Impact of Age, Gender, Socio-Economic Status, and Lifestyle on BMI Distrubution

Yining Jin, James Robb, Kiran Sandhu, Andrew Speirs

1 Introduction

2 Exploratory Analysis

Table 1 shows the Mean, Median and Standard deviation of the BMI results from 2013 - 2016. Looking at Table 1 Below, we can see that the mean and median values for the BMI have a slight increase from 2013 to 2016, which would suggest that there is a population-wide trend of increasing BMI values.

Also, the standard deviation results provide information about the variability or spread of the BMI values within each of the years. The standard deviation results from Table 1 also increase from the 2013 - 2016, indicating that there is more diversity in the BMI results in 2015 and 2016.

Table 1: Mean, Median and Standard deviation of the BMI from 2013-2016

Year	Mean	Median	SD
2013	27.81	27.08	5.28
2014	27.84	27.17	5.21
2015	28.02	27.26	5.53
2016	28.03	27.28	5.79

Figure 1 shows that despite the similar median and quartile values, there are notable differences in the spread and variability of the BMI results across the 4 years, with 2015 and 2016 having the most outliers. The presence of many outliers suggests that there is substantial variability and dispersion of BMI results for each year.

Also, Looking at Figure 1, we can see that the histogram is slightly right skewed with the majority of the observations on the left-hand side of the histogram and a long tail extending towards the right. A right skewed histogram is also known as a positively skewed histogram and the peak of the graph can be found on the left-hand side and more specifically, between 20-30 BMI from the results.

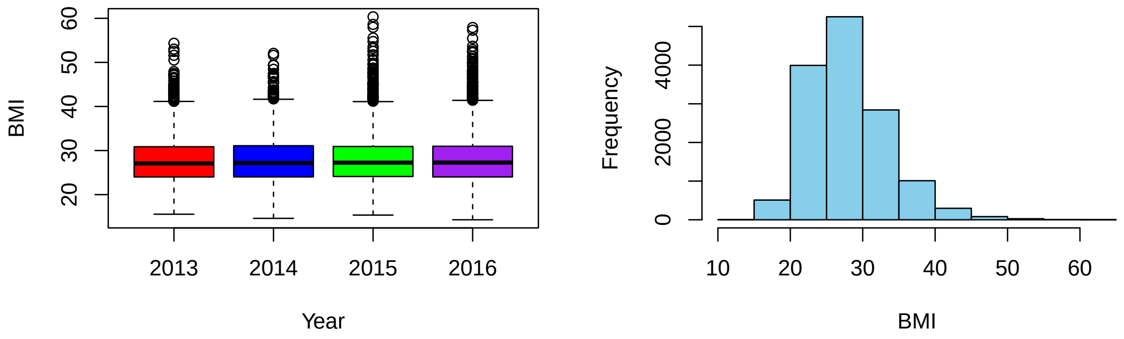

Figure 1: Boxplot for the BMI results for each year(Left) and Histogram of the BMI(Right)

3 Formal Analysis

Analysis was conducted using R Studio and various packages. The significance level for all model tests was $\alpha = 0.05$.

3.1 Question 1

We start by fitting a linear model about the relationship between BMI in Scotland and the given years. To assess H_0 : "Year is not a significant predictor for BMI", the following linear model, model1, was fitted:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \beta_3 \cdot x_{3i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, ..., 14017, \\ &= \beta_0 + \beta_{2014} \cdot \mathbb{1}_{2014}(x) + \beta_{2015} \cdot \mathbb{1}_{2015}(x) + \beta_{2016} \cdot \mathbb{1}_{2016}(x) + \epsilon_i, \end{aligned}$$

Where

- β_0 is the intercept of the regression line for the baseline year (2013);
- β_{year} is the additional term added to β_0 to get the intercept of the regression line for the specified year;
- $\mathbb{1}_{\text{year}}(x)$ is an indicator function indicating the chosen year.

Table 2: Estimates of the model1 coefficients.

term	estimate	std_error	statistic	p_value
intercept	27.810	0.090	310.180	0.000
Year.2014	0.029	0.128	0.224	0.823
Year.2015	0.213	0.128	1.672	0.095
Year.2016	0.223	0.132	1.697	0.090

Results show that the 'Year' predictor is not statistically significant with BMI as all of the p-values are greater than 0.05. So, we are unable to reject the null hypothesis and we conclude that there might be no difference in average BMI across the given years. Now, let's look at the assumptions of this model.

Figure 2 displays the residuals versus fitted values (Left) and histogram of the Residuals (Right). Although there appears to be a greater number of data points distributed above the zero line in the graph, given the size of the dataset, we can make the assumption that the mean of the data is centered around zero. Therefore, we can conclude that the assumptions regarding the residuals, specifically that they have mean zero and constant variance across all of the fitted values, are satisfied. Now, looking at the histogram, the residuals appear to be bell-shaped and they seem to be centered at zero. Even though there seems to be some right skewness in the histogram, due to the large dataset, we can still conclude that the assumptions of constant variance and mean zero for the residuals can be satisfied.

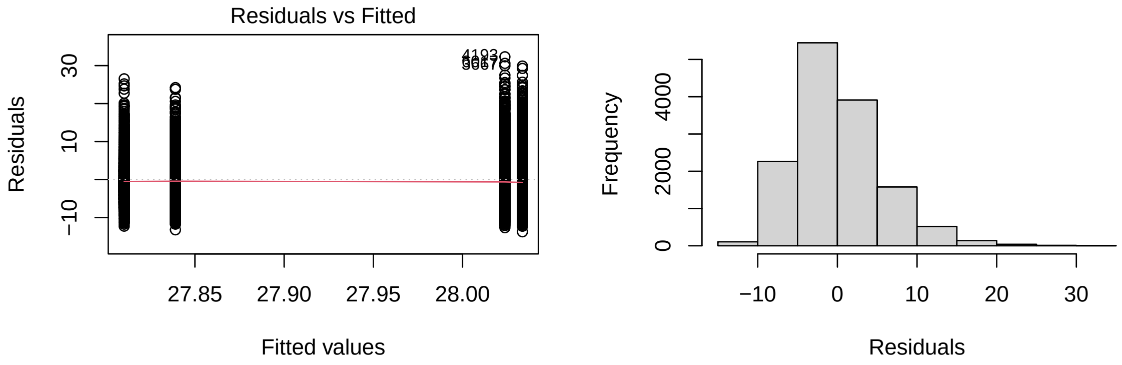

Figure 2: Residuals V Fitted values(Left) and Histogram of Residuals(Right)

3.2 Question 2

For the rest of the variables, namely age, sex, highest education qualification attained, sufficient vegetable intake and sufficient fruit intake, we can fit linear regression models to each of these individually to see whether they have a significant impact on the BMI of an individual. The table below represents the results from all 5 of these models that we have fitted.

Table 3: estimates of the model coefficients

Term	estimate	p-value
intercept	25.636	0.000
Age	0.045	0.000
intercept	27.898	0.000
Sex: Male	0.056	0.544
intercept	27.234	0.000
Education: Higher grade or equiv	0.357	0.012
Education: HNC/D or equiv	0.929	0.000
Education: No qualifications	1.756	0.000
Education: Other school level	1.440	0.000
Education: Standard grade or equiv	0.800	0.000
intercept	28.018	0.000
Veg: Yes	-0.122	0.267
intercept	27.752	0.000
Fruit: Yes	0.248	0.013

Hence, from Table 3 we obtain the following regression lines:

$$\widehat{\text{BMI}} = 25.64 + 0.05 \cdot Age$$

$$\begin{aligned}\widehat{\text{BMI}} = {} & 27.23 + 0.36 \cdot (Education : Higher\ grade\ or\ equiv) + 0.93 \cdot (Education : HNC/D\ or\ equiv) + \\ & 1.76 \cdot (Education : No\ qualifications) + 1.44 \cdot (Education : Other\ School\ level) + \\ & 0.80 \cdot (Education : Standard\ grade\ or\ equiv)\end{aligned}$$

$$\widehat{\text{BMI}} = 27.75 + 0.25 \cdot (Fruit : Yes)$$

From all of these models that have been fitted in table 5 , we can observe the p-values of each of these, and we get the following results from them. The variables Age, Education and Fruit are all statistically significant with BMI distribution whilst Sex and Veg intake are not.

Now, since we have chosen three models that are statistically significant (Age, Education and Fruit), we can now test for the assumptions of each of these models.

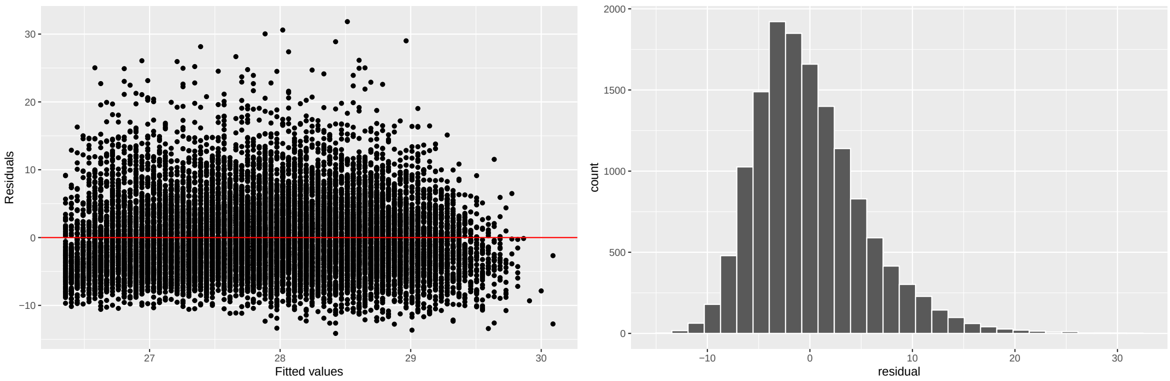

Figure 3: Assumption checking for Age model

We shall begin with the Age model

As we can see from Figure 3 , we have that there is an even spread of the residuals above and below the zero line in both the graph of residuals against Age (left) and against fitted values (middle), with a few outliers above, however due to n being large we can ignore these. In the histogram of residuals we also have the residuals are centered at zero however they are slightly right skewed, nevertheless, due to the large sample size n, we can say this model meets the assumptions.

We will now look at the assumptions for the Education model:

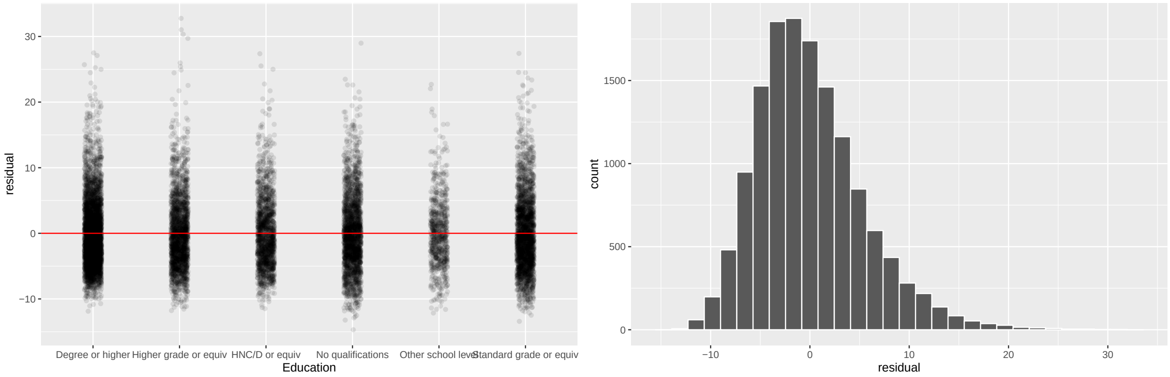

Figure 4: Assumption checking for Education model

In the plot of residuals against highest education qualification attained in Figure 4 ,(left), we can see in the there is an even split of the residuals above and below the zero line, thus implying that the residuals have mean zero. In the histogram of residuals (right), the residuals are centered at zero however there is a slight right skew once again, however since n is large enough we can say that this satisfies the assumptions.

Finally, we can look at the assuptions for the fruit model:

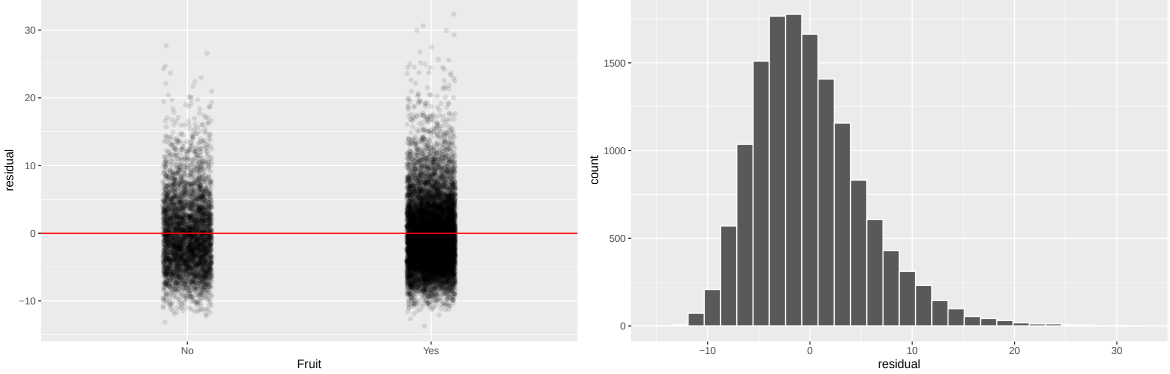

Figure 5: Assumption checking for fruit model

In the plot displaying residuals against the highest education qualification attained (left) in @fig-fruitmodelass, we observe an equal distribution of residuals above and below the zero line, indicating a mean residual value of zero. In the histogram of residuals (right), although the residuals are centered around zero, there is a slight right skew. However, given the sufficiently large sample size (n), we can conclude that this meets the underlying assumptions.

4 Conclusions