Has Obesity Prevalence in Scotland Changed Over Time, and What Factors Influence It?

Jo Scott-Stephen, Eleanor Mills and Sophie Brown

1 Introduction

Obesity rates have been observed to be increasing in many developed countries. These rising rates are concerning as obesity has been linked to numerous health issues, specifically cardiovascular issues, diabetes and cancer. Using data from the Scottish Health Survey, we will analyse obesity trends in the Scottish population from 2013 to 2016 to determine whether Scotland aligns with this pattern. We will further investigate whether any specific socio-economic characteristics and lifestyle factors indicate a higher probability of obesity, aiming to determine if there is a correlation between lifestyle choices and health.

First we will undergo exploratory data analysis in Section 2, followed by a formal analysis in Section 3 and finally we will present our conclusions in Section 4.

2 Exploratory Analysis

We begin by conducting some exploratory data analysis to investigate whether obesity rates have changed over the years.

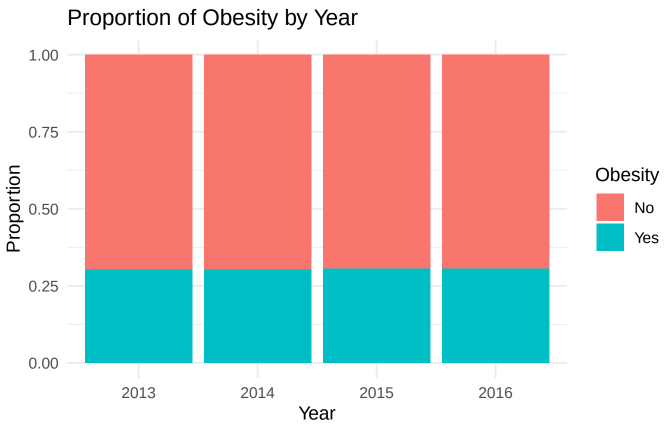

Figure 1: Proportion of Obesity by Year

Figure 1 displays the proportion of obesity by the each year the questionnaire has been taken. There appears to be very little, if any, difference in the change of obesity levels over the years.

Next, we investigate the influence of certain lifestyle factors and socioeconomic status on obesity rates.

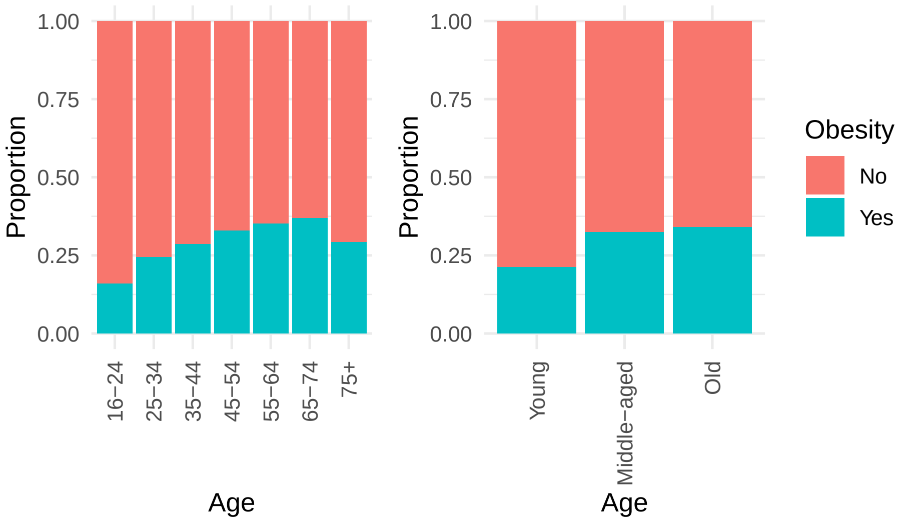

Figure 2: Obesity Proportions against Age (left) and Age Categories (right)

We first look at Figure 2 to see if there is a significant relationship between obesity and age, i.e. can the age of a person change their likelihood of being obese. The plot indicates an increase in obesity rate as age increases, however it then the rate drops in the last category (75+). This decline may be attributed to factors such as weight loss due to health conditions, or survivorship bias, where individuals with obesity-related health risks are less likely to reach older age. The difference in ages is made clearer in the right-hand plot where the age catgegories have been refined, and we see that young people are least likely to be obese. Middle aged are much more likely to be obese but the elderly are the most likely to be obese.

We now investigate whether obesity rates vary by gender using Figure 3. The plot reveals a higher proportion of obese women, suggesting that women may have a greater likelihood of obesity compared to men.

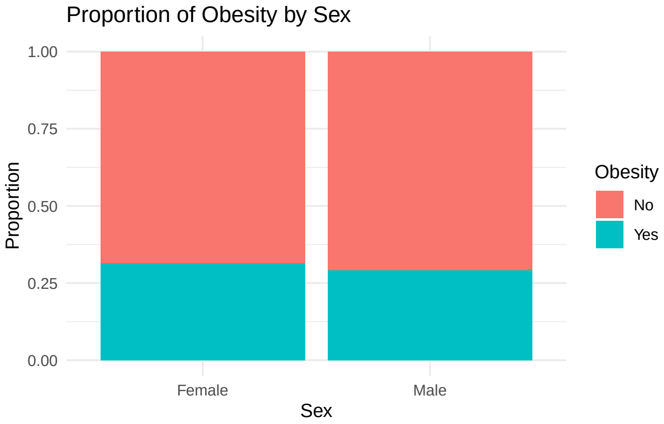

Figure 3: Proportion of Obesity by Gender

Using Figure 4 we investigate the impact of employment status on obesity rates. We have refined employment status into three categories, ‘employed/full time education’, ‘retired’, ‘unemployed/other’. We can see that the least amount of obesity occurs in the employed/full time education category, whilst the unemployed/other category has the highest obesity rates which is interesting as from the previous graphs looking at the impact of age, we would have expected that those who are retired to have the highest obesity rates, however perhaps the slight drop-off that we observed for those over 75 has effected this.

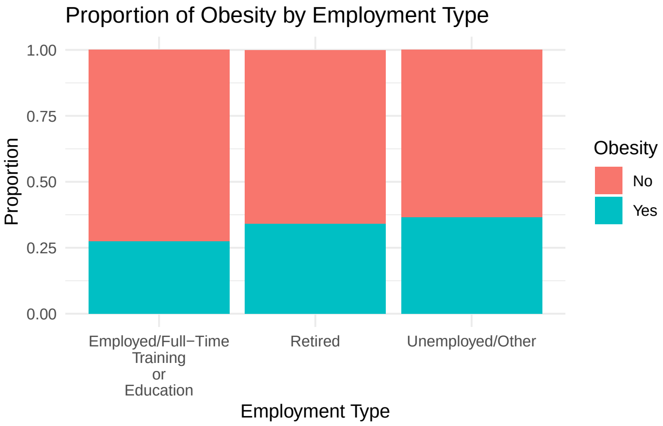

Figure 4: Proportion of Obesity by Employment Type

We then investigate the effect of people's dietary choices. Figure 5 displays a bar plot comparing obesity rates based on whether individuals consumed the recommended daily amount of fruit and vegetables in a day (lifestyle factors). We might have expected this to show there is a significant impact on obesity rates if people ate their '5 a day' however there seems to be no difference in obesity depending on a persons fruit and veg intake.

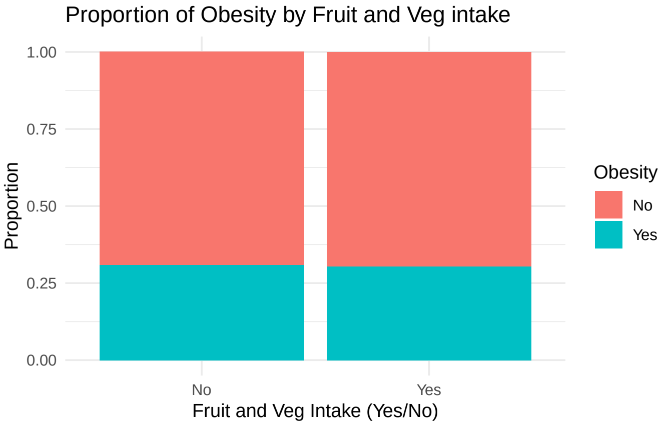

Figure 5: The Proportion of Obesity Against Consumption of the Required Amount of Fruit or Vegetables

3 Formal Analysis

We begin by fitting the full regression model containing all explanatory variables to examine whether the obesity rate has changed over the years. Since our response variable is binary, we fit a logistic regression model without interactions, as our focus is on the individual effects of each covariates. The model can be written as:

$$Obese \sim Bernoulli(p_i)$$

$$logit(p_i) = \alpha + \beta_{\text{sex}} \cdot \mathbb{1}_{\text{sex}} + \beta_{\text{year}} \cdot \text{Year} + \beta_{\text{fv}} \cdot \mathbb{1}_{\text{FV}} + \beta_{\text{middle-aged}} \cdot \mathbb{1}_{\text{middle-aged}} + \beta_{\text{old}} \cdot \mathbb{1}_{\text{old}} + \beta_{retired} \cdot \mathbb{1}_{\text{retired}} + \beta_{unemployed} \cdot \mathbb{1}_{\text{unemployed}}$$

where p_i is the probability of being obese and

- α is the intercept (the baseline log-odds of obesity when all predictors are at their reference category),
- β_{sex} is the coefficient for sex,
- $\mathbb{1}_{\text{sex}}(x)$ is an indicator function such that

$$\mathbb{1}_{\text{sex}}(x) = \begin{cases} 1 & \text{if gender of } x \text{ is male,} \\ 0 & \text{if gender of } x \text{ is female.} \end{cases}$$

- β_{year} is the coefficient for survey year,
- β_{fv} is the coefficient for whether an individual consumes the daily intake of either fruit, vegetables or both,
- $\mathbb{1}_{\text{FV}}(x)$ is an indicator function such that

$$\mathbb{1}_{\text{FV}}(x) = \begin{cases} 1 & \text{if } x \text{ consumes the recommended daily intake of fruit and/or vegetables,} \\ 0 & \text{otherwise.} \end{cases}$$

- $\beta_{\text{middle-aged}}$ is the coefficient for being middle-aged (35-64 years old),
- $\mathbb{1}_{\text{middle-aged}}(x)$ is an indicator function such that

$$\mathbb{1}_{\text{middle-aged}}(x) = \begin{cases} 1 & \text{if } x \text{ is middle-aged (35-64 years old),} \\ 0 & \text{otherwise.} \end{cases}$$

- β_{old} is the coefficient for being old (65+ years),
- $\mathbb{1}_{\text{old}}(x)$ is an indicator function such that

$$\mathbb{1}_{\text{old}}(x) = \begin{cases} 1 & \text{if } x \text{ is old (65+ years old),} \\ 0 & \text{otherwise.} \end{cases}$$

- β_{retired} is the coefficient for being retired,
- $\mathbb{1}_{\text{retired}}(x)$ is an indicator function such that

$$\mathbb{1}_{\text{retired}}(x) = \begin{cases} 1 & \text{if } x \text{ is retired,} \\ 0 & \text{otherwise.} \end{cases}$$

- $\beta_{\text{unemployed}}$ is the coefficient for being unemployed or in another non-working status,
- $\mathbb{1}_{\text{unemployed}}(x)$ is an indicator function such that

$$\mathbb{1}_{\text{unemployed}}(x) = \begin{cases} 1 & \text{if } x \text{ is unemployed or in another non-working status,} \\ 0 & \text{otherwise.} \end{cases}$$

Given the number of covariates and the complexity of our model, we will assess the significance of each covariate in Table 1 to determine if there is potential for refinement.

Table 1: Estimates of the full fitted model's coefficients to 4 d.p.

Term	Estimate	p-value
Intercept		0.9592
Sex: Male		0.0056
Year		0.9289
Eating Recommended Fruit and Veg: Yes		0.2416
Age Category: Middle-aged		0.0000
Age Category: Old		0.0000
Employment Status: Retired		0.1528
Employment Status: Unemployed/Other		0.0000

From Table 1 we see that many of our covariates do not have significant p-values, including the variable 'year'. This means that there is insufficient evidence to suggest that obesity rates have changed over the years. The AIC for this model is 17011.99, which is relatively high, and several covariates have insignificant p-values.

To improve the fit stepwise regression with backward selection will be used to determine whether the full model can be reduced based on the AIC. The model yielding the lowest AIC will be the final model fitted to the data. The regression coefficients from this final model are shown in Table 2. The new model we fit is defined as follows:

$$logit(p_i) = \alpha + \beta_{\text{sex}} \cdot \mathbb{1}_{\text{sex}} + \beta_{\text{middle-aged}} \cdot \mathbb{1}_{\text{middle-aged}} + \beta_{\text{old}} \cdot \mathbb{1}_{\text{old}} + \beta_{retired} \cdot \mathbb{1}_{\text{retired}} + \beta_{unemployed} \cdot \mathbb{1}_{\text{unemployed}}$$

Table 2: Estimates of the fitted model's coefficients to 4 d.p.

Term	Estimate	p-value
Intercept	-1.3402	0.0000
Sex: Male	-0.1014	0.0069
Age Category: Middle-aged	0.5658	0.0000
Age Catergory: Old	0.6249	0.0000
Employment Status: Retired	0.1036	0.1551
Employment Status: Unemployed / Other	0.4006	0.0000

From Table 2 we see that all the covariates except from being retired are statistically significant, indicating sex, age and being unemployed all influence obesity rates. In particular age is seen to be extremely significant. In Scotland being classified as middle-aged causes an increase in 76% increase in odds while being classified as old causes an 87% increase in the odds of being obese compared to young people, holding all else constant.

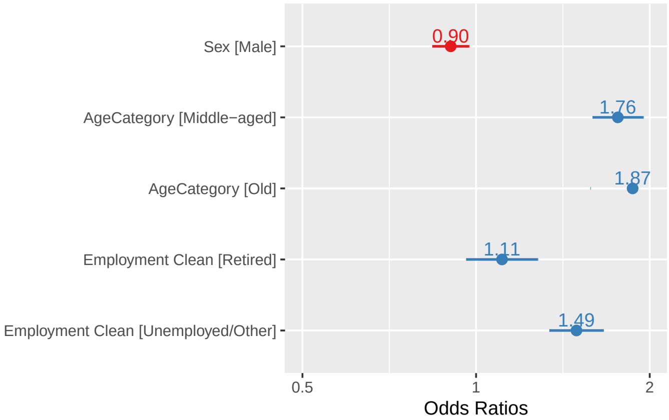

Figure 6: Odds (Obesity)

Furthermore, all else constant, there is a significant 49% increase in the odds of obesity associated with being unemployed. Furthermore, we have a significant 10% decrease in the odds associated with being male rather than female. These findings align with the patterns we observed in Section 2. These results can be visualised in Figure 6, which displays the odds ratios. We note that the figure agrees with our previous conclusion that being retired does not significantly impact the likelihood that someone is obese.

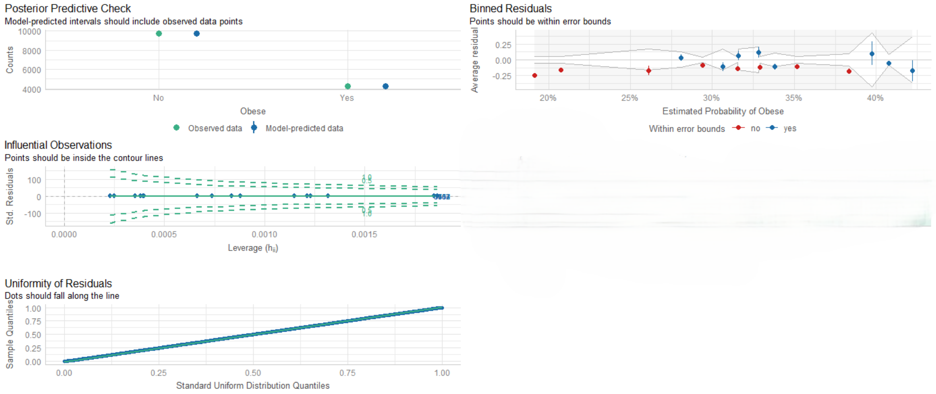

Figure 7: GLM diagnostic plots for obesity in Scotland

Figure 7 allows us to assess if our model assumptions are met. From the diagnostic plots it appears the model's predicted intervals include the observed data points and none of these points are deemed influential. Furthermore, our residuals seem fairly uniform. However, over half of the binned residuals lie outwith the error bounds, indicating an poorly fitted model. This conclusion is further supported by the minimal decrease in the AIC of the reduced model, from 17011 to 17009. Therefore, our previous conclusions and interpretations of the variables may not be appropriate.

4 Conclusions

In conclusion, it appears that overall obesity rates have not changed over the years 2013-16. However, this analysis only looks at the population as a whole, and perhaps certain subgroups of the population dependent on age or socioeconomic status have experienced some sort of change in levels of obesity over these three years, especially as certain characteristics may impact the likelihood that someone is obese. There seems to be some suggestion that being unemployed, aging and being female all increase someone's likelihood of being

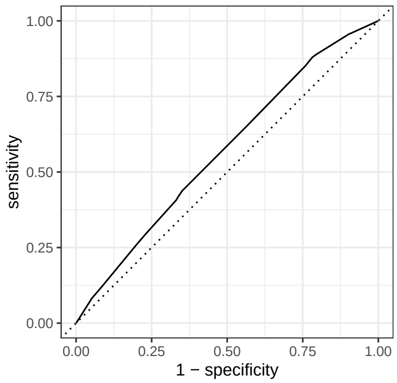

Figure 8: ROC curve for Stepwise Regression Model

To assess how well the model predicts obesity, we examine Figure 8. The ROC curve is fairly close to the diagonal line, indicating our model is only marginally better than random guessing, further suggesting a poor model fit.

obese. Interestingly, it does not seem that eating recommended amounts of fruit and vegetables significantly effect obesity rates. This could be as we looked at the effects of eating recommended amounts either fruit or vegetables, and perhaps these variables should be considered individually, especially as fruit generally has a higher sugar content. Furthermore, being retired does not have a significant effect, which is surprising given the notable similarities in lifestyles between being unemployed and retired. This may be due to the observed drop-off in obesity rates for those aged 75+ possibly due to health conditions. This suggests that age related factors may be influencing obesity rates in a way the fitted model doesn't fully account for. It may be worth considering an interaction term between retirement and age to capture the potential joint effect of these variables on obesity likelihood.

However, as our model assumptions have been violated with less than 95% of binned residuals falling within the error bounds, these conclusions may not be appropriate and should be investigated more thoroughly. The poor performance evident in the ROC curve indicates our model predicts obesity not much better than random chance. Therefore, as our findings are limited due to our model assumptions being violated, another model should be fitted, perhaps with implementing some sort of transformation or a different model altogether.