Title: Social Determinants of Adolescent Obesity: An Empirical Study using both Statistical and Supervised Learning Approaches

Author: I-Ming Chiu, Ph.D.
Rutgers University - Camden
Department of Economics
311 N 5th St.
Camden, NJ 08102
Email: ichiu@camden.rutgers.edu
Tel (o): 856-225-6012

First Draft (9/25/2022): Please don't share and cite the paper. This is a project in progress.

ABSTRACT

Obesity in children and adolescents may lead to other health related issues such as cardiovascular diseases in adulthood and results in rising healthcare cost. In this study, a logistic regression model was estimated based on the pooled cross-sectional National Survey on Drug Use and Health (NSDUH) data from 2013 to 2019 to achieve two goals. First, we aimed at uncovering the direction and strength of the association between adolescent obesity and a list of social determinants of health comprising age, gender, ethnicity, household income, insurance type, family structure, parenting style, and school experiences. Secondly, the logistic model was also served as a classifier (i.e., a supervised learning algorithm) to predict adolescents with potential obesity health issue.

The adjusted odds ratios, found from the training data, revealed that male adolescents had higher risk of being obese comparing to female adolescents. Older-age (16-17) group had a higher risk of having obesity than younger-age (14-15 and 12-13) group. Black or Hispanic adolescents had higher risk of having obesity than White adolescents. In contrast, Asian/Pacific Island adolescents had a lower risk of having obesity comparing to White. Living in a single-parent home, having no siblings under 18, less authoritative parenting style, and having bad school experiences were all associated with higher risk of being obese. Most importantly, adolescents growing up in a lower income environment were more likely to develop obesity, a social gradient in health. Additionally, this study found that adolescents with Medicaid/CHIP insurance had a higher risk of having obesity comparing to those without insurance coverage at all.

Depending on the choice of the threshold value (a probabilistic value), the predictive model in the training data was able to achieve a high recall rate at the expense of a lower accuracy rate and vice versa. Our strategy in terms of model performance was to find a balance between recall and accuracy rate; maximizing the recall rate while maintaining a reasonable accuracy rate. We were able to achieve a rate of about 59.90% for both measurements. The model performance using the same metrics (recall and accuracy rate) remained about the same in the test data, which was an indication of no overfitting problem. Results from this study successfully pre-identified adolescents at risk of having obesity. The early identification

of potential obesity cases according to their social determinants may help bring down the rising healthcare cost via early prevention and intervention programs.

JEL Classification Numbers: I14; I18
Keywords: Adolescent Obesity, Logistic Regression Model; National Survey on Drug Use and Health (NSDUH); Supervised Machine Learning; Social Gradient in Health, Recall & Accuracy Rate

INTRODUCTION

Obesity in childhood and adolescence is a serious public health issue. According to CDC, 1 in 5 children and adolescents are affected in the US [1]. Be more specific, the prevalence of obesity was 19.7% and affected about 14.7 million children and adolescents in the US [2]. This high obesity rate deserves our attention since obesity during pre-adult period often leads to obesity in adulthood and results in many baneful consequences. On the micro level, adults with obesity are often associated with higher risk of having high blood pressure, bad cholesterol, diabetes, and sleep apnea, etc. On the macro level, the aggregate obesity related annual healthcare cost in the US was estimated to be about $173 billion in terms of 2019 US dollar [3]. While the causes of obesity can vary, the surrounding environments we live in (or equivalently the social determinants of health) can be an important contributing factor.

Prior studies revealed that social determinants such as age, gender, family composition, and parenting style, etc. may be associated with obesity among adolescents [4, 5, 6]. However, most studies in this area have focused on examining the direction and strength of such associations using traditional inference-based regression models [7, 8, 9], and very few has attempted to build a predictive model using machine learning algorithms to identify adolescents who are at higher risk of obesity. For those who applied machine learning methods to predict adolescent obesity, the contributing factors were mostly genetics related [10, 11] and ignored the possible impacts from social determinants of health.

In this project, there were two goals. First, we examined the associations between adolescent obesity and a group of broadly defined social determinants of health obtained from a survey

data, National Survey on Drug Use and Health (NSDUH), using logistic regression model. Secondly, based on the empirical findings, we used the same model as a classifier to predict and identify adolescents with obesity issues within the training data set and evaluated the model performance using the test data.

METHODS

Data

The NSDUH is an annual cross-sectional survey sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA) of the US Department of Health and Human Services. By using a stratified multistate area probability sampling method, the NSDUH survey provides nationally representative data for the civilian, noninstitutionalized population aged 12 or older from all 50 states and the District of Columbia in the U.S.. The survey is administered in English and Spanish, and interviews are conducted using computer-assisted interviewing. Our sample was retrieved from NSDUH and including adolescents aged 12 to 17 from the year 2013 to 2019.

Measurement

In the NSDUH data, Body Mass Index (BMI) variable is available for the majority of observations. We transformed the BMI into a binary variable of obesity (normal vs. obese) based on the data charts provided by the World Health Organization (WHO) [12]. Accordingly, in the age group 12-13, obesity was identified for girls whose BMI is greater than 26.2 and boys whose BMI is greater than 24.8. In the age group 14-15, these two boundary values are 28.2 for girls and 27 for boys. In the age group 16-17, these two values are 29.8 for girls and 27.9 for boys. We presented the prevalence of adolescent obesity over time in Figure 1. It showed that the annual trend in obesity was a positive one between 2014 and 2018, there was a big rise from 2017 to 2018, and then plateaued in 2019. The prevalence rates were in the range of 15.96% to 18.5% in the sample period.

Social determinant of health variables include adolescents' age, sex, race/ethnicity, type of insurance coverage, household income, family composition, parenting stye, and school experiences. All of the social determinant of health variables are categorical. Their

individual's categories and relative frequency were reported in Table 1. The prevalence of obesity conditioned on income variable was presented in Figure 2. Interestingly, among the four income groups, the prevalence of obesity was quite stable for the "< $20 K" & "$50-75 K" income groups. Both had been steady in the sample period. On the other hand, there were positive trends for the other two income groups (i.e., "$20-50 K" & "> $75 K"). Further analysis is required to find these unique growth patterns in various income groups.

Methods

We briefly address the logistic model and explain how it can be used as a classifier to detect depression in this section. The logistic model can be represented using the following equation:

$$E(Y \mid X) = P(Y = 1 \mid X) = \frac{\exp(X*\beta)}{1+\exp(X*\beta)} \text{ ; exp: exponential function.} \quad (1)$$

Where Y is a binary variable that represents two states, obesity and normal, and can be encoded using numeric values one and zero. X is a column of vector that represents the collection of social determinants, and β is the associated coefficient vector. The conditional mean of Y given X, $E(Y \mid X)$, is equivalent to the probability that Y is in the obesity state, $P(Y=1 \mid X)$, which can be linked to a logistic function. The purpose to use logistic function is to transform the infinite range of linear predictor values in X*β to a confined interval between zero and one, an interval that represents the probability to have a depression. The maximum likelihood estimator is used to uncover the unknown coefficient vector β. In practice, the results are often presented using the following log of odds equation. So β_i can be interpreted as the effect of one-unit change in X_i on log of odds. Alternatively, we may take the exponentiation of β_i and this value can be explained as an odds ratio of comparing the relative risk of having depression in a category to its reference category.

$$\log\left(\frac{P}{1-P}\right) = X*\beta \quad (2)$$

$$\frac{P}{1-P} \text{ (Odds)} = \exp(X*\beta) \text{ or } e^{X*\beta}\text{; exp: exponential function.} \quad (3)$$

Once the estimation of β ($\hat{\beta}$) is found, the values of X* $\hat{\beta}$ can be used as an input in equation (4) in order to find the predicted probability $\hat{P}$ (a.k.a. score).

$$\hat{P}\ (Y = 1 \mid X) = \frac{\exp(X*\hat{\beta})}{1+\exp(X*\hat{\beta})} \quad (4)$$

The key to build a successful classifier or equivalently a predictive model hinges on the choice of θ, a threshold value used to categorize (predict) the cases into obesity or normal states. In other words, we specify each case as "Obese" if $\widehat{P} > \theta$ or "Normal" otherwise. By combining with the actual obesity and normal cases, we can form a two-by-two confusion matrix as shown in the following table.

Confusion Matrix

Predicted \ Actual	0	1
0	TN	FN
1	FP	TP

TP: True Positives (the predicted depressive cases are indeed cases with depression)
TN: True Negatives (the predicted depressive-free cases are indeed cases without depression)
FP: False Positives (the predicted depressive cases are actually cases without depression)
FN: False Negatives (the predicted depressive-free cases are actually cases with depression)

Accuracy = $\frac{TP + TN}{TP + FN + TN + FP}$ (the proportion of correctly identified cases (obesity & normal) in the entire sample)

Recall = $\frac{TP}{TP + FN}$ (the proportion of correctly identified obesity cases in the total of actual depressive cases)

Each cell in the confusion matrix represents whether the state of predicted cases matches the state of actual cases; and they are true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). To measure the performance of a predictive model, we adopted two metrics, accuracy rate and recall rate. When the data is balanced with about the similar amount of obesity and normal cases, accuracy rate alone can be used as a performance measurement. However, this measurement can be misleading if the data is imbalanced [12]. As indicated in Table 1, the proportion of obesity cases accounts for 16.96% of the data. Therefore, a high accuracy rate can be attributed mostly by correctly identified normal cases because they represent 83.04% of the data. To overcome this problem, we take another performance metric, recall rate, into consideration. Recall rate is also called sensitivity. Our strategy is to choose an optimal threshold value by maximizing

the recall rate while maintaining the highest possible accuracy rate. There are different methods to choose an optimal threshold value under the issue of imbalanced data. Our strategy is subjective since the true cost of false identifications (FN and FP) is unknown.

RESULTS

We applied a three to one split to our sample; 75% of the observations (67,960 cases) was used for estimation purpose and the rest of 25% of observations (22,654 cases) were used to examine the validity of the predictive model. The empirical results of the logistic regression model were reported in Table 2. The adjusted odds ratio (AOR) can be used to examine the relative risk of a group of each social determinant comparing to its reference group. For example, when considering the gender variable, being a girl has a lower risk to be obese (0.6729 to 1) than being a boy (the reference group). Overall, being a girl had lower risk of having obesity than a boy (AOR = 0.6729, $p < 0.01$). Compared with adolescents in the 12-13 age group, those in the 14-15 and 16-17 age group were more likely to have obesity (AOR = 1.0350, $p = 0.2110$; AOR = 1.0624, $p < 0.05$). However, the AOR for 14-15 age group was not significant. It can be observed from the table where some race/ethnicity groups may also be associated with obesity. Comparing to White, Black and Hispanic had a higher risk of being obese (AOR = 1.3080, $p < 0.01$; AOR = 1.2738, $p < 0.01$). In contrast, Asian/NHPIs had a lower risk of being obese (AOR = 0.6710, $p < 0.01$) comparing to White. With respect to family related factors and school experiences, living in a single-father or single-mother households significantly increased the chance of having obesity (AOR = 1.0593, $p < 0.05$; AOR = 1.0746, $p < 0.05$). Low and Medium authoritative parenting style (comparing to high authoritative parenting) had a higher chance of becoming obese (AOR = 1.1705, $p < 0.01$; AOR = 1.0527, $p < 0.05$). Bad and OK school experiences (comparing to good school experiences) also increased the likelihood to have obesity (AOR = 1.2721, $p < 0.01$; AOR = 1.0951, $p < 0.01$). Among those significant ($p < 0.05$) social determinants, gender, race and school experiences had the larger size effect. It was also noticed that adolescents with Medicaid/CHIP insurance coverage had a higher risk of being obese comparing to those without insurance at all.

As explained in the Methods subsection, we searched for an optimal threshold value by

maximizing the accuracy rate while maintaining the highest accuracy rate. The selection criteria can be found using Figure 3. Both accuracy and recall rates were plotted using threshold values ranged from 0.4 to 0.05 (a descending order on horizontal axis). As threshold value decreased, the recall rate was getting higher while the accuracy rate was getting lower. When the threshold value was at about 0.1, 93.37% of recall rate and 31.11% of accuracy rate were achieved. The corresponding confusion matrix (when $\theta = 0.1$) was reported in Table 3(A). At the intersection point of both rates, the threshold value θ is 0.171, and approximately 59.90% of accuracy and recall rate can be obtained. We also checked both recall and accuracy rate using the test data. They were 58.68% and 59.49%, respectively. Both recall and accuracy rates were calculated using confusion matrices that were reported in Table 3(B) and 3(C). More detailed explanations on both rates were provided in the discussion section. Since these two rates are about the same in both training and test data, it indicates that there is no overfitting issue in the predictive model. A four-hold cross-validation method was applied to examine the general performance of the predictive model, and the corresponding recall and accuracy rates are all about the same.

DISCUSSIONS

Using the NSDUH survey data, the logistic regression model was used as a statistical tool to investigate the association of social determinants and adolescent obesity. The outcomes from the model were used a guidance to build a predictive model in which the logistic regression model was served as a classifier; a popular supervised machine learning algorithm to identify a binary target variable (Normal vs. Obese in adolescents). According to the data, the prevalence of obesity among adolescents was about 16.97%, a measure similar to previous findings [number it]. The empirical findings using logistic regression as a statistical tool uncovered the direction and strength of association of each social determinant with the obesity variable. It was difficult to compare our results with previous findings directly. But in general, the majority of the social determinants available in this study had significant associations with adolescent obesity. This indicates that social determinants of health can be an important contributing factor to obesity and requires more careful study.

When using the logistic model as a classifier, the common challenge is how to select model performance metrics. This challenge is more pronounced when dealing with imbalanced data. In our case, the obesity cases accounted for a much smaller portion of the data comparing to normal cases (16.96% vs. 83.04%). Our strategy to maximize recall rate while maintaining the highest possible accuracy rate may not be the best option. An alternative strategy is to specify the cost functions for misclassifications (false positives & false negatives) in order to decide a better threshold value θ. In addition, there exist many other supervised machine learning algorithms that may give rise to better prediction outcomes. Our nest step in this project is to combine other machine learning algorithms with the logistic model in order to build a more powerful classifier. This ensemble approach has been commonly used in other natural science related areas [insert references] but still new to social science studies.

CONCLUSIONS

Obesity in adolescent is a serious health problem. While the causes of obesity can vary, our empirical findings indicate that social determinants of health such as gender, age, race/ethnicity, family composition, parenting style, and school experiences are associated with adolescent obesity significantly. The size effects point out girls, Hispanic/Black, and bad school experiences may contribute more to the chance of higher risk of having obesity among adolescents. Preventive measures can be introduced to focus on the above aforementioned social determinants of health. Special attention can be paid to the more vulnerable groups. For example, more coordination between school and family is needed to provide better school experiences to reduce the chance of having obesity. Lastly, guided by these empirical findings, the logistic model can help pre-identify potential obese cases using selected features (i.e., social determinants of health). The resulting predicted probability can be converted into a risk score for evaluating and detecting the potential obesity cases in adolescents.

Human Subject Statement

Conflict of Interest Statement

References

Figure 1

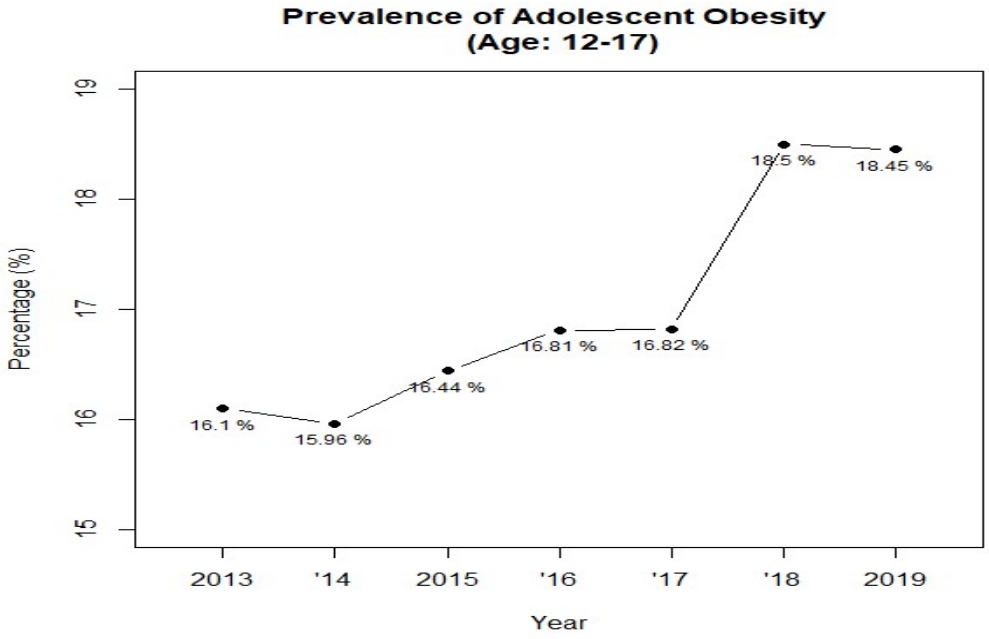

Figure 2

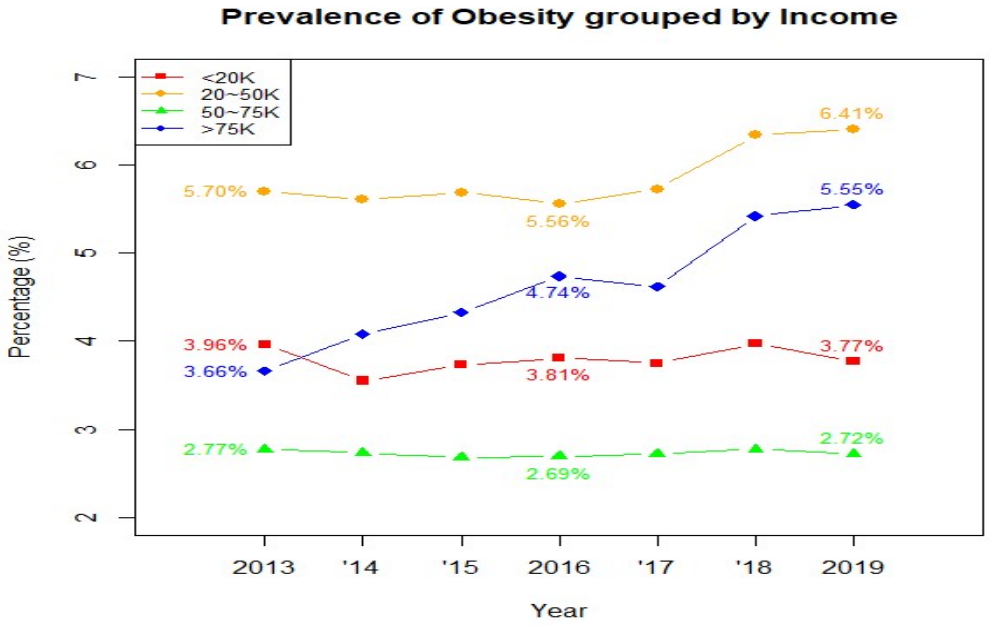

Figure 3

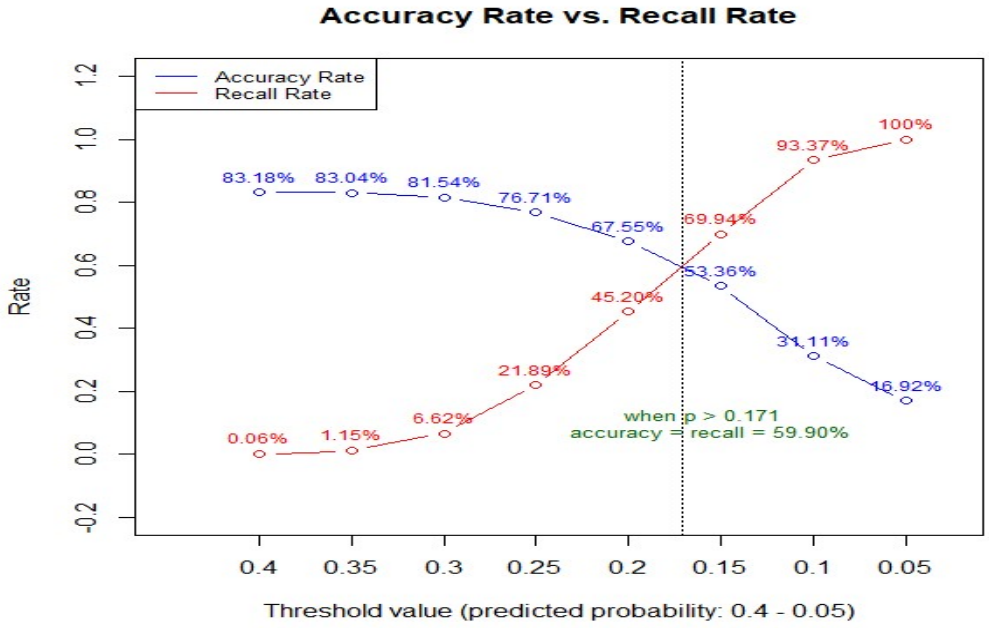
Accuracy Rate vs. Recall Rate
Accuracy Rate
Recall Rate
Rate
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
83.18%
83.04%
81.54%
76.71%
67.55%
53.36%
31.11%
16.92%
0.06%
1.15%
6.62%
21.89%
45.20%
69.94%
93.37%
100%
when p > 0.171
accuracy = recall = 59.90%
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
Threshold value (predicted probability: 0.4 - 0.05)

Table 1. Relative Frequency of Obesity & Social Determinants (N = 90,614)

Obesity	
Normal	83.04%
Obese	16.96%
Gender	
Male	51.27%
Female	48.73%
Age	
12-13	29.84%
14-15	34.66%
16-17	35.50%
Race/Ethnicity	
White	56.90%
Hispanic	19.68%
Black	12.88%
Asian/NHPIs	4.05%
Other	6.49%
Insurance type	
Private	59.26%
Medicaid/CHIP	33.61%
Other	1.52%
No	5.61%
Household income	
<20,000	15.62%
20,000-49,999	28.10%
50,000-74,999	15.80%
>75,000	40.48%
Father in household	
Yes	72.65%
No	27.35%
Mother in household	
Yes	91.51%
No	8.49%
Siblings under 18 in household	
Yes	68.28%
No	31.72%
Authoritative parenting	
High	56.66%
Medium	37.84%
Low	5.49%
School experience	
Good	49.89%
OK	43.40%
Bad	6.71%

Table 2. Logistic Regression Results

Variables	Estimates (AOR)	95% CI (Lower Bound)	(Upper Bound)
(Intercept)	-	-	-
Year	1.0366***	1.0259	1.0475
Gender(ref: Boy)	0.6729***	0.6449	0.7022
Age (ref: Age 12-13)			
Age14-15	1.0350	0.9807	1.0923
Age16-17	1.0624**	1.006	1.1219
Race (ref: White)			
Hispanic	1.2738***	1.2056	1.3456
Black	1.3080***	1.2246	1.3967
Asian/NHPIs	0.6710***	0.5894	0.7612
Other	1.2204***	1.1233	1.3246
Insurance (ref: No)			
Private	0.9485	0.8637	1.0429
Medicaid/CHIP	1.2208***	1.1125	1.3413
Other	0.9452	0.7776	1.1436
Income (ref: < $20,000)			
$20,000 - 49,999	0.8939***	0.8408	0.9505
$50,000 - 74,999	0.7987***	0.7383	0.8639
$75,000 or more	0.5376***	0.4976	0.5809
Father in household (ref: Yes)	1.0593**	1.0076	1.1137
Mother in household (ref: Yes)	1.0746**	0.9999	1.1541
Siblings in household (ref: Yes)	1.1865***	1.1334	1.2421
Authoritative parenting (ref: High)			
Medium	1.0527**	1.0054	1.1121
Low	1.1705***	1.0672	1.2824
School experience (ref: Good)			
OK	1.0951***	1.0465	1.1460
Bad	1.2721***	1.1690	1.3833

*p < 0.1; **p < 0.05; ***p<0.01 (0.05 was used to decide the significance of estimates
AOR: adjusted odds ratio
Null deviance: 58353 on 64446 degrees of freedom
Residual deviance: 56287 on 64425 degrees of freedom
AIC: 56331
McFadden's Pseudo R^2: 14.22% (recalculate)

Table 3. Confusion Matrix

(A)

θ = 0.5 (training)		
Predict\Actual	0 (Normal)	1 (Obese)
0 (Normal)	9942 (TN)	718 (FN)
1 (Obese)	43677 (FP)	10110 (TP)

recall rate = 93.37%; accuracy rate = 31.11%

(B)

θ = 0.171 (training)		
Predict\Actual	0 (Normal)	1 (Obese)
0 (Normal)	32063 (TN)	4337 (FN)
1 (Obese)	21556 (FP)	6491 (TP)

recall rate ≅ accuracy rate = 59.90%

(C)

θ = 0.171 (test)		
Predict\Actual	0 (Normal)	1 (Obese)
0 (Normal)	10701 (TN)	1495 (FN)
1 (Obese)	7239 (FP)	2123 (TP)

recall rate = 58.68%; accuracy rate = 59.49%