Investigating the effect of age, gender, socio-economic status, lifestyle factors and a changing time period has on BMI Distrubution

Yining Jin, James Robb, Kiran Sandhu, Andrew Speirs

Introduction

INTRO

The Question of interests are:

1) Has the body mass index (BMI) in Scotland changed over the given years of the Scottish Health Survey?
2) Are there any differences in the BMI distribution by age, gender, socio-economic status or lifestyle factors?

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.

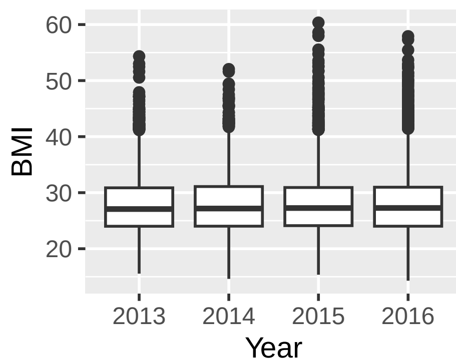

Figure 1: Boxplot for the BMI results for each year

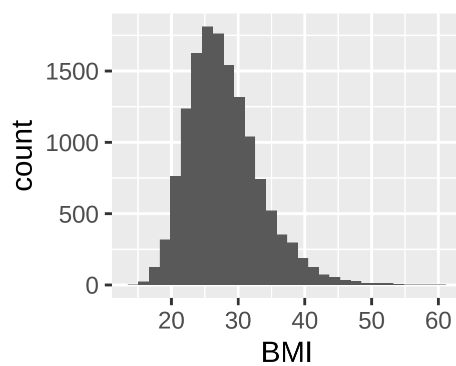

Figure 2: Histogram of the BMI results

Looking at Figure 2 below, we can see that it 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.

Table 2 Shows the correlation between BMI and the other variables in the data. Here, we can see that all the correlation results are of very low correlation which suggests that the variables under consideration are not strongly related in a linear manner. Maybe looking at each year separately would show more accurate results and help find variables that are of stronger correlation.

Table 2: Correlation between BMI and other variables

Variables	Correlation
Age	0.14537
Sex	0.00512
Education	−0.00150
Veg	−0.00937
Fruit	0.02094

Now we will look at the formal analysis for each of the questions of interest below.

Formal Analysis

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

The general linear model was used to assess the trend between Year and BMI distributions. To assess H_0 : "Year is not a significant predictor for BMI", the following linear model, model1, was fitted:

$$BMI = \beta_0 + \beta_1 \cdot Year$$

Where

- β_0 is the intercept of the regression line.
- β_1 is the slope of the regression line.

Table 3: Estimates of the model1 coefficients.

term	estimate	std_error	statistic	p_value
intercept	−145.306	83.521	−1.740	0.082
Year	0.086	0.041	2.074	0.038

Results show that the 'Year' predictor is statistically significant for predicting BMI ($\beta = 0.086$, $s.e. = 0.041$, $t_{(1,14015)} = 2.074$, $p = 0.038$). We reject the null hypothesis and thus, for every unit increase in ‘Year’, the predicted BMI increases by approximately 0.09 units. Hence, from Table 3 we obtain the following regression line:

$$BMI = -145.306 + 0.086 \cdot Year$$

However, it is important to note that the adjusted R^2 result is very low ($F_{(1,14015)} = 4.302$, $p = 0.038$, $R^2_{adjusted} = 0.00024$), indicating that the ‘Year’ predictor only explains a very small fraction of the variability in BMI (0.02%). Given these results, while the year appears to have a statistically significant effect on BMI in this particular dataset, the practical significance of this effect may be limited.

Now, we will compute t-tests to compare the means of the BMI results for each of the 4 years to see if there is a significant difference between them over the time period. The results of each of the t-tests are shown in Table 4 below:

Table 4: t-test results for each of the given years

Years	P_values
2013-2014	0.81616
2013-2015	0.09222
2013-2016	0.09632
2014-2015	0.14704
2014-2016	0.14911
2015-2016	0.94309

From Table 4, we can see that none of the years which were compared had any significant difference in BMI results so we can claim that the BMI distribution has not significant changed from the years 2013 - 2016.

Descriptive analysis suggested that BMI may be affected by Age, Sex, Education, Veg intake and Fruit intake.

QUESTION OF INTEREST 2 FORMAL ANALYSIS

AIC TESTS FOR THE MODELS

ASSUMPTION CHECKING FOR THE BEST MODEL

Conclusions

CONCLUSION