Statistical Analysis of Past-Year Marijuana use in U.S. General Population: A Negative Binomial Regression Model

Qin Zhao[1*], Kesheng Wang[2], Ying Liu[3]

[1]School of Electronics and Information Engineering, Binjiang College of Nanjing University of Information Science and Technology, Wuxi 214105, China
[2]School of Nursing, Health Sciences Center, West Virginia University, Morgantown, WV 26506, USA
[3]Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, Johnson City, TN 37614, USA

Abstract: Marijuana is the most frequently reported illicit substance in the United States. However, limited studies have delved into the analysis of marijuana use as a count variable, in which the distribution often exhibits overdispersion and notable occurrences of zero values. This study encompassed a total of 58,034 individuals, with12,528 having reported marijuana use in the past year from the 2021 National Surveys on Drug Use and Health data. Marijuana use was measured by number of days used in the past year. Three distributions were compared including normal distribution, Poisson, and Negative Binomial (NB) distributions. The Akaike information criterion (AIC), corrected AIC (AICC), consistent AIC (CAIC), and the Bayesian information criterion (BIC) statistics were used to select the best distribution. The overall prevalence of past-year marijuana use was 21.6%. The NB regression model proved to be the best with lowest AIC, AICC, CAIC, and BIC values compared with linear and Poisson models. According to the NB model, African American and age 18-64 years were associated with increased days of marijuana use, whereas, females, rural living, Asian and Hispanic were associated with decreased days of marijuana use. The findings can guide healthcare providers when screening for marijuana use in general population.

1. Introduction

Marijuana is the most frequently reported illicit substance in the United States (U.S.) and research indicates an upward trend in the prevalence of use, frequent use, and cannabis use disorders have been increasing in the U.S. [1-3]. When examining substance use, one crucial aspect is the quantification of outcomes such as the number of days used in the past month or year. This type of data falls into the category of count variables, characterized by distributions often displaying overdispersion (i.e., larger variance) and frequently containing zero values (i.e., zero-inflated). Consequently, the assumption of normality is frequently violated, rendering standard linear regression models unsuitable for analyzing count data [4].

Poisson regression has traditionally been commonly used in analysis of count data across various fields. However, the assumption of equal mean and variance is often violated. As an alternative, the negative binomial (NB) regression model includes a dispersion parameter allowing for the variance to exceed the mean and is gained popularity. In marijuana use, Poisson regression [1,4,5] and NB regression [4,6-8] have been used for days in past marijuana use. However, previous studies in marijuana use have primarily focused on clinical study and small sample size data set [6]. One study used the U.S. 2002-2014 National Surveys on Drug Use and Health (NSDUH) data and NB regression model [6]. Furthermore, there has been limited exploration into the comparison of different count models on marijuana use frequency in a large survey data. For example, one study describes different methods for analyzing counts on cigarette and marijuana smoking data using 67 subjects [4]. The present study compared 3 commonly used count model using a large national representative NSDUH 2021 data in the U.S.

1.1. Distributions in count data

Let Y denote the continuous outcome variable with k predictors Xs, the linear regression model with a normal data structure can be written in equation (1).

$$Yi \mid \beta 1, \beta 2, \ldots \beta k \sim N(\mu i, \sigma 2) \; with \; \mu i = \beta 0 + \beta 1 Xi1 + \cdots + \beta k Xik \quad (1)$$

The Poisson regression model is commonly used to model discrete counts of events, especially in cases in which counts are right-skewed and do not follow a normal distribution. The equation of Poisson regression model can be written in (2).

$$Yi \mid \beta 1, \beta 2, \ldots \beta k \sim Pois\left((\lambda i) \; with \log(\lambda i) = \beta 0 + \beta 1 Xi1 + \cdots + \beta k Xik\right) \quad (2)$$

The NB regression model is for modeling over-dispersed count data when the variance exceeds the mean. It can be considered as a generalization of Poisson regres-

*E-mail: zhaoqq567@outlook.com

sion since it has the same mean structure as Poisson regression and it has an extra parameter (r) to model the over-dispersion. The NB model with mean parameter μ and reciprocal dispersion parameter r can be written in (3):

$$Yi \mid \beta 1, \beta 2, \ldots \beta k, r \sim NegBin\ (\mu i, r)\ with \log\ (ui) = \beta 0 + \beta 1Xi1 + \cdots + \beta kXik\ and\ r = EXP\ (3) \quad (3)$$

2. Subjects and methods

2.1. Study sample

Data were extracted from the latest 2021 NSDUH data. The NSDUH is conducted by the Substance Abuse and Mental Health Services Administration of the United States Department of Health and Human Services. The NSDUH is a survey of the civilian non-institutionalized individuals (≥12 years old) in the U.S. to provide annual population estimates of substance use and health. The NSDUH measures the prevalence and correlations of drug use in the U.S. quarterly and annually. Information from the survey contains data on the use of illicit drugs (e.g., marijuana, cocaine, hallucinogens, heroin, and inhalants), alcohol and tobacco, and nonmedical use of prescription drugs (e.g., pain relievers, tranquilizers, and sedatives). The total sample size of the 2021 NSDUH data is 58,034. Details of the survey design and data collection methods are published elsewhere [9]. There was an Institutional Review Board exemption due to secondary data analysis.

2.2. Measures

The primary outcome was number of days of marijuana use in the past year as a count variable. The past-year marijuana use was coded as binary variable (Yes, No). Demographic factors included participants' age (12-17, 18-25 years, 26-49 years, 50-64 years, and 65 years or older), sex (male and female), and race/ethnicity (Non-Hispanic White, Non-Hispanic African Americans, Asian, Hispanics, and others). There were four categories of marriage status: married; widowed; divorced/separated, and never been married. Education has 5 levels: Less high school, High school grad, Some college/Assoc Degree, College graduate, and 12- to 17-year-olds. Urban-rural status was coded as large metro, small metro and nonmetro. General health has four levels: Excellent, very good, good, and fair/poor.

2.3. Statistical Analysis

The categorical variable was presented in raw value along with the proportions. Continuous variable was presented with mean and standard deviation (SD). Initially, general linear model (GLM) was used to detect the association of each risk factor with the days of past-year marijuana use as a continuous variable. Then Poisson regression model and NB regression model were built in the generalized linear model procedure.

Fit Statistics.

The Akaike information criterion (AIC) is a measure of goodness of model fit that balances model fit against model simplicity. AIC has the form (4)

$$AIC = -2LL + 2p \quad (4)$$

where p is the number of parameters estimated in the model, and LL is the log likelihood evaluated at the value of the estimated parameters.

The corrected AIC (AICC) is given by (5)

$$AICC = -2LL + 2pn/(n - p - 1) \quad (5)$$

where n is the total number of observations used.

Consistent AIC (CAIC) can be written in (6)

$$CAIC = -2LL + p[log(n) + 1] \quad (6)$$

The Bayesian information criterion (BIC) is a similar measure and defined by

$$BIC = -2LL + plog(n) \quad (7)$$

Model selections using AIC, AICC, CAIC, and BIC statistics were recently discussed [10-12]. Models with smaller values of these statistics represented better model fit.

All analysis was conducted with SPSS (IBM Corp. Released 2019. IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp).

3. Results

3.1. Prevalence of past-year marijuana use across demographic variables

Table 1 provides an overview of the prevalence of ER visits in the past year among various demographic variables. The overall prevalence of past-year marijuana use was 21.6% with a breakdown of 22.6% for males and 20.7% for females. Within different age groups, the highest prevalence in adults occurred among aged 18 to 25 years reaching the highest 34.0%, while the second-highest prevalence was observed in 26 to 34- year age group at 29.3%.

3.2. Number of days of past-year marijuana use across demographic variables

Table 1 also describes data on the number of days of past-year marijuana use across various demographic variables. The overall mean number of days of past-year marijuana use was 30.24 days. Notably, the skewness and kurtosis values were 3.02 and 7.77, respectively, indicating a departure from a normal distribution. On average males reported more days of marijuana use (34.73 days) compared to females (26.49 days). The highest mean number of days was recorded among adults aged 18 to 25 years, amounting to 49.12 days.

3.3. Comparison of different count distributions

Table 2 shows a comparison of three models using fit statistics, including AIC, AICC, CAIC, and BIC. The NB regression model emerged as the superior choices with lowest AIC, AICC, CAIC, and BIC values (442154.78, 442154.80, 442371.92, and 442349.92, respectively). In contrast, both linear (620395.53, 620395.55, 620622.54, and 620599.54, respectively) and Poisson models

(6187407.70, 6187407.72, 6187624.84, and 6187602.84, respectively) lagged behind in terms of fit.

Table 1. Past-year marijuana use across demographic variables.

Variable	Total No. (%)	No. Marijuana use	Prevalence (%)	Days Mean ± SD	*t*-statistic, *P* value[a]
Gender (Ref=Male)	26,391 (45.5)	5968	22.6	34.73±93.61	
Female	31,643 (54.5)	6560	20.7	26.49±82.02	-11.3, <0.001
Age group (Ref=12-17 years)	10,743(18.5)	1094	10.2	8.75±44.27	
18-25 years	13,979 (24.1)	4747	34	49.12±106.58	36.56, <0.001
26-34 years	9,588 (16.5)	2813	29.3	44.54±105.54	29.60, <0.001
34-49 years	12,561 (21.6)	2660	21.2	31.13±90.37	19.79, <0.001
50-64 years	5,725 (9.9)	806	14.1	19.37±71.62	7.55, <0.001
65+ years	5,438 (9.4)	408	7.5	8.32±45.71	-0.297, 0.767
Race/Ethnicity (Ref=White)	34,791 (59.9)	7736	22.2	30.33±87.87	
African American	6,743 (11.6)	1552	23.0	37.00±95.36	5.75, 0.001
Asian	3,234 (5.6)	308	9.5	8.00±44.87	-13.72, <0.001
Hispanic	9,929 (17.1)	1923	19.4	26.48±81.78	-3.88, <0.001
Other	3,337 (5.8)	1009	30.2	48.35±108.29	11.39, <0.001
Education (Ref=Less high school)	4,473 (7.7)	1104	24.7	46.07±107.65	
High school	11,189 (19.3)	2884	25.8	44.97±106.15	-0.72, 0.472
Some college	14,251 (24.6)	3935	27.6	41.94±101.39	-2.79, 0.005
College	17,378 (29.9)	3511	20.2	20.36±70.97	-17.75, <0.001
12-17 years old	10,743 (18.5)	1094	10.2	8.75±44.27	-24.29, <0.001
Marital status (Ref=Married)	19,778 (37.6)	2886	14.6	18.70±70.93	
Widowed	1,416(2.7)	121	8.5	12.05±57.87	-2.67, 0.008
Divorced or separated	4,875 (9.3)	1220	25.0	38.64±98.35	13.79, <0.001
Never married	26,519 (50.4)	8116	30.6	44.04±102.48	29.93, <0.001
Urban_rural (Ref= Large Metro)	25,956 (44.7)	5723	22.0	29.63±86.29	
Small Metro	21,883(37.7)	4721	21.6	30.47±88.14	1.04, 0.299
Nonmetro	10,195 (17.6)	2084	20.4	31.30±89.60	1.63, 0.104
General health (Ref=Excellent)	13,030 (22.5)	2091	16.0	17.74±66.48	
Very good	21,931(37.8)	4673	21.3	27.13±82.58	9.76, <0.001
Good	16,870 (29.1)	4002	23.7	36.92±96.60	18.89, <0.001
Fair/poor	6,184 (10.7)	1760	28.5	49.41±110.26	23.56, <0.001
Overall	58034	12528	21.6	30.24±87.58	

Abbreviation: SD = Standard deviation.

[a] *t*-values are based on the general linear model when comparing means across levels of each variable.

Table 2. Model selection in generalized linear model for number of days of past-year marijuana Use.

Distribution	AIC	AICC	CAIC	BIC
Normal distribution	620395.531	620395.552	620622.540	620599.540
Poisson distribution	6187407.699	6187407.718	6187624.838	6187602.838
NB	442154.783	442154.802	442371.922	442349.922

NB: Negative binomial distribution; AIC: Akaike information criterion statistic; AICC: Corrected AIC; Consistent AIC (CAIC); BIC: Bayesian information criterion statistic

3.4. Negative binomial regression analysis of past-year marijuana use.

Utilized the NB model, African American individuals (β=0.034, P=0.022) and those aged 18-64 years (β=1.495, 1.576, 1.272, and 0.694 with P values <0.001 for age 18-25, 26-34, 35-49, and 50-64 years, respectively) exhibited an associated with increased days of marijuana use. Conversely, females (β=-0.39, P<0.001), individuals residing in rural area (β=-0.052 and -0.122 with P values <0.001, for age small metro and nonmetro, respectively), as well as those of Asian and Hispanic backgrounds (β=-1.432 and -0.371 with P values <0.001, respectively) displayed an association with a decreased days of marijuana use. As show in table 3.

Table 3. NB regression model in past-year marijuana use.

Variable	Days β ± SE	$\chi 2$, *P* value[a]
Gender (Ref=Male)		
Female	-0.399±0.009	1891.16, <0.001
Age group (Ref=12-17 years)		
18-25 years	1.495±0.021	5246.41, <0.001
26-34 years	1.576±0.019	6569.22, <0.001
34-49 years	1.272±0.018	5073.45, <0.001
50-64 years	0.694±0.020	116.65, <0.001
65+ years	0.64±3.77	0.85, 0.845
Race/Ethnicity (Ref=White)		
African American	0.034±0.015	5.26, 0.022
Asian	-1.432±0.021	4594.96, <0.001
Hispanic	-0.371±0.013	820.28, <0.001
Other	0.294±0.020	218.56, <0.001
Education (Ref=Less high school)		
High school	0.046±0.018	6.57, 0.010
Some college	0.108±0.018	37.49, <0.001
College	-0.405±0.018	512.48, <0.001
12-17 years old	0.244±0.029	72.49, <0.001
Marital status (Ref=Married)		
Widowed	-0.078±0.030	6.62, 0.010
Divorced or separated	0.695±0.017	1752.73, <0.001
Never married	0.685±0.013	2944.67, <0.001
Urban_rural (Ref= Large Metro)		
Small Metro	-0.052±0.010	26.79, <0.001
Nonmetro	-0.122±0.013	85.65, <0.001
General health (Ref=Excellent)		
Very good	0.411±0.012	1110.57, <0.001
Good	0.695±0.013	2838.75, <0.001
Fair/poor	0.994±0.017	3389.30, <0.001

Abbreviation: SE = Standard error.

[a] $\chi 2$-values are based on the multivariate NB regression analysis.

4. Discussion

This study marks the first application of the NB regression model, alongside comparisons with linear model and Poisson model in modelling the days of past-year marijuana use using extensive survey data. The finding in this study not only updated the prevalence of past-year marijuana use but also affirm the NB regression model is the most suitable choice for modelling the number of days of past-year marijuana use. Moreover, we uncover several demographic variables associated with the days of past-year marijuana use.

This study updated the prevalence of past-year marijuana use in 2021 across demographic factors. The overall prevalence was 21.5%. One previous study has shown that the marijuana use increased from 10·4% to 13·3% in adults in the U.S. from 2002 to 2014 [6]. Another study found that the prevalence was 14.70% in 2016 [2]. The prevalence of past-year marijuana use among age group of 18-25 years was 34% is lower than those of college students (19-26 years) in 2020 (44-47%) which was the highest it has been in the past 35 years [3]

While Poisson regression models have been used in analyzing the number of days of marijuana use [1,4,5]. For example, Poisson regression model was used to determine the associations between neighbourhood problems, social cohesion, and marijuana use among 119 majority African American emerging adult men in a small urban community [5]. Another study used Poisson regression to examine reliability of self-reported lifetime marijuana use among 794 electronic dance music (EDM) party attendees - a high-risk population for drug use [1]. However, the assumption of equal variance and mean is atypical in real-world scenarios. In contrast, the NB regression model accommodates over-dispersion in count data relative to a Poisson model. A feature observed in several studies of marijuana use [4, 6-8]. For example, NB regression model was used to describe the number of days of marijuana use in the previous year used the US 2002-2014 NSDUH data [6]. Another study used NB model to examine the association with marijuana use with child welfare-affiliated maltreated youth (n=216) and comparison youth (n=128) from the same community [7]. Furthermore, a zero-inflated NB regression models was used to predict marijuana use in a community sample of 187 U.S. young adults [8]. Interestingly, one study compared six different count methods including Poisson, zero-inflated Poisson (ZIP), hurdle Poisson (HUP), NB, zero-inflated NB (ZINB), and hurdle negative binomial (HUNB) for analyzing count data and applied these methods in marijuana use using 69 subjects between 16 and 20 years old [4]. In the current study, we extended this analysis by comparing three count models including linear regression, Poisson model, and NB regression model available in SPSS software in analysis of past-year marijuana use using a large survey data and highlighted the superior performance of NB model over lineal and Poisson models. It has been recommended that avoiding Poisson regression in most practical cases and advocate for using NB regression as a baseline model for count data, even when the data contain many zeros [4]. Furthermore, the present study further added that several demographic factors were associated with the days of past-year marijuana use in 2021 using the NB regression model.

This study boasts several strengths, including (1) the use of the most recent 2021 NSDUH data, (2) a large sample size along with adjustment for numerous factors, (3) the application of fit statistics (AIC, AICC, CAIC and BIC) for model selection. Additionally, this study contributes to the scientific understanding the pattern of past-year marijuana use frequency across demographic factors. However, it's essential to acknowledge certain limitations. Firstly, the reliance on self-reported data in NSDUH may introduce recall and social desirability bias. Secondly, defining the days of past-year marijuana use based on self-reported age has its limitations. Thirdly, this study limited itself to comparing three different count models available in SPSS software, while more advanced statistical software options, such as SAS or R packages, allow to employ zero-inflated NB, hurdle Poisson and NB regression that accommodates to zero-generating processes. Furthermore, it's important to note that this study did not incorporate control measures for other substance use variables and COVID-19 in the multivariate regression model, variables

that could potentially be linked to past-year marijuana use in 2021.

5. Conclusions

The present study underscores that enduring prevalence of the past-year marijuana use in 2021, coupled with evident health disparities in prevalence and usage pattern across demographic factors. Furthermore, it highlights the superiority of the NB regression model over linear and Poisson models in analyzing the number of days of past-year marijuana use, particularly in cases with extensive zero values. The findings pave the way for the development of more robust statistical models to address the challenges posed by large numbers of zeros in the analysis of past-year marijuana use. Additionally, the multivariate NB regression model reveals significant associations between demographic variables and the number of days of past-year marijuana use, emphasizing the need for targeted interventions across different demographic groups.

Data availability

The data that support the findings of this study are publicly available at Substance Abuse & Mental Health Data Archive, the National Survey on Drug Use and Health 2021 data.

Acknowledgements

The authors would like to thank the support of data from the 2021 National Survey on Drug Use and Health (NSDUH) dataset.

References

1. Palamar JJ, Le A 2020 Am. J. Drug Alcohol Abuse 46(6):708-717
2. Yu B, Chen X, Chen X, Yan H 2020 BMC Public Health 20(1):156
3. Schulenberg JE, Patrick ME, Johnston LD, O'malley PM, Bachman JG, Miech RA 2021 Monitoring the future national survey results on drug use 1975-2020: Volume II, College students and adults ages 19–60. http://monitoringthefuture.org/pubs.html#monographs.
4. Pittman B, Buta E, Krishnan-Sarin S, O'Malley SS, Liss T, Gueorguieva R 2020 Nicotine Tob. Res. 22(8):1390-8
5. Taggart T, Brown AL, Kershaw T 2018 Am. J. Mens Health 12(4):944-951
6. Compton WM, Han B, Jones CM, Blanco C, Hughes A 2016 Lancet Psychiatry 3(10):954-964
7. Schneiderman JU, Kennedy AK, Negriff S, Jones J, Trickett PK 2016 J. Child Fam. Stud. 25(12):3481-3487
8. Ramirez JJ, Lee CM, Rhew IC, Olin CC, Abdallah DA, Lindgren KP 2020 J. Stud. Alcohol Drugs 81(1):81-88
9. Center for Behavioral Health Statistics and Quality 2022 2021 National Survey on Drug Use and Health Public Use File Codebook, Substance Abuse and Mental Health Services Administration, Rockville, MD
10. Dziak JJ, Coffman DL, Lanza ST, Li R, Jermiin LS 2020 Brief Bioinform. 21(2):553-565
11. Heo J, Lee JY, Kim W 2020 Commun. Stat. Appl. Methods 27:301-311
12. Al Hakmani R, Sheng Y 2023 Behaviormetrika 50:93–120