Project 1 Obesity Prevalence

AUTHOR
Shihong Wang, Jamie Watt, Mackenzie Whitelock, Irene Hanna Anu

1 Introduction

Obesity is a significant public health concern that affects morbidity and mortality. Monitoring its prevalence is essential for evaluating public health policies and interventions. The Scottish Health Survey, which samples individuals from private households across Scotland, provides an opportunity to study trends in obesity. The primary goal of this analysis is to determine whether there has been a statistically noticeable change in obesity prevalence in Scotland over the period 2008–2012. The secondary goal of this analysis is to assess whether there is a difference in obesity levels in other variables recorded in the Scottish Health Survey between 2008 and 2012.

2 Exploratory Analysis

2.1 Prevalence of Obesity Through The Years

The prevalence of obesity for each year was computed using the formula:

$$\text{Prevalence (in \%)} = \frac{N_{\text{obese}}}{N_{\text{total}}} \times 100$$

First we will produce a line plot with data points to illustrate the trend in obesity prevalence over the years.

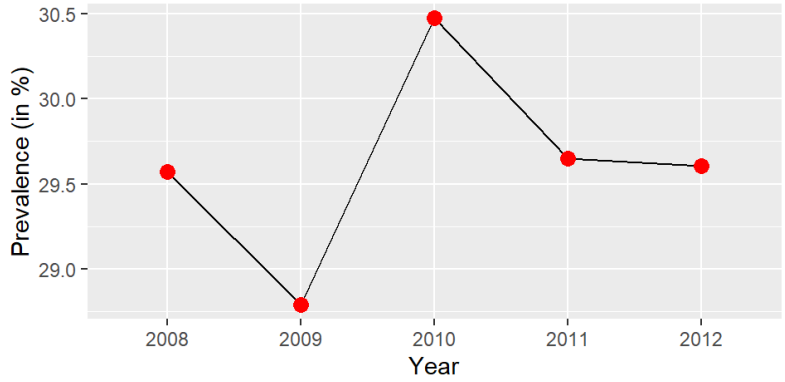


Figure 1: Obesity Prevalence Over Time in Scotland

From the graph we can see that the prevalence appears to fluctuate year to year, rather than moving in a strictly upward or downward path. This suggests that obesity rates may have been influenced by a range of factors that vary from year to year (e.g., sample composition, lifestyle changes, policy impacts).

Also we will produce a bar plot to display the count of individuals classified as obese versus not obese for each survey year.

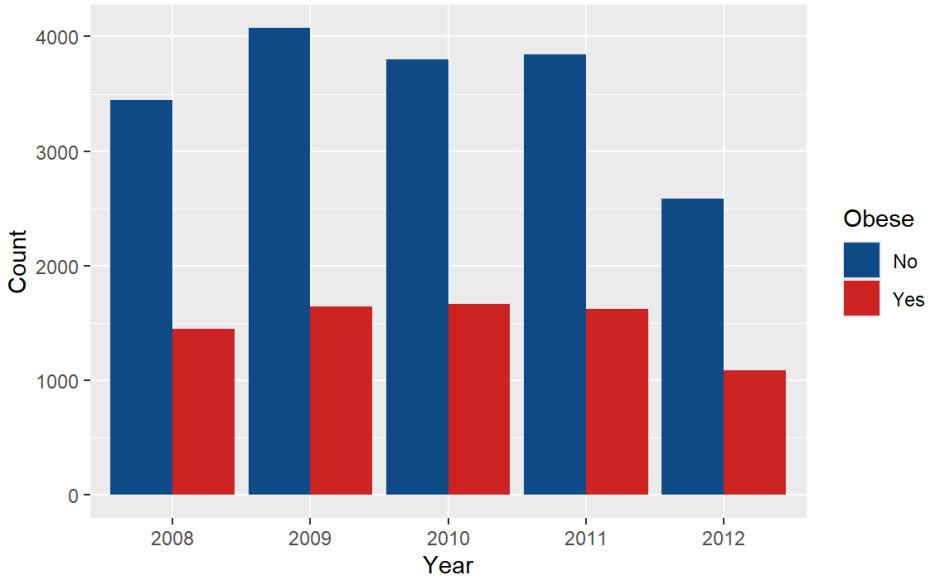
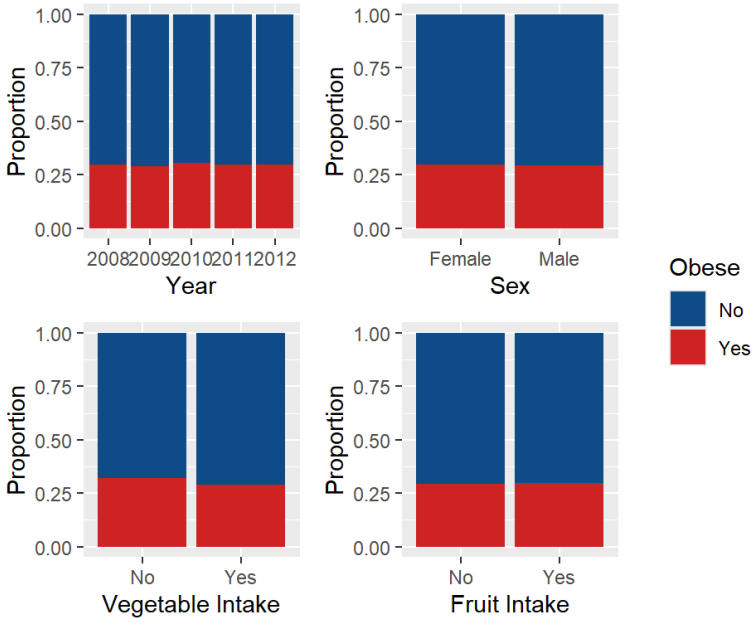


Figure 2: Count of Obesity Classification by Year

We can observe that in each year, the "No" bar (non-obese) is taller than the "Yes" bar (obese). However, the gap between these two bars may differ slightly from year to year.

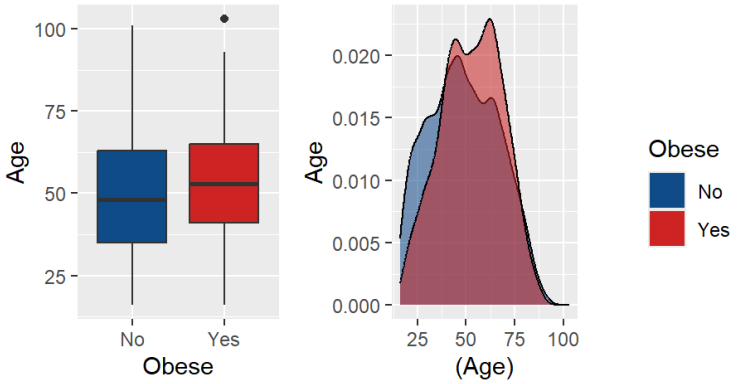
2.2 Differences in Obesity Through The Years

We want to see whether initial plots may give an indication to if there is in fact a difference in obesity levels across the other variables in the survey. We will begin by looking at fruit and veg intake as well as the year and sex to see if there are differences in these areas.



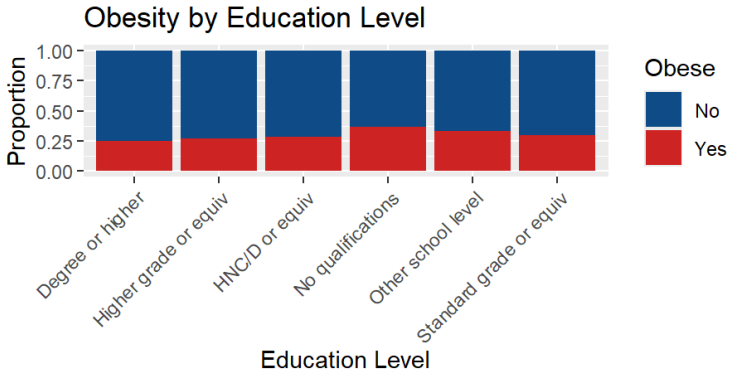
Proportion of Obesity by Year, Sex, Veg Intake and Fruit Intake

We can see here that the proportions of obesity appear to be almost in the same in fruit in take, although there may be a slight difference in whether a person consumes vegetables. We can see that the proportion of obesity across years 2008-2012 do fluctuate slightly but not a lot so we will need to look at this further. There appears to be almost no difference in levels of obesity across males and females though, as these proportions are almost identical. Next we shall look at whether there are differences in obesity levels across ages.



Boxplot and Density plot of age

We can see here that the age of obese people tends to be slightly older than those who are not obese, however we cannot tell whether this is a significant difference from these plots. Finally, we will look at obesity levels across different levels of education.



Proportion of Obesity across Education Levels

From this we can see that the proportion of obesity does seem to vary across education levels, with the higher that level, the lower the proportion of obesity appears to be.

3 Formal Analysis

3.1 Prevalence of Obesity Through the Years

The logistic regression model is given by:

$$y_i \sim \text{Bin}(n_i, p_i)$$

$$logit(p_i) = \alpha + \beta_{year2009} \cdot \mathbb{I}_{\text{year2009}} + \beta_{year2010} \cdot \mathbb{I}_{\text{year2010}} + \beta_{year2011} \cdot \mathbb{I}_{\text{year2011}} + \beta_{year2012} \cdot \mathbb{I}_{\text{year2012}}$$

where $\mathbb{I}_{\text{year}}(x)$ is an indicator function such that

$$\mathbb{I}_{\text{year}}(x) = \begin{cases} 1 \text{ if the xth observation belongs to the particular year,} \\ 0 \text{ otherwise} \end{cases}$$

Where:
- $logit(p_i)$ is the log-odds of the probability of being classified as obese.
- α is the intercept (baseline log-odds when predictors are zero, i.e., the Year 2008).
- $\beta_1, \beta_2, \beta_3, \beta_4$ are the coefficients for the indicator variables corresponding to Years 2009, 2010, 2011, and 2012, respectively.

Logistic Regression Summary

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.868	0.031	-27.716	4.43e-169	-0.930	-0.807
Year2009	-0.038	0.043	-0.882	0.378	-0.122	0.046
Year2010	0.043	0.043	1.003	0.316	-0.041	0.127

term	estimate	std.error	statistic	p.value	conf.low	conf.high
Year2011	0.004	0.043	0.088	0.929	-0.081	0.088
Year2012	0.002	0.048	0.036	0.971	-0.092	0.095

The fitted model is:

$$logit(p_i) = -0.867901 - 0.037769 \cdot \mathbb{I}_{\text{year2009}} + 0.043062 \cdot \mathbb{I}_{\text{year2010}} + 0.003814 \cdot \mathbb{I}_{\text{year2011}} + 0.001743 \cdot \mathbb{I}_{\text{year2012}}$$

- We have fitted a logistic regression model to observe the relation between the years and obesity prevalence in Scotland over the years 2008-2012.
- The baseline category is taken to be Year 2008. From the output odds model summary, we can see that the odds of being obese in Year 2008 was 0.42.
- The coefficients for Year2009, Year2010, Year2011, Year2012 represents the change in odds of being obese compared to the baseline year 2008.
- For Year 2009, the odds ratio of obesity compared to the baseline year (2008) is 0.963. This would indicate that the odds of being classified as obese in 2009 were 3.7% lower than in 2008.
- For Year 2010, the odds ratio of obesity compared to the baseline year (2008) is 1.044. This indicates that the odds of being classified as obese in 2010 were 4.4% higher than in 2008. In other words, individuals in the year 2010 had a slightly increased likelihood of being obese compared to those in the baseline year.

Here are the odds plot of our logistic regression:

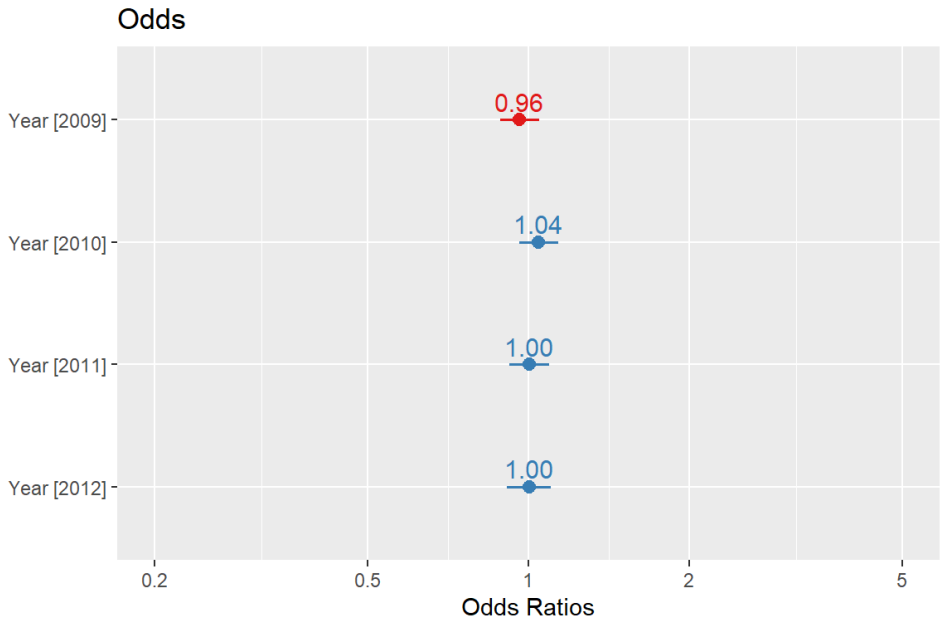


Figure 3: Odds (being obese compared to Year 2008)

We can observe from Figure 3 that:

The plotted values (0.96, 1.04, 1.00, and 1.00) are all very close to 1. This suggests that, compared to the reference year, the odds of being obese in these specific years do not differ dramatically.
Because the odds ratios are near 1, there is likely no strong evidence that obesity prevalence changed significantly across these years. Any differences (e.g., 0.96 vs. 1.04) are quite small and may not be statistically or practically meaningful.

3.2 Differences in Obesity Through The Years

We can attempt to fit a logistic regression model to the data set. With the `stepAIC` function from the `MASS` library, we can perform stepwise model selection to help decide which predictors to include in our GLM.

From this, we see that one potential model fit would be to include the variables Age, Education, and Veg in our GLM. We can also use the `regsubsets` function from the `leaps` package, specifying that evaluations are to be carried out with the Mallow's CP criterion.

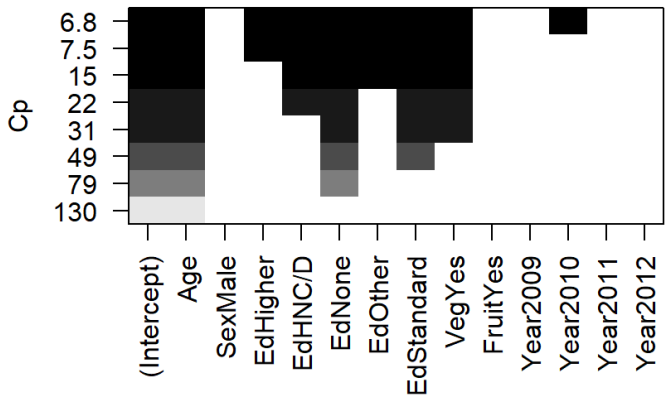


Figure 4: Mallow's CP output

As these two processes are in agreement, we fit the following model to investigate potential differences in obesity levels by age, socio-economic status or lifestyle factors. Both these processes agree that gender does not seem to be a significant predictor, which combined with the findings in our exploratory analysis means we caan feasibly disregard it.

$$y_i \sim \text{Bernoulli}(p_i)$$

$$\log(\frac{p_i}{1-p_i}) = \alpha + \beta_{\text{age}} \cdot x_i + \beta_{\text{veg}} \cdot \mathbb{I}_{\text{veg}} + \beta_{\text{none}} \cdot \mathbb{I}_{\text{none}} + \beta_{\text{standard}} \cdot \mathbb{I}_{\text{standard}} + \beta_{\text{higher}} \cdot \mathbb{I}_{\text{higher}} + \beta_{\text{HNC/D}} \cdot \mathbb{I}_{\text{HNC/D}} + \beta_{\text{other}} \cdot \mathbb{I}_{\text{other}}$$

where

- p_i is the probability of observation i being classified as obese
- α is the intercept term of our model
- β_{age} is the coefficient of our continuous variable, age
- x_i is the age of observation i
- β_j is the additional intercept of an observation that falls into category j
- $\mathbb{I}_j$ is the indicator variable for category j such that

$$\mathbb{I}_j = \begin{cases} 1 & \text{for observations that fall into category j} \\ 0 & \text{otherwise} \end{cases}$$

With our model now fitted, we can obtain our model summary table of the odds estimates of each predictor.

Table 1: Model output of odds, with 95% confidence interval

term	estimate	std.error	statistic	conf.low	conf.high
Intercept	0.211	0.058	−26.718	0.188	0.236
Age	1.012	0.001	12.740	1.010	1.014
Higher or equivalent	1.159	0.047	3.154	1.057	1.270
HNC/D or equivalent	1.272	0.052	4.632	1.148	1.408
No qualifications	1.471	0.042	9.167	1.354	1.597
Other school level	1.245	0.059	3.736	1.109	1.395
Standard grade or equivalent	1.356	0.043	7.158	1.248	1.474

term	estimate	std.error	statistic	conf.low	conf.high
VegYes	0.870	0.034	−4.130	0.814	0.929

From this, we see that our model predicts that a one year increase (all other variables held constant) in age brings 1.012 times the odds of obesity. Similarly, we see that a person who consumes the recommended amount of vegetables per day has 0.87 times the odds of obesity, compared to a person who does not. Finally, we see that those with degrees are predicted to have lower odds of obesity when compared to every other education class.

We note that none of the 95% confidence intervals of our chosen predictors contain 1, which is a positive indicator of the predictors' significance.

We can examine our model's predictive performance by first examining the below plots of predicted obesity probabilities against each explanatory variable, as seen in Figure 5.

Figure 5: Predictive plots of obesity against each explanatory variable

We first note that our findings from the model output table relating to trends in odds / probabilities for each variable are confirmed here. However, we also can note that our model seems hesitant to assign a probability greater than 0.5 to any observation in our data set.

4 Conclusions

The exploratory analysis, through line and bar plots, suggested that while obesity prevalence fluctuated across the years, there was no clear trend. This observation was supported by the logistic regression analysis. The model's odds ratios for subsequent years (2009–2012) were all very close to 1 and none of the differences reached statistical significance. Hence we can conclude that over the period studied there was no evidence of a significant change in obesity prevalence in Scotland. Our model struggled to effectively differentiate between obese and non-obese observations based on the provided predictors. Our exploratory analysis suggested no notable difference in obesity levels between genders, which was confirmed when fitting our model. While our model appears to suggest that age, education level and vegetable consumption are all statistically significant, it is possible that another model with extra explanatory variables could have greater predictive power, which could be worth future analysis. Future studies should examine the weak residuals we found in the model and further examine if there is a link with the binned residuals and distribution of obese to non-obese cases in the study and see if classifiers with a different threshold may help improve the residuals.