# Obesity in Scotland in relation to factors included in the Scottish Health Surveys

Group 9 (2718740K Kyle Kim)

## 1 Introduction

Obesity represents a significant public health issue globally, linked to numerous chronic diseases and a major cause of premature morbidity and mortality. Scotland has the highest obesity rate in the UK in addition to having one of the highest in the developed world which clearly emphasises the significance of the problem. The Scottish health surveys from 2008 and 2012 collected data from individuals regarding their Age, Sex, Education, Body Mass Index (BMI) and whether they consumed the recommended daily fruit and vegetable intake in order to examine the prevalence and determinants of obesity in Scotland. It aims to identify trends in obesity rates over the years and explore how demographic factors, socio-economic status, and lifestyle choices such as diet impact obesity levels. Through descriptive statistics and logistic regression analysis we will be investigating whether the prevalence of obesity in Scotland changed over the given years (2008 to 2012). Additionally, whether the obesity status of individuals is effected differently in relation to their age, gender, socio-economic status or lifestyle factors. From this, we can offer evidence-based recommendations for public health strategies to mitigate this growing health concern.

Section 2 consists of an exploratory analysis of the Scottish health survey data and explores the stated questions of interest. Section 3 contains the results from fitting a multiple regression model to the data, as well as the assessment of the model assumptions. Concluding remarks are given in Section 4.

## 2 Exploratory data analysis

The Figure 2 data indicates minor fluctuations in Scotland's obesity rates from 2008 to 2012, with a notable peak at 30.47% in 2010. The graph underscores this trend, showcasing the temporary surge in 2010 against a backdrop of overall stability.

The Figure 1 indicate a correlation between age and obesity prevalence in Scotland, showing an increase in obesity rates with age, peaking at 35.9% among those aged 60-70. This pattern highlights this age group as having the highest obesity proportion. Beyond this peak, obesity percentages decline with further age increase, with individuals aged 90 and above exhibiting the lowest rate at 9.52%.

The Figure 1 shows a slight disparity in obesity rates between genders in Scotland, with females at 29.80659% and males slightly lower at 29.36790%. This minor difference is visually depicted in the barchart, where the female bar is marginally taller than the male bar, underscoring the small

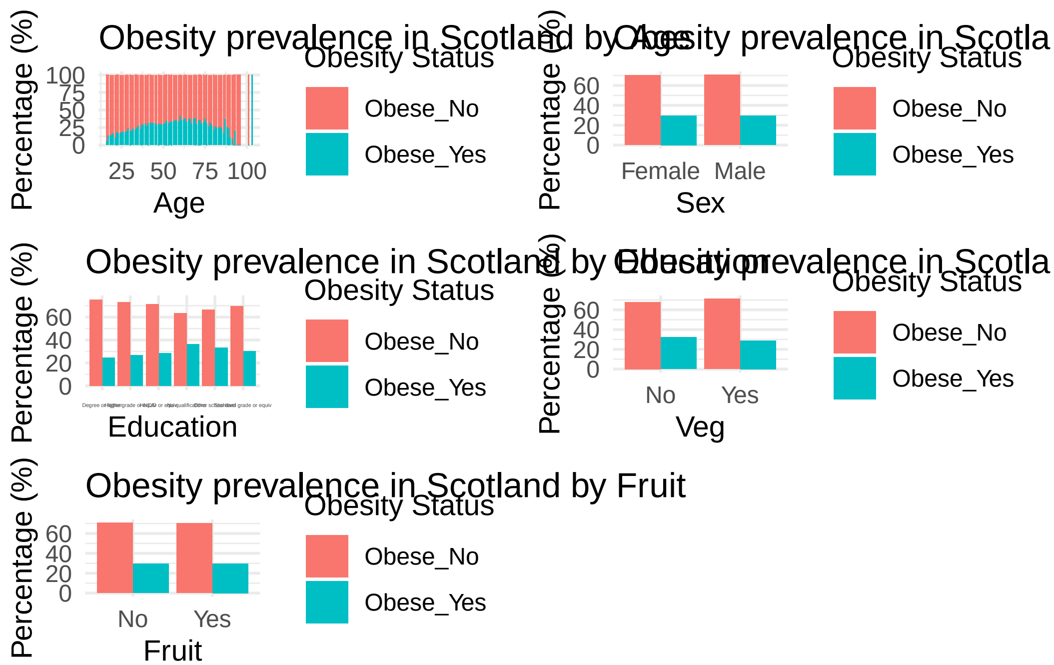


Figure 1: Scotland's Obesity prevalence over the factors

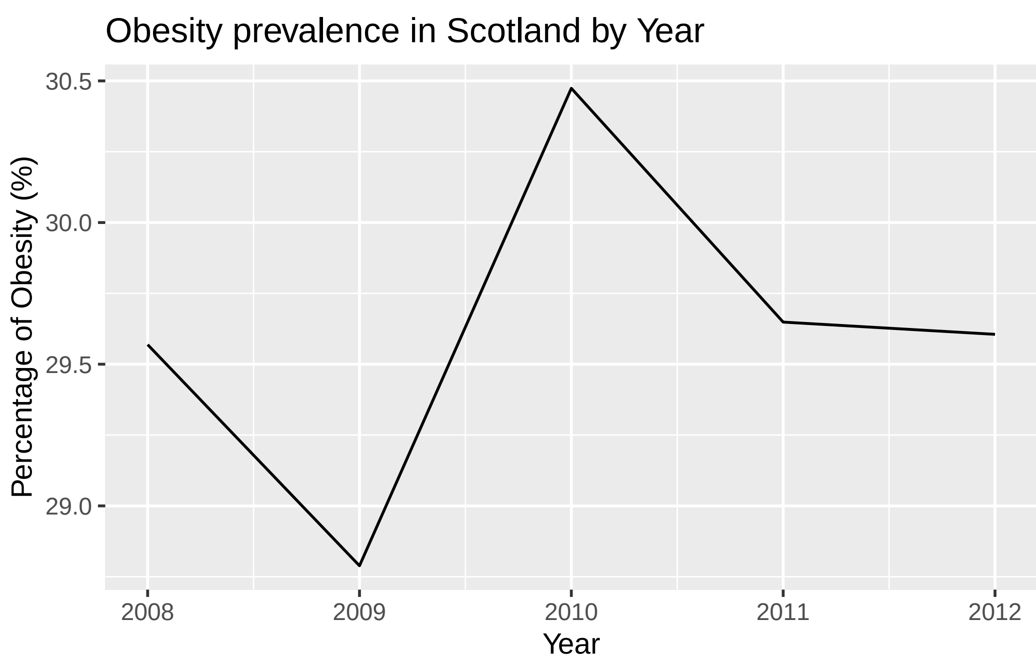


Figure 2: Percentage of obese people in Scotland by year

yet apparent gap. The graph illustrates that obesity is a substantial health concern for both sexes, indicating high prevalence rates across genders.

The Figure 1 display a trend where obesity rates in Scotland decrease with increasing educational levels, from the highest rate among those without qualifications (36.46659%) to the lowest in individuals with degrees or higher (24.77144%), suggesting an inverse correlation between education and obesity.

The Figure 1 show that in Scotland, non-vegetarians have a higher obesity rate (32.06191%) compared to vegetarians (28.93873%), with the bar for non-vegetarians visibly taller in the graph. This suggests a correlation between diet choice and obesity prevalence.

The Figure 1 illustrate a slight difference in obesity rates between fruit consumers (29.72168%) and non-consumers (29.36099%), with nearly equal bar heights in the graph. This suggests a minimal impact of fruit consumption on obesity prevalence, highlighting the complex interplay of dietary and lifestyle factors in determining obesity.

# 3 Formal data analysis

## 3.1 Prevalence of obesity from 2008 to 2012

Next, we formally analyse the data by considering each objectives in turn. Firstly, The logistic regression model for obesity prevalence from 2008 to 2012 which will be fitted is given below :

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Year(x) + \epsilon, \quad \epsilon \sim N(0,\sigma^2) i \ = \ 1, \cdots, 25224$$

where

- $p$ denotes the probability of the outcome being obese (outcome 1)
- $\beta_0$ denotes the intercept of the regression line for the baseline Year (2008)
- $\beta_1$ denotes the coefficients for the specified Year
- $\epsilon$ denotes random error component which are normally distributed with mean zero
- $\sigma^2$ denotes variance

The analysis of the data presented in Table 1 reveals that the coefficients 0 and 1 are -12.13 and 0.0056, respectively. Despite these values, the significance tests indicate that they are not statistically significant at the 5% level, with p-values of 0.5581 and 0.5865, respectively. This aligns with the preliminary observations discussed in Section 2, reinforcing the conclusion that the evidence is inadequate to suggest any significant variation in obesity prevalence in Scotland from 2008 to 2012. Thus, the initial hypothesis suggesting a change in obesity rates within the observed period is not supported by the empirical analysis.

Table 1: Estimates of the regression model coefficients (year).

| Groups | Estimates | pvalues |
|---|---|---|
| Intercept | -12.1300429 | 0.5580535 |

| Groups | Estimates | pvalues |
|---|---|---|
| Year | 0.0056043 | 0.5864945 |

### 3.2 Prevalence of Obesity on the explanatory variables Age, Sex, Education and dietary habits

The model was established by including an array of independent variables such as age, sex, various educational levels, and dietary habits encompassing vegetable and fruit intake. This logistic regression model was selected due to the binary nature of the dependent variable—obesity, categorized as 'obese_yes' or 'obese_no'. The model sought to express the log-odds of the probability of being obese as a linear combination of the predictors, as illustrated by the logistic function:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0+\beta_1 Age(x_1)+\beta_2 Sex(x_2)+\beta_3 Education(x_3)+\beta_4 Veg(x_4)+\beta_5 Fruit(x_5)+\epsilon,\ \ \epsilon \sim N(0,\sigma^2) i\ =\ 1,\cdots,25224$$

where

- $p$ denotes the probability of the outcome being obese (outcome 1)
- $\beta_0$ denotes the intercept
- $\beta_1, \cdots, \beta_5$ denotes the coefficients of the predictor variables
- $\epsilon$ denotes random error component which are normally distributed with mean zero
- $\sigma^2$ denotes variance

we can fit the logistic regression model as follows **?@tbl-regmodel1**:

Upon evaluation of the `model_factors`, `Sex (Male)` and `Fruit` consumption were found to have p-values exceeding the threshold of 0.05, indicating that they do not significantly contribute to the prediction of obesity. Consequently, these variables were removed to refine the model. The refined model coefficients were obtained and are presented in Table 2, which excludes the aforementioned insignificant factors.

To evaluate the goodness of fit for the logistic regression model, the stepwise selection procedure known as stepAIC was employed. The application of stepAIC to the logistic regression model resulted in the exclusion of variables 'Sex' and 'Fruit' intake, as their presence did not contribute to a reduction in the AIC score, suggesting that their inclusion did not improve the model's predictive ability significantly. The final model, which excludes these variables, is considered to be more parsimonious and theoretically sound for predicting the probability of obesity in the studied population.

Table 2: Estimates of the regression model coefficients.

| Groups | Estimates | pvalues |
|---|---|---|
| Intercept | -1.5562531 | 0.0000000 |
| Age | 0.0116245 | 0.0000000 |
| Education Higher | 0.1476566 | 0.0016097 |

| Groups | Estimates | pvalues |
|---|---|---|
| Education HNC/D | 0.2404329 | 0.0000036 |
| Education no | 0.3856073 | 0.0000000 |
| Education other | 0.2187677 | 0.0001871 |
| Education Standard | 0.3046388 | 0.0000000 |
| Vegetable intake | -0.1395564 | 0.0000362 |

The logistic regression equation derived from the model refinement using the stepAIC process, each independent variable's coefficient is integral to the calculation of the log-odds of the probability, p, of an individual being obese. The logistic function is represented as follows::

$$\log\left(\frac{p}{1-p}\right) = -1.556 + 0.012Age + 0.148Education_{Higher} + 0.240Education_{HNC/D} + 0.386Education_{No} + 0.219Education_{O}$$

Further analysis and interpretation of this model should take into account the odds 95% confidence intervals for these coefficients to assess the precision of the estimates. These intervals are essential for understanding the range within which the true value of the coefficients is likely to lie, with a 95% level of confidence.

Then, the odds ratios and their corresponding 95% confidence intervals provide valuable insights into the factors associated with obesity. The confidence intervals for the odds ratios of all predictors do not encompass the value of 1 (see **?@tbl-the_odds** and Figure 3). This observation is critical as it implies that the odds of obesity are significantly different from the null hypothesis value (odds ratio = 1) for each predictor. Therefore, we can conclude with 95% certainty that age, various levels of education, and vegetable intake are statistically significant factors in the prediction of obesity within our model. Specifically, age and lower educational attainment are associated with increased odds of obesity, whereas vegetable intake is associated with decreased odds, underscoring the importance of these factors in public health interventions aimed at reducing obesity prevalence.

For the results, the model aimed to predict probabilities of obesity was developed considering various factors, including 'Sex' and 'Fruit' consumption. However, upon evaluation, both 'Sex' and 'Fruit' were identified as statistically insignificant predictors within the model. This inclusion of insignificant factors could potentially lead to inaccuracies in the predicted probabilities of obesity. Therefore, we have omitted these variables, to enhance the accuracy and reliability of the obesity probability predictions.

Figure 4 presents the predicted probabilities of obesity across different ages, indicating a nearly linear relationship. The graph shows a consistent increase in the probability of obesity as age advances, starting from around 20% in the youngest age group and ascending to close to 45% in the oldest age group represented. The trend line is almost straight, suggesting a steady rate of increase in the likelihood of obesity with age.

The shaded region around the trend line represents the confidence interval, which exhibits a slight increase in width as age progresses. This suggests a small increase in uncertainty of the predictions for the older age groups. Nonetheless, the relatively uniform width of the confidence interval across the age spectrum implies that the variance in the predicted probabilities of obesity does not drastically change with age.

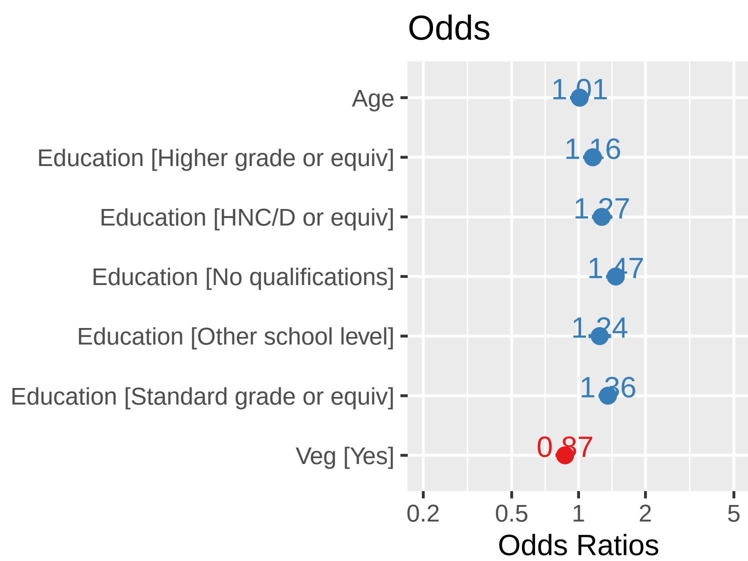


Figure 3: odds of each parameter

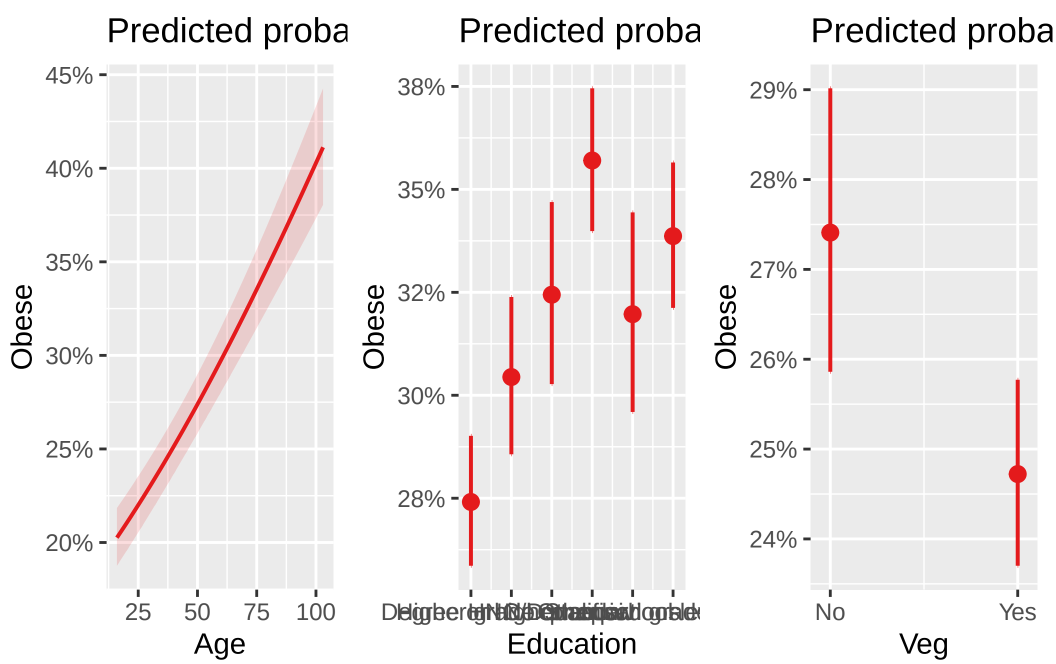


Figure 4: Predicted probabilities of obesity

Figure 4 presents a statistical analysis on the correlation between educational levels and the projected likelihood of obesity. The data suggests a distinct inverse relationship between educational attainment and the propensity for obesity. Specifically, the graph indicates that the segment of the population without any educational qualifications registers the highest mean predicted probability for obesity. In stark contrast, individuals who have obtained a degree or higher education are attributed with the lowest mean predicted probability of being classified as obese.

Additionally, an examination of the variance in the predicted probabilities reveals significant heterogeneity within the 'Other school level' category, which surpasses the uniformity observed in both the 'No qualifications' and 'Degree or higher' categories. This variability may suggest a complex interplay between education and other socio-economic factors that influence obesity rates. These findings highlight the importance of educational attainment in public health strategies aimed at combating obesity.

Figure 4 illustrates the predicted probabilities of obesity with respect to vegetable consumption, as indicated by the binary categories 'Yes' and 'No'. The graph suggests that individuals who do not consume vegetables have a higher predicted probability of obesity, marked by a probability just over 28%. In contrast, the predicted probability for obesity among those who do consume vegetables is significantly lower, indicated by a probability just under 25%.

The error bars, which represent the confidence intervals, are notably longer for the 'No' category, implying greater uncertainty in the prediction for individuals who do not consume vegetables. The shorter error bars for the 'Yes' category suggest more confidence in the prediction for individuals who do consume vegetables.

This visual data suggests a potential inverse relationship between vegetable consumption and the likelihood of obesity, indicating that vegetable intake may be associated with a lower probability of being obese.

# 4 Conclusions

To conclude, from the investigation into the prevalence and likely factors of Obesity in Scotland using the data provided by the Scottish Health surveys between the years 2008 and 2012, the prevalence of obesity has marginally increased by approximately 0.04%. Despite the overall increase, the increase was not consistent over all years the Survey was conducted as the only increase was found between 2009 and 2010 whereas the prevalence decreased for all other years. In addition, from the analyses it was found that gender along with fruit intake do not effect obesity prevalence whereas age, socio-economic status and vegetable intake do. Specifically, those between the ages 60 and 70 have the highest percentage of obesity. Furthermore, those with no qualifications have a 36.5% of obesity, which was the highest out of those with any form of qualification. Finally, Scottish individuals who don't consume the recommended daily fruit intake have a higher proportion of obesity when compared to those who do.