Posible Title*

Amos Okutse[1] Monica Colon-Vargas[1]
[1] School of Public Health, Brown University, Providence, RI.

Abstract

Introduction: hi
Objective: hi **Methods:** hi
Results: hi
Conclusion: hi
Keywords: *Obesity,* Pandemic, Health risk, Lower socioeconomic groups.*

Introduction

The World Health Organization (WHO) defines obesity as a chronic and intricate disease characterized by an accumulation of excess fat that poses significant health risks. This condition is not confined to isolated cases but has become alarmingly prevalent across numerous countries worldwide, warranting recognition as a global pandemic (CITE PAPER1). Among its many complications, obesity significantly elevates the risk of various health issues, including diabetes, fatty liver disease, hypertension, cardiovascular events like heart attacks and strokes, cognitive decline, joint ailments such as osteoarthritis, disrupted sleep patterns due to conditions like obstructive sleep apnea, and an increased susceptibility to certain types of cancer (CITE PAPER1).

Literature suggests that the likelihood of obesity is influenced by a range of factors beyond individual characteristics, including demographic attributes, community infrastructure, socioeconomic conditions, and specific environmental factors within communities (CITE PAPER 2). In certain countries, particularly among lower socioeconomic groups, obesity rates have surged dramatically due to urbanization, shifts in diet and food availability, and reduced physical activity. This rise in obesity is linked to a significant increase in mortality from chronic diseases like type 2 diabetes, heart disease, and certain cancers, potentially shortening life expectancy by up to 20 years. Given the preventable nature of obesity and its associated health risks, early detection is crucial to mitigate the development of serious conditions such as cardiovascular issues, diabetes, and asthma. Obesity's complex origins involve various factors including socioeconomic status, occupation, and lifestyle habits like smoking and physical activity levels. Physical activity and eating habits are key in preventing obesity, as it primarily stems from an imbalance between calories consumed and expended. Weight loss typically involves reducing calorie intake, increasing energy expenditure, or both. When individuals consume more energy than needed, the excess is stored as fat, leading to obesity. Therefore, maintaining a healthy weight relies on a balanced diet and regular physical activity. (CITE PAPER3)

This work is centered in identifying determinants associated with obesity, with particular emphasis on exploring the interplay between socioeconomic indicators and lifestyle behaviors. In Latin American obesity rates have reached alarming levels, posing serious health risks and placing a substantial burden on healthcare systems CITE(paper4). Therefore this work will address the global health issue of obesity, with a specific focus on

*Replication files are available on the author's Github account (https://github.com/MCV20/GLM-project.git). **Current version**: April 23, 2024

Table 1: Variable Description

Variable	Description	Values
Nobeyesdad	Obesity Level	Insufficient/Normal ObesityI/ObesityII ObesityIII/OverweightI/OverweightII
Gender	What is your gender?	Female/Male
Age	What is your age?	Numeric Value
Height	What is your height?	Numeric Value (m)
Weight	What is your weight?	Numeric Value (kg)
family_history_with_overweight	Has a family member suffered or suffers from overweight?	Yes/No
FAVC	Do you eat high caloric food frequently?	Yes/No
FCVC	Do you usually eat vegetables in your meals?	Never/Sometimes/Always
NCP	How many main meals do you have daily?	Numeric Value
CAEC	Do you eat any food between meals?	No/Sometimes/Frequently Always
Smoker	Do you smoke	Yes/No
CH2O	How much water do you drink daily	Numeric Value
SCC	Do you monitor the calories you eat daily	Yes/No
FAF	How often do you do physical activity	Numeric Value
TUE	How much time do you use technological devices such as cell phone, videogames, television, computer and others?	Numeric Value
CALC	How often do you drink alcohol	Never/Sometimes/ Frequently/Always
MTRANS	Which transportation do you usually use	Automobile/Motorbike Bike/Public/Walking

the diverse populations of Mexico, Peru, and Colombia. By examining individuals' eating habits and physical condition, this research aims to understand the factors contributing to obesity prevalence in these regions and understand the underlying mechanisms driving the escalating rates of obesity.

Methodology

Data

The dataset, accessible online at the UC Irvine Machine Learning Repository, includes information pertinent to estimating obesity levels among $(n = 498)$ individuals aged 14 to 61 hailing from Mexico, Peru, and Colombia. Collected via a web-based survey administered to anonymous respondents, the dataset comprises 17 attributes and was available online for a duration of 30 days. Data collection involved posing questions as delineated in Table 1, which also details the variables under study and the methodology employed in the data collection process.

Data Pre-processing

Given the relatively modest sample size of our dataset, we anticipate potential challenges in conducting robust analyses. Our outcome variables comprise seven distinct levels, encompassing three tiers of obesity (Type I, II, III), two tiers of overweight (I, II), as well as normal weight and underweight categories. However, certain categories contain a limited number of individuals, necessitating the consolidation of some for better statistical reliability. Accordingly, we introduce a novel class termed "Obese," which consolidates the three levels of obesity, alongside a separate class denoted as "Overweight," housing the two levels thereof, while retaining the remaining categories unchanged. Furthermore, certain categories within the covariates exhibited minimal representation, with some even lacking individuals in certain categories. Consequently, we undertook additional aggregation of these classes. For instance, within the alcohol consumption variable, we combined the "always" and "frequently" classes into a singular category. Similarly, in the transportation variable, we merged "motorbike" and "automobile" into a consolidated class termed "motor vehicle," while amalgamating "walking" and "bike" into a unified category.

Statistical Modeling

Since the outcome variable exhibits an ordinal nature, where the categories can be ordered as underweight < normal < overweight < obese, we opted for a modeling approach suited to such data characteristics. Specifically, we employed the generalized linear model framework, fitting a proportional odds model. This modeling technique allows us to account for the ordinal nature of the outcome variable and the inherent ordering of its categories. By utilizing the proportional odds model, we can assess the relationship between the predictors and the ordinal outcome variable while accommodating the cumulative nature of the categories.

Variable Selection

In light of the considerable number of variables present within the dataset, an approach to variable selection opted. To this end, we employed two distinct methodologies for variable selection. Primarily, variables were selected based on their significance as determined by p-values derived from the proportional odds model. Additionally, we conducted variable selection utilizing a proportional odds model with lasso penalization, a technique that inherently incorporates variable selection by penalizing less influential predictors. Subsequently, interactions among selected variables were explored, and the resultant models were meticulously compared to discern the most parsimonious and informative model structure.

Class Imbalance

Given the class imbalance within the outcome variable, we address this issue by employing an upsampling technique. The goal is to decrease the disparity by augmenting the instances of minority classes to achieve parity with the majority class, thereby fostering a more balanced representation across all categories. Subsequently, within this augmented dataset, we applied the same two variable selection techniques previously described. This comprehensive approach ensures that the variable selection procedures are conducted on a dataset that reflects a more equitable distribution of observations across the various outcome categories, thereby mitigating potential biases and bolstering the reliability of the ensuing models. We additionally proceed to fit an additional model using the augmented dataset,incorporating solely those covariates that demonstrated statistical significance within the original dataset.

Model Comparisons

The majority of models analyzed in this study were nested, with assessments primarily relying on information criteria such as the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). Notably, the BIC

Table 2: Summary Statistics Grouped by Outcome Category

Characteristic	**Overall**, N = 498	**Normal_Weight**, N = 287	**Obese**, N = 61	**Overweight**, N = 116	**Underweight**, N = 34
Gender					
Female	227.00 (45.58%)	141.00 (49.13%)	23.00 (37.70%)	46.00 (39.66%)	17.00 (50.00%)
Male	271.00 (54.42%)	146.00 (50.87%)	38.00 (62.30%)	70.00 (60.34%)	17.00 (50.00%)
Age	23.15 (6.72)	21.74 (5.10)	26.46 (8.22)	25.77 (8.39)	20.15 (4.03)
Height	1.69 (0.10)	1.68 (0.09)	1.71 (0.10)	1.69 (0.10)	1.70 (0.11)
Family history overweight	300.00 (60.24%)	155.00 (54.01%)	53.00 (86.89%)	76.00 (65.52%)	16.00 (47.06%)
High-calorie food consumption	348.00 (69.88%)	208.00 (72.47%)	44.00 (72.13%)	75.00 (64.66%)	21.00 (61.76%)
Number of meals veggie consumption					
1	32.00 (6.43%)	18.00 (6.27%)	4.00 (6.56%)	7.00 (6.03%)	3.00 (8.82%)
2	272.00 (54.62%)	155.00 (54.01%)	36.00 (59.02%)	69.00 (59.48%)	12.00 (35.29%)
3	194.00 (38.96%)	114.00 (39.72%)	21.00 (34.43%)	40.00 (34.48%)	19.00 (55.88%)
Daily main meals					
1	108.00 (21.69%)	52.00 (18.12%)	16.00 (26.23%)	35.00 (30.17%)	5.00 (14.71%)
3	344.00 (69.08%)	206.00 (71.78%)	44.00 (72.13%)	73.00 (62.93%)	21.00 (61.76%)
4	46.00 (9.24%)	29.00 (10.10%)	1.00 (1.64%)	8.00 (6.90%)	8.00 (23.53%)
Eating between meals					
Frequently	189.00 (37.95%)	118.00 (41.11%)	16.00 (26.23%)	37.00 (31.90%)	18.00 (52.94%)
no	20.00 (4.02%)	10.00 (3.48%)	2.00 (3.28%)	5.00 (4.31%)	3.00 (8.82%)
Sometimes	289.00 (58.03%)	159.00 (55.40%)	43.00 (70.49%)	74.00 (63.79%)	13.00 (38.24%)
Smoke	32.00 (6.43%)	13.00 (4.53%)	10.00 (16.39%)	8.00 (6.90%)	1.00 (2.94%)
Water Intake (L)					
1	135.00 (27.11%)	83.00 (28.92%)	17.00 (27.87%)	25.00 (21.55%)	10.00 (29.41%)
2	266.00 (53.41%)	164.00 (57.14%)	25.00 (40.98%)	61.00 (52.59%)	16.00 (47.06%)
3	97.00 (19.48%)	40.00 (13.94%)	19.00 (31.15%)	30.00 (25.86%)	8.00 (23.53%)
Calorie intake monitoring	55.00 (11.04%)	30.00 (10.45%)	3.00 (4.92%)	16.00 (13.79%)	6.00 (17.65%)
Frequency of days of physical activity (wk)					
0	162.00 (32.53%)	80.00 (27.87%)	28.00 (45.90%)	44.00 (37.93%)	10.00 (29.41%)
1	158.00 (31.73%)	97.00 (33.80%)	15.00 (24.59%)	40.00 (34.48%)	6.00 (17.65%)
2	113.00 (22.69%)	69.00 (24.04%)	12.00 (19.67%)	18.00 (15.52%)	14.00 (41.18%)
3	65.00 (13.05%)	41.00 (14.29%)	6.00 (9.84%)	14.00 (12.07%)	4.00 (11.76%)
Technology use time					
0	243.00 (48.80%)	129.00 (44.95%)	33.00 (54.10%)	68.00 (58.62%)	13.00 (38.24%)
1	181.00 (36.35%)	122.00 (42.51%)	16.00 (26.23%)	30.00 (25.86%)	13.00 (38.24%)
2	74.00 (14.86%)	36.00 (12.54%)	12.00 (19.67%)	18.00 (15.52%)	8.00 (23.53%)
Frequency of alchohol intake					
Frequently	46.00 (9.24%)	19.00 (6.62%)	9.00 (14.75%)	17.00 (14.66%)	1.00 (2.94%)
no	179.00 (35.94%)	107.00 (37.28%)	22.00 (36.07%)	36.00 (31.03%)	14.00 (41.18%)
Sometimes	273.00 (54.82%)	161.00 (56.10%)	30.00 (49.18%)	63.00 (54.31%)	19.00 (55.88%)
Transportation					
Motor_Vehicle	110.00 (22.09%)	51.00 (17.77%)	22.00 (36.07%)	34.00 (29.31%)	3.00 (8.82%)
Public_Transportation	326.00 (65.46%)	200.00 (69.69%)	35.00 (57.38%)	66.00 (56.90%)	25.00 (73.53%)
Walking/Bike	62.00 (12.45%)	36.00 (12.54%)	4.00 (6.56%)	16.00 (13.79%)	6.00 (17.65%)

[1] n (%); Mean (SD)

was preferred over the AIC due to its heightened penalty with larger sample sizes, leading to more stringent significance thresholds CITE(BIC). Model comparisons were conducted using the likelihood ratio test (LRT), aimed at examining competing models by testing the null hypothesis that the simpler and more complex models are equally effective. Specifically, this test assesses whether the additional parameters in the larger model significantly improve fit, implying that their effect sizes are statistically indistinguishable from zero.

Results

Descriptive Statistics

Model Results

Model Diagnostics

Discussion

Code Appendix

```r
knitr::opts_chunk$set(echo = FALSE,
                      message = FALSE,
                      warning = FALSE,
                      fig.align="center")
## Load libraries
library(naniar)
library(dplyr)
library(ggplot2)
library(corrplot)
library(caret)
library(tidymodels)
library(naivebayes)
library(gtsummary)
library(nnet)
library(coefplot)
library(ggpubr)
library(pROC)
library(MASS)
library(stargazer)
library(brant)
library(DescTools)
library(ordinalNet)
library(pomcheckr)
library(ROSE)
library(kableExtra)
theme_set(theme_minimal())
## Load data
dat <- read.csv("ObesityDataSet_raw_and_data_sinthetic.csv")
dat <- dat[1:498,] #take original data before SMOTE
## Data Preprocessing (covariates)
dat$CALC[dat$CALC == "Always"] <- "Frequently"
dat$CAEC[dat$CAEC == "Always"] <- "Frequently"
dat$MTRANS[dat$MTRANS == "Motorbike"] <- "Motor_Vehicle"
dat$MTRANS[dat$MTRANS == "Automobile"] <- "Motor_Vehicle"
dat$MTRANS[dat$MTRANS == "Bike"] <- "Walking/Bike"
dat$MTRANS[dat$MTRANS == "Walking"] <- "Walking/Bike"
```

```r
dat$NObeyesdad <- ifelse(dat$NObeyesdad %in% c("Obesity_Type_I",
 ↪  "Obesity_Type_II", "Obesity_Type_III"), "Obese", dat$NObeyesdad)
dat$NObeyesdad <- ifelse(dat$NObeyesdad %in% c("Overweight_Level_I",
 ↪  "Overweight_Level_II"), "Overweight", dat$NObeyesdad)
dat$NObeyesdad[dat$NObeyesdad == "Insufficient_Weight"] <- "Underweight"

## Factor variables
dat <- dat %>%
  mutate_if(is.character, as.factor)

dat <- dat %>% mutate(
  FCVC = factor(FCVC),
  NCP = factor(NCP),
  CH2O = factor(CH2O),
  FAF = factor(FAF),
  TUE = factor(TUE)
)

## Remove weight variable and rename fam history
dat <- dat %>% dplyr::select(-Weight)
dat <- dat %>% rename(fam_hist = family_history_with_overweight)

## Summary Statistics
dat %>% tbl_summary(digits = list(everything() ~ c(2)),
                    statistic = list(all_continuous() ~ "{mean} ({sd})"),
                    by = NObeyesdad,
                    missing = "no",
                    label = list(
                      fam_hist ~"Family history overweight",
                      FAVC ~ "High-calorie food consumption",
                      FCVC ~ "Number of meals veggie consumption",
                      NCP ~ "Daily main meals",
                      CAEC ~ "Eating between meals",
                      SMOKE ~ "Smoke",
                      CH2O ~ "Water Intake (L)",
                      SCC ~ "Calorie intake monitoring",
                      FAF ~ "Frequency of days of physical activity (wk)",
                      TUE ~ "Technology use time",
                      CALC ~ "Frequency of alchohol intake",
                      MTRANS ~ "Transportation"))%>%
  add_overall() %>%
  bold_labels() %>%
  as_kable_extra(booktabs = TRUE, caption = "Summary Statistics Grouped by
   ↪  Outcome Category") %>%
  kableExtra::kable_styling(latex_options = "scale_down")
```