Hindawi
Advances in Civil Engineering
Volume 2023, Article ID 7116057, 31 pages
https://doi.org/10.1155/2023/7116057

Research Article

Identifying the Smoking and Smokeless Tobacco-Related Predictors on Frequencies of Heavy Vehicle Traffic Accidents in Bangladesh: Linear and Binary Logistic Regression-Based Approach

Md. Anwar Uddin,[1] **Mithun Debnath**,[2] **Sumit Roy**,[3] **Saima Adiba**,[1] **and Mohammad Mahbub Alam Talukder**[4]

[1]*Department of Civil Engineering, Military Institute of Science and Technology, Dhaka, Bangladesh*
[2]*Department of Civil Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh*
[3]*Department of Civil Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh*
[4]*Accident Research Institute (ARI), Bangladesh University of Engineering and Technology, Dhaka, Bangladesh*

Correspondence should be addressed to Md. Anwar Uddin; anwar066.buet13@gmail.com

Received 30 October 2022; Revised 8 January 2023; Accepted 10 January 2023; Published 24 January 2023

Academic Editor: Gen Li

Smoking is responsible for ninety percent of all premature deaths worldwide. Its prevalence is increasing in developing countries such as Bangladesh. Road traffic accidents (RTAs) have risen dramatically in recent years, with tobacco use accounting for 4–5 million fatalities each year. This trend will likely continue as more bus and truck drivers smoke in Bangladesh. Therefore, our study attempts to identify predictors that may be directly related to the frequency of RTAs and smoking. The study included 424 bus and truck drivers and ten key informant interviews (KIIs). Then, a linear regression (LR) analysis model was used to determine how various smoking-related predictors contribute to the frequency of accidents. Furthermore, a binary logistic regression (BLR) model was used to examine the likelihood of a driver being involved in an accident related to various smoking-related predictors. This study demonstrates a strong association between the incidence of accidents and the number of times a person smokes, smokes while driving, and uses smokeless tobacco (SLT) daily. The result has been taken from the second BLR model, which fits with the data more than the LR model. According to that model, a driver is more likely to be in an accident if the number of days per year that he smokes cigarettes increases and if he smokes while driving. Additionally, it stresses the need for more research to make a more accurate forecast.

1. Introduction

In developed nations, smoking is the leading preventable cause of death and disability. Moreover, smoking is injurious to health, especially cigarette smoking. According to Chen et al. [1], cigarette smoking is an addictive behavior causing different health, social, and behavioral problems. Despite these problems, developing countries' smoking rate is rising fast [2]. Currently, tobacco use results in 4–5 million fatalities worldwide, with 2 million deaths occurring in developed nations and 2–3 million in developing countries [3]. Among the fatalities, smoking-related motor vehicle crashes are highly positively correlated [4]. Furthermore, previous research found that smoking is associated with higher rates of motor accidents [5]. So, this study contributes to identifying the predictors for road traffic accidents (RTAs) regarding smoking and SLT use in a developing country for better traffic safety and drug control policies. Moreover,

a new modeling approach is incorporated to predict the likelihood of accidents due to smoking and smokeless tobacco, which was previously ignored.

The South-East Asia region has high tobacco use rates among adults and youth, with a complex consumption behavior [6]. However, only a tiny percentage of young adults quit smoking [3]. Because tobacco products are generally less expensive than cigarettes for almost all demographic categories, many teens and adults prefer them to cigarettes. As a result, smokeless and smoked tobacco products produced domestically are relatively prevalent among adults (more than 43.3% in Timor-Leste, 43.3% in Bangladesh, 34.6% in India, and 34.6% in Thailand) [6]. With such a high smoking rate and SLT usage, the impact of driving behavior on drivers is critical for this region, particularly Bangladesh, as driving is a strenuous activity that needs continual focus and rapid processing of a large amount of data to arrive at quick decisions and actions. The ability, speed, and quality of decision making, in which brain functions, judgment, preference, and choice are integrated to shape the final decision and quickly put it into action, are all greatly influenced by one's mental, emotional, and physical state [7]. So, the research on human behavior in traffic accidents has been extensive, particularly in recent years, as intelligent autos and automotive control technologies have advanced for developed countries [8]. However, there is minimal literature on the effects of smoking on decision making and the impact of traffic accidents in developing countries.

Most studies have demonstrated that human driving behavior increases the risk of an accident [9]. The number of RTAs has increased recently, and smoking rates among bus-truck drivers in Bangladesh are rising. A study by Talukder et al. [10] reveals that RTAs and smoking while driving are statistically significantly associated ($p = 0.003$). They also demonstrate that with a trust period ranging from 1.12 to 13.72, unlicensed drivers are around 3.8 times more likely to be in RTAs than licensed drivers. Other findings from this study show that drivers are hooked to smoking and that smoking habits, smoking knowledge, driving privileges, personal income, and education affect RTAs significantly. Another study by Goon and Bipasha [11] shows that bus drivers smoked 93% of the time on average and spent 20% of their daily salary on cigarettes in Bangladesh. Although most drivers (32.3%) began smoking before entering the driving profession, excessive smoking was encouraged by workplace and environmental stress brought on by a demanding work schedule. However, to the authors' knowledge, there is little research on the impact of smoking and the use of smoked and smokeless tobacco on road safety. However, Talukder et al. [10] tried to show the pattern of smoking and SLT use among drivers with RTAs. However, developing models in any studies have not identified associated smoking-related predictors and parameters. This study has also covered this gap for better RTA prediction and safety policy implication.

Our study attempted to identify predictors that may be directly related to the frequency of RTAs and smoking. In this regard, we investigated the impact of smoking and SLT on traffic accidents in Bangladesh by developing linear and binary logistic regression (BLR) models. This method is unique, and investigating traffic accidents from the standpoint of smoking has never been attempted in a developing country. The first linear regression (LR) model has been developed in this study to understand how various smoking and driving-related factors affect the frequency of road traffic accidents. Because we considered so many variables, it was our challenge to identify the most significant predictors for a clearer picture. The significance of various independent variables related to accidents and driving behaviors is examined in this analysis based on the first LR. Then, in the second regression analysis, only the significant independent variables are considered to obtain more specific predictors. However, while LR produces continuous results, logistic regression produces discrete results. So, if it is necessary to determine whether an accident occurred based on various independent variables, BLR is the best option. Furthermore, the significant independent variables are determined based on the BLR coefficients. Different independent variables were used in the first BLR, as in LR. However, only the significant variables were considered in the second binary logistic model. The BLR model is also used to determine the frequency of accidents in relation to these variables.

In the following part, we review several pieces of literature to find the relevancy of different smoking-related factors associated with RTAs. Section 3 establishes the appropriate methodology for developing models to predict the likelihood of smoking-related predictors on RTAs. Section 4 analyzes the developed models in depth to identify the predictors that directly impact road traffic fatalities. Section 5 then elaborates on the findings and discusses the proper traffic safety policies. Finally, in Section 6, we conclude this work.

2. Literature Review

According to an estimate from the World Health Organization (WHO) (2004) [12], road accidents result in up to 50 million injuries and 1.25 million fatalities annually and nearly 3400 road fatalities every day. However, not all countries experience the same traffic injuries. The likelihood of dying in a car accident varies by country of residence. For instance, nearly 90% of traffic fatalities occur in low and middle-income countries (LMICs). The rate is less than 9 in high-income countries (HICs) but about 20 in LMICs, with the most significant rate in Africa (26.6%). Moreover, less than 3 to over 40 people die for every 100,000 people worldwide, or what is known as the mortality rate. Although there have been improvements in road safety in HICs over the past few decades, there have been no improvements in LMICs. It is anticipated that there will be more traffic deaths in the upcoming years [13].

In Bangladesh, the number of traffic collisions on the roads is rising, which has been acknowledged by the people, government, and administration [14]. According to government statistics, there are more than 60 fatalities in road accidents for every 10,000 cars in Bangladesh. About eight people each day pass away in car accidents. The actual mortality rate is probably substantially greater. According to

Debnath et al. [15], there are still issues with the accident reporting system and the data it produces, and official numbers are prone to underreporting. Between 1982 and 2000, the number of accidents climbed by 43%, while the number of fatalities increased by over 400%, showing that accidents are happening more frequently but more severely [16].

As a higher rate of traffic accidents is observed worldwide as mentioned earlier, there is a need to investigate safety criteria concerning the incidents. Typically, a wide range of contributory elements, both from within the relevant organizations and dysfunctional interactions between them in a broader perspective, can be explored during investigating significant incidents in safety-critical systems [17]. Furthermore, the sociotechnical systems' perspective contends that accidents are not the result of a single, intentional act but rather highly interactive group processes that are influenced by decision and policymakers at all relevant societal levels [18].

Though many factors are responsible for road accidents, one of the most prevalent issues is the risky behavior of drivers [19]. More than 90% of drivers engage in unsafe driving activities in some way. Males ($p < 0.0001$), those who used alcohol or cannabis in adolescence ($p < 0.0001$), those involved in violent or property crimes ($p < 0.01$), and those connected to delinquent or substance-using peers ($p < 0.05$) were the groups most likely to demonstrate frequent dangerous driving behaviors. The likelihood of traffic accidents was strongly correlated ($p < 0.0001$) with the level of risky driving behavior. Young people frequently engage in risky driving habits, especially those prone to externalizing behaviors (substance abuse, crime, and affiliations with deviant peers). Driving at risk increases the likelihood of a traffic collision [20].

Although it is widely researched that smoking substances have disease consequences, there is not much documentation of the association of smoking with an increased risk of injury [21]. So, we will discuss some significant studies on smoking-related RTAs for our research. For instance, we can mention a study done by Saadat and Karbakhsh [22] showing that road traffic crashes (RTCs) had an annual incidence of 14.9%: 14.0% included motor vehicle collisions with other cars and 0.9% included pedestrians. In their univariable analysis, there was a correlation between the RTC and male gender, defensive driving technique (DDT), smoking cigarettes, smoking waterpipes, and driving maneuvers (DMs). Moreover, after adjusting for DDT, multivariable analysis from their study revealed a substantial correlation between RTC and cigarette and waterpipe smoking. They also developed a Poisson regression model and thus showed that smoking a waterpipe, smoking cigarettes, or both, and DDT were independent predictors of the frequency of traffic accidents.

Another study by Koushki and Bustan [23] showed that among young drivers who had never worn a seatbelt, smokers outnumbered nonsmokers by a factor of more than two. The same was true for involvement in traffic accidents: the number of accidents experienced by young drivers was higher for smokers and those who never buckled up. The causes of these may be related to drivers who disobey the seat belt requirement and may also disobey other traffic laws and regulations, which raises the risk of being involved in an accident. They also enlightened that the drivers' distraction typically was brought on by taking a cigarette out of the packet, lighting it, and holding it. At the same time, smoking impairs a driver's coordination, response time, physical mobility, and concentration, which are crucial for preventing accidents. Their study also found that the ignorance of the distractions that smoking causes and the amount of carbon monoxide in cigarette smoke were alarming. Moreover, different literature illustrates that deficient levels of carbon monoxide severely affect a driver's alertness, reaction time, and ability to judge distance and speed [23]. Nevertheless, Grout et al. [24] showed a potential link between the smoking habits of drivers involved in injury-producing traffic accidents and the hours of darkness. They suggested that drivers who smoke have a higher risk of being involved in an injury accident during the hours of darkness than drivers who do not smoke but are also involved in injury accidents. The study also revealed a statistically significant relationship between smoking and seat belt use, with smokers less likely to utilize seat belts than nonsmokers.

Another study by Buñuel Granados et al. [25] found that the risk of accidents is higher among single men under 45. According to their study, smokers are involved in accidents twice as often as nonsmokers. There are no statistically significant differences between smokers who do and do not smoke while driving a car; smoking makes it more likely that the driver will be in a car accident, even if he does not smoke while driving.

Finally, we will discuss another study by Tzortzi et al. [26] that showed the risk of accidents from the distraction of certain lifestyle choices, such as smoking and drinking while driving. This study focused on the factors associated with drivers' distractive behavior in Greece. Drivers' behavior differed depending on their age, gender, social class, and area of residence. For example, male drivers were more likely to engage in drunk driving. In contrast, professional drivers were likelier to use their cell phones for calls and texting, set the GPS, and smoke while driving. They also highlighted the relationship between smoking, alcohol, and road accidents.

In this paper, we collected data through questionnaire surveys of drivers of heavy vehicles (e.g., bus and truck) at six major bus and truck terminals in Dhaka. As our study mainly concerns the different predictors of RTAs of heavy vehicles, we focus on this study group for data collection. However, as discussed in different studies on the variables contributing to RTAs, we tried to find the gaps first. So, we have tried to contribute to filling these gaps for RTAs regarding smoking and tobacco use. Firstly, no significant studies on smoking and tobacco-related RTAs have been conducted in developing countries, especially Bangladesh. Secondly, smoking-related sociodemographic factors for RTAs have been conducted. However, proper predictors related to smoking and SLT have not been well developed. Thirdly, no proper methodology is available for the smoking-related predictors that can be modeled to show

precise results. So, our studies have tried to fill those gaps and helped develop the right statistical models, such as the linear and binary logistics models, to find important predictors of smoking and SLT use for safety and drug use policies in developing countries like Bangladesh.

3. Materials and Methods

3.1. Study Design, Area, and Period. For this study, cross-sectional data collection was conducted over nine months (March–December 2017) at six different locations in Dhaka, Bangladesh (bus terminals: Mohakhali, Gabtoli, and Jatrabari; truck terminals: Tejgaon, Aminbazar, and Doyagonjbazar). The areas were chosen because of their high traffic accidents, substantial intercity vehicle movements, high traffic volume, severe traffic congestion, and intermodal facilities. In addition, we have selected these study regions to understand what factors in Dhaka city led to RTAs explicitly.

3.2. Participants and Sample Selection. The poll and interviews included all heavy vehicle drivers (bus and truck drivers), administrators, and spokespeople from many institutions, such as bus and truck owner groups, labor unions, and transportation authorities. Proper statistical methods have been employed for data sampling, reducing the collected data's bias. From several bus and truck ports (bus drivers: 212 and truck drivers: 212), 424 samples were evenly chosen. Here, the sample population has been taken, maintaining equality. The target population is from different social statuses. We have approached the government and private organizations mentioned above and the target group (vehicle drivers) with proper authoritative permissions.

3.3. Data Collection. A semistructured questionnaire survey was used to provide a clear picture of the smoking and SLT use behaviors of heavy vehicle drivers—bus and truck drivers—with a total of 424 completed questionnaires. In addition, face-to-face interviews were used to collect the respondents' responses.

3.4. Analysis Approach. When the data collection was complete, a data entry operator was engaged to enter the information and was tasked with teaching everyone about using coded responses and multiple-choice questions. After obtaining the data file from the data entry team, another round of reviews was performed to ensure the data were entered correctly.

Here, LR and BLR models have been developed for different RTA predictors. LR analysis is typically used to predict a variable's value based on the value of other variables. The variable whose value needs to be predicted is called the "dependent variable," and the variables that are used to predict the value of the other variable are called "independent variables" [27]. If Y_i is a dependent variable and $x_1, x, \ldots..., x_n$ are independent variables, then an LR can be developed as follows:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \ldots + \beta_p X_{ij} + \varepsilon_i, \quad (1)$$

where for each observation, $i = 1, ..., n$. In the formula above, it is considered that there are n observations of one dependent variable and p independent variables. Thus, Y_i is the i^{th} observation of the dependent variable, X_{ij} is the i^{th} observation of the j^{th} independent variable, and $j = 1, 2, ..., p$. The values βj represent parameters to be estimated, and ε_i represent the i^{th} independent identically distributed normal error. B is a $(p+1)$-dimensional parameter vector, where β_0 is the intercept term.

However, the standard logistic function σ: $R \longrightarrow (0, 1)$ and the general logistic function P: $R \longrightarrow (0, 1)$ can be written as follows:

$$\sigma(Y_i) = P(x_i) = \frac{e^{y_i}}{1 + e^{y_i}} = \frac{e^{y_i}}{1 + e^{-y_i}}. \quad (2)$$

$P(x)$ in the logistic model is defined as the probability of the dependent variable Y_i being a success or case. Moreover, in the binary logistic model, the probability of the dependent variable Y_i equals success/failure or yes/no. Therefore, if yes/no is the case, P (YES) can be used as a dependent variable. For example, suppose the value of P more than 0.5 is taken as yes, and the value of P less than 0.5 is taken as no. In that case, the new dependent variable can be either yes or no, which will be categorical, and it can be named P (YES) [27].

Based on (1) and (2), LR and BLR models with categorical predictor coding were performed to evaluate the link between RTAs and driver smoking behavior. In the LR analysis, the relevance of the predictors was assessed concerning the 95% confidence interval. It is called variable selection. After that, Pareto analysis, multicollinearity detection, goodness-of-fit analyses, fits and diagnostics for unusual observations, analysis of variance (ANOVA), residual analysis, frequency versus residual plot analysis, versus order plot analysis, and normal probability plot analysis were performed. In the case of BLR analysis, significance of the predictor is analyzed by the Wald test. Besides that, goodness-of-fit analyses and odds ratio analyses have been performed. The analysis was conducted using the most recent version of the Statistical Package for Social Sciences (SPSS version 25.0; IBM Corp., Armonk, NY, USA) and MS Excel.

4. Results and Findings

This section provides a detailed modeling approach for identifying the predictors of RTAs relating to smoking and SLT use. Both LR and BLR models will be developed. However, thirteen predictors have been identified for the study purpose, where the response is AccidentFreq (frequency of accidents). They have been denoted in short forms in the LR and BLR models. Table 1 depicts the elaboration of the response AccidentFreq and the thirteen predictors used in both models.

Talukder et al.'s study [10] focused on smoking behavior and driver involvement in RTA in Bangladesh. A relation between these two phenomena was established by univariate, bivariate, and multivariate analysis. That study used these

Table 1: Elaboration of the response and different predictors for RTAs.

Ser. No.	Response/predictors	Elaboration	Value range
1	AccidentFreq	Frequency of accidents	
2	AvgCigPerDay	The average number of cigarettes smoked each day	Natural numbers
3	NosmokePerDrivHr	Number of smoking per hour of driving	Natural numbers
4	SmokeDurDrive	Status of whether a driver smokes while he is driving	Yes, no
5	Foursmokemore	Status of smoking four or more cigarettes per day	Yes, no
6	Age	Age of each driver	Natural numbers
7	DrivingHour	Number of hours taken to drive	Natural numbers
8	CigUseYr	Number of days per year that people smoke cigarettes	Natural numbers
9	NoSmokeDurDrive	Number of smoking while driving	Natural numbers
10	SmokeorTobaccoStatus	Status of smoking (yes or no)	Yes, no
11	SmokeDaily	Status of whether a driver smokes every day	Yes, no
12	TypeofSmoke	Number of the type of smoking a driver uses	Natural numbers
13	TypofSmklessTobc	Type of tobacco-free smoking	Gul, Jarda, Khoini, White-pata, no consumption
14	SmkLessTobDailyStat	Status of a driver's less-than-daily tobacco use	No consumption, daily, sometimes

thirteen variables in different analyses, which have been used as predictors in this linear and binary logistic regression-based approach. Besides, different studies on the relationship between RTA and smoking worked with different predictors mentioned in this study.

Pederson et al. [28] considered current cigarette smoking, driving exposure, and sociodemographic factors to find their relationship with motor vehicle collisions. Those factors include age, number of hours taken to drive, the status of smoking, indication of the number of days per year that people smoke cigarettes, the status of whether a driver smokes every day, and the status of a driver's less-than-daily tobacco use.

Choi et al.'s study [29] considered daily smokers classified into light, moderate, and heavy daily smokers. Furthermore, it says that the risk of unintentional injury increases monotonically with increasing levels of smoking. Compared with nonsmokers, former smokers (PRR, 1.30, 95% CI: 1.23 to 1.37), light daily smokers (PRR 1.34, 95% CI: 1.25–1.42), moderate daily smokers (PRR 1.47, 95% CI: 1.39–1.55), and heavy daily smokers (PRR 1.58, 95% CI: 1.42–1.75) had increased risk for unintentional injuries. In our study, we also considered the status of smoking four or more cigarettes per day (Fourormoresmoke) with the status of whether a driver smokes every day (SmokeDaily) to take a higher level of daily smoking into account like that study.

Koushki and Bustan [23] considered the number of smoking while driving as an important factor for RTA. Of course, smoking an increased number of cigarettes is bad for driving. Nevertheless, the study also mentioned that the distraction is caused by removing the cigarette from the package, lighting it, and holding it. At the same time, smoking affects drivers' concentration, coordination, reaction time, and physical maneuverability.

4.1. First Linear Regression Analysis. The LR analysis has been performed from the collected data, considering accident frequency, driving hours, and daily cigarette consumption. As a result, the LR equation (3) found by the categorical predictor coding (1, 0) is given below:

$$\begin{aligned} AccidentFreq = {} & 3.09 + 0.0346\,x_1 - 1.972\,x_2 + 0.0\,x_{3NO} + 0.871\,x_{3YES} \\ & + 0.0\,x_{4NO} + 5.11\,x_4 YES + 0.0018\,x_5 \\ & - 0.0158\,x_6 + 0.0065\,x_7 + 0.1332\,x_8 \\ & + 0.0x_{9NO} - 1.135x_{9YES} + 0.0x_{10} \\ & + 0.0x_{11Gul} - 0.80x_{11Jarda} - 3.38x_{11Khoini} - 1.42x_{11No}\,consumption \\ & - 1.29x_{11White} - pata + 0.0x_{12Daily} \\ & + 0.694x_{12No}\,consumption \\ & - 0.488x_{12}\_Sometimes \end{aligned}, \tag{3}$$

where x_1 = AvgCigPerDay, x_2 = NosmokePerDrivHr, x_3 = SmokeDurDrive, x_4 = Foursmokemore, x_5 = *Age*, x_6 = DrivingHour, x_7 = CigUseYr, x_8 = NoSmokeDurDrive, x_9 = SmokeorTobaccoStatus, x_{10} = SmokeDaily, x_{11} = TypofSmklessTobc, and x_{12} = SmkLessTobDailyStat.

Here coefficients and model summary of the LR and fits and diagnostics for unusual observations are presented in a tabular form in Table 2.

4.1.1. Variable Selection. The analysis that involves selecting significant predictors based on a p value threshold of 0.05 is typically called variable selection. In a variable process, the goal is to identify a subset of the predictor variables most strongly associated with the response variable while controlling for the false positive rate (the probability of rejecting the null hypothesis when it is true). A statistical test, such as the t-test or the F-test, can be used to calculate the p value for each predictor variable. The p value reflects the probability of obtaining the observed test statistic if the null hypothesis (that the predictor is not associated with the response) is true. Here, the coefficients show how the frequency of accidents changes when one predictor in a regression line changes while other predictors remain constant. In a linear regression model, the coefficients represent the change in the response variable for a one-unit change in the predictor variable, holding all other variables constant. A positive or negative coefficient value denotes an increase or decrease in the frequency of accidents when predictors change. Here SE Coef is the standard error (SE) of a coefficient. In a linear regression model, it is a measure of the variability of the coefficient estimate. It represents the standard deviation of the sampling distribution of the coefficient. It is used to construct confidence intervals for the coefficient, which estimates the range of values within which the true value of the coefficient is likely to fall. In an LR analysis, the test statistic, known as the t value, is a measure of the size of the difference between the estimated coefficient (also called the parameter estimate) and the hypothesized value of the coefficient (also called the null value). It is used to determine the statistical significance of the coefficient. A large t value indicates that the estimated coefficient is significantly different from the hypothesized value and suggests that the relationship between the predictor variable and the response variable is not due to chance.

Table 2: Coefficients and model summary of the linear regression and fits and diagnostics for unusual observations.

	Coef	Standard error (SE) coef	*T* value	*p* value	Variance inflation factor (VIF)
Coefficients					
Term					
Constant	3.09	1.53	2.03	0.043	
AvgCigPerDay	0.0346	0.0206	1.68	0.094	2.33
NosSmokePerDrivHr	−1.972	0.867	−2.27	0.024	16.00
SmokeDurDrive					
YES	0.871	0.341	2.56	0.011	2.33
Foursmokeormore					
YES	5.11	2.03	2.52	0.012	1.60
Age	0.0018	0.0141	0.13	0.898	1.54
DrivingHour	−0.0158	0.0389	−0.41	0.684	1.34
CigUseYr	0.0065	0.0201	0.32	0.748	2.85
NosSmokeDurDrive	0.1332	0.0756	1.76	0.079	16.18
SmokeorTobaccoStatus					
YES	−1.135	0.731	−1.55	0.121	8.08
SmokeDaily					
Not at all	−0.605	0.788	−0.77	0.443	10.55
TypofSmklessTobc					
Jarda	−0.80	1.06	−0.75	0.456	7.66
Khoini	−3.38	2.13	−1.58	0.114	1.77
No consumption	−1.42	1.17	−1.22	0.225	13.32
White-pata	−1.29	1.18	−1.09	0.274	3.16
SmkLessTobDailyStat					
No consumption	0.694	0.955	0.73	0.467	8.91
Sometimes	−0.488	0.764	−0.64	0.523	1.44
Model summary	S	R-sq	R-sq (adj)	R-sq (pred)	
	2.26280	8.74%	5.15%	≤0.001%	

Table 2: Continued.

Fits and diagnostics for unusual observations	Obs	AccidentFreq	Fit	Residuals	Std residuals	Type of residuals	
	1	2.00	−1.02	3.02	2.36	R	X
	20	2.00	1.05	0.95	0.46		X
	34	9.00	1.57	7.43	3.31	R	
	36	12.00	8.98	3.02	2.36	R	X
	37	2.00	1.62	0.38	0.19		X
	44	9.00	1.96	7.04	3.13	R	
	54	7.00	2.44	4.56	2.04	R	
	55	9.00	2.09	6.91	3.09	R	
	60	9.00	1.36	7.64	3.42	R	
	68	12.00	1.45	10.55	4.70	R	
	72	≤0.001	1.15	−1.15	−0.55		X
	74	1.00	2.02	−1.02	−0.49		X
	94	≤0.001	0.66	−0.66	−0.33		X
	97	≤0.001	0.71	−0.71	−0.36		X
	100	≤0.001	1.44	−1.44	−0.71		X
	102	9.00	2.33	6.67	2.99	R	
	104	9.00	1.23	7.77	3.48	R	
	109	≤0.001	1.09	−1.09	−0.54		X
	121	7.00	3.90	3.10	1.62		X
	124	12.00	2.63	9.37	4.20	R	
	133	10.00	1.71	8.29	3.69	R	
	140	7.00	1.71	5.29	2.41	R	
	141	3.00	2.01	0.99	0.51		X
	150	2.00	5.02	−3.02	−2.36	R	X
	151	1.00	1.19	−0.19	−0.09		X
	169	2.00	3.50	−1.50	−0.75		X
	191	≤0.001	1.79	−1.79	−0.85		X
	205	1.00	1.44	−0.44	−0.21		X
	216	12.00	2.09	9.91	4.42	R	
	217	8.00	1.63	6.37	2.83	R	
	218	10.00	1.63	8.37	3.72	R	
	236	2.00	1.25	0.75	0.35		X
	242	8.00	1.95	6.05	2.71	R	
	246	1.00	2.02	−1.02	−0.52		X
	267	≤0.001	0.22	−0.22	−0.10		X
	279	≤0.001	−0.06	0.06	0.03		X
	284	3.00	3.63	−0.63	−0.31		X
	289	≤0.001	−0.63	0.63	0.31		X
	294	5.00	1.09	3.91	1.87		X
	295	9.00	1.68	7.32	3.25	R	
	298	2.00	2.86	−0.86	−0.43		X
	301	3.00	1.05	1.95	0.96		X
	308	7.00	2.35	4.65	2.07	R	
	313	1.00	2.37	−1.37	−0.70		X
	320	1.00	3.01	−2.01	−1.01		X
	332	6.00	3.11	2.89	1.43		X
	341	≤0.001	0.74	−0.74	−0.36		X
	359	≤0.001	−0.05	0.05	0.02		X
	370	2.00	1.04	0.96	0.47		X
	379	12.00	2.67	9.33	4.17	R	
	380	9.00	1.69	7.31	3.25	R	
	388	1.00	1.59	−0.59	−0.28		X
	406	9.00	1.93	7.07	3.20	R	
	408	7.00	2.03	4.97	2.22	R	
	411	10.00	3.20	6.80	3.08	R	

R = large std residuals; X = unusual std residuals.

On the other hand, a small t value indicates that the estimated coefficient is not significantly different from the hypothesized value and suggests that the relationship between the predictor and response variables may be due to chance. The t value is usually compared to a critical value from the t distribution, a statistical distribution used to determine the probability of observing a given t value by chance. In an LR analysis, the p value measures the statistical significance of the estimated coefficients (also called the parameters). It represents the probability of observing the estimate by chance, given that the null hypothesis is true. It is calculated based on the t value of the estimate. It is used to determine the statistical significance of the coefficient. A small p value (usually less than 0.05) indicates that the estimated coefficient is significantly different from the hypothesized value and suggests that the relationship between the predictor variable and the response variable is not due to chance. On the other hand, a significant p value (greater than 0.05) indicates that the estimated coefficient is not significantly different from the hypothesized value and suggests that the relationship between the predictor and response variables may be due to chance. This study indicates a predictor's relevance in causing RTA by its p value.

In Table 2, the p values of NosSmokePerDrivHr, SmokeDurDrive (YES), and Foursmokeormore (YES) are lower than 0.05, so they are statistically significant predictors with a 5% confidence interval. It means an increase in smoking per driving hour, smoking during driving, and smoking more than four cigarettes increases the accident frequency. Other predictors are not significant.

4.1.2. Multicollinearity Detection. The variance inflation factor (VIF) measures the amount of multicollinearity in the model. Multicollinearity occurs when two or more highly correlated predictor variables affect the regression model's accuracy and interpretability. For example, it can cause the standard errors of the coefficients to be inflated. A VIF of 1 indicates no multicollinearity between the model's predictor and the other variables. A VIF greater than 1 suggests multicollinearity between the model's predictor and the other variables. The magnitude of the VIF reflects the severity of the multicollinearity. Multicollinearity can cause problems in a linear regression model, such as unstable coefficient estimates and inaccurate statistical inferences. Suppose the VIF for a predictor variable is found to be high (usually greater than 5 or 10). Removing the variable from the model may be necessary or using a different statistical model may be necessary. The regression analysis found significant predictors, though the model did not fit well. Because it calculates how much the variance of a regression coefficient is inflated because of multicollinearity in the model, the variance inflation factor, or VIF, is present and has a negative impact on the regression findings. When the VIF is greater than 10, there is a significant association and grounds for concern. Numerous predictors with VIFs above ten are observed in the regression model; NosSmokePerDrivHr is one of these and has a variance inflation factor of 16. Additionally, because the regression model could not adequately match the data, a new regression model was created using just two significant predictors with a variance inflation factor of less than 10.

4.1.3. Goodness-of-Fit Analyses. The analyses that consider the R-squared (R-sq), the adjusted R-squared (R-sq (adj)), and the predicted R-squared (R-sq (pred)) in a multiple linear regression model are typically referred to as model fit analyses or goodness-of-fit analyses. All three measures, the R-squared, the adjusted R-squared, and the predicted R-squared, can be used to evaluate the fit of a multiple linear regression model and to compare the fit of different models. In an LR analysis, the R-squared (R-sq) value measures the model's goodness of fit. It represents the proportion of the variance in the response variable explained by the predictor variables. The R-sq value is calculated as the ratio of the sum of squares explained by the model (ESS) to the total sum of squares (TSS). The TSS is the sum of the squared differences between the mean of the response variable and each data point. It reflects the total variability in the response variable. The ESS is the sum of the squared differences between the predicted values of the response variable and the mean of the response variable. It reflects the variability in the response variable that the model explains. The adjusted R-squared or R-sq (adj) is a modified version of the R-squared (R-sq) value used to evaluate an LR model's fit. The R-sq (adj) value considers the number of predictor variables in the model. It adjusts for the fact that the R-sq value tends to increase as the number of predictors increases, even if the additional predictors do not improve the model fit.

The predicted R-squared or R-sq (pred) measures the predictive accuracy of a linear regression model. It represents the proportion of the variance in the response variable explained by the predictor variables in a new dataset. The R-sq (pred) value is calculated using a prediction equation derived from the model. It is based on the predicted values of the response variable for the new dataset. The R-sq, R-sq (adj), and R-sq (pred) value range from 0 to 1, and a higher value indicates a better fit of the model (for R-sq (pred), it indicates a better fit of the model to the new dataset). Their value of 0 indicates that the model does not explain any of the variances in the response variable. In contrast, their value of 1 indicates that the model explains all of the variances in the response variable.

In this LR model, Table 2 shows that the R-squared value is 8.74%, the adjusted R-squared value is 5.15%, and the predicted R-squared value is 0%. These data indicate that the model is not a good fit for the data, as they are extremely low values. A low R-squared value (adjusted or not) suggests that the model cannot explain a large proportion of the variance in the dependent variable. Moreover, a low predicted R-squared value indicates that the model is not expected to perform well in future observations, including predicting any variance in response.

4.1.4. Fits and Diagnostics for Unusual Observations. Fits and diagnostics for unusual observations in multiple linear regression analysis are used to identify any unusual

observations that may have a large impact on the model. These observations are identified by their fitted values, standard errors, confidence intervals, and standardized residuals. It is expected that there will be some unusual observations. For example, based on the criteria for large residuals, roughly 5% of observations would be expected to be flagged as having a large residual. In this paper, they have been identified by residuals and standardized residuals. In multiple linear regression analysis, the residuals (also called "resid") and standardized residuals (also called "Std Resid") are used to identify unusual observations and assess the fit of the model. Residuals differ between the observed and the dependent variables' predicted values. Standardized residuals have been standardized to have a mean of zero and a standard deviation of one. Both residuals and standardized residuals can be used to identify unusual observations in multiple linear regression analyses. Unusual observations often stand out as points far from the trend in residual and leverage plots. Residuals and standardized residuals can also assess the model's fit. The model is likely a good fit if the residuals are randomly distributed around zero. However, suppose the residuals are systematically skewed or show patterns. In that case, it may indicate that the model is not a good fit and that alternative models should be considered. Overall, residuals and standardized residuals are useful tools for identifying and dealing with unusual observations in multiple linear regression analysis and assessing the model's fit. Examining the residuals can provide useful information about how well the model fits the data. Generally, the residuals should be randomly distributed with no obvious patterns and exceptional values. Standardized residuals greater and less than two are usually considered large.

In the fits and diagnostics for unusual observations in Table 2, all the values of Std Resid are either large or unusual. Suppose every value of the standardized residuals (Std Resid) is large or unusual. In that case, it may indicate that the model is not a good fit for the data. Standardized residuals are calculated by dividing the residuals (i.e., the difference between the observed and predicted values) by the standard deviation of the residuals. Unusual observations differ significantly from the rest of the data and may significantly influence the model. Unusual observations often stand out as points with large, standardized residuals, as large, standardized residuals indicate that the residuals are significantly larger or smaller than expected if the model were a good fit.

4.1.5. Analysis of Variance. In an LR analysis, the analysis of variance (ANOVA) is a statistical test used to determine whether the LR model is a good fit for the data. The ANOVA test compares the variance of the residuals (the residual sum of squares or RSS) to the variance of the response variable (the total sum of squares or TSS). It calculates a test statistic known as the F-statistic. This test is based on the null and alternative hypotheses. The null hypothesis means that the LR model does not explain variances in the response variable. The alternative hypothesis means that the LR model explains some or all of the variance in the response variable. The adjusted sum of squares (Adj SS) and the adjusted mean squared error (Adj MSE) are measures of the variability in the data that the model explains. They have adjusted versions of the sum of squares (SS) and the mean squared error (MSE), which account for the number of parameters estimated in the model. The Adj SS and Adj MSE are used in the ANOVA table to compare the explained variance (Adj SS) to the residual variance (Adj MSE). The explained variance represents the variability in the data explained by the model. In contrast, the residual variance represents the variability in the data that the model does not explain. The Adj SS and Adj MSE are used to calculate the F-statistic, a measure of the statistical significance of the model. The F-statistic is calculated as F = Adj SS/Adj MSE.

The F-statistic follows an F-distribution with p and $n - p - 1$ degrees of freedom, where p is the number of parameters or degrees of freedom estimated in the model and n is the sample size. The p value of the ANOVA test is calculated as the probability of observing an F-statistic as large or larger than the observed F-statistic by chance, given that the null hypothesis is true. A small p value (usually less than 0.05) indicates that the null hypothesis can be rejected, and it can be concluded that the model is a good fit for the data. On the other hand, a large p value (greater than 0.05) indicates that the null hypothesis cannot be rejected. The error term represents the variability in the data that the model does not explain. It is also known as the residual variance or the within-group variance. The lack-of-fit term represents the variability in the data that is not explained by the model but is not due to random error. The pure error (error variance or the within-group variance) is the data's variability due to experimental error, such as measurement error or variation in the experimental conditions. The total variance, representing the total variability in the data, is the sum of the explained variance (between-group variance), the pure error, and the lack-of-fit variance. The sum of squares (SS) is a measure of the total variance, which is calculated as the sum of the squared differences between the observed data values and the mean of the data.

Here, Table 3 depicts the analysis of the variance of the different predictors to identify the most significant ones. However, SmokeDurDrive (p value = 0.011), Foursmokeormore (p value = 0.012), and NosSmokePerDrivHr (p value = 0.024) all have p values less than 0.05, making them statistically significant with a 95% confidence interval. The most important predictor is smoking while driving, followed by smoking more than four times per day and the number of cigarettes smoked per hour of driving. This order also reflects the result from the statistical model selection.

4.1.6. Pareto Analysis. A Pareto chart of the standardized effects is a graphical representation of the standardized regression coefficients in a multiple linear regression model. It is used to visualize the relative importance of the predictor variables in their effect on the response variable. In a Pareto chart of the standardized effects, the predictor variables are plotted in descending order of their standardized coefficients, with the most important predictor at the top. This allows us to identify the most influential predictor variables

Table 3: Analysis of variance.

Term	DF	Adjusted sum of square (SS)	Adjusted mean square (MS)	F value	p value
Regression	16	199.55	12.4718	2.44	0.002
AvgCigPerDay	1	14.47	14.4663	2.83	0.094
NosSmokePerDrivHr	1	26.46	26.4591	5.17	0.024
SmokeDurDrive	1	33.52	33.5174	6.55	0.011
Foursmokeormore	1	32.54	32.5396	6.36	0.012
Age	1	0.08	0.0842	0.02	0.898
DrivingHour	1	0.85	0.8467	0.17	0.684
CigUseYr	1	0.53	0.5294	0.10	0.748
NosSmokeDurDrive	1	15.87	15.8686	3.10	0.079
SmokeorTobaccoStatus	1	12.35	12.3541	2.41	0.121
SmokeDaily	1	3.01	3.0136	0.59	0.443
TypofSmklessTobc	4	18.05	4.5132	0.88	0.475
SmkLessTobDailyStat	2	7.30	3.6493	0.71	0.491
Error	407	2083.95	5.1203		
Lack of fit	371	1954.45	5.2681	1.46	0.081
Pure error	36	129.50	3.5972		
Total	423	2283.50			

and compare their relative importance. The analysis of a Pareto chart of the standardized effects in a multiple linear regression model is typically referred to as a Pareto analysis or a Pareto ranking.

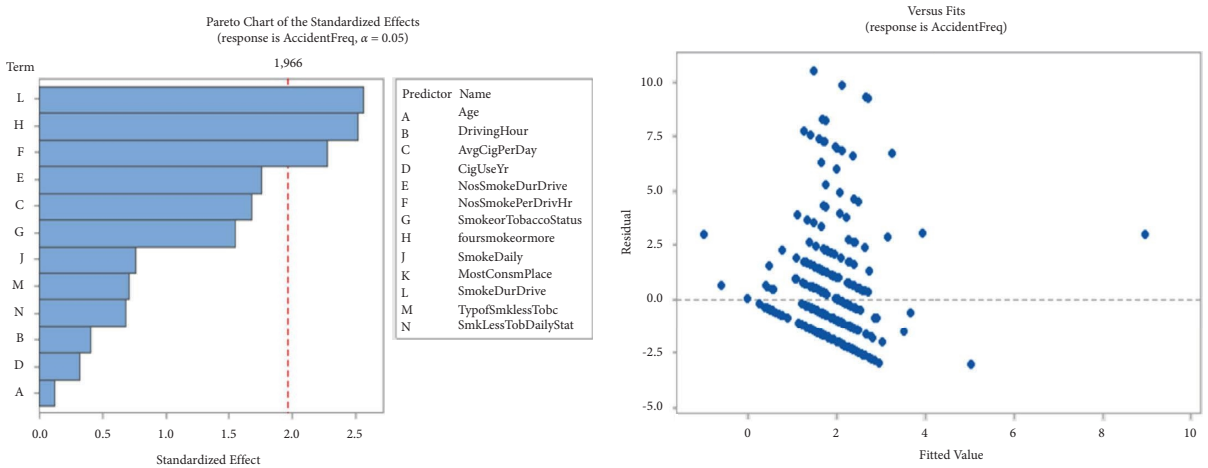
Now standardized effects of the LR model are presented in a Pareto chart in Figure 1(a). From the most significant to the most negligible effect, the Pareto chart displays the absolute values of the standardized effects. The figure also depicts a reference line showing statistically significant impacts. The significance level, which in this instance is 5 in this case, determines the reference line for statistical significance, which is 1.966. Only driving while smoking, smoking more than four cigarettes, and the number of cigarettes smoked per driving hour are shown in this chart as statistically significant. Driver's smoking while driving is the most important of all three predictors, while smoking per driving hour is the least important.

4.1.7. Residual Analysis. A residual versus fitted value plot analysis is typically referred to as residual analysis. Residual analysis is a statistical method used to evaluate the fit of a linear regression model by examining the residuals (i.e., the observed minus predicted values). A residual versus fitted value plot is a graphical representation of the residuals in a linear regression model. It is useful for evaluating the model's fit and identifying patterns or trends in the residuals. A residual versus fitted value plot is created by plotting the residuals on the y-axis and the predicted values or the predictor variables on the x-axis. It is a helpful tool for evaluating the fit of a linear regression model because it allows one to visualize the residuals and identify any patterns or trends in the data. For example, suppose the residuals are randomly dispersed around the horizontal line at zero. In that case, it indicates that the model fits the data well. However, suppose the residuals show a pattern or trend. In that case, it may indicate that the model is not a good fit and that additional predictors or a different model may be needed.

When examining a residual versus fitted value plot, there are mainly three characteristics to look for: randomness, constant variance, and normality. First, the residuals should be randomly dispersed around the horizontal line at zero. If the residuals show a pattern or trend, it may indicate that the model is not a good fit. Second, the spread of the residuals should be roughly constant over the range of the predicted values. For example, suppose the variance of the residuals increases or decreases as the predicted values increase or decrease. In that case, it may indicate that the model is not a good fit. Third, the residuals should be approximately normally distributed. If the residuals are not normally distributed, it may indicate that the model is not a good fit or that the data have been transformed incorrectly.

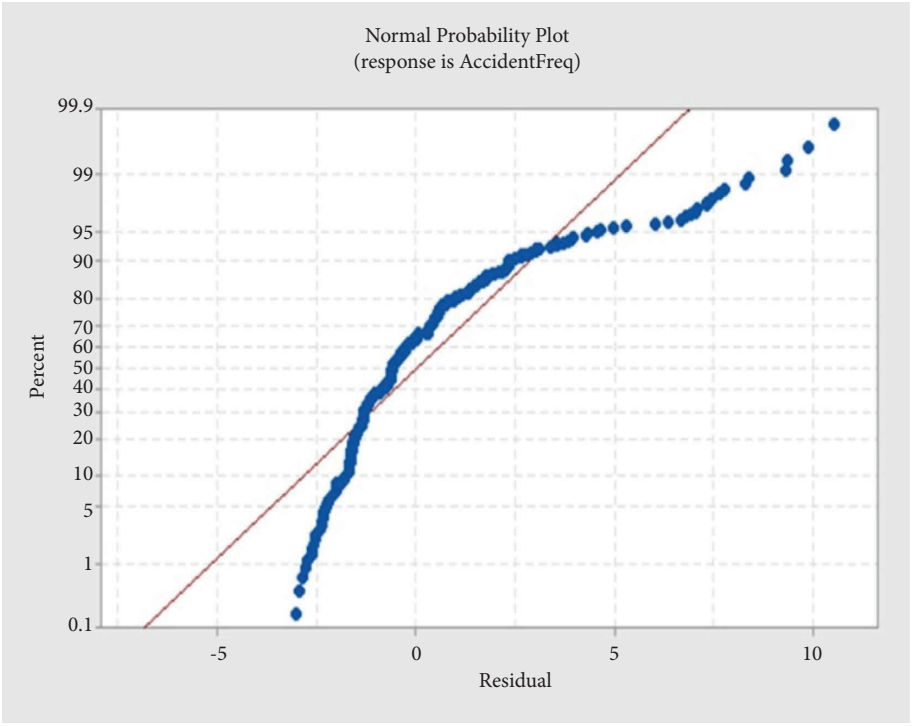
The regression model's "against fits" or "residuals versus fits" plot is displayed in Figure 1(b). Firstly, A pattern is found in the plot, so there is no randomness. Secondly, the plot has no constant variance, as the variance in the plot is not constant over the range of the predicted value. Moreover, thirdly, from visual inspection, it can be said that the plot does not follow a normal distribution, and it is skewed to the right. So, it can be said that the model is not a good fit. The details of the normality would be found in the frequency versus residual plot from Figure 2(b).

4.1.8. Normal Probability Plot Analysis. A normal probability plot is a graphical technique used to assess the normality of the residuals in a linear regression model. It is based on the idea that if the residuals are normally distributed, the plot of the residuals against the quantiles of a normal distribution should be linear. A normal probability plot is a useful tool for assessing the assumptions of a linear regression model and ensuring that the model is appropriate for the data.



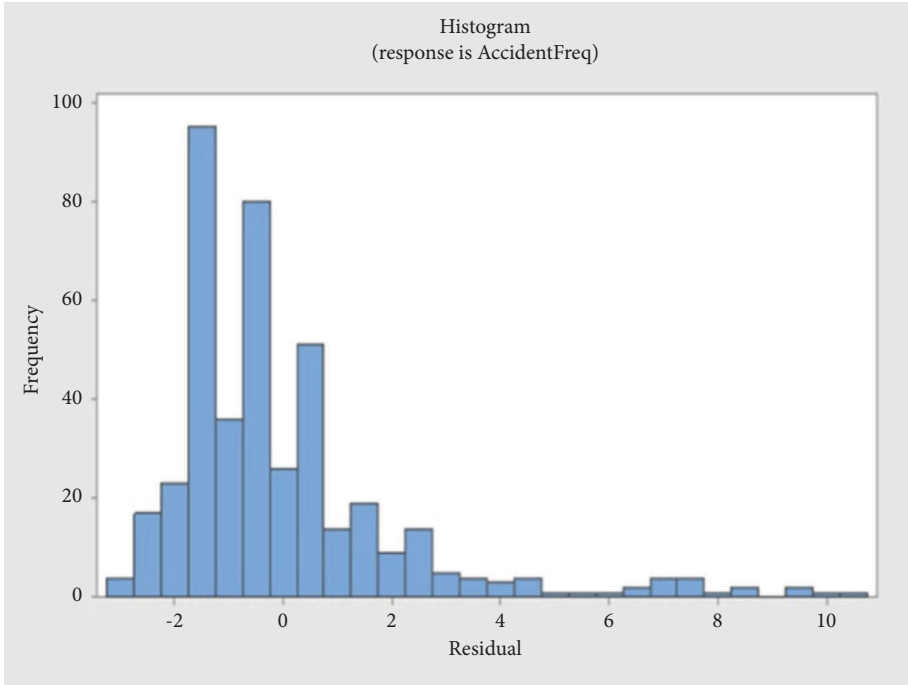
(a) (b)

Figure 1: (a) Pareto chart of the standardized effects and (b) versus fits plot of the first linear regression model.

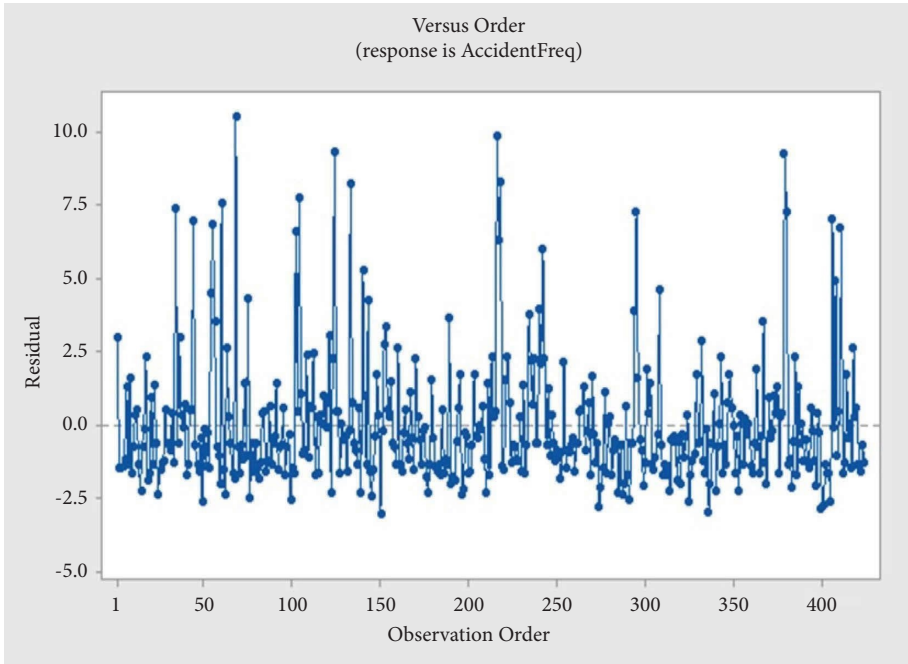


(a)

Figure 2: Continued.



(b)



(c)

Figure 2: (a) Normal probability plot, (b) histogram of frequency versus residual for accident frequency, and (c) versus order plot.

If the plot is linear, likely, the residuals are normally distributed. However, if the plot is not linear, it indicates that the residuals are not normally distributed. This may indicate a problem with the fit of the linear regression model, and using a different model or transforming the data might need to be considered. For example, Figure 2(a) shows that the plot is not linear, so the model does not fit well, and the residuals are not normally distributed.

4.1.9. Frequency versus Residual Plot. A frequency versus residual plot is a graphical representation of the residuals in a linear regression model. It is a type of residual plot that displays the frequency (i.e., the number of observations) of the residuals on the y-axis and the residual values on the x-axis.

Frequency versus residual plot is useful for evaluating the fit of a linear regression model and identifying patterns or trends in the residuals. For example, suppose the residuals

are randomly dispersed around the horizontal line at zero. In that case, it indicates that the model fits the data well. However, suppose the residuals show a pattern or trend. In that case, it may indicate that the model is not a good fit and that additional predictors or a different model may be needed.

Figure 2(b) illustrates how the data are skewed to the right. For example, suppose a frequency versus residual plot is unimodal and right-skewed. In that case, the residuals are concentrated around a single mode (i.e., a peak in the distribution). As a result, the distribution is skewed to the right, with a long tail extending towards higher residual values. This type of distribution may indicate that the model is not a good fit for the data and that additional predictors or a different model may be needed.

4.1.10. Versus Order Plot. In multiple linear regression analysis, a versus order plot can be used to visualize the relationship between the residuals (the differences between the observed and predicted values of the response variable) and the observation order (the order in which the data points were collected). A versus order plot of the residuals versus observation order can be useful for identifying patterns or trends in the residuals that indicate a problem with the fit of the multiple linear regression model. For example, suppose the residuals show a systematic pattern (e.g., increasing or decreasing over the observation order). This could indicate that the model is not accurately capturing the relationship between the predictor and response variables. In this case, adding additional predictor variables to the model or using a different model should be considered.

A versus order plot of residuals can be useful for determining the goodness of fit of a statistical model. A good fit is typically characterized by randomly distributed residuals and approximately constant variance. If the residuals show a pattern or trend, this may indicate that the model is not a good fit or that there are unusual points in the data. There are mainly three characteristics of a versus order plot of residuals that can be used to determine the goodness of fit of a statistical model: randomness, constant variance, and normal distribution. First, suppose the residuals are randomly distributed across the plot. In that case, this suggests that the model is a good fit for the data.

On the other hand, if the residuals show a pattern or trend, this may indicate that the model is not a good fit. Second, suppose the residuals' variance is approximately constant across the plot. In that case, this suggests that the model fits the data well. If the variance of the residuals increases or decreases as a function of the predictor variable, this may indicate that the model is not a good fit. Third, if the residuals are approximately normally distributed, this suggests that the model is a good fit for the data. If the residuals are heavily skewed or have heavy tails, this may indicate that the model is not a good fit.

Figure 2(c) shows the versus order plot of this LR model. Firstly, the data do not seem random, rather a pattern can be found here as residuals are increasing and decreasing with increasing observation order. Secondly, the residuals' variance is far from constant across the plot, as there are many spikes in the residuals. Moreover, thirdly, the plot is not normally distributed as the points are not symmetrical around the center of the plot, and there are many extreme outliers. So, the model is not a good fit.

4.1.11. Major Findings from the First LR Analysis. All of the diagnostic tools for assessing the fit of this LR indicate that the model is not a good fit. So, as our first option, it can be helpful to remodel the multiple linear regression model with only the predictors whose p values are less than 0.05 and whose variance inflation factor (VIF) is less than 10. This can help improve the model by reducing the number of variables and eliminating variables that are not significant or correlated with each other.

p values are used to assess the statistical significance of the coefficients in the multiple linear regression model. A p value less than 0.05 indicates that the coefficient is significantly different from zero and is, therefore, likely to be related to the dependent variable. Remodeling the model with only the predictors whose p values are less than 0.05 can help to reduce the number of variables and improve the model's ability to predict the dependent variable. The variance inflation factor (VIF) measures the multicollinearity among the predictors in the model. Multicollinearity occurs when two or more predictors are highly correlated with each other. High multicollinearity can lead to unstable coefficients and make it difficult to interpret the model's results. Remodeling the model with only the predictors whose VIF is less than ten can help to reduce multicollinearity and improve the stability and interpretability of the model. Retaining the multiple linear regression model with only the predictors whose p values are less than 0.05 and whose VIF is less than ten can improve the model.

According to Table 2, the p values of NosSmokePerDrivHr, SmokeDurDrive (YES), and Foursmokeormore (YES) are lower than 0.05. However, numerous predictors with VIFs above ten are observed in the regression model. NosSmokePerDrivHr is one of these and has a variance inflation factor of 16. So, remodeling with only two predictors, SmokeDurDrive (YES) and Foursmokeormore (YES), should be done in the second LR analysis.

4.2. Second Linear Regression Analysis. In this section, a new regression analysis (4) has been performed considering two predictors, as mentioned earlier: Foursmokeormore and SmokeDurDrive.

$$\text{AccidentFreq} = 1.458 + 0.0\,x_{1\_}NO + 4.89\,x_{1\_}YES + 0.0\,x_{2\_}NO + 0.652\,x_{2\text{YES}}, \tag{4}$$

where x_1 = Foursmokeormore and x_2 = SmokeDurDrive.

4.2.1. Variable Selection. Regression analysis of AccidentFreq versus Foursmokeormore and SmokeDurDrive is shown in Table 4, where both variables had p values lower than 0.05 (even lower than 0.01). Therefore, smoking more than four times a day and smoking during driving are important predictors, with the same reference line of 1.966.

4.2.2. Goodness-of-Fit Analyses. In this LR model, Table 4 shows that the R-squared value is 4.33%, the adjusted R-squared value is 3.87%, and the predicted R-squared value is 0%. These data also indicate that the model is not a good fit for the data, as they are extremely low values. So, this model also fails to explain a large proportion of the variance in the dependent variable and is not expected to perform well in future observations, including predicting any variance in response.

4.2.3. Fits and Diagnostics for Unusual Observations. In the fits and diagnostics for unusual observations in Table 4, all the values of Std Resid are large, and two are unusual. Suppose every value of the standardized residuals (Std Resid) is large in fits and diagnostics for unusual observations. In that case, likely, the model does not fit the data well. This could be because of an inappropriate model form, wrong variable selection, or other problems.

4.2.4. Prediction for Response. In the prediction for response (AccidentFreq) in Table 4, the prediction interval (PI) of Foursmokeormore and SmokeDurDrive predictors in different settings has been analyzed. It also indicates if the model is a good fit and if there are unusual observations and outliers. Here "NO" means the predictor is not considered, and "YES" means the predictor is present in the equation, which would play a role in changing the value of the response (AccidentFreq). A prediction interval (PI) is a range of values likely to contain an individual predicted value with a certain confidence level (in this model, the confidence level is 95%). Generally, a prediction interval should contain the predicted value with a certain confidence level, regardless of whether the predicted value is positive or negative. It is noteworthy that if both the upper and lower limits of the PI have the same sign, it may indicate that the prediction interval is unusual because it does not contain the predicted value with the specified confidence level. There are several possible reasons why both the upper and lower limits of the PI may have the same sign, like the model not being a good fit for the data, unusual observations, or outliers.

According to the prediction for AccidentFreq results, if the setting of Foursmokeormore is YES (for both the settings of SmokeDurDrive), then the PI becomes unusual, as both the extremes of the PI are of the same sign. On the other hand, if the setting of Foursmokeormore is No, then good PIs are found, i.e., if the setting of SmokeDurDrive is YES. The PI is (−2.38080, 6.60045), and if the setting of SmokeDurDrive is NO, then the PI is (−3.02885, 5.94452). Turning the setting of SmokeDurDrive from YES to No makes the PI narrower, meaning that the predictor SmokeDurDrive is a significant factor in predicting AccidentFreq. However, as two unusual PIs are found in the analysis, the model is not a good fit for the data, unusual observations, or outliers.

4.2.5. Pareto Analysis. Figure 3(a) displays the Pareto chart of the standardized effects of the new regression model, where both Foursmokeormore and SmokeDurDrive exhibit strong significance.

4.2.6. Residual Analysis. The regression model's "against fits" or "residuals versus fits" plot is displayed in Figure 3(b). The goodness of fit is poor in this model, so the data do not match the regression line. Firstly, a pattern is found in the plot, two perpendicular lines on the X-axis, so there is no randomness. Secondly, the plot has no constant variance (between the two lines). Thirdly, from visual inspection, the plot does not follow a normal distribution, which is proved in Figures 4(a) and 4(b).

4.2.7. Analysis of Variance. The analysis of variance was done. Foursmokeormore and SmokeDurDrive show the exact p values in the variable selection, proving their high significance. Table 5 depicts the variance of accident frequencies related to two predictors (Foursmokeormore and SmokeDurDrive). For these predictors, the p value is less than 0.005, indicating that all of them are statistically significant for 95% confidence intervals.

4.2.8. Normal Probability Plot Analysis. Figure 4(a) also shows that the plot is not linear. So, the model does not fit well, and the residuals are not normally distributed. In that case, this model does not fit well, and using a different model or transforming the data might need to be considered.

4.2.9. Frequency versus Residual Plot. Figure 4(b) illustrates that this model's data are skewed to the right. So, the residuals are concentrated around a single mode (i.e., a peak in the distribution). The distribution is skewed to the right, with a long tail extending towards higher residual values. Of course, the residuals are not normally distributed. Moreover, it means the model is not a good fit for the data and that additional predictors or a different model may be needed.

4.2.10. Versus Order Plot. Figure 4(c) shows the versus order plot of this LR model. It seems quite the same as Figure 2(c) of the first LR model. Firstly, the data do not seem to be random, rather a pattern can be found here as residuals are increasing and decreasing with increasing observation order. Secondly, the residuals' variance is far from constant across the plot. There are many spikes in the residuals, and the variance seems even more variable than in the previous model. Furthermore, thirdly, the plot is not normally distributed as the points are not symmetrical around the center

Table 4: Coefficients and the model summary of the regression, fits, and diagnostics for unusual observations and prediction for response.

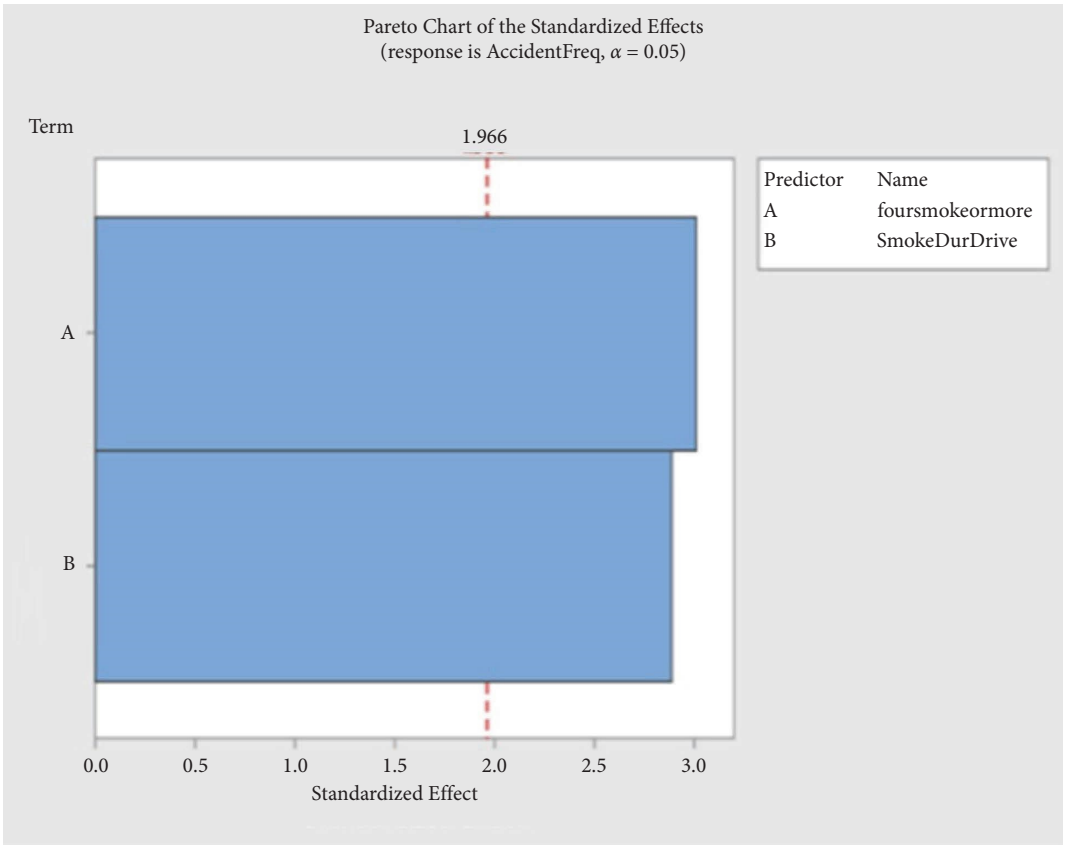
	Term	SE coef	T value	p value	VIF
Coefficients	Constant	0.144	10.10	≤0.001	
	Foursmokeormore YES	1.62	3.02	0.003	1.01
	SmokeDurDrive YES	0.225	2.89	0.004	1.01

Model summary	R-sq	R-sq (adj)	R-sq (pred)
	4.33%	3.87%	≤0.001%

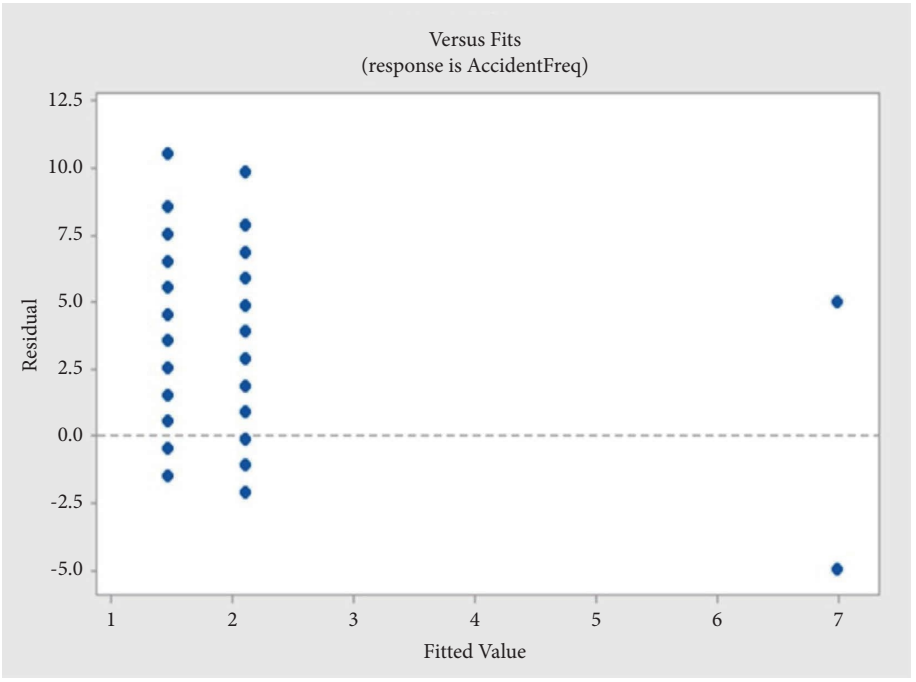
Fits and diagnostics for unusual observations	Obs	AccidentFreq	Fit	Residuals	Std residuals	Type of residuals
	34	9.000	1.458	7.542	3.32	R
	36	12.000	7.000	5.000	3.10	X
	44	9.000	2.110	6.890	3.03	R
	54	7.000	2.110	4.890	2.15	R
	55	9.000	2.110	6.890	3.03	R
	60	9.000	1.458	7.542	3.32	R
	68	12.000	1.458	10.542	4.64	R
	102	9.000	2.110	6.890	3.03	R
	104	9.000	1.458	7.542	3.32	R
	121	7.000	2.110	4.890	2.15	R
	124	12.000	2.110	9.890	4.35	R
	133	10.000	1.458	8.542	3.76	R
	140	7.000	1.458	5.542	2.44	R
	150	2.000	7.000	−5.000	−3.10	X
	216	12.000	2.110	9.890	4.35	R
	217	8.000	1.458	6.542	2.88	R
	218	10.000	1.458	8.542	3.76	R
	242	8.000	2.110	5.890	2.59	R
	295	9.000	1.458	7.542	3.32	R
	308	7.000	2.110	4.890	2.15	R
	379	12.000	2.110	9.890	4.35	R
	380	9.000	1.458	7.542	3.32	R
	406	9.000	2.110	6.890	3.03	R
	408	7.000	2.110	4.890	2.15	R
	411	10.000	2.110	7.890	3.47	R

Prediction for AccidentFreq	Variable	Setting	Fit	SE fit	95% CI	95% PI	
	Foursmokeormore SmokeDurDrive	NO NO	1.4578	0.144363	(1.17407, 1.74159)	(−3.02885, 5.94452)	XX
	Foursmokeormore SmokeDurDrive	YES NO	6.3480	1.62650	(3.15092, 9.54509)	(0.84608, 11.8499)	XX
	Foursmokeormore SmokeDurDrive	YES YES	7	1.61080	(3.83379, 10.1662)	(1.51596, 12.4840)	XX
	Foursmokeormore SmokeDurDrive	NO YES	2.1098	0.173194	(1.76939, 2.45026)	(−2.38080, 6.60045)	XX

**R = large std residuals, X = unusual std residuals, and XX denotes an extremely unusual point relative to predictor levels used to fit the model.



(a)



(b)

Figure 3: (a) Pareto chart of the standardized effects and (b) versus fits plot of the new regression model.

of the plot as before, and there are many extreme outliers. So, the model is not a good fit too.

4.2.11. Major Findings from the Second LR Analysis. Even in the second LR analysis, many diagnostic tools for assessing the model's fit indicate that the model is still not a good fit. It may be necessary to consider alternative models. There are many different regression models, such as linear regression, logistic regression, and polynomial regression. It may be worth trying different models to see if they better fit the data. In this study, the binary logistic regression model has been used.

4.3. First Binary Logistic Regression. In this section, BLR analysis has been done using these variables: AccidentEver versus Age, DrivingHour, AvgCigPerDay, CigUseYr, NosSmokePerDrivHr, TypeofSmoke, NosSmokeDurDrive, SmokeorTobaccoStatus, SmokeDaily, MostConsmPlace, TypofSmklessTobc, SmkLessTobDailyStat, and SmokeDurDrive. Categorical predictor coding (1, 0) was applied in this instance. Logit was the link function. The response was AccidentEver, indicating whether the accident happened or not. Among the 424 people surveyed, 274 reported having experienced at least one accident, whereas 150 reported never having experienced one. Currently, the goal of this BLR is to determine the relationship between the accident and independent variables or predictors such as Age, DrivingHour, Average Cigar Per Day, Cigar Use Year, No Smoke Per Drive Hour, Type of Smoke, No Smoke During Driving, Smoke or Tobacco Status, Smoke Daily, Most Conscious Place, Type of Smokeless Tobacco, SmkLess Tob Daily Stat, and SmokeDurDrive. Here AccidentEver is the dependent variable or response whose possible values are yes or no. The regression equations (5) and (6) are given below:

$$P(YES) = \frac{\exp\left(Y'\right)}{1 + \exp\left(Y'\right)}, \tag{5}$$

$$Y' = \begin{cases} 12 + 0.0182x_1 + 0.0068x_2 + 0.0065x_3 + 0.0350x_4 \\ +0.0633x_5 - 1.256x_6 + 0.0x_7\_\text{NO} - 1.612x_7\_\text{YES} \\ +0.174x_8 + 0.0x_9\_\text{Daily} - 0.47x_9\_\text{Not at all} + 0.0x_{10}\_\text{Gul} - 12x_{10}\_\text{Jarda} - 0.0x_{10}\_\text{Khoini} - 13x_{10}\_\text{No consumption} \\ -12x_{10}\_\text{White} - \text{pata} + 0.0x_{11}\_\text{Daily} + 1.105x_{11}\_\text{No consumpion} - 0.432x_{11}\_\text{Sometimes} + 0.0\,x_{12}\_\text{No} + 1.105\,x_{12}\_\text{YES} \end{cases} \tag{6}$$

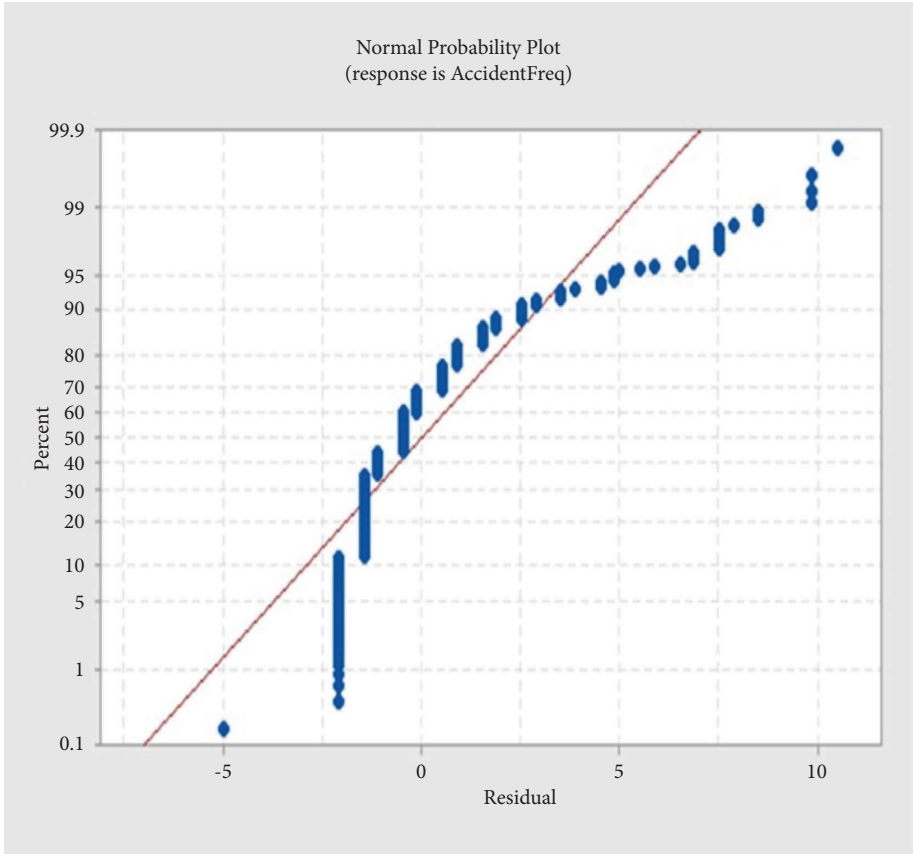
where x_1 = Age, x_2 = DrivingHour, x_3 = AvgCigPerDay, x_4 = CigUseYr, x_5 = NosSmokeDurDrive, x_6 = NosSmokePerDrivHr, x_7 = SmokeorTobaccoStatus, x_8 = TypeofSmoke, x_9 = SmokeDaily, x_{10} = TypofSmklessTobc, x_{11} = SmkLessTobDailyStat, and x_{12} = SmokeDurDrive.

4.3.1. Coefficients and VIF. Table 6 depicts the detailed model summary, coefficients, and fitness test of the developed model. Here, negative coefficients indicate that increasing NosSmokePerDrivHr and SmokeorTobaccoStatus (Yes) would reduce the likelihood of an accident. Conversely, other factors have positive coefficients, meaning increasing them would make accidents more likely. In addition, certain variables offer extremely high VIF.

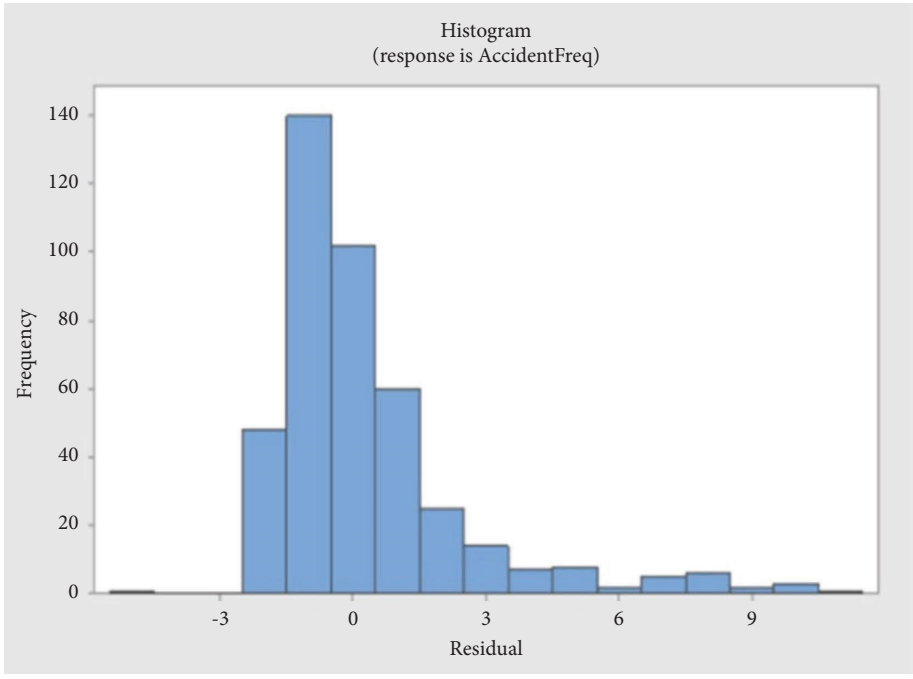
4.3.2. Goodness-of-Fit Analyses. In a binary logistic regression model, deviance measures the difference between the model and the data. Deviance R-squared (Deviance R-sq) measures the proportion of the variance in the response that the model explains. Deviance R-sq (adj) is the adjusted version of Deviance R-sq, which adjusts for the number of predictors in the model. In general, a higher Deviance R-sq value indicates that the model explains a larger proportion of the variance in the response. In contrast, a lower Deviance R-sq value indicates that the model explains a smaller proportion of the variance in the response.

Similarly, a higher Deviance R-sq (adj) value indicates that the model explains a larger proportion of the variance in the response. In comparison, a lower Deviance R-sq (adj) value indicates that the model explains a smaller proportion of the variance in the response. In binary logistic regression, Deviance R-sq and Deviance R-sq (adj) can be used to compare the fit of different models and assess the predictors' importance in the model. For example, a model with a high Deviance R-sq or Deviance R-sq (adj) value may be considered a better fit to the data than a model with a low Deviance R-sq or Deviance R-sq (adj) value. However, it is important to remember that these measures have limitations and may not always accurately reflect the model's fit.

According to Table 6, the Deviance R-sq is 7.98%, and the Deviance R-sq (adj) is 5.08%. These values indicate that the model explains a relatively small proportion of the variance in the response. However, it is difficult to determine whether a model fits well based solely on the Deviance R-squared (Deviance R-sq) and Deviance R-sq (adj) values. These measures have limitations and may not accurately

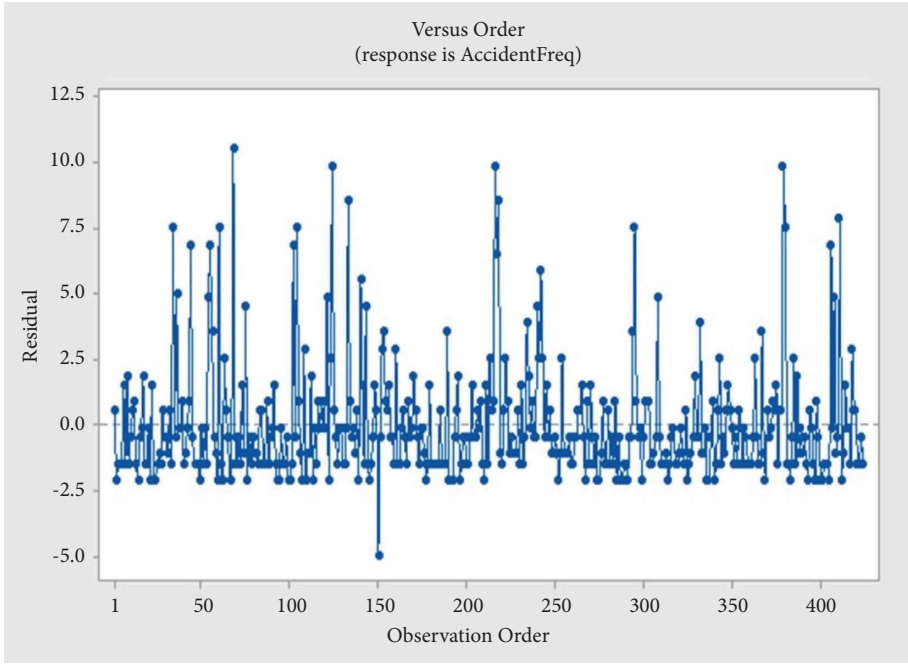


(a)



(b)

Figure 4: Continued.



(c)

Figure 4: (a) Normal probability plot, (b) histogram of frequency versus residual of the new regression model, and (c) versus order plot.

Table 5: Analysis of variance of AccidentFreq versus Foursmokeormore and SmokeDurDrive.

Term	DF	Adj SS	Adj MS	F value	p value
Regression	2	98.78	49.390	9.52	≤0.001
Foursmokeormore	1	47.28	47.281	9.11	0.003
SmokeDurDrive	1	43.39	43.393	8.36	0.004
Error	421	2184.72	5.189		
Total	423	2283.50			

reflect the model's fit. AIC, AICc, and BIC are measures of the fit of a statistical model that is used to compare the performance of different models. These measures are commonly used in binary logistic regression, which is a statistical model used to predict the probability of a binary response (i.e., a response that can take on only two possible values, such as "yes" or "no"). AIC (Akaike Information Criterion) is a measure of the fit of a model that balances the goodness of fit with the model's simplicity. A lower AIC value indicates a better fit. AICc (Corrected Akaike Information Criterion) is a modified version of AIC that is more appropriate for small sample sizes. Like AIC, a lower AICc value indicates a better fit. BIC (Bayesian Information Criterion) is another measure of the fit of a model that balances the goodness of fit with the complexity of the model. BIC places a higher penalty on models with more parameters, making it more conservative than AIC. Therefore, a lower BIC value indicates a better fit. In general, AIC, AICc, and BIC are used to compare different models' fit and select the best model for a given dataset. In binary logistic regression, these measures can be used to compare the fit of different models with different sets of predictors or model configurations.

According to Table 6, the AIC, AICc, and BIC values are 541.01, 542.52, and 609.86, respectively. So, it may indicate that the model's fit is relatively poor, as a lower AIC, AICc, or BIC value indicates a better fit. So, the values of 541.01, 542.52, and 609.86 suggest that the model does not fit the data well. However, it is important to remember that these measures may not always accurately reflect the model's fit, and other measures may also be used to assess the model's fit. For example, in binary logistic regression, deviance, Pearson's chi-squared statistic, and the Hosmer–Lemeshow test are all measures of goodness of fit that can be used to assess the fit of the model to the data. Deviance is a measure of the difference between the model and the data. A lower deviance value generally indicates a better fit. Pearson's chi-squared statistic measures the difference between the observed and expected frequencies in the data. Finally, the Hosmer–Lemeshow test is a goodness-of-fit test that compares the observed and expected frequencies of the response variable in different groups, or deciles, of the predicted probabilities. Unfortunately, the dataset's binary response/frequency format makes deviation or the Pearson test ineffective. The Hosmer–Lemeshow test, however, will be relevant. Here, a low p value and chi-square value of under ten would indicate that the model is insufficient. However, its p value of 0.250 (high) and chi-square value of 10.22 show otherwise. Therefore, the test indicates that the goodness of fit is acceptable.

4.3.3. Fits and Diagnostics for Unusual Observations. In the fits and diagnostics for unusual observations in Table 6, all the values of Std Resid are large or unusual. In that case,

Table 6: Coefficients, model summary, goodness-of-fit tests of the regression, and fits and diagnostics for unusual observations.

	Coef	SE coef	VIF
Coefficients			
Term			
Constant	12	125	
Age	0.0182	0.0141	1.41
DrivingHour	0.0068	0.0372	1.29
AvgCigPerDay	0.0065	0.0200	2.13
CigUseYr	0.0350	0.0208	2.63
NosSmokeDurDrive	0.0633	0.0743	15.38
NosSmokePerDrivHr	−1.256	0.848	16.08
SmokeorTobaccoStatus			
YES	−1.612	0.741	9.10
TypeofSmoke	0.174	0.699	10.50
SmokeDaily			
Not at all	−0.47	1.10	22.13
TypofSmklessTobc			
Jarda	−12	125	111918.84
Khoini	−0	261	1.30
No consumption	−13	125	141242.34
White-pata	−12	125	35901.36
SmkLessTobDailyStat			
No consumption	1.105	0.946	8.34
Sometimes	−0.432	0.784	1.48
SmokeDurDrive			
YES	1.105	0.337	2.31

Model summary	Deviance R-Sq	R-sq (adj)	Deviance AIC	AICc	BIC
	7.98%	5.08%	541.01	542.52	609.86

Goodness-of-fit tests	Test	DF	Chi-square	P-value
	Deviance	407	507.01	0.001
	Pearson	407	425.55	0.253
	Hosmer–Lemeshow	8	10.22	0.250

Fits and diagnostics for unusual observations	Obs	Observed probability	Fit	Residuals	Std residuals	Type of residuals
	1	1.000	1.000	0.003	≤0.001	X
	20	1.000	0.726	0.801	0.87	X
	36	1.000	0.869	0.530	0.70	X
	37	1.000	0.698	0.848	0.95	X
	49	≤0.001	0.869	−2.018	−2.04	R
	72	≤0.001	0.485	−1.152	−1.27	X
	74	1.000	0.645	0.937	1.08	X
	94	≤0.001	0.399	−1.010	−1.16	X
	97	≤0.001	0.499	−1.175	−1.41	X
	100	≤0.001	0.481	−1.145	−1.30	X
	107	≤0.001	0.493	−1.166	−1.25	X
	109	≤0.001	0.530	−1.230	−1.40	X
	111	1.000	0.377	1.396	1.49	X
	121	1.000	0.713	0.823	0.99	X
	141	1.000	1.000	0.003	≤0.001	X
	150	1.000	1.000	0.002	≤0.001	X
	151	1.000	0.577	1.048	1.18	X
	161	≤0.001	0.692	−1.534	−1.65	X
	169	1.000	0.859	0.552	0.65	X
	191	≤0.001	0.766	−1.705	−1.82	X
	236	1.000	0.711	0.826	0.89	X
	246	1.000	1.000	0.003	≤0.001	X
	267	≤0.001	0.295	−0.836	−0.90	X
	274	≤0.001	0.921	−2.251	−2.27	R
	277	1.000	0.697	0.850	0.92	X
	289	≤0.001	0.234	−0.730	−0.81	X
	294	1.000	0.616	0.985	1.08	X
	301	1.000	0.402	1.351	1.54	X
	313	1.000	1.000	0.004	≤0.001	X
	314	≤0.001	0.908	−2.182	−2.21	R
	320	1.000	0.739	0.777	0.89	X
	336	≤0.001	0.863	−1.995	−2.02	R
	341	≤0.001	0.729	−1.616	−1.76	X
	370	1.000	0.481	1.210	1.40	X
	388	1.000	0.664	0.905	0.99	X

R = large std residuals; X = unusual std residuals.

likely, the model does not fit the data well. This could be because of an inappropriate model form, wrong variable selection, or other problems.

4.3.4. Odds Ratio Analysis. In a binary logistic regression analysis, the odds ratio (OR) measures the association between a predictor variable and the response variable. It is the ratio of the odds of the response occurring in one group (e.g., those with a certain value of the predictor variable) to the odds of the response occurring in another group (e.g., those without that value of the predictor). For example, suppose the odds ratio for a predictor variable is 2. In that case, it means that the odds of the response occurring in one group (e.g., those with a certain value of the predictor) are twice the odds of the response occurring in another group (e.g., those without that value of the predictor). Odds ratios are often used in binary logistic regression to interpret the strength of the association between a predictor variable and the response. For example, 1 for the odds ratio indicates no association between the predictor and the response.

In contrast, values greater than 1 indicate a positive association, and values less than 1 indicate a negative association. The values of continuous predictors are continuous, but categorical predictors fall into categories like yes and no. For example, continuous predictors with odds ratios greater than one show that the event is more likely to happen as the predictor rises. Conversely, suppose the odds ratio is less than 1. In that case, as the predictor gets stronger, the likelihood of the event happening decreases. So, Table 6 shows that the following variables have positive odds ratios: DrivingHour, AvgCigPerDay, CigUseYear, NosSmokeDurDrive, and TypeofSmoke, which suggests that when these variables rise, the likelihood that a motorist has had a traffic accident increases.

The odds ratio for categorical predictors compares the likelihood that an event will occur at two levels of the predictor, denoted by level A and level B. The factor's reference level is level B. A level A event is more likely if the odds ratio is more than 1, whereas a level A event is less likely if the odds ratio is less than 1. For example, in Table 7, a driver's likelihood of being involved in a traffic collision decreases with a reduction in SmokeorTobaccoStatus (Yes), SmokeDaily (not at all), and SmkLessTobDailyStat (sometimes) and a rise in SmkLessTobDailyStat (no consumption) with a rise in SmokeDurDrive (Yes).

4.3.5. Wald Test Analysis. The Wald test is a statistical test that can assess the significance of individual predictor variables in a statistical model. The Wald test is based on the Wald statistic, calculated as the estimated coefficient for a predictor divided by its standard error. The Wald statistic follows a chi-squared distribution, and the p value for the Wald test is calculated based on this distribution. The Wald test is often used in regression analysis, including binary logistic regression, to assess the significance of individual predictors in the model. For example, the Wald test is often used in binary logistic regression to assess the significance of individual predictors in the model. It is similar to the analysis of variance (ANOVA) in other types of regression analysis. For example, suppose the p value for a predictor is less than the predetermined significance level (e.g., 0.05). In that case, it is considered statistically significant, and the predictor is considered an important contributor to the model. On the other hand, suppose the p value is greater than the significance level. In that case, the predictor is not considered statistically significant and may be removed from the model.

Table 8 displays the Wald test results along with the dataset. According to the weighted difference between the unrestricted estimate and its hypothesized value under the null hypothesis, where the weight reflects the estimate's precision, the statistical Wald test evaluates limitations on statistical parameters. Under the null hypothesis, it exhibits an asymptotic 2-distribution, which can be used to assess statistical significance. Table 8 shows that SmokeorTobaccoStatus and SmokeDurDrive exhibit significance with 95% CI. SmokeDurDrive demonstrates extreme relevance.

4.3.6. Normal Probability Plot Analysis. A normal probability plot of a binary logistic regression model is a graph that shows the observed deviance residuals plotted against the expected normal deviance residuals. The plot should be approximately linear if the residuals are normally distributed. If the plot is not linear, it may indicate that the residuals are not normally distributed, and the model may not be a good fit for the data. It is true that, in a binary logistic regression model, the response variable is binary (i.e., it takes on only two values, such as "yes" or "no"), and the residuals are not normally distributed. However, the expected deviance residuals (used in the normal probability plot) are based on a normal distribution, so the plot is still useful for assessing the model's goodness of fit. Suppose the observed deviance residuals are approximately linear on the plot. In that case, it suggests that the model fits the data well. If the plot is not linear, it may indicate that the model is not a good fit, and further investigation is needed. Figure 5(a) demonstrates that the dataset is an S-curve rather than following the normal probability plot, as it is not linear. In that case, this model does not fit well, and using a different model or transforming the data might need to be considered. However, the inverted S-curve implies that the distribution is short-tailed.

4.3.7. Frequency versus Deviance Residual Plot. In a binary logistic regression analysis, the frequency versus residual plot is a graphical tool that can be used to assess the model's fit. The plot shows the residuals of the model on the y-axis and the frequency of the residuals on the x-axis. For example, suppose the model is a yes/no-based binary logistic regression model. In that case, the plot can be separated into two groups based on the response: one for the "yes" and one for the "no" responses.

The goodness of fit of a binary logistic regression model can be determined by looking at the frequency versus deviance residual plot. If the two groups (yes and no

Table 7: Odds ratios for continuous and categorical predictors.

	Term		Odds ratio*	95% CI
Continuous predictors	Age		1.0184	(0.9905, 1.0470)
	DrivingHour		1.0069	(0.9361, 1.0829)
	AvgCigPerDay		1.0065	(0.9679, 1.0467)
	CigUseYr		1.0357	(0.9943, 1.0787)
	NosSmokeDurDrive		1.0654	(0.9209, 1.2324)
	NosSmokePerDrivHr		0.2847	(0.0541, 1.4994)
	TypeofSmoke		1.1905	(0.3024, 4.6860)
Categorical predictors	Level A	Level B	Odds ratio*	95% CI
	SmokeorTobaccoStatus			
	YES	NO	0.1995	(0.0467, 0.8529)
	SmokeDaily			
	Not at all	Daily	0.6227	(0.0715, 5.4250)
	TypofSmklessTobc			
	Jarda	Gul	≤0.001	(≤0.001, $4.13843E+101$)
	Khoini	Gul	0.9601	(≤0.001, $2.96379E+222$)
	No consumption	Gul	≤0.001	(≤0.001, $1.71199E+101$)
	White-pata	Gul	≤0.001	(≤0.001, $4.97055E+101$)
	Khoini	Jarda	0.9887	(≤0.001, $2.11697E+200$)
	No consumption	Jarda	0.4137	(0.0536, 3.1936)
	White-pata	Jarda	1.1992	(0.2368, 6.0732)
	No consumption	Khoini	≤0.001	(≤0.001, $5.03528E+189$)
	White-pata	Khoini	≤0.001	(≤0.001, $1.45674E+190$)
	White-pata	No consumption	2.8989	(0.3655, 22.9941)
	Level A	Level B	Odds ratio*	95% CI
	SmkLessTobDailyStat			
	No consumption	Daily	3.0185	(0.4725, 19.2821)
	Sometimes	Daily	0.6491	(0.1398, 3.0149)
	Sometimes	No consumption	0.2150	(0.0313, 1.4759)
	SmokeDurDrive			
	YES	NO	3.0184	(1.5590, 5.8441)

*Odds ratio for level A relative to level B.

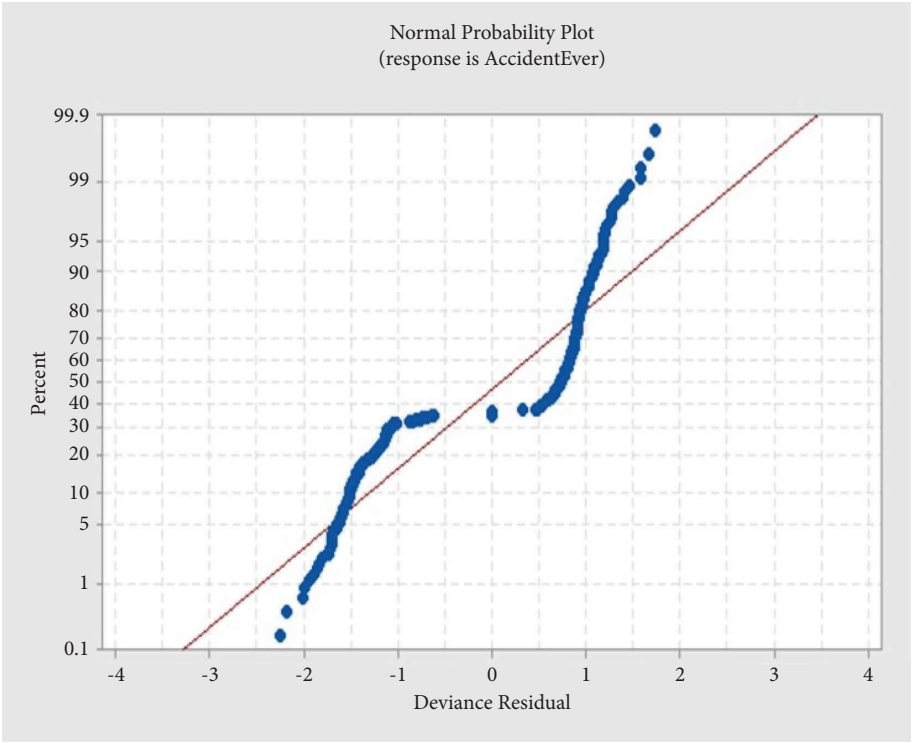
Table 8: Wald test.

Term	DF	Chi-square	p value
Regression	16	31.76	0.011
Age	1	1.66	0.198
DrivingHour	1	0.03	0.854
AvgCigPerDay	1	0.11	0.744
CigUseYr	1	2.84	0.092
NosSmokeDurDrive	1	0.73	0.394
NosSmokePerDrivHr	1	2.20	0.138
SmokeorTobaccoStatus	1	4.73	0.030
TypeofSmoke	1	0.06	0.803
SmokeDaily	1	0.18	0.668
TypofSmklessTobc	4	1.07	0.899
SmkLessTobDailyStat	2	2.47	0.291
SmokeDurDrive	1	10.74	0.001

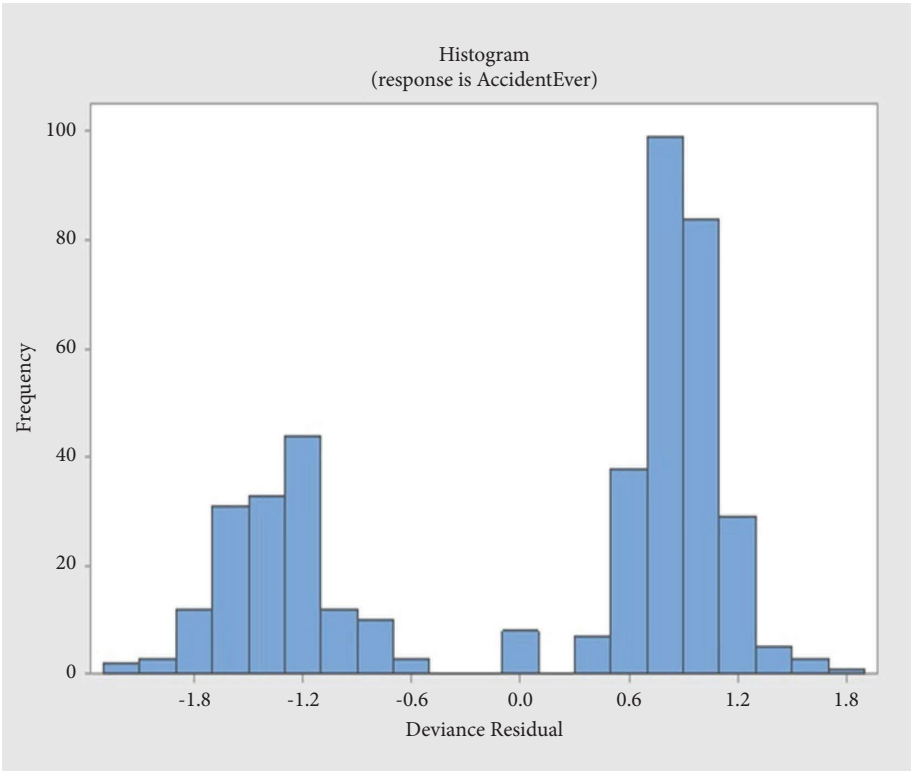
responses) are separated, the model is a good fit. Additionally, if the residuals are randomly distributed around the zero line, it indicates that the model is a good fit. It means that, for a good-fitting model, the points on the frequency versus residual plot should have a random pattern, with the points evenly distributed around the x-axis. On the other hand, if the points on the plot are not evenly distributed for one or both of the "yes" and "no" groups, it may indicate that the model is not a good fit for the data. For example, suppose the points are concentrated on one side of the x-axis for one of the groups. In that case, it may indicate that the model is over or under-predicting the response for that group. So, the deviance residuals should be randomly distributed around zero, with similar frequencies for both positive and negative values. This indicates that the model is a good fit and that the residuals are not systematically biased in one direction.

Figure 5(b) shows the histogram of the frequency versus deviance residual. First, the two groups are not separated and are spread out here. Moreover, second, the residuals are not randomly distributed around the zero line, with the points evenly distributed around the x-axis, as more deviance residuals are concentrated in the yes group. So, the model has poor goodness of fit.

4.3.8. Versus Order Plot. In a binary logistic regression analysis, the versus order plot is a graphical tool that can assess the model's fit and identify patterns in the residuals. Here the versus order plot is the deviance residuals versus observation order plot. The plot shows the deviance residuals of the model on the y-axis and the observation order on the x-axis. If the model is a good fit, the points on the versus order plot should be randomly distributed around the y-axis, with no discernible patterns or trends. On the other hand, if

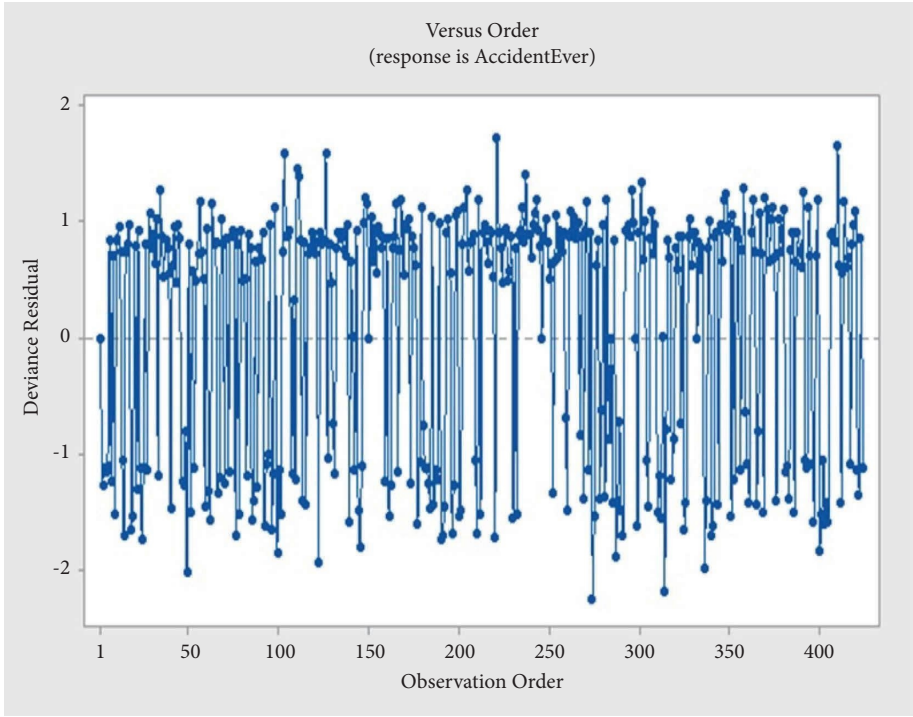


(a)

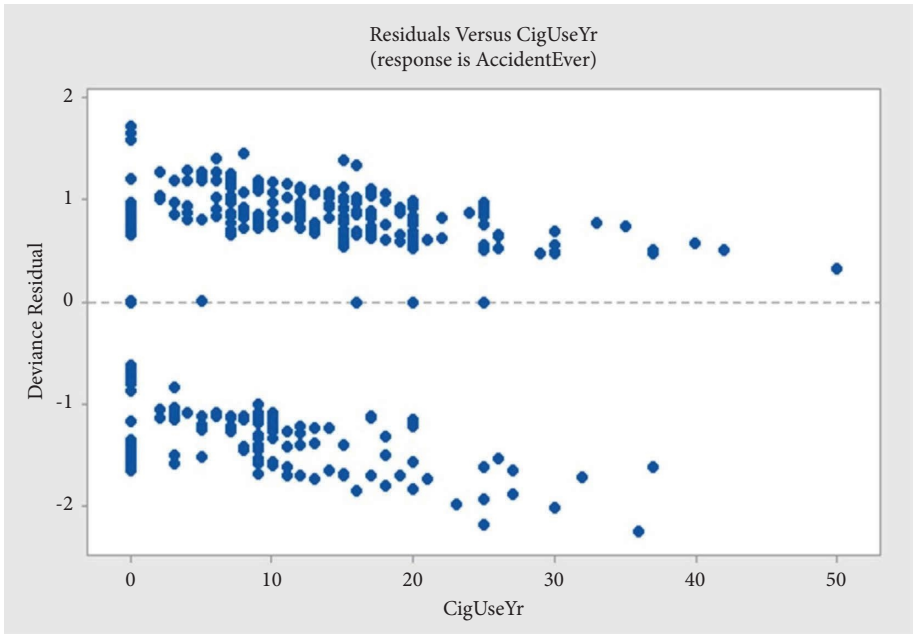


(b)

Figure 5: Continued.



(c)



(d)

Figure 5: (a) Normal probability plot, (b) histogram of frequency versus deviance residual of the binary logistic regression model, (c) versus order plot, and (d) deviance residuals versus CigUseYr.

there are patterns or trends in the residuals, it may indicate that the model is not a good fit for the data.

Suppose the model is a yes/no-based binary logistic regression model. In that case, the plot can be separated into two groups based on the response: one for the "yes" and one for the "no" responses. For a good-fitting model, the points on the versus order plot should be randomly distributed around the y-axis for both the "yes" and "no" groups, with no discernible patterns or trends for either group. If the points on the plot are not randomly distributed for one or both of the "yes" and "no" groups, it may indicate that the model is not a good fit for the data. For example, suppose the points are concentrated on one side of the y-axis for one of the groups. In that case, it may indicate that the model is over or under-predicting the response for that group. The constant variance of the residuals is also important in a binary

logistic regression model. Suppose the variance of the residuals is not constant. In that case, it may indicate that the model is not a good fit for the data or that other factors affect the residuals' variance.

Figure 5(c) shows the versus order plot of this model. The deviance residuals are randomly distributed in both groups, and no discernible pattern or trends are found in either group. However, the points are mostly concentrated in the "yes" group more than the "no" group. Besides, the variance in both groups is not constant. So, the model is not a very good fit.

4.3.9. Deviance Residual Analysis. The deviance residual analysis is done with the deviance residual versus fitted value plot. In a binary logistic model, the regression line against increasing or decreasing values of an independent variable jumps from the negative response to the positive response. There will be two regression lines, one on the positive side and one on the negative side of zero, on the deviance residual versus fitted value plot. The line on the negative side should increase in value on the x-axis and go from zero to a higher negative value on the y-axis. On the other hand, the line on the positive side should follow the increasing value on the x-axis and move from a higher positive value to zero on the y-axis.

In a binary logistic regression analysis, the deviance residual versus fitted value plot is a graphical tool that can assess the model's fit. The plot shows the deviance residuals of the model on the y-axis and the fitted values (predicted probabilities of the response) on the x-axis. If the model is a good fit, the points on the plot should have a random pattern, with the points evenly distributed around the y-axis. For example, suppose the model is a yes/no-based binary logistic regression model. In that case, the plot can be separated into two groups based on the response: one for the "yes" and one for the "no" responses. For a good-fitting model, the deviance residual versus fitted value plot points should have a random pattern, with the points evenly distributed around the y-axis for both the "yes" and "no" groups. On the other hand, if the points on the plot are not evenly distributed for one or both of the "yes" and "no" groups, it may indicate that the model is not a good fit for the data. For example, suppose the points are concentrated on one side of the y-axis for one of the groups. In that case, it may indicate that the model is over or under-predicting the response for that group. In such cases, it may be necessary to try alternative modeling approaches or transform the data to achieve a better fit.

Figure 5(d) shows deviance residuals versus the CigUseYr plot for this model. The points have a random pattern and are evenly distributed around the y-axis for both the "yes" and "no" groups. So, this model is a good fit for the data.

4.3.10. Major Findings from the First BLR Analysis. Only two predictors in this regression model show significance with a 95% CI in the Wald test since their p values are less than 0.05. Other predictors have significantly higher p values. However, CigUseYr has a p value of 0.092, close to the 0.05 threshold. Besides, some tests show that the model is not a good fit for the data. So, to get better results, a new BLR model should be developed using these three predictors.

4.4. Second Binary Logistic Regression. Here the used variables are AccidentEver versus CigUseYr, SmokeorTobaccoStatus, and SmokeDurDrive. The same method and response information have been used in this new model (Table 9).

4.4.1. Odds Ratio Analysis. Table 10 depicts the odds ratios of the second BLR model. The continuous predictor, CigUseYr, has an odds ratio of 1.0545, meaning that a driver's risk of an accident rises as it rises. SmokeorTobaccoStatus (Yes) and SmokeDurDrive (Yes) have odds ratios that are lower and higher than one, respectively. This means that as SmokeorTobaccoStatus (Yes) and SmokeDurDrive (Yes) go down and up, respectively, the likelihood that a driver has experience increases.

4.4.2. Wald Test. Again, the Wald test (Table 11) has been performed for the new regression model, which shows that the p values of all three predictors are less than 0.01. So, all three predictors are highly significant, with a 99% CI.

4.4.3. Normal Probability Plot Analysis. Figure 6(a) demonstrates that the dataset displays an S-curve rather than following the normal probability plot, like the first BLR model. In that case, this model does not fit well, and using a different model or transforming the data might need to be considered. However, the inverted S-curve implies that the distribution is short-tailed.

4.4.4. Frequency versus Deviance Residual Plot. Figure 6(b) shows this new BLR model's frequency versus deviance residual histogram. Nevertheless, this time, the frequency distributions for "yes" and "no" are clearer and less spread out. The points are not evenly distributed around the x-axis, as more deviance residuals are concentrated in the "yes" group. So, it can be concluded that the goodness of fit improved in the second BLR model from the frequency versus deviance residual plot perspective.

4.4.5. Versus Order Plot. Figure 6(c) shows the versus order plot of this model. The deviance residuals are randomly distributed in both groups, and no discernible pattern or trends are found in either group like in the first BLR model. On the other hand, the points are mostly concentrated in the "yes" group more than the "no" group. However, the positive thing is that the variance in both groups is more constant now. So, the model became quite a good fit.

4.4.6. Deviance Residual Analysis. Figure 6(d) shows deviance residuals versus the CigUseYr plot for this model. The points are in a less random pattern than the first BLR model,

Table 9: Regression equation, coefficients, model summary, goodness-of-fit tests of the regression, and fits and diagnostics for unusual observations.

		$P(YES) = (\exp(Y'')/1 + \exp(Y''))$				
		SmokeorTobaccoStatus	SmokeDurDrive			
Regression equation		NO	NO	$Y'' = 0.7376 + 0.05308CigUseYr$		
		NO	YES	$Y'' = 1.422 + 0.05308CigUseYr$		
		YES	NO	$Y'' = -0.4451 + 0.05308CigUseYr$		
		YES	YES	$Y'' = 0.2389 + 0.05308CigUseYr$		
Coefficients		Term		Coef	SE coef	VIF
		Constant		0.738	0.212	
		CigUseYr		0.0531	0.0161	1.63
			SmokeorTobaccoStatus			
		YES		−1.183	0.318	1.70
			SmokeDurDrive			
		YES		0.684	0.245	1.28
Model summary		Deviance R-sq		R-sq (adj)	Deviance AIC	AICc	BIC
		4.69%		4.15%	533.12	533.22	549.32
Goodness-of-fit tests of the regression		Test		DF	Chi-square	p value	
		Deviance		420	525.12	≤0.001	
		Pearson		420	429.22	0.367	
		Hosmer–Lemeshow		7	13.31	0.065	
Fits and diagnostics for unusual observations	Obs	Observed probability	Fit	Residuals	Std residuals	Type of residuals	
	49	≤0.001	0.8619	−1.9900	−2.00	R	
	205	1.0000	0.8427	0.5851	0.60		X
	250	1.0000	0.8562	0.5572	0.57		X
	274	≤0.001	0.8957	−2.1261	−2.14	R	
	341	≤0.001	0.8204	−1.8531	−1.88		X

R = large std residuals; X = unusual std residuals.

Table 10: Odds ratios of the second binary logistic regression model.

	Term	Odds ratio*	95% CI	
Continuous predictors	CigUseYr	1.0545	(1.0218, 1.0882)	
Categorical predictors	Level A	Level B	Odds ratio*	95% CI
			SmokeorTobaccoStatus	
	YES	NO	0.3065	(0.1643, 0.5715)
			SmokeDurDrive	
	YES	NO	1.9818	(1.2257, 3.2045)

*Odds ratio for level A relative to level B.

Table 11: Wald test of the new regression model.

Term	DF	Chi-square	p value
Regression	3	23.62	≤0.001
CigUseYr	1	10.93	0.001
SmokeorTobaccoStatus	1	13.83	≤0.001
SmokeDurDrive	1	7.78	0.005

but they are evenly distributed around the y-axis for both the "yes" and "no" groups. So, this model is not a good fit for the data like the first BLR model.

4.4.7. Major Findings from the Second BLR Model. The predictor *SmokeorTobaccoStatus* has a negative coefficient, which means not consuming tobacco would increase the chance of RTA, which seems counterintuitive. However, the goodness of fit in the second BLR model is proved to be well by the Hosmer–Lemeshow test. Furthermore, the Wald test shows that these three predictors are significant. Besides, the frequency versus deviance residual plot and versus order plot show that the goodness of fit improved much from the first BLR model. So, it can be concluded that the model fits with the data.

5. Discussion

According to Talukder et al. [10], age, marital status, income, and education are important socioeconomic and demographic factors that affect cigarette usage. However, we had a complete picture thanks to investigating driving behavior and accidents in Bangladesh. The study discovered a statistically significant connection between SLT use, smoking, and car accidents.

However, the regression results from the LR analysis show that Foursmokeormore and SmokeDurDrive have significance, and the ANOVA results show that NosSmokePerDrivHr shows significance as well. So, it can be concluded that the accident frequency of a driver increases if the frequency of his smoking a certain amount of smoking tobacco per driving hour decreases, he smokes during driving, and he smokes six or more cigarettes/bidis sticks while driving.

Moreover, the binary logistic analysis results show that CigUseYr, SmokeorTobaccoStatus (No), and SmokeDurDrive (Yes) have more than one odds ratio. Moreover, the Wald test shows that these three predictors have p values less than 0.01, indicating their high significance with a 99% confidence interval. So, it can be concluded that the chance of the driver having the status of involved in an accident ever increases if the number of his cigarette or bidi consumption days per year is increased and if he smokes during driving. On the other hand, SmokeorTobaccoStatus (No) or negative smoking status has high significance, but it is counterintuitive. So, further research is necessary in this case.

A study done in China had a total of 8990 ride-hailing drivers participate in the poll, and 5024 of them, or 55.9%, were current smokers. 32.2% of smokers smoked in their vehicles. The outcomes of the logistic regression analysis were as follows: male drivers (OR = 0.519, 95% CI [0.416, 0.647]), central and eastern regions (OR = 1.172, 95% CI [1.049, 1.309]), working both day and night (OR = 1.287, 95% CI [1.164, 1.424]), nonfixed time (OR = 0.847, 95% CI [0.718, 0.999]), central and eastern regions (OR = 1.330, 95% CI [1.194, 1.480], ages of 35–54 years (OR = 0.585, 95% CI [0.408, 0.829]), current drinker (OR = 1.663, 95% CI [1.526, 1.813]), irregular eating habits (OR = 1.370, 95% CI [1.233, 1.523]), the number of days in a week of engaging in at least 10 min of moderate or vigorous exercise ≥ 3 (OR = 0.752, 95% CI [0.646, 0.875]), taking the initiative to acquire health knowledge occasionally (OR = 0.882, 95% CI [0.783, 0.992]) or frequently (OR = 0.675, 95% CI [0.591, 0.770]), and underweight (OR = 1.249, 95% CI [1.001, 1.559]) and overweight (OR = 0.846, 95% CI [0.775, 0.924]) have an association with the prevalence of current smoking among online ride-hailing drivers. It was found that the smoking rate of ride-hailing drivers was high. Sociodemographic and work-related characteristics and health-related factors affected their smoking behavior [30].

Another study in Bangladesh shows that male and female passive smoker prevalence was 74.3% and 25.8%, respectively. Among those who smoke only secondhand, 22.2% said they had a policy allowing smoking at home, while 29.8% said they had none. As an alternative, 26.0% of passive smokers said it was permitted, while 27.5% said their workplace had no smoking policy. According to a logit regression analysis, the probability of allowing smoking at home was 4.85 times greater for the tobacco smoker group than for the nonsmoker respondents (OR = 4.85, 95% CI = 4.13, 5.71), 1.18 times more likely to be permitted at home in rural than urban areas (OR = 1.18, 95% CI = 1.06, 1.32), and 0.35 times less likely to be permitted at home if the respondent has completed college or university or has a higher education than none (OR = 0.35, 95% CI = 0.24, 0.52). On the other hand, smoking was less likely to be permitted for respondents who had completed college or

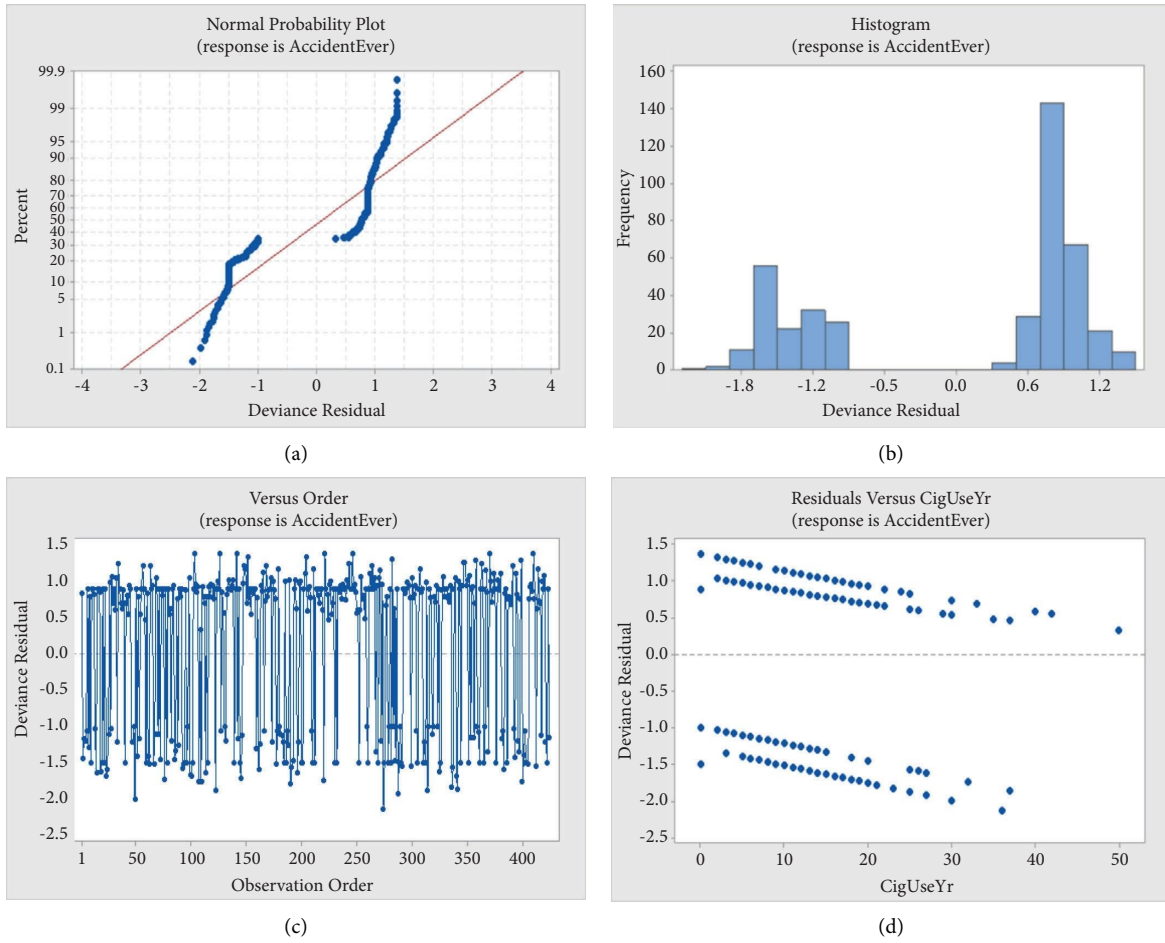


Figure 6: (a) Normal probability plot, (b) histogram of frequency versus deviance residual of the new binary logistic regression model, (c) versus order plot, and (d) deviance residual versus CigUseYr plot.

university and (or) had a higher education than respondents without any formal schooling (OR = 0.26, 95% CI = 0.14, 0.45) and was 1.70 times more likely to be permitted for tobacco smokers than their counterpart respondents at work (OR = 1.70, 95% CI = 1.36, 2.14) [31].

Nevertheless, another study conducted in Bangladesh on the frequency distributions of smoking shows that 24.4% of smokers experienced headache issues because of smoking and cigarette fumes, 68.8% of smokers believe smoking causes gastric problems, 48.8% of smokers feel smoking causes air pollution, 51.3% believe smoking complicates breathing for nonsmokers, 86.3% of smokers learned to smoke from friends, 48.8% of smokers smoke due to addiction, and 25.6% smoke for depression. Usually, 80.6% of smokers light up after eating. The chi-square test shows that the class of smokers was strongly related to heartbeat frequency and that starting to smoke at a certain age level was significantly related to having ailments. At the 1% significance level, smoking by category was significantly correlated with having a sickness, smoking by class was significantly correlated with reasons why people smoke, and smoking by age was strongly correlated with smoking by profession. At the 1% level, a significant odds ratio was discovered (OR = 6.363, 95% CI: 1.918–21.104, p 0.01) for the occupation group of students/labor; their outcomes for contracting illnesses, including gastric issues and fever/headache/others, were 6.363 times higher in the group of smokers who work in services or other occupations [32].

These research findings show the relationship between the ever involvement of road traffic accidents and smoking by drivers. In the regression analyses, the R-sq (adj) values indicated that further research is necessary to have better models with good fitness, which can predict better.

6. Conclusions

First, this study demonstrates that smoking is not the only significant issue. Another important problem is that many smokers are unaware of how RTAs might result from smoking. Second, this study demonstrates a strong association between the incidence of accidents and the number of times a person smokes, smokes while driving, and uses SLT daily. Finally, the result has been taken from the second BLR model, as it fits with the data more than others. According to that model, a driver is more likely to be in an accident if the number of days per year that he smokes cigarettes increases and if he smokes while driving. Additionally, it stresses the need for more research to make a more accurate forecast.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

None of these sources were involved in the design, analysis, data interpretation, writing, or decision to publish the publication. The writers are responsible for the content, which may or may not reflect the official views of the Johns Hopkins Bloomberg School of Public Health and Bangladesh Center for Communication Programs (BCCP).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Md. Anwar Uddin was responsible for conceptualization, investigation, visualization, and original draft preparation. Mohammad Mahbub Alam Talukder curated the data. Mithun Debnath was responsible for methodology and formal analysis. Saima Adiba reviewed and edited the manuscript. Sumit Roy was responsible for project administration and writing draft.

Acknowledgments

The Research Grant Program on Tobacco Control, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA, and Bangladesh Center for Communication Programs (BCCP) collaborated on this research.

References

[1] H. Chen, S. Saad, S. L. Sandow, and P. P. Bertrand, "Cigarette smoking and brain regulation of energy homeostasis," *Frontiers in Pharmacology*, vol. 3, p. 147, 2012.

[2] R. Edwards, "The problem of tobacco smoking," *BMJ*, vol. 328, no. 7433, pp. 217–219, 2004.

[3] R. Peto and A. D. Lopez, "The future worldwide health effects of current smoking patterns," *Tobacco and public health: Science Policy*, vol. 281, no. 6, pp. 281–286, 2004.

[4] M. J. Becker and T. J. Zlatoper, "Relationship between smoking and motor vehicle death rates in the U.S," *Atlantic Economic Journal*, vol. 50, no. 1–2, pp. 53–65, 2022.

[5] R. Sehsah, A.-H. El-Gilany, and A. M. Ibrahim, "Personal protective equipment (PPE) use and its relation to accidents among construction workers," *La Medicina del Lavoro*, vol. 111, no. 4, pp. 285–295, Aug. 2020.

[6] D. N. Sinha, I. Rolle, S. Rinchen, K. Palipudi, and S. Asma, "Tobacco use among youth and adults in member countries of South-East Asia region: review of findings from surveys under the Global Tobacco Surveillance System," *Indian Journal of Public Health*, vol. 55, no. 3, pp. 169–176, 2011.

[7] C. R. M. McKenzie, "Judgment and Decision Making," *Handbook of cognition*, Sage, Los Angeles, UK, 2005.

[8] N. Lin, C. Zong, M. Tomizuka, P. Song, Z. Zhang, and G. Li, "An overview on study of identification of driver behavior characteristics for automotive control," *Mathematical Problems in Engineering*, vol. 2014, pp. 1–15, Article ID 569109, 2014.

[9] K. Young, M. Regan, and M. Hammer, "Driver distraction: a review of the literature," *Distracted driving*, vol. 2007, pp. 379–405, 2007.

[10] M. M. A. Talukder, M. Mia, N. S. Chowdhury et al., "Smoking behavior and driver's involvement in road traffic accidents in Bangladesh," *International Journal of Mental Health and Addiction*, pp. 1–17, 2022.

[11] S. Goon and M. S. Bipasha, "Prevalence and Pattern of Smoking among Bus Drivers of Dhaka, Bangladesh," *Tobacco Use Insights*, vol. 7, p. S13966, 2014.

[12] World Health Organization (Who), *World Report on Road Traffic Injury Prevention*, WHO, Geneva, 2004.

[13] F. Wegman, "The future of road safety: a worldwide perspective," *IATSS Research*, vol. 40, no. 2, pp. 66–71, 2017.

[14] O. F. Hamim, M. Shamsul Hoque, R. C. McIlroy, K. L. Plant, and N. A. Stanton, "A sociotechnical approach to accident analysis in a low-income setting: using Accimaps to guide road safety recommendations in Bangladesh," *Safety Science*, vol. 124, p. 104589, 2020.

[15] M. Debnath, S. Hasanat-E-Rabbi, O. F. Hamim et al., "An investigation of urban pedestrian behaviour in Bangladesh using the Perceptual Cycle Model," *Safety Science*, vol. 138, p. 105214, 2021.

[16] K. M. Maniruzzaman and R. Mitra, "Road accidents in Bangladesh," *IATSS Research*, vol. 29, no. 2, pp. 71–73, 2005.

[17] S. Lee, Y. B. Moh, M. Tabibzadeh, and N. Meshkati, "Applying the AcciMap methodology to investigate the tragic Sewol Ferry accident in South Korea," *Applied Ergonomics*, vol. 59, pp. 517–525, 2017.

[18] J. C. Le Coze, "Reflecting on Jens Rasmussen's legacy. A strong program for a hard problem," *Safety Science*, vol. 71, pp. 123–141, 2015.

[19] M. Debnath, *Investigation of Pedestrian and Bus Drivers' Decision-Making Behavior Using Schema Theory and Perceptual Cycle Model*, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, 2021.

[20] D. Fergusson, N. Swain-Campbell, and J. Horwood, "Risky driving behaviour in young people: prevalence, personal characteristics and traffic accidents," *Australian and New Zealand Journal of Public Health*, vol. 27, no. 3, pp. 337–342, 2003.

[21] J. J. Sacks and D. Nelson, "Smoking and injuries: an overview," *Preventive Medicine*, vol. 23, no. 4, pp. 515–520, 1994.

[22] S. Saadat and M. Karbakhsh, "Association of waterpipe smoking and road traffic crashes," *BMC Public Health*, vol. 10, pp. 1–639, 2010.

[23] P. A. Koushki and M. Bustan, "Smoking, belt use, and road accidents of youth in Kuwait," *Safety Science*, vol. 44, no. 8, pp. 733–746, 2006.

[24] P. Grout, D. K. S. Cliff, M. L. Harman, and D. Machin, "Cigarette smoking, road traffic accidents andseat belt usage," *Public Health*, vol. 97, no. 2, pp. 95–101, 1983.

[25] J. M. Buñuel Granados, R. Córdoba García, M. d. Castillo Pardo Md, J. L. Álvarez Pardo, A. Monreal Hijar, and F. Pablo Cerezuela, "Smoking and nonfatal traffic accidents," *Atención Primaria*, vol. 31, no. 6, pp. 349–353, 2003.

[26] A. Tzortzi, M. Kapetanstrataki, V. Evangelopoulou, and P. Behrakis, "Driving behavior that limits concentration: a nationwide survey in Greece," *International Journal of Environmental Research and Public Health*, vol. 18, no. 8, p. 4104, 2021.

[27] D. A. Freedman, *Statistical Models: Theory and Practice*, Cambridge University Press, Berkeley, CA, USA, 2nd edition, 2009.

[28] L. L. Pederson, J. Koval, E. Vingilis et al., "The relationship between motor vehicle collisions and cigarette smoking in Ontario: analysis of CAMH survey data from 2002 to 2016," *Preventive medicine reports*, vol. 13, pp. 327–331, 2019.

[29] K. c. Choi, S. A. Kim, N. R. Kim, and M.-H. Shin, "Association between smoking and unintentional injuries among Korean adults," *Chonnam medical journal*, vol. 54, no. 3, pp. 184–189, 2018.

[30] X. Chen, X. Gu, T. Li et al., "Factors influencing smoking behaviour of online ride-hailing drivers in China: a cross-sectional analysis," *BMC Public Health*, vol. 21, no. 1, pp. 1326–1411, 2021.

[31] P. Sultana, M. T. Rahman, D. C. Roy et al., "Tobacco control policies to promote awareness and smoke-free environments in residence and workplace to reduce passive tobacco smoking in Bangladesh and its correlates," *PLoS One*, vol. 13, no. 6, pp. 01989422–e199012, 2018.

[32] M. Kamruzzaman, A. Hossain, and E. Kabir, "Smoker's characteristics, general health and their perception of smoking in the social environment: a study of smokers in Rajshahi City, Bangladesh," *Journal of Public Health*, vol. 30, no. 6, pp. 1501–1512, 2022.