DOUBLE DEGREE PROGRAMME IN STATISTICAL SCIENCES

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA
DEPARTMENT OF STATISTICAL SCIENCES "PAOLO FORTUNATI"
First Cycle Degree in Statistical Sciences
Curriculum Stats & Maths

UNIVERSITY OF GLASGOW
SCHOOL OF MATHEMATICS AND STATISTICS
Bachelor of Science with Honours in Statistics

MODELLING ALCOHOL INTAKE WITH THE NATIONAL DIET AND NUTRITION SURVEY FOR THE UK ADULT POPULATION

Presented by:
Davide Perez
000097218

Supervisor:
Dr. Claus-Dieter Mayer
Co-Supervisor:
Prof. Angela Montanari

Summer Graduation Session
Academic Year 2022-2023

Contents

1 INTRODUCTION

The abuse of alcohol has critical effects on the health of a human being, given its substantial role in irreversibly damaging living organs like the liver and the cardiovascular system. It may be helpful then to be able to detect what might be associated to different levels of alcohol consumption and, hopefully, being able to estimate how it changes in a population provided a set of covariates.

In this report, a statistical analysis will be carried out to investigate what are the potential driving factors and to figure out suitable models to predict the alcohol intake distribution for UK residents. To start off, there will be a presentation of the data source and an overview of the involved sampling techniques (SECTION 1.1), followed by a neat explanation of the outcome of interest (SECTION 1.2) and an emphasis on the aim of the project (SECTION 1.3).

1.1 National Diet and Nutrition Survey (NDNS)

Assessing the dietary habits, sustenance and nutritional status among people in a nation is crucial in order to make considerations about public health. The UK government boasts the **National Diet and Nutrition Survey (NDNS)** as a tool to fulfil this very duty.

Developed as a 'continuous rolling programme in the form of a cross sectional study', details about food and drink intakes are gathered for citizens living in private household across Great Britain and Northern Ireland aged over 1.5 years (*Public Health England, 2016*). Having started in 2008 under the current guidelines and ongoing since then, every year the UK government publishes the results which have been the core evidence employed to launch targeted nutritional policies and to keep track of changes in dietary behaviours (*MRC Epidemiology Unit, n.d.*).

The collected data comprises quantitative records of food and nutrients, blood and urine samples, plus exhaustive background information about the individuals so that insights on the relationship between social-economic features of the population and general diet could be provided.

Each yearly cycle, a representative sample of around 1000 participants is taken, evenly split between adults and children and designed as a follow-up study mostly in four consecutive days. The daily measurements relies on self-report which might seem prone to bias and measurement errors, however observing for multiple days accounts for day-to-day variation.

The sampling approach makes use of a clustered stratified design, plus an intensive fieldwork model and a weighting strategy in order to adjust for issues like seasonality and to enhance the quality of representativeness for the UK general population. The complex plan behind the sampling strategy is thoroughly examined in Venables et al., 2022.

1.2 Modelling Alcohol Intake (AI)

As stated before, the aim of the project is to appropriately model alcohol consumption in the UK adult population. Therefore, it is fundamental to consider how to accomplish this goal according to the available data.

Information on drinking habits was obtained through the use of interviews and the dietary diary of subjects where they specifically recorded the consume of wine, beer, liquors and equivalents (*Henderson et al., 2003*). Additionally, the percentage of total energy derived from alcohol has also been reported, likewise for the food.

Given that a significant number of respondents did not regard themselves as drinkers or did not record any AI in the dietary record, statistics are presented whether including non-consumers (i.e. the total sample) or excluding them (for consumers only).
Differences might be observed between subgroups of the population as clear distinctions could be caught for sex, region and age clusters, hence they gain prominence as indispensable predictors for the model.

The knowledge about drinking behaviours is collected in many different ways in the NDNS:

- Dual distinction between drinkers and non-drinkers (**dichotomous variable**) ;
- Categorisation of usage in scale with ordered levels like none, low, medium or high intake (**polytomous variable**) ;
- Exact quantities expressed both in grams and as a percentage of energy taken from the diary (**continuous framework**). By the nature of this phenomenon, many zeroes are expected to be observed in the distribution of AI.

The various formats for the response lead to many options for the statistical models:

A. **Binary Logistic Regression** in the case of dichotomous outcome;
B. **Ordinal Regression** to deal with categorical responses having more than two levels and a meaningful inherent order;
C. While the above two methods have the desirable advantage of keeping the analysis inside the well-known framework of *Generalised Linear Models (GLM)*, it also results in a loss of information. Indeed, besides the extensive presence of null quantities in the AI, the rest of the data is continuously spread, thus a **Mixture Model** would allow to exploit the available continuous data - energy and grams intakes - after adjusting for the zero inflation.

All in all, more than one strategy can be explored that leads towards a variety of possible models to fit.

1.3 Aims of the study

It was just mentioned that multiple models are going to be carried out for the gathered variables, and an evaluation of each of them is due. In particular, it is beneficial to highlight the strengths, frailties, adherence and goodness of each one in the context of the data from the survey.
At the end of the study a comparison between the models will be outline in order to address the following questions:

- What are the pros and cons of the different modelling approaches to alcohol intake (binary, ordinal, mixture model with mass in zero)?
- What are the key variables associated with AI and potential drivers of it?
- What could be improved about the research?
- Which are the suggestions for further work?

2 DATA AND METHODOLOGY

The NDNS records can be accessed from the government site. This project will take into consideration the findings from year 1 to 9 (from 2008/09 to 2016/17).
After an in-depth look on the variables that are going to be used for the analysis (SECTION 2.1), the applied models will be introduced (SECTION 2.2).

2.1 Data set and variables

The survey data is gathered in three ways:

- Individual intake data on daily level. It comprises the subjects' diary intakes during the four consecutive days;
- Individual intake data summarised across the 4 days. The dietary records are averaged for each subject;
- Background information about the individual. It includes sensitive, social-demographic and economic data.

The observations are merged across all the stages and tidied up in order to obtain unique data set, moreover underage participants are filtered out of the study since the target is the adult population. The selected variables are summarised in Table 2.1, while the complete list with full description can be consulted in *UK Data Year 1-4; Year 5-6; Year 7-8; Year 9*.

A few notes on the meaning of these variables:

- *seriali* will not be considered when fitting the model, it is just needed to identify the units during the preliminary stages so that each covariate is matched to the same subject.
- *Alcoholg* is actually the summarised AI of the follow-up over consecutive days.
- *pregnancy* is a variable stating whether the subject is pregnant or breastfeeding at present. However, from now only the former word will be used to identify this combination for the sake of simplicity.
- The geographic areas denoting *region* are "ENGLAND: NORTH", "ENGLAND: CENTRAL/MIDLANDS", "ENGLAND: SOUTH", "SCOTLAND", "WALES", "NORTHERN IRELAND".
- *start_day* comes into play as long as *Alcoholg* is the outcome of interest, given that both variables are taken from the individual dietary information while all the others are part of the background.
- The adjective 'valid' is associated to *height, weight, BMI* because some measurements may be unreliable. Further details on those will be given throughout CHAPTER 3.

2.2 Models to fit

Now, the aforementioned models will be thoroughly outlined. The theoretical descriptions aid to understand the notions carried by each model and will be the rudiments to evaluate the validity of the analysis at the final stage, besides an overview of the application in the context of the AI data set.

VARIABLE NAME	TYPE	DETAILS
seriali	chr	Individual Identifier
drink.binary	chr	Status: drinker / non-drinker
drink.ordinal	chr	Frequency alcoholic beverage last 12 months, 8 levels: "Almost every day" "5-6 per week" "3-4 per week" "1-2 per week" "1-2 per month" "Once every two months" "1-2 per year" "Not at all / Non-Drinker"
Alcoholg	num	Alcohol intake in grams (g)
age	int	Age of the individual in years
sex	chr	Male / Female
pregnancy	chr	Status: pregnant / non-pregnant
region	chr	Country of residence
start_day	chr	Day of the week in which the follow-up started
eq_income	num	Equivalised Household Income
height	num	Valid height measurement (cm)
weight	num	Valid weight measurement (kg)
BMI	num	Valid Body Mass Index (kg/m^2)

Table 2.1: *Selected variables in the data set; the type abbreviations in the second column stand for: character strings (chr), numeric (num), integers (int).*

2.2.1 Binary Logistic Model

The dependent variable admits two categories: DRINKER or NON-DRINKER (as a reply to the question 'whether drinking nowadays'). Let y_i be the alias for the response which takes value 1 if the subject is a drinker, 0 otherwise:

$$y_i = \begin{cases} 1 & \text{if unit } i \text{ is a drinker} \\ 0 & \text{otherwise.} \end{cases}$$

The y_i's are realisations of a **Bernoulli Random Variable** Y_i defined as (*Mann, 2012*):

$$Y_i \sim \text{Ber}(\pi_i) \qquad \text{where} \quad \mathbb{P}(Y_i = 0) = 1 - \pi_i \quad ; \quad P(Y_i = 1) = \pi_i \tag{2.2.1}$$

$\pi_i \in [0, 1]$ is the probability of being a drinker for an adult person with the same features as the ones observed in the i^{th} unit.
Furthermore, it follows from the theoretical distribution of [2.2.1] that $\mathbb{E}[Y_i] = \pi_i$.

The model aims at predicting the π_i's for all the potential combinations between the covariate values. However, a linear regression cannot be effective in this case because the responses are bounded in [0,1] them being probabilities.
Therefore, a **binary logistic regression** is implemented, giving a sigmoid shape to the response function.

The following assumptions must be made *(Kutner et al., 2005; Hosmer and Lemeshow, 2000)*:

1. $Y_i \sim \text{Ber}(\pi_i) \quad \forall i \quad$ independent;
2. $$\mathbb{E}[Y_i] = \pi_i = \frac{\exp\{X_i^T \beta\}}{1 + \exp\{X_i^T \beta\}} \tag{2.2.2}$$
3. The regressors are not affected by sampling variability.

In [2.2.2] X_i is the combination of the covariate values of the i^{th} unit and β is the vector of regression coefficients. The estimator b is computed by means of *Maximum Likelihood Estimation* and numerical search procedures exploiting the observed y_i's and the dependent variables described in the table 2.1. Once the estimates are reckoned, It is possible to obtain the fitted probabilities.

The concept of **Odds Ratio** is crucial for a proper interpretation of the regression coefficients. Let the *odds* of an event be the ratio of the probability of a success ($\mathbb{P}(Y_i = 1)$) to the probability of a failure ($\mathbb{P}(Y_i = 0)$); in the context of the NDNS project, the odds of being a drinker are expressed as $\frac{\pi_i}{1-\pi_i}$.
Indeed, the structural assumption in 2.2.2 can be equivalently rewritten as a linear combination of the regressors exploiting the notion of odds:

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = X_i^T \beta \tag{2.2.3}$$

The left-hand side of the equation takes the name of **logit function** and it is in fact linear in the X_i's. This way, it is easier to understand what is the proper interpretation of the model coefficients likewise to linear models. Thus, β_k is how much the log-odds of being a drinker changes after a unitary increase for the k^{th} regressor, holding all the other predictor values constant.

It can be observed how β_k is the difference between the log-odds after and before the unitary increase:

$$\exp \beta_k = \frac{odds(x_k + 1)}{odds(x_k)} = OR_k \tag{2.2.4}$$

[2.2.4] is the definition of the **odds ratio** for predictor x_k, showing that the coefficients represents a multiplicative effect on the odds.

Doing inference on the regression parameters gives the opportunity to get insights about the AI distribution in the population. Indeed, the sampling properties of the maximum likelihood estimators enable the researchers to make use of the *Wald test* and the *Likelihood Ratio test* for assessing respectively significance of single estimates or nested models.

Of course, as an overall evaluation of the adequacy of the model it is appropriate to examine the diagnostics of the fitted logistic regression such as linearity of the logits with the predictors, influential values and multicollinearity.

2.2.2 Ordinal Regression Model

Alternatively, the AI is also classified in eight ordered categories as denoted in table 2.1. The different levels provide a qualitative gain by adding a ranking than just a binary choice, so it would be convenient to address the improvement in some way.
Ordinal logistic regression (also known as **proportional odds model** and **cumulative logit model**) is particularly handy to deal with categorical responses carrying a meaningful order.

In the NDNS data set, there are $k = 8$ levels representing drinking frequencies ranging from "NOT AT ALL / NON-DRINKER" to "ALMOST EVERY DAY".

Define *(Väisänen, 2020)*:

- $p_k \in [0, 1]$, probability of being in k^{th} category;
- C_{p_k}, cumulative probability of being in k^{th} category or lower;
- $\frac{C_{p_k}}{1-C_{p_k}}$, *cumulative odds* of being at least in k^{th} category.

The cumulative probabilities are estimated up to the second-to-last category (basically, for $k = 1, ..., 7$), since C_{p_8} - corresponding to the category "ALMOST EVERY DAY" - naturally equals 1. Cumulative odds larger than 1 mean higher chances of belonging to the lower categories while values between 0 and 1 are more likely to exceed.

The model is given by 7 intercepts always ordered in size, i.e. $\alpha_1 < \alpha_2 < ... < \alpha_{K-1}$. These can be regarded as a way to categorize a continuous variable into 8 categories.

$$\ln\left(\frac{C_{p_k;i}}{1 - C_{p_k;i}}\right) = \alpha_k - X_i^T\beta \qquad \text{for } k = 1, 2, .., 7 \tag{2.2.5}$$

On the other hand, the vector β in equation 2.2.5 does not change for different k. This is called **proportional odds assumption** which means the effect of the regressors is the same for all the response levels and the odds ratio stays fixed across categories.

$$OR_{ij} = \ln\left(\frac{C_{p_k;i}}{1 - C_{p_k;i}}\right) - \ln\left(\frac{C_{p_k;j}}{1 - C_{p_k;j}}\right) = -(X_i - X_j)^T\beta \tag{2.2.6}$$

Equation 2.2.6 highlights how the odds ratio for two different units depends only on the value of the independent variables, so the effect on the odds will be common for each comparison between categories *(Kutner et al., 2005)*.

Note how this parametrisation puts ahead of the β coefficients a negative sign, suggesting that larger values of the regressors are associated with greater odds of being in a high response category rather than a lower one.

The cumulative nature of the model also extends to the odds ratio interpretation; indeed, the results are presented in terms *cumulative odds ratios* whose values comprise all the levels up to that rank of the categorical variable.

The predictions can also be presented in the form of fitted cumulative probabilities:

$$C_{p_k;i} = \frac{\exp\{\alpha_k - X_i^T\beta\}}{1 + \exp\{\alpha_k - X_i^T\beta\}} \qquad \text{for } k = 1, 2, .., 7 \tag{2.2.7}$$

Then, the response probabilities can be easily computed as $p_k = C_{p_k} - C_{p_{-1}}$. Of course, the estimates will vary according to the fixed set of predictor values.

The *likelihood ratio test* will be exploited to evaluate the variables plus the validity of the proportional odds assumption must be checked by the *test of parallel lines*.

2.2.3 Mixture Model

Everything that has been seen until now disregards the quantitative data on AI. Nevertheless, it should be taken into account that a substantial proportion of the UK adult population is teetotal, meaning that many zeroes will be observed. Therefore, the continuous measures of alcohol intake may be disrupted and assuming all the observations come from the same underlining process might be too restrictive.

A suitable option is considering a **mixture of two models**: one that is discrete in order to adjust for the null amounts and a continuous distribution for the rest of the data; a probability of selection is assigned to each component.
In this case, the AI will have a *semi-continuous* behaviour shaped as a continuous distribution with a probability mass for zero values.

Generally, a mixture variable Y_i is composed of K components. Let $Z_i \in 1, ..., K$ be a label associated to Y_i specifying the component which generates the observation; pay attention to the fact that these labels are not always observed (i.e. there is no way of knowing their origin) thus they are often regarded as latent variables.

The model can be formalised in the following way *(Bonakdarpour, 2016)*:

$$\mathbb{P}(Y_i = y) = \sum_{k=1}^{K} \mathbb{P}(Y_i = y|Z_i = k)\mathbb{P}(Z_i = k) \tag{2.2.8}$$

where $\mathbb{P}(Y_i = y|Z_i = k)$ is the mixture component representing the distribution of Y_i conditioned on its belonging to the k^{th} component and $P(Z_i = k)$, also denoted as π_k, is the mixture weight indicating the probability for that variable to belong to the k^{th} component.

In the specific case of modelling *Alcoholg* the approach will be a **two-part model** : the fit occurs in two stages with two possibly different equations where 'The first stage refers to whether the response outcome is positive. Conditional on its being positive, the second stage refers to its level' as stated by *Min and Agresti, 2002*. Therefore, for the first stage:

$$\ln\left(\frac{\mathbb{P}(Y_i = 0)}{1 - \mathbb{P}(Y_i = 0)}\right) = X_i^T \beta_1 \tag{2.2.9}$$

which will lead to the second stage :

$$\ln(y_i|y_i > 0) = X_i^T \beta_2 + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2) \tag{2.2.10}$$

Note how the vector of regression coefficients are not the same for both equations, having a different subscript.
The estimation is done by means of *Maximum Likelihood Method*, so that *likelihood ratio tests* can be carried out also for this model strategy.

3 PRELIMINARY ANALYSIS

Here, we are going to deal on the implementation of the aforementioned parametric models fitted for the NDNS data. The software used for implementation is R. This chapter includes everything that needs to be done in a statistical analysis like an exploratory analysis of the data, dealing with anomalies in the data set (such as missing values) and fitting of statistical models.

3.1 Exploratory Analysis

Every statistical analysis worth of this name must start with a preliminary look on the data set which is exploited for the research.
As briefly mentioned before, the data used to answer the questions of interest comes from the NDNS stages of the first nine years (from 2008 to 2017). The aggregated number of subjects across all the years was 13,350; after filtering for the adult population (≥ 18 years, the target population of the AI intake investigation), the resulting sample size is 7112. There are no duplicates among the data.

It is useful to have a look on what is going on with the data set, if there is enough variation among the variables or particular pattern hiding some meaning, plus the situation about missingness of the data. It is especially important to see how the outcome - in all its forms - is shaped and the way it varies according to the explanatory variables.
First, we look into the summary of all the variables of interest in the data set, from there we move on to study the categorical variables for alcohol intake and see how they interact with the rest of the variables and finally we end by having a closer look on the quantitative distribution of the response according to its predictors.

3.1.1 General features of the data set

After making sure that the data do not present any duplicated rows, it is useful to get a general idea on what are the main features of the data set. Particularly interesting is the output of the summary statistics for the numerical variables in the NDNS data, shown in table 3.1.

What is most alarming at first sight is the occurrence of "-1" values as the minimum value for the *weight, height, BMI and equivalised income*. Obviously, a negative quantity has no meaning in the context of those variables, therefore it is evident those measurements must be regarded as invalid. In the **list of variables for UK data** the reasons why some invalid measurements for body measures may be found are specified: *Not usable, refused, attempted but not obtained, not attempted or pregnant(relatively to BMI and weight)*.
On the other hand, no information is available on why income also presents missing values, nonetheless it is reasonable to believe they are a consequence of refusal to provide this disclosure. From now on, those invalid observations will be regarded as missing values for the statistical analysis, and later on this chapter an in-depth study of the nature of these missing values will be conducted together with an explanation of the imputation strategy.

Another relevant fact is the gap between the 3rd quartile and the maximum value for both the equivalised income and the alcohol consumption, which is also reflected by a substantial difference among the median and the average alcohol consumption). This suggests to be careful in case these likely outliers would be influential units.

Statistics	Age	Eq. Income	Height	Weight	BMI	Alcohol
Min	18.00	-1	-1	-1	-1	0.00
1^{st} Quartile	34.00	11,823.99	158.80	61.80	22.65	0.00
Median	47.00	21,959.12	165.85	73.40	26.21	0.33
Mean	48.53	26,873.28	156.17	71.67	25.22	11.65
3^{rd} Quartile	63.00	36,884.43	173.50	86.20	30.08	16.05
Max	96.00	184,425.41	199.30	159.20	52.94	401.76

Table 3.1: *Results of the summary statistics for the numerical variable in the data set produced by R*

3.1.2 Categorical outcomes and relationship with other variables

The AI is available in the form of categorical variables, as a dichotomous variable (drinker/non-drinker) and with multiple levels defining a measure of the frequency of drinking. While the former will be the response for a logistic model, the latter will come handy in what we defined above.
It is worthy to have some insights on how these two variable are shaped up and related to the other variables.

In figure 3.1 the box plots of the variables *age, BMI* and *equivalised income* varying for drinking status are shown. In the sample, teetotallers are usually older at parity of quantiles, for instance three quarters of the sample units are aged below 70, while drinkers are at maximum 57 for the same proportion.
The BMI levels do not seem to vastly change across the two categories, just a slight more spread of the central values for the non drinkers, while in the other category it seems more concentrated around its median and then heavier tails.
For both categories, the distribution of the economic measures seems to be positively skewed because the extreme values on the right seem to be quite far from the median and the upper quartiles, simultaneously the income distribution for non-drinkers is more concentrated , with roughly 75% of the sample units being below 50 thousand pounds, for the drinker the situation looks less compact.

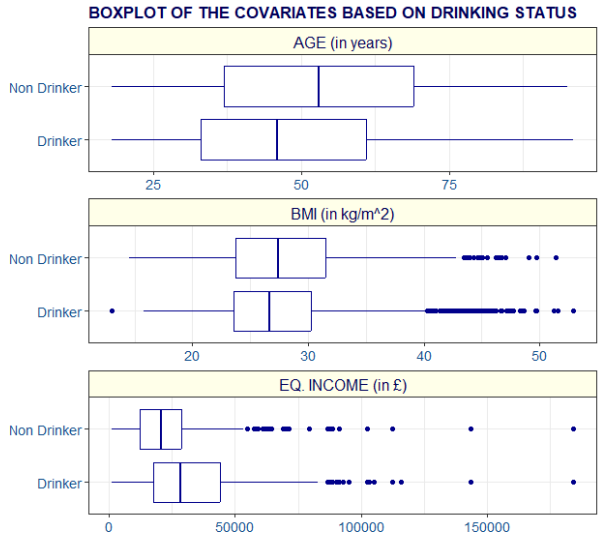


Figure 3.1: *Caption*

In addition, further material is available in Appendix B where frequency tables between the categorical predictors and the drinking status can be consulted (like tables B.1 and B.2).

It is also interesting to see some summary statistics about the recorded intakes of alcohol consumption in table 3.2 for the binary variable in order to observe if there is anything particularly contradictory about the statement of drinking status and then the actual recorded quantities. Of

course, it is reasonable to expect much lower intakes in the non-drinking category than the opposite, however it is always important to remember that some mistakes can be made in the collection stage of the data or even the follow-up study may have happened in a period of time when units experienced a different exposure to alcohol than the usual approach they might have with these substances.

Drinking Status	Count	Min	Median	Mean	Max
Drinker	5717	0	5.67	14.399	401.76
Non-Drinker	1389	0	0	0.347	54.824

Table 3.2: *Summary statistics of recorded alcohol intake (in g) for self-declared drinkers and non-drinkers*

Moving forward, a preliminary study is also carried out for the polytomous variable showing the different levels of AI consumption with a meaningful order which denotes a rough measure of how often a subject drank an alcoholic beverage in the past 12 months. of particular interest is the relationship with some of the other variables.
The figure 3.2 depicts the marginal averages of the plotted numeric covariates stratified by the eight levels of AI frequency as denoted in table 2.1. Those plots can be useful in assessing if the predictors behaves in an ordinal fashion with respect to the different categories of the outcome variable.
Generally, none of the plots shows a monotonic trend, the worst one being the age grid where there is a sort of parabolic behaviour: the people who declared to drink almost every are on average 60 years old, followed by a decrease in both the age and the frequency up to people having a couple of drinks per month aged less than 45 and then the average age goes up again for the meagrest drinking frequencies since people who did not drink in the last 12 months were aged between 52 - 55 .
Apart the categories at the extreme, as more the people drink the more financial resource they have, while BMI worsen for higher quantities.
The size of the dots depends on how many units are part of that category, therefore bigger points mean greater cardinality of the stratum relative to the size of the others.
Furthermore, the actual strata sizes are reported in the second column of table 3.3 together with some summary statistics of the recorded quantitative AI values for each level of the ordinal variable, showing if there is compliance between the quantities recorded in the dietary recap and what the subject declares as his general frequency of having a drink.

Keep in mind this is just an exploratory analysis which could be helpful in later stages of the analysis to maybe explain and/or justify something that comes up in the fitting of the model,so even if none of the graphs in the figure seems to show AI behaving in an ordinal way to the respect of these predictors it could be a situation specific of the sample and of the nature of the variable. The results always need to be contextualised.

3.1.3 Quantitative AI relationship with other variables

Finally, the continuous variable for AI is examined to understand how to properly use it in the mixed model and get a glimpse of its relationship with the model covariates. This can be observed in the figure 3.3.

In 3.3a the categorical predictors of the analysis are shown, in particular the region of origin of the subject, the sex including the instance of females who are pregnant or currently breastfeeding, and the day in which the follow-up study of the subject has begun; the latter is introduced for the continuous model because the dietary intakes may change based on which part of the week is under study, thus allowing to adjust for differences in the weekdays and the festive ones. The

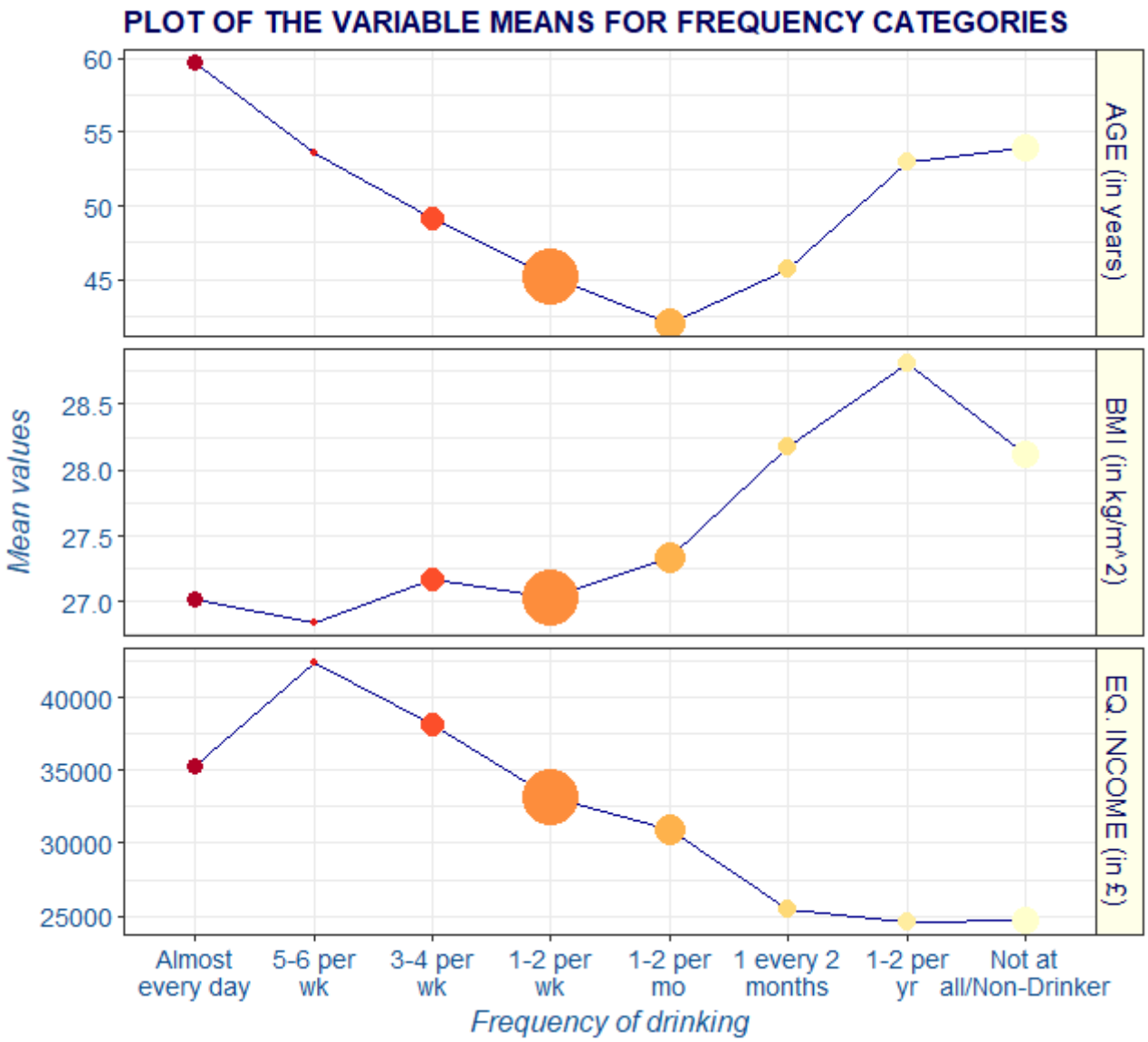


Figure 3.2: *Caption*

Frequency of drinking	Count	Min	Median	Mean	Max
Not at all/Non-Drinker	572	0	0	0.113	27.2
1-2 per year	233	0	0	0.615	40.544
Once every two months	837	0	0	1.55	68.64
1-2 per month	2101	0	0	4.37	93.328
1-2 per week	1110	0	6.634	12.025	172.6
3-4 per week	652	0	20.062	26.092	180.4
5-6 per week	607	0	24.312	30.286	147.42
Almost every day	993	0	27.428	38.94	401.76

Table 3.3: *Summary statistics of recorded alcohol intake (in g) for declared frequency of drinking*

sample frequency of the observations for each level of the factor are displayed below the boxes (note that the frequency in each panel sum up to 7112).
Sex and Pregnancy are considered altogether since it can recognised how the results obtained for pregnant women are not trivial therefore it would be safe to regard it as an additional factor for the analysis. From now on, the two variables will be blended resulting in a categorical variable

with the three levels highlighted in the graph: Males, non-pregnant females and pregnant ones.

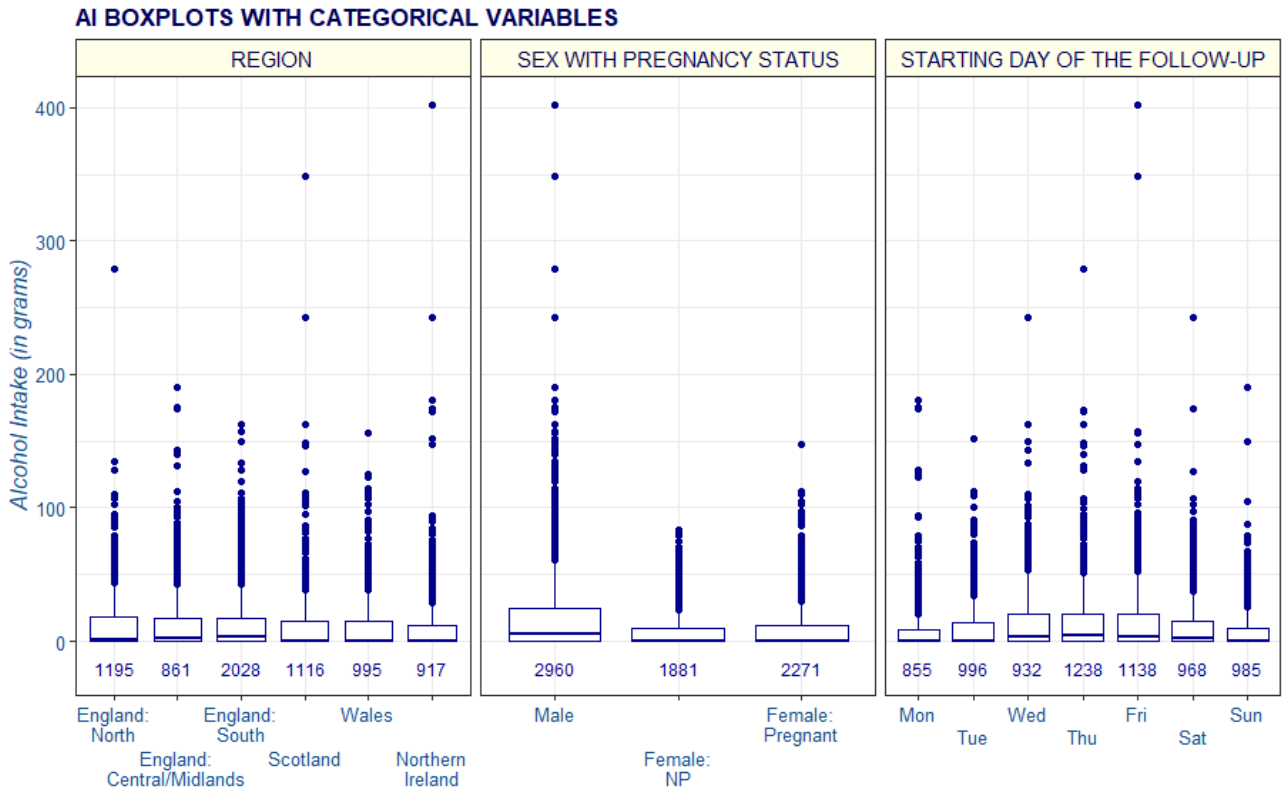
From the graph, it is possible to notice a few interesting facts; for instance, the highest recorded AI corresponds to a Northern Irish man, gathering dietary intakes during the weekend (since the starting day was Friday).

The stacked observations in the income come from the fact that we have an equivalised measure and so some observations had common scores weighted for the household composition and income values.
The highest value of the Alcohol consumption comes up in the model do not correspond to uncommon values for the regressors , but they are located in the middle where also less extreme values are observed for AI. This could suggest that we do not expect some units to be influential at least as a leverage value since they do no point out in the horizontal sense but just for the outcome variable distribution. Of course this is just a guess and it needs more rigorous study.

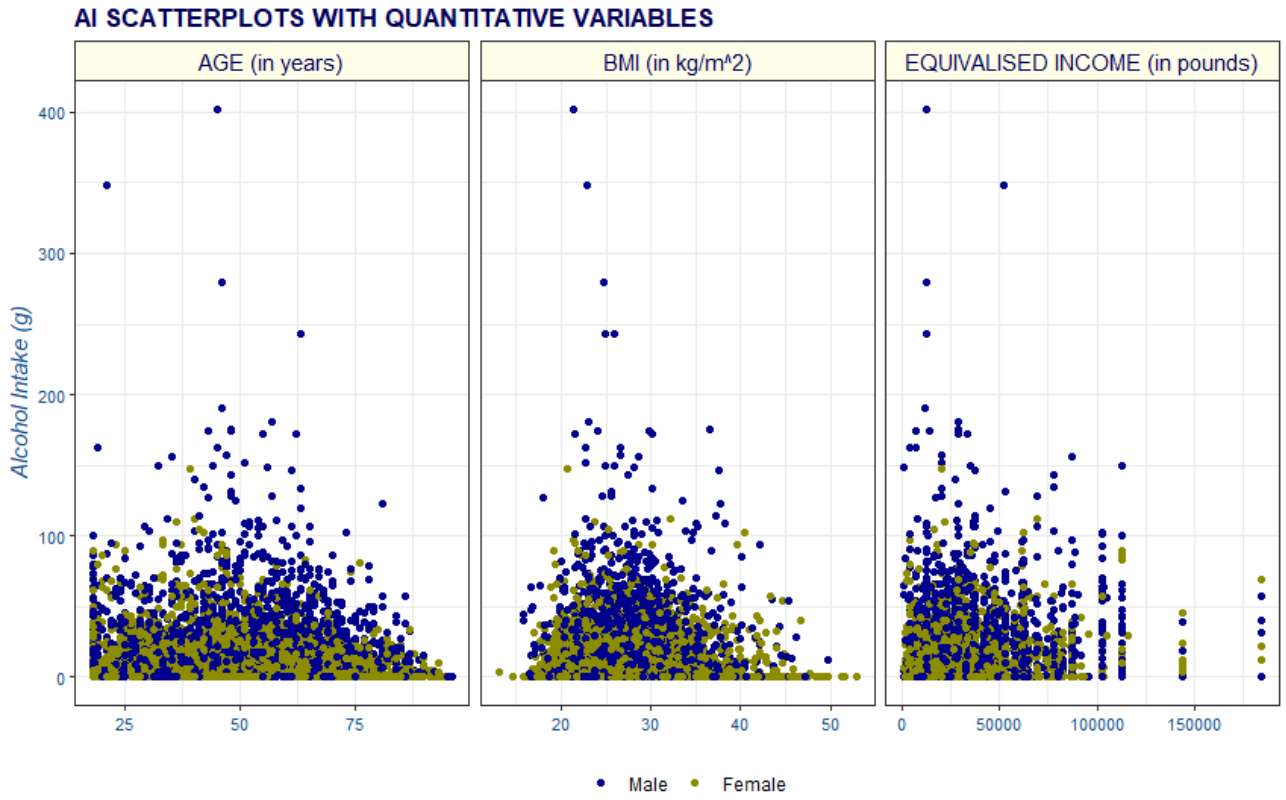
Roughly, we can observe that people who drink more are middle-aged (even if people having the minimum age to drink have a slight more evident spike, while older people drink less and less), around 20 and 30 of BMI and it also seems that people with lower economic status are heavier-drinkers.

Geographically speaking, the median is always around 0, except for the Central and the Southern part of England showing a slighter inflated distribution for the 1st half of the data. On the other hand the most extreme cases of AI correspond to isolated observations coming from Northern Ireland and Scotland, followed by the northern areas of England.
Males show significant higher values than women where pregnant ones have wider distribution that non pregnant ones.
Higher volumes of alcohol are recorded for cohorts starting the follow-up dietary intake in Thursday and Friday which of course would extend up to the end of the week.

Always recall these observations come just from an exploration of the sample data and cannot be enlarged to the general population without an underlining statistical modelling theory.



(a) *Box plots for the categorical variables*



(b) *Scatter plots showing the relationship between the numeric predictors and AI, differentiated by the sex of the subject*

Figure 3.3: *Plots showing the relationship between Alcohol Intake and the predictors*

Let's have a look now on the density distribution of AI where there is an inflation caused by the

large presence of zeroes.
This feature of the variable is highlighted in the top plot of figure 3.4 where the bins around zero stand out to the rest of the distribution. It happens as a natural consequence of the general fact that a portion of the population is considered to be teetotal, therefore it is expected for this type of measure to have a high inflation of zeroes.
The plot below represent the distribution of AI excluding the null observations through a histogram overridden by a density plot highlighting where values are observed on the x-axis by the usage of rugs. Those ones are especially helpful to catch where outlying values are located, like they were identified in the scatter plots of figure 3.3b, otherwise it would be really difficult to notice them just by looking at the histogram since the density is really low in those regions.

These graph are really helpful in making a decision for the second fitted model in the two-part mixed model; indeed, it seems clever to log transform the AI intakes and then consider a linear regression with normality assumption to account for the major spread around lower values (instead of the theoretical symmetry around the expectation of a Normal Distribution) and also adjust for the bigger magnitude of the extreme values in the AI.

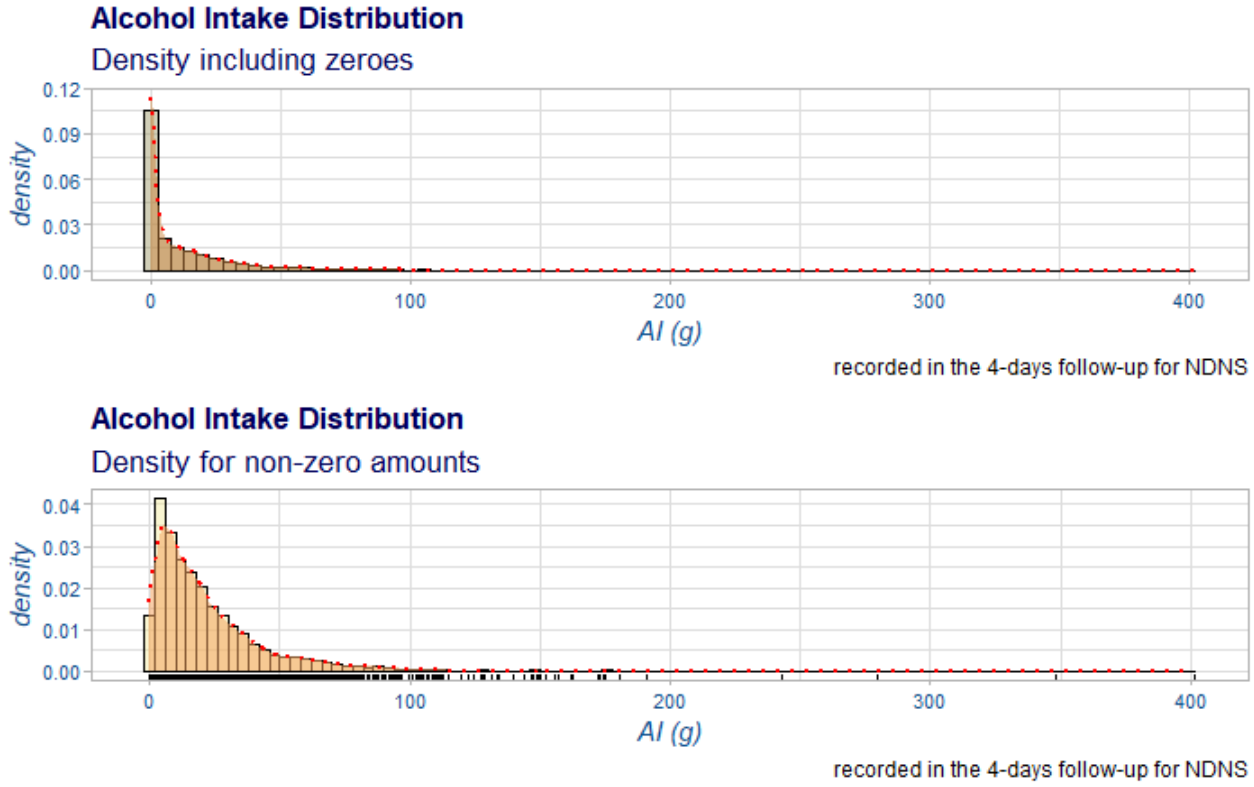


Figure 3.4: *Histograms representing the distribution of the alcohol intake in two versions: including the zeroes (on top) and removing them(below)*

3.2 Imputation of missing values

In 3.1.1 invalid observation were identified, therefore those could be regarded as part of the missing information implicit to the data set. We should find a way to deal with those NA.
An useful tool to deal with missing data analysis in R is the package MICE ("Multiple Imputation by Chained Equations") providing loads of instruments to understand patterns of missing values and also to implement in many ways the technique of multiple imputation *(Austin et al., 2021; van Buuren, 2018).*

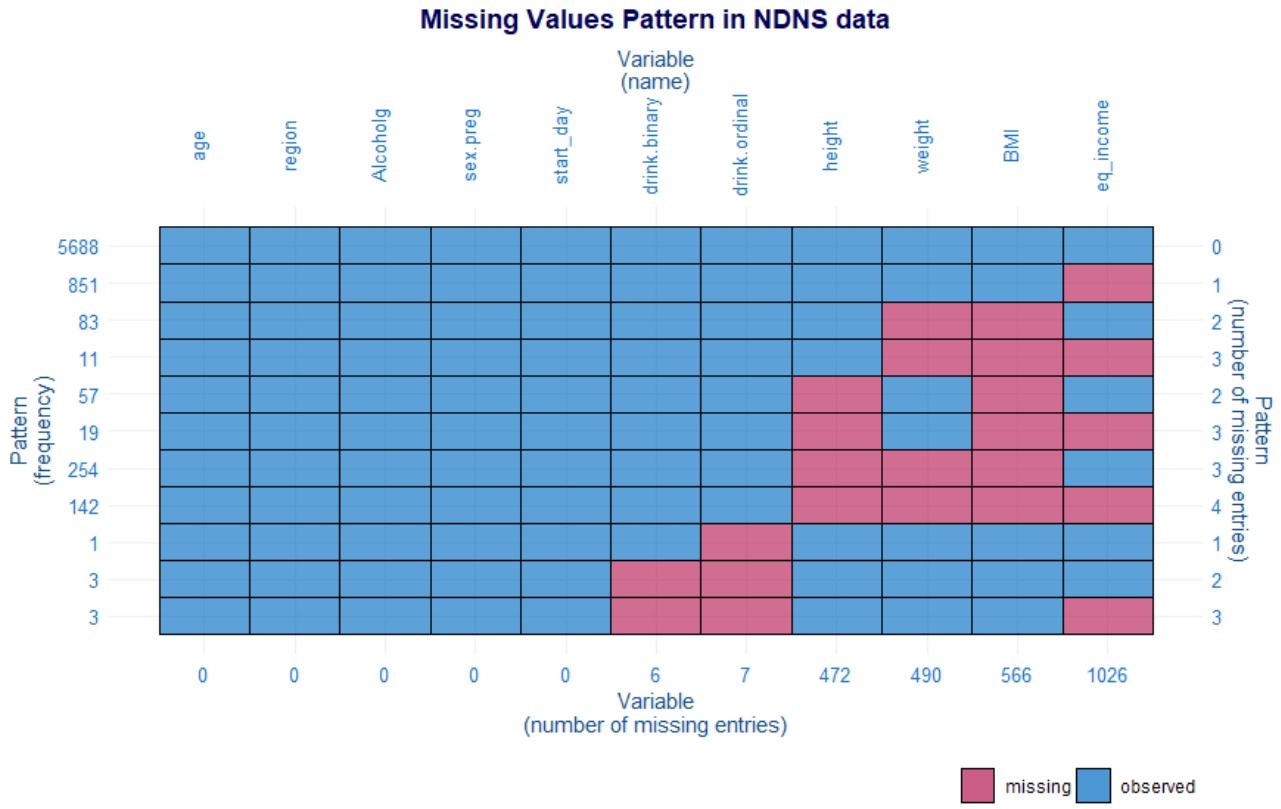


Figure 3.5: *Patterns of the missing values in the data set*

Following, we will see what are the NA pattern in the data set under-study, next a brief theoretical insight on what is the meaning of multiple imputation and how it will be implemented and at the end how it translates with our study.

3.2.1 Missingness in the NDNS data set

The figure 3.5 is a nice way of illustrating how missing values occur in the model, providing many tips on how to set up an imputation analysis.
The plot is composed of a graph made of blue (for observed entries) or red blocks (for missing ones) and numeric values at the margins. Each row represent the units showing that particular combination of observed and non-observed missing values with the frequency specified explicitly in the left margin, while on the right-most column there is the number of red block for that particular pattern, meaning how many variables are not observed for that particular pattern.
On the bottom line, the number of missing values in the data set for the variables specified on top. Immediately, the missing values occur especially for *Equivalised Income* (having the most consistent number of NA's with 1026), *Height, Weight* and consequently *BMI*.

Notice how only the variables useful for the imputation are regarded; for example *Seriali* and *start_day* were disregarded because they would not help in any way to assess why we observe missing values from certain variables, since they have a role in the design process of the study which does not affect in any way how to recover information for income and BMI. At the same time, it is important to include weight and height (even if they are not going to be used in the statistical analysis model) because BMI which is a relevant variable is a by-product of those two, therefore also the reasons why it is missing may be shared with the ones originating it.
Indeed, it can be observed how wherever BMI is missing, weight, height or both are missing too.

While *drink.binary* and *drink.ordinal* do not have many missing entries so discarding those units

wouldn't allow the analysis to loose that much degrees of freedom, for the covariates mentioned before the missing values are quite a lot and go for the complete-case analysis would be a waste of available resources, plus we do not know if the real reasons behind the missingness can be considered to be totally at random so that excluding units would not introduce bias.
For those reasons it seems better to actually find suitable estimates for the missing entries and work around the uncertainty they could bring to the final model estimation.

3.2.2 Multiple Imputation : methods and predictors

In order to fill the missing data we rely on **multiple imputation**: basically, NA entries are filled in more than one time, creating many different plausible imputed data sets. This way, it is possible to provide a measure of the uncertainty in the estimation of the imputed data, which would not be possible with just a single imputation since there is no variation.
Broadly, it happens in three stages as explained in details by van Buuren, 2018:

- generate replacement values and repeat procedure many times, say *m* times;
- fit the analysis model to each imputed data set;
- combine the m parameters estimated in just one, and provide an estimate of the variance which considers both the usual sampling variation (called *within-imputation variance*) and the extra variance caused by the multiple imputation process (*between-imputation variance*).

There is a vast literature on how to properly choose and what is the qualitative influence m, i.e. the number of multiple imputations to perform. In this case, having moderate empty data it feels safe to stick with a lower number of imputation as it is the default for his case, so we will stick to m=5. Furthermore, the choice of m is also reflected on the total variance since the between-imputation variance is enlarged by a factor of $1/m$.

Since there is more than one column with missing values, we are going to deal with methods of multivariate imputation. An important question how to use as much as possible information to fill missing values using variables with missing data to also fill other columns. The strategy we rely on for this particular study is **Fully conditional specification** also known as **chained equations**. imputes multivariate missing data on a variable-by-variable basis (Van Buuren et al. 2006; Van Buuren 2007a). The method requires a specification of an imputation model for each incomplete variable, and creates imputations per variable in an iterative fashion.

To create imputation **predictive mean matching** procedure will be employed. Roughly it works in the following way:

> For each missing entry, the method forms a small set of candidate donors (typically with 3, 5 or 10 members) from all complete cases that have predicted values closest to the predicted value for the missing entry. One donor is randomly drawn from the candidates, and the observed value of the donor is taken to replace the missing value. The assumption is the distribution of the missing cell is the same as the observed data of the candidate donors

Recall donors are units whose observed values are used to fill the missing values of another unit. Usually, the chosen donors are believed to be somewhat similar to the units with missing data. In the case of predictive mean matching this concept of 'similarity' is given by homogeneity of the predicted values from the imputation models. By default, the numbers of donors in *mice* are 5.
It is very efficient because there is consistency on the way the type of the data is expected to behave. Since the imputed values are fundamentally taken from other observations, in no way impossible values are going to occur, they will be realistic (you would not experience negative values for variables which are never less than 0). Therefore imputation will not produce never seen before values for the incomplete variables, but values that are already observed for another unit, so no

originality or creation of a new level. Also take notice of the fact that the set of donors is not always the same for different number of data sets, so in each of them the pool of choices may differ, allowing more variation.

Now it is of interest to define what will be the imputation model we are going to use and most importantly what will be the predictors for our target variables. We must specify the methods to use and the predictor matrix.

The used method as said before it is predictive mean matching for weight, height and equivalised income while for BMI a better choice must be made. This because BMI is a composite variable, a transformation from weight and height so we would like to keep consistency between the imputed values of height and weight and the result of BMI, exactly given by its formula : $\frac{weight(kg)}{height(m)^2}$; since in the provided data height is in meters we need to divide by 100 the observations. By specifying this formula in the methods argument of the, we are still able to use the BMI as predictor for the other variables in the imputation model but when it needs to be filled the values of height and weight will be used to fill in the empty observation.

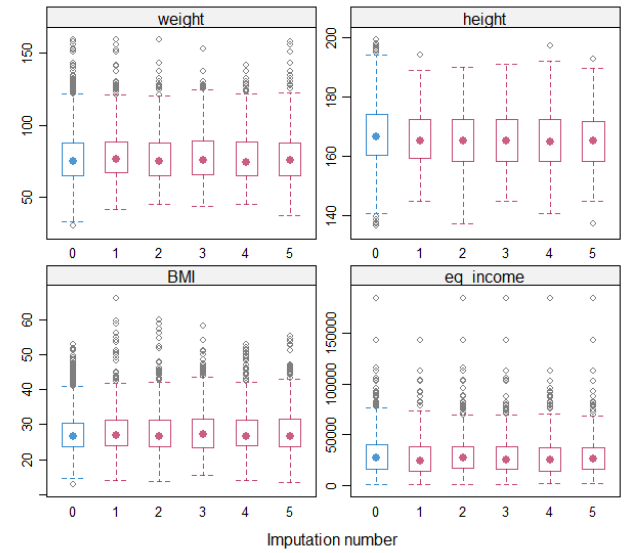
Another important thing is to define the predictor matrix. Basically, it specifies the set of variables to use in order to predict (i.e. used in the imputation model) the target variables.
In general we would like to take care of a couple of points: select all the variables which will be used in the analysis model, especially the outcomes in order to keep the correlation sense of the study , so even if drink.binary and drink.ordinal are not as important to impute because just few observations are non-observed, it is important to include them because they might be relevant to impute the variables which will be later used in the logistic and ordinal regression models. Also include any variable which might improve the variance and it is related to the outcome of the model to be seen later.
So in our specific case we will consider all the variables which are going to be used in the model or related in any way to the AI consumption, excluding starting day of the follow (we believe does not have any role in assessing weight, height or income) and other variables which could be summarised in other variables (so we just consider sex.preg instead of sex and pregnancy alone). Then we rely on the pre-built R function *quickpred()* which aids the user in selecting predictors with simple statistics in order not to use huge number of predictors for each imputation model. Those statistics are based on the correlation between the predictor and the target variable and on the correlation between the indicator of presence of missing values for both the predictor and the target variables.

3.2.3 Quality of the imputation

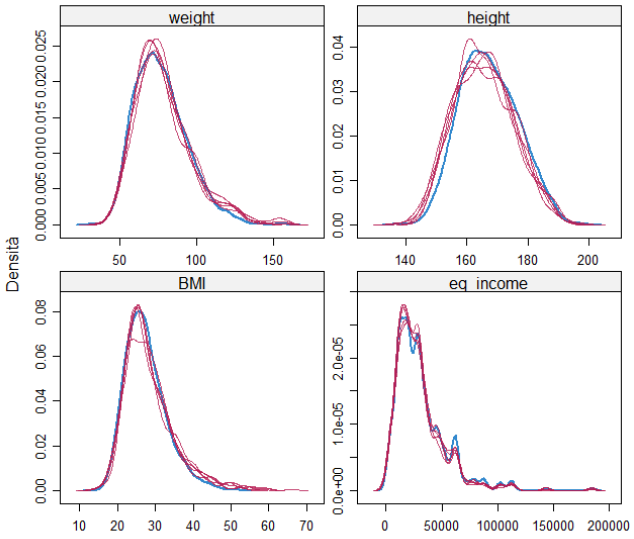
The focus is on distributional discrepancy, i.e. the difference between the observed and imputed data. Basically, it is like studying the "fit" of the imputation model from the observed data.
The general idea is that good imputations have a distribution similar to the one of the observed data, so it can be seen as imputed values could have been real had they been observed. Plots are good to check these differences as the ones in figure 3.6.

More or less, we can observe how there are no huge discrepancies from what the imputations have produced with respect to the the observed data. Take notice of 3.6a where for all the facets the red boxes are included in the range and they do not go outside of what is depicted by the blue box, except for BMI. This is because while all the other variables are imputed by predictive mean matching which assures that assigned values to missing entries will be values already observed elsewhere, BMI has been passively imputed by specifying a formula involving height and weight so that consistency would have been kept. This is good because it allows to enlarge the scope of

BMI with plausible values which risked to be too shortened if considering only the complete data.



(a)



(b)

Figure 3.6: *Graphs useful to check the quality of the imputations, where distributional discrepancy is observed by comparing the difference between the observed data (in blue) and the five imputed datasets (in red).* a *depicts box and whisker plots,* b *depicts kernel density plots of the distributions comparing the red lines (imputed data sets) from the observed represented by the blue line*

4 RESULTS AND DISCUSSION

In this chapter, the results of the analysis will be presented after fitting the various type of the models with an insight on the quality of the fit, checking the diagnostics and discussing the estimated obtained from this analysis.

4.1 Logistic Fit

To model *drink.binary* a logistic regression model is with binomial family and logit link, thanks to the MICE *(van Buuren, 2018)* and GLM functions in R.
The selected model includes the terms *Age, BMI, Equivalised Income, Sex/Pregnancy status, Region* and the interaction terms are evaluated between Age and BMI, Age and Sex/Pregnancy status because this model in studies about dietary consumption are scientifically meaningful to see their interaction effect. The terms are also chosen in order to keep consistency with the other analyses. The median AIC is 6693.208 and its range is [6669.570,6698.952].

	Estimate	Std. Error	p-value
Intercept	-1.335	0.451	0.003
Age	-0.006	0.008	0.481
BMI	-0.010	0.016	0.542
Eq. Income (*in thousands*)	-0.019	0.002	<0.0001
Sex.Preg [Female: NP]	-1.567	0.379	<0.0001
Sex.Preg [Female: Preg.]	0.123	0.259	0.636
Region [England: Central]	0.288	0.115	0.013
Region [England: South]	0.151	0.097	0.120
Region [Scotland]	0.084	0.110	0.446
Region [Wales]	-0.086	0.116	0.454
Region [Northern Ireland]	0.363	0.111	0.001
Age : BMI	0.001	0.0003	0.088
Age : Sex.Preg [Female: NP]	0.030	0.006	<0.0001
Age : Sex.Preg [Female: Preg.]	0.001	0.007	0.898

Table 4.1: *model estimates of the fitted binary logistic model*

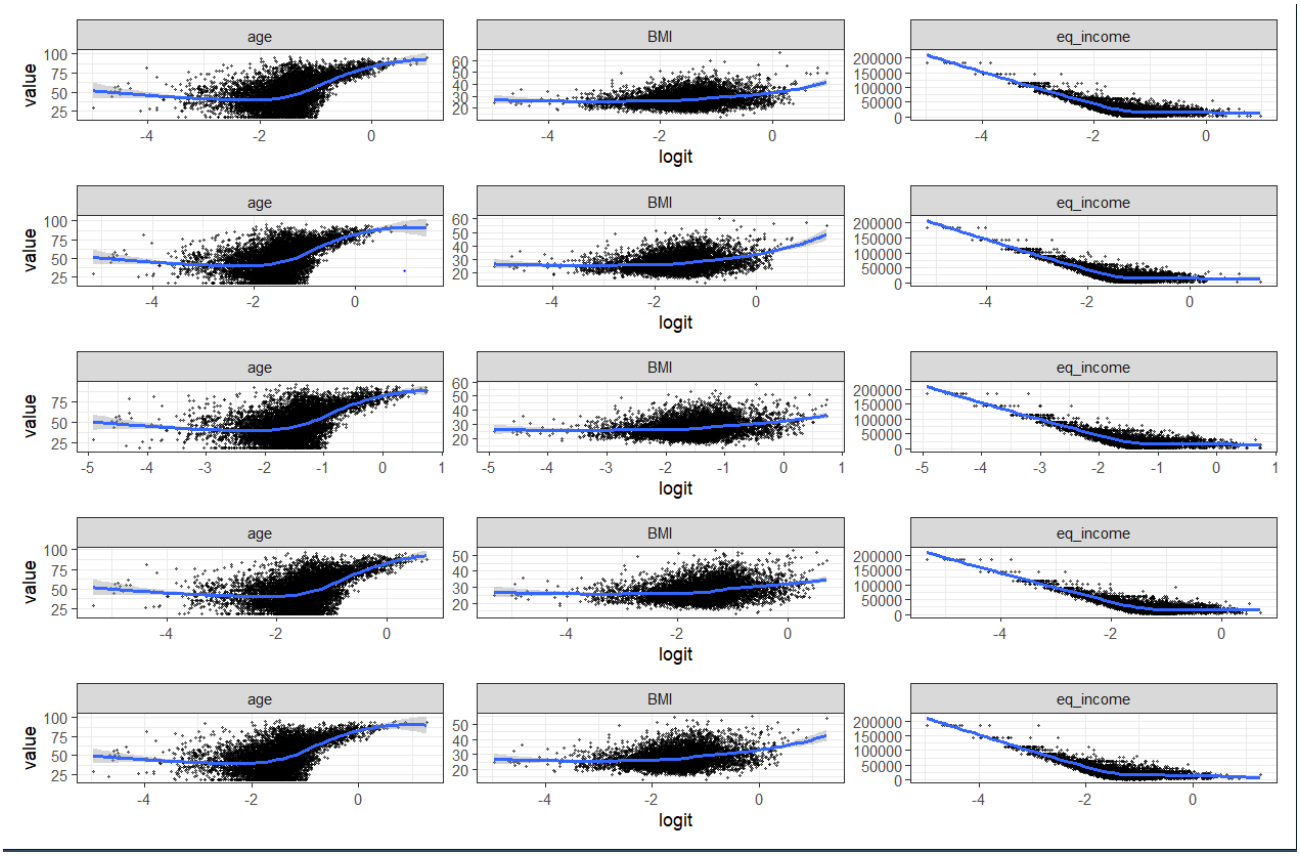


Figure 4.1: *Linearity assumption between the fitted logits of the function and the numerical predictors for the binary logistic model*

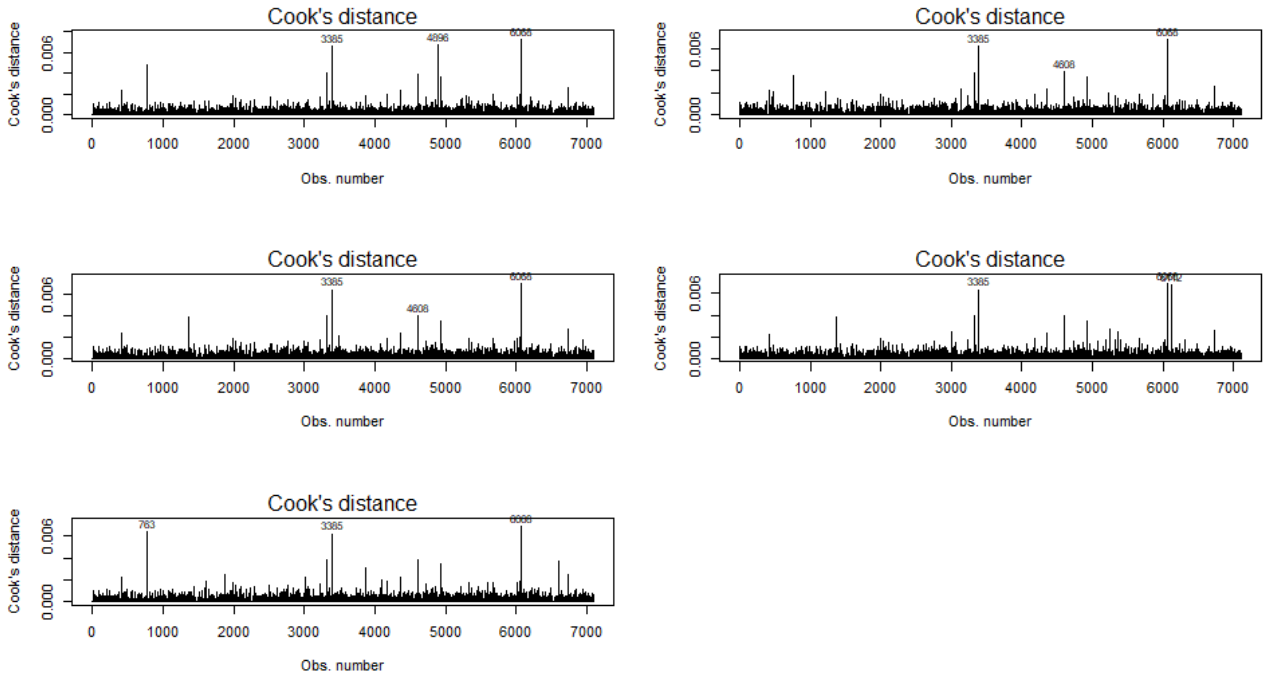


Figure 4.2: *Cook's distance for the logistic regression models in each imputed data set*

The first step in assessing the validity of a model is checking the diagnostics *(Kassambara, 2018)*.

The figure 4.1 shows the logits plotted against the numerical predictor values. Those scatter plots are especially good to assess if the assumption of linearity of the independent variables in equation 2.2.3 is sensible. Having performed the logistic regression on a multiple imputed data set, the diagnostics are checked for all five fitted models; if the assumptions hold more or less for each of them, then also the pooled model will likely be valid.
in the AGE panels is difficult to say with certainty how the trend is varying because of the high concentration of points around -2 and -1 logit values. However, there is clearly a cluster in the top-right part of the plot, corresponding to high values of the logits and age; since those peaks are rarely reached by the left-most points, it does not seem to be too far of a stretch to assume linearity.
BMI has probably the best look out of the three covariates, since there is no alarming sign of deviation from linearity. Indeed, the trend seems mostly steady with a slight nod of positive gradient. EQ_INCOME instead has a clear linear behaviour until right after -2 logit and then it flatlines.
Overall, there is no reason to believe the assumption of linearity is contradicted.

Multicollinearity does not raise any warning flag since it never goes over 2 in any of the imputed data sets.

The last assumption requires checking presence of influential values. Figure 4.2 illustrates the *Cook's distances* for the observation in the various imputed data sets. The fact that units like 3385 and 6068 are highlighted in multiple panels is a consequence of having actual observed data for those observations instead of imputed ones and it shows the consistency of the model throughout the multiple imputations. The magnitude of the distances is not high enough to raise any concern about influential values.

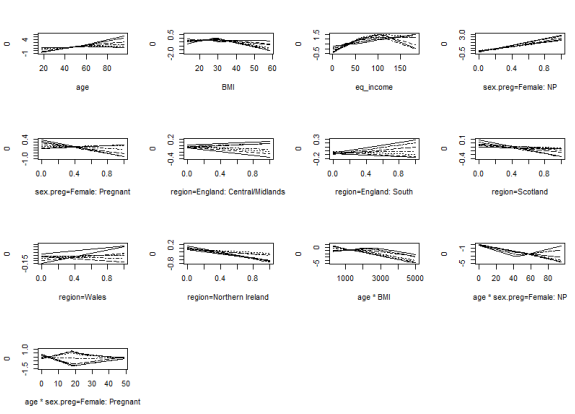
4.2 Ordinal Fit

To model *drink.ordinal* a proportional odds model with logistic family for the cumulative distribution function is fitted thanks to the RMS package in R *(Harrell, 2015)*.
The selected model considers additive effect of *Age, BMI, Equivalised Income, Sex/Pregnancy status, Region* and interaction terms are evaluated between Age and BMI, Age and Sex/Pregnancy status. The terms are also chosen in order to keep consistency with the other analyses. the RMS package provides an overall measure of the AIC, which for this models equals 26658.51.

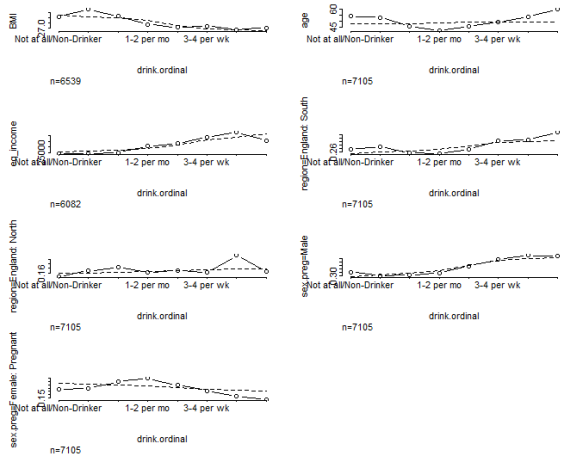
Figure 4.3 shows very helpful plots *(Harrell, 2015)* in order to verify if the proportional odds assumption holds in the context of the NDNS data set.
In 4.3a it is possible to check linearity and the PO assumption at the same time, where each line represent one of the fitted equations for the categories of *drink.ordinal*; if they are parallel then PO is adequate. Unfortunately, what is observed suggests inadequacy, especially in the interaction terms where the lines take really peculiar shape but very distant to any form of parallelism. Even the main effects *eq_income* and *age* do not encourage such idea and the former shows also non-linear trends for some categories.
4.3b recalls what was done in the exploratory analysis with figure 3.2, here with the supplement of what is expected in case PO holds. Indeed, while solid lines connect simple stratified means, the dashed lines connect the estimated expcted values of the AI for different predictors when PO holds. Also here, *age* does not match very well what is expected under the fundamental assumption, therefore being discordant with both variable might suggests the poor sensitivity of the ordinal model. Still, it is worth to stress that *drink.ordinal* features many categories and this makes the assessment of the model extremely complicated to be fairly good. In future studies it might be worthy to keep a lower number of categories or to merge the ones available.

	Estimate	Std. Error	p-value
Non-Drinker	-1.476	0.315	<0.0001
1-2 per year	0.871	0.315	0.006
Every 2 months	0.377	0.315	0.231
1-2 per month	-0.332	0.315	0.291
1-2 per week	-1.744	0.316	<0.0001
3-4 per week	-2.649	0.318	<0.0001
5-6 per week	-3.040	0.319	<0.0001
Age	0.029	0.007	<0.0001
BMI	-0.008	0.011	0.462
Eq. Income	0.013	0.001	<0.0001
Sex.Preg [Female: NP]	2.023	0.281	<0.0001
Sex.Preg [Female: Preg.]	-0.262	0.162	0.105
Region [England: Central]	-0.021	0.080	0.791
Region [England: South]	0.008	0.065	0.900
Region [Scotland]	-0.137	0.073	0.063
Region [Wales]	-0.015	0.075	0.839
Region [Northern Ireland]	-0.379	0.078	<0.0001
Age : BMI	-0.0005	0.0002	0.029
Age : Sex.Preg [Female: NP]	-0.044	0.005	<0.0001
Age : Sex.Preg [Female: Preg.]	-0.003	0.004	0.548

Table 4.2: *model estimates of the fitted proportional odds model*



(a) *Smoothed partials residuals for each predictor*



(b)

Figure 4.3: *Diagnostics of the proportional odds assumption for the ordinal regression model*

4.3 Two-Part Fit

The last model to consider is the two-part model where the quantitative AI in grams is the response variable. As a response the transformed ln(*Alcohol*g + 1) is regarded in order to adjust for the very small quantities in the data. Unfortunately there is no package available to deal with mixture regression for multiple imputed data sets, so the study could only be carried out for the complete-cases data set.

The terms of the regressions are the same of the previous model but this the variable *start_day* is

added to the model because it becomes fundamental when dealing with data gathered from the 4-days follow-up dietary intakes, allowing variation from day to day.
The same models were fitted for the first and second stage. The former had AIC value of 7400 while the latter showed AIC of 8038.1 .

	Estimate	Std. Error	p-value
Intercept	-0.531	0.423	0.209
Age	0.016	0.009	0.055
BMI	-0.032	0.015	0.032
Eq. Income	0.017	0.001	<0.0001
Sex.Preg [Female: NP]	2.134	0.394	<0.0001
Sex.Preg [Female: Preg.]	-0.707	0.223	0.002
Region [England: Central]	0.072	0.104	0.490
Region [England: South]	0.018	0.086	0.837
Region [Scotland]	-0.137	0.099	0.168
Region [Wales]	-0.135	0.102	0.187
Region [Northern Ireland]	-0.182	0.102	0.073
Age : BMI	0.00002	0.0003	0.958
Age : Sex.Preg [Female: NP]	-0.044	0.006	<0.0001
Age : Sex.Preg [Female: Preg.]	0.014	0.006	0.018
start_day [Tue]	0.486	0.111	<0.0001
start_day [Wed]	0.680	0.113	<0.0001
start_day [Thu]	0.811	0.106	<0.0001
start_day [Fri]	0.714	0.107	<0.0001
start_day [Sat]	0.605	0.110	<0.0001
start_day [Sun]	0.149	0.109	0.174

Table 4.3: *model estimates of the two-part model of the first stage*

In figure 4.4 the diagnostics of the two-part model are shown. Focusing on the second part we can see how the main assumption are all appropriately covered.
Indeed, there is no odd behaviour of the residuals which seem evenly spread around 0, in addition the assumption of normality is very on point and makes this model a good candidate for valid inferences. The scale location plot does not suggest the homoskedasticity assumption is unreasonable for this data set.

	Estimate	Std. Error	p-value
Intercept	2.519	0.296	<0.0001
Age	0.003	0.006	0.582
BMI	0.014	0.011	0.177
Eq. Income	0.002	0.001	0.004
Sex.Preg [Female: NP]	0.436	0.257	0.089
Sex.Preg [Female: Preg.]	-0.594	0.148	<0.0001
Region [England: Central]	-0.072	0.062	0.243
Region [England: South]	-0.116	0.051	0.024
Region [Scotland]	-0.065	0.061	0.285
Region [Wales]	0.011	0.063	0.858
Region [Northern Ireland]	-0.106	0.064	0.102
Age : BMI	-0.0002	0.0002	0.381
Age : Sex.Preg [Female: NP]	-0.016	0.004	0.0001
Age : Sex.Preg [Female: Preg.]	0.006	0.004	0.128
start_day [Tue]	0.087	0.072	0.231
start_day [Wed]	0.315	0.072	<0.0001
start_day [Thu]	0.323	0.068	<0.0001
start_day [Fri]	0.284	0.069	<0.0001
start_day [Sat]	0.150	0.071	0.035
start_day [Sun]	-0.050	0.074	0.494

Table 4.4: *model estimates of the two-part model of the second stage*

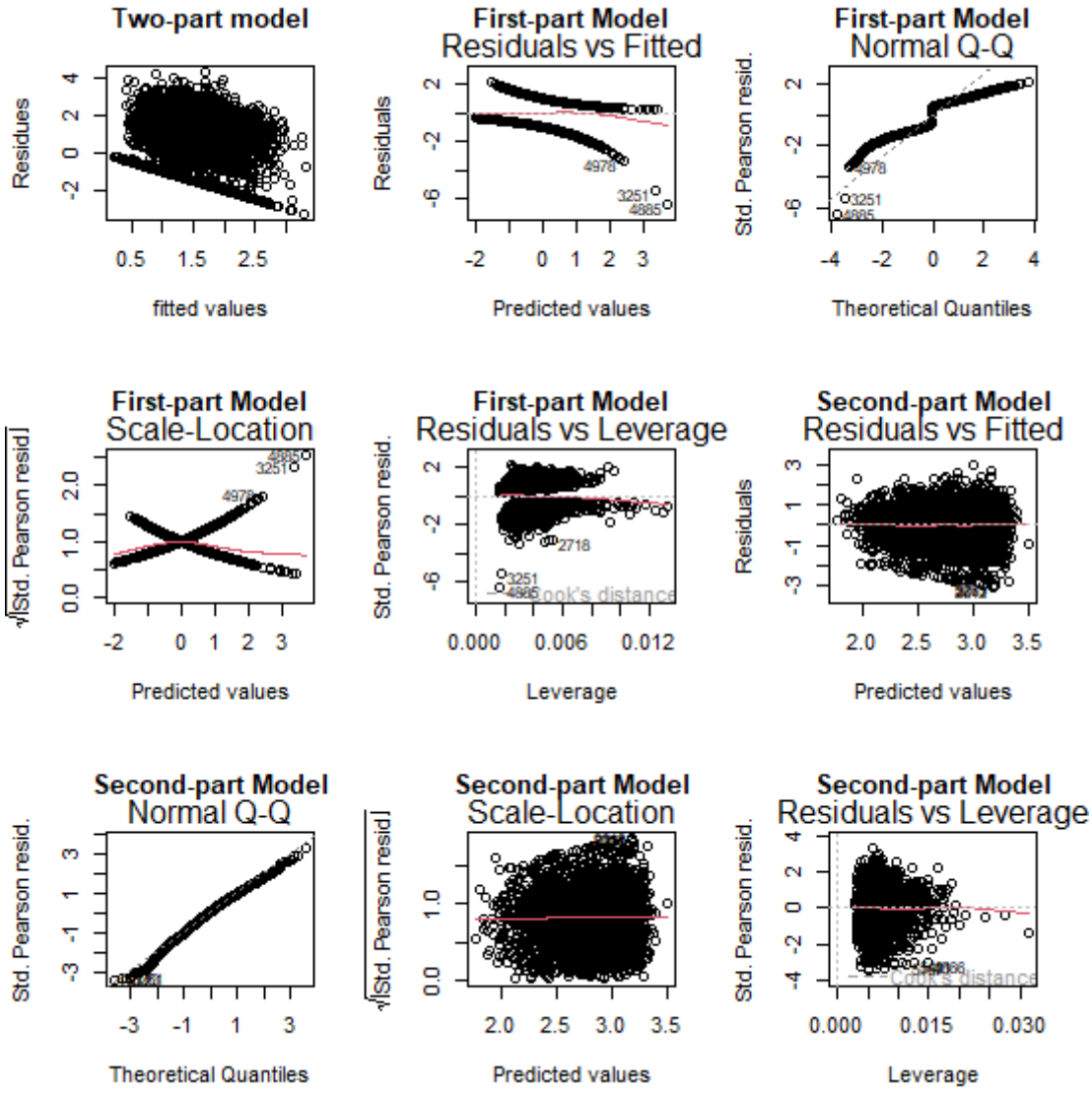


Figure 4.4: *diagnostics for the fitted two-part model*

5 CONCLUSION

At the end of this report, a lot of questions are still not fully covered in details but the knowledge gained from the study is definitely an efficient starting point for further work.

5.1 Summary

The goal of the research was to find suitable strategies to address the following question: "*How can the alcohol intake be modelled for the UK population?*". In order to address it the source of the data for the research has been immediately presented.
The National Diet and Nutrition Survey is one of the most important official statistics instruments at the service of the UK government to make pondered choices on target policies.

To better understand its role on the research, details about the design of the survey and the sampling characteristics were mentioned, together with an overview of the context and the variables at disposal.
Subsequently, the variables involved on the AI data were introduced and an exploratory analysis of the most prominent features was carried out detailing odd situations for measures like equivalised income and BMI. This was a consequence of the presence of many missing values in the data records, therefore it was essential to make an in-depth study on how those issues could be resolved. Indeed, the study relied on multiple imputation to fill out the empty entries making sure that the process was trustworthy enough to make the investigation on AI for the general adult UK population still valid. Predictive mean matching was the chosen method to impute equivalised income , height and weight. The latter variables were especially crucial for the appropriate imputation of BMI exploiting the real-life formula of the index starting exactly from them. Since the imputations were evaluated to be of good quality, the study could proceed.

The different choices of the model depended on the different natures of the AI variable. Indeed, three possibilities have been evaluated:

- a binary logistic model for the drinker / teetotaller status;
- an ordinal regression model for the ordered categorical outcomes with 8 levels of drinking frequencies;
- a two-part mixture model to give the chance of using the continuous measurement of the AI but still be able of taking care of the zero inflation for its distribution. In particular, a logistic model is used to assess if the observation are positive or not, then for the ones which actually are greater than 0 a log-Normal model is fitted.

The results of those models have been presented and their complexities were highlighted. Overall, the model who would perform worse is the ordinal one, since the assumption are not really met. Even if the binary logistic model showed a lower AIC, the general goodness of fit was not as good as for the two-part model, where the required assumptions are fulfilled successfully. Also, the binary logistic model showed many non significant variables, so estimates might no be spotted on. However, it must be mentioned how for the mixture the model the imputed data sets could not be used and the study relied on a simple listwise deletion of the missing data, which might not be the best setting for this study since the risk is of having poor amount of data.

5.2 Further work

A lot of work is still needed to accurately answer the key questions of the study.

The main goal would be to have a better understanding of the nature of missingness in the NDNS so that the imputation models could be improved and a have a stronger theoretical foundation. It would also be ideal to implement a new package in R able to deal with mixture models had them being still a new concept in the broad spectrum of multiple imputation.

A PEER REVIEW REFLECTION

REFLECTION ON THE FEEDBACK The peer review received for this projected highlighted a good engagement of the introduction chapter; therefore, this characteristic of in-depth explanations and wider-scope of the statistical language was attempted also for the other chapters of the report.

The most criticised aspects were the lack of a section dealing solely with the aims of the project instead of making them known from the beginning of the abstract. Therefore, it seemed more appropriate to do so and a section on the aims was added where the key question were explicited.

Seeking better captions for the tables presented up to that point made me really focus on making the figures and any additional feature outside of the text fully captioned with clear and concise descriptions. The goal is to make the graphs understandable already by their comments and not necessarily needing to look for them in the data.

Lastly, it was suggested to use a different style for the bibliography and increase the use of references and quotation. As soon as I started getting more confident with the report structure and template, those characteristics became a fundamental aspect of the overall outline.

REFLECTION ON THE PROCESS Putting myself in a constructive critical position towards my peer really helped me in being at the same time fairer and focused on my actual project. Indeed it is really interesting to see how other people with more or less the same background as yours address some transversal topics that were present in many other projects. This helps in keeping the mind open to inspirations and getting more insight so that it is possible to have an idea of the bigger picture when evaluating statistical analysis.

In particular, I have found really important to properly explain with care the scientific notions behind the practicalities of the project, otherwise it will be really difficult for people not experience in a particular field of study to be able to understand even the foundation points of a project.

B ADDITIONAL TABLES

Region	# Drinkers	# Non-Drinkers	# NA
England: North	980	214	1
England: Central/Midlands	674	186	1
England: South	1645	382	1
Scotland	902	213	1
Wales	820	175	0
Northern Ireland	696	219	2

Table B.1: *Sample contingency table for region and drinking status*

Sex + Pregnancy status	# Drinkers	# Non-Drinkers	# NA
Male	2475	481	4
Female: Non-Pregnant	1371	510	0
Female: Pregnant	1871	398	2

Table B.2: *Sample contingency table for sex and drinking status*

Bibliography

Austin, P. C., White, I. R., Lee, D. S. and van Buuren, S. 2021, 'Missing data in clinical research: A tutorial on multiple imputation', *The Canadian Journal Cardiology* **37**(9), 1322–1331.

Bonakdarpour, M. 2016, 'Introduction to mixture models', `https://stephens999.github.io/fiveMinuteStats/intro_to_mixture_models.html`. Last updated: 31 March 2019; Last accessed: 15 March 2023.

Harrell, F. E. 2015, Ordinal logistic regression, *in* 'Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis', second edn, Springer Series in Statistics, Springer.

Henderson, L., Gregory, J., Irving, K. and Swan, G. 2003, The national diet & nutrition survey: adults aged 19 to 64 years. volume 2: Energy, protein, carbohydrate, fat and alcohol intake, Technical report, Social Survey Division of the Office for National Statistics.

Hosmer, D. W. and Lemeshow, S. 2000, *Applied Logistic Regression*, Wiley Series in Probability and Statistics, second edn, John Wiley & Sons.

Kassambara, A. 2018, 'Logistic regression assumptions and diagnostics in r', `http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/`. Last accessed: 15 March 2023.

Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. 2005, *Applied Linear Statistical Models*, fifth edn, McGraw-Hill/Irwin.

Mann, P. S. 2012, *Introductory Statistics*, eighth edn, John Wiley & Sons.

Min, Y. and Agresti, A. 2002, 'Modeling nonnegative data with clumping at zero: A survey', *Journal of the Iranian Statistical Society* **1**(1-2), 7–33.

MRC Epidemiology Unit n.d., 'National diet and nutrition survey', `https://www.mrc-epid.cam.ac.uk/research/studies/ndns/`. Last accessed: 15 March 2023.

National Diet and Nutrition Survey, Year 9 (2016/17), List of Variables for UK Data n.d.. UK Data Archive Study Number 6533 - National Diet and Nutrition Survey.

National Diet and Nutrition Survey, Years 1-4 (2008/09-2011/12), List of Variables for UK Data n.d.. UK Data Archive Study Number 6533 - National Diet and Nutrition Survey.

National Diet and Nutrition Survey, Years 5-6 (2012/13-2013/14), List of Variables for UK Data n.d.. UK Data Archive Study Number 6533 - National Diet and Nutrition Survey.

National Diet and Nutrition Survey, Years 7-8 (2014/15-2015/16), List of Variables for UK Data n.d.. UK Data Archive Study Number 6533 - National Diet and Nutrition Survey.

Public Health England 2016, 'National diet and nutrition survey', https://www.gov.uk/government/collections/national-diet-and-nutrition-survey#full-publication-update-history. Last updated: 22 September 2021; Last accessed: 15 March 2023.

van Buuren, S. 2018, *Flexible Imputation of Missing Data*, second edn, Chapman & Hall.

Venables, M. C., Roberts, C., Nicholson, S., Bates, B., Jones, K. S., Ashford, R., Hill, S., Farooq, A., Koulman, A., Wareham, N. J. and Page, P. 2022, 'Data resource profile: United kingdom national diet and nutrition survey rolling programme (2008–19)', *International Journal of Epidemiology* **51**(4), e143—-e155.

Väisänen, H. 2020, 'Ordinal logistic regression', https://www.ncrm.ac.uk/resources/online/all/?main&id=20762. Last accessed: 15 March 2023.