:

- describe the percentage of the total consumption in a week that is contributed by each beverage type (`per_spirit_units`, `perc_wine_units`, `perc_beer_units`, `perc_rtd_units`).
- describe whether or not someone shows a clear beverage preference e.g. `does_not_drink_spirits`, `drinks_some_spirits`, `mostly_drinks_spirits`, where "mostly drinks" is defined by a single beverage comprising more that 50% of an individuals average weekly consumption.

The processing is done by the function `alc_weekmean_adult()`. The calculation has the following steps:

- Convert the categorical variables to numeric variables for the frequency with which each beverage is typically consumed (normal beer, strong beer, spirits, sherry, wine, alcopops).
- Convert the reported volumes usually consumed (e.g. small glass, large glass) into volumes in ml, using the beverage size assumptions above. In doing so, variations in recording among years and between the interview and self-complete questionnaire are accounted for.

- Combine the volumes (ml) usually consumed with the frequency of consumption to give the average volume of each beverage type drunk each week (assuming constant consumption across the year).
- Convert the expected volumes of each beverage consumed each week to UK standard units of alcohol consumed, using the alcohol content assumptions above.
- Collapse normal and strong beer into a single "beer" variable by summing their units. Collapse wine and sherry into a single "wine" variable by summing their units.
- Calculate total weekly units but summing across beverage categories.
- Cap the total units consumed in a week at 300 units, assuming that above this already very high level of consumption estimates of variation in consumption are less reliable.

3.2.3 Adult consumption on the heaviest drinking day in the last week

The function `alc_sevenday_adult()` processes the information from the questions on adult (16 or more years old) drinking in the last seven days:

- Number of days drank on in the last seven, `n_days_drink`.
- the characteristics of drinking on the heaviest drinking day

We estimate the number of UK standard units of alcohol drunk on the heaviest drinking day (`peakday`) by using the data on how many of what size measures of different beverages were drunk, and combining this with our standard assumptions about beverage volume and alcohol content. We further estimate their total units drunk of each beverage type on the heaviest drinking day (`d7nbeer_units`, `d7sbeer_units`, `d7spirits_units`, `d7sherry_units`, `d7wine_units`, `d7pops_units`).

Binge drinking status is then categorised into the variable `binge_cat`, with levels `did_not_drink`, `binge` and `no_binge`, where a binge day in defined by males drinking over 8 units and females drinking over 6 units.

Note that in 2007 new questions were added asking which glass size was used when wine was consumed. Therefore the post HSE 2007 unit calculations are not directly comparable to previous years' data.

Missing data is imputed using the means of people who did drink in the last seven days, stratified by year, sex, IMD quintile and age category (0-1, 2-4, 5-7, 8-10, 11-12, 13-15, 16-17, 18-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85-89, 90+).

3.2.4 Children's consumption in the last week

The function `alc_sevenday_child()` processes the information on drinking by children (ages 13-15) in the last seven days. The data on children's drinking comes in the form of survey questions on whether or not they have drunk each beverage type in the last week, and if so, how much of each was drunk. The main output is the variable `total_units7_ch` - the total units drunk in the last seven days.

We estimate the number of UK standard units of alcohol drunk in the last 7 days by using the data on how many of what size measures of different beverages were drunk, and combining this with our standard assumptions about beverage volume and alcohol content.

The information from this question is also used to update the `drinks_now` variable to make it a variable that describes whether or not adults and children drink.

Due to high missingness in this variable, we assume that anyone who has missing data for this variable does not drink. This means that we are likely to under-estimate the number of children who drink.

4 Tobacco consumption

For the Sheffield Tobacco Policy Model (STPM) we use HSE data from years 2001 to the latest available. We use these data to inform the trends in smoking prevalence, the socio-demographic variation in smoking prevalence, and as inputs to a procedure that we use to infer the age-specific probabilities of smoking initiation and quitting (see our smktrans R package). Our upper age limit is 89 years, but otherwise we make use of all ages.

The purpose of this vignette is to explain how we use the HSE data to inform the patterns of tobacco smoking, and to explain how `hseclean` supports this.

Questions about cigarette smoking have been asked of adults aged 16 and over as part of the HSE series since 1991 - we use data from 2001 to the latest year available. We use data on children (12-15 years) and adults (16+ years). There is often a special section in the annual HSE report devoted to describing trends in cigarette smoking e.g.HSE 2015.

4.1 Cigarette smoking status

The function `smk_status()` categorises cigarette smoking into current, former and never regular cigarette smokers. If some smokes either regularly or ocassionally, then they are classified as a current regular cigarette smoker. People who used to smoke regularly or ocassionally are classified as former smokers; people who have only tried a cigarette once or twice are classified as never smokers. We create a smoking status variable for children aged 8-15 years and adults aged >= 16 years. Ever-smokers are people who are either current or former smokers.

4.2 Quitting

The function `smk_quit()` is in development, and will process the data on the motivation to quit smoking, the reasons for quitting smoking, and the support used to stop smoking. It currently produces only one variable - whether someone wants to quit smoking (y/n).

4.3 Former smoking

The function `smk_former()` cleans the data for former smokers on the time since quitting and time spent as a regular smoker. The main issue to overcome is that in the HSE 2015+, time since quit and time spent as a smoker is provided in categories rather than single years. We simulate the single years by just picking a value at random within the time interval, using `num_sim()`. We then fill missing data for these variables as follows:

- For children 8-15 years, we assume that missing values for former smokers' time since quitting and time spent as a former smoker = 1 year.
- For adults, we fill missing values for former smokers' time since quitting and time spent as a former smoker with the average value for each age, sex and IMD quintile subgroup.

4.4 Smoking life-histories

The function `smk_life_history()` cleans the data on the ages when smokers started and stopped being regular cigarette smokers. For each individual smoker, the data recorded in the HSE implies a single age at which a smoker started to smoke and, if they stopped, an age at which they did so. This provides a simplified view of what might be a complicated life history of smoking, e.g. smoking to different frequencies or levels, or starting and stopping multiple times.

Both the start age and stop age will have error in them e.g. due to uncertainty in respondent recall, and, for years 2015+, due to the reporting in categories of time intervals rather than single years, which we then impute introducing random error. Start age is likely to be biased towards earlier ages, because for adult smokers and former smokers with missing values we use the age first tried a cigarette, and for children the reported start age does not necessarily mean the start of regular smoking, it is just the age at which they started to smoke.

We also create a variable for the age at which an individual was censored from our data sample - this is their age at the survey + 1 year.

Any missing data is assigned the average start or stop age for each age, sex and IMD quintile.

4.5 Amount and type of cigarette smoked by current smokers

The function `smk_amount()` cleans the data that describe how much, what and to what level of addiction people smoke. The main variable is the average number of cigarettes smoked per day. For adults, this is calculated from questions about how many cigarettes are smoked typically on a weekday vs. a weekend (this is a weighted average to account for more weekdays in a week than weekends). For children, this is based on asking how many cigarettes were smoked in the last week. Missing values are imputed as the average amount smoked for an age, sex and IMD quintile subgroup.

We categorise cigarette preferences based on the answer to 'what is the main type of cigarette smoked'. For years 2013 and later of the HSE, questions were added that ask how many handrolled vs. machine rolled cigarettes are smoked on a weekday vs. a weekend.

We also categorise the amount smoked, and use information on the time from waking until smoking the first cigarette (this latter variable has a high level of missingness). Together these two variables allow calculation of the heaviness of smoking index.

4.6 Summarise data

Taking the survey design into account is important when estimating the mean and confidence intervals around summary statistics computed from the data i.e. it is not possible to accurately estimate sampling error without accounting for survey design. The `survey` R package (Thomas Lumley 2019) has a collection of functions that incorporate survey design into the calculation of summary statistics. The `survey` package is used by the function `prop_summary()` in `hseclean` to estimate the uncertainty around proportions calculated from a binary variable - `prop_summary()` was designed to simplify the process of estimating smoking prevalence from the HSE data, stratified by a specified set of variables.

Using `prop_summary()`, calculate the proportion of smokers, stratified by year, sex and quintiles of the Index of Multiple Deprivation.

5 Covariate data

The Health Survey for England (HSE) is a series of annual surveys covering health and health-related behaviours. For the Sheffield Tobacco Policy Model (STPM) we use data from years 2001 to the latest available. Our upper age limit is 89 years, but otherwise we make use of all ages.

One important thing to note is that the suppliers of the HSE data introduced tighter information governance rules in 2015, which meant that they stopped providing variables that could be used to identify the age in single years of an individual, and also stopped providing information on number of children in the household. These variables can still be obtained, but only after applying for the secure-access version of the data, which we do not do. Therefore, in our processing of the standard-access version of the data, we use imputation methods to overcome the added restrictions.

`hseclean` is a collection of functions to read and process the HSE data into a suitable form for use in our modelling. Here we describe how we use it to clean and calculate the covariates used in our analyses.

5.1 Survey design variables

The first thing to consider is the influence of survey sampling design, which is variable among years. The variables that describe the sampling structure are `cluster` and `PSU` (probabilistic sampling unit).

In most years there are also survey weights, which are calculated after the survey data has been collected, that when applied are supposed to make the survey sample representative of the general population e.g. if a particular subgroup has been under-sampled, then it receives a higher survey weight. As we understand the HSE methods, the survey weights supplied with the data consider only the age and sex distribution of the population, and do not consider the distribution of socio-economic or health characteristics. The definition and structure of the survey weights provided with the data varies between years, and is described in the dataset documentation for each year of data. For example, some key changes

- For 2001, there were not survey weights for adults but there were for children (to correct for the sampling design that not all children in the household being surveyed).
- For 2002, there were different weights for children, young adults (< 25 years) and older adults. These weights again were just to correct for the sampling design.
- In 2003, non-response weighting was introduced to the HSE data for children and adults.
- Thereafter, weights made the additional corrections for the various boost samples in each year.

`hseclean` contains separate functions for reading the survey data for each year, e.g. `read_2001()`, and a description of the survey weights has been added to the help files of those functions. Any processing or combining of survey weights is done in the functions that read each year of data. The function `clean_surveyweights()` assigns any missing weights the average weight for each year, and standardises the weights to sum to 1 within each year. The resulting survey weight variable for each year is `wt_int`.

5.2 Age

From 2015 onwards, the HSE no longer supplies age in single years (to prevent individual identification). For our modelling, we require age in single years, so we apply a method that randomly assigns an age in single years to individuals for who we only have an age category. The age categories we work with are: 0-1, 2-4, 5-7, 8-10, 11-12, 13-15, 16-17, 18-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80-84, 85-89, 90+. These categories are the finest scale version of age that is available for years 2015+. We then select only individuals younger than 90 years for our modelling.

This processing is done by the function `clean_age()` that calls the function `num_sim()` to simulate single years of age. For years 2015+, we also use `num_sim()` to convert the categorical variables for years since quitting smoking and years spent as a smoker to single years of age.

5.3 Other demographic variables

The function `clean_demographic()` creates variables for ethnicity, sex and quintiles of the Index of Multiple Deprivation (IMDq).

5.3.1 Sex

1 = Male, 2 = Female.

5.3.2 IMD quintiles

5_most_deprived, 4, 3, 2, 1_least_deprived.

5.3.3 Ethnicity

Previous SAPM modelling has used a simple white/non-white classification. The ONS recommend a harmonised ethnicity measure for use in social surveys (ONS, 2017). The use of ethnicity measures is also discussed in Connelly et al. 2016, who recommend testing the sensitivity of analyses to different specifications. We try to map the HSE categories to the ONS recommended groups for England. However, over the years, the HSE is not clear or consistent in how they have categorised chinese and arab as 'asian' or 'other'. In an attempt to harmonise, we have pooled the asian and other categories.

- White (English, Irish, Scottish, Welsh, other European)
- Mixed / multiple ethnic groups
- Asian / Asian British (includes African-Indian, Indian, Pakistani, Bangladeshi), plus Other ethnic group (includes Chinese, Japanese, Philippino, Vietnamese, Arab)
- Black / African / Caribbean / Black British (includes Caribbean, African)

Following inspection of the data, the white/non-white classification does look appropriate, especially given the likely limited sample sizes - so the 2 level variable has also been created. Previous Sheffield modelling in the Sheffield Alcohol Policy Model has also used the white/non-white classification.

5.4 Townsend quintiles of deprivation

Individuals in the HSE are not assigned a Townsend quintile of deprivation, but for a project that investigated the cost of alcohol to primary care in England, we needed to predict the Townsend quintile of each individual so that we could use it to stratify our summary of alcohol consumption.

The function `use_townsend()` adds a Townsend variable to the data. It produces a version of the Health Survey for England data that has the Townsend Index in it, based on the probabilistic mapping between the 2015 English Index of Multiple Deprivation and the Townsend Index from the 2001 census.

It does so based on a matrix (stored in `hseclean::imdq_to_townsend`) that maps quintiles of the Index of Multiple Deprivation onto the Townsend Index of Deprviation. To produce this we used area-level Office for National Statistics data to estimate the statistical association between the two metrics of deprivation. We used estimates of the Townsend Index from 2001 Census data at Ward level, and the Index of Multiple Deprivation 2015 (IMD 2015) at Lower-layer Super Output Area (LSOA) level. First, we mapped the 2001 definitions of Wards to the 2001 definitions of LSOAs. Second, we mapped the 2001 definitions of LSOAs to the 2011 definitions of LSOAs that are used by the IMD 2015.

5.5 Economic status

The function `clean_economic_status()` creates a variety of variables to classify economic status.

The issues around using occupation-based social classifications for social survey research are discussed by Connelly et al. (2016). They advise using a range of alternative measures, and not creating new measures beyond what is already established.

The classifications considered are:

- Employed / in paid work or not.

- The NS-SEC measure which was constructed to measure the employment relations and conditions of occupations (i.e. it classifies people based on their employment occupation). It is therefore not that good at classifying people who are not employed for various reasons.

- The NRS social grade system. This measure is the one used in the Tobacco and Alcohol Toolkit studies, but is not reported in the Health Survey for England. We create this variable by recategorising the NS-SEC 8 level variable. This is important to facilitate the link of analysis to the Toolkit Study.

- Manual vs. non-manual occupation. In the 2017 Tobacco control plan for England, there was a specific target to reduce the difference in rates of smoking between people classified with a manual or non-manual occupation. We create this variable from the 3 level NS-SEC classification by grouping Managerial and professional with intermediate occupations to give the non-manual group.

- Economic status - retired / employed / unemployed.

- Activity status for last week that adds more detail such as 'in education' and 'looking after home or family'.

5.6 Education

The main education variable produced by the function `clean_education()` is a four category description of the age at which someone finished full-time education. The categories are:

- never went to school,

- left at 15 years or younger,

- left at 16-18,

- left at 19 years or over.

If someone was still in full time education at the time of the survey, then if they were younger than 18 years, we assumed they would leave at 16-18, and if they were older than 18 years, we assumed they would leave at 19 years or over.

A further education variable is also produced - which indicates whether an individual reached a degree as their top qualification or not. Here a degree is defined as an "NVQ4/NVQ5/Degree or equiv".

5.7 Family

The function `clean_family()` processes the data on the number of children in the household and the relationship status of each respondent.

5.7.1 Number of children in the household

`kids` is the number of children aged 0-15 years who live in the household. If a 3 year old lives in a household with 2 siblings, aged 6 and 8 years, then we might expect them to be recorded as living in a household with 3 children under age 15 years. The variable is created by combining the HSE data on children and infants in the household. It is categorised into: 0, 1, 2, 3+ children under age 15 years.

The problem with the Health Survey for England is that from 2015 onwards, the number of children in the household is not provided as this information could be identifiable (you can get it if you apply and pay for a

secure dataset). Therefore, for years 2015+, the number of children in the household is completely missing and needs to be imputed.

We impute the number of children for years 2015+ automatically in the function `clean_family()`, based on the correlation between the number of children and a range of demographic and socioeconomic variables in 2012-2014, the last three years for which data on kids is available. This imputation is based on the fit of a multinomial model in `package(nnet)`. The model object is saved in the `hseclean` package as the object `hseclean::impute_kids_model`, and is drawn upon by the `clean_family()` function to impute the data as needed. This imputation won't work unless the required demographic and socio-economic variables have already been cleaned prior to running `clean_family()`. There will still be missing values in `kids` if there are missing values in the predictor variables required by the model. These missing values can be taken care of in a multiple imputation procedure (see vignette(missing_data)).

5.7.2 Relationship status

In previous versions of modelling for the Sheffield Alcohol Policy Model, relationship status has been described as married/not-married. Here, we include more detail by using:

- single
- married, civil partnership or cohabiting
- separated, divorced, widowed

5.8 Income

The function `clean_income()` processes the data on income.

There are a few different options for classifying income - the need to have a measure that is consistent across years of the Health Survey for England has led us to use equivalised income quintiles only. (Past SAPM modelling has used years of the HSE for which a continous variable for equivalised income was provided - and calculated our own income groups - but in later years, this continuous income variable is not available.)

In the past SAPM modelling, a measure of in "poverty" vs. "not in poverty" has been used, where the poverty threshold is defined as 60% of the median income for any year. For years in which we only have income quintiles available, it is not possible to make an exact calculation of poverty, but being in poverty will coincide approximately with the lowest 2 income quintiles.

It would also be possible from the Health Survey for England to classify people as being in receipt of benefits or not, but this is not currently implemented in `hseclean`, and would have to have some thought on how to deal with the changing definitions of benefits over time.

5.9 Health and biometric variables

The function `clean_health_and_bio()` cleans data on presence/absence of certain categories of health condition, and on height and weight.

5.9.1 Health conditions

There are a set of 15 categories of long-lasting illnesses (occurring for or expected to last at least 12 months) that are ascertained consistently across all years of the HSE. These are:
- Cancer
- Endocrine or metabolic condition

- Mental health condition
- Nervous system condition
- Eye condition
- Ear condition
- Heart or circulatory system condition
- Respiratory condition
- Digestive condition
- Genito-urinary condition
- Skin condition
- Musculo-skeletal condition
- Infectious disease
- Blood and related organs condition
- Other complaints

5.9.2 Height and weight

Height (cm) and weight (kg). Weight is estimated above 130kg. Missing values of height and weight are replaced by the mean height and weight for each age, sex and IMD quintile. BMI is calculated according to kg / m^2.

6 Missing data

To prepare for the process that imputes missing data, run the full set of functions to read and clean the data. It is important to note that there has already been some filling-in of missing data done by these cleaning functions - using simple rules,

- For 2015+, the `clean_age()` function has randomly assigned single years of age within each age category.
- If someone is classified as a current smoker, is younger than 16 and has missing data on the time between waking and having their first cigarette, we assume that this time is one hour or more.
- If someone is younger than age 16 and has missing data on their number of children (for survey years prior to 2015), we assume that they had no children.
- If someone of any age is classified as a current smoker but has a missing value for the amount smoked, then we fill that missing value with the average amount smoked within a year, age, sex and IMD quintile subgroup.
- Information on education, employment and socioeconomic status is triangulated across several variables.

The number of children in the household is missing for years 2015+. This is imputed in the function `clean_family()` based on the fit of a multinomial model to years 2012-2014 (see `vignette("covariate_data")`).

The function `select_data()` has the option to filter the data to retain only complete cases for certain variables. To prepare the data, the example code below filters out any incomplete data on key survey variables (age, sex, year, quarter, psu, cluster, imd_quintile). It also filters out any incomplete information on the key smoking and drinking variables, “cig_smoker_status” and “drinks_now”.

7 Multiple imputation

The variable with the most missingness in the data is income5cat (19% missing). In this example, the other variables to be imputed are: kids, ethnicity_4cat, eduend4cat, degree, relationship_status, nssec3_lab,

activity_lstweek.

To conduct the multiple imputation, we use the R package `mice` (Stef van Buuren and Karin Groothuis-Oudshoorn 2011). The process of running the multiple imputation can take a long time and consume a lot of RAM. There is a range of `mice` documentation and tutorials online.

In `hseclean`, multiple imputation is implemented in a basic way by the `impute_data_mice()` function.

`mice` fits a chained series of regression equations that predict the missing values of variables based on their relationships with other selected variables in the data. The `impute_data_mice()` function currently only imputes categorical variables, which could be one of three types: "logreg" - binary Logistic regression; "polr" - ordered Proportional odds model; "polyreg" - unordered Polytomous logistic regression.

In running the multiple imputation, the number of iterations of the imputed data is selected (choosing a small number e.g. < 5 helps keep the size of the resulting imputed data manageable), and the variables to either be predicted or to inform the prediction are selected. If a variable is just going to inform the prediction of the other variables but is not going to be predicted itself, then the model type is set to "", otherwise to one of"logreg","polr" or "polyreg".

References

Connelly, Roxanne, Vernon Gayle, and Paul S. Lambert. 2016. "A Review of Occupation-Based Social Classifications for Social Survey Research." Journal Article. *Methodological Innovations* 9. https://doi.org/10.1177/2059799116638003.

Stef van Buuren and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45 (3): 1–67. %7Bhttps://www.jstatsoft.org/v45/i03/%7D.

Thomas Lumley. 2019. *survey: analysis of complex survey samples.*