

Estimating Obesity Levels Based on Eating Habits and Physical Condition Using Ordinal Logistic Regression

Akhil Ghosh, Rachel Gordon

Introduction

Obesity levels are something many people have sought to explain and address in efforts to improve society's understanding of health and promote a healthy lifestyle. Obesity levels are often defined in terms of the body mass index, or BMI, which is calculated by dividing a person's weight by the square of their height (Formula 1). This calculation takes into consideration the fact that someone who is taller would naturally weigh more and provides a method for calculating whether or not someone is overweight or not by taking their height into consideration as well.

Formula for body mass index (BMI):

$$BMI = \frac{weight}{height^2}$$

Formula 1

Many factors can influence a person's BMI including genetics, body type, eating habits, and other aspects of their lifestyle. According to the CDC, the number of people who have a BMI level of 30 or greater, the threshold for being considered obese, has risen greatly since the 1970s. As a result, an exploration of lifestyle factors may help shed some light onto why this trend has occurred, and provide insight on how to alleviate this issue in the future. The main factors that many people claim affect people's weight include how many calories they consume on a daily basis, the quality of the foods they eat (such as fast food versus a lot of fruits and vegetables), and the amount of exercise they get regularly. There are an endless number of diets and workout regimens available that claim to help people lose weight by changing these aspects of their lifestyle.

This paper seeks to address these claims by modeling the relationship between these lifestyle factors and a person's BMI. A publicly available dataset of these lifestyle factors, height, weight, and BMI was used to address this question and the dataset is described in detail in the next section. Ordinal logistic regression is used to model the relationship between BMI and several lifestyle factors. This method is utilized to determine which predictors may be significant in determining a person's BMI level and whether or not these predictors can accurately predict BMI. Therefore, this paper seeks to address the following research question: *Excluding height and weight, can individuals' eating habits, physical condition, and family history be utilized to accurately predict obesity levels?*

Dataset

The dataset was collected by individuals at The Coast University in Colombia and consists of 2111 observations collected through surveys of undergraduate students at

Table 1: BMI classification according to WHO and Mexican normativity (DO, 2010)

| BMI Classification | |
|--------------------|----------------|
| Underweight | Less than 18.5 |
| Normal | 18.5 to 24.9 |
| Overweight | 25.0 to 29.9 |
| Obesity I | 30.0 to 34.9 |
| Obesity II | 35.0 to 39.9 |
| Obesity III | Higher than 40 |

universities in Mexico, Colombia, and Peru (Palechor 2019). The data includes information regarding the students' weight level (insufficient, normal, overweight level I, overweight level II, obese type I, obese type II, or obese type III), which is determined using BMI. In addition, the dataset includes the students' height, weight, gender, age, family history of being overweight, consumption of high caloric food (FAVC), consumption of vegetables (FCVC), number of main meals (NCP), consumption of food between meals (CAEC), consumption of water (CH20), consumption of alcohol (CALC), overall calorie consumption (SCC), physical activity frequency (FAF), time using technology devices (TUE), transportation used (MTRANS), and smoking habits. Therefore, this dataset may be helpful in determining how eating habits, physical condition, and family history affect obesity levels.

Methods

Data Processing. The first step in processing and cleaning the dataset was to check for missing values, although no data appeared to be missing. Next, the obesity level response variable was converted from seven levels to four such that the different levels of overweight and obese were combined. Initially, the new levels included: underweight, normal weight, overweight, and obese. Following troubles with model building, these levels were further reduced to just normal, overweight, and obese levels. Following the data cleaning, the dataset was randomly split into a training and testing set for model building in which 70% of the data was used for the training set and 30% of the data was used for the test set.

Finally, some initial plots to visualize the relationships present in the dataset were created. Figure 1 shows boxplots of age for each BMI level that are stratified by gender and transportation method. This plot shows that the median age for people who are of normal weight or underweight is lower than the

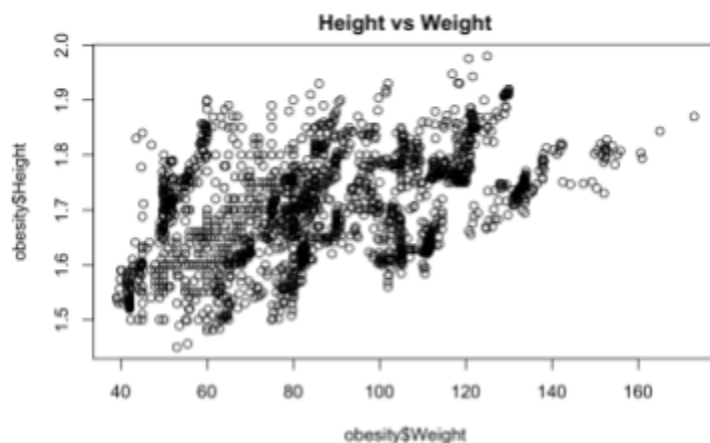


Figure 2

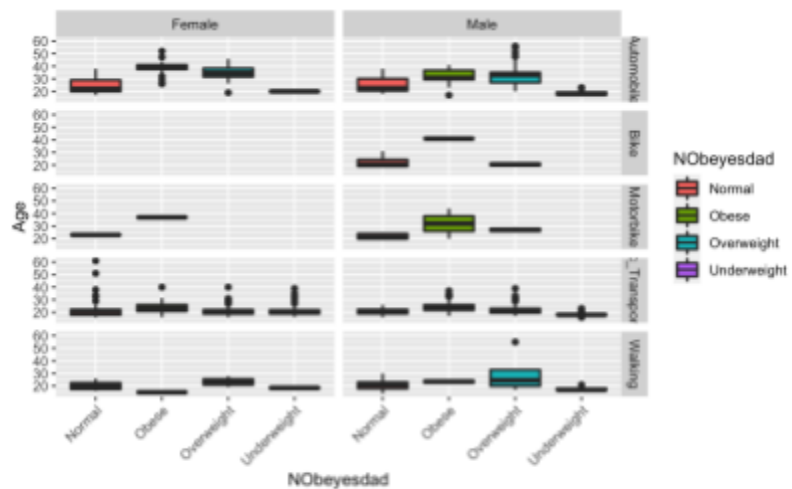


Figure 1: Box Plots of Age by Gender, Transportation, and BMI Level

median age for people who are overweight or obese. Additionally, there do not appear to be any females who use a bike as their main form of transportation in the dataset and there appears to be a greater range of ages for males as opposed to females who are either overweight or obese.

Figure 2 shows a scatterplot of height vs weight, demonstrating that height tends to increase as weight tends

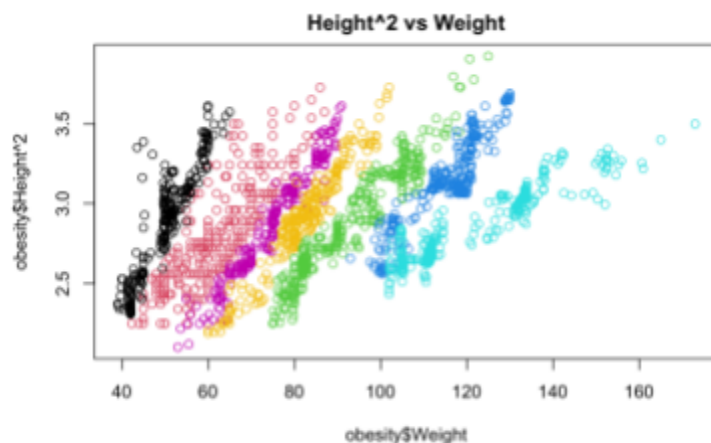


Figure 3

to increase, which explains the use of BMI to consider obesity levels by taking a person's height into account along with a weight. Figure 3 shows the relationship of a person's weight and the square of their height according to BMI. A clear grouping of BMI levels can be seen due to the fact that it is calculated by dividing weight by height-squared. The relationship shown in this plot appears very similar to the one shown in Figure 2, although the scale of the y-axis is different.

Modeling. Due to the ordinal nature of the chosen response levels, with weight increasing from one level to the next, ordinal logistic regression was used to attempt to predict obesity

levels. The "polr" function was used to create a full model using the training data and all available predictors in the dataset except for height and weight, and the resulting accuracy was about 46.85%. Height and weight were excluded from the model due to the fact that the response variable, BMI is calculated using height and weight. Therefore, if height and weight were included in the model, the accuracy may increase significantly and the true importance of the other factors in predicting BMI would be difficult to determine.

| | predictobesity | | | |
|-------------|----------------|-------|------------|-------------|
| | Normal | Obese | Overweight | Underweight |
| Normal | 9 | 40 | 30 | 11 |
| Obese | 2 | 290 | 6 | 0 |
| Overweight | 5 | 135 | 27 | 2 |
| Underweight | 0 | 12 | 54 | 11 |
| [1] | 0.4684543 | | | |

Figure 4

After creating the full model, step AIC was used for model selection to select the best predictors to include in the model. Based on this process, gender, consumption of vegetables (FCVC), number of main meals (NCP), consumption of water (CH20), and consumption of alcohol (CALC) were removed from the model.

Next, the p-values of the remaining predictors were compared and a significance level of 0.05 was used to remove insignificant predictors. As a result, smoking, physical activity frequency (FAF), and overall calorie consumption (SCC) were removed as they all had p-values that were greater than 0.05. However, after this selection process was complete, the model accuracy still did not really improve and the accuracy remained

| Residual Deviance: 2138.487 | | | | |
|-----------------------------------|------------|------------|------------|--------------|
| AIC: 2176.487 | | | | |
| | Value | Std. Error | t value | p-value |
| GenderMale | 0.2218397 | 0.12738646 | 1.7351903 | 8.270711e-02 |
| Age | 0.1462868 | 0.01338787 | 10.9268137 | 0.000000e+00 |
| family_history_with_overweightyes | 2.5689914 | 0.19063232 | 13.4761587 | 0.000000e+00 |
| FAVCyes | 1.0888327 | 0.20097762 | 5.3778759 | 7.536974e-08 |
| FCVC | 0.7278671 | 0.12229410 | 5.9517755 | 2.652490e-09 |
| NCP | -0.1225180 | 0.07686420 | -1.5939542 | 1.109463e-01 |
| CAECFrequently | -1.4793131 | 0.48229116 | -3.0672615 | 2.168298e-03 |
| CAECno | 1.7400106 | 0.53722343 | 3.2388956 | 1.199935e-03 |
| CAECSometimes | 1.9303404 | 0.42700167 | 4.5206858 | 6.163962e-06 |
| CH20 | 0.3283245 | 0.10534994 | 3.0405757 | 2.361264e-03 |
| SCCYes | -0.6364884 | 0.30558861 | -2.0828276 | 3.726694e-02 |
| FAF | -0.3654174 | 0.07393831 | -4.9421928 | 7.724875e-07 |
| TUE | -0.1788767 | 0.09906362 | -1.7975999 | 7.224044e-02 |
| MTRANSBike | 0.6555261 | 1.20068133 | 0.5459617 | 5.850922e-01 |
| MTRANSMotorbike | 1.0533194 | 0.88684696 | 1.1877127 | 2.349466e-01 |
| MTRANSPublic_Transportation | 1.8048188 | 0.19268685 | 9.3665985 | 0.000000e+00 |
| MTRANSWalking | 0.1666957 | 0.42988405 | 0.3877690 | 6.981870e-01 |
| Normal Overweight | 9.5638249 | 0.77416774 | 12.3526523 | 0.000000e+00 |
| Overweight Obese | 11.7914347 | 0.80231882 | 14.6966945 | 0.000000e+00 |

Figure 5

around 46.85%, so other methods needed to be explored.

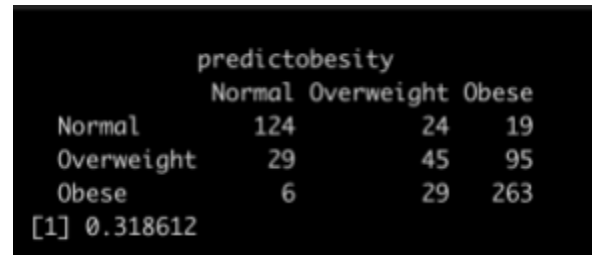
The ordinal logistic regression model described above had the BMI levels in the following order: normal, obese, overweight, underweight. This

ordering may have been throwing off the results as underweight was considered the highest level in the model, although it was the lowest BMI level.

Attempting to refit the model with the BMI in the underweight, normal, overweight and obese levels lead to issues in model building. Therefore, the model was refitted excluding the underweight category and the accuracy improved significantly to about 67.2%.

The same model selection process was then repeated using step AIC and smoking habits and alcohol consumption were removed from the model. The p-values of the remaining predictors were then compared and gender, number of main meals, and time using technology devices were removed from the model because they all had p-values greater than 0.05. The combination of these two model selection techniques improved the accuracy of this new model slightly to about 68.2%.

Finally, penalized multinomial logistic regression for all four BMI levels was implemented using the neural net package. The L2 penalty was used for this model and all of the predictors were included in the full model. This model had significantly greater accuracy than the ordinal logistic regression model that included the underweight level, as the accuracy was about 62.93%. Cross validation was also used to verify the accuracy of this model using five folds and the average of the error rates was about 37.39%, which is very similar to the error rate of 37.07% that was found using the training and testing sets.



| | predictobesity | | |
|--------------|----------------|------------|-------|
| | Normal | Overweight | Obese |
| Normal | 124 | 24 | 19 |
| Overweight | 29 | 45 | 95 |
| Obese | 6 | 29 | 263 |
| [1] 0.318612 | | | |

Figure 6

Results

The ordinal logistic regression model that was found to perform the well only when three BMI levels were classified: normal, overweight, and obese. It had predictors age, family history of being overweight, consumption of high caloric food (FAVC), consumption of vegetables (FCVC), consumption of food between meals (CAEC), consumption of water (CH20), overall calorie consumption (SCC), physical activity frequency (FAF), and transportation used (MTRANS) and an overall accuracy of 67.67%. These results seem to indicate that the insufficient weight class shares similarities to the overweight and obese classes, as the model was unable to perform well when the four classes were separated. It was only until the insufficient bmi label was removed that the ordinal logistic regression was able to perform well, indicating that these classes are more alike than different.

However, the best model for all four BMI levels was found to be the penalized multinomial logistic regression model with an L2 penalty, as this model had the highest accuracy of 62.93% while including all four BMI levels. Furthermore, when building the model with only the three classes as done for the ordinal logistic regression, the model's performance was again improved to an accuracy of 69.25%. Overall, this model seemed to perform much better than the ordinal logistic regression in all facets. This model was then compared to the same model but with height and weight included as predictors and the accuracy drastically increased to 97.48%. This makes sense due to the nature of the response variable and the fact that it was originally derived from height and weight, although these predictors will not be included in the final model because they distract from answering the research question, which is interested in other lifestyle factors. Having the multinomial logistic regression perform

better than the ordinal logistic regression was also interesting, as it points to the conclusion that the underweight class having much more in common with the overweight and obese class than the normal class.

Overall, it seems like while much of the variance could be explained through the models built, when it comes to accurately predicting BMI classes the results indicate that these other lifestyle factors are not nearly as predictive as they are claimed to be. In addition, the need to remove the underweight class in order to have the model function properly indicates that the lifestyle factors between people who lie on the underweight BMI class share many of the same properties as the people on the overweight and obese classes. This is interesting because it opens the discussion on how despite these classes being on the opposite ends of the spectrum, they share similarities on lifestyle indicators that lead to being outside the normal class labels. Future work could explore this further, to see how the health markers of these two classes compare, and if there are any underlying reasons for why these similarities exist despite being on opposite ends of the scale.

References

- Palechor, Fabio Mendoza, and Alexis de Manotas. "Dataset for Estimation of Obesity Levels Based on Eating Habits and Physical Condition in Individuals from Colombia, Peru and Mexico." *Data in Brief*, vol. 25, 2019, p. 104344., <https://doi.org/10.1016/j.dib.2019.104344>.
- "About Adult BMI." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 3 June 2022, www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html.