Multiple logistic regression analysis of cigarette use among high school students

Joseph Adwere-Boamah
Alliant International University

ABSTRACT

A binary logistic regression analysis was performed to predict high school students' cigarette smoking behavior from selected predictors from 2009 CDC Youth Risk Behavior Surveillance Survey. The specific target student behavior of interest was frequent cigarette use. Five predictor variables included in the model were: a) race, b) frequency of cocaine use, c) initial cigarette smoking age, d) feeling sad or hopeless, and e) physically inactive behavior. The results of the logistic regression analysis showed that the full model, which considered all the five independent variables together, was statistically significant. . The strongest predictors of youth smoking behavior were race, frequency of cocaine use and physically inactive behavior. For example, the odds of smoking are increased by a factor of 5.0 if the student is White compared to an African American, controlling for other variables in the model. The logistic model employed explained about 31% of the variance in current frequent cigarette use among the high school students. It correctly classified 93% of the cases. The key finding is that the selected variables are important correlates of frequent cigarette use among high school students.

Keywords: Logistic regression, CDC, Youth risk behavior surveillance system, cigarette smoking.

“Tobacco use is the single most preventable cause of disease, disability, and death in the United States. Each year, an estimated 443,000 die prematurely from smoking or exposure to secondhand smoke, and another 8.6 million have a serious illness caused by smoking” (CDC, 2010. p.1).

INTRODUCTION

The above quotation from CDC documents the harmful effects of cigarette smoking. Despite the risks associated with smoking, CDC (2010) estimates 46 million U.S. adults smoke cigarettes. The Department of Health and Human Services’ Center for Disease Control (CDC) and Prevention maintains cigarette use is the leading preventable cause of death in the United States. The CDC lists tobacco use by young adults as one of its priority health-risk behaviors (CDC, MMWR, 2004). Because most smokers initiate cigarette use during adolescence (Hersch, 1998), the prevalence of both past-year and past-month smoking peeks during smokers' late teens or early twenties. The adverse long-term effects of cigarette smoking as documented above by the CDC are inversely correlated with earlier initial smoking ages. Concomitantly, the CDC has set a national objective to reduce the prevalence of cigarette use among high school students to less than 16% for the year 2010 (CDC, MMWR, 2006).

The CDC has developed a Youth Risk Behavior Surveillance System (YRBSS) to monitor the health risk behaviors among American youth. The system monitors six categories of high-risk behaviors that according to the CDC contribute substantially to "the leading causes of death, disability, and social problems among youth and adults in the United States" (CDC, MMWR, 2010 page 1). The categories of high-risk behaviors are: 1. Behaviors that contribute to unintentional injury and violence, 2. Alcohol and other drug use, 3. Sexual behaviors that contribute to unintended pregnancy and sexually transmitted diseases including HIV infection, 4. Physical inactivity, 5. Obesity and dietary behaviors, and, most importantly for this study, 6. Tobacco use.

The purpose of this study was to assess the impact of a set of predictors on cigarette smoking behavior of high school students. Specifically, the target outcome behavior of interest is Current Frequent Cigarette Use (CFCU) among the youth. This study sought to answer the question: Can CFCU among the youth be accurately predicted from Race, Frequency of Cocaine Use (FCU), age at which the youth smoked a whole cigarette for the first time [Initial Cigarette Smoking Age (ICSA)], feeling of sadness or hopelessness, [Felt Sad or Hopeless (FSH)] and Physically Inactive Behavior (PIB). See Figure 1 for the specific CDC questionnaire items and the predictors or independent variables of this study.

The paper is structured as follows: First, I briefly review the methodological strategy employed and the data source for this study. Subsequently, I present the results of the data analysis through descriptive statistics and a logistic regression indicating the significant predictors. Finally, I summarize the study results.

METHOD

The dependent or the outcome variable of interest, Current Frequent Cigarette Use was constructed as a ‘yes/no’ dichotomous indicator of smoking status based on the response to CDC (2009) survey questionnaire item: "During the past 30 days, on how many days did you smoke cigarettes?" Respondents who answered they smoked 20 to 29 days or all the 30 days were classified as current frequent cigarette user (coded 'yes' otherwise 'no' for less frequent or non cigarette users).

The categorical dependent variable of the study necessitated the use of multiple logistic regression model for investigating whether the likelihood of current frequent cigarette use among the youth was related to the selected predictors above. (Menard, 2010; Hosmer, & Lemeshow, 2000). The specific logistic regression model fitted to the data was:

Logit (CFCU) = $b_o + b_1$ (Hispanic) + b_2 (White) + b_3 (ICSA) + b_4 (FSH) + b_5 (FCU) + b_6 (PIB).

(Where b_0 is a constant. $b_1, b_2, \ldots b_6$ are logistic coefficients or estimates for the parameters, $\beta_1, \beta_2, \ldots \beta_6$).

Race was a design variable coded, Hispanic = 1, White = 2, African American was the reference category. PIB and FSH are categorical variables and were dummy coded 0, 1.

DATA SOURCE

This study drew from the CDC's 2009 National Youth Risk Behavior Survey (YRBS), a questionnaire containing items designed to elicit information from high school students about the fore-mentioned categories of health-related risk behaviors, along with basic demographic information.

The sampling frame for the survey consisted of all public and private schools with 9-12 grade students. Representative samples of students were drawn from those grades. All the questionnaires were self-administered. Student response rate for the national survey was 88%, approximately 16460 students. For this study, usable data from 11683 students were analyzed.

DATA ANALYSIS.

Prior to analysis of the data the dependent variable of interest, current frequent cigarette use which had binary response of Yes/No, was recoded 0, 1 (No/Yes). The coding change was made to reflect the predicted target category, current frequent cigarette use.

As suggested by Menard (2010), preliminary analysis of the data was performed to check the assumptions of logistic regression with respect to the selected predictors of the study. ICSA, FSH, FCU, PIB and Race were subjected to Linear regression analysis to evaluate multicollinearity among the predictors or the independent variables. Multicollinearity among predictors in logistic regression creates problems for the validity of the model for the investigation. In particular, it affects the validity of the statistical tests of the regression coefficients by inflating their standard errors. (Garson, 2010). The results of the analysis showed that the data did not violate the multicollinearity assumption. The tolerance value of each independent variable was greater than .720 which exceeded the suggested criteria of below .10. (Pallant, 2007). Lack of multicollinearity among the independent variables was also supported by the obtained variance inflation factor (VIF) values. They were all well below the cut-off value of 10. (Field, 2005). The VIF values of the variables ranged from 1.013 to 1.376. After the preliminary analysis of the data, the binary logistic regression procedure in SPSS was used to perform the analysis to determine whether the likelihood of CFCU could be predicted from the independent variables. Data from 11683 high school students were included in this analysis.

RESULTS

Sample description: The age distribution of the students ranged from 14 to 18 years old. Among the respondents, approximately 70% were White, 17% African Americans and 13% Hispanics. About 11% of the students smoked a whole cigarette for the first time before age 13. About 46% of the students have ever smoked cigarette.

The prevalence of “Ever smoked” cigarettes was higher among Hispanic students (51.6%) than White (46.1%) and African Americans (43.5). Most “Ever smokers” first smoked at age 13 or 14. The percentage of “Ever smokers” that were 13 or 14 years old was 25%. The corresponding percentages for “Ever smokers” that were 11 or 12, 15 or 16 years old were 12% and 23% respectively. The results of the data analysis showed that the proportion of students who have tried smoking increased with age. For example, by the age of 18, approximately 53% of the youth have tried smoking. In contrast to the relatively high prevalence (51.6%) of “Ever Smoked Cigarettes” among Hispanic youth, a relatively small percentage of Hispanics have ever smoked cigarettes daily, i.e., had ever smoked at least one cigarette every day for 30 days. For example, the prevalence of ever smoked cigarettes daily was higher among White (13.7%) than African American (4.3%) and Hispanic (6.3%).

The results of the logistic regression analysis show that the full model which considered all the five independent variables together was statistically significant, $\chi2$ = 1700.966, df = 6, N = 11424, p < .001.This implies that the odds for a high school student to indicate that he was a current frequent cigarette user were related to the five independent variables, Race, ICSA, FSH, FCU, and PIB.

The model correctly classified approximately 93% of the cases. The “pseudo” R estimates indicate that the model explained between 13% (Cox & Snell R Squared) and 31% (Nagelkerke R Squared) of the variance in current frequent cigarette use. Table 1 presents a summary of the raw score binary logistic regression coefficients, Wald statistics, odds ratios [(Exp (B)] along with a 95% CI. Wald statistics indicate that all the variables significantly predict current frequent cigarette use. The strongest predictor of CFCU was race. In particular, white. The odds ratio for white was 5.0 i.e., the odds of a high school student indicating that he is a current frequent cigarette user are increased by a factor of 5.0 if the student is White compared to African American adjusting for the effects of the other predictors in the model.

Other predictors that made significant contribution to the model (CFCU) were frequency of cocaine use, physical inactive behavior, feeling sad or hopeless and initial cigarette smoking age. The older the youth before he smoked a whole cigarette for the first time, the more likely he would report that he is a current frequent cigarette user. The predictor (ICSA) recorded an odds ratio of 1.6. Thus, the odds of smoking frequently compared to not smoking cigarettes frequently increase by a factor of 1.6 for a unit increase in age from when the youth smoked cigarette for the first time. In other words, the odds of current frequent cigarette use increase by 60% for each unit increase in ICSA. (Warner, 2008).

Cocaine use and cigarette smoking behavior of young people were strongly related. Frequent cocaine use (FCU) predicts smoking behavior. ($p < .001$). For a unit increase in the number of times the youth uses any form of cocaine, including powder, crack, or freebase, the odds for smoking cigarettes frequently i.e., smoking 20 to 29 days or 30 days in one month are increased by a factor of 3.5 when all other variables are held constant. Feeling sad or hopeless recorded an odds ratio of 1.7. This indicates that the odds of smoking cigarettes frequently were about 1.7 times higher for high school students who felt sad or hopeless than for those who did not feel that way. As shown in Table 1, physically inactive behavior is implicated in student smoking. It recorded an odds ratio of about 1.7. In other words, students who were physically inactive have higher odds of smoking frequently (more than 1.7 times as high) compared with students who are physically active, controlling for other variables in the model.

SUMMARY

Multiple logistic regression was used to jointly examine the influence of race, frequency of cocaine use, physically inactive/active behavior, initial cigarette smoking age and feeling sad or hopeless The key finding is that the selected variables are important correlates of current frequent cigarette use among high school students. The strongest predictors of youth smoking behavior are race, frequency of cocaine use and physically inactive behavior. For example, the odds of smoking are increased by a factor of 5.0 if the student is White compared to an African American. The logistic model employed explained about 31% of the variance in current frequent cigarette use among the high school students. It correctly classified 93% of the cases.

Table 1- Logistic regression predicting the likelihood of high school students reporting frequent cigarette use.

Predictor	B	S.E	Wald	Df	P	Odds Ratio	95% C.I. Lower	95% C.I Upper
Race White	1.62	.18	77.75	1	.00	5.4	3.52	7.23
Hispanic	-.46	.27	2.95	1	.09	.63	.38	1.07
FCU	1.26	.07	285.62	1	.00	3.52	3.04	4.07
FIB	.54	.08	41.44	1	.00	1.71	1.45	2.02
FSH	.53	.08	43.65	1	.00	1.70	1.45	1.98
ICSA	.49	.02	712.411	1	.00	1.64	1.58	1.70
Constant	-7.23	.22	1050.87	1	.00	.00		

Figure 1- Selected predictors from CDC's 2009 National Youth Risk Behavior Survey (YRBS).

CDC questionnaire item	Variable name	Acronym
Q5. What is your race	Race	
Q23. During the past 12 months, did you feel ever so sad of hopeless almost every day for two weeks or more in a row that you stopped doing some usual activities?	Felt sad or hopeless	FSH
Q29. How old were you when you smoked a whole cigarette for the first time?	Initial cigarette smoking age	ICSA
Q50. During the past 30 days, how many times did you use any form of cocaine, including powder, crack, or freebase?	Frequency of cocaine use	FCU
Q80. During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day?	Physically inactive/active behavior	PIB

REFERENCES

Center for Disease Control and Prevention, (2010). *Tobacco Use: Targeting the Nation's Leading Killer*. Available at: www.cdc.gov/chronicdisease/resources/publications/aag/osh.htm. Accessed on Sept. 2010.

Center for Disease Control and Prevention, (2009). *Youth Risk Behavior Survey.* Available at: www.cdc.gov/yrbs. Accessed on Sept. 2010.

CDC, (2004). Methodology of the Youth Risk Behavior Surveillance System. *MMWR*. Sept. 24, vol.53. RR-12

CDC, (2006). Cigarette Use Among High School Students – United States, 1991-2005. *MMWR*. 55 (26); 724-726.

CDC, (2010). Youth Risk Behavior Surveillance – United States 2009. *MMWR*. June 4, vol.59/SS-5.

Chao-Ying; Pen, J; & Task-Shing, H. (2002). Logistic Regression Analysis and Reporting: A Primer. *Understanding Statistics: statistical Issues in Psychology, Education, and the Social Sciences. 1(1),* 31-70.

Field, A. (2009). *Discovering Statistics Using SPSS*. London: Sage.

Garson, D. (2010). *Logistic Regression: Footnotes, from North Carolina State University*. Available at: http:/faculty.chass.ncsu.edu/garson/PA765/logistic.htm. Accessed on June 2010.

Hersch, J. (1998). Teen Smoking Behavior and the Regulatory environment. *Duke Law Journal, Vol*. 47:1143

Hosmer, D.W; & Lemeshow, S. (2000). *Applied Logistic Regression, second edition.* New York: Wiley

Menard, S. (2010). *Logistic Regression: From Introductory to Advanced Concepts and Applications.* Thousand Oaks, CA: Sage.

Pallant, J. (2007). *SPSS Survival Manual.* Open University Press. England: Berkshire.

Pampel, F.C. (2000). *Logistic Regression: A Primer.* Sage Quantitative Applications in the Social Sciences Series #132. Thousand Oaks, CA. Sage Publications.

Tabachnic, B. G; & Fidell, L.S. (1996). *Using Multivariate Statistics, 3rd ed.* New York: Harper Collings.

Warner, R. (2008). *Applied Statistics: From Bivariate Through Multivariate Techniques*. Thousand Oaks, CA. Sage Publications.