Kevin Skelly

Examining Predictors of Obesity Using National Health Interview Survey Data

Introduction:

Obesity is a growing public health concern with significant health and economic consequences. Understanding the factors that contribute to obesity is crucial for developing effective prevention and management strategies. This study utilizes data from the National Health Interview Survey (NHIS), a comprehensive national survey in the United States, to examine the predictors of obesity in the adult population, specifically information on age, sex, BMI, and weekly exercise frequency. In this report, I will utilize categorical dependent variable regression to develop a comprehensive predictor for obesity based on the given variables.

Methods:

The dataset used in this report contains five features, with each row representing a single response to the NHIS. Two binary variables, *bmi2* and *female,* represent whether or not the respondent is obese and female, respectively. In both cases, 1 represents the affirmative case, and 0 the negative case. An ordinal variable *bmiC* provides 4 categories for obesity (1 denoting underweight, 2 average weight, 3 overweight, and 4 obese). Finally, two ratio variables, *exfreqwR* and *age*, denote how often the respondent exercises per week, and the respondent's age.

Analysis was performed by first using descriptive statistics to summarize the different variables in the dataset and generate a mental image of the respondents. Descriptive statistics were also used to analyze the distribution of the *bmiC* variable. Following this, I estimated the effect of *female, age*, and *exfreqwR* on BMI using the categorical dependent variable *bmi2*. This was next done using the quadratic term *age^2* to check for significant non-linear effects. I also completed regressions on the *bmiC* feature, first attempting ordinal logistic regression and then

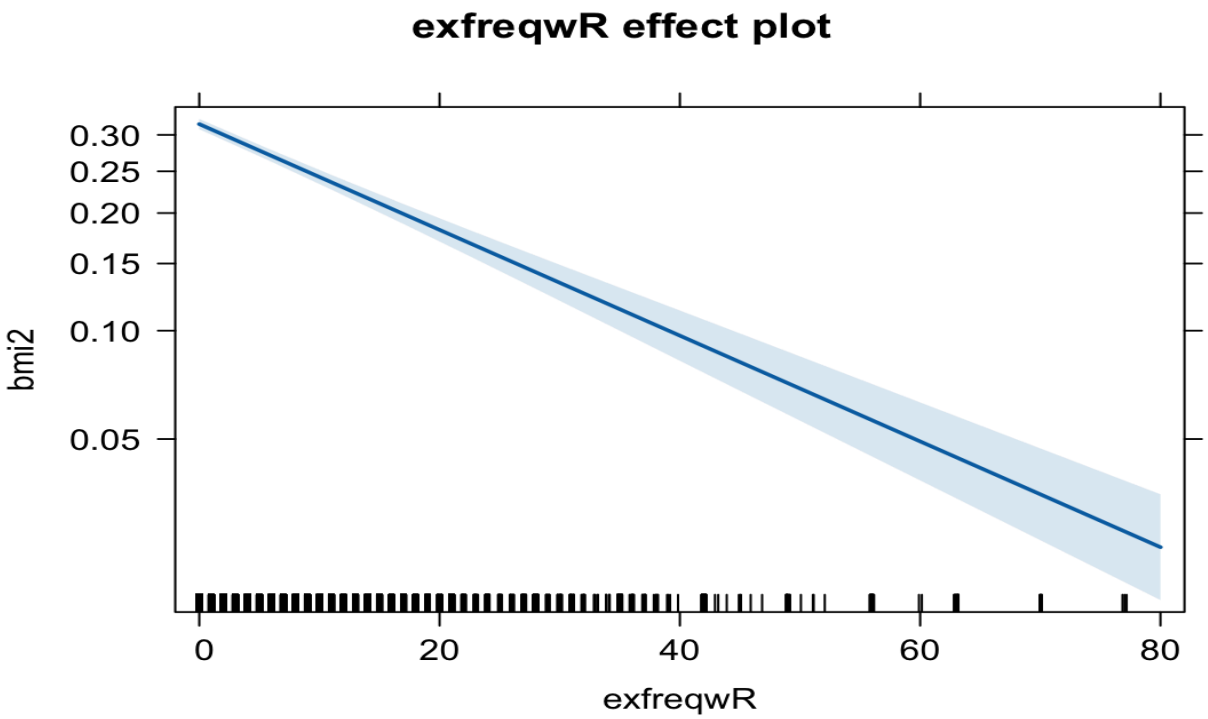
using multinomial logistic regression after assumptions failed. Because of this, I will only be discussing the results of the multinomial logistic regression in this report.

Results:

Descriptive statistics paint the picture of the respondents for 2009 NHIS. 55% of the respondents were female, with a mean age of 47.25(median 46). The standard deviation for *age* is 17.82 years. Regarding exercise frequency, the average respondent exercised 5.30 times per week, with a standard deviation of 7.03 workouts. The mean of 3.0 workouts indicated a right skew, possibly due to one respondent who answered that they completed 84 workouts per week. The distribution of *bmiC* returned a positive skew, indicating that more people are overweight or obese compared to average or underweight.

Initial logistic regression proved somewhat effective in predicting whether or not an individual would be obese (*bmi2 == 1).* The most significant feature proved to be exercise frequency, with an odds ratio of 0.96. This odds ratio indicates a decrease in the chance of obesity by 4% with each additional workout performed weekly. The effect of exercise frequency determined by initial logistic regression is presented in Figure 1. The intercept of 0.3 indicates that at 0 workouts per week, a respondent has a predicted 30 percent chance of being obese.

Figure 1: Plot of the effect of weekly exercise frequency on obesity



To test for hidden effects, logistic regression was next performed with a quadratic effect on age. As seen in Table 1, this change significantly increased the significance of the age variable, increasing its odds ratio from 1.002 to 1.100. To further evidence a weaker original model, the Wald Test returned that Model 2 significantly improves the model fit compared to Model 1.

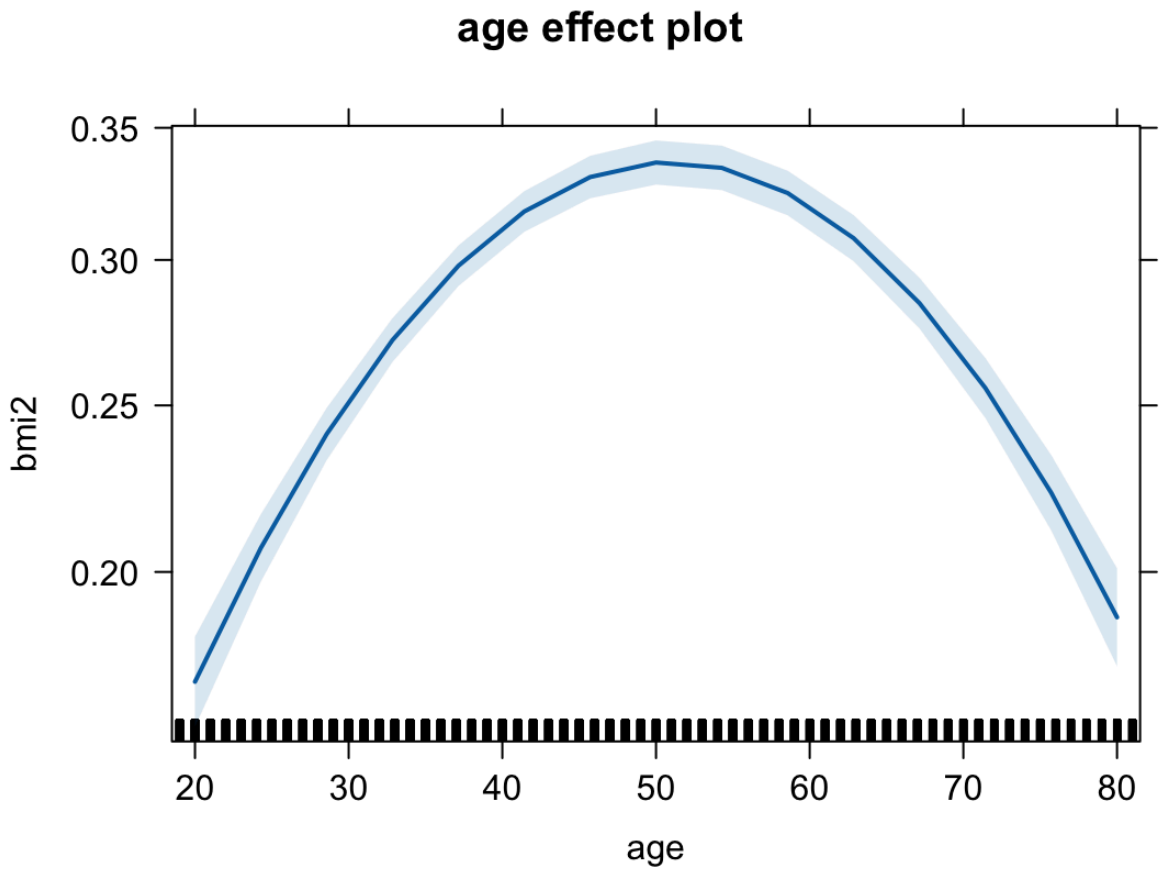
Table 1: Results of regressions on bmi2 using the NHIS_2009 dataset

Dependent Variable: BMI2 (1 if Obese, 0 if not)		
Independent Variable	**Logistic Regression**	**Logistic Regression with Quadratic Term**
Intercept	**0.423***** (0.044)	**0.054***** (0.112)
Female	**1.005** (0.028)	**1.024** (0.028)
Weekly Exercise Frequency	**0.964***** (0.003)	**0.963***** (0.002)
Age	**1.002*** (0.001)	**1.100***** (0.005)

Age^2	-	**0.999***** (0.000)
Summary Statistics		
Adjusted R^2	0.009	0.023
n	25541	25541

Additionally, the age^2 effect indicated a very significant slightly negative effect on obesity, indicating that the odds of being overweight or obese might increase with age initially, but then start to level off or even decrease at older ages. This plot is demonstrated in Figure 2, which seems to show the age where the chances of obesity peak to be around 50 years of age.

Figure 2: Effect of age on obesity as indicated by Model 2



After the parallel lines assumption was violated on ordinal regression for *bmiC,* regression was continued forward with multinomial regression. The results in Table 2 demonstrate the factor with which a one-unit increase in the independent variable will increase the category of the dependent variable. For example, women are 3.431 times more likely to be underweight versus obese compared to men.

Table 2: Odds ratios for body mass index (BMI) using multinomial regression, with reference category of obesity

Dependent Variable: Body Mass Index			
Independent Variable	**1 = underweight versus obesity**	**2 = average weight versus obesity**	**3= overweight versus obesity**
Intercept	6.431*** (.000167)	16.855*** (.000663)	3.560*** (.000578)
female	3.431*** (.000162)	1.403*** (.0148)	.678*** (.0156)
age	.777*** (.00402)	.877*** (.00144)	.953*** (.00141)
Number of times per week a person engages in physical activity	1.0172*** (.00844)	1.0426* (.00267)	1.0355*** (.00266)
Age^2	1.0024*** (.0000555)	1.0012*** (.0000215)	1.0005*** (.0000216)
Summary Statistics			
n	25541	25541	25541

Discussion:

Analysis of the 2009 National Health Interview Survey data revealed several key factors influencing body mass index and the likelihood of obesity. The general profile of respondents reported by descriptive statistics indicated a slightly female majority and an average age of 47.25 years. The distribution of BMI categories indicated a noticeable skew towards overweight and obesity, highlighting a potential national public health concern.

Categorical regression confirmed the expected negative relationship between exercise frequency and obesity rates. Each additional workout performed per week was correlated with a 4% decrease in chances of obesity, in both the quadratic and non-quadratic logistic regression. This insight serves to emphasize the importance of working out to maintain physical health.

The inclusion of the quadratic term in logistic regression revealed a more nuanced relationship with the chances of obesity. While the initial model suggested that there was a slightly positive correlation, the quadratic term indicated a more significant non-linear effect. As age increases, the model suggests that obesity chances will increase until about the age of 50, peaking at about a 30% chance of obesity. After this, obesity chances will begin to decline again in a parabolic fashion.

The multinomial model suggested a high degree of significance in gender, age, age^2, and workout rate in predicting BMI. This all serves to underline previous findings. This regression also indicates a more noticeable relationship between sex and obesity concerns. Further investigation into this relationship would be warranted, if the study were to continue.

In conclusion, the study found the need for specialized and targeted interventions for obesity based on age and gender, as well as a need for increased physical activity in American civilians to prevent obesity. In discussing limitations, it remains important to note that the responses were self-reported, leaving the door open for inconsistencies such as the respondent who reported 84 workouts per week.