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

Group 9 (2718740K Kyle Kim, 2559983C Rachel Crossan, 2572501G Matthew Gillen, 2652154K Scott Kerr)

1 Introduction

Obesity is a significant global public health concern, associated with various chronic diseases and premature morbidity and mortality. Scotland's obesity rate is the highest in the UK and among the highest in the developed world, highlighting the severity of the issue. The Scottish health surveys of 2008 and 2012 collected data on Age, Sex, Education, Body Mass Index (BMI), and adherence to recommended fruit and vegetable intake to examine obesity prevalence and determinants. Using descriptive statistics and logistic regression analysis, we will explore changes in obesity prevalence from 2008 to 2012 and assess differences in obesity status based on age, gender, socioeconomic status, and lifestyle. These findings will inform evidence-based recommendations for public health strategies to address this growing health issue.

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 1 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.

Figure 2 indicates 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, after this point the obesity prevalence drops off. Figure 3 shows a slight disparity in obesity rates between genders in Scotland, with females at 29.8% and males slightly lower at 29.4%, indicating obesity is an issue for both sexes. Figure 4 displays a trend where obesity rates in Scotland decrease with increasing educational levels, from the highest rate among those without qualifications (36.5%) to the lowest in individuals with degrees or higher (24.8%), suggesting an inverse correlation between education and obesity. Figure 5 shows that those who do not consume the recommended daily intake of vegetables have a higher obesity rate (32.1%) than those who do (28.9%). Figure 6 illustrates a slight difference in obesity rates between

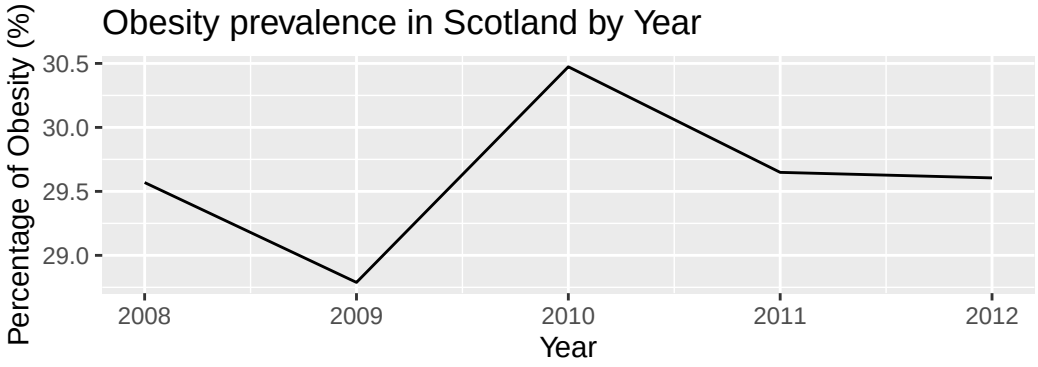

Figure 1: Percentage of obese people in Scotland by year

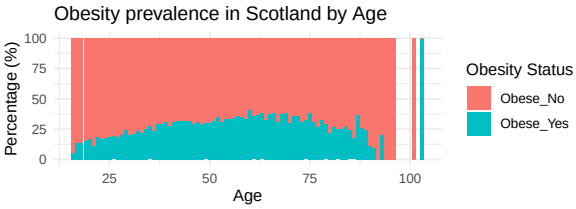

Figure 2: Obesity by Age

Figure 3: Obesity by Sex

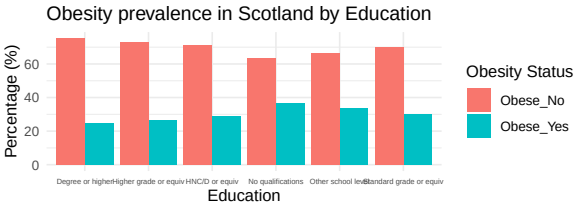

Figure 4: Obesity by Education

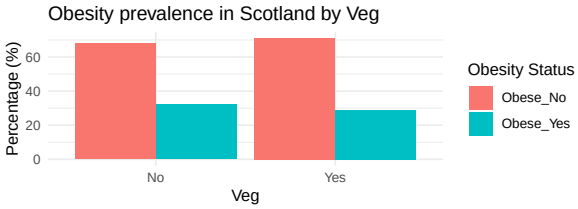

Figure 5: Obesity by Veg

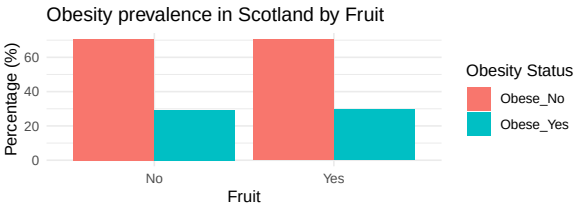

Figure 6: Obesity by Fruit

fruit consumers (29.7%) and non-consumers (29.4%), indicating a minimal impact of fruit consumption on obesity prevalence.

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)
- β_0 denotes the intercept of the regression line for the baseline Year (2008)
- β_1 denotes the coefficients for the specified Year
- ϵ denotes random error component which are normally distributed with mean zero
- σ^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
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)
- β_0 denotes the intercept
- $\beta_1, \cdots, \beta_5$ denotes the coefficients of the predictor variables
- ϵ denotes random error component which are normally distributed with mean zero
- σ^2 denotes variance

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.

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
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:

$$\log\left(\frac{p}{1-p}\right) = -1.556+0.012 Age+0.148 Education_{Higher}+0.240 Education_{HNC/D}+0.386 Education_{No}$$
$$+0.219 Education_{Other}+0.305 Education_{Standard}-0.140 Veg+\epsilon,\quad \epsilon \sim N(0,\sigma^2)$$

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.

The confidence intervals for the odds ratios of all predictors do not encompass the value of 1 (see Figure 7). 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, with age and lower educational associated with increased odds of obesity, and vegetable intake with decreased odds.

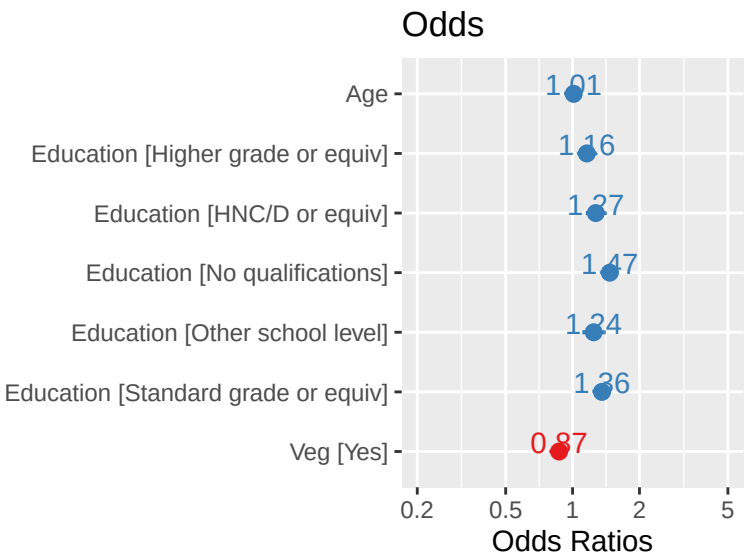

Figure 7: odds of each parameter

Figure 8 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. 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.

Figure 9 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.

Figure 10 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

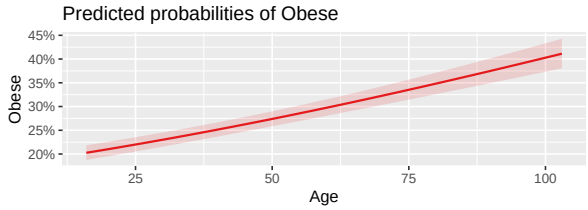

Figure 8: Predicted probability by Age

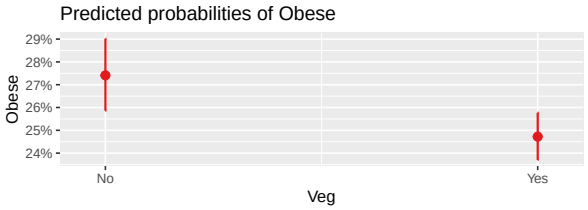

Figure 9: Predicted probability by Veg

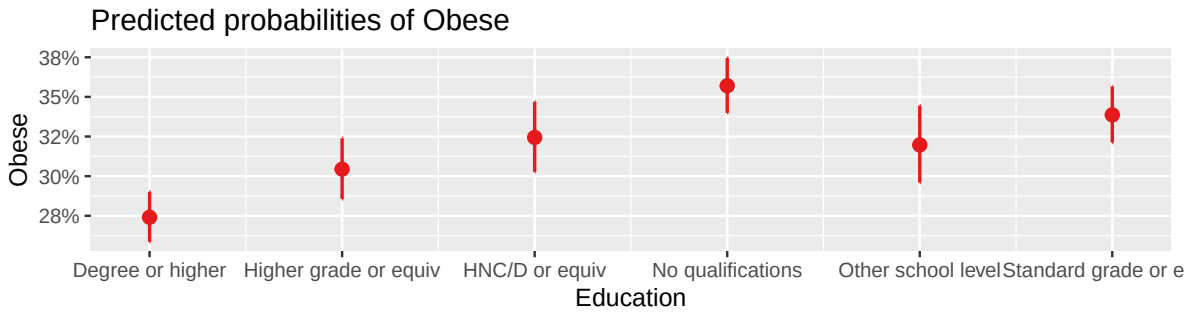

Figure 10: Predicted probability by Education

confidence intervals, are notably longer for the 'No' category, implying greater uncertainty in the prediction for individuals who do not 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, the analysis of obesity prevalence and contributing factors in Scotland based on data from the Scottish Health Surveys spanning 2008 to 2012 reveals a slight overall increase in obesity rates, approximately 0.04%. However, this increase was not consistent across all survey years; while most years showed a decrease in obesity prevalence, there was a notable spike between 2009 and 2010. Sex and fruit intake were not found to significantly impact obesity prevalence, whereas age, socio-economic status, and vegetable intake did. Specifically, individuals aged between 60 and 70 exhibited the highest obesity rates, and those with no qualifications had the highest prevalence at 36.5%. Furthermore, individuals who did not meet the recommended daily vegetable intake had a higher proportion of obesity. Future research could extend the investigation over a longer period to uncover more recent and long-term trends in obesity. Additionally, delving deeper into lifestyle factors such as diet, physical activity, alcohol consumption, and smoking status would provide valuable insights. Examining obesity prevalence at a regional level within Scotland could pinpoint areas with the highest and lowest rates. Lastly, evaluating intervention strategies targeting identified significant factors aims to tackle the obesity issue in Scotland effectively.