Adolescent Obesity using Logistic Regression

Prepared by: Emilie Campos

5/11/2018

Introduction

Obesity is a rising problem in America's youth and with multiple diseases and conditions stemming from weight gain, a serious effort has been made to slow the obesity rate [1]. The National Health and Nutrition Examination Survey (NHANES), performed by the National Center for Health Statistics (NCHS), aims to assess the health and nutritional status of adults and children in the United States [2]. The survey is unique in that it combines interviews with physical examinations. Our goal is to predict adolescent obesity using sex, age, race, family income, education, and physical activity. We include predictors for socioeconomic status and demoographic information because, as research has shown, obesity disproportionately affects different communities [1]. We aim to investigate the association between these predictors and obesity using logistic regression models.

Methods

Sample

The NHANES survey was done in two parts: a home interview and a health examination. The home interview included demographic, socioeconmoic, dietary, and health-related questions, while the physical health examination and laboratory tests performed medical, dental, and physiological tests. The entire suvery examines a nationally representative sample of approximately 5,000 persons each year from across the country [2]. After removing many missing values of predictors that pertained to our question, we were left with 570 adolescents, aged 12-15. The characteristics of this subsample are reported in Table 2 and the descriptions of each predictor, namely the suvery questions, are given in Table 1.

Table 1: Variable Descriptions

Variable	Question
Obese	Calculated from BMI: 0 = BMI < 30.0, 1 = BMI >= 30.0
Gender	Gender of the participant.
Age	Age in years of the participant at the time of screening.
Race	Recode of reported race and Hispanic origin information, with Non-Hispanic Asian Category
Family Income	Total family income (reported as a range value in dollars)
Time spent sitting	Do not include time spent sleeping. How much time {do you/does SP} usually spend sitting on a typical day?
Days physically active for at least 60 minutes (past 7 days)	During the past 7 days, on how many days {were you/was SP} physically active for a total of at least 60 minutes per day? Add up all the time {you/he/she} spent in any kind of physical activity that increased {your/his/her} heart rate and made {you/him/her} breathe hard some of the time.
Hours of TV or videos per day on average (past 30 days)	Now I will ask you first about TV watching and then about computer use. Over the past 30 days, on average how many hours per day did {you/SP} sit and watch TV or videos?
Hours of computer use per day on average (past 30 days)	Over the past 30 days, on average how many hours per day did {you/SP} use a computer or play computer games outside of school? Include Playstation, Nintendo DS, or other portable video games.

Statistical analysis

Initially the data was comprised of 9,255 persons. However, in order to use the predictors in the proposed research question, it was necessary to remove the observations where many of the variables were set to either missing because a person refused to answer, did not know the answer, or any other reason. Additionally, some persons had their family income coded simply as "Less than $20,000" or "$20,000 or Over". Since this did not agree with the other categories of family income, we removed these observations as well. This left us with a data set with 570 observations.

In order to have model parsimony, we needed to collapse the categories of many of the variables. The variable `income` became values from $0-$24,999 became the poverty category, $25,000-$74,999 for middle, and $75,000 and over for high income [3]. We also combined "Mexican" and "Other Hispanic". After viewing the histograms for hours of sedentary activity, days phyiscally active, screen time, and computer use, we used reasonable values near the median as cutoff values for high and low.

Each of the predictors were considered categorical, thus each of the pairwise Chi-Square tests of independence were performed. The p-values from these tests are supplied in Table 4. From these results, we see there is evidence of dependencies between: gender and race, gender and days physically active, gender and computer use, gender and sedentary activity, race and family income, race and days physically active, sedentary activity and hours of screen time, and screen time and computer time. This causes us to believe there is a possible route for variable selection in order to make better predicitons.

In addtion to checking for possible dependence between predictors, we fit logistic models that regressed the outcome, obesity, on each predictor singly. From these individual models, we see that the null deviances provided in Table 5, which are distributed $\chi^2(569)$, indicate the data is consistent with the model in all of the simple logistic regressions. We then fit a logistic model that regressed the outcome, obesity, on all of the untransformed predictors. Without continuous variables, there was no need to assess linearity in the log odds. Using residual analysis, we found that all of the studentized residuals were within 3 standard deviations. Also, the confidence interval displacement statistic indicated there were no influential observations.

Results

The multiple logistic regression model takes the form:

$$\begin{aligned} log\left(\frac{\pi}{1-\pi}\right) = &- 2.169 - 0.078 * female - 0.194 * age_{13} + 0.451 * age_{14} \\ &+ 0.784 * age_{15} - 0.369 * White + 0.699 * Black - 1.669 * Asian \\ &+ 0.292 * Multi - Racial - 0.449 * (MiddleIncome) - 1.063 * (HighIncome) \\ &+ 0.203 * (MediumHours) + 0.846 * (HighHours) - 0.086 * (HighDays) \\ &+ 0.021 * (HighScreenTime) + 0.132 * (HighComputerUse) \end{aligned}$$

where π represents the probability of an adolescent being obese. With all of the categories of all of the predictors, the resulting models has 16 coefficients. The unadjusted odds ratios are supplied in Table 3. We interpret these values as the odds ratio between the category and reference category. For example, the value 2.139 for age 15 indicates 15 year olds are approximately 2 times more likely to be obese than 12 year olds. We note that many of the odds ratios in Table 3 are insignificant, implying the odds ratio between the category and the reference group does not differ from 1 significantly. This is also supported by their confidence intervals, which contain 1. The parameters that are significant at the $\alpha = 0.05$ level in the full model are age 15, Non-Hispanic Black, and High Income. High income is protective in that adolescents from families with high income are 65% less likely to be obese as adolescents from families below the poverty line. Non-Hispanic Black is injurious as they are 2 times as likely to be obese as Hispanic adolescents.

Conclusion

Based on the significance of the odds ratios, reported as p-values in Table 3, when controlling for the other predictors, only age, race, and income played a role in determining the odds of obesity when controlling for the other factors. This is surprising in that, one would think activity level would affect the odds of being obese. However, it is important to note that socioeconomic status can greatly affect the availability of healthy foods. We would want to further investigate the diets of these adolescents to understand the relationship with obesity.

Limitations

Without removing the dependence between variables, many of the odds ratio estimations do not differ from 1 significantly in the multiple logistic regression model. In addition to this, many of the measurements used, namely the survey responses, were most likely inaccurate. This forced us to bin the responses into larger bins and without more precision, it is difficult to know if these factors really are insignificant.

Tables and Figures

Table 2: Sample Characteristics

Variable	Proportion of Adolescents, $n = 570$
Obese (BMI >= 30.0)	
0	87.5
1	12.5
Gender	
Male	53.5
Female	46.5
Age (in years)	
12	23.7
13	25.3
14	26.7
15	24.4
Race	
Hispanic	32.8
Non-Hispanic White	29.5
Non-Hispanic Black	22.3
Non-Hispanic Asian	8.8
Other Race - Including Multi-Racial	6.7
Family Income	
Poverty	21.2
Middle	49.5
High	29.3
Hours of Sedentary Activity	
Low	12.8
Medium	73.7
High	13.5
Days physically active at least 60 minutes	
Low	55.6
High	44.4
Hours watching TV or videos in past 30 days	
Low	62.8
High	37.2
Hours of computer use in past 30 days	
Low	43.3
High	56.7

Table 3: Unadjusted and Adjusted Odds Ratios

	Unadjusted			Adjusted		
Variable	OR	95% CI	p	OR	95% CI	p
Gender						
Male	ref	-	-	ref	-	-
Female	0.822	(0.49, 1.35)	0.445	0.925	(0.53, 1.6)	0.782
Age						
12	ref	-	-	ref	-	-
13	1.017	(0.44, 2.34)	0.968	0.824	(0.35, 1.97)	0.659
14	1.735	(0.84, 3.76)	0.148	1.57	(0.73, 3.5)	0.255
15	2.139	(1.04, 4.61)	0.043	2.189	(1.03, 4.88)	0.047
Race						
Hispanic	ref	-	-	ref	-	-
Non-Hispanic White	0.522	(0.24, 1.06)	0.08	0.691	(0.31, 1.49)	0.355
Non-Hispanic Black	1.921	(1.06, 3.52)	0.033	2.013	(1.08, 3.8)	0.029
Non-Hispanic Asian	0.139	(0.01, 0.68)	0.056	0.188	(0.01, 0.97)	0.111
Other Race - Including Multi-Racial	1.273	(0.44, 3.2)	0.626	1.339	(0.45, 3.53)	0.573
Family Income						
Poverty	ref	-	-	ref	-	-
Middle	0.562	(0.32, 0.99)	0.044	0.638	(0.35, 1.17)	0.14
High	0.245	(0.11, 0.52)	<0.001	0.345	(0.14, 0.79)	0.014
Hours of Sedentary Activity						
Low	ref	-	-	ref	-	-
Medium	1.188	(0.55, 2.98)	0.686	1.225	(0.54, 3.18)	0.65
High	2.671	(1.07, 7.32)	0.042	2.331	(0.88, 6.76)	0.101
Days physically active at least 60 minutes						
Low	ref	-	-	ref	-	-
High	0.793	(0.47, 1.31)	0.37	0.918	(0.53, 1.58)	0.758
Hours watching TV or videos in past 30 days						
Low	ref	-	-	ref	-	-
High	1.192	(0.71, 1.97)	0.497	1.021	(0.58, 1.76)	0.94
Hours of computer use in past 30 days						
Low	ref	-	-	ref	-	-
High	1.052	(0.64, 1.75)	0.844	1.141	(0.66, 2)	0.642

Table 4: Chi-Square GOF p-values

	gender	age	race	income	sedentary	days	screens	computer
gender								
age	0.223							
race	0.037	0.693						
income	0.505	0.522	<0.001					
sedentary	0.187	0.025	0.207	0.03				
days	0.001	0.107	0.008	0.114	0.006			
screens	0.72	0.814	0.513	0.156	<0.001	0.028		
computer	<0.001	0.084	0.567	0.749	0.631	0.131	0.001	
obese	0.523	0.084	<0.001	0.001	0.021	0.442	0.583	0.946

Table 5: Individual Model Deviance

Variable	Deviance
Gender	427.96
Age	421.90
Race	406.72
Family Income	414.38
Hours of Sedentary Activity	421.83
Days physically active at least 60 minutes	427.73
Hours watching TV or videos in part 30 days	428.09
Hours of computer use in past 30 days	428.51

References

1. Obesity Rates & Trends - The State of Obesity [Internet]. [cited 2018 May 9]. Available from: https://stateofobesity.org/rates/

2. NHANES - National Health and Nutrition Examination Survey Homepage [Internet]. [cited 2018 May 2]. Available from: https://www.cdc.gov/nchs/nhanes/index.htm

3. Middle Class Income: Definition, Types, Range [Internet]. [cited 2018 May 10]. Available from: https://www.thebalance.com/definition-of-middle-class-income-4126870